IDENTIFICACIÓN DE FACTORES TECNOLÓGICOS PARA LA APLICACIÓN DE BIG DATA

JUAN FELIPE CASTAÑEDA MEDINA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

PEREIRA

2016

IDENTIFICACIÓN DE FACTORES TECNOLÓGICOS PARA LA APLICACIÓN DE BIG DATA

JUAN FELIPE CASTAÑEDA MEDINA

Monografía para optar al título de Ingeniero de Sistemas y Computación

Asesor

Carlos Augusto Meneses Escobar

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

PEREIRA

2016

CONTENIDO

INTRODUCCIÓN

La información siempre ha sido el factor clave en la toma decisiones en todos los ámbitos de la actividad humana. El ascenso de las tecnologías informáticas, con su capacidad para procesar grandes volúmenes de datos ha facilitado considerablemente su análisis, y la oportunidad de obtener información valiosa. Empresas y organizaciones de todos los sectores están tomando conciencia del valor potencial que tienen todos los datos que manejan y que eran ignorados.

Este trabajo contiene una descripción de lo que es Big Data y sus principales características, para luego comenzar a detallar algunas técnicas y herramientas actuales de almacenamiento, procesamiento y análisis, visualización y seguridad necesarias para aplicar correctamente Big Data y obtener el mayor valor posible de los datos.

1. GENERALIDADES

1.1 PLANTEAMIENTO DEL PROBLEMA

La tendencia del crecimiento de datos y el aumento de la capacidad necesaria para su procesamiento, es una situación a la que se enfrentan actualmente todas las organizaciones, pues se encuentran algo abrumadas ante la cantidad de información que las rodea y a la que no saben hacer frente ni sacar provecho, por lo que es absolutamente necesario disponer de herramientas para almacenar y gestionar dicha información.

Algunas estimaciones calculan que la cantidad de datos almacenados se sitúa en el entorno de varios zetabytes (1 Zetabyte= $10^{21\ bytes}$), cantidad que va aumentando exponencialmente de forma imparable.

Ante esta avalancha de datos generados diariamente, se hace necesario identificar qué factores tecnológicos se deben tener en cuenta para abordar y organizar dichos datos y convertirlos en información útil para una entidad **[2]**.

1.2 JUSTIFICACIÓN

Los seres humanos estamos creando y almacenando información continuamente y cada vez más en cantidades enormes. Según el **OBS** (Online Business School), en los últimos 10 años se ha creado más información que en toda la historia de la humanidad, hecho motivado principalmente por el desarrollo de los dispositivos móviles con conexión a Internet, del comercio electrónico y de las redes sociales **[1]**.

Cuando no hay control sobre los datos en la organización, la productividad se ve impactada negativamente, mientras que una correcta gestión de éstos, permite extraer el verdadero valor que representan para la compañía.

La información que circula en la web de forma constante, traducida en cifras puede ser aprovechada por las empresas para, por ejemplo, detectar tendencias en el mercado y orientar las acciones que se van a llevar a cabo, lo cual ayuda a tomar mejores decisiones y conseguir los resultados deseados.

Este documento va dirigido, especialmente, a todas las personas que están involucradas en tecnologías de información, Ingenieros de Sistemas, científicos de datos, analistas, directores de TI y gerentes que vean en Big Data un elemento competitivo y que brinde un valor añadido en la organización.

1.3 OBJETIVOS

1.3.1. OBJETIVO GENERAL

Crear un documento (guía) que sirva de consulta para identificar qué factores tecnológicos son necesarios para la correcta aplicación de técnicas de Big Data en condiciones que lo requieran.

1.3.2. OBJETIVOS ESPECÍFICOS

- Analizar técnicas de almacenamiento de información digital.
- Analizar técnicas de procesamiento de datos
- Analizar técnicas de visualización de datos.
- Analizar técnicas de seguridad de la información.
- Identificar factores tecnológicos relevantes para la aplicación de técnicas de Big Data.
- Realizar guía de consulta.

2. ESTADO DEL ARTE

Es tal la cantidad de datos que se está generando que las empresas están intensificando sus esfuerzos para gestionarlos y ser capaces de extraer valor para su negocio. Casi el 49% está inmerso en un proyecto de Big Data o lo va a estar próximamente. El poder alimentar la inteligencia de negocio con datos en tiempo real, a la vez que se mejora la proactividad hacia el cliente disponiendo de información para trabajar con escenarios predictivos, es una ventaja enorme.

Desde el punto de vista de tamaño de empresa, las grandes llevan ventaja a las pequeñas y medianas en la implantación de proyectos Big Data. Entre las principales limitaciones que éstas se encuentran están: la disponibilidad de presupuestos y la selección de trabajadores cualificados para analizar y gestionar los datos. Los perfiles que mayormente buscan son: científicos (27%), arquitectos de datos (24%), analistas de datos (24%), visualizadores de datos (23%), analistas de investigación (21%), y analistas de negocio (21%). **[3]**

Sin embargo, el uso intensivo del Big Data se está haciendo mayoritariamente en organizaciones digitales donde es crítico el análisis para poder tocar al 'cliente virtual'. Estas compañías han conseguido ventajas competitivas que les ha situado en posiciones dominantes en el mercado. Si se piensa en el ROI (retorno de la inversión), un proyecto Big Data debe ser multifuncional, desde marketing hasta finanzas, si no, se caerá en una iniciativa costosa y poco eficaz. La tecnología está disponible, la dificultad está en la reestructuración y transformación digital de las organizaciones.

Big Data está abriendo una valiosa ventana de información, desde los hábitos de compra del consumidor, hasta el inventario disponible. Pero esta visión interna es muy limitada si consideramos el crecimiento del negocio digital.

Gartner[A] ha identificado tres tendencias que describen la capacidad de la gestión de la información para transformar los procesos de negocio en los próximos años:

A. Gartner Inc. Empresa consultora y de investigación de las tecnologías de la información (EE.UU)

I. En el 2020, la información se utilizará para reinventar, automatizar o eliminar el 80% de los procesos de negocio y productos de la década actual.

El desarrollo del internet de las cosas proporcionará nuevos tipos de datos en tiempo real suministrando información a la cadena de valor y facilitando la automatización de los procesos con decisiones programadas.

II. En 2.017, más del 30% de las empresas accederán a un amplio Big Data a través de intermediarios de servicios de datos que añadirán contexto a las decisiones de negocio.

El negocio digital cada vez más exigirá tiempo real. Para poder anticiparse, las organizaciones deberán saber que pasa dentro y fuera, y la información actual no es suficiente. Se necesita el contexto para soportar las exigencias del negocio digital. Un contexto muy disperso y voluminoso, distribuido en cientos de miles de web, dispositivos y medios sociales.

Surgirán una nueva categoría de servicios en la nube cuyo objetivo será proporcionar al negocio los datos que necesita para sus operaciones y procesos de toma de decisiones.

III. En 2.017, más del 20% de las evaluaciones que los clientes realicen sobre los productos aprovecharán el internet de las cosas.

La disponibilidad de información por parte del consumidor (a través del móvil, medios sociales, la nube,…) sobre sus vendedores, creará un nuevo estilo de valoración basado en los datos de miles de sensores integrados en los productos. Esta información, mucho más objetiva, puede ser un factor diferencial clave para las marcas.

Cerca del 40% de las organizaciones todavía no están pensando en implementar Big Data, pero sin duda esto será una de las claves de supervivencia de las empresas en el futuro inmediato, pues el poder obtener información potencialmente valiosa será vital para mantener e incrementar la competitividad. **[3]**

2.1. BIG DATA

2.1.1. ¿Qué es Big Data?

Desde que se originó el concepto de Big Data, existen diversas definiciones y acotaciones del mismo. Según Gartner Inc, "Son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones."**[4]**

IBM, uno de los principales actores tecnológicos, también nos da una aproximación interesante de lo que es Big Data. Menciona que éste término describe enormes cantidades de datos que no pueden ser procesados o analizados usando procesos o herramientas tradicionales. Generalmente, para utilizar la expresión Big Data, se deben hablar de *petabytes* de datos (1 Petabyte = 10^{15} bytes).

Pero no solo el gran volumen es lo que principalmente identifica a Big Data, existen otros elementos fundamentales que se conocen como las **4 V** del Big Data: Volumen, Variedad, Velocidad y Veracidad.

La variedad hace referencia a los diferentes tipos y fuentes de datos, incluyendo datos estructurados, semi-estructurados y no estructurados, los cuales se generan y presentan de diversas formas como texto, audio, video, dispositivos móviles, sistemas de GPS, sensores, datos web y de redes sociales, etc.

La velocidad se refiere al tiempo con que se crean, procesan y analizan los datos. Actualmente los datos se generan a una velocidad que los sistemas convencionales no pueden soportar. Además, aplicaciones que analizan datos de posicionamiento, movimiento, temperatura y procesos similares en los que el tiempo es fundamental, se requiere que la velocidad de respuesta sea lo suficientemente rápida para obtener la información correcta en el momento preciso.

Y la veracidad es el nivel de fiabilidad asociado a los tipos de datos. Dicho de otra forma, es el valor de incertidumbre ante ciertos datos, como los sentimientos de las personas, sensores que

presentan interferencias, condiciones climáticas o indicadores económicos. A pesar de la fluctuación, todos estos datos contienen y proporcionan información valiosa. **[5]**

2.1.2. ¿De dónde proceden los datos?

