Universidad Libre de Colombia

Implementación de proyectos de Big Data

Informe Final Monográfico

Juan Camilo Gomez Pinzon
1 de Septiembre de 2017

Resumen

En esta época de información y datos, las empresas tienen que usar todos los recursos que tienen disponibles, para poder tener ventajas que les permitan hacerse con más clientes y, así mismo, mantener la confianza que se tiene con los clientes antiguos. Uno de estos recursos son los datos, de todo tipo, información que para la compañía sin ningún tipo de transformación o análisis no tiene un gran valor, sin embargo, con tecnologías que no son nuevas, pero que en años recientes han sido más y más usadas como lo es el Big Data, se puede generar un valor considerable para cualquier tipo de compañía que utilice los datos como su principal fuente de información.

La tecnología de Big Data tiene que pasar por diferentes etapas y cubriendo varios aspectos que son relevantes para la implementación de un proyecto de este tipo, para esto usaremos como caso de uso un vendedor de software/ licencias de uso de software, y que deberían tener en cuenta para un proyecto enfocado en Big Data, pasando por todas sus etapas y conceptos principales.

Palabras clave: Big Data, Hadoop, MapReduce, NoSQL, Análisis de datos, Modelo de datos, ETL.

Introducción

Big data es un tema que ha estado siendo utilizado más y más por diferentes sectores de la economía global, ya que encuentran un gran valor agregado en información que antes se veía desperdiciada o guardada por mucho tiempo y sin valor alguno, en el mercado actual el big data ha estado creciendo en uso, como por ejemplo en optimización del riesgo, nuevos modelos de negocio y colaboración de empleados además de ser usados en marketing y en el área de la salud, ayudando a los médicos a generar mucho más valor a sus pacientes y poder encontrar soluciones a sus enfermedades y alteraciones con la ayuda de todos los análisis realizados en la antigüedad.

Con esto en mente y buscando ampliar el uso de la tecnología de Big data en general, en este informe se hablará de cómo el Big data puede ser usado en la industria de venta, licenciamiento y soporte de software, generando valor agregado al cliente y ayudando así a la generación de nuevos clientes y más oportunidades de negocio.

Implementación de Big data

Recientemente la venta de software, licencias y soporte de estas ha visto un auge debido al nuevo uso de “Cloud Computing” y la facilidad de brindar servicios a través de internet con máquinas que de otra manera no podrían hacerlo, ya sea por su capacidad o su rendimiento. El objetivo principal de las compañías es poder vender sus productos a la mayor cantidad de gente posible y así obteniendo las mayores ganancias, Big data puede ayudarle a este mercado no solo a vender productos, sino también a la optimización de recursos del cliente y, por consiguiente, la retención de los mismos, dándoles un valor agregado que un proveedor de software común no les ofrecería.

La inclusión de Big data como componente para generar valor a cliente ha generado otra consecuencia, esta ha sido la presión para recortar aún más los tiempos de decisión. En un entorno complejo, las oportunidades se abren y se cierran en espacios de tiempo cortísimos y las organizaciones se ven obligadas a tomar decisiones casi en tiempo real (García, 2013). Lo que hace la hace una herramienta fundamental para el crecimiento y desarrollo de una compañía en esta época y por lo cual es una herramienta vital para la competencia con otras compañías.

A lo largo de los últimos años el término “Big Data” se ha estado usando con más fuerza en el mercado, sin embargo, muchas personas aún no entienden el concepto de “Big Data”, Big Data puede ser definido como las técnicas y tecnologías que hacen que sea económico hacer frente a los datos a una escala extrema teniendo en cuenta 3 aspectos:

“1) Las técnicas y la tecnología, lo que significa que la empresa tenga personal, el cual tenga gran representación y análisis de datos para tener un valor agregado con información que no ha sido manejada.

2) Escala extrema de datos que supera a la tecnología actual debido a su volumen, velocidad y variedad.

3) El valor económico, haciendo que las soluciones sean asequibles y ayuden a la inversión de los negocios" (Forrester,2011)

Desde esta definición podemos observar que los retos más grandes para una compañía que quiera usar Big Data, recae en usar la información apropiadamente, definiendo un objetivo claro y utilizando técnicas tanto para analizar esta gran cantidad de información de una manera rápida y para poder mostrar esta información de una forma clara y que pueda ser usada para tomar decisiones de negocio que ayuden a la compañía de una manera significativa.

