# BIG DATA EN LAS EMPRESAS: UNA NUEVA ERA DE LA INFORMACIÓN


**GIOVANNI COCK GOMEZ**

**UNIVERSIDAD MILITAR NUEVA GRANADA**
**FACULTAD DE CIENCIAS ECONÓMICAS**
**ESPECIALIZACIÓN EN ALTA GERENCIA**
**BOGOTÁ D.C. NOVIEMBRE DE 2014**

# BIG DATA EN LAS EMPRESAS: UNA NUEVA ERA DE LA INFORMACIÓN

**GIOVANNI COCK GOMEZ**

**Trabajo presentado como requisito para optar al grado**
**De ESPECIALISTA EN ALTA GERENCIA**

**Asesor metodológico**
**JESÚS SALVADOR MONCADA CERÓN**

**UNIVERSIDAD MILITAR NUEVA GRANADA**
**FACULTAD DE CIENCIAS ECONÓMICAS**
**ESPECIALIZACIÓN EN ALTA GERENCIA**
**BOGOTÁ D.C. NOVIEMBRE DE 2014**

## Resumen

La continua evolución del mercado y la rápida convergencia entre la data, información y conocimiento hace que las empresas busquen nuevos mecanismos tecnológicos para generar un valor diferencial de negocio, fomentando el uso de nuevas tecnologías como Big Data para mejorar la eficiencia operativa, calidad de productos y servicios personalizados que transformen el comportamiento de los usuarios y creando nuevos modelos de negocio. **El presente artículo describe las bases y usos prácticos de Big Data en entornos empresariales.**

## Palabras clave

Data, información, tecnología, empresa, Big Data, modelos de negocio, Hadoop.

## Abstract

The continuing evolution of the market and the rapid convergence of data, information and knowledge makes companies seek new technological mechanisms to generate a differential business value, encouraging the use of new technologies such as Big Data to improve operational efficiency, product quality and personalized services that transform the user behavior and creating new business models. **This paper describes the basis and practical uses of Big Data in enterprise environments.**

## Key Words

Data, information, technology, business, Big Data, business models, Hadoop.

## Introducción

**El presente escrito tiene como objetivo describir la relación de la tecnología, la información y las comunicaciones "TIC[1]" con la alta gerencia y el mundo empresarial, identificar los conceptos básicos, tecnologías de Big Data, su aplicación en mercado y presentar una arquitectura de alto nivel para una solución de Big Data basada en Hadoop[2] para entornos empresariales.**

La estructura general del presente escrito está dispuesta de la siguiente forma: el papel que juega la tecnología y la información en el mundo empresarial, tendencias tecnológicas empresariales, concepto de Big Data, tecnologías y fabricantes en el mercado, y su aplicación al mundo empresarial bajo tres perspectivas: Solución empresarial de Big Data (arquitectura empresarial propuesta), casos de aplicación y mejores prácticas de la industria para proyectos de BI/BA[3] y Big Data.

## La tecnología y la información: la nueva era en el mundo empresarial

La tecnología y la información son esenciales en la operación y cumplimiento de los objetivos de una organización. Para sustentar la importancia de la tecnología en las empresas se utilizó como referencia el articulo La tecnología en la empresa (Carr, 2013)[4], en donde se describe la evolución de la tecnología y su importancia a lo largo de los años desde algunos puntos de vista tales como: cambio cultural en los gerentes y personas encargadas de la operación de las compañías, uso de tecnologías propietarias y el incremento anual de inversión respecto al presupuesto anual de IT. Dicha tasa de crecimiento es proporcional al grado de madurez de la tecnologías de información, lo que ha generado un diferencial estratégico para las compañías respecto a su competencia en el mercado.

Con el pasar de los años la posición privilegiada de IT ha cambiado, denominado por el autor como "desvanecimiento de la ventaja", catalogado por la introducción de tecnologías estándar, homogéneas, visibilidad nula de la tecnología frente a la estrategia empresarial y la adopción de tecnologías en la industria. Con el pasar de los años la posición de IT en el concepto empresarial y estratégico, ha cambiado lo que ha convertido las tecnologías de IT en un commodity, cambiando su posición privilegiada en el mercado a una posición defensiva y altamente competitiva. Los elementos que describe el autor para identificar IT como un commodity son:

- ✓ Alta estandarización y homogenización
- ✓ Escalabilidad de las funciones
- ✓ Proyección de aplicaciones no propietarias

Algunos elementos notables que identifican el inicio de la comoditización del mundo de IT pueden ser la alta oferta tecnológica, alta capacidad de satisfacer necesidades particulares sin incurrir en costes adicionales y el cambio en los modelos de negocio de los fabricantes

[1] Tecnologías de la información y la comunicación (TIC)
[2] Proyecto Big Data de código Abierto
[3] BI/BA: Business Inteligence / business analytics
[4] Carr, N. G. (2013). IT Doesn't Matter. Hardvard Business Review, r0305b.

de IT. Dichos cambios han generado un cambio radical en la oferta y demanda de los mercados. Un claro ejemplo de esto es la incursión de nuevos modelos de negocio basados en web tales como IaaS[5], SaaS[6], PaaS[7], Daas[8], UcaaS[9], entre otros.

Cada cambio en los mercados genera nuevas oportunidades y retos, ej.: pasar de una posición ofensiva a una posición defensiva, lo cual no significa que las tecnologías de IT dejen de ser básicas en las operaciones del día a día de una empresa o sector. La tecnología pasa de ser un elemento diferenciador en la estrategia empresarial a un elemento básico para la buena operación, lo que genera nuevas reglas y prioridades en el entorno empresarial frente a las tecnologías de IT, algunos elementos son:

- ✓ Optimización de gastos y montos de inversión de IT
- ✓ Inversión en tecnologías maduras
- ✓ Enfocar las prioridades de inversión a gestionar posibles vulnerabilidades (Gestión del Riesgo)

La información al ser el corazón de cualquier negocio genera que la tecnología sea el backbone dentro de las arquitecturas empresariales. Los fabricantes e integradores de tecnología tendrán que utilizar toda su experiencia, conocimiento y creatividad para buscar nuevas formas de negocio que le permitan a las empresas a enfocar sus esfuerzos en el core de negocio, la tercerización[10] una salida? Que pasa con la información confidencial de negocio? tendremos que encontrar puntos de equilibrio y adecuar aquellas tendencias tecnologías a necesidades puntuales del negocio, a fin de garantizar que el retorno de inversión sea el esperado por la organización y establecer así a la tecnología como un facilitador del cumplimiento de los objetivos empresariales.

### Tendencias tecnológicas Empresariales (Gartner, 2012)

Gartner[11] define *The Nexus of Forces* como la convergencia e interdependencia de cuatro tendencias fundamentales: *cloud,* social, movilidad y la información **"big data"** (grandes cantidades de datos transformados en información). La dependencia entre estas fuerzas está transformando el comportamiento de los usuarios y creando nuevos modelos de negocio.