Los seres humanos cada día creamos y almacenamos información en cantidades astronómicas. Una parte es recogida en llamadas telefónicas, transacciones bancarias y demás operaciones procedentes de nuestros dispositivos móviles, que según OBS, en 2020 sumarán 30.000 millones conectados a Internet. En un minuto, en Internet se generan 4,1 millones de búsquedas en Google, se escriben 347.000 tuits, se comparten 3,3 millones de actualizaciones en Facebook, se suben 38.000 fotos a Instagram, se visualizan 10 millones de anuncios, se suben más de 100 horas de vídeo a YouTube, se escuchan 32.000 horas de música en *streaming*, se envían 34,7 millones de mensajes instantáneos por Internet o se descargan 194.000 apps. En total, más de 1.570 terabytes de información por minuto. **[6]**

Otra parte que suministra una cantidad considerable de datos son los sensores que monitorizan objetos y sectores como transporte, industria, servicios, etc. Estos comunican a través de la red la información de los datos capturados, y se conoce como el Internet de las Cosas, el cual según Capgemini, generará el 40% de la totalidad de los datos creados. **[7]**

Y otro conjunto de datos es el que proviene de la seguridad, defensa y servicios de inteligencia. Son generados por lectores biométricos como escáneres de retina, huellas digitales o lectores de cadenas de ADN. Se analizan para obtener mecanismos de seguridad y generalmente son custodiados por ministerios de defensa y departamentos de inteligencia. **[8]**

2.1.3. Tipos de datos en Big Data

Datos estructurados:

Son aquellos datos que tienen bien definido su longitud y formato, como las fechas, números, cadenas de caracteres y se almacenan en tablas. Un ejemplo de estos son las bases de datos relacionales y hojas de cálculo. **[51]**

Datos no estructurados:

Son los datos que carecen de un formato específico, por lo que se encuentran en el formato tal y como se recolectaron. No están contenidos en una base de datos o tipo de estructuras de datos. Se generan en mensajes de correo electrónico, documentos de texto, PDFs, software de colaboración y documentos multimedia (imágenes, archivos de audio y video. **[9]**

Datos semi-estructurados:

Son aquellos datos que siguen una especie de estructura implícita, pero no tan regular como para poder ser gestionada y automatizada como la información estructurada. Este tipo de datos se genera por ejemplo en las páginas web, solicitudes de empleo, avisos legales, señales de tráfico.**[10]**

2.1.4. Tecnologías de Big Data

Procesar las enormes cantidades de información que se generan con rapidez, no es posible con las herramientas y métodos tradicionales. Para que las instituciones y organizaciones obtengan el máximo potencial que ofrece Big Data se debe utilizar la infraestructura tecnológica adecuada para almacenar, procesar y analizar estos grandes volúmenes de datos.

Para hablar sobre estos factores, se debe comenzar por MapReduce, que es la base de la programación de las diferentes herramientas de software y continuar con Hadoop, uno de los software más destacados. **[52]**

2.1.4.1. MAPREDUCE

Es el modelo de programación utilizado por Google para escribir aplicaciones que estén en la capacidad de procesar grandes cantidades de datos en numerosos grupos de componentes de hardware de manera confiable.

El nombre del marco proviene de dos importantes funciones en programación funcional: Map y Reduce. Map toma un conjunto de datos y se convierte en otro conjunto de datos, en el que los elementos se dividen en tuplas (pares clave/valor). Y Reduce toma la salida de un Map como entrada y combina los datos tuplas en un conjunto más pequeño de tuplas.

MapReduce permite escalar fácilmente procesamiento de datos en múltiples nodos. Esto es posible debido a que al escribir una aplicación en MapReduce, la escala de ésta que se ejecuta en cientos, miles o más máquinas en un clúster es simplemente un cambio de configuración. **[11]**

2.1.4.2. HADOOP

Es considerada la herramienta perfecta para gestionar Big Data. Bajo la dirección de la Fundación Apache, Hadoop es una biblioteca de software de código abierto que soporta el procesamiento

distribuido de grandes conjuntos de datos a través de clústers, permitiendo hacer consultas complejas sobre las bases de datos existentes y obteniendo los resultados con rapidez.

El principal objetivo de Hadoop es solucionar el problema de almacenar y procesar la información que supera la capacidad de una sola máquina permitiendo el almacenamiento de los datos en diferentes ordenadores conectados a través de una red de modo que la complejidad de su gestión sea transparente para el usuario.**[12]**

Hadoop implementa el paradigma computacional Map/Reduce, donde la aplicación se divide en muchos pequeños fragmentos de trabajo, cada uno de los cuales se pueden ejecutar o reiniciar en cualquier nodo del clúster. Además, proporciona un sistema de archivos distribuido (HDFS) que almacena los datos en los nodos de cómputo, produciendo un alto ancho de banda agregado en todo el clúster. **[13]**

Existen otros proyectos relacionados con Hadoop:

- **Avro:**
  Es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido dentro del archivo. **[14]**

- **Cassandra:**
  Apache Cassandra es una base de datos no relacional distribuida y basada en un modelo de almacenamiento "clave-valor". Permite grandes volúmenes de datos en forma distribuida. Está diseñado como un sistema distribuido, para el despliegue de un gran número de nodos a través de múltiples centros de datos, tener redundancia y recuperarse antes fallos. Además se pueden agregar nuevos nodos sin necesidad de interrumpir la ejecución de la aplicación ó reemplazar los que presenten fallos, manteniendo la disponibilidad. Está integrado con Apache Hadoop para soportar MapReduce. **[15]**

- **Flume**

  Igualmente hace parte de Hadoop y surge para subir datos de aplicaciones al sistema de archivos de Hadoop (HDFS). Su Arquitectura se basa en flujos de streaming de datos, ofrece mecanismos para asegurar la entrega y mecanismos de recuperación. **[16]**

- **HBase:**

  Es un sistema de bases de datos orientado a columnas que se ejecuta en HDFS y a diferencia de los sistemas de bases de datos relacionales, HBase no soporta un lenguaje de consulta estructurado como SQL. Cada tabla contiene filas y columnas como una base de datos relacional y cada tabla tiene definida su clave principal de acceso. HBase permite que muchos atributos sean agrupados llamándolos familias de columnas, de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto.**[17]**

- **Hive:**

  Es una infraestructura de data warehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language(HQL), estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el cluster de Hadoop.**[18]**

- **Jaql:**

  Query Language for Javascript Object Notation (JSON) - es un lenguaje funcional y declarativo que facilita la explotación de datos en formato JSON, e incluso en archivos semi-estructurados de texto plano y diseñado para procesar grandes volúmenes de información. El objetivo de JAQL es que el desarrollador de aplicaciones de Hadoop pueda concentrarse en qué quiere obtener, y no en cómo lo tenga que obtener.

  Distribuye el Query en procesos map y reduce según sea necesario, para reducir el tiempo de desarrollo respectivo en analizar los datos. **[19]**

- **Lucene:**

  Provee de librerías para indexación y búsqueda de texto y también es utilizado en la implementación de motores de búsqueda. Su funcionamiento es sencillo, los documentos son divididos en campos de texto y se genera un índice sobre éstos. La indexación es el componente clave de Lucene, lo que le permite realizar búsquedas rápidamente independientemente del formato del archivo. **[20]**

2.2. SISTEMAS DE ALMACENAMIENTO

2.2.1. Almacenamiento NoSQL

Son sistemas de almacenamiento que no cumplen con el modelo entidad-relación, siendo más flexibles y concurrentes, con el objetivo de tener una mejor escalabilidad al manipular grandes volúmenes de información y de manera más rápida que las bases de datos relacionales. Se debe aclarar que NoSQL significa "Not Only SQL", por lo que no quiere decir que se descarte el paradigma del modelo relacional, sino que con base en el reconocimiento de la naturaleza de la información y al uso que se haga de ella, es preferible un paradigma u otro. **[21]**

Se pueden clasificar cuatro grupos de bases de datos NoSQL:

A. Almacenamiento Clave-Valor

Se crean pares clave-valor por cada entrada en la base de datos, donde se accede al dato a partir de la clave que es única. Los valores son aislados e independientes entre ellos y no son interpretados por el sistema, es decir, la base de datos no sabe lo que se almacena dentro del campo "valor". Esto es responsabilidad de la aplicación que explota los datos, y es lo que facilita que sean escalables y con un alto rendimiento. **[2]**

B. Almacenamiento Documental

Similares a las bases de datos Clave-Valor, diferenciándose en el dato que guardan. En este tipo de almacenamiento se guardan datos semiestructurados, que pasan a llamarse documentos y pueden ser de clase XML, JSON, BSON o la que acepte la misma base de datos. Todos los documentos tienen una clave única con la que puede ser accedido e identificado explícitamente. **[2]**

C. Almacenamiento en Grafo

Se basan en la teoría de grafos, donde se representa la información como los nodos y sus relaciones son utilizadas para recorrer la base de datos. Se usa principalmente en casos de relacionar grandes cantidades de datos que pueden ser muy variables y en representar relaciones en el ámbito social, geográfico, software de recomendación y controles de acceso. **[51]**

D. Almacenamiento orientado a Columnas

Es el más similar a las bases de datos relacionales. La información se almacena en filas cuyas columnas pueden ser diferentes entre ellas, facilitando la información no estructurada. Orientado a almacenar datos con tendencia a escalar horizontalmente, lo que admite guardar varios atributos y objetos bajo una misma clave, pero que no serán interpretables directamente por el sistema. **[51]**

2.3. DATA WAREHOUSE

Traduce literalmente Almacén de Datos. William Harvey Inmon es considerado el padre de los Data Warehouse, cuando a comienzos de la década de los 90 expuso la siguiente definición de éstos: "Una colección de datos que sirve de apoyo a la toma de decisiones, organizados por temas, integrados, no volátiles y en los que el concepto de tiempo varía respecto a los sistemas tradicionales". **[52]**

Según Inmon, estas son las principales características de un Data Warehouse:

- **Integrado:** el Data Warehouse se construye a partir de los datos de las diversas fuentes de datos de una organización, por lo que aquellos datos deben integrarse en una estructura consistente, con diferentes niveles de detalle para adecuarse a las distintas necesidades de los usuarios y eliminando las inconsistencias existentes entre los distintos sistemas operacionales. **[22]**
- **Temático:** los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, se pueden tener todos los datos organizados por clientes, proveedores, productos, etc., independientemente de la aplicación que los vaya a utilizar y siendo las peticiones más fáciles de responder dado que la información solicitada reside en el mismo lugar. **[22]**
- **Histórico:** el tiempo es parte implícita de la información contenida en un Data Warehouse. Los datos almacenados sirven para realizar análisis de tendencias y no sólo para reflejar comportamientos del negocio en el momento presente. El Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones. **[22]**
- **No volátil:** la información contenida en un Data Warehouse está disponible para ser leída, pero no modificada. Se debe recordar que uno de los objetivos de estos almacenes de datos es dar soporte a la toma de decisiones, por lo que se pueden necesitar análisis de datos de diferentes momentos de tiempo para realizar comparaciones. La información es permanente y no recibe actualizaciones sino que se mantienen diferentes versiones de dichos datos. **[52]**