Cabe aclarar que las cantidades de datos a las cuales llamamos "Big data", pueden variar dependiendo de la cantidad de procesamiento que tengamos, la cantidad de información que necesitemos extraer y el tiempo que tengamos para hacerlo.

Pasos de implementación de Carrasco

Teniendo en cuenta estos aspectos, podemos entrar en la industria de venta de software y cómo estos pueden aprovechar este nuevo concepto, Para esto, Aquí algunos pasos que se deben tener en cuenta en cuanto a la implementación de Big data (Carrasco,2013):

Entender el negocio y los datos. Este primer paso pide un análisis detallado con las personas que hoy laboran y entienden los procesos y los datos que la empresa maneja.

El segundo paso consiste en determinar los problemas y cómo los datos pueden ayudar. Al momento de conocer los procesos es muy posible que se encuentren los problemas de la empresa o del negocio.

Establecer expectativas razonables, es decir, definir metas alcanzables; esto se puede lograr si al implementar la solución de un problema éste no presenta alguna mejora, y se debe buscar otra solución.

Inicios de la implementación

Existe una recomendación especial, y es que cuando se inicia un proyecto de Big Data es necesario trabajar en paralelo con el sistema que hoy está funcionando.

Al tratar de implementar un proyecto de Big Data se debe ser flexible con la metodología y las herramientas; esto se debe a que las dos anteriores son recientes y pueden llegar a presentar problemas al implementarlas. Esto se puede solucionar realizando investigación e inversión en este tipo de tecnología.

Es importante mantener el objetivo de Big Data en mente; esto porque el proceso es pesado y porque no es tedioso, máxime cuando los métodos y herramientas que usan Big Data para el análisis de datos aún pueden presentar problemas, y la idea es que se mantenga en mente la meta final del proyecto sin desanimarse pronto.

Siguiendo los pasos de Carrasco, es necesario que entendamos los datos que más se utilizan en la venta de software hoy en día, ya sea como una licencia o como un servicio (SaaS). En el caso de las licencias no solo se tienen los tradicionales contratos y el periodo en el que se acaban los mismos, sino que también se cuenta con cotizaciones popularidad del programa en general, redes sociales de los productos y en general la vista de los productos/compañías en el ojo público, esta información en general se encuentra en varios tipos: audio, texto, videos, mensajes en redes sociales (Facebook, Twitter) todo esto información debe ser tomada en cuenta en la recolección de datos, para luego ser analizados con las diferentes técnicas de análisis de las cuales hablaremos más adelante. Por otro lado, en el caso de los SaaS, se tienen estadísticas más importantes además de las anteriormente descritas, ya que estos servicios son controlados, al menos en parte, por la compañía que presta estos servicios por esto se tienen estadísticas de uso general, programas que se ejecutan, la versión de cada producto, periodo del servicio, reporte de uso por máquina y otros más que

pueden ayudarle a la extracción de información valiosa tanto para la compañía como para el cliente.

La mayoría de empresas, guarda infinidad de información que no está siendo usada de una manera eficiente para la compañía, esta información está usualmente almacenada en data warehouses (centros de datos de información), y es de estos warehouses que contienen toda esta información que se hará la extracción de datos, cabe aclarar que los datos que se extraen de estos centros de información tienen un valor mucho mayor al que antes se encontraba en estos. Ya que:

La digitalización de prácticamente “todo” da lugar a nuevos tipos de grandes datos en tiempo real en un amplio abanico de sectores. “Muchos de ellos son datos no normalizados: por ejemplo, datos en streaming, geoespaciales o generados por sensores que no encajan bien en los warehouses relacionales, tradicionales y estructurados” (IBM, 2012).

“Las tecnologías y técnicas de análisis avanzado de hoy en día permiten a las empresas extraer conocimientos de los datos con un nivel de sofisticación, velocidad y precisión nunca antes visto” (IBM, 2012).

