UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Revista Científica

http://revistas.udistrital.edu.co/ojs/index.php/revcie/index

CIDC

Artículo de investigación

Arquitectura tecnológica para Big Data

Architecture technology for Big Data

Tecnologia da Arquitetura para Data Big

Juan José Camargo Vega[1]
Jonathan Felipe Camargo Ortega[2]
Luis Joyanes Aguilar[3]

Fecha de recepción: junio 2014 **Fecha de aceptación:** noviembre 2014

Para citar este artículo: Camargo, J. J. Camargo, J. F. y Aguilar, L. (2015). Arquitectura tecnológica para Big Data. *Revista Científica*, 21, 7-18. **Doi:** 10.14483/udistrital.jour.RC.2015.21.a1

Resumen

El termino Big Data cada día que pasa, se torna más importante, es por esto que en la presente investigación se estudia, analiza y da a conocer de manera exhaustiva las diferentes arquitecturas de Big Data, con sus características, herramientas, tecnologías y estándares relacionadas con dicho termino, con el único objetivo de brindar una ayuda al sector empresarial para una posible implementación de Big Data.
Palabras Clave: Big Data, Hadoop, NoSQL, calidad de datos, modelo de datos, arquitectura.

Abstract

The term Big Data with each passing day, it becomes more important, which is why in this research is studied, analyzed and disclosed in a comprehensive manner the different architectures of Big Data, with its features, tools, technologies and standards related to that end, for the sole purpose of providing assistance to the business sector for possible implementation of Big Data.
Keywords: Big Data, Hadoop, NoSQL, data quality, data model, architecture.

Resumo

O termo Big Data cada dia se torna mais importante, que é por isso que nesta pesquisa são estudados, analisados e divulgados de uma forma abrangente as diferentes arquiteturas de Big Data, com seus recursos, ferramentas, tecnologias e padrões, com o objectivo para prestar assistência ao sector empresarial para possível implementação.
Palavras-chave: Big Data, o Hadoop, NoSQL, qualidade de dados, modelo de dados, arquitetura.

[1] Universidad Pontificia de Salamanca campus Madrid–España. Contacto: jjcamargovega@uptc.edu.co
[2] Universidad del Bosque, Bogotá- Colombia. Contacto: jfcamargo@unbosque.edu.co
[3] Universidad Pontificia de Salamanca campus Madrid–España. Contacto: luis.joyanes@upsam.es

Introducción

El concepto de Big Data apenas se empieza a conocer en el medio organizacional, por lo que se ignoran aspectos relevantes del tema. Las empresas desconocen qué hacer con el gran volumen de datos que generan y les llega de otras. Por lo anterior, el presente trabajo estudia las arquitecturas posibles de Big Data. Un problema de Big Data es la forma como crece cada día en volumen, velocidad y variedad, debido al desarrollo algorítmico de las tecnologías de información.

Uno de los objetivos de la presente investigación es analizar las diferentes arquitecturas de Big Data, como características, beneficios y desventajas, además de estudiar herramientas, tecnologías, modelos y estándares. El resultado de la presente investigación puede servir a las empresas, independientemente de su misión y tamaño, que desconozcan el uso de Big Data.

Definición De Big Data

Según Zikopoulos, Eaton, DeRoos, Deutsch y Lapis (2012), Big Data refiere a la información que no puede ser procesada o analizada mediante procesos tradicionales.

Para zdnet.com (2010), Big Data son "cantidades masivas de datos que se acumulan con el tiempo que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos".

Big Data también alude "al tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales" (Dans, 2001), así como a "las herramientas, procesos y procedimientos que permitan a una organización crear, manipular y administrar grandes conjuntos de datos e instalaciones de almacenamiento", en palabras del analista Dan Kusnetzky, del Grupo Kusnetzky (Preimesberger, 2011).

En [5] "Forrester define Big Data como las técnicas y tecnologías que hacen que sea económico para hacer frente a los datos a una escala extrema. Big Data trata de tres cosas: 1. Las técnicas y la tecnología, lo que significa que la empresa necesita gente que sabe qué hacer con los datos para obtener valor. 2. Escala extrema de datos que supera a la tecnología actual debido a su volumen, velocidad, y variedad. 3. El valor económico, haciendo que las soluciones sean asequibles y ayuden a la inversión de los negocios".

Otra definición señala que "Big Data es la frontera de la capacidad de una empresa para almacenar, procesar y acceder (ZEPA) todos los datos que necesita para operar con eficacia, tomar decisiones, reducir los riesgos y atender a los clientes" (Gualtieri, 2012).