El Data Warehouse contiene datos relativos a los datos, algo que se conoce como metadatos. Estos permiten mantener información de la procedencia de la información, la periodicidad de refresco, su fiabilidad, forma de cálculo, etc., relativa a los datos contenidos en el almacén. Le brindan soporte al usuario final, ayudándolo a acceder al Data Warehouse, indicando que información está disponible y que relevancia tiene. Además le asiste en la construcción de consultas, informes y análisis mediante las herramientas de navegación. En cuanto a los encargados técnicos, obtienen soporte en aspectos de auditoría, gestión de la información histórica y demás labores de administración del Data Warehouse. **[22]**

Habiendo expuesto las características del Data Warehouse, se pueden resaltar los siguientes beneficios:

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio. **[22]**
- Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente. **[22]**
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Facilita el acceso a la información corporativa: Los contenidos del data warehouse deben ser entendibles, navegables y su acceso debe estar caracterizado por el alto rendimiento.**[23]**
- Actúa como "seguro de vida" para proteger toda la información de la organización, de forma que ésta quede accesible, entendible, estructurada y completa. **[23]**

2.4. DATAMART

Es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica, con el objetivo de ayudar a tomar las mejores decisiones dentro de dicha

área. Un Datamart puede ser alimentado desde los datos de un Data Warehouse, o integrar por sí mismo un compendio de distintas fuentes de información. Es una opción ideal para pequeñas y medianas empresas que no tienen la capacidad económica de tener a su disposición un Data Warehouse.

Para crear el Datamart de un área funcional de una empresa es necesario identificar la estructura óptima para el análisis de su información, la cual puede ser sobre una base de datos OLPT (Procesamiento de Transacciones En Línea) u OLAP (Procesamiento Analítico en Línea). La elección de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento. **[24]**

I. **Datamart OLAP:**

   Se basan en los populares cubos OLAP, que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares. **[24]**

II. **Datamart OLTP:**

   Puede basarse en un simple extracto del data warehouse, aunque lo común es introducir mejoras en su rendimiento (las agregaciones y los filtrados suelen ser las operaciones más usuales) aprovechando las características particulares de cada área de la empresa. **[24]**

2.5. CLOUD COMPUTING

Es un concepto y una tecnología joven al igual que Big Data, que hace referencia a el almacenamiento de información, comunicación entre ordenadores, ofrecer servicios y acceso y uso de recursos informáticos, todo esto ocurriendo en la nube, o sea, a través de Internet.

Internet, de una manera sencilla, se puede entender como un conjunto de ordenadores distribuidos por el mundo y unidos por una tupida malla de comunicaciones, que ofrece espacios de

información a todo el que tenga acceso. Todo lo que allí ocurre es totalmente transparente para el usuario, es decir, para él no es relevante el lugar en el que esté alojada físicamente la información y no necesita conocimiento técnico para utilizarla. Por ello se puede representar a Internet como una nube a la que se accede en busca de información y servicios.

En cuanto a los servicios que se destacan en Cloud Computing, se encuentran los de *hosting*, que permiten guardar información fuera del ordenador de un usuario, en servidores que están en la nube y a los que se accede a través de una red de comunicaciones. Igualmente, el uso de correo electrónico es un ejemplo de servicio de cloud computing, pues tanto la aplicación que se utiliza como los datos que se intercambian con los destinatarios, están almacenados en la nube. **[25]**

A. Infraestructura como servicio (IaaS, Infrastructure as a Service)

Ofrece al cliente espacio de almacenamiento o capacidad de procesamiento en sus servidores. Así el usuario tendrá a su disposición "un disco duro de capacidad ilimitada" y un procesador de rendimiento casi infinito, solo restringido a su capacidad económica de contratación del servicio. Además del espacio en servidores virtuales, IaaS, abarca aspectos como el de las conexiones de red, ancho de banda, direcciones IP y balanceadores de carga. Los recursos de hardware disponibles proceden de una multitud de servidores y redes, generalmente distribuidos entre numerosos centros de datos, de cuyo mantenimiento se encarga el proveedor del servicio.

Los proveedores de infraestructura deben permitir que los clientes puedan aumentar o disminuir los recursos de cómputo y almacenamiento a medida que cambian los requerimientos. Así mismo, que pueda disminuir los recursos, lo que debería bajar el costo de la solución. **[26]**

Ventajas de Infraestructura como servicio:

- **Escalabilidad:** recursos disponibles de la manera y en el momento en que el cliente los necesita

- **Menor costo:** servicio accesible a demanda, el usuario sólo paga por los recursos que realmente utiliza.

- **Independencia de la localización:** acceso al servicio desde cualquier lugar, siempre y cuando disponga de una conexión a Interntet y el protocolo de seguridad del servicio lo permita.

- **No hay puntos únicos de fallo:** si falla un servidor, el servicio global no se verá afectado, gracias a la gran cantidad restante de recursos de hardware y configuraciones redundantes. **[26]**

B. Plataforma como servicio (PaaS, Platform as a Service)

El servicio de Plataforma pone a disposición de los usuarios herramientas para la realización de desarrollos informáticos, de manera que aquellos pueden construir sus aplicaciones o piezas de software sin necesidad de adquirir e implantar en sus ordenadores locales dichas herramientas. Este servicio tiene dos claras ventajas para el desarrollador de aplicaciones: no tiene que adquirir las costosas licencias para desarrollo de las herramientas de mercado y, por otra parte, el proveedor de servicios se encarga de que dichas herramientas estén en óptima situación de mantenimiento. **[27]**

Ventajas de Plataforma como servicio:

- **Flexibilidad:** los usuarios tienen control sobre las herramientas que se instalan dentro de sus plataformas y pueden elegir las características que consideren necesarias

- **Adaptabilidad:** Las características se pueden modificar si las circunstancias lo ameritan.

- **No se necesita invertir en infraestructura física:** el usuario sólo debe alquilar los recursos y la infraestructura virtual que necesite.

- Menores costos, manejo financiero más flexible y eficiente y valor agregado. **[27]**

C. Software como servicio (SaaS, Software as a Service)

Permite al cliente alquilar y usar un software en línea, en vez de comprarlo y descargarlo en sus propios equipos de cómputo. De esta forma, todo el trabajo de procesamiento y almacenamiento de archivos se realiza en servidores remotos a los que se accede a través de Internet, utilizando un navegador web.

El principal beneficio de SaaS es que reduce los costos para el usuario, al evitar que pague altos precios en la compra de software que podría quedar obsoleto en unos años y simplemente puede alquilar uno que siempre esté actualizado.

Es escalable, si el usuario determina que necesita más espacio de almacenamiento o servicios adicionales, puede acceder a ellos sin necesidad de instalar nuevo software o hardware. **[28]**

Cloud Computing ofrece diferentes tipos de privacidad que pueden elegir los usuarios. Por ello se plantean varios modelos de almacenamiento en la nube:

I. Público:

Los usuarios acceden a los servicios de manera compartida sin que exista un exhaustivo control sobre la ubicación de la información que reside en los servidores del proveedor. El almacenamiento en la nube pública utiliza un mismo conjunto de hardware para realizar el almacenamiento de la información de varias personas, con medidas de seguridad y espacios virtuales para que cada usuario puede ver únicamente la información que le corresponde. Este servicio es alojado externamente, y se puede acceder mediante Internet, y es el que usualmente una persona individual puede acceder, por su bajo costo y el bajo requerimiento de mantenimiento. **[29]**

II. Privado:

Para los clientes que necesiten una infraestructura, plataforma y aplicaciones de su uso exclusivo, por la criticidad de la información que manejan. Este tipo de almacenamiento en la nube puede ser presentado en dos formatos: on-premise (en la misma oficina o casa) y alojado externamente. Generalmente es más usado por empresas que por usuarios individuales y éstas tienen el control administrativo, por lo que les es posible diseñar y operar el sistema de acuerdo a sus necesidades específicas. **[29]**

III. Híbrido:

Combina características de las dos anteriores, de manera que parte del servicio se puede ofrecer de manera privada, como la infraestructura y otra parte de manera compartida, como las herramientas de desarrollo. De esta forma el usuario puede personalizar las funciones y aplicaciones que mejor se adapten a sus necesidades y los recursos que se utilizan. **[29]**

3. TÉCNICAS DE ANÁLISIS

3.1. Minería de datos (Data Mining)

La minería de datos es un conjunto de técnicas y tecnologías que permiten explorar grandes volúmenes de datos para tratarlos, encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto y convertir esto en información útil que le permita tomar decisiones a la organización. Es una de las vías clave de explotación del Data Warehouse, pues este es su entorno natural de trabajo. **[30]**

La siguiente es la definición de minería de datos dada por Fayyad en 1996: "Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos". **[31]**

En términos generales, el proceso de Data Mining se compone de cuatro etapas principales:

I. **Determinación de los objetivos:** Delimitar los objetivos del cliente bajo la orientación del especialista en Data Mining.
II. **Preprocesamiento de los datos:** Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de Data Mining.
III. **Determinación del modelo:** Se inicia realizando unos análisis estadísticos de los datos y luego se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación.
IV. **Análisis de los resultados:** En esta etapa se verifican si los resultados obtenidos son coherentes y se cotejan con los obtenidos por los análisis estadísticos y de visualización gráfica. Con base en éstos el cliente determina si son novedosos y le aportan nuevo conocimiento para tomar decisiones. **[30]**

3.1.1. Técnicas de Data Mining

Para el proceso de Data Mining, se dispone de una amplia gama de técnicas que asisten en cada una de las fases de dicho proceso:

3.1.1.1. Análisis estadístico

Se utilizan las siguientes herramientas.