Tipos de datos en Big data

Por esta misma razón se deberán usar diferentes herramientas que pueden guardar estos datos de una forma que más adelante nos genere valor, en esta parte del proceso es importante que se tengan algunos conceptos claros como los son datos estructurados, no estructurados y semi estructurados. En los cuales se dividirá toda la información que extraigamos sin importar la técnica de extracción que usemos, como lo dice IBM, la digitalización de todo hace que podamos guardar videos imágenes y otro tipo de elementos que antes no se le podía determinar un valor en una base de datos relacional, estos se consideran datos no estructurados, los datos deben poder ser organizados, clasificados, almacenado. Estos datos se pueden observar a diario en correos electrónicos, archivos de texto, un documento de algún procesador de palabra, hojas electrónicas, una

imagen, un objeto, archivos de audio, blogs, mensajes de correo de voz, mensajes instantáneos, contenidos Web y archivos de video, entre otros (Joyanes, 2014)

Además de estos datos se encuentran los datos semiestructurados, que son datos que no tienen un formato definido, los que “tienen son etiquetas que facilitan separar un dato de otro. Un dato de estos se lee con un conjunto de reglas de cierto nivel de complejidad” (Paulsen, 2011) como, por ejemplo:

Nombre	Teléfono	Sexo	Correo
Pedro Pérez	2127409	M	pedroperez@gmail.com

Nombre	Apellido	Teléfono	Correo
Mario	Rodríguez	0987526221	mario@gmail

Primer apellido	Segundo apellido	Nombres	Correo	Teléfono
Martínez	Arévalo	Julio	aremar@gmail.com	24356712

Tablas de información de ejemplo, Juan José Camargo-Vega, 2014

Como se puede ver estas tres tablas contienen información que no puede ser relacionada en una base de datos común, ya que las tablas agrupan la información de forma diferente lo que pude causar problemas en una base de datos convencional.

Y finalmente los datos estructurados, este tipo de datos se dividen en estáticos (array, cadena de caracteres y registros) y dinámicos (listas, pilas, colas, árboles, archivos). “Se puede definir que los datos estructurados son aquellos de mayor facilidad para acceder, pues tienen una estructura bien especificada” (Karl, 2011)

Al obtener toda esta cantidad de datos puede parecer imposible extraer información valiosa de estos pero debemos recordar que:

“En todos los sectores el caso de negocio de big data está orientado en gran medida a abordar objetivos centrados en el cliente” (IBM, 2012)

“Una base de gestión de la información escalable y extensible es un requisito fundamental para el avance de big data” (IBM, 2012)

“Para que las empresas puedan obtener el máximo valor de big data son necesarias funcionalidades analíticas avanzadas” (IBM, 2012)

Almacenamiento de datos

Para el proceso de análisis estos datos deben estar guardados en bases de datos y sea bases de datos SQL o NOSQL, para ser procesadas y extraer un resultado que será guardado en las mismas y procedería a un análisis de resultados y posterior modelado de resultados.

El SQL (Structure Query Language) es un lenguaje estándar de cuarta generación que “se utiliza para definir, gestionar y manipular la información contenida en una Base de Datos Relacional desarrollado por Microsoft Corporation” (IBM,2012)

El modelo SQL ha sido usado desde los inicios de la computación para relacionar objetos y datos de una manera organizada y consultable con diferentes sentencias o “querys” y es usada en la actualidad por la mayoría de bases de datos relacionales en el mercado.

Mientras que las bases de datos NOSQL son más modernas, se definen como, “Una categoría amplia con un grupo de soluciones de persistencia que no siguen el modelo de datos relacional, y que no utilizan SQL como lenguaje de consulta estas se pueden definir en 4 principales categorías” (Hansel Garcia, 2012)

• Orientadas a clave-valor (Key-Value stores)

• Orientadas a columnas (Wide Column stores)

• Orientadas a documentos (Document stores)

• Orientadas a grafos (Graph databases)

Estas 4 categorías representan la mayoría de datos que podemos encontrar en la

web y otras fuentes de información no convencionales de las cuales se tienen grandes cantidades de datos.

En nuestro caso específico se tendría información relevante en cada una de estas categorías

Para las orientadas a clave-valor (contratos, información general de clientes y usuarios, etc.) que son para datos que se tengan en una volumen muy grande como para ser guardados en bases de datos normales, en estas bases , simplemente se almacenan datos que se agrupan dependiendo de un valor general, también llamado etiqueta que permite consultar datos específicos de esta manera, cabe aclarar que estas bases de datos son muy malas en cuanto concierne a integridad de los datos, ya que el volumen de estas bases comúnmente es muy grande como para hacer una verificación de este tipo.

