ideca

27.07.19

UAECD

Guía para el Análisis Espacial y no Espacial Haciendo Uso de las Herramientas Tecnologías Como Big Data.

—

Instructivo guía para el análisis espacial y no espacial haciendo uso de las herramientas tecnológicas como Big Data

www.
ideca.
gov.
co

ideca

Análisis Espacial y no Espacial Haciendo Uso de las Herramientas Tecnologías como Big Data

Instructivo guía para el análisis espacial y no espacial haciendo uso de las herramientas tecnológicas como Big Data

Fecha de creación: Marzo de 2016
Página web: www.ideca.gov.co
Correo electrónico: ideca@catastro.gov.co

Autores: Unidad Administrativa Especial de Catastro Distrital - Gerencia IDECA

UAECD

Contenido

Objetivo y Alcance

1.1 Objetivo

El presente instructivo tiene por objeto presentar una orientación para el uso de técnicas y tecnologías que permitan almacenar, procesar y analizar datos e información estructurada, no estructurada y semiestructurada de carácter espacial y no-espacial que posibiliten el desarrollo de tecnologías interoperables para el mantenimiento de datos geoespaciales.

1.2 Alcance

El presente instructivo pretende abarcar a todas las entidades del orden Distrital que forman parte de IDECA, y como complemento al desarrollo de las mismas actividades en la materia dentro de IDE de Bogotá, buscando orientar el desarrollo de iniciativas que promuevan el uso de técnicas y tecnologías para almacenar, procesar y analizar datos e información estructurada, no estructurada y semiestructurada de carácter espacial y no-espacial.

Definiciones, Siglas y Abreviaturas

A

Agnóstico

Es una disciplina que se basa en experiencias y observaciones, entonces, todo aquello que no puede ser experimentado u observado de manera directa será declarado imposible e inaccesible[i].

Apache Avro

Es un sistema de serialización de datos. En los proyectos en Hadoop, suele haber grandes cuantidades de datos, así que la serialización se usa para procesar y almacenar estos datos, de forma que el rendimiento en tiempo sea efectivo. Esta serialización puede ser en texto en plano, JSON, en formato binario[ii].

Apache CouchDB

Es una base de datos NoSQL, que carece de un esquema o de estructuras de datos predefinidas como las tablas en las bases de datos relacionales, la información almacenada son documentos JSON, la estructura de los datos o documentos puede acomodarse a las necesidades de cambio o a la evolución del software. CouchDB es una base de datos que abarca completamente la red, utiliza documentos en JSON para guardar los datos, permite acceder a los datos desde un navegador web a través del protocolo http, permite realizar operaciones utilizando JavaScript[iii].

Apache Flume

Es un sistema distribuido para capturar de forma eficiente, agregar y mover grandes cuantidades de datos log de diferentes orígenes (diferentes servidores) a un repositorio central, simplificando el proceso de recolectar estos datos para almacenarlos en Hadoop y poder analizarlos[iv].

Apache Pig

Inicialmente desarrollado por Yahoo, permite a los usuarios de Hadoop centrarse más en el análisis de los datos y menos en la creación de programas MapReduce. Proporciona un lenguaje procedural de alto nivel

y su lenguaje de programación Pig está pensado para poder trabajar en cualquier tipo de datos[v].

Apache Sqoop

Apache Sqoop traduce "Sql-to-Hadoop", es una herramienta diseñada para transferir de forma eficiente bulk data entre Hadoop y sistemas de almacenamiento con datos estructurados, como bases de datos relacionales. Algunas de sus características son: permite importar tablas individuales o bases de datos enteras a HDFS, generar clases Java que permiten interactuar con los datos importados, además de permitir importar de las bases de datos SQL a Hive[vi].

Apache ZooKeeper

Es un proyecto de Apache que proporciona una infraestructura centralizada y de servicios que permiten la sincronización del Clúster. ZooKeeper mantiene objetos comunes que se necesiten en grandes entornos de Clúster[vii].

B

Base de datos

Una base de datos o banco de datos (en inglés: database) es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso[viii].

Base de datos NoSQL

Representan una evolución en la arquitectura de aplicación del negocio, están diseñadas para proveer el almacenamiento de datos confiables, escalables y disponibles a través de un conjunto de sistemas configurables que funcionan como nodos de almacenamiento[ix].

Big Data

Es la tendencia en el avance de la tecnología hacia un nuevo enfoque de entendimiento y toma de decisiones, 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[x].

C

Chukwa Es un sistema de captura de datos y framework de análisis que trabaja con Hadoop para procesar y analizar grandes volúmenes de logs. Incluye herramientas para mostrar y monitorizar los datos capturados[xi].

Clúster Según Strauch[xii], es otro enfoque para la partición de datos que se esfuerza por la transparencia hacia los clientes que deberían manejar un grupo de servidores de bases de datos en lugar de un único servidor.

Customer Centricity Es una estrategia cuyo objetivo primordial es alinear la conceptualización, desarrollo y comercialización de los productos y servicios de una marca, con las necesidades y deseos de sus clientes más valiosos. Esta estrategia tiene un fin muy específico: maximizar los beneficios de la marca a largo plazo[xiii].

D

DBaaS Es un servicio emergente de Cloud Computing, en el cual el Cliente usará el servicio de un Sistema de Bases de Datos sin la necesidad de preocuparse por los elementos necesarios, tanto de hardware como de software, para que dicho servicio funcione correctamente[xiv].

G

Grafo Conjunto, no vacío, de objetos llamados vértices o nodos y una selección de pares de vértices llamados aristas.

H

Hadoop Es un marco de desarrollo de código abierto que permite el procesamiento de grandes conjuntos de datos, de manera distribuida a través de un grupo o clúster de computadoras, usando un modelo de programación sencillo[xv].

Hbase Es el sistema de almacenamiento no relacional para Hadoop. Es una base de datos de código abierto, distribuido y escalable para el

almacenamiento de Big Data. Está escrita en Java e implementa el concepto de Bigtable desarrollado por Google[xvi].

HDFS Es el sistema de almacenamiento, es un sistema de ficheros distribuido, creado a partir del Google File System (GFS). El HDFS se encuentra optimizado para grandes flujos y trabajar con ficheros grandes en sus lecturas y escrituras. Su diseño reduce la E/S en la red. La escalabilidad y disponibilidad son otras de sus claves, gracias a la replicación de los datos y tolerancia a los fallos[xvii].

Hive Es un sistema de Data Warehouse para Hadoop que facilita el uso de la agregación de los datos, ad-hoc queries, y el análisis de grandes datasets almacenados en Hadoop[xviii].

HTML HTML, sigla en inglés de HyperText Markup Language (lenguaje de marcas de hipertexto), hace referencia al lenguaje de marcado para la elaboración de páginas web. Es un estándar que sirve de referencia del software que conecta con la elaboración de páginas web en sus diferentes versiones, define una estructura básica y un código (denominado código HTML) para la definición de contenido de una página web, como texto, imágenes, videos, juegos, entre otros[xix].

HTTP Hypertext Transfer Protocol - Protocolo de transferencia de hipertexto, es el protocolo de comunicación que permite las transferencias de información en la World Wide Web[xx].

I

Instructivo Documento que detalla la forma de llevar a cabo una generalidad o una actividad de un proceso o un procedimiento[1].

[1] 03-01-PR-01 Procedimiento Administración Documental – UAECD.

J

JAVA Es un lenguaje de programación de propósito general, concurrente, orientado a objetos que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible[xxi].

JSON Es un formato ligero de intercambio de datos, basado en un subconjunto del lenguaje de programación JavaScript[xxii].

L

Lenguaje SQL Es el idioma para bases de datos relacionales. Se trata de una consulta declarativa, SQL es estandarizada por el Instituto Americano de Estándares Nacionales (ANSI) y Organización Internacional de Normalización (ISO) a partir de 1986 y tiene varias revisiones desde entonces[xxiii].

Linked Data Los Datos Enlazados es la forma que tiene la Web Semántica de vincular los distintos datos que están distribuidos en la Web, de forma que se referencian de la misma forma que lo hacen los enlaces de las páginas web[xxiv].

M

Machine-to-Machine (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[xxv].

Mahout Es un proyecto para crear aprendizaje automático y data mining usando Hadoop. Mahout puede ayudar a descubrir patrones en grandes datasets. Tiene algoritmos de recomendación, clustering y clasificación[xxvi].

MapReduce Es el core de hadoop, es el paradigma de programación que permite escalabilidad a través de cientos y miles de servidores en un clúster hadoop[xxvii].

MongoDB Es un sistema de base de datos multiplataforma orientado a documentos, de esquema libre, esto significa que cada entrada o registro puede tener un esquema de datos diferentes, con atributos o "columnas" que no tienen por qué repetirse de un registro a otro[xxviii].

N

Nube La computación en la nube, conocida también como servicios en la nube, (del inglés cloud computing), es un paradigma que permite ofrecer servicios de computación a través de una red, que usualmente es Internet[xxix].

P

PHP Lenguaje de programación gratuita y multiplataforma, se ejecuta en el servidor web justo antes de enviar la página web a través de internet al cliente.

Protocolo En redes, un protocolo de comunicaciones o protocolo de red es la especificación de una serie de reglas para un tipo particular de comunicación. La red Internet se basa en el modelo de referencia TCP/IP (Transmission Control Protocol/Internet Protocol) que toma su nombre de los dos principales protocolos que regulan la comunicación a través de esta red[xxx].

T

TCP/IP La familia de protocolos de Internet es un conjunto de protocolos de red en los que se basa Internet y que permiten la transmisión de datos entre computadoras. En ocasiones se le denomina conjunto de protocolos TCP/IP, en referencia a los dos protocolos más importantes que la componen, que fueron de los primeros en definirse, y que son los dos más

	utilizados de la familia: TCP (Transmission Control Protocol), Protocolo de Control de Transmisión, e, IP (Internet Protocol), Protocolo de Internet[xxxi].
W	
World Wide Web Consortium	World Wide Web Consortium (W3C), es un consorcio internacional que genera recomendaciones y estándares que aseguran el crecimiento de la World Wide Web a largo plazo[xxxii].

Generalidades

El análisis de información en grandes volúmenes, de diversas fuentes, y a gran velocidad, de todos los datos disponibles con los que cuenta una organización, se está convirtiendo en el contexto de Big Data, como una tendencia que surge para generar valor de negocio con los datos.

Teniendo en cuenta lo anterior, se puede observar que la implementación de tecnologías de Big Data adquiere mayor importancia en las organizaciones, dado la necesidad de procesar y analizar grandes cantidades información para la toma de decisiones.

Para comprender esta nueva tendencia de la Big Data y conducir su aplicabilidad en los dominios de conocimiento georreferenciado, con el contenido de los datos e información del Distrito Capital, el presente instructivo busca dar claridad en algunos conceptos que rodean el desarrollo de este tipo de iniciativas.

3.1 ¿Qué es Big Data?

Big Data surgió fruto de la necesidad ciertas empresas (Yahoo! y Google) para tratar de resolver sus problemas empresariales y explorar nuevas oportunidades de negocio aplicando el marco de trabajo Hadoop.

Aunque no existe unanimidad en la definición de Big Data, existe cierto consenso en la fuerza disruptiva que suponen los grandes volúmenes de datos y la necesidad de su captura, almacenamiento y análisis.

Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad con el objeto de extraer valor económico de ellos.

Así entonces, se define Big Data Según Gartner[xxxiii], como una referencia a aquellos sistemas de información que manejan conjuntos de datos de gran volumen, de alta velocidad, de veracidad, de valor y de gran variedad de recursos, que demandan formas rentables e innovadoras de procesamiento de la información para mejorar la comprensión y la toma de decisiones.

De esto se puede obtener beneficios como[xxxiv]:

- ✓ Optimizar el cálculo y la precisión algorítmica para reunir, analizar, enlazar y comparar conjuntos de grandes datos.
- ✓ Identificar patrones para la toma de decisiones en los ámbitos económico, social, técnico y legal.