I. **Análisis de la Varianza:** contrasta si existen diferencias significativas entre las medidas de una o más variables continuas en grupo de población distintos.
II. **Regresión:** define la relación entre una o más variables y un conjunto de variables predictoras de las primeras.
III. **Ji cuadrado:** contrasta la hipótesis de independencia entre variables.
IV. **Componentes principales:** permite reducir el número de variables observadas a un menor número de variables artificiales, conservando la mayor parte de la información sobre la varianza de las variables.
V. **Análisis clúster:** permite clasificar una población en un número determinado de grupos, en base a semejanzas de perfiles existentes entre los diferentes componentes de dicha población.
VI. **Análisis discriminante:** método de clasificación de individuos en grupos que previamente se han establecido, y que permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto identificar cuáles son las variables que mejor definan la pertenencia al grupo. **[32]**

3.1.1.2. Métodos basados en árboles de decisión

Es un análisis que genera un árbol de decisión para predecir el comportamiento de una variable, a partir de una o más variables predictoras. Es útil en aquellas situaciones en las que el objetivo es dividir una población en distintos segmentos basándose en algún criterio de decisión.

El árbol se construye partiendo el conjunto de datos en dos o más subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado utilizando el mismo algoritmo. Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta. **[32]**

3.1.1.3. Algoritmos genéticos

Son métodos numéricos de optimización, en los que la variable o variables que se pretenden optimizar junto con las variables de estudio constituyen un segmento de información. Las configuraciones de las variables de análisis que obtengan mejores valores para la variable de respuesta, corresponderán a segmentos con mayor capacidad reproductiva. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Se puede además introducir elementos aleatorios para la modificación de las variables (mutaciones). Al cabo de cierto número de iteraciones, la población estará constituida por buenas soluciones al problema de optimización. **[32]**

3.1.1.4. Redes neuronales

Son métodos de proceso numérico en paralelo, en el que las variables interactúan mediante transformaciones lineales o no lineales, hasta obtener unas salidas. Estas salidas se contrastan con los que tenían que haber salido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red se reconfigura, hasta obtener un modelo adecuado. **[32]**

3.1.1.5. Series temporales

Es el conocimiento de una variable a través del tiempo para, a partir de ese conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones. Suelen basarse en un estudio de la serie en ciclos, tendencias y estacionalidades, que se diferencian por

el ámbito de tiempo abarcado, para por composición obtener la serie original. Se pueden aplicar enfoques híbridos con los métodos anteriores, en los que la serie se puede explicar no sólo en función del tiempo sino como combinación de otras variables de entorno más estables y, por lo tanto, más fácilmente predecibles. **[32]**

3.1.2. Metodología de aplicación

Para utilizar correctamente estas técnicas es necesario aplicar una metodología estructurada al proceso de Data Mining:

- **Muestreo**

  Se refiere a extraer la muestra de la población sobre la que se va a aplicar el análisis. Puede tratarse de una muestra aleatoria, pero puede también ser un subconjunto de datos del Data Warehouse que cumplan unas condiciones determinadas. No se trabaja con toda la población para simplificar el estudio y disminuir la carga de proceso. La muestra más óptima será aquella que teniendo un error asumible contenga el número mínimo de observaciones. **[32]**

- **Exploración**

  Una vez determinada la muestra de la población que sirve para la obtención del modelo se deberá determinar cuáles son las variables explicativas que van a servir como entradas al modelo. Para ello es importante hacer una exploración por la información disponible de la población que nos permita eliminar variables que no influyen y agrupar aquellas que repercuten en la misma dirección. Se pueden utilizar herramientas que permitan visualizar de forma gráfica la información utilizando las variables explicativas como dimensiones y técnicas estadísticas que nos ayuden a poner de manifiesto relaciones entre variables. **[32]**

- **Manipulación**

  Tratamiento realizado sobre los datos de forma previa a la modelización, en base a la exploración realizada, para definan claramente las entradas del modelo a realizar. **[32]**

- **Modelización**

  Permite establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibilitan inferir el valor de las mismas con un nivel de confianza determinado.

- **Valoración**

  Análisis de la bondad del modelo contrastando con otros métodos estadísticos o con nuevas poblaciones muestrales.

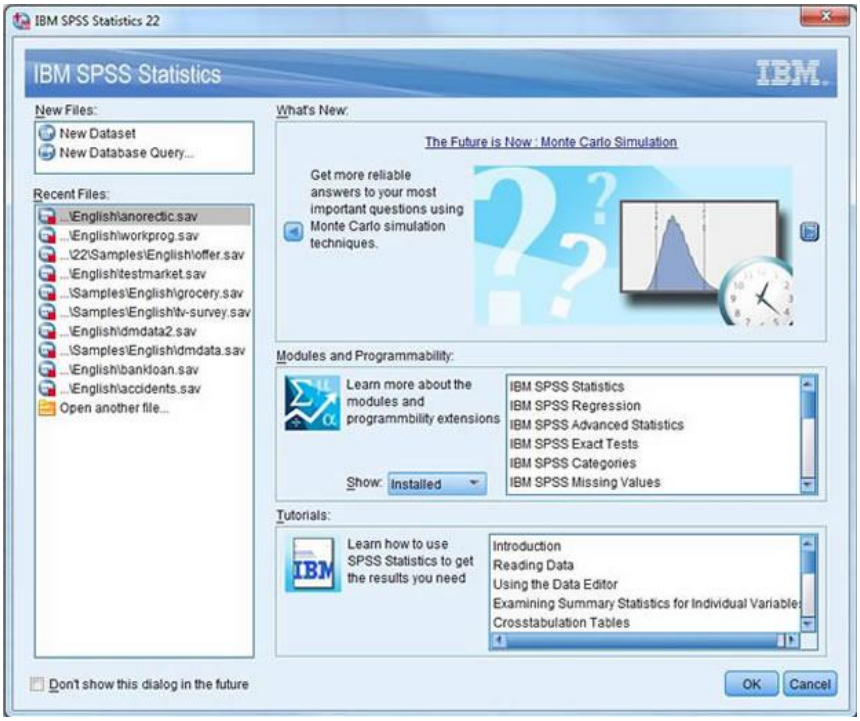
3.1.3. Software de minería de datos

I. IBM SPSS Statistics

Es un completo conjunto de datos y herramientas de análisis predictivo fácil de utilizar para usuarios empresariales, analistas y programadores estadísticos. **[33]** La versión Standard ofrece los procedimientos estadísticos principales que los gestores y los analistas necesitan para tratar las cuestiones empresariales y de investigación básicas. Proporciona herramientas que permiten a los usuarios consultar datos y formular hipótesis para pruebas adicionales de forma rápida, así como ejecutar procedimientos para ayudar a aclarar las relaciones entre variables, crear clústeres, identificar tendencias y realizar predicciones. **[34]**

La versión Professional está diseñada para usuarios que realizan varios tipos de análisis en profundidad y que necesitan ahorrar tiempo mediante la automatización de tareas de preparación **[35].** Y la versión Premium ayuda a los usuarios a completar las tareas con facilidad en cualquier

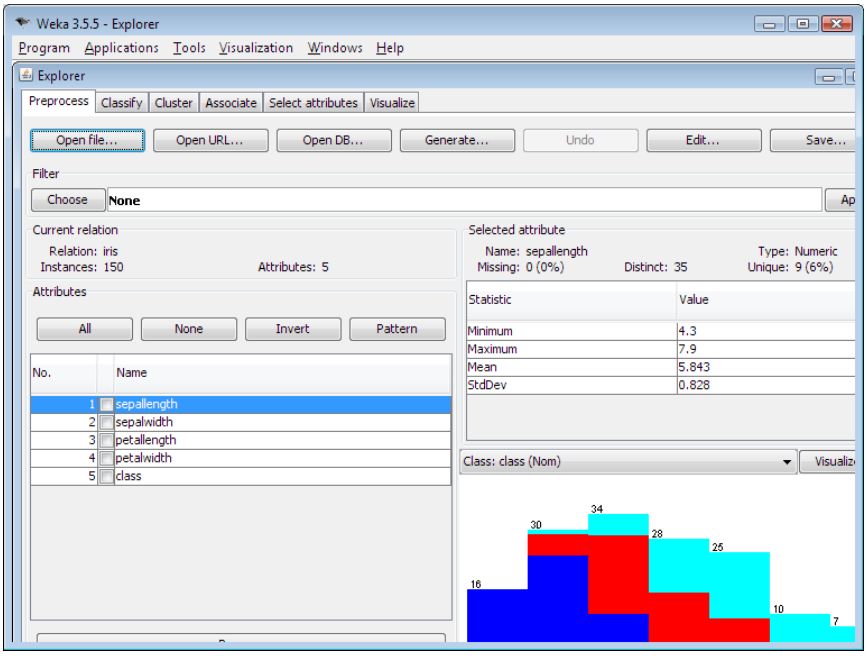
fase del proceso analítico e incluye una variedad de funciones integradas para la ejecución de tareas analíticas especializadas en toda la empresa. **[36]**