5 Infrastructure as a Service (Iaas): En español Infraestructura como Servicio.

6 Software as a Service (SaaS): En español Software como Servicio

7 Platform as a Service (PaaS): En español Plataforma como Servicio

8 Data as a Service (DaaS): En español Datos como servicio

9 Unified Communications as a Service (UCaaS): En español Comunicaciones Unificadas como servicio

[10] Tercerización (outsourcing) es un proceso utilizado por una empresa en la que otra empresa u organización es contratada para desarrollar una determinada área de la empresa.

[11] Gartner Inc. es una empresa consultora y de investigación de las tecnologías de la información.

**Diagrama No. 1 Gartner: "Nexus of Forces"**


Si analizamos la cuarta fuerza "Información", esta está estrechamente relacionada con analítica y Big data, razón por la cual se expondrá alguna visión general de la cuarta fuerza transformadora que está y estará presente en cualquier organización a fin de tener bases suficientes para la toma adecuada de decisiones empresariales.

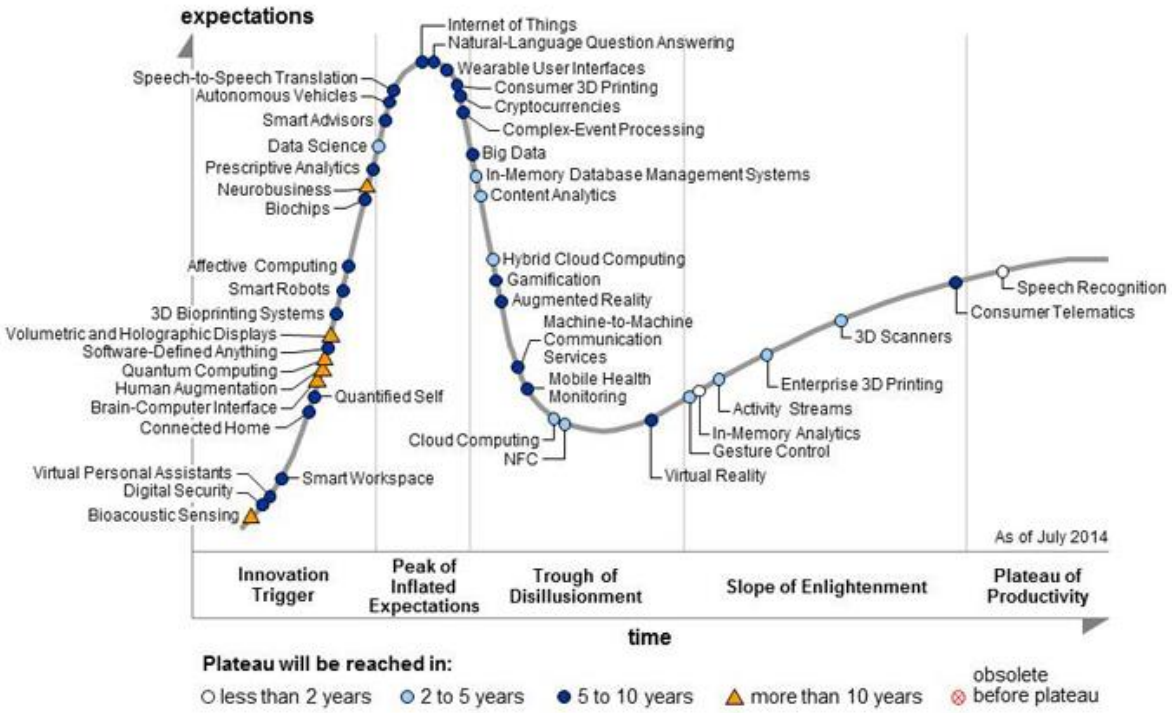
"La información no se almacena en cualquier lugar en particular . Más bien , se almacena en todas partes. La información debe ser entendida como un principio de " evocación " en vez de ser un principio de "encuentro" (Norman, 1981)

"La información será la fuerza más visible a los usuarios finales . la analítica avanzada de Big Data será clave para permitir la transfomacion de los modelos de negocio".[12]

Gartner ha definido el "Hype Cycle for Emerging Technologies", como herramienta para destacar el comportamiento y posición de tecnologías en el mercado según dos variables, expectativa y tiempo.

[12] Gartner. www.gartner.com

## Diagrama No. 2 Gartner hype cycle of emerging technologies 2014

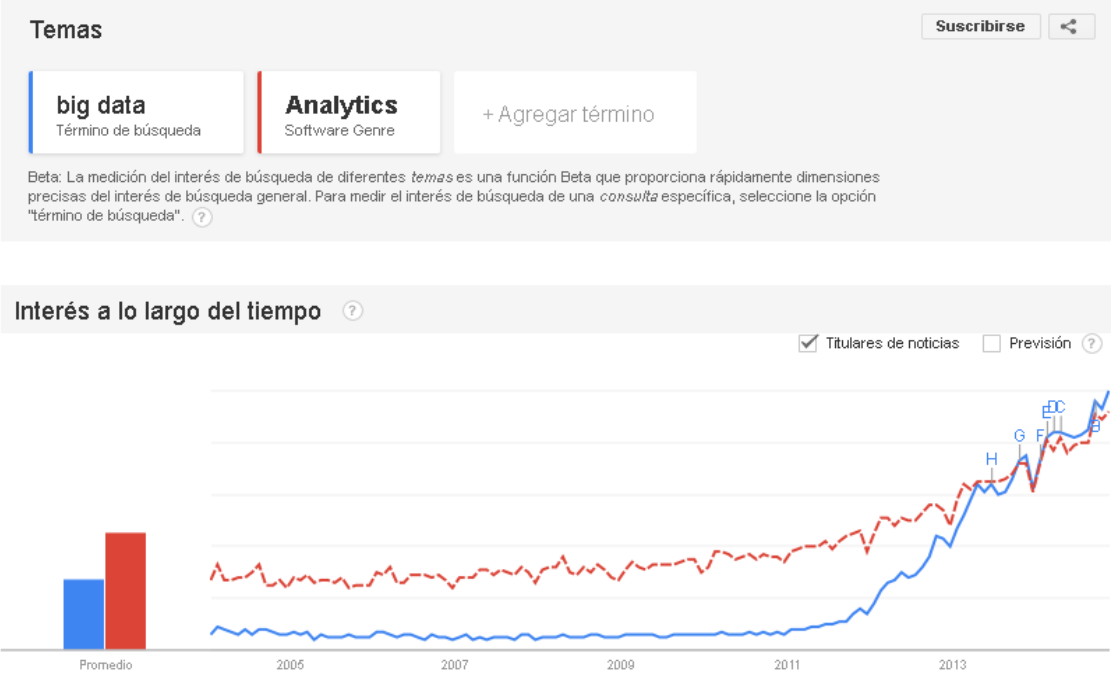



Como tecnología emergente encontramos a Big Data como elemento clave de la fuerza descrita por Gartner como "Información" cuyo objetivo es integrar data estructurada, semi estructurada y no estructurada dentro de la toma de decisiones gerencial. Big data se encuentra en el cuadrante como aquella tecnología que genera unas expectativas relevantes en el mercado.