En este esfuerzo de afrontar un proyecto Big Data se da atención a los tres principales retos, como es la **Variedad**, la **Velocidad** y el **Volumen** de información, para luego hoy día atender a la **Variabilidad**, la **Veracidad**, la **Visualización** y el **Valor** que aportan esos datos a la organización, tal cual como se detallan a continuación[xxxv]:

- ✓ **Variedad:** representa todos los tipos de datos, y supone un desplazamiento fundamental en el análisis de los requisitos desde los datos estructurados tradicionales hasta la inclusión de los datos en bruto (sin procesar), semiestructurados y no estructurados como parte del proceso fundamental de la toma de decisiones.
- ✓ **Velocidad:** la idea de velocidad no se asocia a la tarea de crecimiento de los depósitos o almacenes de datos, sino que se aplica la definición al concepto de los datos en movimiento, es decir, la velocidad a la cual fluyen los datos junto con la frecuencia de las actualizaciones de las grandes bases de datos.
- ✓ **Volumen:** las empresas amasan grandes volúmenes de datos, desde terabytes hasta petabytes; luego hacia exabytes de información y se espera que para el 2020 entremos en la era del zettabyte, a manera de referencia: solo Twitter genera más de 9 terabytes (TB) de datos cada día, Facebook, 10 TB. Así pues, el volumen de datos disponibles en las organizaciones hoy día está en ascenso mientras que el porcentaje de datos que se analiza está en disminución.
- ✓ **Variabilidad:** La variabilidad se refiere a los datos cuyo significado está en constante cambio. Este es particularmente el caso cuando la recolección de datos se basa en el procesamiento del lenguaje.

- ✓ **Veracidad:** la veracidad o fiabilidad de Big Data supone un gran reto a medida que la variedad y las fuentes de datos crecen, donde el actuar con esta información veraz permitirá la correcta toma de decisiones.

- ✓ **Visualización:** Desde el modelo tradicional de gestión de los datos se comenzó heredando el formato de informe o CUBE. Ahora con la gran cantidad de datos masivos que son necesarios para realizar una muestra, una vez que se ha procesado, es necesario disponer de una manera de representar la información de un modo más accesible y fácil de leer, aquí es donde entra en juego el concepto de la Visualización. Las visualizaciones pueden contener decenas de variables y parámetros, muy lejos de la variables x e y de la barra estándar de coordenadas. Encontrar una manera de presentar esta información que haga ver los resultados de una manera clara es uno de los desafíos de Big Data.

- ✓ **Valor:** supone para las organizaciones obtener información de los grandes datos de una manera rentable y eficiente, como es el caso del software Hadoop, que procesa grandes volúmenes de datos a través de un clúster de centenares, o incluso millares de computadores de un modo muy económico.

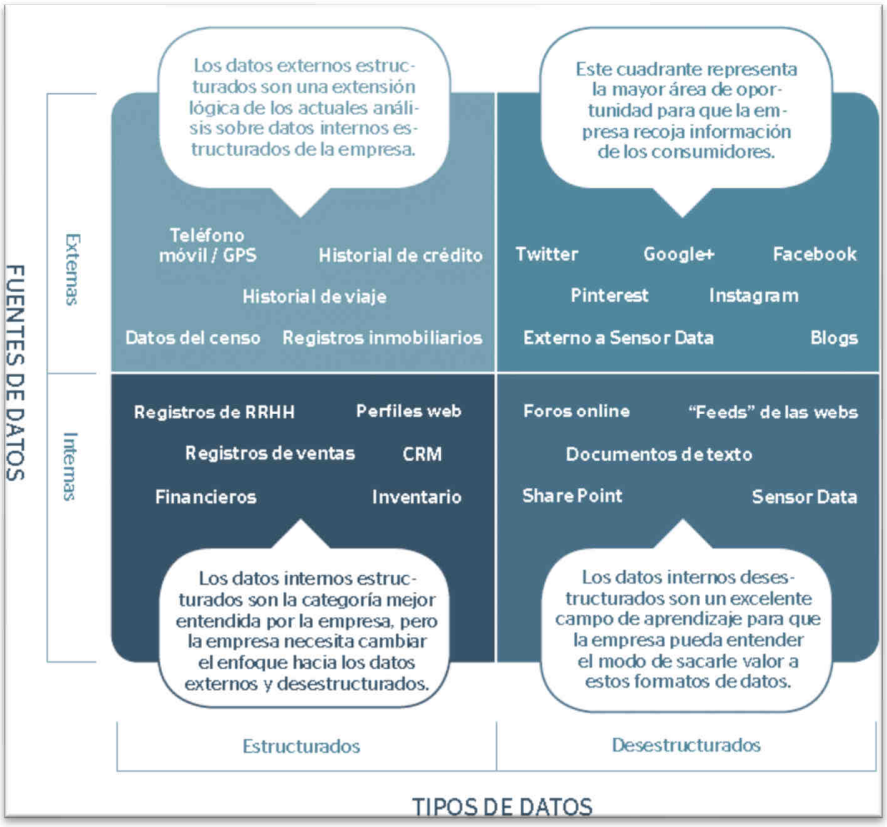
Las nuevas herramientas de manipulación de Big Data han originado nuevas categorías dentro de los tipos de datos no estructurados como son datos semiestructurados y datos no estructurados propiamente dicho, frente a los datos estructurados (datos tradicionales que almacenan datos estructurados en las bases de datos relacionales).

- ✓ Datos estructurados: datos con formato o esquema fijo que poseen campos fijos. Son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos, fundamentalmente.

- ✓ Datos semiestructurados: datos que no tienen formato fijo, pero contienen etiquetas y otros marcadores que permiten separar los elementos datos, ejemplos típicos son el texto de etiquetas de XML y HTML.

- ✓ Datos no estructurados: datos que no tienen campos fijos: audio, video, fotografías, o formatos de texto libre como correos electrónicos, mensajes instantáneos SMS, artículos, libros, mensajes de mensajería instantánea tipo WhatsApp, Viber, etc.

G1 Gráfico 1 Fuentes y tipos de datos Big Data [xxxvi]
Fuente: Booz & Company. Benefitting from Big Data, 2012.



La integración de estos datos facilita a cualquier organización la combinación de los Big Data con los datos transaccionales tradicionales para generar valor y conseguir la mayor eficacia posible, lo cual supone una gran oportunidad de negocio para organizaciones y empresas.

3.2 ¿Cómo ha sido la evolución de Big Data?

El surgimiento de Big Data se remonta al nacimiento de las primeras herramientas informáticas que llegaron en 1940. En esa misma década comenzaron a aparecer programas que eran capaces de predecir posibles escenarios futuros. Por ejemplo, el equipo del Proyecto Manhattan (1944)

que realizaba simulaciones por ordenador para predecir el comportamiento de una reacción nuclear en cadena[xxxvi].

Según Artaza[xxxvii], no fue hasta la década de los 70 en la que se popularizó el análisis de datos. En 1978 se crea Black-Scholes, un modelo matemático que permitía predecir el precio de acciones futuras, pero con la llegada de Google en 1998 y el desarrollo de algoritmos para mejorar las búsquedas en la web, es cuando se produce realmente el estallido de Big Data.

Ahora, con la entrada del nuevo siglo, este concepto se acuña y recoge todo el significado que se le otorga en la actualidad. Según los analistas en la materia, hoy en día se generan 2,5 trillones de bytes relaciones con el Big Data. Además, surge la demanda de aquellos perfiles profesionales que sean capaces de gestionar herramientas de análisis.

3.3 ¿Cuál es la arquitectura de Big Data?

Los Big Data han generado el advenimiento de nuevos tipos de datos y tecnologías emergentes tales como Hadoop, NoSQL, "en-memoria" o analítica de Big Data. Para aprovechar las ventajas de estos desarrollos, las organizaciones necesitan crear una arquitectura de referencia que integre estas tecnologías emergentes en las infraestructuras existentes. Esta arquitectura de referencia de Big Data consta de dos componentes fundamentales: arquitectura y gobierno de Big Data, que debe integrarse con las infraestructuras existentes y coexistir con las acciones del gobierno de los datos tradicionales.

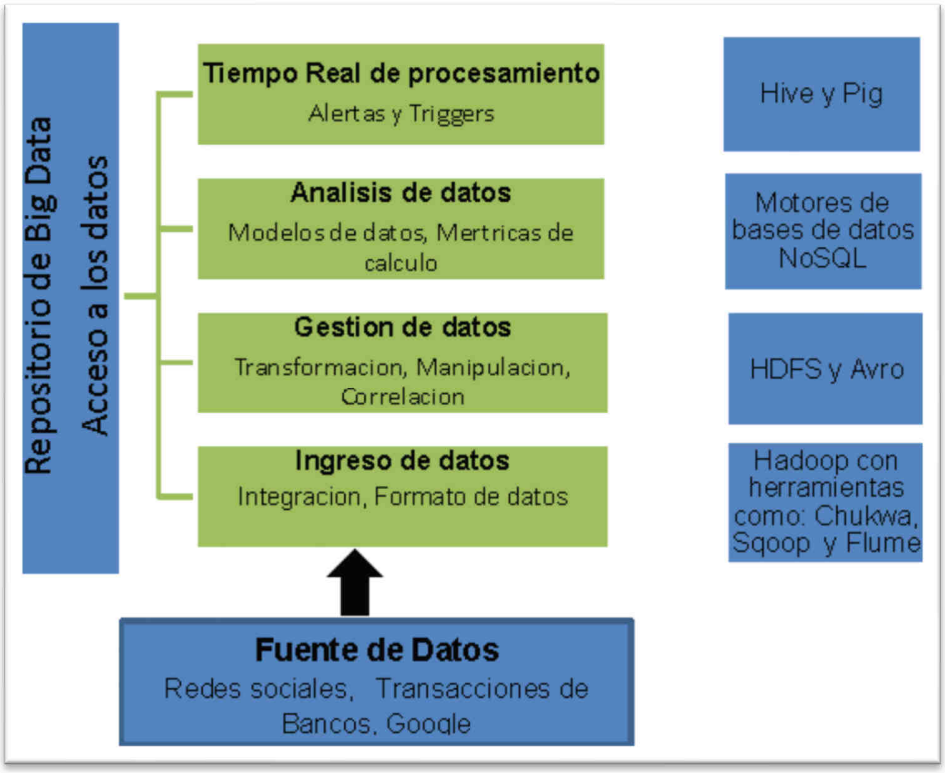
Por lo cual, la gestión y procesamiento de Big Data es un problema abierto y vigente que puede ser manejado con el diseño de una arquitectura de 5 niveles, la cual está basada en el análisis de la información y en el proceso que realizan los datos para el desarrollo normal de las transacciones. A continuación, se pueden ver los niveles que contienen un ambiente Big Data y la forma en que se relacionan e interactúan entre ellos:

Ingreso de datos: el Ingreso de datos es el procedimiento de obtener e importar información para su posterior uso o almacenamiento en una base de datos. Consiste en coleccionar datos de muchas fuentes con el objetivo de realizar un análisis basado en modelos de programación[xxxviii].

Gestión de datos: la administración de datos es el desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos con el fin de gestionar las necesidades del ciclo de vida de información de una empresa de una manera eficaz. Es un enfoque para administrar el flujo de

datos de un sistema a través de su ciclo de vida, desde su creación hasta el momento en que sean eliminados. La administración de Big data es la forma en que se organizan y gestionan grandes cantidades de datos, tanto de información estructurada como no estructurada para desarrollar estrategias con el fin de ayudar con los conjuntos de datos que crecen rápidamente, donde se ven involucrados terabytes y hasta peta bytes de información con variedad de tipos[xxxix].

G2 Gráfico 2 Arquitectura de un ambiente de Big Data [xxxix]
Fuente: Defining the Big Data Architecture Framework (BDAF



Análisis de datos: es el proceso de examinar grandes cantidades de datos para descubrir patrones ocultos, correlaciones desconocidas y otra información útil[xl]. Esta información puede proporcionar ventajas competitivas y resultar en beneficios para el negocio, como el marketing para generar mayores ingresos.

Según Rouse[xli], el objetivo principal del análisis de datos es ayudar a las empresas a tomar mejores decisiones de negocios al permitir a los científicos y otros usuarios de la información analizar grandes volúmenes de datos transaccionales, así como otras fuentes de datos que puedan haber quedado sin explotar por la inteligencia del negocio convencional.

