IBM **Bluemix** ¡Desarrolla en la nube en un click!

Comience su prueba gratuita

¿Qué es Big Data?

Todos formamos parte de ese gran crecimiento de datos

Debido al gran avance que existe día con día en las tecnologías de informació
han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubri
lo que sus herramientas tradicionales reportan sobre su información, al mismo
últimos años el gran crecimiento de las aplicaciones disponibles en internet (ge
sociales, etc.) han sido parte importante en las decisiones de negocio de las er
artículo tiene como propósito introducir al lector en el concepto de Big Data y c
características de los componentes principales que constituyen una solución d

Ricardo Barranco Fragoso es IT Specialist para Information Management de IBM Software Gr
diciembre de 2010. De las principales soluciones en las que está enfocado se encuentran: Ma
Identity Insight y Big Data; y como este rol participa activamente apoyando las actividades de
de experiencia y previo a su incorporación a IBM, Ricardo ha participado en diversos proyect
Desarrollador Senior, Líder Técnico y Arquitecto Jr. utilizando principalmente tecnología Java
Internet Applications. Ricardo cuenta con una Licenciatura en Computación por la Universida
Iztapalapa.

18-06-2012

1. Introducción

El primer cuestionamiento que posiblemente llegue a su mente en este momento es ¿Qué es Big Data y porqué se ha vuelto tan importante? pues bien, en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología 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 para su análisis. De tal manera que, el concepto de Big Data aplica información que no puede ser procesada o analizada utilizando pro tradicionales. Sin embargo, Big Data no se refiere a alguna cantida usualmente utilizado cuando se habla en términos de petabytes y e Entonces ¿Cuánto es demasiada información de manera que sea e procesada y analizada utilizando Big Data? Analicemos primerame

Gigabyte = 10^9 = *1,000,000,000*
Terabyte = 10^{12} = *1,000,000,000,000*
Petabyte = 10^{15} = *1,000,000,000,000,000*
Exabyte = 10^{18} = *1,000,000,000,000,000,000*

Además del gran ***volumen*** de información, esta existe en una gran pueden ser representados de diversas maneras en todo el mundo, dispositivos móviles, audio, video, sistemas GPS, incontables sens industriales, automóviles, medidores eléctricos, veletas, anemómet pueden medir y comunicar el posicionamiento, movimiento, vibració y hasta los cambios químicos que sufre el aire, de tal forma que las analizan estos datos requieren que la ***velocidad*** de respuesta sea lograr obtener la información correcta en el momento preciso. Estas principales de una oportunidad para Big Data.

Es importante entender que las bases de datos convencionales sor relevante para una solución analítica. De hecho, se vuelve mucho m en conjunto con la plataforma de Big Data. Pensemos en nuestras

derecha, cada una ofrece fortalezas individuales para cada tarea er
un beisbolista sabe que una de sus manos es mejor para lanzar la
atraparla; puede ser que cada mano intente hacer la actividad de la
resultado no será el más óptimo.

2. ¿De dónde proviene toda esa información?

Los seres humanos estamos creando y almacenando información c
vez más en cantidades astronómicas. Se podría decir que si todos
del último año fueran guardados en CD's, se generaría una gran to
la Luna y de regreso.

Esta contribución a la acumulación masiva de datos la podemos en
industrias, las compañías mantienen grandes cantidades de datos t
reuniendo información acerca de sus clientes, proveedores, operac
manera sucede con el sector público. En muchos países se adminis
datos que contienen datos de censo de población, registros médico
todo esto le añadimos transacciones financieras realizadas en línea
móviles, análisis de redes sociales (en Twitter son cerca de 12 Tera
diariamente y Facebook almacena alrededor de 100 Petabytes de f
geográfica mediante coordenadas GPS, en otras palabras, todas ac
mayoría de nosotros realizamos varias veces al día con nuestros "s
hablando de que se generan alrededor de 2.5 quintillones de bytes

