UNIVERSIDAD SIGLO 21

CARRERA DE INGENIERÍA EN SOFTWARE

Trabajo Final de Graduación

Proyecto de Aplicación Profesional (PAP)

INTEGRACIÓN DE TECNOLOGÍAS BIG DATA EN SOLUCIONES DE ENTERPRISE DATA WAREHOUSE

Autor: César Zamboni
Prof. seminario: Adriana Pérez
Prof. tutoría: Ana Carolina Ferreyra

2017

Resumen

En este trabajo se propone una solución para un caso de uso de análisis de archivos de logs (registro de eventos de un servidor web) que incorpora tecnologías de Big Data al Enterprise Data Warehouse (EDW) actualmente en uso, en particular se utiliza Apache Hadoop y algunos productos open-source de su ecosistema, como Flume para la incorporación de la nueva fuente de datos, MapReduce (Pig y Hive) y HDFS para procesarlos y almacenarlos de manera distribuida, Sqoop para integrarlos al EDW y Qlikview para la visualización y análisis.

A su vez se describe la problemática actual de las pequeñas y medianas empresas con respecto al manejo de sus datos para el análisis y la toma de decisiones en todos los niveles de la organización (operativas, tácticas y estratégicas) y la necesidad de incorporar técnicas de Analítica avanzada, con el objetivo de seguir siendo competitivas en el contexto de la Transformación digital que las empresas pioneras en adoptar nuevas tecnologías disponibles están llevando a cabo.

Se analiza el nuevo paradigma de Data Lake que las empresas deberían adoptar para seguir siendo competitivas, en particular pequeñas y medianas empresas que actualmente poseen soluciones de EDW, y se propone la adopción de tecnologías de Big data y procesamiento en la nube para que un cliente hipotético pueda adecuarse al nuevo escenario.

Palabras clave: Big data, Enterprise Data Warehouse, Apache Hadoop, Open-source, Analítica avanzada, Computación en la nube, Transformación digital.

Abstract

This dissertation proposes a solution for a use case of log file analysis (web server event log) that incorporates Big Data technologies into the Enterprise Data Warehouse (EDW) currently in use, using Apache Hadoop and some open-source ecosystem products such as Flume for incorporating the new data source, MapReduce (Pig and Hive) and HDFS for distributed processing and storage, Sqoop for integration into EDW and Qlikview for visualization and analysis.

At the same time, it describes the current problems of small and medium-sized enterprises about the management of their data for analysis and decision making at all levels of the organization (operational, tactical and strategic) and the need to incorporate Advanced analytics, with the objective of remaining competitive in the context of the Digital Transformation that companies pioneered to adopt new technologies available are carrying out.

It analyzes the new Data Lake paradigm that companies should adopt to remain competitive, particularly small and medium-sized companies that currently own EDW solutions, and proposes the adoption of Big Data technologies and processing in the cloud so that a customer can adapt to the new scenario.

Keywords: Big data, Enterprise Data Warehouse, Apache Hadoop, Open-source, Advanced Analytics, Cloud computing, Digital Transformation.

Tabla de contenido

Tabla de imágenes

1. Introducción

“La transformación digital implica un completo cambio de paradigma en la forma en la que encaramos los negocios hoy en día. Las empresas que logren incorporar la tecnología adecuada y consigan reinventarse, serán disruptores y emergerán como auténticos referentes, evitando ser desplazados de sus mercados” indicó Aníbal Carmona, Presidente de CESSI (Cámara de Empresas de Software y Servicios informáticos de Argentina), en el marco de la conferencia "Panorama de la Transformación Digital de los Negocios" en junio de 2016.

Tras indagar sobre la Trasformación Digital y sus nuevos paradigmas para los negocios, el presente trabajo se propone demostrar de manera práctica cómo incorporar tecnologías más adecuadas al nuevo paradigma de Data Lake para la adquisición, almacenamiento y análisis de datos para la toma de decisiones. Los destinatarios principales de esta propuesta son directivos en general de pequeñas y medianas empresas, pero también cualquier miembro de estas organizaciones que actualmente tome decisiones a partir de tecnologías tradicionales de Bussiness Intelligence basadas en el Enterprise Data Warehouse.

Partiendo de la premisa de que todos los negocios de hoy en día están dirigidos por sus datos, y que por lo tanto estos son el factor diferenciador, se verá cómo el nuevo panorama de la Transformación Digital convierte en una necesidad vital para la organización incorporar las nuevas tecnologías hoy disponibles que permiten capturar aquellos datos que las tecnologías tradicionales hasta ahora no han podido manejar adecuadamente, es decir con efectividad y a un costo razonable.

Por lo tanto, no se trata solo de incorporar nuevas tecnologías sino de exponer la importancia del cambio cultural que se requiere en las organizaciones para utilizar estas tecnologías y evitar ser desplazados de los mercados.

2. Problemática

A lo largo de la historia en las empresas se ha dado una constante lucha entre la cultura del negocio y la cultura de IT (Information Technology). Es decir, entre quienes trabajan en la empresa cumpliendo los roles propios del negocio (finanzas, RRHH, marketing, ventas, etc.) y quienes trabajan en el área de tecnología en la misma empresa (responsables de mantener la infraestructura tecnológica, tanto el software como el hardware).

Por un lado, el negocio necesita tener vistas, perspectivas y contexto locales, es decir información específica de su área o departamento, mientras que por el otro IT intenta dar una solución centralizada para toda la organización con el fin de facilitar el mantenimiento, optimizar los recursos y para que la gobernabilidad, el modelado y la carga de datos sólo tenga que hacerse una vez, garantizando accesibilidad, integridad y confidencialidad de la información.

En esta lucha el negocio generalmente se inclina por el uso de recursos y soluciones aisladas (Ej. bases de datos propias, datos en archivos Excel de manera local), mientras que IT propone la adopción de una base de datos unificada que alimente a las bases departamentales, con el fin de resolver los requerimientos de datos de los directivos que necesitan una visión global, pero también la de usuarios de mandos medios y operaciones con necesidades específicas de sus áreas o departamento.

Esta búsqueda de una solución centralizada de IT es la que derivó en lo que conocemos como el Enterprise Data Warehouse (EDW), que es una tecnología consolidada y que ha venido madurando por más de 30 años. El EDW es el núcleo de los proyectos de Business

Intelligence o Inteligencia de Negocios (BI), que permiten reunir, depurar y transformar datos de sus sistemas transaccionales en información estructurada para dar soporte a la toma de decisiones.

Aunque más adelante en este trabajo se hará una descripción más detallada podemos decir que un EDW es un almacén o base de datos unificada que contiene toda la información del negocio de una organización y lo hace accesible en toda la empresa.

Con el tiempo, sin embargo, además de permanecer la ya mencionada lucha entre el negocio e IT, han aparecido nuevos desafíos:

• **Costos**: A medida que los datos crecen también lo hacen los costos. A menudo las organizaciones luchan para justificar el costo de almacenamiento contra el valor proporcionado por estos datos.

• **Complejidad**: Muchos EDWs han crecido más allá de sus requerimientos funcionales iniciales, para pasar a manejar transformaciones de datos necesarias y algunas veces complejas, que se sobrecargan más allá de su uso previsto para realizar análisis de datos.

• **Escalabilidad**: En los últimos años han aparecido nuevas fuentes de datos, muchas veces de datos masivos. Estos nuevos datos, con su estructura variable y de gran escala, significan un gran reto para los actuales EDW.
(A Hortonworks White Paper, March 2015)

Pero a su vez han surgido cambios dramáticos en la tecnología disponible. La tecnología actual reduce el costo del almacenamiento de información, permite análisis de datos en tiempo real y le proporciona al usuario información a mayor velocidad.
Son estos nuevos retos y el impacto de las nuevas tecnologías los que han dado lugar a un nuevo modelo o paradigma: *Data Lake*.

Un Data Lake, o Business Data Lake, es un repositorio de almacenamiento que contiene una gran cantidad de datos en bruto en su formato original, incluyendo datos estructurados, semi-estructurados y no estructurados. La estructura y los requisitos de datos no requieren ser definidos sino hasta que se necesitan dichos datos. (Dull, Big Data Cheat Sheet on Hadoop, 2015)

La tecnología hoy disponible que puede soportar el nuevo enfoque de Data Lake es Apache Hadoop, entre otras. El proyecto Apache Hadoop desarrolla software de código abierto para la computación escalable, fiable, y distribuida y actualmente es la plataforma más utilizada en proyectos de Big Data (datos masivos).

En pocas palabras los proyectos de Big Data se encargan de capturar, almacenar y procesar datos masivos, que llegan a grandes velocidades y cuya estructura es muy variable que las organizaciones pueden analizar para optimizar su proceso de toma de decisiones. Apache Hadoop es un framework que permite el procesamiento distribuido de grandes conjuntos de datos a través de clusters de computadoras que utilizan modelos de programación simples. Está diseñado para pasar de servidores individuales a miles de máquinas, cada una ofreciendo computación y almacenamiento local.

Esta tecnología permite almacenar todo tipo de datos de una forma más barata y también procesar estos datos más rápido, y está llevando a las empresas y organizaciones pioneras en adoptarlas al nuevo paradigma cultural y tecnológico conocido como Trasformación Digital.

Ante estos desafíos y problemática, en el marco del presente trabajo, es que se plantean los siguientes interrogantes pensando principalmente en pequeñas y medianas empresas:

¿Comprenden los ejecutivos de las pequeñas y medianas empresas de nuestra región la importancia de incorporar las nuevas fuentes de datos a sus EDW?

¿Están estas organizaciones en condiciones de afrontar los costos de incorporar nuevas tecnologías y contratar los recursos humanos para implementar proyectos de Big Data?

¿Es posible incorporar de manera progresiva tecnologías de Big Data como complemento de una solución existente de EDW?

¿Conocen los ejecutivos los riesgos y las consecuencias de no estar preparados para los nuevos paradigmas en el manejo de las nuevas fuentes de datos?

¿Podemos llevar a cabo en estas organizaciones pequeños proyectos piloto con tecnologías de Big que se adapten a sus presupuestos y recursos?

2.1 Justificación

Según el informe presentado en Julio de 2015 por el *Centro Global para la Transformación Digital de los Negocios*, en los próximos cinco años la transformación digital desplazará del mercado a cerca del 40 por ciento de las actuales compañías. (Global Center for Digital Business Transformation, 2015)

Transformación digital es un concepto que abarca todo tipo de cambios que se van a producir en las empresas y en la economía para hacer el mejor uso posible a la acumulación de tecnologías de estas últimas décadas, y para estos cambios, tal como cita Mar Castaño (*Data Scientist Director* de la consultora española Territorio creativo) "la información y conocimiento extraídos a partir de los datos constituyen un factor clave de éxito y la materia prima fundamental para:

- La búsqueda de nuevas fuentes de ingresos
- El desarrollo de nuevos productos y servicios
- La optimización de procesos
- La reducción de costes
- La mejora del flujo de trabajo de los empleados y el incremento de la productividad
- El conocimiento de los clientes y la mejora de su experiencia
- La adquisición, retención y fidelización de los clientes
- La detección de tendencias de consumo y nuevos patrones de comportamiento." (Castaño, 2016)

Las empresas siempre han tenido "sed de información", y esa sed ha evolucionado desde ser meramente útil para las operaciones, hasta la posición que tiene hoy en día como la base de la estrategia y el éxito. (Capgemini, 2013)

Existe un acceso sin precedentes al poder de procesamiento y almacenamiento informático y los usuarios cada vez más exigen un mejor y mayor acceso a todos sus datos, sin importar de dónde vengan y cuáles sean sus características.

Las empresas que utilizan EDW generalmente manejan datos estructurados en bases de datos relacionales. Sin embargo, existe desde hace mucho tiempo una gran cantidad de datos no estructurados o semi-estructurados que las tecnologías de EDW no son lo suficientemente eficaces como para manejarlos. A estos últimos tipos de datos son a los que nos referiremos por ahora como Big Data.

Se trata de datos con los que por décadas hemos contado pero que no teníamos la tecnología apropiada para procesarlos, almacenarlos y mezclarlos con los datos estructurados. Hoy podemos hacerlo con tecnologías de Big Data como Apache Hadoop que nos permite mezclar y combinar Big Data con los tradicionales datos estructurados en una fracción del tiempo y costo que requieren los sistemas de EDW.

Estas tecnologías además de almacenar y procesar Big Data de una manera más barata y rápida que el EDW permiten realizar Advanced Analytics (Analítica avanzada), para lo cual se requiere de personal técnico especializado (data scientist o científico de datos) pero que ya están siendo tendencia en las empresas pioneras en emplear estas tecnologías, y que permiten a las organizaciones brindar mejores servicios a sus clientes y mejorar la toma de decisiones.

Por lo tanto, es importante para que las pequeñas y medianas empresas sigan siendo competitivas en el mediano plazo que empiecen a incorporar tecnología Big Data y se familiaricen con ella, ya que incluso pueden sacar provecho, aunque aún no cuenten con Big Data para procesar, y aunque aún no puedan asumir el costo para contar con el personal técnico especializado.

Es posible emplear Hadoop simplemente como una extensión del EDW para procesar y almacenar datos estructurados de una manera más barata y rápida, y luego con el tiempo ir incorporando otras aplicaciones del ecosistema de Hadoop para realizar análisis cada vez más complejos y avanzados.

3. Objetivo general del proyecto

Implementar tecnologías de Big Data dentro de una arquitectura tradicional de Enterprise Data Warehouse de pequeñas y medianas empresas.

4. Objetivos específicos del proyecto

- Advertir sobre la Trasformación Digital en marcha para ver la importancia de incorporar tecnologías de Big data.
- Analizar los sistemas tradicionales de Enterprise Data Warehouse (EDW) para detectar las limitaciones actuales.
- Presentar el nuevo paradigma de Data Lake para el EDW como una oportunidad para incorporar tecnologías de Big Data.
- Advertir sobre la importancia de incorporar Advanced Analitycs o Analítica Avanzada para analizar los datos en el nuevo contexto y paradigmas de la Transformación Digital.
- Proponer una forma gradual y de bajo impacto económico para incorporar las tecnologías de Big Data en PYMES como complemento de sus EDW.
- Realizar una aplicación que demuestre como complementar el EDW con Big Data.

5. Objetivo general del sistema

- Incorporar, almacenar y procesar mediante tecnología Big Data datos que actualmente no están disponibles en el EDW, con el fin de integrarlos a los datos ya existentes y proporcionar información adicional sobre los posibles clientes, productos y servicios.

6. Límite

Desde la adquisición de datos hasta la visualización de los resultados del análisis que requieren los distintos niveles de la organización (directivos, gerentes, operarios) con herramientas tradicionales de BI.

7. Alcance del trabajo

- Adquisición de datos procedentes de nuevas fuentes con herramientas de Big Data.
- Procesamiento y almacenamiento de dichos datos con herramientas de Big Data.
- Almacenamiento de los nuevos datos en EDW
- Análisis con herramientas de Big Data y de BI.
- Visualización con herramientas de BI.

8. No Contempla

Analítica avanzada. Adquisición y procesamiento de datos no estructurados.

9. Marco Teórico

El presente trabajo pretende ser útil para empresas que actualmente hacen uso de EDW y que sus requerimientos de datos exigen incorporar nuevas fuentes de datos no estructurados, mayor espacio para almacenamiento y un mayor poder de procesamiento a menor costo. Por lo tanto, se comenzará describiendo los conceptos más importantes ya mencionados en las secciones anteriores:

- Transformación digital (IoT y Big Data).
- BI y Enterprise Data Warehouse (EDW).
- Data lake.
- Hadoop.
- Advanced analytics.

9.1 Transformación digital

Llamamos Transformación digital a la profunda y acelerada transformación de las actividades de negocio, procesos, competencias y modelos para aprovechar plenamente los cambios y oportunidades de las tecnologías digitales y su impacto en la sociedad de una manera estratégica.

Si bien las nuevas tecnologías digitales siempre han tenido impacto en los negocios y la sociedad, uno de los cambios que se ven en los últimos años es la velocidad a la que todo está sucediendo. Las evoluciones tecnológicas y los cambios que traen sobre nosotros están acelerando rápidamente, mostrando un crecimiento exponencial y trayendo consecuencias.

Y como reflejan algunos estudios que se referencian más adelante en este trabajo la Transformación digital se está convirtiendo en la estrategia clave para las empresas que quieran mantenerse competitivas en el mercado actual globalizado que exige reinventar la manera de hacer las cosas. Algunos analistas estiman que el 40% de las actuales empresas no sobrevivirán a este gran cambio y, por tanto, es necesario conocerlo en profundidad.

A diferencia de transformaciones anteriores ya no alcanza con el uso de herramientas digitales para automatizar y mejorar la forma de trabajo existente, sin modificar sustancialmente la manera de hacer las cosas y sin nuevas reglas de juego. La transformación actual tiene más que ver con cambiar por completo la manera de trabajar.

Galen Gruman, Editor ejecutivo de la revista Infoword, especializada en Tecnologías de Información desde 1978, enfatiza que el concepto más importante de la trasformación digital es “fungibilidad”. Con este término se refiere a la capacidad intrínseca de algo para ser cambiado sustantivamente. Es decir que los ítems diseñados deberían tener esta capacidad adaptativa, ya sean productos o servicios como así también los procesos para crearlos, administrarlos y entregarlos. (Gruman, 2016)

Entonces si bien en libros de otras épocas sobre Transformación Digital ya se hablaba de concepto de fungibilidad, aunque no se pusiera en práctica, la mayoría de la gente se detenía en la parte de Digitalización de la definición y se perdía lo relacionado con Transformación.