A continuación se presenta un análisis del comportamiento del termino en Google Trends[13], como fundamento del continuo interés del mercado en los términos de "Big Data" y "analítica de Datos".

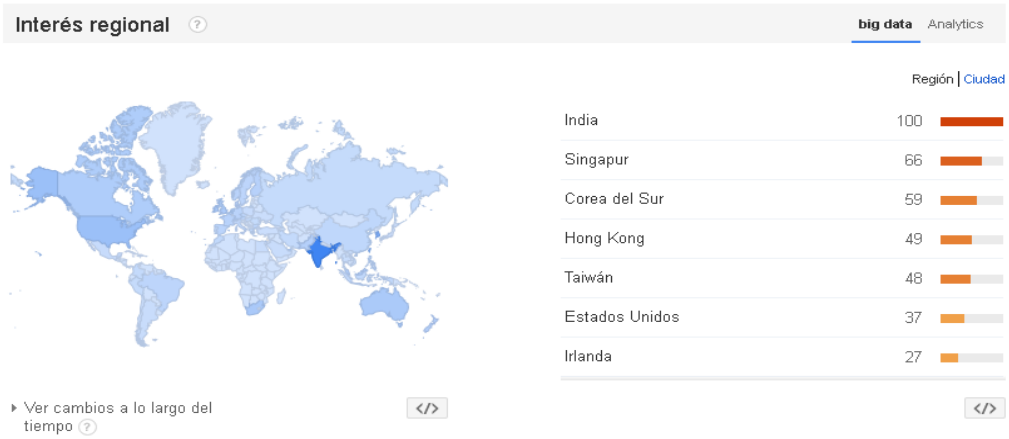
[13] Google Trends es una herramienta de Google Labs que muestra los términos de búsqueda más populares del pasado reciente

## Diagrama No. 3 Google Trends: Big data



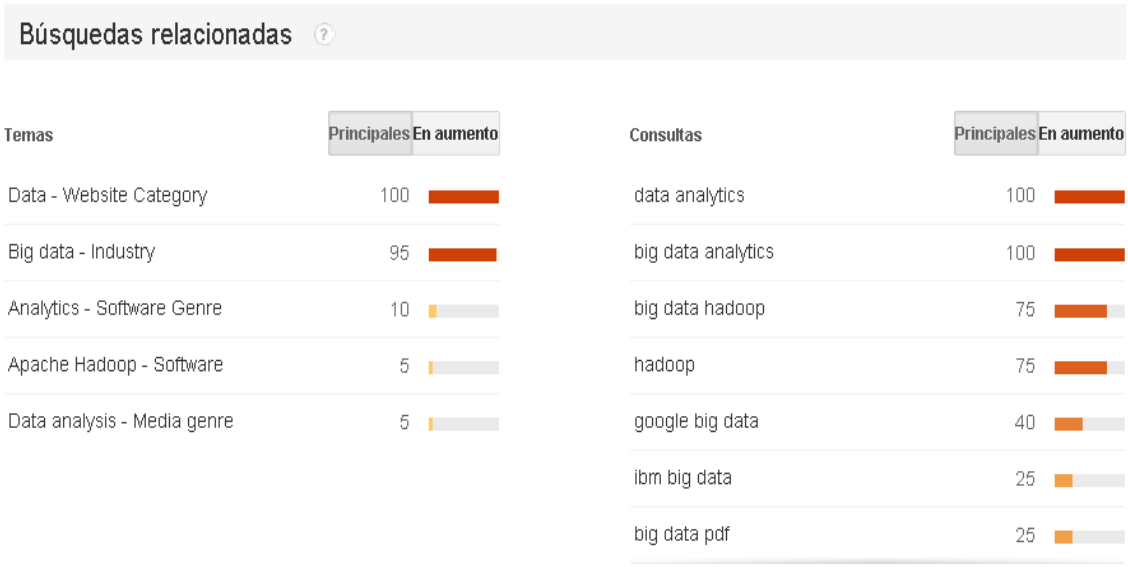

Si analizamos el comportamiento y búsquedas del concepto en Google Trends, encontramos que el termino de Big Data presenta un crecimiento exponencial durante los últimos dos años, lo que demuestra claramente que se posiciona con una tecnología importante en la toma de decisiones empresarial mediante la recopilación, correlación[14] y análisis de data.

## Diagrama No. 4 Google Trends: Big data Interés Regional




[14] Correlación: indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos o más variables estadísticas

## Diagrama No. 5 Google Trends: Big Búsquedas relacionadas




Dentro de las búsquedas relacionadas con Big Data, encontramos que las economías con mayor desarrollo tecnológico tienen un especial interés sobre el termino sin que esto suponga que las económicas en desarrollo no lo tengan. Los términos como Hadoop y Data Analytics[15] se posicionan como elementos con una relevancia importante dentro de Big Data.

## Preparación al concepto de Big Data: Una nueva era de la información

(IBM, 2014)[16] "En términos generales podríamos referirnos a Big Data como a la tendencia tecnológica que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis.

El concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes[17] y exabytes[18] de datos.

Además del gran volumen de información, esta existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones

[15] Data Analytics: es la ciencia de examinar los datos en bruto con el fin de extraer conclusiones acerca de esa información y convertirla en conocimiento.
[16] http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/
[17] Petabyte = 10^15 = 1,000,000,000,000,000
[18] Exabyte = 10^18 = 1,000,000,000,000,000,000

que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data.

Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho más vital cuando se usa en conjunto con la plataforma de Big Data.

Los seres humanos estamos creando y almacenando información constantemente y cada vez más en cantidades astronómicas. Se podría decir que si todos los bits y bytes de datos del último año fueran guardados en CD's, se generaría una gran torre desde la Tierra hasta la Luna y de regreso.

Esta contribución a la acumulación masiva de datos la podemos encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto le añadimos transacciones financieras realizadas en línea o por dispositivos móviles, análisis de redes sociales (en Twitter son cerca de 12 Terabytes de tweets creados diariamente y Facebook almacena alrededor de 100 Petabytes de fotos y videos), ubicación geográfica mediante coordenadas GPS, en otras palabras, todas aquellas actividades que la mayoría de nosotros realizamos varias veces al día con nuestros "smartphones", estamos hablando de que se generan alrededor de 2.5 quintillones[19] de bytes diariamente en el mundo.

De acuerdo con un estudio realizado por Cisco[20], entre el 2011 y el 2016 la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78%, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta. Las naciones unidas proyectan que la población mundial alcanzará los 7.5 billones para el 2016 de tal modo que habrá cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 Exabytes mensuales o 130 Exabytes anuales. Este volumen de tráfico previsto para 2016 equivale a 33 billones de DVDs anuales o 813 cuatrillones de mensajes de texto.

