¿Qué es Big Data?

Todos formamos parte de ese gran crecimiento de datos

Ricardo Barranco Fragoso 18-06-2012

Debido al gran avance que existe día con día en las tecnologías de información, las organizaciones se han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubrir y entender más allá de lo que sus herramientas tradicionales reportan sobre su información, al mismo tiempo que durante los últimos años el gran crecimiento de las aplicaciones disponibles en internet (geo-referenciamiento, redes sociales, etc.) han sido parte importante en las decisiones de negocio de las empresas. El presente artículo tiene como propósito introducir al lector en el concepto de Big Data y describir algunas características de los componentes principales que constituyen una solución de este tipo.

1. Introducción

El primer cuestionamiento que posiblemente llegue a su mente en este momento es ¿Qué es Big Data y porqué se ha vuelto tan importante? pues bien, en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia 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. 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, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos. Entonces ¿Cuánto es demasiada información de manera que sea elegible para ser procesada y analizada utilizando Big Data? Analicemos primeramente en términos de bytes:

Gigabyte = 10^9 = 1,000,000,000

Terabyte = 10^{12} = 1,000,000,000,000

Petabyte = 10^{15} = 1,000,000,000,000,000

Exabyte = 10^{18} = 1,000,000,000,000,000,000

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, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., 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.

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. Pensemos en nuestras manos izquierda y derecha, cada una ofrece fortalezas individuales para cada tarea en específico. Por ejemplo, un beisbolista sabe que una de sus manos es mejor para lanzar la pelota y la otra para atraparla; puede ser que cada mano intente hacer la actividad de la otra, mas sin embargo, el resultado no será el más óptimo.

2. ¿De dónde proviene toda esa información?

Los seres humanos estamos creando y almacenando información constantemente y cada vez más en cantidades astronómicas. Se podría decir que si todos los bits y bytes de datos del último año fueran guardados en CD's, se generaría una gran torre desde la Tierra hasta la Luna y de regreso.

Esta contribución a la acumulación masiva de datos la podemos encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto le añadimos transacciones financieras realizadas en línea o por dispositivos móviles, análisis de redes sociales (en Twitter son cerca de 12 Terabytes de tweets creados diariamente y Facebook almacena alrededor de 100 Petabytes de fotos y videos), ubicación geográfica mediante coordenadas GPS, en otras palabras, todas aquellas actividades que la mayoría de nosotros realizamos varias veces al día con nuestros "smartphones", estamos hablando de que se generan alrededor de 2.5 quintillones de bytes diariamente en el mundo.

1 quintillón = 10^{30} *= 1,000,000,000,000,000,000,000,000,000,000*

De acuerdo con un estudio realizado por Cisco[1], entre el 2011 y el 2016 la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78%, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta. Las naciones unidas proyectan que la población mundial alcanzará los 7.5 billones para el 2016 de tal modo que habrá cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 Exabytes mensuales o 130 Exabytes anuales. Este volumen de tráfico previsto para 2016 equivale a 33 billones de DVDs anuales o 813 cuatrillones de mensajes de texto.