El siguiente es un ejemplo muy pequeño de lo que significa la fungibilidad de los ítems diseñados. Imaginemos que se aplica a la seguridad informática: en lugar de dar respuestas predefinidas a amenazas y vulnerabilidades ya conocidas de antemano tanto a en los aparatos o dispositivos como en el software, utilizar técnicas como Machine Learning para adaptar no

sólo las respuestas, sino también los modelos de amenazas subyacentes a medida que las amenazas cambian. Como veremos más adelante Machine Learning (aprendizaje automático) es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Es decir, una técnica de predicción englobada dentro de lo que también más adelante definiremos como *Advanced Analytics* o analítica avanzada de datos.

Esto nos muestra que en este contexto de Transformación digital las empresas u organizaciones deberán estar preparadas para sacar el mayor provecho posible de sus datos a través de la Analítica avanzada.

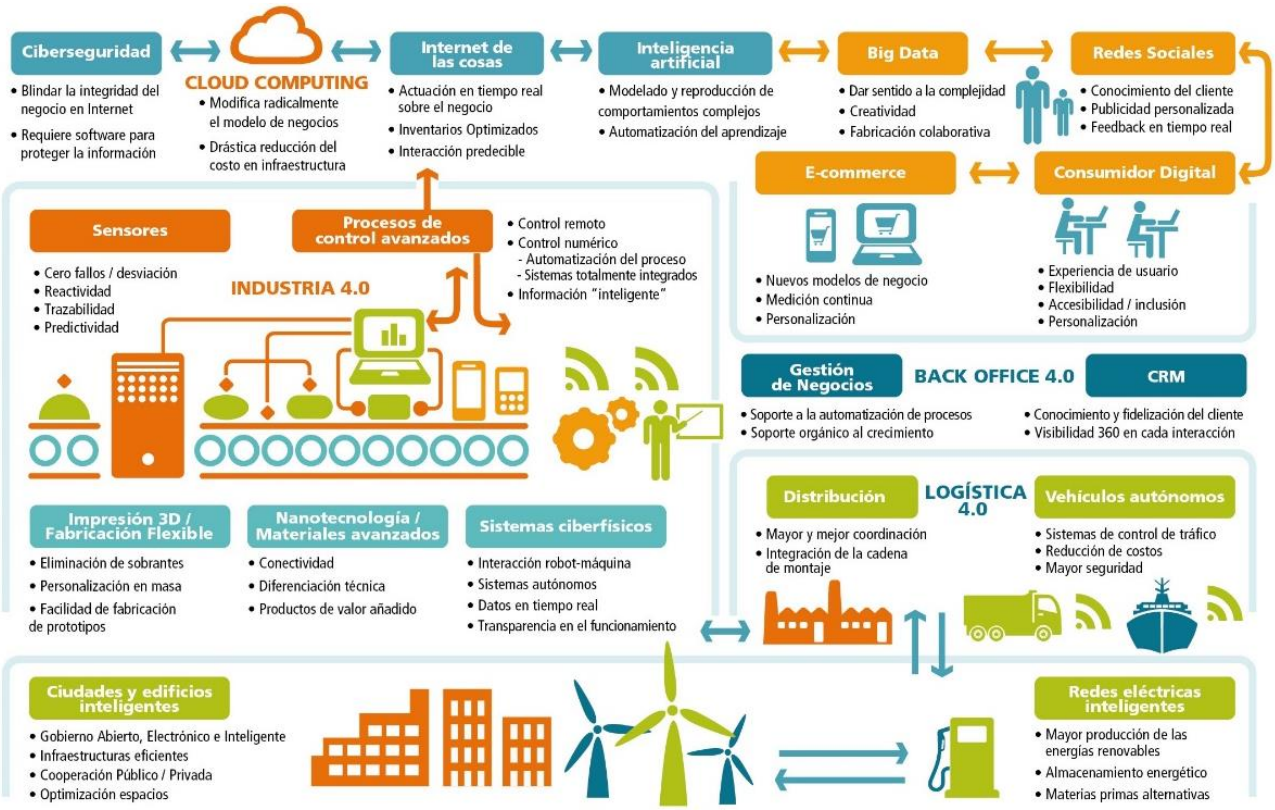
Es en este panorama en el cual las organizaciones y sus clientes generan cada vez más datos, y donde no son solo las personas las que los generan sino que también son cada vez más los objetos o dispositivos (Internet de las Cosas), donde se torna más importante la dimensión de la inteligencia para dar sentido a todo este mar de datos cada vez más complejo y creciente, para poder generar significado y valor de la información, aprovechando todo esto en el momento justo y sobre todo por los motivos correctos.

Mapa de la transformación digital

Para comprender el impacto de lo que estamos tratando a continuación vemos el Mapa de la Transformación Digital que fue presentado el pasado 14 de junio de 2016 en la Conferencia "Panorama de la Transformación Digital de los Negocios" organizado por CESSI (Cámara de Empresas de Software y Servicios informáticos de Argentina) y con el auspicio del Ministerio de la Producción de la Nación, el Ministerio de Comunicaciones de la Nación y la Unión Industrial Argentina. Este mapa diferencia los principales aportes que la tecnología puede realizar en cada una de las industrias y los actores que podrían verse beneficiados a

través de las diferentes herramientas como ciberseguridad, cloud computing, internet de las cosas e inteligencia artificial:

Imagen 1 – Mapa de la Transformación Digital en Argentina



Fuente: (CESSI, Mapa de la Transformación Digital, 2016)

"La transformación digital implica un completo cambio de paradigma en la forma en la que encaramos los negocios hoy en día. Las empresas que logren incorporar la tecnología adecuada y consigan reinventarse, serán disruptores y emergerán como auténticos referentes, evitando ser desplazados de sus mercados" indicó Aníbal Carmona, Presidente de CESSI, en el marco de la conferencia antes citada.

A continuación se da una breve definicion de algunas tecnologías que se observan en el mapa de arriba, obtenidas del glosario del sitio web de la consultora Gartner (Gartener IT Glossary, 2017):

Cloud Computing: computación *en la nube* es un estilo de computación en el que las capacidades escalables y elásticas de IT se entregan como un servicio utilizando las tecnologías de Internet.

Internet de las cosas: El Internet de las Cosas (IoT) es la red de objetos físicos que contienen tecnología integrada para comunicarse y sensar o interactuar con sus estados internos o el ambiente externo.

Inteligencia artificial (AI): es una tecnología que aparece para emular la actuación humana mediante el aprendizaje, llega a sus propias conclusiones, para entender contenidos complejos, mediante la participación en diálogos naturales con las personas, para mejorar el rendimiento cognitivo humano (también conocido como computación cognitiva) o para la sustitución de la gente en la ejecución de las tareas no rutinarias . Las aplicaciones incluyen los vehículos autónomos, reconocimiento automático del habla y la generación y detección de nuevos conceptos y abstracciones (útil para detectar posibles nuevos riesgos y ayudar a los seres humanos a entender rápidamente grandes masas de información en constante cambio).

Big Data: Son activos de información de alto volumen, alta velocidad y/o gran variedad que exigen formas rentables e innovadoras de procesamiento de esta información que proporcionan una visión mejorada, mejoras en las tomas de decisiones y la automatización de procesos.

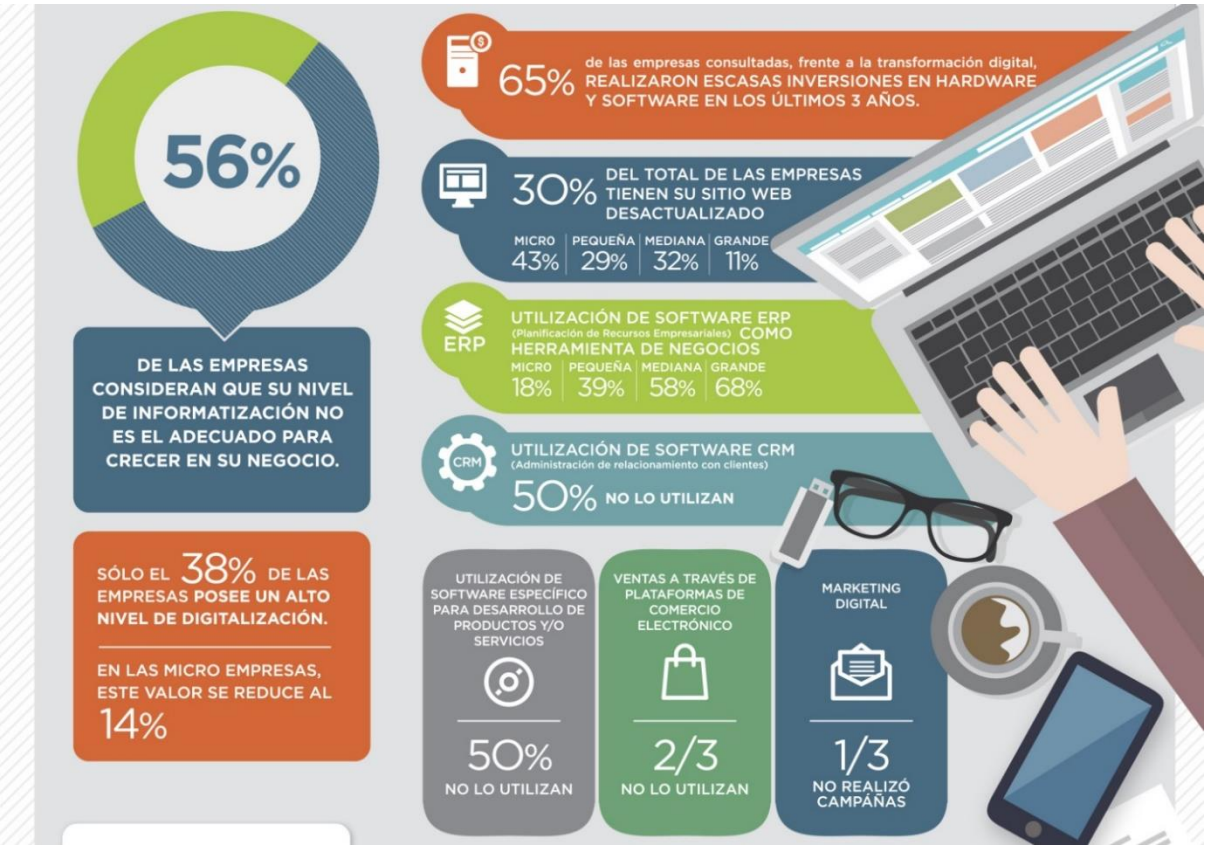
Las oportunidades surgen como AI, móviles (IoT), cloud computing y las redes sociales cambian la forma en que las organizaciones, consumidores y empleados interactúan entre sí. Las empresas y los gobiernos pueden, por ejemplo, crear nuevas experiencias para los clientes, mejorar la atención al ciudadano y entregar mejores resultados a los pacientes, además de reducir los costos y aumentar la productividad de su personal.

Y el combustible para energizar toda esta Transformación Digital es precisamente el Big Data aplicado a la inteligencia de negocios.

Transformación digital de las empresas en Argentina

A continuación, se muestran los datos de la *"Encuesta sobre el nivel de transformación digital de las empresas en Argentina"* realizada por el Observatorio Permanente de la Industria del Software y Servicios Informáticos (OPSSI), en los meses de mayo y junio de 2016, en la cual participaron CEOs de todos los sectores de la industria:

Imagen 2 - Encuesta sobre el nivel de transformación digital de las empresas en Argentina



Fuente: (CESSI, Estado y Perspectivas de la Transformación Digital en las Empresas Argentinas, 2016)

Como se puede ver en el gráfico existe una gran brecha entre la situación actual de informatización, digitalización o uso de nuevas tecnologías y lo que las empresas estiman que van a necesitar en cuanto a inversiones para el año que viene. Esta situación es más marcada aun en las pequeñas empresas.

Y es aquí donde Big Data surge no sólo con su dimensión tecnológica, sino también con una dimensión social, económica, política y cultural. Se abre lugar para nuevas oportunidades a partir de cambios en la organización del trabajo, mejoras en la toma de decisiones, la forma de producción y el acceso a los mercados que permiten reducir brechas de productividad y generar mejoras en la competitividad, en la calidad de vida y en la inclusión social.

Big Data permite a las empresas hacer ajustes significativos y estratégicos que reducen al mínimo los costos y maximizan los resultados. Si se conoce lo que los consumidores y los empleados están haciendo actualmente, se pueden crear proyecciones para lo que van a hacer en el futuro, y empezar a aplicar los cambios para hacer frente a estas necesidades y objetivos. Por lo que una transformación digital no está completa a menos que una empresa adopte Big Data.

Estudios sobre la adopción de Big Data en las empresas

A continuación, se muestran datos de un estudio elaborado por la empresa Accenture Analytics en abril de 2014 llamado "Big success from Big Data" en el cual además de datos estadísticos podemos ver recomendaciones para abordar un proyecto de Big Data. Accenture encuestó a más de 1.000 empresas de siete industrias distintas y con sede en 19 países que habían completado al menos una implementación de Big Data.

El reporte dice que las organizaciones que empiezan y completan proyectos de Big Data ven resultados prácticos y un valor significativo. Los ejecutivos reportan que Big Data ofrece resultados empresariales para un amplio espectro de objetivos corporativos estratégicos, desde nuevas fuentes de ingresos y el desarrollo de nuevos mercados hasta la mejora de la experiencia del cliente y el mejoramiento del rendimiento en toda la empresa. Las organizaciones consideran Big Data como extremadamente importante y central para su estrategia digital.

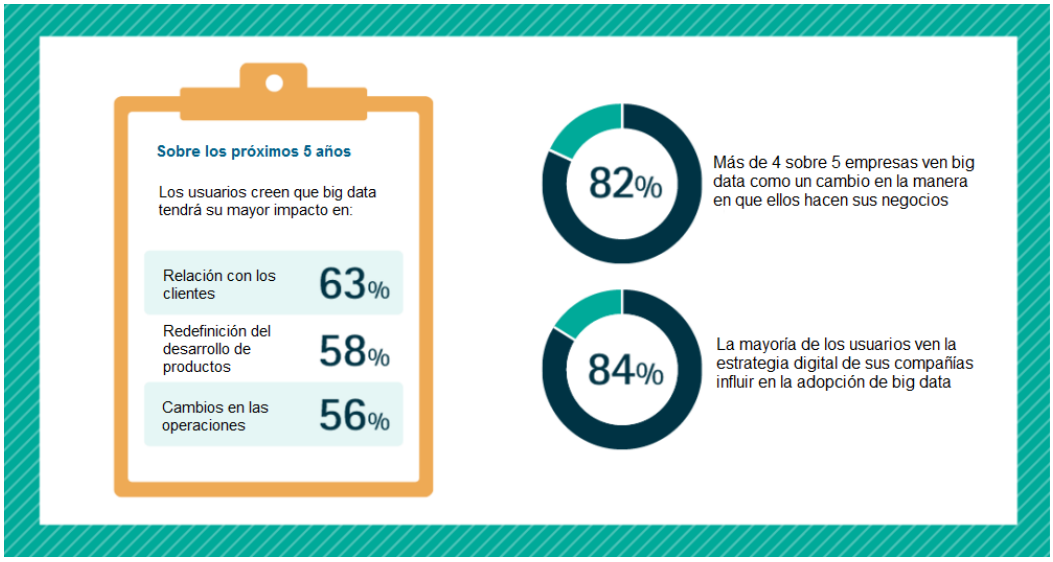
Los usuarios que han completado al menos un proyecto están muy satisfechos con sus incursiones iniciales en Big Data. La gran mayoría informa que están satisfechos con los resultados del negocio y que su iniciativa de Big Data está satisfaciendo sus necesidades.

Imagen 3 - Big success from Big Data

Fuente: (Accenture, Big Success with Big data, 2014)

Los pioneros en adoptar esta tecnología están viendo Big Data como el facilitador para transformar sus organizaciones en empresas digitales. Pensar en los datos como un activo requiere que las organizaciones cambien sus mentalidades, se vuelvan más centradas en los datos y ensamblen y adquieran las herramientas y habilidades necesarias para administrar los datos a velocidad y a escala.

Imagen 4 - Big success from Big Data



Fuente: (Accenture, Big Success with Big data, 2014)

"Las empresas han llegado a un punto de transición en el que, en lugar de hablar sobre los resultados que podrían conseguir con Big Data, están empezando a ver beneficios reales como aumentos de ingresos, una mayor fidelización de los clientes y operaciones más eficientes… Están empezando a ver que el Big Data es una de las piedras angulares de la transformación digital.", según Narendra Mulani, senior managing director global de Accenture Analytics.

