Análisis de las posibilidades de uso de Big Data en las organizaciones

Máster en Empresas y Tecnologías de la Información y la Comunicación

Curso: 2012-1013

Autor: David López García

Email: dlg88.medina@gmail.com

Tutora: Prof. Rocío Rocha Blanco

Analysis of the possibilities of use of Big Data in organizations

Master in Business and Information Technology

Grade: 2012-1013

Author: David López García

Email: dlg88.medina@gmail.com

Professor: Rocío Rocha Blanco

Análisis de las posibilidades de uso de Big Data en las organizaciones

David López García

dlg88.medina@gmail.com

Tutora: Rocío Rocha Blanco

Resumen:

En estos tiempos que corren denominados "la era de la información" en la cual, la sociedad, los clientes y las empresas están cambiando. Estos tres grupos cada vez generan e intentan procesar más y más datos, cantidades que para muchos son imposibles de imaginar. Para lograr adquirir y analizar tanta información surge el término *Big Data*. Un término joven que presenta confusión respecto a su alcance.

En este trabajo se tratará de aclarar en qué consiste, su alcance, como lo utilizan las empresas y en qué situación se encuentra. Además también se abarcará otros términos relacionados con *Big Data*, como pueden ser la minería de datos, el *Cloud Computing* o el *Data Warehouse*. Igualmente también se aclarara porqué surge *Big Data*, de donde procede y por que para muchos tecnólogos sugiere un cambio de etapa en el mundo de las Tics.

Palabras clave:

Big Data, Data Warehouse, Data Mining, Cloud Computing, Hadoop.

Analysis of the possibilities of use of Big Data in organizations

David López García

dlg88.medina@gmail.com

Tutor: Roció Rocha Blanco

Abstract:

In these times called "the information age" in which society, customers and businesses are changing. These three groups are generating and try to process more and more data, which amounts too many, are impossible to imagine. In order to acquire and analyse as much information arises the term Big Data. A young term that produce confusion about its scope.

In this work we attempt to clarify what its scope, as used by businesses and what the situation is. In addition cover other terms related to *Big Data*, such as Data Mining, Cloud Computing or the Data Warehouse.

In addition also clarify that arises Big Data, from which and that to many technologists stage suggests a change in the world of information technology.

Keywords:

Big Data, Data Warehouse, Data Mining, Cloud Computing, Hadoop

ÍNDICE DE CONTENIDOS

Introducción

En la actualidad ha surgido un concepto que para muchas personas ha tenido gran importancia, ya que se ha eliminado una limitación a la tecnología actual. Dicho término se denomina *Big Data* y para innumerables tecnólogos ha nacido para marcar el siguiente gran paso que va a dar el mundo de las Tics. En este trabajo se tratara de explicar dicho término y otros términos estrechamente relacionados con él.

El término Big Data, actualmente sigue generando confusión, es una palabra a la cual se le atribuyen multitud de usos de entre ellas pueden destacar: análisis de redes sociales, análisis de datos en tiempo real, análisis de grandes repositorios de datos, NoSQL… Pero realmente ¿Qué es *Big Data*? Es todo esto y mucho más.

Para pensar en *Big Data*, se tiene que saber que actualmente se vive en la era de la información, con un teléfono móvil en cada bolsillo, un ordenador portátil en cada mochila y grandes sistemas de tecnología funcionando diariamente mandando datos y datos cada segundo, se ve claramente que el mundo tiene más datos que nunca, pero esto no es todo, ya que día a día crece aún más. Un ejemplo claro de esto es el del telescopio *Sloan Digital Sky Survey* construido en el año 2000 en Nuevo México. Durante las primeras semanas este telescopio recopiló más información de los que se habían acumulado en toda la historia de la astronomía, pero esto no es más que un pequeño ejemplo de la gran avalancha que sufrimos en la actualidad. Gracias a esto *Big Data* se está revolucionando el mundo, organizaciones, personas y tecnología

Objetivos del trabajo

El objetivo general de este trabajo es explicar en qué consiste el término *Big Data* y a qué hace referencia. Al ser un término tan amplio en los objetivos generales se explicarán otra serie de términos y cuestiones vinculados estrechamente con él: de donde procede la información, con que tecnología está relacionada, como se utiliza dicha tecnología, *Data Mining*, *Cloud Computing*…

Una vez comprendido que es esté término tan confuso denominado *Big Data* y a que hace referencia, este trabajo se centrará en unos objetivos más específicos que buscan averiguar si es una tecnología que perdurará en el futuro, en qué momento se encuentra y como lo utilizan las grandes compañías actualmente para obtener ventajas competitivas ante sus competidores**.**

Justificación

La justificación de la realización de dicho proyecto, es debido a la gran notoriedad que está teniendo esta tecnología actualmente. Cualquier persona sin o con conocimientos tecnológicos, se pregunta cómo se almacena toda la información que se genera en el mundo: en *Facebook, Twitter, Smartcities* o como *Google* es capaz de manejar todas las transacciones que se hacen a diario. Pero no solo se queda aquí, ya que *Big Data* alcanza todos los ámbitos: bolsa, climatología, astronomía, la cantidad de datos que se genera actualmente es abrumadora y solo el hecho de saber cómo se consigue captar y analizar dicha información me parece una justificación bastante razonable.

Además cuando tuve conocimiento de dicha tecnología, me recordó a la tecnología *Data Warehouse*, la cual me impresionó junto con la utilización de *Data Mining* y *Business Intelligence* en grandes organizaciones utilizando un software tan complejo como es el SAP y que con ello consiguen obtener ventajas competitivas. Visto esto y sabiendo que yo vengo de la rama de empresariales me entusiasmo la idea de averiguar cómo las organizaciones utilizaban *Big Data* y para qué.

Estado del Arte

Una vez vistos los objetivos generales y específicos de esté trabajo, para alcanzar los se ha cogido como referencia diferentes estudios:

En primer lugar a destacar el estudio realizado conjuntamente por *IBM Institute for Business Value* y la Escuela de negocios *Saïd* en la Universidad de Oxford, el cual estuvo basado en el uso de *Big Data* en el mundo real con las empresas más innovadoras. Con esta referencia se mostraran ejemplos de cómo se ve el *Big Data* actualmente por las organizaciones y de hacia dónde se dirige.

Otro estudio que ha servido como referencia es el realizado por *TicBeat* en Octubre de 2012 para así tomar conciencia del término *Big Data*.

El estudio realizado sobre *McAfee* sobre la seguridad de los datos y las brechas también fue un estudio a tener muy en cuenta.

También se pasó a analizar la situación tanto a nivel español como a nivel mundial de diferentes redes sociales para así ver la cantidad de datos que se maneja en ellos.

Por ultimo nombrar el libro "*Big Data* la Revolución de los Datos Masivos" que me hizo comprender de la cantidad de Datos que se generan.

Marco teórico

Conceptos clave para comprender el Big Data

Debido al gran avance que se ha experimentado a lo largo de los últimos años en las tecnologías, más en concreto en el mundo de las tecnologías de la información y la comunicación, lo que comúnmente se denomina Tics, las empresas han tenido que adaptarse a diferentes desafíos, pero existe uno que ha cobrado gran importancia a lo largo de los últimos años. Este desafío consiste en como manipular, administrar, almacenar, buscar y analizar grandes volúmenes de datos. Con el termino *Big Data* hacemos referencia a este gran desafío de las empresas consistente en el tratamiento y análisis de grandes repositorios de dato.

Por lo tanto la primera cuestión a resolver será: ¿Qué es el *Big Data*?, surgiendo luego muchas otras cuestiones como por ejemplo ¿De dónde salen todos esos datos o información?, ¿Cómo llegan *al Big Data?*, ¿Cómo se procesan?, ¿Qué tipo de Software se utiliza? Son preguntas cuyas respuestas se encuentran expuestas con claridad en este documento con el fin de mostrar la importancia *de Big Data.*

¿Qué es el Big Data?

Desde la presentación del término por el MGI (McKinsey Global Insitute) en Junio de 2011 han existido diversos intentos de acotación del concepto.

(Manyika, J y otros, 2011) definen *Big Data* como el conjunto de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos.

Una de las aproximaciones más completas de *Big Data* es la facilitada por Gartner (2012): "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."

Según Wikipedia *Big Data* es término aplicado a: "Un conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable".

Según el artículo "In Perspective" de *Fidelity Worldwide Investment*[2] (2012) *Big Data* es: "el término inglés que designa los conjuntos de datos de gran tamaño y generalmente desestructurados que resultan difíciles de manejar usando las aplicaciones de bases de datos convencionales".

El informe de *TicBeat* (2012) define *Big Data* como: "la enorme cantidad de datos que desde hace unos años se genera constantemente a partir de cualquier actividad."; más adelante dicho informe recalca que: "el *Big Data* bien entendido en la búsqueda del mejor camino para aprovechar dicha avalancha de datos".

Sin embargo un estudio realizado *por IBM Institute for Business Value* junto con la colaboración de *Saïd Business School* (2012) el cual consistió en dar a los encuestados (más de 1144 negocios y profesionales de TI de 95 países y docenas de expertos en la materia) una serie de características sobre *Big Data* para que escogieran las dos que mejor describiera el concepto. El resultado es el visible en Figura 1 "Definición de *Big Data*"

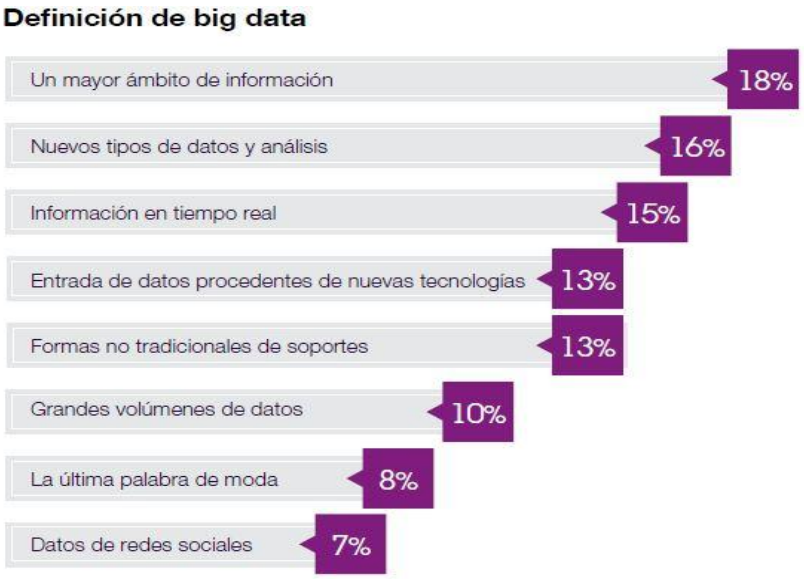

Figura 1 "Definición de Big Data" Fuente: IBM Institute for Business Value

Como se puede observar en la Figura 1, según los encuestados la definición de Big Data tiene sentido con un mayor ámbito de actuación de información y nuevos tipos de datos de análisis mientras que no tiene sentido con datos en redes sociales.

Otros definiciones hacen referencia a la tendencia en el avance de la tecnología que han abierto las puertas hacían un nuevo enfoque del entendimiento y toma de decisiones.

[2] Fidelity Worldwide Investment: es una gestora internacional de fondos de inversión.

Como se puede observar existe gran variedad de definiciones de *Big Data* todas con cierto parecido, pero que en conjunto puede producir cierta confusión sobre el término. Desde mi punto de vista la definición más clara no tiene que hacer ni referencias a nuevas tecnologías como dicen algunos autores, ni a cambios de software. La definición de *Big Data* es: la que se centra en el tratamiento y análisis de grandes volúmenes de datos.

Importancia del Big Data

Con el término *Big Data* se hace referencia a la tendencia del avance de las tecnologías que han abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos que llevaría demasiado tiempo cargarlos en una base de datos relacional para su posterior análisis. Por lo tanto, el *Big Data* se aplicará para toda aquella información que no pueda ser procesada por los métodos tradicionales.

Una base de datos es un conjunto de datos interrelacionados. Cuando se habla de base de datos relacional se hace referencia a la teoría del modelo de datos relacional obra del investigador de IBM *Edgar Codd* en 1970 y que goza de una fuerte base matemática. El modelo relacional se caracteriza a muy grandes rasgos por disponer que toda la información que debe de estar contenida en tablas, y las relaciones entre datos deben ser representadas explícitamente de ese mismo modo. Lo que se consigue con este modelo es trabajar siempre sobre tablas relacionadas entre sí. Evitando duplicidad de registros y garantizando la integridad referencial, es decir que si se elimina un registro, se eliminan todos los relacionados. El gran inconveniente que presenta es el tiempo necesario para manejar grandes cantidades de datos, pero esto se logra gracias al *Big Data.* Por otro lado lo que se consigue al trabajar con bases de datos es combinar diferentes tipos de datos y de una manera formalizada.

Por lo tanto las ventajas de una base de datos relacional se podrían definir en:

- Integridad referencial (sin duplicidad…).
- Normalización (surgen estándar SQL...).
- Permite establecer roles (permisos de entradas a tablas).

No obstante también surgen desventajas de la utilización de bases de datos relacionales, aunque en este trabajo solo se va a nombrar tres:

- Cantidad de manejo de datos limitada.
- Lectura exclusiva de lenguajes estructurados.

- Orientadas a satisfacer objetivos de aplicaciones anteriores.

Estas tres desventajas nombradas anteriormente las resuelve *Big Data*, gracias a que su estructura es capaz de almacenar y procesar grandes cantidades de datos y de los tres tipos de datos posibles (estructurados, semi-estructurados y sin estructurar) además es una arquitectura orientada a los programas actuales.

Beneficios del Big Data

Una vez que se sabe la importancia de *Big Data* sobre todo gracias a la mejora que supuso respecto a los modelos relaciónales se citaran los beneficios más habituales del *Big Data*, no obstante estos benéficos no se tienen porque aplicar a todas las organizaciones, ya que cada organización tiene y actúa en diferentes condiciones.

A continuación se citan los beneficios e inconvenientes más relevantes que han sido extraídos de un artículo publicado en Eureka-startups (2013) por *Vauzza*:

- Gestión del cambio**:**
  - Búsqueda de nuevas oportunidades de negocio a través de segmentación mejorada y venta cruzada de productos (mejora de la estrategia).
  - Mediante la aplicación de análisis y modelado predictivo a los datos de cuentas de clientes e historial de transacción, la solución permite a los agentes llevar a cabo una segmentación basada en la probabilidad de que el cliente contrate servicios o productos complementarios, o contratar servicios de mayor valor (mejora de segmentación).
  - Mediante el análisis de consumo de los servicios y productos de los clientes, la empresa puede optimizar las estrategias de venta cruzada, afinar mensajes de marketing y proporcionar ofertas específicas. Se puede predecir con mayor exactitud qué productos son los más apropiados para cada cliente (mejora de la estrategia).
  - Ofrecer la combinación adecuada de servicios y productos mejora la eficacia y la eficiencia de la fuerza de ventas de la compañía, mientras que el toque más personalizado ayuda a los agentes a forjar lazos más estrechos con clientes, lo cual mejora la lealtad (mejora de la estrategia).

    - Mejoras Operativas: Mayor capacidad de visibilidad del negocio a través de informes más detallados.
- Análisis de navegación web y hábitos de consumo online:
    - Análisis de Redes Sociales: Determinar los círculos sociales de los clientes a partir de interacciones telefónicas y redes sociales online genera una visión completa de los clientes, identificando el papel que desempeñan en sus círculos y su grado de influencia.
    - Marketing Viral (marketing que explota redes sociales…): Detecta clientes más influyentes, roles sociales… para maximizar la difusión de tus productos y servicios (mejor conocimiento de clientes y del mercado en redes sociales).
    - Análisis de datos de navegación: Analiza la navegación Web y hábitos de consumo online: extrae nuevas y valiosas perspectivas de los clientes. Se identifica al usuario (localización, estado del terminal, servicios de acceso), se monitorizan sitios y búsquedas por palabra, urls visitadas, tiempo de navegación, etc. (mejor conocimiento del cliente).
    - Cuadro de Mandos en tiempo real, la información siempre está disponible sin esperas de actualización de los datos (información en tiempo real).
- Anticipación a los problemas:
    - Un sistema predictivo de análisis y cruce de datos nos permite poder anticiparnos a posibles problemas que puede surgir en el futuro, como por ejemplo una predicción de riesgo de catástrofes que permitiría ajustar la política de precios y aprovisionar fondos para posibles pagos (utilidad para ver la veracidad de los datos ante datos imprecisos) .
- Mejoras de Procesos:
    - Permite la simplificación de procesos actuales y control del negocio (reducción de costes).
    - Análisis de Seguridad. Analítica proactiva que permite la reducción de riesgos y pérdidas frente a fraudes (reducción de costes).
    - Permite detectar patrones complejos de fraude en tiempo real analizando los datos históricos, el patrón de uso de información de geolocalización, análisis de transacciones y operaciones sospechosas (reducción de costes).
- Soporte a la toma de decisiones a través de algoritmos automáticos.

- Una analítica sofisticada que analice todos los informes y datos, ayuda a la toma de decisiones, reduciendo los riesgos y descubre información que antes podría estar oculta, pero a la vez importante (ayuda a la toma de decisiones).

- Reducción de costes.
- Reducción de tiempos.
- Desarrollo de nuevos productos.
- Ofertas optimizadas y personalizadas.
- Tomas de decisiones más inteligentes que con los anteriores sistemas *Business Intelligence.*
- Filtros inteligentes de seguridad en el negocio electrónico

Todas estas ventajas se pueden agrupar en una principal que se derivan en todas las demás ventajas: "obtener más información/conocimiento" de los clientes de la propia empresa, inclusive de la propia empresa y la competencia para obtener una ventaja competitiva respecto a los competidores ofreciendo a los clientes lo que quieren o incluso a crear una necesidades que los clientes aun no tienen.

Cuando se hace referencia a "obtener más información/conocimiento" no se refiere a una gran cantidad de datos, sino que hay que diferenciar entre datos-Información-conocimiento.