Según [7], Big Data "se refiere a las herramientas, los procesos y procedimientos que permitan a una organización crear, manipular y gestionar conjuntos de datos muy grandes y las instalaciones de almacenamiento".

"Big Data es un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable" (Wikipedia, 2013).

Gartner define el Big Data como: "un gran volumen, velocidad o variedad de información que demanda formas costeables e innovadoras de procesamiento de información que permitan ideas extendidas, toma de decisiones y automatización del proceso" (searchdatacenter.techtarget.com, 2013).

Por último, para Krishnan (2013), Big Data son grandes volúmenes de datos disponibles con desiguales grados de complejidad, con diferentes velocidades y con gran ambigüedad. Datos que no se pueden procesar con tecnologías tradicionales, además, son generados desde redes sociales, redes de sensores, dispositivos de rayos X, centrales nucleares, aviones, ventas, marketing, compras y finanzas personales.

Arquitectura De Big Data

Existen varios modelos de arquitectura de Big Data, de los cuales se mostrarán tres:

Arquitectura de procesamiento de Big Data propuesta por Krishnan

La arquitectura presentada por Krishnan (2010), representada en la Figura 1, es similar al tratamiento de la gestión de datos que hace bastante tiempo se conoce, el cual consiste en cuatro etapas: recolección o recopilación, carga, transformación y extracción de datos. Para este caso, el autor lo llama enfoque de procesamiento de Big Data (Demchenko, Ngo y Membrey, 2012).

Etapa 1. Recolección o recopilación o fuentes de datos de Big Data

En la primera etapa, los datos son recibidos de diferentes orígenes o fuentes, que pueden ser: páginas web, redes sociales, máquina a máquina (M2M), transacciones, biometría o generados por el ser humano (Krishnan, 2013).

Páginas web. Puede registrar las huellas en páginas web, con el seguimiento de clics que realice el usuario (Soares, 2012a; Mysore, Khupat y Jain, 2013).

- Redes sociales. De medios o redes como Facebook, Twitter, LinkedIn y blogs. Por ejemplo, como menciona Soares (2012a):
- Muchas compañías de seguros ahora utilizan los medios sociales para investigar las denuncias. Sin embargo, la mayoría de los reguladores todavía no permiten a las aseguradoras usar los medios sociales para establecer las tasas de política durante el proceso de suscripción. Por ejemplo, si una aseguradora de vida que ve el perfil de Facebook de un solicitante indica que ella es una estudiante de aviación, la aseguradora no puede usar ese conocimiento para aumentar sus primas debido a que podría ser considerado un alto riesgo.
- M2M. Son datos que se generan de máquina a máquina (*hardware* a *hardware*), tales como: lecturas de sensores, medidores y otros dispositivos. Por ejemplo, como lo menciona Soares (2012a):

  Varias utilidades están desplegando contadores inteligentes para medir el consumo de agua, gas y electricidad a intervalos regulares de una hora o menos. Estos contadores inteligentes generan grandes cantidades de datos de intervalo que necesitan ser gobernados de manera apropiada. Las utilidades deben salvaguardar la privacidad de estos datos de intervalo, ya que potencialmente puede revelar las actividades del hogar del abonado, así como cuándo un dueño de casa puede estar lejos.
- Transacciones. Son datos que pueden provenir de registros detallados de llamadas (*charging data record*, CDR) de las telecomunicaciones, registros de facturación de servicios públicos que están cada vez más disponibles en formatos semiestructurados y no estructurados. Por ejemplo, como lo menciona Soares (2012a):
- Un gran plan de salud procesa más de 500 millones de reclamaciones por año, con cada registro de las reclamaciones que consta de 600 a 1.000 atributos. El plan utiliza el análisis predictivo para determinar si se requieren ciertas medidas proactivas para un pequeño subconjunto de los miembros. Sin embargo, el equipo de inteligencia de negocios encontró que los médicos estaban usando los códigos de procedimiento inconsistentes para presentar reclamaciones, lo que limitó la efectividad de los análisis predictivos. El equipo de inteligencia de negocios también cuestionó el texto dentro de documentos de reclamos.

  Biometría. Este tipo de datos incluye huellas digitales, la genética, la escritura a mano, escáner de retina, entre otros. Por ejemplo, como menciona Soares (2012a):
- Un informe de marzo de 2012 de la Comisión Federal de Comercio de EE. UU. detalla cómo los minoristas podrían potencialmente utilizar la tecnología de reconocimiento facial en combinación con una foto de los medios sociales para hacer ofertas personalizadas a los clientes en función de su

comportamiento de compra y la ubicación. Dado que esta información podría tener un gran impacto en los programas de fidelización de los minoristas, sino que también tendría consecuencias graves de la privacidad. Los minoristas tendrían que hacer las revelaciones de privacidad adecuadas antes de implementar estas aplicaciones.

