CENTRO DE CIENCIAS BÁSICAS

DEPARTAMENTO DE SISTEMAS DE INFORMACIÓN

TESIS

PROPUESTA DE UN MÉTODO PARA ANALIZAR DATOS DE TWITTER CONSIDERADOS COMO BIG DATA PARA GENERAR NUEVA INFORMACIÓN

PRESENTA

César Alejandro Pedroza García

PARA OBTENER EL GRADO DE MAESTRO EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES

TUTOR

Dr. Juan Muñoz López

COMITÉ TUTORAL

Mtro. Abel Alejandro Coronado Iruegas

Dra. María Dolores Torres Soto

Aguascalientes, Ags., 30 de Mayo de 2016

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

POSGRADOS UAA

DICTAMEN DE REVISIÓN DE LA TESIS / TRABAJO PRÁCTICO

DATOS DEL ESTUDIANTE	
NOMBRE: CESAR ALEJANDRO PEDROZA GARCIA	ID (No. de Registro): 44767
PROGRAMA: MAESTRÍA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES	ÁREA: La ingeniería / Enfoque de sistemas para el mejoramiento de los procesos organizacionales usando SI/TI
TUTOR/TUTORES: DR. JUAN MUÑOZ LÓPEZ	
TESIS (X)	TRABAJO PRÁCTICO ( )
OBJETIVO: PROPONER UN MÉTODO QUE PERMITA EXPLORAR LOS DATOS REGISTRADOS EN LAS CONVERSACIONES DE LOS USUARIOS DE LA RED SOCIAL DE TWITTER, CONSIDERADOS COMO FUENTE DE BIG DATA, PARA GENERAR INFORMACIÓN ACTUALIZADA Y COHERENTE PARA POSTERIORMENTE PROBARLO EN TEMAS DE INTERÉS NACIONAL COMO LA MOVILIDAD COTIDIANA Y EL IMPACTO DE EVENTOS DE LA VIDA REAL.	

DICTAMEN	
CUMPLE CON CRÉDITOS ACADÉMICOS:	(X)
CONGRUENCIAS CON LAS LGAC DEL PROGRAMA:	(X)
CONGRUENCIA CON LOS CUERPOS ACADÉMICOS:	(X)
CUMPLE CON LAS NORMAS OPERATIVAS:	(X)
CONINCIDENCIA DEL OBJETIVO CON EL REGISTRO:	(X)

Aguascalientes, Ags. a 30 de MAYO de 2016

FIRMAS

DRA. MARIA DOLORES TORRES SOTO
CONSEJERO ACADÉMICO DEL ÁREA

DR. JOSÉ MANUEL MORA TAVAREZ
SECRETARIO TÉCNICO DEL POSGRADO

DR. EN C. BIOQ. JUAN JÁUREGUI RINCÓN
SECRETARIO DE INVESTIGACIÓN
Y POSGRADO

Código: FO-040200-23
Revisión: 00
Emisión: 21/02/11

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

CESAR ALEJANDRO PEDROZA GARCÍA
MAESTRÍA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES
P R E S E N T E.

Estimado alumno:

Por medio de este conducto me permito comunicar a Usted que habiendo recibido los votos aprobatorios de los revisores de su trabajo de tesis y/o caso práctico titulado: **"Propuesta de un método para analizar datos de Twitter considerados como Big Data para generar nueva información",** hago de su conocimiento que puede imprimir dicho documento y continuar con los trámites para la presentación de su examen de grado.

Sin otro particular me permito saludarle muy afectuosamente.

A T E N T A M E N T E
Aguascalientes, Ags., a 27 de mayo de 2016
"Se lumen proferre"
EL DECANO

M. en C. JOSE DE JESUS RUIZ GALLEGOS

c.c.p.- Archivo.

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

FORMATO DE CARTA DE VOTO APROBATORIO

M. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS
P R E S E N T E

Por medio del presente como Tutor designado del estudiante **CESAR ALEJANDRO PEDROZA GARCÍA** con ID *44767* quien realizó *la tesis* titulada: **PROPUESTA DE UN MÉTODO PARA ANALIZAR DATOS DE TWITTER CONSIDERADOS COMO BIG DATA PARA GENERAR NUEVA INFORMACIÓN**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a *imprimirla*, y así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a *26* de *Mayo* de 2016.

Dr. Juan Muñoz López
Tutor de *tesis*

c.c.p.- Interesado
c.c.p.- Secretaría de Investigación y Posgrado
c.c.p.- Jefatura del Depto. de Sistemas de Información
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES
FORMATO DE CARTA DE VOTO APROBATORIO

M. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS
P R E S E N T E

Por medio del presente como Asesor designado del estudiante **CESAR ALEJANDRO PEDROZA GARCÍA** con ID *44767* quien realizó *la tesis* titulada: **PROPUESTA DE UN MÉTODO PARA ANALIZAR DATOS DE TWITTER CONSIDERADOS COMO BIG DATA PARA GENERAR NUEVA INFORMACIÓN**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a *imprimirla*, y así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a *26* de *Mayo* de 2016.

M. Abel Alejandro Coronado Iruegas
Asesor de *tesis*

c.c.p.- Interesado
c.c.p.- Secretaría de Investigación y Posgrado
c.c.p.- Jefatura del Depto. de Sistemas de Información
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

FORMATO DE CARTA DE VOTO APROBATORIO

M. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS
P R E S E N T E

Por medio del presente como Asesor designado del estudiante **CESAR ALEJANDRO PEDROZA GARCÍA** con ID *44767* quien realizó *la tesis* titulada: **PROPUESTA DE UN MÉTODO PARA ANALIZAR DATOS DE TWITTER CONSIDERADOS COMO BIG DATA PARA GENERAR NUEVA INFORMACIÓN**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a *imprimirla*, y así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a *26* de *Mayo* de 2016.

Dra. María Dolores Torres Soto
Asesor de *tesis*

c.c.p.- Interesado
c.c.p.- Secretaria de Investigación y Posgrado
c.c.p.- Jefatura del Depto. de Sistemas de Información
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

Agradecimientos

Quiero agradecer en primer lugar a Dios por todas las bendiciones que me ha dado en la vida en especial la de hace un par de días que me mando a mi hermosa hija Aria, y por aquellos retos que me ha puesto y que me han hecho crecer de manera personal y profesional.

También me gustaría expresar mi más sincero agradecimiento al maestro Félix Patlán y al doctor Julio Dena que fueron quienes me alentaron a ponerme el reto de estudiar un posgrado.

A mi asesor, el doctor Juan Muñoz por sus consejos y todo el apoyo personal y profesional que me ha brindado antes y durante el tiempo de preparación de esta investigación; por su acertada orientación y disposición para la realización de este trabajo considerando las saturadas cargas de trabajo que tiene.

A mis sinodales, el maestro Abel Coronado y a la doctora María Dolores por su paciencia, dedicación, conocimiento y orientación brindados para la realización de este trabajo.

A los profesores que compartieron sus conocimientos, vivencias y experiencias a lo largo de todo el posgrado en sus diferentes especialidades haciendo más ameno el extenuante horario de clase.

Dedicatorias

Este trabajo se lo dedico en primer lugar a Dios quién supo guiarme por el buen camino durante estos poco más de dos años, darme las fuerzas para seguir adelante y enfrentar los problemas que se presentaban enseñándome a encarar las adversidades y por rodearme de personas valiosas que me han brindado su apoyo de manera incondicional en mi crecimiento personal y profesional.

A mi esposa Berenice por su paciencia, su apoyo y los ánimos que me brinda día con día para lograr mis metas personales y profesionales. También por el amor y comprensión que me brinda ya que este tiempo de preparación también implicó muchos sacrificios para ella.

A mi hija Aria, que al saber de su llegada me motivo a seguir adelante en este reto de estudiar la maestría y por hacerme dar cuenta sobre lo que realmente importa en la vida.

A mis padres porque sin su ejemplo, dedicación y cariño me hubiese sido posible llegar hasta el punto en el que estoy, personal y profesionalmente.

A mis hermanos Marco, Rodrigo y Daniel por darme los ánimos y consejos para seguir adelante.

A Martha y a Paula por brindarme sus consejos y darme la seguridad que necesitaba en tiempos difíciles.

Índice General

Índice de Tablas

Índice de Figuras

Resumen

Las oficinas nacionales de estadísticas son los organismos que por ley se encargan de recolectar, procesar, almacenar y difundir información estadística de manera oficial proveniente de las fuentes de datos tradicionales como los censos, las encuestas y los registros administrativos. En México, la oficina nacional de estadística es el INEGI quien aparte de realizar sus actividades cotidianas, ha estado explorando los diferentes mecanismos para extraer información de temas de interés nacional proveniente de otras fuentes de datos, como es el caso de la red social de Twitter que por las características que tiene entra dentro de la categoría de Big Data, y así poder determinar si se cuenta con los recursos necesarios para trabajar en este tipo de proyectos.

La presente investigación propone y prueba un método para trabajar con los datos recolectados de la red social de Twitter para darle un orden y llevar un control del proceso de generación de información de lo que se publica en México mediante la identificación de los elementos involucrados y sus características, la forma en la que interactúan, las actividades relacionadas con el proceso y el conjunto de conocimientos y habilidades de las personas que estarán trabajando en los proyectos. Para la realización del método se estudió la literatura del estado del arte, las técnicas y herramientas (estas últimas utilizando un proceso de prueba y error), y las experiencias y el conocimiento adquirido tras desarrollar otras soluciones dentro del mismo dominio en el INEGI para determinar qué elementos son considerados importantes en el análisis de Big Data. El método integra dichos aspectos con la finalidad de enriquecerlo y perfeccionarlo para poder utilizarlo en nuevos proyectos de manera que pueda ajustarse y ser útil en diferentes entornos organizacionales.

Para probar el método propuesto se diseñaron dos casos prácticos, el primero fue un análisis de movilidad cotidiana de los usuarios de Twitter que publican dentro del territorio nacional mediante el que cual se puede obtener los patrones de desplazamiento que han tenido en un determinado tiempo. Y por otro lado, se realizó un análisis de impacto de eventos de la vida real que permitió conocer la forma en la que un evento impacta en la sociedad mediante el estudio del número de menciones que tienen en las publicaciones de los usuarios de Twitter contra el tiempo de duración de dichos eventos.

Abstract

National statistical offices are agencies which by law are responsible for collecting, processing, storing and disseminating statistical information gathered officially from traditional data sources such as censuses, surveys and administrative records. In Mexico, the NSO is INEGI who besides performing daily activities, has been exploring different mechanisms for extracting useful information about national interest topics from other data sources, such as the social network Twitter having features that fit within the category of Big Data, in order to determine if it has the resources needed to work in thus kind of projects.

This research proposes and evaluates a method for working with data collected from the social network Twitter to arrange and control the process of generating statistics of what is published in Mexico by identifying the elements involved and their characteristics, the way they interact among them, the activities related to the process and the knowledge and skills of people who will be working on these projects. To accomplish the method it was needed to study the literature of the state of art, techniques and tools (the latter using a test and error process), and the experiences and knowledge gained from developing other solutions within the same domain in INEGI in order to determine which elements are considered important in the analysis of Big Data. The method integrates these aspects in order to enrich and perfect to use in new projects so that it can adjust and be useful in different organizational environments.

To test the method proposed there were designed two case studies, the first was an analysis of daily mobility of Twitter users who publish within the national territory obtaining displacement patterns that have taken in a given time. On the other hand, an impact analysis of events was developed and in order to know how an event of real life impacts on society by studying the number of mentions that are published by Twitter users against duration of these events.

Introducción

Big Data se ha convertido en un tema relevante donde organizaciones y gobiernos de varios países están interesados en descubrir las ventajas que se pueden obtener al analizar los diferentes tipos de datos generados dentro y fuera de las mismas. Sin embargo, actividades como la administración, el seguimiento, el control y el análisis de los datos pueden llegar a ser complicadas en el contexto de Big Data si no se cuenta con la infraestructura, el conocimiento y las habilidades necesarias para desarrollarlas. También es necesario tener en cuenta que, basándose en la experiencia, la posibilidad de éxito de cualquier empresa siempre es mayor cuando se cuenta con una serie de métodos que permitan guiar el desarrollo de dichas actividades.

El término de Big Data se ha estado esparciendo con gran rapidez en dominios como el cuidado de la salud, la mercadotecnia, las finanzas, el entretenimiento y otros, sin embargo en el campo de las estadísticas oficiales aún es un tema de discusión por los retos que implica asegurar la calidad y la accesibilidad de los datos junto con su almacenamiento y su procesamiento (Scannapieco Monica, Virgillito Antonino, & Zardetto Diego, 2013).

Para hacer frente a esta situación, las Oficinas Nacionales de Estadística (ONE por las siglas) en conjunto con organismos internacionales están formando grupos para confrontar estos retos además de determinar qué productos se pueden ofrecer considerando las diferentes fuentes de Big Data disponibles.

En México, el Instituto Nacional de Geografía y Estadística (INEGI por sus acrónimo) es el organismo encargado de captar, procesar y difundir información de calidad, veraz y oportuna del territorio, la población, la economía, etc. (INEGI, 2014). Esta institución ha manifestado gran interés en conocer el potencial de utilizar Big Data para la generación de información actualizada y coherente en temas como los anteriormente mencionados, con la finalidad de proporcionar mejores insumos para tomar decisiones bien sustentadas y desarrollar políticas públicas más eficientes.

Uno de los proyectos exploratorios en los que ha sobresalido el INEGI a nivel mundial y que es reconocido por la Comisión Económica de las Naciones Unidas para Europa, (UNECE por su acrónimo en inglés de United Nations Economic Commission for Europe) es el análisis de tweets con el que se genera información relacionada con la migración de

la población, el turismo, etc. De la misma manera y con la misma fuente de datos se está trabajando en realizar análisis de sentimiento el cual tiene la intención de conocer el estado de ánimo de la gente que publica tweets sobre temas específicos de interés nacional.

El presente trabajo proporciona una propuesta de un método informático para realizar análisis con datos de Twitter para generar información nueva en temas de interés nacional. Para probar y validar la utilidad del mismo se realizaron dos casos prácticos, un análisis de movilidad cotidiana y otro para análisis de impacto de eventos de la vida real en México, y al mismo tiempo se pretende que sirvan de guía para las organizaciones que tienen pensado incursionar en el desarrollo de proyectos de análisis de Big Data.

La estructura que tiene este trabajo de investigación es descrita en siete capítulos. En el primero de ellos se explica la situación problemática por la cual se decidió realizar esta investigación tomando en cuenta la posición en la que se encuentran las ONE para lidiar con los retos que implica trabajar con Big Data. También se define el tipo de investigación, los objetivos y preguntas de investigación que se resolvieron al desarrollar este trabajo.

En el capítulo dos se encuentra la información referente a la revisión del estado del arte donde se presentaron las teóricas propuestas por diversos autores que han trabajado en investigaciones de índole similar y del mismo modo se estudiaron las contribuciones y limitaciones de diversas metodologías de análisis de Big Data y procesos de ciencia de datos que fueron tomadas para este trabajo de investigación.

En el capítulo tres se hace referencia al marco teórico donde se identificaron las áreas de conocimiento presentando algunos conceptos, técnicas, herramientas y tecnologías para solucionar este tipo de problemas y del mismo modo, se revisó la literatura para los casos prácticos con los que posteriormente se probó el método propuesto.

En el capítulo cuatro se describe la metodología que se siguió para atacar al problema planteado en los capítulos anteriores donde se menciona la forma en la que se desarrollaron cada uno de los capítulos de esta investigación.

El capítulo cinco es el más importante de esta investigación ya que es donde se describen las fases que conforman el método propuesto tomando como base las teorías revisadas en el marco teórico y los análisis exploratorios realizados en el INEGI. La

aportación directa del tesista consistió en observar y conocer el proceso que seguía el personal del INEGI en la elaboración de algunos estudios de análisis, posteriormente se revisaron diversas metodologías de análisis de Big Data y procesos de ciencia de datos para identificar que actividades y en qué orden de ejecución podrían contribuir y complementar el proceso original del Instituto brindándole mayor claridad y entendimiento. Después de esto se procedió a crear una versión inicial del método a proponer el cual fue haciéndose cada vez más robusto cuando se realizaron los casos prácticos con los que se probó su utilidad.

En el capítulo seis se detalla cómo se siguió el método propuesto, fase por fase, para resolver los casos prácticos de análisis de movilidad cotidiana y de impacto de eventos de la vida real con el que se probó la utilidad del método mediante su aplicación en la práctica con los datos recolectados de la red social de Twitter para generar información actualizada y coherente en los temas de interés nacional antes mencionados.

En el capítulo siete se detallan los resultados de la investigación y del desarrollo de los casos prácticos realizados.

Finalmente, se detallan las conclusiones de la investigación donde se explica cómo se alcanzaron los objetivos planteados, se detallan los retos y limitaciones que implicó desarrollar este trabajo y se proponen algunas consideraciones sobre trabajos a futuro para continuar avanzando y mejorando el método propuesto.

Capítulo 1. Estructura de la investigación

Descripción de la problemática particular

El organismo encargado de recolectar, procesar y difundir información relacionada con el territorio, la población, la economía, etc. del país (INEGI, 2014) es el INEGI el cual ha puesto gran interés en conocer el potencial de explotar los datos generados en la red social de Twitter, considerada como fuente de Big Data, con fines exploratorios, obteniendo aquellos datos que tengan las características útiles para generar nueva información actualizada y coherente en temas de interés nacional que puedan complementar las generadas por las ONE y por ende apoyar a las distintas organizaciones a tomar de decisiones más eficientes.

Uno de los proyectos en los que ha sobresalido el INEGI a nivel mundial es el análisis de tweets con el que se genera información relacionada con la migración de la población, el turismo, etc. De la misma manera y con la misma fuente de datos se está trabajando junto con otras instancias nacionales e internacionales para realizar análisis de sentimientos el cual tiene la intención de conocer el estado de ánimo de la gente cuando publica tweets sobre temas específicos de interés nacional. Sin embargo, cuando el INEGI empezó a incursionar en el desarrollo de estos proyectos se topó con muchos problemas, ya que el proceso que se seguía no era del todo claro por lo que se utilizaba el método de prueba y error realizando los ajustes que se fueran presentando dependiendo de la exigencia del análisis. Con el paso del tiempo el personal involucrado empezó a obtener experiencia y pese a no tener documentado estos procesos ha logrado desarrollar los proyectos previamente comentados. De aquí nace la necesidad de crear un método informático que permita conocer la manera de realizar estos procesos para que pueda ser utilizado como guía y lograr un desarrollo estandarizado de estas soluciones en todo el INEGI.

Planteamiento del problema

Si bien es cierto que las organizaciones privadas son las que van un paso adelante en la generación de soluciones de Big Data, algunas ONE incluyendo el INEGI de México, han estado trabajando arduamente para explorar los beneficios que se pueden obtener en el campo de las estadísticas oficiales, ya sea generando nuevas estadísticas o para ofrecer datos que ayuden a complementar las estadísticas obtenidas de analizar las fuentes de datos tradicionales, para identificar los retos que Big Data presenta antes, durante y después de solucionar el problema; y para evitar el riesgo que existe de ser reemplazados por organizaciones no oficiales que generan grandes cantidades de datos y que están listas para explotar las oportunidades de Big Data (Scannapieco Monica et al., 2013). Aunado a esta situación, y por ser un tema relativamente reciente, muchas de ellas aún están en etapas tempranas de implementación llegando a tener dificultades por situaciones metodológicas, de calidad, de tecnología, de acceso a datos, de legislación y privacidad, etcétera haciendo que sus desarrollos se compliquen o simplemente no se puedan concluir.

Tal es la necesidad de que las ONE exploren las fuentes de Big Data que la Comisión de Estadística de las Naciones Unidas aceptó la creación, en Marzo de 2014, del Grupo de Trabajo Global (GWG por las siglas en ingles de Global Work Group) de Big Data en las Estadísticas Oficiales, donde durante su primer reunión el 31 de octubre de 2014, reconoció lo siguiente:

"*El uso de Big Data para generar estadísticas oficiales es una obligación de la comunidad estadística basada en el principio fundamental de satisfacer las expectativas de la sociedad de productos mejorados con mejores y más eficientes maneras de trabajar*" (UN, 2015).

Esto ayuda al cumplimiento de la agenda para el desarrollo post-2015, (proceso dirigido por las Naciones Unidas que tiene como objetivo ayudar a definir un marco común de acción y cooperación global para el desarrollo sustentable con adopción desde el año 2000) y con los Objetivos de Desarrollo Sostenible (SDG por las siglas en ingles de Sustainable Development Goals) se realzó la necesidad de enfrentar los desafíos estadísticos relacionados con la incorporación de nuevas fuentes de datos con enfoques

innovadores (Data-Pop Alliance(Harvard Humanitarian Initiative, MIT Media Lab y Overseas Development Institute), 2016).

La necesidad de proporcionar un método que sirva como guía desde etapas tempranas y durante el desarrollo de soluciones Big Data es el objetivo particular que tiene este trabajo. Si se revisa la definición básica de un método se tiene que es un conjunto de pasos ordenados para hacer algo, de aquí el interés por conocer cuáles son los elementos y el camino científico a seguir para realizar análisis con datos de manera regular provenientes de Twitter.

El problema que este proyecto pretende resolver es satisfacer la necesidad de contar con un método que permita identificar de manera clara y sencilla "qué elementos son necesarios y que características deben de tener", "cómo es la forma en la que interactúan", "qué actividades son esenciales" y "qué roles participan en los proyectos" para generar estadísticas útiles utilizando los datos recolectados de la red social de Twitter. El método proporciona un panorama general sobre los componentes a utilizar incluyendo un conjunto de técnicas y herramientas para explotar este tipo de datos.

Tipo de Investigación

Esta investigación se considera como un trabajo de diseño conceptual descriptivo de acuerdo a lo expuesto por (Mora, 2004), (Mora, Gelman, Paradice, & Cervantes, 2008) y por (Hevner, March, Park, & Ram, 2004), un método empírico en el que se desarrolla un objeto conceptual para obtener y analizar los datos generados en redes sociales que cumplan con características que sean útiles para generar información actualizada en temas de interés nacional.

El estudio tuvo un enfoque exploratorio ya que estableció una serie de aspectos útiles que en base a la observación del cómo se realizan los análisis de Big Data en el INEGI y la revisión y estudio de otros trabajos realizados de índole similar por otros investigadores permitieron diseñar el método propuesto. Luego de esto, el método entró en una fase de prueba mediante la elaboración de un par de casos prácticos de soluciones de análisis Big Data (análisis de movilidad cotidiana y de impacto de eventos de la vida real) para generar información de estos temas.

Objetivos generales y específicos

OBJETIVO GENERAL:

- Proponer un método que permita explorar los datos registrados en las conversaciones de los usuarios de la red social de Twitter, consideradas como fuente de Big Data, para generar información actualizada y coherente para posteriormente probarlo en temas de interés nacional como el análisis de movilidad cotidiana y análisis de impacto de eventos de la vida real.

OBJETIVOS ESPECÍFICOS:

- Identificar los elementos necesarios para producir información estadística basada en datos provenientes de Twitter.
- Determinar cómo interactúan los elementos para la producción de información basada en datos de Twitter como base para el desarrollo de un método.
- Probar el método para analizar la capacidad de determinar patrones de movilidad mediante el análisis de los metadatos de las conversaciones originadas en Twitter utilizando los parámetros de posición geográfica, tiempo y frecuencia.
- Probar el método para examinar la capacidad de medir el impacto de eventos en las conversaciones registradas en Twitter a través de los parámetros de frecuencia de menciones y longitud en el tiempo.

Preguntas de investigación

PREGUNTA PRINCIPAL

- ¿Cuál es la serie de pasos a seguir para obtener información actualizada haciendo uso de los datos registrados en las conversaciones de los usuarios de la red social de Twitter generadas en México?

PREGUNTAS SECUNDARIAS

- ¿Cuáles son los elementos que se necesitan para trabajar con datos provenientes de Twitter para producir información actualizada?
- ¿Cómo es que los elementos interactúan entre sí?
- ¿Qué clase de patrones de movilidad se pueden descubrir al aplicar el método propuesto para analizar los metadatos de las conversaciones originadas en Twitter utilizando los parámetros de posición geográfica, tiempo y frecuencia?
- ¿Cómo afecta un evento de la vida real en la red social de Twitter con relación a los parámetros de frecuencia de menciones y longitud en el tiempo registrados en las conversaciones de los usuarios?

Justificación

Con las nuevas y diversas fuentes de Big Data es posible generar conocimiento de una manera más rápida y económica. Estas fuentes de datos tienen gran utilidad en el campo de las estadísticas oficiales ya que pueden ayudar, entre otras cosas, a llenar espacios vacíos donde los datos son escasos o difíciles de obtener mediante las fuentes de datos tradicionales por cuestiones de costos, tiempo de recolección, procesamiento, análisis, etcétera. Del mismo modo, llega a tener utilidad para el logro de los objetivos de la agenda para el desarrollo post-2015 de las Naciones Unidas entre los que se encuentra la lista de SDG. Solo por mencionar algún ejemplo, para lograr el objetivo uno del SDG que consiste en erradicar la pobreza de manera global, es necesario recolectar datos principalmente a través de costosas encuestas a hogares que pueden incluir pequeñas unidades geográficas alejadas de las zonas urbanas como pueblos, comunidades y aldeas que implica la movilización de entrevistadores a dichos lugares. Una posible fuente de Big Data que ayudaría a resolver esta situación podrían ser los datos recolectados de teléfonos móviles los cuales tienen por lo general una alta penetración en la población (Data-Pop Alliance(Harvard Humanitarian Initiative, MIT Media Lab y Overseas Development Institute), 2016).

Ciertamente es importante comentar que se han desarrollado muchos trabajos y publicaciones que utilizan Big Data para realizar distintos tipos de análisis como la minería de opiniones (Hodeghatta, 2013), (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011), (Hao Wang, Can, Abe Kazemzadeh, François Bar, & Shrikanth Narayanan, 2012) y el análisis de movilidad humana (Zagheni, Garimella, Weber, & State, 2014), (Gabrielli Lorenzo, Rinzivillo Salvatore, & Ronzano Francesco, 2014), sin embargo uno de los principales problemas es que las técnicas de minería de datos para las estadísticas oficiales están en pleno desarrollo (Buelens, Daas, & van den Brakel, 2012).

Considerando los aspectos vistos en el planteamiento del problema, esta investigación pretende proporcionar, en primer lugar, al INEGI un método informático que permita guiar y estandarizar el proceso a seguir para el desarrollo de soluciones que utilicen Twitter, que es considerada como una fuente de Big Data, para obtener información actualizada y coherente en temas de interés nacional. Y en segundo lugar, a pesar de que el método está enfocado en esta problemática muy particular se pretende que pueda ser utilizado por cualquier organización, pública o privada, que esté interesada en analizar los datos

gratuitos provenientes de Twitter o por el otro lado utilizarlo como base para utilizar otras fuentes de Big Data.

Las expectativas que tiene el personal del INEGI respecto al desarrollo de este tipo de proyectos con Twitter y otras redes sociales consideradas como fuentes de Big Data, son muy altas donde para poder compararlas y/o complementar estadísticas oficiales, en el año 2015 se logró incluir una pregunta respecto al uso de redes sociales en la Encuesta Nacional sobre Disponibilidad y Uso de las Tecnologías de Información y Comunicaciones en Hogares (ENDUTIH por sus siglas) que trata de obtener información sobre disponibilidad y uso de la tecnología por individuos de seis años y más en México.

Los casos prácticos que fueron seleccionados para probar el método propuesto tienen un alto impacto social. El primero de ellos, el análisis de movilidad humana, ayuda a conocer entre una variedad de estudios, los patrones de desplazamiento de los usuarios de Twitter mediante la obtención de la referencia geográfica en la que se encuentran al momento de publicar un tweet. Con este tipo de estudios se ven beneficiados, por ejemplo, los proceso de planeación de transporte, carreteras y vialidades que llevan a cabo las dependencias correspondientes. Un estudio de este tipo normalmente suelen ser algo tardado dependiendo de las características que delimitarán el proyecto como el tamaño de la zona geográfica y la variedad de temas a cubrir llegando a encontrar trabajos que pueden tardar de seis a doce meses y ocupar una gran cantidad de personas para realizar las actividades de levantamiento, procesamiento y análisis de los datos. Un estudio de esta clase fue el realizado por el INEGI en la Zona Metropolitana del Valle de México (ZMVM por sus siglas), abarcando fechas desde Octubre de 2006 a Marzo de 2007 para realizar una prueba piloto; del 12 de Mayo al 16 de Junio de 2007 para el levantamiento de los datos y del 28 de Mayo al 7 de Agosto para su procesamiento y llegó a necesitar de un total de 1,416 personas. Otro estudio fue el de la empresa USTRAN en la zona conurbada de Guadalajara en el año 2002-2003 que tardó ocho meses en llevarse a cabo y ocupo 800 meses hombre (http://www.ustran.com/casos/proyecto43.htm).

Por otro lado, el análisis de impacto de eventos permite conocer el grado en el que un evento de la vida real puede afectar las actividades diarias de la sociedad que tiene cuenta en la red social mediante el estudio de los mensajes que los usuarios publican contra el tiempo de duración del evento. Si bien, es posible encontrar diversas organizaciones privadas que otorgan este tipo de estudios donde dependiendo del tipo de

servicio que se contrate se puede definir un conjunto de características como la cantidad máxima de tweets a analizar y durante cuánto tiempo (uno o dos meses) y la cantidad máxima de palabras a buscar; además algunas de ellas cuentan con herramientas como la generación de reportes históricos por hasta uno o dos meses y en tiempo real, etc. Los costos de estos servicios se pueden encontrar desde los $16 dólares mensuales con lo más básico hasta $679 el más completo (Twitterbinder (https://www.tweetbinder.com/), TweetReach (https://tweetreach.com/), follow the hashtag (http://www.followthehashtag.com/)). A pesar de que en la aplicación del método solo se obtuvó el porcentaje de tweets que pone Twitter de manera gratuita es posible realizar estos estudios sin hacer alguna inversión fuerte, bastando únicamente con tener un conocimiento en lenguajes de programación y algunas técnicas estadísticas.

El trabajo queda abierto para trabajos futuros en los que se puede tomar como base ya sea para compararlo con nuevos métodos para trabajar con fuentes de Big data que para generar estadísticas hasta para probar la generación de algún modelo estadístico que resuelva alguna problemática en particular.

Capítulo 2. Estado del Arte

De acuerdo a una encuesta de Gartner los empresarios están invirtiendo, o tienen pensado hacerlo, en proyectos de Big Data debido a que han visto las oportunidades que conlleva analizarlos (Mousannif, Sabah, Douiji, & Sayad, 2014). Esto se ve reflejado en un sin número de trabajos de investigación que en las últimas fechas se han estado realizando teniendo como origen principal los datos generados en las distintas fuentes disponibles de Big Data, como el internet de las cosas (Singh, Pandey, Shankar, & Dumka, 2015) (Yang, Hu, Cheng, Miao, & Zheng, 2014a), los datos abiertos y las redes sociales (Abbasi et al., 2014) (He, Zha, & Li, 2013) (Azmandian, Singh, Gelsey, Chang, & Maheswaran, 2013), por mencionar algunas; brindando información de mucho interés para las organizaciones tanto privadas como públicas que deciden desarrollarlos.

Cabe resaltar que no todos los proyectos de Big Data son iguales ya que dependen directamente de las necesidades propias de la investigación, sin embargo, todos los proyectos tiene el mismo objetivo en común: obtener descubrimientos en los datos que tienen a su disposición. Aunado a esta necesidad, la escasez de personal con el conocimiento y las habilidades, además de la falta de un método que guie a los departamentos de TI en la planeación y ejecución de un proyecto de Big Data complican aún más la situación de las organizaciones a implementar este tipo de proyectos.

Una metodología es una guía que se sigue con la finalidad de realizar las acciones propias de una investigación indicando qué hacer y cómo actuar ante un problema recurrente trayendo consigo su solución de una forma veraz, sistemática y disciplinada. Todo esto basándose en la infraestructura tecnológica, los procesos, las herramientas y las habilidades que debe tener el personal en temas de modelado analítico y en estrategias de decisión (Sheikh, 2013).

El análisis de Big Data involucra el uso de técnicas y herramientas mediante las cuales es posible detectar patrones en el comportamiento de los datos. Debido a la naturaleza propia de Big Data es necesario dar un enfoque diferente a los procesos de análisis de datos tradicionales. Un ejemplo claro se puede observar en el análisis del streaming de datos donde se requiere de una gran cantidad de almacenamiento disponible, procesamiento en altas velocidades, aplicar técnicas de análisis rápidas (incluso en tiempo real) para satisfacer las necesidades de información de la organización, etc.

Metodologías de análisis de Big Data

A continuación se muestra la revisión de literatura de algunas metodologías existentes para trabajar con Big Data.

Metodología de análisis de datos de Mohanty y sus colegas

El trabajo realizado por (Mohanty, Jagadeesh, & Srivatsa, 2013) presenta una metodología de análisis de datos cíclica bastante completa formada por siete fases. La diferencia que tiene esta metodología con otras implementaciones de análisis de datos y metodologías de inteligencia de negocios es el número de iteraciones que deben de hacer los diseñadores para resolver el problema en procesos de gran escala.

Una pequeña descripción de lo que se realiza en cada una de las fases de la metodología son los siguientes:

- Analizar y evaluar los casos de uso del negocio. Es el equivalente a realizar una prueba de concepto donde se detectan cuestiones como la falta de valor en los datos, la falta de infraestructura para soportar los diferentes formatos de datos, etcétera. También se empiezan a definir los objetivos del análisis de los datos detectando cuestiones como el problema, el comportamiento, las complicaciones, el impacto hacia el negocio, los antecedentes y las condiciones en las que suceden los problemas.
- Desarrollar las hipótesis del negocio. Se realizan análisis de datos exploratorios para resolver el problema de la empresa con más certeza en cuanto a la suficiencia y utilidad de los resultados, para luego realizar el análisis completo. Es recomendable seguir las iteraciones hasta que todos los involucrados están de acuerdo en los requisitos técnicos y del negocio.
- Desarrollar el enfoque de análisis. Se definen las técnicas y el método de análisis para resolver el problema del negocio, así como también se establecen el tipo de salidas que tendrá el sistema.
- Construir y preparar los conjuntos de datos. Los datos pasan por un proceso de extracción, limpieza, transformación y carga para dejar únicamente los que se van a requerir para su análisis.

- Seleccionar y construir los modelos de análisis. Como los modelos de análisis pueden llegar a ser complejos es común ver prototipos desarrollados con algoritmos utilizando muestras significativas, siendo este un proceso iterativo hasta encontrar el mejor algoritmo para después ejecutarlo en un ambiente de producción.
- Construir el sistema de producción. Se crea una aplicación con el fin de cubrir dos enfoques distintos, la exploración de los datos y el procesamiento de los datos.
- Medir y monitorear. Para poder medir el resultado es indispensable poder interpretar la información, además de identificar posibles errores, bugs o modelos que pudiesen afectar y que provoquen datos no útiles. El sistema de análisis debe tener herramientas para ofrecer a los usuarios la capacidad de interpretar los resultados del análisis y repetirlo tantas veces sea necesario con diferentes supuestos, parámetros, o conjuntos de datos.

En la Figura 1 se puede apreciar cómo es que está planteada esta metodología.

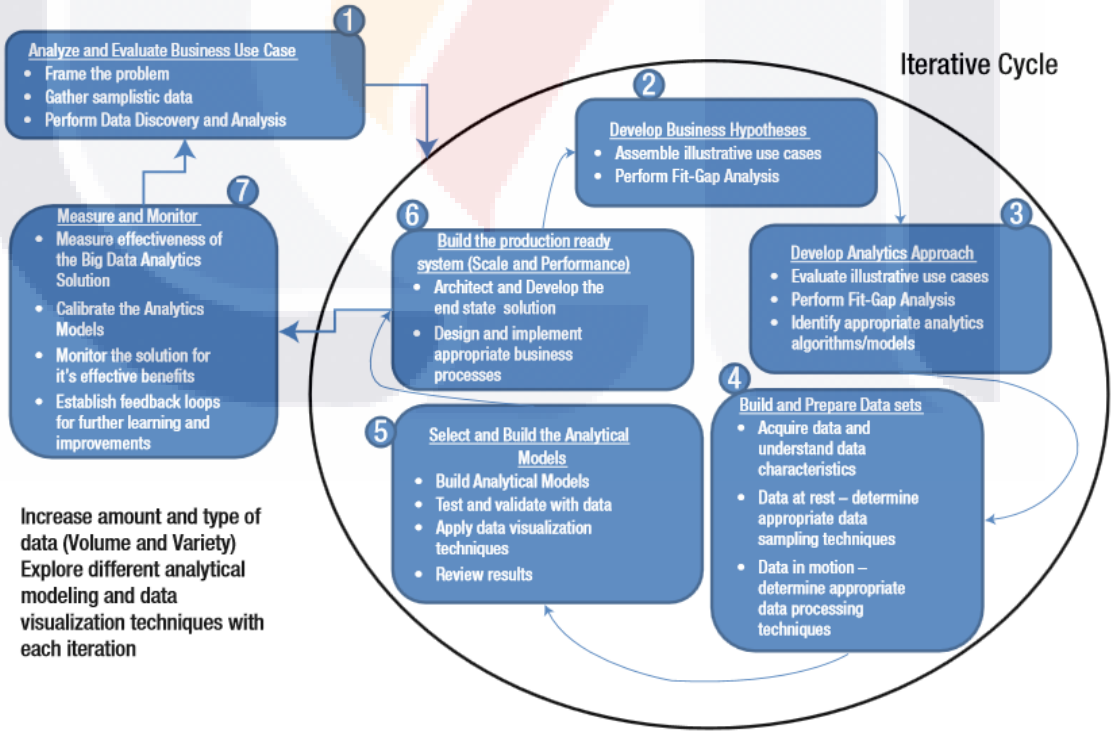


Figura 1. Metodología de análisis de Big Data. Fuente (Mohanty et al., 2013)

Metodología de análisis de datos de Raghupathi y sus colegas

Otra metodología para analizar fuentes de Big Data es la presentada por (Raghupathi & Raghupathi, 2014). Algo muy peculiar de esta metodología es que está enfocada y aplicada en el campo del cuidado de la salud donde el análisis de Big Data ha tenido mucha importancia y muchas aportaciones, llegando a encontrar investigaciones como la detección de brotes de influenza mediante el monitoreo de los registros de consulta del buscador de Google (Ginsberg et al., 2009), otro estudio similar es el de (Carneiro & Mylonakis, 2009) donde presentan la aplicación de Google Trends como una herramienta basada en la web para la vigilancia en tiempo real de los brotes de enfermedades, siendo de 7 a 10 días más rápido que el Centro para el Control y Prevención de Enfermedades (CDC por sus acrónimo en inglés Centers for Desease Control and Prevention).

La metodología está formada por cuatro fases entre las cuales existen ciclos de retroalimentación para minimizar el riesgo de fracaso, estas son:

- Fase 1: Declaración de concepto. Es donde establecen las necesidades del análisis de Big Data tomando como base las características con las que se generan y se revisan las ventajas y desventajas en las opciones de los costos, la escalabilidad, etcétera.
- Fase 2: Definición del propósito del estudio. Identifica preguntas que serán respondidas mediante el análisis del estudio tratando de ser lo más precisas posible ya que este tipo de investigaciones son complejas y costosas. Además de esto se proveen los antecedentes del estudio.
- Fase 3: Implementación del estudio. Realizan una serie de proposiciones basadas en la declaración concepto. Identifican variables dependientes e independientes, definen las fuentes de datos con las que van a trabajar para procesarlas y tenerlas listas para su análisis, siendo este último un proceso iterativo basado en análisis what-if. Realizan la evaluación y selección de las plataformas para después aplicar una serie de técnicas de análisis a los datos.
- Fase 4: Desplegado: Se prueban, se validan y se evalúan los modelos y los resultados con los involucrados en el negocio.

En la Figura 2 se puede apreciar cómo es que está planteada esta metodología.

Step 1	Concept statement
	• Establish need for big data analytics project in healthcare based on the "4Vs".
Step 2	Proposal
	• What is the problem being addressed?
	• Why is it important and interesting?
	• Why big data analytics approach?
	• Background material
Step 3	Methodology
	• Propositions
	• Variable selection
	• Data collection
	• ETL and data transformation
	• Platform/tool selection
	• Conceptual model
	• Analytic techniques
	-Association, clustering, classification, etc.
	• Results & insight
Step 4	Deployment
	• Evaluation & validation
	• Testing

Source: Adapted from [Raghupathi & Raghupathi, [9]].

Figura 2. Metodología de análisis de Big Data en el campo de cuidado de la salud. Fuente (Raghupathi & Raghupathi, 2014)

Metodología de análisis de datos de Mousannif y sus colegas

El trabajo realizado por (Mousannif et al., 2014) muestra una metodología para construir proyectos de Big Data. Está formada por 3 fases principales y estas a su vez están integradas por una o varias etapas.

- La primer fase está dedicada para la planeación del proyecto, donde se pueden ver las actividades necesarias para definir el problema que se va a resolver, el alcance del estudio, las fuentes de los datos que se van a requerir, etcétera. Del mismo modo se toman en consideración aspectos como la protección de datos sensibles y la seguridad de la red.
- La siguiente fase es la implementación del proyecto que abarca desde la recolección de los datos hasta la representación y visualización de los resultados

finales del procesamiento de los datos mostrando para cada actividad un conjunto de herramientas y tecnologías que facilitan su entendimiento. También cubre aspectos como la selección de algoritmos para el análisis de los datos.

- En la última fase, la post implementación, se encuentran las actividades donde se identifica y se extrae a tiempo información de interés para determinar las acciones que pueden brindar una ventaja competitiva a la organización. Posteriormente se evalúa el proyecto respondiendo preguntas que ayudan a medir los resultados esperados y su calidad.

En la Figura 3 se puede apreciar cómo es que está planteada esta metodología.

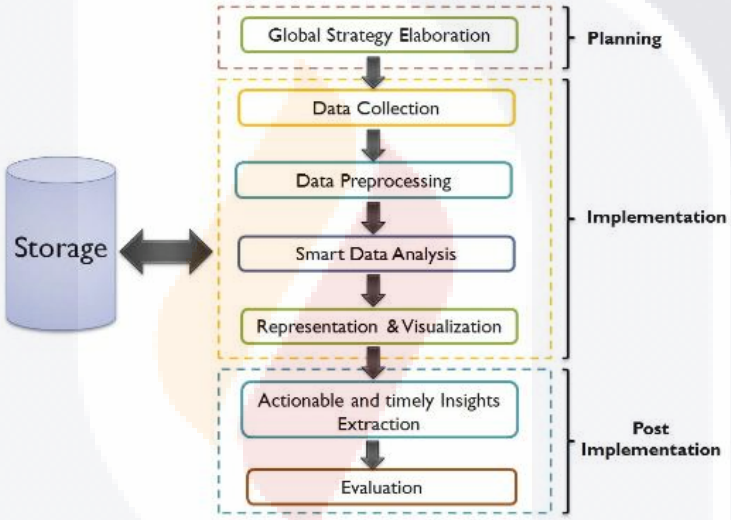


Figura 3. Metodología de un proyecto de Big Data. Fuente (Mousannif et al., 2014)

Metodología de análisis de datos de Sheikh

También es posible encontrar metodologías de análisis de datos que tienen como base enfoques tradicionales de desarrollo de software. Uno de estos trabajos es el presentado por (Sheikh, 2013) donde utiliza un enfoque tradicional de cascada y está formado por las siguientes fases:

- Requerimientos. En esta fase se realizan las actividades relacionadas con la recolección de los requerimientos del usuario y es en donde se obtiene

información relacionada con el problema y los objetivos de la investigación, se definen los requerimientos y se identifican las fuentes de las que se extraerán, del mismo modo se definen los requerimientos para el modelo, de las estrategias de decisión y de la integración operacional.

- Análisis. Se describen los requerimientos y la utilización de los datos con mayor detalle en el contexto del proceso de negocio identificando posibles huecos que deberán considerarse para tener un mejor entendimiento del problema y su solución. Los siguientes elementos son revisados a detalle: planteamiento y objetivo del problema, perfiles y datos (identificación de interdependencias y correlación entre los campos en el contexto del negocio), el modelo y la estrategia de decisión, integración operacional y auditoría y control de la solución.
- Diseño. En esta fase se diseñan las elementos que fueron revisados en el análisis, teniendo los siguientes componentes: extensión del almacén de datos, análisis de variables, análisis del data mart, diseño de las estrategias de decisión, diseño de la integración operativa, auditoría y control.
- Implementación. Esta fase consiste en obtener los datos desde su origen, manipularlos y cargarlos (extracción, transformación y carga, ETL por las siglas en inglés de Extraction, Transformation and Load) utilizando software de limpieza, formateo y agregación. El modelo de análisis es implementado en una herramienta de minería de datos y es probado, validado y evaluado. La estrategia de decisión puede ser desarrollada utilizando ambientes tradicionales de programación o mediante herramientas de administración de procesos de negocio (proceso de administración de negocio, BPM por las siglas en ingles de Bussiness Process Management) para posteriormente integrarla con el sistema operacional y auditarla y controlarla para reportar los resultados con herramientas de procesamiento analítico en línea (OLAP por las siglas en ingles de Online Analitical Process) como los tableros de control.
- Despliegue. El modelo de análisis es probado y validado para que ayude a la toma de decisiones integrando transacciones de datos y ejecutando estrategias con las salidas del mismo. Se realizan ejercicios de prueba y error al modelo para ajustarlo antes de ejecutar las estrategias en el negocio. Finalmente se realiza una prueba de integración de la solución incluyendo la grabación y revisión de datos de auditoría y control.

- Ejecución y monitoreo. En esta fase se pone en ejecución la solución y es monitoreada de cerca con la finalidad de detectar a tiempo posibles cambios de optimización al modelo. Para este caso utiliza la noción de estrategias de decisión campeón-retador donde una estrategia existente es denominada estrategia campeón y una nueva estrategia como estrategia retadora en donde esta última se pone a prueba para ver si mejora los resultados. Si después de una observación parece que la estrategia retadora está funcionando mejor, ésta sustituye la estrategia campeón.

La Figura 4 muestra la metodología antes mencionada.

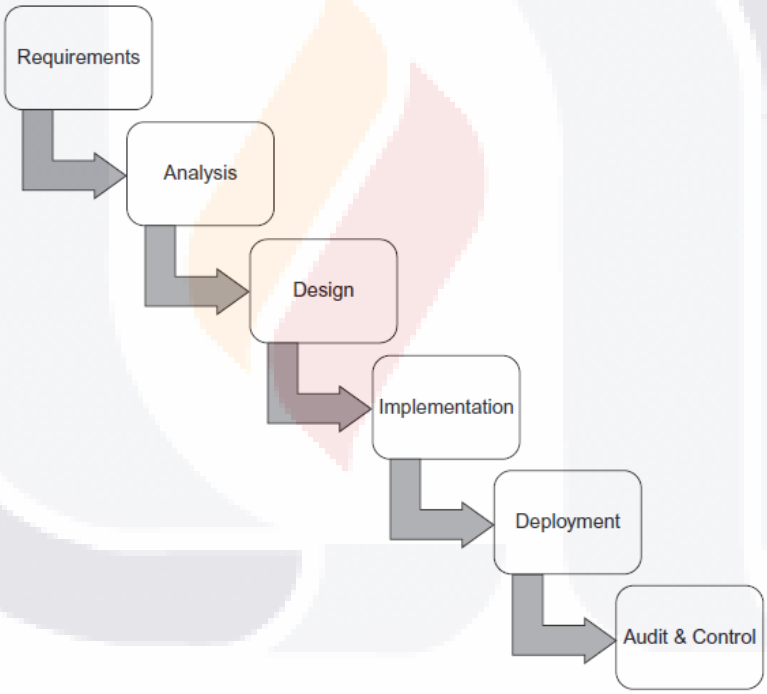


Figura 4. Metodología para el análisis de datos. Fuente (Sheikh, 2013)

Procesos de ciencia de datos

Otro tema importante por mencionar es la ciencia de datos la cual se puede definir básicamente como un campo interdisciplinario de sistemas y procesos para extraer

conocimiento de datos que pueden estar organizados o no y que provienen en distintas formas y de distintas fuentes.

Diversos autores (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) (Saltz, 2015) comentan acerca de la estrecha relación que existe entre la ciencia de datos y el descubrimiento de conocimiento en bases de datos (KDD por las siglas en ingles de Knowledge Discovery in Databases), donde este último se refiere al proceso de descubrir conocimiento e información útil dentro de los datos contenidos en algún repositorio de información. Teniendo esto en consideración, algunos investigadores han diseñado distintos procesos para proporcionar una mejor comprensión de las tareas implicadas en el análisis de datos. A continuación se comentan algunos de estos trabajos.

Proceso de ciencia de datos de Jagadish y sus colegas

Jagadish (Jagadish et al., 2014) describe un proceso iterativo formado por cinco etapas en las que ejemplifica, para cada una de ellas mediante un caso de estudio, algunos de los retos a los que se puede enfrentar el equipo de ciencia de datos. Las etapas son las siguientes:

- Adquisición de los datos. En esta etapa se hace referencia a la identificación de los datos que se necesitaran para hacer la investigación donde se hace un filtrado para separar la información que es útil de la que no lo es.
- Extracción de la información y la limpieza. En esta etapa se realizan actividades para extraer los datos y de homogenizarlos en caso de que provengan en distintos formatos. Estas actividades se realizan tomando en consideración la confiabilidad de los datos con los que se esté trabajando.
- Integración de datos. Análisis de datos a gran escala requieren colecciones de datos heterogéneas para que sean efectivos. En esta etapa se utiliza una serie de técnicas y herramientas para transformar e integrar los datos para poder interpretarlos en un formato estandarizado.
- Modelado y análisis. Consta de métodos para consulta y minería de Big Data.
- Interpretación. Etapa en la que se revisan los resultados por un tomador de decisiones teniendo en consideración los supuestos hechos al iniciar el proyecto para dos cosas, la primera para redirigir el análisis en caso de no haber

encontrado lo que se buscaba utilizando diferentes parámetros, fuentes de datos o el modelo de análisis; o la segunda, para reportar y tomar acciones para resolver la situación problemática.

Proceso de ciencia de datos de Guo

Guo proporciona un proceso iterativo (Guo, 2013) típico en la ciencia de datos y está formado por cuatro fases de alto nivel, como son la preparación de los datos, el análisis, la reflexión y difusión de los resultados. Cada una de estas fases se menciona a continuación:

- Preparación. Consta de las actividades de adquisición de los datos y la de formateo y limpieza de los mismos.
    - Adquisición de los datos. El primer paso es la obtención de los datos donde de acuerdo con Guo, los problemas más comunes a los que se enfrentan los equipos de ciencia de datos son: la procedencia, la administración y el almacenamiento de los datos.
    - Formateo y limpieza de los datos. Lo más probable es que los datos crudos vengan en distintos formatos por lo que el equipo de ciencia de datos debe de crear scripts o editarlos manualmente para estandarizarlos e integrarlos para poder realizar el análisis. También en esta capa se puede encontrar errores de semántica, errores de entradas y de inconsistencias entre formatos en los datos por lo que es importante limpiarlos.
- Análisis. La etapa de análisis consiste en escribir, ejecutar, inspeccionar y refinar los programas informáticos para obtener ideas de datos. El programador se involucra en un ciclo de iterativo de edición, ejecución, inspección de los archivos de salida para obtener información y el descubrimiento errores, su depuración y re-edición.
- Reflexión. Consiste en pensar y comunicar las salidas de los análisis. Se toman notas sobre los experimentos en formatos físicos y digitales; se realizan reuniones para discutir los resultados y planear los siguientes pasos para el análisis; se realizan comparaciones y se exploran alternativas entre las salidas para ajustar los códigos de programación o la ejecución de parámetros.

- Difundir resultados. En esta etapa se dan a conocer los resultados mediante reportes escritos como memorandos, presentaciones, o como publicaciones de investigaciones académicas. En algunos casos se distribuye el software desarrollado para que se reproduzcan los resultados o se realicen sistemas prototipos.

La Figura 5 muestra gráficamente los pasos de este flujo de trabajo.