A continuación se muestra las diferencias de esos 3 elementos: datos, información y conocimiento (Figura 2):

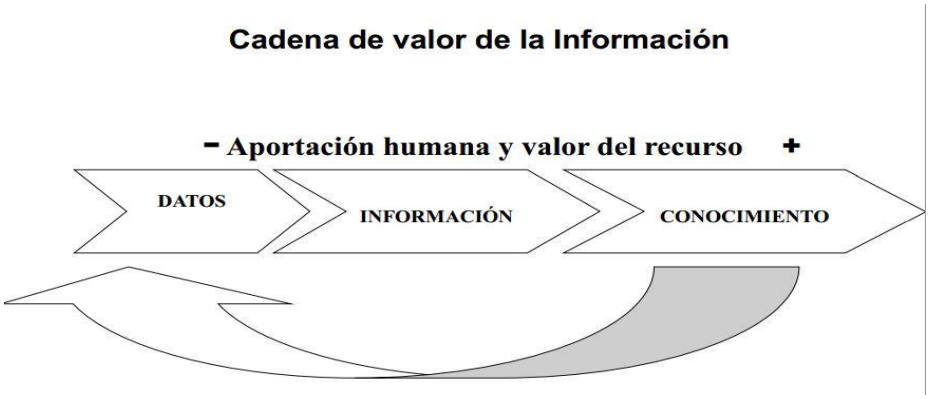

Figura 2. El conocimiento. Fuente: Máster Empresas y Tecnologías de la Información, Daniel Pérez (apuntes de clase 2013/1014)

- *Dato* es un elemento primario de información que por sí solos son irrelevantes para la toma de decisiones. La manera más clara de verlo es con un ejemplo. Un número de teléfono o un nombre de una persona, son datos, que sin un propósito o utilidad no sirven para nada.

- *La información* se puede definir como un conjunto de datos procesados y que tiene relevancia o propósito y que por lo tanto son de utilidad para las personas que la utilizan para la toma de decisiones.
- *El conocimiento* es una mezcla de experiencias, valores, información y know-How que aplicaran los conocedores de este para la toma de decisiones.

Donde realmente entra en juego *Big Data* es en el proceso de encontrar la información la cual puede ser transformada en conocimiento entre esas grandes cantidades de datos recolectadas por las organizaciones y no en cómo se recolectan esos datos. La visión optimista de un *Big Data* perfecto será aquel en el que las empresas serían capaces de obtener datos de cualquier fuente, aprovechar esos datos y obtener la información que se convertiría en conocimiento útil para la organización permitiendo incorporar todas las ventajas anteriormente nombradas.

Inconvenientes Big Data

No obstante no hay que olvidarse de los inconvenientes del Big Data. Siendo el principal de ellos el proceso de adopción de *Big Data*: software y hardware necesario y su coste. Pero además existen otros muchos de menor peso como por ejemplo:

- Rechazo por parte del personal.
- Gasto de formación.
- Colaboración necesaria por parte de todos los departamentos.
- La denominada "Toma de decisiones pasivas", esto hace referencia antes de la instalación de Big Data, a que las empresas primero esperan a que lo instalen sus competidores para ver que errores comentes con la creencia de que ellos lo podrán adoptar mucho más rápido.
- Coste.
- Problemas de privacidad
- Problemas de información desactualizada.
- Filtrado (no todos los datos son información).

A parte de estos, hay que considerar un gran inconveniente antes de realizar un proyecto de *Big Data* y que es tan sencillo como saber sí: ¿Es realmente útil para la organización? ¿La empresa tiene necesidad de *Big Data*? ¿Se cuenta con los recursos necesarios para afrontar un proyecto de *Big Data*? ¿Cuánto costará?, es decir, ¿Mi empresa realmente necesita *Big Data*? a pesar de todos los beneficios

que me puede proporcionar. Sí la respuesta es “SI” los inconvenientes no deberían importar puesto que las ventajas que se obtienen serán mucho mayores.

Aplicaciones del Big Data

Las gran cantidad de aplicaciones de *Big Data*, solo viendo el alcance que tiene puede ser incalculable, no obstante, el análisis realizado por IBM en la Figura 3, muestra las 5 orientaciones preferentes a la hora de aplicar *Big Data* en organizaciones en la que el 49% de las organizaciones prefieren aplicar *Big Data* para centrarse en el cliente, el 18% en optimización operativa, el 15% en gestión financiera y de riesgo, el 14% en el nuevo modelo empresarial y un 4% en colaboración empresarial.

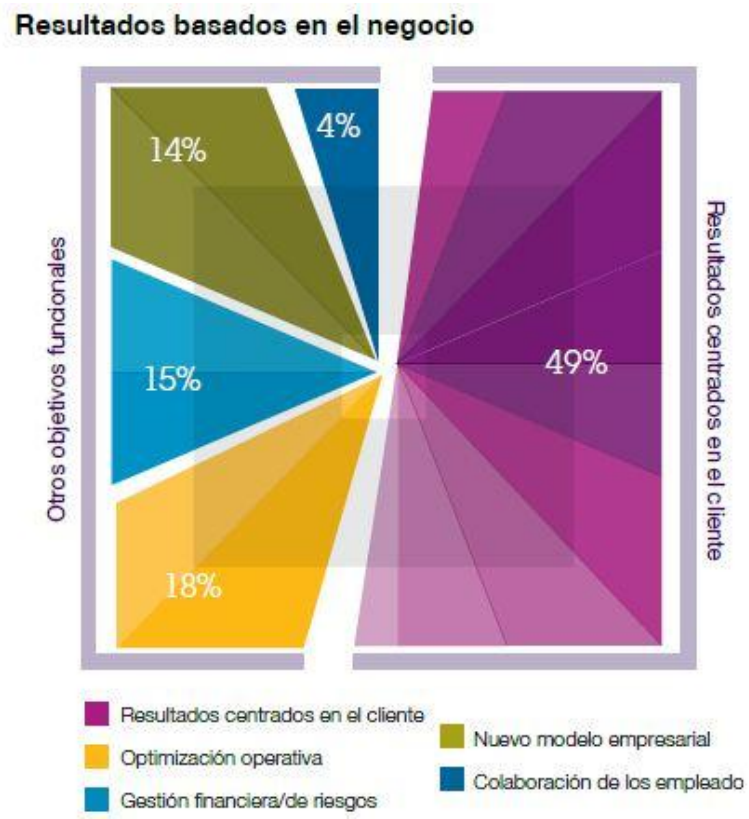

Figura 3 Orientación Big Data. Fuente: IBM. Año 2012

Pero para no limitarnos a esas 5 orientaciones dedicadas de *Big Data*, se nombraran a continuación diferentes aplicaciones:

- Determinar las causas de los fracasos, los problemas y defectos en tiempo casi-real, potencial el ahorro de miles de millones de dólares anuales.
- Optimizar las rutas de muchos miles de vehículos de entrega de paquetes mientras están en la carretera.
- Genera cupones de compras en el punto de venta en base a compras anteriores y actuales de los clientes.
- Enviar las recomendaciones a la medida para dispositivos móviles, mientras que los clientes están en el área de derecho de aprovechar las ofertas.
- Todo Recalcular carteras de riesgo en cuestión de minutos.
- Identificar rápidamente los clientes que más importan.
- Utilice el análisis de clics y la minería de datos para detectar comportamientos fraudulentos.
- Consulta y generación de informes.
- Extracción de datos.
- Visualización de datos.
- Analítica geoespacial, de Streming, de video, de voz, texto de lenguaje natural.

¿Qué cantidades de datos hacen referencia a *Big Data*?

Como bien se ha mencionado con anterioridad, con el término *Big Data*, se hace referencia al tratamiento y análisis de grandes repositorios de datos, estos repositorios varían su tamaño, estaríamos hablando para hacernos una idea en términos en bytes de Tabla 1:

Gigabytes= 1000000000 = 10^9 bytes
Terabytes=1000000000000 = 10^{12} bytes
Peta bytes=1000000000000000 = 10^{15} bytes
Exabytes=1000000000000000000 = 10^{18} bytes
Zettabytes=1000000000000000000000 = 10^{21} bytes
Yottabyte=1000ZB

Tabla 1 “tamaños en bytes” Fuente IBM (2012); Raul G. Beneyto (2013)

Como ya se ha dicho con anterioridad, actualmente se está en la época de la información y lo que hoy parece mucha información, en unos pocos años parecerá poca, por este motivo, este apartado solo puede servir como referencia actualmente. Además del gran volumen de información, existe una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por

ejemplo: dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo suficientemente rápida para lograr obtener la información correcta en el momento preciso.

Para hacerse una idea de lo abrumadora que es la cantidad de datos que se generan pongo el ejemplo de la empresa Domo[3], la cual hizo un análisis en el año 2012 de la cantidad de información que los internautas dan de uso a la red cada minuto: "Cada minuto que pasa, los 2.700 millones de personas con acceso a Internet que se calcula que hay actualmente en el mundo, envían más de 200 millones de correos electrónicos, realizan 2 millones de consultas a Google, suben 48 horas de vídeo a YouTube, escriben más de 100.000 mensajes en Twitter, publican casi 30.000 nuevos artículos en sitios como *Tumblr* o *WordPress*, suben más de 6.000 fotografías a *Instagram* y *Flickr*, se descargan 47000 aplicaciones del sistema operativo IOS…"

A continuación se muestra un dibujo explicativo (Figura 4).

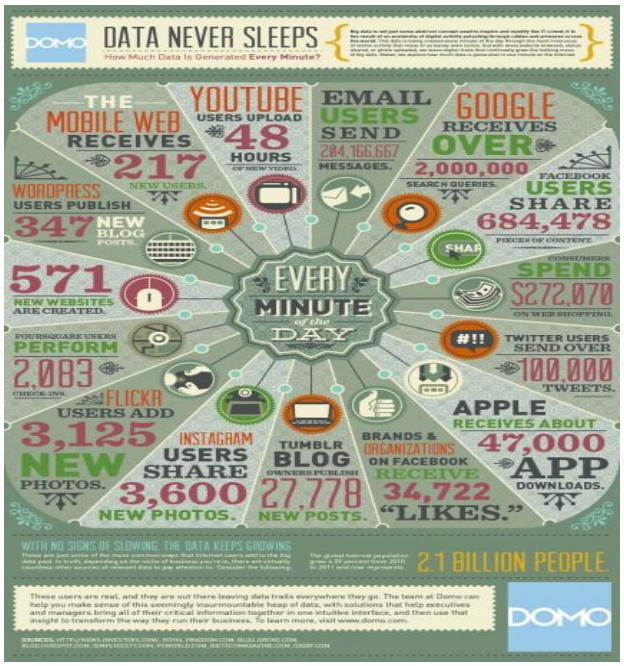

Figura 4. "Every Minute of the Day" Fuente: Domo 2012.

[3] Fuente: FRIDAY, JUNE 8, 2012 http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/ Josh James Founder, CEO & Chairman of the Board Big Data

¿De dónde proviene toda la información que obtendremos mediante el *Big Data*?

Los seres humanos crean día a día cada vez más y más información, toda esta información proviene de diferentes lugares: redes sociales, *Smartphones*, *Smart Cities*, las empresas, las denominadas comunicaciones M2M (*machine to machine*), sensores digitales ya sean de medición eléctrica o de temperatura, sísmicos, se estima que existen más de 30 millones de sensores interconectados en diferentes sectores y se espera que este número crezca anualmente un 30%. Como se observa en el análisis realizado por la empresa Domo, cada minuto se genera multitud de información.

A continuación se verán ejemplos para que quede patente la importancia de *Big Data* realizadas por la agencia de marketing online Concepto 05 en 2013 en una estadística de redes sociales en España (2013).

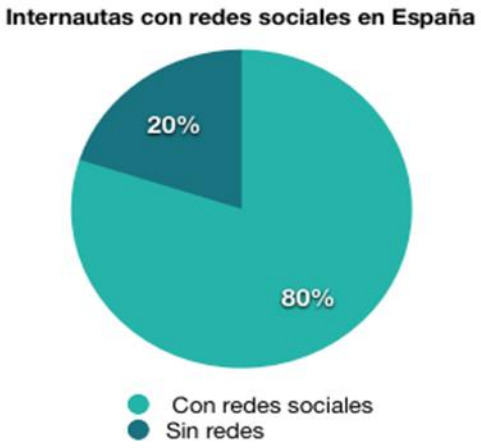

Figura 5. Estadística redes sociales. Fuente: Concepto 05

Como se puede observar en esta publicación de la figura 5, realizada en marzo de 2013, solo el 20 % de los internautas españoles no tiene ninguna red social. Para ser más exactos como se muestra en la figura Nº6 en 2012 el 67,9% de los hogares españoles estaban conectados a Internet, según los resultados de la encuesta sobre equipamiento y uso de tecnologías de la información y comunicación en los hogares publicados por el INE. Este indicador continúa su tendencia de crecimiento de los últimos años, y en el último año el porcentaje de hogares conectados a Internet ha crecido 4 puntos porcentuales.

En España existen 10,4 millones de viviendas familiares que tienen acceso a Internet, con un aumento de medio millón de hogares respecto al año 2011.

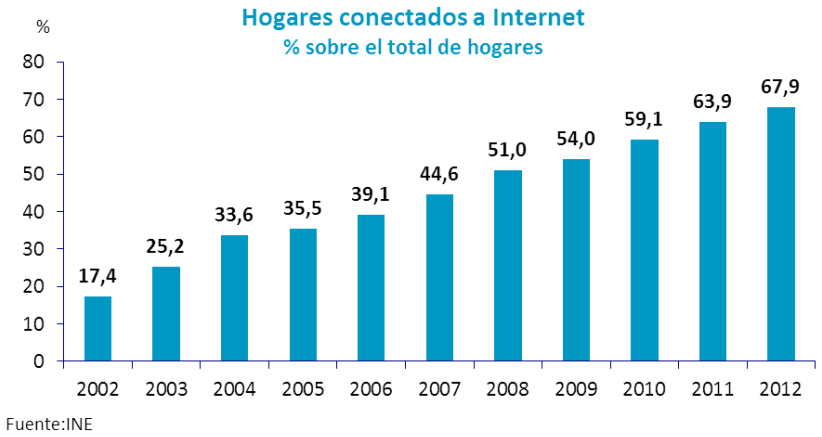

Figura 6. Hogares conectados a Internet. Fuente: Concepto 05; INE 2013; ONTSI

Datos redes sociales en todo el mundo

Facebook, tras superar los 800 millones de usuarios en todo el mundo, empieza a facilitar datos sobre usuarios activos. Gracias a ello, sabemos que desde el pasado mes de marzo esta red cuenta con 1.110 millones de usuarios activos mensuales en todo el mundo. Por su parte, *Twitter* no facilita datos de usuarios desde 2011, año en que rozaba los 200 millones de usuarios en el mundo, de los cuales 100 millones eran activos. Algunos estudios sostienen que a finales de 2012 habría superado los 485 millones de usuarios, de los cuales 288 serían usuarios activos.

Al llegar el verano de 2012 *Tuenti* rebasaba los 14 millones de usuarios. En esa misma fecha lanza sus primeros anuncios como distribuidora de telefonía móvil, superando en los primeros meses de 2013 los 100.000 clientes.

Desde su aparición en fase beta hacia abril del 2011, *Google Plus* ha experimentado uno de los mayores crecimientos en número de usuarios, llegando a los 500 millones a finales del 2012. Además asegura ostentar una tasa de usuarios activos del 47%. Con este dato se despejan algunas de las dudas sobre la supervivencia de esta red social, que en su primer año solo alcanzaba un 14% de usuarios activos, unos 9 millones. Otra de las redes que ha duplicado su número de usuarios en el último año es *LinkedIn*, llegando a los 225 millones.

Datos de redes sociales en España:

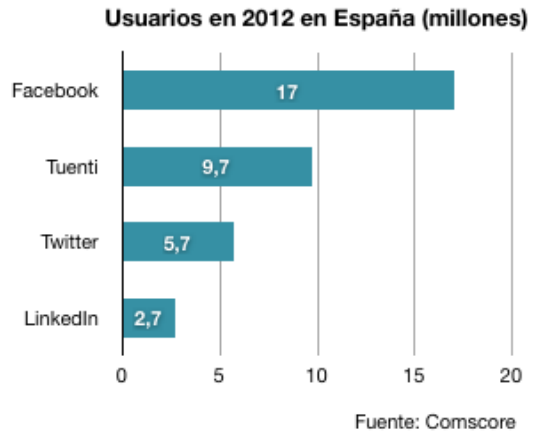

Figura 7. Usuarios 2012 en España. Fuente: Concepto 05; ComScore (2012)

Según **Comscore**[5](2012)**,** empresa de medición en Internet. Sus datos son útiles, ya que ofrecen una buena aproximación del número de usuarios activos anual de las diferentes redes sociales en España. Como se puede observar en la Figura 7 *Facebook* va en cabeza con 17 millones de usuarios activos, seguido por *Tuenti*, *Twitter* y *LinkedIn.*

Con las cantidades de usuarios activos que están en las redes sociales se trata de dar a conocer gran cantidad de información que se almacena.

Pero recordar que *Big Data* no son solo redes sociales o Smartphones, sino que abarca a todos los ámbitos imaginables, un claro ejemplo es el citado por *Kenneth Cukier* (2010) que también podemos encontrar en el libro de *Victor Mayer-Schönberger* (2013): el telescopio de *Sloan Digital Sky Survey* en Nuevo México construido en 2000. Durante las primeras semanas este telescopio recopiló más información de los que se habían acumulado en toda la historia de la astronomía. Para el año 2010 el

[5] ComScore: comScore es una empresa líder en la medición de internet que proporciona análisis para el Mundo Digital™. ComScore mide cómo navegan las personas en el mundo digital – y convierte estos datos en información y acciones para que nuestros clientes maximicen el valor de sus inversiones digitales. comScore fue fundada en 1999 por el Presidente y CEO Magid Abraham y el Chairman Gian Fulgoni. comScore se convirtió en una empresa pública en junio 2007.

archivo del proyecto constaba de 140 terabytes de datos. Sin embargo, su futuro sucesor, el telescopio previsto para 2016 denominado *Sinóptico* acopiara esa cantidad de datos en 5 días.