- Generados por el ser humano. Son datos producto de grabaciones de voz, correo electrónico, documentos en papel, las encuestas y registros electrónicos. Por ejemplo, como lo menciona Soares (2012a):
- El departamento de análisis en un hospital construyó un modelo predictivo basado en 150 variables y 20.000 encuentros con el paciente para determinar la probabilidad de que un paciente sea readmitido dentro de los 30 días de tratamiento para la insuficiencia cardíaca congestiva. En un ejemplo de la eficacia del modelo de predicción, el equipo de análisis identifica la condición de fumador del paciente como una variable crítica. Al principio, sólo el 25 por ciento de los datos estructurados alrededor de la condición de fumador se rellena con binarios respuestas sí/no. Sin embargo, el equipo de análisis se incrementó la tasa de la población para el consumo de tabaco y el 85 por ciento de los encuentros mediante el uso de análisis de contenidos en base a los registros médicos electrónicos que contengan notas del médico, resúmenes de alta, y el paciente exámenes físicos-que permite al equipo de análisis para mejorar la calidad de la estructura baja densidad de población de datos mediante el uso de fuentes de datos no estructurados.

La anterior clasificación de los datos puede estar en formatos estructurados, no estructurados o semiestructurados. La velocidad y volumen de los datos depende del origen o fuente de los datos [Soares, 2012a, 2012b; Joyanes, 2013; Sun y Heller, 2012).

También se consideran fuente de datos: "registros, trayectorias de navegación, datos de redes sociales, transferencias de noticias, emails, salidas

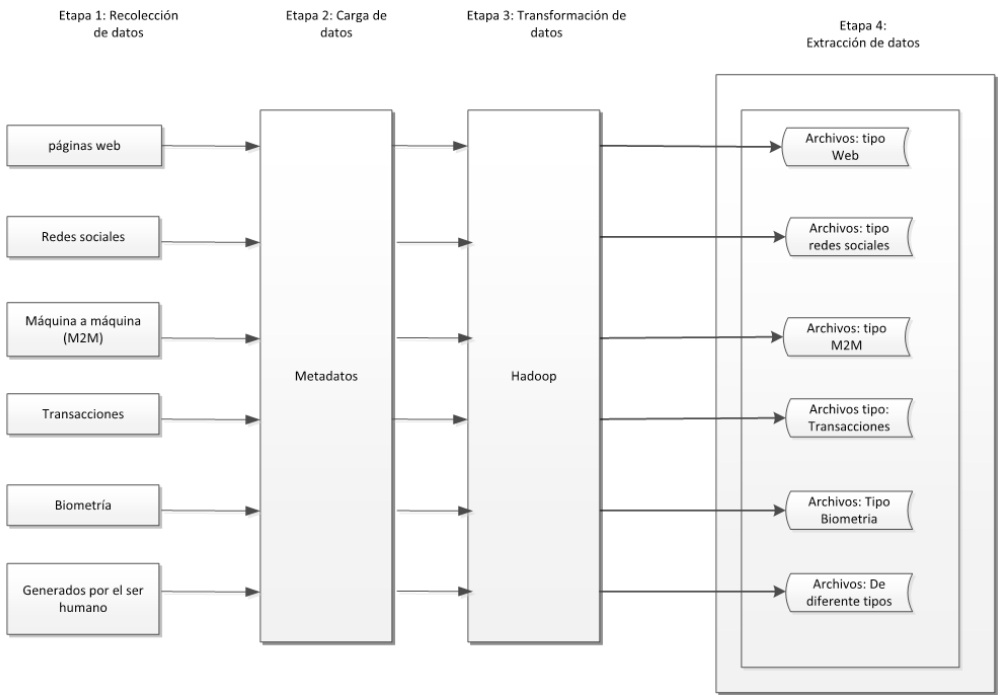

Figura 1. Arquitectura de Big Data propuesta por Krishnan

Fuente: Krishnan (2013)

de sensores electrónicos e incluso algunos datos transaccionales" (Tucker, 2013).