1 quintillón = 10^{30} = *1,000,000,000,000,000,000,000,000,000*

De acuerdo con un estudio realizado por Cisco[1], entre el 2011 y e
tráfico de datos móviles crecerá a una tasa anual de 78%, así como
dispositivos móviles conectados a Internet excederá el número de h
Las naciones unidas proyectan que la población mundial alcanzará

2016 de tal modo que habrá cerca de 18.9 billones de dispositivos c
escala mundial, esto conllevaría a que el tráfico global de datos mó
Exabytes mensuales o 130 Exabytes anuales. Este volumen de tráf
equivale a 33 billones de DVDs anuales o 813 cuatrillones de mens

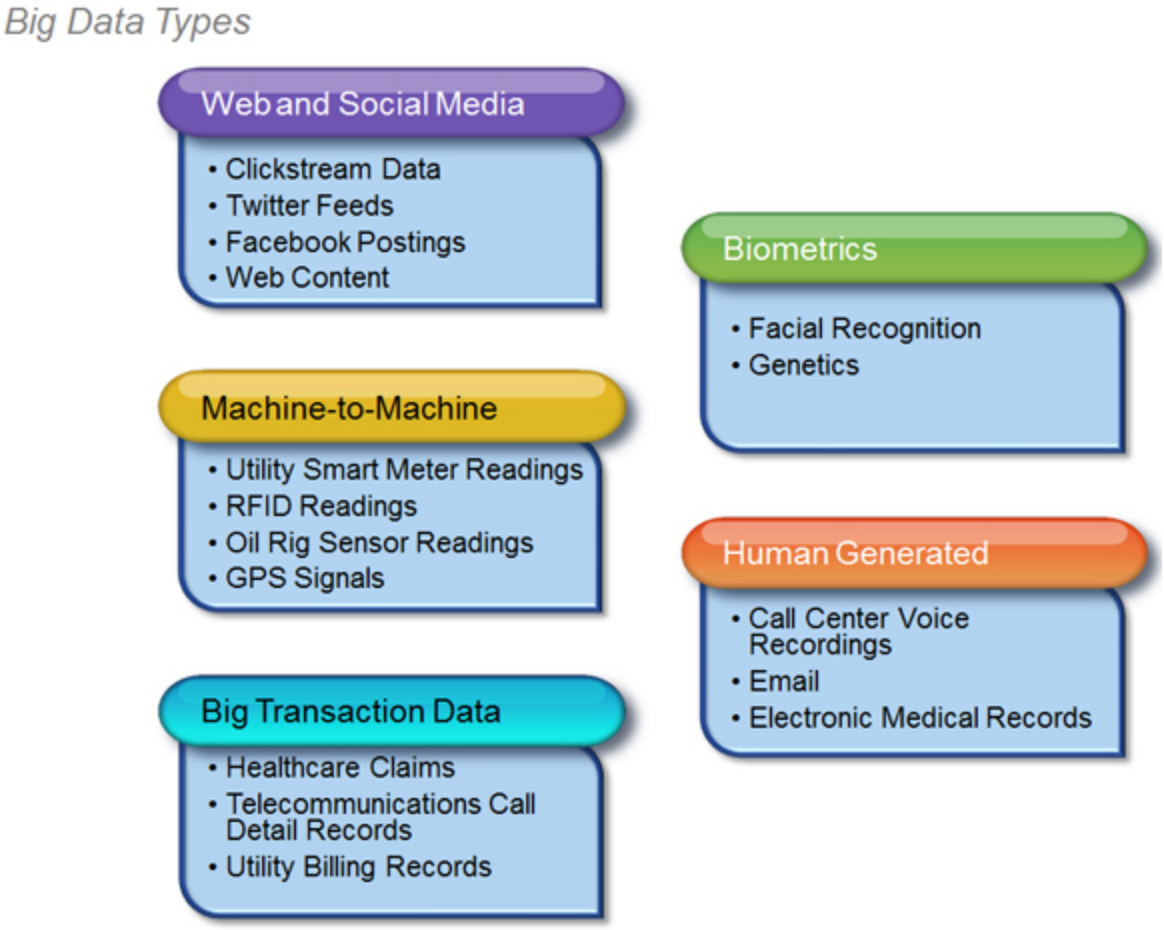
Pero no solamente somos los seres humanos quienes contribuimos
enorme de información, existe también la comunicación denominad
(M2M machine-to-machine) cuyo valor en la creación de grandes ca
también es muy importante. Sensores digitales instalados en conte
la ruta generada durante una entrega de algún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

