Big Data Analytics: propuesta de una arquitectura

Jonathan Solano Rodriguez y Estefany Leiva Valverde

Escuela de Ingeniería,
Universidad Latinoamericana de Ciencia y Tecnología,
ULACIT, Urbanización Tournón, 10235-1000
San José, Costa Rica
jsolanor678@ulacit.ed.cr, xleivav759@ulacit.ed.cr
http://www.ulacit.ac.cr

Abstract. Las organizaciones generan grandes volúmenes de datos de una amplia variedad de tipos y a altas velocidades, lo cual provoca inconvenientes para su análisis y gestión con herramientas y métodos convencionales. Esto ha motivado el surgimiento de nuevas tecnologías para realizar un procesamiento adecuado de dichos datos. En este trabajo Se realiza una revisión sistemática de arquitecturas de Big Data Analytics creadas por empresas de la industria de software, así como varias propuestas académicas. Esta revisión tiene como fin realizar una comparación para identificar las tecnologías y métodos se están utilizando en la gestión de Big Data. Con base en la comparación efectuada se propone una arquitectura para Big Data Analytics que se divide en cuatro capas: origen de datos, recuperación y transformación, análisis y visualización. La arquitectura propuesta tiene como objetivo proporcionar información para orientar sobre las herramientas que pueden ser utilizadas para llevar a cabo Big Data Analytics.

Keywords: Big Data, Big Data Analytics, Arquitectura Big Data.

1 Introducción

En la actualidad la información que generan las organizaciones se caracteriza por la velocidad con la cual es producida, el volumen que es generado en poco tiempo y la variedad de los tipos de datos. Esto ha sido motivado por el surgimiento de las nuevas tecnologías que en la actualidad se encuentran presentes en todos los niveles de las vidas de las personas: abarcando los procesos de producción, comercialización, seguridad y entretenimiento de los individuos. Por lo tanto, los tipos de información que se generan pueden ser estructurados, semi-estructurados y no-estructurados, lo que dificulta su almacenamiento, gestión y análisis.

Esto ha originado que las organizaciones tengan dificultades a la hora de gestionar y analizar sus datos, lo cual motivó el surgimiento de nuevos métodos y técnicas para la gestión y análisis, aprovechando las grandes capacidades de procesamiento y almacenamiento de los equipos de cómputo y la drástica reducción de su costo en los últimos años.

En este contexto surgió Big Data Analytics como un conjunto de tecnologías, métodos y técnicas para el análisis de datos, con el fin de transformar datos en conocimiento y ofrecer ventajas competitivas al permitir una adecuada toma de decisiones en los negocios. Big Data Analytics hace uso de técnicas mejoradas para la extracción de información relevante con base en métodos de inteligencia de negocios y minería de datos.

En esta investigación se propone una arquitectura para Big Data Analytics, a partir del estudio de las propuestas que existen en la industria y académicamente, tomando lo más importante de ellas y combinándolas para así crear una arquitectura, identificando las tecnologías, métodos y técnicas utilizadas para el almacenamiento, análisis de datos y la presentación de los resultados del análisis.

La metodología de trabajo que se implementará en este artículo, consistirá en una revisión de literatura de la documentación existente sobre las arquitecturas de Big Data; tanto las que son propuestas por el mercado actual como por los académicos, para poder proponer una arquitectura en base a ellas.

2 Estado del arte

Existen diferentes tecnologías que se relacionan con Big Data, como lo es el Framework Hadoop que permite trabajar con grandes cantidades de datos ya sean terabytes, petabytes, entre otros. Está conformado por Hadoop Common, Hadoop Distributed File System (HDFS) y Hadoop MapReduce que permite el procesamiento de datos.

2.1 Big Data

El término de Big Data hace referencia a grandes cantidades de datos de diferentes tipos, entre los que se incluyen datos estructurados, semi-estructurados y no estructurados. En general, el procesamiento de grandes cantidades de datos no puede ser gestionado o analizados a grandes velocidades usando métodos y técnicas convencionales. La figura 1 hace referencia a las 4 propiedades que conforman Big Data.

2.2 Big Data Analytics

Es el proceso de examinar los datos para encontrar patrones entre ellos e información útil que se puede utilizar para la toma de decisiones. Con el Big Data Analytics los encargados de analizar los datos pueden analizar grandes volúmenes que con un análisis convencional y con soluciones de inteligencia de negocios no pueden ser tratados. La figura 2, indica los beneficios de obtenidos al procesar toda esa información.