Etapa 2. Carga de datos de Big Data

En esta etapa, los datos se cargan aplicando el concepto de metadatos (datos que describen otros datos). Además de la carga como tal, es la primera vez que los datos se estructuran (Krishnan, 2013) . Es de aclarar que los metadatos son información que describe características de cualquier dato, como el nombre, la ubicación, la importancia percibida, la calidad y sus relaciones con otros objetos de datos que la empresa considere digno de la gestión (Soares, 2012b). La siguiente tarea es la capacidad de usar metadatos, bibliotecas semánticas y datos maestros. Se busca vincular los datos entre el conjunto de datos estructurados y no estructurados con metadatos y datos maestros. Se deben transformar los datos no estructurados en estructurados. Es importante acudir a la integridad referencial, la cual ayuda inicialmente con la clave principal y las demás relaciones en una base de datos tradicional (Soares, 2012b).

Etapa 3. Transformación de datos de Big Data

En este punto, los datos se transforman mediante la aplicación de las reglas del negocio y el procesamiento de los datos. Respecto al procesamiento, en cada etapa producen resultados intermedios que se pueden almacenar para un posterior examen. El resultado de esta etapa son unas cuantas claves de metadatos con modelo clave-valor (Krishnan, 2013).

Etapa 4. Extracción de datos de Big Data

El objetivo de la extracción es obtener datos para su posterior análisis, generar informes operativos y su posible visualización y, por último, y no el más importante, para su almacenamiento (Krishnan, 2013) .

Arquitectura de Big Data propuesta por Bob Marcus

El modelo de arquitectura por niveles fue propuesto por Bob Marcus (Tabla 1).

Tabla 1. Arquitectura Big Data propuesta por Marcus

F. Aplicaciones e interfaces de usuario	
E. Análisis e interfaces de bases de datos	
D. Bases de datos operacionales y de analítica	G. Servicios de Apoyo
C. Fundación altamente escalable	
B. Secuencia y procesamiento ETL	
A. Fuentes de datos externos	

Fuente: [18]

Fuentes de datos externas

Como se aprecia en la Tabla 1, el componente A, es parte de la arquitectura de datos que suministra las entradas de datos externas y la producción de los componentes internos de Big Data [18].

Secuencia y procesamiento ETL

Las tareas que se desarrolla en el componente B son filtrar y transformar los flujos de datos provenientes de los recursos externos. El procesamiento de datos que se realiza es el llamado "en movimiento" entre los almacenes de datos [18].

Fundación altamente escalable

Según la Tabla 1, y respecto al componente C, existen tres tipos de escalonamiento:

- El primero, a nivel de la infraestructura, existe con el fin de poder atender el almacenamiento y procesamiento de grandes volúmenes de datos.
- El segundo se refiere a los almacenes de datos. Tal como lo menciona Marcus, "es la esencia de la arquitectura Big Data", la cual sucede en forma de "escalabilidad horizontal [que] usando componentes menos caros puede apoyar el crecimiento ilimitado de almacenamiento de datos". Sin embargo debe haber capacidades de tolerancia a fallas disponibles para manejar los fallos de sus componentes.
- Por último, el autor propone escalonar el procesamiento buscando "aprovechar las ventajas de los almacenes de datos distribuidos escalables, [por lo cual] es necesario contar con el procesamiento distribuido en paralelo escalable con tolerancia a fallos similares".

Bases de datos operacionales y de Analíticas

En el componente D se proponen tres clases de bases de datos:

- Bases de datos analíticas. El análisis de bases de datos toma los datos procesados y escalonados de la sección anterior. Son bases de datos altamente optimizadas para sola lectura (por ejemplo, columnas de almacenamiento, amplia indexación y desnormalización). A menudo es aceptable para las respuestas de base de datos por tener una latencia alta (por ejemplo, invocar el procesamiento por lotes escalable sobre grandes conjuntos de datos).
- Bases de datos operacionales. Estas bases de datos mantienen una excelente operación en lectura y escritura en general de forma eficiente. Por ejemplo, las bases de datos NoSQL, son de uso frecuente en las arquitecturas de datos grandes en esta capacidad. Los datos pueden ser posteriormente transformados y cargados en las bases de datos analíticas para soportar aplicaciones analíticas.
- En la memoria de datos Grids. Se refieren a los datos ubicados en memorias cachés, que buscan minimizar escribir en disco los datos. Se pueden utilizar para aplicaciones en tiempo real a gran escala que requieren acceso transparente a los datos.