Características Big Data

¿Por qué es tan revolucionario el término Big Data? Existen diversas razones pero tiene tres características que destacan sobre todas las demás y que lo hacen ser único:

- Volumen
- Velocidad
- Variedad

Las denominas 3 V del *Big Data*:

Volumen: Suele utilizarse como sinónimo de Big Data. A pesar de ser uno de los aspectos más llamativos, no es el único. El reto relacionado con el volumen de datos se ha puesto de manifiesto recientemente, debido a la proliferación de los sistemas de información e inteligencia, el incremento del intercambio de datos entre sistemas y dispositivos nuevos, nuevas fuentes de datos, y el nivel creciente de digitalización de los medios de comunicación que antes sólo estaban disponibles en otros formatos, tales como texto, imágenes, videos y audio.

La cantidad de datos día a día será muy superior a las que actualmente existen con lo cual se obtiene un valor añadido d. Las empresas están cubiertas de una cantidad cada vez mayor de datos de todo tipo, acumulando fácilmente terabytes, incluso peta bytes, de información.

Velocidad: Se asocia con la proliferación de nuevas fuentes de datos, y la necesidad de utilizar estos datos más rápidamente. Fuentes de datos automatizados, tales como sensores, RFID, GPS generan datos cada fracción de segundo para varias métricas diferentes y, junto con otros equipos de la empresa, causan un flujo constante de datos que se generan con el tiempo. Los dispositivos que generan datos a intervalos más largos, tales como los teléfonos inteligentes, también terminan generando corrientes constantes de datos que necesitan ser ingeridos rápidamente. Por otro lado, todos estos datos tienen poco o ningún valor si no se convierten rápidamente en información útil.

Variedad: Los grandes volúmenes de datos incluyen cualquier tipo de datos, estructurados y no estructurados como texto, datos de sensores, audio, vídeo, secuencias de clic o archivos de registro, entre otros. Al analizar estos datos juntos se encuentra información nueva.

Para explicar esta característica, la mejor opción es imaginar la creciente cantidad de información que almacena *Facebook* sobre sus usuarios y lo diversa que es esta. En su base de datos se puede encontrar la edad, el sexo o el país de millones de personas. Con *Big Data* esto es posible de hacer. Está característica está relacionada con la organización de los datos. Esta organización se divide básicamente en datos estructurados, semi-estructurados y no estructurados. Los datos estructurados son los datos tradicionalmente presentes en los sistemas corporativos (bases de datos, archivos jerárquicos y secuenciales, etc.), los datos semi-estructurados suelen estar disponibles a través de los registros del sistema (servidores web, CDR, etc.) y los datos no estructurados se relacionan principalmente, con el contenido digital más reciente, y se pusieron a disposición previamente en un formato no digital, tales como archivos de imagen, audio, texto, entre otros. El universo del *Big Data* contempla la posibilidad de utilizar todos los datos disponibles a través de correos electrónicos, documentos, mensajes, imágenes, grabaciones de audio, registros, videos, etc.

Existe la posibilidad para alguno autores de una cuarta "V" la Veracidad, hace referencia al nivel fiabilidad asociado a ciertos tipos de datos, es decir, *Big Data*, lleva asociado un factor de incertidumbre ante ciertos datos, por ejemplo, los sentimientos y sinceridad de los seres humanos, condiciones climatológicos. Para dejar más claro está cuarta "v" vamos al sector de la producción energética, en muchos países existe la normativa de que parte de la producción energética proceda de fuentes renovables, pero la climatología no se puede predecir con precisión, no se puede saber la cantidad de viento que soplara para que los molinos eólicos generen X cantidad de energía. Para solucionar esto los analistas de *Big Data* tienen diferentes soluciones para así obtener unos datos más fiables. Esta soluciones son: matemáticas avanzadas (Técnicas de optimización…), fusión de datos de múltiples fuentes menos fiables, etc.

Otra característica importantes q define al *Big Data* es la Complejidad está relacionada con la forma de tratar con todas las características mencionadas anteriormente, para brindar información útil de manera eficiente.

Muchos proveedores explotan sus características técnicas para almacenar grandes volúmenes de datos y se centran en las características aisladas de *Big Data*, sin mostrar cómo hacerlo de una manera integrada y sencilla. Otros piensan en Big Data como *Data Warehouse* o *Business Intelligence*, pero el mayor potencial de *Big Data*, es la capacidad de hacer el análisis avanzado de estos datos, que también se llama *Big Data Analytics*.

Otros conceptos relacionados con Big Data

Ahora que ya se sabe en qué consiste *Big Data* y a qué hace referencia se hablará de tecnología estrechamente relacionada con él. Tan relacionada que se puede decir que ha surgido a partir de ella, en el caso de *Data Warehouse* o *Data Mining* y otra que la complementa como es el *Cloud Computing*.

Data Warehouse [6]

El *Data Warehouse* es una evolución de los sistemas de bases de datos relacionales, es un proceso, no un producto. En 1988 los investigadores de IBM *Barry Devlin y Paul Murphy* inventaron el término *Warehouse* de información, aunque el considerado padre de los Data Warehouse es *William Harvey Inmon*.

Los *Data Warehouse* fueron creados en la década de los 90e y son un conjunto de datos que las organizaciones utilizan de apoyo para la toma de decisiones y que las cuales al mismo tiempo pueden consultar mediante las tecnologías de los *Data Mining*.

La definición de William Harvey Inmon dice: "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".

Características:

De esta definición podemos destacar 4 características principales: Organizado por temas, integración, no volátil y temporalidad.

- Organizado por temas:

La organización por temas hace referencia al hecho de que los datos en el Data Warehouse no se organizan acorde con las aplicaciones que los usan, sino que lo hacen acorde con su semántica, independientemente de que aplicación los utilice. Por ejemplo, una compañía podría tener datos organizados por clientes, proveedores, productos, etcétera, independientemente de la aplicación que los vaya a utilizar.

- La integración (característica más importante según su autor). Un *Data Warehouse* se construye a partir de los datos de las diversas fuentes de datos de una organización, lo que hace necesario un esfuerzo para "poner en común" todo este dato proveniente de las diferentes fuentes.

[6] Esté artículo está recogido de BuenasTareas de un Trabajo realizado por mí, junto con Víctor Pacheco para el Máster en Empresas y Tecnologías de la Información (2012-2013).

Cada una de las fuentes de datos de la organización tendrá sus propios modelos de datos, sus propias políticas de asignación de nombres a campos, de codificación de valores, y un largo etcétera de diferencias que hacen que el hecho de recolectar los datos de ellas para unirlos en un esquema común suponga un gran esfuerzo, tanto computacional como humano.

La información de *los Data Warehouse* proviene de: El sistema transaccional (contabilidad, ventas…).

- Datos de fuentes externas. Toma como punto de partida la información recogida en el sistema transaccional convirtiéndola en datos históricos y no modificables, sobre los que se realizaran las tendencias y provisiones.
- No volatilidad: Existen varias razones por las que los datos de un Data Warehouse no sean volátiles. Las más importantes son:

Un *Data Warehouse* se construye para dar soporte a la toma de decisiones, y este tipo de tareas pueden requerir el análisis de datos de diferentes momentos del tiempo para realiza análisis comparativos. Mantener diferentes versiones temporales de los datos permite recuperar el estado de los datos de la organización en cualquier instante, de modo que se pueden deshacer efectos indeseados de procesamientos erróneos. Por lo tanto un los datos de un *Data Warehouse* no sufren actualizaciones sino que se mantienen diferentes versiones de dichos datos.

- Temporalidad: En los sistemas tradicionales, la caducidad de los datos, o su validez no suele exceder de dos o tres meses. En muchos casos los datos varían todos los días. Mientras tanto, los datos del Data Warehouse tienen un horizonte temporal de años (Entre 5 y 10 años). En los sistemas de gestión, los datos con los que se trabaja son los datos actuales, mientras que los datos del Data Warehouse pueden verse como una serie de "*snapshots*" tomados en un momento del tiempo, que no sufren actualizaciones.

La estructura de los datos operacionales puede contener, o no, alguna referencia temporal. En cambio, la fecha siempre forma parte de la clave de los datos en el *Data Warehouse*, para distinguir las diferentes versiones de los datos, como ya se había mencionado.

Por último no debemos olvidarnos de que los sistemas *Data Warehouse* son un sistema de apoyo en la toma de decisiones que junto con los *Data Mining* son una importante herramienta empresarial actualmente.

Ventajas e inconvenientes Data Warehouse

Con base en lo anterior podemos obtener los siguientes beneficios:

- Apoyo en la toma de decisiones de la empresa a cualquier nivel jerárquico.
- Proporcionar mejores productos al mercado a través de la optimización de tiempos de producción y toma de decisiones.
- Analizar información de ventas a diario permitiendo agilizar la toma de decisiones que puedan afectar el desempeño o proyección de la empresa.
- Complemento fundamental de la minería de datos.

Por otra parte, su empleo supone los siguientes inconvenientes:

- Implementar un *Data Warehouse* implica un alto costo y no suele ser estático necesita mantenimiento que su costo es elevado. Costos de adaptación de la empresa, formación, mantenimiento, coste del Software y hardware.
- Incluso pueden quedar obsoletos en cualquier momento. Se confunde con sistemas operacionales por que cumplen con algunas funciones parecidas al *Data Warehouse* pero puede resultar peor por algunas funciones son muy caras o que no se usen muy repetidamente.
- Capacidad limitada.

El principal inconveniente del *Data Warehouse* es la necesidad de adaptar toda la empresa para acoger al Data Warehouse lo cual resulta enormemente costoso.

Aplicaciones y funciones en la empresa del Data Warehouse

Dentro de las empresas que empleen este tipo de software para optimizar sus tareas, es importante, dentro la tecnología *Data Warehouse*, diferenciar dos tipos fundamentales de sistemas de información que se dan en todas las organizaciones. Nos referimos ahora a los sistemas técnico - operacionales y los sistemas de soporte de decisiones. Cabe destacar que es este último el que conforma toda la base del software *Data Warehouse*.

En primer lugar vamos a comentar los sistemas técnico - operativos, que cubren el núcleo de operaciones tradicionales de captura masiva de datos (Data Entry) y servicios básicos de tratamiento de datos, con tareas predefinidas (contabilidad, facturación, almacén, presupuesto, personal y otros sistemas administrativos). Gracias al *Data Warehouse*, estos sistemas están evolucionando junto con la irrupción de censores, autómatas, sistemas multimedia y bases de datos relacionales más avanzadas.

Para continuar, mencionaremos la importancia de los Sistemas Estratégicos, orientados a soportar la toma de decisiones, facilitando la labor de la dirección, proporcionándole un soporte básico, en forma de mejor información, para la toma de decisiones. Se caracterizan porque son sistemas sin carga periódica de trabajo, es decir, su utilización no es predecible, al contrario de los casos anteriores, cuya utilización es periódica.

Destacan entre estos sistemas: los Sistemas de Información Gerencial (MIS), Sistemas de Información Ejecutivos (EIS), Sistemas de Información Georeferencial (GIS), Sistemas de Simulación de Negocios (BIS y que en la práctica son sistemas expertos o de Inteligencia Artificial - AI).

Conviene señalar también los Sistemas Tácticos, los cuales fueron diseñados para soportar las actividades de coordinación de actividades y manejo de documentación, definidos para facilitar consultas sobre información almacenada en el sistema, proporcionar informes y, en resumen, facilitar la gestión independiente de la información por parte de los niveles intermedios de la organización.

Destacan entre ellos: los Sistemas Ofimáticos (OA), Sistemas de Transmisión de Mensajería (Correo electrónico y Servidor de fax), coordinación y control de tareas (*Work Flow*) y tratamiento de documentos (Imagen, Trámite y Bases de Datos Documentales).

Finalmente aparecen los Sistemas Interinstitucionales, los cuales está surgiendo recientemente como consecuencia del desarrollo organizacional orientado a un mercado de carácter global. Dicho mercado obliga a pensar e implementar estructuras de comunicación más estrechas entre la organización y el mercado (Empresa Extendida, Organización Inteligente e Integración Organizacional), todo esto a partir de la generalización de las redes informáticas de alcance nacional y global (Internet), que se convierten en vehículo de comunicación entre la organización y el mercado sin importar la distancia.

Por otra parte, hay otras funciones dentro de la empresa que tienen que ver con el planeamiento, previsión y administración de la organización. Estas funciones son también críticas para la supervivencia de la organización, especialmente en nuestro mundo de rápidos cambios.

Para concluir con este apartado vamos a comentar las funciones basadas en el conocimiento, formadas por los sistemas de toma de decisiones ya que son estos sistemas sobre los que se basa la tecnología *Data Warehouse*.

Las funciones como "planificación de marketing", "planeamiento de ingeniería" y "análisis financiero", requieren de sistemas de información que las soporten.

Estos sistemas están relacionados con el análisis de los datos y la toma de decisiones, frecuentemente, decisiones importantes sobre cómo operará la empresa, ahora y en el futuro. Estos sistemas tienen un alcance bastante grande al almacenar una cantidad casi incontable de datos.

Los datos para el soporte de decisiones, con frecuencia, toman un número de áreas diferentes y necesita cantidades grandes de datos operacionales relacionadas, de ahí que se empleen, cada vez con mayor grado de importancia, los ya mencionados y aclarados almacenes de datos denominados Data Warehouse.

Tecnologías y software de Data Warehouse

Optimizar la elaboración de las tareas dentro de la empresa, así como llevar a cabo una satisfactoria toma de decisiones depende de muchos e innumerables factores. Para simplificar todo este dilema, distintas y múltiples empresas han elaborado software. El inconveniente, como ya vimos anteriormente, sigue siendo el importante desembolso económico que supone su aplicación.

Con el objeto de argumentar lo expuesto en el párrafo anterior a continuación se presentan varios ejemplos de algunos de este software desarrollados por destacadas empresas.

POWERCENTER INFORMÁTICA	
PowerData ofrece con *Powercenter* Informática la posibilidad de contar con una vista única y exhaustiva de los activos de la información crítica de la empresa y gestionar la complejidad de esos datos, incluyendo variables como volúmenes, latencias, múltiples formatos y estructuras, gracias a *PowerCenter* de Informática Corporation, líder mundial en tecnologías para la gestión de datos. *PowerCenter* Informática facilita que los datos estén disponibles en el momento y en la forma precisa para aumentar así la eficiencia operativa de su compañía. Precio: 140.000 $ http://www.informatica.com/es/products/enterprise-data-integration/powercenter/	

BI4Dynamics NAV	

BI4Dynamics NAV es un software de Business Intelligence que se considera el mejor en relación a tareas de reporte de datos y análisis. El sistema BI4Dynamics NAV soluciona este hecho utilizando las herramientas de Data Warehouse, el "corazón" de la solución. La mayoría de los expertos de inteligencia de negocio consideran esta solución como la mejor infraestructura para apoyar iniciativas estratégicas.

Precio: 100.000$ http://www.bi4dynamics.com/

LITEBI	

Litebi es un software completo de Business Intelligence ofrecido en modalidad SaaS, para poder contrarrestar el elevado coste y complejidad del software BI tradicionales. Con el sistema Litebi, usted podrá crear soluciones analíticas en tres pasos:

- *Define qué quieres analizar y construye un Data Warehouse completo con nuestros editores web.*
- *Carga datos complejos desde cualquier origen con liteIntegrator.*
- *Analiza tu información utilizando herramientas web de análisis potentes y fáciles de utilizar.*

Precio del alquiler: 150 euros al mes.

http://www.parqueinnova.com/pages/empresas/tec.-informacioacuten/litebi-s.l.php

ORACLE BUSINESS INTELLIGENCE SUITE	

Oracle Business Intelligence Suite es una solución integrada de productos de inteligencia empresarial (BI) con cuadros de mando, un completo sistema de consultas ad hoc,

alertas e información proactiva, informes financieros y corporativos, datos predictivos en tiempo real y análisis desconectado entre otras funciones.
Precio: 28.000 euros.

http://www.oracle.com/us/solutions/ent-performance-bi/enterprise-edition-066546.html

Como conclusión se puede decir que el empleo de los *Data Warehouse* es un avance importante en grandes empresas multinacionales que manejen un gran volumen de datos procedentes de diversas bases de datos. A pesar del gran desembolso que supone la implantación de esta herramienta, a largo plazo, (entre 5 y 10 años), supone una inversión eficiente desde el punto de vista de la toma de decisiones. No obstante este tipo de herramientas se consideran innecesarias en empresas que manejen un reducido volumen de datos. En el hipotético caso de que dichas empresas manejen gran volumen de datos, tendrían la posibilidad de subcontratarlo en la modalidad de renting o inclusive utilizar otro tipo de modelo de base de datos, más apropiada a su tamaño.

Data Mining o Minería de Datos

La minería de datos es una herramienta que permite extraer conocimiento de los datos que tenemos almacenados para tratarlos y convertirlos en información útil y objetiva que ayudará al empresario tomar las decisiones más adecuadas.

La definición de la minería de datos dada por (Fayad y otros 1996): "es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos"

Según el portal *Daedalus* la Minería de Datos se define como la extracción no ligera de información implícita, previamente desconocida y potencialmente útil, a partir de datos. En la actual sociedad de la información, la minería de datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz para los objetivos de cualquier organización.

Características de la minería de datos:

- Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.
- El entorno de la minería de datos suele tener una arquitectura cliente-servidor.

- Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados.
- El usuario del *Data Mining* es muchas veces un usuario final con poca o ninguna habilidad de programación.
- Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.
- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
- La minería de datos produce cinco tipos de información: asociaciones, secuencias, clasificaciones, agrupamientos y pronósticos.

Algoritmos y técnicas de explotación de datos:

Antes de hablar de los algoritmos y técnicas de exploración hay que explicar las diferentes fases que tiene un Data Mining sin entrar plenamente en dichas fases. A groso modo en una primera fase se obtendría la selección de los datos que podría estar contenida en un *Data Warehouse*. A continuación se procesaría la información, es una etapa de limpieza por así decirlo donde se eliminan los datos no necesarios. Por último aplicaríamos los diferentes modelos o técnicas de explotación para así obtener información útil. Estos son algunos de los modelos y técnicas de explotación:

- Predicción
- Asociación

  Detecta asociaciones. Es muy utilizado en los supermercados y grandes superficies.

  Ejemplo análisis del carro de la compra, ofrecer recomendaciones a los compradores…
- Clustering

  Búsqueda de elementos afines en un conjunto.

