Big Data, análisis de datos en la nube

Verónica Idalia Rosa-Urrutia
Docente Utec
veronica.rosa@utec.edu.sv

José Guillermo Rivera-Pleitez
Docente Utec
joseguillermo.rivera@gmail.com

Recibido: 07/04/2016 - Aceptado: 05/05/2016

Resumen

Big Data en El Salvador es novedoso, por lo que el objetivo de esta investigación es elaborar una guía metodológica en la que se reflejara el uso de herramientas *Big Data* para almacenar, procesar y analizar grandes cantidades de datos, con el fin de poder obtener conclusiones que ayuden en la toma de decisiones. Para esta investigación se hizo uso de dos dataset con información sobre registro de productos alimenticios y medicamentos. Los datos fueron almacenados y procesados por medio de herramientas *Big Data*, tales como Hadoop y Hive, R para análisis estadístico, finalizando con la creación de visualizaciones en Google chart, Jqplot y D3. La investigacion se llevó a cabo durante los meses de febrero a noviembre de 2015.

Palabras clave

Big Data, visualizaciones, conjunto de datos, análisis estadístico.

Abstract

Big Data in El Salvador is novel, so that the objective of this research is to elaborate a methodological guide, in which the use of *Big Data* reflect tools to store, process and analyze large amounts of data, in order to draw conclusions that can help in making decisions. For this research, two dataset containing information on registration of foodstuffs and medicines were used. The data was stored and processed through Big Data tools such as Hadoop and Hive, R for statistical analysis, ending with the creation of visualizations in Google chart, Jqplot and D3. The research was conducted during the months of February to November 2015.

Keywords

Big Data, visualizations, dataset, statistical analysis

1 Universidad Tecnológica de El Salvador, Ingeniera en Sistemas y Computación, Máster en Docencia Universitaria, Máster en Visual Analitycs y Big Data. Docente tiempo completo y administradora de Laboratorio de Informática Utec.

2 Universidad Tecnológica de El Salvador, Ingeniero en Sistemas y Computación, Máster en Docencia Universitaria, Máster en Recursos Humanos. docente hora-clase Utec.

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

Introducción

Big Data es un término aplicado a conjuntos de datos que superan la capacidad del *software* habitual para ser capturados, gestionados y procesados en un tiempo razonable. Se considera un conjunto de datos que crecen rápidamente y que no pueden ser manipulados por las herramientas de gestión de bases de datos tradicionales (Aguilar, 2013).

En el 2001 se realizó un informe de investigación en el que el analista Doug Laney del META Group (ahora Gartner)[3] definía "el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen, la velocidad y la variedad".

"*Big Data* es desde hace unos años el término de moda dentro del mundo de la informática. Dicho de otra manera, durante 2012 y parte de 2013 el 60 % de los artículos de opinión de tecnología avanzada hablan de *Big Data* como la nueva estrategia indispensable para las empresas de cualquier sector, declarando, poco menos, que aquéllos que no se sumen a este nuevo movimiento se quedarán "obsoletos" en cuanto a la capacidad de reacción en sus decisiones, perdiendo competitividad y oportunidades de negocio contra su competencia."[4]

Es tanta la información que se genera a diario en la web a través de las redes sociales, buscadores, almacenamiento de datos en la nube, etc. Lo cual resulta abrumador y solo el hecho de saber cómo se consigue captar y analizar dicha información es sorprendente.

También se sabe que las redes sociales hoy en día, aportan mucha información relevante que los usuarios comparten libre y públicamente en la web. Todo esto es aprovechable por las empresas para detectar tendencias en el mercado y enfocar las acciones que se van a llevar a cabo, algo que ayuda a tomar mejores decisiones y a que los resultados sean beneficiosos. Las ventajas las obtendrán aquellas empresas que sepan cómo hacerlo, al utilizar herramientas que faciliten el procesado masivo de datos y en poco tiempo.

Todo lo anterior se debe a la necesidad de crear nuevas formas de comunicación y almacenar dicha información de manera constante, siendo esta de rápido crecimiento. Esta contribución a la acumulación masiva de datos se puede encontrar en diversas industrias. Las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores y operaciones; de la misma manera sucede con el sector público.

