Herramientas y Antecedentes Big Data

24 de mayo del 2014
B. Sarmiento[1], M. Hernández[2] y X .Gómez[3]
{B. Sarmiento[1], M. Hernández[2], X .Gómez[3]}@unisimon.edu.co

Resumen- En este articulo la idea principal consiste en dar un conocimiento sobre el concepto de Big Data, el gran beneficio que tienen las empresas u organizaciones con la implementación de las diferentes plataformas que hacen parte de Big Data, entre la cual se encuentran el almacenamiento de información de gran volumen para que se puedan convertir en objetivos concretos. Analizaremos también el manejo de la información de las herramientas más utilizadas por las organizaciones en la actualidad por medio de ilustraciones, sus ventajas y desventajas desde el punto de vista del ámbito empresarial y social, como guardan la información, de donde proviene la información, los datos que explora el concepto de Big Data, También daremos a conocer la aplicación que Microsoft ha implementado para una plataforma Big Data.

Palabras clave—Big Data, Herramientas, Plataformas, bases de datos relacionales.

Abstract- In this article the main idea is to provide an understanding of the concept of Big Data, the great benefit which companies or organizations with the implementation of the different platforms that are part of Big Data, among which are the information storage large volume so that they can become specific targets. Also analyze the information management tools used by most organizations today by way of illustration, their advantages and disadvantages from the point of view of the business and social environment, as store information, hence the information, data that explores the concept of Big data, also we will present the application that Microsoft has implemented for a Big data platform.

I. INTRODUCCIÓN

Cuando escuchamos la palabra Big Data nos referimos a: procesamiento y análisis de enormes cantidades de información, muy grandes de que resulta casi imposible analizarlos con una base de datos relacional. En si la esta tendencia nos lleva a un ambiente que se nos hace muy familiar: el incremento de las páginas Web, la información de las redes sociales, aplicaciones de imágenes y videos, dispositivos móviles, sensores, etc. Capaces de procesar más de 2.5 quintillones de bytes al día [1], hasta el punto de decir que el 90% de los datos creados en el mundo se han procesados en los últimos 2 años. En si hablamos de un ámbito muy relevante desde diferentes puntos de vista, ya sea para el análisis del clima o datos geográficos, para el entorno de la salud, o por el entorno empresarial, que precisamente es donde en este entorno se utiliza más este concepto.

II. DEFINICIÓN DEL BIG DATA

Big Data es una base de datos no convencional la cual tiene como función principal analizar datos que se han vuelto tan grandes que no se pueden procesar, almacenar y analizar mediante métodos convencionales.

Una manera de caracterizar estos datos que se usan es recurriendo a lo que dicen las 3 *V* [2]en referencia a volumen, variedad y velocidad:

- Volumen: el universo digital sigue expandiendo sus fronteras y se estima que ya hemos superado la barrera del zetta byte.
- Velocidad: la velocidad a la que generamos datos es muy elevada, y la Proliferación de sensores es un buen ejemplo de ello. Además, los datos en tráfico –datos de vida efímera, pero con un alto valor para el negocio– crecen más deprisa que el resto del universo digital.
- Variedad: los datos no solo crecen sino que también cambian su patrón De crecimiento, a la vez que aumenta el contenido desestructurado.

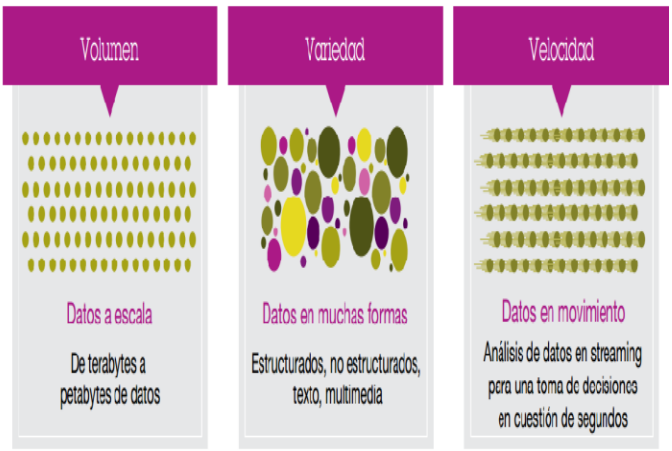

Fig. 2 Dimensión de los Datos [3]

A veces también se añade otra V, la de valor. [2]Extraer valor de toda esta información marcará la próxima década. El valor lo podremos encontrar en diferentes formas: mejoras en el rendimiento del negocio, nuevas fuentes de segmentación de clientes, automatización de decisiones tácticas, etc.

Como ya hemos visto, el origen de los datos para una empresa puede ser diverso.