En cuanto a las orientadas a columnas, Las bases de datos orientadas a columnas son en realidad lo que se podría suponer, tablas de datos donde las columnas de valores de datos representan el almacenamiento estructural. “Los datos son almacenados como secciones de las columnas de datos en lugar de filas de datos, como en la mayoría de los gestores relacionales” (Hansel Garcia, 2012) es por esto que estas sirven para encontrar una característica especifica de una manera esencialmente más rápida que las demás, en nuestro caso se podría usar a la hora de encontrar información determinada de un programa o licencia adquirida por cientos o miles de clientes y como cada de uno de ellos la usa.

Siguiendo con las orientadas a documentos, estas bases de datos son consideradas por muchos como un escalón superior ante los simples gestores de llave-valor, puesto que permiten encapsular pares de llave-valor en estructuras más complejas denominadas documentos. Por otra parte, “no existe un esquema estricto a seguir para definir estos documentos, lo cual simplifica sustancialmente su uso"(Osmel Yanes, 2012), estas bases serian el resultado de una exploración de toda la información contenida en las bases de dato llave-valor, después de ser analizadas en busca de coincidencias y una congruencia, para ser usadas en análisis más fácilmente.

Por ultimo al tener un análisis de la información y poder crear relaciones de esta información estas se podrían guardar en este tipo de bases ya que, este tipo de base de datos está diseñada para los datos cuyas relaciones son bien representadas en forma de grafo, o sea, “los datos son elementos interconectados con un número no determinado de relaciones entre ellos debe estar absolutamente normalizada, esto quiere decir que cada tabla tendría una sola columna y cada relación tan solo dos, con esto se consigue que cualquier cambio en la estructura de la información tenga un efecto tan solo local”-Hansel 2012

De esta manera se guardaría la información desde la extracción inicial de información con poco o mínimo valor para las compañías hasta llevarla a un punto en el que la información es considerada para tomar decisiones importantes para el negocio y que, además, le den valor al cliente.

Es importante que conozcamos las herramientas que se deben usar en este proceso, en su mayoría gestores de bases de datos no relacionales que nos ayudan a tener esta información almacenada de manera organizada e integra (si es el caso):

Hadoop Distributed File System (HDFS)	HDFS facilita el almacenamiento para los clusters Hadoop. Divide la información en partes pequeñas y la distribuye entre los servidores/nodos disponibles
MapReduce	MapReduce provee la interfaz para la distribución de sub-tareas y la recolección de resultados además controla el progreso de cada una de las tareas asociadas a los servidores/nodos.
PIG and PIG Latin	El lenguaje de programación Pig está diseñado para aceptar cualquier tipo de dato. Está compuesto de dos partes, el lenguaje, el PigLatin, y la ventana de ejecución donde el cogido es ejecutado.

Hive	Hive es una arquitectura de soporte de Hadoop que equilibra el uso de SQL con la plataforma Hadoop, permite a programadores de SQl usar sentencias HQL, que en estructura son muy parecidas a SQL.
Jaql	Jaql es un lenguaje funcional de sentencias, diseñado para facilitar procesos paralelos, este consiste en traducir sentencias en tareas MapReduce.
Zookeeper	Zookeeper permite a una infraestructura centralizada con varios servicios, proveer sincronización a través de un cluster de servidores, aplicaciones de análisis de Big data usan estos servicios para coordinar procesamiento paralelo entre clusters.
HBase	HBase es un Sistema de manejo de bases de datos orientado en columnas usa un lenguaje NoSQL
Cassandra	Cassandra es un sistema de bases de datos distribuido designado para poder manejar grandes volúmenes de datos. Este es un sistema NoSQL
Oozie	Oozie, es un proyecto de código libre, que facilita el flujo de trabajo y la coordinación entre tareas.

ETL

En este punto ya tenemos la información, diferenciada en varias bases de datos, ya sean SQL o no SQL, y tenemos un objetivo claro, tomar decisiones frente a la compañía que no solo del den valor a la compañía sino al mismo tiempo al cliente. Para poder empezar con el análisis de información es necesario entender el concepto de ETL, que se usara constantemente no solo durante el proceso de análisis sino en el proceso de extracción de la información.