Pero no solamente los seres humanos son los que contribuyen al crecimiento enorme de información. Existe también la comunicación denominada máquina a máquina (M2M, machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante (Fragoso, 2014)

En la siguiente infografía se observa una representación de *Big Data*.

Figura 1. Infografía *Big Data*
Fuente: http://n-economia.com/notasalerta/trasformacion-digital-big-data-infografia/

Por otro lado, están los *dataset* públicos, que son archivos que se encuentran alojados en la nube de forma pública, los cuales están en distintos formatos y es allí donde también surge el problema, cuando los datos ya no son estructurados como comúnmente se ha acostumbrado a utilizarlos en las bases de datos relacionales tradicionales,

3 http://www.gartner.com/analyst/40872/Douglas-Laney
4 Airtor Moreno. Responsable de inteligencia artificial de Ibermática.

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

pues estos se encuentran en formatos tales como json, csv, dat, arff, ncol, etc. En estos casos, se hace necesario el uso de herramientas que permitan almacenar y procesar ese tipo de ficheros.

Por esto es que se considera importante una propuesta metodológica en la que se haga uso de herramientas propias de *Big Data*, para el procesamiento masivo de información, análisis de los resultados y visualización de los datos. Para ello se trabajó con dos *dataset* públicos los cuales contenían una gran cantidad de datos sobre productos alimenticios, bebidas y medicamentos de fabricantes de muchos países del mundo, permitiendo tener un control detallado de esos productos y realizar los análisis necesarios.

Lo que se pretendía con esa información es que las empresas gubernamentales y privadas conocieran los procedimientos necesarios y las herramientas que serán útiles para solventar el problema de trabajar con datos masivos y obtener resultados en menor tiempo.

Se tomó a bien hacer uso de dos *dataset* para aprovechar las herramientas y mostrar lo que se podía hacer con la información contenida.

Materiales y métodos

Como lo que se realizó al final fue una guía metodológica para el procesamiento, análisis y visualización de datos, se hizo uso de las siguientes herramientas *Big Data*:

Hadoop (White, 2012) En la actualidad, Hadoop es un proyecto de *software* libre, con licencia Apache, cuya finalidad es prestar una plataforma para la gestión de grandes cantidades de datos. Los principales componentes que constituyen Hadoop son el sistema de archivos HDFS y el motor MapReduce.

HDFS (Hadoop Distributed File System) es un sistema de archivos distribuido, inspirado en el Google File System (GFS) de Google, que permite distribuir los datos entre distintos nodos de un clúster (llamados datanodes), gestionando la distribución y la redundancia de forma transparente para el desarrollador que vaya a hacer uso de esos datos.

En la figura 2, se ejemplifica como los bloques de datos son escritos hacia HDFS. Se observa que cada bloque es almacenado tres veces y al menos un bloque se almacena en un diferente rack para lograr redundancia.

El motor MapReduce es un sistema que gestiona los mecanismos para ejecutar tareas MapReduce de forma distribuida entre los diferentes nodos del clúster Hadoop. De nuevo, la forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador (Dean & Ghemawat, 2004). En la figura 3, se ejemplifica un proceso sencillo de MapReduce.

Además, el ecosistema de Hadoop se compone de otros proyectos que, sin ser vitales para su funcionamiento, permiten realizar determinadas tareas de un modo más sencillo o más eficiente.

Existen varias distribuciones de Hadoop, tales como Hortonworks, Cloudera, MapR, pero para esta investigación se utilizó Hortonworks (Hortonworks, 2011-2016).

Dentro de Hadoop, se hizo uso de la herramienta Hive (Capriolo, Wample, & Rutherglen, 2012), el cual es un proyecto que forma parte del ecosistema Hadoop y por ello, viene incluido en muchas distribuciones de Hadoop, incluyendo la distribución Hortonworks.

El propósito de Hive es, en cierto modo, emular un sistema de bases de datos relacional encima de Hadoop.

Así, el usuario podrá crear tablas e insertar datos (o crearlas a partir de ficheros existentes en HDFS), para posteriormente consultarlas empleando un lenguaje de modelado y de consulta muy similar a Structured Query Language (SQL).

Es importante entender que esta lógica funciona bien cuando trabajamos con datos que son estructurados, puesto que el concepto de tablas en el modelo relacional estructura los datos en columnas (campos) y en filas (registros).