Tiempo real de procesamiento: según Halim[xlii], es un proceso que automatiza e incorpora el flujo de datos en la toma de decisiones, este aprovecha el movimiento de los datos para acceder a la información estática y así lograr responder preguntas a través de análisis dinámicos. Los sistemas de procesamiento de flujo se han construido con un modelo centrado que funciona con datos estructurados tradicionales, así como en aplicaciones no estructuradas, como vídeo e imágenes. Así, el procesamiento de flujos es adecuado para aplicaciones que tienen tres características: calcular la intensidad (alta proporción de operaciones de E/S), permitir paralelismo de datos y por último la capacidad de aplicar los datos que se introducen de forma continua.

Dentro de esta arquitectura en un ambiente Big Data se pueden utilizar diferentes herramientas, donde cada una de estas cumple un papel importante para la implementación. A continuación se realiza una descripción de cada una de las tecnologías Big Data de código abierto incluidas en esta arquitectura.

Apache Hadoop[2]: es un marco de desarrollo de código abierto que permite el procesamiento de grandes conjuntos de datos, de manera distribuida a través de un grupo o clúster de máquinas informáticas, usando un modelo de programación sencillo.

MapReduce[3]: es un modelo de programación para el procesamiento de grandes conjuntos de datos. MapReduce es usado para hacer computación distribuida sobre clúster de servidores.

Storm: es un sistema de computación distribuida en tiempo real, libre y de código abierto, nacido con el Twitter. Storm hace fácil procesar de manera fiable flujos no estructurados de datos, desarrollándose en el ámbito del procesamiento en tiempo real, lo que hizo de Hadoop para el procesamiento por lotes[4].

[2] http://hadoop.apache.org/
[3] http://research.google.com/archive/mapreduce.html
[4] http://storm-project.net/

HBase[xliii]: es el sistema de almacenamiento no relacional para Hadoop. HBase es una base de datos de código abierto, distribuido y escalable para el almacenamiento de Big Data. Está escrita en Java e implementa el concepto de Bigtable desarrollado por Google. Así como Bigtable aprovecha el almacenamiento de datos distribuidos proporcionado por el sistema de archivos de Google, Apache HBase Bigtable proporciona capacidades similares sobre Hadoop y HDFS[5].

Hive[6]: es un Sistema Data Warehouse para Hadoop que facilita resúmenes de datos, consultas ad-hoc, y el análisis de grandes conjuntos de datos almacenados en los sistemas de archivos compatibles con Hadoop[xliv].

Pig[7]: fue desarrollado inicialmente en Yahoo! para permitir a los usuarios de Hadoop centrarse más en el análisis de grandes conjuntos de datos y dedicar menos tiempo a tener que escribir programas mapper y reducer. El lenguaje de programación Pig está diseñado para manejar cualquier tipo de datos, de ahí el nombre. Pig está formado por dos componentes: el primero es el lenguaje en sí mismo, que se llama PigLatin, y el segundo es un entorno de ejecución donde los programas PigLatin se ejecutan[xlv].

R[8]: es el lenguaje de programación líder en el mundo para el análisis estadístico y la realización de gráficos. R, además de ser un lenguaje para la minería de datos es un entorno de programación[xlvi].

Sqoop[9]: es una aplicación con interfaz de línea de comandos para la transferencia de datos entre bases de datos relacionales y Hadoop.

Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido, escalable y portátil escrito en Java para el framework Hadoop.

[5] http://hbase.apache.org/
[6] http://hive.apache.org/
[7] http://pig.apache.org/
[8] http://www.r-project.org/
[9] http://sqoop.apache.org/

3.4 ¿Qué son base de datos NoSQL?

Como se ha mencionado los tipos de Big Data proceden de numerosas fuentes de datos: datos tradicionales de empresas, datos generados por máquinas (M2M) y de Internet de las cosas, datos sociales, datos de biometría y genética, datos personales o generados por las personas.

Estos volúmenes de datos se vienen almacenando, normalmente, en los almacenes de datos de las empresas (EDW – Enterprise Data Warehouse) principalmente para el procesamiento de grandes conjuntos de datos, y los almacenes de datos especiales (data marts) subconjuntos o conjuntos especializados de Data Warehouse, y se almacenan en base de datos relacionales.

Por otra parte, aparecerán nuevos almacenes de datos para tratar los grandes volúmenes que conformarán las bases de datos NoSQL y "en-memoria", su tratamiento requerirá el uso de herramientas ETL (Extraction, Transformation, Load) que preparen los datos procedentes de las fuentes de datos y los guarden en los almacenes de datos.

Así, en lo que respecta a base de datos existen múltiples categorías:

- ✓ SQL, base de datos relacionales tradicionales.
- ✓ NoSQL, Not only SQL - base de datos no relacional, distribuida, de alto rendimiento y altamente escalable.
- ✓ "In-memory", base de datos que realiza todo su procesamiento en memoria principal.
- ✓ Base de datos heredadas (Legacy).

Adicionalmente, a estas base de datos, se destacan aquellas que funcionan en la nube y que están dando a la tendencia DBaaS (Database as a Service), como son: Amazon RDS, DynamoDB, SimpleDB, PostgreSQL, Xeround (MySQL), Microsoft SQL Azure Database (SQL Server), Google App Engine (NoSQL), Salesforce Database.com (Oracle), ClearDB (MySQL), Cloudant (CouchDB).

Sin embargo, las bases de datos analíticas son las más utilizadas en la actualidad, estas son: las bases de datos NoSQL y las bases de datos "en-memoria".

El término NoSQL empieza aparecer en los 90´s y su primer uso se da en el 2009 por Eric Vans[xlvii], con el objeto de dar una solución a las problemáticas planteadas anteriormente, dando una

posibilidad de abordar la forma de gestionar la información de una manera distinta a como se venía realizando.

Así, las bases de datos NoSQL (Not only SQL) son una categoría de sistemas de gestión de bases de datos que no utilizan SQL como lenguaje de consulta principal. Estas bases de datos no requieren esquemas de tablas fijas, y no soportan operaciones Join. Están optimizadas para operaciones de lectura/escritura escalables en lugar de pura consistencia.

Las bases de datos NoSQL son la siguiente generación de bases de datos que tiene las siguientes características[xlviii]:

- ✓ **Distribuido:** Sistemas de bases de datos NoSQL a menudo distribuidos donde varias máquinas cooperan en grupos para ofrecer a los clientes datos. Cada pieza de los datos se replica normalmente durante varias máquinas para la redundancia y alta disponibilidad.
- ✓ **Escalabilidad horizontal:** a menudo se pueden añadir nodos de forma dinámica, sin ningún tiempo de inactividad, lo que los efectos lineales de almacenamiento logran capacidades de procesamiento general.
- ✓ **Construido para grandes volúmenes:** muchos sistemas NoSQL fueron construidos para ser capaz de almacenar y procesar enormes cantidades de datos de forma rápida.
- ✓ **Modelos de datos no relacionales:** los modelos de datos varían, pero en general, no son relacional. Por lo general, permiten estructuras más complejas y no son tan rígida que el modelo relacional.
- ✓ **No hay definiciones de esquema:** la estructura de los datos generalmente no se define a través de esquemas explícitos. En su lugar, los clientes almacenan datos como deseen, sin tener que cumplir con algunas estructuras predefinidas.

Las bases de datos NoSQL se clasifican en cuatro grandes categorías, de acuerdo con la taxonomía propuesta por Scofield y Popescu[xlix]:

- ✓ **Orientadas a Clave-Valor[l]:** Key-Value, son bases de datos más simples en cuanto su uso, dado que simplemente almacena valores identificados por una clave. Normalmente, el valor guardado se almacena como un arreglo de bytes (BLOB). De esta forma el tipo de

contenido no es importante para la base de datos, solo interesa la clave y el valor que tiene asociado. Aplicaciones de este tipo son Cassandra y DynamoDB.

- ✓ **Orientadas a documentos[li]:** según Camacho, son como un almacén Key-Value, a diferencia que la información no se guarda en binario, sino como un formato que la base de datos pueda leer, como XML, permitiendo realizar consultas avanzadas sobre los datos almacenados. Aplicaciones de este tipo son MongoDB y CouchDB.

- ✓ **Orientada a grafos[lii]:** estas bases de datos manejan la información en forma de grafo, dando una mayor importancia a la relación que tiene los datos. Con esto se obtiene consultas de forma óptima comparada con un modelo relacional. Aplicaciones de este tipo son Neo4J, InfiniteGraph, AllegroGraph, OpenLink, HyperGraphDB.

- ✓ **Orientada a columnas[liii]:** BigTable (tabla o columnas), estas bases de datos guardan la información en columnas en lugar de renglones, logrando una mayor velocidad en realizar la consulta. Esta solución es conveniente en ambientes donde se presenten muchas lecturas como en Data Warehouses y Sistemas de Business Intelligence. Aplicaciones de este tipo son Apache HBase, HyperTable y Cassandra.

Estas cuatro grandes categorías, almacenan todas las ventajas de las bases de datos no relacionales, permiten la escalabilidad y la rápida analítica que necesitan las actuales aplicaciones de grandes datos. Los grandes precursores del movimiento NoSQL son Google BigTable y Amazon Dynamo, ambas de código abierto.

Por su parte, las bases de datos “en-memoria” se basan en las tecnologías de computación “en memoria”, que es una tecnología que permite el procesamiento de cantidades masivas de datos en memoria principal para proporcionar resultados inmediatos de las transacciones y del análisis.

Para conseguir el rendimiento deseado, la computación en-memoria se apoya en tres conceptos fundamentales:

- ✓ Mantenimiento de los datos en memoria para aumentar la velocidad de acceso a los datos.

- ✓ Minimizar el movimiento de los datos para potenciar el concepto de almacenamiento en columna, comprensión y ejecución de cálculos al nivel de base de datos.

- ✓ Aprovecha la arquitectura multinúcleo de los modernos procesadores y de los servidores multiprocesador, mediante técnicas distribuidas que proporcionan mejores resultados que un único servidor.

Los tres grandes proveedores de tecnología que ofrecen plataformas de bases de datos in-memory son:

- ✓ SAP con su herramienta HANA,
- ✓ Oracle con sus herramientas Exadata,
- ✓ y Exalytics, Microsoft. Aunque IBM, HP, EMC ofrecen soluciones hardware-software para estas bases de datos en memoria.

En la sección de **Anexos**, se puede encontrar un cuadro comparativo sobre algunos motores de bases de datos NoSQL, que se pueden utilizar para la construcción de un ambiente Big Data, teniendo en cuenta su taxonomía; esto a manera de orientación para el desarrollo de una futura iniciativa con Big Data.

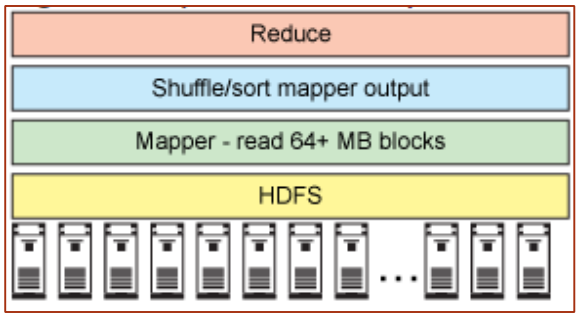
3.5 ¿En qué consiste el marco de trabajo Hadoop?

Para la manipulación de los datos Big Data surge el marco de trabajo Hadoop, cuyo origen se remonta con publicaciones en la materia hechas por Google. Se describen en ellos técnicas para la indexación de información en la Web, la distribución en miles de nodos, y la presentación al usuario como un conjunto significativo.

De tal forma que, Hadoop es un proyecto de software de código abierto de Apache para obtener valor de volumen/velocidad/variedad, increíbles de datos acerca de una organización. Es una herramienta que se integra con un ecosistema existente de Inteligencia de Negocios.

Hadoop tiene dos componentes centrales, el almacenamiento de archivos llamado Hadoop Distributed File System (HDFS), y la infraestructura de programación llamada MapReduce, como se ilustra en la gráfica 3[liv]:

G3 Gráfico 3 Arquitectura Hadoop
Fuente: ¿Cuáles son los componentes de Hadoop? – Portal IBM



Hadoop, al contrario que los sistemas tradicionales, está diseñado para explorar a través de grandes conjuntos de datos y producir sus resultados, mediante un sistema de procesamiento distribuido por lotes (batch). Así, el marco de trabajo Hadoop comprende tres componentes básicos:

- ✓ Hadoop Distributed File System (HDFS), como sistema de archivos distribuido que proporciona acceso high-througput a datos de aplicaciones.
- ✓ Hadoop MapReduce. Como sistema de procesamiento masivamente paralelo de conjuntos de grandes datos.
- ✓ Hadoop Common, utilidades típicas que soportan a los restantes módulos de Hadoop.