Analítica e interfaces de bases de datos

El componente E consta de tres partes:

- Análisis de interfaces de procesos en lotes. Se refiere al tipo de interfaz usada para el procesamiento de datos que provienen en lotes o Batch (p. ej. Map-Reduce). También hace referencia a la interfaz de usuario para acceder a los datos en almacenes de datos escalables (por ejemplo del sistema de archivos Hadoop).
- Análisis de interfaces interactivas. "Los almacenes de datos pueden ser bases de datos escalables horizontalmente sintonizados para las respuestas interactivas (p. ej. HBase) o lenguajes de consulta en sintonía con los modelos de datos".
- Análisis de interfaces en tiempo real. Se deben analizar con cuidado las interfaces que atienden o son parte de un sistema de tiempo real, pues los eventos son complejos tanto en el procesamiento como almacenamiento de los datos.

Aplicaciones e interfaz de usuario

El componente F se refiere a las aplicaciones e interfaces de usuario, las cuales no deben ser algoritmos complejos, al usar grandes cantidades de datos distribuidos.

Servicios de apoyo

Es el componente G. Estos servicios hacen referencia a los componentes necesarios para la implementación y gestión de sistemas robustos de Big Data, los cuales se pueden discriminar as:

- Diseñar, desarrollar e implementar herramientas. Consiste en tener a la mano herramientas bien desarrolladas, es decir, de alto nivel de calidad, de manera que sirvan para implementar soluciones Big Data.
- Seguridad. Aspecto importante a nivel de controles de seguridad de grandes volúmenes de datos, pues en la actualidad son escasos o limitados. Se busca ampliar la seguridad en el Big Data.
- Gestión de procesos. "Los distribuidores comerciales son el suministro de herramientas de gestión de procesos para aumentar las implementaciones de código abierto".
- Gestión de recursos de datos. En [18] se hace hincapié en "Herramientas de control de datos de código abierto que son todavía inmaduras. Estos serán aumentados en un futuro por los proveedores comerciales".

  – Administración del sistema. El autor se refiere a que las "herramientas de gestión de sistemas de código abierto también son inmaduras (p. ej. Ambari). Afortunadamente sólidas herramientas de administración del sistema están disponibles en el mercado para la infraestructura escalable (por ejemplo, basado en la nube)".

Arquitectura de Big Data propuesta por Microsoft

El modelo consta de cuatro componentes (figura 2): fuentes de datos (*Data Sources*), transformación de datos (*Data Transformation*), infraestructura de datos (*Data Infrastructure*) y uso de datos (*Data Usage*) [18].

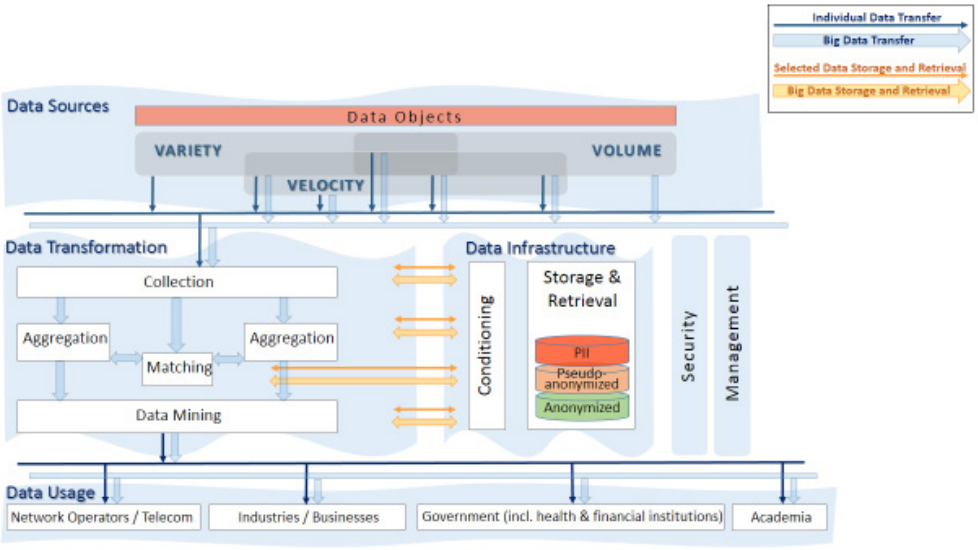

Figura 2. Arquitectura Big Data propuesta por Microsoft

Fuente: [18]

- Fuentes de datos.
  Estos datos presentan tres características que definen el Big Data: volumen, velocidad y variedad. Como aspecto relevante, este tipo de datos son independientes de su contenido o del contexto. "Por lo general, los datos de 'big data' se recogen para un propósito específico". Además, "una vez que se recogen los datos, se pueden volver a utilizar para una variedad de propósitos, algunos potencialmente desconocidos en el tiempo de recogida" [18].