En el informe también se hacen algunas recomendaciones a los ejecutivos de empresas, para obtener mejores resultados en sus proyectos de Big Data y no caer en los problemas arriba planteados:

- Explorar todo el ecosistema de Big Data y actuar con rapidez, ya que la velocidad de cambios tanto en los datos como en las tecnologías se incremente día a día.
- Empezar poco a poco: para demostrar la importancia de Big Data es recomendable empezar solo en un área de negocio a través de un programa piloto o una prueba de concepto, y no en todo el negocio a la vez.
- Desarrollar conocimientos: Las empresas deben reforzar sus programas de capacitación a empleados, ya que existe un gran déficit de ingenieros con el rol de *data scientist* o científico de datos, indispensable para sacar el mayor provecho de los proyectos de Big Data a través de la analítica avanzada. El informe de Accenture arroja que "el 54% de los ejecutivos afirman que sus empresas ya ofrecen oportunidades de formación técnica interna a sus empleados. La mayor parte de las organizaciones también recurren a expertos externos; apenas un 5% de los encuestados aseguran que sus empresas utilizan exclusivamente recursos internos para sus proyectos de Big Data".

Del estudio también se desprende que las grandes empresas (con más de 10.000 millones de dólares en ingresos anuales) enfocan el Big Data de un modo distinto a las pequeñas empresas (con menos de 500 millones de dólares de ingresos anuales):

- En cuanto a si consideran al Big Data como extremadamente importante, el 67% de los ejecutivos de grandes empresas respondieron afirmativamente, mientras que en empresas más pequeñas solo un 43% de los encuestados lo afirman.
- En las grandes empresas tienen una percepción del Big Data más amplia que en las pequeñas y utilizan más fuentes de datos en sus proyectos de Big Data, como datos de redes sociales (54% frente a 29%), datos de visualización (50% frente a 29%) o datos no estructurados (49% frente a 36%).
- El 62% de los ejecutivos de grandes empresas afirman que los altos directivos comprenden y apoyan las iniciativas de Big Data, frente al 42% de los encuestados en pequeñas empresas.

Los ejecutivos revelan que las principales barreras que han debido superar en la adopción de Big Data son:

1. la seguridad (el mayor problema, citado por el 51%),
2. las limitaciones presupuestarias (citado por el 47%),
3. la falta de expertos en Big Data (citado por el 41%),
4. la falta del uso continuado de Big Data y analítica (citado por el 37%),
5. y la integración en los sistemas existentes (citado por el 35%). (Accenture, Las empresas consideran Big Data fundamental para su Transformación Digital, 2014)

9.2 Business Intelligence

Según la definición de la consultora Gartner Business Intelligence (BI) es un término paraguas que incluye aplicaciones, infraestructura y herramientas, y las mejores prácticas que posibilitan el acceso a la información y su análisis para mejorar y optimizar el proceso de toma de decisiones en los negocios. (Gartner, Business Intelligence (BI), 2015)

Con el uso de tecnologías y las metodologías de Business Intelligence se pretende convertir datos en información, y a partir de la información ser capaces de descubrir conocimiento.

BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un EDW), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones.

El proceso de Business Intelligence incluye entre sus actividades la comunicación de los descubrimientos obtenidos tras el análisis, y efectuar los cambios necesarios tras conocer dichos descubrimientos (*insights* es el término utilizado en inglés para los descubrimientos).

9.3 Enterprise Data Warehouse (EDW)

El Enterprise Data Warehouse (EDW) actualmente es el componente estándar en las arquitecturas de datos empresariales, ya que proporciona una visión del valor de negocio y un poderoso análisis de datos para la toma de decisiones tanto de operarios, ejecutivos, directores, científicos de datos y desarrolladores de software. El EDW es parte del modelo de Business Intelligence (BI), el cual proporciona un conjunto de métodos y técnicas que las organizaciones usan para la toma de decisiones tácticas y estratégicas. A lo largo del tiempo podemos destacar una serie de hitos en la evolución del BI que nos llevan hasta la situación actual:

- A finales de los 1960's el inglés Edgar Frank Codd, mientras trabajaba en IBM, inventa el Modelo Relacional, el cual dio lugar a las Bases de Datos Relacionales.
- En los años 1970's se desarrollan las primeras bases de datos y las primeras aplicaciones empresariales ERP (Enterprise Resource Planning) y CRM (Customer Relationship Management) tales como SAP, JD Edwards, Siebel y PeopleSoft. Estas aplicaciones permitieron aumentar la información disponible en los sistemas, pero no fueron capaces de ofrecer un acceso rápido y fácil a dicha información.
- Ya en la década de los 1980's aparece el concepto Data Warehouse de la mano de Ralph Kimball y Bill Inmon, y la aparición de los primeros sistemas de *reporting*. Ya había potentes sistemas de bases de datos, pero aun no existían aplicaciones que facilitasen su explotación.
- Por 1989 aparece el término de *Business Intelligence* que se le atribuye a Howard Dresner, por entonces un analista de Gartner Group.
- En los 1990s aparece Business Intelligence 1.0. que es el inicio para la creación del conocimiento y es el comienzo a través de herramientas del apoyo a la toma de decisiones de las organizaciones. Está dirigido a expertos del negocio y del área de IT (Information Technology).
- En los 2000s surge el *Business Intelligence 2.0*. En esta versión del modelo se consolidan las aplicaciones BI en unas pocas plataformas (Oracle, SAP, IBM, Microsoft). Aparte de la información estructurada, se empieza a considerar otro tipo de información y documentos no estructurados. Es más amigable que la versión 1.0 y aprovechando la Web 2.0 ya no solo está dirigido solamente a expertos sino también a usuarios finales. (Microsoft, 2009)

El Enterprise Data Warehouse (EDW) se ha convertido en la base para la toma de decisiones dentro del mundo de los negocios y las soluciones EDW actuales son maduras y extremadamente efectivas tanto para el *reporting* como para el análisis de datos.

Un EDW es una arquitectura de almacenamiento diseñada para almacenar datos extraídos de los sistemas transaccionales (ERP, CRM, entre otros), bases de datos operacionales y fuentes externas. El EDW entonces combina estos datos realizando agregación y sumarización para que queden de una forma adecuada para el análisis de datos de toda la organización y los reportes predefinidos para las necesidades del negocio.

Un Data Warehouse contiene datos organizados en áreas temáticas con versiones variables en el tiempo de un mismo registro, con el apropiado nivel de granularidad o detalle para que resulte útil para diferentes tipos de análisis. (Gartner, Data Warehouse, 2015)

9.4 Datamart

Un Datamart es una versión reducida de un Data Warehouse, especializada en el almacenamiento de los datos de un área de negocio específico de un departamento. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento.

9.5 Ciclo de vida para proyectos de BI

Para el desarrollo de aplicaciones de BI existe una metodología propuesta por (Larissa T. Moss, 2003) la cual cuenta con las 6 etapas comunes a todos los proyectos de ingeniería de software:

1. Justificación
2. Planificación
3. Análisis de negocio
4. Diseño
5. Construcción
6. Despliegue

A continuación, se describen las actividades de cada fase:

1. Justificación.
    a. **Business Case Assessment**. Aquí evaluaremos un caso de uso, para lo cual inicialmente se realizará una búsqueda en empresas locales donde se detecte una necesidad u oportunidad de mejora basándonos en el objeto de estudio del presente trabajo. En su defecto se trabajará con un cliente virtual buscando un caso de uso apropiado.
2. Planificación.
    a. **Enterprise Infrastructure Evaluation**. Aquí se analizará la infraestructura con la que cuenta la empresa donde se realizará el trabajo, tanto de hardware como de software.
    b. **Project Planning**. Se realiza una planificación de tiempos y recursos para lograr implementar el proyecto.

3. Análisis de negocio.
    a. **Project Requirements Definition**. Se relevarán y especificarán los requerimientos de datos del cliente.
    b. **Data Analysis**. Se analizarán los datos con los que actualmente cuenta el cliente, como así también la necesidad de incorporar nuevas fuentes de datos al Enterprise Data Warehouse.
    c. **Application Prototyping**. En esta etapa se construirá un prototipo de la solución propuesta que permita realizar un mejor análisis y especificación de requerimientos.
    d. **MDR Analysis**. Aquí se analizará el repositorio (base de datos) de la meta-data necesaria para la solución, es decir toda aquella información necesaria para poder consultar los datos y realizar el mantenimiento de los mismos.
4. Diseño.
    a. **Database Design**. Diseño de la base de datos necesaria para la solución propuesta
    b. **ETL Design**. Diseño del sistema de extracción, transformación y carga de datos en el EDW.
    c. **MDR Design**. Diseño del repositorio de meta-data.
5. Construcción.
    a. **ETL Development**. Desarrollo del sistema de extracción, transformación y carga de datos en el EDW.
    b. **Application Development**. Desarrollo de aplicaciones que permitan analizar y visualizar información.

c. **Data Mining**. Desarrollo de aplicación de Data Mining en caso de que el problema lo amerite.
d. **MDR Development**. Construcción del repositorio de meta-data.

6. Despliegue.
    a. **Implementation**. Aquí desplegamos todos los desarrollos finalizados quedando disponibles para la evaluación del cliente.
    b. **Release Evaluation**. En este punto se evaluará si la entrega es correcta y se ajusta a lo esperado por el cliente.

Abreviaciones: ETL (Extract/Transform/Load); MDR (Meta Data Repository).

9.6 Data Lake

Data lake es el concepto o enfoque que nos lleva a introducir las tecnologías Big Data, con la cual además de cubrir los requerimientos de datos antes mencionados le permitirá a las pequeñas y medianas empresas estar preparadas para la transformación digital e ir incorporando de manera progresiva herramientas de análisis avanzados de datos.

Según James Dixon, fundador y director de tecnología de Pentaho (suite de herramientas open source para BI) a quien se le atribuye el término de Data Lake, “si pensamos un Datamart como un recipiente de agua embotellada – depurada, envasada y lista para consumir – un Data Lake es como un gran lago o estanque natural de agua. El contenido del Data Lake proviene de una fuente natural que llena el lago, y varios consumidores pueden venir a examinar, bucear o tomar muestras de este contenido.”

Un *Data Lake*, o *Business Data Lake*, al igual que el EDW es un repositorio de almacenamiento, pero que contiene una gran cantidad de datos en bruto en su formato original, incluyendo datos estructurados, semi-estructurados y no estructurados. La estructura y los requisitos de datos no requieren ser definidos sino hasta que se necesitan dichos datos. (Dull, Big Data Cheat Sheet on Hadoop, 2015).

A continuación, se muestra una tabla que ayuda a definir el concepto de Data Lake, resaltando las diferencias claves con el EDW y explicándolas más abajo:

Imagen 5 – Data Warehouse vs Data Lake

DATA WAREHOUSE	**VS.**	**DATA LAKE**
Estructurados, procesados	**DATOS**	Estructurados/semi-estructurados/no-estructurados, sin procesar
Schema-on-write	**PROCESAMIENTO**	*Schema-on-read*
Caro para grandes volúmenes de datos	**ALMACENAMIENTO**	Diseñado para almacenamiento de bajo costo
Menos ágil, configuración fija	**AGILIDAD**	Altamente ágil, se configure y reconfigure cuando se necesita
Maduro	**SEGURIDAD**	Madurando
Profesionales de negocio	**USUARIOS**	Científicos de datos

Fuente: (Dull, Big Data Cheat Sheet on Hadoop, 2015)

- **Datos**. Un EDW solo almacena datos que han sido modelados, mientras que un Data Lake almacena todo tipo de datos: estructurados, structured, semi-estructurados y and no-estructurados.

- **Procesamiento**. Para poder cargar datos en EDW primero se necesita darles forma o modelarlos. Esto es lo que se llama *schema-on-write*. Mientras que con Data Lake se guardan los datos en su forma original, y se les da la forma y estructura cuando se necesita usarlos. Esto se llama *schema-on-read*. Por lo tanto, tenemos dos enfoques diferentes.
- **Almacenamiento**. Para Data Lake existen tecnologías open source que utilizan hardware *commodity* o de uso masivo para el almacenamiento en forma de clusters.
- **Agilidad**. El EDW es un repositorio altamente estructurado, por lo que cambiar la estructura de los datos ante un requerimiento del negocio se vuelve un proceso muy lento. En contraste el Data Lake carece de la estructura del EDW, lo que le permite configurar y reconfigurar los modelos más fácilmente.
- **Seguridad**. Aquí EDW lleva la ventaja debido a los años de madurez de esta tecnología.
- **Usuarios**. Aquí también por el distinto grado de madurez de las dos tecnologías en EDW es más fácil de usar, mientras que Data Lake de momento requiere de usuarios técnicos.

Tal como se mencionó previamente la tecnología hoy disponible que puede soportar el nuevo enfoque de Data Lake es Big Data (datos masivos) con proyectos como el de Apache Hadoop, que desarrolla software de código abierto para la computación escalable, fiable, y distribuida y actualmente es la plataforma más utilizada en proyectos de Big Data.

9.7 Big Data

Anteriormente se mencionó en el presente trabajo que la tecnología de Big Data permite almacenar todo tipo de datos de una forma más barata y también procesar estos datos más rápido. Cuando se habla de todo tipo de datos nos referimos a datos estructurados, semi-estructurados y no-estructurados.

Si bien todos estos tipos de datos siempre existieron, el problema es que el EDW está diseñado para manejar datos estructurados en bases de datos relacionales, y no es bueno para el manejo de los otros dos tipos de datos.

A continuación, vemos una tabla con ejemplos de los distintos tipos de datos. Los recuadros naranjas representan a los datos estructurados, mientras que el resto (en azul) son los que las nuevas tecnologías Big Data permiten almacenar y procesar de una manera más rápida y barata:

Imagen 6 – Tipos de datos

POS DATA	CRM	FINANCIAL DATA	LOYALTY CARD DATA	TROBLE TICKETS
EMAIL	PDF FILES	SPREAD-SHEETS	WORD PROCESSING DOCUMENTS	RFID TAGS
GPS	WEB LOG DATA	PHOTOS	SATELLITE IMAGES	SOCIAL MEDIA DATA
BLOGS	FORUMS	CLICK-STREAM DATA	VIDEOS	XML DATA
MOBILE DATA	WEBSITE CONTENT	RSS FEEDS	AUDIO FILES	CALL CENTER TRANSCRIPTS

Fuente: (Tamara Dull, SAS, 2016)

Como vemos el 20% de los datos con los que trabajamos diariamente son estructurados y el 80% restante (cuadros azules) son datos semi-estructurados y no estructurados.

Hoy contamos con tecnologías para colectar, procesar, almacenar y analizar todos estos tipos de datos juntos. Con Hadoop podemos mezclar y vincular los datos que vemos en los cuadros naranjas con los azules necesitando solo una fracción del costo y tiempo del que requieren los sistemas tradicionales de bases de datos relacionales, y a esto lo podemos integrar en nuestro EDW.

En cuanto a la pregunta de ¿qué es Big Data?, por tratarse de un fenómeno relativamente reciente no existe una única definición al respecto, pero si hay ciertas coincidencias entre la comunidad académica, técnicos y profesionales. En la publicación (McKinsey Global Institute, 2011) encontramos una primera definición dinámica, ya que depende de las limitaciones tecnológicas del momento: “Big Data refiere a un conjunto de datos cuyo tamaño está más allá de la capacidad que tienen el software de base de datos para capturar, almacenar, administrar y analizar”. Por lo tanto, a medida que se va corriendo la barrera tecnológica también lo harán los tamaños que se consideran Big Data. Hablamos de Big Data cuando el tamaño se vuelve el principal problema.

Además del volumen existen otros puntos en común entre los especialistas a la hora de definir Big Data que incorporan más dimensiones al análisis. Se refieren a las características de la variedad y la velocidad de la información, incluso dándole mayor importancia a estas últimas que al volumen, conformando así las tres **V**’s que suelen utilizarse para definir el concepto de Big Data (**V**olumen, **V**ariedad y **V**elocidad). (Gartner, Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, 2011).

Con variedad se suele hacer referencia a la heterogeneidad de las fuentes y estructuras de datos (no estructurados), los cuales por lo general no están listos para ser integrados a una aplicación; mientras que con velocidad se hace referencia tanto a la frecuencia de llegada de los datos como al tiempo de respuesta o *feedback* para que estos, los datos, sean útiles en la toma de decisiones.

Algunas definiciones incluyen una cuarta **V** que remite a **V**alor, variabilidad o virtual (Armah, 2013). Según la publicación (White paper, 2012) faltaría incluir las características de privacidad de las personas y facilidad de uso (usabilidad) para tener una definición más completa del concepto.

Para poder obtener valor de este volumen creciente de datos necesitamos poder procesarlos más rápidamente, para lo cual en un esquema tradicional de datos se necesitan equipos cada vez más potentes (Ej. más CPU y RAM), lo cual requiere una continua inversión en hardware con el consiguiente esfuerzo económico que esto conlleva para las pequeñas y medianas organizaciones.

Sin embargo, esta solución tradicional ha encontrado un importante cuello de botella: el acceso a disco duro. Es decir que, aunque seamos capaces de procesar datos cada vez más rápido la velocidad de acceso a disco no mejora lo suficiente.

Es mediante el procesamiento distribuido como se viene a resolver este problema, a través de arquitecturas en clúster en vez de arquitecturas monolíticas, en donde la responsabilidad del procesado no depende de una única máquina sino de los nodos que forman el clúster, los cuales constituyen un sistema más fácilmente escalable para ganar capacidad de procesamiento y no requiere de servidores de última generación.

Por lo tanto, las soluciones de Big Data existentes en el mercado, como Apache Hadoop, tienen en común que utilizan arquitecturas distribuidas a través de clusters de *hardward commodity* (o de uso masivo) en lugar de costosos servidores.