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

Hive es una herramienta adecuada para usuarios que estén familiarizados con las bases de datos relacionales. Permite crear tablas y hacer consultas sobre ellas empleando un lenguaje similar a SQL, si bien estas consultas se traducirán automáticamente a rutinas MapReduce.

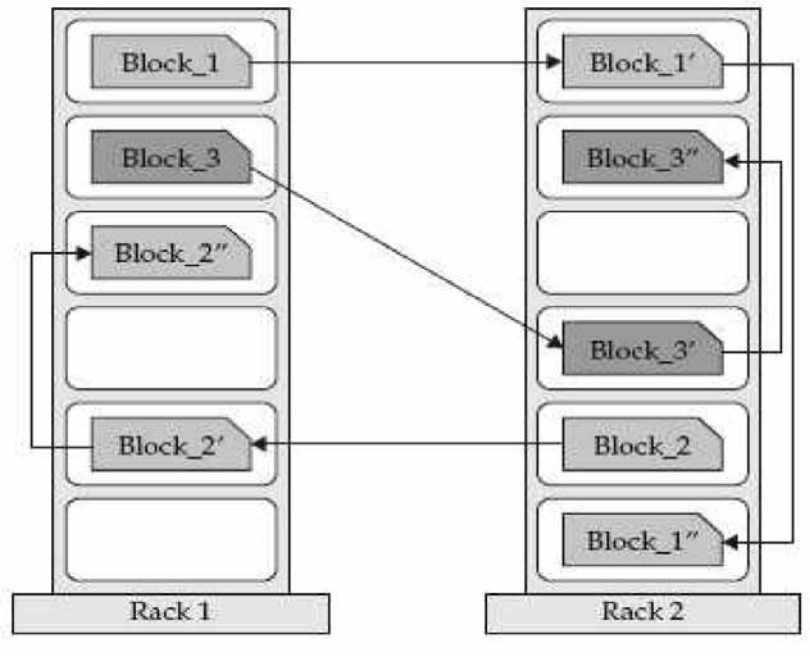

Figura 2. Ejemplo de HDFS
Fuente: Sitio web IBM.

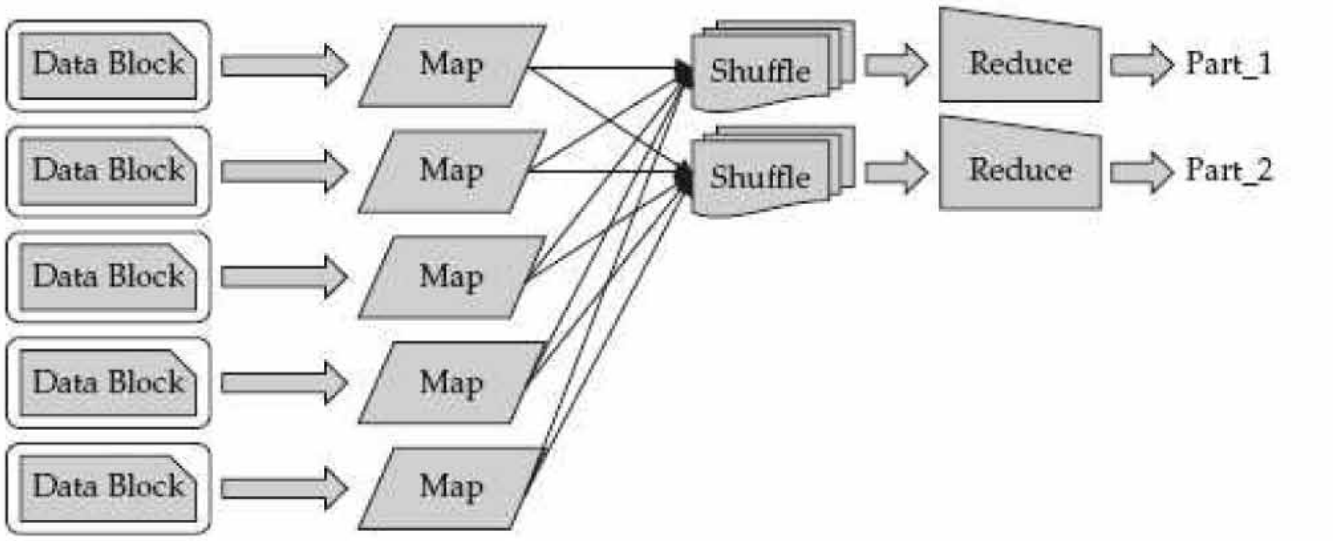

Figura 3. Ejemplo de MapReduce
Fuente: Sitio web IBM.