3. ¿Qué tipos de datos debo explorar?

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué infor
analizar?, sin embargo, el cuestionamiento debería estar enfocado
el que se está tratando de resolver?.[2]

Si bien sabemos que existe una amplia variedad de tipos de datos a
clasificación nos ayudaría a entender mejor su representación, aun
estas categorías puedan extenderse con el avance tecnológico.

Figura 1. Tipos de datos de Big Data[2]



1.- *Web and Social Media*: Incluye contenido web e información que sociales como Facebook, Twitter, LinkedIn, etc, blogs.

2.- *Machine-to-Machine (M2M)*: M2M se refiere a las tecnologías qu otros dispositivos. M2M utiliza dispositivos como sensores o medid evento en particular (velocidad, temperatura, presión, variables met químicas como la salinidad, etc.) los cuales transmiten a través de r inalámbricas o híbridas a otras aplicaciones que traducen estos eve significativa.

3.- *Big Transaction Data*: Incluye registros de facturación, en teleco detallados de las llamadas (CDR), etc. Estos datos transaccionales formatos tanto semiestructurados como no estructurados.

4.- *Biometrics*: Información biométrica en la que se incluye huellas d

retina, reconocimiento facial, genética, etc. En el área de seguridad biométricos han sido información importante para las agencias de ir

5.- *Human Generated*: Las personas generamos diversas cantidade información que guarda un call center al establecer una llamada tel correos electrónicos, documentos electrónicos, estudios médicos, e

4. Componentes de una plataforma Big Data

Las organizaciones han atacado esta problemática desde diferente montañas de información han generado un costo potencial al no de asociado. Desde luego, el ángulo correcto que actualmente tiene el popularidad para analizar enormes cantidades de información es la abierto *Hadoop.*

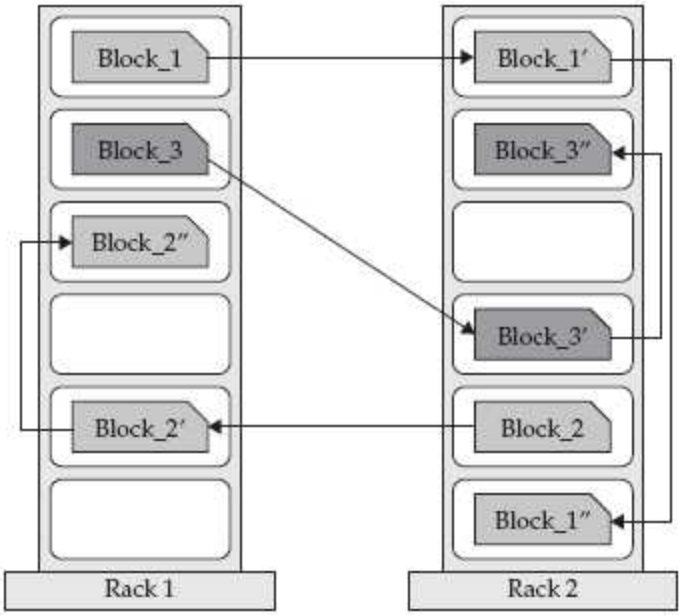
Hadoop está inspirado en el proyecto de Google File System(GFS) programación *MapReduce*, el cual consiste en dividir en dos tareas para manipular los datos distribuidos a nodos de un clúster logrand procesamiento.[5] Hadoop está compuesto de tres piezas: *Hadoop* (HDFS), *Hadoop MapReduce* y *Hadoop Common.*

Hadoop Distributed File System(HDFS)

Los datos en el clúster de Hadoop son divididos en pequeñas pieza distribuidas a través del clúster; de esta manera, las funciones `map` ejecutadas en pequeños subconjuntos y esto provee de la escalabil procesamiento de grandes volúmenes.