A continuación se presentan brevemente algunas herramientas fundamentales para la programación de Hadoop[iv]:

- ✓ **HDFS:** como se ha mencionado, es el sistema de ficheros distribuido utilizado por Hadoop. Las dos ideas principales de HDFS es por un lado que sea un sistema de ficheros que facilite una alta escalabilidad tolerante a fallos. Por otro lado Hadoop necesita que los problemas que se estén intentando solucionar involucren un gran número de datos. Así, HDFS debe garantizar un alto rendimiento de datos para que Hadoop sea capaz de procesar.
- ✓ **Avro:** es un sistema de serialización de datos; contiene: estructuras de datos, formato de datos binario, un archivo contenedor para almacenar datos persistentes, llamada a procedimiento remoto (RPC), y es de fácil integración con lenguajes dinámicos.

- ✓ **Chukwa:** es un sistema de recopilación de datos de código abierto para el seguimiento de grandes sistemas distribuidos. Se construye en la parte superior del sistema de archivos distribuido Hadoop (HDFS) y Map/Reduce. Chukwa también incluye un conjunto de herramientas flexibles para la visualización, seguimiento y análisis de resultados de los datos recogidos.

- ✓ **Sqoop:** es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales. Sqoop importa los datos de un sistema de gestión de bases de datos relacionales (RDBMS) como MySQL u Oracle al sistema de archivos distribuido Hadoop (HDFS), donde transforma los datos y luego los exporta de nuevo a un RDBMS.

- ✓ **Flume:** es un servicio distribuido, confiable y disponible para recolectar, agregar y mover grandes cantidades de datos de registro eficientemente. Cuenta con una arquitectura simple y flexible basada en transmisión de flujos de datos. Es robusto y tolerante a fallos con los mecanismos de fiabilidad, conmutación por error y los mecanismos de recuperación.

- ✓ **Hive:** es la infraestructura de almacenamiento de datos construida sobre Apache Hadoop para proporcionar el resumen de datos, consultas ad-hoc y análisis de grandes conjuntos de datos. Proporciona un mecanismo para proyectar en la estructura de los datos en Hadoop y consultar los datos utilizando un lenguaje similar a SQL llamado HiveQL (HQL). Hive facilita la integración entre Hadoop y herramientas para la inteligencia de negocios y la visualización. Hive permite al usuario explorar y estructurar los datos, analizarlos y luego convertirla en conocimiento del negocio.

- ✓ **Pig:** es una plataforma para el análisis de grandes conjuntos de información que consiste en un lenguaje de alto nivel para la expresión de los programas de análisis de datos, junto con la infraestructura necesaria para la evaluación de estos programas. La propiedad más importante es que su estructura es susceptible de paralelismo, que a su vez les permite manejar grandes conjuntos de datos.

- ✓ **HBase:** almacenamiento de valor de clave escalable. Funciona similarmente a un hash-map persistente. No es una base de datos relacional pese al nombre HBase.

- ✓ **Zookeeper:** usado para gestionar sincronización por clúster.

- ✓ **Mahout:** aprendizaje de máquina para Hadoop. Usado para análisis predictivos y otros análisis avanzados.

3.6 ¿En qué consiste el modelo de programación MapReduce?

MapReduce[lvi] es una técnica para la indexación de información en la Web, el cual consiste en dividir el proceso de manipulación de los grandes datos en dos tareas (mapper y reducer), esto para manipular los datos distribuidos a nodos de un cluster, logrando un alto paralelismo en el procesamiento. La entrada a este modelo de programación MapReduce es un conjunto de pares clave/valor y la salida es otro conjunto de pares clave/valor.

- ✓ **Función Map:** a partir del conjunto de pares clave/valor de entrada se genera un conjunto de datos intermedios. La función Map asocia claves idénticas al mismo grupo de datos intermedios. Cada grupo de datos intermedios estará formado por una clave y un conjunto de valores, por lo tanto, estos datos intermedios van a ser a su vez la entrada de la función de Reduce.

- ✓ **Función Reduce**: la fase de Reduce se encargará de manipular y combinar los datos provenientes de la fase anterior para producir a su vez un resultado formado por otro conjunto de claves/valores.

3.7 ¿Qué es HDFS como sistema de ficheros distribuido utilizado por Hadoop?

Como se ha mencionado anteriormente, HDFS (Hadoop Distributed File System) es un sistema de almacenamiento y ficheros distribuido. Creado a partir del Google File System (GFS). HDFS se encuentra optimizado para grandes flujos y trabajar con ficheros grandes en sus lecturas y escrituras. Su diseño reduce la E/S en la red. La escalabilidad y disponibilidad son las características relevantes, gracias a la replicación de los datos y tolerancia a los fallos. Los siguientes son los componentes de un clúster HDFS[lvii]:

- ✓ **NameNode:** regula el acceso a los ficheros por parte de los clientes. Mantiene en memoria la metadata del sistema de ficheros y ejerce control de los bloques de fichero que tiene cada DataNode.

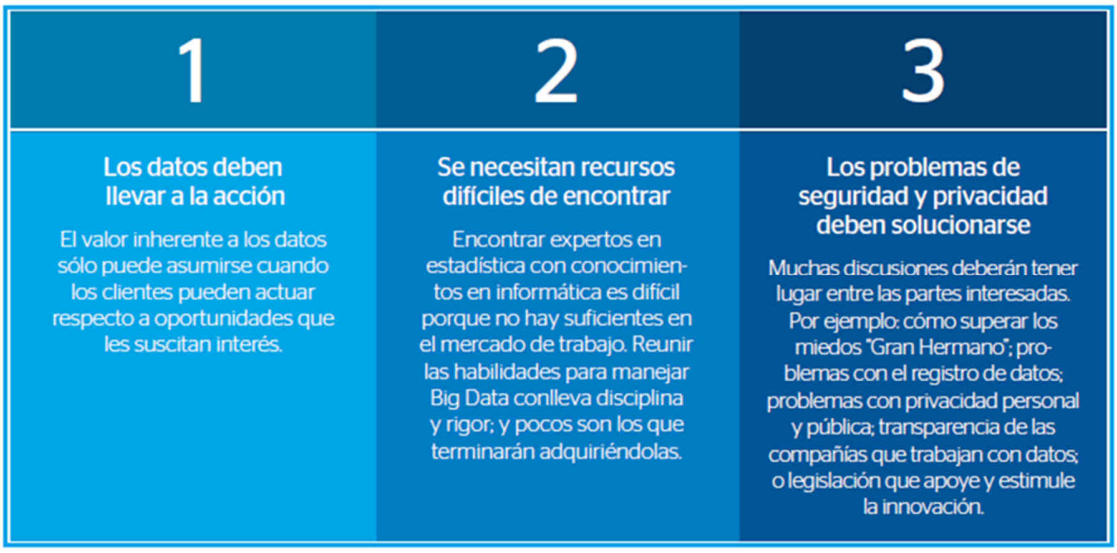
- ✓ **DataNode:** son los responsables de leer y escribir las peticiones de los clientes. Los ficheros están formados por bloques, estos se encuentran replicados en diferentes nodos.

3.8 ¿Cuáles son los aspectos claves hacia la creación de valor?

A partir de la definición de Big Data y analizando los resultados obtenidos en proyectos de implementación, existe un gran potencial en las organizaciones para explotar el valor de los datos en relación con el gran volumen de datos de hoy en día, a los nuevos tipos de datos y análisis, y a la necesidad de más análisis de información en tiempo real para la toma de decisiones oportuna.

Debido a lo anterior, se han planteado en el medio que para la creación de valor se deben tomar en consideración los siguientes aspectos:

G4 Gráfico 4 El camino hacia la creación de valor con tres aspectos claves [lx]
Fuente: The Financial Brand

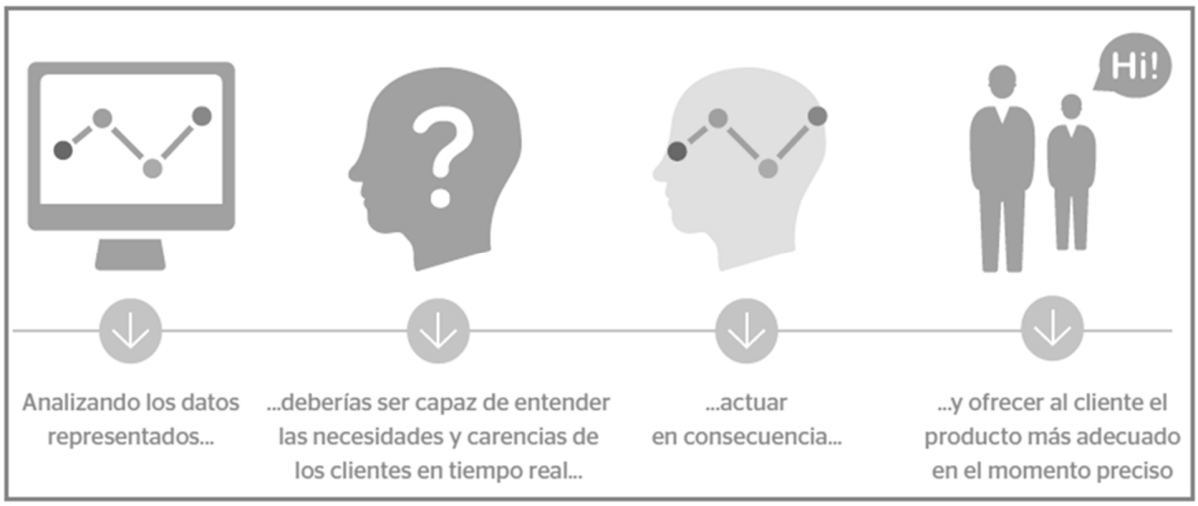


Frente a la coyuntura sobre qué datos generan mayor rendimiento habría que analizar los datos desde una perspectiva de la visualización del consumidor. Sin embargo, el valor de dichos datos no se puede concretar a no ser que proporcionen una oportunidad para el actuar. En este proceso se debe analizar que las organizaciones deban ofrecer a los consumidores el producto o servicio

adecuado en el momento adecuado; en otras palabras, necesitan entender las necesidades y deseos actuales de los clientes y tener capacidad para actuar en consecuencia.

Por este motivo, la rápida comprensión de los diversos flujos de datos y de la subsiguiente información extraída con Big Data son procesos críticos que deben orientar a generar valor añadido a los productos o servicios dentro de una organización. En la siguiente figura, se presenta este flujo de interrogantes obligados a responder que conduzcan a la creación de valor.

G5 *Gráfico 5 Flujo de pasos para la creación de valor lxi*
Fuente: Centro de Innovación BBVA. Big Data.



Así, la estrategia de Big Data debe contar con la capacidad de colisión entre la imaginación y la tecnología, donde tanto las organizaciones y profesionales deberían de ponerse como objetivo adquirir capacidades para el análisis de flujos de datos en tiempo real mediante fuentes multiestructuradas y con herramientas para grandes volúmenes de datos; sin embargo, desglosando cada una de las siguientes consideraciones se puede alcanzar esta estrategia:

- ✓ **No se embarque en tareas demasiado ambiciosas,** es decir, prioriza las inversiones en tecnología.
- ✓ **Desarrolla una hoja de ruta,** buscando asesoramiento sobre cuáles son las mejores tecnologías en las que debe invertir en función de las actuales estrategias empresariales e inversiones.

- ✓ **Encuentra el valor desde dentro,** es decir, audita y potencia información que ya existe en las fuentes de datos corporativas; el entender los activos de datos ya existentes puede ayudar impulsar casos de uso de Big Data más optimizados.
- ✓ **Sé un líder en la revolución social,** buscando datos en nuevas fuentes, yendo más allá de las tradicionales fuentes de datos estructuradas.
- ✓ **Promueve un centro de competencia,** creando un grupo de interés que promueva la colaboración, la comunicación abierta y la alineación de negocios y tecnología.
- ✓ **La gestión del cambio es crucial,** asegurando que se usan métodos y procedimientos estandarizados para minimizar el impacto en la organización.
- ✓ **Gestión del riesgo,** incluyendo en los proyectos a analistas de datos con un enfoque empresarial y asegurando de que tienen el apoyo de TI y de los responsables de los datos en la empresa "Oficial de Seguridad de los Datos", para que ayuden a alinear las necesidades del negocio con las iniciativas de Big Data.