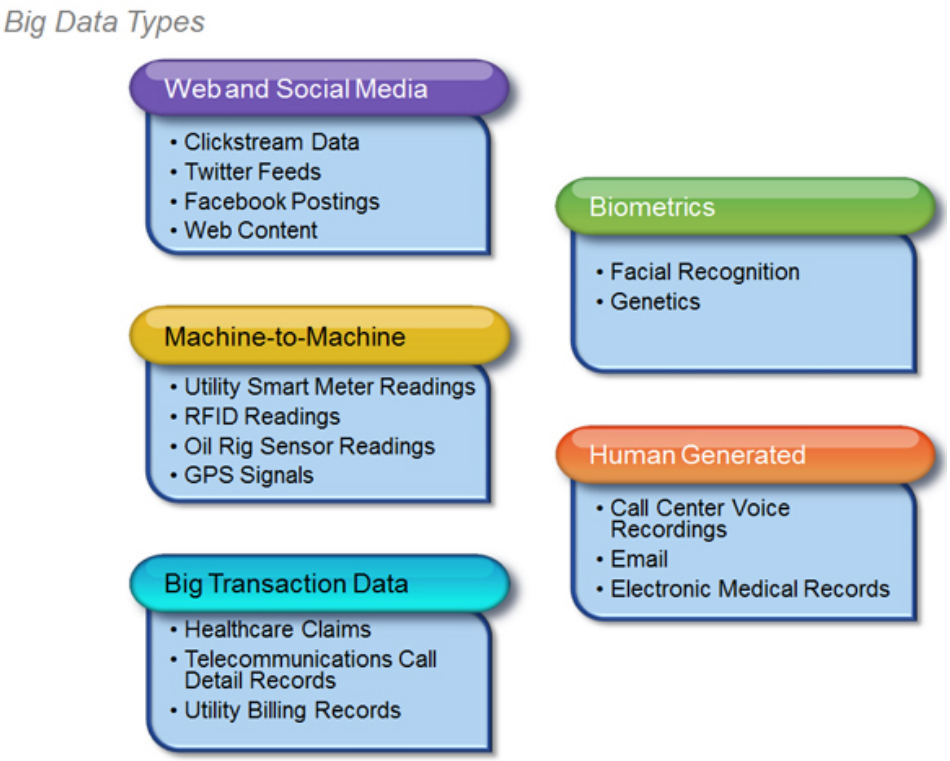
Pero no solamente somos los seres humanos quienes contribuimos a este crecimiento enorme de información, existe también la comunicación denominada máquina a máquina (M2M machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante. Sensores digitales instalados en contenedores para determinar la ruta generada durante una entrega de algún paquete y que esta información sea enviada a las compañías de transportación, sensores en medidores eléctricos para determinar el consumo de energía a intervalos regulares para que sea enviada esta información a las compañías del sector energético. Se estima que hay más de 30 millones de sensores interconectados en distintos sectores como automotriz, transportación, industrial, servicios, comercial, etc. y se espera que este número crezca en un 30% anualmente.

3. ¿Qué tipos de datos debo explorar?

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué información es la que se debe analizar?, sin embargo, el cuestionamiento debería estar enfocado hacia ¿qué problema es el que se está tratando de resolver?.[2]

Si bien sabemos que existe una amplia variedad de tipos de datos a analizar, una buena clasificación nos ayudaría a entender mejor su representación, aunque es muy probable que estas categorías puedan extenderse con el avance tecnológico.

Figura 1. Tipos de datos de Big Data[2]



1.- *Web and Social Media*: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc, blogs.

2.- *Machine-to-Machine (M2M)*: M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3.- *Big Transaction Data*: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4.- *Biometrics*: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5.- *Human Generated*: Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

4. Componentes de una plataforma Big Data

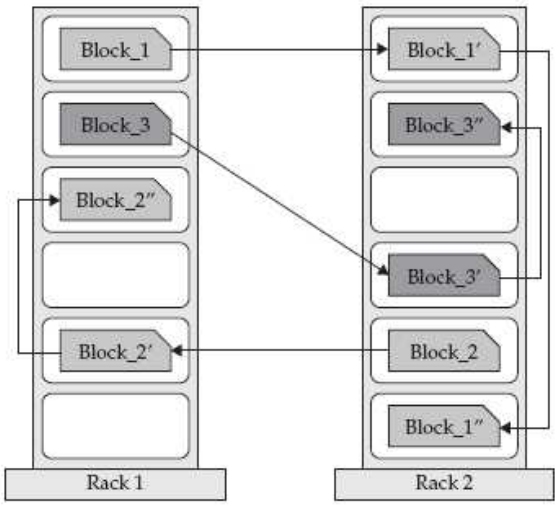
Las organizaciones han atacado esta problemática desde diferentes ángulos. Todas esas montañas de información han generado un costo potencial al no descubrir el gran valor asociado. Desde luego, el ángulo correcto que actualmente tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto *Hadoop.*