2.3 Algunas arquitecturas existentes de la industria son:

Arquitectura IBM

Fig. 1. Las 4v de Big Data.

Una de las arquitecturas de Big Data estudiada es la definida por IBM (IBM, 2013) donde se presentan cuatro capas que procesan la información antes de ser presentados a los usuarios finales, las cuales son el origen de datos, la recuperación de datos y la transformación para que estos puedan ser analizados y una última capa que realiza la presentación de datos.

En la capa de origen de datos se define de donde se van a obtener los datos para ser procesados. Los cuales son divididos en 10 grupos los cuales son:

1. Información geográfica
2. Contenido generado por humanos
3. Sensores
4. Sistemas heredados por las organizaciones
5. Registros (Logs)
6. Bases de datos
7. Sistemas de gestión de datos (DMS)
8. Dispositivos inteligentes
9. Sensores de captación de datos
10. Otros proveedores de datos externos

Dentro de las arquitecturas estudiadas, solamente IBM realiza una clasificación para los datos de entrada.

Debido a que la información que proviene del origen de datos puede ser muy variable, se definió la recuperación y transformación con el fin de poder decidir si deben de ser transformados o se pueden almacenar en su formato actual, esta capa se divide en tres grupos los cuales son:

Fig. 2. Big Data Analytics.

1. Adquisición de datos: Captura los datos de las fuentes de datos y se encarga de transferirlos al grupo de almacenamiento o de transferirlos al grupo de transformación.
2. Transformación de datos: Realiza una transformación de los datos para que puedan ser procesados como un solo conjunto.
3. Almacenamiento de datos distribuidos: Este grupo es el encargado de almacenar los datos ya procesados.

Cuando la información ya se encuentra almacenada es necesario realizar un análisis de los datos, para esto la capa de análisis se encargara extraer información relevante para el negocio, la cual es dividida en tres grupos los cuales son:

1. Motor de análisis: Uso de componentes para identificación de entidades, administración de modelos y algoritmos de análisis).
2. Administración de modelos: Mantiene modelos estadísticos.
3. Identificación de entidades: Identifica y llena las entidades contextuales.

La última capa de esta arquitectura es la de consumo de datos la cual muestra el resultado del análisis realizado y son presentados para la visualización de los interesados, esto con el fin de que los datos puedan ser interpretados y se puedan tomar decisiones o crear informes en la organización. Este proceso se divide en cuatros partes:

1. Interceptor de transacciones
2. Procesos de gestión para el negocio
3. Monitoreo en tiempo real
4. Informes

Una particularidad de esta arquitectura es que cuenta con cuatro capas que abarcan las capas lógicas anteriores (origen, lectura, análisis y consumo), con el fin de obtener una mejor definición, las cuales son:

1. Integración de la información
2. Gestión de Procesos de Negocio
3. Gestión del sistema
4. Calidad del servicio

Arquitectura Oracle

En la arquitectura de Oracle (Oracle, 2013) el origen de datos se encuentra en la capa de información la cual está conformada por los componentes de gestión de información como lo son almacenes de datos. También existen otros componentes para la recuperación, integración, procesamiento y virtualización de los datos.

Además se definen dos grupos de almacenes, los datos que se han cargado con fines específicos como lo son los datos operacionales, sistemas de gestión de contenidos y estos almacenes a su vez se integran con el DataWarehouse lógico. Por otro lado dentro de esta capa se encuentran los componentes que dan tratamiento y detección de eventos para los distintos tipos de datos.

La transformación de datos se encuentra en la capa de información la cual está conformada por el procesamiento de los datos y la detección de eventos.

El análisis de la información se encuentra en la capa de procesamiento de la arquitectura, la cual contiene las actividades de procesamiento de alto nivel. Entre las aplicaciones que se encuentran en esta capa son:

- Aplicaciones de análisis basadas en los procesos.
- Aplicaciones de análisis basadas en la industria.
- Análisis personalizado de aplicaciones.

La presentación de los resultados se encuentra en la capa de interacción que está compuesta por las aplicaciones con la que los usuarios finales interactúan. Además incluye las herramientas utilizadas por los analistas para realizar las tareas de análisis.