9.8 Hadoop

Apache Hadoop es una plataforma de software de código abierto para el almacenamiento y procesamiento distribuido de datos masivos en clústeres de computadoras construidos a partir de hardware *commodity*. Los servicios de Hadoop proporcionan almacenamiento de datos, procesamiento de datos, acceso a datos, gobierno de datos, seguridad y operaciones.
En otras palabras, se trata de un *framework open source* que permite escribir aplicaciones robustas, eficientes y distribuidas que necesitan procesar una gran cantidad de datos. Está escrito en Java y permite que las aplicaciones se ejecuten en clústeres de miles de nodos.

La génesis de Hadoop vino del *paper* "*Google File System*" publicado en octubre de 2003. Este paper generó otro trabajo de investigación de Google - *MapReduce: "Simplified Data Processing on Large Clusters"*. El desarrollo se inició en el proyecto *Apache Nutch*, pero se trasladó al nuevo subproyecto Hadoop en enero de 2006. Hadoop 1.0 fue lanzado en abril de 2006 y sigue evolucionando gracias a los muchos contribuyentes al actual proyecto Apache Hadoop, de la Fundación Apache.
Una de las razones por la que la mayoría de las organizaciones utilizan Hadoop es por su capacidad para almacenar, administrar y analizar grandes cantidades de datos estructurados y no estructurados de forma rápida, fiable, flexible y de bajo costo:

- **Escalabilidad y rendimiento**: el procesamiento distribuido de datos locales a cada nodo en un clúster permite a Hadoop almacenar, gestionar, procesar y analizar datos a escala de *petabytes* (10^{15} bytes).

- **Confiabilidad**: los grandes clusters de computación tienden a fallar en sus nodos individuales. Hadoop es fundamentalmente resistente: cuando un nodo falla, el procesamiento es redirigido a los nodos restantes del clúster y los datos se replican automáticamente como preparación para futuras fallas de otros nodos.
- **Flexibilidad**: a diferencia de los sistemas de gestión de bases de datos relacionales tradicionales, no es necesario crear esquemas estructurados antes de almacenar datos. Se pueden almacenar datos en cualquier formato, incluidos los formatos semiestructurados o no estructurados, y luego analizar y aplicar esquema o formato a los datos cuando se leen y necesitan ser usados.
- **Bajo Costo**: a diferencia del software propietario, Hadoop es de código abierto (*open source*) y funciona con hardware de bajo costo (*commodity hardware*). (Hortonworks, 2016)

Hadoop no es un tipo de base de datos, sino más bien un ecosistema de software que permite la computación masiva en paralelo. Y es un facilitador para la incorporación de ciertos tipos de bases de datos distribuidas **NoSQL** (como HBase), las cuales permiten que los datos sean distribuidos en miles de servidores sin reducir de manera significativa el rendimiento.

Un elemento básico del ecosistema Hadoop es MapReduce, un modelo informático que básicamente toma procesos de datos intensivos y extiende el procesamiento a través de un número potencialmente infinito de servidores (generalmente referido como un clúster Hadoop). Este ha significado un cambio de paradigma en el soporte a las necesidades de procesamiento de Big Data; un procedimiento de datos masivos que podría tomar hasta 20 horas de tiempo de procesamiento en un sistema de base de datos relacional centralizado, solo podría tomar 3 minutos cuando se distribuye a través de un clúster de servidores *commodity* Hadoop, que procesan todo en paralelo.

9.9 NoSQL

Más arriba se hacía referencia a que Hadoop es un facilitador para la incorporación de ciertos tipos de bases de datos distribuidas *NoSQL* (como HBase), por lo que se hará una explicación de este concepto y su importancia.

El sistema de gestión de base de datos relacional (RDBMS por sus siglas en inglés) ha sido el estándar de facto para la gestión de bases de datos en toda la era de Internet. La arquitectura detrás de RDBMS es tal que los datos se organizan de una manera muy estructurada, siguiendo el modelo relacional.

Sin embargo, RDBMS se considera ahora como una tecnología de base de datos en declive. Mientras que la organización precisa de los datos mantiene el EDW muy ordenado, la necesidad de que los datos sean bien estructurados en realidad se convierte en una carga sustancial con volúmenes de datos extremadamente grandes, lo que resulta en la disminución de rendimiento a medida que el tamaño se hace más grande.

Por lo tanto, RDBMS generalmente no se considera como una solución escalable para satisfacer las necesidades de Big Data.

La eficiencia de NoSQL se puede lograr porque a diferencia de las bases de datos relacionales que son altamente estructuradas, las bases de datos NoSQL son no estructuradas por naturaleza, sacrificando requisitos de estricta consistencia en pos de la agilidad y velocidad. NoSQL se centra en el concepto de bases de datos distribuidas, donde los datos no estructurados pueden ser almacenados en varios nodos de procesamiento, y a menudo a través de múltiples servidores. Esta arquitectura distribuida permite a las bases de datos NoSQL ser escalables horizontalmente; a medida que los datos se incrementan, sólo hay que añadir más hardware para mantener el rendimiento.

La infraestructura de base de datos distribuida NoSQL ha sido la solución para el manejo de algunos de los más grandes almacenes de datos en el planeta - por ejemplo, algunos de la talla de Google, Amazon y la CIA.

9.10 Casos de uso de Hadoop

Muchas de estas nuevas tecnologías relacionadas con Hadoop todavía están madurando. No tienen las décadas de investigación y desarrollo detrás de ellos como los sistemas relacionales existentes de EDW. Por eso los especialistas coinciden en que Hadoop y su ecosistema aún no están 100% listos para ser utilizado en las empresas. (Tamara Dull, SAS, 2016)

Si simplemente se desea usarlo como un repositorio de almacenamiento adicional (o alternativo) y / o como un procesador de datos de corto plazo, entonces definitivamente Apache Hadoop está listo para ser implementado.

Sin embargo, si se desea ir más allá del almacenamiento y procesamiento de datos y se está buscando algunas de las mismas capacidades de análisis y gestión que tiene actualmente el EDW, Apache Hadoop por sí solo no lo cubrirá.

Se necesitará obtener asistencia técnica de IT y desarrolladores, internos y externos, para explorar el vasto ecosistema de productos y proyectos open source y propietarios relacionados con Hadoop para lograr estos objetivos.

Contrariamente a la creencia popular, Hadoop no es sólo para Big Data. Es cierto que Hadoop fue originalmente desarrollado para abordar las necesidades de Big Data de las compañías web/media, pero hoy en día, se está utilizando en todo el mundo para hacer frente

a un conjunto más amplio de necesidades de datos, grandes y pequeños, por prácticamente todas las industrias.

A continuación, se describen seis casos comunes de uso de Hadoop, tres de los cuales no requieren Big Data para aprovechar al máximo esta herramienta y de aquí surgirán las propuestas para el cliente tipo del presente trabajo:

1. **Aterrizaje de los datos estructurados (Stage)**. Utilice Hadoop como plataforma de almacenamiento de datos para su EDW. ¿Qué pasa si Hadoop procesa y transforma sus datos operativos antes de cargarlos en su EDW? La ventaja es que debido al bajo costo del almacenamiento de Hadoop, usted podría almacenar ambas versiones de los datos en Hadoop: los datos crudos, nativos (raw data) y los datos transformados. Sus datos estarían ahora en un solo lugar, facilitando la administración, el procesamiento y el análisis posterior.
2. **Procesamiento de datos estructurados.** Utilice Hadoop para actualizar los datos en su EDW y / o en los sistemas operacionales. En lugar de usar recursos costosos de BI para actualizar los datos en el EDW, ¿por qué no enviar los datos necesarios a Hadoop, dejar que este los procese y luego enviar los datos actualizados de vuelta al EDW? Este caso de uso no sólo se aplica al procesamiento de los datos del EDW, sino también a los datos de cualquiera de sus sistemas operacionales o analíticos. Aproveche la capacidad de procesamiento de bajo costo de Hadoop para que sus sistemas relacionales se liberen para hacer lo que mejor saben hacer y para lo que fueron diseñados.
3. **Archivar todos los datos**. Utilice Hadoop para archivar todos sus datos de manera local o en la nube. Dado que Hadoop se ejecuta en *hardware commodity* que se puede escalar fácil y rápidamente, las organizaciones ahora pueden almacenar y archivar

muchos más datos a un costo mucho menor. Por ejemplo, ¿qué pasa si no tiene que destruir los datos después de su vida planificada para ahorrar en los costos de almacenamiento? ¿Qué pasa si usted puede mantener todos sus datos de manera fácil y rentable? O tal vez no se trata sólo de mantener los datos a mano, sino más bien, ser capaz de analizar más datos. ¿Por qué limitar su análisis a los últimos tres, cinco o siete años, cuando puede almacenar y analizar datos de décadas?

4. **Procesar cualquier dato.** Utilice Hadoop para aprovechar los datos que actualmente no están disponibles en su ecosistema de EDW. Este caso de uso se centra en dos categorías de datos: (1) fuentes de datos estructurados que no se han integrado en su almacén de datos y (2) fuentes de datos no estructuradas y semiestructuradas. En términos más generales, se trata de datos que actualmente no forman parte de su EDW y que podrían proporcionar información adicional sobre sus clientes, productos y servicios. Debido a que Hadoop puede almacenar y procesar cualquier dato, puede encargarse de los datos que su EDW no puede o no manejar bien.
5. **Acceda a cualquier dato (a través del EDW)**. Utilice Hadoop para ampliar su EDW y mantenerlo en el centro del universo de datos de su organización. Este caso de uso
6. está dirigido a empresas que quieren mantener el EDW como el sistema de registro de facto, al menos por ahora. Como componente complementario, Hadoop puede utilizarse para procesar e integrar cualquier tipo de estructura de datos, semiestructurada y no estructurada, y cargar lo que se necesita en el EDW. Esto permite a las empresas continuar utilizando sus herramientas actuales de BI / analytics con su ecosistema de EDW.
7. **Acceda a cualquier dato (vía Hadoop).** Utilice Hadoop como plataforma de aterrizaje (stage) para todos los datos y aproveche las fortalezas tanto del EDW como de Hadoop. Como se mencionó anteriormente, una ventaja de la captura de datos en

> Hadoop es que se puede almacenar en su estado crudo, nativo (raw data). No necesita ser formateado de antemano como con los datos tradicionales estructurados; Se puede formatear en el momento de la consulta de datos, cuando estos son requeridos. Este caso de uso es el que más apoya el concepto de usar Hadoop como un "data lake".

Por lo tanto, no hay que cometer el error de creer que Hadoop es sinónimo de Big Data, porque no lo es. Es sin dudas una de las más populares tecnologías de Big Data actuales que se pueden utilizar incluso si no se tiene Big Data, como puede verse en los primeros tres casos de uso anteriores. (Dull, A Non-Geek's Big Data Playbook, 2016)

9.11 Advanced analytics

Advanced Analytics o Analítica avanzada es el examen autónomo o semiautónomo de datos o contenidos utilizando sofisticadas técnicas y herramientas, típicamente más allá de las de la inteligencia empresarial tradicional (BI), para descubrir ideas más profundas, hacer predicciones o generar recomendaciones. Las técnicas de Advanced Analytics incluyen *data/text mining (minería de datos), machine learning* (aprendizaje automático), *pattern matching* (correlación de patrones), *forecasting (*pronóstico*)*, visualización, semantic y sentiment analysis (se trata de análisis en las redes sociales de cómo sus clientes ven su empresa y los motivos), análisis de redes y clusters, multivariate statistics (estadísticas multivariantes), análisis de gráficos, simulación, procesamiento de eventos complejos y redes neuronales.

Si bien no está en el alcance del presente trabajo aplicar o profundizar en Analítica avanzada sin dudas es hacia donde apunta toda línea de trabajo futuro ya que se trata de la meta a alcanzar una vez familiarizados con las tecnologías de Big Data, cuando el cliente objetivo de este trabajo cuente con los recursos humanos con las habilidades y competencias

correspondientes, como científicos de datos (Data Scientist), ya que es a partir de los resultados de estas técnicas avanzadas de análisis que las organizaciones pioneras están obteniendo valores agregados, y será la manera de diferenciarse y subsistir en la era de la transformación Digital, tal como lo reflejan reportes y encuestas que se compartirán más adelante en este informe.

9.12 Data Scientist

El rol de *data scientist* (científico de datos) es fundamental para las organizaciones que buscan extraer conocimiento de los activos de información para iniciativas de Big Data y requiere una amplia combinación de habilidades que se pueden cumplir mejor en equipo. Combina conocimientos de matemática, estadísticas y programación, y se lo puede utilizar para mejorar la eficiencia de cualquier negocio o actividad estatal. La colaboración y el trabajo en equipo son necesarios en este rol para interactuar con las partes interesadas en una organización y entender los asuntos del negocio de una empresa en particular.

Las habilidades analíticas y de modelado de decisiones que la función del *data scientist* son necesarias para descubrir relaciones dentro de los datos y para detectar patrones. También se requieren habilidades de gestión de datos para extraer o construir el subconjunto de datos relevantes utilizado para el análisis.
Ya en 2012 la publicación Harvard Business Review catalogó al rol de Data Scientist como el trabajo más atractivo del siglo XXI, aunque actualmente existe un gran déficit de estos profesionales. (HBR, 2012)

9.13 T.I.C (Tecnología de la Información y Comunicación)

A modo de referencia a continuación se presenta un cuadro con algunas las soluciones existentes en el mercado que implementan tecnología Big Data con plataforma Hadoop:

Imagen 7 – Soluciones con plataforma Hadoop

Solución	Empresa
ElasticMap Reduce	Amazon
Big Data Analytics	Teradata
HDInsight	Microsoft
InfoSphere Big Insights	IBM
HortonWorks Data Plattform	HortonWorks. Distribución con versión community*
Cloudera Impala	Cloudera. Distribución con versión community*
MapRtechnologies	MapR. Distribución con versión community*

Fuente: Recopilación de los sitios web de las distintas empresas mencionadas en el cuadro.

* Las **distribuciones Hadoop** son paquetes que combinan en una única instalación la configuración de un clúster Hadoop con los productos del ecosistema más populares, y además algunas ofrecen una versión gratuita (*community*).

Cabe mencionar que como alternativa a la plataforma Hadoop también existen soluciones con arquitectura de base de datos NoSQL como CouchBase, MongoDB, MongoDB o Amazon DynamoDB.

Para la implementación de los casos de uso en este trabajo se optó por el uso de la plataforma Hadoop, ya que es el más utilizado y por el amplio abanico de herramientas con el que cuenta actualmente su ecosistema. En particular se opta por la distribución de *HortonWorks*, por su

fácil instalación y ya que cuenta con una versión gratuita para proyectos piloto, lo cual será un aliciente para el cliente objetivo a quien se dirige el presente trabajo.

Por lo tanto, las tecnologías utilizadas serían:

-Big Data

-VMWare (virtualización)

-HortonWorks Data Plattform

-HDFS (Sistema de archivos distribuidos Hadoop)

-Herramientas de visualización de BI

-SQL y lenguaje de scripting para ETL

-Aplicaciones en java, javascript y css basadas en web browser y móviles para análisis y visualiación.

9.14 Ecosistema Hadoop

Hadoop es potente porque es extensible y es fácil de integrar con cualquier componente. Apache Hadoop no es en realidad un solo producto sino una colección de varios componentes. Cuando todos estos componentes se fusionan, hacen que Hadoop sea muy fácil de usar.

A continuación, se hace una breve introducción a los principales componentes del ecosistema.

MapReduce

MapReduce es un framework de software para escribir fácilmente aplicaciones que procesan grandes cantidades de datos en paralelo en grandes clusters de commodity hardware de una manera fiable y tolerante a fallos. En términos de programación, hay dos funciones que son más comunes en MapReduce.

- **La función Map:** La computadora o nodo maestro toma la entrada y la convierte o divide en partes más pequeñas, y la distribuye en otros nodos de trabajo. Todos los nodos de trabajo solucionan su propio problema pequeño y dan respuesta al nodo maestro.
- **La función Reduce:** El nodo maestro combina todas las respuestas procedentes de los nodos de trabajo y lo convierte en alguna forma de salida, que es la respuesta de nuestro gran problema distribuido.

Generalmente tanto la entrada como la salida están guardadas en un sistema de archivos. El framework es responsable de programar tareas, supervisarlas e incluso volver a ejecutar las tareas fallidas. (Apache, MapReduce Tutorial, 2013)

Hadoop Distributed File System (HDFS)

HDFS es un sistema de archivos distribuido que proporciona acceso de alto rendimiento a los datos. Cuando los datos se envían a HDFS, este los divide automáticamente en varios bloques y almacena / replica estos datos, lo que garantiza una alta disponibilidad y tolerancia a fallos. Un archivo consiste en muchos bloques (grandes bloques de 64MB y más).

Estos son los principales componentes de *HDFS***:**

- **NameNode:** Actúa como el nodo maestro del sistema. Mantiene el sistema de nombres, es decir, directorios y archivos y administra los bloques que están presentes en los DataNodes.
- **DataNodes:** Son los nodos esclavos que se despliegan en cada máquina y proporcionan el almacenamiento real. Son responsables de servir las peticiones de lectura y escritura para los clientes.