Captura de pantalla SPSS Statitics 22

II. Weka

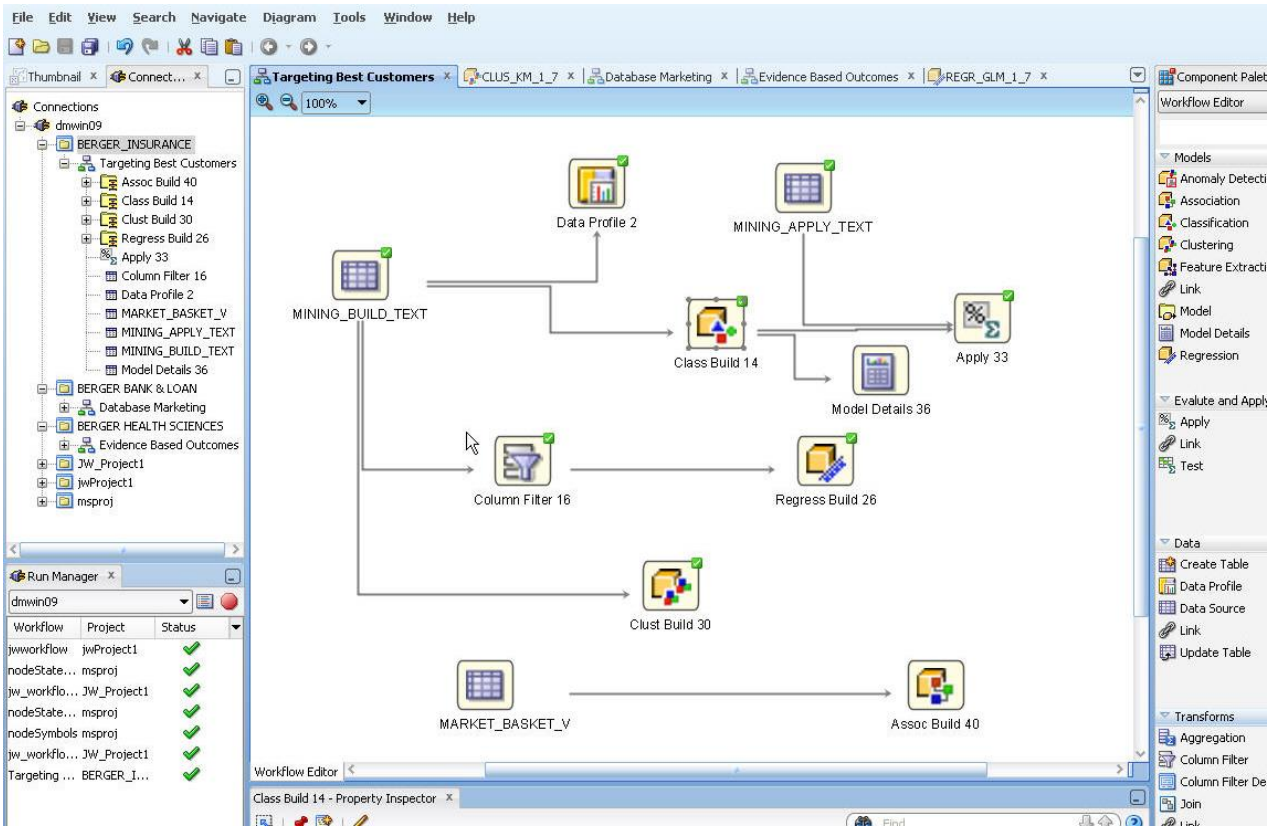
Es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y distribuido bajo licencia GNU-GPL. Contiene herramientas para el procesamiento previo de datos, clasificación, regresión, clustering, reglas de asociación y visualización. **[37]**



Captura de pantalla Weka 3.5.5

III. Oracle Data Miner

Proporciona potentes algoritmos de minería de datos que permiten encontrar tendencias, hacer predicciones y aprovecharlos en la organización. Estos algoritmos se implementan como funciones SQL y aprovechan las fortalezas de la plataforma Oracle. Cuenta con una interfaz gráfica de usuario que le permite a los analistas de datos, negocios y científicos de datos trabajar intuitiva y directamente con los datos utilizando comandos como “arrastrar y soltar”. **[38]**



Captura de pantalla Oracle Data Miner

3.1.4. Extensiones del Data Mining

- **Web Mining**

  Consiste en aplicar las técnicas de la minería de datos a documentos y servicios en la Web**.** Al visitar sitios en la Red, se dejan huellas digitales (direcciones IP, navegadores, cookies, etc) que los servidores automáticamente almacenan en una bitácora de accesos llamados log. Las herramientas de web mining analizan y procesan estos logs para producir información significativa, como por ejemplo saber cuál es la navegación de un cliente antes de hacer una compra en línea. **[39]**

- **Text mining**

  En vista de que el mayor porcentaje de información de una organización está almacenada en forma de documentos, técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, apoyan al text mining (minería de texto). **[40]**

  Consiste en examinar una colección de documentos y descubrir información no contenida en otro documento individual de este conjunto, es decir, obtener información sin haber partido de algo. **[41]**

3.2 BUSINESS INTELLIGENCE

Es una herramienta o estrategia empresarial que tiene el objetivo de transformar datos en información útil y relevante para optimizar el proceso de toma de decisiones en la organización. Desde un punto de vista más teórico, Business Intelligence se podría definir como el conjunto de metodologías, aplicaciones y tecnologías que permiten obtener, depurar y transformar datos de los sistemas transaccionales e información interna y externa a la empresa, para explotar o analizar dicha información y así convertirla en conocimiento útil que ayude a la toma de decisiones.

La inteligencia de negocio actúa como un factor estratégico para una empresa u organización, generando una potencial ventaja competitiva: el proporcionar información privilegiada para responder a los problemas de negocio, como entrada a nuevos mercados, promociones u ofertas de productos, control financiero, reducción de costos, análisis de perfiles de clientes. Los proyectos de Business Intelligence suelen iniciarse a través de la alta gerencia, los departamentos de planeación estratégica, finanzas o mercadeo. **[42]**

Los productos más destacados de Business Intelligence que existen hoy en día son:

A. Cuadros de Mando Integrales (CMI)

Es una herramienta de control empresarial que permite establecer y monitorizar los objetivos de una empresa y de sus diferentes áreas, desde el punto de vista estratégico y con una perspectiva general. Así, con la información periódica obtenida del seguimiento en el cumplimiento de los objetivos, la toma de decisiones resulta más sencilla y eficaz, y se pueden corregir las desviaciones a tiempo. **[43]**

Un Cuadro de Mando Integral se compone de cuatro perspectivas en las que se establecen los objetivos estratégicos:

- **Perspectiva de aprendizaje y crecimiento:** relacionada con los recursos más importantes en el proceso de creación de valor: materiales (tecnología) y las personas. Incide sobre la importancia que tiene el concepto de aprendizaje por encima de lo que es en sí la formación tradicional.
- **Perspectiva interna:** recoge indicadores de procesos internos que son críticos para el posicionamiento en el mercado y para llevar la estrategia a buen término. Éstos proporcionan información valiosa acerca del grado en que las diferentes áreas de negocio se desarrollan correctamente.
- **Perspectiva financiera:** incorpora la visión de los accionistas y mide la creación de valor de la empresa. En síntesis, esta perspectiva refleja uno de los objetivos más importantes de una organización con ánimo de lucro: sacar máximo provecho de las inversiones realizadas.

- **Perspectiva del cliente:** relacionada con el posicionamiento de la organización en el mercado o en los segmentos donde se quiere competir y reforzará o debilitará la percepción del valor de la marca por parte del consumidor.

CMI ofrece una amplia visión para un seguimiento detallado de la marcha del negocio, que engloba muchos aspectos y permite observar otras variables determinantes en el buen desarrollo

de la empresa. Además, facilita la planificación de estrategias a mediano y largo plazo y genera la información necesaria para tomar decisiones útiles. **[43]**

B. Sistemas de Soporte a la Decisión (DSS)

Son una herramienta de Business Intelligence enfocada al análisis de los datos de una organización. Permiten resolver gran parte de las limitaciones de los programas de gestión. Estas son algunas de sus características principales:

- **No requiere conocimientos técnicos**: Un usuario común puede crear nuevos gráficos e informes y navegar entre ellos, haciendo uso por ejemplo, de "arrastrar y soltar". Por tanto, para examinar la información disponible o crear nuevos indicadores no es imprescindible buscar un experto técnico.
- **Rapidez en el tiempo de respuesta:** como la base de datos subyacente suele ser un Data warehouse corporativo o un datamart, con modelos de datos en estrella o copo de nieve, están optimizadas para el análisis de grandes volúmenes de información. **[44]**

- **Restricciones de usuario:** se refiere a que no todos los usuarios tengan acceso a toda la información, sino de que tenga acceso a la información que necesita para que su trabajo sea lo más eficiente posible.

- **Disponibilidad de información histórica:** en estos sistemas está a disposición comparar los datos actuales con información de otros períodos históricos de la compañía, con el fin de analizar tendencias, fijar la evolución de parámetros de negocio, etc.

El principal objetivo de los Sistemas de Soporte a Decisiones es explotar al máximo la información residente en una base de datos corporativa, mostrando informes muy dinámicos y con gran potencial de navegación, pero siempre con una interfaz gráfica amigable, vistosa y sencilla para el usuario. **[44]**

C. Sistemas de Información Ejecutiva EIS

Son una herramienta de software, basados en DSS, que proveen a los gerentes de un acceso sencillo a información interna y externa de su organización, y que es relevante para sus factores clave de éxito. Su propósito principal es que tengan a su disposición un panorama completo del estado de los indicadores de negocio que le afectan en tiempo real, manteniendo también la posibilidad de analizar con detalle aquellos que no estén cumpliendo con las expectativas establecidas, para determinar las acciones a realizar más adecuadas

Un modelo adecuado de BI conseguirá que la información sea íntegra, represente la realidad del negocio y soporte la toma de decisiones basada en datos confiables. Si la Inteligencia de Negocios se aplica correctamente, el resultado no solo redundará en ganancias, crecimiento y eficiencia, sino que permitirá a la organización vigilar el desempeño de sus negocios y desarrollar acciones de mejora. **[45]**