Sobre las características de volumen, según Pettey y Goasduff (2011):

> el aumento de los volúmenes de datos dentro de sistemas de la empresa es causado por el volumen de transacciones y otros tipos de datos tradicionales, así como por los nuevos tipos de datos. El exceso de volumen es un problema de almacenamiento, pero el exceso de datos también es un tema de análisis masivo (Pettey y Goasduff, 2011).

En cuanto a la variedad,

> los líderes de TI siempre han tenido un problema: traducir grandes volúmenes de información transaccional en las decisiones —ahora hay más tipos de información para analizar— provenientes principalmente de los medios sociales y móviles (sensible al contexto). Variedad incluye datos tabulares (bases de datos), datos jerárquicos, documentos, correo electrónico, datos de medición, video, imágenes fijas, audio, datos *ticker*, transacciones financieras y más.

Sobre la última característica, la velocidad,

Se trata de flujos de datos, la creación de registros estructurados y disponibilidad para el acceso y la entrega. Velocidad significa tanto cómo se está produciendo de datos rápido y qué tan rápido los datos se tratarán de satisfacer la demanda (Pettey y Goasduff, 2011).

- Transformación de datos
  Consta de cuatro subetapas. Cada una de ellas puede tener su etapa de preprocesamiento específico, creación de metadatos, utilizar diferentes infraestructuras de datos especializados de acuerdo con sus necesidades, tener su propia privacidad y, por último, sus propias políticas [18].

- Colección
  En la transformación de datos aparece un proceso llamado colección, donde se recogen datos en diferentes tipos y formas, se pueden allegar datos de fuentes y estructuras similares o iguales, o combinadas. Además, se crean metadatos para que sea más fácil hacer una búsqueda de los datos [18].

- Agregación
  Consiste en adicionar datos a una colección más grande, cuando uno o varios metadatos tengan claves iguales. "Como resultado, la información acerca de cada objeto se enriquece o el número de objetos en la colección crece" [18].

- Congruencia
  En esta etapa se recogen datos con metadatos sin importar si son diferentes y se unen a una colección

Tabla 2. Comparación arquitectura de: Krishnan y Marcus

Krishnan	Marcus
Etapa 1. Recolección de datos	(A) Fuentes de datos externas
Etapa 2. Carga de datos	(B) Secuencia y procesamiento ETL
Etapa 3. Transformación de datos	(B) Secuencia y procesamiento ETL
Etapa 4. Extracción de datos	(C) Fundación altamente escalables (D) Bases de datos operacionales y analíticas

Tabla 3. Comparación arquitectura de Krishnan y Microsoft

Krishnan	Microsoft
Etapa 1. Recolección de datos	Fuentes de datos
Etapa 2. Carga de datos	Transformación de datos
Etapa 3. Transformación de datos	Transformación de datos
Etapa 4. Extracción de datos	Infraestructura de Big Data

Tabla 4: Comparación arquitectura de Marcus y Microsoft

Bob Marcus	Microsoft
(A) Fuentes de datos externas	Fuentes de datos
(B) Secuencia y procesamiento ETL	Transformación de datos
(C) Fundación altamente escalables (D) Bases de datos operacionales y analíticas	Infraestructura de Big Data
(E) Analítica e interfaces de bases de datos (F) Aplicaciones e interfaz de usuario	Uso de los datos

más grande. Al final se obtiene que cada objeto sea enriquecido.

- Minería de datos
  La minería o *data mining* es una extracción de datos para luego poder hallar relaciones entre ellos. Existen "dos tipos de minería de datos: descriptivo, que proporcione información sobre los datos existentes, y predictivo, lo que hace que los pronósticos basados en los datos" [18].

- Infraestructura de Big Data
  Se considera la infraestructura como "un paquete de almacenamiento de datos o software de base de datos, servidores, almacenamiento y redes utilizados en apoyo de las funciones de transformación de datos y de almacenamiento de datos según sea necesario" [18] .

- Uso de los datos
  Respecto al uso de los datos, depende del usuario y sus necesidades particulares, pero estos se pueden presentar en diferentes formatos y bajo ciertas consideraciones de seguridad [18].

Comparación de arquitecturas propuestas por Krishnan y Marcus

Para empezar, la arquitectura que presenta Krishnan muestra cuatro etapas, mientras que la de Marcus muestra siete niveles, pero, para la presente comparación se tendrá en cuenta tan solo cinco de ellas (Tabla 2).