La siguiente figura ejemplifica como los bloques de datos son escrit que cada bloque es almacenado tres veces y al menos un bloque s diferente rack para lograr redundancia.

Figura 2. Ejemplo de HDFS



Hadoop MapReduce

MapReduce es el núcleo de Hadoop. El término MapReduce en rea procesos separados que Hadoop ejecuta. El primer proceso `map`, el datos y lo convierte en otro conjunto, donde los elementos individua `tuplas` (pares de llave/valor). El proceso `reduce` obtiene la salida entrada y combina las tuplas en un conjunto más pequeño de las m intermedia es la denominada `Shuffle` la cual obtiene las tuplas del que nodo procesará estos datos dirigiendo la salida a una tarea `red`

La siguiente figura ejemplifica un flujo de datos en un proceso senc

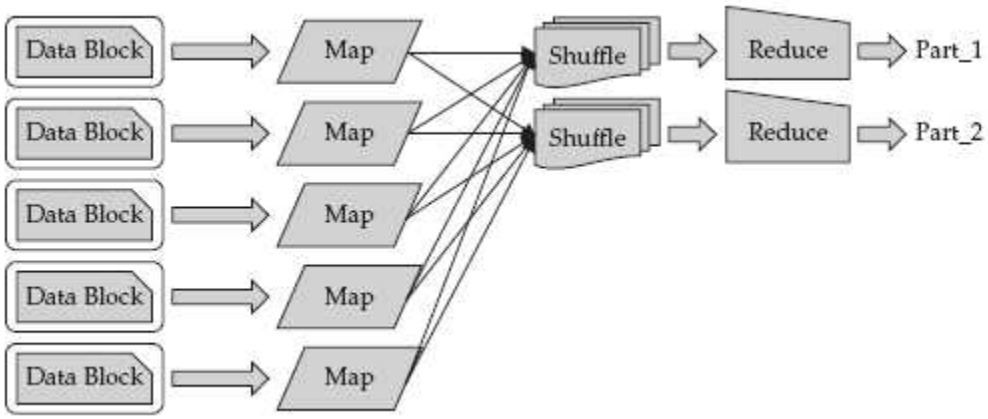


Figura 3. Ejemplo de MapReduce

Hadoop Common

**Hadoop Common Components son un conjunto de lib
varios subproyectos de Hadoop.
Además de estos tres componentes principales de Ha
proyectos relacionados los cuales son definidos a co**

Avro

**Es un proyecto de Apache que provee servicios de se
se guardan datos en un archivo, el esquema que defir
guardado dentro del mismo; de este modo es más ser
aplicación leerlo posteriormente puesto que el esquer
dentro del archivo.**

Cassandra

**Cassandra es una base de datos no relacional distribu
modelo de almacenamiento de <clave-valor>, desarro
Permite grandes volúmenes de datos en forma distrib
de las empresas que utiliza Cassandra dentro de su p**

Chukwa

**Diseñado para la colección y análisis a gran escala de
toolkit para desplegar los resultados del análisis y mo**

Flume

Tal como su nombre lo indica, su tarea principal es di fuente hacia alguna otra localidad, en este caso hacia Hadoop. Existen tres entidades principales: `sources`, `sinks`. **Un** `source` **es básicamente cualquier fuente de destino de una operación en específico y un** `decorato` **dentro del flujo de datos que transforma esa informac manera, como por ejemplo comprimir o descomprimir otra operación en particular sobre los mismos.**