- **NameNode secundario:** Es responsable de realizar controles periódicos (*checkpoints*). En caso de un fallo del NameNode, se puede reiniciar utilizando el *checkpoint*. (Apache, HDFS Architecture Guide, 2013)

Hive

Hive es parte del ecosistema de Hadoop y proporciona una interfaz similar a SQL (like-SQL) para Hadoop. Es un sistema de data warehouse para Hadoop que facilita la sumarización de datos, las consultas ad-hoc y el análisis de grandes conjuntos de datos almacenados en sistemas de archivos compatibles con Hadoop.

Proporciona un mecanismo para proyectar la estructura en estos datos y consultarlos usando un lenguaje similar al SQL llamado HiveQL. Hive también permite que los programadores tradicionales de map/reduce conecten sus *mappers* y *reducers* personalizados cuando sea inconveniente o ineficiente expresar esta lógica en HiveQL.

Los principales bloques constitutivos de Hive son:

1. **Metastore** – Para almacenar metadatos sobre columnas, particiones y catálogo de sistemas.
2. **Driver** – Para administrar el ciclo de vida de una sentencia HiveQL
3. **Query Compiler** – Para compilar HiveQL en un grafo acíclico dirigido.
4. **Execution Engine** – Para ejecutar las tareas en el orden correcto que son producidas por el compilador.
5. **HiveServer** – Para proporcionar una interfaz Thrift y un servidor JDBC / ODBC. (Apache, APACHE HIVE TM)

HBase (Hadoop DataBase)

HBase es una base de datos distribuida, orientada a columnas y utiliza HDFS para el almacenamiento. HDFS trabaja con el patrón de "en escribir una vez y leer muchas veces", pero esto no siempre se da de esta forma. También podemos requerir un acceso aleatorio de lectura / escritura en tiempo real de grandes conjuntos de datos, y es aquí donde HBase entra en acción. HBase se basa en HDFS y se distribuye en una base de datos orientada a columnas.

Estos son los principales componentes de *HBase***:**

- **HBase Master:** es el responsable de negociar el equilibrio de carga (load balancing) en todos los RegionServers y mantiene el estado del clúster. No forma parte de la ruta de recuperación o almacenamiento de datos reales.
- **RegionServer:** Se despliega en cada máquina y recibe datos y procesa solicitudes de entrada/salida. (Apache, Welcome to Apache HBase)

Zookeeper

ZooKeeper es un servicio centralizado para mantener información de configuración, nombrar, proporcionar sincronización distribuida y proporcionar servicios de grupo que son muy útiles para una variedad de sistemas distribuidos. HBase no puede funcionar sin ZooKeeper. (Apache, Welcome to Apache ZooKeeper)

Mahout

es una biblioteca escalable de *machine learning* (aprendizaje automático, una técnica de analítica avanzada) que implementa diferentes enfoques de aprendizaje automático. En la actualidad Mahout contiene cuatro grupos principales de algoritmos:

- *Recommendations*, también conocidos como *Collective filtering*
- *Classifications*, también conocidos como *Categorization*
- *Clustering*
- *Frequent itemset mining*, también conocidos como *parallel frequent pattern mining*.

Los algoritmos en la biblioteca de Mahout pertenecen al subconjunto que se puede ejecutar de una manera distribuida y se han escrito para ser ejecutables en MapReduce. (Apache, What is Apache Mahout?)

Sqoop (SQL-to-Hadoop)

Sqoop es una herramienta diseñada para transferir de forma eficiente datos estructurados de SQL Server y SQL Azure a HDFS y luego usarlos en trabajos MapReduce y Hive. Incluso se puede usar para mover datos de HDFS a SQL Server. (Apache, Apache Sqoop)

Apache Spark

Apache Spark es un motor de cálculo general que ofrece un rápido análisis de datos a gran escala. Spark se basa en HDFS, pero elimina MapReduce y en su lugar utiliza su propio marco de procesamiento de datos. Los casos de uso común para Apache Spark incluyen consultas en tiempo real, procesamiento de secuencias de eventos, algoritmos iterativos, operaciones complejas y *machine learning* (aprendizaje automático).

Pig

Pig es una plataforma para analizar y consultar grandes conjuntos de datos que consisten en un lenguaje de alto nivel para expresar los programas de análisis de datos, junto con la infraestructura para evaluar estos programas. Las operaciones integradas de Pig pueden dar sentido a los datos semi-estructurados, como los archivos de log, y el lenguaje es extensible utilizando Java para agregar soporte para tipos de datos personalizados y transformaciones.

Pig tiene tres propiedades clave principales:

- Permite extensibilidad.
- Optimización de recursos.
- Es fácil de programar.

La propiedad más destacada de los programas de Pig es que su estructura es susceptible de una gran paralelización, lo que a su vez le permite manejar conjuntos de datos muy grandes. En la actualidad, la capa de infraestructura de Pig se compone de un compilador que produce secuencias de programas MapReduce. (Apache, Welcome to Apache Pig!, 2016)

Flume

Flume es un framework para recolectar, agregar y mover enormes cantidades de datos de archivos de log o de texto dentro y fuera de Hadoop. Los agentes se encuentran en toda la infraestructura de IT dentro de servidores web, servidores de aplicaciones y dispositivos móviles. Flume tiene un motor de procesamiento de consultas, por lo que es fácil transformar cada nuevo lote de datos antes de que se envíe al receptor previsto. (Apache, Welcome to Apache Flume)

Ambari

Ambari fue creado para ayudar a administrar Hadoop. Ofrece soporte para muchas de las herramientas del ecosistema de Hadoop, incluyendo Hive, HBase, Pig, Sqoop y Zookeeper. La herramienta cuenta con un panel de control de administración que realiza un seguimiento de la salud del clúster y puede ayudar a diagnosticar problemas de rendimiento. (Apache, Ambari)

Visualización

QlikView es una herramienta de visualización de la compañía Qlik para realizar análisis sobre los datos del EDW. Gestiona las asociaciones entre los conjuntos de datos a nivel de máquina, no a nivel de aplicación, almacenando tablas individuales en su motor asociativo, en memoria.

Cada dato de cada campo del conjunto analítico de datos está asociado a todos los demás datos del conjunto total de datos. Por conjuntos de datos entendemos cientos de tablas, con miles de campos. (Qlik, 2017)

9.15 Competencia

Spark vs. Hadoop?

Al igual que Hadoop, Spark es un framework open source de herramientas Big Data de la Fundación Apache que surge para superar las limitaciones del componente MapReduce de Hadoop. Se trata de un motor de procesamiento de datos hasta 100 veces más rápido que Hadoop y en los últimos años se ha generado un debate en los círculos de gestión de datos en relación con Spark vs. Hadoop.

Si bien es posible que Spark se convierta en el reemplazante natural de Hadoop, dada la superioridad en cuanto a potencia para analítica avanzada en tiempo real, de momento también pueden ser herramientas complementarias, ya que Spark no cuenta con un sistema de almacenamiento distribuido como Hadoop, por lo que pueden usarse juntos.

Dado que el objetivo del presente trabajo no es dar soluciones de analítica avanzada se descarta a Spark como una solución superadora o que pueda ser considerada competencia en el cliente objetivo al que se dirige este trabajo.

10. Diseño Metodológico

10.1 Relevamiento o recolección de información

El presente proyecto utiliza como instrumentos para el relevamiento, investigaciones de terceros basadas en entrevistas y encuestas que poseen relación directa a la temática estudiada. En particular datos surgidos de encuestas realizadas este año en Argentina y EEUU con respecto al estado de la transformación digital y Big Data en las empresas, los cuales están detallados en el apartado de Transformación digital dentro del marco teórico.

10.2 Metodología de desarrollo de software

Se utilizará un ciclo de vida siguiendo la guía *Business Intelligence Roadmap* descripto en el Marco de Referencia de la autora Larisa Moss, es decir las 6 etapas y 16 pasos descriptos previamente en este informe.

Se elige este ciclo de vida iterativo e incremental frente a los modelos de desarrollo tradicionales en cascada y con despliegue big-bang ya que las aplicaciones de BI requieren de continuas mejoras basadas en el feedback de la comunidad del negocio. Las soluciones para la toma de decisiones son aplicaciones que atraviesan todas las áreas de las organizaciones, por lo tanto, se requiere integración de todos sus sistemas, y para ello un ciclo de vida como el seleccionado permite un despliegue más controlado contribuyendo a una mejor calidad del producto debido a que se realizan pequeñas integraciones continuas con impactos más reducidos.

10.3 Herramientas a utilizar

Para la *etapa de Justificación* se utilizará como herramienta principal la investigación del panorama actual de las PYMES con respecto a la Transformación digital, y a partir de allí se caracterizará un cliente tipo y sus necesidades al cual se le propondrán las mejoras mediante el presente trabajo.

Para la *etapa de Planificación* se utilizará la herramienta de Microsoft Project 2013.

En la *etapa de Análisis de negocio* se hará un análisis pormenorizado del caso de uso para ganar un entendimiento sólido de los requerimientos de negocio para una solución potencial. Se construirá un prototipo inicial para un caso de uso utilizando la arquitectura que integre BI con Big Data y que le permita al cliente un primer acercamiento a la solución propuesta.

En la *etapa de Diseño* se concebirá el producto que resuelve el problema de negocio. Para ello se utilizará el lenguaje UML 2.0 que permitirá generar los diagramas necesarios que describen la arquitectura requerida. Para el modelado de los procesos de negocio utilizaremos diagramas de flujo desarrollándolos en la Microsoft Visio 2013.

En la *etapa de Construcción* se evaluará la conveniencia de utilizar una plataforma open source en lugar de una propietaria, que nos permita la incorporación de tecnologías de Big Data a la arquitectura de BI del cliente. También se evaluará la posibilidad de utilizar una distribución que facilite la construcción rápida de un prototipo.

Para la parte de EDW utilizaremos SQL Server de Microsoft, la versión Express, ya que es gratuita y a su vez es el motor de bases de datos más difundido tanto para las bases de datos transaccionales del negocio como para el Data Warehouse y Datamarts.

Para la visualización de la información utilizaremos la herramienta QlikView, por tratarse de una de las aplicaciones más utilizadas y que cuenta con una versión gratuita. Además de ser una aplicación para BI proporciona integración con distintas fuentes de datos como las de Big Data.

En la etapa de Despliegue pondremos a funcionar toda la solución y se realizarán pruebas para comprobar que se cumplió con los requerimientos del cliente. Para ello se desarrollará un plan de pruebas en la herramienta online Trello (https://trello.com/), donde también se especificarán todas las tareas del proyecto. Se elige esta herramienta ya que es gratuita y muy utilizada en proyectos agiles y ciclos de vida iterativos.

10.4 Diagrama de Gantt

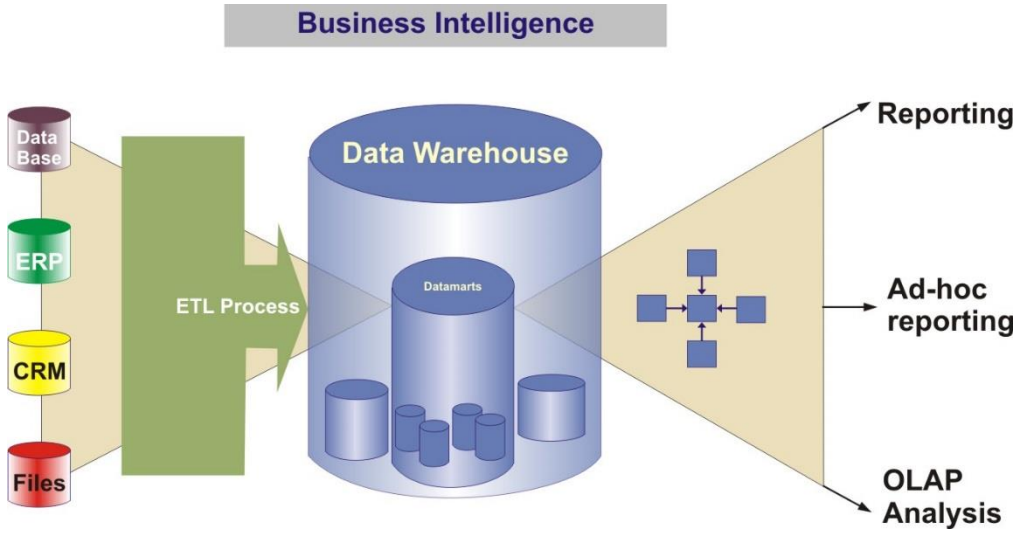
Imagen 8 - Gantt

Task Mode	Task Name	Duration	Start	Finish	Predecessors
	Analizar las alternativas Open Source de tecnologías BIG DATA	**8 days**	**Mon 1/2/17**	**Wed 1/11/17**	
	Buscar al menos dos alternativas del mercado	5 days	Mon 1/2/17	Fri 1/6/17	
	Seleccionar la opción mas adecuada	3 days	Mon 1/9/17	Wed 1/11/17	3
	Analizar y seleccionar herramientas de BI para visualizar	**8 days**	**Thu 1/12/17**	**Mon 1/23/17**	
	Buscar al menos dos alternativas del mercado	5 days	Thu 1/12/17	Wed 1/18/17	4
	Seleccionar la opción mas adecuada	3 days	Thu 1/19/17	Mon 1/23/17	6
	Analizar y rediseñar arquitectura BI	**14 days**	**Tue 1/24/17**	**Fri 2/10/17**	
	Presentar una arquitectura clásica de BI	4 days	Tue 1/24/17	Fri 1/27/17	7
	Analizar qué elementos pueden ser reemplazados por BIG DATA	10 days	Mon 1/30/17	Fri 2/10/17	9
	Implementar la arquitectura diseñada	**21 days**	**Mon 2/13/17**	**Mon 3/13/17**	
	Instalar plataforma BIG DATA en VMware	11 days	Mon 2/13/17	Mon 2/27/17	10
	Integrar Plataforma BIG DATA con BI	10 days	Tue 2/28/17	Mon 3/13/17	12
	Generar datos de entrada	**10 days**	**Tue 3/14/17**	**Mon 3/27/17**	
	Automatizar fuentes de datos de entrada	10 days	Tue 3/14/17	Mon 3/27/17	13
	Integrar y analizar los datos existentes	**28 days**	**Tue 3/28/17**	**Thu 5/4/17**	
	Almacenar en staging datos de las fuentes	7 days	Tue 3/28/17	Wed 4/5/17	15
	Extraer datos de las respectivas fuentes	7 days	Thu 4/6/17	Fri 4/14/17	17
	Transformar los datos para análisis	7 days	Mon 4/17/17	Tue 4/25/17	18
	Cargar los datos en EDW	7 days	Wed 4/26/17	Thu 5/4/17	19
	Visualizar la información obtenida	**32 days**	**Fri 5/5/17**	**Mon 6/19/17**	
	Conectar las herramientas para visualizar	7 days	Fri 5/5/17	Mon 5/15/17	20
	Personalizar las visualizaciones en función del usuario	15 days	Tue 5/16/17	Mon 6/5/17	22
	Demostrar integración de los datos de EDW y BIG DATA	10 days	Tue 6/6/17	Mon 6/19/17	23

10.5 Cliente objetivo

El cliente tipo o cliente objetivo para el cual se realizará el proyecto es una pequeña o mediana empresa que ya cuenta con una arquitectura tradicional de BI en funcionamiento con un EDW que almacena datos provenientes de las bases de datos de sus aplicaciones transaccionales, y cuenta con algunas aplicaciones para la visualización de datos, análisis y *reporting (QlikView)*, que cumplen la función de dar soporte a la toma de decisiones, pero que aún no maneja Big Data ni cuenta con recursos económicos y humanos como para migrar definitivamente a esta tecnología, corriendo el riesgo de volverse poco competitivo en los próximos años.

Imagen 9 – Enterprise Data Warehouse



Fuente imagen: (EDW, 2017)

Como podemos observar en el gráfico anterior se representan los sistemas transaccionales a la izquierda, de cuyas bases de datos se extraen los datos necesarios según los requerimientos, se limpian (transforman) y se cargan en el EDW o en Datamarts, y desde allí otras aplicaciones consumen los datos cargados para realizar análisis o reportes.

10.6 Relevamiento

Como el cliente de este Proyecto es un cliente tipo o hipotético se relevan las características tradicionales del EDW que nos permiten más adelante identificar oportunidades de mejora.