Pero no solamente somos los seres humanos quienes contribuimos a este crecimiento enorme de información, existe también la comunicación denominada máquina a máquina (M2M machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante. Sensores digitales instalados en contenedores para determinar la ruta generada durante una entrega de algún paquete y que esta información sea enviada a las compañías de transportación, sensores en medidores eléctricos para determinar el consumo de energía a intervalos regulares para que sea enviada esta información a las compañías del sector energético. Se estima que hay más de 30 millones de sensores interconectados en distintos sectores como automotriz, transportación, industrial, servicios, comercial, etc. y se espera que este número crezca en un 30% anualmente."

[19] 1 quintillón = 10^30 = 1,000,000,000,000,000,000,000,000,000,000

[20] Cisco Systems es una empresa global con sede en San José, principalmente dedicada a la fabricación, venta, mantenimiento y consultoría de equipos de telecomunicaciones

## Definiciones de Big Data: Perspectivas

**Big Data se denomina todo conjunto de datos que, debido a sus características, excede ampliamente la capacidad de procesamiento de los sistemas tradicionales de gestión de datos dados los grandes volúmenes que se generan a gran velocidad, por múltiples canales y en distintos formatos. "Big Data ha irrumpido en el sector de la tecnología y la información como la mejor solución para recogerlos, almacenarlos, buscarlos, compartirlos, analizarlos, visualizarlos, procesarlos y entenderlos".**

**IDC**[21]: "*Big Data es una nueva generación de tecnologías y arquitecturas diseñadas para extraer el valor económico de grandes volúmenes de una amplia variedad de datos al permitir a alta velocidad de captura, descubrimiento y / o análisis."*

A continuación se expondrán definiciones diferentes de Big Data de reconocidos fabricantes a fin de establecer puntos comunes.

**Cloudera**[22]: "*En términos generales, el término Big Data se refiere a cualquier dato que por cualquier razón (no sólo volumen) no se pueden gestionar asequible por sus sistemas tradicionales. Big Data es un concepto relativo, y es altamente contextual para el medio ambiente. Por ejemplo, incluso si su organización no se acumula datos en una escala similar a Facebook, o incluso si se recoge principalmente un solo tipo de datos, bien puede tener datos grandes desafíos, así como oportunidades*.

*Big Data presenta una gran oportunidad para las empresas en todas las industrias. Al hacer uso de nuevos volúmenes y variedades de los datos, las organizaciones pueden hacer preguntas acerca de sus clientes y su negocio como nunca antes. Por ejemplo, las organizaciones están utilizando los datos para ofrecer una mejor experiencia del cliente, lo que resulta en una base de clientes más leales de la que pueden obtener un mayor valor. Al mismo tiempo, con una mejor comprensión de las operaciones de negocio, es posible identificar áreas de ineficiencia que, si se dirigió, pueden ayudar a reducir los costos de operación."*

**Teradata**[23]: "*Si se hace correctamente, es la unión de negocio y de TI para producir resultados que diferencian, de que el poder que hacia adelante y reducir los costos. Big Data es menos sobre el tamaño de los datos y la información sobre la capacidad de manejar una gran cantidad de diferentes tipos de datos y la aplicación de técnicas de análisis de gran alcance"*.

[21] Empresa de investigación , análisis y de asesoría , especializa en la tecnología de la información , telecomunicaciones y tecnología.

[22] Compañía de software con sede en América que ofrece basada en Hadoop Apache software, soporte y servicios, y capacitación para clientes empresariales

[23] Teradata Corporation es una empresa estadounidense especializada en herramientas de data warehousing y herramientas analíticas empresariales.

**IBM**[24]: *"Todos los días, creamos 2,5 trillones de bytes de datos - tanto que el 90% de los datos en el mundo de hoy se ha creado en los últimos dos años. Estos datos vienen de todas partes: sensores utilizados para recopilar información sobre el clima, los mensajes a sitios de medios sociales, fotos digitales y videos, registros de transacciones de compra, y las señales de teléfono celular GPS para nombrar unos pocos. Estos datos son Big Data".*

Algunas **definiciones complementarias**[25] son:

| | |
|---|---|
| El Big Data original | "Big Data como los tres Vs: Volumen, Velocidad, y la variedad. Esta es la definición más venerable y bien conocida, primero acuñado por Doug Laney de Gartner hace más de doce años. Desde entonces, muchos otros han intentado llevarlo a 11 con adicional Vs incluyendo validez, veracidad, valor y visibilidad." |
| Big Data como Tecnología | "Nuevas tecnologías, y en particular, el aumento rápido de las tecnologías de código abierto tales como Hadoop y otras formas NoSQL de almacenar y manipular datos. Los usuarios de estas nuevas herramientas necesitaban un término que los diferenciaba de las tecnologías anteriores, Big Data." |
| Big Data como señales | "Este es otro enfoque de negocio y que divide al mundo por la intención y el momento en lugar del tipo de datos, por cortesía de SAP, Steve Lucas. El 'viejo mundo' es acerca de las transacciones, y para cuando se registran estas operaciones, ya es demasiado tarde para hacer algo acerca de ellos: las empresas están constantemente 'manejando desde el espejo retrovisor ". En el "nuevo mundo", las empresas pueden utilizar en su lugar los nuevos datos "señal" para anticipar lo que va a suceder, e intervenir para mejorar la situación". |
| Big Data como Oportunidad | Matt Aslett en términos generales define big data como "datos de análisis que fueron ignorados previamente debido a las limitaciones de la tecnología pasada" Técnicamente, Matt utiliza el término "Datos Oscuro" en vez de "Big Data." |
| Big Data como metáfora | En su maravilloso libro El Rostro Humano de Big Data, periodista Rick Smolan dice datos grande es "el proceso de ayudar al planeta crezca un sistema nervioso, una en la que somos más que otro, humano, tipo de sensor." |

Existen algunas características básicas dentro del concepto de Big data, como **volumen** orientado a la cantidad de información, **velocidad** orientado a que tan rápido se genera la información, y la **variedad** orientado a la cantidad de tipos de datos (estructurados, semi estructurados y no estructurados).

[24] IBM es una empresa multinacional estadounidense de tecnología y consultoría
[25] http://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html

En el 2012 IBM definió las características principales de un sistema Big Data a través de 3 V´s:

**Volumen**: Un sistema Big Data debe almacenar y trabajar con grandes cantidades de datos. Este nuevo escenario obliga a que se deban utilizar nuevas infraestructuras, más escalables y distribuidas.

**Velocidad**: Los datos se generan cada vez más rápidamente y además necesitamos transformar esos datos en información útil en un menor tiempo. Los tiempos necesarios en recibir, almacenar y procesar los datos se deben reducir para poder dar una respuesta en el menor tiempo, acercándonos lo más posible al "tiempo real".