La primera etapa en el modelo de Krishnan Krish es llamado "recolección de datos"; en Bob Marcus se tiene como "fuentes de datos externas", pero la tarea a desarrollar es muy similar, por no decir, que igual.

Las actividades propuestas por Krishnan en la etapa "carga de datos" y "transformación de datos" son cubiertas de forma similar en la arquitectura de Bob Marcus, en algo llamado "secuencia y procesamiento ETL".

Por último, en la arquitectura de Krishnan "Etapa 4: Extracción de datos" es cubierta en las fases "(C) Fundación altamente escalables" y "(D) Bases de datos operacionales y Analíticas", del modelo propuesto de Marcus.

Comparación de arquitecturas propuestas por Krishnan y Microsoft

La arquitectura que presenta Krishnan muestra cuatro etapas, mientras que Microsoft muestra cuatro componentes (Tabla 3).

Las tareas desarrolladas en la etapa uno son las mismas que se hacen en "fuente de datos" de la arquitectura presentada por Microsoft. En el modelo de Krishnan, las actividades de la etapa dos y tres son muy similares a las actividades desarrolladas en la etapa llamada "transformación de datos", de la arquitectura de Microsoft. Mientras que la extracción de datos (Krishnan), en Microsoft es llamado "infraestructura de Big Data", las dos con actividades similares.

Comparación de arquitecturas propuestas por Marcus y Microsoft

La arquitectura que presenta Marcus tiene siete niveles y la Microsoft, cuatro componentes (Tabla 4).

En la comparación de las arquitecturas presentadas por Marcus y Microsoft, se encuentra que las funciones son similares en cada etapa o paso, pero, en cada modelo con nombres diferentes.

Conclusiones

Después de revisar las tres propuestas de arquitectura de Big Data, se pueden concluir varios aspectos:

Tomar lo mejor de cada modelo y presentar una propuesta de arquitectura, con el fin de buscar una futura implementación con el menor esfuerzo por parte de las personas encargadas de dicha labor.

Del modelo de Krishnan se tomaría en su orden la etapa uno "Recolección de datos", pues muestra de manera explícita y con claridad los sitios de dónde y cómo pueden provenir los datos del Big Data, al que se agregaría el modelo de Marcus, que tiene en cuenta la recolección de los datos que se producen en las organizaciones para el Big Data.

La segunda fase se puede tomar igualmente de la propuesta de Krishnan, "Carga de datos", porque, además de cumplir el cargar los datos, también realiza la labor de estructurar los datos, es decir, hace uso de Hadoop, combinando con bibliotecas semánticas, y como aspecto relevante acude a la integridad referencial, la cual ayuda inicialmente con la clave principal y las demás relaciones en una base de datos tradicional.

La tercera etapa se le puede llamar la de transformación de los datos; es recomendable acudir al modelo de arquitectura presentado por Microsoft, pues contiene cuatro subtareas por desarrollar: primera, colección, que recoge datos en diferentes tipos y formas, lo más importante de esta labor es que crea metadatos, lo cual ayuda de manera ágil a la búsqueda de datos. La segunda es la agregación, que permite adicionar datos a la colección ya existente del paso anterior, siempre y cuando uno o varios metadatos presenten claves iguales. La siguiente subtarea es la congruencia, que permite unir datos con metadatos, sin importar si son diferentes y se unen a una colección más grande. Y, para cerrar con broche de oro, la última subtarea es la minería de datos o *data mining*, que después de extraer datos permite hallar relaciones entre los datos.

La cuarta o ultima labor por realizar en el Big Data es la extracción de datos, para lo que se puede acudir al modelo de arquitectura de Krishnan, pues es allí donde, después de extraer los datos, se tienen listos para su análisis, también para generar informes operativos, su posible visualización y, por último, el almacenamiento.

De otra parte, en el modelo de Marcus aparece una etapa "(G) Servicios de apoyo", de la cual es relevante tener en cuenta el aspecto "seguridad", que no es contemplado en las otras arquitecturas estudiadas en el presente artículo. Como se sabe, los datos deben tener seguridad, en fin, lo que se debe buscar es tener un buen nivel de controles de seguridad en Big Data.

También un aspecto del modelo de Marcus, "Diseñar, desarrollar e implementar herramientas",

en la etapa "(G) Servicios de Apoyo", no debe ser parte de la labor de la persona que implementa un Big Data, pues hoy existen herramientas que ayudan con esta gestión. Es decir, no es necesario seguir en este aspecto el modelo de Marcus.