Entre los componentes más comunes son:

- Dashboards.
- Herramientas para el análisis inteligente.
- Reportes.
- Herramientas avanzadas de análisis.
- Gráficos
- Hojas de cálculos.

Aparte de las capas mencionadas anteriormente, Oracle en su arquitectura propone 3 capas más que no se encuentran en ninguna de las otras arquitecturas estudiadas, como lo son la capa de la infraestructura compartida que incluye el hardware y las plataformas en las que se ejecutan los componentes de Big Data, lo cual incluye:

- Infraestructura de bases de datos.
- Infraestructura de Big Data.
- Infraestructura para análisis.

La capa de servicios incluye los componentes que prestan servicios de uso común. Incluye servicios como:

- Presentación de servicios.
- Servicios de información.
- Actividades de monitoreo del negocio.
- Reglas del negocio.
- Manejo de eventos.

Los dos primeros están orientados a servicios de arquitectura (SOA). Los demás proporcionan servicios para la capa de procesamiento.

Y por último la capa de distribución por múltiples canales, la cual los resultados pueden ser entregados por distintos medios ya sea por computadoras de escritorio, laptops, teléfonos celulares, tablets, emails, entre otras.

Cabe recalcar que la arquitectura puede soportar distintos componentes que afectan las capas de la arquitectura, esto incluye:

- Modelado
- Análisis
- Seguimiento
- Gestión
- Seguridad de la información

Arquitectura Cloudera

La arquitectura descrita por Cloudera (Awadallah, n.d.), (Cloudera, n.d.) es posible dividirla en cuatro capas, las cuales abarcan todo el proceso que se ha visto en arquitecturas anteriores las cuales corresponde a origen de datos, transformación, procesamiento o análisis y datos de consumo (Visualización).

La capa de origen de datos describe que se pueden cargar datos estructurados, semi-estructurados y no estructurados.

En la etapa de recuperación y transformación los datos pueden ser recuperados en dos posibles formas de como cargar los datos las cuales son por:

- Lotes
- Eventos

La carga por lotes, es para fuentes de datos estructurados y se recomienda la utilización de Apache Sqoop para realizar la carga y el caso de los se archivos se menciona el uso de:

- The Hadoop Distributed File System (HDFS)

Por otro lado la carga por eventos son recomendados cuando se necesita extraer información en tiempo real como lo son las transacciones o logs, entre las herramientas que proponen para aplicar la carga son:

– Apache Flume
– Apache Spark
– Apache HBase

Una vez que ya han sido capturados por Hadoop para su procesamiento, son almacenados con un formato y tamaño específico. Para realizar esto se divide en tres secciones:

– Formatos de almacenamiento
– Particionamiento de datos
– Control de Acceso

Posteriormente se procede al procesamiento de datos el cual se divide en:

– Transformación de datos, se realizará la transformación de datos en otros formatos o se aplicaran algoritmos de compresión.
– Analítica, se aplicarán modelos matemáticos para las la agrupación y recomendación de datos, por lo que se utilizarán sistemas de procesamiento como:
  • MapReduce
  • Apache Spark
  • Apache Giraph

Cuando la información es procesada es necesario convertirlo en un flujo de trabajo optimizado, esto es posible mediante el sistema de flujos de trabajo Apache Oozie.

En la última etapa cuando los datos ya se encuentran preparados, por lo que es necesario exponer dicha información mediante Apache Solr e interfaces JDBC que los usuarios SQL y herramientas de BI puedan utilizar.

Además incluye tres capas que abarcan todo el proceso anterior las cuales son:

1. Canalización paralela
2. Framework de Hadoop
3. Hardware

Arquitectura Microsoft

Por último la arquitectura propuesta por Microsoft (Microsoft, n.d.), está compuesta en gran parte por las herramientas de Inteligencia de Negocios, utilizando tecnología propia de Microsoft, esta se conforma de cinco capas las cuales son: orígenes de datos, integración, almacenamiento de datos, modelado y análisis de datos y la parte de visualización y reportaría.

En la capa de origen de los datos ya que se encuentran las fuentes de la cual se obtendrá la información, las cuales son:

– Aplicaciones de negocio.
– Servicios de datos maestros.