Por ejemplo [2], le pueden llegar de sus propios sistemas de información de apoyo a las ventas o de interacción con sus clientes, así como estar generados por las máquinas o sensores incrustados en cualquier tipo de dispositivo o producto de la empresa. Y no olvidemos la información que circula por las redes sociales sobre una determinada empresa, que sin duda es muy valiosa para esta.

Pero hay otro origen muy importante de los datos, representado por las plataformas de información que varios

gobiernos están abriendo. Estos datos públicos pueden ser informes, mapas, estadísticas, estudios, análisis, creados y gestionados por la administración en todos los ámbitos (sanidad, economía, educación, población, etc.), que son de gran interés público. [2]

Según una encuesta realizada por IBM [3]explica que hay cierta confusión en la definición del Big Data, yaqué a los encuestados le pidieron que eligiera dos características de Big Data y no hay una características que predomine sobre el resto si no que los encuestados dividieron opiniones acerca de describir las características del concepto de Big Data.

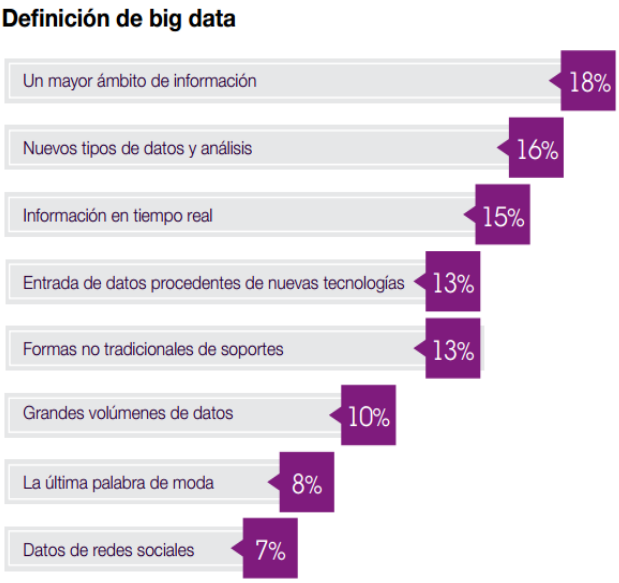

Fig. 1 Características del concepto de Big Data [3]

Hay diferentes datos de Big data. El primer dato a una cantidad mínima de datos llamados "duros" (números o hechos) descritos por Alex 'Sandy' Pentland, profesor del Instituto Tecnológico de Massachusetts, de los Estados Unidos, como "migajas digitales". [4]Se dice que son 'estructurados' porque constituyen conjuntos de datos de variables que pueden ser fácilmente etiquetados, categorizados y organizados (en columnas y filas por ejemplo) para un análisis sistemático.

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 "Smartphone", estamos hablando de que se generan alrededor de 2.5 quintillones de bytes diariamente en el mundo. [1]

III. ALGUNAS HERRAMIENTAS

Existen varias herramientas para el análisis y procesamiento de datos de mayor volumen, la cual tienen diferentes características a la hora de su ejecución, entre las más comunes podemos encontrar:

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. [1]

Hadoop MapReduce (HMR): 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).

Cassandra: Cassandra es una base de datos no relacional distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java. Su objetivo principal es la escalabilidad lineal y la disponibilidad.

La arquitectura distribuida de Cassandra está basada en una serie de nodos iguales que se comunican con un protocolo P2P con lo que la redundancia es máxima. Permite grandes volúmenes de datos en forma distribuida. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma. [1]

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. [1]

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 Map Reduce ejecutados en el cluster de Hadoop. [1]

Apache Drill: Apache Drill es una de código abierto marco de software que soporte intensivo de datos de aplicaciones distribuidas para el análisis interactivo de los conjuntos de datos de gran escala. Taladro es la versión de código abierto de Google de Dremel sistema que está disponible como un servicio de infraestructura denominado Google BigQuery. Uno se indique explícitamente objetivo de diseño es que Drill es capaz de escalar a 10.000 servidores o más y ser capaz de procesar peta bytes de datos y miles de millones de registros en cuestión de segundos. [5]

Rapid Miner: RapidMiner67 (anteriormente, YALE, Yet Another Learning Environment) es un programa informático de código abierto para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales y su licencia es la AGPL. [6]