  Ejemplo: Segmentación de mercados
- Arboles de decisiones

  Herramienta de clasificación.
- Series Temporales

  Algoritmo específico para predecir los valores de magnitudes en función del tiempo.

  Ejemplo de utilización serian en análisis bursátiles.

- Algoritmo Naive Bayes

  Suele ser el primer algoritmo para explorar datos. Lo que hace es buscar correlaciones entre atributos (Características).

- Redes neuronales

  Resuelve problemas de clasificación y regresión (igual clasificación pero prediciendo una magnitud continua) al igual que los árboles de decisión. Suele utilizarse como alternativa a los árboles de decisiones.

Software de minería de datos

- ***IBM SPSS Statistics*** http://www.01.ibm.com/software/es/analytics/spss/products/statistics/

El software estadístico líder mundial para empresas, gobierno, organizaciones de investigación y académicas. *IBM SPSS Statistics* es un completo conjunto de datos y herramientas de análisis predictivo fácil de utilizar para usuarios empresariales, analistas y programadores estadísticos. Produce: Los modelos de clústeres, Árboles de decisión, Regresión general, Redes Neuronales, *Naive Bayes* (producido únicamente por *SPSS Statistics Server*). También consume conjunto de reglas y modelos de apoyo de máquinas de vectores.

- ***ELVIRA*** http://www.ia.uned.es/~elvira/manual/manual.html

El programa Elvira es fruto de un proyecto de investigación financiado por la CICYT y el Ministerio de Ciencia y Tecnología, en el que participan investigadores de varias universidades españolas y de otros centros. El programa Elvira está destinado a la edición y evaluación de modelos gráficos probabilistas, concretamente redes bayesianas y diagramas de influencia. Elvira cuenta con un formato propio para la codificación de los modelos, un lector-intérprete para los modelos codificados, una interfaz gráfica para la construcción de redes, con opciones específicas para modelos canónicos (puertas OR, AND, MAX, etc.), algoritmos exactos y aproximados (estocásticos) de razonamiento tanto para variables discretas como continuas, métodos de explicación del razonamiento, algoritmos de toma de decisiones, aprendizaje de modelos a partir de bases de datos, fusión de redes, etc. Elvira está escrito y compilado en Java, lo cual permite que funcione en diferentes plataformas y sistemas operativos (MS-DOS/Windows, Linux, Solaris, etc.).

- ***WEKA*** http://www.cs.waikato.ac.nz/ml/weka/

Se trata de un entorno genérico de minería de datos. Se destaca por ser una herramienta multiplataforma y de código abierto, desarrollado por la universidad de Waikato, Nueva Zelanda. Los tipos de modelo en que se apoyan son: Regresión y Regresión general, Redes Neuronales, Artículo modelos establecidos, Árboles de decisión. Implementa algoritmos de aprendizaje para su aplicación. Es de destacar que en todos ellos se puede realizar una combinación con los métodos de selección de variables. Magnífica suite de minería de datos de libre distribución.

- ***SAS Enterprise Miner*****:** http://www.sas.com/technologies/analytics/datamining/miner/

Una de la principales características de SAS Enterprise Miner18 es que está diseñada pensando en su utilización por parte de los responsables de negocio -a través de una interfaz de usuario sumamente intuitiva- a la vez que cumple las expectativas de los responsables de Sistemas de Información y de los analistas: "el trabajo en equipo de estos tres colectivos de profesionales permite a las empresas la reducción de costes en el desarrollo de soluciones *Data Mining*". Para más información acerca de SOFTWARE http://www.dmg.org/products.html

Ventajas e inconvenientes de la minería de Datos

Desde mi punto de vista la Minería de Datos o *Data Mining* es una herramienta muy valiosa para las organizaciones ya que nos ofrece herramientas que ayudan a la toma de decisiones como pueden ser las agrupaciones o predicciones. No obstante a pesar de todas las ventajas que nos proporciona la minería de datos con la información que proporciona a la empresas, existen también una serie de inconvenientes como la falta de privacidad de los datos, errores en los modelos y patrones obtenidos, dificultades de escalabilidad y manejo de software hacen que la minería de datos tenga que mejorar.

Cloud Computing

Es una tecnología joven al igual que *Big Data* que nos brinda la posibilidad de ofrecer servicios a través de internet. Esta nueva tecnología busca tener todos nuestros archivos e información en internet sin preocuparnos de tener la capacidad suficiente para almacenar dicha información.Cloud Computing coge fuerza cuando la provisión de hardware se convierte en un problema, ya que dicho

hardware tiene además de costes monetarios los tiene de espacio, escalabilidad es aquí donde Cloud Computing es una gran alternativa.

Ventajas Cloud Computing

- Reducción de costes.
- Mayor velocidad de trabajo.
- Ahorro en tiempo de instalación.
- Acceso multiplataforma.
- Información en tiempo real.

Desventajas Cloud Computing

- Dependencia de proveedores y de conexión a internet
- Datos sensibles se encuentran fuera de la empresa
- Seguridad
- Escalabilidad a largo plazo

Servicios Cloud Computing

- Amazon Web Services

- Amazon Web Services le ofrece un conjunto completo de servicios de infraestructuras y aplicaciones que le permiten ejecutar prácticamente todo en la nube, desde aplicaciones empresariales y proyectos de grandes datos hasta juegos sociales y aplicaciones móviles.
- Fuente: http://aws.amazon.com/es/

- Rackspace cloud

- Conjunto de productos y servicios Cloud Computing: aplicación web de hospedaje, almacenamiento en la nube, servidores privados virtuales, copias de seguridad, monitoreo.

Business Intelligence

Business Intelligence se puede considerar es una herramienta empresarial con la habilidad de transformar datos en información y después esa información en conocimiento. Desde un punto de vista más teórico se podría definir como el conjunto de metodologías, aplicaciones, y tecnologías que tienen como objetivo obtener, depurar y modificar datos de los sistemas transaccionales para la explotación de dicha información por la empresa convirtiéndola en conocimiento útil.

Dicho de otro modo lo que se logra con *Business Intelligence* es obtener una ventaja competitiva para las organizaciones obteniendo información y ayudando a la toma de decisiones de la alta dirección. Business Intelligence consta de 3 partes fundamentales que se pueden observar en la Fig. 8:

- Minería de datos → Se utilizará para realizar los análisis.
- *Data Warehouse* → Se integraran las diferentes bases de datos que tenga las empresas, históricas, clientes, contabilidad, etc.
- Data Mart → Es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica.

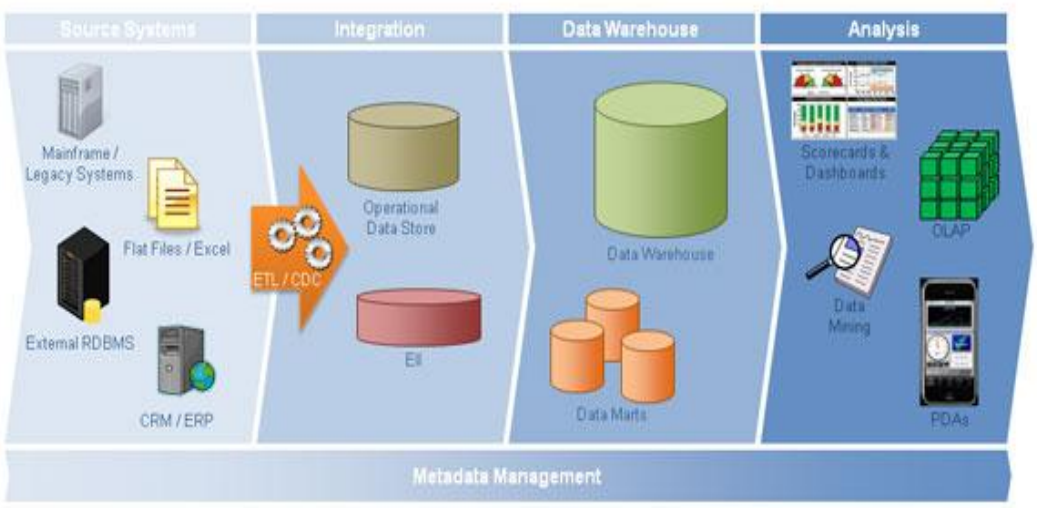

Figura 8. Fuente: http://www.lingaro.com/lingaro/competencies/business-intelligence

Se puede observar el funcionamiento de manera clara, como de diferentes bases de datos estructuradas se almacenan en un *Data Warehouse* y finalmente se analizan con herramientas como *Data Mining*. Esto Constituirá un *Business Intelligence.*

Big Data Analytics

Es una nueva herramienta empresarial la cual nos permitirá examinar grandes repositorios de datos de Big Data, con el objetivo de ayudar en la toma de decisiones descubriendo patrones ocultos, correlaciones desconocidas, predicciones y otra información útil y así permitir lograr ventajas competitivas para las empresas u organizaciones que lo posean. Según el artículo de *Gustavo Tamaki* (2012 "La hora del *Big Data*"): "Algunos analistas confirman que las empresas que adoptan Big Data Analytics tendrán una ventaja competitiva de 20% en todas las métricas financieras sobre sus competidores."

El objetivo principal de Big Data Analytics es ayudar en la toma de decisiones de negocio al permitir analizar grandes volúmenes de datos de bases de datos transaccionales, así como otras fuentes de datos que pueden quedar sin explotar por la inteligencia de negocio (BI). Estas otras fuentes de datos pueden incluir registros de servidor Web y los datos de seguimiento de clics en Internet, informes de actividades de medios sociales, los registros detallados de llamadas de teléfonos móviles, la información captada por los sensores, correo electrónico, *tweets*…

No se debe asociar exclusivamente Big Data Analytics a grandes volúmenes de datos y análisis de datos grandes con datos no estructurados ya que también tienen en cuenta Bases de datos con datos estructurados, es decir, bases de datos relacionales.

Big Data Analytics si solo analizaría bases de datos estructuradas se podría realizar herramientas más conocidas de análisis predictivo y minería de datos. No obstante como también recoge datos de fuentes no estructuradas o semi-estructuradas. Como resultado, una nueva clase de tecnología. Las tecnologías relacionadas con *Big Data Analytics* incluyen *NoSQL* bases de datos, *Hadoo*p y *MapReduce* . Estas tecnologías forman el núcleo de la plataforma de software de código abierto que soporta el procesamiento de grandes volúmenes de datos a través de sistemas en clúster.

Diferencias Business Intelligence y Big Data Analytics

Relacionado a lo visto en los puntos anteriores, se procederá a observar las principales diferencias entre *Business Intelligence* y *Big Data Analytics* para que de este modo se comprendan mejor ambos términos.

	Business Intelligence	***Big Data Analytics***
Velocidad	Menor velocidad de análisis	Mayor velocidad de Análisis gracias a nuevas tecnologías.
Capacidad de análisis	Menor capacidad. Almacén de datos de menor capacidad.	Big Data, grandes repositorios de datos. Zeta bytes, Peta bytes…
Tipos de datos	Estructurados, semi-estructurados, no estructurados.	Solo estructurados.
Herramientas	*Data Warehouse, data Mining…*	Herramientas de BI más herramientas de análisis de datos semi o sin estructurar: *Hadoop*, *MapReduce*, Pentaho
Escalabilidad	Menor.	Mayor

Fuente: Elaboración propia

Existe la creencia de que el termino *Business Intelligence* y *Big Data Analytics* no están completamente definidos y por ello hay discusiones sobre su significado o en qué consisten.

Desde mi punto de vista son prácticamente es lo mismo o una evolución, una herramienta empresaria de apoyo a la toma de decisiones empresariales, pero que gracias al marketing se intenta vender como un producto totalmente nuevo.

Me refiero a evolución porque ambos tienen el mismo objetivo ayudar a la toma de decisiones permitiendo así conseguir ventajas competitivas, no obstante existen ciertas mejoras *en Big Data*

Analytics gracias a nuevas tecnologías que lo hacen más diverso y potente. La principal diferencia es que *Big Data Analytics* es capaz de procesar un mayor número de datos de diferentes fuentes (estructuradas, semi-estructuras y no estructuras) y con mayor velocidad gracias a la utilización de *Big Data*, mientras que *BI* solo lo podría realizar con datos de fuentes estructuradas y con menor capacidad de análisis. Por esto *Big Data Analytics* es más potente y sustituirá a *Business Intelligence*.

Tipos de datos Big Data

Se puede decir que básicamente hay tres tipos de datos en *Big Data*:

- Datos estructurados
- Datos no estructurados
- Semi-Estructurados

Datos estructurados

Los datos estructurados son aquellos datos que tienen bien definido su longitud y su formato. Suelen ser fechas números, cadenas de caracteres y están almacenados en tablas. En las empresas estos datos los encontramos en información obtenida a partir de CRM, ERP etcétera. Estos datos suelen estar guardada en un Data Warehouse si contienen mucha información y si el negocio o la empresa no generan tal cantidad de datos tendrán una base de datos relacional. Para consultar estos datos se realizan mediante consultas SQL. La mayoría de los casos de uso *Business Intelligence* y *Business Analytics* trabajan con este tipo de datos.

De donde obtenemos los datos estructurados:

A. Datos generados por maquinas.
   - Datos procedentes de sensores: existen múltiples ejemplos como los procedentes de un GPS, contadores eléctricos, tacómetros, equipos médicos, etc....
   - Web Log Data: servidores, redes, aplicaciones, etc.. generan grandes cantidades de datos estructurados.
   - Datos procedentes de puntos de venta: basta con pensar en un hipermercado con una cajera pasando códigos de barras por un lector.

- Datos financieros: muchas operaciones bancarias y bursátiles son de datos estructurados generados automáticamente.

B. Datos generados por personas.

Los datos estructurados generados por personas también son variados y pasan desde los registros de una contabilidad en un ERP pasando por el hecho de cumplimentar un formulario en una web o incluso nuestros movimientos en uno de esos juegos on-line que ahora nos encontramos en Facebook.

Los datos estructurados son el pilar de las bases de datos relacionales. En los modelos relacionales, toda la información esta guardada en un esquema de tablas y dichas tablas tendrá definidas unos campos y relaciones entre ellas.

Datos no estructurados

Son lo opuesto a los datos estructurados, es decir, carecen de un formato específico. Al igual que los datos estructurados son generados:

A. Datos generados por máquinas y computadoras.

- Imágenes de satélites.
- Datos científicos: gráficos sísmicos, atmosféricos, etc...
- Fotografía y vídeo: por ejemplo cámaras de vigilancia.
- **Datos recopilados de sónar y radar'(posicionamiento smarthphone…)**

B. Datos generados por personas, o sea, datos picados por personas en un ordenador.

- Textos incluidos dentro de los sistemas de información internos de las organizaciones: basta con pensar en documentos, presentaciones, correos electrónicos, etc...
- Datos provenientes de redes sociales: *Twitter*, *Facebook*, *LinkedIn*, *Flickr*, *Instragram*, Tuenti. El número de redes sociales crece cada día, cada vez es más común ver diferentes redes sociales que hacen referencia a diferentes grupos.
- Datos provenientes de nuestros dispositivos móviles: pensemos en los mensajes que enviamos con nuestros teléfonos móviles.
- Contenido de sitios web: podemos ir desde vídeos de YouTube contenidos de páginas web o incluso blogs.

Se puede decir que el 80% de los datos de una empresa son no estructurados, y que gracias al Big Data ahora se pueden analizar y obtener información útil para las organizaciones. No obstante cabe recalcar que el *Big Data* no solo se centra en los datos no estructurados, sino que los hace en todos, tanto estructurados como no estructurados y sin olvidarnos de los semi-estrucuturados.

Datos semi-estructurados

Los datos semi-estructurados son una mezcla de los estructurados y no estructurados, es decir, estos datos siguen una especie de estructura implícita, pero no tan regular como para poder ser gestionada y automatizada como la información estructurada. Un ejemplo de esto son las "páginas webs". Estos datos tienen la peculiaridad de que manteniendo esa pequeña estructura se puede sacar información útil. El formato va evolucionando hasta convertirse en un protocolo o fórmula generalmente aceptada, con una serie de características definitorias. Se podría decir que estos datos semi-estructurados poseen sus propios "metadatos semi-estructurados", que describen los objetos que trata el texto y las relaciones que se pueden inferir. Ejemplos de estos datos son las notas de defunción, las solicitudes de empleo, los listados de propiedades inmobiliarias, avisos legales o los nombres de cuentas bancarias.

Utilización del *Big Data*

Bastante difícil encontrar información en España sobre como las empresas utilizan *Big Data*, mientras que los datos de EEUU son más fáciles de obtener sobre la creación de empleo a consecuencia de la implantación de *Big Data* en las empresas en EEUU. Estos datos son ofrecidos por la empresa norteamericana *Icrunchdata*[7]. Esta empresa lo que ha hecho concretamente es desarrollar un índice sobre la demanda de puestos de trabajo asociados a *Big Data* y el BI/BA (Big Data Job Índex), este índice ya presenta un buen uso del *Big Data* ya que se ha generado gracias a él.

Los datos obtenidos a Agosto de 2013 observados en la Figura 9 son una demanda de aproximadamente 575.506 puestos de trabajo repartidos en:

- Analistas: 36,4%
- *Big Data*: 22,3%
- Científicos de datos: 13,5%

[7] Incrunchdata: es un interesante portal de empleo para todos aquellos que quieran trabajar en Big Data o en BI en los EEUU. http://www.icrunchdata.com/

- Desarrollo de *software* (para *BI* y *Big Data*): 13%
- Estadísticos: 10,1%
- *Business Intelligence*: 4,6%

Se estima que para el año que viene solo en EEUU estaríamos hablando de más de un millón de puestos y para 2015 será de 1,9 millones.