Hadoop está inspirado en el proyecto de Google File System(GFS) y en el paradigma de programación *MapReduce*, el cual consiste en dividir en dos tareas (`mapper` – `reducer`) para manipular los datos distribuidos a nodos de un clúster logrando un alto paralelismo en el procesamiento.[5] Hadoop está compuesto de tres piezas: *Hadoop Distributed File System* (HDFS), *Hadoop MapReduce* y *Hadoop Common.*

Hadoop Distributed File System(HDFS)
Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas *bloques* y distribuidas a través del clúster; de esta manera, las funciones `map` y `reduce` pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes.

La siguiente figura ejemplifica como los bloques de datos son escritos hacia HDFS. Observe que cada bloque es almacenado tres veces y al menos un bloque se almacena en un diferente rack para lograr redundancia.

Figura 2. Ejemplo de HDFS



Hadoop MapReduce

MapReduce es el núcleo de Hadoop. El término MapReduce en realidad se refiere a dos procesos separados que Hadoop ejecuta. El primer proceso `map`, el cual toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en `tuplas` (pares de llave/valor). El proceso `reduce` obtiene la salida de `map` como datos de entrada y combina las tuplas en un conjunto más pequeño de las mismas. Una fase intermedia es la denominada `Shuffle` la cual obtiene las tuplas del proceso `map` y determina que nodo procesará estos datos dirigiendo la salida a una tarea `reduce` en específico.

La siguiente figura ejemplifica un flujo de datos en un proceso sencillo de MapReduce.

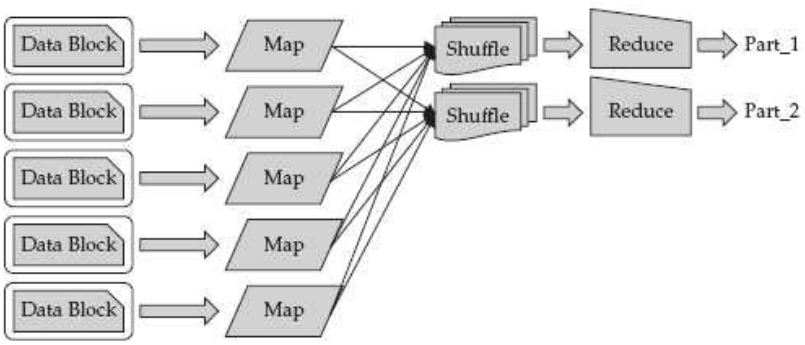


Figura 3. Ejemplo de MapReduce

Hadoop Common

Hadoop Common Components son un conjunto de librerías que soportan varios subproyectos de Hadoop.
Además de estos tres componentes principales de Hadoop, existen otros proyectos relacionados los cuales son definidos a continuación:

Avro

Es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema esta definido dentro del archivo.

Cassandra
Cassandra es una base de datos no relacional distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java. Permite grandes volúmenes de datos en forma distribuida. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma.

Chukwa
Diseñado para la colección y análisis a gran escala de "logs". Incluye un toolkit para desplegar los resultados del análisis y monitoreo.

Flume
Tal como su nombre lo indica, su tarea principal es dirigir los datos de una fuente hacia alguna otra localidad, en este caso hacia el ambiente de Hadoop. Existen tres entidades principales: `sources, decorators y sinks.` Un `source` es básicamente cualquier fuente de datos, `sink` es el destino de una operación en específico y un `decorator` es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos.

HBase
Es una base de datos columnar (column-oriented database) que se ejecuta en HDFS. HBase no soporta SQL, de hecho, HBase no es una base de datos relacional. Cada tabla contiene filas y columnas como una base de datos relacional. HBase permite que muchos atributos sean agrupados llamándolos *familias de columnas,* de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto. Eso es distinto a las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto. Facebook utiliza HBase en su plataforma desde Noviembre del 2010.