Las bases de datos relacionales necesitan apoyo

En la actualidad, el escenario de las bases de datos ha estado marcado por lo que se conoce como SQL10 (lenguaje de consulta estructurado). En el mercado encontramos muchas

opciones consolidadas, como DB2, Oracle, MySQL, Informix, Microsoft SQL Server, PostgreSQL, etc. Estas bases de datos, que se suelen denominar relacionales, siguen las reglas ACID, hecho que les permite garantizar que un dato sea almacenado de manera inequívoca y con una relación definida sobre una estructura basada en tablas que contienen filas y columnas. Pero en el mundo del big data aparece el problema de que las bases de datos relacionales no pueden manejar el tamaño, ni la complejidad de los formatos, ni la velocidad de entrega de los datos que requieren algunas de las aplicaciones de hoy en día, por ejemplo, aplicaciones en línea con miles de usuarios concurrentes y millones de consultas al día. [2]

En el mundo de Big Data las bases de datos relacionales no manejan con eficacia el tamaño de los datos, ni la complejidad de la información, ni la velocidad de entrega de la información que requieren las aplicaciones hoy en día, como por ejemplo las aplicaciones basadas en la web que requieren de miles de usuarios concurrentes y millones de consulta por día.

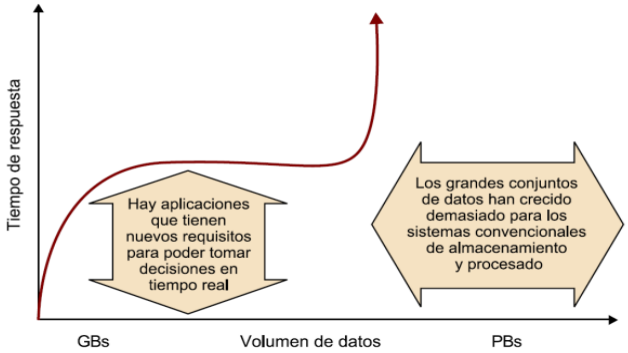

Fig. 3-comportamiento de los sistemas tradicionales de almacenamiento de la información. [2]

Por un lado, no pueden atender los requerimientos, como el de requerir un tiempo de respuesta muy bajo, que actualmente presentan algunas aplicaciones. [2]

En estos casos, hoy por hoy se están poniendo los datos en memoria (lo que se conoce por in-memory). Este tipo de aplicaciones se verán beneficiadas por las mejoras en almacenamiento que se prevén, comentadas en el apartado anterior.

Por otro lado, aparece el problema de que determinadas aplicaciones requieren un volumen de datos que ya no pueden ser almacenados y procesados usando bases de datos tradicionales. [2]

Algunas aplicaciones que se están desarrollando en la actualidad

Recientemente Microsoft presentó su plataforma para la gestión de Big Data. La principal novedad es la unión de la gestión y el análisis de datos en una única plataforma, utilizando SQL Server 2014 con módulos con capacidad analítica, como Intelligent Systems Service (IIS) y Analytics Platform System (APS). [7]

Microsoft Azure Intelligent System Service (ISS) es un nuevo servicio de Azure la cual se encarga de ayudar a los clientes a implementar nuevas soluciones de Internet of Things a través de una conexión segura y una captura de datos ordenadas, la cual es el nuevo servicio de Azure que ayuda a l independiente de que la información sea generada por maquinas, sensores, dispositivo .etc. y del sistema operativo que cuenta. La plataforma que ofrece Microsoft se basa el SQL Server 2014, que ha aumentado la velocidad de procesamiento, envío y recepción de datos, para mejorar su rendimiento. Por otro lado, se integra con Microsoft Azure para ofrecer soporte de backup y potenciar la escalabilidad desde la Nube. [7]

Creamos una transformación PDI para poder enlistar las direcciones ip: [8]

1. En la clase "hdfs_ip_list" vamos a leer un archivo tipo HDFS, por lo que agregamos el nodo "Hadoop Input File".
2. Agregamos el archivo al "Hadoop Input File"
3. Definimos el contenido del archivo, la cual la definimos de tipo csv, le desactivamos la opción de encabezado y la definimos de formato UNIX.
4. Agregamos los campos del archivo: Ordenamos las filas del archivo HDFS con el nodo "sort file", después conectamos la entrada de hadoop con ordenar filas.
5. Entramos al nodo "sort file" para editar sus propiedades, al campo client_ip sus datos los vamos ordenar de orden ascendente.
6. Agregamos el nodo "Dummy" para que sea nuestra salida del programa.

Crear la transformación HDFS

En esta clase crearemos la transformación PDI para obtener los datos para el informe HDFS.

1. Añadimos un parámetro en la transformación para poder obtenerlo en el informe.
2. Agregamos el nodo "Hadoop Input File" para cargar el archivo HDFS, ingresamos a las propiedades del nodo y cargamos nuestro archivo

   Después nos desplazamos a la opción "fields" y definimos los campos
3. Agregamos el nodo "Get Variables" y añadimos el parámetro que hemos creado anteriormente.
4. Ahora agregamos el nodo "Filters rows" ya que nosotros deseamos filtrar las filas que coincidan con la dirección ip seleccionada.
5. Agregamos el nodo "Sort Rows" para que las filas estén ordenadas por año.