HBase

Es una base de datos columnar (column-oriented data ejecuta en HDFS. HBase no soporta SQL, de hecho, H de datos relacional. Cada tabla contiene filas y colum de datos relacional. HBase permite que muchos atribu llamándolos *familias de columnas,* de tal manera que una familia de columnas son almacenados en un solo distinto a las bases de datos relacionales orientadas a las columnas de una fila dada son almacenadas en co utiliza HBase en su plataforma desde Noviembre del 2

Hive

Es una infraestructura de data warehouse que facilita conjuntos de datos que se encuentran almacenados e distribuido. Hive tiene definido un lenguaje similar a S

**Query Language(HQL), estas sentencias HQL son sep
servicio de Hive y son enviadas a procesos MapRedu
cluster de Hadoop.**
**El siguiente es un ejemplo en HQL para crear una tabl
obtener información de la tabla utilizando Hive:**

```
CREATE TABLE Tweets (from_user STRING, userid BIGINT, tweettext STRING, retweets INT)
COMMENT "This is the Twitter feed table"
STORED AS SEQUENCEFILE;
LOAD DATA INPATH "hdfs://mode/tweetdata" INTO TABLE TWEETS;
SELECT from_user, SUM(retweets)
FROM TWEETS
GROUP BY from_user;
```

Jaql

**Fue donado por IBM a la comunidad de software libre.
Javascript Object Notation (JSON) es un lenguaje fun
que permite la explotación de datos en formato JSON
procesar grandes volúmenes de información. Para ex
Jaql reescribe los queries de alto nivel (cuando es ne
de "bajo nivel" para distribuirlos como procesos Map**

**Internamente el motor de Jaql transforma el query en
`reduce` para reducir el tiempo de desarrollo asociado
en Hadoop. Jaql posee de una infraestructura flexible
analizar datos semiestructurados como XML, archivos
planos, datos relacionales, etc.**

Lucene

Es un proyecto de Apache bastante popular para reali

sobre textos. Lucene provee de librerías para indexac texto. Ha sido principalmente utilizado en la implemer de búsqueda (aunque hay que considerar que no tiene "crawling" ni análisis de documentos HTML ya incorp a nivel de arquitectura de Lucene es simple, básicame (*document*) son divididos en campos de texto (*fields*) y sobre estos campos de texto. La indexación es el com Lucene, lo que le permite realizar búsquedas rápidam independientemente del formato del archivo, ya sean HTML, etc.

Oozie

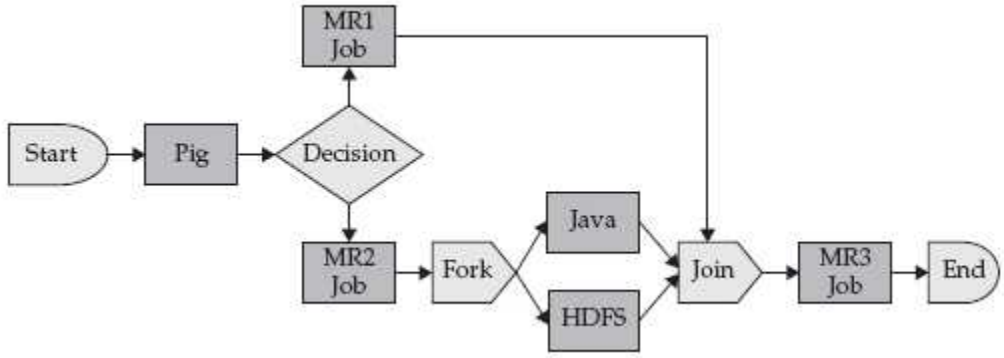
Como pudo haber notado, existen varios procesos qu distintos momentos los cuales necesitan ser orquesta las necesidades de tan complejo análisis de informaci

Oozie es un proyecto de código abierto que simplifica y la coordinación entre cada uno de los procesos. Per pueda definir acciones y las dependencias entre dicha