3.9 ¿Cómo las empresas se pueden beneficiar de Big Data?

En este camino que recorren las organizaciones para capturar el valor que los datos encierran, tanto para mejorar los procesos de negocio actuales, como para crear nuevos productos basados en datos, no se debe perder el foco de buscar nuevas perspectivas en el uso de estos datos y combinarlos con otros de forma que tanto la organización como otras instituciones, empresas o personas puedan tomar mejores decisiones.

El beneficio en el uso de Big Data se debe revisar y evaluar en varios aspectos dentro de los siguientes dos ámbitos:

Internamente, usando los datos para beneficio de la organización:

- ✓ Mejora de la experiencia de cliente utilizando los productos o servicios.
- ✓ Mejora de la eficiencia de los procesos de negocio dentro de la organización.
- ✓ Análisis de riesgos, permitiendo llegar a obtener una visión más amplia de otras organizaciones productoras o consumidoras de información.

- ✓ Adecuación de la oferta de productos en función de las necesidades reales de la comunidad o "customer centricity".

Externamente, de forma que sean otras organizaciones las que se beneficien del valor de los datos:

- ✓ Ayudar a miembros de la organización a entender mejor su rendimiento, sus clientes y su contexto geográfico y temporal.
- ✓ Ayudar a los gestores urbanos a tomar mejores decisiones gracias a un mejor conocimiento de la ciudad.
- ✓ Medir el impacto real[10] de eventos o de decisiones concretas.
- ✓ Permitir que terceros creen nuevos servicios de valor sobre datos anónimos y agregados proporcionados por la organización, en combinación o no con otras fuentes de datos.

3.10 ¿Cuáles son los modelos de negocio emergentes en el contexto de Big Data?

El Big Data está generando grandes oportunidades en pro de la efectividad operativa, acentuado en:

- ✓ Análisis de datos operativos aprovechando abundantes datos de producción para mejorar los procesos y la calidad del producto.
- ✓ Mejora de la planificación y predicción aprovechando la cantidad de datos de procesos históricos, recursos y productos.

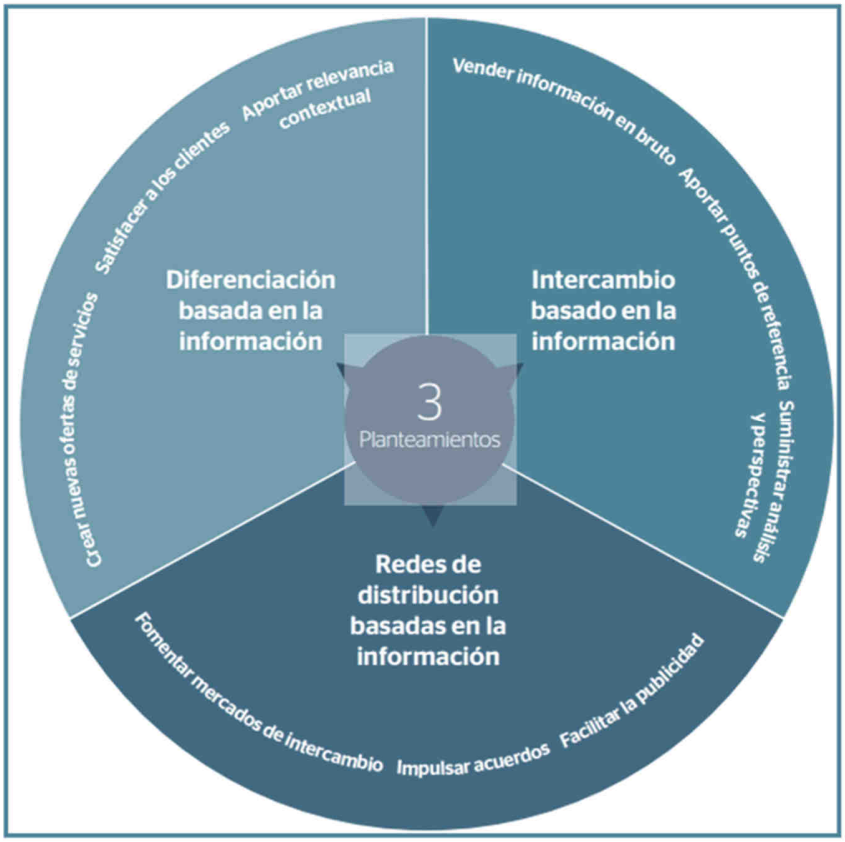
Por otra parte, genera importantes oportunidades a partir del análisis de la actividad del consumidor, relacionado con:

[10] www.mwcimpact.com

- ✓ Análisis de la actividad del cliente, es decir, almacenando las preferencias del cliente para personalizar lo que se muestra, monitorizar el uso para evaluar las métricas de la web.
- ✓ Monitorizar los medios sociales, analizando los sentimientos del consumidor hacia la marca y sus productos en redes sociales.

Lo anterior, plantea nuevos modelos de negocio en el mundo de Big Data, destacando en general tres grandes planteamientos, como se muestra en el siguiente gráfico.

G6 Gráfico 6 Modelos de negocio emergentes en Big Data
Fuente: Centro de Innovación BBVA. Big Data



En resumen, estos grandes planteamientos van orientados a:

- ✓ Fomentar mercados de intercambio e impulsar acuerdos
- ✓ Suministrar análisis y perspectivas
- ✓ Vender información en bruto
- ✓ Satisfacer a los clientes
- ✓ y crear nuevas ofertas de servicios.

3.11 ¿Cuáles son los retos en la gestión empresarial con el potencial de Big Data?

Las necesidades de Big Data están creando nuevas expectativas en las organizaciones, demandando nuevas responsabilidades y la necesidad de afrontar retos y oportunidades que se ofrecerán alrededor del análisis de los grandes volúmenes de datos así como la necesidad de pensar y contratar los nuevos roles y perfiles de trabajo, entre ellos el científico de datos.

Otro aspecto relevante cuando se toma la decisión de implantar las tecnologías de Big Data en una organización, se debe pensar es por el diseño de la correspondiente estrategia, la integración de Big Data en la empresa misma, y estudiar la presencia de los grandes datos en la empresa y cómo medirlos atendiendo a sus características fundamentales como: volumen, velocidad, variedad y valor; es por ello que a continuación se presentan cinco (5) retos que se deben afrontar en la gestión empresarial para hacer frente a este potencial de Big Data:

- ✓ **Liderazgo:** se deben fijar objetivos claros y definir qué métricas son valiosas para conocer mejor el cliente, progresar y mejorar los resultados del negocio.
- ✓ **Gestión del talento:** emergen el nuevo rol de analistas de Big Data, en especial los científicos de datos, que sean capaces de analizar la información y destacar en ella el valor añadido que darán a las organizaciones.
- ✓ **Tecnologías:** surgen herramientas para gestionar las características del modelo de las 3V de la información que se genera a diario (volumen, velocidad, variedad), donde gran parte de este software utilizado es de código abierto, basado principalmente en el marco de Hadoop y en base de datos NoSQL e in-memory. Además, nuevas tecnologías basadas en

la nube (cloud computing) que permiten ofrecer maneras muy eficientes en costo de escalar (extender) la capacidad de almacenamiento y procesamiento demandado por Big Data.

- ✓ **Toma de decisiones:** para conseguir que la información facilite la toma de decisiones es necesario **construir modelos predictivos** que **optimicen los resultados de negocio**, **mejorar las operaciones**, **la experiencia de cliente y la estrategia**.
- ✓ **Cultura corporativa:** el uso de los Big Data en las organizaciones precisa de un nuevo cambio organizacional que requerirá: desarrollar analíticas relevantes que muestren con sencillez la evolución del negocio, crear herramientas de analíticas sencillas de utilizar por parte del personal de la empresa, y desarrollar las capacidades necesarias para obtener el máximo rendimiento de Big Data.

Así pues, el nuevo rol del científico de dato se constituye en una mezcla de analista, científico (físico, matemático, estadístico, biólogo) e ingeniero de sistemas, y tiene entre sus capacidades fundamentales, la inclinación a marcar o detectar tendencias, sacando conclusiones a partir de grupos de datos no categorizados, recopilados por las empresas.

Por otra parte, aunque el concepto de inteligencia de negocios se concibe como una colección de tecnologías y sistemas de información que soportan la toma de decisiones empresariales, proporcionando información de operaciones internas y externas; son aplicaciones que consideran cómo analizar los datos del usuario, cómo se presentan los resultados de sus análisis y cómo los gerentes y ejecutivos implementan estos resultados; esta inteligencia de negocios se deberá adaptar a los Big Data de modo que las herramientas de reporting, consultas, visualización, y analítica de datos deberán permitir el tratamiento e integración de todo tipo de datos, estructurados y no estructurados.

3.12 ¿Cuál es el Impacto con la Convergencia Entre Nube, Big Data Y Datos Abiertos?

La convergencia entre Hadoop, Big Data y Analitica de Datos registran tendencias en alza junto con la nube (cloud), lo social (social media y social business), y la movilidad (tecnologías, dispositivos y redes).

Según Gartner[lviii], esta convergencia e interdependencia llamado "Nexo de las fuerzas" se proyectan para transformar el comportamiento de los usuarios, creando nuevos modelos de negocio.

Con ello, las organizaciones comienzan a dar cuenta del valor y del poder que ofrecen los datos, sopesando las limitaciones presupuestarias con los grandes beneficios en estas nuevas tecnologías haciendo uso de las numerosas infraestructuras, de software propietario y de software abierto que el mercado ofrece.

Otro efecto que se da con esta convergencia es el cambio en la plataforma tecnológica, ahora denominada "Tercera Plataforma", donde la primera plataforma estaba relacionada con el mainframe, y la segunda plataforma giraba en torno a la llegada de la PC y la arquitectura cliente/servidor, pero esta tercera plataforma está sustentada en:

- ✓ Dispositivos y aplicaciones móviles, la computación móvil (movilidad)
- ✓ Se consolidarán los servicios en la nube (cloud computing)
- ✓ Redes de banda ancha con análisis de grandes cantidades de información (Big Data)
- ✓ Tecnologías de plataformas sociales en la Web, donde el poder de los "social media" no estará sólo enfocado en el mercado de consumo y el consumidor final, sino en la importancia de crear una plataforma de conversación con clientes, proveedores y empleados, sustentado en la adopción de soluciones analíticas, soluciones que le permitan a los clientes visualizar los datos que tienen y que seguirán captando de una manera que haga sentido para la toma de decisiones.

También, con esta convergencia tendrá un mayor protagonismo la "analítica predictiva", en una amplia variedad de funciones empresariales, con énfasis en ventas y marketing, intensificando

nuevos canales de marketing y de negocios, que se pueden emplear para dirigirse de manera más eficaz a los clientes existentes, alcanzar nuevos mercados y coordinar esfuerzos. Con ello, las organizaciones que exploten las posibilidades de análisis predictivo serán más competitivas y podrán predecir mejor los productos y servicios que desean sus clientes, cuáles deben ser sus acciones online y offline, y qué han de hacer para mantenerlos fieles.

Por último, lo anterior junto a la incorporación de funcionalidades de análisis de sentimientos en sus plataformas, facilitarán a las organizaciones la capacidad para extraer información como parte de las actividades de análisis existentes.

Instrucción

Hasta el momento se han dado las generalidades para profundizar en los conceptos entorno al uso de técnicas y tecnologías como Big Data, sin embargo, en esta sección se presenta una serie de pasos que orientan un proceso de construcción de soluciones bajo los principios de Big Data, y con ello buscar examinar y descubrir el valor en los datos, en el sentido de ofrecer valor añadido a los productos que una organización ofrezca a la comunidad.