Hive
Es una infraestructura de data warehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language(HQL), estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el cluster de Hadoop.
El siguiente es un ejemplo en HQL para crear una tabla, cargar datos y obtener información de la tabla utilizando Hive:

```
CREATE TABLE Tweets (from_user STRING, userid BIGINT, tweettext STRING, retweets INT)
COMMENT 'This is the Twitter feed table'
STORED AS SEQUENCEFILE;
LOAD DATA INPATH 'hdfs://node/tweetdata' INTO TABLE TWEETS;
SELECT from_user, SUM(retweets)
FROM TWEETS
GROUP BY from_user;
```

Jaql

Fue donado por IBM a la comunidad de software libre. Query Language for Javascript Object Notation (JSON) es un lenguaje funcional y declarativo que permite la explotación de datos en formato JSON diseñado para procesar grandes volúmenes de información. Para explotar el paralelismo, Jaql reescribe los queries de alto nivel (cuando es necesario) en queries de "bajo nivel" para distribuirlos como procesos MapReduce.

Internamente el motor de Jaql transforma el query en procesos `map` y `reduce` para reducir el tiempo de desarrollo asociado en analizar los datos en Hadoop. Jaql posee de una infraestructura flexible para administrar y analizar datos semiestructurados como XML, archivos CSV, archivos planos, datos relacionales, etc.

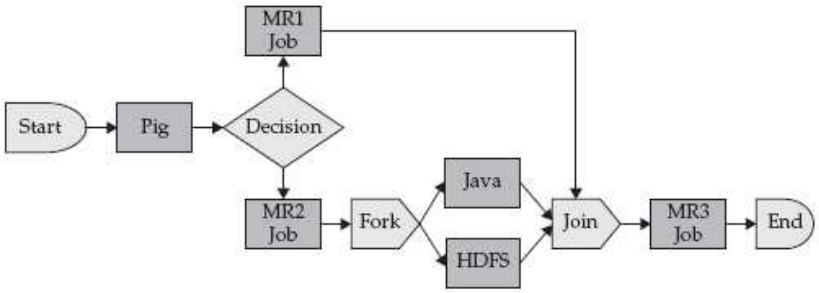
Lucene
Es un proyecto de Apache bastante popular para realizar búsquedas sobre textos. Lucene provee de librerías para indexación y búsqueda de texto. Ha sido principalmente utilizado en la implementación de motores de búsqueda (aunque hay que considerar que no tiene funciones de "crawling" ni análisis de documentos HTML ya incorporadas). El concepto a nivel de arquitectura de Lucene es simple, básicamente los documentos (*document*) son divididos en campos de texto (*fields*) y se genera un índice sobre estos campos de texto. La indexación es el componente clave de Lucene, lo que le permite realizar búsquedas rápidamente independientemente del formato del archivo, ya sean PDFs, documentos HTML, etc.

Oozie
Como pudo haber notado, existen varios procesos que son ejecutados en distintos momentos los cuales necesitan ser orquestados para satisfacer las necesidades de tan complejo análisis de información.

Oozie es un proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos. Permite que el usuario pueda definir acciones y las dependencias entre dichas acciones.

Un flujo de trabajo en Oozie es definido mediante un grafo acíclico llamado *Directed Acyclical Graph (DAG),* y es acíclico puesto que no permite ciclos en el grafo; es decir, solo hay un punto de entrada y de salida y todas las tareas y dependencias parten del punto inicial al punto final sin puntos de retorno. Un ejemplo de un flujo de trabajo en Oozie se representa de la siguiente manera:

Figura 4. Flujo de trabajo en Oozie



Pig
Inicialmente desarrollado por Yahoo para permitir a los usuarios de Hadoop enfocarse más en analizar todos los conjuntos de datos y dedicar menos tiempo en construir los programas MapReduce. Tal como su nombre lo indica al igual que cualquier cerdo que come cualquier cosa, el lenguaje *PigLatin* fue diseñado para manejar cualquier tipo de dato y *Pig* es el ambiente de ejecución donde estos programas son ejecutados, de manera muy similar a la relación entre la máquina virtual de Java (JVM) y una aplicación Java.