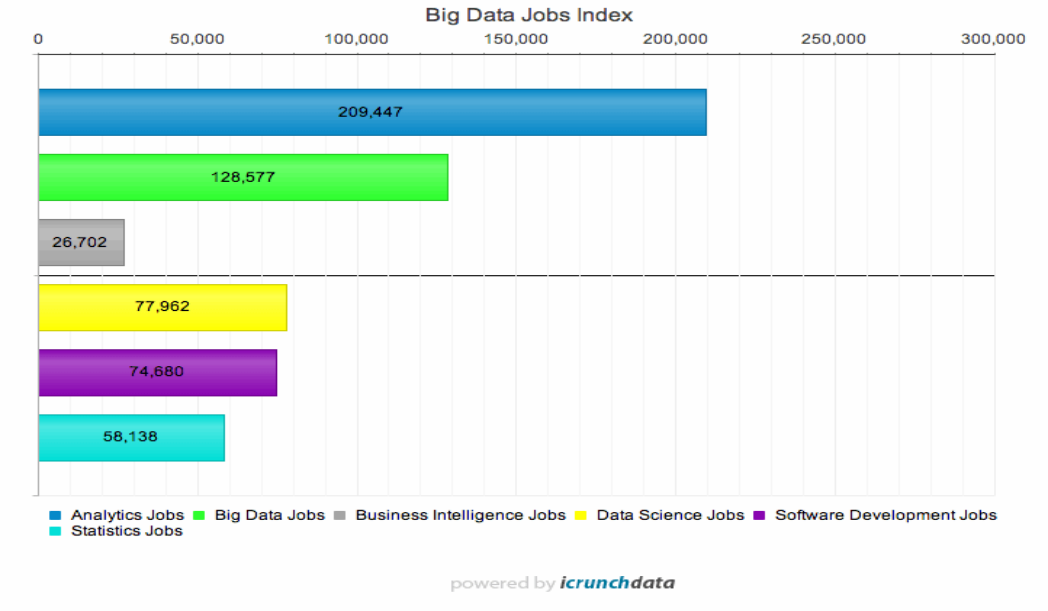

Figura 9. Big Data Job Index. Fuente. Icruchdata año 2013

Utilización del Big Data en España

Es de sobra conocido que los avances tecnológicos tienden a implantarse de forma considerablemente tardía en nuestro país en comparación con otros países y a todo esto le podemos añadir la actual situación de crisis económica que sufrimos y que parece que no mejora.

A pesar de todo esto, lo cierto es que en España existen en la actualidad iniciativas de éxito sobre Big Data. Según un estudio de IDC España, patrocinado por EMC, *JasperSoft*, *Microsoft* y *Sybase*, el mercado de Big Data está en auge en nuestro país. Los datos, recabados a partir de 502 entrevistas a expertos españoles, lo confirman: un 4,8% de las empresas ya han incorporado estos procesos a su negocio y las previsiones indican que en 2014 la adopción será del 19,4%, lo que supone un incremento del 304% con respecto a 2012. Como vemos, España está todavía en la fase inicial, aunque con estas cifras el Big Data se empieza a mostrar como un factor imprescindible en las empresas españolas.

IDC (principal proveedor global de inteligencia de mercado, servicios de asesoría y organización de eventos para los mercados de tecnologías de la información y las comunicaciones). IDC ayuda a los profesionales de TI, ejecutivos de negocio y a la comunidad de inversores a tomar decisiones basadas en hechos sobre adquisiciones tecnológicas y estrategias de negocio. Más de 1.000 analistas de IDC en 110 países proporcionan su experiencia global, regional y local en tecnologías y oportunidades y tendencias sectoriales. Desde hace más de 49 años, IDC proporciona información estratégica para ayudar a sus clientes a conseguir sus objetivos de negocio. IDC es una subsidiaria de IDG, la empresa líder mundial en medios tecnológicos, investigación y eventos. Los beneficios tampoco son desdeñables. Ya en el 2010 esta tecnología generaba entorno a los 3.200 millones de dólares en todo el mundo. Según las estimaciones de IDC, esta cifra podría llegar a alcanzar los 16.900 millones de euros en 2015.

Las cifras demuestran que, a pesar de la crisis, las empresas están interesadas por tecnologías que generan una mayor eficiencia organizacional y que proporcionan nuevas oportunidades de negocio.

Por otro lado el *Big Data* cobra sentido cuando hablamos de empresas con un alto volumen de información, generada muy rápidamente, procedente de diversas fuentes, con distintos formatos y con datos de calidad sin estas características no tendría sentido, se podrían utilizar otras tecnologías mucho más económicas y menos complejas, con esto queremos recalcar que con el gran número de Pymes que existe en nuestro país dificultara la incorporación de *Big Data*.

Dificultades para implantar *Big Data*

Los responsables de TI consideran que hay diversos obstáculos para adoptar soluciones de *Big Data* se pueden observar en la Figura 10, siendo la seguridad la principal preocupación seguida de las carencias de presupuesto y personal.

A escala global, más de uno de cada cuatro consultados (27 %) consideran la seguridad de los datos y la gestión del riesgo sus principales retos en proyectos de *Big Data*, debido a múltiples factores:

- Enorme volumen de datos.
- Las distintas formas de acceso a dichos datos.
- La falta de presupuesto para seguridad.

Las preocupaciones de seguridad están más patentes en China (45 %), India (41 %), Estados Unidos (36 %) y Brasil (33 %).

La falta de presupuesto (16 %) y la falta de tiempo para estudiar el fenómeno Big Data (14 %) constituyen los principales obstáculos para las dos terceras partes de los encuestados.

Casi uno de cada cuatro consultados (el 23 %) citan la carencia de suficientes profesionales de TI (13 %) o de expertos en Big Data (10 %) como la mayor barrera, especialmente en Japón (31 %) y Brasil (30 %).

Pero no son solo estas las princípiales dificultades para implantar Big Data, existen otras derivadas de las herramientas e infraestructuras necesarias y otras derivadas de la inversión para desarrollar y mantener un proyecto de *Big Data*.

Además existen otras preocupaciones que pueden llegar a dificultar la impanación de nuestro proyecto si nos e tienen en cuenta:

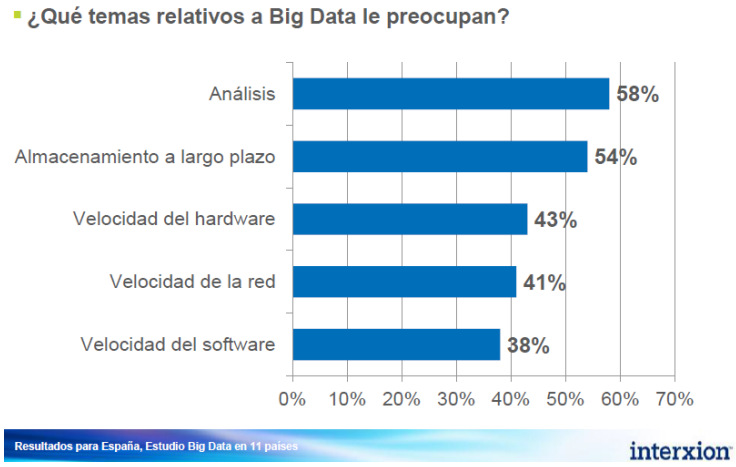

Figura 10. Temas relativos que preocupan de Big Data. Fuente: Interxion, Año 2013

Desde mi punto de vista la mayor dificultad para implantar *Big Data*, es el presupuesto debido a la crisis actual, una vez has superado este punto, tendríamos que centrarnos en la seguridad, ya que con ese volumen de datos que manejamos tenemos que tener especial cuidado, si nos roban los datos que tanto hemos intentado conseguir no nos servirán de nada, perderemos esa ventaja que hemos adquirido y si se tratan de datos confidenciales estaremos comprometiendo información privilegiada con la correspondiente sanción/consecuencia. Cuando superemos estos dos puntos, lo tercero pero no menos importante sería obtener una velocidad de análisis correcta para que esos datos que tengamos la convirtamos en información útil.

Plataformas y software para tratamiento de Big Data

En la actualidad existen diferentes herramientas, software para el tratamiento de la tecnología *Big Data.* A la hora de hablar del software de tratamiento de grandes almacenes de datos lo primero que se debe hacer es hablar de *MapReduce* que es la base de la programación de los diferentes herramientas y software. Se continuara por Hadoop que es el software más utilizado, seguidamente se hablara de los *Appliances* y por último de *Pentaho* una plataforma Open Source que está ganando multitud de seguidores.

MAPREDUCE

(Cristina Requena 2010) define MapReduce como un *framework*, es decir, representa una arquitectura de Software (lenguaje de programación), muy usado en la programación de funciones de alto nivel:

- map(list[], oper) aplica la operacionoper a la lista list, retornando una nueva lista cuyos elementos han sido operados, individualmente, por oper. Por ejemplo, si disponemos de la lista list[1,2,3] y de la operación suma x = x+1 y realizamos la operación map(list, suma) el resultado de la misma será [2,3,4].
- fold (list[], oper) aplica la operación oper a la lista list, retornando un elemento producto de la operación de los elementos de list entre sí. Por ejemplo, si disponemos de la lista list[1,2,3] y de la operación suma (x:xs) = x+suma(xs) y realizamos la operación fold(lista, suma), el resultado de la misma será 6.

Debido a la posibilidad de que la operación no sea asociativa, los lenguajes de programación ofrecen, normalmente, dos operaciones: una “hacia la derecha”, o foldr y otra “hacia la izquierda”, o foldl. En general, en el área de la computación distribuida *Map* se utiliza para fraccionar una operación compleja entre varios nodos y *Fold/Reduce* para recoger los resultados y unificarlos. Por su parte, los *frameworks MapReduce* toman la base de las operaciones mencionadas anteriormente para crear una operación genérica y más compleja, cuyo funcionamiento es realmente útil para las bases de datos *NoSQL*: en vez de usarse sobre listas de valores unidimensionales, ésta toma como parámetros entrantes una lista de tuplas de tipo (clave, valor) y devuelve una lista de valores. Entre las operaciones

map (distribuida) y *reduce* (normalmente localizada) se genera una lista de *tuplas* (clave, valor) con valores temporales, de las que reduce filtra solamente las que tengan una determinada clave.

En definitiva, *MapReduce* es fundamental en las bases de datos *NoSQL* para permitir la utilización de funciones de agregación de datos, ya que al carecer de esquema son mucho más complicadas que en las bases de datos relacionales clásicas.

Ejemplo de uso de *MapReduce*: MongoDB→es un sistema de base de datos *NoSQ*L orientado a documentos, desarrollado bajo el concepto de código abierto y nacida en 2007. En vez de guardar los datos en tablas como se hace en las base de datos relacionales, MongoDB guarda estructuras de datos en documentos tipo JSON con un esquema dinámico (MongoDB llama ese formato BSON), haciendo que la integración de los datos en ciertas aplicaciones sea más fácil y rápida. Esta base de datos es altamente utilizada en las industrias, por ejemplo son muy utilizadas por MTV *Network,Foursquare*.

Nota: JSON, acrónimo de JavaScript *ObjectNotation*, es un formato ligero para el intercambio de datos. JSON es un subconjunto de la notación literal de objetos de JavaScript que no requiere el uso de XML.

HADOOP

Apache Hadoop (http://hadoop.apache.org/) es una solución de software libre diseñada para el tratamiento de hasta exabytes de datos distribuidos en múltiples nodos. Hadoop se ha convertido en un estándar sobre el que se desarrollan herramientas comerciales por compañías tradicionales. Se puede decir que es la solución tecnología sobre el procesamiento de Big Data que más destaca.

La solución Hadoop se basa en un desarrollo de Google del año 2009 denominado MapReduce, y que actúa en dos fases. La primera fase, Map, y la segunda Reduce.

Cabe especificar que Hadoop no es un programa en sí, es decir, no podemos descargar un programa denominado Hadoop directamente, ya que Hadoop es un ecosistema de productos bajo el paraguas de la Apache Software Foundation[8]. De esta forma hay dos productos principales que conforman el núcleo de cualquier aplicación Hadoop. Estos son el HDFS y MapReduce. Pero además de estos productos

[8] Apache Apache Software Foundation (ASF) es una organización no lucrativa (en concreto, una fundación) creada para dar soporte a los proyectos de software bajo la denominación Apache, incluyendo el popular servidor HTTP Apache.

básicos, existen multitud de productos o iniciativas *opensource* que modifican o complementan el núcleo de Hadoop. Los más utilizados en los proyectos de *BI* y *Big Data* posiblemente serán:

- *PIG*→*Apache Pig* es una plataforma para el análisis de grandes conjuntos de datos que consta de un lenguaje de alto nivel para la expresión de programas de análisis de datos, junto con la infraestructura para la evaluación de estos programas. La mejor característica de los programas de *Pig* es que su estructura es susceptible de paralización sustancial, lo que permite el manejo de grande cantidades de conjuntos de datos.
- *HIVE*→Gestiona los datos almacenados En hdfs y proporciona un lenguaje de consulta basada en SQL para generar datos.
- *HBASE*→Base de datos distribuida no relacional
- *ZOOKEEPER*→Servicio centralizado para mantener la información de configuración, denominación, proporcionando sincronización distribuida y la prestación de servicios de grupo.
- *SQOOP*→ Una herramienta eficiente para la transferencia de datos de una BD relacional al *HDFS*
- *MAHOUT*→*Apache Mahout* tiene implementaciones de una amplia gama de algoritmos de aprendizaje automática y minería de datos: agrupaciones, clasificación, filtrado colaborativo y patrón de la minería frecuente.

Un punto que se tiene que tener claro, es que Hadoop, es un subconjunto de programas o plataformas, que son capaces de colaborar entre sí y crear sinergias, debido a sus diferentes funcionalidades.

Características de Hadoop:

- Económico.
- Escalable (Adaptable).
- Eficiente (muy veloz, dado que realiza su trabajo en forma de paralelo).
- Confiable (mantiene automáticamente copias los datos en nodos para la prevención de fallos).

Breve historia de Hadoop

- Empieza en 2002 *con Doug Cutting* y *Mike Cafarella.*
- Inspirado por los papers de Google en MapReduce y Google File System.
- Proyecto nombrado a partir del elefante de peluche amarillo del hijo de Doug (de ahí el logo).
- Empieza como parte de la manera de manejar los datos de un motor de búsqueda web (Notch).

- Proyecto Apache Hadoop inicia – 2006.
- Desarrollado y bastante usado en Yahoo!.
- Usado también en *LastFM*, *Facebook* y *The New York Times*.
- 1 TB sort benchmark - 209 seg. – 2008.
- Minute sort - 500 GB en 59 seg. (1400 nodos).
- 100 TB sort benchmark - 173 min. (3400 nodos) – 2009.

Arquitectura Hadoop

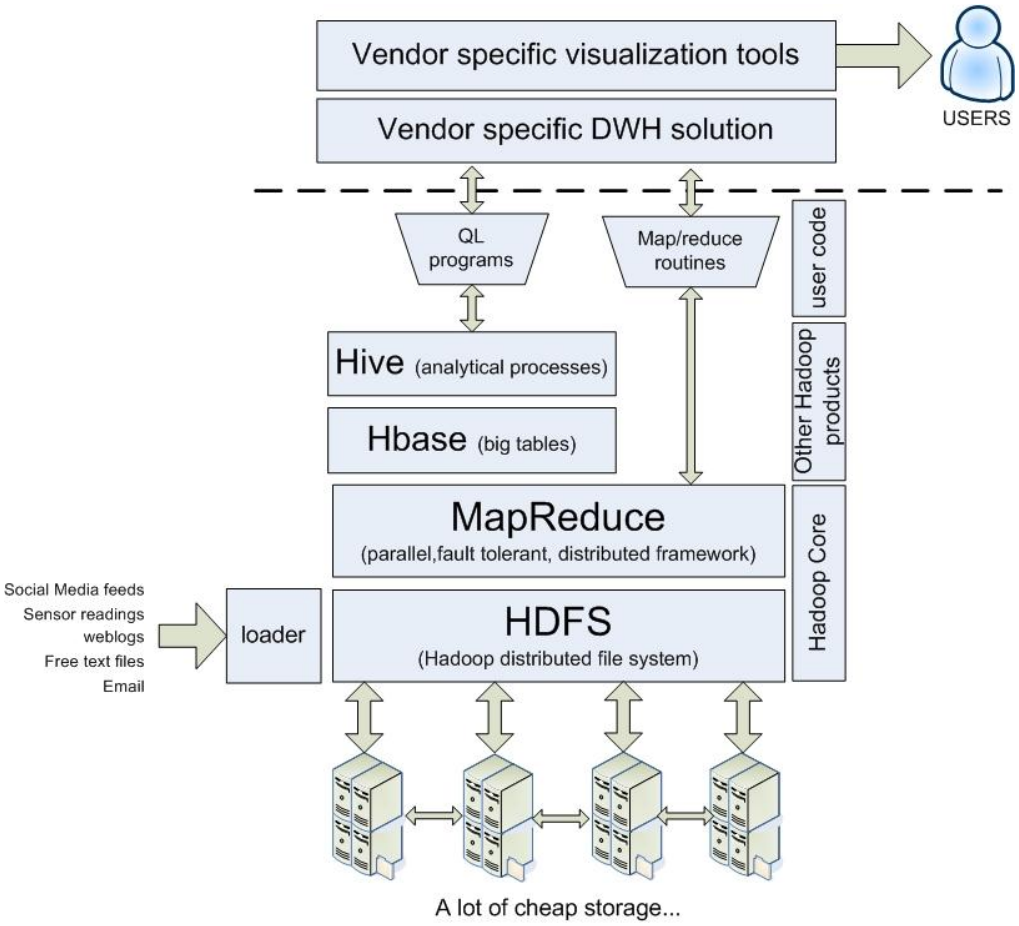

Figura 11 Arquitectura *Hadoop.*

La arquitectura fundamental de Hadoop, Figura 11, se basa en tres componentes fundamentales:

A. HDFS→Un *filesystem* distribuido que proporciona un alto rendimiento de acceso a datos de la aplicación.
B. *Hadoop MapReduce*→ La plataforma por excelencia para el procesamiento distribuido de grandes conjuntos de datos.

C. *Hadoop Commons*→ Utilidades comunes sobre las cuales se apoyan sub-proyectos Hadoop constituyendo de este modo sinergias. Por ejemplo: Uno de los productos de *Hadoop*, *HDFS*, es un sistema de archivos. Y generalmente es la primera piedra de un proyecto *Hadoop*. *HDFS* es altamente distribuido y tolerante a fallos, y está especialmente pensado para correr clústeres de pc's de escritorio, ya que es súper escalable. Pero claro, un sistema de archivos no es eficiente a la hora de recuperar información (no al menos al nivel de un RDBMS), es lento y proporciona pocas herramientas de búsqueda. Aquí es donde entran en juego otros de los productos Hadoop, como *HBase* (creado a partir de *google'sBigTable*), que nos ofrece una capa de acceso a la información en el *HDFS* mucho más eficiente. Aquí es donde se producen las sinergias de *Hadoop* mezclando todas sus utilidades o sub-productos.