El proceso de ETL, se enfoca en 3 procesos: Extraer, Transformar, y Cargar (Load), en la etapa de extracción, "los datos son recibidos de diferentes orígenes o fuentes, que pueden ser: páginas web, redes sociales, máquina a máquina (M2M), transacciones, biometría o generados por el ser humano" (Krishnan 2013), Como vimos anteriormente nuestros datos provienen de este tipo de fuentes y algunas otras descritas anteriormente.

En la etapa de carga "los datos se cargan aplicando el concepto de metadatos (datos que describen otros datos). Además de la carga como tal, es la primera vez que los datos se estructuran"-Krishnan,2013, se estructuran buscando un objetivo específico, o agrupándolos por características que, a nosotros, como compañía, nos interesen.

Por último, "El objetivo de la extracción es obtener datos para su posterior análisis, generar informes operativos y su posible visualización y, por último, y no el más importante, para su almacenamiento"Krishnan 2013, es importante aclarar que el valor de la información que se extraiga de estos constantes procesos es definido por las decisiones que tomen con esto, es decir, que las compañías deben usar esta información de forma retroactiva, para que esta le de valor a la compañía.

Análisis específicos

A la hora de realizar el proceso de ETL, es importante realizar un análisis final de la información y mostrarla a las personas encargadas de tomar decisiones en la compañía para que estas puedan tenerla en cuenta para los siguientes pasos de la compañía.

Para esto es necesario comprender los análisis que se pueden realizar a la información y que análisis puede servir mejor para nuestro caso específico:

Para empezar se tienen los análisis cuantitativos, que se enfocan en números y estadísticas, en nuestro caso podemos enfocarnos en la usabilidad de ciertos productos o la venta de los mismos, para tener en claro el universo que tiene cada uno de estos productos.

Otro de los análisis que se pueden llegar a realizar es un análisis cuantitativo orientado a describir cualidades de varios datos por medio de palabras. En contraste con el análisis de datos cuantitativo, esto implica analizar una pequeña muestra con mayor profundidad, en nuestro caso implica buscar características que tengan productos en comparación a otros, ya sean propios o de la competencia, en busca de un diferenciador que esté generando o pueda generar valor adicional. Está enfocado en la interpretación humana de ciertos datos que pueden o no llegar a generar valor en un futuro.

Uno de los posibles análisis, y es uno de los más usados en este ámbito de Big data, es el de la minería de datos, “La minería o data mining es una extracción de datos para luego poder hallar relaciones entre ellos. Existen dos tipos de minería de datos: descriptivo, que proporcione información sobre los datos existentes, y predictivo, lo que hace que los pronósticos basados en los datos” (Tucker 2013), este es uno de los principales análisis a realizar, buscando información pertinente a productos específicos, contratos y demás que faciliten nuestras decisiones en el futuro sobre estos mismos.

Además se cuentan con otros tipos de análisis como lo son:

El Análisis Confirmatorio de Datos es un enfoque deductivo en el cual se propone previamente la causa del fenómeno investigado. La causa o suposición que se propone se conoce como hipótesis. Luego, los datos son analizados para comprobar o refutar la hipótesis y proporcionar respuestas definitivas a preguntas específicas.

Analítica diagnóstica: la analítica diagnóstica tiene como objetivo determinar la causa de un fenómeno que ocurrió en el pasado, usando preguntas que se enfocan en la razón del evento.

Analítica predictiva: la analítica predictiva se ejecuta en un intento por determinar el resultado de un evento que podría ocurrir en el futuro.

Estas últimas técnicas se basan en una de las primeras 3 mencionadas, estas técnicas pueden ser usadas dependiendo de nuestro objetivo con la información y tienen que ser consideradas a la hora de iniciar con el proceso de ETL.

Al tener en nuestro poder información que pasó por un proceso de ETL, y que además fue posteriormente analizada se puede proceder al modelado de esta.