Finalmente, el modelo de arquitectura de Marcus, en el paso llamado "(G) Servicios de Apoyo", muestra algo denominado "Gestión de recursos de datos", donde menciona las "Herramientas de control de datos de código abierto que son todavía inmaduras"; este no se debe tener presente para abordar proyectos de Big Data, pues realmente lo que se necesita son herramientas para el Big Data que muestren altos niveles de calidad.

Referencias

Dans, E. (2011). *Big Data: una pequeña introducción*. [En línea]. Recuperdao de http://www.enriquedans.com/2011/10/big-data-una-pequena-introduccion.html [Consultado el 28 de abril de 2014].

Demchenko, Y., Ngo, C., y Membrey, P. (2012). *Architecture Framework and Components for the Big Data Ecosystem*. *Draft Version 0.2*. [En línea]. Recuperado de http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

es.wikipedia.org, [En línea] Disponible en: http://es.wikipedia.org/wiki/Big_data,(2013). [Consultado: 25-Mar-13].

Gualtieri, M. (2012). *The Pragmatic Definition of Big Data*. [En línea]. Recuperado de http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data [Consultado el 4 de enero de 2013].

Hopkins, B. (2011). *Beyond the Hype of Big Data*. [En línea]. Recuperado de http://www.cio.com/article/692724/Beyond_the_Hype_of_Big_Data [Consultado el 1 de mayo de 2012].

Joyanes, L. (2013). *Big Data: análisis de grandes volúmenes de datos en organizaciones* (1ª edición). México: Alfaomega.

Krishnan, K. (2013). *Data warehousing in the age of big data*. Morgan Kaufmann Series on Business Inteligence. USA: Elsevier.

Mysore, D., Khupat, S., y Jain S. (2013). *Big data architecture and patterns, Part 1: Introduction to big data classification and architecture*. [En línea]. Recuperado de http://www.ibm.com/developerworks/library/bd-archpatterns1/index.html?ca=drs [Consultado el 11 de mayo de 2014].

NIST Big Data Program. (2013). Big Data Architecture Models: A Survey Version 1.2 [En línea] Disponible en: http://bigdatawg.nist.gov/show_InputDoc.php [Consultado: 01-Jun-14].

Pettey C., y Goasduff L. (2011). *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. [En línea]. Recuperado de: http://www.gartner.com/newsroom/id/1731916 [Consultado el 17 de mayo de 2014].

Preimesberger, Chris. eWeek. 28. 13. (2011), [En línea] Disponible en: http://search.proquest.com/docview/885430073/1366B171EE72EDB474F/1?accountid=43790, [Consultado el 30 de abril de 2014].

searchdatacenter.techtarget.com. (2013). *Reporte de Gartner analiza "big data" alrededor de tecnología de datos*. [En línea]. Recuperado de http://searchdatacenter.techtarget.com/es/noticias/2240171952/Reporte-de-Gartner-analiza-big-data-alrededor-de-tecnologia-de-datos [Consultado el 30 de diciembre de 2014].

Soares, S. (2012a). *A Framework that Focuses on the "Data" in Big Data Governance*. [En línea]. Recuperado de http://ibmdatamag.com/2012/06/a-framework-that-focuses-on-the-data-in-big-data-governance/ [Consultado el 5 de mayo de 2014].

Soares, S. (2012b). *Big Data Reference Architecture*. [En línea]. Recuperado de http://sunilsoares.wordpress.com/2012/07/22/big-data-reference-architecture-2/ [Consultado el 5 de mayo de 2014].

Sun, H. y Heller, P. (2012). Oracle Information Architecture: An Architect's Guide to Big Data [En línea] Disponible en: http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf [Consultado: 11-May-14].

Tucker, T. (2013). *Developing a big data application for data exploration and discovery. Tips, techniques, and practical guidelines to help you get started*. [En línea]. Recuperado de http://www.ibm.com/developerworks/library/bd-exploration/index.html?ca=dbg-twodw [Consultado el 11 de mayo de 2014].

zdnet.com, CBS Interactive, (2013), [En línea] Disponible en: http://translate.google.com.co/translate?hl=es&sl=en&tl=es&u=http%3A%2F%2Fwww.zdnet.com%2Fblog%2Fvirtualization%2Fwhat-is-big-data%2F1708&anno=2 [Consultado el 11 de marzo de 2013].

Zdnet.com. (2010). [En línea]. Recuperado de: http://www.zdnet.com/search?q=big+data [Consultado el 3 de abril de 2014].

Zikopoulos, P., Eaton C., DeRoos D., Deutsch T. y Lapis G. (2012). *Understanding Big Data*. United States of America: McGraw-Hill.