ZooKeeper
ZooKeeper es otro proyecto de código abierto de Apache que provee de una infraestructura centralizada y de servicios que pueden ser utilizados por aplicaciones para asegurarse de que los procesos a través de un cluster sean serializados o sincronizados.
Internamente en ZooKeeper una aplicación puede crear un archivo que se persiste en memoria en los servidores ZooKeeper llamado *znode.* Este archivo *znode* puede ser actualizado por cualquier nodo en el cluster, y cualquier nodo puede registrar que sea informado de los cambios ocurridos en ese *znode*; es decir, un servidor puede ser configurado para "vigilar" un *znode* en particular. De este modo, las aplicaciones pueden sincronizar sus procesos a través de un cluster distribuido actualizando su estatus en cada *znode*, el cual informará al resto del cluster sobre el estatus correspondiente de algún nodo en específico.

Como podrá observar, más allá de Hadoop, una plataforma de Big Data consiste de todo un ecosistema de proyectos que en conjunto permiten simplificar, administrar, coordinar y analizar grandes volúmenes de información.

5. Big Data y el campo de investigación

Los científicos e investigadores han analizado datos desde ya hace mucho tiempo, lo que ahora representa el gran reto es la escala en la que estos son generados.

Esta explosión de "grandes datos" está transformando la manera en que se conduce una investigación adquiriendo habilidades en el uso de Big Data para resolver problemas complejos relacionados con el descubrimiento científico, investigación ambiental y biomédica, educación, salud, seguridad nacional, entre otros.

De entre los proyectos que se pueden mencionar donde se ha llevado a cabo el uso de una solución de Big Data se encuentran:

- El *Language, Interaction and Computation Laboratory (CLIC)* en conjunto con la Universidad de Trento en Italia, son un grupo de investigadores cuyo interés es el estudio de la comunicación verbal y no verbal tanto con métodos computacionales como cognitivos.
- Lineberger Comprehensive Cancer Center - Bioinformatics Group utiliza Hadoop y HBase para analizar datos producidos por los investigadores de *The Cancer Genome Atlas(TCGA)* para soportar las investigaciones relacionadas con el cáncer.
- El PSG College of Technology, India, analiza múltiples secuencias de proteínas para determinar los enlaces evolutivos y predecir estructuras moleculares. La naturaleza del algoritmo y el paralelismo computacional de Hadoop mejora la velocidad y exactitud de estas secuencias.
- La *Universidad Distrital Francisco Jose de Caldas* utiliza Hadoop para apoyar su proyecto de investigación relacionado con el sistema de inteligencia territorial de la ciudad de Bogotá.
- La *Universidad de Maryland* es una de las seis universidades que colaboran en la iniciativa académica de cómputo en la nube de IBM/Google. Sus investigaciones incluyen proyectos en la lingüistica computacional (machine translation), modelado del lenguaje, bioinformática, análisis de correo electrónico y procesamiento de imágenes.

Para más referencias en el uso de Hadoop puede dirigirse a :
http://wiki.apache.org/hadoop/PoweredBy

El *Instituto de Tecnología de la Universidad de Ontario (UOIT)* junto con el Hospital de Toronto utilizan una plataforma de big data para análisis en tiempo real de IBM (*IBM InfoSphere Streams*), la cual permite monitorear bebés prematuros en las salas de neonatología para determinar cualquier cambio en la presión arterial, temperatura, alteraciones en los registros del electrocardiograma y electroencefalograma, etc., y así detectar hasta 24 horas antes aquellas condiciones que puedan ser una amenaza en la vida de los recién nacidos.