4. TÉCNICAS DE VISUALIZACIÓN

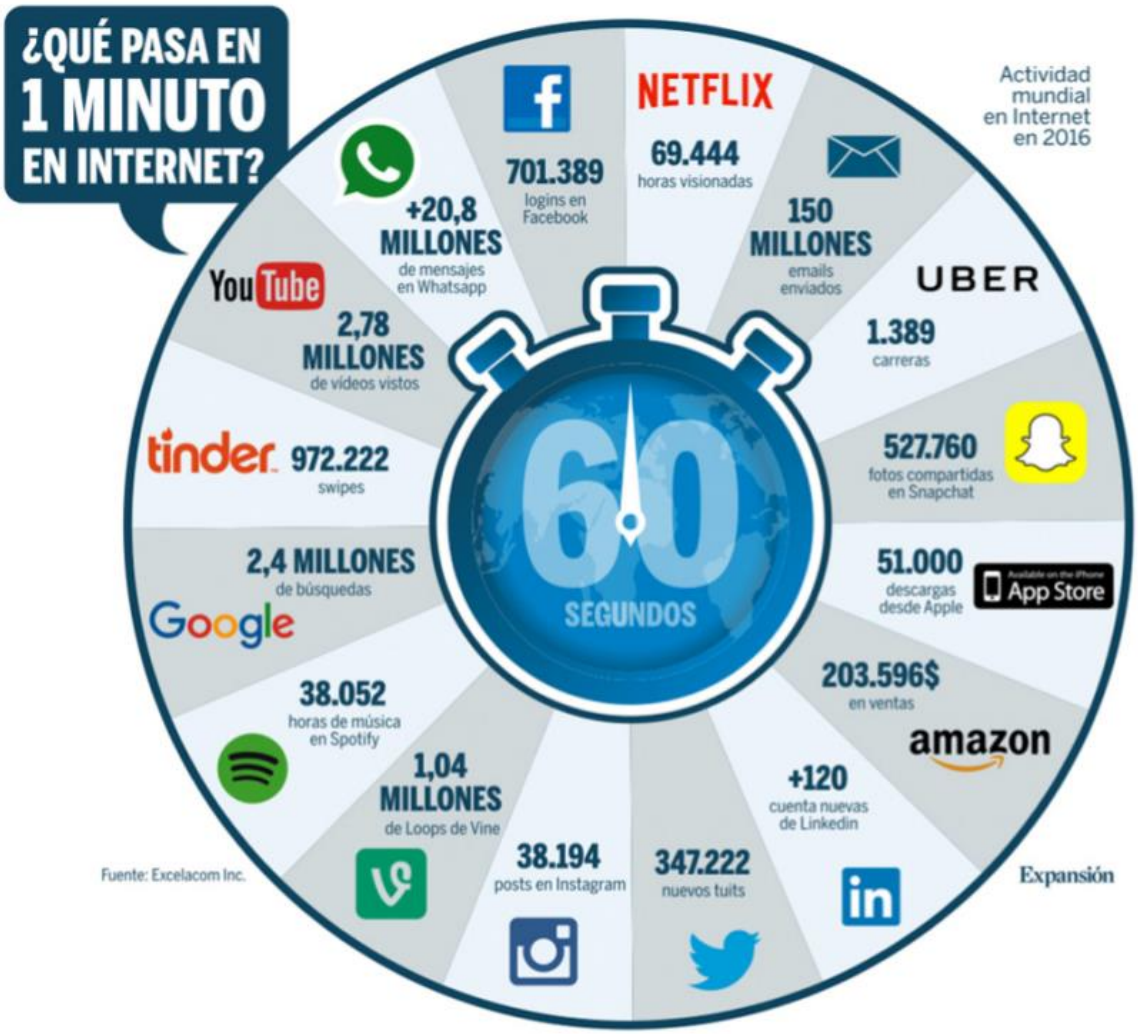
Diariamente las organizaciones necesitan analizar y comprender cantidades enormes de datos para poder tomar decisiones de negocio. Las herramientas de visualización de datos permiten representar cualquier tipo de información de una forma visual y sencilla, hecho de vital importancia en la era del Big Data, donde se debe estar en la capacidad de extraer el valor de millones de datos en el menor tiempo posible. **[46]**

Las siguientes son unas de las más destacadas herramientas de visualización de información:

- **Infografía**

  La infografía es una combinación de textos e imágenes resumidas, explicativas y fáciles de entender con el fin de comunicar información de manera visual para facilitar su

transmisión. Son un medio muy poderoso para representar información que a través de texto puro sería muy complicado entender y facilita que se recuerde por más tiempo. **[47]** La infografía acelera la asimilación de conceptos y conlleva la toma de las mejores decisiones para obtener resultados exitosos en la organización. Mediante ella, por ejemplo, se puede convencer a los clientes para que adquieran determinado producto ó también para comunicarse de forma rápida y eficaz con los empleados. **[48]**



Ejemplo de infografía: Qué pasa en 60 segundos en Internet. Fuente: **Excelacom Inc.** Obtenido de https://ticsyformacion.com/2016/04/25/que-pasa-en-un-mi

- **Tableau Software**

Es un software de Business Intelligence que debido a su gran capacidad visual de análisis, facilita la comprensión de datos. Su funcionamiento es muy intuitivo, permitiendo la creación de visualizaciones de alto nivel, informes y tableros de control con tan sólo arrastrar los datos y así poder ver los cambios en tiempo real, mientras éstos se van realizando. Permite mostrar diferentes representaciones de los datos en un mismo tablero y añadir información extra, por ejemplo, a través de documentos o páginas web. **[49]**

- **Qlick View**

Es una herramienta que le ofrece al usuario la posibilidad de recopilar datos procedentes de múltiples orígenes, manipularlos y organizarlos, según su objetivo, para presentarlos de manera muy visual. Una de sus particularidades está en poseer los datos integrados con el propio cuadro de mandos, trabajando así desconectado de los orígenes de los datos. No requiere una formación previa avanzada para manejar la herramienta y en poco tiempo es dominada por el usuario. **[49]**

- **Many Eyes**

Many Eyes es la herramienta de visualización de datos diseñada y creada por IBM. Tiene como particularidad el hecho de tratarse de una herramienta de uso público, donde todas las visualizaciones que se creen podrán ser vistas, comentadas y valoradas por el resto de usuarios.

Ofrece varias posibilidades de personalización y su funcionamiento es sencillo: deben subirse los datos (preparados previamente en la misma u otra plataforma) y una vez subidos, el usuario escoge el tipo de visualización que desea. **[49]**

- **Google Fusion Table**

Google también cuenta con su propia herramienta para la visualización de datos y sólo se necesita tener una cuenta de Google para utilizarla. Permite compartir los datos de forma abierta y construir visualizaciones personalizadas en función de cómo desee representarlo el usuario. Entre las posibilidades que ofrece están diagramas de dispersión, líneas de tiempo, gráficos de barras e incluso mapas geográficos a través del servicio de Google Maps. Es totalmente gratuita y se puede almacenar en Google Drive para compartirla y seguir trabajando de manera colaborativa. **[49]**

- **D3, Data Driven Documents**

Es una herramienta capaz de ofrecer visualizaciones interactivas online muy avanzadas con complejos conjuntos de datos. Se trata de una librería de Javascript que ofrece la posibilidad de crear diagramas bastante impresionantes y gráficos a partir de una amplia variedad de fuentes de datos. Es de código abierto, y utiliza los estándares web, lo que la convierte en muy accesible. Un aspecto negativo de D3 es que debido a que es muy complejo, se necesitan conocimientos de programación y su lenguaje en concreto. Por tanto, no es tan sencillo de utilizar como otras herramientas. **[49]**

- **CartoDB**

Es una herramienta lanzada por la firma española CartoDB, que facilita la creación de mapas, visualizaciones y análisis de datos. Permite a cualquier empresa visualizar sus bases de datos de una manera muy sencilla en función de criterios geográficos. CartoDB ofrece su servicio online mediante un modelo de negocio freemium: si se necesita plasmar en un mapa interactivo unos pocos datos, con hasta 50 megas y hasta 5 tablas, CartoDB dispone de un paquete gratuito. A partir de ahí se tienen diferentes opciones de pago en función de las necesidade del usuario. Entre sus clientes destacados se encuentran las Naciones Unidas, National Geographic o la Nasa, que han visto en CartoDB una

manera muy sencilla de crear mapas interactivos utilizando la nube como soporte y su software de código abierto. **[50]**

5. SEGURIDAD EN BIG DATA

Como se ha visto a lo largo de este documento, Big Data representa una excelente oportunidad para las organizaciones de todos los sectores. Mediante el aprovechamiento de nuevos volúmenes y variedades de datos, científicos, ejecutivos, gerentes de productos, comercializadores y una amplia variedad de personas pueden comenzar a elaborar estrategias y tomar decisiones más acertadas, descubrir nuevas oportunidades de optimización y ofrecer soluciones innovadoras.

Sin embargo, para que la aplicación de Big Data no se vuelva un problema, se deben implementar métodos adecuados de cifrado y seguridad. **[53]**

Algunos de los retos para la seguridad que deben tenerse en cuenta son:

- La adopción de la tecnología que permita manejar Big Data debe ser pensada específicamente. Estructuras de cómputo distribuido, en los cuales intervienen múltiples plataformas y sistemas deben tener consideraciones especiales de seguridad, pues tanta diversidad puede dar a lugar a que queden agujeros de seguridad explotables por delincuentes informáticos. **[54]**

- Respecto al almacenamiento y procesamiento en la nube, se requiere contar con las garantías necesarias para que se mantenga la confidencialidad de la información.

- La educación de los empleados, para que se incorporen hábitos seguros en el manejo de la información, apoyados en soluciones que aseguren el acceso y manipulación de los datos.