- Datos de navegación y registros web.
- Tienda Azure: Es un mercado unificado para clientes y socios de Microsoft

La parte de transformación de datos que se encuentra en la capa es la de integración, en esta se encuentran los componentes y herramientas de procesamiento e integración de datos. Los cuales son utilizados para extraer y transformar información, entre ellas se encuentran:

- SQL Server Integration Services.
- Servicios de calidad de datos.
- Servicios inteligentes de sistemas de Azure.
- Microsoft StreamInsight

Cuando la información se encuentra integrada y almacenada se procede al análisis de los datos se menciona la utilización del motor analítico de Microsoft el cual es Analysis Services, el cual tiene integrado la posibilidad de realizar y tabular módulos, multidimensionales y minería de datos.

La presentación de los datos se realiza en la capa de visualización y reporteo de datos, esta capa se encargara de presentar los datos analizados a los usuarios finales o interesados, esto se puede realizar por medio de herramientas o servicios como lo son:

- SQL Reporting Services.
- Excel, Power BI.
- SharePoint.
- En la nube (Power BI para Offices 365).

Una característica de la arquitectura de Microsoft que no tienen las demás arquitecturas analizadas es la capa de almacenamiento e integración, de forma paralela se pueden integrar tecnología como:

- Plataformas de análisis de sistemas
- Almacenes de datos paralelos
- HdInsight azure: Este es la introducción de Hadoop para el procesamiento de grandes datos en la nube.
- HortonWorks ayuda para el procesamiento distribuido mediante hadoop.

Otra característica de esta capa es que una vez que los datos en los repositorios ya que se encuentran integrados son analizados posteriormente, se pueden utilizar herramientas de almacenamiento como:

- SQL Server.
- HBase.
- Almacenamiento en Azure.
- Base de datos SQL Azure.
- SQL Server Parallel DataWarehouse (PDW Server SQL).

Arquitectura	IBM	Oracle	Cloudera	Microsoft
Origen de datos	– Estructurados. – Semi-estructurados. – No-estructurados.	– Sistema de Generación de datos. – Datos externos. – Datos operacionales.	– Estructurados. – Semi-estructurados. – No-estructurados.	– Aplicaciones de negocio. – Datos de navegación y registros web. – Datos maestros. – Sensores de dispositivos y fuentes de datos de streaming.
Recuperación y Transformación	– Adquisición de datos. – Almacenamiento de datos distribuidos. – Integración de la información.	– Procesamiento de datos. – Detección de eventos.	– Cargas por lotes y eventos. – Formatos de almacenamiento – Particionamiento de datos. – Control de Acceso.	– SQL Server Integration Services. – Servicios de calidad de datos. – Servicios de sistemas inteligentes de Azure.
Análisis	– Administración de modelos. – Identificación de entidades. – Gestión de procesos de negocio.	– Aplicaciones basadas en los procesos. – Aplicaciones basadas en la industria. – Análisis personalizado de aplicaciones. – Planificación y estrategias de negocios.	– Analítica de datos. – Flujos de trabajo.	– SQL Server Analysis Services. – Modelos multidimensionales y tabulares. – Minería de datos.
Presentación	– Interceptor de transacciones. – Procesos de gestión para el negocio. – Monitoreo en tiempo real. – Informes.	– Dashboards. – Reportes. – Herramientas avanzadas analíticas. – Gráficos. – Análisis guiado. – Hojas de cálculo.	– Apache Solr – Interfaces JDBC – Inteligencia de negocios.	– SQL Server Reporting Services. – SharePoint. – Excel, Power BI. – Cloud (Power BI Offices 365).

Table 1. Comparación de Arquitecturas para Big Data Analytics

Además se realizó el análisis de dos arquitecturas académicas las cuales son:

La arquitectura descrita en (Chan, 2013) el origen de datos se divide en (semi-estructurados - no estructurados) y (datos estructurados).

En la capa de recuperación y transformación describe dos entradas de datos son:

- DataWarehouse esto bajo la utilización del proceso ETL (Extracción, transformación y carga).
- Información en bases de datos NoSql, donde son extraídos por modelos como Hadoop cluster, HDFS y MapReduce.

El análisis de datos se realiza mediante tecnologías como MapReduce e Inteligencia de Negocios (BI), cuando este proceso es completado la información se encuentra en un estado ya procesable, por lo que es capturada por aplicaciones operacionales y analíticas o es capturada por aplicaciones que procesan información en tiempo real.