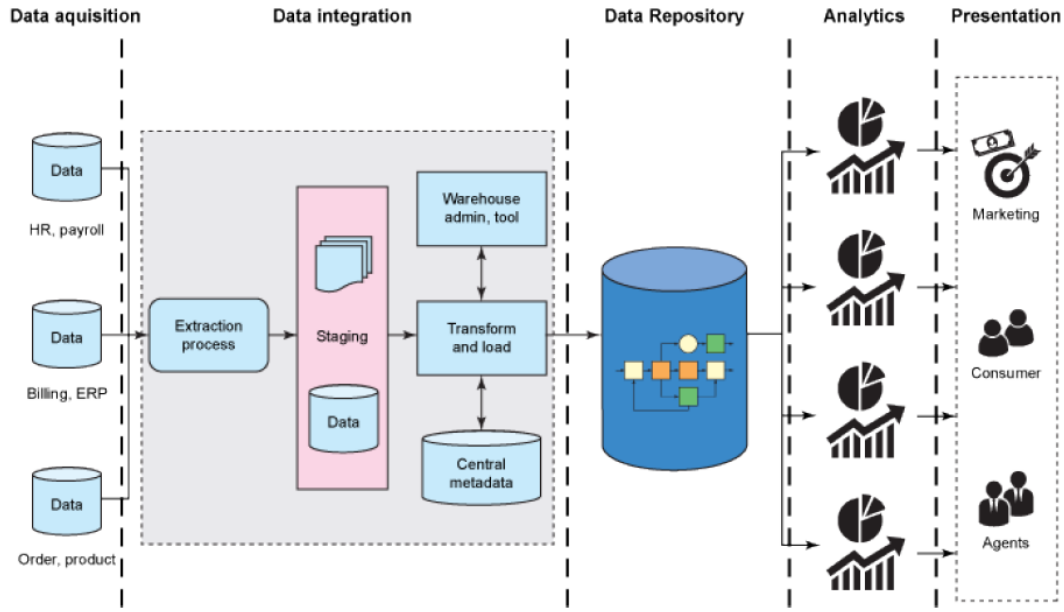
Antes de entrar en el relevamiento estructural y funcional de un proyecto típico de BI a continuación se describen algunas características relevadas que son propias de los EDW tradicionales, y que justifican las propuestas de solución más adelante.

Tradicionalmente los EDW analizan datos estructurados y transaccionales que provienen de bases de datos relacionales.

Estos EDW establecen KPI's (indicadores clave de performance) y una arquitectura guiada por el modelado de datos (model-driven architecture).

En la siguiente figura podemos ver cómo es el flujo de los datos en una estructura tradicional de EDW:

Imagen 10 - Arquitectura tradicional de referencia de EDW



Fuente (IBM, 2014)

En el diagrama vemos cómo:

• Los sistemas de procesamiento de transacciones en línea (OLTP), como los sistemas de RRHH, de finanzas o gestión de clientes, apoyan los procesos empresariales de la empresa.

• Los almacenes de datos operativos (ODS) acumulan las transacciones comerciales para respaldar los informes de las operaciones y permiten realizar consultas sencillas antes de integrar sus datos en el EDW.

• Los EDWs acumulan y transforman las transacciones comerciales para apoyar tanto la toma de decisiones operativas como las estratégicas.

Cada capa de esta arquitectura realiza una función particular:

- Capa de Adquisición de datos (Data acquisition): Consiste en componentes para obtener datos de todos los sistemas fuente, como recursos humanos, finanzas y facturación. Generalmente las fuentes de datos son internas a la organización y los datos son estructurados.

- Capa de Integración de datos (Data integration): Consiste en componentes de integración para el flujo de datos desde las fuentes hasta la capa de repositorio de datos en la arquitectura. Los datos extraídos se alojan en una zona de *staging* o aterrizaje de datos y allí se trasforman antes de ser cargados en el repositorio central. En esta capa se requiere mucha capacidad de procesamiento ya que los procesos de transformación y limpieza de datos pueden ser complejos.

- Capa de Repositorio de datos (Data repository): Almacena datos en un modelo relacional para mejorar el rendimiento y la extensibilidad de las consultas. Se requiere mucha capacidad de almacenamiento si se desea contar con grandes históricos.

- Capa de Análisis (Analytics): Almacena datos en formato de cubo OLAP (OnLine Analytical Processing o procesamiento Analítico en Línea) para facilitar a los usuarios realizar análisis de lo que pasa en la organización o negocio.

- Capa de presentación (Presentation): Aplicaciones o portales que dan acceso a diferentes grupos de usuarios. Las aplicaciones y los portales consumen los datos a través de páginas web y *portlets* (módulos Web reutilizables) que se definen en la herramienta de informes o mediante servicios web. (IBM, 2014)

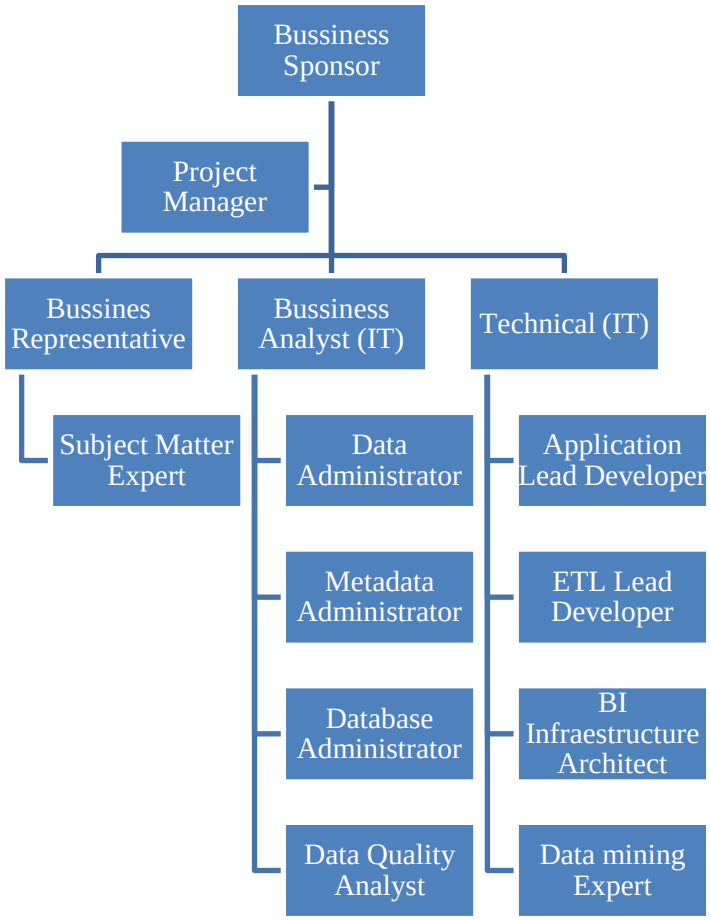
10.7 Relevamiento Funcional

Organigrama

El presente proyecto se enmarca como emprendimiento personal, por lo que no se cuenta con un organigrama organizacional.

Solo a modo informativo a continuación definimos los roles que podemos encontrar en un proyecto típico de BI que sigue la metodología propuesta por Larisa Moss, y con quienes tendríamos que interactuar a la hora de llevar adelante el presente trabajo:

Imagen 11 – Organigrama de roles



Funciones de los roles

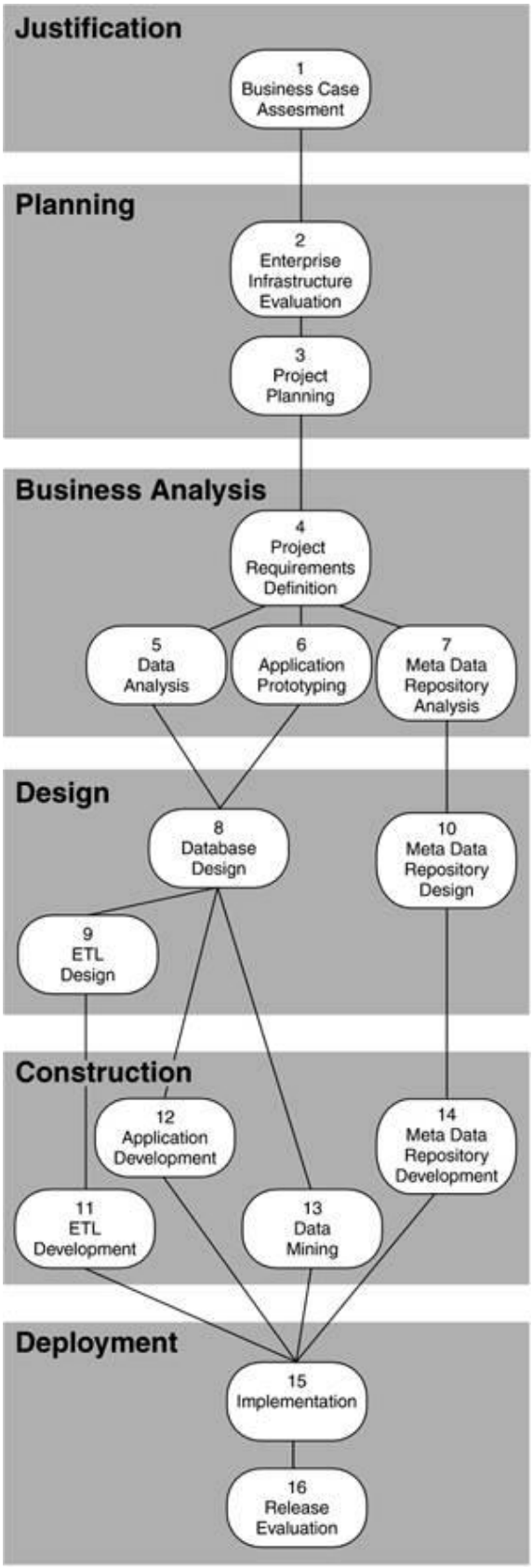
Rol	**Responsabilidades principales**
Business sponsor	Promover la iniciativa de BI y eliminar los obstáculos relacionados con el negocio para el equipo del proyecto BI.
Application lead developer	Diseñar y supervisor el desarrollo del acceso y análisis de las aplicaciones (ej., reportes, queries).
BI infrastructure architect	Establecer y mantener la infraestructura técnica de BI.
Business representative	Participar en las sesiones de modelado, proveer definiciones de datos, escribir casos de prueba, tomar decisiones de negocio, resolver disputas entre unidades de negocio, y mejorar la calidad de los datos bajo el control de la unidad de negocio representada por su rol.

Data administrator	Realización de análisis de datos de toda la Organización, la creación de los modelos de datos lógicos específicos del proyecto, y la fusión de los modelos de datos lógicos en un modelo de datos lógicos de la empresa.
Data mining expert	Elegir y ejecutar la herramienta de minería de datos; debe tener un background estadístico.
Data quality analyst	Evaluar la calidad de datos de origen y la preparación de especificaciones de limpieza de datos para el proceso ETL.
Data base administrator	Diseñar, cargar, monitorear y ajustar las bases de datos de destino de BI.
ETL lead developer	Diseñar y supervisar los procesos ETL.
Meta data administrator	Construir u obtener la licencia (comprar), mejorar, cargar, y mantener el repositorio de meta data.
Project manager	Definir, planificar, coordinar, controlar y revisar todas las actividades del proyecto; seguimiento y reporte del progreso; la resolución de problemas técnicos y comerciales; tutoría del equipo; la negociación con los proveedores, el representante de negocios, y el promotor del negocio; tiene la responsabilidad general del proyecto.
Subject matter expert	Proporcionar conocimiento del negocio sobre datos, procesos y requisitos.

Procesos de negocios

A continuación, mostramos a modo de diagrama de procesos las 6 etapas descriptas en la metodología para llevar adelante el proyecto:

Imagen 12 – Ciclo de vida



Fuente: (Larissa T. Moss, 2003)

Proceso de Análisis de datos

Las actividades realizadas durante el análisis de datos, paso 5 del gráfico anterior, están orientadas hacia la comprensión y corrección de las discrepancias existentes en los datos de la empresa. Por lo tanto, el análisis de datos es una actividad centrada en el negocio, no es una actividad centrada en el sistema.

Actividades:

1. Analizar las fuentes de datos externas.

   Además de requerir datos de origen operacional internos, muchas aplicaciones de BI necesitan datos de fuentes externas. La fusión de los datos externos con los datos internos presenta un conjunto de desafíos. Los datos externos son a menudo sucios e incompletos, y por lo general no se sigue el mismo formato o estructura clave que tienen los datos internos. En este paso se busca identificar y resolver estas diferencias.
2. Refinar el modelo de datos lógico.

   Un modelo lógico de datos de alto nivel específico para el proyecto debería crearse durante uno de los pasos anteriores. Además, algunos o todos los datos internos y externos pueden haber sido modelados en otros proyectos y pueden ya ser parte del modelo de datos lógico de la empresa. En ese caso, se extrae la parte representativa del modelo de datos lógicos de la empresa y se amplía con los nuevos objetos de datos, nuevas relaciones de datos, y nuevos elementos de datos. Si los datos requeridos no se han modelado previamente, se crea un nuevo modelo de datos lógicos para el alcance de este proyecto de BI. Debe incluir todos los datos internos, así como los elementos de datos externos.

3. Analizar la calidad de las fuentes de datos.

   Al mismo tiempo que el modelo lógico de datos se crea o se expande, la calidad de los archivos de origen interno y externo y bases de datos fuente deben ser analizadas en detalle. Es bastante común que los datos operativos existentes no se ajusten a las reglas de negocio establecidas y las políticas comerciales. Muchos elementos de datos se utilizan para múltiples propósitos o simplemente se dejan en blanco. Aquí se identifican todas estas discrepancias y se incorporan en el modelo de datos lógicos.

4. Expandir el modelo de datos lógico de la empresa.

   Una vez que el modelo lógico de datos del proyecto es relativamente estable, se une al modelo lógico de datos de la empresa. Durante este proceso de fusión se pueden identificar discrepancias o inconsistencias adicionales de datos. Estas deberán ser enviadas de nuevo al proyecto de BI para su resolución.

5. Resolver las discrepancias de datos.

   Ocasionalmente las discrepancias en los datos descubiertos durante el análisis de datos implican el involucramiento de otros representantes del negocio de otros proyectos. En ese caso, se convoca a los demás representantes del negocio, así como a los propietarios de los datos para resolver sus diferencias.

6. Escribir las especificaciones para la limpieza de datos.

   Una vez que todos los problemas se identifican y los datos son modelados, se escriben las especificaciones de cómo limpiar los datos. Estas especificaciones deben estar en un lenguaje llano para que puedan ser validadas por el dueño de los datos y por los usuarios del negocio.

Productos de las actividades de Análisis de datos:

1. Modelo de datos lógico normalizado y con todos los atributos.

   Este modelo de datos lógicos específico el proyecto es un diagrama entidad-relación totalmente normalizada mostrando entidades del kernel, entidades asociativas, entidades características, cardinalidad, opcionalidad, identificadores únicos, y todos los atributos.

2. Meta datos del negocio.

   Las entidades de negocio y atributos del modelo de datos lógico deben ser descritos con metadatos. Los componentes de los meta-datos incluyen nombres de datos, definiciones de datos, relaciones de datos, identificadores únicos, tipos de datos, longitudes de datos, dominios, reglas de negocio, políticas y propiedad de los datos. Estos son capturados por lo general en el repositorio de meta-datos que suele ser una herramienta CASE (computer-aided software engineering).

3. Especificaciones de limpieza de datos (Data-cleansing)

   Este documento describe la lógica de limpieza que se debe aplicar a los datos de origen con el fin de ponerlos en conformidad con las normas técnicas de conversión de datos, las reglas de dominio de datos y las reglas de integridad de datos de negocio. Este documento se utiliza para crear las especificaciones de transformación en el documento de mapeo "fuente de datos- base de datos destino" en proceso de Diseño del ETL.

4. Modelo lógico de datos empresarial expandido.

   Este entregable es producido por el área de administración de datos o por el grupo responsable de la arquitectura empresarial al momento de hacer la integración del modelo lógico de datos del actual proyecto con el modelo de datos lógicos de la empresa. Cualquier exclusión de entidades o atributos y las discrepancias entre los modelos serán presentadas al equipo del proyecto de BI para su resolución. (Larissa T. Moss, 2003)

Proceso de Extracción, transformación y carga de datos (ETL)

Este proceso se encuentra subcontratado a una empresa consultora. Cada vez que nuestro cliente requiere modificaciones en los procesos ETL actuales o tiene nuevas necesidades de análisis de datos traslada estos requerimientos al personal especializado externo. Para ello se sigue un procedimiento documentado de requerimientos relevados por parte de analistas funcionales del cliente que luego son trasladados a la empresa subcontratada.

Luego la empresa consultora sigue un proceso de desarrollo de software iterativo e incremental hasta lograr cumplimentar los requerimientos con un equipo de desarrolladores que trabajan en las oficinas del cliente.

Esta etapa de construcción o modificación del ETL es la que más esfuerzos, tiempo y recursos requiere en todo el ciclo de un proyecto de BI y requiere mucha interacción de las distintas áreas de negocio con los especialistas subcontratados.

Proceso de diseño de base de datos

Para este proceso nos limitamos a relevar lo concerniente al almacenamiento de datos históricos. Observamos que por normativas internas y legales nuestro cliente se ve obligado a mantener datos de los últimos 5 años.

Para ello realizan backups o copias de seguridad comprimiendo la información y almacenándola en un servidor dedicado a tal fin y también en dispositivos de almacenamiento externos, como discos duros externos y memorias de estado sólido, dependiendo del departamento. No poseen una metodología unificada y los datos almacenados requieren de una restauración en las bases de datos para poder ser accedidos si fuera necesario.

El acceso a estos datos no ocurre habitualmente ya que, para el proceso de análisis y toma de decisiones en sus niveles operacional, táctico como estratégico se limitan a trabajar con los datos del último año en curso.

11. Diagnóstico

A continuación, se dará el diagnóstico para los tres procesos relevados en el apartado anterior, en los cuales se observa oportunidad de mejora mediante el presente trabajo:

1) Proceso: Análisis de datos

Problema: No se está considerando la oportunidad de incorporar fuentes de datos no estructurados o semi-estructurados para obtener valor de negocio incorporándolos al EDW.