D. Otras Partes:
   - *DAtaNodes.*
   - *SecondaryNamenode.*
   - Balanceador.
   - *JobTracker y el Tasktraker.*

Funcionamiento Hadoop

HDFS (The Hadoop Distributed File System) es un sistema de archivos que trata de recopilar toda la información posible. Se puede definir como un sistema de archivos distribuido, escalable y portátil escrito en Java para *el framework Hadoop*.

El funcionamiento consiste en que cada nodo en una instancia Hadoop típicamente tiene un único nodo de datos; un clúster de datos forma el clúster HDFS. La situación es típica porque cada nodo no requiere un nodo de datos para estar presente. Cada nodo sirve bloques de datos sobre la red usando un protocolo de bloqueo específico para HDFS. El sistema de archivos usa la capa TCP/IP para la comunicación; los clientes usan RPC para comunicarse entre ellos. El *HDFS* almacena archivos grandes (el tamaño ideal de archivo es de 64 MB), a través de múltiples máquinas. Consigue fiabilidad mediante replicado de datos a través de múltiples hosts, y no requiere almacenamiento *RAID* en ellos. Con el valor de replicación por defecto, 3, los datos se almacenan en 3 nodos: dos en el mismo rack, y otro en un rack distinto. Los nodos de datos pueden hablar entre ellos para reequilibrar datos, mover copias, y conservar alta la replicación de datos. *HDFS* no cumple totalmente con *POSIX* porque los

requerimientos de un sistema de archivos *POSIX* difieren de los objetivos de una aplicación Hadoop, porque el objetivo no es tanto cumplir los estándares *POSIX* sino la máxima eficacia y rendimiento de datos. *HDFS* fue diseñado para gestionar archivos muy grandes. *HDFS* no proporciona alta disponibilidad.

Cuándo usar HDFS?

- Archivos muy, muy grandes (GB o más).
- Necesidad de particionar archivos.
- Fallo de nodos sin perder información.
- Una escritura, muchas lecturas.

¿Cuándo no usar HDFS?

- Baja latencia.
- Muchos archivos pequeños.
- Múltiples "escritores".
- Modificaciones arbitrarias a los archivos.

¿Por qué Hadoop?

- Más rápido que un RDBMS para grandes volúmenes de datos (especialmente datos no organizados).
- Más rápido que un HPC tradicional, ya que implementa optimizaciones teniendo en cuenta la topología de la red (optimiza el uso de la red).
- Evita la pérdida de información a través de replicación.
- API fácil de aprender.
- Posibilidad de trabajar con lenguajes diferentes a Java.

Arquitectura *HDFS:*

Se muestran en la Figura 12 y 13 a continuación*:*

- *Namenode*→Parte principal del sistema de archivos HDFS, alojado en el master se encarga de gestionar los metadatos del sistema de archivos, *namespace*.
- *DataNodes*→Proporcionan Servicios de Almacenamiento de bloque de datos para el sistema de archivos compartido y servicios de recuperación.
- *SecondaryNamenode*→ Se encarga de la copia de seguridad del *NameNode* en tiempo Real, recordemos que una de las características era la fiabilidad.

- Balanceador→Equilibra el uso de espacio en disco en un clúster *HDFS* cuando algunos *datanodes* se llenan o cuando nuevos nodos vacíos se unen al clúster.
- *JobTracker* y el *TaskTraker*: procesos que se encargan de la gestión de los *JobsMapReduce.*

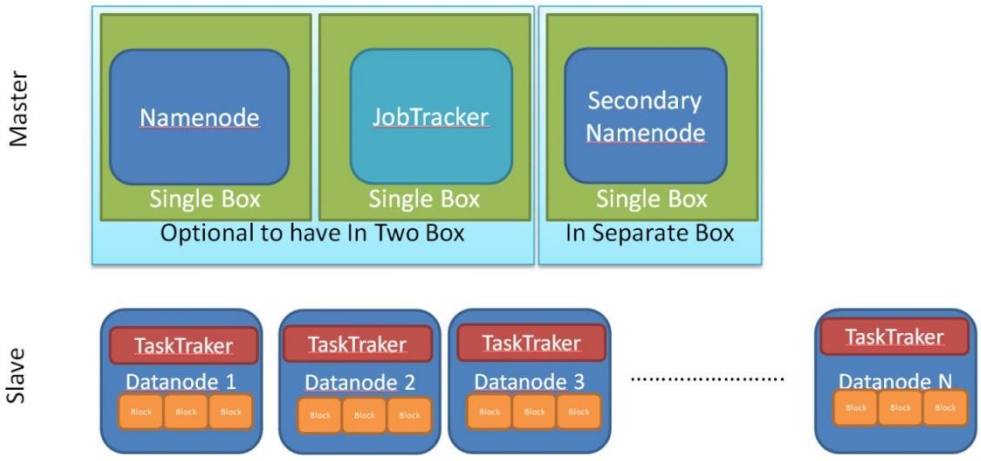

Figura 12 Funcionamiento *HDFS.*

Fuente http://eventos.citius.usc.es/bigdata/workshops/hadoop-taller.pdf

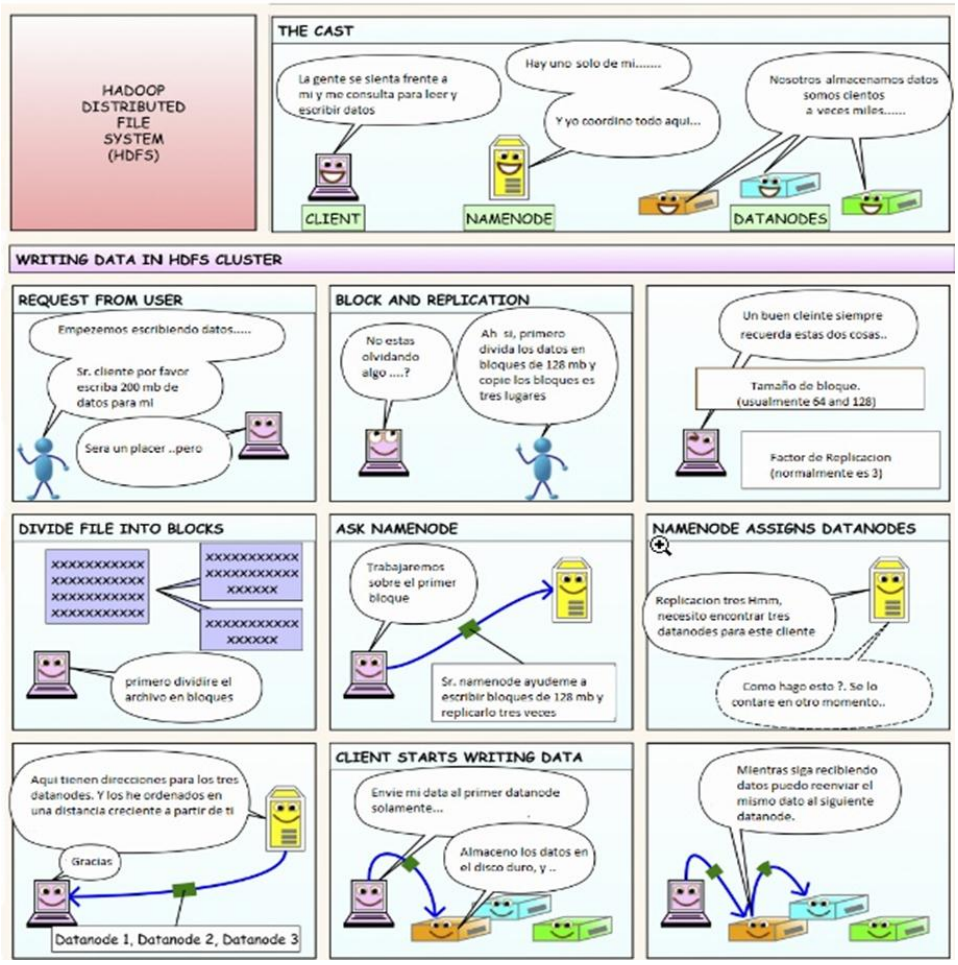

Figura 13. Proceso de escritura.

Fuente: http://nosql.mypopescu.com/post/15561851616/hadoop-distributed-file-system-hdfs-a-cartoon-is-worth

Ejemplos de empresas que utilizan Hadoop[9]

- ***Facebook***:

  Utiliza Hadoop para almacenar copias de los datos internos y fuentes de datos de grandes dimensiones, así como de fuente de informes/análisis y aprendizaje…

- ***Google***:

  Uno de los pilares clave para el desarrollo de *Hadoop*. Junto con *IBM*, lleva a cabo la iniciativa "*UniversityInitiativetoAddress Internet-Scale Computing Challenges*" que intenta mejorar el

[9] http://wiki.apache.org/hadoop/PoweredBy
http://prezi.com/0zlpwnwpvsb0/presentacion-hadoop/
http://www.bi.dev42.es/2011/12/17/hadoop-en-proyectos-de-business-intelligence/

conocimiento de los estudiantes sobre computación paralela para adaptarse a modelos de programación y paradigmas como MapReduce.

- ***ClouderaInc***:

  Es una de las principales compañías que dan soporte y formación en Hadoop. Tiene su propia distribución de Hadoop y uno de sus trabajadores "*Tom White*", ha escrito un libro referencia de Hadoop.

- ***Twitter****:*

  Utiliza *Hadoop* para almacenar y procesar *tweets* y ficheros de log. Además utiliza *Pig* para trabajos programados y ad-hoc.

- ***Yahoo****:*

Otra de las empresas promotoras de Hadoop, lo utilizan en más de 25000 clústeres para temas de búsquedas en web y sistemas auxiliares.

LOS APPLIANCES

Además de *Hadoop* existen otras alternativas para el manejo de grandes repositorios de datos, es el caso de los Appliances.

Un *Appliance*, es un término común de la lengua anglosajona, en lenguaje cotidiano se suele traducir como "aplicación" o incluso como "dispositivo", no obstante en el mundo de las Tics, este término toma un carácter más técnico. Un *Appliance* se caracteriza porque tiene una interacción óptima entre el hardware y el software, es decir, están adaptados perfectamente o lo que es lo mismo fabricados el uno para el otro. Los *Appliances* están destinados a un solo campo de actividad, el cual, le dominan perfectamente. Un ejemplo claro para entender esto, es imaginarnos una lavadora, con una lavadora no podemos ni telefonear, ni calentar comida, ni ver la tv, solo sirve para lavar y en ese aspecto es lo mejor.

Por lo tanto podemos definir un *Appliance* en el mundo del *Big Data*, como la aplicación (*hardware* y *software*) que tiene el objetivo único y exclusivo de manejar, recopilar y analizar grandes repositorios de datos.

Un ejemplo de esto lo tiene Oracle: Oracle Big Data Appliance:

- Oracle Big Data Appliance es un sistema optimizado para adquirir, organizar y cargar datos no estructurados en Oracle Database 11g. Combina componentes de hardware optimizados con nuevas soluciones de software para ofrecer la solución de grandes datos más completa.

- *Oracle Big Data Appliance* es un sistema de ingeniería optimizada para adquirir, organizar y cargar los datos no estructurados en Oracle Database 11g. Se combina componentes de hardware optimizado con nuevas soluciones de software para datos grandes para ofrecer la solución más completa de datos grande. Oracle Gran Data *Appliance* incluye una distribución de código abierto de Apache Hadoop ™ ™, Oracle *NoSQLDatabase*, adaptador de *Aplicación Oracle Data Integrator* para *Hadoop*, Oracle cargador para Hadoop, y una distribución de código abierto de R.

Existen otras empresas que ofrecen soporte Big Data por ejemplo Amazon[10] con un gran catálogo de servicios Big Data o relacionados con él.

- Amazon Elastic Compute cloud (EC2): Capacidad informática en la nube.
- Amazon Elastic MapReduce : Procesar grandes cantidades de datos.
- Amazon DynamoDB: Gestión bd NoSql.
- Amazon Simple Storage Service (S3): Almacenamiento masivo.

Otra empresa conocida que ofrece servicios es Telefónica, con su servicio Instant Server similar al EC2.

Pentaho

Pentaho es una alternativa Open Source para Business Intelligence. El modelo de negocio que utiliza Pentaho es de código libre y comercial, por lo que elimina las licencias de software y `proporciona soporte mediante suscripciones anuales.

Pentaho está orientado al *Business Intelligence* o *Big Data Analytics* por lo que su objetivo principal es ayudar en la toma de decisiones cuando se tienes grandes repositorios de datos. Proporciona una interfaz interactiva fácil y multiplataforma para permitirle acceso a grandes repositorios de datos, crear e interactuar con informes, análisis de datos.

Al igual que pasaba con Hadoop, Pentaho es un subconjunto de programas o plataformas, que son capaces de colaborar entre sí y crear sinergias, debido a sus diferentes funcionalidades.

[10] Fuente: http://aws.amazon.com/es/products-solutions/

Compañías como Telefónica lo utilizan y actualmente el Banco Santander concretamente en su CPD de Solares pretende instalar esta alternativa.

Business Case del *Big Data*

Business Case es un instrumento estratégico que existe para valorar y tomar la mejor decisión respecto a un proyecto de inversión. Para ser más precisos el Business Case es un conjunto de métodos que nos van guiando para medir y evaluar, de forma eficiente y concreta, cuál es el impacto financiero y/o económico de tomar una u otra decisión, así como para documentar y presentar estructuradamente dicho análisis, de tal forma que la persona que lo analice cuente con todos los elementos(tanto financieros como no financieros) para tomar una decisión sin depender de la persona que realiza el análisis de la propuesta, no se trata exclusivamente de hacer un análisis financiero, sino de llegar más allá. Se podría decir que es un plan de negocios o análisis de viabilidad.

El Business Case es muy utilizado en proyectos de IT, dada la complejidad de estos y el rápido ciclo de vida que tienen, por ello la medula del Business Case es la línea temporal de tiempo, la cual podrán ser días, semanas, años y así mostrarnos el escenario de trabajo para la implementación de las estrategias financieras, las cuales ayudaran a tomar decisiones referentes a reducir costos, incrementar y/o acelerar utilidades. Para realizar es necesario calcular los flujos de efectivo mediante:

- Van: Es aquel que permite determinar la valoración de una inversión en función de la diferencia entre el valor actualizado de todos los cobros derivados de la inversión y todos los pagos actualizados originados por la misma a lo largo del plazo de la inversión realizada. Se puede observar la fórmula matemática en la Figura Nº14 a continuación:

$$\text{VAN} = -I + \frac{R}{(i-g)}$$

Figura 14. Formulación Van.

- Tir: La tasa interna de retorno de una inversión o proyecto es la tasa efectiva anual compuesto de retorno o tasa de descuento que hace que el valor actual neto de todos los flujos de efectivo (tanto positivos como negativos) de una determinada inversión igual a cero. Se puede observar la formula en la Figura Nº15.

$$VAN = \sum_{t=1}^{n} \frac{F_t}{(1+TIR)^t} - I = 0$$

Figura 15. Formulación Tir.

- Payback o periodo de recuperación: método por el cual una empresa al realizar una inversión o un proyecto dicta o dice cuánto tarda en recuperar dicha inversión sin tener en cuenta los flujos de caja el Payback lo único que tiene en cuenta es el tiempo por lo tanto es el tiempo que tarda antes en recuperarse dicha inversión.
- Aparte de estos términos hay que tener encuentra muchos otros como el ROI, el riesgo, análisis de sensibilidades…

El esquema general de un Business Case, sin olvidarnos que puede tener modificaciones dependiendo del proyecto sería:

- Esquema General
- Sumario ejecutivo
- Introducción
- Métodos y Análisis de Datos
- Alcances y Límites
- Supuestos
- Modelo Costo /Beneficio
- Fuente de Datos y Métodos Empleados
- Modelo Financiero
- Riesgos, sensibilidad y contingencias
- Conclusiones y Recomendaciones

En resumen el *Business Case* es una herramienta operativa que no solamente debería ser útil para evaluar las inversiones -aspecto clave- antes de tomar una decisión sino para el seguimiento posterior de los resultados de dicha inversión. Esto no solamente aportará información sobre las diferencias en los resultados sino también conocimiento empírico para futuros casos y debería ser obligatoria antes de empezar cualquier proceso de *Big Data*.

A la hora de aplicarlo a un proyectó de *Big Data*, tendrá mucha importancia, ya que veremos si nos va ser viable realizar dicha inversión. Obviamente las empresas que quieran implantarlo, su objetivo será obtener la mayor cantidad de información posible que añada valor a su empresas, sino irían en busca de dicha información no tendría sentido implantar un proyecto *Big Data*. El valor añadido que se introduzca en la empresa será mayor en cuanto más completa sea la información, de no ser así los

recursos dedicados al *Big Data* carecerían de valor. Por ello es de gran importancia analizar en qué punto está la empresa antes de comenzar un proyecto como este.

Según un estudio de *Interxion* de marzo de 2013 Figura 16: "El 25% de las empresas han explorado y elaborado un *Business Case* para *Big Data*" Sin embargo, a pesar del clamor mediático e industrial con respecto a *Big Data*, relativamente pocas empresas han conseguido encontrar un lugar para ello en sus propias operaciones: sólo la cuarta parte de los negocios han explorado y encontrado un **Business Case** viable para *Big Data*. No obstante, su aplicación se está teniendo en amplia consideración y un 81% de organizaciones ya han estudiado las posibilidades de *Big Da*ta o tienen intención de hacerlo.

Figura 16. Interxion. Fuente: Estudio 2013 de Interxion "*Big Data* más allá del ruido"

En mi opinión con respecto a los aportes vistos anteriormente, existe gran número de las empresas están interesadas en *Big Data* y como es lógico primeramente lo analizan mediante Business Case para ver si dicho proyecto puede ser beneficiario para las empresas, no obstante existe un gran número de pequeñas y medianas empresas "PYMES" las cuales no puedes llevar un proyecto de tales envergaduras, ya sea porque es un costo demasiado elevado y podrían utilizar otras alternativas o porque simplemente debido a su tamaño no les es necesario.