Además para los datos provenientes de Bases de datos NoSql (no estructurados y semi-estructurados) es posible realizar todo proceso análisis, procesamiento y captura de información en tiempo real.

La arquitectura propuesta por Barlow (Barlow, 2013), consiste de 4 capas: datos, análisis, integración y decisión.

Capa de datos: La componen bases de datos RDBMS, NoSQL, Hbase, datos no estructurados que se encuentran en MapReduce de Hadoop y los flujos de datos en la web. En él se sugiere dividir la capa de datos en dos subcapas, una para el almacenamiento y la otra para el procesamiento.

Capa de análisis: Se encuentra encima de la capa de datos, incluye un entorno de producción para análisis dinámicos en tiempo real, un entorno de desarrollo para la construcción de modelos y un Datamart local que actualiza periódicamente desde la capa de datos.

Capa de integración: Se encuentran las aplicaciones del usuario final y los motores de análisis, y por lo general incluye un motor de reglas y una API dinámica para análisis de "brokers" de comunicación entre los desarrolladores de aplicaciones y los científicos de datos.

Realizando una comparación entre las arquitecturas se aprecia que la primer arquitectura se le es posible agregar una capa de presentación para incluir tecnología de visualización para los usuarios finales y ambas concuerdan con el uso del framework de hadoop para el procesamiento de datos así como análisis en tiempo real.

Arquitectura	An Architecture for Big Data Analytics	Real-Time Big Data Analytics: Emerging Architecture
Origen de datos	– Estructurados – Semi-Estructurados – No Estructurados	– Bases de datos: • RDBMS • NoSQL • Hbase – Flujos de datos en la web
Recuperación y Transformación	– DataWarehouse – Bases de datos NoSql	No mencionado
Análisis	– MapReduce – Inteligencia de Negocios (BI)	– Análisis dinámicos en tiempo real – Datamart local
Presentación	No mencionado	API dinámica para análisis de "brokers"

Table 2. Comparación de Arquitecturas para Big Data Analytics

3 Desarrollo

En base a las cuatro arquitecturas de la industria analizadas anteriormente, se desarrolló una propuesta de arquitectura donde se toma lo mejor de cada de una de ellas para formar una arquitectura propia indicando las herramientas que se pueden utilizar en una de las capas y otras herramientas no mencionadas en las arquitecturas. Lo anterior con el fin de que sirva como base para aquellas organizaciones que quieran implementar Big Data, por lo que se indica cuales herramientas o métodos se pueden utilizar para ser implementada.

La arquitectura se dividirá en 4 capas las cuales son:

- Origen de datos
- Recuperación y Transformación
- Análisis
- Visualización

Primera capa - Origen de datos

Esta capa describe de donde provienen los datos que van a alimentar la arquitectura, estos pueden ser de distintas fuentes y tipos de datos. Los posibles datos a incluir pueden provenir de diferentes orígenes como lo son:

1. Estructurados
   - RDBMS (SQL Server, Sysbase, DB2, PosgreSQL, MySql)
   - DataWarehouses
   - CRM
   - ERP
   - Semi-estructurados:
     - Datos generados por máquinas
     - XML
     - JSON
     - EDI
     - Correo electrónico
2. No Estructurados
   - Redes sociales
   - Multimedia
   - QR
   - Texto
   - Datos análogos
   - GPS

Segunda capa – Recuperación y Transformación

Esta etapa se encarga de recuperar y transformar los datos, para que sean fácilmente integrados, procesados y almacenados para un análisis posterior.

La extracción de los datos se puede realizar por medio de dos métodos ya sea por lotes o eventos, por lo que es posible utilizar herramientas como:

1. Lotes

  - Apache Sqoop
2. Eventos
  - Apache Flume
  - Apache Storm
  - Apache Spark
  - Apache Chukwa

Para la transformación de los datos se pueden utilizar las siguientes herramientas:

- Apache Camel
- Data Integration ó Kettle
- SQL Server Integration Services
- Microsoft StreamInsight
- Data Quality Services (DQS)
- Oracle Warehouse Builder

Para el almacenamiento de los datos una vez transformados se pueden utilizar las siguientes bases de datos:

- Apache Hive
- Apache Cassandra
- Neo4j
- MongoBD
- Apache HBase
- Azure Sql DataBase
- Sql Server
- PostgreSQL
- IBM DB2