- En muchas ocasiones se opta por soluciones que simplifiquen los requerimientos de seguridad para mantener las funcionalidades requeridas, pero se debe tener cuidado porque aunque facilitan las operaciones en el corto plazo, pueden surgir problemas más adelante. **[54]**

- La información recopilada que se incluye en los proyectos de Big Data guardará relación con otros conjuntos de datos, que podrán generar nueva información o alterar los datos originales de diferentes maneras, a menudo impredecibles. Las organizaciones deben asegurarse de que todos los requisitos de seguridad y privacidad que se aplican a los conjuntos originales de datos sean monitorizados y mantenidos en los procesos de Big Data a lo largo del ciclo de vida de la información, desde la recopilación de los datos hasta su divulgación o destrucción. **[55]**

El almacenamiento en la nube ofrece servicios con herramientas que en la actualidad son básicas para millones de personas en todo el mundo, que confían sus archivos y demás, para disfrutar de grandes beneficios como la sincronización de archivos entre diferentes dispositivos. Sin embargo, estos archivos deben estar protegidos, mediante cifrado, tanto del contenido como en la transferencia de éste. Por lo general, el cifrado se realiza una vez el archivo ha llegado al servidor. **[56]**

Se pueden distinguir dos tipos de cifrado respecto al almacenamiento en la nube:

- **Cifrado en servidor**

  Cifrado en servidor o cifrado del lado del servidor es el método que utilizan la mayoría de servicios de almacenamiento de archivos en la nube. Esto quiere decir que los archivos llegan al servidor sin cifrar, y allí son cifrados, normalmente, con la contraseña del usuario.

  El nivel de seguridad de este método es perfectamente aceptable, siempre que la transferencia de los archivos se haga a través de una conexión segura (HTTPS / SSL), debido a que los archivos viajan del equipo al servidor sin cifrar. Los sitios web de estos servicios y las aplicaciones para PC y móvil fuerzan la conexión segura.

  Se debe resaltar que aunque los datos están seguros, no se puede garantizar su privacidad, pues el administrador del servidor puede acceder a ellos y/o a las claves de cifrado. [56]

- **Cifrado en cliente**

Cifrado en cliente o cifrado del lado del cliente consiste en cifrar los archivos antes de que salgan del equipo, por lo general también, con la contraseña del usuario. Igualmente, lo ideal en este caso es que la contraseña nunca salga del cliente. Dicho de otro modo: los responsables del servicio solo almacenan y sincronizan datos cifrados, cuyo contenido no pueden descifrar.

El cifrado en cliente tiene varias ventajas para las dos partes:

- Para el usuario es más privado, pues solo en su equipo permanecen sus archivos descifrados.
- Cualquier robo de datos en el servidor o durante la transferencia del archivo solo obtendrá archivos cifrados.
- El servicio no se hace responsable de los contenidos que aloje el usuario y no tiene acceso a ellos.

Una limitación en este tipo de cifrado, es que el usuario perderá el acceso a su información si pierde su contraseña, ya que ésta solo se guarda en el cliente. **[56]**

5.1 Tecnologías y soluciones destacadas

5.1.1. HP Atalla

Las soluciones de seguridad y cifrado de datos de HP Atalla (ofrecido por Hewlett-Packard, empresa de tecnología estadounidense) protegen los datos confidenciales a través de todo su ciclo de vida – estén en reposo, en movimiento o en uso– en las instalaciones o en entornos de la nube o móviles, garantizando la protección continua mientras gestionan un rendimiento y una flexibilidad óptimos.

Un producto notable que ofrece seguridad en la nube es **Enterprise Secure Key Manager (ESKM)**, para garantizar la seguridad de los datos. Es una solución completa para la gestión de claves empresariales y protección de datos para asegurar servidores y almacenamiento en la nube

contra pérdidas, gestión deficiente y ataques administrativos y operativos. **[57]** Unifica y automatiza los controles de encriptación para evitar el acceso no autorizado a datos sensibles, reduce el costo y la complejidad de la gestión de claves a través de una infraestructura distribuida, con los controles de seguridad consistentes y servicios clave automatizados y un único punto de gestión y realiza y gestiona las claves de cifrado sin disminuir el rendimiento del servidor. También es totalmente compatible con Key Management Interoperability Protocol (KMIP) (Protocolo de Interoperabilidad de administración de claves), estándar abierto para la gestión de claves de encriptación. **[58]**

5.1.2. Vormetric

Vormetric, una compañía estadounidense, también se destaca en el sector ofreciendo completas soluciones de seguridad de datos que abarca entornos físicos, virtuales y la nube. La Plataforma de seguridad de datos de Vormetric ofrece capacidades para el cifrado de datos masivos, la administración de claves y el control de accesos, que incluye varias ofertas de productos que comparten una infraestructura común y expansible. Además, la solución genera inteligencia de seguridad respecto del acceso a los datos por parte de usuarios, procesos y aplicaciones. Permite a las organizaciones maximizar los beneficios del análisis de datos masivos e incrementar al máximo la seguridad de los datos sensibles en los entornos de Big Data. **[59]**

Los siguientes productos son parte del portafolio que ofrece Vormetric:

- **Protección de fuentes de Big Data**

  Las organizaciones pueden aprovechar los datos provenientes de una amplia variedad de fuentes, tanto estructuradas como no estructuradas, para satisfacer sus iniciativas de datos masivos. Los datos provenientes de bases de datos, almacenes de datos, registros de sistema, hojas de cálculo y muchos otros sistemas variados pueden incorporarse en un entorno de datos masivos.

  Para implementar la seguridad de datos en estas fuentes heterogéneas, las organizaciones pueden emplear las siguientes soluciones de Vormetric:

- **Cifrado transparente de Vormetric:** cifra y controla el acceso en el nivel del sistema de archivos. La solución de cifrado es sencilla de implementar porque no requiere ningún cambio en las aplicaciones. **[59]**

- **Cifrado de aplicaciones de Vormetric:** permite cifrar columnas específicas en una aplicación antes de que esta escriba el campo en una base de datos. Al cifrar una columna específica, puede asegurarse de que un campo confidencial específico permanezca ilegible, incluso luego de su importación y su procesamiento dentro del entorno de datos masivos.

❖ **Protección de marcos de trabajo de Big Data**

En los entornos de Big Data, los datos se replican y migran de forma rutinaria entre una gran cantidad de nodos. Además, la información confidencial se puede almacenar en registros de sistema, archivos de configuración, cachés de disco, registros de errores, etc. El Cifrado transparente de Vormetric protege con eficacia los datos en todas estas áreas y proporciona cifrado, control de acceso de usuarios con privilegios e inteligencia de seguridad. **[59]**

❖ **Protección de análisis de Big Data**

Los resultados del análisis de los datos masivos tienen muchos formatos, los cuales incluyen reportes bajo demanda, informes automatizados y consultas ad hoc. Con mucha frecuencia, estos resultados contienen propiedad intelectual que resulta muy valiosa para una organización y constituye un posible blanco de ataque. Los equipos de seguridad pueden emplear las siguientes soluciones con el fin de proteger el análisis de Big Data:

- **Cifrado transparente de Vormetric:** puede implementarse fácilmente en los servidores, en los cuales puede cifrar resultados de datos masivos, y controlar y supervisar quién tiene acceso a ellos.

- **Cifrado de aplicaciones de Vormetric:** se puede emplear para proteger campos específicos que se pueden crear en aplicaciones de análisis. **[59]**

❖ **Administración de Claves**

La Plataforma Vormetric Data Security centraliza claves de cifrado de terceros y almacena certificados de manera segura. Brinda una administración de claves de cifrado empresarial de alta disponibilidad, basada en estándares para el Cifrado de base de datos transparente (TDE) y los dispositivos compatibles con el Protocolo de Interoperabilidad con la Administración de Claves (KMIP) y ofrece almacenamiento e inventario de certificados. La consolidación de la administración de claves de cifrado empresarial ofrece una implementación de políticas uniforme entre los sistemas y reduce los costos de capacitación y mantenimiento. **[60]**

Otras características de la administración de claves de Vormetric son las siguientes:

- **Reduce el tiempo de inactividad:** la alta disponibilidad, las notificaciones proactivas de certificados y la caducidad de las claves de cifrado reducen el tiempo de inactividad de la aplicación y del usuario.
- **Centraliza los informes:** genera informes consolidados para cumplimiento y auditorías simplificadas de claves de cifrado y uso de certificados.
- **Operaciones multitenencia:** administración basada en roles para la gestión compartimentada de las políticas de seguridad de datos, las claves de cifrado de datos y los registros de auditoría.
- **Cifrado de base de datos transparente (TDE):** consolida la administración de claves de cifrado para Oracle y SQL Server de Microsoft. **[60]**

❖ **Cifrado de aplicaciones**

El Cifrado de aplicaciones de Vormetric es una biblioteca que simplifica la integración del cifrado de nivel de aplicaciones en aplicaciones corporativas existentes. Esta

biblioteca de cifrado ofrece un conjunto de API basados en estándares documentados que se utilizan para realizar operaciones de administración de claves criptográficas y de cifrado. Permite a los desarrolladores elegir el cifrado basado en el estándar AES (Advanced Encryption Standard) ó el esquema que mantiene el Cifrado con preservación de formato. El cifrado de aplicaciones de Vormetric elimina la complejidad y el riesgo de implementar una solución interna de administración de claves y cifrado. **[61]**

- **Cifrado transparente**

Permite el cifrado de datos estáticos, el control de acceso de usuarios privilegiados y la recolección de registros de inteligencia de seguridad sin rediseñar aplicaciones, bases de datos o infraestructura. Su instalación es simple, escalable y rápida. Los Agentes de cifrado transparente de Vormetric se instalan en la parte superior del sistema de archivos en los servidores o máquinas virtuales para ejecutar la seguridad de datos y las políticas de cumplimiento. **[62]**

Algunos de sus atributos principales son:

- **Despliegue transparente:** No requiere desarrollo o cambios en la experiencia del usuario, aplicaciones o infraestructura.
- **Limita el riesgo de los usuarios privilegiados:** El software de encriptación de datos detiene a usuarios privilegiados de root, de sistema, de la nube, de almacenamiento y otros administradores de accesar datos mientras mantiene su habilidad para llevar a cabo las actividades administrativas usuales.
- **Amplio soporte a S.O y aplicaciones heterogéneas:** Los agentes de cifrado soportan plataformas Windows, Linux y Unix plataformas así como la mayoría de bases de datos y archivos no estructurados.
- **Mantiene acuerdos de niveles de servicio (SLA):** Al distribuir agentes optimizados para sistemas de archivos específicos y hardware de aceleracion de encripción en los servidores resulta en una latencia muy baja y poca carga adicional a los sistemas. **[62]**

6. CONCLUSIONES

- La información tiene cada vez más importancia en el desarrollo de los negocios y las organizaciones. Su uso es cada vez más relevante y clave en los escenarios de evolución de las empresas.