Un flujo de trabajo en Oozie es definido mediante un g llamado *Directed Acyclical Graph (DAG),* y es acíclico permite ciclos en el grafo; es decir, solo hay un punto salida y todas las tareas y dependencias parten del pu final sin puntos de retorno. Un ejemplo de un flujo de

representa de la siguiente manera:

Figura 4. Flujo de trabajo en Oozie



Pig

Inicialmente desarrollado por Yahoo para permitir a los usuarios de
en analizar todos los conjuntos de datos y dedicar menos tiempo er
MapReduce. Tal como su nombre lo indica al igual que cualquier ce
cosa, el lenguaje *PigLatin* fue diseñado para manejar cualquier tipo
ambiente de ejecución donde estos programas son ejecutados, de
relación entre la máquina virtual de Java (JVM) y una aplicación Ja

ZooKeeper

ZooKeeper es otro proyecto de código abierto de Apache que prove
centralizada y de servicios que pueden ser utilizados por aplicacion
que los procesos a través de un cluster sean serializados o sincron
Internamente en ZooKeeper una aplicación puede crear un archivo
memoria en los servidores ZooKeeper llamado *znode.* Este archivo
actualizado por cualquier nodo en el cluster, y cualquier nodo puede
informado de los cambios ocurridos en ese *znode*; es decir, un serv
configurado para "vigilar" un *znode* en particular. De este modo, las
sincronizar sus procesos a través de un cluster distribuido actualiza
znode, el cual informará al resto del cluster sobre el estatus corresp

en específico.

Como podrá observar, más allá de Hadoop, una plataforma de Big l ecosistema de proyectos que en conjunto permiten simplificar, adm analizar grandes volúmenes de información.

5. Big Data y el campo de investigación

Los científicos e investigadores han analizado datos desde ya hace ahora representa el gran reto es la escala en la que estos son gene

Esta explosión de "grandes datos" está transformando la manera en investigación adquiriendo habilidades en el uso de Big Data para re complejos relacionados con el descubrimiento científico, investigaci biomédica, educación, salud, seguridad nacional, entre otros.

De entre los proyectos que se pueden mencionar donde se ha lleva solución de Big Data se encuentran:

El *Language, Interaction and Computation Laboratory (CLIC)* en c de Trento en Italia, son un grupo de investigadores cuyo interés es comunicación verbal y no verbal tanto con métodos computaciona

Lineberger Comprehensive Cancer Center - Bioinformatics Group para analizar datos producidos por los investigadores de *The Can* para soportar las investigaciones relacionadas con el cáncer.

El PSG College of Technology, India, analiza múltiples secuencias determinar los enlaces evolutivos y predecir estructuras molecular algoritmo y el paralelismo computacional de Hadoop mejora la vel estas secuencias.

La *Universidad Distrital Francisco Jose de Caldas* utiliza Hadoop p

de investigación relacionado con el sistema de inteligencia territori

La *Universidad de Maryland* es una de las seis universidades que académica de cómputo en la nube de IBM/Google. Sus investigaci en la lingüistica computacional (machine translation), modelado de análisis de correo electrónico y procesamiento de imágenes.

Para más referencias en el uso de Hadoop puede dirigirse a : http://wiki.apache.org/hadoop/PoweredBy

El *Instituto de Tecnología de la Universidad de Ontario (UOIT)* junto Toronto utilizan una plataforma de big data para análisis en tiempo *InfoSphere Streams*), la cual permite monitorear bebés prematuros neonatología para determinar cualquier cambio en la presión arteria alteraciones en los registros del electrocardiograma y electroencefa detectar hasta 24 horas antes aquellas condiciones que puedan se de los recién nacidos.