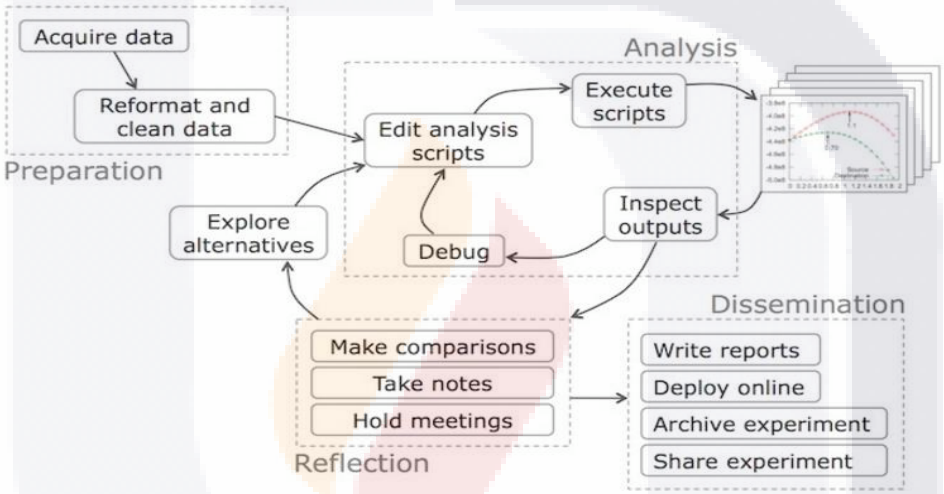


Figura 5. Flujo de trabajo de ciencia de datos de Guo. Fuente (Guo, 2013)

Proceso estándar de la industria cruzada para minería de datos (CRISP-DM)

En otros casos se ha utilizado el proceso estándar de la industria cruzada para minería de datos (CRISP-DM por las siglas en ingles de Cross Industry Standard Process for Data Mining) como un primer paso hacia la definición de una metodología de la ciencia de datos (Saltz, 2015). CRISP-DM es un modelo iterativo neutral desarrollado por la industria junto con el aporte de usuarios y de proveedores de herramientas y de servicios de minería de datos (Shearer, 2000) capaz de ofrecer a las organizaciones una visión más estructurada de un problema de análisis de datos promoviendo un conjunto de mejores prácticas para realizarlos de una mejor y más rápida manera. Las fases de este proceso son:

- Entendimiento del negocio. Esta fase se enfoca en entender el problema desde la perspectiva del negocio y en desarrollar un plan que permita resolverlo. Se desarrollan algunos pasos como la determinación de los objetivos del negocio, evaluación de la situación, determinación de los objetivos de la minería de datos y la producción del plan del proyecto.
- Entendimiento de los datos. Esta fase comienza desde que los datos son recolectados y los analistas empiezan a familiarizarse con ellos. Algunas fuentes de datos estarán disponibles de forma gratuita, mientras que otros necesitaran un esfuerzo mayor para obtenerlas teniendo que estimar la relación costo y beneficio de aquellas fuentes con las que se desea trabajar. Se compone de cuatro pasos que son: la recolección inicial, la descripción, la exploración y la verificación de la calidad de los datos.
- Preparación de los datos. Una vez que los analistas tienen una buena comprensión de datos, se procede a manipularlos, convertirlos y limpiarlos de tal forma que se puedan utilizar en algún modelo para obtener resultados específicos. Los cinco pasos en la preparación de datos son la selección, la limpieza, la construcción, la integración y el formateo de los datos.
- Modelado. En esta fase se seleccionan y se aplican las técnicas de modelado con los que se podrán analizar los datos. Estos modelos trabajan con un número determinado de parámetros que necesitan ser calibrados una y otra vez hasta obtener el valor más óptimo. Los pasos que se realizan en esta fase son la selección de la técnica de modelado, la generación de diseño de pruebas, la creación y la evaluación de los modelos.
- Evaluación. El propósito de esta fase es la de evaluar los resultados obtenidos de modelo de minería de datos para dar la confianza de que son válidos y fiables y que se satisfacen los objetivos de negocio originales. Los pasos que conforman esta fase son la evaluación de los resultados, la revisión de procesos, y la determinación de los próximos pasos.
- Despliegue. Dependiendo de los requerimientos iniciales, los resultados pueden ser organizados y presentados a los interesados para tomar decisiones o si es un proyecto que afecta un proceso determinado pueden ser puestos en uso real con el fin de tener algún retorno de la inversión. En esta fase se pueden ver pasos

como son la implementación del plan, la supervisión y el mantenimiento del plan, la producción del informe final, y la revisión del proyecto.

La Figura 6 muestra gráficamente las fases del modelo CRISP-DM y como se relacionan entre sí para hacer el proceso de manera iterativa.

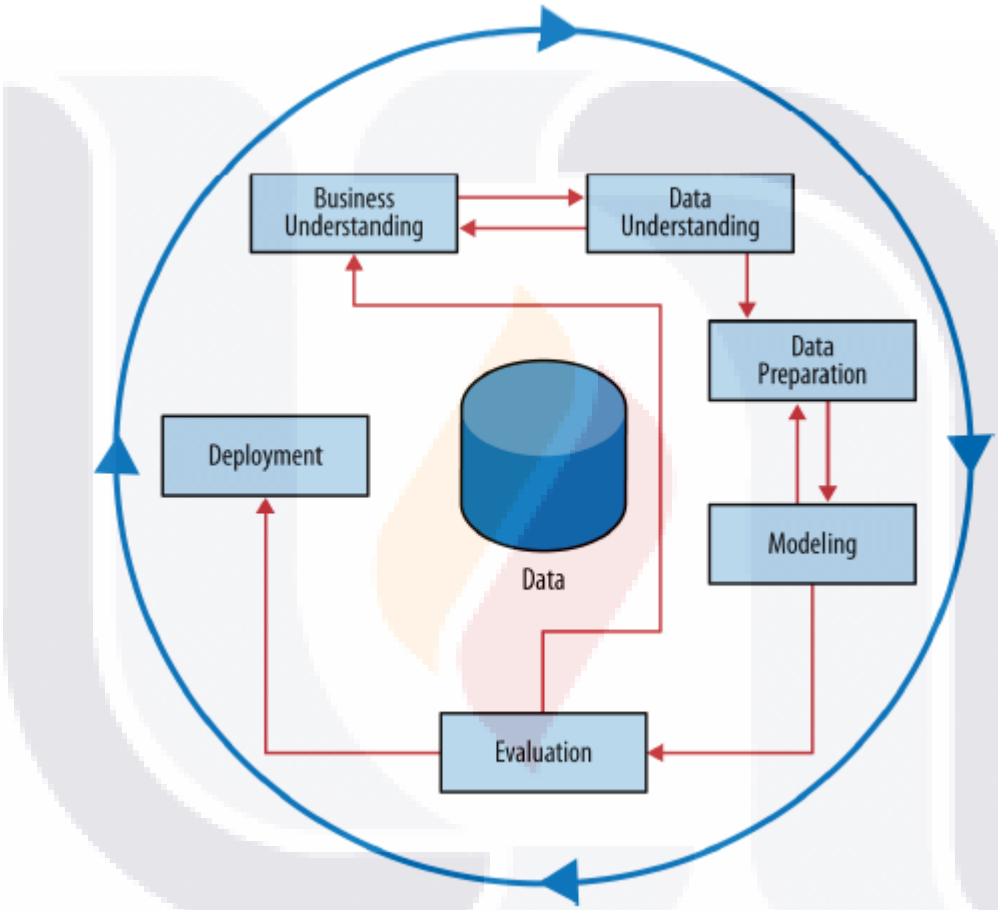


Figura 6. Modelo CRISP-DM para minería de datos. Fuente (Shearer, 2000)

Proceso de análisis de datos de Cuesta

De acuerdo con (Cuesta, 2013) es posible hacer predicciones de un evento si se tiene una buena comprensión del mismo. El proceso de análisis de datos que propone de un conjunto de actividades a realizar agrupadas en cinco fases las cuales son:

1. El problema. Esta fase consiste en la identificación de la situación problemática contestando preguntas de alto nivel que ayudaran a comprender los objetivos y requisitos desde una perspectiva de dominio. Otras preguntas ayudaran a determinar el enfoque que tendrá es estudio teniendo que pueden ser de tipo inferencial, profético, descriptivo, exploratorio, causal, correlacional, etcétera.
2. Preparación de datos. En esta fase se realizan las actividades necesarias para obtener, limpiar, normalizar, y transformar los datos de la manera más óptima y cuidando aspectos de calidad. Si se trabaja con datos con poca calidad puede ocasionar que el análisis arroje resultados engañosos y poco confiables. Para que los datos puedan considerarse buenos necesita tener las siguientes características: completos, coherentes, inequívocos, contable, correctos, estandarizado y no redundantes.
3. Exploración de datos. Se revisan los datos mediante métodos gráficos o estadísticos con la finalidad de encontrar patrones, conexiones y relaciones entre ellos.
4. Modelo predictivo. Se seleccionan o se crean modelos estadísticos para predecir de mejor manera el comportamiento de los datos y del problema. Posteriormente se evalúa el modelo para asegurarse de que el análisis no es demasiado optimista o demasiado ajustado. El autor presenta dos formas diferentes de validar el modelo, la validación cruzada donde se dividen los datos en subconjuntos del mismo tamaño y se prueba el modelo para estimar cómo se va a realizar en la práctica; y la validación Hold-Out que consiste en dividir aleatoriamente en tres subgrupos: juego de entrenamiento, set de validación y prueba.
5. Visualización de los resultados. En esta fase se identifica la forma en la que se presentan los resultados como por ejemplo: informes tabulares, gráficos 2D, cuadros de mando, etcétera. Si el resultado del análisis es un producto, entonces se identifica el proceso y se despliega mediante interfaces de escritorio, en la web, dispositivos móviles, etcétera. Cada elección dependerá de los de los datos y del tipo de análisis. Esta fase tiene el objetivo de identificar nuevos patrones o relaciones contenidas en los datos y tiene que ser lo visualmente significativa para ayudar en la toma de decisiones.

El proceso de análisis de datos que propone se compone de las actividades mostradas en la Figura 7.

FASE	ACTIVIDAD
El problema	Declarar del problema
Preparación de los datos	Obtener los datos
	Limpiar los datos
	Normalizar los datos
	Transformar los datos
Exploración de los datos	Estadísticas exploratorios
	Visualización exploratorio
Modelado predictivo	Modelado Predictivo
	Validar el modelo
Visualización de los resultados	Visualizar e interpretar sus resultados
	Despliegue su solución

Figura 7. Proceso de análisis de datos de Cuesta. Fuente este documento tomando como base lo expuesto por (Cuesta, 2013)

Resumen de similitudes de los principales trabajos revisados

En la Tabla 1 se presentan las similitudes que tienen las metodologías y los procesos de ciencia de datos revisadas anteriormente con la finalidad de identificar que fases o etapas corresponden entre cada una de ellas. El objetivo de incluir los procesos de ciencia de datos es porque, a pesar de que no cubren aspectos relacionados con temas relacionados con la organización y el planteamiento del estudio, las actividades que se realizan pueden ayudar a entender con mayor facilidad las fases finales de las metodologías revisadas debido al grado de detalle con el que se explican.

Tabla 1. Comparación de metodologías de análisis y procesos de ciencia de datos. Elaboración propia.

(Mohanty, 2014)	(Raghupathi & Raghupathi, 2014)	(Mousannif et al., 2014)	(Jagadish et al., 2014)	(Guo, 2013)	(Shearer, 2000)	(Cuesta, 2013)
Analizar y evaluar los casos de uso del negocio	Declaración de concepto	Elaboración de estrategia global			Entendimiento del negocio	El problema
Desarrollar las hipótesis del negocio	Definición del propósito del estudio					
Desarrollar el enfoque de análisis						
Construir y preparar los conjuntos de datos	Implementación del estudio	Implementación	Adquisición de los datos	Adquisición de los datos	Entendimiento de los datos	Preparación de los datos
			Extracción y limpieza de los datos	Extracción y limpieza de los datos	Preparación de los datos	
			Integración de los datos			
						Exploración de datos
Seleccionar y construir los modelos de análisis			Modelado y análisis	Análisis	Modelado	Modelo predictivo
					Evaluación	
Construir el sistema de producción	Desplegado	Post Implementación	Interpretación	Reflexión de resultados	Despliegue	Visualización de los resultados
				Difundir resultados		
Medir y monitorear						

Resumen de contribuciones y limitaciones de los principales trabajos relacionados

A continuación se presenta un resumen de las contribuciones y limitaciones de las metodologías para hacer análisis de Big Data y procesos de ciencia de datos vistos, algunas de ellos cubren aspectos muy parecidos entre sí y otros que no, por lo que resultan ser útiles para esta investigación con la finalidad de retomarlos y complementarlos para aproximar el objetivo principal de este trabajo. Es importante mencionar que este resumen no pretende comparar las metodologías porque puede ser el caso de que fueran planteadas para un negocio en particular por lo que aplicarla a otra organización requeriría hacerle cambios para adaptarla a las necesidades muy específicas de esa otra organización. El objetivo de este análisis es más bien determinar el grado en que éstas pueden ayudar a cubrir las necesidades planteadas para resolver el problema del trabajo de investigación tomando en consideración la elaboración de algunos estudios realizados anteriormente en el Instituto con datos de la red social de Twitter.

PROPUESTA POR (Sheikh, 2013)

CONTRIBUCIONES

Es una metodología bastante completa formada por seis fases que a su vez están compuestas de un conjunto de actividades a desarrollar secuencialmente (18 actividades) teniendo como base el modelo en cascada utilizado en el desarrollo de software. Esto mejora el entendimiento del personal como desarrolladores, analistas y arquitectos de software para participar en un proyecto de análisis de datos ya que lo pueden equiparar con las actividades que realizan cotidianamente.

Otro de aspecto importante es la consideración de auditorías y actividades de control al final de cada fase con el objetivo de garantizar que el trabajo realizado durante ese periodo cumple con lo esperado por parte de los interesados en el proyecto.

LIMITACIONES

A pesar de ser una metodología que se basa en el desarrollo de sistemas, el amplio conjunto de actividades a realizar hace que el desarrollo de una solución pueda ser

tardada de culminar. Otra parte que no se ajusta a los casos de estudio del INEGI es la necesidad de que el método sea iterativo en algunas de sus capas con la finalidad de agilizar el desarrollo de la solución y de perfeccionar la idea en caso de necesitar regresar a fases anteriores para obtener los datos esperados.

PROPUESTA POR (Mohanty et al., 2013)

CONTRIBUCIONES

Esta metodología tiene un amplio grado de detalle ya que está formada por siete etapas que van desde la detección de la situación de interés de la empresa hasta la medición y monitoreo de la solución generada. Algunas de estas etapas están altamente relacionadas ya que son ejecutadas iterativamente lo que permite modificar y expandir la solución gradualmente tanto en el tamaño de las muestras como en la aplicación de diferentes técnicas analíticas para solucionar el problema del negocio. Es importante enfatizar que tiene un gran parecido con algunas metodologías de análisis tradicionales y procesos de ciencia de datos con la diferencia de estar enfocado a trabajar con grandes cantidades de datos. Por las características que posee esta metodología se puede decir que es aplicable en muchas áreas de investigación solamente con hacerle algunos ajustes propios al problema a resolver.

LIMITACIONES

Una de las limitaciones que puede llegar a presentarse al utilizar esta metodología es la mala delimitación de cada una de las muestras pudiendo ocasionar que el número de iteraciones que se realicen en todo el proceso sea algo considerable y por consecuencia provoque que el desarrollo de la solución tarde más de lo provisto.

PROPUESTA POR (Raghupathi & Raghupathi, 2014)

CONTRIBUCIONES

A pesar de ser creada originalmente para realizar análisis de Big Data en el campo del cuidado de la salud no le pide nada a las otras dos metodologías vistas en la revisión de

literatura considerando que podría aplicarse en otro campo diferente. Una ventaja de esta metodología es que en un proyecto en el campo del cuidado de la salud la mayor parte de datos tradicionalmente son archivos estáticos de papel, películas de rayos X, y scripts, por lo que se puede decir que la metodología contempla la parte de los tratamientos necesarios para trabajar con ellos. Por otro lado, hay enfermedades que requieren constante monitoreo por lo que la velocidad con la que se obtienen los datos es mayor y en tiempo real llegando a la misma conclusión que con los datos tradicionales, tal es la importancia de trabajar con estos datos que en el campo en cuestión una decisión tomada un segundo más tarde puede ser la diferencia entre la vida y la muerte de un paciente; ejemplo de esto son múltiples mediciones diarias de glucosa en diabéticos, lecturas de presión arterial, los monitores cardíacos, electrocardiogramas, etc. Otro punto a favor de esta metodología son los ciclos de retroalimentación entre cada fase con la finalidad de minimizar el riesgo de que el proyecto no se realice satisfactoriamente.

LIMITACIONES

A pesar de ser una metodología que contempla un conjunto amplio de actividades y que al finalizar cada una de estas pasa por un proceso de retroalimentación para detectar errores es necesario, para el punto de vista del INEGI, contar con una metodología lo suficientemente interactiva para poder regresar a fases anteriores y probar los avances realizados con distintas parámetros para así perfeccionar la solución. Tampoco incluye la parte de roles de los involucrados en el proyecto para la realización de las distintas actividades y fases para encontrar la solución.

PROPUESTA POR (Mousannif et al., 2014)

CONTRIBUCIONES

Muestra una serie de pasos secuenciales que cubren el ciclo de vida de un proyecto de Big Data (como se mostró en la revisión de literatura). La principal ventaja de este trabajo es que se muestra para cada una de las capas, un conjunto de herramientas y tecnologías con las que se pueden realizar las actividades que conforman las etapas de dichas capas. Esto permite tener una idea más clara al momento de determinar que

infraestructura tecnológica (hardware y software) y que habilidades debe de tener la persona que estará trabajando en el proyecto.

LIMITACIONES

Hay una parte que no se ajusta a los casos de estudio del Instituto que es la necesidad de que el método sea iterativo en algunas de sus capas con la finalidad de perfeccionar la idea y culminar el proyecto con los datos esperados. Tampoco incluye la parte de roles de los involucrados en el proyecto para la realización de las distintas actividades y fases.

PROPUESTA POR (Shearer, 2000)

CONTRIBUCIONES

Pese a que el modelo CRISP- DM está enfocado a los procesos de minería de datos, debido a su completes diversos autores lo han tomado como referencia para proponer tanto metodologías de análisis de datos como procesos de ciencia de datos.

LIMITACIONES

A pesar de ser un modelo de minería de datos diseñado a finales de 1996 (Shearer, 2000), las 6 fases de alto nivel de CRISP-DM siguen siendo una buena descripción para el proceso de análisis, sin embargo hay algunos aspectos que necesitan ser actualizados para que pueda mantenerse y adaptarse ante los nuevos retos del Big Data y la ciencia de datos moderna. Tampoco incluye la parte de roles de los involucrados en el proyecto para la realización de las distintas actividades y fases.

PROPUESTA POR (Cuesta, 2013)

CONTRIBUCIONES

Está formada por un conjunto de fases que de manera global muestra el proceso a seguir para poder hacer predicciones del comportamiento de los datos cubriendo aspectos muy interesantes que van desde la identificación del problema hasta la

interpretación y despliegue de la solución. Del mismo modo presenta un conjunto de técnicas y herramientas para realizar distintos tipos de análisis.

LIMITACIONES

A pesar de ser proceso que contempla un conjunto amplio de actividades no cubre algunos aspectos importantes del proceso como las retroalimentaciones por parte de los interesados para verificar que el proceso está generando la información esperada y la iteratividad que este comportamiento requiere. Tampoco incluye la parte de roles de los involucrados en el proyecto para la realización de las distintas actividades y fases.

PROPUESTA POR (Guo, 2013)

CONTRIBUCIONES

Presenta las actividades involucradas en un proceso iterativo para la obtención de información que van desde la adquisición de los datos hasta la presentación de resultados llegando a tener un grado de detalle en el que se puede apreciar actividades como la ejecución de scripts y la inspección de sus salidas.

LIMITACIONES

Por ser un proceso de ciencia de datos no incluye la parte de la situación del negocio ni como se beneficia la organización con la solución del Big Data que se quiere resolver con el análisis de los datos. Tampoco incluye la parte de roles de los involucrados en el proyecto para la realización de las distintas actividades y fases.

Capítulo 3. Marco teórico

En este capítulo se comenta y se profundiza en las áreas de conocimiento que contextualizan el problema de investigación presentadas en varios apartados:

En la sección de las estadísticas oficiales se presenta una idea general de lo que son y, para que sirven, que temas abarcan, los principios que las rigen para generar información veraz y de calidad de un país y quien se encarga de generarlas.

En la siguiente sección que es llamada fuentes de datos tradicionales se mencionan que eventos originan los datos que se han utilizado durante años para generar estadísticas oficiales

La sección llamada Big Data, es en donde se explica porque surgió el término, qué es, qué características deben tener los datos para considerarlos dentro de esta categoría y la razón por la cual diversas organizaciones están apostándole para trabajarlo y obtener un beneficio para su negocio.

En la sección de tipos y fuentes de Big Data, se mencionan algunos ejemplos de las fuentes de datos a través de las cuales se puede obtener Big Data identificando y explicando si entran en la categoría de fuentes de datos estructuradas o no estructuradas.

En la sección de tecnologías base para trabajar Big Data, se enlistan las tecnologías base sobre las cuales trabaja Big Data que son los sistemas de archivos distribuidos, el marco de trabajo MapReduce y las bases de datos NOSQL.

En la sección de arquitecturas de referencia de Big Data, se muestra un conjunto de arquitecturas de referencia de algunas organizaciones privadas donde se pueden apreciar los diferentes elementos que pueden intervenir en la implementación de una solución Big Data

En la sección llamada Científico de Datos, se describen las habilidades y características que tienen las personas del área de ciencia de datos para obtener nuevas ideas al trabajar con Big Data.

En la siguiente sección, la de análisis de Big Data, se presentan algunos estudios que se han realizado utilizando diferentes fuentes de Big Data en diferentes áreas de interés con la finalidad de demostrar que es aplicable a organizaciones de cualquier giro.

Para la sección de visualización de Big Data, se describen un conjunto de características que se aconseja que tengan las presentaciones de los datos producto del análisis para que sean entendibles por las personas que tomaran decisiones.

En la sección de herramientas de código abierto para trabajar con Big Data, se presenta un conjunto de herramientas con las que se puede recolectar, almacenar y procesar Big Data.

Para la sección de fuente de Big Data: Twitter, se presenta la red social de Twitter como fuente de Big Data gracias a puso a disposición del público en general y de manera gratuita, los mensajes que publican sus usuarios mediante el uso de API desarrolladas por la misma organización. También se presentan algunos trabajos de análisis de tweets realizados por diversos investigadores.

Para la sección de trabajos anteriores de análisis de movilidad cotidiana, se presentan estudios de movilidad realizados con datos de diferentes fuentes, incluyendo Twitter para conocer el proceso que se siguió para lograr los objetivos planteados en dichas investigaciones.

En la sección trabajos anteriores de análisis de impacto de eventos de la vida real se presentan estudios realizados con los datos extraídos de Twitter para conocer el proceso que se siguió para lograr los objetivos planteados en dichas investigaciones.

1. Estadísticas Oficiales

De acuerdo con (ONU, 2013) las estadísticas oficiales son aquellas que permiten explicar de manera más clara la situación económica, demográfica, social y ambiental en la que se encuentra un país teniendo como principales interesados la sociedad y el Estado.

En 1994 la Comisión de Estadística aprobó un conjunto de principios fundamentales para estadísticas oficiales y que en el año de 2013 fue reafirmado nuevamente (ONU, 2013). Estos principios pretenden hacer hincapié en la importancia de la generación de estadísticas de calidad soportadas por marcos jurídicos e institucionales en todos los niveles políticos con la finalidad de dar la confianza necesaria que las estadísticas representan la realidad.

Las ONE son los encargados de captar, procesar y difundir información de calidad, veraz y oportuna del territorio, la población, la economía, etc. de un país (INEGI, 2014) permitiéndole tomar decisiones en tiempo real con el fin de servir a sus ciudadanos y poder lidiar con los retos nacionales que se presenten (como en el sector salud, la economía, el estado del tiempo, etc. (McKinsey & Company, 2011).

2. Fuentes de datos tradicionales

Dentro de las fuentes de datos que tradicionalmente utilizan las ONE para generar estadísticas oficiales se encuentran los censos, las encuestas y los registros administrativos(INEGI, 2014). Un censo es el conjunto de actividades realizadas para recolectar datos demográficos, sociales y económicos de toda la población de un país en un momento determinado para ser analizados y evaluados para posteriormente publicar los resultados obtenidos (INEGI, 2014)(“Censos Bolivia”, 2012). En pocas palabras es la movilización civil de mayor envergadura que encara un país, por la cantidad de recursos humanos y materiales que involucra.

Las encuestas son operaciones de recolección de datos provenientes de una muestra de la población de estudio sobre temas específicos y realizados de manera personalizada por cada país, de acuerdo a sus necesidades. Ejemplo de estas son: la encuesta anual de servicios 2013, y la encuesta anual de la Industria de la construcción 2013 elaboradas por (“IBGE :: Instituto Brasileiro de Geografia e Estatística”, 2015)

Los registros administrativos se utilizan para producir información estadística generada mediante los trámites que realizan instituciones públicas. Un ejemplo de estos se obtienen son los nacimientos, defunciones generales y fetales, matrimonios y divorcios provenientes del registro civil (INEGI, 2014).

Por lo general, las estadísticas de carácter social o económico, como son los censos y las encuestas, están sujetas a la colaboración de la población quienes son los que proporcionan información personal o de sus negocios de forma confidencial por lo que es de vital importancia la construcción de una "conciencia estadística" (Lindenboim, 2011) en la sociedad con la finalidad de que la información recabada sea de calidad y las estadísticas generadas representen efectivamente la situación del país.

3. Big Data

En los últimos años diversas organizaciones públicas y privadas han puesto gran interés en las grandes cantidades de datos que se generan de manera continua y acelerada a través de distintos sistemas y dispositivos electrónicos, muchos de los cuales están conectados a internet, con la finalidad de obtener conocimiento acerca de su situación actual dentro de un entorno de negocios que cambia de manera constante. Anteriormente, este era un problema porque se carecía de herramientas y procesos para trabajar con estas cantidades de datos en el tiempo en el que eran necesarios para tomar decisiones importantes en la organización, sin embargo con el paso del tiempo y con los nuevos avances tecnológicos se pudo mejorar notablemente los tiempos de procesamiento y análisis de datos trayendo consigo un enfoque que impulsa el desarrollo de organizaciones que toman decisiones basadas en datos. Teniendo esto como base, apareció el término Big Data y se utiliza para describir a las grandes cantidades de datos que no pueden ser procesados o analizados utilizando herramientas y procesos tradicionales.

Algo a tener en cuenta es que conforme pase el tiempo y se desarrollen nuevas tecnologías de almacenamiento y procesamiento, el Big Data será cada vez mayor, tal cual lo expresa Jim Pickerman de Microsoft Research quien lo expone de la siguiente manera: "*El Big Data de hoy será, Short Data dentro de 10 años*"(Wilson, 2014).

Diversos autores han presentado sus propias definiciones, sin embargo todas tienen algunos aspectos en común. Una de ellas es la presentada en el trabajo de (TechAmerica Foundation's Federal Big Data Commission, 2014) que lo plantea de la siguiente manera: "*Big Data es un término que describe un enorme volumen de datos complejos obtenidos a gran velocidad, que requieren técnicas y tecnologías avanzadas*

para permitir la captura, almacenamiento, distribución, gestión y análisis de la información".

La organización (IBM, 2012) lo describe como "*grandes cantidades de datos que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis aplicándose a toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales*".

Los autores (Beyer, M. A., & Laney, D., 2014) definen Big Data como "*un conjunto de activos de información generados en grandes volúmenes provenientes de diversas fuentes a grandes velocidades que requieren nuevas formas de procesamiento para permitir toma de decisiones mejoradas, descubrimiento conocimiento y optimización de procesos*".

En la Figura 8 se muestra la línea de tiempo de búsquedas en el motor de búsqueda de Google referentes a Big Data.

Figura 8. Tendencias de búsqueda en Google del término Big Data. Fuente https://www.google.com/trends

Doug Laney y Fan Wei (Fan Wei & Bifet Albert, 2013) fueron los primeros en hablar de algunas de las características necesarias en el campo de la gestión de grandes volúmenes de datos que son el volumen, la variedad y la velocidad a las que denominaron como las 3Vs:

1. Volumen: La disponibilidad de infraestructura junto con la reducción de precios de la misma ha hecho que el tamaño de los datos aumente día con día de manera masiva.
2. Variedad: Hay diferentes tipos de datos procedentes de la web y las redes sociales, de las imágenes, videos, etc. De aquí que los tipos de datos se categoricen en estructurados y no estructurados.
3. Velocidad: Los datos se generan a gran velocidad como flujos de datos teniendo como objetivo analizarlos en tiempo real.

Conforme se hizo más popular el término de Big Data, diversos autores (Fan Wei & Bifet Albert, 2013) (Chandarana, and, & Vijayalakshmi, 2014)(TechAmerica Foundation's Federal Big Data Commission, 2014) agregaron más características:

4. Veracidad: La confiabilidad y exactitud de los datos dependerá de la calidad obtenida mediante la revisión de características como la consistencia en los datos, la detección de duplicados, etc.
5. Valor: El valor de negocio que representa a la organización una ventaja competitiva.

De acuerdo con (IBM, 2012) la información que proviene principalmente de los dispositivos conectados a Internet (como tabletas, teléfonos inteligentes, laptops, etc.) tendría una tasa de crecimiento anual del 78% entre los años de 2011 a 2016 estimando que dichos dispositivos excederían para este último año el número de habitantes en el planeta. De acuerdo a las Naciones Unidas (IBM, 2012), la población mundial alcanzará los 7.5 miles de millones en 2016 teniendo aproximadamente 18.9 mil millones de dispositivos conectados a Internet. Con toda esta cantidad de información y con las herramientas y técnicas necesarias es posible encontrar tendencias ocultas que anteriormente podrían detectarse con procesos muy costosos.

Según (Sánchez J. M., 2013), los expertos confirman que las empresas estarán inmersas en un futuro en el mundo de Big Data teniendo el reto de almacenar, buscar, compartir y agregar valor a los datos que hasta la fecha son inaccesibles. Se han realizado encuestas dirigidas a distintos sectores donde se reafirman estas conclusiones (CIO, 2012).

4. Tipos y fuentes de Big Data

Una de las principales características de Big Data es la capacidad de integrar datos estructurados y no estructurados provenientes de distintas fuentes.

Los datos estructurados representan aproximadamente del 15 al 20 por ciento de la información existente (TechAmerica Foundation's Federal Big Data Commission, 2014) (Hurwitz, Nugent, Halper, & Kaufman, 2013) teniendo como principal característica un tamaño y un formato definido, es principalmente depositada en bases de datos relacionales (RDBMS de las siglas de Relational Data Base Management System) utilizando comúnmente el lenguaje de consultas estructurado (SQL por las siglas en inglés Structured Query Language) para poder ser almacenados, consultados y analizados. Como fuentes de este tipo de información se encuentra la generada de manera automática por computadoras y la generada por humanos(Hurwitz et al., 2013). Ejemplo de estas son: para la generada por computadoras, el uso de sensores (RFID, GPS) que transmiten datos de ubicaciones de diferentes dispositivos y los logs de información web generada por servidores y aplicaciones para dejar registro de su operación; para la generada por humanos se encuentra la información capturada en formularios y el flujo de clics en un sitio web.

La información no estructurada, equivalente aproximadamente del 80 al 85 por ciento (Hurwitz et al., 2013) es aquella que no sigue un formato específico por lo que es más complicada de analizar. De la misma forma en la que se puede categorizar la generación de información estructurada, la información no estructurada presenta las mismas características (Hurwitz et al., 2013). Ejemplo de la generada por computadoras se pueden encontrar las imágenes satelitales, imágenes sísmicas, información atmosférica, etc.; para la generada por humanos se encuentra la información capturada en las redes sociales, correo electrónico, videos, blogs, foros, etcétera (TechAmerica Foundation's Federal Big Data Commission, 2014).

La disposición de las nuevas fuentes de datos está abriendo un sin número de posibilidades para obtener conocimiento sobre temas específicos de una población determinada, como pueden ser estudios de sentimiento de opinión pública, de movilidad humana, de salud pública, etcétera. Un aspecto que se debe considerar con estas fuentes de datos es que la velocidad con la que se generan es tan alta que ha hecho que el

desarrollo de proyectos se realice de forma ágil e interactiva para poder tomar decisiones acertadas en el momento que se requiera. Big Data se ajusta a este proceso donde se utilizan ciclos de tiempo cortos con resultados rápidos y la participación constante del usuario para hacer entregas graduales a una solución de negocio (Hurwitz et al., 2013).

5. Tecnologías base para trabajar Big Data

Hasta la fecha se han creado un sin número de tecnologías para trabajar en las fases de agregación, manipulación, administración, análisis y visualización para encontrar soluciones a problemáticas que solo con Big Data se puede resolver. Tecnologías como los sistemas de archivos distribuidos, el marco de trabajo MapReduce, bases de datos NOSQL, procesamiento masivamente paralelo (MPP por las siglas en inglés de Massive Parallel Processing), computación en la nube y la minería de datos son algunas de las utilizadas (Patel, Birla, & Nair, 2012).

Sistema de archivos distribuido

Un sistema de archivos distribuidos está formado por un conjunto de computadoras interconectadas entre sí y su propósito es el almacenamiento de datos. El proyecto Hadoop (https://hadoop.apache.org/), un software de código abierto para cómputo distribuido confiable y escalable, tiene uno de éstos. El sistema de archivos distribuido de Hadoop (HDFS por el acrónimo en inglés de Hadoop File System) es la implementación de código abierto del sistema de archivos de Google (GFS de las siglas en ingles de Google File System) y es el sistema de almacenamiento primario utilizado por las aplicaciones de Hadoop. Tiene la capacidad de manejar archivos muy grandes (cientos de megabytes, gigabytes o terabytes de tamaño), dar acceso al streaming de datos mediante la idea de que el patrón de procesamiento más eficiente es que se escriban una vez y que se consulte tantas veces sea necesario, es tolerable a fallas en hardware mediante la implementación de réplicas de los datos a nivel de nodos y a nivel de racks, entre otras.

HDFS ha sido diseñado con una arquitectura maestro / clúster de esclavos y consiste en un nodo nombre (NN por las siglas en inglés de Name Node) y una serie de nodos de datos (DN por las siglas en inglés de Data Node).

El NN se conforma de dos elementos principales que exponen el sistema de archivos a los usuarios y permite la recuperación del almacenamiento de datos. Estos elementos son: el sistema de archivos de nombres de espacio (FSN por su siglas en inglés de File Space Name) y el sistema de archivos de metadatos (FSM por su siglas en inglés de File System Metadata).

- El FSN se encarga de llevar un mapeo de las ubicaciones físicas de los diferentes DN donde se almacenan cada uno de los bloques en los que es dividido un archivo. También soporta la organización de archivos jerárquica tradicional al igual que otros sistemas de archivos con operaciones básicas de archivos (crear, borrar, mover o cambiar el nombre).
- FSM. HDFS utiliza un registro de transacciones llamado EditLog depositado en el NN para grabar persistentemente en el FSM cada cambio que se produce. La creación de un archivo nuevo en el HDFS provoca que el NN inserte un registro en el EditLog, por mencionar un ejemplo. Además del EditLog, también existe un archivo llamado FsImage el cual almacena todo el FSN, la asignación de bloques a archivos y las propiedades del sistema de archivos. Tanto el FsImage EditLog y forman parte del FSM.

El DN de HDFS almacena los datos en archivos en su propio sistema de archivos local en bloques de tamaño configurables que normalmente son de 128 MB (Hurwitz et al., 2013). Almacena cada bloque de datos en un archivo separado mediante el uso de heurísticas para determinar el número óptimo de archivos por directorio además de crear subdirectorios si fuera necesario. Cuando el DN se inicia, envía un Blockreport NN, que es una lista de todos los bloques de datos HDFS ubicado en su sistema de archivos local. Cada bloque replica a través de varios DN para crear un sistema de archivos distribuido que serán fiables si se produce un fallo de DN.

En la Figura 9 se muestran los elementos y la forma en la que interactúan los elementos de la arquitectura del sistema de archivos distribuido de Hadoop.

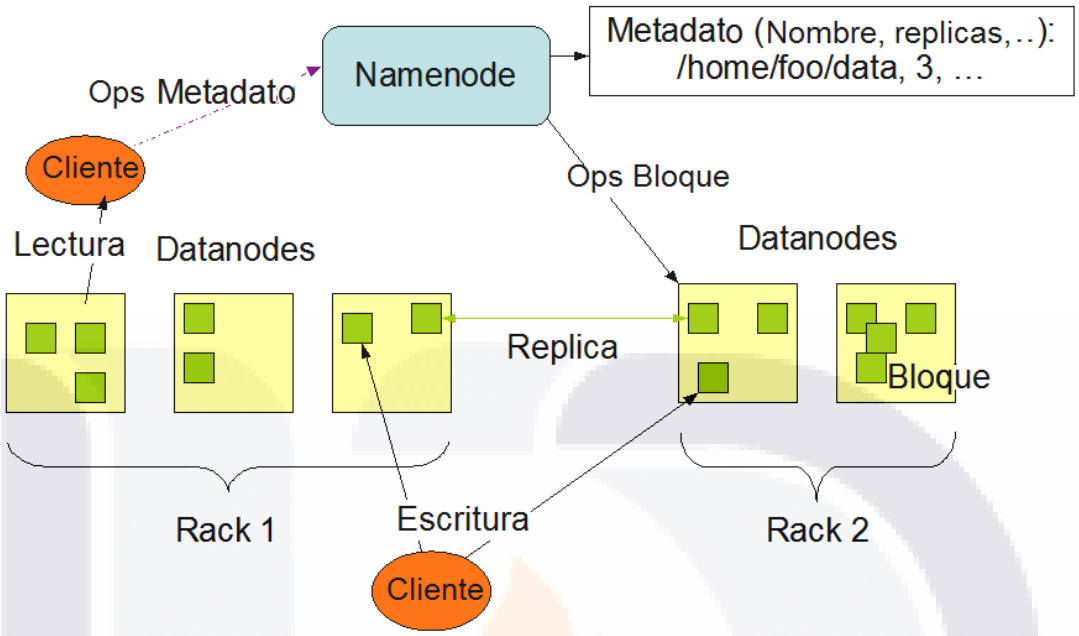


Figura 9. Arquitectura de HDFS fuente ("Apache Hadoop 2.7.1 – HDFS Architecture", 2015)

Marco de trabajo MapReduce.

Otra de las tecnologías involucradas es MapReduce, el modelo de programación más utilizados por Big Data para el procesamiento de grandes cantidades de información (estructurada y no estructurada) de manera paralela en un sistema de cómputo distribuido. Inicialmente utilizado por Google para el indexado de sitios Web para el servicio de búsquedas (Dean & Ghemawat, 2008), se basa en dos conceptos utilizados en lenguajes de programación funcional: Mapear y Reducir.

La arquitectura física de MapReduce está compuesta por un equipo maestro y un grupo de trabajadores, y su funcionamiento interno está basado en las siguientes operaciones (Hameurlain, Rahayu, Taniar, 2012):

- Las entradas están formadas por grupos de pares [clave, valor] y son procesadas utilizando funciones para mapear definidas por el usuario para generar un grupo intermedio de pares [clave, valor] que posteriormente son procesados por la

función de reducción que mezcla de manera ordenada los valores de los grupos por claves.

- Una vez hechos los grupos, las tareas (de mapeo o reducción) son repartidas por el equipo maestro entre los trabajadores que están desocupados.
- Las salidas de las tareas de mapeo se dividen en tantas particiones como tareas de reducción existan. Las entradas con la misma clave serán asignadas a la misma partición para garantizar su correcta ejecución.
- Después de esto los pares [clave, valor] de las particiones se ordenan y clasifican, y se envían a los trabajadores para su la ejecución de la tarea correspondiente.
- Se obtiene la salida final del cómputo de los datos.

Bases de datos NOSQL

Las bases de datos NOSQL o no relacionales no dependen del modelo tabla/llave ni utilizan SQL (de aquí su nombre NOSQL, not only SQL) como los sistemas manejadores de base de datos relacionales (RDBMS por las siglas de Relational Data Base Management System).

Algunas de las características que poseen las bases de datos NOSQL son:

- Escalabilidad, la base de datos tiene la capacidad de expandirse o contraerse de manera transparente al usuario tomando como base los flujos de datos.
- Diseño de persistencia, al igual que los sistemas RDBMS y al estar trabajando con datos que se generan a grandes velocidades y en grandes cantidades la persistencia juega un papel muy importante en las bases de datos NOSQL.
- Diversidad de interfaces, cuentan con una gran variedad de mecanismos de conexión para los programadores (la mayoría utiliza API restful).

Actualmente en el mercado es posible encontrar distintos tipos de bases de datos NOSQL.

- Bases de datos de documentos. Almacenan cualquier tipo de documento que contenga un grupo de llaves y un valor asignado a estos. La mayoría de las bases de datos de este tipo almacena documentos en formato JSON. CoouchDB,

MongoDB, Elasticsearch son bases de datos NOSQL que utilizan este modelo de almacenamiento.

- Bases de datos columnares. Estas bases de datos ofrecen gran flexibilidad, rendimiento y escalabilidad ya que los datos son almacenados en filas identificadas por una llave y son el equivalente a una tabla relacional. En este tipo de base de datos se pueden encontrar HBase y Cassandra.
- Bases de datos geográficas. Los elementos básicos de estas bases de datos son puntos, líneas y polígonos que representan un área geográficamente definida; son comúnmente utilizadas para el desarrollo de sistemas de información geográficos. Postgis es un plugin de Postgresql e incluye las características antes mencionadas.

6. Arquitecturas de referencia de Big Data

Una arquitectura de referencia es un diseño conceptual que muestra la forma en la que todos los elementos (componentes de hardware y software, patrones, principios, mejores prácticas, etc.) que conforman una solución se comunican entre sí de manera clara y ordenada, además de que sirven como plantillas para guiar a las organizaciones en la creación de arquitecturas propias que se ajusten al contexto del negocio y de las tecnologías de información de la misma. Emprender el camino de trabajar con Big Data sin tener en cuenta alguna arquitectura hace más complicada la obtención de información de valor y por ende la toma de decisiones. Según el artículo publicado por (Olavsrud Thor, 2014), Richard Daley, uno de los fundadores y director de estrategia de Pentaho dice que hay miles de arquitecturas de referencia de Big Data las cuales empezarían fusionarse para 2014. A continuación se presentan algunas arquitecturas encontradas en la literatura de Big Data

Arquitectura de Sawant y Shah

La arquitectura para la solución de aplicaciones de Big Data presentada por (Sawant & Shah, 2013) identifica los componentes que deben considerarse en la infraestructura tecnológica de la organización donde hacen hincapié en que los elementos que forman la

arquitectura pueden ser de código abierto o bajo licencia con la finalidad de sacarle provecho a todas sus bondades.

Esta arquitectura está formada por nueve capas, las cuales son:

1. Capa de las fuentes de datos. En esta capa se encuentran las fuentes de datos estructuradas y/o no estructuradas, internas y/o externas a la organización disponibles para trabajar con Big Data.
2. Capa de consumo. Esta capa tiene la responsabilidad de separar la información útil de la que no es relevante para el análisis, que es aproximadamente en una proporción del diez al noventa por ciento de los datos, mediante procesos de validación, limpieza, transformación, reducción e integración de los datos para poder almacenarla y utilizarla para su análisis.
3. Capa de almacenamiento distribuido (Hadoop). Como su nombre lo dice, la capa de almacenamiento distribuido es la encargada de almacenar los datos que se generan en grandes velocidades. Una de sus ventajas es la capacidad de prever la falta de disponibilidad de los datos debido a que son almacenados de manera redundante para agilizar su procesamiento.
4. Capa de infraestructura de sistemas de archivos distribuidos. Es la capa física que da soporte a la capa de almacenamiento. La capa de infraestructura física de sistema de archivos distribuidos está basada en un modelo computacional por lo que los datos son almacenados a través de la red.
5. Capa de gestión de la plataforma de sistema de archivos distribuidos. Es la capa que tiene la función de brindar el acceso a las bases de datos NOSQL utilizando el sistema de archivos distribuido.
6. Capa de seguridad. La capa de seguridad juega un rol importante en la arquitectura, el hecho de tener un nodo de trabajo no confiable generará resultados no confiables. Otro aspecto en el que se debe tener en consideración es que las tecnologías recientes se convierten en blancos para hackers por lo que es necesario contar con reglas de seguridad como la autenticación de nodos, la encriptación de archivos, el establecimiento de una comunicación segura entre los nodos utilizando secure socket layers SSL, TLS, etc.
7. Capa de monitoreo. Es la capa encargada de proporcionar información sobre el estado en el que se encuentran los componentes de la arquitectura.

8. Capa del motor de análisis. Algunos tipos de análisis podrán requerir de herramientas tradicionales para poder llevarse a cabo mientras que otros necesitaran métodos más sofisticados.
9. Capa de visualización. Es la capa utilizada por analistas y científicos de datos en la que pueden obtener información de interés mediante la revisión de modelos visuales.

En la Figura 10 se puede observar la arquitectura presentada por (Sawant & Shah, 2013) en la que se muestran las capas y las tecnologías que podrían utilizarse para implementar cada una de ellas.

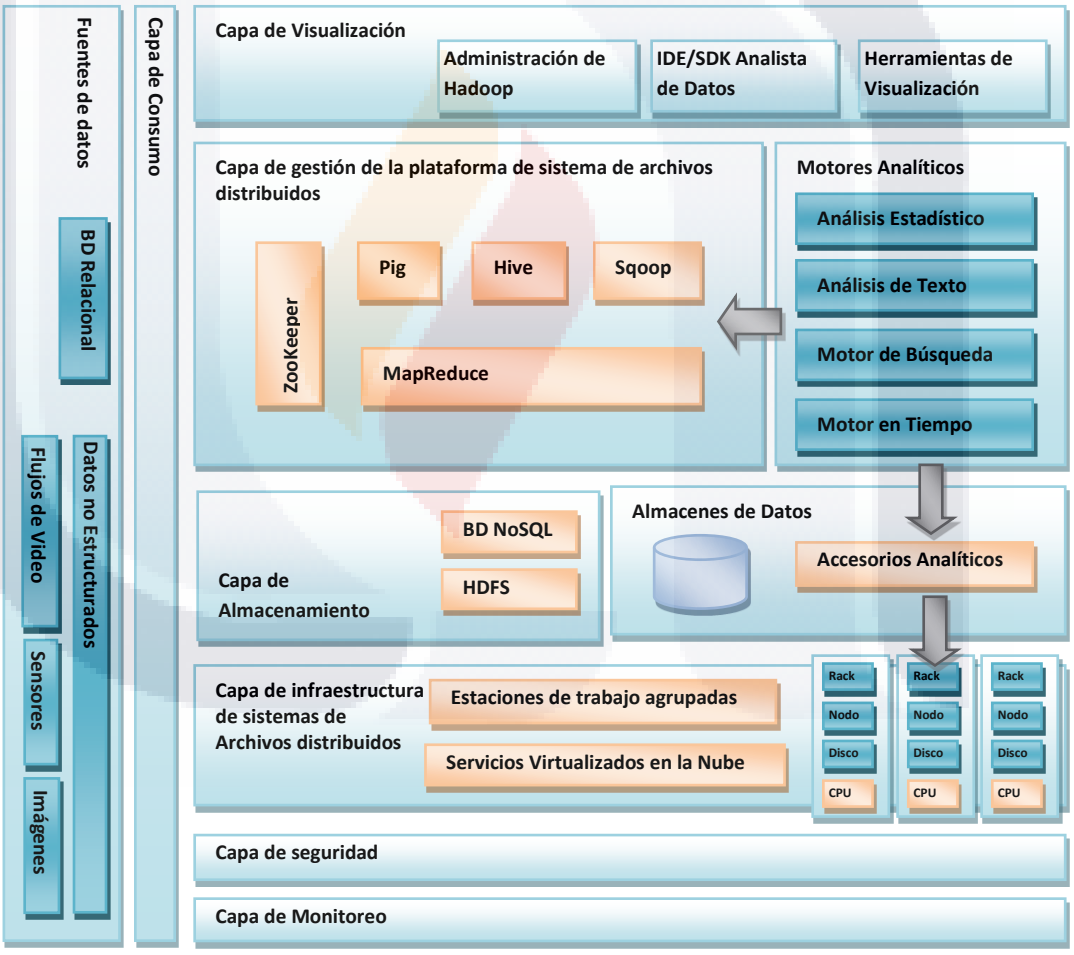


Figura 10. Arquitectura de Big Data elaborada por (Sawant & Shah, 2013)

De igual manera, la iniciativa privada que se dedica a la generación de hardware y software para trabajar con Big Data ha implementado y dado a conocer las arquitecturas con las que trabajan con el fin de promover sus componentes (Oracle, 2015), (Microsoft, 2015), (IBM, 2013).

Arquitectura de IBM

La arquitectura definida por (IBM, 2013) está formada por cuatro capas lógicas y cuatro capas denominadas verticales. Las primeras tienen el objetivo de proveer una manera de organizar los componentes que integran la solución de Big Data, mientras que las capas verticales hacen referencia a aspectos que afectan a los componentes de las capas lógicas.

Las capas lógicas son:

1. Capa de fuentes de Big Data. Esta capa está formada por los tipos de Big Data disponibles (por su formato, la velocidad con la que se generan, el volumen, etc.) por lo que es necesario determinar cuáles de todos ellos serán útiles para las necesidades de análisis de la organización.
2. Capa de preparación de los datos y almacenamiento. Además de ser la capa responsable de la adquisición y almacenamiento de los datos provenientes de las distintas fuentes de Big Data, se encarga de convertirlos, si es necesario, a un formato adaptable a cómo los datos se van a analizar.
3. Capa de análisis. Es la capa que utiliza los datos almacenados en la capa anterior para analizarlos y generar el conocimiento que el negocio espera obtener.
4. Capa de consumo. Esta capa utiliza la salida proporcionada por la capa de análisis para obtener ideas de los datos mediante el uso de aplicaciones de visualización, de monitoreo en tiempo real, motores de recomendaciones, la mejora de un proceso de negocio o un servicio de la organización.

Las capas verticales son:

1. Integración de la información. Es la capa encargada de la conexión entre las diversas fuentes de Big Data las cuales requieren de conectores, aceleradores,

adaptadores, API, etcétera que se utilizaran dependiendo del origen de los datos a analizar.

2. El gobierno de Big Data. El gobierno de Big Data tiene el objetivo de definir un conjunto de reglas que ayudan a las empresas a tomar las decisiones correctas sobre los datos que permitan guiar procesos sólidos para supervisar, estructurar, almacenar y proteger los datos desde el momento en que entran a la organización.
3. Gestión de sistemas. La gestión de sistemas es fundamental para una solución de Big Data porque de esta depende que se cumplan los objetivos garantizando que ningún componente de las capas lógicas falle. En esta capa se puede encontrar actividades como la gestión de logs de los sistemas, máquinas virtuales, aplicaciones y otros dispositivos; el monitoreo de alertas y notificaciones en tiempo real; el uso de dashboard en tiempo real que muestra varios parámetros actuales de los sistemas; la administración de la capacidad de almacenamiento, etc.
4. Calidad del servicio. Esta capa es responsable de definir criterios para determinar la calidad de datos, políticas de privacidad y seguridad de los datos, la frecuencia con la que se actualizan los datos, y los filtros que se utilizarán para limpiar los datos.

La Figura 11 muestra los elementos que conforman la arquitectura desarrollada por IBM.

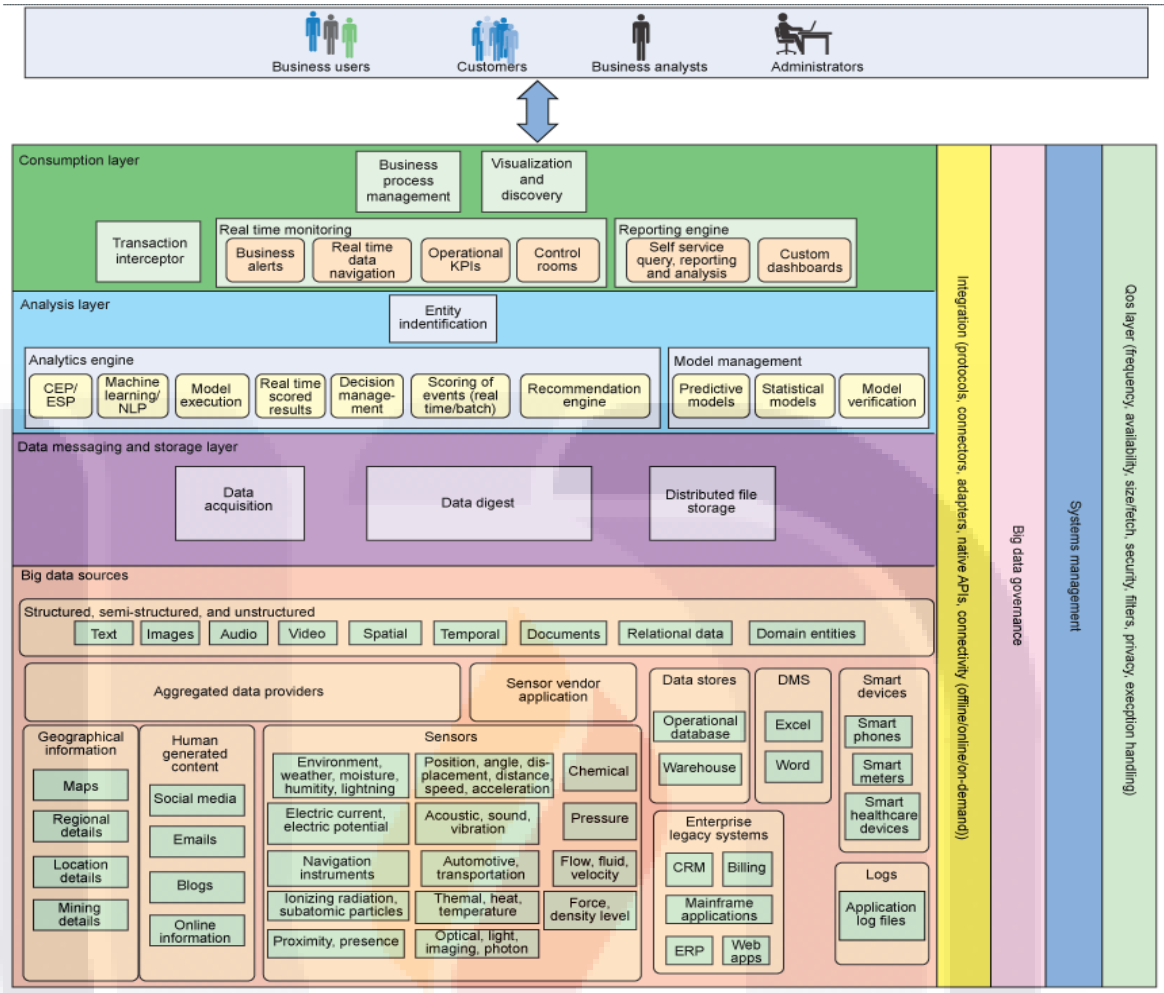


Figura 11. Arquitectura de Big Data desarrollada por IBM (IBM, 2013)

Arquitectura de Microsoft

Microsoft es una de las organizaciones que está desarrollando herramientas para la generación de soluciones de Big Data entre las que se encuentra el HDInsight una solución basada en el marco de trabajo de Apache Hadoop que forma parte de la línea de productos Microsoft Business Intelligence (BI) and Analytics (Microsoft, 2015).