**Variedad**: El formato de los datos es más heterogéneo y no se dispone de tiempo ni de medios para homogeneizar los mismos. Nos alejamos del modelo soportado por datos estructurados (Bases de Datos relacionales) y nos adentramos a un nuevo escenario con una mayor cantidad de datos en formatos diferentes (bases de datos, HTML, XML, texto plano, imágenes, video, audio, código fuente, etc.)

Desde 2012 el Big Data ha evolucionado muy rápidamente y estas 3 V´s se han quedado cortas en su definición, por lo que IBM introdujo una nueva: la Veracidad.

**Veracidad:** Los datos que se analizan y procesan deben ser veraces y con ello confiables.

Actualmente cada vez está más aceptada una 5 V:

**Valor**. El objetivo final de todo sistema Big Data debe ser obtener valor de todos los datos disponibles a través de un almacenamiento y procesamiento eficiente y al menor coste posible.

## Big Data en la organización

Dentro del top 3 de prioridades de los CIOs[26] según Gartner del 2013 se encuentran en orden de importancia los siguientes elementos, Analítica e Inteligencia de Negocio, tecnologías móviles y computación en la nube. A continuación se presentan las prioridades tecnológicas de los últimos 6 años, con el fin de tener un referente importante de la relevancia de las tecnologías en la alta gerencia.

Dentro de la analítica e Inteligencia de negocio encontramos a Big Data como referente en la recolección, almacenamiento, análisis, visualización, procesamiento y entendimiento de la data a fin de generar información y convertir esta en conocimiento para las organizaciones a fin de tomar decisiones adecuadas.

[26] CIOs: Chief Information Officer

## Diagrama No. 6 Gartner Prioridades tecnológicas

| Prioridades Tecnológicas | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 |
|---|---|---|---|---|---|---|
| Analytics and Business Intelligence | 1 | 1 | 5 | 5 | 1 | 1 |
| Mobile Technologies | 2 | 2 | 3 | 6 | 12 | 12 |
| Cloud Computing (SaaS, Iaas, Paas) | 3 | 3 | 1 | 2 | 16 | * |
| Collaboration Technologies(workflow) | 4 | 4 | 8 | 11 | 5 | 8 |
| Legacy Modernization | 5 | 6 | 7 | 15 | 4 | 4 |
| IT Management | 6 | 7 | 4 | 10 | * | * |
| Customer Relationship Management | 7 | 8 | 18 | * | * | * |
| Virtualization | 8 | 5 | 2 | 1 | 3 | 3 |
| Security | 9 | 10 | 12 | 9 | 8 | 5 |
| ERP Applications | 10 | 9 | 13 | 14 | 2 | 2 |

Algunos de los beneficios que Big Data puede generar a una organización son:

- ✓ Mejorar la capacidad para adquirir y organizar datos.
- ✓ Mejora en la capacidad de análisis, descubrimiento, predicción, y planificación.
- ✓ Mejores decisiones, reacción rápida, mayor innovación y ventaja competitiva.
- ✓ Obtener una visión completa de los clientes actuales y potenciales a través de múltiples canales.
- ✓ Implementar análisis predictivo para ser más eficaz y proactivo
- ✓ Generar estrategias de marketing personalizadas utilizando analítica avanzada.
- ✓ Reducir la latencia en los procesos críticos de la organización a fin de tener en tiempo real el comportamiento de las variables requeridas para la toma de decisiones.
- ✓ Entender los datos para mejorar la toma de decisiones.
- ✓ Visión de 360 grados para extender la visión del cliente incorporando nuevos canales
- ✓ Seguridad, disminuir el riesgo y lograr la detección de fraudes
- ✓ Análisis de datos para mejorar los resultados del negocio
- ✓ Integrar Big Data al Data warehouse[27] para aumentar la eficiencia.

Algunas de las ventajas que se pueden obtener al utilizar Big Data (sin limitarse a) son:

| Sector | Ventaja |
|---|---|
| Entretenimiento | Análisis de redes sociales<br>Identificar tendencias |
| Medicina y Salud | Análisis de estudios clínicos<br>Prevención de enfermedades |
| Servicios Públicos | Análisis de Smart-meters[28]<br>Predicción de consumo energético |
| Finanzas | Detección de fraude |

[27] Copia de las transacciones de datos específicamente estructurada para la consulta y el análisis

[28] El término de medidor inteligente a menudo se refiere a un contador de electricidad , pero también puede significar un dispositivo de medición de gas natural o el consumo de agua

| | |
|---|---|
| | Patrones de comportamiento de tarjetas de crédito |
| Comercio | Marketing<br>Programas de lealtad de clientes<br>Ofertas personalizadas |
| Gobierno | Seguridad<br>Contra terrorismo |
| Telecomunicaciones | Reducción de deserción de clientes<br>Análisis de CDRs<br>Redes sociales y transacciones |
| Tecnología | Desarrollo de nuevos productos |

## Mercado: Fabricantes y tecnologías

Para la elaboración del cuadro comparativo de soluciones empresariales de Big Data, se procedió a identificar las soluciones más relevantes según consultores especializados en tecnología, entre los cuales se destaca Forrester.

Una de las tecnologías más relevantes y con mayor proyección dentro del ecosistema de Big Data es Hadoop, su principal ventaja radica en el uso de hardware commodity y el uso de una arquitectura escalable horizontal. Para efectos del presente análisis se tomara Hadoop como referente a fin de establecer su posicionamiento en el mercado dado que los grandes fabricantes utilizan el proyecto de código abierto como base de sus soluciones de Big Data.

"*Apache ™ Hadoop® es un proyecto de software de código abierto que permite el procesamiento distribuido de grandes conjuntos de datos a través de las agrupaciones de servidores de productos básicos . Está diseñado para escalar desde un único servidor a miles de máquinas , con un muy alto grado de tolerancia a fallos . En lugar de confiar en el hardware de gama alta , la capacidad de recuperación de estos grupos proviene de la capacidad del software para detectar y manejar las fallas en la capa de aplicación*"[29].

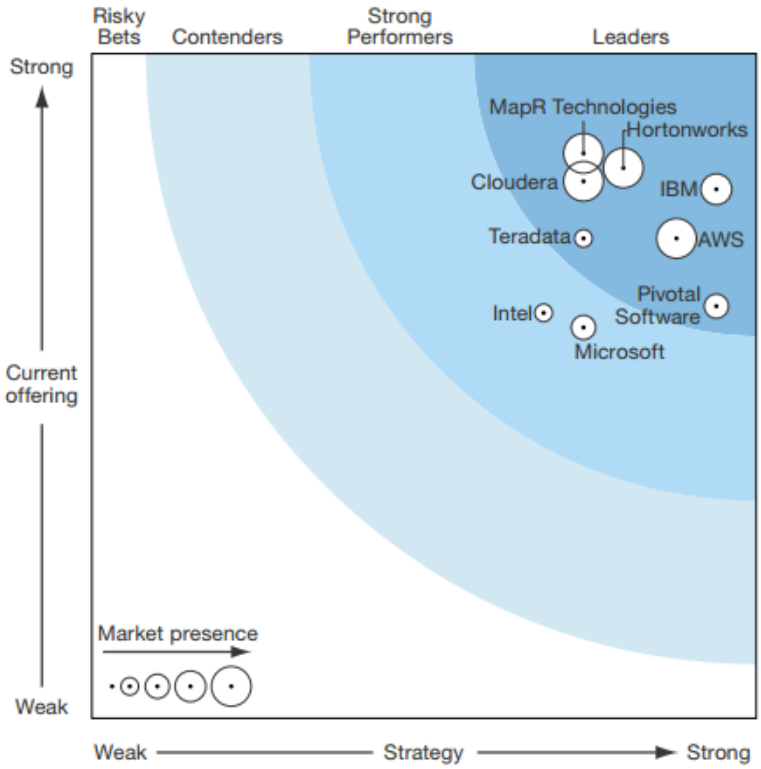
Forrester Research[30] publicó un análisis de Hadoop, en su informe , The Forrester Wave ™: Big Data Hadoop Solutions, Q1 2014, donde se presentan los fabricantes más relevantes en el uso de Hadoop como herramienta clave para soluciones de Big Data.