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

R (R, s.f.) Se puede definir R desde dos perspectivas:

- R es un entorno de *software*
- R es un lenguaje de programación

Fundamentalmente R, puede ser definido como un entorno *software* para el análisis matemático y estadístico de datos, en cierto sentido similar a herramientas tales como Microsoft Excel. A través del entorno de R, somos capaces de manipular datos (por ejemplo, cargarlos desde ficheros, editarlos, volverlos a almacenar...), realizar análisis sobre esos datos y presentar los resultados gráficamente para facilitar su interpretación.

El entorno *software* viene acompañado de un lenguaje de programación que pone a nuestra disposición las funcionalidades típicas de un lenguaje de propósito general (manejo de variables, tipos y estructuras de datos, operadores, mecanismos de control del flujo de ejecución, funciones, etc.) combinadas con librerías y herramientas específicas para facilitar el análisis de datos. Utilizando este lenguaje es relativamente sencillo implementar nuestras propias funciones y *scripts* para automatizar el procesamiento de ciertos datos.

En la práctica, estas dos perspectivas están muy relacionadas así, por ejemplo, para interactuar con el entorno de R se utilizaron expresiones escritas en el lenguaje R.

¿Por qué se utilizó R?

Actualmente existe una amplia gama de herramientas que pudiéramos pensar en utilizar a la hora de llevar a cabo análisis de datos como (por ejemplo, Microsoft Excel, S-PLUS una versión comercial del lenguaje S, SAS, SPSS de IBM, etc.) Así pues, una de las cuestiones que las empresas pudieran plantear en esta metodología es por qué elegir R como herramienta de análisis de datos.

Algunas de las razones que podría emplear a la hora de justificar la decisión incluyen lo siguiente:

- R es *software* de código libre con licencia GNU GPL (General Public License).
- Mientras que las principales herramientas de análisis son de pago (algunas con precios bastante elevados), R es completamente gratuito.
- Existen versiones para los sistemas operativos más comunes: Windows, Mac OS X y Linux.
- Posee una comunidad de usuarios amplia y muy activa, con lo que va a resultar relativamente sencillo encontrar documentación o ayuda en foros si resulta necesario.
- El entorno es fácilmente extensible, mediante el desarrollo de paquetes. Debido a esto, evoluciona rápidamente: nuevos algoritmos y técnicas de análisis se incorporan con regularidad.
- R y sus extensiones nos ofrece una gran variedad de herramientas de análisis y visualización de datos. Actualmente existen más de 5000 paquetes disponibles para ser instalados en el entorno.

Herramientas de visualización

Google Chart (Developers, 2016) Es una aplicación de Google para realizar estadísticas web, de fácil uso para desarrolladores de *software* web, usado en muchos campos como Google Analytics, se puede usar con diferentes formatos: Json, Javascript y *pluggings* que se pueden integrar con varios lenguajes de programación.

Esta herramienta permite realizar gráficos atractivos y existe una gran variedad de galerías disponibles en el sitio de Google para poder utilizarlos y adaptarlos a las necesidades de análisis de cada persona.

Jqplot (Jqplot pure javascript plotting, s.f.) Es un *framework* para el trazado de gráficos y *plugging* jQuery Javascript, jqPlot. Produce hermosas líneas, barras y gráficos circulares con muchas características:

- Numerosas opciones de estilo gráfico.
- Fecha, ejes con formato personalizable.
- Hasta 9 ejes Y.
- Texto eje girado.
- Cálculo automático de la línea de tendencia.
- La información sobre herramientas y punto de datos resaltado.
- Valores predeterminados razonables para facilitar su uso.

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

D3.js (D3js.org, 2015) "Es una biblioteca JavaScript para manipular documentos basados en los datos. D3 ayuda a llevar los datos a la vida usando HTML, SVG y CSS. El énfasis de D3 en estándares web ofrece todas las capacidades de los navegadores modernos sin atar a sí mismo a un marco exclusivo, que combina componentes de visualización de gran alcance y un enfoque impulsado por los datos a la manipulación del Document Object Model (DOM)."