La arquitectura está formada por cinco capas que son:

1. Capa de orígenes de datos. Igual que en la capa fuentes de Big Data de la arquitectura anterior, en esta capa se encuentran las fuentes de la cual se obtendrá la información.

2. Capa de integración. Es la capa que se encarga de la recolección, limpieza y transformación de los datos utilizando distintas herramientas ofrecidas por Microsoft.
3. Capa de almacenes de datos. Además de ser la capa responsable del almacenamiento de los datos y el análisis en paralelo de los mismos.
4. Capa de modelado de datos y análisis. En esta capa se trabaja con componentes que permiten modelar los datos para soportar análisis de inteligencia de negocio (BI). También se tiene integrado la posibilidad de realizar módulos tabulares y multidimensionales además de la minería de datos.
5. Capa de visualización y reportes. La presentación de los datos se realiza en la capa de visualización con la finalidad de facilitar el análisis de los datos a los usuarios finales mediante el uso de algunos servicios y herramientas.

La Figura 12 muestra la visión general de la plataforma de datos Microsoft con las herramientas a utilizar en cada una de las capas.

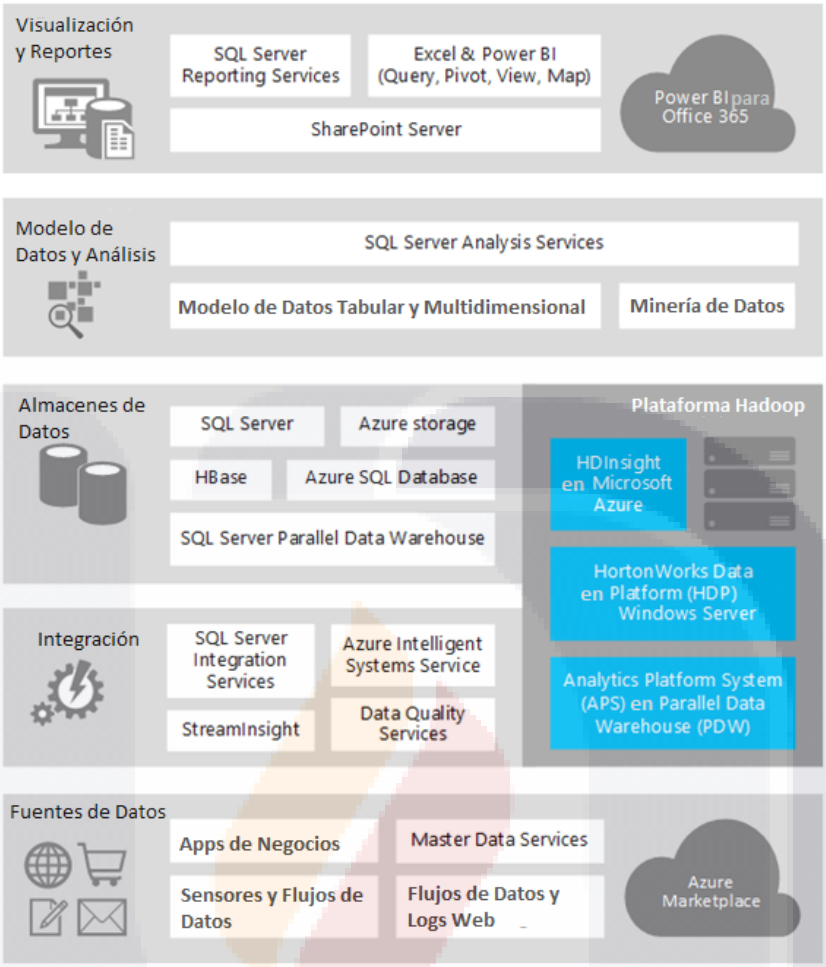


Figura 12. Visión general de la plataforma de datos Microsoft (Microsoft, 2015)

A pesar de existir diferencias entre las tres arquitecturas presentadas debido al grado de detalle que algunas tienen es posible detectar ciertas similitudes entre ellas mediante su análisis. A simple vista se puede observar que la arquitectura realizada por (Sawant & Shah, 2013), a diferencia de la arquitectura de IBM y Microsoft, presenta capas para la infraestructura y la gestión de sistemas de archivos distribuidos además de contar con dos capas que dan apoyo al resto de las capas.

La arquitectura de IBM presenta en una capa la parte de preparación y almacenamiento de los datos, actividad que se realiza en capas distintas en las otras dos arquitecturas. Además de esta característica, también cuenta con cuatro capas de apoyo que dan mayor grado de formalidad a sus soluciones al garantizar la calidad del servicio,

el gobierno, la gestión de los sistemas y la integración de la información entre todas las capas.

Por último y no por eso la menos importante esta la arquitectura de Microsoft que contiene los elementos necesarios para generar una solución de Big Data. Está compuesta por cinco capas y para cada una de ellas la empresa ha desarrollado herramientas para facilitar su implementación.

En la Tabla 2 se puede apreciar con mayor facilidad las diferencias y similitudes entre las arquitecturas.

Tabla 2. Diferencias y similitudes entre las arquitecturas estudiadas. Elaboración propia.

	Arquitecturas		
	Sawant y Shah(2013)	**IBM (2013)**	**Microsoft (2015)**
Capas lógicas	Capa de fuentes de datos	Capa de fuentes de datos	Capa de orígenes de datos
	Capa de consumo de datos	Capa de preparación de los datos y almacenamiento	Capa de integración
	Capa de almacenamiento distribuido		Capa de almacenamiento de datos
	Capa de infraestructura de Hadoop	-	-
	Capa de gestión de Hadoop	-	-
	Capa del motor de análisis	Capa de análisis	Capa de modelado y análisis
	Capa de visualización	Capa de consumo	Capa de visualización y reportes
Capas de apoyo	Capa de seguridad	Capa de calidad de servicio	-
	Capa de Monitoreo	Capa de Gobierno	-
	-	Capa de gestión de sistemas	-
	-	Capa de Integración de información	-

7. Científico de Datos

Algunos de los obstáculos a los que se puede enfrentar una organización para el análisis de Big Data, además de contar con el hardware y software necesario, es encontrar al personal que tenga las habilidades necesarias para entender el comportamiento de los datos para extraer información útil. De acuerdo con (Patil, 2011), algunas de las características que deben tener estas personas son:

- Experiencia técnica, es necesario que tenga antecedentes en diferentes campos como la informática, las matemáticas y la estadística; junto con un amplio conocimiento del negocio para que, una vez recolectados los datos, pueda utilizarlos y aplicarle distintos procesos para obtener el máximo valor de los datos al analizarlos (Tascón, 2013).
- Curiosidad, resolviendo distintas cuestiones que les permiten hacer múltiples descubrimientos al analizar los datos.
- Inteligente, es la capacidad de poder resolver un problema mediante diferentes formas creativas
- Habilidad de comunicación, no solo en el ámbito de la tecnología, sino también en el campo de negocios para que un directivo pueda comprender fácilmente los resultados obtenidos del análisis ofreciendo a las organizaciones una ventaja competitiva (Davenport & Patil, 2012).

Esta profesión recibe el nombre de científico de datos (término introducido por dos de los científicos de datos originales, DJ Patil y Jeff Hammerbacher, cuando estaban trabajando en LinkedIn y Facebook (Krishnan, 2013)). Entre las actividades que realizan se encuentran las necesarias para extraer datos desde diferentes fuentes y trabajarlos como un conjunto; transformarlos y organizarlos para darle sentido a los datos sin importar las restricciones de hardware y software; aplicar técnicas de minería para analizarlos y encontrar conocimientos y relaciones inesperadas entre ellos; identificar las técnicas de visualización que mejor se ajusten a las características de los datos para facilitar su transmisión a los involucrados en el negocio (Liu et al., 2009) (Davenport & Patil, 2012).

El trabajo de científico de datos es considerado como un trabajo prometedor debido a la escasez de personas con este perfil y a la necesidad que tienen las empresas no sólo de conocer su entorno sino también de conocerse a sí mismas mediante el análisis de Big Data, esta es una de las razones por las que estas personas son bien remuneradas económicamente y solo unas cuantas grandes organizaciones como Twitter, Facebook, Amazon, Google (por mencionar algunas) pueden pagar sus salarios.

Uno de los cuidados que debe tener en consideración un científico de datos, cuando esté trabajando en una organización, es mantenerse actualizado y manejar un conjunto

de herramientas que estén dentro del estado del arte con la finalidad de proveer mejores resultados.

En una encuesta realizada por (Harris, Murphy, & Vaisman, 2013) a través del internet a científicos de datos se encontró que hay distinción entre ellos de acuerdo a sus habilidades, experiencias, nivel de educación, auto identificación y su presencia en la web identificando cuatro grupos diferentes:

- Desarrollador de datos, que es el que se encarga de resolver problemas técnicos y gestionar los datos (recolectar, almacenar y aprender de los datos);
- Investigador de datos, tiene la función de entender procesos complicados sin importar el tipo de negocio gracias a sus conocimientos académicos y realizar análisis estadísticos con los datos;
- Innovador de datos, Al igual que el investigador de datos, cuenta con un amplio conocimiento académico y manejan un amplio repertorio de herramientas y tecnologías para resolver problemas de análisis de datos de la mejor manera.
- Persona de negocios de datos, son aquellos que se enfocan en la organización y como obtener ganancia de los proyectos de datos.

Del mismo modo se clasificaron las habilidades y se agruparon por conocimiento que ellos mismos consideraron importantes para realizar su trabajo, teniendo las siguientes categorías:

- Habilidades de negocio. Conocimiento en desarrollo de productos y conocimiento del negocio.
- Habilidades de machine learning / Big Data. Conocimiento en datos estructurados y no estructurados, técnicas de machine learning y conocimiento en Big Data y Distributed Data.
- Habilidades matemáticas/ investigación de operaciones. Conocimiento en optimización, matemáticas, modelos gráficos, algoritmos y simulación.
- Habilidades de programación. Conocimiento en administración de sistemas y programación en back end y front end.
- Habilidades estadísticas. Conocimiento en técnicas de visualización, estadísticas temporales, encuestas, análisis espacial, manipulación de datos, etc.

Y finalmente empataron la clasificación de científicos de datos con las habilidades y conocimientos obtenidos mediante la elaboración de un gráfico (Figura 13) de mosaico donde se muestra la proporción donde quedaron los encuestados de acuerdo a sus respuestas.

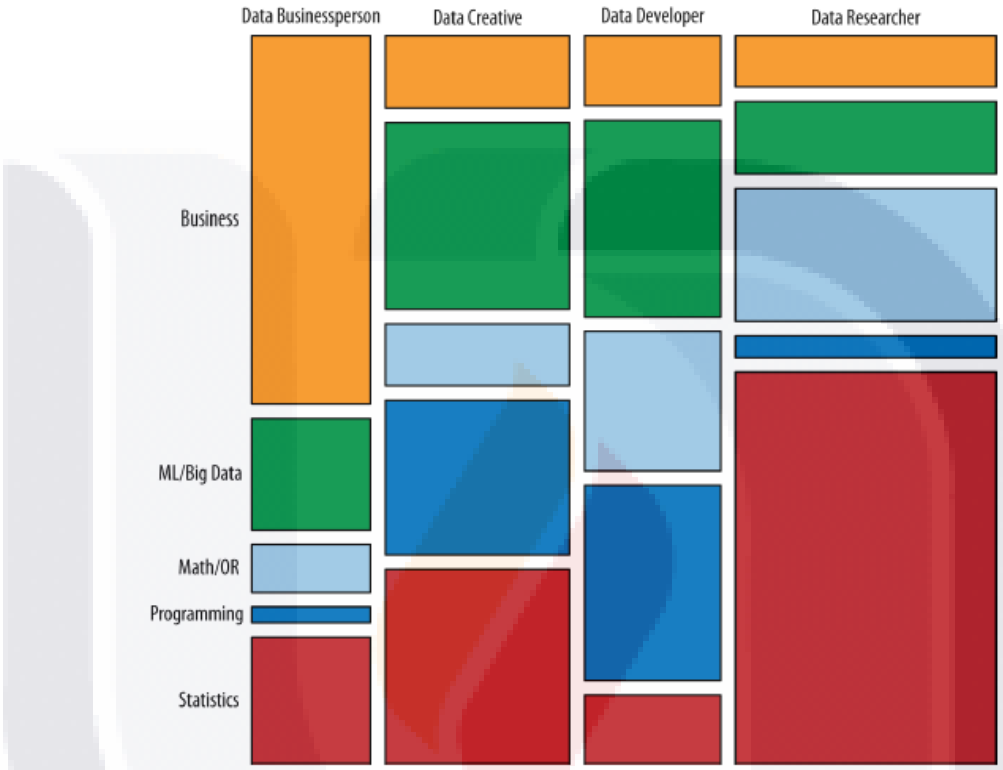


Figura 13. Relación de la clasificación de científicos de datos y las habilidades que realizan. Fuente (Harris et al., 2013)

De una forma similar a la que presenta Harris y colegas respecto a las categorías de los científicos de datos, se encuentra el trabajo realizado por (Saltz & Shamshurin, 2015) donde explora el proceso de desarrollar un proyecto de ciencia de datos identificando los roles involucrados en el mismo:

- Grupo de Gerentes. Su principal interés es determinar si el proyecto de ciencia de datos puede hacerle frente a los retos de la organización.
- Equipo de ciencia de los datos: Prepara y analiza los datos utilizando métodos de minería y técnicas de visualización de datos.

- Equipo de operaciones de datos: Se encarga de la preparación y transformación de los datos para su análisis.
- Equipo de desarrollo de software: Desarrollan de herramientas de software para ayudar a que el equipo de ciencia de datos realice el análisis de datos;
- Clientes. Son aquellos en los que están inspirados los proyectos de ciencia de datos.

Independientemente de la categoría en la que se encuentre un científico de datos, se estima que para 2018 la demanda de personas con este perfil podría superar la oferta que se produce actualmente por 140,000 y 190,000 puestos (McKinsey & Company, 2011). En respuesta a esta problemática, varias universidades a nivel mundial como la universidad de Berkeley (con la Maestría en Ciencias de la Información y Datos en línea) y la universidad de Georgia (con la Maestría en Ciencias en Analítica) están planeando lanzar programas de ciencia de datos o modificar los programas existentes en estadísticas añadiendo cursos y ejercicios de Big Data (Davenport & Patil, 2012).

8. Análisis de Big Data

Uno de los principales desafíos a los que se enfrentan las organizaciones es poder obtener conocimiento de los datos a los que tienen acceso buscando la respuesta a la pregunta “¿qué es lo que dicen los datos?” para así obtener una ventaja competitiva. El análisis de los datos es la tarea que hace frente a esta situación y consiste en obtener conclusiones basadas en la revisión de un conjunto de datos.

Detección de datos válidos. Este punto es importante ya que la información que se publica en blogs, noticias, mensajes, redes sociales, etcétera puede verse influenciada por actores o factores (como la percepción de la gente, el dialecto, el sarcasmo, la ironía, etc.) provocando que los datos recolectados sean inexactos y engañosos.

Diversos investigadores han trabajado dentro del campo de Big Data generando una amplia variedad de tipos de análisis, dependiendo del alcance y las necesidades de información, .es posible encontrar análisis de movilidad humana, análisis de impacto de

eventos, análisis predictivos, análisis de sentimientos mediante la minería de opiniones, prevención y detección de fraudes, análisis de amenazas, análisis de migración, etc.

Gracias a la capacidad que tienen diversos dispositivos de poder identificar las coordenadas geográficas de su ubicación mediante el sistema de posicionamiento global (GPS por las siglas en ingles de Global Positioning System) es posible realizar exploraciones de flujos de migración entre distintos países, la actividad turística y social de una ciudad determinada, etcétera. Este tipo de estudios reciben el nombre de movilidad humana en los que dependiendo del interés propio de la investigación se pueden realizar sobre áreas geográficas específicas y así detectar patrones de movimiento (Naaman, 2011) (Gabrielli Lorenzo et al., 2014) mediante la identificación y reconstrucción de los movimientos que sigue un usuario en particular en un espacio y tiempo determinado. Anteriormente este tipo de investigaciones se realizaban indirectamente con la información que generaban las ONE mediante los levantamientos censales, tomando a consideración que los resultados arrojaban información con un retraso ocasionado por las fechas en las que se obtuvieron las estadísticas oficiales de las que se sostenían (Zagheni et al., 2014).

Con el análisis del impacto de eventos es posible conocer como los hechos que ocurren en el día a día pueden llegar a afectar la percepción de los mismos en la gente. Esto se hace gracias a la facilidad con la que es posible dar a conocer cualquier evento que ocurre en un lugar y en un tiempo determinado por de agencias de noticias y promotores de eventos quienes han encontrado en las redes sociales una plataforma de distribución y publicación de información.

El análisis predictivo podría ser definido como la combinación del análisis tradicional con técnicas de minería de datos (como el machine learning, las redes neuronales y el análisis de texto por mencionar algunas) que permiten modelar y predecir el comportamiento de las entidades que se están estudiando (Krishnan, 2013) (Hurwitz et al., 2013) yendo más allá de responder preguntas como ¿qué paso?, ¿cuándo paso? y ¿qué impacto tuvo? llegando a obtener estimaciones realistas de las condiciones en las que se desenvolverán en el futuro de una organización y así poder tomar decisiones que permitan maximizar las oportunidades o minimizar los riesgos detectados en las predicciones (Bari Anasse, Chaouchi Mohamed, & Jung Tommy, 2014). Un ejemplo claro

de una organización experta en realizar análisis predictivo con los datos que recolecta mediante sus distintas aplicaciones es Amazon, la cual a través de revisar el historial de compras de un cliente puede detectar patrones y hacer sugerencias de productos que le pueden ser de interés para el cliente (Hurwitz et al., 2013).

La minería de opiniones es un problema muy acotado del procesamiento de lenguaje natural (NPL por las siglas en inglés de Natural Process Language) donde el sistema necesita identificar si una sentencia presenta expresiones con sentimientos positivos o negativos dando brecha para hacer progresos tangibles en todos los frentes del NPL y tener un alto impacto práctico(Cambria Erick, Schuller Björn, Xia Yunqing, & Havasi Catherine, 2013). Tradicionalmente se realizaba esta actividad haciendo encuestas a los clientes por cada producto o función, en la que se tenía que depender completamente de la buena voluntad de las personas para elaborar la encuesta. Este método se empezó a reemplazar mediante la recopilación de información de forma automática desde la World Wide Web donde se puede hallar una variedad de fuentes de datos a través de distintas plataformas como Twitter y Facebook donde se puede encontrar toneladas de información que reflejan las opiniones y actitudes de las personas, los cuales son publicados y compartidos entre los usuarios todos los días en tiempo real.

A pesar de las grandes ventajas que obtienen las organizaciones financieras con la adopción de tecnologías de la información y de telecomunicaciones recientes están propensas a la realización de actividades ilícitas por parte de usuarios malintencionados que logran infiltrarse rompiendo distintas normas de seguridad haciendo que las organizaciones tengan perdidas de millones de dólares anuales (Cardenas, Manadhata, & Rajan, 2013). Una de las áreas propensas a este tipo de ataques son las transacciones realizadas en línea. Un ejemplo de estas son las realizadas por sitios web que siguen un modelo de negocio (B2C por el acrónimo de Bussiness to Customer) y C2C (por el acrónimo de Customer to Customer) las cuales consisten en tener un grupo de clientes que hacen sus compras directamente en internet con un proveedor o con otro cliente teniendo como casos más comunes el robo de cuentas de usuario o el robo de las cuentas de pagos (Yang, Hu, Cheng, Miao, & Zheng, 2014b). Por este motivo es necesario contar con un plan para la prevención y detección de fraudes, procesos en los que se pueden apoyar de Big Data para transformar los sistemas de seguridad mediante el análisis y el uso de algoritmos para su detección.

Los tipos de análisis vistos en los párrafos anteriores solo son un ejemplo del tipo de conocimiento que se puede obtener de analizar Big Data. En la Tabla 3 se puede observar un listado más amplio de tipos de análisis categorizados por área de uso:

Tabla 3. Tipos de análisis categorizados por áreas de uso. Elaboración propia.

Cuidado de la salud	Mercadotecnia	Social media	Manufactura
• Detección de enfermedades • Monitoreo de pacientes • Medicamento personalizado • Pronósticos de tratamientos • Disminución de errores de diagnósticos	• Análisis de los sentimientos de los clientes • Campañas de análisis • Publicidad personalizada • Predicciones de precios	• Publicidad personalizada • Servicios basados en localización	• Análisis de la cadena de suministro • Mantenimiento predictivo • Análisis de reclamo de garantías • Monitoreo de producción
Transporte y turismo	**Telecomunicaciones**	**Servicios financieros**	**Gobierno**
• Seguimiento de ubicaciones • Detección de patrones de movimiento	• Análisis de los sentimientos de los clientes • Análisis de logs • Mantenimiento predictivo • Análisis y optimización de redes	• Análisis y detección de fraudes • Análisis de riesgo crediticio • Seguimiento de ubicaciones	• Servicios públicos • Análisis y detección de fraudes • Análisis de amenazas • Análisis de patrones de crímenes • Pronósticos de clima
Entretenimiento	**e-commerce**	**Energía**	**Alimentos**
• Análisis de los sentimientos de los clientes • Contenido personalizado • Análisis de comportamiento • Optimización de experiencias en-juego	• Análisis de flujo de paginas • Análisis de comportamiento de compras • Publicidad personalizada • Análisis de sentimiento de los clientes	• Descubrimiento de fuentes de energía no renovables • Análisis de redes • Mantenimiento predictivo	• Monitoreo de cultivos • Análisis de comportamiento de consumo de los clientes • Predicciones de precios • Servicios de entrega de alimentos

9. Visualización de Big Data

La visualización de datos es el paso final de la cadena de suministro de información mediante la cual se provee de un panorama general que permite explorar la situación problemática, hacerse preguntas y descubrir nuevas posibilidades de mejora. En otras

palabras, la visualización de datos es una de las técnicas más importantes para poder analizar de forma más clara la situación actual de la empresa. Dichas visualizaciones tienden a ser altamente interactivas y dinámicas por naturaleza pudiendo aplicarse a través de distintas técnicas como son los mapas mentales, infográficos, diagramas de conexión, etc. (Hurwitz et al., 2013) Uno de los principales problemas para la visualización de Big Data es ocasionada por el volumen de los datos y por la velocidad con la que se generan. La administración de Big Data requerirá técnicas fuera de las formas espacio-temporales de los datos para su visualización, además de que deberán tratar con datos no estructurados como el texto, los gráficos, tablas, etc. (Childs et al., 2013).

10. Herramientas de software libre para trabajar con Big Data

A continuación se presentan algunas de las herramientas de software libre para trabajar con Big Data, es importante comentar que estas herramientas son utilizadas dentro del INEGI para la realización de algunos estudios exploratorios de Big Data.

ElasticSearch

Es un software de código abierto basado en Apache Lucene, una biblioteca de búsquedas de texto completo. Elasticsearch fue desarrollado en el lenguaje de programación Java bajo la licencia de Apache y es utilizado para almacenar, buscar y realizar análisis con grandes volúmenes de datos de forma rápida y casi en tiempo real gracias a su motor de búsquedas de texto en un ambiente distribuido y su interfaz web que utiliza una API de servicios de transferencia de representación de estado (REST por el acrónimo en inglés de REpresentational State Transfer). Los datos son almacenados mediante documentos con formato JSON donde todos sus campos son indexados para poder realizar búsquedas posteriores sobre estos.

Grandes organizaciones están utilizado Elasticsearch en sus actividades diarias como:

- Wikipedia utiliza Elasticsearch para proporcionar búsquedas de texto completo con fragmentos de búsqueda resaltados, búsquedas conforme escribe y sugerencias de tipo lo que usted quiso decir.

- The Guardian utiliza Elasticsearch para combinar los registros de los visitantes con los datos de redes sociales para proporcionar retroalimentación en tiempo real a sus editores sobre la respuesta del público a nuevos artículos.
- StackOverflow combina búsquedas de texto completo con las consultas de geolocalización y utiliza más-como-esta para encontrar preguntas y respuestas relacionadas.
- GitHub utiliza Elasticsearch para consultar 130 mil millones de líneas de código.

Rivers de Elasticsearch

Un river es un servicio conectable ejecutándose dentro del clúster de Elasticsearch para insertar datos indexados y está compuesto de un nombre único y un tipo. Tienen la característica de ser muy fácil de trabajar dando la posibilidad de recabar datos de CouchDB, Wikipedia, RabbitMQ, Twitter, etc.

Las primeras implementaciones de los river eran exitosas y muy útiles, sin embargo poco a poco se empezó a detectar una debilidad, la estabilidad del clúster. Debido a la naturaleza de los rivers de trabajar con sistemas y librerías externas, la utilización de la memoria, sockets y otros recursos, los rivers empezaron a verse sobrecargados causando la inestabilidad del clúster.

El equipo de desarrollo de Elastic decidió depreciar los rivers a partir de la versión 1.5 para eliminarlos por completo en una próxima actualización con el objetivo de darles a los usuarios el tiempo necesario para determinar otro tipo de herramientas para sustituir la funcionalidad que éstos tenían y realizar la migración correspondiente.

Logstash

Logstash es una herramienta basada en jRuby, una implementación realizada con el lenguaje de programación Java y del lenguaje de programación Ruby. Se utiliza para recibir, procesar y dar salida a todo tipo de registros, registros de sistema, registros del servidor web, registros de errores, registros de aplicación, etcétera.

Al utilizar Elasticsearch como almacén de datos y Kibana como una herramienta de información para el usuario, Logstash actúa como un caballo de batalla brindando un arsenal de plugins que permite aprovechar de sus funciones con poco esfuerzo. Los plugins son:

- Entrada desde una fuente de datos.
- Filtros que son acciones para procesar eventos.
- Codecs para la conversión de formatos aceptados por Logstash, y
- Salidas que son los destinos donde los datos serán almacenados.

Con Logstash y una infraestructura distribuida, cada servidor web debe ser configurado para correr Lumberjack (es opcional pero altamente recomendado para economizar recursos). Lumberjack hace un forward de los logs a un servidor corriendo Logstash con una entrada de Lumberjack. Como Lumberjack requiere SSL, los logs van a ser encriptados del servidor web al servidor de logs central.

Apache Spark

Apache Spark es un clúster de sistemas computacionales de propósito general inicialmente desarrollado en la universidad Berkeley en 2009 y posteriormente donado a la fundación Apache. Desde 2009 a la fecha Spark ha sido utilizado por una amplia cantidad de organizaciones para procesar grandes cantidades de datos y tiene como colaboradores a más de 800 desarrolladores de 200 organizaciones.

Algunas de las características que presenta Spark están las siguientes:

- Comparado contra el paradigma de MapReduce de Hadoop, Spark puede ejecutar tareas cien veces más rápidas gracias a su motor DGA que soporta flujos de datos cíclicos en memoria, o diez veces más rápido en disco.
- Proporciona un API de alto nivel para construir aplicaciones paralelas rápidamente en los lenguajes de Java, Scala, Python y R.
- Cuenta con un amplio conjunto de herramientas de alto nivel que pueden ser integradas en una misma aplicación, incluyendo SQL Spark para SQL y

procesamiento de datos estructurada, MLlib para machine learning, GraphX para el procesamiento gráfico y Spark Streaming.

- Puede acceder diversas fuentes de datos como HDFS, Cassandra, HBase y S3. Además de que puede ejecutarse de manera standalone o en la nube, con Hadoop y con Mesos.

Lenguaje de programación Scala

En un inicio la mayoría de las aplicaciones desarrolladas para Hadoop requerían conocimientos en Java, sin embargo, con la creación de herramientas como Apache Spark en el año 2013, Scala se volvió más popular entre los desarrolladores de Big Data.

Scala es un moderno lenguaje de programación que fusiona el paradigma orientado a objetos con la programación funcional diseñado para desarrollar componentes de software utilizando una programación concisa, elegante y con tipos estáticos.

Lenguaje de programación R

El lenguaje de programación y entorno para la generación de cálculos y gráficos estadísticos llamado R es un proyecto GNU de software libre altamente extensible que ofrece una amplia variedad de técnicas estadísticas (análisis de series de tiempo lineal y no lineal de modelado, pruebas estadísticas clásicas, clasificación, agrupación, etc.) para realizar análisis de datos.

R es una de las herramientas que se pueden integrar con apache Hadoop para trabajar con Big Data conformando una potente plataforma para realizar análisis sobre grandes conjuntos de datos. Del mismo modo se encuentra el paquete para Apache Spark, SparkR que proporciona una interfaz ligera para usar Spark Apache desde R. Además, en la más reciente versión de Spark (1.6.1), SparkR provee una implementación que soporta operaciones como la selección, filtrado, agregación, etcétera en grandes conjuntos de datos.

11. Fuente de Big Data: Twitter

Twitter es un servicio de microblogging que ha estado en constante crecimiento desde que fue lanzado en Octubre de 2006 (Java, Song, Finin, & Tseng, 2007). Actualmente se estima que tiene más de 288 millones de usuarios activos en todo el mundo, los cuales llegan a enviar aproximadamente más de 500 millones de tweets cada día (Twitter, 2015b). El componente principal de Twitter son las actualizaciones de estado, también llamados tweets, y están formados por un máximo de 140 caracteres. En el sitio web los definen de la siguiente manera: "*Un Tweet es una expresión de un momento o idea. Puede contener texto, fotos y videos. Millones de Tweets se comparten en tiempo real, todos los días*" (Twitter, 2015a).

La dinámica de Twitter consiste básicamente en la generación de tweets por parte de los usuarios, donde dichos mensajes pueden ser leídos por otros usuarios siempre y cuando se hayan registrado como "seguidores" del usuario que escribió el tweet. El usuario que publica el tweet tiene la posibilidad de configurar si desea que éstos sean públicos, es decir, que toda la gente los pueda ver, que sean visibles solo para los seguidores o también pueden ser enviados de usuario a usuario.

Así como el término "seguidor" tiene gran importancia para entender fácilmente el funcionamiento de Twitter, la compañía ha puesto a disposición un glosario (http://estwitter.com/glosario/) con las palabras que se utilizan entre la comunidad de Twitter. A continuación se enlistan las que se consideran más importantes para este trabajo de investigación:

- @replies / mentions (menciones) – La forma de llamar a un usuario desde un tweet publicado por otro usuario es anteponiendo el símbolo del arroba (@) delante del nombre de usuario (por ejemplo: @estwitter), de forma que aparecerá la pestaña de replies del primero si así lo ha configurado (es opcional).
- RT / RTW / ReTweet – Es como la opción de reenviar en el correo electrónico, reenviar un tweet de otra persona. Otra versión es MT, que es cuando el usuario modifica un tweet enviado por otro usuario pero además modifica el contenido del mismo para expresar la idea que tiene sobre ese tema.

- Hashtag – Etiqueta de Twitter formada por el caracter # (almohadilla) y una palabra (por ejemplo #estwitter, #Elections, #NBAFinals) que permite identificar los tweets y categorizarlos por temas. Twitter convierte estas palabras en búsquedas hacia su motor de búsqueda, por lo que tienen gran popularidad por los usuarios.
- Trending Topic / Trends / Temas del Momento – Son las palabras de Twitter con mayor crecimiento tiempo real, es el resultado de un complejo algoritmo que muestra las palabras más mencionadas en los tweets.

Una de las características que posee Twitter es la capacidad de referenciar geográficamente, bajo permiso del usuario, la ubicación desde la que está compartiendo o publicando información abriendo las puertas a diversos estudios de interés para los investigadores.

API de Twitter

Twitter puso a disposición algunos de sus servicios de manera gratuita mediante la creación de API públicas, que son un conjunto de funciones y procedimientos para ser utilizadas por desarrolladores como una capa de abstracción, y tienen la finalidad de que cualquier persona pueda crear aplicaciones que se comuniquen con la red social. Es importante comentar que dependiendo de las necesidades de las aplicaciones que realice el desarrollador se pueden usar las API de manera individual o hacer combinaciones de estas.

Las API de Twitter son:

- REST API. Esta API está desarrollada con la tecnología REST para el intercambio de datos con los usuarios dando acceso de lectura y escritura de tweets dentro sus aplicaciones.
- Streaming API. Esta API permite acceder a una muestra equivalente al 1 por ciento (si 500 millones son los tweets que se generan globalmente, el 1% son 500 mil tweets diarios) del tráfico de tweets a nivel global en tiempo real y de manera gratuita. Por otra parte, si lo que se desea es acceder al total de los tweets se puede utilizar el Streaming de Twitter llamado Firehose pero ya bajo un costo.

- API de anuncios. Permite a los socios integrar sus propias soluciones de publicidad con la plataforma de publicidad de Twitter.

Twitter utiliza la notación de objetos de javascript (JSON por el acrónimo en inglés de JavaScript Object Notation) como formato para el intercambio datos. JSON tiene la particularidad de ser fácil de generar, de leer y no depende del lenguaje de programación en el que fue escrito al estar basado en el concepto de pares [clave, valor]. Es importante mencionar que no todas las claves disponibles son utilizadas en todos los contextos por lo que considerar una clave nula, vacía o con la ausencia de un valor es lo mismo.

La respuesta enviada por Twitter desde el API de Streaming es obtenida en formato JSON y tiene la estructura mostrada en la Figura 14.


```
{
  "created_at": "Thu Jul 16 06:30:17 +0000 2015",
  "id": 621567510085791700,
  "id_str": "621567510085791744",
  "text": "Cuba logra plaza en cuartos y México siembra dudas en partido loco - Cuba logró hoy la última plaza para los... http://t.co/8TssNlWCA0",
  "source": "<a href=\"http://www.informatesv.com\"rel=\"nofollow\">InformateSV</a>",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 2495023879,
    "id_str": "2495023879",
    "name": "InformateSV",
    "screen_name": "InformateSV",
    "location": "El Salvador, Centroamérica",
    "url": "http://www.informatesv.com",
    "description": "¡Todas las noticias de El Salvador en un solo sitio!",
    "protected":false,
    "verified": false,
    "followers_count": 13634,
    "friends_count": 7609,
    "listed_count": 60,
    "favourites_count": 87,
    "statuses_count": 83470,
    "created_at": "Wed May 14 21:20:36 +0000 2014",
    "utc_offset": null,
    "time_zone": null,
    "geo_enabled": false,
    "lang": "es",
```

```
33.         "contributors_enabled": false,
34.         "is_translator": false,
35.         "profile_background_color": "85D6FF",
36.         "profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png",
37.         "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
38.         "profile_background_tile": false,
39.         "profile_link_color": "0084B4",
40.         "profile_sidebar_border_color": "FFFFFF",
41.         "profile_sidebar_fill_color": "DDEEF6",
42.         "profile_text_color": "333333",
43.         "profile_use_background_image": false,
44.         "profile_image_url": "http://pbs.twimg.com/profile_images/466957034206683136/kJx_H-lw_normal.jpeg",
45.         "profile_image_url_https": "https://pbs.twimg.com/profile_images/466957034206683136/kJx_H-
lw_normal.jpeg",
46.         "profile_banner_url": "https://pbs.twimg.com/profile_banners/2495023879/1400165777",
47.         "default_profile": false,
48.         "default_profile_image": false,
49.         "following": null,
50.         "follow_request_sent": null,
51.         "notifications": null
52.       },
53.       "geo": null,
54.       "coordinates": null,
55.     "place": null,
56.       "contributors": null,
57.       "retweet_count": 0,
58.       "favorite_count": 0,
59.       "entities": {
60.         "hashtags": [],
61.         "trends": [],
62.         "urls": [
63.           {
64.             "url":"http://t.co/8TssNlWCA0",
65.             "expanded_url":"http://www.informatesv.com/cuba-logra-plaza-en-cuartos-y-México-siembra-dudas-en-
partido-loco",
66.             "display_url":"informatesv.com/cuba-logra-pla…",
67.             "indices": [
68.               112,
69.               134
70.             ]
71.           }
72.         ],
73.         "user_mentions": [],
74.         "symbols": []
75.       },
76.       "favorited": false,
77.       "retweeted": false,
78.       "possibly_sensitive": false,
79.       "filter_level": "low",
80.       "lang": "es",
81.       "timestamp_ms": "1437028217227",
82.       "@version": "1",
```

```
83.     "@timestamp": "2015-07-16T06:30:17.000Z",
84.     "type": "test1"
85.   }
```

Figura 14. Descripción técnica de un tweet

Aplicaciones con los datos de Twitter

Desde que Twitter puso a disposición sus distintas API para el desarrollo de aplicaciones y hasta la fecha, se han realizado una gran cantidad de trabajos de investigación en distintos campos gracias a las ventajas que ofrece como el poder disponer de datos muy diversos en tiempo real y de manera gratuita. En dichos estudios se han demostrado que, aunque la mayor parte de los datos generados en la red social de Twitter no son cien por ciento confiables por provenir de fuentes no oficiales, existe un porcentaje que pueden ser comparados con la realidad (Gupta & Kumaraguru, 2012) siendo un retrato generalizado de la sociedad para obtener información de interés, como por ejemplo predicciones en el mercado de valores, la política y los movimientos sociales y culturales (Hao Wang et al., 2012) (J. Paul Michael, and, & Dredze, Mark, 2011).

Esto es logrado gracias a la forma en la que trabaja Twitter donde los usuarios establecen lazos con otros usuarios comunicándose mediante los tweets.

Después de que los datos obtenidos de Twitter son limpiados y formateados a las necesidades del estudio se ha visto que llegan a ser tan parecidos a la realidad que diversos investigadores han realizado comparaciones contra lo que publican organismos oficiales demostrando que existe un alto grado de correlación entre ambas fuentes de datos (J. Paul Michael et al., 2011) (Hawelka Bartosz et al., 2013). Un estudio de estos fue realizado con millón y medio mensajes de Twitter aplicando el Modelo de Aspecto de Temas de Enfermedades (ATAM por sus siglas en inglés) (J. Paul Michael et al., 2011) que tenía como objetivo encontrar la correlación existente entre los tweets con los índices de influenza obtenidos por el CDC en los Estados Unidos mediante un experimento realizado cada semana entre los meses de Agosto de 2009 y Octubre de 2010. En la Figura 15 se puede apreciar que el coeficiente de correlación entre los datos de Twitter y el CFC era alto, aproximadamente de 0.958.

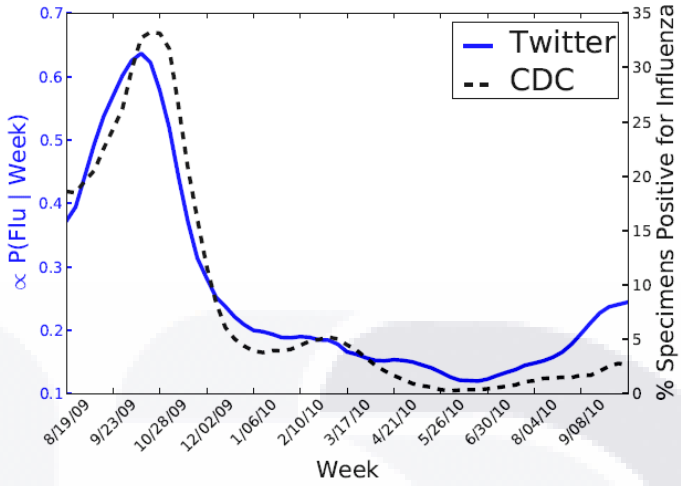


Figura 15. Tasa de influenza medido por el CDC vs ATAM (J. Paul Michael et al., 2011)

Uno de los retos a los que se enfrentan los científicos de datos al analizar los tweets es que el texto difiere mucho el lenguaje hablado en el sentido que incluye argots y palabras mal escritas incluyendo el uso de emoticones (conjunto de caracteres para representar un estado de ánimo), direcciones URL, RT para re-tweet, @ para hacer referencia a lo que un usuario menciona, # de hashtags y repeticiones sin olvidar que todo esto en sólo 140 caracteres.

Tomando como base el tipo de estudios anterior, otro de los trabajos en los que distintas organizaciones han puesto especial interés es la minería de opiniones y de sentimientos de lenguaje natural (Hodeghatta, 2013) ya que a partir de esta pueden obtener información que responda preguntas como ¿qué es lo que piensan los clientes de la organización?, ¿qué les agrada y qué les desagrada de sus productos? (Cuesta, 2013), ¿cuáles estrategias de marketing servirían para determinado sector?, etc. Este tipo de análisis representa un reto que implica tener un amplio conocimiento de las reglas sintácticas y semánticas del lenguaje.

12. Trabajos anteriores de análisis de movilidad humana

Dentro de este tipo de análisis es posible encontrar estudios de exploración de flujos de migración, actividad turística, dispersión de enfermedades y epidemias que anteriormente eran obtenidos mediante el estudio de estadísticas oficiales o con los datos obtenidos en pequeñas observaciones y encuestas.

El trabajo presentado por (Hawelka Bartosz et al., 2013) analiza tweets geo-referenciados con el objetivo de descubrir patrones globales de movilidad humana para después ser comparados con estadísticas globales de turismo y modelos de movilidad presentados por otros autores.

El procedimiento puede ser definido en las siguientes fases:

1. Preparación de los datos y su preprocesamiento. Se limpió la base de datos para evitar la contaminación de las estadísticas de movilidad. Se examinaron todas las localizaciones de un usuario y excluir aquellos lugares en los que se haya desplazado a una velocidad de 1000km/h y se filtraron los tweets realizados por entidades no humanas, como los generados por anuncios y juegos en la web.
2. Definición del país de residencia del usuario. En esta fase es primordial diferenciar entre un usuario turista y uno residente, por lo que se tomó como regla determinar al país de residencia al que tuviera más tweets referenciados en ese lugar por usuario, mientras que los lugares en los que tuviese menor número de tweets serian tomado como visitante de un país.
3. Detección de patrones de movilidad temporal. Patrones de movilidad encontrados a nivel mundial fueron durante las semanas de los meses de Julio y Agosto, utilizados comúnmente para vacacionar junto con algunas semanas del mes de Diciembre para los festejos de navidad y año nuevo.
4. Determinación y particionamiento de conexiones entre países. Se desarrolló una red de movilidad donde cada país se consideró como un nodo La relación de la red era direccional y las conexiones se hicieron considerando el flujo desde el país de residencia al país donde aparecía como visitante. El procedimiento de particionamiento para delimitar a las comunidades cohesivas espacialmente se realizó utilizando un algoritmo propuesto por Sobolevsky (A General Optimization

Technique for High Quality Community Detection in Complex Networks.") y fue aplicado de manera iterativa a las subredes dentro de la comunidad.

Como conclusión de este estudio se obtuvo que Twitter puede ser considerado como una fuente de datos valorable para los estudios de movilidad humana ya que están en línea con las estadísticas oficiales de turismo reportadas en el Foro de Economía Mundial de 2013 (WEF por las siglas en inglés de World Economic Forum), tal cual se muestra en la Figura 16.

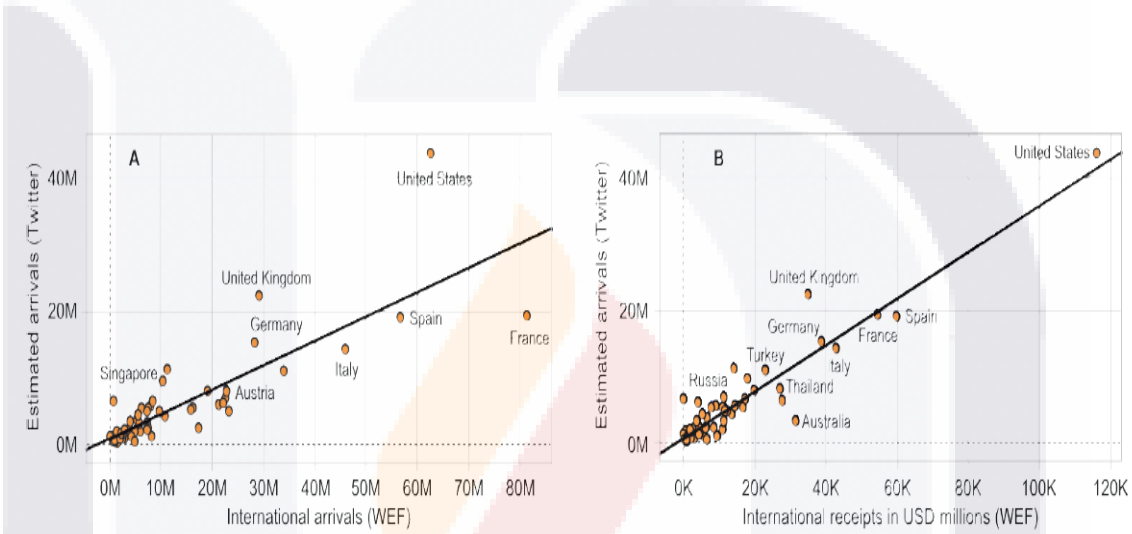


Figura 16. Las llegadas internacionales estimadas con datos de Twitter frente a las llegadas (A) y el valor nominal de los ingresos turísticos (gastos de los visitantes internacionales, B) proporcionados por WEF de 2013. Correlación medida con el estadístico R2 es igual a 0,69 y 0,88 respectivamente (Hawelka Bartosz et al., 2013).

Otra de las áreas de interés del estudio de movilidad humana es la detección de patrones de migración entre distintos países. A pesar de que los investigadores de estas áreas dependían del levantamiento de censos para obtener datos con los que podían estimar indirectamente el movimiento migratorio existía un espacio vacío de datos entre las fechas en las que se realizan los censos provocando que el análisis de la información tenga un retraso.

El trabajo realizado por (Zagheni et al., 2014) presenta un estudio con datos provenientes de Twitter de los países que conforman la Organización para la Cooperación y Desarrollo Económico (OECD por las siglas en inglés de Organisation for Economic Co-

operation and Development) y tiene como principal objetivo complementar las estadísticas de migración existentes además de proponer un método para aprovechar la información pública en línea para mejorar el pronóstico migratorio.

El método utilizado está formado por cuatro fases que se detallan a continuación:

1. Recolección de datos y su preprocesamiento. Se obtuvieron únicamente tweets georreferenciados en un lapso de aproximadamente dos años mediante el API de Twitter; se mapearon por país; se identificaron los que tenían referencias geográficas en un país, en dos países, en tres países, etcétera; se crearon muestras dependiendo del número de personas que potencialmente muestran patrones de movilidad en los y se dividió el conjunto de datos en periodos de cuatro meses por concepto del estudio.
2. Detección de las características demográficas de los usuarios de Twitter en la muestra. Para esta fase se utilizó el software de reconocimiento facial Face++ para estimar el género y la edad de los usuarios tomando como base la imagen de su perfil con la finalidad de proveer una idea general de la población sobre la cual se hizo el estudio y que debe ser considerada con precaución.
3. Estimación de las tendencias del grado de migración con un "enfoque de diferencia en diferencias". Se empleó un estimador de diferencia en diferencias para evaluar los cambios relativos en las tendencias y dependiendo de este se determinaba si se utilizaban series de tiempo para estimar los rangos de migración.
4. Medición de la relación entre movilidad nacional e internacional. Se utilizó el radio de giro para medir la distancia promedio entre los tweets georreferenciados hasta su baricentro.

Según los autores del estudio, las tasas de migración estimadas no son consideradas representativas de los países de la OECD, los datos obtenidos representaron la experiencia de movilidad de los usuarios que postearon regularmente sus tweets de los cuales se obtuvieron ideas interesantes. Por ejemplo, se observó que las tendencias de migración de México a los Estados Unidos van a la baja conforme a los resultados obtenidos por el Centro de Investigación Pew.

En 2009, (Azevedo, Bezerra, Campos, & de Moraes, 2009) presentó una metodología para realizar análisis del comportamiento de movilidad peatonal provenientes de redes inalámbricas y sistemas GPS tomando como contexto las áreas físicas en las que se mueven las personas. La metodología consiste en la ejecución de manera secuencial de 4 fases que son:

1. Definición de escenarios. En esta fase se definieron parámetros como el tipo de escenario (movimiento de automóviles y peatones en una zona urbana, automóviles en una autopista, gente caminando dentro de edificios), el área de interés, el número de dispositivos y la densidad de la red junto con sus respectivas restricciones para cada punto.
2. Captura y procesamiento. Se capturaron datos provenientes de redes inalámbricas y sistemas GPS los cuales tuvieron que ser procesados para garantizar que la señal obtenida indica realmente la ubicación de persona o el sistema GPS.
3. Análisis estadístico de datos. Se realizó el cálculo de varias medidas estadísticas como promedios, desviaciones estándar, varianzas, coeficientes de correlación, etcétera.
4. Detección de patrones de comportamiento. Se analizaron los datos obtenidos de las fases anteriores para obtener información importante de la movilidad humana. Si no se detectaban patrones, se realizaba un reajuste en la configuración de los parámetros de fases anteriores hasta que se obtuvieran los resultados esperados.

La metodología se aplicó en un parque de una ciudad de Brasil durante los meses de Enero a Mayo de 2008 con una participación de ciento veinte personas elegidas al azar para seguir su rastro mediante un dispositivo GPS. Se graficó el comportamiento de las variables y se identificó a que distribución de probabilidad se asemejaba más, y como resultado de este análisis se reportó que existe una leve variación en el cambio de velocidad y el ángulo de dirección que tienen las personas mientras caminan en el parque como se muestra en la Figura 17.

Figura 17. Rastro de las personas medido por GPS. Fuente (Azevedo et al., 2009)

El estudio de trayectorias es una parte importante en el análisis de movilidad ya que permite reconstruir los movimientos que ha realizado un usuario en un espacio y tiempo determinado. El trabajo desarrollado por (Gabrielli Lorenzo et al., 2014) permitió detectar patrones de movilidad urbana con datos públicos referenciados geográficamente en una zona urbana desde la red social de Twitter a través de técnicas de minería de datos.

Para la detección de trayectorias se apoyaron de DBpedia para caracterizar si la trayectoria fue realizada por un turista o por una persona local y se asoció a cada lugar que se visitó a las categorías específicas de FourSquare. Después de detectar las trayectorias realizadas por los usuarios se identificaron lugares clave realizando el análisis de matrices OD (origen- destino) y se utilizaron estadísticas de intermediación del borde para medir la relevancia de cada lugar.

De igual manera se construyó una matriz origen-destino (OD por sus siglas) semántica con las nueve principales categorías de Foursquare lo que permitió caracterizar las actividades llevadas a cabo al principio y al final de cada trayectoria junto con la importancia de estas actividades en relación con el contexto urbano. También se

generaron representaciones tabulares y diagramas de acore para visualizar la matriz OD semántica.

Como resultado final de este estudio se demostró que el análisis de información de fuentes de datos oportunistas constituye una forma relevante y barata de obtener una perspectiva de la movilidad urbana.

La Tabla 4 muestra la comparación entre los distintos estudios revisados de movilidad humana tomando como consideración las capas revisadas en la sección de arquitecturas de referencia de Big Data.

Tabla 4. Tabla comparativa de trabajos anteriores de análisis de movilidad humana. Elaboración propia.

Procedimiento	Método de Zagheni et al., 2014	Método de Azevedo, Bezerra, Campos, & de Moraes, 2009	Método de Gabrielli Lorenzo et al., 2014	Método de Hawelka Bartosz et al.
Definición de reglas	El tiempo del estudio se dividió en periodos de cuatro meses por concepto del estudio.	Se definieron parámetros como el tipo de escenario, el área de interés, el número de dispositivos y la densidad de la red. Definición de restricciones para los parámetros previamente definidos.	-	Se determinó como país de residencia al que tuviera más tweets referenciados en ese lugar por usuario, mientras que los lugares en los que tuviese menor número de tweets serian tomado como visitante de un país.
Recolección de datos	Se obtuvieron únicamente tweets georreferenciados en un lapso de aproximadamente dos años mediante el API de Twitter.	Se capturaron datos provenientes de redes inalámbricas y sistemas GPS.	Se obtuvó tweets referenciados geográficamente de una zona urbana extraídos mediante el API de Twitter.	Se obtuvieron tweets geo-referenciados obtenidos mediante el API de Twitter
Preprocesamiento	Se mapearon por el país y se identificaron los que tenían referencias geográficas en distintos países; se crearon muestras dependiendo del número de personas que potencialmente mostraron patrones de movilidad. Se utilizó un software de reconocimiento facial para estimar el género y la edad de los usuarios tomando como base la imagen de su perfil para proveer una idea general de la población sobre la cual se hizo el estudio.	Los datos fueron procesados para garantizar que la señal obtenida indicara realmente la ubicación de persona o el sistema GPS.	Se apoyaron de DBpedia para caracterizar si la trayectoria fue realizada por un turista o por una persona local y se asoció a cada lugar que se visitó a las categorías específicas de FourSquare.-	Se limpió la base de datos para evitar la contaminación de las estadísticas de movilidad. Se examinaron todas las localizaciones de un usuario y se excluyeron aquellos lugares en los que se haya desplazado a una velocidad de 1000km/h.
Análisis de datos	Medición entre movilidad nacional e internacional. Se utilizó el radio de giro	Se realizó el cálculo de varias medidas	Se construyó una matriz OD (origen- destino) semántica	El procedimiento de particionamiento se realizó

	para medir la distancia promedio entre los tweets georreferenciados hasta su baricentro	estadísticas como promedios, desviaciones estándar, varianzas, coeficientes de correlación, etcétera.	con las nueve principales categorías de Foursquare lo que permitió caracterizar las actividades llevadas a cabo al principio y al final de cada trayectoria.	utilizando un algoritmo y fue aplicado de manera iterativa a las subredes dentro de la comunidad.
Detección de patrones de comportamiento	Estimación de tendencias migratorias. Se empleó un estimador de diferencia en diferencias para evaluar los cambios relativos en las tendencias y dependiendo de este se determinaba si se utilizaban series de tiempo para estimar los rangos de migración.	Se analizaron los datos obtenidos de las fases anteriores para obtener información importante de la movilidad humana. Si no se detectan patrones, se realiza un reajuste en la configuración de los parámetros de fases anteriores hasta que se obtengan los resultados esperados.	-	Los patrones de movilidad encontrados a nivel mundial fueron durante las semanas de los meses de Julio y Agosto, utilizados comúnmente para vacacionar junto con algunas semanas del mes de Diciembre para los festejos de navidad y año nuevo.
Representación	-	-	Se generaron representaciones tabulares y diagramas para visualizar la matriz OD semántica.	Se realizó un grafo direccional donde cada país fue consideró como un nodo de la red y los limites fueron asignados con el número de usuarios de Twitter considerando el flujo desde el país de residencia al país donde aparecía como visitante

13. Trabajos anteriores de análisis de impacto de eventos

Otra clase de estudios que se han realizado es el análisis del impacto que tiene un evento determinado en la sociedad donde algunos de ellos los comparan contra otros medios oficiales de recolección de datos con el objetivo de determinar en qué grado se asemeja a la realidad y ver qué tan aplicables pueden ser para su uso como audiencia social. Es posible encontrar eventos de diferentes y muy variadas índoles como los políticos, sociales (Luis Cesar Torres Nabel, 2014) (Congosto, Deltell Escolar, Claes, & Osteso, 2013), tecnológicos (Luis César Torres Nabel, 2009) y culturales los cuales, aunado al calibre del propio evento, la dispersión del mismo aumenta en gran medida teniendo un mayor número de espectadores gracias a su propagación en las redes sociales.