Los laboratorios *Pacific Northwest National Labs(PNNL)* utilizan de igual manera IBM InfoSphere Streams para analizar eventos de medidores de su red eléctrica y en tiempo real verificar aquellas excepciones o fallas en los componentes de la red, logrando comunicar casi de manera inmediata a los consumidores sobre el problema para ayudarlos en administrar su consumo de energía eléctrica.[3]

La esclerosis múltiple es una enfermedad del sistema nervioso que afecta al cerebro y la médula espinal. La comunidad de investigación biomédica y la *Universidad del Estado de Nueva York (SUNY)* están aplicando análisis con big data para contribuir en la progresión de la investigación, diagnóstico, tratamiento, y quizás hasta la posible cura de la esclerosis múltiple.[4]

Con la capacidad de generar toda esta información valiosa de diferentes sistemas, las empresas y los gobiernos están lidiando con el problema de analizar los datos para dos propósitos importantes: ser capaces de detectar y responder a los acontecimientos actuales de una manera oportuna, y para poder utilizar las predicciones del aprendizaje histórico. Esta situación requiere del análisis tanto de datos en movimiento (datos actuales) como de datos en reposo (datos históricos), que son representados a diferentes y enormes volúmenes, variedades y velocidades.

6. Conclusiones

La naturaleza de la información hoy es diferente a la información en el pasado. Debido a la abundacia de sensores, micrófonos, cámaras, escáneres médicos, imágenes, etc. en nuestras vidas, los datos generados a partir de estos elementos serán dentro de poco el segmento más grande de toda la información disponible.
El uso de Big Data ha ayudado a los investigadores a descubrir cosas que les podrían haber tomado años en descubrir por si mismos sin el uso de estas herramientas, debido a la velocidad del análisis, es posible que el analista de datos pueda cambiar sus ideas basándose en el resultado obtenido y retrabajar el procedimiento una y otra vez hasta encontrar el verdadero valor al que se está tratando de llegar.

Como se pudo notar en el presente artículo, implementar una solución alrededor de Big Data implica de la integración de diversos componentes y proyectos que en conjunto forman el ecosistema necesario para analizar grandes cantidades de datos.
Sin una plataforma de Big Data se necesitaría que desarrollar adicionalmente código que permita administrar cada uno de esos componentes como por ejemplo: manejo de eventos, conectividad, alta disponibilidad, seguridad, optimización y desempeño, depuración, monitoreo, administración de las aplicaciones, SQL y scripts personalizados.
IBM cuenta con una plataforma de Big Data basada en dos productos principales: IBM InfoSphere BigInsights™ e IBM InfoSphere Streams™, además de su reciente adquisición Vivisimo, los cuales están diseñados para resolver este tipo de problemas. Estas herramientas están construidas para ser ejecutadas en sistemas distribuidos a gran escala diseñados para tratar con grandes volúmenes de información, analizando tanto datos estructurados como no estructurados.

Dentro de la plataforma de IBM existen más de 100 aplicaciones de ejemplo recolectadas del trabajo que se ha realizado internamente en la empresa para casos de uso e industrias específicas. Estos aplicativos están implementados dentro de la solución de manera que las organizaciones puedan dedicar su tiempo a analizar y no a implementar.

7. Referencias

1. Cisco, **Internet será cuatro veces más grande en 2016**, Artículo Web http://www.cisco.com/web/ES/about/press/2012/2012-05-30-internet-sera-cuatro-veces-mas-grande-en-2016--informe-vini-de-cisco.html
2. Soares Sunil, **Not Your Type? Big Data Matchmaker On Five Data Types You Need To Explore Today**, Artículo Web http://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/
3. Clegg Dai, **Big Data: The Data Velocity Discussion**, Artículo Web http://thinking.netezza.com/blog/big-data-data-velocity-discussion
4. Kobielus James, **Big Data Analytics Helps Researchers Drill Deeper into Multiple Sclerosis**, Artículo Web http://thinking.netezza.com/blog/big-data-analytics-helps-researchers-drill-deeper-multiple-sclerosis
5. Aprenda más acerca de Apache Hadoop en http://hadoop.apache.org/
6. Zikopolous Paul, Deroos Dirk, Deutsch Tom, Lapis George, **Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data**, McGraw-Hill, 2012

7. Foster Kevin, Nathan Senthil, Rajan Deepak, Ballard Chuck, **IBM InfoSphere Streams: Assembling Continuous Insight in the Information Revolution**, IBM RedBooks, 2011