Big Data

Introducción

El primer cuestionamiento que posiblemente llegue a nuestra 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 ya sean estructurados, no estructurados o 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 = 109 = 1,000,000,000
Terabyte = 1012 = 1,000,000,000,000
Petabyte = 1015 = 1,000,000,000,000,000
Exabyte = 1018 = 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.

¿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.

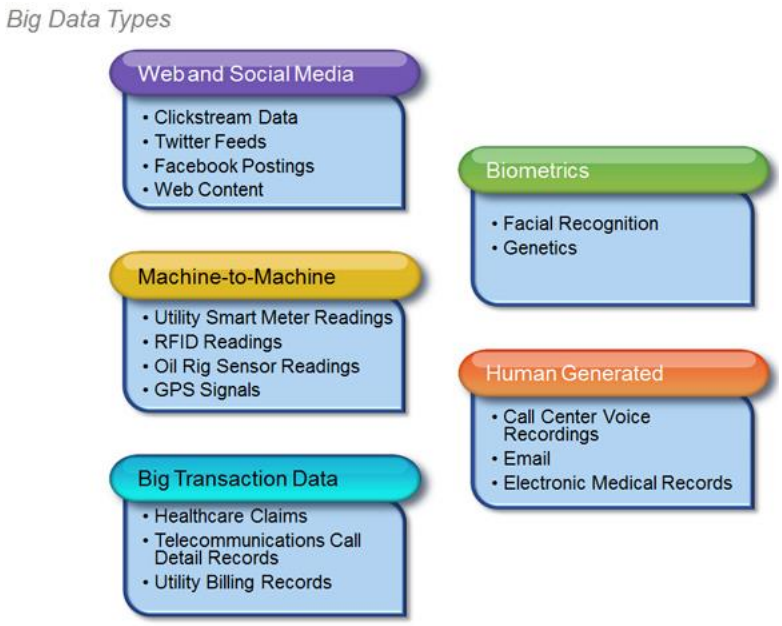
`1 quintillón = 10 30 = 1,000,000,000,000,000,000,000,000,000`

De acuerdo con un estudio realizado por Cisco, 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 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.

¿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?*

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.



Tipos de datos de Big Data

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.

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

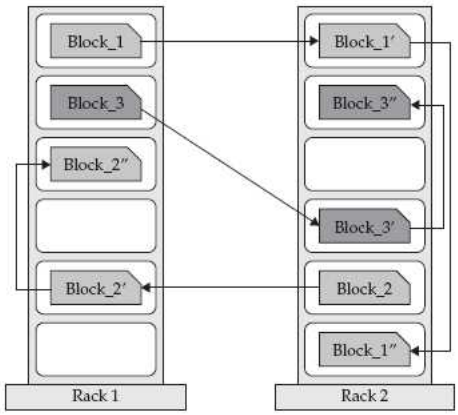
asociado. Desde luego, el ángulo correcto que actualmente tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto Hadoop.

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. 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.

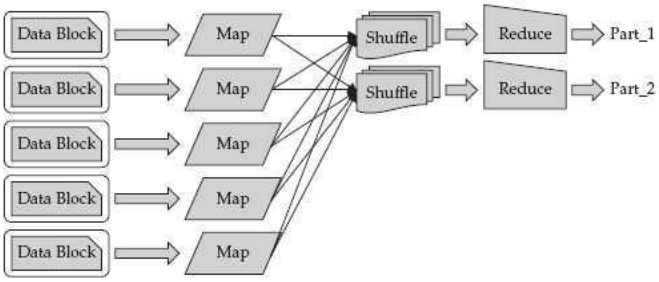
La siguiente figura ejemplifica como los bloques de datos son escritos hacia HDFS. Observe que cada bloque es almacenado tres veces y al menos un bloque se almacena en un diferente rack para lograr redundancia.



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.



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:

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.

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.

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.

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.