   Agregamos el nodo "Row Denormaliser" para enrollar los registros de cada año en una sola fila con un campo por cada mes.

   En el campo primario seleccionamos la variable "month_num", y los campos que componen la agrupación serian "client_ip" y "year" y los campos de destino serían los meses del año.
6.
7. Agregamos 2 nodos "Dummy" para que sea la salida de nuestro programa.

El nodo output seria la salida del archivo con sus campos ordenados y el Dummy2 sería la salida falsa si los datos de las filas no se filtran.

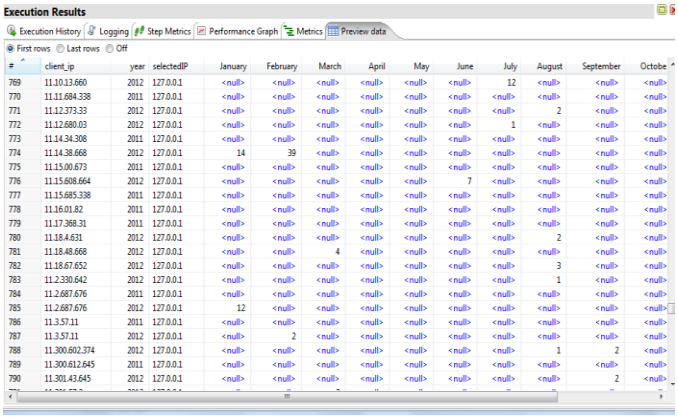

Fig 4- resultados obtenidos

Presentación de resultados

En esta ejecución nos muestra los resultados que nos arrojo nuestro programa, la cual se dividen en varias columnas, donde "client_ip" es la dirección ip del usuario, "year" es el año con la cual el usuario ingreso, "selectedIP" es la dirección por default de la cual vamos a tomar y las demás columnas son los meses del año la cual ingreso el usuario. Como podemos ver los meses del año le asignamos la variable "pageviews" para que los datos los muestres en los campos seleccionados. Los campos null son vacíos.

IV. CONCLUSIONES

En este proyecto se analizaron varias herramientas que procesan Big Data, realizando un estudio intensivo; para de tal manera poder identificar cuál es la plataforma que maneja de una manera más óptima los datos de gran volumen, analizando sus características, sus ventajas y desventajas y su uso en la inteligencia negocios. Con el estudio realizado anteriormente decidimos realizar las pruebas con la plataforma *Pentaho*, nos ayuda a organizar los datos por filas, nos permite proporcionar una consistencia total con la licencia de código abierto utilizado por *Hadoop*, además también tiene la función de automatizar procesos y crear diferentes escenarios para *MapReduce*, operaciones con archivos HDFS y scripts de *Pig.*

Luego de varias pruebas en el software *Pentaho*, se identificaron pasos para cargar datos de gran tamaño a un archivo HDFS, en este caso, además de cargarlos también se pueden agrupar, y ordenar. Para un informe óptimo, se hicieron investigaciones en la web sobre cómo manejar estos tipos de datos, además la implementación de determinados lenguajes de programación para dicho software, su interfaz gráfica de usuario, la cual es bastante sencilla e intuitiva. Una vez cargados estos datos, se puede trabajar con mira hacia al futuro agregándole diferentes herramientas que utiliza *Hadoop,* como lo son *MapReduce*, *Hive y Pig.*

V. RERENCIAS

[1] Ricardo Barranco Fragroso. (2012, junio) IBM(International Bussiness Machines). [Online]. http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

[2] Jordi Torres i Viñals, "Del cloud computing al big data," universitat oberta de catalunya , Barcelona, 2012.

[3] Michael Schroeck, Rebecca Shockley, Janet Smart, Dolores Romero, and Peter Tufano, "Analytics: el uso de big data en el mundo real," IBM Institute for Business Value, Oxford, Informe ejecutivo 2012.

[4] EDGE. (2012, agosto) Edge organization. [Online]. http://edge.org/conversation/reinventing-society-in-the-wake-of-big-data

[5] Hadoop apache. Hadoop. [Online]. http://hadoop.apache.org/

[6] apache drill foutation. (2012, septiembre) wikipedia. [Online]. http://en.wikipedia.org/wiki/Apache_Drill

[7] APPY WEEK. (2014, abril) Appy Week. [Online]. http://www.appy-geek.com/Web/ArticleWeb.aspx?regionid=8&articleid=22290097

[8] pentaho. (2012) pentaho wiki. [Online]. http://infocenter.pentaho.com/help/index.jsp?topic=%2Fpdi_user_guide%2Ftopic_hadoop.html