Paso 1. Identifique los datos que pueden añadir valor a la organización y qué se puede inferir con ellos

Como buena práctica, antes de tomar la decisión de abordar iniciativas aplicando tecnologías Big Data, se hace necesario realizar una identificación al interior de la organización respecto a cuáles serán los objetivos que se buscan alcanzar con la implementación de dichas iniciativas e inferir el grado de beneficio aprovechando la información disponible para servir mejor a los ciudadanos.

Hay que considerar que el enfoque de dichos **beneficios** con el Big Data redunda en la **mejora de la experiencia de cliente** y en la **mejora de la eficiencia de los procesos de negocio** dentro de la organización.

Este proceso de identificación de objetivos y planificación redunda en examinar cuáles son los datos que produce la organización, cuál es la data que se genera en la organización o qué datos son los que recolecta y cuáles son accesibles.

Adicionalmente, con este paso se busca realizar un inventario para conocer qué datos son los que posee y los que realmente necesita la organización, identificando los datos bien sea operacionales, comerciales, financieros, públicos o **los datos provenientes de redes sociales**, los cuales se pueden explorar para obtener nuevas formas de valor.

En este proceso de identificación importante considerar aspectos relevantes al uso de los datos que se dará de manera interna (dentro de la organización) y aspectos con una visión externa, de forma que sean otras organizaciones las que se beneficien del valor de los datos.

Así entonces, en este primer paso, se debe asegurar el recolectar los datos correctos, es decir, aquellos que realmente le proporcionarán valor a la organización; cuidando de recoger todos los datos que pueda para resolver qué hacer con ellos más adelante, ya que demasiados datos podrían poner en riesgo el éxito del proyecto o volverlo demasiado costoso. En este primer paso, también se puede empezar a identificar el tratamiento (Planificación) que se desea dar para el almacenamiento y análisis de los datos recolectados, que muy probablemente no son estructurados (tales como videos, audio o datos provenientes de las redes sociales), cuando aplique.

Paso 2. Analice cómo transformar el negocio a partir de los datos identificados

Este es el paso donde a partir de la creatividad se busca imaginar cómo el negocio puede cambiar a partir de los datos identificados en el paso anterior. Esta transformación del negocio es uno de los aspectos más demandantes en esfuerzo intelectual dado que reinventarse no es tarea fácil.

Así, desde este paso se puede mirar qué está haciendo la competencia y comparar cuáles son las estrategias adoptadas por otras organizaciones, en aras de obtener ideas para nuestro proceso de transformación. Una buena práctica para apoyar este proceso puede ser el aplicar el modelo de pensamiento lateral de Edward de Bono[lix], empleada como una técnica para la resolución de problemas sobre cómo, usando datos de manera imaginativa y creativa, podría cambiar o crear un negocio, modelo que le permita preguntarse, qué puede hacer de diferente e imaginar más allá de sus límites.

Paso 3. Identifique las tecnologías, herramientas de software y requerimientos de hardware necesarios para la implementación de un ambiente de Big Data

Con este paso se analizan cuales herramientas y plataformas dispone o requeriría la organización para manipular los datos estructurados como los datos no estructurados, que se requieren para el montaje, configuración e integración para la construcción de iniciativas en el contexto de Big Data.

En esta identificación se debe tener presente:

- ✓ **Datos estructurados:** datos con formato o esquema fijo que poseen campos fijos. Son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos, fundamentalmente.
- ✓ **Datos semiestructurados:** datos que no tienen formato fijo, pero contienen etiquetas y otros marcadores que permiten separar los elementos de los datos, ejemplos típicos son el texto de etiquetas de XML y HTML.
- ✓ **Datos no estructurados:** datos que no tienen campos fijos: audio, video, fotografías, o formatos de texto libre como correos electrónicos, mensajes instantáneos SMS, artículos, libros, mensajes de mensajería instantánea tipo WhatsApp, Viber, etc.

Así, la integración de estos datos facilita a cualquier organización la combinación de los Big Data con los datos transaccionales tradicionales para generar valor y conseguir la mayor eficacia posible, lo cual supone una gran oportunidad de negocio para organizaciones y empresas.

Como punto de referencia para identificar estas tecnologías a continuación se describen algunos de los proveedores tecnológicos que suministran herramientas para uso como Big Data[lx]:

- ✓ **Software de gestión de terceros:** software para gestionar clústeres Hadoop, cuyos productos son generalmente agnósticos en cuanto a las distribuciones a las que dan soporte.

  - Apache Mesos
  - Corona
  - Stack IQ
  - WANdisco
  - Zettaset

- ✓ **Distribución:** son productos de software empaquetados que pretenden facilitar la implementación y gestión de clústeres Hadoop, frente a tener que descargar las diversas bases de código Apache e intentar concebir un sistema.

  - Cloudera
  - EMC Greenplum
  - Hadoop
  - Hortonworks
  - IBM
  - Intel
  - MapR

- ✓ **Bases de datos operacionales:** son importantes para muchas de las aplicaciones web. Donde sí se está desarrollando grandes negocios en internet, hallar una que se ajuste a los volúmenes de datos y que rinda según las necesidades es vital.

  - Apache Accumulo
  - Apache HBase
  - Drawn to Scale
  - Lily
  - Splice Machine
  - Sqrrl
  - TempoDB

- ✓ **SQL en Hadoop:** Las soluciones SQL en Hadoop incrementan la accesibilidad de Hadoop y permiten a las organizaciones reutilizar la inversión en el aprendizaje de SQL.

  - Apache Drill
  - Apache Giraph
  - Citus Data
  - Hadapt
  - Impala (Cloudera)
  - Lingual (Cascading)
  - Phoenix (Force.com)
  - Pivotal HD (Greenplum)
  - RainStor
  - The Hive
  - The Stinger Initiative (Hortonworks)

- ✓ **Frameworks**: con los frameworks los desarrolladores y científicos de datos pueden sacarle rendimiento a Hadoop de una manera barata y fácil. Los frameworks permiten la expansión de las fuentes y almacenes de datos a los que da apoyo.

  - Apache Hama Project
  - Apache Pig
  - Apache Tez
  - Cascading (Concurrent)
  - Mortar
  - Scalding (Twitter)

- ✓ **Hadoop - como servicio de apps/analíticas:** El modelo de nube deja a los usuarios sacar provecho de la inversión en infraestructura y de la experiencia en materia de un proveedor de servicios sin tener que realizarlo internamente.

  - Birst
  - Cetas (VMWare)
  - Kontagent
  - Packetloop
  - Qubole
  - Treasure Data

- ✓ **Hadoop - infraestructura como servicio:** orientado a ofrecer algo como un Dropbox para Big Data BI (Business Intelligence). Permite reducir el ruido que se encuentra en las infraestructuras a la hora de implementar Hadoop.

    - Amazon Elastic MapReduce
    - GoGrid
    - Infochimps
    - Infosphere BigInsights (IBM)
    - Joyent
    - Mortar Data
    - Skytap
    - Sungard
    - VertiCloud (Beta)
    - Windows Azure (Microsoft)

- ✓ **Aplicaciones analíticas & Plataformas:** tecnologías con tendencia hacia una plataforma más unificada de análisis de Big Data, como es la introducción de consultas en tiempo real; con esto, Hadoop ha dado un gran paso hacia la unificación de la mayoría de las aplicaciones analíticas de Big Data en una plataforma integral.

    - 0xdata
    - Apache Mahout
    - Continuuity
    - Datameer
    - HStreaming
    - Karmasphere
    - NGData
    - PacketPig (Packetloop)
    - Platfora
    - Radoop
    - Tresata
    - WibiData

Sin embargo, existen otros proveedores de tecnología destacados como:

- ✓ **Alternativas HDFS:** como sistema de archivos, el HDFS (Hadoop Distributed File System), es un sistema de archivos distribuido que proporciona acceso high-througput a datos de aplicaciones.

    - Cassandra (via DataStax Enterprise)
    - Ceph
    - Cleversafe (Dispersed Storage Network)
    - EMC (Isilon)
    - IBM (GPFS)
    - NetApp (NetApp Open Solution for Hadoop)
    - Lustre
    - Red Hat (Red Hat Storage/GlusterFS)
    - Quancast File System

- ✓ **Plataformas alternativas:**

    - Disco
    - HPCC Systems
    - Pervasive Software (DataRush)
    - Spark/Shark

- ✓ Hadoop reenvasado:

    - Data Direct Networks
    - Dell
    - HP
    - Microsoft
    - Nutanix
    - SGI
    - Teradata/Aster Data

Paso 4. Análisis, integración e interpretación de los datos

Con las herramientas identificadas el siguiente paso consiste en modelar y analizar los datos para responder preguntas sobre la organización ayudando a encontrar la forma de dar valor a los datos para crear nuevos modelos de negocio.

En el proceso de integración se pretende ofrecer un enriquecimiento de datos mediante su integración coherente, para dar mayor contexto y significado. Para ello se deben responder las siguientes preguntas:

- ✓ ¿Cómo puede la nueva tecnología para el procesamiento de Big Data, utilizar todos los datos y la tecnología disponible?
- ✓ ¿Cómo puede la tecnología y los datos existentes ser mejorados mediante la adición de grandes volúmenes de datos?
- ✓ ¿Cómo pueden las nuevas formas de análisis y aplicaciones usar tanto lo viejo como lo nuevo?

Finalmente, por integración consiste en ofrecer alguna propuesta de unas pautas para facilitar dicha gestión de los datos.

Paso 5. Establezca la mejora de la calidad de los datos

Con la mejora de la calidad de los datos se pretender es definir las políticas de procedencia de los datos, normalizar terminología y asegurar la interoperabilidad, mejorando la integración con recursos Web externos.

La calidad de los datos debe considerar en su mejora el adoptar herramientas de uso Big Data más económicas y prácticas para manejar los grandes volúmenes de tráfico de datos, a la vez que mantengan una alta calidad con la experiencia del usuario.

Paso 6. Seguridad y privacidad de Big Data

Ante el creciente desarrollo de las tecnologías de Big Data se plantea multitud de interrogantes en cuestiones de seguridad, privacidad y temas relacionados, como los problemas de conservación de los datos, usos delictivos, propiedad intelectual, propiedades medioambientales producidos por los centros de datos, protección del anonimato, libertad

de expresión, sumados a otros derechos de los usuarios; es por ello que será preciso estar atentos a la publicación de normativas, directivas, leyes de las agencias de seguridad y protección de datos y privacidad, del Gobierno en el orden distrital, nacional; y también, a las correspondientes normativas internacionales de la Unión Europea, ONU, Unesco y foros de la materia a nivel internacional.

Así entonces, con este paso, la seguridad y privacidad de Big Data busca formular leyes de privacidad de los datos, y leyes para la protección de los datos personales de los entes que aportan información a la organización, más cuando la propiedad de los datos no es clara y las estructuras de gobierno no son efectivas.

Paso 7. Identifique y aplique herramientas de analítica y reporting de Big Data para la mejora del negocio

En esta etapa, se deben revisar tres enfoques para establecer el análisis (Analytics) de información para los grandes volúmenes, de diversas fuentes, a gran velocidad y con una flexibilidad sin precedente para generar valor de negocio con los datos dentro de una organización, como se describe a continuación:

Analítica de Datos

El análisis de datos Big Data es el proceso de examinar, a una gran velocidad, grandes volúmenes de datos, con tamaños desde terabytes hasta petabytes, de una amplia variedad de tipos y de gran valor para descubrir patrones ocultos, correlaciones desconocidas y otra información útil, de modo que los resultados del análisis puedan proporcionar ventajas competitivas a las organizaciones en relación con la competencia y producir beneficios para el negocio, tales como un marketing más efectivo y eficaz, y mayores ingresos.

Así entonces, con esta identificación de herramientas en este enfoque, implica los procesos y actividades diseñados para obtener y evaluar datos para extraer información útil, es decir, examinar datos en bruto (sin ningún procesamiento previo) con el propósito de obtener conclusiones acerca de la información contenida en ellos.

Dentro de las técnicas más utilizadas en analítica de datos son: consultas e informes (quering y reporting), visualización, minería de datos, análisis de datos predictivos, lógica difusa, optimización, streaming de audio, video o fotografía, etc.