Para poder hacer uso de esta herramienta, es necesario conocer de JavaScript. Por consiguiente, es recomendable aprender ese lenguaje de programación, para aprovechar todas sus ventajas de diseño.

Requisitos a considerar

Los requisitos técnicos para poder hacer uso de las herramientas tecnológicas antes mencionadas son los siguientes:

Herramienta tecnológica	Requisito de *hardware*	Requisito de *software*
Hadoop	✓ 5.ª generación del procesador Intel® Core™ i7 ✓ Memoria de 16 GB expandible a 32 GB ✓ Disco duro de 1 a 2 TB	✓ Windows 8.1 ✓ Linux de 64 bits ✓ Oracle VM VirtualBox ✓ CentOS 6.5 o más reciente ✓ Distribución Hortonworks
Programa R	No requiere de características especiales.	✓ Windows 8.1 ✓ R o R Studio
Google chart, jqplot, D3.js	No requiere de características especiales.	✓ Windows 8.1 ✓ Cualquier editor de texto: bloc de notas, sublime text o notepad ✓ Navegador web ✓ Servidor web

Hadoop es la única herramienta que requiere de muchos recursos de *hardware*, sobre todo de memoria RAM y de disco duro y es por ello que lo más recomendable es que esté instalado en un servidor, pero por las limitantes encontradas y no disponer de suficientes recursos en el equipo utilizado, se optó por crear máquinas virtuales para simular un máster y un esclavo y hacer el despliegue en un entorno similar a uno de producción.

Los demás programas utilizados son gratuitos y con lecturas de manuales en la web se pueden llegar a utilizar sin ningún problema.

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

Metodología

La investigación se realizó con el objetivo de conocer sobre el auge de *Big Data*, y como se podrían utilizar algunas herramientas para el procesamiento, análisis y visualización de los datos. Se hizo un estudio exploratorio de tipo descriptivo por medio de una encuesta en la que se seleccionó únicamente a profesionales en el área de Informática, debido a que se necesitaba una opinión bastante real en cuanto a conocimientos con respecto a las nuevas tecnologías.

En cuanto a los participantes, el proceso de recolección de datos se hizo con profesionales en el área de Informática que laboran en el departamento de San Salvador. La muestra fue aleatoria, para la cual participaron 40 personas de ambos géneros y de distintas edades, aunque esos datos no fueron relevantes para la investigación.

El instrumento utilizado fue una encuesta, en la cual se buscaba explorar los conocimientos con los que cuentan actualmente los profesionales en el área de Informática sobre *Big Data*. Dicha encuesta fue elaborada con la herramienta Google Forms, la cual contaba con 12 preguntas cerradas y a cada participante se les proporcionó el *link* para que pudieran contestar en línea, en el momento que dispusieran.

Esta encuesta ayudó a determinar la importancia de la elaboración de una propuesta metodológica en el uso de herramientas de *Big Data*.

Luego de pasar las encuestas, Google Forms también da la facilidad de enviar las respuestas en una hoja de Google Spreadsheets en el *drive* de la cuenta de correo electrónico en la cual se elaboró la encuesta, a parte que proporciona los gráficos respectivos, que son elaborados con Google Chart.

Se aprovechó esta herramienta porque es una forma de hacer uso de nuevas tecnologías en la web, por medio de las cuales se obtuvo resultados que permitieron realizar un análisis y se establecieron conclusiones que describieron información importante para medir el grado de conocimiento sobre las herramientas de *Big Data*.