En la actualidad existen herramientas como Twitterbinder (https://www.tweetbinder.com/), TweetReach (https://tweetreach.com/), follow the hashtag (http://www.followthehashtag.com/) que ayudan a medir el impacto de un evento en Twitter donde algunas de ellas funcionan mediante hashtags o palabras en el tweet permitiendo tener una imagen más clara de lo que ha pasado en dicho evento brindando información como el impacto, alcance, fuente, idioma, los contribuidores más activos, además de descargar informes en PDF y en Excel.

Un estudio de análisis de un evento de carácter político-social es el realizado por (Luis Cesar Torres Nabel, 2014) donde discute el caso originado en Twitter el 27 de Abril de 2013 en México conocido como el caso “Lady Profeco” que en poco tiempo se convirtió en un “trending topic” y llego a culminar con la destitución del titular en turno de la Procuraduría Federal del Consumidor (PROFECO). El objetivo del estudio fue determinar si la hipótesis de influencia de enmarcamiento (“framing” en inglés) de los medios de comunicación, proveniente de la teoría de la agenda setting, puede establecer una fijación cognitiva en la audiencia para considerar al evento relevante en la vida pública.

- Se recolectaron los tweets mediante el empleo de seis hashtags.
- Se detectaron y se analizaron los principales actores de dichas publicaciones para caracterizar la masa ciudadana.
- De los usuarios detectados se extrajeron dos métricas, el grado de influencia y el número de seguidores que tienen utilizando la aplicación tweetlevel (http://www.edelmanberland.com/).

- A partir del análisis anterior, se obtuvieron nuevos indicadores como el número de seguidores, el número de retweets, menciones, popularidad, compromiso, confianza, etc.
- Finalmente, se identificó la jerarquía de cada uno de los actores involucrados teniendo como base los niveles de la aplicación de tweetlevel: observadores (espectadores del evento social), comentadores (su influencia está más en lo colectivo que en lo individual), curadores (recopila, selecciona y filtra la información), iniciadores de ideas (grupo creativo detrás de muchas ideas en internet) y los amplificadores (actores con mucha influencia y seguidores).

Como resultado se obtuvo que el 63,3% de los usuarios analizados tienen alto índice de influencia por ser profesionales en temas públicos teniendo que solo el 33,7% de los actores son usuarios sin ningún tipo de agenda evidente. Por lo menos en este caso *#LadyProfeco* parece ser que los periodistas y algunos políticos profesionales son actores clave para enmarcar el acontecimiento y volverlo un caso de interés público.

Un estudio comparativo entre los datos de la red social y una fuente de datos oficial es el realizado por (Congosto et al., 2013) cuya hipótesis fue demostrar que la información proveniente de ambas fuentes, en este caso la oficial representa la audiencia audimétrica del evento televisivo “los premios Goya” realizados el 17 de Febrero de 2013 en España, contra lo recolectado en Twitter no son comparables. Por ser un evento con alto impacto político y social se planteó encontrar quienes son los líderes de opinión en Twitter durante la transmisión en directo del programa.

- Los datos de Twitter fueron recolectados mediante una aplicación desarrollada por la propia investigadora además de contar con el apoyo de dos empresas españolas, como herramientas de control, que dan seguimiento a la audiencia social televisiva, mientras que los datos oficiales la organización GECA donó la información de la audiencia de la televisión recolectada por audímetros que es el estándar oficial.
- Después de esto se realizó la comparación de ambas audiencias para verificar si es posible obtener una estimación de la audiencia real del programa por medio del flujo de tweets.
- Identificar y analizar los líderes de opinión formados en Twitter durante la transmisión del programa.

- Analizar el comportamiento de la audiencia social y la interacción de los usuarios, cadena de televisión, presentadores, premiados y productora del evento.
- Elaborar un análisis estadístico y semántico de los tweets emitidos por los internautas.
- Identificar los temas de debate en Twitter durante la gala.

Como parte de los reportes de la investigación se encontró que la audiencia social y creativa medida por medio de Twitter no corresponde a la audiencia audimétrica recolectada por los sistemas tradicionales de GECA, salvo en algunos momentos en los que llegan a coincidir llegando a la conclusión de que Twitter no puede sustituir al sistema audimétricopero de acuerdo al estudio sí lo complementa ya que permite encontrar aspectos importantes como las diferentes discusiones generadas sobre el evento, el grado de participación de los televidentes, quienes eran los líderes de opinión y el perfil que tienen. Este tipo de cuestiones sería prácticamente imposible descubrir con otros sistemas de medición.

Capítulo 4. Metodología

A pesar de que las tecnologías y las herramientas para trabajar en proyectos de Big Data han ido evolucionando en los últimos años y con el aumento en el interés en la construcción de soluciones de Big Data, todavía cuesta trabajo obtener conocimiento mediante el análisis de estos datos. La necesidad de contar con una guía que se pueda utilizar desde etapas tempranas y durante el desarrollo de un proyecto efectivo de análisis de Big Data es el interés particular que tiene este trabajo.

El método propuesto en este documento tiene el objetivo de conducir al usuario en la realización de las actividades propias de una investigación indicando qué hacer y cómo actuar ante una situación recurrente trayendo consigo su solución de una forma sistemática y disciplinada para generar información con datos de Twitter.

Según (Mora, 2004) (Mora et al., 2008)y por (Hevner et al., 2004) la propuesta de métodos es considerada como investigaciones de diseño conceptual y tienen carácter exploratorio y/o descriptivo siendo uno de los principales métodos disponibles en el campo de los sistemas de información utilizado para generar nuevas teorías, modelos o esquemas conceptuales que serán probados empíricamente empleando otros métodos de investigación como encuestas, casos de estudio o experimentos de laboratorio. Mora comenta (Mora, 2004) que el desarrollo de este tipo de investigaciones se puede realizar en cuatro fases consistentes en:

- Formulación del problema de investigación.
- Análisis de trabajos relacionados.
- Desarrollo del modelo conceptual.
- Validación del modelo resultante.

De acuerdo con (Hevner et al., 2004), la investigación en la ciencia del diseño de sistemas de información es un proceso para resolver problemas mediante la creación de un artefacto innovador donde las teorías existentes son insuficientes o para solucionarlo de una manera más eficiente y útil para el dominio del problema específico. Para lograr esto propone un marco de trabajo y un conjunto de pasos a seguir para realizar una investigación en la ciencia de diseño. El primero se utiliza para alinear las estrategias de negocio con las estrategias y la infraestructura de las tecnologías de la información; y los segundos se utilizan para ayudar a entender, ejecutar y evaluar una investigación en la ciencia del diseño y consisten en:

1. Diseñar un artefacto en la forma de un constructo, un modelo, un método o una instanciación.
2. Determinar la relevancia del problema.
3. Diseñar la forma en la que se evaluará el artefacto.
4. Describir que contribuciones tendrá la investigación.
5. Aplicar métodos rigurosos en la construcción y evaluación del artefacto de diseño.
6. Diseñar como un proceso de búsqueda.
7. Comunicar el resultado de la investigación

Para el desarrollo de este trabajo de investigación se tomó como base principal el trabajo realizado por (Mora, 2004) y se consideraron algunas de las ideas presentadas por (Hevner et al., 2004), para definir el proceso que se siguió y dio como resultado la creación de un artefacto delimitado en un método informático para trabajar con Twitter, considerada como fuente de Big Data, para generar información actualizada y coherente de interés nacional.

Posteriormente se procedió a evaluar el método mediante el marco de trabajo de Análisis y Diseño de Sistemas Basado en el Modelo de Información Normativo (NIMSAD por su siglas en ingles de Normative Information Model-Based Systems Analysis and Design) propuesto por N. Jayaratna que es una de las técnicas que ha sido utilizada para la evaluación de metodologías en el área de sistemas de información (López, 2008) (Jayaratna & Armstrong, 2005) y que se compone de cuatro elementos clave.

1. La situación problemática que representa el contexto del problema. Parecida a la formulación del problema de investigación de (Mora, 2004)
2. El solucionador destinado del problema que viene siendo el rol que asume un grupo o un individuo que recomienda posibles soluciones. En este caso es todo el personal involucrado que labora en el INEGI y el tesista de este documento.
3. El proceso de solución del problema. Parecida al desarrollo del modelo conceptual. de (Mora, 2004)
4. La evaluación. Consiste en definir criterios para determinar la utilidad del proceso de solución del problema.

Descripción de la metodología

A continuación se describe el proceso que se siguió en este trabajo de investigación, desde la formulación de la problemática de investigación hasta la comunicación de los resultados obtenidos y la elaboración de la documentación del mismo.

1. Formulación del Problema de Investigación

El problema de investigación fue planteado debido a la necesidad del INEGI de obtener conocimiento a partir del análisis de grandes cantidades de datos que tienen a su disposición, haciendo esto de una manera clara y ordenada y que pueda ser replicable obteniendo los mismos resultados. Ante esta situación se decidió aplicar este trabajo de investigación en proponer un método informático que servirá de guía para trabajar nuevos proyectos con fuentes de Big Data para generar información actualizada y coherente en temas de interés nacional.

Posteriormente se procedió a determinar el contexto del problema de investigación para de esta forma identificar sus límites y establecer objetivos y preguntas de investigación. Para esto se tomó en cuenta la revisión del estado del arte referente a métodos o metodologías para la solución de proyectos de Big Data y se revisaron algunos procesos de ciencia de datos con la finalidad de identificar que investigaciones se han realizado con el mismo enfoque que se presenta en este trabajo y determinar qué aspectos de cada una de ellas se apegan más a las necesidades de investigación del INEGI tomando en consideración la elaboración de algunos estudios realizados anteriormente con datos de la red social de Twitter.

2. Determinación de la relevancia de la investigación

Una vez que el problema quedó delimitado, se continuó con determinar la importancia de la investigación. Teniendo en consideración lo que se indicó en la formulación del problema, hay gran interés por parte del INEGI en incursionar en el desarrollo de soluciones de Big Data para generar estadísticas interesantes y que puedan ser utilizadas para la toma de decisiones y diseño de políticas públicas que sirvan para el desarrollo del país en áreas donde anteriormente era complicado por la carencia de fuentes de datos o por los altos costos que implicaban su recolección o su procesamiento.

Adicionalmente, en la revisión de la literatura del tema se encontró que investigadores y tomadores de decisiones de organizaciones públicas y privadas de distintos tamaños tienen la mira puesta en temas como la recolección, el almacenamiento, el procesamiento, la administración, la seguridad y el análisis de Big Data y en el cómo hacerles frente para obtener el mayor beneficio al implantar este tipo de soluciones dentro de una organización. Algunos de los campos que han sacado provecho de estas soluciones son: el área del cuidado de la salud donde se puede extraer conocimiento para mejorar la detección de enfermedades, monitorear de una manera más eficiente a pacientes en estado grave o terminal, también se pueden disminuir errores de diagnósticos y hasta realizar pronósticos de tratamientos. Otro ejemplo es en el área de mercadotecnia donde se puede realizar análisis de los sentimientos de los clientes respecto a cierto servicio o producto, análisis de campañas publicitarias, detección de patrones de compras para envió de publicidad personalizada y hasta predicciones de precios de cualquier producto.

3. Elaboración del Marco Teórico-Conceptual

Se realizó una revisión de literatura con la finalidad de conocer a detalle las teorías y conceptos del campo de estudio, una revisión de términos técnicos y también se identificaron las herramientas y tecnologías aplicadas por el INEGI en proyectos anteriores de análisis de Big Data para utilizarlas en el proceso de solución de los casos prácticos con los que se pretendió evaluar el método informático generado en esta investigación.

4. Construcción de la propuesta del método

Tomando como base las metodologías y los procesos de ciencia de datos vistos en la revisión de la literatura del estado del arte, los trabajos previos del INEGI con algunos proyectos de análisis de Big Data con carácter exploratorio y considerando los elementos y las similitudes que existen entre ellos se construyó la propuesta inicial del método donde se describió a detalle cada uno de los pasos que lo conforman además de realizar su mapa correspondiente para facilitar su entendimiento y su implementación.

5. Evaluación del modelo

Para evaluar el modelo se utilizó el marco de trabajo NIMSAD propuesto por N. Jayaratna. Los primeros tres pasos que comprende NIMSAD son muy parecidos a los propuestos por (Mora, 2004) pero con la excepción de que este último está enfocado en el diseño del artefacto y no a la evaluación, por lo que para este último paso se decidió en utilizar el marco de trabajo antes comentado.

La forma de evaluar el resultado fue mediante la construcción de dos casos prácticos, un análisis de movilidad cotidiana de los usuarios que publican dentro de la República Mexicana en la red social de Twitter y un análisis de impacto de eventos de la vida real en la misma red social donde se aplicó el método propuesto identificando y describiendo cada uno de los pasos que lo integran teniendo la intención de demostrar en ambos casos que la aplicación del método sirve para generar información útil en la vida real.

Los criterios de aceptación de los resultados obtenidos en los dos casos prácticos al aplicar el método fueron definidos por el experto del área del INEGI el cual ha estado trabajando directamente con las técnicas y herramientas para analizar el Big Data, mismas que se utilizaron en esta investigación. Otra de las funciones que tuvo fue la de estar revisando constantemente esta investigación para monitorear los avances y ver que los objetivos fueron alcanzados.

6. Comunicación del resultado

Se presentaron los resultados del método informático tomando en consideración los casos de estudio realizados (análisis de movilidad cotidiana y análisis de impacto de eventos de la vida real) donde se determinó el grado de utilidad del método y se identificaron posibles áreas de oportunidad para realizar distintos tipos de estudio en el contexto de una problemática particular. Y finalmente se realizó el documento de tesis para reportar los resultados de la investigación.

Capítulo 5. Descripción de la solución

En este capítulo se presenta la propuesta del método que busca facilitar el desarrollo de un proyecto de análisis de Big Data mediante la presentación de un conjunto de fases y actividades que permitirán alcanzar las metas planteadas de una manera clara y ordenada.

Para la realización del método se estudiaron: la literatura del estado del arte, las técnicas y herramientas (estas últimas utilizando un proceso de prueba y error) las experiencias y el conocimiento adquirido tras desarrollar otras soluciones dentro del mismo dominio en el INEGI para determinar qué elementos son considerados importantes en el análisis de Big Data. El método integra dichos aspectos junto con la identificación de los roles participantes en la solución con la finalidad de enriquecerlo y perfeccionarlo para poder utilizarlo en nuevos proyectos de manera que pueda ajustarse y ser útil en diferentes entornos organizacionales.

Identificación de roles participantes en la solución

Tomando como base lo visto en la forma de trabajar en el INEGI y la revisión de literatura, los roles que se utilizaron en el método propuesto son:

- Persona de negocios o experto del dominio, son aquellos que tienen la capacidad de tomar decisiones respecto al problema que se va a resolver ya que conocen a fondo los procesos de negocio de la organización y saben cómo obtener ganancias mediante la interpretación de los resultados.
- Investigador de datos, gracias a que poseen un amplio conocimiento académico tiene la habilidad de entender procesos complicados sin importar el tipo de negocio ni los análisis estadísticos a desarrollar. Se encarga de investigar las posibles fuentes de datos a utilizar en el proyecto y de definir una serie de reglas para su recolección, preparación, transformación, y si es necesario, integración y procesamiento de los datos para posteriormente realizar los análisis correspondientes y presentarlos al experto de dominio para su respectiva evaluación.
- Innovador de datos, al igual que el investigador de datos, cuenta con el conocimiento para resolver problemas de análisis de datos, sin embargo, la forma en la que lo hacen es mediante el empleo de un amplio repertorio de herramientas y tecnologías de última generación que permite hacerlo de una mejor manera. Tiene la función de determinar las herramientas y tecnologías a utilizar además de definir los algoritmos

que se necesitaran codificar para preparar, transformar, integrar y procesar los datos.

- Desarrollador de datos, que se encargan de desarrollar, probar y depurar los scripts y herramientas de software que permitan recolectar, almacenar y aprender de los datos además de resolver problemas técnicos relacionados con estas actividades.

Dependiendo de las características de la organización y de los conocimientos de las personas involucradas en el proyecto es posible que uno o más roles puedan ser cubiertos por una misma persona.

Descripción general del método informático

Como se comentó previamente, el método está basado en el cumplimiento de un conjunto de fases y actividades que involucran los pasos necesarios desde la identificación de la situación problemática que la organización desea controlar o erradicar hasta la medición y monitoreo de los resultados producto del análisis de las fuentes de datos seleccionadas, los cuales podrán ayudar a guiar a la organización de una manera más segura mediante la toma de decisiones basadas en datos.

El método consta de la ejecución de ocho fases formando un proceso cíclico donde el número de iteraciones dependerá de las características propias de la investigación, teniendo por ejemplo que si el estudio a realizar forma parte de un análisis exploratorio de Big Data tendrá menor número de iteraciones que un estudio donde su producto final será la implantación de una aplicación para un determinado proceso de negocio de la organización.

Del mismo modo pero a nivel interno, algunas de las actividades que conforman las fases se realizan de forma secuencial y otras de manera iterativa debido a la naturaleza propia del Big Data (las 3Vs vistas en el marco teórico) donde los involucrados del proyecto deben tener la posibilidad de modificarlas y expandirlas sus procesos de manera gradual con el propósito de garantizar en mayor medida los resultados esperados de la investigación.

Es importante comentar que hay actividades que no es necesario describir o documentar y esto dependerá del tipo de estudio que se esté realizando. En el método se describe cuáles de ellas son de carácter obligatorio y cuales son opcionales.

En la Figura 18 se puede apreciar la descripción general del método, donde la ejecución de las fases uno, dos y tres se realizan iterativamente para diseñar y experimentar la viabilidad de la investigación, posteriormente se realiza la fase cuatro que es donde se recolectan los datos con los que se va a trabajar y viene una segunda ejecución de fases iterativas (la fase cuatro y cinco) que consisten en preparar los datos y aplicarle diferentes técnicas de análisis a los datos tantas veces sean necesarias para resolver el problema de negocio. La siguiente fase corresponde a la presentación de los resultados o a la implantación en proceso de negocio de la organización, esto último depende de si la investigación es un análisis exploratorio o una aplicación de Big Data. Y finalmente llega la fase donde se monitorea periódicamente y se miden los resultados después de que el sistema comienza a funcionar en la producción.

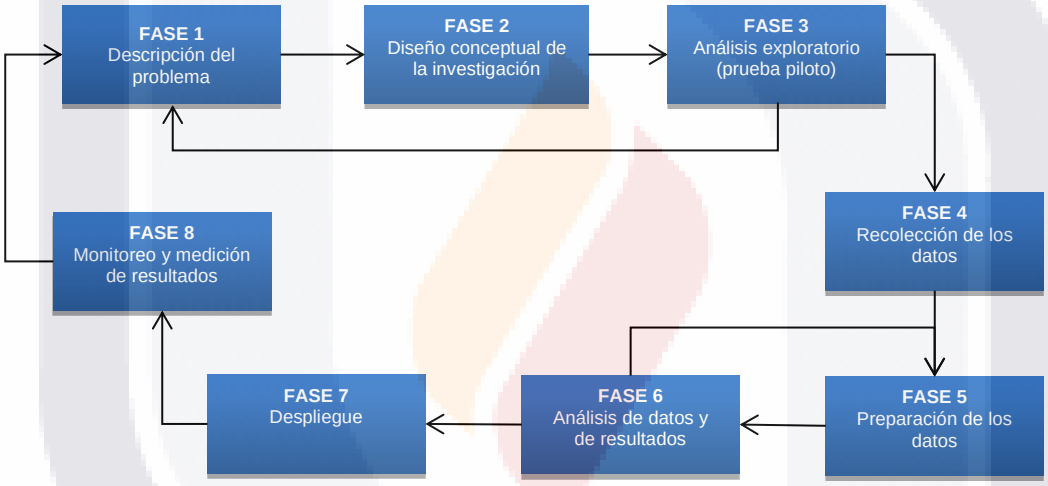


Figura 18. Descripción general del método propuesto. Elaboración propia.

A continuación se describen y se detallan cada una de las fases vistas en la Figura 18.

Fase 1. Descripción del problema.

Es la fase inicial del método y consiste en identificar y describir a detalle los motivos por los cuales se desea indagar en la solución de un proyecto de análisis de Big Data.

Tomando como base los roles de los involucrados en los proyectos de Big Data descritos anteriormente, se tiene que en el desarrollo de esta fase participan las personas de negocio o

expertos del dominio de los datos ya que son los que conocen la forma actual de trabajar de la organización y la forma en la que desean que funcione; y el innovador de datos que será en encargado de encontrar, de manera creativa, la mejor solución al problema.

Esta fase forma parte de un proceso iterativo, consiste en realizar múltiples reuniones entre los involucrados y es aconsejable no pasar a la siguiente fase hasta tenerla bien definida ya que la correcta descripción del problema permitirá que el análisis arroje datos útiles para la solución de la situación. Los datos de entrada, el proceso y los datos de salida para esta fase son los siguientes:

Elementos de entrada:

- Detección de una situación no deseada dentro de la organización o el interés en algún estudio con fines exploratorios.

Proceso de la fase:

Las actividades a realizar dentro de la fase son:

1.1 Identificación del problema *(Actividad necesaria)*. Es una descripción de la situación actual no deseada dentro del ambiente de una organización y puede ser identificada contestando preguntas de alto nivel que posteriormente servirán para definir los objetivos y requisitos de la investigación. De acuerdo con (Mora, 2004) la situación problemática es originada por diversas causas como la falta de conocimiento en un área determinada, la falta de métodos eficientes para generar un producto o servicio y hasta aspectos negativos de tipo social, económico, tecnológico, etcétera, internos o externos a la organización. Mientras más conocimiento se tenga acerca de los procesos, los flujos de materiales, las estructuras, las personas, la tecnología y la información de la organización será más fácil entender y solucionar los problemas que se presenten.

1.2 Impacto al negocio (*Actividad necesaria*). Es una declaración donde se especifica cómo afecta la situación no deseada a cierto proceso, producto o servicio que se realiza en la organización.

1.3 Antecedentes (*Actividad necesaria*). Se realiza un análisis para identificar los hechos generales que fueron los causantes de la situación problemática dentro de la organización. Adicionalmente se realiza una investigación sobre estudios relacionados para analizar posibles soluciones, los límites de sus trabajos y sus resultados con el objetivo de identificar más fácilmente una solución para el problema de la organización.

1.4 Condiciones en las que ocurre (*Actividad necesaria*). Se identifican las condiciones sobre las cuales ocurre la situación problemática para facilitar su comprensión.

1.5 Definición de objetivos (*Actividad necesaria*). Son sentencias donde se especifica lo que se espera obtener al finalizar el proyecto. Es recomendable redactar los objetivos lo más claramente posible y mantenerlos siempre presentes para disminuir la probabilidad de desviarse del proceso de la solución.

1.6 Identificación de las fuentes de datos a utilizar (*Actividad necesaria*). Una vez identificado el problema se procede en determinar si las fuentes de datos disponibles actualmente son capaces de brindar los datos necesarios para llevar a cabo el proyecto. Si no fuese el caso se necesita realizar un análisis costo beneficio para decidir la mejor manera de obtener los datos, algunas alternativas que se pueden encontrar es la compra del conjuntos de datos de interés o la recolección mediante el desarrollo de una aplicación informática, por mencionar algunos ejemplos.

1.7 Definición del alcance (*Actividad necesaria*). Se especifican una serie de reglas que delimitan las fronteras sobre las que se trabajara en la solución con la finalidad de reducir la complejidad poniendo mayor atención en los elementos involucrados para no desviarse de problema que se está tratando de resolver.

Elementos de salida:

- Documento de texto en el que se especifica las actividades realizadas en el proceso de la fase correspondiente (identificación del problema. Impacto al negocio, antecedentes, objetivos y fuentes de dato a utilizar).

Fase 2. Diseño conceptual de la investigación.

Esta fase está orientada a la identificación de los aspectos que determinaran el camino de deberá tener la investigación. Al igual que la fase anterior, forma parte de un proceso iterativo donde se requiere realizar múltiples reuniones entre los involucrados con la finalidad de que las actividades a realizar queden lo mejor definidas para dedicarle el menor tiempo posible.

Los roles que participan en su desarrollo son el innovador de datos que será en encargado de determinar lo que se quiere obtener de la investigación tomando como base los objetivos definidos en la fase anterior, y junto con el investigador de datos, y tomando en cuenta su experiencia determinarán las técnicas, herramientas y plataformas para realizar el análisis de los datos.

Cada que se finalice la fase, el equipo involucrado en la misma y la persona de negocios de los datos evaluarán y determinarán si el diseño conceptual cumple con las expectativas, si se llega a contraponer con algún objetivo habrá que realizarse nuevamente la fase actual. Los datos de entrada, el proceso y los datos de salida para esta fase son los siguientes:

Elementos de entrada:

- Documento de texto generado en la fase anterior con la identificación del problema, impacto al negocio, antecedentes, objetivos y fuentes de dato a utilizar.

Proceso de la fase:

La descripción de las actividades a realizar son:

2.1 Definición de preguntas de investigación (*Actividad necesaria*). Son interrogantes que presentan de manera directa la situación problemática y cuya respuesta representa el

camino a seguir para resolverla. Las preguntas deben ser planteadas lo más precisas posibles sin utilizar términos ambiguos ya que de lo contrario pueden conducir a una mala solución.

2.2 Definición de hipótesis (*Actividad opcional*). Es la definición de preposiciones que suelen verse como posibles respuestas a las preguntas de investigación pudiendo ser aceptadas o rechazadas dependiendo del análisis realizado en el estudio o investigación. Hernández Sampieri y colegas (Hernández Sampieri, Fernández Collado, & Pilar Baptista, 2006) las describen como una explicación tentativa del fenómeno investigado que sirven como guías para un estudio.

2.3 Detección de variables (*Actividad opcional*). Se identifican las variables dependientes e independientes que se van a utilizar para probar las hipótesis del estudio.

2.4 Perfilar variables (Actividad necesaria). Es la identificación de todas las variables involucradas en el estudio, desde las que están relacionadas directamente en el domino de aplicación hasta las creadas para utilizarlas en los modelos analíticos como sumatorias, porcentajes, etcétera y las variables de decisión que son aquellas que se utilizan en las estrategias de decisión una vez que se tengan los resultados del modelo.

2.5 Selección de técnicas y elementos para el análisis (*Actividad necesaria*). En esta actividad se establecen las técnicas de análisis que se aplicaran en los datos con la el propósito de extraer patrones útiles que describan el comportamiento de los mismos. Algunas de las técnicas de análisis que se pueden utilizar son: la estadística descriptiva, minería de datos, optimización, redes neuronales, etc. Del mismo modo se presentan los algoritmos, datos, servicios, sistemas y aplicaciones extras que son necesarios para realizar el análisis del estudio.

2.6 Definición de infraestructura y herramientas de software (Actividad necesaria). Actualmente existe una variedad de herramientas, tanto de software libre como de licencia comercial, entre las que se puede elegir la que más se apegue a las diferentes necesidades de análisis la organización. Algunas de las consideraciones antes de seleccionar las herramientas son:

- Las fuentes de los datos a analizar y sus respectivos formatos.

- La capacidad de almacenamiento de datos que la organización planea utilizar.
- La frecuencia con la que va a utilizar información en tiempo real.
- La velocidad a la que va a procesarse la información.
- El grado de precisión de los datos.
- La capacidad de escalar las aplicaciones.

Dichas necesidades pueden ser cubiertas por los proveedores de plataformas ya que ofrecen componentes de software como: el software de integración de datos, DBMS, software de análisis y de inteligencia de negocios y de procesamiento de flujos de datos.

La selección de plataformas es un paso muy importante ya que la correcta elección de éstas contribuirá a que el estudio se realice sin improvistos facilitando los procesos de análisis exploratorios para la comprensión de los datos, optimizando los resultados y minimizando los errores.

2.7 Definición de la arquitectura. Una vez que ya se cuenta con las técnicas, herramientas y la infraestructura el siguiente paso es determinar cómo se relacionan entre ellos mediante el diseño de una arquitectura que funcione como un pilar, proporcionando un marco de referencia para que los involucrados en resolver el problema de estudio puedan seguir la misma línea de trabajo y se apegue a lo planeado en la construcción de la solución

2.8 Elaboración de programa de trabajo (*Actividad necesaria*). En esta actividad se procede a elaborar un programa de trabajo de manera que el desarrollo del proyecto se realice de manera coordinada y sistemática relacionando entre si y de manera directa los recursos disponibles. Algo a tener en consideración es que como el método propone una serie de actividades que pueden realizarse una o más veces, el enfoque de planeación tradicional en el que se tienen bien identificadas las actividades y la forma de realizarlas puede acercarse poco a las necesidades de la solución donde la estimación de finalización del proyecto puede verse afectada al no tener consideradas el número exacto de iteraciones que puedan tener. Por esta razón se considera que una orientación a una planeación ágil en la que se hagan revisiones cada dos a cuatro semanas facilitará en gran medida a disminuir el número de veces que haya que realiza la misma actividad ya que los involucrados del proyecto estarán en tiempo y forma para hacer cualquier modificación que sea necesaria para cumplir con los objetivos del estudio.

Elementos de salida:

- Documento de texto en el que se especifica las actividades realizadas en el proceso de la fase correspondiente (definición de preguntas de investigación, hipótesis y variables dependientes e independientes además de su respectivo perfilado,
- Selección de técnicas (algoritmos) y elementos para el análisis.
- Definición e instalación de la infraestructura y herramientas de software.

Fase 3. Análisis exploratorio (prueba piloto).

Una vez que se tienen identificadas las fuentes de datos, las técnicas y herramientas, y la infraestructura el siguiente paso consiste en realizar una prueba piloto con una muestra previamente levantada de los datos para explorarlos y verificar si efectivamente contienen el valor suficiente para revelar ideas, verificar si los recursos asignados son suficientes para manejar la cantidad de datos y determinar si las técnicas y las herramientas son las adecuadas para el análisis.

Los roles involucrados en esta fase son el innovador de datos y el desarrollador de los datos donde su principal objetivo es determinar el grado de validez, calidad y veracidad de los datos además de poder determinar si las fuentes seleccionadas tienen suficientemente valor para continuar con las siguientes fases.

Esta fase, junto con las dos anteriores forman parte de un proceso iterativo que utiliza el método de prueba y error por lo que es común encontrar durante las primeras ejecuciones algunas situaciones que devuelvan resultados no esperados. Una vez finalizada la fase, el innovador de los datos y la persona de negocios de los datos se reunirán con el equipo involucrado en el desarrollo de la fase para evaluar y determinar si la prueba piloto cumple con las expectativas planteadas al inicio de la investigación. Si se llega a contraponer con algún objetivo habrá que realizarse nuevamente la fase actual.

Los datos de entrada, el proceso y los datos de salida para esta fase son:

Elementos de entrada:

- Documento de texto generado en la fase anterior con definición de preguntas de investigación, hipótesis y variables dependientes e independientes además de su respectivo perfilado,
- Técnicas (algoritmos) y elementos para el análisis.
- Infraestructura y herramientas de software instaladas.

Proceso de la fase:

3.1 Recolectar una muestra representativa (*Actividad necesaria*).Se obtiene un conjunto de datos de la población empleando alguna de las técnicas de muestreo que permita producir una muestra aleatoria y representativa de toda la población.

3.2 Definir conjunto de criterios para el proceso de carga, limpieza y transformación de los datos (*Actividad necesaria*). Esta actividad forma parte del proceso de conocer los datos los cuales es muy probable que vengan en distintos formatos y con errores de semántica o de inconsistencias, por lo que es necesario determinar diferentes criterios para cargarlos, limpiarlos, transformarlos y verificarlos que contienen el valor suficiente para obtener ideas nuevas antes de aplicar las técnicas o algoritmos identificados para el análisis.

3.3 Aplicar las técnicas o desarrollar los algoritmos identificados en la fase anterior (*Actividad necesaria*). Se aplican las técnicas y/o algoritmos previamente identificados haciendo uso de la infraestructura y herramientas de software con la finalidad de obtener información que permita detectar la situación problemática; o para determinar, si es el caso, si la solución no es viable para continuar con el proyecto.

3.4 Presentar los resultados obtenidos con los datos muestra en graficas o tablas. (*Actividad necesaria*). Se realizan graficas o tablas para mostrar los resultados obtenidos del análisis exploratorio para tener una idea de como se comportan los datos.

3.5 Determinar si las técnicas y las herramientas son las adecuadas para el análisis (*Actividad necesaria*). Tomando como base los resultados presentados en la actividad anterior se determina si las técnicas y herramientas brindan información útil para resolver el problema o en su defecto es necesario aplicar algún ajuste para obtener los resultados esperados.

Elementos de salida:

- Conjunto de criterios para aplicar el proceso de carga, limpieza y transformación de los datos.
- Identificación de ajustes en las técnicas y/o algoritmos para el análisis.
- Visto bueno para proseguir con la siguiente fase o por el contrario rechazar los resultados con la fuente de datos y regresar a la fase anterior.

Fase 4. Recolección de los datos.

En esta fase se procede a obtener los datos de las fuentes identificadas las cuales pueden ser generadas por la misma organización o por terceros. De acuerdo con (Guo, 2013) los problemas más comunes a los que se enfrentan los equipos de ciencia de datos son: la procedencia, la administración y el almacenamiento de los datos.

Para llevar a cabo esta fase se requiere utilizar hardware compuesto por servidores físicos y/o servidores virtuales los cuales pueden estar instalados físicamente en la organización o pueden estar utilizando servicios de la nube. Anteriormente estos servidores se encargaban de almacenar los datos provenientes de sistemas de información (sistemas de soporte de decisiones (DSS por las siglas en ingles de Decision Support System), sistemas de planeación de recursos empresariales (ERP por las siglas de Enterprise Resource Planning) y los sistemas de administración de relación con los clientes (CRM por las siglas de Customer Relationship Management), sistemas de información geográficos, etc.), sin embargo, ahora con el internet de las cosas estos servidores pueden ser suministrados con datos provenientes de diferentes dispositivos electrónicos como los teléfonos inteligentes, sensores electrónicos, satélites, flujo de datos de la web, las redes sociales, sistemas operativos, etc.

Dentro del software que se utiliza en esta capa se puede encontrar a los RDBMS que es en donde por lo general la gran mayoría de las organizaciones almacenan datos estructurados referentes a las actividades que realizan día con día, mientras que las bases de datos NOSQL las utilizan para almacenar información estructurada o no estructurada de otra clase de proyectos.

Si las fuentes de datos son generadas por terceros el proceso de recolección es diferente ya que es necesario aceptar ciertas condiciones que pone la organización proveedora de los datos para poder establecer un medio de comunicación entre ambas organizaciones. Después de haber aceptado las condiciones, el problema de la transferencia de los datos de un lugar a otro lo resuelve la organización proveedora poniendo a disposición el uso de API.

Los roles que trabajan en esta fase son el investigador de datos y el desarrollador de los datos donde su principal objetivo es desarrollar los scripts, definir las técnicas y gestionar las herramientas para recolectar los datos. El innovador de datos deberá supervisar la recolección para agilizar el proceso.

Los datos de entrada, el proceso y los datos de salida para esta fase son:

Elementos de entrada:

- Fuentes de Big Data a trabajar.
- Elementos para conectarse a los servicios y extraer los datos.
- Infraestructura y herramientas de software instaladas.

Proceso de la fase:

4.1 Preparar y configurar las herramientas de software para almacenar los datos utilizando los elementos para conectarse a los servicios del proveedor de la fuente de Big Data.

4.2 Puesta en marcha del servicio para la recolección de los datos.

Elementos de salida:

- Datos recolectados

Fase 5. Preparación de los datos.

Después de haber obtenido los datos, éstos pasan por un proceso que se encarga de transformarlos a un formato aceptable, validarlos y estandarizarlos utilizando software o algoritmos diseñados por terceros o por los mismos investigadores de datos para realizar la limpieza, formateo y agregación para lograr la validez y la congruencia que se necesita para que el análisis tenga los resultados esperados. Por esta razón es importante tener en cuenta los aspectos definidos en la formulación del problema ya que guiarán la preparación de los datos por el camino adecuado.

Para llevar a cabo esta fase es necesario tener en consideración el empleo de sistemas de archivos distribuidos que están formados comúnmente por un nodo maestro interconectado con un conjunto de nodos esclavos, estos almacenan información de manera redundante para obtener una máxima eficiencia y rendimiento de los datos. De lo anterior que también se requiera de una tecnología para procesamiento en paralelo que tiene la finalidad de distribuir el cómputo de los datos en distintos equipos para agilizar el procesamiento siendo esta una característica fundamental en el manejo de Big Data.

Esta es una de las fases más complicadas de realizar ya que cada fuente de datos presenta características que hacen que estos procesos sean únicos generando sus propios retos por enfrentar.

Pasando a la parte del hardware que se necesita en esta capa se encuentran los clústeres de equipos de cómputo que almacenarán los datos de manera redundante para su procesamiento en paralelo. En este punto es importante que la organización tome en cuenta algunas consideraciones como son el rendimiento, la disponibilidad, la escalabilidad, la flexibilidad a fallos, el costo (clústeres locales o la nube), entre otras para preparar los datos. Adicionalmente, es imprescindible tener conexiones de red, en las que, dependiendo del rendimiento y la flexibilidad que se desee tener, cuente con canales redundantes con la suficiente capacidad que permita manejar las grandes cantidades de información que pudiesen llegar a la organización.

Esta fase forma parte de un proceso iterativo que utiliza el método de prueba y error por lo que es común encontrar resultados no esperados durante las primeras ejecuciones. Los roles involucrados son el investigador de datos y el desarrollador de los datos donde su principal objetivo es desarrollar los scripts y aplicar las técnicas de limpieza, normalización, procesamiento e integración de las fuentes de datos; y al igual que en la fase anterior, el innovador de datos deberá supervisar la preparación de los datos para garantizar la calidad y validez de los mismos.

Los datos de entrada, el proceso y los datos de salida para esta fase son:

Elementos de entrada:

- Datos recolectados
- Infraestructura y herramientas de software instaladas.
- Conjunto de reglas para aplicar el proceso de carga, limpieza y transformación de los datos.

Proceso de la fase:

5.1 Aplicar las técnicas o desarrollar algoritmos haciendo uso de la infraestructura y herramientas de software identificados en la fase anterior para cargar, limpiar y transformar los datos para que queden en un formato valido y estandarizado.

Elementos de salida:

- Datos limpios, validados y formateados de acuerdo a las necesidades del estudio.

Fase 6. Análisis de datos y de resultados.

Esta fase tiene como principal objetivo definir los modelos para analizar los datos, estructurados o no estructurados, y posteriormente obtener conocimiento mediante la ejecución de algoritmos y la revisión de resultados. Al igual que la fase anterior, forma parte de un

proceso iterativo donde se requiere realizar múltiples revisiones entre los involucrados con la finalidad de que las actividades a realizar arrojen los valores esperados y la presentación de resultados quede lo más entendible posible. Los roles que participan en su desarrollo son el innovador de datos y el investigador de datos además. Al finalizar el análisis, se reunirá la persona de negocio de datos con el equipo involucrado en la fase con la finalidad de revisar los resultados y hacer sus respectivas retroalimentaciones.

Los datos de entrada, el proceso y los datos de salida para esta fase son:

Elementos de entrada:

- Datos limpios, validados y formateados de acuerdo a las necesidades del estudio.
- Infraestructura y herramientas de software instaladas.
- Técnicas (algoritmos) y elementos ajustados en la fase del análisis exploratorio.

Proceso de la fase:

La descripción de las actividades a realizar son:

6.1 Construcción del modelo (*Actividad opcional*). En esta fase se seleccionan o se crean los modelos estadísticos que permitirán predecir de mejor manera el comportamiento del problema y su solución. Para esta fase se recomienda utilizar muestras significativas de los datos con la finalidad de agilizar el proceso de prueba, validación y evaluación tanto del modelo como de los algoritmos convirtiéndose en un proceso iterativo utilizado el método de prueba y error para recalibrarlos y asegurarse de que el análisis no es demasiado optimista o demasiado ajustado.

El hardware utilizado en esta fase es exactamente el mismo que la anterior, un clúster de equipos de cómputo redundantes interconectados a la red para poder ejecutar algoritmos de procesamiento en paralelo. En la parte del software se pueden encontrar lenguajes de programación y éstos pueden o no implementar las funciones de MapReduce los cuales se pueden utilizar para realizar análisis personalizados.

Además de contar con las herramientas de hardware y software es indispensable tener un amplio conocimiento de técnicas estadísticas con la finalidad de proveer una mejor solución para el tipo de análisis de interés.

6.2 Evaluación del modelo (*Actividad opcional*). Se prueba, se validan y se evalúan los modelos y los resultados con los involucrados en el negocio teniendo en consideración los supuestos hechos en la formulación del problema con la finalidad de que si es necesario redirigir el análisis en caso de no haber ideas útiles se recalibren los parámetros del modelo, se agreguen o cambien fuentes de datos; o para reportar y tomar acciones con los resultados obtenidos para resolver la situación problemática y ayudar a la toma de decisiones.

6.3 Representación y visualización de resultados (*Actividad necesaria*). La presentación de los datos se realiza con la finalidad de facilitar el análisis de los datos a los usuarios finales mediante el uso de algunos servicios y herramientas para hacerlas altamente interactivas y dinámicas.

Elementos de salida:

- Gráficas, tablas, reportes o resúmenes de las cifras obtenidas de analizar los datos.

Fase 7. Despliegue.

En esta fase el modelo de análisis ya fue validado y el negocio comienza a ver los resultados del proyecto mediante la toma de decisiones. Los roles involucrados en esta fase son la persona de negocio de datos y el innovador de los datos quienes buscarán las estrategias que ayuden a resolver el problema.

Los datos de entrada, el proceso y los datos de salida para esta fase son:

Elementos de entrada:

- Gráficas, tablas, reportes o resúmenes de las cifras obtenidas de analizar los datos.

Proceso de la fase:

7.1 Aplicación con los datos (*Actividad opcional*). Si el resultado del análisis es un producto, entonces se identifica el proceso y se implanta mediante interfaces de escritorio, en la web, dispositivos móviles, etcétera con el objetivo de obtener algún retorno de la inversión.

7.2 Comunicación de resultados (*Actividad necesaria*). A pesar de que los resultados del análisis se han estado revisando por parte de los interesados del estudio, en esta fase se realiza la entrega formal de los resultados obtenidos del análisis utilizando las técnicas y herramientas de presentación y visualización definidas en la fase anterior para tomar decisiones.

Elementos de salida:

- Producto para mejorar un proceso de negocio, o
- Comunicación o presentación de los resultados obtenidos del estudio empleando distintas formas como la publicación en algún medio organizacional, revista científica o en un congreso.

Fase 8. Monitoreo y medición de resultados.

En esta fase se monitorea y se miden los resultados obtenidos en la empresa de aplicar las ideas producto del análisis. Si se llega a presentar alguna situación fuera de lo planeado habrá que replantear la solución y hacer los ajustes necesarios para que el modelo vuelva a representar la situación problemática. Si el estudio es realizado con fines exploratorios, las actividades de monitoreo y medición de los resultados consta en la verificación y cumplimiento de los criterios de aceptación establecidos por el experto del área. Los roles involucrados en esta fase son la persona de negocio de datos y el innovador de los datos.

Los datos de entrada, el proceso y los datos de salida para esta fase son:

Elementos de entrada:

- Producto para mejorar un proceso de negocio

Proceso de la fase:

Medición (*Actividad opcional)*. Para poder medir la situación actual de un proceso de negocio es necesario recolectar datos de un conjunto de características que al analizarlas e interpretarlas puedan brindar la información requerida para en un momento dado compararla contra la generada en otra fecha en particular. El tiempo, el costo, la calidad, la satisfacción de los clientes y la frustración de los trabajadores son algunos ejemplos de características que se pueden medir, aunque es importante comentar que la importancia o el valor que brinden estas varían dependiendo del proceso que se desea medir.

8.1 Monitoreo (*Actividad opcional*). Si el resultado del estudio es un producto implantado en un proceso de la organización es necesario que dicho producto sea monitoreado de manera regular con la finalidad de verificar que todo va conforme a lo planeado o para detectar a tiempo posibles cambios que afecten el proceso y se requiera por ende actualizar el modelo. Esta actividad va de la mano con la de medición ya que se utilizan las métricas generadas anteriormente y se determina si el resultado es el esperado o de no ser así es necesario realizar las modificaciones pertinentes con el fin de mejorarlo empleando diferentes estrategias. En este caso es común ver que se utilice la noción de estrategias de decisión campeón-retador donde la estrategia actual es denominada estrategia campeón y una nueva estrategia como estrategia retadora en donde esta última se pone a prueba para ver si mejora los resultados. Si después de una observación parece que la estrategia retadora está funcionando mejor, ésta sustituye la estrategia campeón.

Elementos de salida:

- Identificación de estrategias para mejorar la solución tomando como base las métricas generadas.

Resumen del método propuesto.

A manera de resumen se muestra en la Figura 19 el flujo completo y las actividades de cada una de las fases del método.

Fase 1 **Descripción del problema**	1. **Identificación del problema***(Actividad necesaria)* 2. **Impacto al negocio***(Actividad necesaria)* 3. **Antecedentes***(Actividad necesaria)* 4. **Condiciones en las que ocurre***(Actividad necesaria)* 5. **Definición de objetivos***(Actividad necesaria)* 6. **Identificación de las fuentes de datos a utilizar***(Actividad necesaria)* 7. **Definición del alcance***(Actividad necesaria)*
Fase 2 **Diseño conceptual de la investigación**	1. **Definición de preguntas de investigación***(Actividad necesaria)* 2. **Definición de hipótesis** 3. **Detección de variables** 4. **Perfilar variables***(Actividad necesaria)* 5. **Selección de técnicas y elementos para el análisis***(Actividad necesaria)* 6. **Definición de infraestructura y herramientas de software***(Actividad necesaria)* 7. **Definición de la arquitectura***(Actividad necesaria)* 8. **Elaboración de programa de trabajo***(Actividad necesaria)*
Fase 3 **Análisis exploratorio**	1. **Recolectar una muestra representativa** *(Actividad necesaria)* 2. **Definir conjunto de criterios para el proceso de carga, limpieza y transformación de los datos** *(Actividad necesaria)* 3. **Aplicar las técnicas o desarrollar los algoritmos identificados en la fase anterior** *(Actividad necesaria)* 4. **Presentar los resultados obtenidos con los datos muestra en graficas o tablas** *(Actividad necesaria)* 5. **Determinar si las técnicas y las herramientas son las adecuadas para el análisis** *(Actividad necesaria)*
Fase 4 **Recolección de los datos**	1. **Preparar y configurar las herramientas de software para almacenar los datos utilizando los elementos para conectarse a los servicios del proveedor de la fuente de Big Data** *(Actividad necesaria)* 2. **Puesta en marcha del servicio para la recolección de los datos** *(Actividad necesaria)*
Fase 5 **Preparación de los datos**	1. **Aplicar las técnicas o desarrollar algoritmos para cargar, limpiar y transformar los datos** *(Actividad necesaria)*
Fase 6 **Análisis de datos y de resultados**	1. **Construcción del modelo** 2. **Evaluación del modelo** 3. **Representación y visualización de los datos** *(Actividad necesaria)*
Fase 7 **Despliegue**	1. **Aplicación con los datos** *(Actividad opcional)* 2. **Comunicación de los resultados** *(Actividad necesaria)*
Fase 8 **Monitoreo y medición de los resultados**	1. **Medición** *(Actividad opcional)* 2. **Monitoreo***(Actividad opcional)*

Figura 19. Descripción detallada de las fases junto con las actividades a realizar en el método. Elaboración propia.

Características del método.

Como se puede apreciar, el método comparte mucha similitud con las metodologías y procesos de ciencia de datos revisados en la literatura pero difiere de estos gracias a la iteratividad de ciertas fases que se consideran críticas para determinar la viabilidad de encontrar la solución al problema de investigación o en su defecto atacarlo mediante el empleo de otro tipo de soluciones. Otras características que tiene el método son:

- Integrar un conjunto de actividades que han sido utilizadas en proyectos de análisis y ciencia de datos en muchos dominios para formar un proceso para realizar proyectos de análisis de Big Data.
- Es ajustable a las necesidades de estudio de la organización gracias a que las actividades del método se pueden acomodar con facilidad y realizar únicamente aquellas que se requieran para la solución del problema. Esto trae ventajas como el aumento de la eficiencia de los involucrados en el proyecto al trabajar específicamente sobre lo planeado y el ahorro de recursos como el tiempo y costos de implementación
- Permite aplicar el método científico al llevar un control del proceso realizado para automatizarlo y replicarlo tantas veces sea necesario obteniendo los mismos resultados.
- Debido a que con el análisis de Big Data se puede extraer conocimiento de una situación particular del negocio que anteriormente era complicado de obtener, el empleo del método propuesto ayuda a cambiar de ideología del proceder en la toma de decisiones de la organización por medio de corazonadas por otra que se basa en el resultado de los análisis de los datos.
- Al tener bien identificadas las fases del método y el personal involucrado que se necesita es posible designar de mejor manera los tiempos a las fases más pesadas sin quitarle la importancia a aquellas que por ser más sencillas no dejan de tener un papel sustancial.
- Al tener como origen los estudios exploratorios realizados en el INEGI, los roles presentados en este documento pueden ser la base de un área dedicada para la ciencia de datos en otra organización,
- Encaminar a usuarios principiantes en la obtención de estadísticas mediante el análisis de grandes volúmenes de datos.

Adicionalmente a las características en la Tabla 5 se muestra una comparativa de las actividades que incluyen las metodologías revisadas, los procesos de ciencia de datos y el método propuesto en este trabajo de investigación.

Tabla 5. Actividades de las metodologías revisadas y el método propuesto. Elaboración propia.

Actividades	Trabajos revisados								Este Trabajo
	1	2	3	4	5	6	7	8	9
Analizar y evaluar el negocio	X	X	X	X			X	X	X
Encuadrar el problema	X	X	X	X			X	X	X
Revisar antecedentes del problema	X	X		X					X
Elaborar plan del proyecto							X		X
Identificar las fuentes de datos				X					X
Recolectar datos de muestra	X								X
Realizar análisis y descubrimiento en los datos	X			X			X		X
Establecer acciones para garantizar la seguridad			X						
Desarrollar hipótesis de negocio	X	X		X				X	X
Armar casos de uso	X								
Identificar variables (dependientes e independientes)		X		X					X
Desarrollar el enfoque de análisis	X			X					X
Evaluar casos de uso	X								
Identificar los métodos y algoritmos(técnicas) para el análisis	X	X	X	X			X		X
Diseñar estrategias de decisión de negocio				X					X
Construir y preparar el conjunto de datos	X	X	X		X		X		X
Adquirir los datos	X	X	X		X	X	X	X	X
Perfilar los datos				X					X
Entender los datos	X	X	X	X			X		X
Verificar calidad de los datos							X	X	X
Integrar los datos de diferentes fuentes					X	X	X		
Aplicar técnicas de ETL	X	X	X	X	X	X	X	X	X
Identificar plataformas y herramientas		X							X
Evaluar plataformas y herramientas		X					X		X
Seleccionar y construir los modelos analíticos	X	X	X	X	X		X	X	X
Construir el modelo	X	X		X	X	X	X	X	X
Probar y validar el modelo con los datos	X	X		X	X	X	X	X	X
Definir técnicas de visualización de datos	X		X					X	
Evaluar resultados	X	X	X	X		X	X		
Implementar estrategias de decisión de negocios				X					X
Construir el sistema para producción	X	X	X	X					X
Desarrollar el estado final de la solución	X	X	X	X	X	X	X	X	X
Diseñar e implementar el proceso de negocio	X	X	X	X		X	X	X	X
Medir y monitorear	X	X	X	X					X
Medir la efectividad de la solución	X		X	X	X				X
Calibrar los modelos (mantenimiento)	X			X	X		X		X
Monitorear la solución y sus beneficios	X			X			X		X
Establecer ciclos de retroalimentación para futuras mejoras	X	X	X	X					X

Relación de trabajos revisados contra el método propuesto

1. |Mohanty (Mohanty et al., 2013)
2. Raghupathi (Raghupathi & Raghupathi, 2014)
3. Mousannif (Mousannif et al., 2014)
4. Sheikh (Sheikh, 2013)
5. Jagadish (Jagadish et al., 2014)
6. Guo (Guo, 2013)
7. Shearer (Shearer, 2000)
8. Cuesta (Cuesta, 2013)
9. Este trabajo

Aun cuando la relación de metodologías de análisis y procesos de ciencia de datos revisados tienen diferentes enfoques, tienen muchas actividades que en común y otras que están enfocadas para dominios de estudio específicos.