**Forrester**: "Hadoop es imparable ya que sus raíces provienen de código abierto y que se posiciona profundamente dentro de las empresas como solución reconocida en Big Data. Su enfoque para la gestión de datos está transformando la forma en las empresas almacenan, procesan, analizan y comparten la información". La evaluación general de soluciones de Big Data sobre Hadoop es:

[29] http://www-01.ibm.com/software/data/infosphere/hadoop/

[30] Forrester Research es una empresa independiente de investigación de mercados que brinda asesoramiento sobre el impacto existente y potencial de la tecnología a sus clientes y al público en general

## Diagrama No. 7 Forrester Wave™: Big Data Hadoop Solutions, Q1 '14




Dentro del cuadrante de líderes encontramos Hortnoworks, IBM, AWS, Cloudera, Teradata, Pivotal Software, MapR Tech.

A continuación se presenta la tabla comparativa según la identificación de los fabricantes más relevantes del Mercado a nivel de Big Data segmento empresarial, las variables utilizadas para el análisis comparativo son:

- ✓ Fabricante o vendedor
- ✓ Herramienta o producto utilizado
- ✓ Categoría
- ✓ Características principales
- ✓ Tecnología Base

| ID | Vendedor | Herramienta / producto | Categoria | Caracteristicas principales | Tecnologia Base |
|---|---|---|---|---|---|
| 1 | Hortonworks[31] | Hortonworks Data Platform | Big data Operacional | *comunidad de código abierto<br>* arquitectura tipo ecosistema<br>* Los métodos de procesamiento de lotes - a través interactivo a tiempo real<br>* Plataforma de Datos Hortonworks (HDP)<br>* Implementación para HDP de Linux a Windows, On-Premise a In-Cloud<br>* asociaciones estratégicas clave, como Teradata, Microsoft y SAP | Hadoop |
| 2 | Cloudera[32] | Cloudera Enterprise | Big data Operacional | * Arquitectura Abierta<br>* Procesamiento por lotes, SQL interactivo, búsqueda empresarial y análisis avanzado<br>* Gestión y la vigilancia de herramienta para Hadoop<br>* Motor Analítico SQL para Hadoop, impala<br>* Dentro de la memoria de la máquina de aprendizaje y procesamiento Corriente<br>* procesamiento paralelo masivo (MPP) de la arquitectura<br>* más de 200 clientes de pago | Hadoop |
| 3 | IBM[33] | InfoSphere BigInsights | Big data Operacional | * Proporciona análisis avanzados construidos sobre la tecnología Hadoop<br>* Corriente de computación<br>* Ofrece acceso SQL a los datos almacenados en BigInsights través de Big SQL<br>* Soporte Jaql, para los datos estructurados y no estructurados<br>* El soporte LDAP | Hadoop |

[31] http://hortonworks.com/
[32] http://www.cloudera.com/content/cloudera/en/home.html
[33] http://www.ibm.com/ibm/puresystems/us/en/pd_hadoop.html

| | | | | | |
|---|---|---|---|---|---|
| | | | | * Integración con SPSS analítica avanzada, gestión de carga de trabajo para el alto rendimiento<br>* Más de 100 implementaciones de Hadoop | |
| 4 | Teradata[34] | Teradata Open Distribution for Hadoop (TDH) | Big data Operacional | * Basado en Hortonworks<br>* Solución Appliance<br>* Integración con la herramienta de gestión de Teradata y SQL-H, un motor SQL federada que permite consultar datos a clientes de su almacén de datos y Hadoop<br>* Alto costo / efectiva relación | Hadoop |
| 5 | Amazon Web Services (AWS)[35] | Amazon Elastic MapReduce (Amazon EMR) | Big data Operacional | * Hadoop en la nube (uso Por pago)<br>* Integración con Amazon Kinesis para procesamiento de flujo y almacenamiento de datos Amazon Redshift<br>* autoscaling que cambiar el tamaño de los clústeres basado en políticas<br>* Soporte para bases de datos NoSQL adicionales<br>* Integración de BI con proveedores de terceros | Hadoop |
| 6 | Oracle[36] | Oracle Big Data Appliance | Big data Operacional | * Solución Appliance<br>* Todo el software Cloudera Enterprise Technology incluyendo Cloudera CDH, Cloudera Manager, y Cloudera RTQ (Impala)<br>* Pre-integrado configuración de bastidor completo con 18 de los servidores Sun x86 de Oracle e InfiniBand y Ethernet<br>* Ejecutar diversas cargas de trabajo en el sistema de | Hadoop |

[34] http://www.teradata.com/Teradata-Portfolio-for-Hadoop/
[35] http://aws.amazon.com/es/elasticmapreduce/
[36] http://www.oracle.com/us/products/middleware/data-integration/hadoop/overview/index.html

| | | | | | |
|---|---|---|---|---|---|
| | | | | Hadoop y NoSQL<br>* Las capacidades de seguridad integral, incluyendo autenticación, autorización y auditoría | |
| 8 | HP[37] | Vertica | Big data Operacional | * realizar consultas SQL avanzadas en los datos almacenados a granel en HDFS<br>* entregado como software para plataformas estándar (excluyendo Windows)<br>* disponibles son las AppSystems HP para Vertica, versiones en la nube (a través de HP Cloud Services y Amazon ofrendas Máquina Imagen) y Data Warehouse On Demand - un acogido, servicio gestionado<br>* Remanso de HP, que combina seguridad, Hadoop y datos no estructurados a través de Autonomía | Hadoop |
| 9 | Google[38] | BigQuery | Big data Operacional | * Almacenamiento en columnas<br>* Arquitectura Árbol<br>* Ad hoc y de ensayo y error de consulta interactiva de los datos para un análisis rápido y solución de problemas<br>* Data Mining caso de uso parcial (por ejemplo, el análisis de datos de verificación previa para la minería de datos)<br>* Procesamiento de datos no estructurados parcialmente (expresiones regulares en el texto) | Dremel |