Pero a pesar de todo preveo un crecimiento constante de las empresas que investigan Big Data apoyándose unas en otras y con el objetivo de obtener un análisis comercial de todos los clientes, oportunidades y diferentes mercados de una manera mucho más detallada que la forma actual.

Seguridad en Big Data

La seguridad unos de los aspectos con más controversia del ámbito de las Tics. Es el principal rechazo de los consumidores a la hora de navegar, comprar o realizar diversas transacciones en internet, junto con la protección de datos y esto no es diferente en el mundo de *Big Data*.

Ambos términos los podemos aunar desde el punto de vista del robo de información, es decir, asegurar que solo las personas adecuadas acceden a cierta información, desde mi punto de vista este es uno de los aspectos más importantes a la hora de desarrollar un proyecto de *Big Data*, sin olvidarnos de cómo se utilizan esos datos (apartado que se ve a más adelante junto con la Ley de protección de Datos).

La seguridad importa tanto a las empresas debido la perdida de información que implica, lo que conllevara a perder dicha ventaja competitiva y sus consecuencias relevantes a estas (pérdida de clientes, de información valiosa…), sino que también importa al consumidor que cada vez es más consciente de cómo es utilizada su información y de su valor para las empresas por ello exigen políticas de seguridad que en muchas ocasiones no se tienen en cuenta. Por todo esto es tan importante tener un buen sistema de seguridad para el *Big Data*.

Un estudio de la compañía McAfee http://www.mcafee.com/us/resources/reports/rp-needle-in-a-datastack.pdf de Enero de 2013 llamado " *Needle in a Datastack" dice: "*las empresas son vulnerables a las brechas de seguridad por su incapacidad para analizar o almacenar adecuadamente grandes cantidades de datos, así al menos lo considera el 35% de los directivos entrevistados.*"*

Es más, el 22% de los consultados por el estudio –realizado por la firma de investigación *Vanson Bourne* el pasado mes de enero de 2013 y en el que participaron 500 directores de TI de Estados Unidos, Reino Unido, Alemania y Australia– asevera que su empresa necesitaría un día para identificar una brecha, y un 5% opina que este proceso les llevaría una semana. Las organizaciones reconocen que, como media, reconocer una brecha de seguridad les llevaría 10 de horas.

Aunque el 73% de los participantes en el estudio afirma poder valorar su estado de seguridad en tiempo real, el 74% confía en su capacidad para detectar amenazas internas en tiempo real, el 78% amenazas perimetrales, el 72% ataques de *malware* de 'día zero' y el 80% controlar de cumplimiento de normativas, lo cierto es que **el** 58% de los consultados indica que sus organizaciones han sufrido una brecha de seguridad en el último año**.** Además, solo el 24% se dio cuenta en pocos minutos.

Lo que muestran estos datos, indican desde McAfee, es una desconexión entre los departamentos de TI y los profesionales de la seguridad dentro de las organizaciones**;** organizaciones, por otra parte, cada vez más expuestas a amenazas persistentes y cada vez más avanzadas avanzadas.

Por lo tanto queda patente que es necesario intensificar la seguridad en los modelos de *Big Data*. Para ello ya existe software específico como por ejemplo los de la compañía:

http://www.lookwisesolutions.com/index.php/es/

Lookwise Solutions, compañía "spin-off" del grupo S21sec, está dedicada al desarrollo de productos desde hace 10 años, que dan respuesta a las necesidades de las organizaciones en materia de gestión de la seguridad, *Big Data* y de cumplimiento normativo.

No obstante (Bárbara Madariaga 2013) dice en su artículo: "*Big Data* va a revolucionar el sector de la seguridad, siendo el impulsor de los cambios que se van a realizar en el mismo alimentando los modelos de seguridad basados en la inteligencia. "En consecuencia, se espera que *Big Data* altere seriamente casi todas las disciplinas conocidas dentro de la seguridad de información"". De lo que se puede sacar la conclusión de que el propio *Big Data* va a revolucionar la seguridad aplicando su propia tecnología y así conseguir que las amenazas desaparezcan..

Y no se puede olvidar los diferentes modelos de Seguridad que existen sobre TICS:

- **ISO 17799:**

ISO / IEC 17799:2005 establece los lineamientos y principios generales para iniciar, implementar, mantener y mejorar la gestión de seguridad de la información en una organización. Los objetivos describen ofrecer orientaciones generales sobre las metas comúnmente aceptadas de gestión de seguridad de información. ISO / IEC 17799:2005 contiene las mejores prácticas de los objetivos de control y controles en las siguientes áreas de gestión de seguridad de la información.

- **COBIT**:

El COBIT es precisamente un modelo para auditar la gestión y control de los sistemas de información y tecnología, orientado a todos los sectores de una organización, es decir, administradores IT, usuarios y por supuesto, los auditores involucrados en el proceso. El COBIT es un modelo de evaluación y

monitoreo que enfatiza en el control de negocios y la seguridad IT y que abarca controles específicos de IT desde una perspectiva de negocios.

- **ITIL**:

Information Technology Infrastructure Library (‘Biblioteca de Infraestructura de Tecnologías de Información’), frecuentemente abreviada ITIL, es un marco de trabajo de las mejores prácticas destinadas a facilitar la entrega de servicios de tecnologías de la información (TI) de alta calidad. ITIL resume un extenso conjunto de procedimientos de gestión ideados para ayudar a las organizaciones a lograr calidad y eficiencia en las operaciones de TI. Estos procedimientos son independientes del proveedor y han sido desarrollados para servir de guía para que abarque toda infraestructura, desarrollo y operaciones de TI.

- **ISO SERIE 2700**

ISO/IEC 27000 es un conjunto de estándares desarrollados -o en fase de desarrollo- por ISO (International Organization for Standardization) e IEC (International Electrotechnical Commission), que proporcionan un marco de gestión de la seguridad de la información utilizable por cualquier tipo de organización, pública o privada, grande o pequeña.

Existen más modelos que su implantación dependerá del tipo de empresa, presupuesto… que conseguirán que las empresas que utilicen *Big Data* sean más seguras, no obstante estos modelos deben evolucionar para dar mayor seguridad y ser acompañados con otro software más específico.

Ley de protección de datos y Big Data

El Artículo 1 de la Ley Orgánica 15/2009 de 13 de diciembre dice: “La presente Ley Orgánica tiene por objeto garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar”.

La protección de datos personales (que es un derecho fundamental y que por tanto tiene la máxima protección), está recogida en diversos documentos internacionales. En España está regulado por una Ley Orgánica (que requiere una mayoría cualificada de parlamentarios para ser aprobada -la mitad más uno de todos los diputados de los que se compone el hemiciclo-). Se reservan a este tipo de materias (desarrollo de derechos fundamentales).

Como afecta la protección de datos a *Big Data*, tenemos que recordar que *Big Data* procesa 3 tipos de contenidos diferentes: estructurados, semi-estructurados y no estructurados. Estos últimos, pueden provenir de redes sociales por lo que una fuga de estos supondría una vulnerabilidad de sus derechos aunque previamente dicha información haya sido cedida por dichos usuarios.

A mi parecer *Big Data* ha llegado para quedarse y se va ir integrando cada vez más en todo tipo de empresas, pero para que realmente este bien implementado, deben de mejorarse todos los sistemas de seguridad y modelos de seguridad actualmente vigentes, ya que de no ser así, existirán multitud de brechas que vulneren tanto a los usuarios como a las mismas empresas. Por lo tanto se podría decir que *Big Data* y la seguridad van a ir evolucionando de la mano o al menos es lo que se debe de esperar.

Casos de empresas que utilizan Big Data

YELP [11]

Yelp es una guía urbana y electrónica que ayuda a la gente a encontrar los mejores sitios para comer, hacer compras, beber, relajarse y divertirse. Se basa en las opiniones argumentadas de una comunidad vibrante y activa de residentes locales. Yelp es una forma divertida y fácil de encontrar, comentar y hablar de cosas interesantes (y no tan interesantes) de tu mundo.

Fue fundada en 2004 pero en 6 años se convirtió en un fenómeno internacional. En noviembre de 2010 ya tenía más de 39 millones de visitantes y más de 14 millones de críticas.

Yelp utiliza tecnología *Big Data* para revisar todas las críticas o mensajes que dejan sus usuarios. Con ello consigue realizar un filtro de revisión automatizado para identificar contenidos sospechosos y minimizar la exposición al consumidor. Además con la tecnología Big Data almacena información del consumidor para después poderle ofrecer un catálogo personalizado de ofertas especiales de sitios, eventos…. al igual que Amazon con su tienda virtual. A todo lo anterior nombrado hay que añadir que hay que gestionar todas las cuentas gratuitas q se crean y fotos que se

[11] Fuente http://aws.amazon.com/es/solutions/ Amazon 13 de Junio 2012[11]

[1] http://www.yelp.es/faq#what_is_yelp

suben para n enseñar los lugares. Y si quedan dudas de que manejan poca información hay que añadirle toda la información recopilada por las aplicaciones móviles de Smartphone, tabletas…

La tecnología que utiliza Yelp es propiedad de Amazon:

- Amazon Elastic MapReduce: para obtener información de interés para sus usuarios.
- Amazon Storage Service (Amazon S3): almacenar las fotos y registros de los usuarios diarios alrededor de 100GB.

Shazam [12]

Shazam es una aplicación creada para dispositivos móviles como Smartphone y tabletas. Tiene la función de permitir instantáneamente conocer el título de la canción, autor y álbum de la canción que está sonando, con tan solo escuchar unos segundos con el dispositivo la canción.

Tiene una base de datos de más de 60 millones de canciones, más de 200 millones de usuarios registrados. Está disponible en 33 idiomas diferente en los 200 países en los que está presente.

Como funciona Shazam: Nuevamente Shazam utiliza tecnología *Big Data* de Amazon utiliza concretamente:

- Amazon Web Services: con esto Shazam conseguía un conjunto completo de servicios de infraestructura y aplicaciones que permitían ejecutar todo desde la nube.
- Amazon DynamoDB,: Con esto Shazam conseguía un servicio de bases de datos NoSQL rápido y totalmente gestionado que permite almacenar y recuperar de manera fácil cualquier cantidad de datos.
- Amazon Elastic compute Cloud (Amazon EC2): servicio web que proporciona capacidad informática con tamaño modificable en la nube, para que en los grandes eventos en los que multitud de usuarios utilizan Shazam puedan ser atendidos correctamente.

Con esta tecnología Shazam ha conseguido estar en el ranking de las 10 aplicaciones más descargadas de todos los sistemas operativos portátiles (Android, IOS, Windows Phone, Blackberry,…).

[12] http://aws.amazon.com/es/solutions/case-studies/shazam/
[12] Fuente http://aws.amazon.com/es/solutions/ Amazon 13 de Junio 2012[12]

ETSY[13]

Es un mercado en línea que se especializa en artículos hechos a mano, antigüedades y materiales para manualidades. En Etsy cualquier persona puede subir su propia tienda y ofrecer productos por categorías. Se fundó en 2005 y en 2012 cruentaba con más de 800000 tiendas, más de 1.4 billones de visitas mensuales y más de 18 millones de productos en venta.

Como funciona Etsy: También utiliza tecnología *Big Data* perteneciente a Amazon:

- Amazon EC2: servicio web que proporciona capacidad informática con tamaño modificable en la nube, para los grandes picos de usuarios simultáneos.
- Amazon Web Services : Donde aloja Adtuitve, compañía que adquirió como servidor de anuncios. Con esto consiguió orientar anuncios al por menor a un ritmo de 100 millones de consultas por mes. Que observamos con esto, que se analizaba lo que querían los usuarios de Etsy y se les mostraba esos anuncios.
- Amazon S3: Almacenamiento de información.
- Amazon Elastic MapReduce: AEM es capaz de ejecutar docenas de algoritmos en cientos de máquinas para así obtener recomendaciones de interés a los usuarios.

TELEFÓNICA[14]

Telefónica es una empresa española operadora de servicios de telecomunicaciones (telefonía fija, telefonía móvil, ADSL, FTTH, etc.) multinacional con sede central en Madrid, España.

Lo que consigue telefónico gracias a *Big Data*:

- Consumo de móvil por regiones, con lo que se obtendría el nivel socioeconómico de un país. Esto hace unos años solo se podría realizar mediante encuestas y hoy en día es una realidad gracias a Big Data.
- Solución extremo a extremo *Big Data* que incluye una fase de identificación de fuentes internos y externos, una auditoria para procesar datos erróneos y varios modelos de predicciones dependiendo de su cliente objetivo.

[13] http://decoracion.about.com/od/tiendasdemueblesydecoracion/a/Que-es-Etsy.htm
[13] http://aws.amazon.com/es/solutions/case-studies/etsy/
[14] http://bigdata-hadoop.pragsis.com/pages/2/casos_de_uso
http://www.aunclicdelastic.com/los-retos-del-bigdata/ Lorena de la Flor 14 de Junio de 2013

- Evaluación de riesgos para predecir impagos.
- Estudios de predicción de mejores precios.
- Estudios de identificación de medios para publicitar los productos
- Estudio de localización de apertura de nuevos locales.
- Mejora el tiempo de respuesta y reduce costes.

La tecnología utilizada de Big Data por telefónica es:

- Instant Servers: es similar al EC2 de Amazon ya comentado anteriormente, pero con funcionalidades más limitadas. No obstante esta tecnología también la comercializa Telefónica al igual que Amazon el Amazon EC2.La ventaja que tiene Instan Servers respecto a EC2 es su simplicidad a la hora de gestionar redes privadas mientras que EC2 es más complejo. Otra ventaja es el coste, más barato Instant Servers. No obstante EC2 tiene más funcionalidades y está mejor valorado.

PAYPAL[15]

Es una compañía del grupo EBay la cual ha implementado una forma rápida y segura de pagar por internet sin tener la necesidad de compartir la información financiera de las cuentas de crédito con los vendedores. Opera en 195 países con 25 divisas diferentes y con más de 128[16] millones de cuentas activas.

PayPal al igual que otras compañías como Amazon o Google gracias a la tecnología Big Data han desarrollado patrones de actividad fraudulenta. PayPal tiene una serie de filtros de administración de fraudes. Esto lo consigue recopilando datos de sitios fraudulentos tales como datos financieros, direcciones IP, información del navegador, diferente tipología de información con las cual pueden identificar con antelación prevenir transacciones fraudulentas. Gracias a esto PayPal es la forma de pago más fiable de internet y en la que confían más usuarios en todo el mundo

[15] Informe de O´Reilly 8 de Febrero 2011 http://strata.oreilly.com/2011/02/big-data-fraud-protection-payment.html
[16] https://www.paypal-media.com/es/about Cuentas activas fuente : PayPal 2013

EBAY[17]

Es un portal Web de compra y venta en Internet: un lugar en el que se reúnen compradores y vendedores para intercambiar prácticamente de todo. EBay en 2012 tenía más de 100 millones de usuarios activos según sus informes oficiales.

EBay gracias a la tecnología Big Data, consigue beneficios de:

- Búsquedas y anuncios inteligentes.
- Catalogo inteligente.
- Búsqueda de patrones de vendedores fraudulentos.
- Almacenamiento de cuentas de usuario, anuncios, fotos…
- Búsqueda de artículos más rápidos

Los ingenieros de EBay comunican que gracias a Hadoop son capaces de acceder a más de 300 millones de anuncios de empresas, además de obtener gran cantidad de información histórica con lo que les permite entender a todos los clientes.

Para 2015 prevé un volumen de ventas de 86000 millones de euros este nivel de transacciones no sería posible sin la tecnología Big Data

FACEBOOK

Facebook es una red social que crece día a día y acumula más de 100 peta bytes. Gran parte de los ingresos de FB son gracias a la publicidad. Gracias a *Big Data* FB gestiona toda su publicidad de manera inteligente dirigiéndola a los usuarios que la requieren. Otro ejemplo claro de la utilización de *Big Data* es el etiquetado inteligente que detecta los rostros según los usuarios. Pero el ejemplo más claro es el manejo de las bases de datos de FB donde almacena fotos, perfiles, conversaciones, usuarios en más de 50000 servidores.

Si a todo esto le añadimos que también recopila información de nuestros dispositivos móviles y de los navegadores que utilizamos nos podemos imaginar la gran cantidad de datos que tiene que manejar FB.

[17] http://investor.ebayinc.com/#&panel1-2

Se utiliza Hadoop para almacenar copias de registro interno y las fuentes de datos de dimensiones y lo utilizan como una fuente para la presentación de informes / análisis y aprendizaje automático.

A fecha de 19/06/2013 tienen 2 grandes grupos:

- Un grupo 1100-máquina con 8.800 núcleos y un 12 PB de almacenamiento de crudo.
- Un grupo de 300 máquinas con 2.400 núcleos y unos 3 PB de almacenamiento de crudo.
- Cada nodo (productos básicos) tiene 8 núcleos y 12 TB de almacenamiento.

Fuente: http://wiki.apache.org/hadoop/PoweredBy#F

LastFM

LastFM es un servicio de recomendaciones musicales. Para ello LastFM analiza la música que miles de usuarios están escuchando en cada momento y así recopilar información sobre que canciones le gustan más a cada usuario, con qué frecuencia o momentos del día escuchan las diferentes canciones, etc. Una vez recopilada dicha información se compara con la de millones de oyentes de todo el mundo para así recomendarte la música, los artistas, los álbumes que más les gustan a los usuarios. Además Lo utilizan para el cálculo de tablas, informes sobre canciones más escuchadas en todo el mundo, análisis de funciones de audio a gran escala. Sin olvidarnos que LastFM es capaz de recoger las pistas de MP3 de nuestros dispositivos portátiles y de otros software como Spotify.

Para realizar esto LastFM utiliza Hadoop:

a. Con más de 100 nodos.
b. Dual quad-core Xeon L5520@2.27GHz y L5630@2.13GHz, 24 GB de RAM, 8 TB (4x2TB) / nodo de almacenamiento.