Al momento de conformar aquellas actividades que se consideraron para formar parte del método, de las cuales se tomaron como base las que se realizaban en el INEGI, algunas de ellas se englobaron dentro de otras actividades debido a las características que presentaban para facilitar el seguimiento y la comprensión del método por parte de los involucrados en el proyecto. Por otro lado, un pequeño número de actividades fueron las que no se integraron dentro del método por considerar que no tienen tanto aporte a la investigación como las demás. Dichas actividades fueron: Establecer acciones para garantizar la seguridad, evaluar casos de uso, integrar los datos de diferentes fuentes, definir técnicas de visualización de datos y evaluar resultados.

Aprobación del método

Como parte de la aprobación del método se identificaron dos cosas que tenían que demostrarse al aplicar el método en los casos prácticos, la primera fue en presentar resultados claros y en la obtención de los resultados esperados cumpliendo con las expectativas de los interesados.

Capítulo 6. Resultados obtenidos al aplicar el método propuesto

Para probar el método propuesto se diseñaron dos casos prácticos que tienen un alto impacto social por el conocimiento que se puede obtener de ellos. El primero es el análisis de movilidad cotidiana de los usuarios de la red social de Twitter que publican dentro del territorio nacional mediante el cual se buscó encontrar patrones de desplazamiento que tuvieron en un determinado tiempo. Este tipo de estudios son útiles en los procesos que realizan las dependencias de gobierno para la planeación de rutas de transporte público y la construcción y asignación de los flujos de carreteras y vialidades.

Y por otro lado, el análisis de impacto de eventos se buscó medir la forma en la que un evento de la vida real impacta en la sociedad mediante el estudio del número de menciones que publican los usuarios de Twitter contra el tiempo de duración de dichos eventos. Los eventos a analizar en la red social son acontecimientos que tuvieron un alto número de consultas en el motor de búsqueda de Google teniendo como referencia lo generado con la aplicación de tendencias de google (google trends por su traducción al inglés).

Aplicación del método en el caso práctico de análisis de movilidad

Fase 1. Descripción del problema

1.1 Identificación del problema

Los estudios de exploración de movilidad de una población tienen gran importancia a nivel mundial ya que con ellos es posible conocer el comportamiento de diferentes situaciones como los flujos de migración y de circulación, la actividad turística, transporte y tráfico, dispersión de enfermedades y epidemias, etc. que se dan en determinada área geográfica. Sin embargo, algunas de las principales problemáticas a la que se enfrentan los investigadores que realizan estos estudios es que anteriormente los datos eran obtenidos de dos maneras diferentes. La primera era haciendo uso de las estadísticas oficiales tradicionales como los censos, registros administrativos y encuestas especializadas realizadas por las ONE en fechas muy específicas por lo que era necesario ser muy selectivo en cuanto a cuando realizar el estudio y tener bien identificado en qué áreas geográficas se levantaban esos datos para ver cuáles de ellas se podrían considerar. Y la segunda era mediante la contratación de una empresa que prestara el

servicio para realizar encuestas donde la ventaja era que se podría programar en un tiempo y área geográfica específica pero con la limitante que los costos de este tipo de levantamiento de datos son muy altos y necesitaban muchas horas- hombre. Aunado a esto, como en la República Mexicana cada Estado es autónomo y no hay fronteras internas entre cada uno de ellos es sumamente complicado detectar la dinámica de movimientos de las personas, trayendo en sí que las maneras de recolectar los datos implique tener considerado las situaciones previamente expuestas y de aquí surge el interés de este estudio.

1.2 Impacto al negocio

Esta investigación se realizó en conjunto con otras pruebas piloto como un experimento para determinar si se cuenta con los conocimientos y los elementos necesarios para emprender el desarrollo de otros proyectos de Big Data, y del mismo modo para realizar análisis exploratorios con la fuente de datos seleccionada para diagnosticar el valor que ofrece al INEGI por analizarla.

1.3 Antecedentes

Entre los estudios que ha realizado el INEGI a través de los años en las grandes aglomeraciones urbanas del país se encuentran las EOD levantada para el área metropolitana de la Ciudad de México en 1994 y la EOD 2007 de la ZMVM.

Por otro lado también es posible encontrar en México levantamientos de EOD realizados por empresas privadas como la del año 1998 realizada por la empresa ARHSA para la Zona Conurbada de Tampico-Madero-Altamira (Izquierdo, 2008), la EOD levantada en el año 2000 por la empresa USTRAN (http://www.ustran.com/) para la Zona Metropolitana de San Luís Potosí; y la realización de una EOD en la Avenida de los Insurgentes (Ciudad de México) en el 2004 por la misma empresa (Izquierdo, 2008).

1.4 Condiciones en las que ocurre

La mayor parte de estas investigaciones se nutren de datos recabados por organizaciones privadas o por estudios oficiales donde el principal objetivo está vinculado a programas de

desarrollo urbano, tanto municipales como metropolitanos y estatales (Izquierdo, 2008); y de turismo (Fragoso et al., 2014).

Para el INEGI este problema ocurre tras la necesidad de realizar una serie de análisis exploratorios con fuentes de Big Data para generar información actualizada en el tema de movilidad cotidiana que registran los usuarios de la red social de Twitter. Una de las investigaciones piloto del INEGI con la participación y en coordinación con organismos nacionales como la Secretaría de Turismo fue un proyecto que utilizó la información extraída de Twitter para identificar el turismo en Puebla y Guanajuato los días 1, 2 y 3 de Febrero de 2014 donde posteriormente se comparó con estadísticas oficiales de los observatorios turísticos de ambos Estados.

1.5 Definición de objetivos

El objetivo de este caso práctico fue planteado desde el comienzo de este trabajo y fue formulado de manera precisa y clara de manera que no existiera ambigüedad respecto al tipo de respuesta esperado. El objetivo es declarado en el siguiente párrafo:

"Probar el método propuesto para analizar la capacidad de determinar patrones de movilidad mediante el análisis de los metadatos de las conversaciones originadas en Twitter utilizando los parámetros de posición geográfica, tiempo y frecuencia."

1.6 Identificación de las fuentes de datos a utilizar

Los datos fueron obtenidos de la red social de Twitter porque aparte de que permite identificar el origen geográfico de los tweets publicados, de ser públicos y gratuitos (el uno por ciento del total de tweets generados a nivel mundial) cumple con las características necesarias para considerarlo como Big Data (velocidad, volumen y variedad) por ser una fuente de datos no estructurada al manejar los tweets en formato JSON los cuales se generan a grandes velocidades teniendo que el INEGI recolectó en un total de 41,786,137 de Enero de 2014 a Febrero de 2015. Otra de las razones por las que se decidió elegir esta fuente de datos es por la facilidad que brinda para acceder a los tweets a través de las distintas API para el desarrollo de aplicaciones.

1.7 Definición del alcance

Para la realización de este caso práctico se definieron un conjunto de reglas que sirvieron para determinar lo que está dentro y fuera de las fronteras del estudio para llevar un mejor control y no desviarse de los objetivos planteados.

Los datos fueron obtenidos de la red social de Twitter durante el periodo de Enero de 2014 a Febrero de 2015 utilizando un filtro para recolectar únicamente aquellos tweets que fueron publicados con su respectiva referencia geográfica y que estén dentro de las coordenadas que enmarcan el territorio de la República Mexicana mostrado en la Figura 20 (INEGI, 1991) y son:

- Norte: 32° 43´ 06´´ latitud norte o 32.71865357 en representación decimal, en el Monumento 206, en la frontera con los Estados Unidos de América (3 152.90 kilómetros).
- Sur: 14° 32´ 27´´ latitud norte o 14.53209836 en representación decimal, en la desembocadura del río Suchiate, frontera con Guatemala (1 149.8 kilómetros).
- Este: 86° 42´ 36´´ longitud oeste o -86.71040527 en representación decimal, en el extremo suroeste de la Isla Mujeres.
- Oeste: 118° 27´ 24´´ longitud oeste o -118.40764955 en representación decimal, en la Punta Roca Elefante de la Isla de Guadalupe, en el Océano Pacífico.

Figura 20. Mapa de la República Mexicana con división política y ejes geográficos Fuente: (INEGI, 1991)

Fase 2. Diseño conceptual de la investigación

2.1 Definición de preguntas de investigación

La pregunta de investigación de este caso práctico fue planteada desde el comienzo de este trabajo la cual fue formulada de manera precisa y clara de manera que no existiera ambigüedad respecto al tipo de respuesta esperado. La pregunta es realizada en el siguiente párrafo:

¿Es posible encontrar patrones de movilidad utilizando el método propuesto mediante el análisis de los metadatos de las conversaciones originadas en Twitter utilizando los parámetros de posición geográfica, tiempo y frecuencia?

2.2 Definición de hipótesis

Como en este caso práctico no se trata de pronosticar ningún dato o hecho en las variables sino más bien describir el comportamiento de los datos referente a la movilidad de los usuarios de Twitter, no es necesaria la definición de hipótesis por lo que no aplica este punto del método.

2.3 Detección de variables

Igual que el punto anterior, la actividad relacionada con la detección de variables dependientes e independientes del método no aplica debido a que no existen hipótesis a probar por las características del estudio.

2.4 Perfilar variables

Teniendo como base la estructura del tweet revisado en el marco teórico se detectaron que las siguientes variables son útiles para la elaboración del ejercicio de movilidad:

- Created_at: Variable que contiene información del tiempo universal coordinado (UTC acrónimo en inglés de Coordinated Universal Time) en el que es creado el tweet. El UTC es el estándar de tiempo por el cual el mundo regula la hora tomando como referencia un reloj atómico que se ajusta de acuerdo a las medidas de rotación

de la tierra. Un ejemplo de la manera en la que viene el dato es: Ago 27 13:08:45 +0000 2015.

- User.name: Es el nombre del usuario que publicó el tweet. Este campo solo se utilizará para identificar la movilidad de cada uno de los usuarios y posteriormente se eliminará del proceso de análisis para proteger la identidad de los mismos.
- Place.full_name: Indica el lugar de origen del que fue publicado un tweet. Un ejemplo de la manera en la que viene el dato es: Guadalajara, Jalisco.
- Coordinates. Representa la ubicación geográfica desde donde se publico el tweet.

2.5 Selección de técnicas y elementos para el análisis

El tipo de análisis que se realizó sobre el conjunto de datos recolectados de la fuente de datos seleccionada fue empleando técnicas de estadística descriptiva, la cual tiene como principal objetivo poner de manifiesto las características más importantes de los datos y sintetizarlas en gráficas y tablas.

De los elementos externos que se utilizaron para el análisis y la representación visual de los datos están:

- Servicio web de mapas (WMS por las siglas en ingles de Web Map Service) para consultar información cartográfica del Marco Geoestadístico Nacional (MGE) en las capas de límites estatales y municipales.
- Catálogo único de claves de áreas geoestadísticas estatales, municipales y localidades.

Dos de las técnicas o algoritmos que se utilizaron para la preparación de los datos y su análisis son:

Algoritmo para la obtención de Estados y Municipios partiendo de las referencias geográficas de los tweets

Para identificar el Municipio y el Estado desde donde el usuario de Twitter está enviando un mensaje se utilizó un algoritmo desarrollado por personal del INEGI consiste básicamente en realizar una análisis espacial con la referencia geográfica obtenida de cada uno de los tweets para después buscarla dentro de un conjunto de geometrías representativas de los Estados y

Municipios de la República Mexicana para identificar cuál de ellas encierra el punto correspondiente a la referencia geográfica del tweet y así obtener las claves indicadas para poder realizar los análisis posteriores.

Algoritmo para la detección de la movilidad

El algoritmo que se siguió para determinar si un usuario de la red social viajó entre Municipios y Estados es el siguiente:

1. Obtener todos los tweets que haya publicado un usuario en particular durante todo el tiempo de recolección.
2. Identificar si al menos dos tweets del usuario fueron publicados en diferentes áreas geográficas.
    a. Si se cumple el criterio anterior
        i. Extraer todos los registros de los usuarios junto con las fechas involucradas ordenándolas por este último campo.
        ii. Se itera el total de registros tomando y comparando cada dos registros.
            1. Si los dos registros en el campo de ubicación geográfica a nivel Municipio son iguales.
                a. No hay movilidad, el usuario está en el mismo Municipio.
            2. Si los registros en el campo de ubicación geográfica son diferentes.
                a. Hay movilidad el usuario twitteó en otro Municipio, se agregan los registros en un arreglo exclusivo.
    b. Si no cumple el criterio
        i. No se considera al usuario para el análisis.

2.6 Definición de la infraestructura y herramientas de software

Con la finalidad de realizar distintas pruebas y demostrar que cualquier organización puede almacenar, procesar y analizar los datos de Twitter, se instaló una infraestructura compuesta por seis equipos de cómputo interconectados entre sí y con salida a internet para poder recolectar los datos provenientes del API de Twitter.

Las características técnicas del clúster de equipos de cómputo son:

- Computadora marca HP modelo Compaq 6735B.
- AMD Turion X2 Ultra dual core a 1.4 GHZ.
- Disco duro 160 GB.
- Memoria RAM con 4 GB.
- Sistema Operativo de 64 bits con Linux.

Para realizar la preparación de los datos, los análisis exploratorios, los anáñisis definitivos y la visualización de resultados se ocupó únicamente un equipo de cómputo con las siguientes características técnicas:

- Computadora marca Dell modelo Optiplex 9020.
- Intel Core i7 4790 a 3.6 GHZ con 8 MB de cache.
- Disco duro de estado sólido de 128 GB.
- Disco duro SATA de 1 TB a 7200 RPM.
- Memoria RAM con 16 GB.
- Tarjeta gráfica PCIe con puerto HD de 2 GB.
- Sistema operativo Windows 8.1 Enterprise.

Hablando de hardware, y como se vio anteriormente, es de vital importancia contar para cada una de las capas de la arquitectura con la infraestructura tecnológica ya que es el elemento que se relaciona directamente con todos los procesos proveyendo la comunicación, el almacenamiento y la capacidad de procesamiento de la gran cantidad de datos con la que se cuenta.

A continuación se enlistan las plataformas y las herramientas de software que se utilizó para realizar el estudio.

- Cuenta de usuario de Twitter.
- Aplicación para desarrollador en el API de Twitter.
- Claves de autenticación. (Clave del consumidor, secreto del consumidor, token de acceso, y el acceso secreto del token)

- Software para la recolección de datos, para el caso del INEGI de una fuente externa. El software a utilizar es ElasticSearch.
- Herramientas para la sincronización con los servicios del API de Twitter. El software de Logstash, el plugin de entrada de datos de Twitter con su funcionalidad extendida y el plugin de salida de Elasticsearch para el almacenamiento de los datos.
- La carga se llevó a cabo apoyándose del plugin Spark-csv desarrollado por la compañía Databricks y el almacenamiento de los datos se realizó sobre Apache Spark.
- Para la transformación y procesamiento de los datos se desarrollaron scripts en el lenguaje de programación orientado a funciones, Scala, en conjunto con la librería SQL de Apache Spark.
- Para la visualización de los resultados en forma de mapas se utilizó el sistema de información geográfica de código abierto QGIS y el lenguaje de programación Python.
- Para la generación de gráficas y tablas se utilizó Microsoft Excel.
- Para la generación de estadísticas descriptivas de los datos se empleó el lenguaje de programación R.

2.7 Definición de la arquitectura

Los elementos, técnicas y herramientas vistas en las actividades anteriores son primordiales para definir la arquitectura que se utilizó este caso práctico la cual fue conformada por aquellas capas que tenían en común tanto las revisadas en el marco teórico como la que se estaba utilizando en el INEGI para los análisis exploratorios trabajados teniendo como resultado una arquitectura que se compone de cuatro capas: la capa de fuentes de datos; la de preparación y almacenamiento de los datos; modelado y análisis; y finalmente la capa de visualización y reportes. En la Figura 21 se puede apreciar de manera más clara como es que se relacionan los elementos (plataformas y herramientas) identificados para trabajar con Twitter dentro del INEGI.

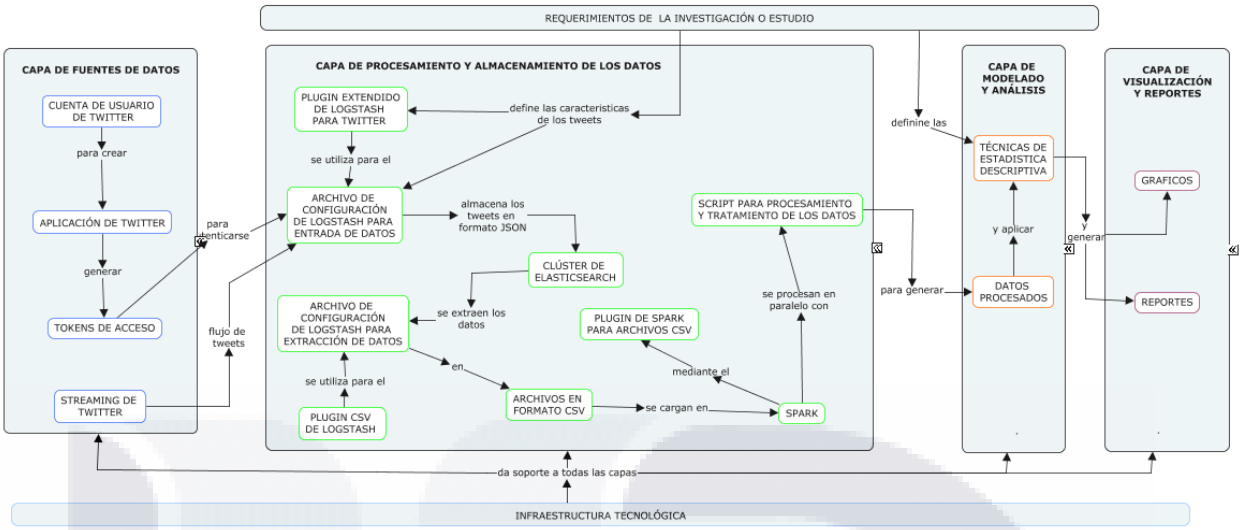


Figura 21. Mapa conceptual de elementos. Elaboración propia.

2.8 Elaboración de programa de trabajo

Como el método contiene fases y actividades que pueden realizarse tantas veces como sea necesario el programa de trabajo consistió en realizar revisiones periódicas cada dos a cuatro semanas en las que se presentaban los avances realizados para determinar si el desarrollo del proyecto iba de acuerdo a lo especificado en los objetivos.

En la Tabla 6 se presenta un cronograma de trabajo en el que se muestran las actividades del método a partir de la Fase 3 especificando que rol de los involucrados en el proyecto la realiza y el tiempo planeado en una iteración para el caso práctico de análisis de movilidad. Para este caso práctico no se consideraron relevantes contemplar las primeras fases del método para el cronograma por ser fases de definición del proyecto, sin embargo pueden llegar a considerarse para otros proyectos. Para la actividad de recolección de la muestra representativa, perteneciente a la fase 3 y toda la fase 4 referente a la recolección de los datos el tiempo que tomó obtenerla para este ejercicio fue de cero días ya que el INEGI ya había obtenido con los tweets anteriormente.

Tabla 6. Cronograma de trabajo con roles asignados y duración aproximada en una iteración, Elaboración propia.

Fase	Actividad	Duración	Rol asignado
Fase 3 Análisis exploratorio	1. Recolectar una muestra representativa	0 días	- Innovador de datos - Desarrollador de los datos
	2. Definir conjunto de criterios para el proceso de carga, limpieza y transformación de los datos	7 días	- Innovador de datos
	3. Aplicar las técnicas o desarrollar los algoritmos haciendo identificados en la fase anterior	7 días	- Desarrollador de los datos
	4. Presentar los resultados obtenidos con los datos muestra en graficas o tablas	7 días	- Innovador de datos - Desarrollador de los datos
	5. Determinar si las técnicas y las herramientas son las adecuadas para el análisis	3 días	- Innovador de datos
Fase 4 Recolección de los datos	1. Preparar y configurar las herramientas de software para almacenar los datos utilizando los elementos para conectarse a los servicios del proveedor de la fuente de Big Data	0 días	- Innovador de datos - Desarrollador de los datos
	2. Puesta en marcha del servicio para la recolección de los datos	0 día	- Desarrollador de los datos
Fase 5 Preparación de los datos	1. Aplicar las técnicas o desarrollar algoritmos para cargar, limpiar y transformar los datos	30 días	- Innovador de datos - Desarrollador de los datos
Fase 6 Análisis de datos y de resultados	1. Construcción del modelo	7 días	- Innovador de datos - Investigador de datos
	2. Evaluación del modelo	14 días	- Innovador de datos - Investigador de datos
	3. Representación y visualización de los datos	21 días	- Innovador de datos - Investigador de datos
Fase 7 Despliegue	1. Aplicación con los datos	3 días	- Investigador de datos
	2. Comunicación de los resultados	3 días	- Persona de negocios - Investigador de datos
Fase 8 Monitoreo y medición de los resultados	1. Medición	3 días	- Persona de negocios - Investigador de datos
	2. Monitoreo	3 días	- Persona de negocios - Investigador de datos

Fase 3. Análisis exploratorio

3.1 Recolectar una muestra significativa.

Para la realización del análisis exploratorio se utilizaron los tweets recolectados durante el mes de Mayo de 2014. El mes fue elegido empleando la técnica de muestreo aleatorio simple ya que permite obtener muestras representativas con gran rapidez y simpleza mediante la generación de números aleatorios donde cada elemento de la población tiene la misma probabilidad de ser elegido.

3.2 Definir conjunto de criterios para el proceso de carga, limpieza y transformación de los datos

Antes de pasar al análisis fue necesario observar los datos con el fin de conocer los dominios de valores que tenía cada una de las variables con los que se trabajó para determinar si existían registros con errores, con inconsistencias o si es necesario reformatearlos.

Uno de los procesos que se realizó inicialmente fue el de obtener las variables del Estado y el Municipio a partir de las coordenadas geográficas que tiene el tweet mediante la implementación de un algoritmo, sin embargo, estas nuevas variables tuvieron un leve problema de formatos de acuerdo a como lo maneja comúnmente el INEGI de acuerdo a lo normado por el Marco Geoestadístico Nacional (MGN por sus siglas), donde por ejemplo para el Estado de Aguascalientes que le corresponde la clave de identificación 01 en el MGN, en algunas extracciones de tweets el algoritmo enviaba la clave 1 presentando el mismo caso para aquellos Estados donde la clave de identificación es menor que 10 por lo que fue necesario normalizarlo. Otra de las variables que se estandarizó fue la fecha debido a la variación que hay en los cuatro husos horarios del territorio nacional se optó por manejar el tiempo de la zona centro de la República Mexicana el cual se obtiene haciendo la sustracción de 6 horas al UTC brindado por Twitter (UTC-6) y posteriormente se le quitaron los datos referentes a los minutos, segundos y milisegundos para formatearlo en DD/MM/AAAA HR Ej. 18/02/2016 11 que corresponde al 18 de Febrero de 2016 a las 11 horas.

Respecto a la limpieza de los datos se quitaron aquellos usuarios de Twitter que no presentaban movilidad a nivel Municipal o Estatal en la publicación de sus Tweets por considerarse irrelevantes para la investigación.

3.3 Aplicar las técnicas o desarrollar los algoritmos identificados en la fase anterior.

Teniendo los datos de la muestra limpios se procedió a desarrollar el algoritmo para la detección de la movilidad.

3.4 Presentar los resultados obtenidos con los datos muestra en graficas o tablas.

Los datos obtenidos al analizar la muestra indican que la mayor cantidad de movimientos a nivel Estatal registrados por los usuarios de Twitter se encuentran entre el Estado de México como origen y la Ciudad de México como destino con 30,772 desplazamientos mientras que en sentido contrario se tiene una ligera disminución de traslados con 30,437. En la Tabla 7 se puede apreciar con mayor claridad lo antes mencionado, agregando que de acuerdo a las cifras brindadas por el análisis, la Ciudad de México representa para este mes el punto en el que concentra la mayor cantidad de desplazamientos tanto de origen como destino de los usuarios de la red social de Twitter.

Tabla 7. Movimientos entre diferentes Estados en el mes de Mayo de 2014, Elaboración propia.

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad	Total de Movimientos
15	México	09	Cd. México	30,772
09	Cd. México	15	México	30,437
17	Morelos	09	Cd. México	3,210
09	Cd. México	17	Morelos	3,028
21	Puebla	09	Cd. México	2,257
09	Cd. México	21	Puebla	2,169
13	Hidalgo	09	Cd. México	1,712
09	Cd. México	22	Queretaro	1,633
09	Cd. México	13	Hidalgo	1,616
22	Queretaro	09	Cd. México	1,609
Resto:				84,433

3.5 Determinar si las técnicas y las herramientas son las adecuadas para el análisis.

Se encontró que las técnicas y las herramientas son útiles para realizar análisis de movilidad y detectar patrones de movimiento generalizado, sin embargo, en este caso práctico se presentó la necesidad de volver a definir las preguntas de investigación por ser un poco ambiguas junto con la definición de los objetivos.

Fase 4. Recolección de los datos

4.1 Preparar y configurar las herramientas de software para almacenar los datos utilizando los elementos para conectarse a los servicios del proveedor de la fuente de Big Data.

Como se mencionó previamente, Twitter tiene a disposición un conjunto de API diseñadas para que los desarrolladores de aplicaciones tengan un amplio repertorio de acciones disponibles como escribir y leer tweets, obtener los hashtags más populares, obtener la lista de favoritos de un usuario, etc. Para cualquiera de las API de Twitter los pasos necesarios para conectarse a los servicios son:

- Crear cuenta de Twitter.
- Crear aplicación de Twitter.
- Generar claves de autenticación. (Clave del consumidor, secreto del consumidor, token de acceso, y el acceso secreto del token)

Para la parte de extracción y almacenamiento de los tweets se utilizó el software conocido como ElasticSearch en conjunto de varias herramientas que trabajan con este fin. Inicialmente se utilizaron los servicios de conexión llamados River desarrollados por Elastic, la misma compañía que desarrollo ElasticSearch para la recolección de los tweets, sin embargo, con el paso del tiempo se percataron de algunas situaciones que provocaban la inestabilidad del clúster haciendo que pudiera colapsar en cualquier momento por lo que la empresa decidió ponerlos obsoletos para futuras versiones.

Tomando en cuenta esta situación, se optó por utilizar Logstash para la recolección de los datos utilizando los plugins que vienen integrados en el software, no obstante, surgió otro problema al tratar de recabar los tweets tal cual los necesita el INEGI y es que estuvieran referenciados geográficamente dentro del territorio nacional. Por tal motivo se eligió extender el funcionamiento del plugin de Logstash para Twitter con la finalidad de agregar la funcionalidad deseada utilizando el lenguaje de programación en el que fue desarrollado, jRuby. Después de haber extendido el plugin se pudo almacenar los tweets del API en el tiempo necesario para desarrollar el experimento en el servidor de Elasticsearch a través de la ejecución de un archivo de configuración de Logstash que contenía información acerca del cómo iban a entrar, como se filtrarían y como saldrían los datos. Para la entrada de tweets se utilizó el plugin extendido de Twitter para la recolección de tweets por referencia geográfica y para la salida se utilizó el plugin de ElasticSearch para generar archivos CSV.

4.1 Puesta en marcha del servicio para la recolección de los datos.

Poner en marcha el servidor de Elasticsearch y lanzar el archivo de configuración de Logstash para empezar la descarga del flujo de datos de Twitter.

Fase 5. Preparación de los datos

5.1 Aplicar las técnicas o desarrollar algoritmos para cargar, limpiar y transformar los datos.

Una vez que el tiempo de recolección llegó a su fin fue necesario extraer los datos del servidor de Elasticsearch para cargarlos y procesarlos en Apache Spark. Esta extracción se realiza almacenando los datos en un determinado número de archivos y puede hacerse de diversas formas dependiendo del conocimiento que se tenga, por ejemplo una manera de hacerlo es mediante scripts desarrollados en el lenguaje de programación Python o mediante un archivo de configuración de Logstash que utilice los plugins de salida de Elasticsearch y CSV respectivamente para generar los archivos en este formato.

Para facilitar la carga de los archivos con formato CSV se utilizó el plugin Spark-CSV, y posteriormente haciendo uso de las librerías de Spark se puedó trabajar con los datos para limpiarlos, transformarlos y validarlos como si fuera una entidad de una base de datos relacional. El proceso que se siguió para realizar estas actividades en Spark fue:

Proceso de carga, limpieza y transformación de los datos.

1. Crear archivo de configuración utilizando el plugin de salida de datos CSV de Logstash.
2. Extraer los datos almacenados en ElasticSearch mediante la generación de archivos de valores separados por comas (CSV por las siglas en ingles de Comma Separated Values), para esto es necesario tomar en consideración el tipo de estudio y los diferentes escenarios que se van a trabajar.
3. Cargar archivo con los datos de los tweets. Mediante el plugin Spark-CSV.
4. Cargar los archivos CSV en el Sistema de Archivos de Hadoop utilizando la librería de Spark CSV.
5. Algoritmo para la obtención de Estados y Municipios partiendo de las referencias geográficas de los tweets

6. Estandarizar el formato de los campos del Estado y Municipio de acuerdo con la siguiente regla:
    a. Entidad ϵ {01,02,03,….,32}
    b. Municipio ϵ {001,002,003,….,999}
7. Formatear la fecha de acuerdo a las necesidades del trabajo de investigación, en este caso se designaron 4 dígitos para el año, seguido por 2 dígitos para hacer referencia al mes y 2 para el día separados por un guion medio (AAAA-MM-DD).
8. Identificar de aquellos usuarios que postearon al menos dos tweets en diferente área geográfica (Algoritmo de detección de movilidad).
9. Extraer los usuarios con todos sus respectivos movimientos entre las áreas geográficas, así como el mes cuando se enviaron los tweets involucradas. Ejemplo:
    a. Registro 1: Usuario X, 09, 2014-01, mensaje 1 del usuario X enviado desde la Entidad 09.
    b. Registro 2: Usuario X, 15, 2014-01, mensaje 2 del usuario X, viajando a la Entidad 15.
10. Formatear los registros de forma que el movimiento entre áreas geográficas quede almacenado en un solo registro. Resultado del ejemplo anterior: Usuario X, 09,15, 2014-01
11. Calcular total de movimientos registrados por mes.
12. Calcular total de movimientos por Entidad y Municipio.
13. Cargar catálogo de Entidades, Municipios y Localidades.
14. Del paso 7 y del paso 8 hacer combinación con el catálogo de entidades para extraer el nombre de la Entidad y del Municipio.
15. Exportar el resultado en archivos con formato de salida CSV.
16. Extraer del catálogo de Entidades, Municipios y Localidades, las coordenadas geográficas.
17. Exportar el resultado en un archivo con formato de salida CSV.

Después de tener los datos en los formatos necesarios se extrajeron aquellos registros donde los usuarios de Twitter tuvieron movilidad ya sea de origen o destino a los Estados y Municipios donde se encuentran las principales ciudades del país (Ciudad de México, Monterrey, Jalisco) mediante la elaboración de algunos algoritmos utilizando el lenguaje de programación Python.

1. Tomar como archivo de entrada el archivo donde se almacenaron los movimientos por Entidad y Municipio.
2. Extraer aquellos registros donde el Estado y Municipio de origen o el Estado y Municipio destino sea igual a alguno donde se encuentran las principales ciudades del país.
3. Exportar el resultado en archivos con formato de salida CSV.
4. Generar distintas tablas de resumen de los datos utilizando Microsoft Excel.

Teniendo los datos limpios y validados se utilizaron para describir o para predecir su comportamiento mediante el empleo de distintas técnicas estadísticas, este trabajo se centrará en el desarrollo de estadísticas descriptivas a través de la generación de gráficos, tablas y reportes.

Fase 6. Análisis de datos y de resultados

6.1 Construcción del modelo

Para este caso práctico no se utilizó un modelo matemático ya que el tipo de análisis a realizar sobre el conjunto de datos generados en la capa anterior será mediante el empleo de técnicas de estadística descriptiva, la cual tiene como principal objetivo poner de manifiesto las características más importantes de los datos y sintetizarlas en gráficas, tablas y mapas.

6.2 Evaluación del modelo

Las pruebas y validaciones de los algoritmos utilizados para generar las estadísticas de este caso práctico fueron realizadas bajo la supervisión de los expertos del área responsable de realizar estos proyectos en el INEGI. Del mismo modo y para comprobar que la información generada es significativa se compararon con la cantidad de habitantes totales y dependiendo del tipo de lugar (ciudad metropolitana, playas turistas y ciudades fronterizas) se determinó si hay algún patrón que explique el movimiento de los usuarios entre dichos Municipios o Estados.

6.3 Representación y visualización de los datos

Proceso para genera las rutas entre el Municipio y el Estado de origen con el Municipio y Estado destino con los archivos generados en el proceso anterior utilizando la herramienta QuantumGis.

1. Iterar cada uno de los registros obtenidos en los archivos del proceso anterior
    1. Buscar la Entidad y Municipio de origen en el archivo donde se almacenaron las coordenadas geográficas de los Municipios y las Entidades.
    2. Poner un punto geográfico donde coincidan los datos del paso anterior.
    3. Buscar la Entidad y Municipio destino en el archivo donde se almacenaron las coordenadas geográficas de los Municipios y las Entidades.
    4. Poner un punto geográfico donde coincidan los datos del paso anterior.
    5. Unir los dos puntos generados para trazar una línea.

Fase 7. Despliegue

7.1 Aplicación con los datos.

Como la elaboración de este caso práctico fue con fines exploratorios en esta fase se especifica la forma en la que se mostraron los resultados que se generaron en el análisis y consistió en el uso de distintas gráficas de resumen de los datos, mapas y tablas con la finalidad de facilitar su interpretación y determinar su utilidad.

7.2 Comunicación de los resultados

Como parte de la comunicación de los resultados se presentan las tablas y graficas generadas con la información extraída categorizada de la siguiente manera: movilidad Nacional por día y por hora, movilidad Nacional general por Municipios, movilidad de la Ciudad de México, movilidad del Estado de Jalisco y la movilidad del Estado de Nuevo León y la movilidad mensual por Estado.

Movilidad Nacional por día y por hora

Como se comentó anteriormente los resultados se presentan en forma de gráficas y tablas con estadísticas descriptivas donde se muestran temas como las horas en la que los usuarios de Twitter publican en diferentes Municipios. En la Tabla 8 y la Tabla 9 se puede apreciar esta situación donde se detectó que la hora a la que se registra menor número de movimientos entre los usuarios de la red social de Twitter es a las 4 am, mientras que se mueven más a las 8 pm.

Tabla 8.Total de movimientos registrados por hora del día

Hora del día	Total de movimientos	Hora del día	Total de movimientos
00	185,013	**12**	326,749
01	105,039	**13**	336,598
02	58,660	**14**	360,298
03	35,903	**15**	369,319
04	25,109	**16**	357,644
05	36,410	**17**	341,448
06	81,531	**18**	346,247
07	155,530	**19**	367,038
08	226,382	**20**	373,421
09	275,243	**21**	354,694
10	299,835	**22**	325,376
11	322,628	**23**	276,331

Tabla 9. Resumen de estadísticas de movimientos registrados por hora del día

Mínima	25,109
1er Cuarto	142,907
Mediana	311,232
Media	247,602
3er Cuarto	348,359
Máxima	373,421

También es posible apreciar lo que se presenta en la Tabla 8 y la Tabla 9 pero gráficamente en la Figura 22.

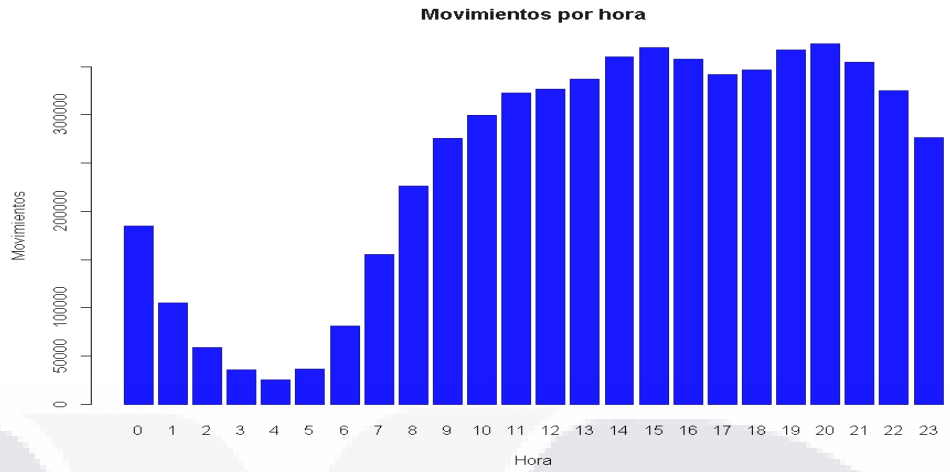


Figura 22. Total de movimientos registrados por hora del día

Un reporte muy similar al anterior, pero con mayor grado de detalle es el presentado en las Tabla 10 y en la Tabla 11 donde se puede apreciar esta misma situación donde se detectó que la hora a la que se registra menor número de movimientos entre los usuarios de la red social de Twitter sigue siendo a las 4 am para todos los días, mientras que se mueven más a las 8 pm con excepción de los martes y los domingos. Con las tablas a este nivel de detalle, se puede observar que los días en los que los usuarios de Twitter publican en diferentes Municipios son durante los días viernes, sábado y domingo a partir de las 9 am, para este último baja a las 7 pm.

Tabla 10. Total de movimientos registrados por día en horas

Hora	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
0	21,228	22,247	19,860	21,965	26,189	34,688	38,836
1	10,671	10,759	9,432	10,544	13,873	22,695	27,065
2	5,040	5,518	4,155	5,031	7,065	14,969	16,882
3	2,735	2,839	2,195	2,625	3,909	9,845	11,755
4	2,493	2,262	1,902	2,179	3,075	5,745	7,453
5	5,743	5,623	4,878	4,498	5,201	5,015	5,452
6	14,220	14,187	12,511	12,322	13,527	8,171	6,593
7	25,728	26,748	24,689	24,081	27,358	16,347	10,579
8	36,134	36,727	35,085	32,839	39,535	28,323	17,739
9	41,369	41,001	39,682	39,734	44,851	39,679	28,927
10	42,099	40,922	39,717	41,381	45,681	48,198	41,837
11	41,335	40,847	42,456	43,936	48,195	53,683	52,176
12	40,409	39,854	39,757	47,309	46,869	54,877	57,674
13	40,652	43,052	42,645	46,846	47,706	54,800	60,897
14	43,642	47,003	46,128	50,833	53,909	57,230	61,553
15	43,268	48,667	45,489	50,667	57,597	60,969	62,662
16	42,445	44,892	41,982	47,988	55,643	61,521	63,173
17	41,904	40,561	39,841	46,795	52,366	59,499	60,482
18	44,726	41,739	41,888	49,569	53,297	59,138	55,890
19	49,698	46,953	47,928	54,907	57,107	57,352	53,093

20	50,265	48,818	50,796	57,628	58,726	57,781	49,407
21	47,364	45,625	48,413	53,505	57,783	56,703	45,301
22	43,953	40,293	43,684	48,654	54,336	53,981	40,475
23	37,025	33,254	35,575	39,690	48,021	48,933	33,833
Total	774,146	770,391	760,688	835,526	921,819	970,142	909,734

Tabla 11. Resumen de estadísticas de movimientos registrados por día en horas

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
N	24	24	24	24	24	24	24
Media	32,256.08	32,099.62	31,695.33	34,813.58	38,409.12	40,422.58	37,905.58
Desviación estándar	16,405.79	16,241.94	16,772.04	18,961.55	19,969.17	20,618.84	20,496.36
Mediana	40,993.5	40,427	39,737	42,658.5	47,287.5	51,308	41,156
mad: desviación absoluta mediana	6,544.2	9,712.51	9,001.61	13,338.95	13,473.13	13,233.69	26,571.16
Mínimo	2,493	2,262	1,902	2,179	3,075	5,015	5,452
Máximo	50,265	48,818	50,796	57,628	58,726	61,521	63,173
Rango	47,772	46,556	48,894	55,449	55,651	56,506	57,721
Skew	-0.77	-0.77	-0.73	-0.62	-0.7	-0.59	-0.29
Kurtosis	-1.1	-1.08	-1.16	-1.27	-1.21	-1.39	-1.5
Error estándar	3,348.82	3,315.37	3,423.58	3,870.51	4,076.19	4,208.8	4,183.8

Lo explicado anteriormente se puede notar de una manera más clara de lo que se presenta en la Tabla 10 y la Tabla 11, mediante la gráfica de la Figura 23.

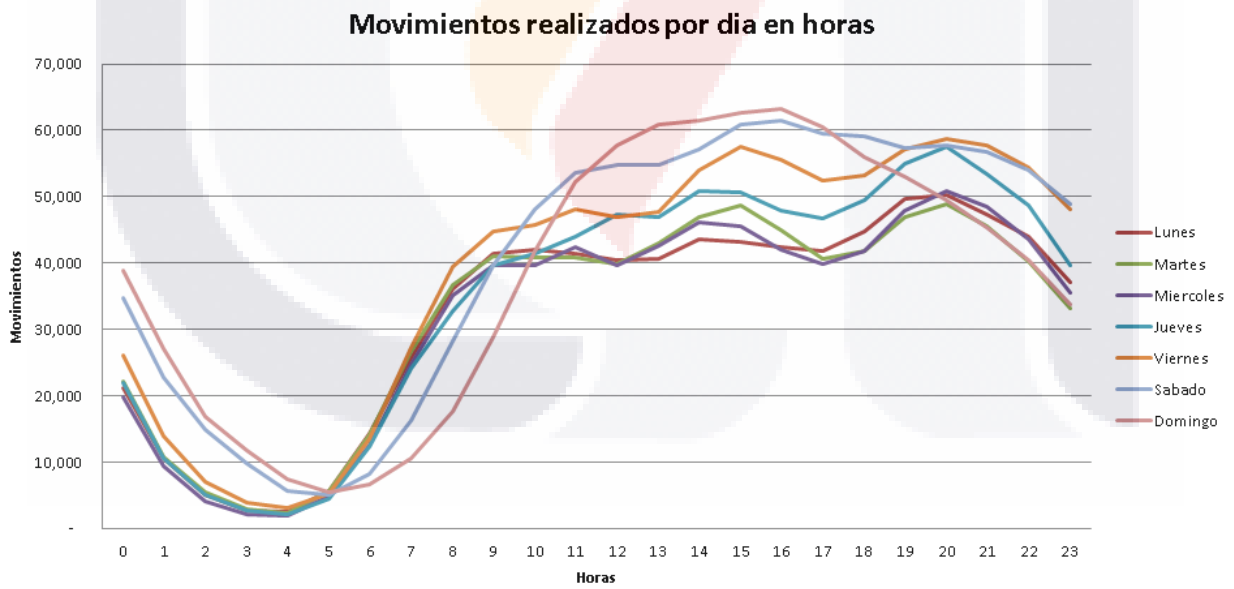


Figura 23. Total de movimientos registrados por día y hora

Movilidad Nacional general por Municipio

La Figura 24 muestra un mapa temático con todos los movimientos realizados por los usuarios de Twitter en la República Mexicana en el periodo seleccionado para esta investigación donde los colores más oscuros representan una mayor frecuencia de la movilidad realizada entre dos Municipios. Conforme los colores se hacen más tenues, la frecuencia de los movimientos entre dichos Municipios es menor.

Una de las consideraciones que se tomó para desarrollar estos mapas fue en mostrar únicamente aquellos Municipios donde la cantidad de movimientos es mayor a cien ya que un alto porcentaje de movimientos caía dentro de esta categoría lo que afectaba considerablemente los tiempos de procesamiento además de considerarse como poco relevantes y no ser representativas. Es importante aclarar también que para el trazo de las líneas no se utilizaron las coordenadas propias de los tweets ya que al momento de ponerlos en el mapa era complicado de leer por lo que en su lugar se identificaron las coordenadas centrales de los polígonos de los Municipios y se partió de este criterio.

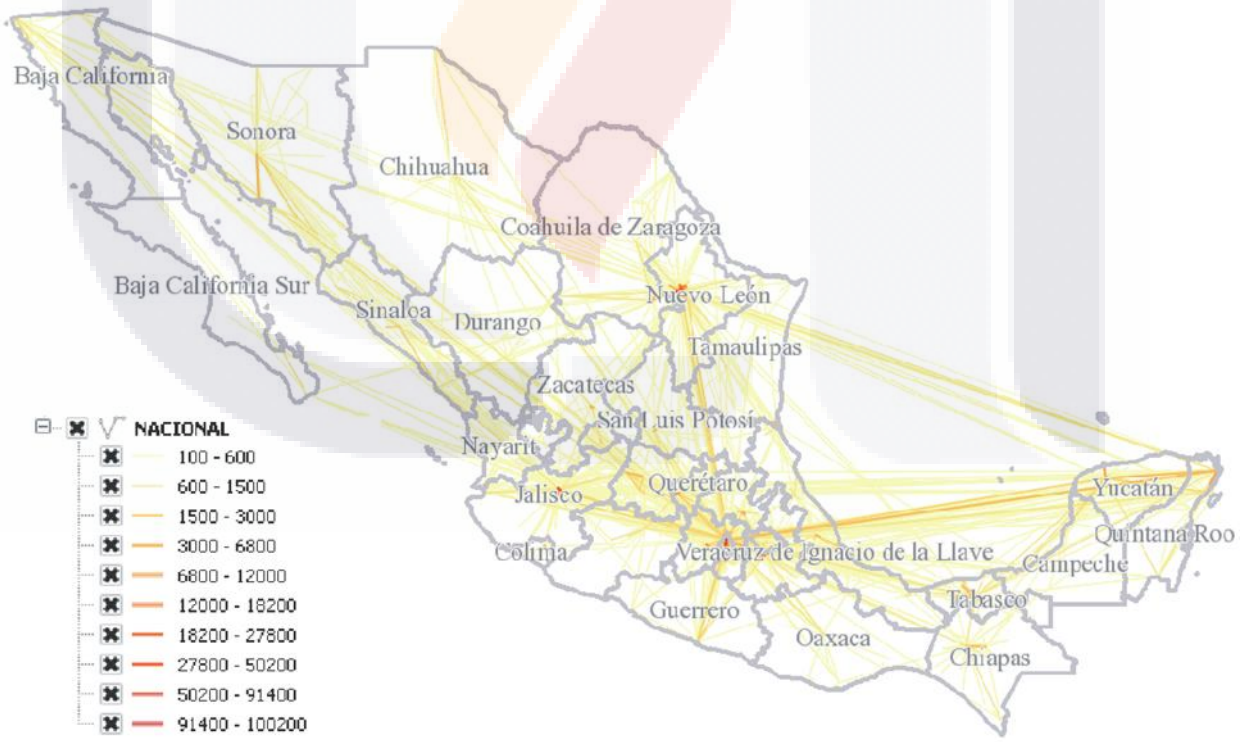


Figura 24. Movimientos registrados a nivel Nacional

Analizando un poco más a detalle el mapa y teniendo en consideración lo comentado anteriormente sobre los colores, es posible apreciar que el mayor flujo de movimientos con origen o destino de los Municipios o Delegaciones dentro y fuera de los Estados registrados a través de los mensajes que publican los usuarios de la red social de Twitter están dentro de las áreas metropolitanas de las ciudades más importantes del país (la Ciudad de México; la ciudad de Monterrey, en Nuevo León y la ciudad de Guadalajara, en Jalisco).

También es posible percibir que hay Municipios con alta afluencia de movimientos en aquellos que tienen lugares turísticos playeros como es caso del Municipio de Benito Juárez y Solidaridad en Quintana Roo que es donde están ubicados la ciudad de Cancún y Playa del Carmen. En el mapa es posible observar que su mayor cantidad de movimientos es entre las principales ciudades del país debido a que sus aeropuertos cuentan con entradas y salidas todo el año a estas ciudades de acuerdo a los horarios publicados en (http://es.cancun-airport.com/llegadas.htm); del mismo modo se puede observar el Municipio de Acapulco de Juárez en Guerrero que es visitado en su mayoría por los usuarios de Twitter de los Municipios pertenecientes al Estado de México, la Ciudad de México, Puebla y Morelos ; los Municipios de Mérida y Progreso en Yucatán presentan un patrón donde el punto de movimientos de los usuarios de Twitter es entre estos dos Municipios; Veracruz y Boca del Rio en Veracruz son el punto de origen o destino de Municipios dentro o fuera del Estado. Estos son algunos casos que se puede deducir solo con observar el mapa a detalle.

Finalmente y del mismo modo que en los Municipios anteriores, es posible ver cuáles de ellos se encuentran cerca de la frontera con el país vecino del norte, Estados Unidos de América, encontrando que Tijuana y Mexicali en Baja California son el centro de origen o destino de una amplia variedad de movimientos tanto del interior del Estado como de otros; el caso de Nogales es un poco distinto, ya que el punto de movimiento de los usuarios de Twitter es entre este Municipio y el Hermosillo en el mismo Estado de Sonora; Ciudad Juárez en Chihuahua y Piedras negras en Coahuila tiene el mismo patrón de movimientos que el anterior de Sonora, mientras que Nuevo Laredo en Tamaulipas presenta una variedad de movimientos que se realizan entre éste y algunos Municipios del Estado de Nuevo León.

En la Tabla 12 se puede apreciar un patrón generalizado que los Municipios y Delegaciones de los Estados en los que mayor número de movimientos registraron los usuarios de Twitter fueron dentro de las áreas metropolitanas del país teniendo a la Ciudad de México en primer lugar, con la movilidad registrada entre las delegaciones de Miguel Hidalgo y la Cuauhtémoc;

en el Estado de Jalisco con el lugar número cinco se encontró que Guadalajara y Zapopan son las que más movimientos tienen, mientras que en el lugar número siete se encuentra Nuevo León con los Municipios de San Pedro Garza García y Monterrey

Tabla 12. Relación de Municipios con mayor número de movimientos realizados por los usuarios de Twitter

Posición	Estado Origen	Municipio Origen	Estado Destino	Municipio Destino	Movimientos realizados
1	Cd. México	Miguel Hidalgo	Cd. México	Cuauhtémoc	100,161
2	Cd. México	Cuauhtémoc	Cd. México	Miguel Hidalgo	99,225
3	Cd. México	Cuauhtémoc	Cd. México	Benito Juárez	91,375
4	Cd. México	Benito Juárez	Cd. México	Cuauhtémoc	90,735
5	Jalisco	Guadalajara	Jalisco	Zapopan	73,918
6	Jalisco	Zapopan	Jalisco	Guadalajara	73,560
7	Nuevo León	San Pedro Garza García	Nuevo León	Monterrey	56,095
8	Nuevo León	Monterrey	Nuevo León	San Pedro Garza García	55,624
9	Cd. México	Coyoacán	Cd. México	Tlalpan	50,712
10	Cd. México	Tlalpan	Cd. México	Coyoacán	50,565
11	Nuevo León	Monterrey	Nuevo León	San Nicolás de los Garza	50,165
12	Nuevo León	San Nicolás de los Garza	Nuevo León	Monterrey	50,071
13	Cd. México	Benito Juárez	Cd. México	Coyoacán	47,257
14	Cd. México	Coyoacán	Cd. México	Benito Juárez	46,820
15	Cd. México	Cuauhtémoc	Cd. México	Coyoacán	46,526
16	Cd. México	Coyoacán	Cd. México	Cuauhtémoc	46,441
17	Cd. México	Miguel Hidalgo	Cd. México	Benito Juárez	41,972
18	Cd. México	Benito Juárez	Cd. México	Miguel Hidalgo	41,814
19	Veracruz de Ignacio de la Llave	Boca del Rio	Veracruz de Ignacio de la Llave	Veracruz	38,086
20	Veracruz de Ignacio de la Llave	Veracruz	Veracruz de Ignacio de la Llave	Boca del Rio	38,010
21	Cd. México	Benito Juárez	Cd. México	Álvaro Obregón	37,515
22	Cd. México	Álvaro Obregón	Cd. México	Benito Juárez	37,382
23	Nuevo León	Monterrey	Nuevo León	Guadalupe	36,316
24	Nuevo León	Guadalupe	Nuevo León	Monterrey	35,761
25	Cd. México	Cuauhtémoc	Cd. México	Álvaro Obregón	35,094
26	Cd. México	Álvaro Obregón	Cd. México	Cuauhtémoc	34,930
27	Cd. México	Cuauhtémoc	Cd. México	Gustavo A. Madero	28,479
28	Cd. México	Gustavo A. Madero	Cd. México	Cuauhtémoc	27,845

29	Puebla	Puebla	Puebla	San Andrés Cholula	27,773
30	Cd. México	Cuauhtémoc	Cd. México	Venustiano Carranza	27,763
31	Puebla	San Andrés Cholula	Puebla	Puebla	27,754
32	Cd. México	Miguel Hidalgo	Cd. México	Álvaro Obregón	26,256
33	Cd. México	Álvaro Obregón	Cd. México	Miguel Hidalgo	26,216
34	Cd. México	Coyoacán	Cd. México	Álvaro Obregón	26,006
35	Cd. México	Álvaro Obregón	Cd. México	Coyoacán	25,794
36	Cd. México	Venustiano Carranza	Cd. México	Cuauhtémoc	25,696
37	Cd. México	Cuauhtémoc	Cd. México	Tlalpan	21,253
38	Cd. México	Tlalpan	Cd. México	Cuauhtémoc	21,117
39	Cd. México	Miguel Hidalgo	Cd. México	Coyoacán	20,232
40	Cd. México	Tlalpan	Cd. México	Benito Juárez	20,202
41	Cd. México	Benito Juárez	Cd. México	Tlalpan	20,149
42	Cd. México	Coyoacán	Cd. México	Miguel Hidalgo	20,096
43	Nuevo León	Monterrey	Nuevo León	Apodaca	18,238
44	Nuevo León	Apodaca	Nuevo León	Monterrey	17,564
45	Cd. México	Miguel Hidalgo	México	Naucalpan de Juárez	16,543
46	México	Naucalpan de Juárez	Cd. México	Miguel Hidalgo	16,451
47	Cd. México	Álvaro Obregón	Cd. México	Tlalpan	16,177
48	Cd. México	Tlalpan	Cd. México	Álvaro Obregón	16,162
49	México	Toluca	México	Metepec	15,989
50	México	Metepec	México	Toluca	15,820

Por otra parte y mostrando un panorama más amplio respecto al origen y destino de los movimientos realizados por los usuarios de Twitter en la República Mexicana, en la Tabla 13 se presenta una matriz origen/destino a nivel Estado donde se puede apreciar los patrones generalizados de movimientos entre dos Estados haciendo el cruce de filas por columnas y a través de las cifras que se encuentran en diagonal principal de la matriz es posible ver la movilidad realizada dentro del mismo Estado pero contabilizando aquellos que publicaron en un Municipio diferente.