Los laboratorios *Pacific Northwest National Labs(PNNL)* utilizan de InfoSphere Streams para analizar eventos de medidores de su red verificar aquellas excepciones o fallas en los componentes de la re casi de manera inmediata a los consumidores sobre el problema pa administrar su consumo de energía eléctrica.[3]

La esclerosis múltiple es una enfermedad del sistema nervioso que médula espinal. La comunidad de investigación biomédica y la *Univ Nueva York (SUNY)* están aplicando análisis con big data para cont la investigación, diagnóstico, tratamiento, y quizás hasta la posible múltiple.[4]

Con la capacidad de generar toda esta información valiosa de difer

empresas y los gobiernos están lidiando con el problema de analiza
propósitos importantes: ser capaces de detectar y responder a los a
de una manera oportuna, y para poder utilizar las predicciones del a
situación requiere del análisis tanto de datos en movimiento (datos
en reposo (datos históricos), que son representados a diferentes y
variedades y velocidades.

6. Conclusiones

La naturaleza de la información hoy es diferente a la información er
abundacia de sensores, micrófonos, cámaras, escáneres médicos,
nuestras vidas, los datos generados a partir de estos elementos se
segmento más grande de toda la información disponible.
El uso de Big Data ha ayudado a los investigadores a descubrir cos
tomado años en descubrir por si mismos sin el uso de estas herram
velocidad del análisis, es posible que el analista de datos pueda ca
basándose en el resultado obtenido y retrabajar el procedimiento u
encontrar el verdadero valor al que se está tratando de llegar.

Como se pudo notar en el presente artículo, implementar una soluc
implica de la integración de diversos componentes y proyectos que
ecosistema necesario para analizar grandes cantidades de datos.
Sin una plataforma de Big Data se necesitaría que desarrollar adici
permita administrar cada uno de esos componentes como por ejem
conectividad, alta disponibilidad, seguridad, optimización y desemp
monitoreo, administración de las aplicaciones, SQL y scripts persor
IBM cuenta con una plataforma de Big Data basada en dos product
InfoSphere BigInsights™ e IBM InfoSphere Streams™, además de
Vivisimo, los cuales están diseñados para resolver este tipo de prob

herramientas están construidas para ser ejecutadas en sistemas dis
diseñados para tratar con grandes volúmenes de información, anali
estructurados como no estructurados.

Dentro de la plataforma de IBM existen más de 100 aplicaciones de
trabajo que se ha realizado internamente en la empresa para casos
específicas. Estos aplicativos están implementados dentro de la sol
organizaciones puedan dedicar su tiempo a analizar y no a impleme

7. Referencias

1. Cisco, **Internet será cuatro veces más grande en 2016**, Artículo
   http://www.cisco.com/web/ES/about/press/2012/2012-05-30-interr
   mas-grande-en-2016--informe-vini-de-cisco.html
2. Soares Sunil, **Not Your Type? Big Data Matchmaker On Five D**
   Explore Today, Artículo Web http://www.dataversity.net/not-your-
   data-matchmaker-on-five-data-types-you-need-to-explore-today/
3. Clegg Dai, **Big Data: The Data Velocity Discussion**, Artículo We
   http://thinking.netezza.com/blog/big-data-data-velocity-discussion
4. Kobielus James, **Big Data Analytics Helps Researchers Drill D**
   Sclerosis, Artículo Web http://thinking.netezza.com/blog/big-data
   researchers-drill-deeper-multiple-sclerosis
5. Aprenda más acerca de Apache Hadoop en http://hadoop.apache
6. Zikopolous Paul, Deroos Dirk, Deutsch Tom, Lapis George, **Unde**
   Analytics for Enterprise Class Hadoop and Streaming Data, M
7. Foster Kevin, Nathan Senthil, Rajan Deepak, Ballard Chuck, **IBM**
   Assembling Continuous Insight in the Information Revolutior

Bluemix

Guías rápidas de inicio y demos de la plataforma abierta en la Nube de IBM.

Súbete a la Nube de IBM

Abre el potencial del cómputo en la nube con productos y servicios de IBM.

Arquitectura de Nube Abierta de IBM

Al cambiar cómo se manejan los negocios y la sociedad, la computación en nube está abriendo gigantescas avenidas de innovaciones.