Tercera capa – Análisis

En esta etapa se analiza la información procesada y almacenada anteriormente y se busca la información más relevante para el negocio, por lo que utiliza tecnologías que aplican patrones matemáticos para retornar información como:

- Spark
- Weka
- Radoop
- SQL Server Analysis Services
- RapidMiner
- Pervasive
- PowerHouse
- Apache Giraph
- Rapid Analytics

Cuarta etapa – Visualización

Esta capa se encargara de presentar los datos analizados a los usuarios finales, de una manera entendible para que puedan ser utilizadas para la toma de decisiones organizacionales o creación de informes. Esto es posible realizarlo de una forma clara y configurable mediante herramienta que presentan la información de forma visual como los son:

- Reporting Services
- Qlik
- Tableau
- Apache Solr
- Splunk
- Pentaho

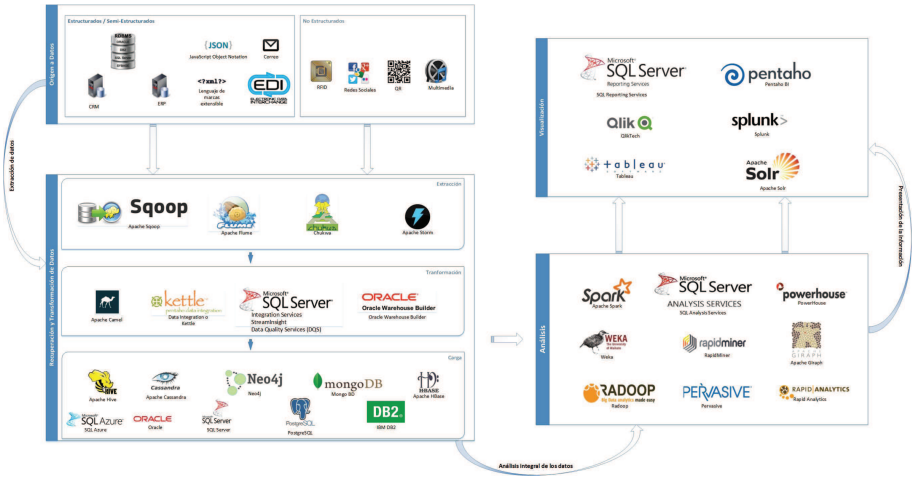

Fig. 3. Diseño de arquitectura Big Data.

4 Conclusiones

Big Data Analytics permite la exploración y análisis de grandes volúmenes de información para facilitar la toma de decisiones mediante el descubrimiento de patrones de comportamiento, tendencias en los mercados o el comportamiento de los clientes. Por lo que ofrecen información para la toma de decisiones, y por consiguiente el beneficio de los negocios. La arquitectura propuesta proporciona un modelo para aquellas organizaciones que deseen implementar un análisis de Big Data, ya que integra herramientas que existen actualmente en el mercado tanto como de software libre, software pagado, bases de datos SQL y NoSQL.

References

Awadallah, A. (n.d.). *The platform for big data.* Retrieved from https://www.hashdoc.com/documents/10968/the-platform-for-big-data pages 6

Barlow, M. (2013). *Real time big data analytics: Emerging architecture.* O'Reilly. pages 10

Chan, j., Joseph O.1. (2013). An architecture for big data analytics. *Communications of the IIMA*, *13*(2), 1 - 13. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=aci&AN=95612792&lang=es&site=ehost-live pages 10

Cloudera. (n.d.). *Information-driven financial services, big data and the enterprise data hub.* Retrieved from http://www.cloudera.com/content/dam/cloudera/Resources/PDF/whitepaper/Cloudera_Financial_Services_Industry.pdf pages 6

IBM. (2013). *Big data architecture and patterns, part 3: Understanding the architectural layers of a big data solution.* Retrieved from http://www.ibm.com/developerworks/library/bd-archpatterns3/ pages 3

Microsoft. (n.d.). *Understanding microsoft big data solutions.* Retrieved from http://msdn.microsoft.com/en-us/library/dn749804.aspx pages 7

Oracle. (2013, Setiembre). *Big data & analytics reference architecture.* Retrieved from http://www.oracle.com/technetwork/topics/entarch/oracle-wp-big-data-refarch-2019930.pdf pages 5