Tabla 13. Tabla origen/destino de movimientos realizados por los usuarios de Twitter. Parte 1

Origen\ Destino	Aguascalientes	Baja California	Baja California Sur	Campeche	Coahuila de Zaragoza	Colima	Chiapas	Chihuahua	Cd. México	Durango	Guanajuato	Guerrero	Hidalgo	Jalisco	Mexico
Aguascalientes	8,182	57	28	11	210	97	37	65	2,522	146	1,602	66	98	2,658	607
Baja California	63	14,567	520	15	95	50	152	200	3,841	52	294	160	50	1,495	484
Baja California Sur	30	537	3,162	14	59	33	31	41	2,117	23	101	41	37	864	395
Campeche	15	16	17	8,076	18	7	234	15	1,852	9	43	36	33	123	186
Coahuila	174	95	50	15	34,319	27	38	597	2,616	7,527	360	78	77	669	476
Colima	95	49	35	6	27	14,496	20	12	957	14	211	30	23	4,114	185
Chiapas	37	182	24	269	51	10	40,660	59	4,290	42	166	103	72	518	561
Chihuahua	70	190	32	23	568	20	69	11,546	2,432	353	155	50	42	563	328
Cd. México	2,484	3,865	2,100	1,906	2,689	955	4,458	2,472	2,016,160	1,139	10,486	13,614	14,901	18,461	283,139
Durango	151	52	17	9	7,501	14	50	337	1,101	7,447	149	15	25	370	157
Guanajuato	1,605	281	72	35	373	210	154	147	10,816	165	65,737	516	736	5,894	3,062
Guerrero	70	134	45	39	79	20	132	52	14,154	24	522	10,818	308	492	3,424
Hidalgo	93	43	36	31	65	19	85	43	15,144	24	714	319	54,833	525	6,812
Jalisco	2,635	1,539	902	132	711	4,104	526	574	18,194	358	5,928	491	533	293,205	3,182
Mexico	633	417	360	190	482	172	479	283	282,676	158	3,084	3,156	6,923	3,094	265,593
Michoacán	164	246	33	24	77	228	87	187	5,538	25	3,029	886	179	4,107	2,384
Morelos	48	73	64	26	71	29	92	60	29,232	29	289	3,152	311	422	4,540
Nayarit	130	99	35	9	73	56	33	50	1,301	50	437	33	54	7,859	272
Nuevo Leon	521	703	445	151	11,483	113	463	1,013	14,218	648	1,683	305	333	3,661	1,984
Oaxaca	30	288	38	94	45	34	1,002	49	6,319	29	177	314	145	366	980
Puebla	136	232	88	212	219	71	1,085	123	22,032	55	938	1,471	2,754	1,093	5,038
Queretaro	446	117	74	67	348	96	161	120	15,812	132	10,932	506	2,415	1,836	6,343
Quintana Roo	153	265	165	1,184	341	54	652	231	13,472	103	652	249	272	1,642	2,192
San Luis Potosi	476	38	25	19	407	57	60	68	3,097	101	1,569	104	216	1,121	804
Sinaloa	90	932	479	16	605	67	69	397	3,560	995	341	69	90	3,678	558
Sonora	60	2,615	192	12	106	38	47	403	2,884	80	196	59	37	1,548	355
Tabasco	31	47	27	2,810	42	14	4,776	44	4,165	3	107	68	93	488	517
Tamaulipas	70	140	45	229	556	24	71	103	3,146	70	429	83	259	682	559
Tlaxcala	19	28	3	23	16	4	26	9	2,593	5	115	59	400	104	622
Veracruz	99	264	65	1,044	158	49	1,153	98	13,847	37	522	318	1,022	979	2,623
Yucatan	46	108	29	5,305	103	17	599	48	5,170	17	252	76	62	567	667
Zacatecas	**1,873**	**49**	**22**	**5**	**471**	**26**	**26**	**137**	**1,345**	**293**	**498**	**123**	**47**	**1,142**	**251**

Tabla 14. Tabla origen/destino de movimientos realizados por los usuarios de Twitter. Parte 2

Origen\ Destino	Michoacán de Ocampo	Morelos	Nayarit	Nuevo Leon	Oaxaca	Puebla	Queretaro	Quintana Roo	San Luis Potosi	Sinaloa	Sonora	Tabasco	Tamaulipas	Tlaxcala	Veracruz de Ignacio de la Llave	Yucatan	Zacatecas
Aguascalientes	151	43	134	528	16	127	437	162	490	85	54	31	77	16	116	49	1,879
Baja California	280	89	121	686	245	240	132	304	57	946	2,574	41	145	25	267	84	53
Baja California Sur	40	74	42	446	37	96	66	201	27	503	206	32	44	5	85	39	27
Campeche	28	31	8	154	81	230	60	1,248	17	29	22	2,779	214	16	971	5,422	10
Coahuila	77	61	68	11,624	56	212	319	343	383	560	88	48	560	13	157	82	485
Colima	237	19	59	121	26	66	98	40	46	77	49	16	27	5	46	24	33
Chiapas	71	130	33	446	931	1,083	175	712	53	70	59	4,883	52	24	1,217	619	32
Chihuahua	195	66	43	973	33	143	105	223	65	404	374	35	131	7	96	39	166
Cd. México	5,581	29,251	1,283	14,257	6,459	21,844	15,770	13,803	3,010	3,660	2,939	4,255	3,172	2,633	13,807	5,365	1,333
Durango	33	25	45	662	25	62	130	98	103	957	96	7	57	2	40	26	300
Guanajuato	3,036	298	439	1,688	201	930	10,706	583	1,582	326	186	100	433	110	516	259	474
Guerrero	879	2,926	35	343	315	1,508	527	269	125	72	55	81	75	68	315	80	111
Hidalgo	182	291	54	334	146	2,698	2,390	232	234	58	43	81	270	432	982	65	46
Jalisco	3,978	412	7,621	3,751	316	1,116	1,829	1,678	1,149	3,617	1,643	479	709	109	1,006	546	1,122
Mexico	2,455	4,434	273	1,870	903	5,057	6,545	1,954	841	501	332	473	538	669	2,527	581	213
Michoacán	20,227	173	114	340	75	329	1,227	220	149	126	101	55	116	36	162	66	54
Morelos	187	38,369	32	293	185	2,077	503	392	92	85	66	81	88	117	354	86	24
Nayarit	117	33	3,582	264	24	92	156	95	104	663	133	28	39	4	67	26	49
Nuevo Leon	357	301	281	597,723	269	961	1,183	3,688	1,827	1,191	885	581	8,112	50	1,386	665	735
Oaxaca	74	178	36	285	44,506	2,696	259	212	60	98	45	289	55	78	2,531	117	34
Puebla	336	2,085	80	924	2,723	136,734	1,119	1,032	294	198	161	1,153	351	8,472	9,676	293	80
Queretaro	1,176	464	157	1,183	255	1,163	63,433	538	1,464	214	130	140	352	173	620	176	232
Quintana Roo	227	411	108	3,687	209	1,059	565	47,570	199	169	294	1,745	331	49	1,549	10,248	85
San Luis Potosi	157	78	117	1,883	63	313	1,398	211	16,916	111	65	48	1,427	23	384	72	501
Sinaloa	132	104	627	1,191	98	219	218	156	116	22,484	2,794	42	108	21	266	58	127
Sonora	88	70	114	889	60	149	163	258	58	2,799	47,989	32	78	6	113	61	77
Tabasco	47	76	25	569	269	1,155	149	1,665	50	44	27	69,392	268	56	4,425	1,865	18
Tamaulipas	97	92	39	8,101	65	347	332	342	1,460	99	79	281	59,940	29	4,212	179	119
Tlaxcala	38	107	3	49	77	8,475	178	42	25	16	7	54	29	18,358	623	31	8
Veracruz	207	365	65	1,367	2,574	9,682	631	1,491	393	252	91	4,466	4,189	610	201,888	581	60
Yucatan	76	96	27	673	123	299	191	10,501	70	69	59	1,841	175	29	529	48,303	25
Zacatecas	**64**	**24**	**50**	**761**	**28**	**70**	**210**	**83**	**508**	**116**	**61**	**17**	**126**	**10**	**47**	**33**	**16,693**

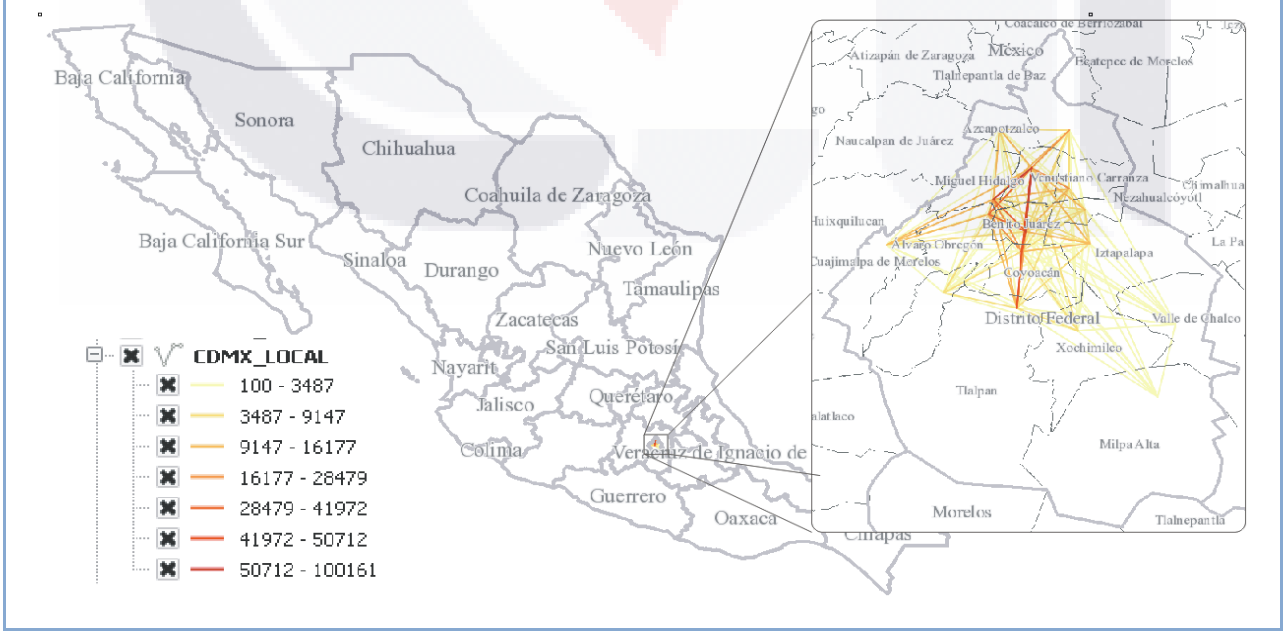
A continuación, se realiza el mismo tipo de análisis pero enfocado en las principales ciudades del país (la Ciudad de México; la ciudad de Monterrey, en Nuevo León y la ciudad de Guadalajara, en Jalisco) mostrando en las siguientes figuras el resultado del procesamiento de los datos de Twitter y las tablas con los Municipios con mayor número de movimientos.

Movilidad en el Ciudad de México

La Ciudad de México es la capital del país. Tiene une extensión territorial que representa el 0.08% del territorio nacional, una población de 8, 918, 653 habitantes que equivale al 7.5% del total del país. Su principal actividad económica es el comercio y tiene una aportación del 16,5% al producto interno bruto (PIB por sus siglas) nacional de acuerdo a cifras del INEGI (INEGI, 2010).

La Figura 25 muestra los movimientos registrados dentro de la Ciudad de México. Del mismo modo que en la figura anterior los colores más oscuros representan una mayor frecuencia de la movilidad realizada entre dos Municipios y los colores más tenues representan una frecuencia menor de movimientos.

Mapa



Tabla

Estado Origen	Nombre Estado	Municipio Origen	Nombre del Municipio	Estado Destino	Nombre Estado	Municipio Destino	Nombre del Municipio	Total de Movimientos
09	Cd. México	016	Miguel Hidalgo	09	Cd. México	015	Cuauhtémoc	100,161
09	Cd. México	015	Cuauhtémoc	09	Cd. México	016	Miguel Hidalgo	99,225
09	Cd. México	015	Cuauhtémoc	09	Cd. México	014	Benito Juarez	91,375
09	Cd. México	014	Benito Juarez	09	Cd. México	015	Cuauhtémoc	90,735
09	Cd. México	003	Coyoacan	09	Cd. México	012	Tlalpan	50,712
09	Cd. México	012	Tlalpan	09	Cd. México	003	Coyoacan	50,565
09	Cd. México	014	Benito Juarez	09	Cd. México	003	Coyoacan	47,257
09	Cd. México	003	Coyoacan	09	Cd. México	014	Benito Juarez	46,820
09	Cd. México	015	Cuauhtémoc	09	Cd. México	003	Coyoacan	46,526
09	Cd. México	003	Coyoacan	09	Cd. México	015	Cuauhtémoc	46,441
Resto								1,345,442

Figura 25. Movimientos registrados dentro del Ciudad de México

En la tabla de la Figura 25 se puede apreciar que la mayor cantidad de movimientos fueron realizados entre las Delegaciones Miguel Hidalgo, Cuauhtémoc, Benito Juárez, Coyoacán y Tlalpan. Comparándolas contra otras estadísticas como el número total de habitantes en el año 2014 (Instituto Nacional de Estadística y Geografía, 2014b) y el número total de planteles, aulas, bibliotecas y talleres a inicios del ciclo escolar 2012-2013 (Instituto Nacional de Estadística y Geografía, 2014b) se observa a simple vista que no presentan ninguna relación ya que las Delegaciones que tienen mayor población son Iztapalapa y Gustavo A. Madero, que por cierto también tienen el mayor número de planteles escolares, posiblemente con otro tipo de análisis en que se tomen en cuenta también las horas en que se generan los tweets se pueda determinar si existe un vínculo entre estas variables y la movilidad. Donde sí es posible contemplar una probable relación es con el número de pasajeros transportados anualmente en el sistema de transporte colectivo metro en la línea cuatro caminos- taxqueña que brinda servicio a las Delegaciones Miguel Hidalgo, Cuauhtémoc, Benito Juárez, Iztacalco y Coyoacán de la Ciudad de México y a Naucalpan de Juárez, un Municipio del Estado de México. La estadística y un mapa con las rutas del metro se muestran en la Figura 26.

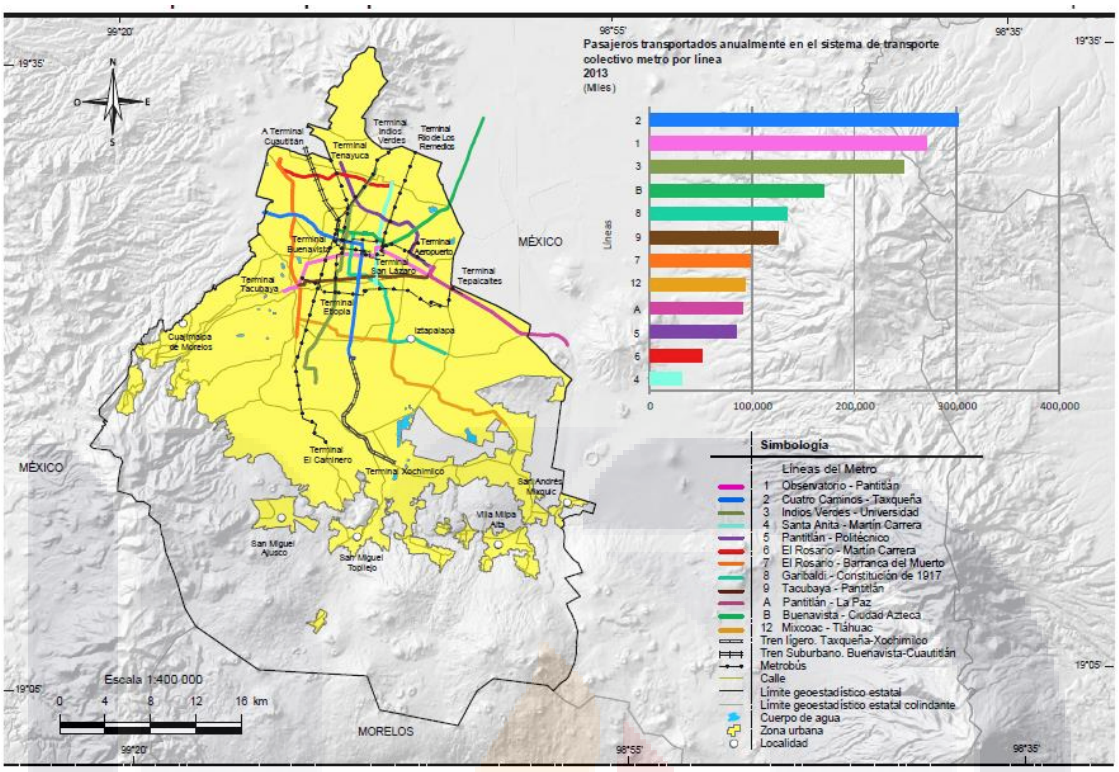


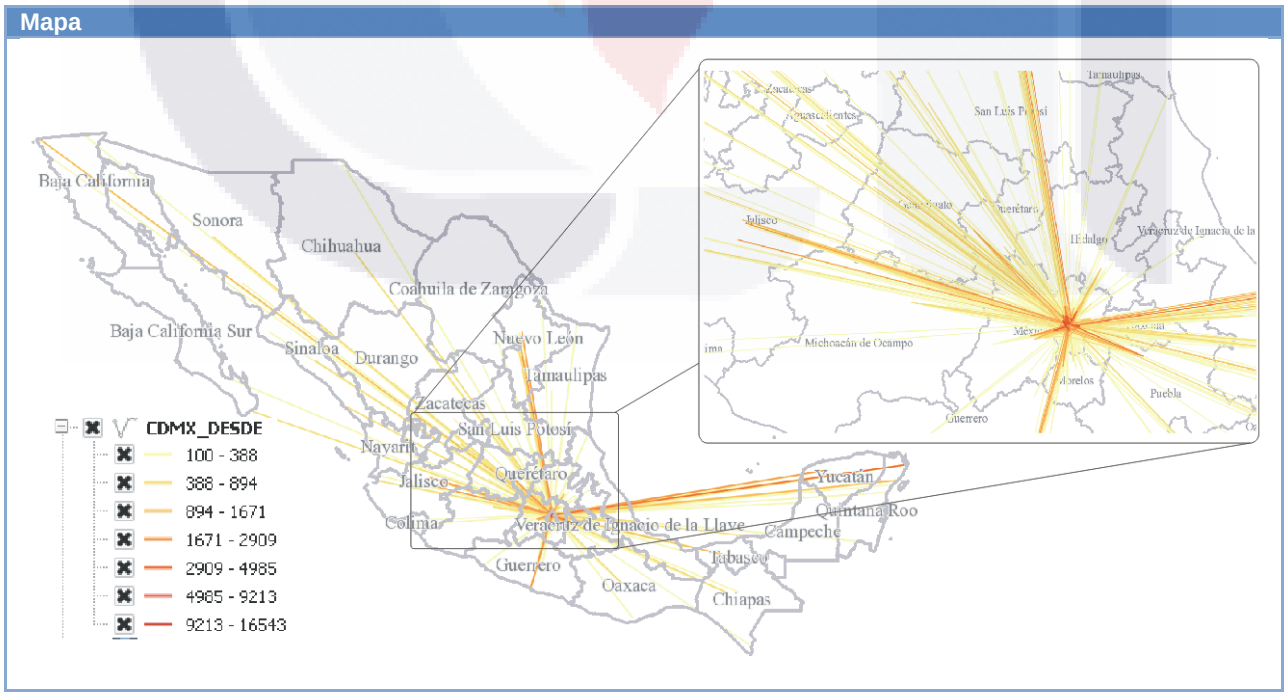
Figura 26. Infraestructura del transporte público de la ciudad de México. Fuente (Instituto Nacional de Estadística y Geografía, 2014b)

Otra variable con la que pudiera tener relación la movilidad es con la cantidad de llegadas de turistas de la Ciudad de México, tanto los residentes del país como los extranjeros, donde la mayor cantidad de llegadas fueron registradas en las Delegaciones Cuauhtémoc, Miguel Hidalgo, Benito Juárez, Venustiano Carranza y Gustavo A. Madero. Los datos totales de la llegada de turistas por delegación categorizados por el criterio si el visitante es residente del país o no, se muestran en la Figura 27.

Delegación	Total	Residentes en el país	No residentes en el país
Distrito Federal	**12 677 217**	**10 298 530**	**2 378 687**
Álvaro Obregón	**478 480**	329 510	148 970
Azcapotzalco	**174 002**	157 447	16 555
Benito Juárez	**986 118**	787 463	198 655
Coyoacán	**212 090**	160 836	51 254
Cuajimalpa de Morelos	**94 362**	69 535	24 827
Cuauhtémoc	**6 861 446**	5 707 146	1 154 300
Gustavo A. Madero	**603 682**	550 781	52 901
Iztacalco	**186 411**	159 609	26 802
Iztapalapa	**301 333**	252 307	49 026
La Magdalena Contreras	**64 656**	43 627	21 029
Miguel Hidalgo	**1 524 924**	1 139 194	385 730
Tláhuac	**18 832**	14 156	4 676
Tlalpan	**293 238**	226 099	67 139
Venustiano Carranza	**788 559**	626 163	162 396
Xochimilco	**89 084**	74 657	14 427

Figura 27. Llegada de turistas a establecimientos de hospedaje por delegación según residencia 2013. Fuente (Instituto Nacional de Estadística y Geografía, 2014b)

Hablando un poco más a nivel nacional, la Figura 28 muestra los movimientos realizados desde la Ciudad de México hacia otros Municipios pertenecientes de otros Estados de la República Mexicana. Del mismo modo que en la figura anterior los colores más oscuros representan una mayor frecuencia de la movilidad realizada entre dos Municipios y los colores más tenues representan la frecuencia menor de movimientos.



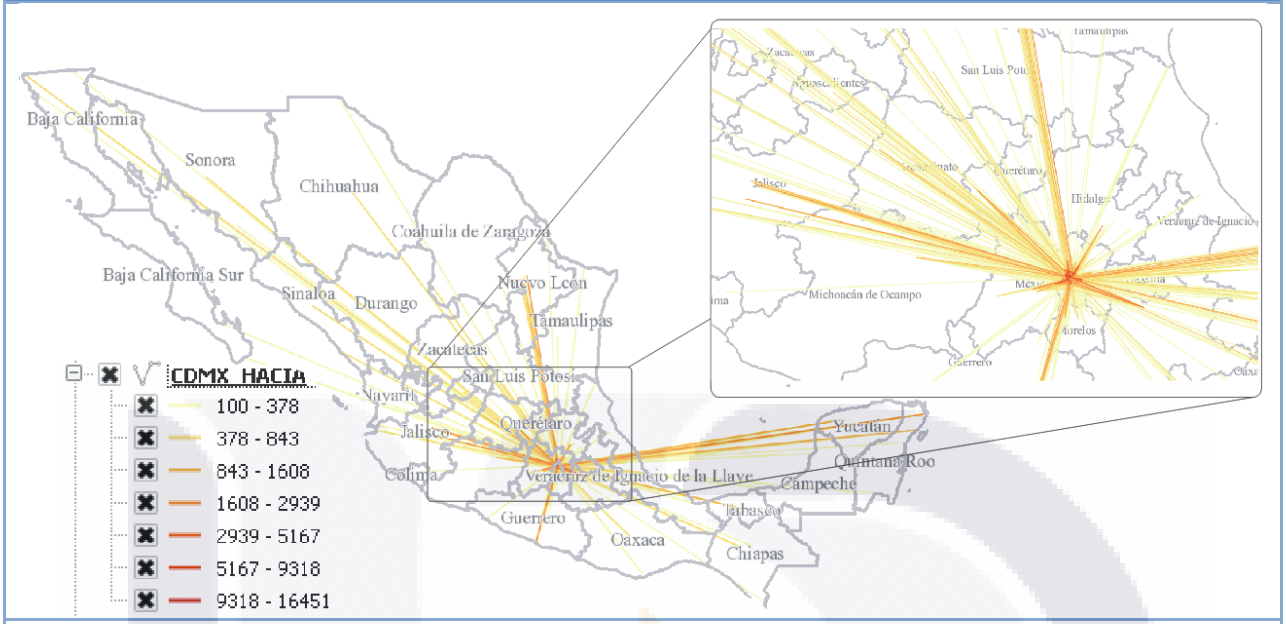
Tabla

Estado Origen	Nombre Estado	Municipio Origen	Nombre del Municipio	Estado Destino	Nombre Estado	Municipio Destino	Nombre del Municipio	Total de Movimientos
09	Cd. México	016	Miguel Hidalgo	15	Edo. México	057	Naucalpan de Juarez	16,543
09	Cd. México	015	Benito Juarez	15	Edo. México	057	Naucalpan de Juarez	14,105
09	Cd. México	015	Benito Juarez	15	Edo. México	104	Tlalnepantla de Baz	9,213
09	Cd. México	016	Miguel Hidalgo	15	Edo. México	104	Tlalnepantla de Baz	7,063
09	Cd. México	015	Benito Juarez	15	Edo. México	033	Ecatepec de Morelos	6,811
09	Cd. México	015	Benito Juarez	15	Edo. México	058	Netzahualcoyotl, B	6,523
09	Cd. México	005	Gustavo A. Madero	15	Edo. México	104	Tlalnepantla de Baz	6,297
09	Cd. México	016	Miguel Hidalgo	15	Edo. México	037	Huixquilucan	5,890
09	Cd. México	014	Benito Juarez	15	Edo. México	057	Naucalpan de Juarez	5,217
09	Cd. México	002	Azcapotzalco	15	Edo. México	104	Tlalnepantla de Baz	4,985
Resto								430,434

Figura 28. Movimientos registrados desde el Ciudad de México

Por ser muy parecido el comportamiento de la movilidad de la Figura 28 y la Figura 29 se explicará después de presentar esta última, la cual consiste en mostrar los movimientos realizados desde los Municipios de otros Estados de la República Mexicana hacia la Ciudad de México. Del mismo modo que en la figura anterior los colores más oscuros de la Figura 29 representan una mayor frecuencia de la movilidad realizada entre los Municipios de dos Estados diferentes y los colores más tenues representan frecuencias menores de movimientos.

Mapa



Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre del Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre del Municipio	Total de Movimientos
15	Edo. México	057	Naucalpan de Juarez	09	Cd. México	016	Miguel Hidalgo	16,451
15	Edo. México	057	Naucalpan de Juarez	09	Cd. México	015	Cuauhtémoc	14,093
15	Edo. México	104	Tlalnepantla de Baz	09	Cd. México	015	Cuauhtémoc	9,318
15	Edo. México	104	Tlalnepantla de Baz	09	Cd. México	016	Miguel Hidalgo	7,020
15	Edo. México	033	Ecatepec de Morelos	09	Cd. México	015	Cuauhtémoc	6,827
15	Edo. México	058	Netzahualcoyotl	09	Cd. México	015	Cuauhtémoc	6,305
15	Edo. México	104	Tlalnepantla de Baz	09	Cd. México	005	Gustavo A. Madero	6,071
15	Edo. México	037	Huixquilucan	09	Cd. México	016	Miguel Hidalgo	5,955
15	Edo. México	057	Naucalpan de Juarez	09	Cd. México	014	Benito Juarez	5,167
15	Edo. México	104	Tlalnepantla de Baz	09	Cd. México	002	Azcapotzalco	4,923
Resto								429,283

Figura 29. Movimientos registrados hacia el Ciudad de México

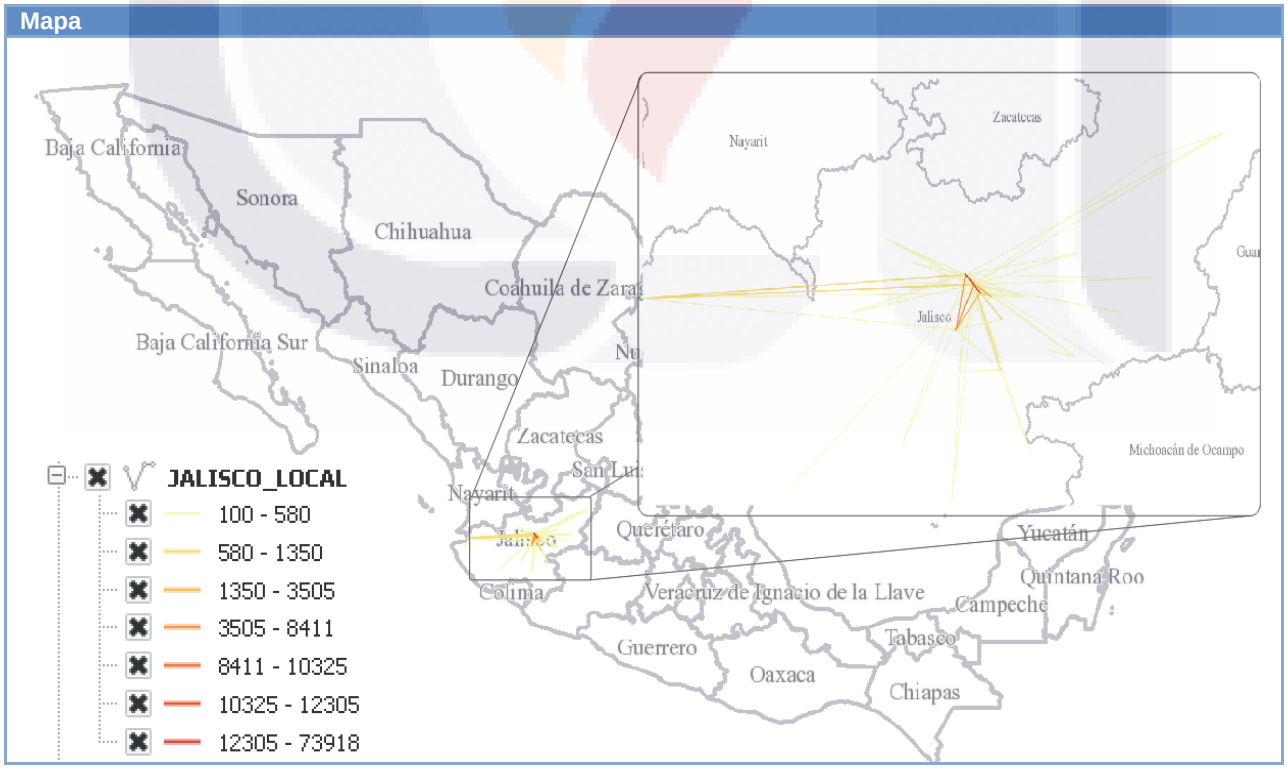
El Estado que más tiene interacción con la Ciudad de México es el Estado de México ya que comparten una alta integración socioeconómica debido a que las Delegaciones de la Ciudad de México y muchos de los Municipios del Estado de México como Naucalpan de Juárez, Tlalnepantla de Baz, Ecatepec de Morelos, Netzahualcóyotl, Huixquilucan entre otros pertenecen a la ZMVM. De acuerdo con (CTS México & ITDP, 2011) la situación de movilidad actual de esta zona fue ocasionada a que durante muchos años se le dio prioridad al transporte

individual sobre el colectivo además del crecimiento demográfico acelerado hacia las periferias de la ciudad fue causando que la población económicamente activa (PEA por su sigla) tuviera que viajar mayores distancias para llegar a sus centros de trabajo.

Movilidad en el Estado de Jalisco

Jalisco es otro de los Estados que tienen mayor actividad en la República Mexicana. Cuenta con une extensión territorial que representa el 4.01% del territorio nacional, una población de 7, 844, 830 habitantes que equivale al 6.6% del total del país. Su principal actividad económica es el comercio y tiene una aportación del 6.,5% del PIB nacional de acuerdo a cifras del INEGI (INEGI, 2010).

La Figura 30 muestra los movimientos realizados entre los Municipios del Estado de Jalisco. Al igual que en el caso anterior de la Ciudad de México, la asignación de colores en el mapa para indicar la frecuencia de movimientos es la misma teniendo que los colores más oscuros representan la mayor frecuencia de la movilidad y los colores más tenues representan una frecuencia menor de movimientos.



Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre del Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre del Municipio	Total de Movimientos
14	Jalisco	039	Guadalajara	14	Jalisco	120	Zapopan	73,918
14	Jalisco	120	Zapopan	14	Jalisco	039	Guadalajara	73,560
14	Jalisco	039	Guadalajara	14	Jalisco	098	San Pedro Tlaquepaque	12,305
14	Jalisco	098	San Pedro Tlaquepaque	14	Jalisco	039	Guadalajara	11,853
14	Jalisco	120	Zapopan	14	Jalisco	098	San Pedro Tlaquepaque	10,325
14	Jalisco	098	San Pedro Tlaquepaque	14	Jalisco	120	Zapopan	10,251
14	Jalisco	039	Guadalajara	14	Jalisco	097	Tlajomulco de Zuñiga	8,411
14	Jalisco	120	Zapopan	14	Jalisco	097	Tlajomulco de Zuñiga	8,315
14	Jalisco	097	Tlajomulco de Zuñiga	14	Jalisco	039	Guadalajara	7,683
14	Jalisco	097	Tlajomulco de Zuñiga	14	Jalisco	120	Zapopan	7,587
Resto								47,683

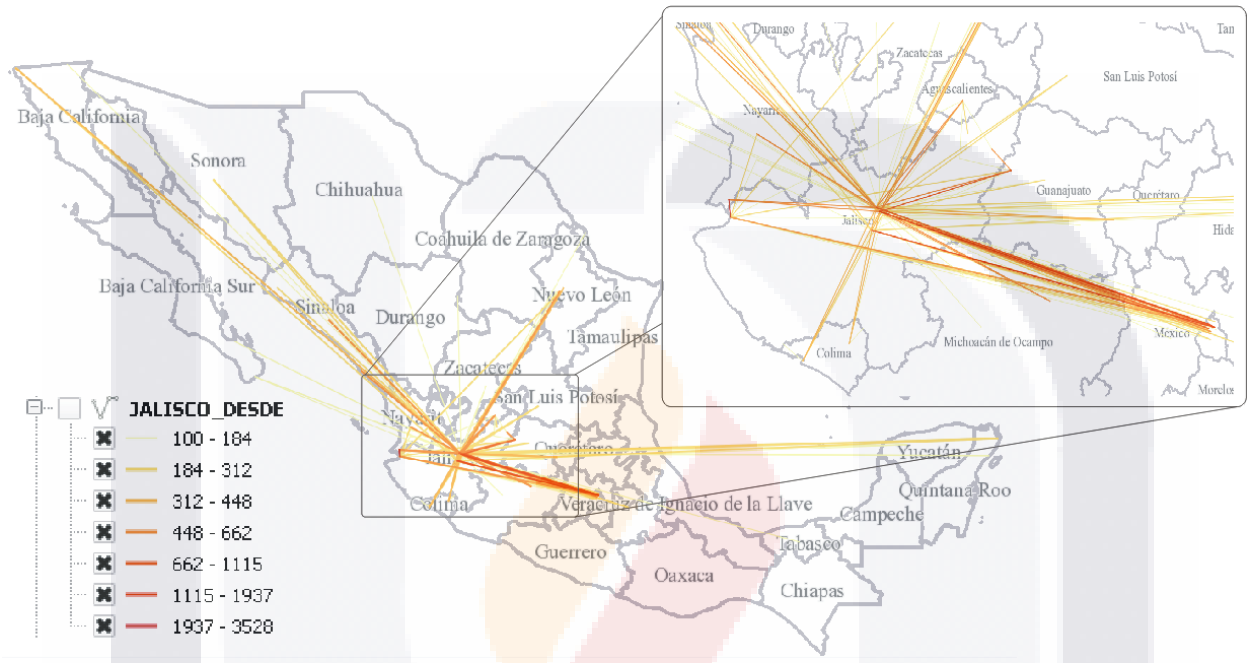
Figura 30. Movimientos registrados dentro del Estado de Jalisco

En la tabla de la Figura 30 se puede observar que la mayor cantidad de movimientos realizados por los usuarios de Twitter en el Estado de Jalisco durante el tiempo de recolección de datos fueron entre los Municipios de Guadalajara, Zapopan, San Pedro Tlaquepaque y Tlajomulco de Zúñiga que en conjunto con Ixtlahuacán de los Membrillos, Juanacatlán, el Salto y Tonalá forman parte de la zona metropolitana de Guadalajara (Consejo Nacional de Población (Mexico) & Instituto Nacional de Estadística, Geografía e Informática (Mexico), 2004).

Si se compara con otras estadísticas como el número total de habitantes, los Municipios que aparecen en el análisis de movilidad son los que contaban con más población en el año 2014 según (Instituto Nacional de Estadística y Geografía, 2014a) teniendo 1,500,821 en Guadalajara, 1,324,360 en Zapopan, 644,491 en San Pedro Tlaquepaque y 523,620 en Tlajomulco de Zúñiga. Del mismo modo considerando el número total de planteles, aulas, bibliotecas y talleres a inicios del ciclo escolar 2012-2013 se observó que son de los Municipios que mayor cantidad tienen con 1,136 planteles y 13,097 aulas para Guadalajara, 960 y 10,895 para Zapopan, 368 y 3,545 en San Pedro Tlaquepaque y 281 planteles y 2,819 aulas en Tlajomulco de Zúñiga. Con estas estadísticas y con otro análisis a mayor detalle podría determinarse si existe relación entre estas variables y la movilidad registrada de los usuarios de la red social de Twitter.

La Figura 31 muestra los movimientos realizados desde el Estado de Jalisco hacia otros Municipios pertenecientes a otros Estados de la República Mexicana. Se sigue aplicando el mismo criterio de asignación de colores en el mapa para indicar la frecuencia de movimientos de los usuarios de Twitter.

Mapa



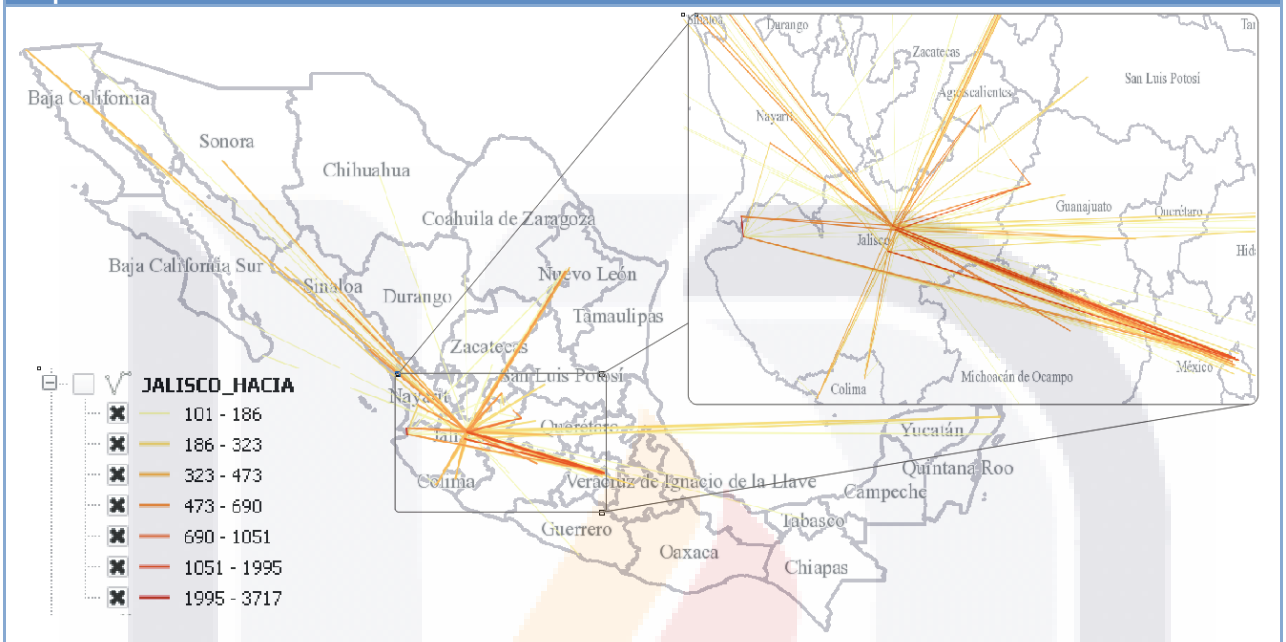
Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre del Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre del Municipio	Total de Movimientos
14	Jalisco	067	Puerto Vallarta	18	Nayarit	020	Bahia de Banderas	3,528
14	Jalisco	039	Guadalajara	09	Cd. México	015	Cuauhtémoc	1,937
14	Jalisco	097	Tlajomulco de Zuñiga	09	Cd. México	017	Venustiano Carranza	1,733
14	Jalisco	120	Zapopan	09	Cd. México	015	Cuauhtémoc	1,115
14	Jalisco	039	Guadalajara	09	Cd. México	017	Venustiano Carranza	1,024
14	Jalisco	039	Guadalajara	09	Cd. México	016	Miguel Hidalgo	809
14	Jalisco	039	Guadalajara	11	Guanajuato	020	León	731
14	Jalisco	120	Zapopan	09	Cd. México	017	Venustiano Carranza	718
14	Jalisco	120	Zapopan	25	Sinaloa	012	Mazatlán	662
14	Jalisco	039	Guadalajara	09	Cd. México	014	Benito Juarez	644
Resto								29,754

Figura 31. Movimientos registrados desde del Estado de Jalisco

Por ser muy parecido el comportamiento de la movilidad de la Figura 31 y la Figura 32 se explicará después de presentar esta última, la cual consiste en mostrar los movimientos realizados desde otros Estados de la República Mexicana hacia Jalisco.

Mapa



Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre del Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre del Municipio	Total de Movimientos
18	Nayarit	020	Bahia de Banderas	14	Jalisco	067	Puerto Vallarta	3,717
09	Cd. México	015	Cuauhtémoc	14	Jalisco	039	Guadalajara	1,995
09	Cd. México	017	Venustiano Carranza	14	Jalisco	097	Tlajomulco de Zúñiga	1,806
09	Cd. México	017	Venustiano Carranza	14	Jalisco	039	Guadalajara	1,432
09	Cd. México	015	Cuauhtémoc	14	Jalisco	120	Zapopan	1,051
09	Cd. México	017	Venustiano Carranza	14	Jalisco	120	Zapopan	1,026
09	Cd. México	016	Miguel Hidalgo	14	Jalisco	039	Guadalajara	793
11	Guanajuato	020	León	14	Jalisco	039	Guadalajara	772
09	Cd. México	017	Venustiano Carranza	14	Jalisco	067	Puerto Vallarta	690
25	Sinaloa	012	Mazatlán	14	Jalisco	120	Zapopan	682
Resto								29,682

Figura 32. Movimientos registrados hacia el Estado de Jalisco

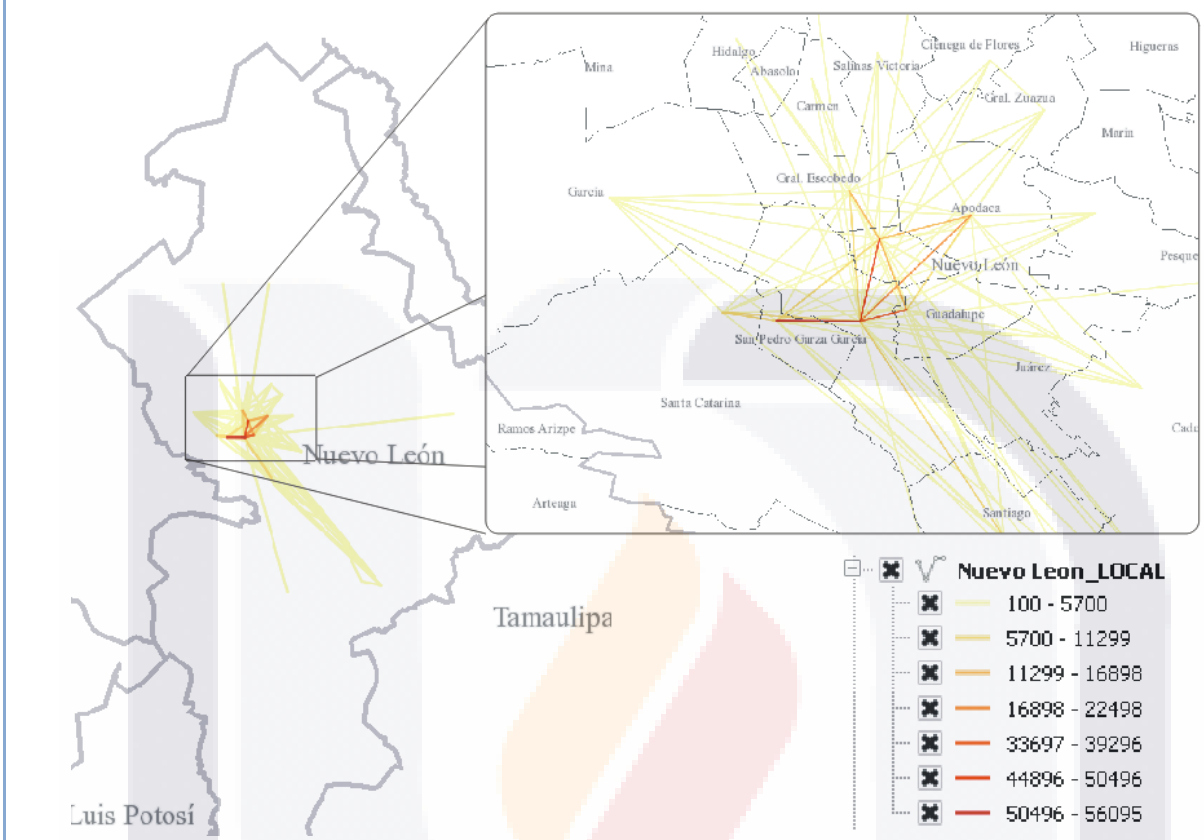
En la tabla de la Figura 31 se puede apreciar que la mayor cantidad de movimientos realizados por los usuarios de Twitter hacia dentro o fuera del Estado de Jalisco es desde el Municipio de Puerto Vallarta al Municipio de Bahía de Banderas en Nayarit con un total de 3,528 movimientos. Estos Municipios se caracterizan por ser destino turístico de playa de la región además de que conforman una de las tres áreas metropolitanas que hay en el Estado de Jalisco y parte de Nayarit. También es posible apreciar que otros Estados con los que tiene bastante relación son la Ciudad de México, Guanajuato y Sinaloa.

Movilidad en el Estado de Nuevo León

Otra de los Estados con gran actividad en el país es Nuevo León que de acuerdo con cifras de INEGI tiene une extensión territorial que representa el 3.27% del territorio nacional con una población de 5, 119, 504 habitantes que equivale al 4.3% del total del país. Su principal actividad económica es el comercio y tiene una aportación del 7.3% al PIB nacional (INEGI, 2010).

La Figura 33 muestra los movimientos realizados dentro del Estado de Nuevo León. Al igual que en las figuras anteriores los colores más oscuros representan una mayor frecuencia de la movilidad realizada entre dos Municipios y los colores más tenues representan la frecuencia menor de movimientos.

Mapa



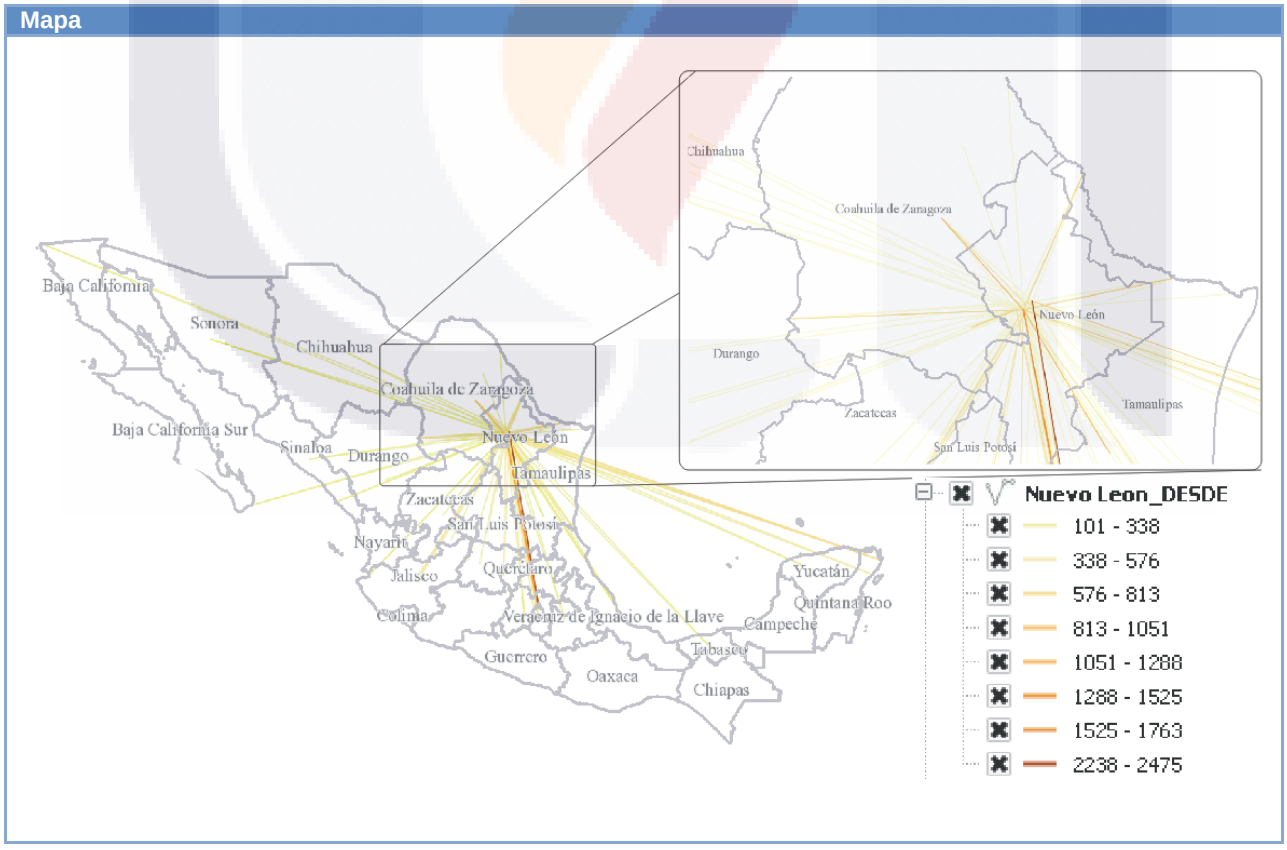
Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre del Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre del Municipio	Total de Movimientos
19	Nuevo León	019	San Pedro Garza Garcia	19	Nuevo León	039	Monterrey	56,095
19	Nuevo León	039	Monterrey	19	Nuevo León	019	San Pedro Garza Garcia	55,624
19	Nuevo León	039	Monterrey	19	Nuevo León	046	San Nicolás de los Garza	50,165
19	Nuevo León	046	San Nicolás de los Garza	19	Nuevo León	039	Monterrey	50,071
19	Nuevo León	039	Monterrey	19	Nuevo León	026	Guadalupe	36,316
19	Nuevo León	026	Guadalupe	19	Nuevo León	039	Monterrey	35,761
19	Nuevo León	039	Monterrey	19	Nuevo León	006	Apodaca	18,238
19	Nuevo León	006	Apodaca	19	Nuevo León	039	Monterrey	17,564
19	Nuevo León	046	San Nicolás de los Garza	19	Nuevo León	021	General Escobedo	13,012
19	Nuevo León	046	San Nicolás de los Garza	19	Nuevo León	006	Apodaca	13,005
Resto								276,646

Figura 33. Movimientos registrados dentro del Estado de Nuevo León

En la tabla de la Figura 33 se puede observar que la mayor cantidad de movimientos realizados por los usuarios de Twitter en el Estado de Nuevo León fueron hechos entre los Municipios de Monterrey, San Pedro Garza García, San Nicolás de los Garza, Guadalupe, Apodaca y General Escobedo, los cuales, junto con García, Juárez, Salinas Victoria, Santa Catarina y Santiago forman parte de la zona metropolitana de Monterrey. De los resultados presentados en la tabla de la Figura 33 se puede apreciar que tienen un patrón de movimientos muy similar los cuales pueden representar la entrada y salida de un Municipio a otro como el caso de San Pedro Garza García y Monterrey que tienen un total de 56,095 movimientos, y en sentido contrario, es decir de Monterrey a San Pedro Garza García tiene 55,624. Este comportamiento se repite para el resto de los Municipios con excepción de San Nicolás de los Garza y General Escobedo.

La Figura 34 muestra los movimientos realizados desde el Estado de Nuevo León hacia otros Municipios pertenecientes a otros Estados de la República Mexicana. Los colores más oscuros representan una mayor frecuencia de la movilidad realizada entre dos Municipios y los colores más tenues representan la frecuencia menor de movimientos.