Resultados

El resultado obtenido fue una guía metodológica, que contiene los siguientes pasos:

1. Se realizó el almacenamiento y procesamiento de un *dataset* público, el cual contenía información sobre registro de medicamentos, haciendo uso de Hadoop.
2. Con la herramienta Hive que viene en la distribución Hortonworks de Hadoop, se hizó las consultas necesarias, ya que esta herramienta es similar a las instrucciones que se utilizan en SQL, por lo que, para los que están acostumbrados a trabajar con base de datos relacionales, les es fácil entender la lógica de cómo trabaja Hive. En El Salvador, SQL es el *software* más utilizado para bases de datos y esa es la razón por la que se seleccionó esta herramienta.
3. Para el análisis de datos estadísticos se utilizó el programa R. Las razones del por qué se seleccionó este programa fueron explicadas en el desarrollo de la investigación. Este programa nos devolvió información sobre datos importantes de los productos que estaban almacenados y a la vez permitió que se realizaran conclusiones en base a los resultados.
4. Después de haber hecho un análisis estadístico y las consultas pertinentes de los datos, se procedió a realizar los gráficos necesarios para una mejor comprensión de los resultados y en base a ello sacar conclusiones y de esa manera tomar decisiones. Las herramientas que se pueden utilizar son Google Chart, Jqplot o D3.js, dependiendo de las habilidades y conocimientos de la persona encargada de trabajar con la creación de los gráficos.

Discusión

De acuerdo con los resultados obtenidos en la encuesta que se pasó a los profesionales, se encontró que hay mucho desconocimiento de elementos importantes de *Big Data* en El Salvador, pero que a la vez hay muchos deseos de querer incursionar y utilizar las herramientas que faciliten el procesamiento y análisis de grandes volúmenes de datos. Por lo tanto, en esta investigación sería de mucha utilidad que se hiciera una guía metodológica que permite describir los pasos necesarios para utilizar

Revista *entorno*, junio 2016, número 61
ISSN: 2218-3345

algunas herramientas de *Big Data* y los requisitos que deben tomarse en cuenta para poder hacer uso de ellas.

Se usaron dos *dataset*, estos ficheros estaban en formato csv y fueron descargados de la web. Uno de los ficheros contenía 33028 registros y 9 campos: Registro sanitario, Nombre del producto, Fecha de vencimiento, Fábrica, Dirección de la Fábrica, Teléfono de la Fábrica, Representante, Teléfono del representante.

El otro fichero tenía 757 registros y 8 campos: N°, Principio Activo, Producto, Registro, Titular, Resolución, Fecha resolución, Uso/tratamiento.

Con la información contenida en los *dataset* se hizo la metodología, en la cual se utilizaron las herramientas Hadoop y Hive. Hadoop es una herramienta muy potente para el procesamiento de los datos y dentro de ella, en la distribución Hortonworks, está Hive, que es utilizada para realizar consultas de forma similar como se realiza en SQL.

Hay que considerar que para poder hacer uso de Hadoop, se debe contar con el recurso adecuado, pues esta requiere de equipo especial con las características mencionadas anteriormente.

Posteriormente se hizo uso del programa R para la realización de análisis estadístico, con esta herramienta se pudieron obtener resultados significativos del *dataset* que se había seleccionado.

Por último, se crearon visualizaciones atractivas de acuerdo con los resultados obtenidos de las consultas con Hive y de los análisis estadísticos con R. Las herramientas utilizadas fueron Google Chart, Jqplot y D3.js aunque no es necesario trabajar con todas, debido a que se puede utilizar la que resulte más conveniente o más práctica según lo que se necesite representar.

Referencias

Aguilar, L. J. (2013). *Big Data*, Analisis de los grandes volumenes de datos. México: Alfaomega.

Big Data, Analisis de los grandes volúmenes de datos. (s.f.).

Capriolo, E. Wample, D., & Rutherglen, J. (2012). Programming Hive. Estados Unidos de Norte América: Mike Loukides and Coutney Nash.

D3js.org. (2015). Obtenido de https://d3js.org/

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters.

Developers, G. (2016). Google Developers. Obtenido de https://developers.google.com/chart

Fragoso, R.B. (19 de marzo de 2014). Evaluando *Software*. com. Obtenido de http://www.evaluandosoftware.com/nota-3684-Que-es-el-Big-Data.html

Hortonworks (2011-2016). Hortonworks inc. Obtenido de http://hortonworks.com/

Jqplot pure javascript plotting (s.f.). Obtenido de http://www.jqplot.com/index.php

R, F. (s.f.). R Home. Obtenido de https://www.r-project.org/about.html

White, T. (Mayo de 2012). Hadoop. The Definitive Guide (3rd Edition). Obtenido de http://download.bigbata.com/ebook/oreilly/books/Hadoop.The.Definitive.Guide.3rd.Edition.May.2012.pdf