Spotify: 

Fuente: http://files.meetup.com/5139282/SHUG%201%20-%20Hadoop%20at%20Spotify.pdf

Spotify es la plataforma musical más conocida actualmente. Spotify ofrece múltiples servicios aunque el principal es ofrecer música en “Streming”, es decir, a través de internet, para ello tiene una base de datos con más de 10000 canciones. No obstante ofrecer música no es su único servicio al igual que LastFM ofrece recomendaciones musicales por gustos, ofrece radios interactivas, posibilidad de conectarse con diversos usuarios para ver la música que escuchan incluso de conectarse a diversas

redes sociales. Y no debemos olvidar que Spotify no es un servicio gratuito y tiene que implementar formas de pago seguras como ya se comentó anteriormente de EBay y PayPal.

A continuación se muestran datos recogidos por Spotify en Hadoop:

- 200 GB de datos comprimidos de usuarios por día.
- 100GB de datos de servicios por día.
- 60GB de datos generados por Hadoop al día.
- 190 nodos por clúster de 4 PB de capacidad de almacenamiento.

En la figura 15 se puede observar la infraestructura interna de Spotify:

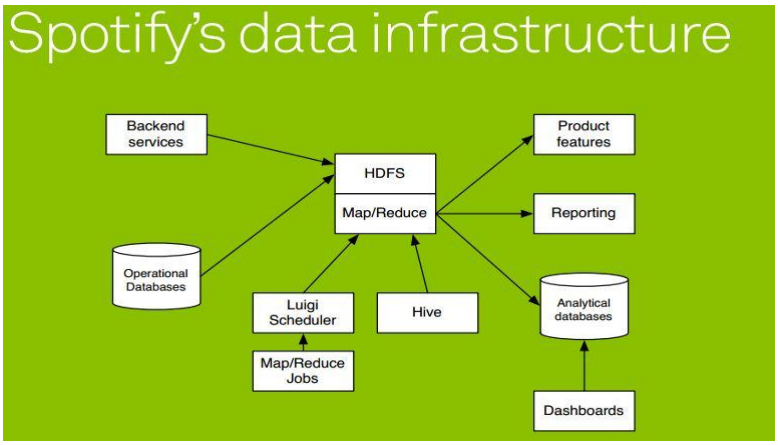

Figura 15 Infraestructura Spotify.

Fuente: http://files.meetup.com/5139282/SHUG%201%20-%20Hadoop%20at%20Spotify.pdf Enero 2013

LinkedIn:

LinkedIn es la mayor red profesional del mundo con más de 225 millones de usuarios con el objetivo de poner en contacto a profesionales del mundo laboral para ayudarles a aumentar su productividad y rendimiento. Cuando un usuario accede a LinkedIn obtiene acceso a personas, empleos, noticias, actualizaciones e información en tiempo real.

La base de datos de LinkedIn es inmensa debido a la cantidad de CV que ingresan los usuarios o empresas en busca de trabajadores. Además realiza análisis en búsqueda de recomendaciones de trabajo según los perfiles de cada usuario, almacenamiento de fotos, correos internos, como se puede observar es una red social con mucho tránsito de datos.

Para realizar esto tiene:

- Hardware:
    - 800 Westmere basado HP SL 170x, con núcleos de 2x4, 24 GB de RAM, 6x2TB SATA.
    - 1900 Westmere basado SuperMicro X8DTT-H, con núcleos de 2x6, 24 GB de RAM, 6x2TB SATA.
    - 1400 basados en puente de arena SuperMicro con 2x6 núcleos, 32GB RAM, 6x2TB SATA.
- Software:
    - RHEL 6.3.
    - Sun JDK 1.6.0_32.
    - Apache Hadoop 0.20.2 + parches y Apache Hadoop 1.0.4 + parches.
    - PIG: analiza los grandes conjuntos de datos, es capaz de manejar cualquier tipo de dato.
    - HIVE: es una infraestructura de Data Warehouse, que facilita administrar grandes volúmenes de datos.

Twitter: 

Twitter es una aplicación web de microblogging, que reúne las ventajas de los blogs, las redes sociales y mensajería instantánea, de este modo los usuarios pueden estar en contacto en tiempo real con personas de su interés con mensajes de no más de 140 caracteres.

Actualmente Twitter tiene almacenados cerca de 12 Terabytes de Tweets creados diariamente, por ello para gestionar tan descomunal número necesita de tecnologías *Big Data.*

Para ello utiliza:

- Web and Social Media: es un tipo de contenido de *Big Data* que se recopila en las redes sociales.
- Cassandra: es una base de datos no relacional distribuida. Permite el manejo de grandes volúmenes de datos.
- Hive: es una infraestructura de Data Warehouse, que facilita administrar grandes volúmenes de datos.

Conclusiones

De este trabajo se han obtenido diferentes conclusiones, empezando por averiguar que la tecnología *Big Data* no solo vale para obtener grandes cantidades de datos, sino que también sirven para analizar esos enormes volúmenes de datos y conseguir así información y conocimiento. También se ha averiguado que es una tecnología que está emergiendo poco a poco y que a pesar de que no todas las empresas tienen porque acudir a ella, ha nacido para quedarse y marcar una nueva etapa en el mundo de las Tics. Gracias a ella en estos momentos hay organizaciones que poseen grandes cantidades de nuestra información privada y debe de ser regulada fuertemente.

Por otro lado ha sido expuesto como antes de instalar en una organización *Big Data* es necesario hacer un *Business Case* para ver si es viable o no el proyecto. Con respecto a las organizaciones que puedan obtener la tecnología *Big Data*, les permitirá crear unas imágenes más complejas de las preferencias y demandas de los clientes, además de sus debilidades y de la de sus competidores con lo que obtendrán una gran ventaja competitiva. No obstantes estas empresas tendrán que tener una infraestructura muy completa para no sufrir robos de datos y sus indemnizaciones correspondientes a sus clientes. Han quedado patentes como importantes compañías como Facebook, Twitter, Shazam o Spotify se han convertido en compañías imbatibles gracias a esta tecnología.

También se ha mostrado innumerable técnicas, algoritmos relacionados con *Big Data* además de métodos relacionadas con Business Intelligence, Data Warehouse y Data Mining y que estas últimas no deben de ser tratadas como una tecnología diferente, sino como una evolución necesaria para la época en la que estamos de la "era de la información".

Por último y no por ello menos importante, a pesar de que hoy en día se diga que estamos en "La era de la Información" en la que se generan enormes cantidades de datos, lo que actualmente parecen enormes cantidades pronto se convertirían en ínfimas. Por ello la tecnología *Big Data* debe y puede seguir evolucionando y de este modo aprovechar la gran avalancha de datos, sin olvidar por otro lado que no todos ellos son útiles, que existe mucho ruido entre ellos. Destacando la importancia de captar únicamente aquellos que puedan ser posteriormente transformados en información y conocimiento.

Agradecimientos

*En primer lugar agradecer a mi tutora **Rocío Rocha Blanco**, por haberme aconsejado durante todo este proyecto, dándome diferentes ideas, puntos de vista e incluso material informativo para completar este trabajo girándome de este modo hacia la culminación del mismo.*

*En segundo lugar a la señorita **Marta Orcajo**, por a verme apoyado durante la realización del trabajo en todo momento.*

*En tercer lugar a mi **familia** quien sin su ayuda no hubiera podido llegar hasta aquí ni asistir a esta universidad.*

*En cuarto lugar a mis **compañeros** del Servicio de Informática que me permitieron escoger las vacaciones para poder acabar con éxito este trabajo.*

*En quinto lugar a mis **amigos**, SIMPRE están ahí.*

*Y por último a la "**Hamburguesería EL PUENTE**", ese lugar de trabajo, que es tan cálido y familiar.*

Bibliografía

- Wikipedia (2013). http://es.wikipedia.org/wiki/Big_data
- Fidelity (2012):https://www.fondosfidelity.es/static/pdfs/informes-fondos/Fidelity_ArgInvSXXI_BigData_Sept12_ES.pdf
- TicBeat (2012). http://www.ticbeat.com/libreriaticbeat/big-data/
- TicBeat (2013). http://bigdata.ticbeat.com/big-data-seguridad-matrimonio-bien-avenido-ignorado-por-las-empresas/
- Eureka-Startups. Artículo escrito por *Vauzza* (2013) "Todo lo que necesitas saber sobre Big Data" http://www.eureka-startups.com/blog/2013/05/28/todo-lo-que-necesitas-saber-sobre-big-data/
- IBM (2012). Artículo de *Ricardo Barranco Fragoso* http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/
- IBM (2013): Articulo de Peter J. Jamack "Analitica de Inteligencia de negocios de Big Data" (2013) http://www.ibm.com/developerworks/ssa/library/ba-big-data-bi/
- DOMO (2012): Artículo escrito por Josh James "How Much Data is Created Every Minute?" http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/
- EMC (2010). artículo especial de "*The Economist"* http://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf
- EMC (2011). Artículo de Bill Schimarzo "*Analisis de Big Data*" http://spain.emc.com/collateral/emc-perspective/h8668-ep-cloud-big-data-analytics.pdf
- CSO ESPAÑA (2013). Artículo de Bárbara Madariaga http://www.csospain.es/El-Big-Data-revolucionara-la-seguridad-de-la-informacion-/seccion-actualidad/noticia-129680
- McAfee (2013). "Needle In a Datastack: The rise of big security data " http://www.mcafee.com/us/resources/reports/rp-needle-in-a-datastack.pdf
- SPOTIFY https://www.spotify.com/es/get-spotify/go/premium/?free_trial=true
- LOOKWISE http://www.lookwisesolutions.com/index.php/es/compania/noticias-y-eventos/noticias/index.php
- DAEDALUS: http://www.daedalus.es/que-tecnologias-nos-diferencian/inteligencia-de-negocio/mineria-de-datos/
- RACKSPACE: http://www.rackspace.com/es/cloud/servers/
- DMG (Data Mining Group). http://www.dmg.org/products.html

- Concepto 05. Artículo de Inés Gómez Plaza: "Analisis redes sociales en España" (2013). http://www.concepto05.com/2013/07/estadisticas-usuarios-redes-sociales-en-espana-2013/
- BBVA. Artículo "En qué punto estamos".(2013) https://www.centrodeinnovacionbbva.com/magazines/innovation-edge/publications/21-big-data/posts/153-big-data-en-que-punto-estamos
- Ontsi. Artículo " Porcentaje de hogares conectados a internet" (2012) http://www.ontsi.red.es/ontsi/es/indicador/hogares-conectados-internet
- Policy Exchange. Articulo (2013) de *Cris Yiu* http://www.policyexchange.org.uk/images/publications/the%20big%20data%20opportunity.pdf
- PayPal https://www.paypal-media.com/es/about
- EBay http://investor.ebayinc.com/#&panel1-2
- O´reilly Starta making data work (2012): http://strata.oreilly.com/2011/02/big-data-fraud-protection-payment.html
- AMAZON http://aws.amazon.com/es/big-data/ ; http://aws.amazon.com/es/solutions/case-studies/etsy/
- RedSeguridad.com. Artículo de Angel Gallego "*Big Data* proyecta la era de la seguridad Inteligente" (2013) http://www.redseguridad.com/empresas/fabricantes/big-data-proyecta-la-era-de-la-seguridad-inteligente
- RedSeguridad.com. Artículo de José Manuel Rodríguez de Llano http://www.redseguridad.com/opinion/articulos/seguridad-ante-el-fenomeno-big-data
- PragSis. Artículo "Casos de Uso de Big Data" (2012) http://bigdata-hadoop.pragsis.com/pages/2/casos_de_uso
- Xataka. Artículo "Big Data, Big Business: Tu vida en una factura telefónica" (2013) http://www.xatakaon.com/entrevistas/big-data-big-business-tu-vida-en-una-factura-telefonica
- Lapastillaroja.net. Artículo de Sergio Montoro Marzo: http://lapastillaroja.net/2013/03/eco-bigdata/ (2013)
- Oracle "Big Data Appliance" (2013) https://shop.oracle.com/pls/ostore/product?p1=OracleBigDataAppliance&p2=&p3=&p4=&sc=ocom_BigDataAppliance
- Revista Cloud Computing. Artículo "Entrevista a Victor Mayer-Schönberger" (2013) http://www.revistacloudcomputing.com/2013/08/entrevista-a-viktor-mayer-schonberger-autor-del-libro-big-data-la-revolucion-de-los-datos-masivos/

- INTERXION:.Artículo “El Big Data eclosiona en España” (2012) http://www.vecdis.es/el-big-data-eclosiona-en-espana/
- INTERXION. Artículo “Big Data más allá del ruido” (2013) http://www.interxion.com/Documents/Whitepapers%20and%20PDFs/Big%20Data/Big_Data-Beyond-hype-es.pdf
- Bloggin Zenith. Artículo “El Big Data según los expertos: implantación retos y situación española” (2013) http://blogginzenith.zenithmedia.es/actualidad/el-big-data-segun-los-expertos-implantacion-retos-y-situacion-espanola-v/
- Vecdis. Artículo “Como crear ventajas competitivas a partir de la información: Big Data” (2012) http://www.vecdis.es/como-crear-ventajas-competitivas-a-partir-de-la-informacion-bigdata-2012/
- El Blog de German. Artículo escrito por German Piñeiro “Big Data ¿Qué es?” (2013) http://www.elblogdegerman.com/2013/03/18/big-data-que-es-ejemplo-de-aplicaciones-del-concepto/
- Tech&Roi. Artículo de Gustavo Tamaki “La hora del Big Data” (2012) http://www.techroi.com.pe/techroi/thechroi/13/82/la-hora-del-big-data
- Documania 2.0. Artículo de Raul G. Beneyto (2013). “Cuanta información se genera en el mundo“ http://documania20.wordpress.com/2013/09/16/cuanta-informacion-se-genera-y-almacena-en-el-mundo/
- NoSQL.es. Artículo de Cristian Requena “NoSql” (2010). http://www.nosql.es/blog/nosql/mapreduce.html
- Aprendiendo Business Intelligence. Artículo de Antonio Rivas “hadopp en proyectos Business Intelligence”. (2011) http://www.bi.dev42.es/2011/12/17/hadoop-en-proyectos-de-business-intelligence/
- Smartes Computing Blog. Artículo de Crystal Anderson “Whats is Big Data” (2013) http://www.smartercomputingblog.com/big-data/what-is-big-data-and-why-does-it-matter/
- Outsourceando. Artículo “tipos de datos” (2013) http://outsourceando.blogspot.com.es/2013/05/tipos.datos.big.data.html
- Pentaho. http://www.pentaho.com/

Referencias:

- *Bárbara Madriaga*. (2013). "El Big Data revolucionará la seguridad de la información". CSO ESPAÑA
- *Bill Schimarzo* (2011). "Análisis de Big Data"
- *Cris Yiu*. (2013) "The Big Data Opportunity".
- *Cristian Requena*. (2010). "NoSql"
- *Edgar Codd* (1970) "A Relational Model of Data for Large Shared Data Banks"
- *Gartner*. (2012). "The Importance of Big Data" ; (2013) "Top Technology Predictions for 2013 and Beyond".
- *Gustavo tamaki*. (2012). "La hora del Big Data"
- *Inés Gómez Plaza*. (2013) "Análisis redes sociales en España"
- *Josh Jam*es. (2012). "How Much Data is Created Every Minute?"
- *Kennet Cukier*. (2010). "The Economist Data, Data Everywhere"
- *Manyika. J ; Chul M. ; Brown M*. (2011). "Big Data: The next frontier for innovation, competition and opportunity"; Mckinsey Global Intitute
- *Michael Schroeck, Rebecca Shockley, Dra. Janet Smart, Dolores Romero-Morales, Peter Tufano.* (2012). "Analytics: el uso de Big Data en el mundo real", IBM Institute for Business Value, Escuela de Negocios Saïd en la Universidad de Oxford.
- *Raul G. Beneyto* (2013). "Cuanta información se genera en el mundo"
- *Ricardo Barranco Fragoso* (2012). "¿Que es Big Data?".
- *Victor Mayer-Schönberger Kenneth Cukier*. (2013). "Big Data: La revolución de los datos masivos", Universidad de Oxford

Organizaciones:

- *ComScore*: es una empresa líder en la medición de internet que proporciona análisis para el Mundo Digital™. ComScore mide cómo navegan las personas en el mundo digital – y convierte estos datos en información y acciones para que nuestros clientes maximicen el valor de sus inversiones digitales. http://www.comscore.com/
- Concepto 05: Agencia de Marketing Online http://www.concepto05.com/nosotros/
- *DOMO*: Compañía de gestión empresarial http://www.domo.com/schedule-a-demo/learn/1

- *EMC:* Empresa multinacional, fabricante de software y sistemas para administración y almacenamiento de información. http://spain.emc.com/index.htm
- Fidelity Worldwide Investment: es una gestora internacional de fondos de inversión.
- *IBM*: Empresa multinacional: http://www.ibm.com/es/es/
- *Icrunchdata*: es un interesante portal de empleo para todos aquellos que quieran trabajar en Big Data o en BI en los EEUU. http://www.icrunchdata.com/
- IDC: principal proveedor global de inteligencia de mercado, servicios de asesoría y organización de eventos para los mercados de tecnologías de la información y las comunicaciones. http://www.idcspain.com/
- *INE*: Instituto nacional de Estadistica www.ine.es
- *Interxion*: Empresa líder en Europa en centros de datos independientes para el alojamiento de equipos Tics. http://www.interxion.com/es/quienes-somos/
- *McKinsey Global Insitute*: Firma de consultoria global McKinsey Global Insitute
- *ONTSI*: Observador nacional de las telecomunicaciones y de las SI. Fuente: http://www.ontsi.red.es/ontsi/
- RED: Entidad pública empresarial adscrita al Ministerio de Industria, Energia y turismo… www.red.es
- Ted: Organización no lucrativa dedicada a la tecnología, entretenimiento y diseño. www.ted.com
- *TICBeat*: Firma de referencia en análisis de Técnologia y Tendencias web en español http://www.ticbea: t.com
- *Vauzza*: Expertos en estrategia e implementación de proyectos tecnológicos. http://vauzza.es/es/
- Vecdis: Empresas varias de servicios y soluciones tecnológicas y gestión del conocimientos http://www.vecdis.es/