Causa: La tecnología que están utilizando actualmente en la arquitectura de BI solo soporta fuentes de datos estructurados provenientes de las bases de datos de los sistemas operacionales.

2) Proceso: Diseño y construcción de ETL

Problema: Capacidad limitada de almacenamiento y procesamiento para los procesos ETL.

Causa: Se están utilizando bases de datos relacionales para la etapa de staging de datos, y los procesos más complejos de ETL requieren de mayor capacidad de procesamiento.

3) Proceso: Diseño y construcción de Bases de datos.

Problema: Existen datos históricos que ya no se utilizan con frecuencia o se consultan cada vez menos, pero incrementan la demanda de almacenamiento en los servidores del EDW y reducen la performance general.

Causa: No se están considerando almacenamientos alternativos para estos datos que liberen a los servidores de EDW pero que al mismo tiempo sigan estando disponibles para los usuarios.

12. Propuestas de solución

A continuación, se proponen mejoras en tres casos de uso que integran tecnología de Big Data, más concretamente la plataforma open-source de Hadoop, con la plataforma de EDW tradicional de nuestro cliente. Esta propuesta está comprendida dentro de una estrategia de adopción progresiva de nuevas tecnologías más adecuadas a los nuevos paradigmas de Data Lake en el manejo de datos.

Como se planteó a lo largo del presente informe el objetivo para líneas futuras de trabajo es poder incorporar en nuestro cliente las técnicas de Analítica avanzada *(Advanced analytics)* e incorporar también en este proceso las fuentes de Big Data, indispensables ambas para evitar ser desplazados de su mercado en el contexto de la Transformación Digital en el que muchas empresas ya están inmersas.

Dado que actualmente nuestro cliente no cuenta con fuentes de Big Data, con la infraestructura que las soporte ni con los recursos humanos que puedan llevar adelante el rol de científico de datos (data scientist) la propuesta de mejora se ajusta a la incorporación de herramientas open-source de Big Data en la infraestructura de hardware existente en la empresa y para que los recursos humanos que actualmente administran y usan el EDW puedan también hacer uso del sistema mejorado.

Por lo tanto las tres mejoras propuestas a continuación, además de los beneficios que en cada una se detallan, permitirán a nuestro cliente ir incorporando nuevas herramientas del ecosistema Hadoop a medida que pueda contratar o capacitar al personal necesario para realizar cada vez análisis más complejos para procesar también datos semi-estructurados o sin estructura, provenientes de nuevas fuentes de datos internas o externas, de dispositivos (IoT), sensores, servidores o páginas web entre otros.

Las propuestas son:

1. **Aterrizaje de los datos estructurados (Stage)**. Utilizar Hadoop y su ecosistema como plataforma de almacenamiento de datos para su EDW. Procesar y transformar datos semi-estructurados antes de cargarlos en el EDW. La ventaja es que debido al bajo costo del almacenamiento de Hadoop, se pueden almacenar ambas versiones de los datos en Hadoop: los datos crudos, nativos (raw data) y los datos transformados. Los datos estarían ahora en un solo lugar, facilitando la administración, el procesamiento y el análisis posterior.

2. **Procesamiento de datos semi-estructurados.** Procesar mediante Hadoop y su ecosistema datos semi-estructurados que actualmente no están disponibles porque el EDW no puede manejar o no maneja bien, con el fin de integrarlos a los datos ya existentes y proporcionar información adicional sobre los clientes, productos y servicios.

3. **Archivar todos los datos**. Utilizar Hadoop y su ecosistema para archivar todos sus datos de manera local o en la nube. Dado que Hadoop se ejecuta en *hardware commodity* que se puede escalar fácil y rápidamente, se podrán almacenar y archivar muchos más datos a un costo mucho menor. La ventaja es que se contará con un backup “vivo”, es decir que se puede seguir consultando fácilmente al igual que el resto de los datos, a diferencia de los backup que se almacenan comprimidos, en soportes extraíbles o incluso en distintas instalaciones edilicias.

Requerimientos de información:

1. Se requiere que los nuevos datos se incorporen en el EDW organizacional para ser analizados junto al resto de la estructura de datos que actualmente maneja la organización en su arquitectura de BI.
2. Se requiere poder analizar y visualizar los nuevos datos mediante las herramientas de BI existentes.
3. Se desea mantener almacenados y accesibles por tiempo indeterminado los datos traídos al staging (es decir los datos que aún no fueron transformados y cargados en el EDW por los procesos de ETL)con el fin de poder satisfacer nuevos requerimientos de datos que impliquen modificaciones en los procesos ETL sin perder historial, y sin que esto implique un compromiso de los recursos de almacenamiento y procesamiento de los servidores de bases de datos actuales dedicados al staging y los procesos ETL.
4. Se desea que cuando los grandes volúmenes de datos históricos en el EDW dejen de ser requeridos por más de 2 años pasen a un lugar de almacenamiento alternativo para liberar recursos en el EDW, pero que queden activos, es decir disponibles para consultas eventuales de los usuarios.
5. Se requiere que las actuales aplicaciones de BI para análisis y visualización permitan consultar tanto los datos que están en el EDW como los que se encuentran en los medios de almacenamiento alternativos, accediendo a ellos directamente sin que tengan que ser incorporados nuevamente al EDW.

13. Desarrollo del Producto/servicio

13.1 Análisis y Diseño

Como hemos desarrollado a lo largo del trabajo, vamos a incorporar Hadoop al sistema existente de EDW de nuestro cliente, que como vemos en el diagrama de abajo tendrá la función de servir como estapa de stage, recolectando datos de los sistemas operacionales del cliente y a su vez nuevas fuentes de datos como logs de servidores de aplicaciones, servidores web y streaming desde redes sociales.

La inclusión de estas nuevas fuentes de datos le permiten al cliente las siguientes funcionalidades:

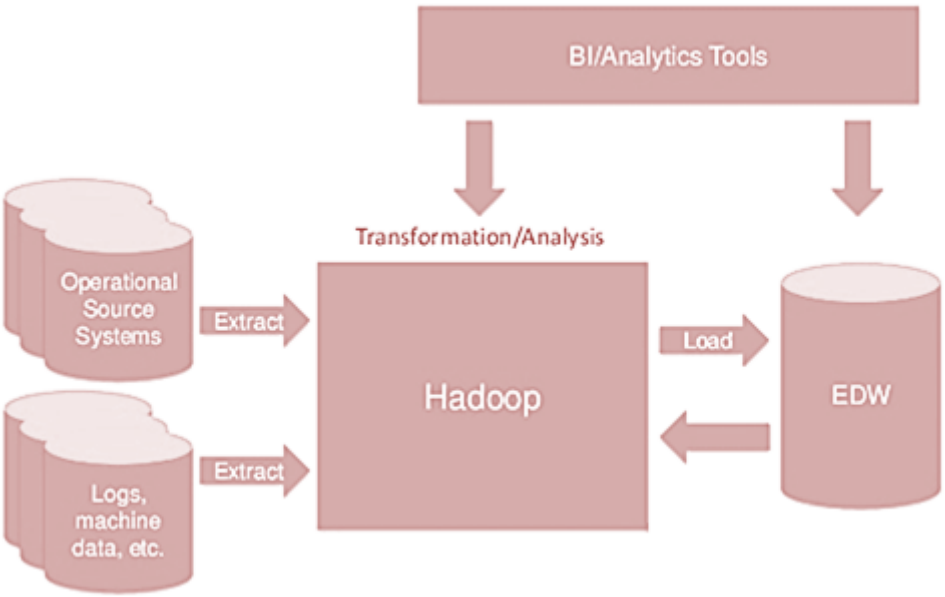
- Identificar en sus aplicaciones errores o vulnerabilidades a través de los logs de aplicaciones.
- Detectar opertunidades de mejora en los canales de venta online, analizando los clicstrems de sus clientes alnavegar por el sitio de compras.
- Obtener feedback de sus productos al analizar streaming de redes sociales como Twitter.
- Integrar las nuevas fuentes de datos con los datos ya existentes de sus bases de datos operativas para obtener nuevos insigths o valor agregado para la toma de desiciones.

A su vez Hadoop cumple el rol del ETL, procesando los datos crudos antes de cargarlos en el EDW, y como etapa de almacenamiento de históricos, tanto de los datos en crudo como de datos procesados. Esto le ofrece al cliente las siguientes mejoras en sus procesos:

- Libera recursos del EDW al procesar y almacenar los datos en Hadoop.

- Permite escalar la solucion integrando servidores comodities al cluster de Hadoop, con el menor costo y facilidad que esto implica comparando con la escalabilidad del EDW.
- Facilita la incorporación de nuevas fuentes de datos no estructurados o semi-estructurados desde Hadoop.
- Agiliza el proceso al momento de requerir cambios en los modelos de datos, ya que se mantienen los datos en crudo en el cluster de Hadoop y se cuenta con un histórico activo.
- Requiere menos programación y soporte de IT que el EDW, ya que Hadoop y su ecosistema permiten a los analistas adquirir, procesar e integrar los nuevos datos mediante configuraciones, consultas de tipo SQL (Hive) y lenguajes de scripting (Pig):

Imagen 13 – EDW con Hadoop



Fuente: Apache Hadoop, 2017

En lugar de hacer una instalación de cero de Hadoop utilizaremos la distribución de Hortonworks, ya que es una de las mas utilizadas junto con Cloudera e integra en una misma instalación el conjunto principal de productos del ecosistema de Hadoop y nos evitamos tener que instalarlos de manera invididual a riesgo de encontrarnos con incompatibilidad de versiones. Además esta distribución cuenta con una edición gratuita.

La edición gratuita está disponible con el nombre de Sandbox en una máquina virtual montada sobre un sistema operativo Linux CentOs que además de un sistema base de Hadoop contiene las siguientes herramientas de su ecosistema:

- HDFS es el sistema de almacenamiento, es un sistema de ficheros distribuido. Fue creado a partir del **Google File System** (GFS). HDFS se encuentra optimizado para grandes flujos y trabajar con ficheros grandes en sus lecturas y escrituras. Aquí almacenaremos los datos en crudo y también los procesados, a modo de repositorio de históricos.

- MapReduce es un proceso batch, creado para el proceso distribuido de los datos. Permite de una forma simple, paralelizar trabajo sobre los grandes volúmenes de datos, como combinar web logs con los datos relacionales de una base de datos OLTP, de esta forma ver como los usuarios interactúan con el website. Sobre esta capa se ejecutan las herramientas como Flume, Sqoop, Pig o Hive, evitándole al cliente tener que escribir Jobs en java, y reemplazando ese esfuerzo por procesos más ágiles de configuración, consutas SQL y scripting.

- Apache Pig, inicialmente desarrollado por Yahoo, permite a los usuarios de Hadoop centrarse más en el análisis de los datos y menos en la creación de programas MapReduce. Para simplificar el análisis proporciona un lenguaje procedural de alto nivel.

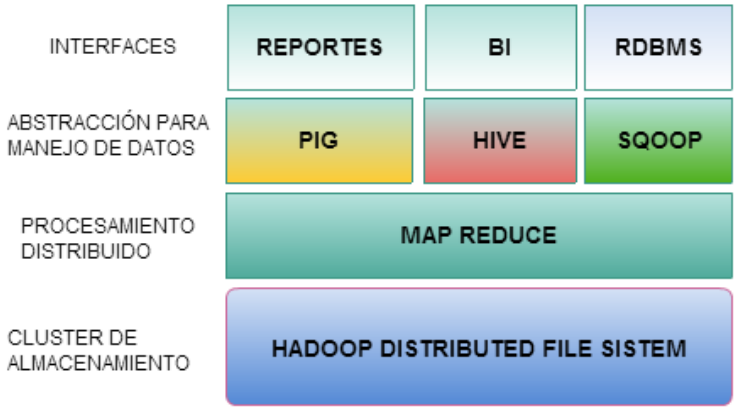
- Hive es un sistema de Data Warehouse para Hadoop que facilita el uso de la agregación de los datos, ad-hoc queries, y el análisis de grandes datasets almacenados en Hadoop. Hive proporciona métodos de consulta de los datos usando un lenguaje parecido al SQL, llamado *HiveQL*. Además, permite de usar los tradicionales Map/Reduce cuando el rendimiento no es el correcto. Tiene interfaces JDBC/ODBC, por lo que empieza a funcionar su integración con herramientas de BI.

- Apache Sqoop ("Sql-to-Hadoop"), es una herramienta diseñada para transferir de forma eficiente bulk data entre Hadoop y sistemas de almacenamiento con datos estructurados, como bases de datos relacionales de los sistemas operacionales del cliente, para luego integrar esos datos con las nuevas fuentes que nos permite integrar Flume.

- Apache Flume es un sistema distribuido para capturar de forma eficiente, agregar y mover grandes cuantidades de datos log de diferentes orígenes (diferentes servidores) a un repositorio central, simplificando el proceso de recolectar estos datos para almacenarlos en Hadoop y poder analizarlos. Permite hacer el streaming en tiempo real tanto de logs como de servidores web o comentarios en redes sociales como Twitter.

13.2 Diagrama de componentes

Imagen 14 – Ecosistema Hadoop.

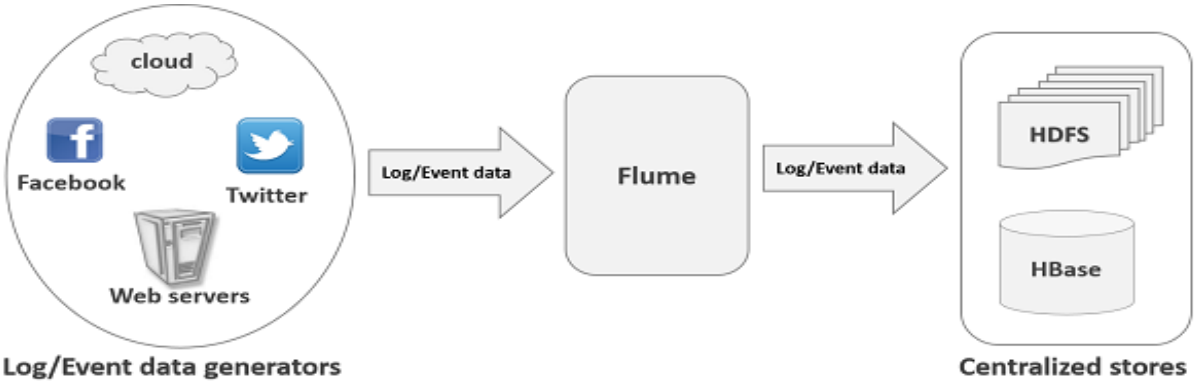


Fuente: Apache Hadoop, 2017

En la demostración del presente trabajo utilizaremos Apache Flume para incorporar al sistema de EDW una nueva fuente de datos semi-estructurada, en concreto archivos de logs de un servidor web de nuestro cliente el cual registra la actividad y navegación de los usuarios por los productos y servicios de la empresa. Como requerimiento candidato tenemos la incorporación de streaming de Twitter, para analizar el feedback que los clientes dan de los productos.

Estos logs serán capturados en tiempo real mediante agentes de Flume instaladores en el servidor web y serán almacenados en Hadoop, directamente en su cluster de almacenamiento HDFS.

Imagen 15 – Flume



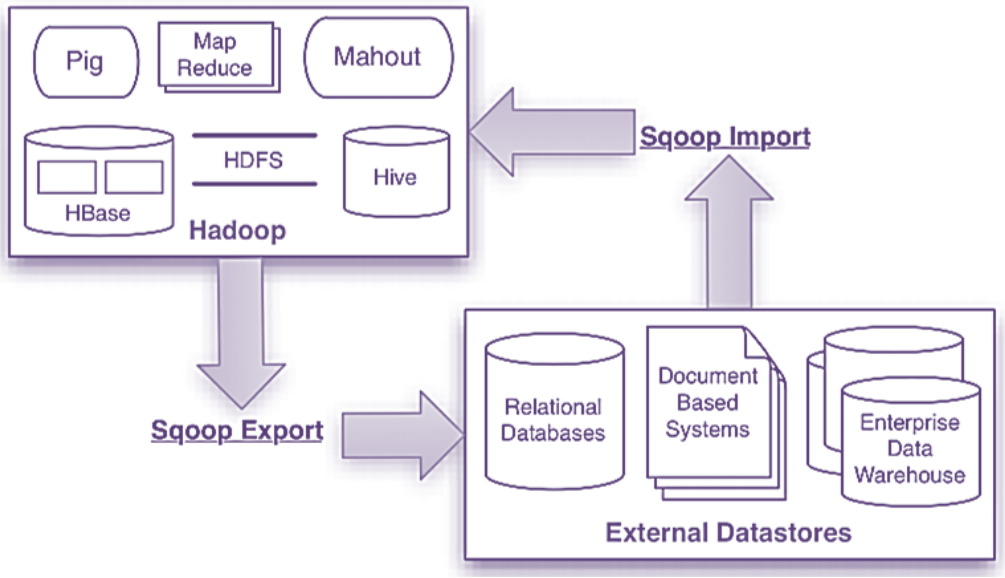
Fuente: Apache Flume, 2017

Una vez obtenidos los archivos de logs mediante Hive se procederá a la creación de los metadatos necesarios en el repositorio de Hive, con el cual crearemos una base de datos y una tabla que luego nos permitirán hacer consultas del tipo SQL.