Por otra parte, considerar en su selección, que estas herramientas de analítica deben permitir a los usuarios:

- ✓ Analizar los grandes datos de un modo rápido y económico
- ✓ Los usuarios deben ser capaces de explorar y visualizar datos masivos mediante gráficos interactivos, cuadros de mando integral (balanced scorecards), tableros de control (dashboards), herramientas de reporting y query (informes y consultas) de resultados, así como herramientas de visualización, en tiempo real cuando sea necesario.

Los proveedores y herramientas de analítica de Big Data propietarias son: Oracle, HP Vertica, IBM, Microsoft, Sybase, SAP, SAS, Teradata, Tableau Software, Kognitio, EMC, Greenplum, Google Big query.

Como herramientas de analítica de Big Data de código abierto están: Hadoop, R, Apache HBase, Pentaho y Jaspersoft.

Este análisis para incursionar con iniciativas en Big Data se realiza con herramientas de software utilizadas normalmente como parte de la disciplina de la analítica avanzada, y las herramientas más usuales son las que tienen:

- Consultas avanzadas en SQL
- Consultas e informes (quering y reporting)
- Análisis estadístico avanzado.
- Visualización de datos.
- Minería de datos, minería de textos, minería web y minería social.
- Análisis y modelado predictivo
- Optimización
- Sensibilización
- Cuadros de control y de mando (dashboard y scorecards)

Si el caso dentro de la organización es la de integrar la infraestructura de analítica de Big Data y la infraestructura de inteligencia de negocios de la misma organización, la mejor forma

de conseguir esta integración es utilizar plataformas de Big Data, fundamentalmente en torno a Hadoop, bases de datos “en-memoria” y NoSQL.

Para esto, una solución es desarrollar un sistema completo de código abierto utilizando el marco de trabajo Hadoop (HDFS y MapReduce), y herramientas tales Zookeeper, Solr, Sqoop, Hive, HBase, Nagios y Cacti. Otra solución sería desarrollar un sistema utilizando herramientas propietarias e inyectores a Hadoop como puede ser el caso de IBM con las herramientas InfoSphere, BigInsights e IBM Netezza. Además de las plataformas anteriores, existen proveedores como SAP con su producto HANA, Oracle con Exadata y Exalytics, entre otros proveedores.

Analítica Web

La analítica web o analítica del tráfico de datos en un sitio web, se centra en el análisis de los datos que fluyen a través de sitios y páginas web, este análisis de datos en la web es lo que se conoce como análisis del tráfico web.

Así, la analítica web es el análisis de datos cuantitativos y cualitativos de un sitio web y de la competencia, para impulsar una mejora continua de la experiencia online que tienen tanto los clientes habituales como los potenciales y que se traduce en unos resultados esperados (online y offline).

Este análisis web se fundamentalmente en el flujo o secuencia de clics “clickstream”. Este flujo de clics permitirá conocer casi todo acerca de los usuarios o consumidores así como disponer de datos suficientes para analizar lo que está sucediendo y las acciones a realizar para mejorar.

Esta secuencia de clics permitirá recopilar, almacenar, procesar y analizar los datos a nivel de clic del sitio web. Esta tarea se podrá obtener con herramientas de analítica web como Google Analytics, Yahoo Analytics, Webtrends, etc; y se podrá obtener la información bien en el propio sitio web o en el servidor web dependiendo de la herramienta de software instalada.

Para un mayor detalle de las métricas más utilizadas en este análisis del flujo secuencial de clics o analítica web, se puede consultar el capítulo de Generalidades de este documento.

Respecto a herramientas de analítica web, existen una gran cantidad de herramientas de analítica Web de pago y también gratuitas; una breve selección puede ser: Coremetrics, Omniture, Piwik, Woopra, Google Analytics, WebTrends, Unica, entre otros.

Analítica Social

La analítica social es el análisis de datos de los medios sociales (blogs, wikis, redes sociales, RSS), permite integrar y analizar los datos no estructurados que se encuentran en el correo electrónico, la mensajería instantánea, los portales web, los blogs y otros medios sociales, usando las herramientas de obtención de datos existentes, los informes de inteligencia de negocios o empresariales, y otras herramientas como los cuadros de mando integral.

Existen numerosas herramientas de analítica social cuyas funcionalidades son muy diversas, entre ellas están:

- ✓ Estadística social, donde los medios sociales (redes sociales, plataformas de blogs, wikis, etc) tienen sus propias herramientas estadísticas que permiten el acceso y análisis de los datos.

- ✓ Herramientas de reputación e influencia social, para tener información sobre aquello que dice la gente sobre la empresa, marca, producto o servicio, para conocer la influencia social de la compañía, así como su reputación digital.

Dentro de este análisis social, es importante el análisis de sentimiento o de sentimientos, también conocidos en algunos ambientes como minería de opinión, se refiere al análisis automático del sentimiento que trata de traducir a indicadores más o menos medibles, las emociones humanas inmersas en los datos sociales, tanto en fuentes externas y autónomas (redes sociales, blogs, microblogs, foros, medios de comunicación, wikis, etc.) como internas o propias de la empresa (interacciones almacenadas en el CRM, transcripciones de conversaciones registradas en el sistema de soporte de incidencias, encuestas realizadas a clientes y empleados).

Así, entre las herramientas más destacadas para medir el análisis de sentimiento, se encuentran: Klout, PeerIndex, Twitalyzer, How Sociable, Viralheat, etc.

Paso 8. Defina el Gobierno de los Big Data

Respecto al gobierno de los Big Data, este debe estar incluido dentro del marco más amplio de **gobierno de la información** y del **gobierno de las TI**. Según la norma internacional ISO 38.500, el gobierno de las TI tiene como principal objetivo "*evaluar, dirigir y monitorear las TI para que proporcionen el máximo valor posible a la organización*", así entonces, los principios del buen **gobierno corporativo de TIC** que define la norma son: responsabilidad, estrategia, adquisición, rendimiento, conformidad, y factor humano. Mientras que el **gobierno de la información**, es un enfoque holístico para la gestión y potenciación de la información en los beneficios del negocio y comprende la calidad de la información, protección y gestión del ciclo de vida de la información.

Por su parte, Sunil Soares[lxi], plantea las siete disciplinas básicas de Big Data que considera emanadas de las disciplinas básicas del gobierno de la información: organización, metadatos, privacidad, calidad de los datos, integración de procesos de negocios, integración de los datos maestros y gestión del ciclo de vida de la información.

Con ello, el definir el gobierno de los datos en Big Data para la organización, debe al menos contemplar en su formulación las disciplinas básicas del gobierno de la información y los principios del buen gobierno corporativo de TIC.

Paso 9. Analice, Determine e Implemente la Convergencia entre Big Data, Datos Abiertos Enlazados y Computación en la Nube

En este último paso lo que se busca es tomar conciencia de la importancia de los datos como recurso estratégico para luego plantear un análisis de las tres corrientes tecnológicas que están hoy día en alza, cada vez más integradas, y que confluyen con el objetivo de generar

modelos de negocio que están basados en datos y en capturar el valor que los propios datos encierran.

Para este análisis se recomienda apoyarse en los documentos instructivos elaborados para las tendencias de Big Data, Datos Enlazados y Computación en la Nube, dentro del subproceso de “Formulación y Mantenimiento de Políticas de Información Geográfica”.

Anexos

T1 Tabla 1 Base de Datos NoSQL orientada a Key-Value [lxvi]
Fuente: Guerrero, Fabián y Rodríguez, Jorge. Universidad Católica de Colombia

Base de Datos NoSQL				
Criterios	**Redis**	**Riak**	**Dynamo**	**Scalaris**
¿Qué es?	Motor de base de datos en memoria, basado en el almacenamiento en tablas de hashes (llave, valor)	Es una, base de datos NoSQL de la aplicación de los principios de Amazon Dynamo	Bases de datos NoSQL que proporciona un rendimiento rápido y fiable con una perfecta escalabilidad	Es un almacén de claves-valor, transaccional distribuido y escalable. Fue la primera base de datos NoSQL, que apoyó las propiedades ACID
Versión actual	Versión 2.4	Versión 1.2	Beta	Versión 0.5
Plataforma operativa	Unix, Linux, Solaris, OS/X, no existe soporte oficial para Windows	Linux, BSD, Mac, OS X, Solaris	Multiplataforma	Linux, Os X
Almacenamiento	Tablas de hashes	Fragmento particiones	Atributos multi-valuados	Múltiples claves
Tipo de índices	Geoespacial	Índices Secundarios y claves compuestas	Índices Secundarios	Índices secundarios
Esquema de replicación y distribución	Maestro-esclavo	Replicación multi-master	Maestro esclavo	Replicación multi-master
Lenguaje de consulta	API Lua	JavaScript REST Erlang	API	API JSON
Herramientas con las que se integra	ActionScript, Clojure, Erlang, Go, Haskell, Javascript, PHP, Python, Ruby	Erlang, HTTP API, PBD API	SDK AWS, CloudWatch	Servidor Web Frambesia
Tipo Licencia	Licencia BSD: software de código abierto	Apache	Propietaria	Apache
Lenguaje creación	C/C++	Erlang y C, Javascript	Java	Erlang
Creado por	Salvatore Sanfilippo and Pieter Noordhuis	Apache	Amazon	Instituto Zuse de Berlín
Protocolo	Telnet-like	HTTP/REST	HTTP/REST	JSON-RPC

Características	Tiene sistemas get/set, incrementos y decrementos de números, operaciones de listas y de conjuntos	Utilizado como una base de datos gráfica de alta escalabilidad, disponibilidad y tolerancia a fallos	No presenta esquemas fijos, y cada elemento puede tener un número diferente de atributos	En un sistema basado en Erlang realizando operaciones de escritura consistentes y distribuidas
Utilidad	Para la gestión de sesiones de usuarios y soluciones de cache, también en mensajería instantánea.	Para desarrolladores de juegos web y móvil.	Para la publicidad digital, juegos sociales y aplicaciones de dispositivos conectados	Para la gestión de archivos en Python y Ruby

T2 Tabla 2 Bases de datos NoSQL orientadas a documentos [lxvii]
Fuente: Guerrero, Fabián y Rodríguez, Jorge. Universidad Católica de Colombia

	Base de Datos NoSQL			
Criterios	**CouchDB**	**MongoDB**	**Base X**	**eXist**
¿Qué es?	Base de datos que abarcan completamente la web	Bases de datos NoSQL orientada a objetos más avanzada y flexible	Es un sistema de gestión de base de datos nativo y ligero sobre XML y XQuery,	Es sistema de gestión de base de datos de código abierto construida enteramente sobre tecnología XML
Versión actual	Versión 1.2	Versión 2.2	Versión 7.7	Versión 2.0
Plataforma operativa	Windows, Mac y Linux	Windows, Linux, OS X y Solaris	Multiplataforma	Multiplataforma
Almacenamiento	B-tree	BSON y JSON	documentos XML y colecciones	Documentos XML
Tipo de índices	Índices secundario y geoespacial	Índices secundario y geoespacial	Búsqueda de texto	Índices secundarios
Esquema de replicación y distribución	Replicación multi-master	Maestro-esclavo	No tiene modelo de replicación	Maestro-esclavo
Lenguaje de consulta	JavaScript REST Erlang	API JavaScript REST	XQuery	XQuery
Tipo Licencia	Apache License 2.0	AGPL(drivers: Apache)	BSD	GNU LGPL
Lenguaje creación	Lenguaje Erlang	C++	Java	Java
Creado por	Apache License 2.0	10gen	Cristian Grün	Wolfgang Meier

Protocolo	HTTP/REST	Custom, binary(BSON)	XML-RPC	XML-RPC
Características	Bidireccionamiento, Vistas: incrustado MapReduce, autenticación posible	Maestro/esclavo de replicación, las consultas son expresiones Javascript, tiene indexación geoespacial	Utiliza una representación tabular de estructuras de árbol XML. La base de datos actúa como un contenedor para un solo documento o una colección de documentos	Sigue estándares W3C XML, como XQuery. Soporta REST interfaces para interactuar con AJAX formularios web
Utilidad	Aplicaciones web altamente concurrentes como los juegos en línea y sistemas CMS y CRM	Es ideal para aplicaciones con estructuras complejas como blogs (post, comentarios, rollbacks, etc) o aplicaciones de analítica (Google analytics).	Para hacer seguimiento de colecciones de objetos, también para reducir informes financieros. Gestión de auditoría, control de calidad y datos de producción.	Ideal para el sector editorial y de los medios de comunicación y para el desarrollo web