[37] http://h17007.www1.hp.com/us/en/converged-infrastructure/converged-systems/bigdata-hadoop.aspx
[38] https://cloud.google.com/solutions/hadoop/

## Solución Empresarial Propuesta: Arquitectura Big Data basada en Hadoop

La arquitectura de Big Data propuesta para entornos empresariales está compuesta por 6 capas, estas son:


A continuación se presentara cada capa, su concepto general y la herramienta propuesta.

| Capa | Concepto General | Herramientas |
|---|---|---|
| Fuentes de datos | Son conjuntos de elementos generados por fuentes diversas, que pueden estar o no organizados y clasificados de la siguiente forma:<br>✓ Datos estructurados (tablas, arreglos, listas: DB data (CRM, ERP, SCM), datos operacionales<br>✓ Datos semi-estructurados: Documentos, emails, web logs, archivos csv, xml<br>✓ Datos no estructurados: Sensores, dispositivos, data geolocalización, SMS, Interacciones sociales, Audio&Video, imagenes, Web, redes sociales. | No aplica. |
| Colección e integración | Compuesta por herramientas que recopilan e integran las fuentes de datos recolectadas, previo a cualquier proceso de extracción, transformación y carga de data. | Las herramientas seleccionadas son Flume para recolección datos en tiempo real y Sqoop para la colección de diferentes bases |

| | | |
|---|---|---|
| | | de datos relacionales y NoSQL. |
| Área de Staging, agregación | Área de almacenamiento intermedio que puede ser temporal, encargada de extraer, transformar y cargar la data colectada e integrada. La data puede ser procesada previa al almacenamiento o puede ser almacenada previo a su procesamiento. | Las herramientas seleccionadas son Hive, Impala para información interactiva y Strorm para streaming. |
| Almacenamiento | Área de la arquitectura encargada de consolidar en un repositorio centralizado o distribuido la información colectada por las diversas fuentes de data. | El almacenamiento a utilizar será HDFS con su respectivo procesamiento MapReduce. |
| Análisis | Es la capta encargada de transformar la data en información, incluye análisis de texto, análisis predictivo y data mining, cuyo objetivo es detectar patrones de comportamiento a través de algoritmos matemáticos. | La plataforma seleccionada es Mahout. |
| Presentación | Es la capa encargada de presentar la información a los analistas de negocio, mediante la exploración visual e interacción de resultados obtenidos en la capa de análisis. | La opción tecnológica seleccionada para esta función es Tableau. |

### Casos de aplicación , empresas que utilizan Big Data

Dentro de los casos de aplicación[39] y la forma en la que algunas empresas utilizan Big Data encontramos los siguientes ejemplos:

| **Aplicación** | **Relacionamiento con clientes** |
|---|---|
| **Uso** | Las Herramientas de datos grandes permiten que el empleado pueda comprobar el perfil del cliente en tiempo real y saber qué productos o servicios de referencia (s) que podría asesorar |
| **Información fuente** | Histórico de compras, información de redes sociales, comentarios en Blogs, páginas web. |
| **Información generada** | Campañas y recomendaciones de compra específicas para cada cliente (tipificación del perfil) |

| **Aplicación** | **Desarrollo de productos y/o procesos** |
|---|---|
| **Uso** | Entender cómo los demás perciben sus productos de manera que pueda adaptarse. |
| **Información fuente** | Análisis de texto no estructurados los medios sociales le permite |

[39] http://datascienceseries.com/stories/ten-practical-big-data-benefits

| | |
|---|---|
| | descubrir los sentimientos de sus clientes e incluso los segmentos en diferentes ubicaciones geográficas o entre diferentes grupos demográficos. |
| **Información generada** | Big Data permite probar miles de diferentes variaciones de diseños asistidos por ordenador, de modo que se pueda comprobar cómo los cambios de menor importancia, por ejemplo, el material afecta a los costes, los plazos de entrega y rendimiento. A continuación, puede aumentar la eficiencia del proceso de producción en consecuencia<br>-Los minoristas son capaces de optimizar sus acciones sobre la base de las predicciones generadas a partir de datos de medios sociales, las tendencias de búsqueda web y las previsiones meteorológicas |

| **Aplicación** | **Análisis de riesgos** |
|---|---|
| **Uso** | Analizar el nivel de riesgo de un individuo, institución, etc. que permita establecer basado en diferentes entradas si se presenta un nivel de riesgo bajo, medio o elevado, y su impacto según data histórica. |
| **Información fuente** | Escanear y analizar los informes de los periódicos o los medios de comunicación social se alimenta de modo que mantenga permanentemente al día sobre los últimos avances en su industria y su entorno |
| **Información generada** | Análisis detallados y alertas tempranas sobre proveedores y clientes. Esto permitirá que usted tome acción cuando uno de ellos está en riesgo por ejemplo: riesgo no pago |

| **Aplicación** | **Seguridad de la información e informática(mantener sus datos seguros)** |
|---|---|
| **Uso** | Detectar la información potencialmente sensible que no está protegido de manera adecuada y asegúrese de que se almacena de acuerdo con los requisitos reglamentarios |
| **Información fuente** | Fuentes de información diversa orientada a reputación de una empresa, logs de equipos de infraestructura tecnológica, monitoreo de data entrante y saliente de una empresa en pro de salvaguardar el activo "información y conocimiento" |
| **Información generada** | Generación de una alarma que pueda comprometer la fuga de información confidencial, ej: búsqueda de 16 números de un dígito - potencialmente de datos de tarjetas de crédito - que puedan ser almacenados o enviados por correo electrónico. |

| Aplicación | Reducción de los costes de mantenimiento |
|---|---|
| **Uso** | Evitar el reemplazo innecesario de todas las piezas de tecnología al cabo de un número de años, incluso aun cuando los dispositivos puedan tener una vida útil mayor |
| **Información fuente** | Las cantidades masivas de datos que el acceso y uso, y su velocidad sin igual pueden detectar en su defecto los dispositivos de la red y predecir cuándo van a presentar algún tipo de fallo. |
| **Información generada** | Estrategia de reemplazo mucho más rentable para la utilidad y menos tiempo de inactividad, como dispositivos defectuosos se realiza un seguimiento mucho más rápido. |

| Aplicación | Ofertas de salud personalizados |
|---|---|
| **Uso** | Cuando alguien es diagnosticado con una enfermedad por lo general se someten a una terapia, y si eso no funciona, los médicos tratan de otra, etc. Pero ¿qué pasa si un paciente podría recibir medicamentos que se adapte a sus genes individuales? Esto daría lugar a un mejor resultado, menos coste, menos frustración y menos miedo. |
| **Información fuente** | Con la cartografía del genoma humano y las herramientas de Big Data, que pronto será un lugar común para que cada uno tenga sus genes mapeados como parte de su expediente médico. |
| **Información generada** | Relación y entrega de la medicina relacionada con determinantes genéticos que causan una enfermedad y el desarrollo de fármacos diseñados expresamente para tratar esas causas. |