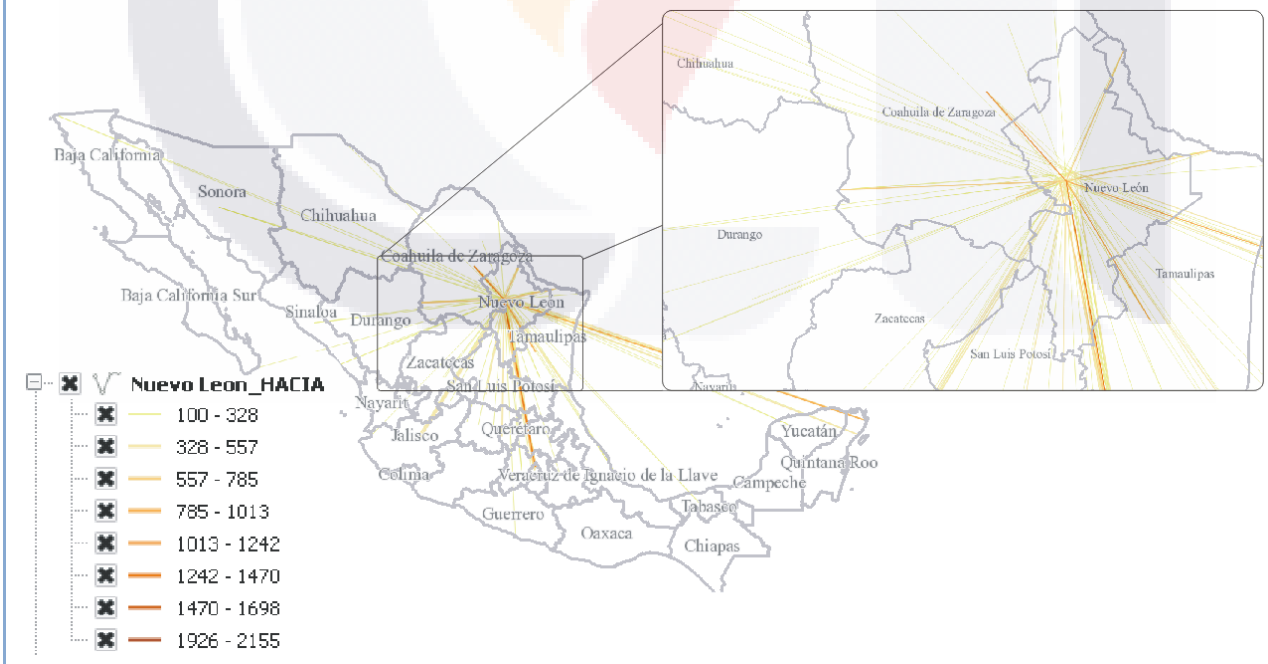
Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre Municipio	Total de Movimientos
19	Nuevo León	006	Apodaca	09	Cd. México	017	Venustiano Carranza	2,475
19	Nuevo León	039	Monterrey	05	Coahuila de Zaragoza	030	Saltillo	2,135
19	Nuevo León	039	Monterrey	09	Cd. México	015	Cuauhtémoc	1,392
19	Nuevo León	039	Monterrey	09	Cd. México	017	Venustiano Carranza	1,195
19	Nuevo León	039	Monterrey	05	Coahuila de Zaragoza	018	Monclova	1,044
19	Nuevo León	039	Monterrey	28	Tamaulipas	041	Victoria	875
19	Nuevo León	039	Monterrey	23	Quintana Roo	005	Benito Juarez	810
19	Nuevo León	039	Monterrey	05	Coahuila de Zaragoza	035	Torreon	771
19	Nuevo León	039	Monterrey	28	Tamaulipas	032	Reynosa	770
19	Nuevo León	039	Monterrey	09	Cd. México	016	Benito Juarez	762
Resto								24,432

Figura 34. Movimientos registrados desde el Estado de Nuevo León

La Figura 35 muestra los movimientos realizados desde los Municipios de otros Estados hacia el Estado de Nuevo León.

Mapa



Tabla

Entidad Origen	Nombre Entidad	Municipio Origen	Nombre del Municipio	Entidad Destino	Nombre Entidad	Municipio Destino	Nombre del Municipio	Total de Movimientos
09	Cd. México	017	Venustiano Carranza	19	Nuevo León	006	Apodaca	2,383
05	Coahuila de Zaragoza	030	Saltillo	19	Nuevo León	039	Monterrey	2,174
09	Cd. México	017	Venustiano Carranza	19	Nuevo León	039	Monterrey	1,762
09	Cd. México	015	Cuauhtémoc	19	Nuevo León	039	Monterrey	1,390
05	Coahuila de Zaragoza	018	Monclova	19	Nuevo León	039	Monterrey	1,083
23	Quintana Roo	005	Benito Juarez	19	Nuevo León	039	Monterrey	902
28	Tamaulipas	041	Victoria	19	Nuevo León	039	Monterrey	856
09	Cd. México	016	Miguel Hidalgo	19	Nuevo León	039	Monterrey	782
05	Coahuila de Zaragoza	035	Torreon	19	Nuevo León	039	Monterrey	766
28	Tamaulipas	032	Reynosa	19	Nuevo León	039	Monterrey	739
Resto								23,766

Figura 35. Movimientos registrados hacia el Estado de Nuevo León

Movilidad mensual por Estado

Por otro lado también se obtuvieron los movimientos realizados en los meses que cubren el periodo de Enero de 2014 hasta Febrero de 2015 entre los distintos Municipios a nivel nacional, (mostrados en la Tabla 15) teniendo que los tres en los que más movimientos registraron los usuarios de Twitter son Mayo, Octubre y Junio.

Tabla 15.Cantidad de movimientos realizados por los usuarios de Twitter durante el periodo de Enero de 2014 hasta Febrero de 2015

Mes	Movimientos realizados	Porcentaje
Enero 2014	100,414	1.69%
Febrero 2014	551,804	9.29%
Marzo 2014	304,956	5.13%
Abril 2014	439,276	7.39%
Mayo 2014	614,305	10.34%
Junio 2014	587,326	9.88%
Julio 2014	530,321	8.92%
Agosto 2014	551,524	9.28%
Septiembre 2014	560,619	9.43%
Octubre 2014	599,749	10.09%
Noviembre 2014	509,373	8.57%
Diciembre 2014	293,768	4.94%
Enero 2015	244,295	4.11%
Febrero 2015	54,716	0.92%
Total	**5,942,446**	**100.00%**

En las siguientes tablas (Tabla 16, Tabla 17, Tabla 18, Tabla 19, Tabla 20, Tabla 21, Tabla 22, Tabla 23, Tabla 24, Tabla 25, Tabla 26, Tabla 27, Tabla 28 y la Tabla 29) se presenta información a nivel Estado las cuales, a pesar de presentar un menor grado de detalle geográficamente hablando no dejan de ser importantes. Analizando las tablas en conjunto se puede observar que los Estados entre los que se detectaron mayor número de movimientos a lo largo del año son la Ciudad de México, México, Morelos, Prueba, Hidalgo y Querétaro. El Estado de Jalisco aparece en el último semestre, mientras que Nuevo León aparece en el último trimestre del año. El Estado de Guerrero se puede encontrar en los meses de Enero, Febrero, Abril, Julio y Agosto y el Estado de Quintana Roo en el mes de Diciembre muchos de estos meses tienen días de vacaciones de acuerdo al calendario escolar de la Secretaria de Educación Pública (SEP por sus siglas).

Tabla 16. Movimientos entre diferentes Estados en el mes de Enero de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	5,269
15	México	09	Cd. México	4,212
09	Cd. México	17	Morelos	836
09	Cd. México	12	Guerrero	407
09	Cd. México	21	Puebla	373
09	Cd. México	13	Hidalgo	313
09	Cd. México	14	Jalisco	282
21	Puebla	09	Cd. México	277
17	Morelos	09	Cd. México	272
09	Cd. México	22	Queretaro	266
Resto:				13,300

Tabla 17. Movimientos entre diferentes Estados en el mes de Febrero de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
15	México	09	Cd. México	27,759
09	Cd. México	15	México	27,344
17	Morelos	09	Cd. México	3,440
09	Cd. México	17	Morelos	3,066
21	Puebla	09	Cd. México	1,930
09	Cd. México	21	Puebla	1,772
13	Hidalgo	09	Cd. México	1,510
12	Guerrero	09	Cd. México	1,414
09	Cd. México	13	Hidalgo	1,380
22	Queretaro	09	Cd. México	1,308
Resto:				67,686

Tabla 18. Movimientos entre diferentes Estados en el mes de Marzo de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
15	México	09	Cd. México	15,279
09	Cd. México	15	México	14,925
17	Morelos	09	Cd. México	1,561
09	Cd. México	17	Morelos	1,381
21	Puebla	09	Cd. México	976
09	Cd. México	21	Puebla	949
14	Jalisco	09	Cd. México	767
09	Cd. México	22	Queretaro	757
09	Cd. México	14	Jalisco	750
22	Queretaro	09	Cd. México	748
Resto:				38,918

Tabla 19. Movimientos entre diferentes Estados en el mes de Abril de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	21,249
15	México	09	Cd. México	20,824
09	Cd. México	17	Morelos	2,682
17	Morelos	09	Cd. México	2,556
09	Cd. México	21	Puebla	1,630
21	Puebla	09	Cd. México	1,591
12	Guerrero	09	Cd. México	1,384
09	Cd. México	12	Guerrero	1,327
09	Cd. México	14	Jalisco	1,257
09	Cd. México	13	Hidalgo	1,231
Resto:				72,098

Tabla 20. Movimientos entre diferentes Estados en el mes de Mayo de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad	Total de Movimientos
15	México	09	Cd. México	30,772
09	Cd. México	15	México	30,437
17	Morelos	09	Cd. México	3,210
09	Cd. México	17	Morelos	3,028
21	Puebla	09	Cd. México	2,257
09	Cd. México	21	Puebla	2,169
13	Hidalgo	09	Cd. México	1,712
09	Cd. México	22	Queretaro	1,633
09	Cd. México	13	Hidalgo	1,616
22	Queretaro	09	Cd. México	1,609
Resto:				84,433

Tabla 21. Movimientos entre diferentes Estados en el mes de Junio de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
15	México	09	Cd. México	29,127
09	Cd. México	15	México	28,996
09	Cd. México	17	Morelos	2,747
17	Morelos	09	Cd. México	2,693
21	Puebla	09	Cd. México	1,896
09	Cd. México	21	Puebla	1,863
14	Jalisco	09	Cd. México	1,616
09	Cd. México	14	Jalisco	1,598
22	Queretaro	09	Cd. México	1,436
13	Hidalgo	09	Cd. México	1,382
Resto:				78,994

Tabla 22. Movimientos entre diferentes Estados en el mes de Julio de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	23,434
15	México	09	Cd. México	23,137
09	Cd. México	17	Morelos	2,646
17	Morelos	09	Cd. México	2,588
09	Cd. México	21	Puebla	1,831
21	Puebla	09	Cd. México	1,744
09	Cd. México	12	Guerrero	1,544
09	Cd. México	14	Jalisco	1,487
12	Guerrero	09	Cd. México	1,480
14	Jalisco	09	Cd. México	1,443
Resto:				90,509

Tabla 23. Movimientos entre diferentes Estados en el mes de Agosto de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
15	México	09	Cd. México	26,412
09	Cd. México	15	México	25,829
17	Morelos	09	Cd. México	2,694
09	Cd. México	17	Morelos	2,429
21	Puebla	09	Cd. México	2,026
09	Cd. México	21	Puebla	1,981
12	Guerrero	09	Cd. México	1,430
14	Jalisco	09	Cd. México	1,410
22	Queretaro	09	Cd. México	1,387
09	Cd. México	14	Jalisco	1,384
Resto:				78,321

Tabla 24. Movimientos entre diferentes Estados en el mes de Septiembre de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
15	México	09	Cd. México	26,805
09	Cd. México	15	México	26,650
09	Cd. México	17	Morelos	2,523
17	Morelos	09	Cd. México	2,427
21	Puebla	09	Cd. México	2,103
09	Cd. México	21	Puebla	2,002
14	Jalisco	09	Cd. México	1,713
09	Cd. México	14	Jalisco	1,701
22	Queretaro	09	Cd. México	1,567
09	Cd. México	22	Queretaro	1,538
Resto:				78,751

Tabla 25. Movimientos entre diferentes Estados en el mes de Octubre de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	28,097
15	México	09	Cd. México	27,064
09	Cd. México	21	Puebla	2,698
09	Cd. México	14	Jalisco	2,664
09	Cd. México	17	Morelos	2,633
21	Puebla	09	Cd. México	2,608
14	Jalisco	09	Cd. México	2,485
17	Morelos	09	Cd. México	2,483
09	Cd. México	19	Nuevo León	2,168
19	Nuevo León	09	Cd. México	2,066
Resto:				101,627

Tabla 26. Movimientos entre diferentes Estados en el mes de Noviembre de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
15	México	09	Cd. México	24,027
09	Cd. México	15	México	23,290
17	Morelos	09	Cd. México	2,617
09	Cd. México	17	Morelos	2,396
21	Puebla	09	Cd. México	2,290
09	Cd. México	21	Puebla	2,222
09	Cd. México	14	Jalisco	2,150
14	Jalisco	09	Cd. México	2,129
19	Nuevo León	09	Cd. México	1,707
09	Cd. México	19	Nuevo León	1,702
Resto:				91,762

Tabla 27. Movimientos entre diferentes Estados en el mes de Diciembre de 2014

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	12,978
15	México	09	Cd. México	12,752
14	Jalisco	09	Cd. México	1,500
09	Cd. México	14	Jalisco	1,492
09	Cd. México	17	Morelos	1,402
09	Cd. México	21	Puebla	1,247
17	Morelos	09	Cd. México	1,211
09	Cd. México	23	Quintana Roo	1,189
21	Puebla	09	Cd. México	1,159
09	Cd. México	19	Nuevo León	1,135
Resto:				57,719

Tabla 28. Movimientos entre diferentes Estados en el mes de Enero de 2015

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad Destino	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	11,911
15	México	09	Cd. México	11,871
09	Cd. México	17	Morelos	1,202
17	Morelos	09	Cd. México	1,189
21	Puebla	09	Cd. México	973
09	Cd. México	21	Puebla	903
09	Cd. México	14	Jalisco	830
14	Jalisco	09	Cd. México	822
23	Quintana Roo	09	Cd. México	789
09	Cd. México	13	Hidalgo	730
Resto:				38,472

Tabla 29. Movimientos entre diferentes Estados en el mes de Febrero de 2015

Clave Entidad Origen	Nombre Entidad Origen	Clave Entidad	Nombre Entidad Destino	Total de Movimientos
09	Cd. México	15	México	2,730
15	México	09	Cd. México	2,635
17	Morelos	09	Cd. México	291
09	Cd. México	17	Morelos	280
09	Cd. México	21	Puebla	204
21	Puebla	09	Cd. México	202
09	Cd. México	22	Queretaro	149
19	Nuevo León	09	Cd. México	144
14	Jalisco	09	Cd. México	144
09	Cd. México	19	Nuevo León	143
Resto:				6,849

En la Figura 36 se muestra a manera de síntesis el contenido de las tablas anteriores con la finalidad de tener un panorama más amplio en relación a los patrones de movilidad encontrados entre los Estados de la República Mexicana en los que se encuentra mayor actividad de los usuarios de Twitter entre las fechas de Enero de 2014 y Febrero de 2015.Si bien, los valores absolutos de movimientos posiblemente no son estadísticamente representativos ya que se toman de una muestra de tweets que de origen ya muestra un sesgo, es posible establecer patrones que son útiles para la toma de decisiones ya que de manera general muestran el comportamiento que se encontró en los datos de otros estudios al identificar los puntos geográficos en los que existen mayor cantidad de movimientos (Hawelka Bartosz et al., 2013) (Zagheni et al., 2014) (Gabrielli Lorenzo et al., 2014). Para este estudio es posible apreciar que hay mucha relación entre el número de movimientos de una ciudad de origen X y otra destino Y con la misma ciudad Y como origen y la X como destino como es el caso del Estado de México- Ciudad de México- Estado de México; Morelos- Ciudad de México - Morelos y Puebla- Ciudad de México – Puebla.

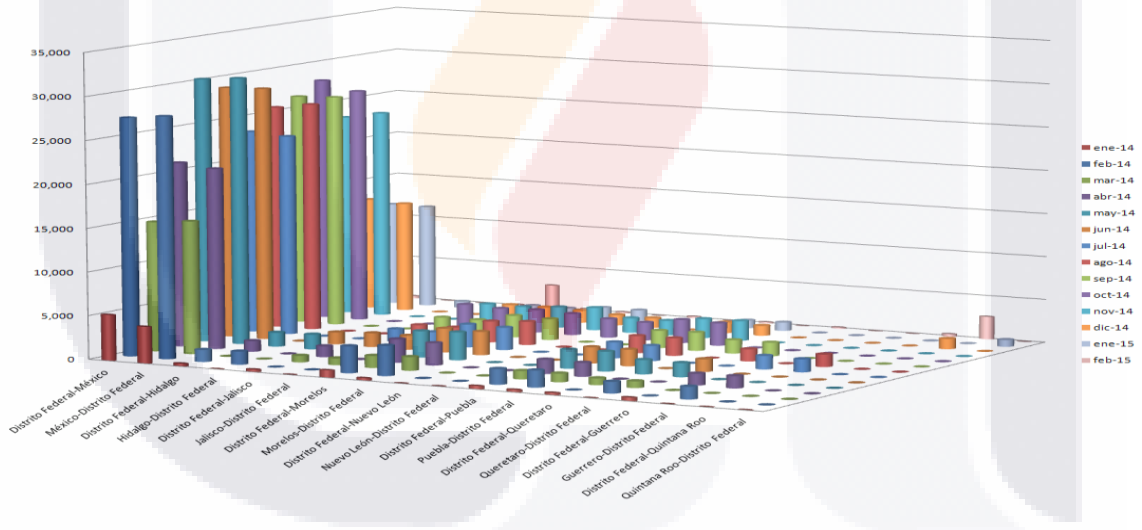


Figura 36. Resumen de movimientos entre Estados con mayor actividad durante todo el año. Elaboración propia.

Fase 8. Monitoreo y medición de los resultados

Como el tipo de estudio del caso práctico es con fines exploratorios, las actividades de monitoreo y medición de los resultados constó en la verificación de los criterios de aceptación y su cumplimiento, además se detectaron áreas de oportunidad para próximos estudios de movilidad humana.

Aplicación del método en el caso práctico de análisis de impacto de eventos de la vida real en la red social

Fase 1. Descripción del problema

1.1 Identificación del problema

El impacto que tienen ciertos eventos como las crisis financieras, epidemias, movimientos sociales, etc. en la sociedad afecta de diversas maneras pudiendo verlas reflejadas en los comentarios que hacen en Internet mediante el uso de sitios públicos como blogs, foros y redes sociales llegando a generar comportamientos colectivos.

La medición del impacto que pueden tener dichos eventos puede ser de gran utilidad para planificar y mejorar las acciones que permitan sacar conclusiones de ellos y aprender para situaciones posteriores.

La problemática que se trata de resolver es determinar si es posible medir el impacto de eventos de diferentes índoles mediante el análisis de los mensajes registrados en Twitter utilizando como base los parámetros de frecuencia de menciones y longitud en el tiempo.

1.2 Impacto al negocio

Esta investigación se realizó en conjunto con otras pruebas piloto como parte de un experimento para determinar si se cuenta con los conocimientos y los elementos necesarios para emprender el desarrollo de otros proyectos de Big Data, y del mismo modo realizar análisis exploratorios con la fuente de datos seleccionada para diagnosticar el valor que ofrece al INEGI al analizarla.

1.3 Antecedentes

No existe registro dentro del INEGI de haber realizado un análisis de impacto de eventos pero es posible encontrar trabajos de investigación recientes con una amplia variedad de temas entre los que están la política, sociales (Luis Cesar Torres Nabel, 2014) (Congosto et al., 2013), tecnológicos (Luis César Torres Nabel, 2009) y culturales. El potencial que tiene esta clase de análisis es tan alto que varias organizaciones han desarrollado herramientas de software que ayudan a medir el impacto que tiene cualquier evento en Twitter y lo ofrecen como un servicio

en el que brindan información como la cantidad de tweets, el alcance, la fuente, idioma, los contribuidores más activos, además de proporcionar informes en distintos formatos. Algunos ejemplos de estas herramientas son: Twitterbinder (https://www.tweetbinder.com/), TweetReach (https://tweetreach.com/), follow the hashtag (http://www.followthehashtag.com/) .

1.4 Condiciones en las que ocurre

Para el INEGI este problema ocurre tras la necesidad de realizar una serie de análisis exploratorios con fuentes de Big Data para generar información actualizada y coherente del impacto que tienen en los usuarios mexicanos de la red social de Twitter de ciertos eventos de interés nacional. Del mismo modo se viene trabajando con otros proyectos piloto con la participación y en coordinación con organismos internacionales.

1.5 Definición de objetivos

El objetivo de este caso práctico fue planteado desde el comienzo de este trabajo y fue formulado de manera precisa y clara de manera que no existiera ambigüedad respecto al tipo de respuesta esperado. El objetivo es declarado de la siguiente manera:

“Probar el método propuesto para examinar la capacidad de medir el impacto de eventos en las conversaciones registradas en Twitter a través de los parámetros de frecuencia de menciones y longitud en el tiempo.”

1.6 Identificación de las fuentes de datos a utilizar

Al igual que el caso práctico anterior, los datos fueron obtenidos de la red social de Twitter por ser públicos y gratuitos además de que cumplen con las características necesarias para considerarlo como Big Data (velocidad, volumen y variedad). Con esta fuente de datos se han realizado una gran cantidad de trabajos de investigación en distintos campos.

1.7 Definición del alcance

Para la realización de este caso práctico se definieron un conjunto de reglas que sirvieron para determinar lo que está dentro y fuera de las fronteras del estudio para llevar un mejor control y no desviarse de los objetivos planteados.

Los datos fueron obtenidos de la red social de Twitter durante el periodo de Junio a Septiembre de 2015 utilizando un filtro para recolectar únicamente aquellos tweets que fueron publicados con su respectiva referencia geográfica y que estén dentro de las coordenadas que enmarcan el territorio de la República Mexicana (INEGI, 1991) y son:

- Norte: 32° 43´ 06´´ latitud norte o 32.71865357 en representación decimal, en el Monumento 206, en la frontera con los Estados Unidos de América (3 152.90 kilómetros).
- Sur: 14° 32´ 27´´ latitud norte o 14.53209836 en representación decimal, en la desembocadura del río Suchiate, frontera con Guatemala (1 149.8 kilómetros).
- Este: 86° 42´ 36´´ longitud oeste o -86.71040527 en representación decimal, en el extremo suroeste de la Isla Mujeres.
- Oeste: 118° 27´ 24´´ longitud oeste o -118.40764955 en representación decimal, en la Punta Roca Elefante de la Isla de Guadalupe, en el Océano Pacífico.

Figura 37. Mapa de la República Mexicana con división política y ejes geográficos Fuente: (INEGI, 1991)

Además de esto se definieron los eventos sobre los cuales se realizó el análisis, para lo cual se tomó en consideración la aplicación web de ("Tendencias de Google - Listas de búsquedas más populares sobre Todas las categorías", 2015) para ver que búsquedas fueron las más populares del año. La lista de los diez eventos más buscados es:

1. Huracán Patricia
2. Joan Sebastián
3. 50 sombras de Grey
4. Lorena Rojas
5. Chapo Guzmán
6. Jurassic World
7. Rápido y furioso 7
8. Mi corazón es tuyo
9. Copa Oro 2015
10. Carmen Aristegui

Para el caso práctico se eligieron dos eventos:

1. El evento de la fuga del narcotraficante Joaquín "El Chapo" Guzmán con interés de seguridad pública.
2. El evento de entrega de premios los Kids Choice Awards México en su edición 2015 que aunque no fue de los temas más buscados en Google si tuvo un impacto en la red social de Twitter al considerarla como medio para votar por los candidatos a ser nominados en las diferentes categorías de premios.

Para el caso del análisis del impacto de la fuga del chapo se hizo una revisión en las publicaciones de sitios de noticias y revistas electrónicas para definir el conjunto de palabras relevantes al evento para buscarlas en los tweets:

- chapo guzm[aá]n
- chapo
- joaqu[ií]n guzm[aá]n
- altiplano
- narcotr[aá]ficante
- cartel de Sinaloa
- guzm[aán] loera

Para el caso del análisis del impacto de los KCA México 2015 se realizó el mismo tipo de investigación de palabras y se consideraron las siguientes para buscar en los tweets:

- kca
- kcamexico
- auditorio nacional
- nick
- nickelodeon
- kids choice
- choice awards m[eé]xico

Del mismo modo y para ambos casos no se contemplaron aquellos juegos de caracteres que representaban emoticones.

Fase 2. Diseño conceptual de la investigación

2.1 Definición de preguntas de investigación

La pregunta de investigación de este caso práctico fue planteada desde el comienzo de este trabajo la cual fue formulada de manera precisa y clara de manera que no existiera ambigüedad respecto al tipo de respuesta esperado. La pregunta se presenta en el siguiente párrafo:

¿Es posible medir el impacto que tienen ciertos eventos con el método propuesto en las conversaciones registradas en Twitter a través de los parámetros de frecuencia de menciones y longitud en el tiempo?

2.2 Definición de hipótesis

Como en este caso práctico no se trata de pronosticar ningún dato o hecho en las variables sino más bien describir el comportamiento de los mismos con relación al impacto que tienen ciertos eventos de la vida real afectando directamente lo que expresan los usuarios en sus publicaciones de la red social de Twitter se consideró irrelevante la definición de hipótesis.

2.3 Detección de variables

Igual que el punto anterior, la actividad relacionada con la detección de variables dependientes e independientes del método no aplica debido a que no existen hipótesis a probar por las características del estudio.

2.4 Perfilar variables

Teniendo como base la estructura del tweet revisado en el marco teórico se detectaron que las siguientes variables son útiles para la elaboración del ejercicio de análisis de impacto de eventos:

- Created_at: Variable que contiene información del tiempo universal coordinado (UTC acrónimo en inglés de Coordinated Universal Time) en el que es creado el tweet. El UTC es el estándar de tiempo por el cual el mundo regula la hora tomando como referencia un reloj atómico que se ajusta de acuerdo a las medidas de rotación de la tierra. Un ejemplo de la manera en la que viene el dato es: Ago 27 13:08:45 +0000 2015.
- Text: Es el texto en formato UTF-8 correspondiente a lo que publicó el usuario en el tweet.

2.5 Selección de técnicas y herramientas para el análisis

El tipo de análisis que se realizó sobre el conjunto de datos recolectados fue empleando técnicas de estadística descriptiva, la cual tiene como principal objetivo poner de manifiesto las características más importantes de los datos y sintetizarlas en gráficas y tablas.

De los elementos externos que se utilizaron para el análisis y la representación visual de los datos están:

- Generador gratuito de nubes de palabras en línea para la presentación de resultados.
- Microsoft Excel para la generación de gráficas y tablas de resumen.

Dos de las técnicas o algoritmos que se utilizaron para la preparación de los datos y su análisis fueron:

Algoritmo para obtener el impacto de un evento

El algoritmo que se siguió para determinar que tweets tienen relación con el evento así como su frecuencia por día fue el siguiente:

1. Definir el conjunto de palabras a buscar.
2. Iterar los tweets del conjunto de datos.
    a. Si el tweet contiene una o más palabras del conjunto definido en el paso anterior.
        i. Incluir tweet para su análisis
    b. Si no lo contiene
        i. Excluir el tweet.
3. Calcular la frecuencia de los tweets de manera diaria.
4. Calcular el total de tweet publicados.

Algoritmo para la extracción de las palabras por frecuencia

Para la extracción de las palabras que más sonaron de un evento determinado en la red social se siguió el siguiente proceso:

1. Definir un conjunto de palabras que quedaran excluidas del análisis por no tener relevancia para el estudio.
2. Dividir el texto de los tweets seleccionados en el algoritmo anterior mediante el uso de un delimitador, en este caso fueron: espacio en blanco, coma, punto, salto de línea y las comillas.
3. Excluir las palabras definidas en el paso 1.
4. Contar palabra por palabra e incrementar aquellas que se repitan n número de veces
5. Iterar los tweets del conjunto de datos.
6. Ordenar los resultados por mayor frecuencia.

2.6 Definición de la infraestructura y herramientas de software

Con la finalidad de realizar distintas pruebas y demostrar que cualquier organización puede almacenar, procesar y analizar los datos de Twitter, se instaló una infraestructura compuesta por seis equipos de cómputo interconectados entre sí y con salida a internet para poder recolectar los datos provenientes del API de Twitter.

Las características técnicas del clúster de equipos de cómputo son:

- Computadora marca HP modelo Compaq 6735B.
- AMD Turion X2 Ultra dual core a 1.4 GHZ.
- Disco duro 160 GB.
- Memoria RAM con 4 GB.
- Sistema Operativo de 64 bits con Linux.

Las características de la red son:

- Conexión a internet con 10 MB de velocidad de descarga.

Para realizar la preparación de los datos, los análisis exploratorios, los análisis definitivos y la visualización de resultados se ocupó únicamente un equipo de cómputo con las siguientes características técnicas:

- Computadora marca Dell modelo Optiplex 9020.
- Intel Core i7 4790 a 3.6 GHZ con 8 MB de cache.
- Disco duro de estado sólido de 128 GB.
- Disco duro SATA de 1 TB a 7200 RPM.
- Memoria RAM con 16 GB.
- Tarjeta gráfica PCIe con puerto HD de 2 GB.
- Sistema operativo Windows 8.1 Enterprise.

Hablando de hardware, y como se vio anteriormente, es de vital importancia contar para cada una de las capas de la arquitectura con la infraestructura tecnológica ya que es el elemento que se relaciona directamente con todos los procesos proveyendo la comunicación, el almacenamiento y la capacidad de procesamiento de la gran cantidad de datos que se requiere.

Se enlistan las plataformas y las herramientas de software que se utilizarán para realizar el estudio.

- Cuenta de usuario de Twitter.
- Aplicación para desarrollador en el API de Twitter.
- Claves de autenticación. (Clave del consumidor, secreto del consumidor, token de acceso, y el acceso secreto del token)
- Software para la recolección de datos, para el caso del INEGI de una fuente externa. El software a utilizar es ElasticSearch.
- Herramientas para la sincronización con los servicios del API de Twitter. El software de Logstash, el plugin de entrada de datos de Twitter con su funcionalidad extendida y el plugin de salida de Elasticsearch para el almacenamiento de los datos.
- La carga se llevó a cabo apoyándose del plugin Spark-CSV desarrollado por la compañía Databricks y el almacenamiento de los datos se realizó sobre Apache Spark.
- Para la transformación y procesamiento de los datos se desarrollaron scripts en el lenguaje de programación orientado a funciones, Scala, en conjunto con la librería SQL de Apache Spark.
- Para la generación de gráficas y tablas se utilizó Microsoft Excel.

2.7 Definición de la arquitectura

Las plataformas y herramientas vistas en la actividad anterior son primordiales para definir la arquitectura que se utilizó para realizar este caso práctico, la cual fue conformada por aquellas capas que tenían en común tanto las revisadas en el marco teórico como la que se estaba utilizando en el INEGI para los análisis exploratorios trabajados teniendo como resultado una arquitectura que se compone de la capa de fuentes de datos: preparación y almacenamiento de los datos, procesamiento de los datos, modelado y análisis, visualización y reportes. En la Figura 38 se puede apreciar de manera más clara como es que se relacionan los elementos (plataformas y herramientas) identificados para trabajar con Twitter dentro del INEGI.

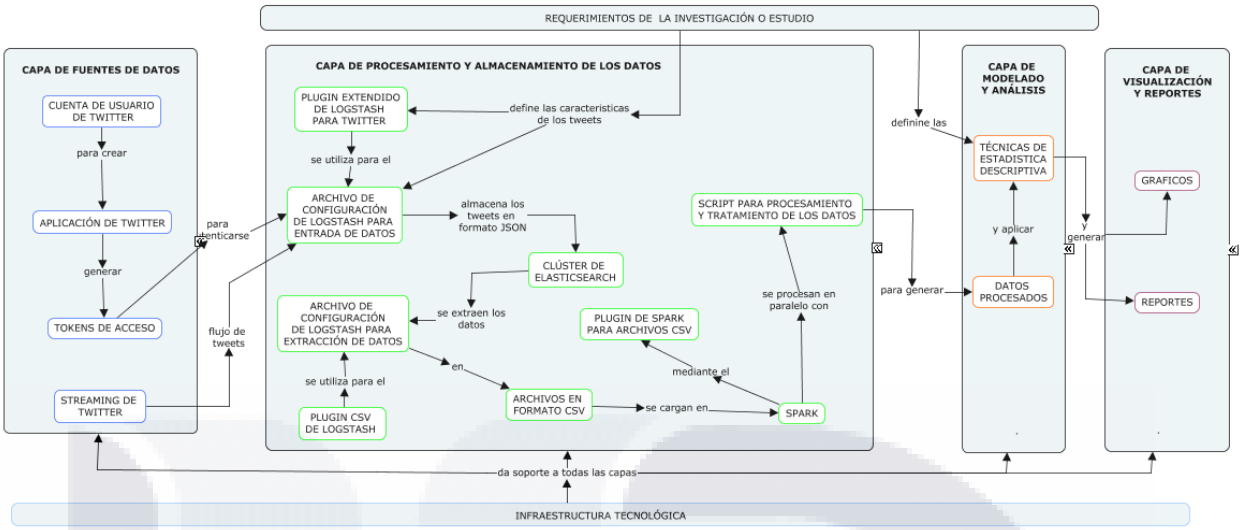


Figura 38. Mapa conceptual de elementos

2.8 Elaboración de programa de trabajo

Como el método contiene fases y actividades que pueden realizarse tantas veces como sea necesario el programa de trabajo consistió en realizar revisiones periódicas cada dos a cuatro semanas en las que se presentaban los avances realizados para determinar si el desarrollo del proyecto iba de acuerdo a lo especificado en los objetivos.

En la Tabla 30 se presenta un cronograma de trabajo en el que se muestran las actividades del método a partir de la Fase 3 especificando que rol de los involucrados en el proyecto la realiza y el tiempo planeado en una iteración para el caso práctico de análisis de impacto de eventos. Para este caso práctico no se consideraron relevantes contemplar las primeras fases del método para el cronograma por ser fases de definición del proyecto, sin embargo pueden llegar a considerarse para otros proyectos. Algo que salta a simple vista de observar la tabla es que para la recolección de la muestra representativa, perteneciente a la fase 3 y toda la fase 4 referente a la recolección de los datos, el tiempo que tomó obtenerla fue de cero días, esto se debe a que personal del INEGI ya había obtenido con los tweets previamente.

Tabla 30. Cronograma de trabajo con roles asignados y duración aproximada en una iteración, Elaboración propia.

Fase	Actividad	Duración	Rol asignado
Fase 3 Análisis exploratorio	1. Recolectar una muestra representativa	0 días	- Innovador de datos - Desarrollador de los datos
	2. Definir conjunto de criterios para el proceso de carga, limpieza y transformación de los datos	7 días	- Innovador de datos
	3. Aplicar las técnicas o desarrollar los algoritmos haciendo identificados en la fase anterior	7 días	- Desarrollador de los datos
	4. Presentar los resultados obtenidos con los datos muestra en graficas o tablas	7 días	- Innovador de datos - Desarrollador de los datos
	5. Determinar si las técnicas y las herramientas son las adecuadas para el análisis	3 días	- Innovador de datos
Fase 4 Recolección de los datos	1. Preparar y configurar las herramientas de software para almacenar los datos utilizando los elementos para conectarse a los servicios del proveedor de la fuente de Big Data	0 días	- Innovador de datos - Desarrollador de los datos
	2. Puesta en marcha del servicio para la recolección de los datos	0 día	- Desarrollador de los datos
Fase 5 Preparación de los datos	1. Aplicar las técnicas o desarrollar algoritmos para cargar, limpiar y transformar los datos	21 días	- Innovador de datos - Desarrollador de los datos
Fase 6 Análisis de datos y de resultados	1. Construcción del modelo	7 días	- Innovador de datos - Investigador de datos
	2. Evaluación del modelo	7 días	- Innovador de datos - Investigador de datos
	3. Representación y visualización de los datos	7 días	- Innovador de datos - Investigador de datos
Fase 7 Despliegue	1. Aplicación con los datos	3 días	- Investigador de datos
	2. Comunicación de los resultados	3 días	- Persona de negocios - Investigador de datos
Fase 8 Monitoreo y medición de los resultados	1. Medición	1 días	- Persona de negocios - Investigador de datos
	2. Monitoreo	1 días	- Persona de negocios - Investigador de datos

Fase 3. Análisis exploratorio

3.1 Recolectar una muestra significativa.

Para la realización del análisis exploratorio se eligió el tema de la fuga de Joaquín “El chapo” Guzmán utilizando los tweets recolectados durante el mes de Julio de 2014. El mes fue elegido empleando la técnica de muestreo aleatorio simple ya que permite obtener muestras representativas con gran rapidez y simpleza mediante la generación de números aleatorios donde cada elemento de la población tiene la misma probabilidad de ser elegido.

3.2 Definir conjunto de criterios para el proceso de carga, limpieza y transformación de los datos

Antes de pasar al análisis fue necesario observar los datos con el fin de conocer los dominios de valores que tenía cada una de las variables con los que se trabajaría para determinar si existían registros con errores, con inconsistencias o si era necesario reformatearlos.

Para el análisis de este caso práctico se contemplaron únicamente aquellos tweets que contenían alguna de las palabras previamente definidas para el evento de la fuga de Joaquín “El Chapo” Guzmán excluyendo del estudio todos aquellos que no cumplan con este criterio. Otra de las variables que se estandarizó fue la fecha debido a la variación que hay en los cuatro husos horarios del territorio nacional se optó por manejar el tiempo de la zona centro de la República Mexicana el cual se obtiene haciendo la sustracción de 6 horas al UTC brindado por Twitter (UTC-6) y posteriormente se le quitaron los datos referentes a los minutos, segundos y milisegundos para formatearlo en DD/MM/AAAA HR Ej. 18/02/2016 11 que corresponde al 18 de Febrero de 2016 a las 11 horas.

3.3 Aplicar las técnicas o desarrollar los algoritmos identificados en la fase anterior.

Teniendo los datos de la muestra limpios se procedió a desarrollar el algoritmo para obtener el impacto de un evento y el de extracción de palabras por frecuencia.

3.4 Presentar los resultados obtenidos con los datos muestra en graficas o tablas.

En la Figura 39 es posible ver el impacto que tuvo la fuga del chapo tomando como base la cantidad de mensajes publicados por los usuarios de la red social de Twitter. Al realizar el análisis del evento seleccionado se obtuvo que la cantidad de tweets relacionados con el chapo antes de su fuga tenían una frecuencia constante mientras que repentinamente el día 12 de Julio, uno posterior a la fuga, se incrementaron en grandes proporciones. Conforme fue pasando el tiempo la frecuencia fue bajando hasta tener casi el mismo comportamiento que antes del evento.

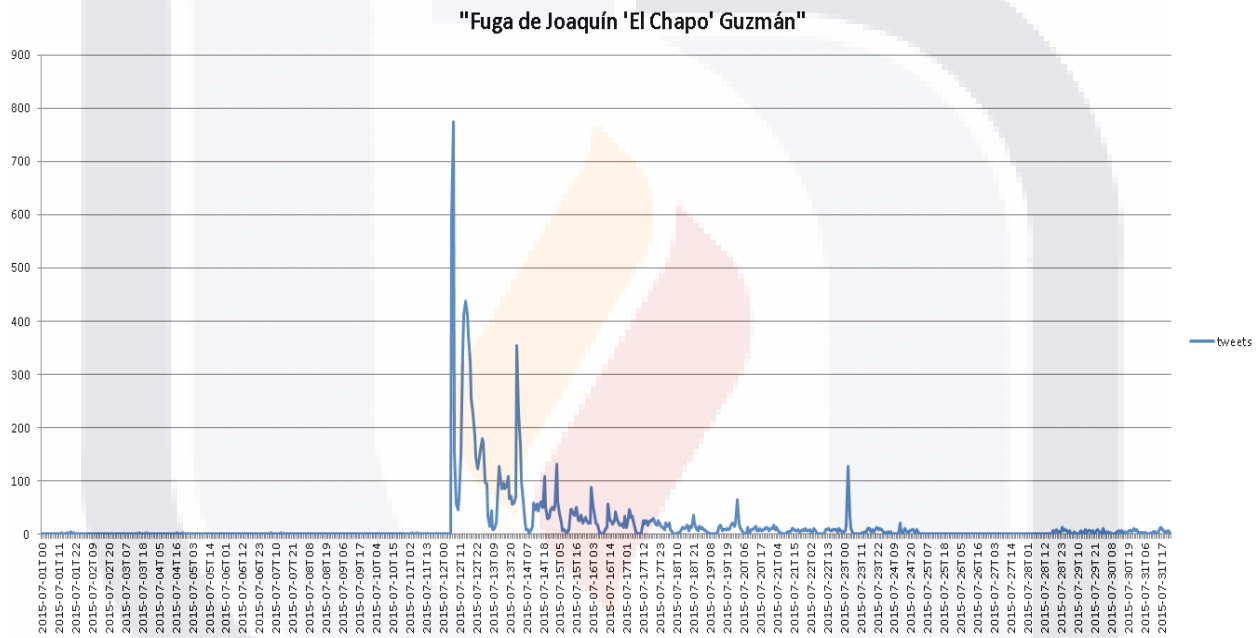


Figura 39. Frecuencia de los tweets relacionados con la fuga de "El Chapo"

Dentro de los tweets obtenidos en el proceso anterior se procedió a obtener que palabras eran las que tenían más frecuencia utilizando el algoritmo para la extracción de las palabras por frecuencia definido en la sección Selección de técnicas y herramientas para el análisis obteniendo palabras como Chapo con 10,044 veces, fuga con 2,473 y Guzmán fueron las que más ocasiones se repitieron. En la Tabla 31 se listan las diez palabras con mayor frecuencia.

Tabla 31. Diez palabras con mayor frecuencia en los tweets relacionados con la fuga de "El Chapo"

Palabra	Frecuencia
chapo	10,044
fuga	2,473
guzmán	1,312
penal	645
chapo"	634
más	557
túnel	550
escapó	518
altiplano	515
@epn	497

Estas diez palabras junto con otras cuarenta son presentadas en una nube de palabras en la Figura 40 donde las que tienen mayor frecuencia se presentan con un tamaño más grande mientras las que se repitieron en menor grado están más pequeñas.

Figura 40. Nube de palabras extraída de los tweets relacionados con la fuga de "El Chapo"

3.5 Determinar si las técnicas y las herramientas son las adecuadas para el análisis.

Como el fin del análisis exploratorio y en general de este caso práctico es mostrar el comportamiento de los datos, se encontró que las técnicas y las herramientas son útiles para obtener el impacto de un evento de la vida real y a su vez saber de qué temas, personas y acciones están publicando los usuarios de Twitter. Sin embargo hubo la necesidad de redefinir las palabras y términos debido a que había ocurrencias de estos en tweets que no estaban relacionados directamente con el evento, quedando la lista con los siguientes términos: chapoguzm[aá]n, chapo guzm[aá]n, el chapo, elchapo, joaqu[ií]n guzm[aá]n, joaqu[ií]nguzm[aá]n, altiplano, narcotr[aá]ficante, cartel de Sinaloa, guzm[aán] loera, guzm[aán]loera.

Fase 4. Recolección de los datos

4.1 Preparar y configurar las herramientas de software para almacenar los datos utilizando los elementos para conectarse a los servicios del proveedor de la fuente de Big Data.

Igual que en el caso práctico anterior, la forma de acceder a los datos es mediante el API de Twitter utilizando los datos de autenticación provistos por la organización.

Para la parte de extracción y almacenamiento de los tweets se utilizó el software conocido como ElasticSearch en conjunto de varias herramientas que trabajan con este fin, como es el caso de Logstash para la recolección de los tweets utilizando los plugins que vienen integrados en el software.

4.2 Puesta en marcha del servicio para la recolección de los datos.

Poner en marcha el servidor de Elasticsearch y lanzar el archivo de configuración de Logstash para empezar la descarga del flujo de datos de Twitter.

Fase 5. Preparación de los datos

5.1 Aplicar las técnicas o desarrollar algoritmos para cargar, limpiar y transformar los datos.

Una vez que el tiempo de recolección ha llegado a su fin es necesario extraer los datos del servidor de Elasticsearch para cargarlos y procesarlos en Spark. Para la extracción en este caso práctico se realizó almacenando los datos en un determinado número de archivos Parquet, que es un formato de almacenamiento columnar disponible en cualquier proyecto en el ecosistema Hadoop, facilitando en gran medida el paso de información entre Elasticsearch y Hadoop.

Para realizar la carga de los archivos con formato parquet, las librerías de Spark vienen con comandos que permiten trabajar con ellos como si fuera una entidad de una base de datos relacional para posteriormente limpiarlos, transformarlos y validarlos. El proceso que se siguió para realizar estas actividades en Spark fue:

Proceso de carga, limpieza y transformación de los datos.

1. Crear archivo Parquet conectando Apache Spark con Elasticsearch mediante el archivo de java (JAR por el acrónimo en inglés de Java Archive) correspondiente.
2. Cargar archivo Parquet con los tweets completos.
3. Obtener las variables identificadas para el estudio desde los tweets.
4. Convertir formato de fecha de acuerdo a las necesidades del trabajo de investigación, en este caso se designaron cuatro dígitos para el año, seguido por dos dígitos para hacer referencia al mes, dos dígitos para el día separados por un guion medio, y otros dos dígitos para la hora separado de los anteriores con un espacio en blanco (AAAA-MM-DD HH).
5. Formatear los textos de los tweets de forma que al momento de buscar las palabras no haya problemas al encontrarlas si viene con mayúsculas o minúsculas.
6. Identificar un conjunto de palabras que se relacionan con los eventos estudiados con el fin de encontrar a todos aquellos tweets que las contengan utilizando el algoritmo de la fase de selección de técnicas y herramientas para el análisis.
7. Excluir aquellos tweets que no tienen relación con el evento.
8. Excluir los emoticones encontrados en los tweets seleccionados del paso anterior.
9. Exportar el resultado en archivos con formato de salida CSV.

Fase 6. Análisis de datos y de resultados

6.1 Construcción del modelo

Para este trabajo el tipo de análisis se realizó sobre el conjunto de datos generados en la capa anterior empleando técnicas de estadística descriptiva, la cual tiene como principal objetivo poner de manifiesto las características más importantes de los datos y sintetizarlas en gráficas.

6.2 Evaluación del modelo

Las pruebas y validaciones de los algoritmos utilizados para generar las estadísticas de este caso práctico fueron realizadas bajo la supervisión de los expertos del área que laboran en el INEGI. Del mismo modo y para comprobar que la información generada al aplicar los algoritmos en los datos de Twitter es significativa se compararon con las gráficas generadas

con la herramienta Tendencias de Google (Google Trends por su traducción al inglés) para determinar si hay algún patrón de comportamiento similar.

6.3 Representación y visualización de los datos

El proceso para generar las tablas, gráficas y nubes de palabras es relativamente sencillo utilizando herramientas como Microsoft Excel y el generador gratuito de nubes de palabras en línea tagul.

1. Cargar los archivos resultantes del procesamiento de los datos en Microsoft Excel.
2. Acomodar los datos de acuerdo a las necesidades de las tablas y gráficas.
3. Seleccionar los datos y generar las tablas y gráficas necesarias.

Para generar las nubes de palabras en la herramienta Tagul basta con generar un archivo CSV con el conjunto de palabras obtenidas en el procesamiento de los datos, la frecuencia con la que se repiten y otra serie de parámetros indicados en el sitio web como el color, el ángulo y la fuente que tendrá la palabra.

Fase 7. Despliegue

7.1 Aplicación con los datos.

Como la elaboración de este caso práctico fue con fines exploratorios en esta fase se especifica el tipo de resultados que se generaron utilizando distintas gráficas de resumen de los datos para poder interpretarlos y determinar su utilidad.

7.2 Comunicación de los resultados

Como parte de la comunicación de los resultados se presentan las tablas y graficas generadas con la información extraída de los eventos seleccionados. Primero se revisó el impacto que tuvo el evento de la fuga del chapo a través de los parámetros de frecuencia de menciones y longitud en el tiempo.

Impacto del evento Fuga del Chapo Guzmán por fecha

El evento que se utilizó en esta fase fue la segunda fuga de Joaquín el Chapo Guzmán de una prisión de máxima seguridad en México el día 11 de Julio de 2015 alrededor de las 9:00 p.m. haciendo uso de un túnel que lo llevaría a una vivienda a casi 1.5 kilómetros de la prisión. Este hecho representó un duro golpe para el gobierno actual de México poniendo a las autoridades a trabajar para recapturar a uno de los hombres más buscados del mundo.

Rápidamente el evento ocasionó un alza de notas informativas a través de noticieros de televisión y radio, en periódicos y revistas y sin dejar de lado los comentarios de usuarios de Internet donde se vio reflejada la misma situación en blogs, foros y las redes sociales donde Twitter no fue excepción.

Para este análisis se pudo observar que el comportamiento que los usuarios de la red social de Twitter tenían antes del 11 de Julio era muy reservado respecto a escribir tweets sobre el chapo. Sin embargo conforme se dio a conocer la noticia el día siguiente las reacciones no se dejaron esperar convirtiéndolo en uno de los eventos más mencionados para esas fechas. El evento tuvo su mayor número de publicaciones durante un lapso de aproximadamente 15 días presentando alzas esporádicamente. En la Figura 41 se puede lo comentado anteriormente.

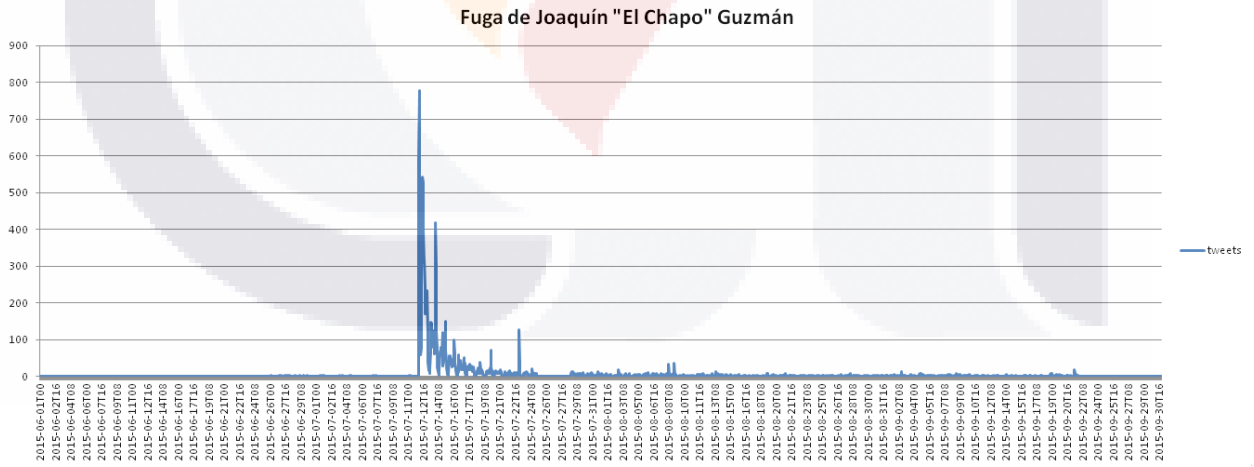


Figura 41. Impacto en Twitter del evento la fuga de Joaquín "El Chapo" Guzmán

Dentro de los tweets obtenidos en el proceso anterior en el lapso de tiempo en el que se presentó el evento diversas palabras fueron las que se repitieron con mayor frecuencia dando

una idea general de los temas que preocupan a los usuarios que publican sus puntos de vista en la red social de Twitter. Para obtener dichas palabras se utilizó el algoritmo para la extracción de las palabras por frecuencia definido en la sección Selección de técnicas y herramientas para el análisis obteniendo palabras con las que se muestran en la Tabla 32.

Tabla 32. Diez palabras con mayor frecuencia en los tweets relacionados con la fuga de "El Chapo"

Palabra	Frecuencia
Fuga	3,285
#chapoguzman	2,024
#elchapo	1,052
@epn	970
Túnel	767
México	669
Escape	641
Gobierno	611
#elchapoguzman	547
Ahora	527

Estas diez palabras junto con otras cuarenta son presentadas en una nube de palabras en la Figura 42 donde las que tienen mayor frecuencia se presentan con un tamaño más grande mientras las que se repitieron en menor grado están más pequeñas.

Figura 42. Nube de palabras extraída de los tweets relacionados con la fuga de "El Chapo"

El mismo evento fue revisado en la herramienta de tendencias de Google teniendo como resultado que la semana del 12 al 18 de Julio las consultas realizadas en la República Mexicana el número de búsquedas tuvo un alza repentina con temas como : el chapo, el chapo guzmán, chapo guzmán fuga, fuga del chapo, chapo guzmán 2015, noticias chapo guzmán y memes chapo guzmán , respectivamente. En la Figura 43 se muestra el comportamiento de los usuarios comentado anteriormente.

Figura 43. Impacto en Google Trends del evento la fuga de Joaquín "El Chapo" Guzmán

Analizando a detalle la gráfica generada mediante los datos recolectados de la red social de Twitter y la gráfica generada por la herramienta de tendencias de Google es posible ver que el comportamiento de ambas aplicaciones es muy similar. Para determinar si estadísticamente presentan o tienen alguna relación los resultados obtenidos sería conveniente realizar algún análisis estadístico a fondo, pero eso ya es un tema ajeno a este trabajo de investigación.