T3 Tabla 3 Bases de datos NoSQL orientadas a columnas [lxviii]
Fuente: Guerrero, Fabián y Rodríguez, Jorge. Universidad Católica de Colombia

Base de Datos NoSQL				
Criterios	**Cassandra**	**Hbase**	**Hypertable**	**Jackrabbit**
¿Qué es?	Base de datos Apache que brinda escalabilidad y alta disponibilidad sin comprometer el rendimiento	Es una bases de datos distribuida de fuente abierta, modelada después de Google BigTable	Base de datos de código abierto escalable, similar al modelo de BigTable, propiedad de Google	Es un repositorio de código abierto de contenido para la plataforma Java
Versión actual	Versión 1.1.5	Versión 0.94	Versión 0.96	Versión 2.6
Plataforma operativa	Multiplataforma	Multiplataforma	Multiplataforma	Multiplataforma
Tipo de índices	Índice secundario	Claves compuestas	Índice secundario	Búsqueda de texto
Esquema de replicación y distribución	Maestro-esclavo	Maestro-esclavo	Maestro-esclavo	No tiene modelo de replicación
Lenguaje de consulta	API CQL	API REST XML	API	API

Herramientas con las que se integra	Facebook, Twitter, dig, rockspace	Facebook	Baidu, Rediff.com o Zvents	Facebook
Tipo Licencia	Apache License 2.0	Apache License 2.0	GPL 2.0	Apache License 2.0
Lenguaje creación	JavaScript	Java	C++	Java
Creado por	Apache	Apache	Zvents	Apache
Protocolo	Thrift & Custom binary CQL3	HTTP/REST	Thrift, C++ library, orHQL shell	HTTP/REST
Características	Consulta por columna, ajustable para la distribución y la reproducción, compatibilidad con múltiples centros de datos de replicación	Utiliza Haddop HDFS como almacenamiento, optimización de las consultas en tiempo real, se compone de varios tipos de nodos	Implementa diseño de BigTable de Google, la búsqueda se puede limitar a intervalos de clave/columna. Conserva los últimos valores históricos	Acceso al contenido fino y de grano grueso. Contenidos jerárquica, contenido estructurado, propiedades binarias consultas XPath, consultas SQL
Utilidad	Diseñado para aplicaciones en la industria financiera	Para realizar consultas rápidas de forma interactiva sobre un conjunto de datos definido	ideal para aplicaciones que necesitan soportar una gran demanda de datos en tiempo rea	Para transacciones, flujos BPM y en sistemas de planificación de recursos empresariales

T4 Tabla 4 Bases de datos NoSQL orientadas a gafos [lxix]
Fuente: Guerrero, Fabián y Rodríguez, Jorge. Universidad Católica de Colombia

	Base de Datos NoSQL			
Criterios	**Neo4j**	**DEX**	**HyperGraphDB**	**AllegroGraph**
¿Qué es?	Es una base datos enfocada a grafos, debido su modelo grafico es muy ágil y rápido para las operaciones de datos	Es una base de datos orientada a grafos que permite analizar grandes volúmenes de datos	Es una base de datos de gráficos, diseñada específicamente para la inteligencia artificial y los proyectos web de semántica	Base de datos gráfica moderna, utiliza memoria en combinación con el almacenamient o basado en disco
Versión actual	Versión 1.5	Versión 4.8	Versión 1.2	Versión 4.1
Plataforma operativa	Multiplataforma	Multiplataforma	Unix, Linux Windows, Mac	Amazon EC2, Linux
Almacenamiento	Local	Memoria volátil	Nodos	Nodos local

Tipo de índices	Índice secundario	No posee índices	No posee índices	geoespacial
Esquema de replicación y distribución	No tiene modelo de replicación	No tiene modelo de replicación	No tiene modelo de replicación	No tiene modelo de replicación
Lenguaje de consulta	Tinkerpop Gremlin Cypher	API	API	SparQL
Herramientas con las que se integra	PHP, JavaScript, Ruby	Facebook, Twitter	JSON, XML	SOLR y MongoDB, Python
Tipo Licencia	GPL	Uso personal / Uso comercial	LGPL	Propietaria
Lenguaje creación	Java	C++	Java	Java, C#
Creado por	Neo Technology	Sparsity-Technologies		Franz Inc.
Protocolo	HTTP/REST	HTTP/REST	HTTP/REST	HTTP/REST
Características	Optimizado para lecturas, Operaciones en el Api de Java, Indexación de nodos y relaciones	Capacidad de almacenamiento de datos y rendimiento, con órdenes de magnitud de miles de millones de nodos, aristas y atributos	Gráfico orientado de almacenamiento. Recorridos por caminos y consultas de tipo relacional. Indexación personalizable	Commit, Rollback y puntos de control, basado en multiprocesamiento, reducción de paginación, un mejor rendimiento
Utilidad	Para buscar rutas en las relaciones sociales, transporte público, mapas de carreteras, o topologías de red.	Redes de seguridad y detección de fraudes, También en redes físicas como transporte y electricidad	En aplicaciones Java del lado del servidor, en bioinformática, en redes de investigación	Se utiliza en bioinformática, en agentes inteligentes y web semántica

Referencias

i. < http://www.definicionabc.com/general/agnostico.php>
ii. <http://avro.apache.org/>
iii. APACHE. Apache CouchDB. Los Ángeles: Apache Software Foundation. Disponible en: <http://couchdb.apache.org/>
iv. <http://flume.apache.org/>
v. <http://pig.apache.org/>
vi. <http://sqoop.apache.org/>
vii. <http://zookeeper.apache.org/>
viii. ALDANA, Luis. Introducción a las bases de datos. 1 ed. Puebla. p.7.
ix. ORACLE. Oracle NoSQL Database. California: Oracle. Disponible en: <http://www.oracle.com/technetwork/database/database-technologies/nosqldb/overview/index.html>
x. Barranco, Ricardo. IBM. ¿Qué es big data?. México D.F. 2013. Disponible en <http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/>
xi. <http://incubator.apache.org/chukwa/>
xii. STRAUCH, Christof. NoSQL Databases. 1 ed. New York, 2011. 149 p.
xiii. < http://thinkandsell.com/blog/customer-centricity-los-mejores-clientes-como-centro-de-la-estrategia-de-marketing/>
xiv. < https://www.nessys.es/que-son-las-bases-de-datos-como-servicio-dbaas/>
xv. IBM. ¿Qué es Big data?, Op.cit.
xvi. Cenaltic, Badajoz. CENATIC. Open Smart Cities II: Big Data de Código Abierto. Disponible en: <http://observatorio.cenatic.es/index.php?option=com_content&view=article&id=808:open-smart-cities-ii-open-big-data-&catid=94:tecnologia&Itemid=137>
xvii. TICOUT. Introducción a Hadoop y su ecosistema. Madrid. Disponible en: <http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/>

xviii. <http://hive.apache.org/>
xix. < https://es.wikipedia.org/wiki/HTML>
xx. < https://es.wikipedia.org/wiki/Hypertext_Transfer_Protocol>
xxi. < https://es.wikipedia.org/wiki/Java_(lenguaje_de_programaci%C3%B3n)>
xxii. JSON. Introducing JSON. Disponible en: <http://www.json.org/>
xxiii. SUTINEN, Olly. NoSQL–Factors Supporting the Adoption of Non-Relational Databases. University of Tampere. Department of Computer Sciences. M.Sc. thesis, 2010, p. 21.
xxiv. < http://www.w3c.es/Divulgacion/GuiasBreves/LinkedData>
xxv. IBM. ¿Qué es Big data?, Op.cit.
xxvi. <http://mahout.apache.org/>
xxvii. IBM. What is MapReduce?. New York. Disponible en: <http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/>
xxviii. Paramio, Carlos. Una introducción a MongoDB. Madrid. Disponible en:<http://www.genbetadev.com/bases-de-datos/una-introduccion-a-mongodb>
xxix. < https://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube>
xxx. < https://es.wikibooks.org/wiki/Tecnolog%C3%ADas_de_Internet/Protocolos>
xxxi. < https://es.wikipedia.org/wiki/Familia_de_protocolos_de_Internet>
xxxii. < https://es.wikipedia.org/wiki/World_Wide_Web_Consortium>
xxxiii. GARTNER. Big Data. Disponible en : <http://www.gartner.com/it-glossary/big-data/>
xxxiv. ROUTLEDGE. Critical questions for Big data. Cambridge: Danah Boyd & Kate Crawford. Disponible en: <http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878 >
xxxv. Joyanes,Luis. Big Data. Análisis de grandes volúmenes de datos en organizaciones. 2013. Marcombo, S.A.
xxxvi. Booz & Company. Benefitting from Big Data, 2012.
xxxvii. HP. La era del Big Data. California. 2013. Disponible en: <http://h30499.www3.hp.com/t5/Infraestructura-Convergente-de/La-Era-del-Big-Data/ba-p/6151357#.UkiHs4ZLMvL/>
xxxviii. Ibid.

xxxix. Demchenko, Yuri. Defining the Big Data Architecture Framework (BDAF). Ámsterdam. UNIVERSITY OF AMSTERDAM. 2013. Disponible en: <http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf/>
xl. Rouse, Margaret. TECHTARGET. Data Ingestion. Massachusetts. 2013. Disponible en: <http://whatis.techtarget.com/definition/data-ingestion>
xli. Rouse, Margaret. TECHTARGET. Data management. Massachusetts. 2013. Disponible en: <http://searchdatamanagement.techtarget.com/definition/data-management>
xlii. Rouse, Margaret. TECHTARGET. Análisis de "big data". Massachusetts. 2013. Disponible en: <http://searchdatacenter.techtarget.com/es/definicion/Analisis-de-big-data/>
xliii. Ibid.
xliv. Halim, Nagui. IBM. Stream processing. New York. 2013. Disponible en: <http://www.ibm.com/smarter-computing/us/en/technical-breakthroughs/stream-processing.html/>
xlv. <http://research.google.com/archive/bigtable.html>
xlvi. < http://www-01.ibm.com/software/data/infosphere/hadoop/hive/>
xlvii. <http://www-01.ibm.com/software/data/infosphere/hadoop/pig/>
xlviii. < http://cloudcomputing.sys-con.com/node/2325498>
xlix. <https://apachebigdata2016.sched.org/speaker/john.eric.evans>
l. NASHOLM, Petter. Extraer datos de bases de datos NoSQL. 1 ed. Gotemburgo, 2012. p.10.
li. STRAUCH, Christof. NoSQL Databases. 1 ed. New York, 2011. p.26.
lii. Barranco, Ricardo. IBM. ¿Qué es big data?. México D.F. 2013. Disponible en <http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/>
liii. IBM. ¿Qué es big data?, Op.cit
liv. IBM. ¿Qué es big data?, Op.cit
lv. IBM. ¿Qué es big data?, Op.cit
lvi. < https://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209hadoopbigdata/>
lvii. UCC. Diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia, Op.cit
lviii. < https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/>

lix. <http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/>
lx. The Financial Brand, Big Data: Big Opportunity In Banking… Or Big B.S.?, noviembre 2012
lxi. Centro de Innovación BBVA. Big Data. Es hora de generar valor de negocio con los datos. 2013. Disponible en: <https://www.centrodeinnovacionbbva.com/sites/default/files/bigdata_spanish.pdf>
lxii. < http://www.gartner.com/technology/research/nexus-of-forces/>
lxiii. < https://es.wikipedia.org/wiki/Pensamiento_lateral>
lxiv. BBVA. Big Data: ¿En qué punto estamos?.Bogotá. Disponible en: <https://www.centrodeinnovacionbbva.com/innovation-edge/21-big-data/p/153-big-data-en-que-punto-estamos>
lxv. < http://www.dataversity.net/contributors/sunil-soares/>
lxvi. Guerrero, Fabián y Rodríguez, Jorge. Diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia. Universidad Católica de Colombia. Tesis. 2013
lxvii. UCC. Diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia, Op.cit
lxviii. UCC. Diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia, Op.cit
lxix. UCC. Diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia, Op.cit

www.
ideca.
gov.
co

Latitud: 4.603557, Longitud: -74.094105

Bogotá, Cundinamarca, Colombia.