- Big Data permite una mayor transparencia y una mejor utilización de la información, creando valor agregado que la organización usa para la toma de decisiones y estrategias.

- El almacenamiento y el análisis de datos masivos facilitará el surgimiento de nuevas formas de organizaciones y de relaciones entre altos directivos y empleados, con lo que se mejorará la eficiencia empresarial.

- Las organizaciones se han dado cuenta de que Big Data no sólo es la tendencia del momento, sino de que llegó para quedarse y es una necesidad para ser competitivos donde las predicciones y toma de decisiones son una condición para lograr el éxito.

- Las organizaciones deben adoptar estrategias personalizadas de seguridad para minimizar la ocurrencia de riesgos

- Este documento puede ser útil para científicos, analistas y visualizadores de datos, analistas de negocios, gerentes e involucrados con las tecnologías de información. Además, puede ser útil en el desarrollo de proyectos con las tecnologías y herramientas expuestas a lo largo del documento.

7. BIBLIOGRAFÍA

[1] **OBS Business School.** *El volumen de datos generados por smartphones crecerá un 63% los próximos cuatro años*. 2014. Disponible en http://www.obs-edu.com/noticias/informe/el-volumen-de-datos-generado-por-smartphones-crecera-un-63-los-proximos-cuatro-anos/

[2] Álvarez, Bernabéu Auban y Peñarrubia Carrión, (2015). ***Big Data el valor de los datos:*** *Estado actual y tendencias del Big Data como nuevo activo en la economía europea* (1ª ed). España. Disponible en https://www.coiicv.org/publicaciones/send/23-monografias/467-big-data-el-valor-de-los-datos.

[3] **Big Data: estado del arte y tendencias.** Obtenido de http://www.innovan.do/2015/03/20/big-data-estado-del-arte-y-tendencias/

[4] **Gartner Inc,** IT Glossary, obtenido de http://www.gartner.com/it-glossary/big-data/

[5] **IBM,** Big Data, obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

[6] Obtenido de http://www.obs-edu.com/noticias/estudio-obs/en-2020-mas-de-30-mil-millones-de-dispositivos-estaran-conectados-internet/

[7] **Capgemini**, Fundamentos del Internet de las cosas, obtenido de https://www.mx.capgemini.com/fundamentos-del-internet-de-las-cosas

[8] Kohlwey, Edmund; Sussman, Abel; Trost, Jason; Maurer, Amber (2011). «*Leveraging the Cloud for Big Data Biometrics»*. *IEEE World Congress on Services*.

[9] **Lantares Solutions,** Información no estructurada: lo que nos enseñan los datos. Obtenido de http://www.lantares.com/blog/informacion-no-estructurada-lo-que-nos-ensenan-los-datos

[10] **Marketing directo,** Datos simples semiestructurados, obtenido de http://www.marketingdirecto.com/marketing-general/marketing/datos-simples-semiestructurados/

[11] **MapReduce.** Obtenido de http://www.tutorialspoint.com/es/hadoop/hadoop_mapreduce.htm

[12] **¿Qué es el Big Data?** Obtenido de http://www.fundacionctic.org/sat/articulo-que-es-el-big-data

[13] **Hadoop.** Obtenido de https://es.wikipedia.org/wiki/Hadoop#cite_note-1

[14] **Qúe es Big Data?.** Obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

[15] **Apache Cassandra.** Obtenido de https://es.wikipedia.org/wiki/Apache_Cassandra#cite_note-5

[16] **Apache Flume.** Obtenido de https://unpocodejava.wordpress.com/2012/10/25/que-es-apache-flume/

[17] **Qué es HBase.** Obtenido de http://www-01.ibm.com/software/data/infosphere/hadoop/hbase/

[18] **Qué es Big data?** Obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

[19] **Utilizando JAQL para analizar Big Data.** Obtenido de https://www.ibm.com/developerworks/ssa/local/im/utilizando-jaql-para-analizar-big-data/

[20] **Qué es Big data?** Obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

[21] **Almacenamiento de datos estructurados.** Obtenido de https://www.ibm.com/developerworks/ssa/library/bd-almacenamiento-datos/

[22] **Qué es un Data Warehouse?** Obtenido de http://www.dataprix.com/que-es-un-datawarehouse

[23] **Stratebi.** Datawarehouse. Obtenido de http://www.stratebi.com/datawarehouse

[24] **Datamart**. Obtenido de http://www.sinnexus.com/business_intelligence/datamart.aspx

[25] **Qué es Cloud Computing?** Definición y concepto. Obtenido de http://www.ticbeat.com/cloud/que-es-cloud-computing-definicion-concepto-para-neofitos/

[26] **Qué es IaaS?** Obtenido de **http://www.interoute.es/what-iaas**

[27] **Qué es PaaS?** Obtenido de **http://www.interoute.com/what-paas**

[28] **Qué es Saas?.** Obtenido de **http://www.interoute.com/what-saas**

[29] **Qué es Almacenamiento en la nube?.** Obtenido de http://aprenderinternet.about.com/od/La_nube/g/Almacenamiento-en-la-nube.htm

[30] **Sinnexus.** Datamining. Obtenido de http://www.sinnexus.com/business_intelligence/datamining.aspx

[31] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds., *Advances in Knowledge Discovery and Data Mining***, MIT Press, 1996.**

[32] **Sinnexus.** Datamining. Obtenido de http://www.sinnexus.com/business_intelligence/datamining.aspx

[33] **IBM.** IBM SPSS Statistics. Obtenido de http://www-01.ibm.com/software/co/analytics/spss/products/statistics/

[34] **SPSS Statitics Standard.** Obtenido de http://www-03.ibm.com/software/products/es/spss-stats-standard

[35] **SPSS Statitics Professional.** Obtenido de http://www-03.ibm.com/software/products/es/spss-stats-professional

[36] **SPSS Statitics Premium.** Obtenido de http://www-03.ibm.com/software/products/es/spss-stats-premium

[37] **Weka 3**. Obtenido de http://www.cs.waikato.ac.nz/ml/weka/

[38] **Oracle Data Miner.** Obtenido de http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html

[39] **Kosala & Blockeel 2000.** Obtenido de

http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

[40] **Data Mining.** Obtenido de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

[41] **Text analysis and knowledge mining system.** Nasukawa, Nagano, 2001. Obtenido de http://webcache.googleusercontent.com/search?q=cache:39rfmuhoguoJ:gcc.uni-paderborn.de/www/wi/wi2/wi2_lit.nsf/94c2e6f98cf9af83c1256bc900524c42/01b957b8f83cd103c1256bbb0044edb0/%24FILE/nasukawa.pdf+&cd=2&hl=es&ct=clnk&gl=co

[42] **Business Intelligence.** Obtenido de http://www.sinnexus.com/business_intelligence/

[43] **Cuadro de mando integral.** Obtenido de http://www.lantares.com/blog/bid/331346/Cuadro-de-Mando-Integral-Todo-lo-que-Debes-Saber

[44] **Sistemas de Soporte a la Decisión.** Obtenido de http://www.sinnexus.com/business_intelligence/sistemas_soporte_decisiones.aspx

[45] **Sistemas de Información Ejecutiva**. Obtenido de http://www.sinnexus.com/business_intelligence/sistemas_informacion_ejecutiva.aspx

[46] **Visualización de datos.** Obtenido de http://www.synergicpartners.com/que-hacemos/disciplines/visualizacion-datos/

[47] **Infografías.** Obtenido de_http://www.ofifacil.com/ofifacil-infografias-que-es-definicion-como-se-hacen.php

[48] **Por qué utilizar Infografías?** Obtenido de https://ernestoolivares.es/por-que-infografia/

[49] **Visualización de datos.** Obtenido de http://www.e-interactive.es/blog/visualizacion-de-datos-10-potentes-herramientas-que-debes-conocer/#axzz49SAnWRZq

[50] **CartoDB y 4 herramientas más de visualización.** Obtenido de http://blogthinkbig.com/visualizacion-de-datos/

[51] **Big Data.** Recuperado de https://es.wikipedia.org/wiki/Big_data#cite_note-20

[52] López García D. ***Análisis de las posibilidades de uso de Big Data en las organizaciones***, Santander, España, 75 p. Trabajo de grado (Máster en Empresas y Tecnologías de

la Información y la Comunicación).2012. Universidad de Cantabria.

[53] **Seguridad de Big Data.** Casos de uso sobre sobre seguridad de Vormetric. Obtenido de http://es.vormetric.com/data-security-solutions/use-cases/big-data-security

[54] **Lo que representa Big data para la seguridad de la información.** Obtenido de http://www.welivesecurity.com/la-es/2014/01/29/que-representa-big-data-seguridad-informacion/

[55] **Big Data: cinco grandes retos en seguridad y privacidad.** Obtenido de http://www.valoresdigital.es/big-data-cinco-grandes-retos-en-seguridad-y-privacidad/

[56] **Almacenamiento en la nube.** Obtenido de http://muyseguridad.net/2014/01/22/nube-cifrado-cliente-servidor/

[57] **Enterprise Secure Key Manager.** Obtenido de http://www8.hp.com/co/es/software-solutions/eskm-enterprise-secure-key-management/

[58] **HP Atalla: seguridad y encriptamiento de los datos.** Obtenido de http://www.cioal.com/2014/06/10/hp-atalla-seguridad-y-encriptamiento-en-todo-el-ciclo-vital-de-los-datos/

[59] **Seguridad de Big Data.** Obtenido de http://es.vormetric.com/data-security-solutions/use-cases/big-data-security

[60] **Administración de claves.** Obtenido de http://es.vormetric.com/products/vormetric-key-management

[61] **Cifrado de aplicaciones.** Obtenido de http://es.vormetric.com/products/vormetric-application-encryption

[62] **Transparent Encryption.** Obtenido de http://es.vormetric.com/products/transparent-encryption