Impacto del evento Kids Choice Awards México 2015

Los Kids' Choice Awards México es la edición mexicana de los populares Nickelodeon's Kids Choice Awards realizados en los Estados Unidos donde los jóvenes eligen a sus artistas favoritos de TV, música, redes sociales, teatro, radio, entre otras. La edición México 2015

corresponde a la sexta y se llevó a cabo por primera vez en el Auditorio Nacional el día 15 de Agosto teniendo como anfitriones a Mario Bautista y la ex RBD Maite Perroni y presentaciones musicales de CD9, Heffron Drive, Urband5 y muchos más.

Como novedad para ese año implementaron una forma adicional a la que usaban anteriormente (a través de los sitios web mundonick.com, trendybynick.com y la aplicación móvil Nick App) en la que los fans pudieran votar por sus artistas favoritos utilizando la red social de Twitter a través del hashtag #KCAMexico y el nombre de su predilecto teniendo como fecha límite el 17 de Julio para posteriormente el 12 de Agosto anunciar a los nominados finales.

Para este análisis se pudo observar que el comportamiento que los usuarios de la red social de Twitter fue muy activo teniendo aproximadamente 6 alzas de publicaciones relacionadas con el evento. En la Figura 44 se puede observar cómo fue la reacción de los fans a lo largo del evento.

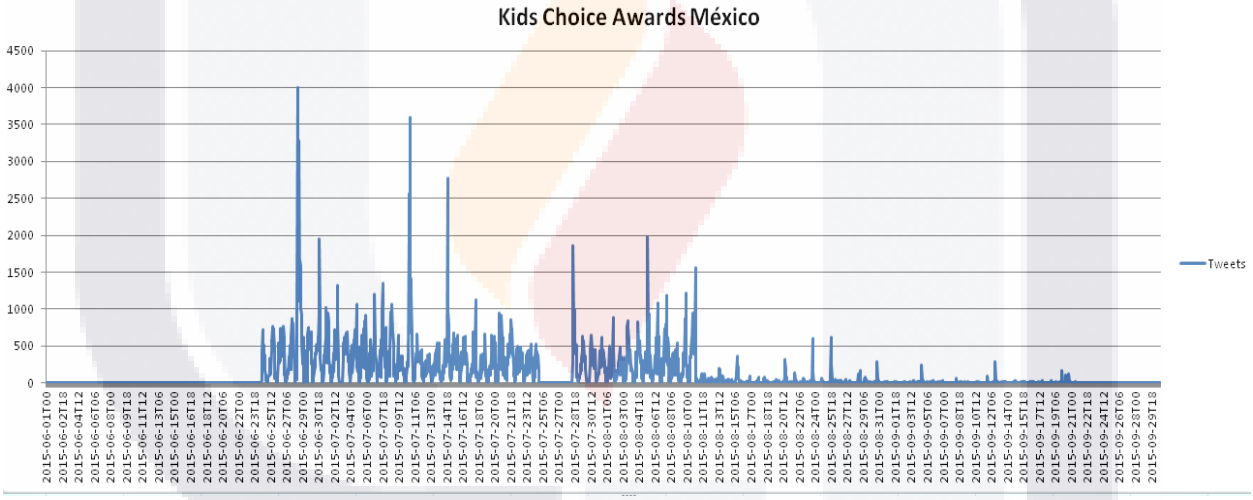


Figura 44. Impacto en Twitter del evento Kid's Choice Awards México 2015

Dentro de los tweets obtenidos en el proceso anterior en el lapso de tiempo en el que se realizó el evento diversas palabras fueron las que se repitieron con mayor frecuencia dando una idea general de los temas que preocupan a los usuarios que publican sus puntos de vista en la red social de Twitter respecto al evento. Para obtener dichas palabras se utilizó el algoritmo para la extracción de las palabras por frecuencia definido en la sección Selección de técnicas y herramientas para el análisis obteniendo palabras como las que se muestran en la Tabla 33.

Tabla 33. Diez palabras con mayor frecuencia en los tweets relacionados con el evento KCA México 2015

Palabras	Frecuencia
#kcamexico	250,193
voto	200,087
@mariobautista_	132,094
#venabailar	114,453
#mariobautista	113,120
#mariobautistaredes	111,783
Rt	69,879
#kcaméxico	66,803
#cd9	63,278
#meequivoque	61,133

Estas diez palabras junto con otras cuarenta son presentadas en una nube de palabras en la Figura 45 donde las que tienen mayor frecuencia se presentan con un tamaño más grande mientras las que se repitieron en menor grado están más pequeñas.

Figura 45. Nube de palabras extraída de los tweets relacionados con la fuga de "El Chapo"

Al observar la Figura 44 y para tener mayor entendimiento de lo que paso en el evento, surgió la necesidad de saber qué era lo que se estaba diciendo durante las alzas de tweets publicados por lo que se aplicó nuevamente el algoritmo para la extracción de palabras por frecuencia pero limitándolo únicamente a los días donde se registró mayor actividad los cuales

fueron el 28 y 30 de Junio, el 10 y el 14 de Julio y el 6 y el 10 de Agosto. Como resultado (Figura 46) de aplicar el algoritmo se extrajeron las diez palabras que más se repitieron durante los seis días sin tener mucha variación entre éstas sino en su frecuencia.

28/06/2015		30/06/2015		10/07/2015	
Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia	Palabra
17,633	Voto	5,202	#kcamexico	8,211	voto
17,080	#kcamexico	3,619	voto	7,914	#kcamexico
12,232	@mariobautista_	2,913	#kcaméxico	7,006	@mariobautista
11,722	#venabailar	2,749	#cd9	7,004	#venabailar
11,551	#mariobautistaredes	2,737	#meequivoque	6,748	#mariobautistaredes
6,508	Siguentes	2,412	rt	6,632	#mariobautista
5,788	Categorias	1,559	@mariobautista_	4,896	rt
4,259	1voto	1,556	#dannapaola	4,258	cada
4,104	#dannapaola	1,448	#mariobautista	3,919	siguentes
2,997	voto!	1,444	#venabailar	3,862	categorias

14/07/2015		05/08/2015		10/08/2015	
Frecuencia	Palabra	Frecuencia	Palabra	Frecuencia	Palabra
3,302	#kcamexico	7,594	#kcamexico	7,238	#kcamexico
2,664	Voto	6,911	voto	3,760	voto
1,774	#kcaméxico	5,819	@mariobautista_	2,760	#meequivoque
1,706	#mariobautistaredes	4,584	#mariobautista	2,756	#cd9
1,675	@mariobautista_	4,561	#venabailar	1,666	@mariobautista_
1,656	#mariobautista	4,508	#mariobautistaredes	1,339	#kcaméxico
1,584	#venabailar	2,821	cada	1,333	#saak
1,028	#cd9	2,740	1voto	1,062	@saakmx
981	#meequivoque	2,193	Rt	937	rt
851	#llevamedespacio	1,823	siguentes	911	juntos

Figura 46. Palabras con mayor frecuencia los días que se publicaron más tweets relacionados con los KCA México 2015

El mismo evento fue revisado en la herramienta de tendencias de Google entre los que sobresalen la semana del 28 de Junio al 4 de Julio y la semana del 16 al 22 de Agosto donde esta última representó el alza más importante de consultas realizadas en la República Mexicana haciendo búsquedas de temas como: 2015 Kids' Choice Awards - Award ceremony,

Nickelodeon Kids' Choice Awards - Award, respectivamente. En la Figura 47 se muestra el comportamiento de los usuarios comentado anteriormente.

Figura 47. Impacto en Google Trends del KCA México 2015

Analizando a detalle la Figura 44, generada con los datos de Twitter, y la Figura 47 generada mediante la herramienta de tendencias de Google es posible ver que el comportamiento de ambas aplicaciones presentan un alza de tweets y búsquedas en los últimos días del mes de Junio, sin embargo, el incremento registrado en las consultas hechas en el buscador de Google no concuerda con lo descubierto en Twitter, posiblemente sea porque la participación de los fans consistió en elegir a los candidatos en las diferentes categorías de premios hasta la fecha límite el 17 de Julio. Para determinar si estadísticamente presentan o tienen alguna relación los resultados obtenidos sería conveniente realizar algún análisis estadístico a fondo, pero eso ya es un tema ajeno a este trabajo de investigación.

Fase 8. Monitoreo y medición de los resultados

Como el tipo de estudio del caso práctico es con fines exploratorios, las actividades de medición y monitoreo de los resultados constó en la verificación del cumplimiento de los criterios de aceptación.

Capítulo 7. Análisis de resultados

El análisis de los resultados obtenido en este trabajo de investigación comienza desde la revisión de la literatura para la elaboración de investigaciones conceptuales de (Mora, 2004) , (Mora et al., 2008) y (Hevner et al., 2004) donde comentan la importancia de realizar un conjunto de actividades que al aplicarlas en este trabajo permitieron encontrar una solución al problema que trata esta investigación el cual es proponer un método informático para trabajar con fuentes de Big Data y generar estadísticas de acuerdo a las necesidades del INEGI.

Posteriormente se revisó el estado del arte de metodologías de análisis de Big Data y procesos de ciencia de datos para detectar las similitudes y áreas de oportunidad que tenía el proceso que se estaba utilizando en ese momento para que a partir de eso se diseñara la versión inicial del artefacto que se propone en este documento, la cual se fue refinando poco a poco mediante un proceso iterativo en el que se iban detectando nuevos elementos de interés que lo fueron llevando desde lo más abstracto hasta llegar a un grado de detalle que permitió hacerlo fácil de entender y fácil de aplicar para resolver proyectos de análisis de Big Data.

De acuerdo con (Mora, 2004), se realizaron pruebas de concepto a través de dos casos de uso para darle validez al instrumento generado, uno enfocado a realizar análisis de movilidad humana y el otro en analizar el impacto de eventos de la vida real con los datos extraídos de la red social de Twitter, ambos en la República Mexicana. Los casos prácticos ayudaron a determinar la utilidad del método dando la confianza necesaria para utilizarlo extrayendo información de interés del comportamiento de la sociedad mediante las conversaciones que tienen en la red social. Para el caso de movilidad se pudo apreciar que es posible detectar patrones de comportamiento de movilidad generalizado a través de los metadatos de la posición geográfica y tiempo registrados en los mensajes originados en Twitter para obtener la frecuencia de movimientos por día, hora y mes a nivel Municipal, también se extrajeron los movimientos realizados en las tres principales zonas metropolitanas del país categorizando estos últimos en movimientos dentro de Municipios del mismo Estado, movimientos con origen en los Municipios del Estado en cuestión hacia otros Municipios de cualquier otro Estado y movimientos con destino en los Municipios del Estado en hacia otros Municipios de cualquier otro Estado de la República Mexicana para posteriormente poder compararlos con algunas estadísticas oficiales. Por otro lado, en el estudio de impacto de eventos se pudo medir y apreciar cómo afecta un evento determinado en la sociedad que participa en las conversaciones registradas en Twitter a través de los parámetros de frecuencia de menciones

de determinadas palabras relacionadas con el evento y longitud en el tiempo que se estuvieron mencionando los cuales fueron comparados contra lo que consulta la gente en el motor de búsquedas de google mediante la herramienta de Google Trends llegando a la conclusión que ambas herramientas presentan un comportamiento muy parecido. También se crearon nubes de palabras para poder apreciar de manera más rápida que palabras eran las que más veces se repitieron en los tweets que tenían que ver con los eventos estudiados.

La aplicación del marco de trabajo de NIMSAD ayudó a evaluar el instrumento generado permitiendo verificar que el proceso de diseño del método ha sido correcto y la comprobación a través de su aplicación en las situaciones presentadas en los casos prácticos ha permitido establecer que resulta una solución práctica.

Conclusiones

Los métodos de investigación conceptual revisados en esta investigación han probado ser de gran utilidad para llevar a cabo un trabajo en el que se ha creado un instrumento complejo que consistió en proponer un método para analizar los datos de Twitter que son considerados como una fuente más de Big Data para generar información actualizada. Para lograr esto se identificó un conjunto de elementos existentes en el universo de Big Data entre los que se encontraron diferentes técnicas, algoritmos, tecnologías y herramientas que al integrarlas en un mismo proyecto permitieron realizar las actividades de recolección, carga, procesamiento y análisis de la porción de tweets que ofrece la red social de manera gratuita. La forma en la que interactúan todos estos elementos se puede apreciar de manera más clara en las arquitecturas presentadas en los casos prácticos donde se muestran la relación directa que hay entre ellos mediante la identificación del flujo de entradas y salidas. Del mismo modo, otro de los elementos que fueron identificados son los roles que pueden tener las personas que trabajan en la solución de estos proyectos ayudando a darle mayor claridad y facilidad de uso al método propuesto.

Una vez generado el instrumento se probó su utilidad con el uso de pruebas de concepto a través de dos casos prácticos donde se generó información de movilidad humana y análisis de impacto de eventos en la República Mexicana, lo que ayudó a determinar que el método generó el conocimiento esperado dando la confianza necesaria para utilizarlo en el futuro. Del mismo modo, el uso del marco de trabajo NIMSAD permitió verificar que el proceso de diseño del método fue correcto gracias a la supervisión realizada en los mismos casos prácticos por los usuarios expertos que hay dentro del INEGI.

Del análisis realizado a los resultados de los casos prácticos se puede concluir que se han cumplido los objetivos específicos de este trabajo donde es posible demostrar que el instrumento conceptual desarrollado es una solución útil al problema planteado en esta investigación. Para el caso de movilidad se pudo apreciar que los metadatos de las conversaciones originadas en Twitter permitió detectar patrones de movimiento generalizados utilizando los parámetros de posición geográfica, tiempo y frecuencia de los usuarios que publicaban en el mismo Estado y Municipio de origen; mismo Estado y Municipio destino (y viceversa, visto gráficamente en los mapas de manera lineal); mismo Estado y diferente Municipio de origen, y el mismo Estado y Municipio destino (y viceversa, viéndolo gráficamente en el mapa a manera de estrella). Al compararlos con algunas estadísticas oficiales se

encontró entre otras cosas que los movimientos se repiten con mayor frecuencia en las principales zonas metropolitanas del país y algunos lugares turísticos. También se obtuvo el día y la hora en la que se registran estos movimientos.

Por otro lado, con el estudio de impacto de eventos se pudo medir y apreciar cómo afecta un evento determinado a la sociedad que participa en las conversaciones registradas en Twitter a través de los parámetros de frecuencia de menciones y longitud en el tiempo los cuales fueron comparados contra lo que consulta la gente en el motor de búsquedas de Google mediante la herramienta de Google Trends teniendo un comportamiento parecido en determinados momentos. También se encontró que la variabilidad del tiempo de vida de un evento depende de la naturaleza del mismo presentando altas y bajas de menciones en las que el parecer de los usuarios pueden presentar el mismo patrón de opinión o uno completamente distinto. Si bien no fue posible analizar la totalidad de mensajes publicados en la red social por las características propias de la investigación se encontró que con la cantidad de tweets disponibles en el flujo de datos de Twitter de manera gratuita es posible ver la magnitud e importancia que tiene un evento determinado en la comunidad de usuarios.

Los desarrollos de los casos prácticos muestran un ejemplo de la variedad de soluciones que se pueden resolver aplicando el método los cuales sirvieron para evaluar la aplicabilidad del método permitiendo realizarlos de una manera clara y concisa sin tener que estar utilizando un proceso constante de prueba y error como se realizaba anteriormente para resolver el problema (salvo en la fase de procesamiento o preparación de los datos que es la que toma entre el cincuenta y el setenta por ciento del tiempo de desarrollo de todo el proyecto (Shearer, 2000). Al igual que el método, los casos prácticos se realizaron bajo la supervisión de los usuarios expertos que hay dentro del INEGI.

Aportaciones

La principal aportación se hace a través de la creación del método que cubre varias necesidades presentadas en el INEGI empezando con establecer una serie de pasos que de manera controlada conduzcan a la generación de estadísticas utilizando el red social de Twitter, misma que puede ser aplicada en cualquier organización gracias a su fácil acceso y sin necesidad de contar con una robusta infraestructura tecnológica. Otra de las aportaciones es que el método propuesto y la demostración de su utilidad con los casos prácticos proveen un

panorama general a aquellas organizaciones que no tienen experiencia y que deseen explorar el mundo de Big Data presentando los elementos, las actividades, los roles que se necesitan y las ventajas que se pueden obtener con el desarrollo de este tipo de proyectos. Aunado a lo anterior, la identificación de los roles de los involucrados junto con las actividades que se presentan en el método pueden utilizarse como base para la formación de equipos de trabajo dedicados a realizar análisis con fuentes de Big Data.

Retos y limitaciones

Al emprender el desarrollo de este trabajo se presentaron una serie de retos a los que se tuvo que hacer frente para desarrollar esta investigación. En primer lugar y tal vez sea la misma situación en la que se encuentran muchas organizaciones, fue que el conocimiento con el que se contaba de todo lo relacionado con Big Data era muy básico sin tener una idea de la cantidad de aspectos que son importantes a considerar en la solución de este tipo de proyectos. Posteriormente a la revisión de la literatura se desarrolló la versión inicial del método, sin embargo al momento de probarlo en los casos prácticos surgió otro problema que fue el desconocimiento en la práctica de las tecnologías, técnicas y herramientas que habría que utilizar.

Una de las limitaciones que se encontraron en el camino fue al momento de definir los eventos que se iban a estudiar en el caso práctico del análisis de impacto de eventos reales, ya que el servicio de la recolección de los tweets estaba intermitente por lo que se tuvo que buscar eventos donde esta situación se presentara lo menos posible. Esto fue causado principalmente por carácter propio de los estudios (por parte del INEGI), que tiene fines exploratorios, por lo que no se contaba con la infraestructura necesaria para responder ante estas eventualidades y sobretodo estar funcionando al cien por ciento durante el tiempo que lleva el proyecto en desarrollo. Para resolver este problema se decidió implementar una estrategia para recolectar de manera redundante los tweets en un servidor alojado en la nube.

Consideraciones a futuro

El presente trabajo fue realizado sobre una serie de características organizacionales y tecnológicas especificas por lo que sería conveniente realizar otros estudios con el fin de

analizar el comportamiento y los resultados de aplicar el método en otros entornos que derivan en nuevos retos y restricciones causados por el constante cambio en las características antes mencionadas. Del mismo modo se hace la recomendación en probar la validez del método utilizando otro tipo de fuentes de datos, incluso integrando varias de éstas y de distintos tipos de datos estructurados y no estructurados para probar la valides Motivado por esto es necesario tener en consideración en un futuro extender las fases del método incorporando este tipo de aspectos mediante ajustes y anotaciones que permitan seguir utilizándolo, robustecerlo y sacarle el mayor provecho posible conforme se vayan ocupando.

Otro buen ejercicio que se podría hacer a corto plazo es comparar los resultados obtenidos en este trabajo de investigación en el caso práctico del análisis de movilidad para la zona metropolitana de la Ciudad de México contra la encuesta origen destino (EOD) que realizará el INEGI en los hogares de la Ciudad de México 2016 (EOD 2016) entre las fechas del 25 de Abril al 03 de Junio de 2016.

Glosario

Servicio WMS. Un servicio web de mapas (WMS por las siglas en ingles de Web Map Service) es un estándar para publicar cartografía en Internet cuyas especificaciones están recibidas en el Open Geoespatial Consortium (OGC), este servicio permite generar mapas de forma dinámica a partir de coordenadas geográficas en un formato de imagen como PNG, GIF o JPEG, facilitando con ello la construcción de mapas personalizados a partir de datos tomados de distintas fuentes.

Un servicio WMS se utiliza para consultar información cartográfica vía internet. Su consulta puede realizarse a través de Sistemas de Información Geográfica (SIG) en equipos de escritorio o para la construcción de aplicaciones hibridas en WEB (Mashups).

Fuente (http://www.inegi.org.mx/inegi/contenidos/serviciosweb/infogeografica.aspx)

Marco Geoestadístico Nacional. El Marco Geoestadístico es un sistema único y de carácter nacional, diseñado y creado por el INEGI en 1978, para referenciar correctamente la información estadística de los censos y encuestas con los lugares geográficos correspondientes. Proporciona la ubicación de las Localidades, Municipios y Estados del país, utilizando coordenadas geográficas.

Divide al territorio nacional en áreas con límites identificables en campo, denominadas Áreas geoestadísticas, con tres niveles de desagregación: Estatal (AGEE), Municipal (AGEM) y Básica (AGEB), ésta puede ser urbana o rural, dependiendo de las diferencias de densidad de población y uso del suelo.

Las áreas geoestadísticas cuentan con una clave única en el territorio nacional, lo que permite relacionar los datos estadísticos específicos del lugar.

La información del Marco Geoestadístico constituye un auxiliar en la delimitación entre Estados y Municipios, sobre todo, en los lugares en que los límites político administrativos se encuentran indefinidos.

Fuente: (http://www.inegi.org.mx/geo/contenidos/geoestadistica/m_geoestadistico.aspx)

CSV. Los archivos CSV (del inglés comma-separated values) son un tipo de documento en formato abierto en las que los valores se separan por comas y las filas por saltos de línea.

Modelo de negocio Bussiness to Customer (B2C). Se refiere a la estrategia que desarrollan las empresas comerciales para llegar directamente al cliente o consumidor final.

Modelo de negocio Consumer to Consumer (C2C). Modelo de negocio utilizado en comercio electrónico para definir una estrategia de cliente a cliente

Referencias Bibliográficas

Abbasi, A., Adjeroh, D., Dredze, M., Paul, M. J., Zahedi, F. M., Zhao, H., … Ross, A. (2014). Social Media Analytics for Smart Health. *IEEE Intelligent Systems*, *29*(2), 60–80. http://doi.org/10.1109/MIS.2014.29

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. En *Proceedings of the Workshop on Languages in Social Media* (pp. 30–38). Stroudsburg, PA, USA: Association for Computational Linguistics. Recuperado a partir de http://dl.acm.org/citation.cfm?id=2021109.2021114

Apache Hadoop 2.7.1 – HDFS Architecture. (2015, junio 29). Recuperado el 16 de noviembre de 2015, a partir de http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

Azevedo, T. S., Bezerra, R. L., Campos, C. A. V., & de Moraes, L. F. M. (2009). An Analysis of Human Mobility Using Real Traces. En *IEEE Wireless Communications and Networking Conference, 2009. WCNC 2009* (pp. 1–6). http://doi.org/10.1109/WCNC.2009.4917569

Azmandian, M., Singh, K., Gelsey, B., Chang, Y.-H., & Maheswaran, R. (2013). Following Human Mobility Using Tweets. En L. Cao, Y. Zeng, A. L. Symeonidis, V. I. Gorodetsky, P. S. Yu, & M. P. Singh (Eds.), *Agents and Data Mining Interaction* (pp. 139–149). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-642-36288-0_13

Bari Anasse, Chaouchi Mohamed, & Jung Tommy. (2014). *Predictive Analytics For Dummies*.

Beyer, M. A., & Laney, D. (2014). The Importance of'Big Data': A Definition - Google Académico. Recuperado el 12 de noviembre de 2015, a partir de https://scholar.google.com.mx/scholar?q=The+Importance+of%27Big+Data%27%3A+A+Definition&btnG=&hl=es&as_sdt=0%2C5

Buelens, B., Daas, P., & van den Brakel, J. (2012). Data Mining for Official Statistics: Challenges and Opportunities. En *2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW)* (pp. 915–915). http://doi.org/10.1109/ICDMW.2012.37

Cambria Erick, Schuller Björn, Xia Yunqing, & Havasi Catherine. (2013). New Avenues in Opinion Mining and Sentiment Analysis. Recuperado a partir de https://scholar.google.es/scholar?q=big+data+%22sentiment+analysis%22&btnG=&hl=es&as_sdt=0%2C5

Cardenas, A. A., Manadhata, P. K., & Rajan, S. P. (2013). Big Data Analytics for Security. *IEEE Security Privacy*, *11*(6), 74–76. http://doi.org/10.1109/MSP.2013.138

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *49*(10), 1557–1564. http://doi.org/10.1086/630200

Censos Bolivia. (2012). Recuperado el 12 de abril de 2015, a partir de http://www.ine.gob.bo:8081/censo2012/quees.aspx

Chandarana, P., and, & Vijayalakshmi. (2014). Big Data analytics frameworks. En *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)* (pp. 430–434). http://doi.org/10.1109/CSCITA.2014.6839299

Childs, H., Geveci, B., Schroeder, W., Meredith, J., Moreland, K., Sewell, C., … Bethel, E. W. (2013). Research Challenges for Visualization Software. *Computer*, *46*(5), 34–42. http://doi.org/10.1109/MC.2013.179

CIO. (2012, febrero 24). Big Data Causes Concern and Big Confusion. Recuperado el 22 de noviembre de 2014, a partir de http://www.cio.com/article/2399015/big-data/big-data-causes-concern-and-big-confusion.html

Congosto, M. L., Deltell Escolar, L., Claes, F., & Osteso, J. M. (2013). Análisis de la audiencia social por medio de Twitter. Caso de estudio: los premios Goya 2013. *Revista ICONO14. Revista científica de Comunicación y Tecnologías emergentes*, *11*(2), 53. http://doi.org/10.7195/ri14.v11i2.577

Consejo Nacional de Población (Mexico), & Instituto Nacional de Estadística, Geografía e Informática (Mexico) (Eds.). (2004). *Delimitación de las zonas metropolitanas de México* (1. ed). México, D.F. : Aguascalientes, Ags: Secretaría de Desarrollo Social : Consejo Nacional de Población ; Instituto Nacional de Estadística, Geografía e Informática.

CTS México, & ITDP. (2011). *10 Estrategias de Movilidad para un Estado de México Competitivo, Seguro y Sustentable 10 Estrategias de Movilidad para un Estado de México Competitivo, Seguro y Sustentable: Hacia una Red Integrada de Transporte en la Zona Metropolitana del Valle de México* (p. 89). Mexico DF. Recuperado a partir de http://mexico.itdp.org/wp-content/uploads/EDOMEX_VF.pdf

Cuesta, H. (2013). *Practical Data Analysis*. Packt Publishing Ltd.

Data-Pop Alliance(Harvard Humanitarian Initiative, MIT Media Lab y Overseas Development Institute). (2016). Oportunidades y requerimientos para aprovechar el uso de Big Data para las estadísticas oficiales y los Objetivos de Desarrollo Sostenible en América Latina, 83.

Davenport, T. H., & Patil, D. J. (2012, octubre). Data Scientist: The Sexiest Job of the 21st Century. Recuperado el 9 de noviembre de 2015, a partir de https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM*, *51*(1), 107–113. http://doi.org/10.1145/1327452.1327492

Fan Wei, & Bifet Albert. (2013). Mining Big Data: Current Status, and Forecast to the Future, *14*(2).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fragoso, P. M. L., Salinas, J. C. S., Leyva, G., Bustos, A., Muñoz, J., Fraustro, S., & Coronado, A. (2014). USO PRODUCTIVO DE BIG DATA Y REDES SOCIALES EN EL SECTOR TURISMO. Recuperado a partir de http://datatur.sectur.gob.mx/Documentos%20Publicaciones/2014_1_DocInvs.pdf

Gabrielli Lorenzo, Rinzivillo Salvatore, & Ronzano Francesco. (2014). From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns. En Villatoro Daniel (Ed.), *Citizen in Sensor Networks* (pp. 26–35). Springer International Publishing. Recuperado a partir de http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-319-04178-0_3

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. http://doi.org/10.1038/nature07634

Guo, P. (2013, octubre). Data Science Workflow: Overview and Challenges. Recuperado el 5 de enero de 2016, a partir de http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

Gupta, A., & Kumaraguru, P. (2012). Credibility Ranking of Tweets During High Impact Events. En *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* (p. 2:2–2:8). New York, NY, USA: ACM. http://doi.org/10.1145/2185354.2185356

Hao Wang, Can, D., Abe Kazemzadeh, François Bar, & Shrikanth Narayanan. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. En *Proceedings of the ACL 2012 System Demonstrations* (pp. 115–120). Stroudsburg, PA, USA: Association for Computational Linguistics. Recuperado a partir de http://dl.acm.org/citation.cfm?id=2390470.2390490

Harris, H., Murphy, S., & Vaisman, M. (2013). *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. O'Reilly Media, Inc.

Hawelka Bartosz, Sitko Izabela, Beinat Euro, Sobolevsky Stanislav, Kazakopoulos Pavlos, & Ratti Carlo. (2013). Geo-located Twitter as the proxy for global mobility patterns. Recuperado a partir de http://arxiv.org/ftp/arxiv/papers/1311/1311.0680.pdf

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. http://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hernández Sampieri, R., Fernández Collado, C., & Pilar Baptista, L. (2006). *Metodologia de la investigación* (Cuarta).

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information System Research, *28*(1), 75–105.

Hodeghatta, U. R. (2013). Sentiment Analysis of Hollywood Movies on Twitter. En *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1401–1404). New York, NY, USA: ACM. http://doi.org/10.1145/2492517.2500290

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data For Dummies* (1st ed.). For Dummies.

IBGE :: Instituto Brasileiro de Geografia e Estatística. (2015). Recuperado el 12 de abril de 2015, a partir de http://www.ibge.gov.br/espanhol/estatistica/calendario_estudos2015.shtm

IBM. (2012, junio 18). ¿Qué es Big Data? [CT316]. Recuperado el 26 de noviembre de 2014, a partir de http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

IBM. (2013, octubre 15). Big data architecture and patterns, Part 3: Understanding the architectural layers of a big data solution [CT316]. Recuperado el 9 de junio de 2015, a partir de http://www.ibm.com/developerworks/library/bd-archpatterns3/

INEGI. (1991). *DATOS BASICOS DE LA GEOGRAFIA DE MEXICO* (segunda edición). México: INEGI. Recuperado a partir de

http://www.inegi.org.mx/inegi/SPC/doc/INTERNET/DATOS%20BASICOS%20DE%20LA%20GEOGRAFIA%20DE%20MEXICO.pdf

INEGI. (2010). Resumen. Ciudad de México. Recuperado el 20 de mayo de 2016, a partir de http://www.cuentame.inegi.org.mx/monografias/informacion/df/default.aspx?tema=me&e=09

INEGI. (2010). Resumen. Jalisco. Recuperado el 20 de mayo de 2016, a partir de http://www.cuentame.inegi.org.mx/monografias/informacion/jal/default.aspx?tema=me&e=14

INEGI. (2010). Resumen. Nuevo León. Recuperado el 20 de mayo de 2016, a partir de http://www.cuentame.inegi.org.mx/monografias/informacion/nl/default.aspx?tema=me&e=19

INEGI. (2014, noviembre). Acerca del INEGI. Recuperado el 29 de noviembre de 2014, a partir de http://www.inegi.org.mx/inegi/acercade/default.aspx

Instituto Nacional de Estadística y Geografía. (2014a). *Anuario estadístico y geográfico de Jalisco 2014*. México. Recuperado a partir de http://www.datatur.sectur.gob.mx/ITxEF_Docs/JAL_ANUARIO_PDF.pdf

Instituto Nacional de Estadística y Geografía. (2014b). *Anuario estadístico y geográfico del Distrito Federal. 2014.* México. Recuperado a partir de http://www.datatur.sectur.gob.mx/ITxEF_Docs/DF_ANUARIO_PDF.pdf

Izquierdo, J. M. C. (2008). Estudios sobre movilidad cotidiana en México. *Scripta Nova: Revista electrónica de geografía y ciencias sociales*, (12), 273-.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and Its Technical Challenges. *Commun. ACM*, *57*(7), 86–94. http://doi.org/10.1145/2611567

J. Paul Michael, and, & Dredze, Mark. (2011). You Are What You Tweet: Analyzing Twitter for Public Health.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. En *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (pp. 56–65). New York, NY, USA: ACM. http://doi.org/10.1145/1348549.1348556

Jayaratna, N., & Armstrong, H. (2005). Applying the NIMSAD Framework to Evaluating IA Education Projects.

Krishnan, K. (2013). Chapter 13 - Big Data Analytics, Visualization, and Data Scientists. En K. Krishnan (Ed.), *Data Warehousing in the Age of Big Data* (pp. 251–255). Boston: Morgan Kaufmann. Recuperado a partir de http://www.sciencedirect.com/science/article/pii/B9780124058910000131

Lindenboim, J. (2011). Las estadísticas oficiales en Argentina ¿Herramientas u obstáculos para las ciencias sociales?: Useful tools or obstacles for the social sciences? *Trabajo y sociedad*, (16), 19–38.

Liu, L., Zhang, H., Li, J., Wang, R., Yu, L., Yu, J., & Li, P. (2009). Building a Community of Data Scientists: An Explorative Analysis. *Data Science Journal*, *8*(0). http://doi.org/10.2481/dsj.008-004

López, J. M. (2008). METODOLOGÍA PARA DISEÑAR MODELOS ARQUITECTÓNICOS DE REFERENCIA QUE INTEGRAN SISTEMAS HEREDADOS. http://doi.org/10.13140/2.1.1429.2801

McKinsey & Company. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Recuperado a partir de http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Microsoft. (2015, junio). Understanding Microsoft big data solutions. Recuperado el 11 de junio de 2015, a partir de https://msdn.microsoft.com/en-us/library/dn749804.aspx

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). Big Data Analytics Methodology. En *Big Data Imperatives* (pp. 197–220). Apress. Recuperado a partir de http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-1-4302-4873-6_7

Mora, M. (2004). *Descripción del Método de Investigación Conceptual, versión 2* (pp. 1–17). Aguascalientes, México: UAA.

Mora, M., Gelman, O., Paradice, D., & Cervantes, F. (2008). The case for Conceptual Research in Information Systems (pp. 1–10). Niagara Falls, Ontario, Canada.

Mousannif, H., Sabah, H., Douiji, Y., & Sayad, Y. O. (2014). From Big Data to Big Projects: A Step-by-Step Roadmap. En *2014 International Conference on Future Internet of Things and Cloud (FiCloud)* (pp. 373–378). http://doi.org/10.1109/FiCloud.2014.66

Naaman, M. (2011). Geographic Information from Georeferenced Social Media Data. *SIGSPATIAL Special*, *3*(2), 54–61. http://doi.org/10.1145/2047296.2047308

Nabel, L. C. T. (2009). Ciberprotestas y consecuencias políticas:: Reflexiones sobre el caso de internet necesario en México. *Razón y palabra*, (70), 49.

Nabel, L. C. T. (2014). El poder de las redes sociales: la “mano invisible” del framing noticioso. El caso de #LadyProfeco. *Revista ICONO14. Revista científica de Comunicación y Tecnologías emergentes*, *12*(2), 318. http://doi.org/10.7195/ri14.v12i2.625

Olavsrud Thor. (2014, febrero). 2014, ¿Año del Big Data? Recuperado el 9 de junio de 2015, a partir de http://www.computerworld.es/sociedad-de-la-informacion/es-2014-el-ano-de-la-arquitectura-big-data

ONU. (2013). Fundamental Principles of Official Statistics.

Oracle. (2015). *An Enterprise Architect’s Guide to Big Data Reference Architecture Overview*.

Patel, A. B., Birla, M., & Nair, U. (2012). Addressing big data problem using Hadoop and Map Reduce. En *2012 Nirma University International Conference on Engineering (NUiCONE)* (pp. 1–5). http://doi.org/10.1109/NUICONE.2012.6493198

Patil, D. J. (2011). *Building data science teams the skills, tools and perspectives behind great data science groups*. Sebastopol, CA: O'Reilly.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 1. http://doi.org/10.1186/2047-2501-2-3

Saltz, J. S. (2015). The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness. En *2015 IEEE International Conference on Big Data*. Recuperado a partir de http://www.midp.info/uploads/1/0/6/5/10650753/position_paper_-_final.pdf

Saltz, J. S., & Shamshurin, I. (2015). Exploring the process of doing data science via an ethnographic study of a media advertising company. En *Big Data (Big Data), 2015 IEEE International Conference on* (pp. 2098–2105). IEEE. Recuperado a partir de http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7363992

Sánchez J. M. (2013, diciembre 9). Big Data: presente y futuro para las empresas. Recuperado el 12 de abril de 2015, a partir de http://www.abc.es/tecnologia/informatica-soluciones/20131207/abci-data-analisis-201312051430.html

Sawant, N., & Shah, H. (2013). *Big Data Application Architecture Q&A: A Problem - Solution Approach* (1 edition). New York, NY: Apress.

Scannapieco Monica, Virgillito Antonino, & Zardetto Diego. (2013). Placing Big Data in Official Statistics: A Big Challenge.

Shearer, C. (2000). The CRISP-DM Model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Sheikh, N. M. (2013). *Implementing analytics: a blueprint for design, development, and adoption*. Amsterdam: Elsevier.

Singh, S., Pandey, S., Shankar, R., & Dumka, A. (2015). Application of Big Data Analytics to optimize the operations in the upstream petroleum industry. En *2015 2nd International*

Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1074–1079).

Tascón, M. (2013). Introducción: Big Data. Pasado, presente y futuro. *Telos: Cuadernos de comunicación e innovación*, (95), 47–50.

TechAmerica Foundation's Federal Big Data Commission. (2014). Demystifying Big Data: A Practical Guide To Transforming The Business of Government.

Tendencias de Google - Listas de búsquedas más populares sobre Todas las categorías. (2015). Recuperado el 16 de febrero de 2016, a partir de https://www.google.com.mx/trends/topcharts#vm=cat&geo=MX&date=2015&cid

Twitter. (2015a). Story of a Tweet. Recuperado el 7 de abril de 2015, a partir de https://about.twitter.com/es/what-is-twitter/story-of-a-tweet

Twitter. (2015b, marzo). About Twitter, Inc. | About. Recuperado el 12 de marzo de 2015, a partir de https://about.twitter.com/company

UN. (2015). Big Data for Official Statistics. Recuperado el 24 de mayo de 2016, a partir de http://unstats.un.org/unsd/bigdata/

Wilson, L. A. (2014). Big Data sessions bring variety of opinions. *MRS Bulletin*, *39*(4), 376–376. http://doi.org/10.1557/mrs.2014.78

Yang, Q., Hu, X., Cheng, Z., Miao, K., & Zheng, X. (2014a). Based Big Data Analysis of Fraud Detection for Online Transaction Orders. En V. C. M. Leung, R. X. Lai, M. Chen, & J. Wan (Eds.), *Cloud Computing* (pp. 98–106). Springer International Publishing. Recuperado a partir de http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-319-16050-4_9

Yang, Q., Hu, X., Cheng, Z., Miao, K., & Zheng, X. (2014b). Based Big Data Analysis of Fraud Detection for Online Transaction Orders. En V. C. M. Leung, R. X. Lai, M. Chen, & J. Wan (Eds.), *Cloud Computing* (pp. 98–106). Springer International Publishing.

Recuperado a partir de http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-319-16050-4_9

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring International and Internal Migration Patterns from Twitter Data. En *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion* (pp. 439–444). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. http://doi.org/10.1145/2567948.2576930

Anexos

Anexo 1. Guía para crear cuenta de Twitter para descargar tweets

- Contar con una cuenta de usuario de Twitter. Para acceder a los datos desde el API de Twitter es necesario contar con cuatro token que son obtenidos al crear una aplicación a través de una cuenta de usuario de Twitter existente.
- Adquirir los token para autenticarse en el API de Twitter. Para poder acceder al streaming de Twitter, es necesario contar con una cuenta de usuario y contraseña válida para después visitar el sitio de desarrolladores https://dev.twitter.com/apps y poder crear una aplicación nueva. Para poder registrar esta aplicación se solicitan los siguientes datos que se muestran en la Figura 48:
    1. Nombre: El nombre de la aplicación, puede ser cualquier nombre que usted desee, sin embargo, no se puede utilizar la palabra Twitter en el nombre.
    2. Descripción: Breve descripción del funcionamiento de la aplicación (puede ser cualquier cosa que le guste).
    3. Website: Puede ser tu blog personal o sitio web.
    4. Callback URL: Puede dejarse en blanco.

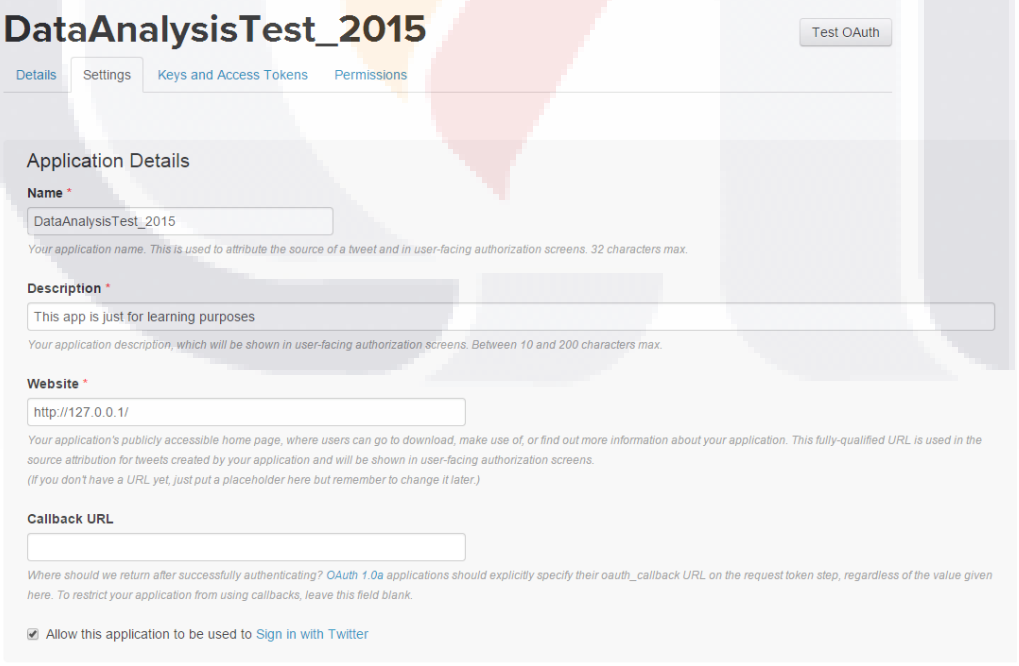


Figura 48. Creación de aplicación de Twitter

La aplicación ha sido creada con las siguientes características, Figura 49.

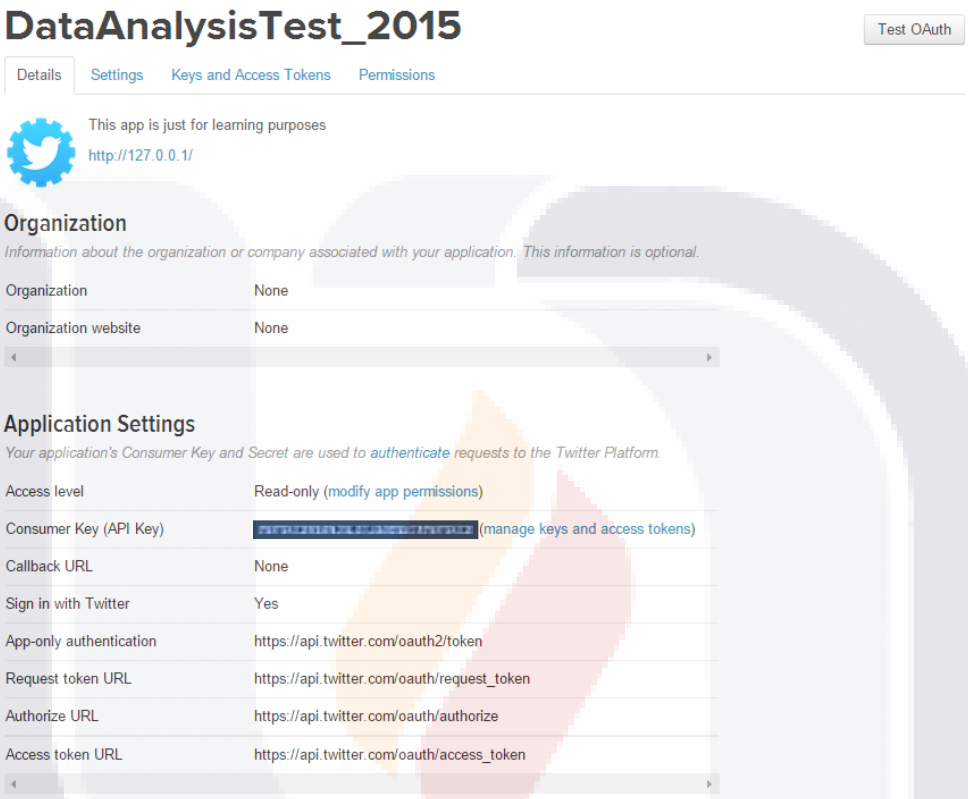


Figura 49. Características de la aplicación de Twitter

Por último, es necesario crear los token de acceso, para ello tendrá que hacer clic en Create my access token que se muestra en la Figura 50 siguiente.

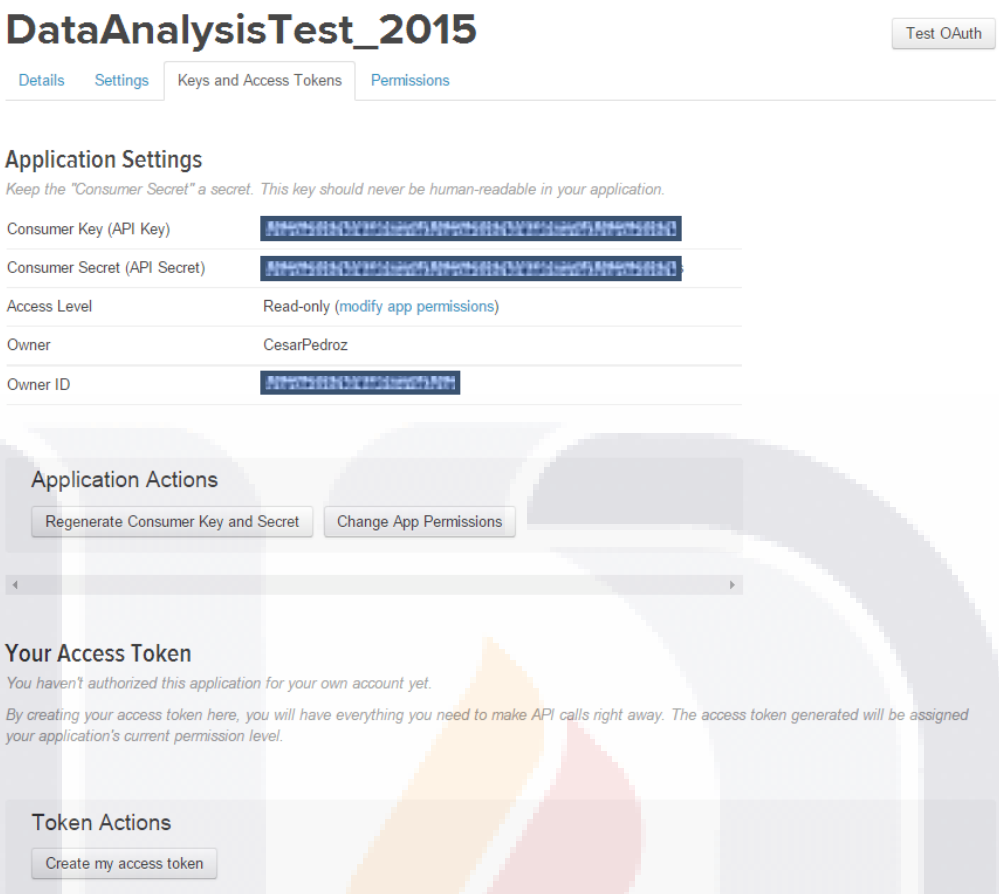


Figura 50. Pantalla para la creación de token

En la siguiente captura de pantalla se pueden ver los cuatro token para la autenticación; Consumer key, Consumer secret, Access token, and Access token secret. Twitter notificará si los token fueron creados de manera satisfactoria.

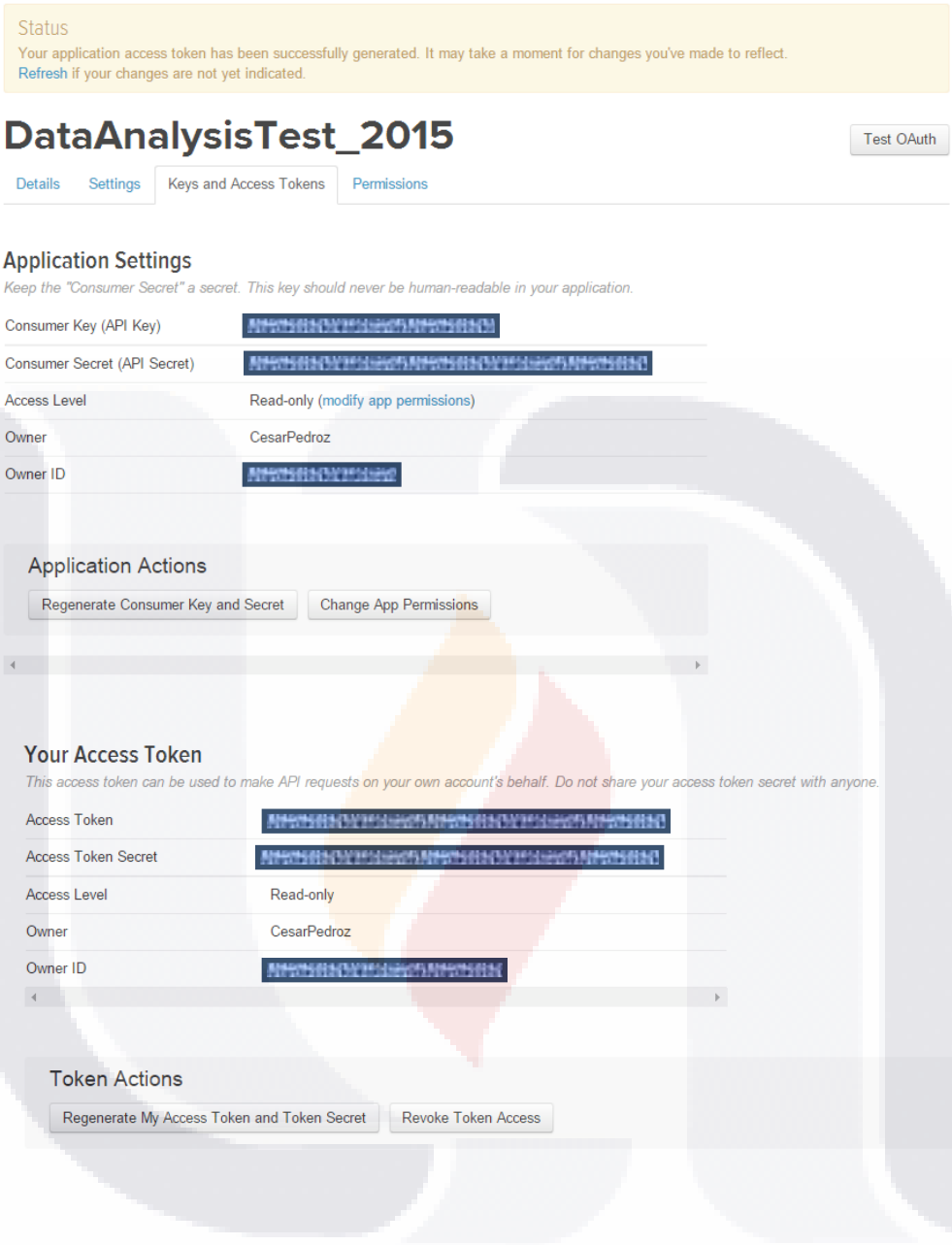


Figura 51. Pantalla con los token de acceso al API de Twitter

Anexo 2. Guía para la descarga e instalación de las herramientas utilizadas en los casos prácticos en el sistema operativo Windows 8.1

En esta sección se detallaran cada uno de los pasos que se siguieron para el desarrollo del método informático. Se cubren aspectos desde la instalación del software necesario hasta los procesos de carga transformación y limpieza de los datos.

Descargar e instalar Elasticsearch

Los pasos para tener funcionando Elasticsearch son relativamente sencillos:

1. Primero que nada es necesario descargar Elasticsearch accediendo directamente al sitio web https://www.elastic.co/downloads eligiendo la versión de interés (para este trabajo se utilizó la versión 1.5.1).
2. Después de esto es necesario descomprimir el archivo descargado en alguna carpeta (por ejemplo, c:\elasticsearch-1.5.1).
3. Ejecutar el archivo batch llamado elasticsearch.bat ubicado en la carpeta bin de la carpeta previamente descomprimida (c:\elasticsearch-1.5.1\bin\ elasticsearch.bat).
4. Para ver funcionando Elasticsearch se puede acceder desde un navegador web mediante la url *http://localhost:9200/*

Descargar e instalar Logstash

Los pasos para tener funcionando Logstash son relativamente sencillos, a continuación se describen:

1. Descargar Logstash accediendo al sitio web https://www.elastic.co/downloads eligiendo la versión de interés (para este trabajo se utilizó la versión 1.5.1).
2. Después de esto es necesario descomprimir el archivo descargado en alguna carpeta (por ejemplo, c:\logstash-1.5.1).
3. Preparar el archivo de configuración que lea los datos del streaming de Twitter y los almacene en Elasticsearch. Para este trabajo se utilizó el archivo Logstash-twitterLocation.conf.
4. Para empezar a descargar tweets se tiene que ejecutar Logstash desde ventana de comandos en la carpeta bin de Logstash la siguiente instrucción: c:\logstash-1.5.1\bin\logstash -f Logstash-twitterLocation.conf.

Descargar e Instalar Spark

Para instalar el modo Standalone Spark, sólo tiene que colocar una versión compilada de Spark en cada nodo del clúster. Puede obtener versiones pre-construidos de Spark con cada lanzamiento o construirlo por usted mismo.

Se puede ejecutar una aplicación en el clúster Spark desde su shell interactivo, simplemente pase la URL spark://IP: PUERTO del maestro como constructor del SparkContext como se muestra en el siguiente comando:

./bin/spark-shell --master spark: // IP: PUERTO

Para ejecutar la librería CSV de Spark es necesario tener instalado una versión de Scala y ejecutar la llamada del Shell interactivo de Spark con la siguiente instrucción (para la versión 1.11 de Scala):

$SPARK_HOME/bin/spark-shell --packages com.databricks:spark-csv_2.11:1.2.0