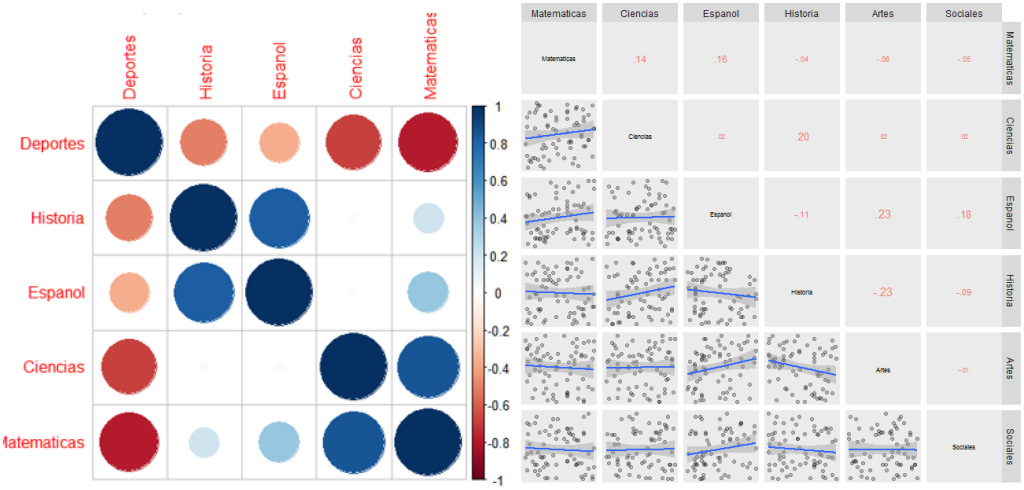
Modelado de Información

La información extraída y analizada debe ser modelada correctamente, para esto existen varias herramientas que nos permiten modelar la información de diferentes formas y a su vez generar reportes fáciles de entender para una persona que no estuvo en el proceso ETL. En esta parte del proceso usaremos la herramienta R y Rstudio, las cuales son herramientas de análisis y modelado de datos fáciles de usar.

Para nuestro caso específico debemos tener en cuenta que la información y el análisis se basa en el cliente y los productos que les vendemos diariamente, por esto las variables se van a centrar en, principalmente: tiempo de uso, licencias activas, maquinas que han estado estos programas en los últimos meses, maquinas que tienen el producto instalado y licenciado y, sin embargo, no lo han usado nunca. Esto con el fin de proveer al cliente con soluciones eficientes, que les ayuden a disminuir uso y costos, ganando así confianza con el cliente y una relación mucho más duradera.

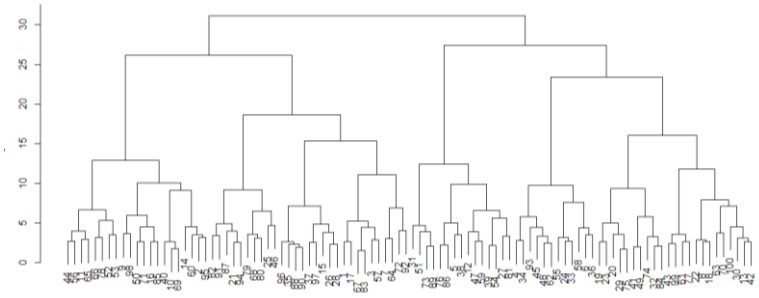
Siendo así, hay varios programas que nos ayudan a verificar todas estas variables con diferentes graficas que les muestran a las personas en cargo de las decisiones del negocio, como también a los clientes, la información analizada y así tomar decisiones efectivas, a continuación, se muestran las gráficas más usadas para tomar decisiones basadas en big data:

- **Correlación de Pearson:** el coeficiente de correlación de Pearson es una medida de la relación lineal entre dos variables aleatorias y cuantitativas. La diferencia entre el coeficiente de correlación y la covarianza no es otra que a éste primero no le afecta la escala de medida de las variables



Grafos tomados de la aplicación R, módulo rattle.