| Aplicación | Deportes: McLaren Formula 1 |
|---|---|
| **Uso** | En McLaren Mercedes, todas las decisiones comienzan con inmensas cantidades de datos en bruto. Análisis de datos es parte de todo el equipo que hace, desde el diseño del coche hasta el día de la carrera. |
| **Información fuente** | El coche McLaren F1 está equipado con más de 160 sensores que constantemente transmiten datos a un equipo de ingenieros y encargados de tomar decisiones que producen un gigabyte de datos en bruto por carrera. |
| **Información generada** | Estadísticas de uso, funcionamiento, correlación de sensores, etc. que ayudan a la toma de decisiones rápida durante una carrera y diseño del coche futuro. |

| Aplicación | Deportes: Brazil 2014 World Cup |
|---|---|
| **Uso** | la Asociación Alemana de Fútbol desplegó tecnologías de analítica avanzada para proporcionar datos a los entrenadores, directivos y jugadores para analizar las mejoras individuales y de equipo. |

| | |
|---|---|
| **Información fuente** | Utilizando el sistema, los datos se recogen a partir de las ocho en las cámaras de campo que compactan millones de puntos de datos por segundo. |
| **Información generada** | Los datos se convierten entonces en simulaciones y gráficos que se pueden ver en una tableta o un teléfono inteligente, lo que permite instructores, entrenadores y jugadores para identificar y evaluar las situaciones clave en cada partido. |

## Mejores prácticas y lecciones aprendidas para proyectos de Big Data

Algunos de los elementos (mejores prácticas)[40] y (lecciones aprendidas)[41] que deben tenerse en cuenta a la hora de emprender un proyecto de Big Data son:

| | |
|---|---|
| Mejores Practicas | ✓ Establecer un mapa de ruta de Big Data en la organización, de corto, medio y largo plazo.<br>✓ Determina las metas y objetivos de la empresa en términos de uso de la data, uso de la información y generación de conocimiento<br>✓ Asegúrese de que su hoja de ruta tiene puntos de referencia que sean razonables y alcanzables.<br>✓ Comience por embarcarse en un proceso de descubrimiento. Usted necesita tener una idea de lo que los datos que ya tiene, dónde está, quién posee y controla, y la forma en que se utiliza actualmente.<br>✓ Conozca claramente las fuentes de datos que tiene y cómo funciona la superposición de la data<br>✓ Averiguar qué datos grande que no tienen, e identificación de posibilidades de mejora<br>✓ Comprender las opciones tecnológicas del mercado en temas de Big Data y analítica, previo a cualquier selección<br>✓ Familiarizarse con las herramientas y técnicas que están surgiendo como parte del gran ecosistema de Big Data<br>✓ Probar y validar continuamente los supuestos que se tengan.<br>✓ Probar continuamente, si se están consiguiendo los resultados que parecen difícil de creer, es importante para evaluar los resultados de forma periódica. |
| Lecciones aprendidas | ✓ Suponer que el usuario medio de negocios tiene el know-how o el conocimiento técnico requerido el momento de utilizar las herramientas de Analítica y Big Data<br>✓ Permitir que Excel se convierta por de defecto en |

[40] http://www.dummies.com/how-to/content/five-big-data-best-practices.html
[41] Recomendaciones en pro de garantizar el éxito de una actividad o proceso

| | |
|---|---|
| | plataformas de analítica y análisis de datos<br>✓ Suponer que un almacén de datos va a resolver todos los accesos de la información, aseguramiento de calidad y definición de los requisitos de negocio o de data<br>✓ Selección de una herramienta de Big Data o Analítica sin una necesidad de negocio específica<br>✓ Asociar que la cantidad de columnas a la calidad del sistema<br>✓ Implementar soluciones o herramientas tecnológicas de forma individual (por departamento)<br>✓ Evaluar, adquirir e implementar las herramientas con una orientación a usuarios avanzados<br>✓ Permitir que los usuarios acumulen su propia información en sus hojas de cálculo personales a un punto en que su información se convierte en una fuente de datos críticos<br>✓ Comenzar un proyecto de BI o BA, antes de que la necesidad de información ha sido identificada, sólo para encontrar que han incurrido en otro gasto y no han resuelto un solo problema<br>✓ Compra de software de BI para el "análisis de propósito general."<br>✓ Intentar cumplir con todas las demandas de su empresa de forma simultánea, implementación de Big Bang.[42] |

[42] Adopción del big bang es un método de migración de hardware o software que consiste en deshacerse del sistema existente y la transferencia de todos los usuarios al nuevo sistema de forma simultánea.

## Conclusiones

- ✓ La tecnología y la información son esenciales en la operación y cumplimiento de los objetivos estratégicos de una organización, por esta razón se hace necesario implementar tecnologías que apalanquen la toma de decisiones en tiempo real y que garanticen la calidad de la información.
- ✓ La información será la fuerza más visible a los usuarios finales. la analítica avanzada de Big Data será clave para permitir la transformación de los modelos de negocio.
- ✓ Para obtener valor de negocio real a partir de Big Data, se requieren herramientas adecuadas para capturar y organizar una amplia variedad de tipos de datos de diferentes fuentes, y ser capaz de analizar fácilmente que en el contexto de todos los datos de la empresa.
- ✓ Big Data se denomina como todo conjunto de datos que, debido a sus características, excede ampliamente la capacidad de procesamiento de los sistemas tradicionales de gestión de datos dados los grandes volúmenes que se generan a gran velocidad, por múltiples canales y en distintos formatos.
- ✓ Una de las tecnologías más relevantes y con mayor proyección dentro del ecosistema de Big Data es Hadoop, su principal ventaja radica en el uso de hardware commodity y el uso de una arquitectura escalable horizontal.
- ✓ El análisis de datos de Big Data puede revelar nuevas fuentes de ingresos, proporcionar nuevas ideas en el comportamiento del cliente e identificar las tendencias del mercado, lo que supone un desafío para los departamentos de IT dado que se requieren herramientas tecnológicas que permitan recoger, almacenar, buscar, compartir, analizar, visualizar, procesar y entender diferentes tipos de datos, con comportamientos no normalizados.

## Bibliografía


Carr, N. G. (2013). IT Doesn't Matter. *Hardvard Business Review*, r0305b.

Forum, M. C. (2012). *Big Data, Analytics, and the Future of Marketing & Sales.* Amazon.

Gartner. (23 de 10 de 2012). *Gartner Identifies the Top 10 Strategic Technology Trends for 2013*. Obtenido de http://www.gartner.com/newsroom/id/2209615

IBM. (2014). *IBM*. Obtenido de http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Needham, J. (2010). *Disruptive Possibilities: How Big Data Changes Everything.* O´Reilly.

Norman, D. R. (1981). *D. Rumelhart and D. Norman.*

Weill, J. W. (2002). Six IT Decisions Your IT People Shouldn't Make. *Hardvard business review*, Product 5895.