- **Clustering Jerarquico**: “Un procedimiento estadístico multivariante que se basa en un conjunto de información de una muestra de individuos, a partir del cual se intenta reorganizar a estos individuos en grupos relativamente homogéneos, con la particularidad que los individuos dentro de un grupo sean lo más similares posibles entre sí y entre distintos grupos sean lo más disímiles entre sí” Aldenderfer y Blashfield (1984)



Seguridad en Big data:

El tema de seguridad en Big data es un tema muy concurrido en los últimos años, este tema le importa a clientes y empresas por igual ya que una pérdida de información conlleva a perder una ventaja competitiva y además hacer perder confianza con el cliente, el cual cada vez es más consciente de lo que se hace con su información y esto puede llevar a afectar relaciones que en otras circunstancias podrían haber sido mucho más duraderas, es por eso que este tema no se debe dejar atrás en cualquier empresa que esté considerando usar Big data.

Para esto es recomendable tener en cuenta los diferentes modelos de seguridad que las empresas están usando en este momento:

ISO 17799: ISO / IEC 17799:2005 establece los lineamientos y principios generales para iniciar, implementar, mantener y mejorar la gestión de seguridad de la información en una organización.

COBIT: El COBIT es precisamente un modelo para auditar la gestión y control de los sistemas de información y tecnología, orientado a todos los sectores de una organización,

ITIL: Information Technology Infrastructure Library (‘Biblioteca de Infraestructura de Tecnologías de Información’), frecuentemente abreviada ITIL, es un marco de trabajo de las mejores prácticas destinadas a facilitar la entrega de servicios de tecnologías de la información (TI) de alta calidad.

ISO SERIE 2700 ISO/IEC 27000 es un conjunto de estándares desarrollados -o en fase de desarrollo- por ISO (International Organization for Standardization) e IEC (International Electrotechnical Commission), que proporcionan un marco de gestión de la seguridad de la información utilizable por cualquier tipo de organización, pública o privada, grande o pequeña.

Inconvenientes/Desventajas en Big Data

Por Ultimo es importante hablar de algunas de las desventajas o inconvenientes que pueden aparecer al momento de realizarlo o planearlo, en nuestro caso en específico, el control de licencias y productos de software debe enfocarse correctamente para no malgastar recursos de manera que no le den algún tipo de ganancia a la empresa, entre otros inconvenientes se encuentran:

- Gastos iniciales y de formación
- Problemas de privacidad y uso de información
- Coste total

En esencia estos son unos de los mayores inconvenientes del uso de big data, sin embargo, si la empresa considera útil esta técnica de análisis y toma de decisiones y después de analizar todos los inconvenientes no hay ningún problema, el resultado de esta actividad será mucho más gratificante que los inconvenientes que posee

Conclusión

La información y la forma en la que manipulamos esta, ha cambiado drásticamente desde el último siglo y todo gracias a los avances tecnológicos que nos permiten manipular grandes cantidades de información. Uno de los más grandes avances es Big Data, que consiste en un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis.

Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho más vital cuando se usa en conjunto con la plataforma de Big Data.

De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico

Además del gran volumen de información, esta existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, video, sistemas GPS, incontables sensores digitales en equipos industriales y demás, los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data.

“Entre el nacimiento del mundo y el año 2003, hubo cinco exabytes de información creada. Actualmente creamos cinco exabytes cada dos días” Eric Schmidt, ex CEO de Google.

Bibliografía

- J Nagler, JA Tucker ,Drawing inferences and testing theories with big data, 2015

- F. Carrasco, Los 6 pasos que su organización debe seguir para confiar en Big Data. América Latina. Disponible en: http://www.cioal.com/2013/07/31/los-6-pasos-que-su-organizacion-debe-seguir-para-confiar-en-big-data/, 2013.

- ibm.com, ¿Qué es Big Data? Disponible en: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html, 2012

- es.wikipedia.org, RFID. Disponible en: http://es.wikipedia.org/wiki/RFID, 2010.

- E. Redmond, & J. Wilson, Seven Databases in Seven Weeks, USA: O'Reilly Media, Inc., Pragmatic Programmers, LLC. 2012

- Emc.com, Big Data transforms Business. Disponible en: http://www.emc.com/microsites/ebook/index.htm#/slide-intro, 2012

- T. Olavsrud, Big Data Causes Concern and Big Confusion. Disponible en: http://www.cio.com/article/700804/Big_Data_Causes_Concern_and_Big_Confusion?page=2&taxonomyId=3002, 2012

- hadoop.apache.org, Disponible en: http://hadoop.apache.org/, 2013.

- K Krishnan , Data warehousing in the age of big data, 2013