Con Pig podremos hacer transformaciones más complejas de los datos cargados en la base de datos en Hive para luego cargarlos en el EDW e integrarlos con el resto de los datos actuales, o para dejarlos almacenados en el sistema de archivos distribuido de Hadoop.

Para obtener datos del EDW o de los sistemas transaccionales utilizamos la herramienta de Import de Sqoop. Mientras que para cargarlos en las bases de datos relacionales del EDW una vez procesados utilizamos Export:

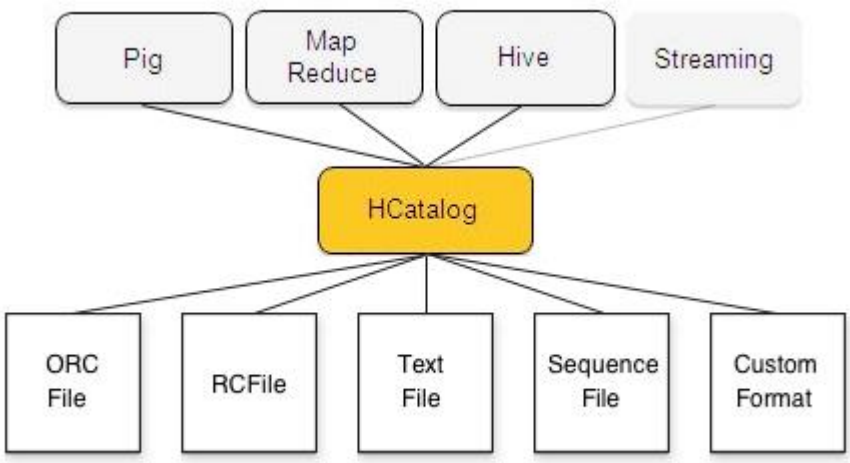
Imagen 16 – Sqoop.



Fuente: (Sqoop, 2017)

Sqoop permite incorporar los meta datos necesarios en H-Catalog para que luego las demás aplicaciones como Mapreduce, Hive, Pig puedan leer y escribir fácilmente a las tablas importadas Hadoop, independientemente de la estructura de los datos.

Imagen 17 – H-Catalog.

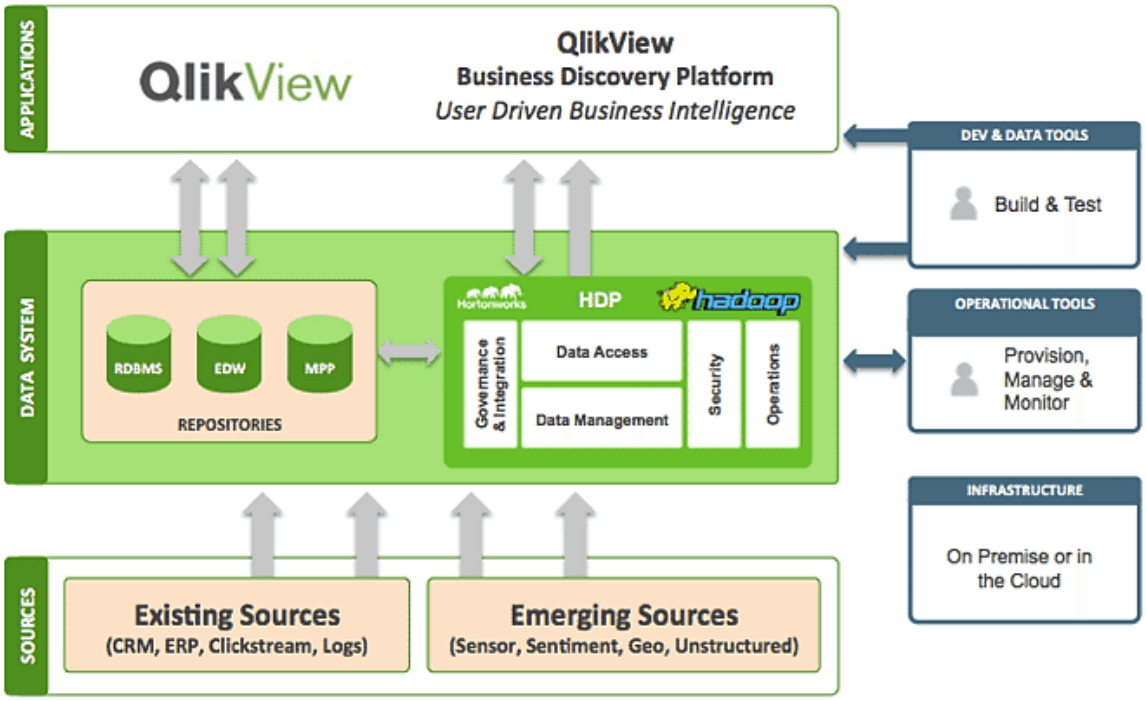


Fuente: (Apache H.-C. , 2017)

La Interfaces de salida pueden ser cualquier aplicación que se conecte mediante **JDBC/ODBC** (Herramientas de BI, Excel con PowerView, aplicaciones) para visualizar la información.

Para el presente trabajo se utilizará QlikView, ya que es la herramienta de BI que actualmente usa nuestro cliente como salida del EDW, que puede conectarse directamente al EDW, pero también al cluster HDFS de Hadoop:

Imagen 18 – QlikView



Fuente: QlikView, 2017

Mediante QlikView se demostrará la facilidad para acceder directamente a los datos en Hadoop y realizar distintos tipos de análisis que permitan encontrar nuevos patrones en los comportamientos de los clientes y en los procesos de negocio, que a través del proceso de toma de decisiones permiten a nuestro cliente mejorar sus productos y servicios.

13.3 Administración del Proyecto

Calidad

En los proyectos de BI la calidad pasa principalmente por la calidad de los datos. Es a partir de datos limpios que podemos combinarlos y agregarlos podemos responder a preguntas de negocio sencillas y a partir de esa información consolidada obtener conocimiento para la toma de decisiones.

Por lo tanto, es importante que para el trabajo actual tenga un rol protagónico el rol del analista de calidad de datos a lo largo del todo el ciclo de vida de desarrollo.

Debido a que el alcance del proyecto no contempla incorporar fuentes de datos externas vamos a reducir el esfuerzo necesario de las tareas de calidad de datos ya que la nueva fuente de datos semi-estructurados a incorporar se trata de logs de un servidor de aplicaciones de la misma organización que ya están estandarizados.

Por consiguiente, el esfuerzo principal pasará por transformarlos de manera que sean consistentes con el resto de los datos del EDW y que sean lo suficientemente completos como para aportar información de valor para el negocio y conocimiento nuevo para la toma de decisiones.

De todas formas, tal como se especificó en el ciclo de vida de desarrollo, en la etapa de Despliegue está contemplada la evaluación de la entrega donde se debe validar que los datos capturados y procesados sean los correctos para responder las preguntas que el negocio requiere.

Calidad de los datos

El analista de calidad de datos se encarga de buscar y analizar datos sucios en los archivos fuente. Dado que es imposible limpiar todos los datos sucios, para el actual proyecto la

organización debe establecer procedimientos de limpieza y directrices de priorización. El analista de calidad de datos es el administrador de los estándares de calidad de datos.

Mientras que todo el mundo en el lado de los negocios y en el departamento de IT quiere calidad, rara vez se asigna el tiempo suficiente para lograrlo porque la calidad y el esfuerzo son limitaciones polarizadas. Mayor calidad requiere más esfuerzo y por lo tanto más tiempo para entregar. Dado que los factores de tiempo impulsan a la mayoría de las organizaciones, el esfuerzo es su principal restricción (la más alta prioridad), seguido por el alcance, el presupuesto y los recursos (normalmente en ese orden); Y la calidad se empuja al fondo del montón (prioridad más baja). Las restricciones de proyectos de BI nunca deben estar en este orden, sino justamente en el inverso. (Larissa T. Moss, 2003)

Costos

El presente trabajo propone una inversión progresiva por parte del cliente. Por lo tanto, para hacer el proyecto piloto solo se requiere de una mínima inversión inicial que se describirá en este apartado para el almacenamiento y procesamiento en la nube y para las horas hombre de un profesional interno o externo que llevará adelante las tareas planificadas de desarrollo. El nuevo software a incorporar es de licencia *open-source* e inicialmente no se prevé contratar soporte para la distribución utilizada de Hortonworks.

Si se decidiera optar por un profesional interno, que podría ser parte del personal que actualmente administra el EDW, será necesario tener en cuenta los esfuerzos iniciales de capacitación y aprendizaje de las nuevas tecnologías. Pero dicho esfuerzo puede considerarse como parte del plan actual de capacitaciones constantes de la empresa, y luego se puede ir ampliando de manera progresiva a medida que la empresa requiera de análisis más complejos de sus datos.

Si el cliente contara con servidores locales con recursos disponibles podría ahorrarse el costo de los servicios en la nube, ya que la solución propuesta no requiere de un hardware específico o con características especiales para la tecnología a emplear, es decir que puede utilizarse *hardware commodity*.

A continuación, se detallan los ítems a tener en cuenta a la hora de pensar en costos iniciales, que como veremos impactan principalmente en los servicios.

- Software:

  - Sistema Operativo open-source basado en Linux (Centos).
  - Ecosistema de Hadoop open-source.

- Servicios:

  - Implementación y mantenimiento Hardware.
  - Implementación, mantenimiento y desarrollo de Software.
  - Capacitación de personal técnico y usuarios.
  - Contratación de soporte del sistema operativo y del ecosistema Hadoop.
  - Consultoría en procesos de negocios para definir nuevos requerimientos.
  - Contratación para el almacenamiento y procesamiento de Hadoop en la nube.

- Hardware

  - Nuevos servidores *commodity* para el *cluster* Hadoop (opcional para futuro escalamiento de la solución con servidores locales).

Asumiremos que inicialmente los únicos costos serán por la contratación de servicios en la nube para alojar el sandbox de Hortonworks y por el esfuerzo (horas/hombre) que implica el proyecto según la siguiente estimación de actividades y tareas a llevar a cabo:

Task Name	Duration
Proyecto BI + BIG DATA	**121 days**
Analizar las alternativas Open Source de tecnologías BIG DATA	**8 days**
Buscar al menos dos alternativas del mercado	5 days
Seleccionar la opción mas adecuada	3 days
Analizar y seleccionar herramientas de BI para visualizar	**8 days**
Buscar al menos dos alternativas del mercado	5 days
Seleccionar la opción mas adecuada	3 days
Analizar y rediseñar arquitectura BI	**14 days**
Presentar una arquitectura clásica de BI	4 days
Analizar qué elementos pueden ser reemplazados por BIG DATA	10 days
Implementar la arquitectura diseñada	**21 days**
Instalar plataforma BIG DATA en VMware	11 days
Integrar Plataforma BIG DATA con BI	10 days
Generar datos de entrada	**10 days**
Automatizar fuentes de datos de entrada	10 days
Integrar y analizar los datos existentes	**28 days**
Almacenar en staging datos de las fuentes	7 days
Extraer datos de las respectivas fuentes	7 days
Transformar los datos para análisis	7 days
Cargar los datos en EDW	7 days
Visualizar la información obtenida	**32 days**
Conectar las herramientas para visualizar	7 days
Personalizar las visualizaciones en función del usuario	15 days
Demostrar integración de los datos de EDW y BIG DATA	10 days

Teniendo en cuenta que el tiempo estimado es de 121 días, es decir 4 meses, y considerando el honorario mensual promedio de un Analista Programador de $35.000, según los datos del Consejo Profesional de Ciencias Informáticas de la Provincia Córdoba actualizados a marzo de 2017, nos da un costo de horas/hombre de $140.000. (CPCIPC, 2017)

A continuación, se dejan como referencia los precios de la contratación de servicios de procesamiento y almacenamiento en la nube de **Microsoft Azure.**

La siguiente tabla muestra los precios por mes para un servicio denominado *General purpose Virtual Machines* de Microsoft Azure, apropiada los requisitos mínimos de hardware para la plataforma Hortonworks (sandbox) que usaremos en este trabajo:

INSTANCIA	NÚCLEOS	MEMORIA	CAPACIDAD	PRECIO
A0	1	0.75 GiB	**20 GB**	~$13.40/por mes
A1	1	1.75 GiB	**40 GB**	~$18.60/ por mes
A2	2	3.50 GiB	**60 GB**	~$63.24/ por mes
A3	4	7.00 GiB	**120 GB**	~$139.88/ por mes
A4	8	14.00 GiB	**240 GB**	~$279.75/ por mes

El servicio de Microsoft Azure ofrece además el primer mes gratuito, y como puede verse en la tabla de arriba es posible contratar el servicio A0 desde unos $13 dólares para la etapa de desarrollo y pruebas de un proyecto piloto, y luego ir escalando a medida que se requiere más capacidad de procesamiento y almacenamiento.

La opción A3 de aproximadamente $140 dólares por mes es una configuración adecuada para los requerimientos mínimos de la máquina virtual con el sandbox de Hortonworks. (Azure, 2017)

Costo total

Teniendo en cuenta los ítems mencionados anteriormente, para un proyecto piloto con las características propuestas en el presente trabajo, el costo estimado total es de $140.000 de servicios profesionales más $6.720 por la utilización de los servicios en la nube de Microsoft Azure, opción A3 de la tabla de arriba, teniendo en cuenta que el primer mes es gratuito.

Riesgos

Como se menciona en el marco teórico las principales barreras que enfrentan las organizaciones para llevar a buen puerto un proyecto de Big data son:

1. La seguridad
2. Las limitaciones presupuestarias
3. La falta de expertos en Big Data
4. La falta del uso continuado de Big Data y analítica
5. La integración en los sistemas existentes

Para sortear estas barreras y para prevenir la mayoría de los riesgos que implican se ha planificado un proyecto con un alcance bien acotado que no requiere de un presupuesto significativo frente a los costos actuales del EDW de nuestro cliente.

Sin embargo, una vez evaluados y especificados los casos de uso propuestos es importante que se tengan presentes sus riesgos y que exista un plan de contingencia en caso de que se hagan presentes durante el proyecto, para evitar o minimizar el impacto.

A continuación, se definen los riesgos principales para los tres casos de uso propuestos en el proyecto:

R01 – Seguridad de los datos.

Magnitud: Alta

Descripción: Al almacenar los datos que habitualmente se almacenaban en el EDW en los servidores del cluster de Hadoop, y ante la falta de conocimientos de los recursos humanos para la correcta gestión de la seguridad y control de acceso en la plataforma Hadoop puede agregar una vulnerabilidad al sistema completo.

Impacto: La vulnerabilidad puede ser explotada y usuarios no autorizados pueden tener acceso a información sensible de la empresa, clientes o productos.

Indicador: No posee

R02 – Falta de capacidad de almacenamiento.

Magnitud: Media

Descripción: Ante la necesidad de almacenar más datos y de ampliar el cluster de servidores, ya sea con hardware propio o en la nube, no existe presupuesto disponible para la compra de equipos o para contratar servicios en la nube.

Impacto: No se contará con el histórico adecuado de datos para hacer los análisis requeridos.

Indicador: Espacio en disco de los equipos actuales.

R03 – Falta de recursos humanos.

Magnitud: Media

Descripción: El cambio de prioridades en los proyectos actuales limitan el tiempo disponible para implementar el proyecto actual.

Impacto: Retrasos en el calendario planificado.

Indicador: Puntos de control del calendario planificado.

R04 – Imposibilidad de integración con el EDW existente.

Magnitud: Baja

Descripción: No es posible integrar en el EDW los nuevos datos capturados y procesados por Hadoop.

Impacto: No se podrán usar las aplicaciones actuales de BI para obtener información de valor de productos, clientes y servicios a partir de los nuevos datos.

Indicador: Accesos al EDW denegados a Hadoop por cambios de política en la gestión de datos.

Matriz de Riesgos:

Id.	Prob. de Ocurrencia	Nivel de Impacto	Evaluación del Riesgo	Acciones de Prevención	Acción de Corrección
R01	40	5	200	- Capacitar acerca de la administración de la seguridad en Hadoop. - No almacenar eh Hadoop datos sensibles de los sistemas operacionales.	Restringir los accesos que no estaban controlados.
R02	50	3	150	- Almacenar los datos estrictamente necesarios. - Definir política de eliminación de datos.	Depurar y borrar los datos menos utilizados y más viejos.
R03	40	3	120	- Hacer una planificación realista de acuerdo a la disponibilidad de los RRHH en los próximo 4 meses.	Re-planificar.
R04	10	2	20	- Hacer participar en el proyecto a los usuarios de negocio y administradores de sistemas claves para obtener autorización de antemano de los accesos necesarios.	Utilizar aplicaciones que puedan consultar los nuevos datos directamente desde Hadoop.

Columnas de la matriz:

Id.: Identificador de Riesgo

Probabilidad (1 a 100): Grado de probabilidad de que el Riesgo finalmente se produzca.

Nivel de Impacto: Grado de Impacto en el Proyecto en el caso de que el Riesgo finalmente se produjera. Se mide en una escala de 1 a 5, siendo 1=poco influyente, hasta 5=fuertemente influyente.

Evaluación del Riesgo: Valor numérico resultante del producto del Grado de Probabilidad por el Grado de Impacto. Este producto dará la prioridad que tendrá la gestión de este Riesgo y la implantación de sus medidas preventivas o correctoras.

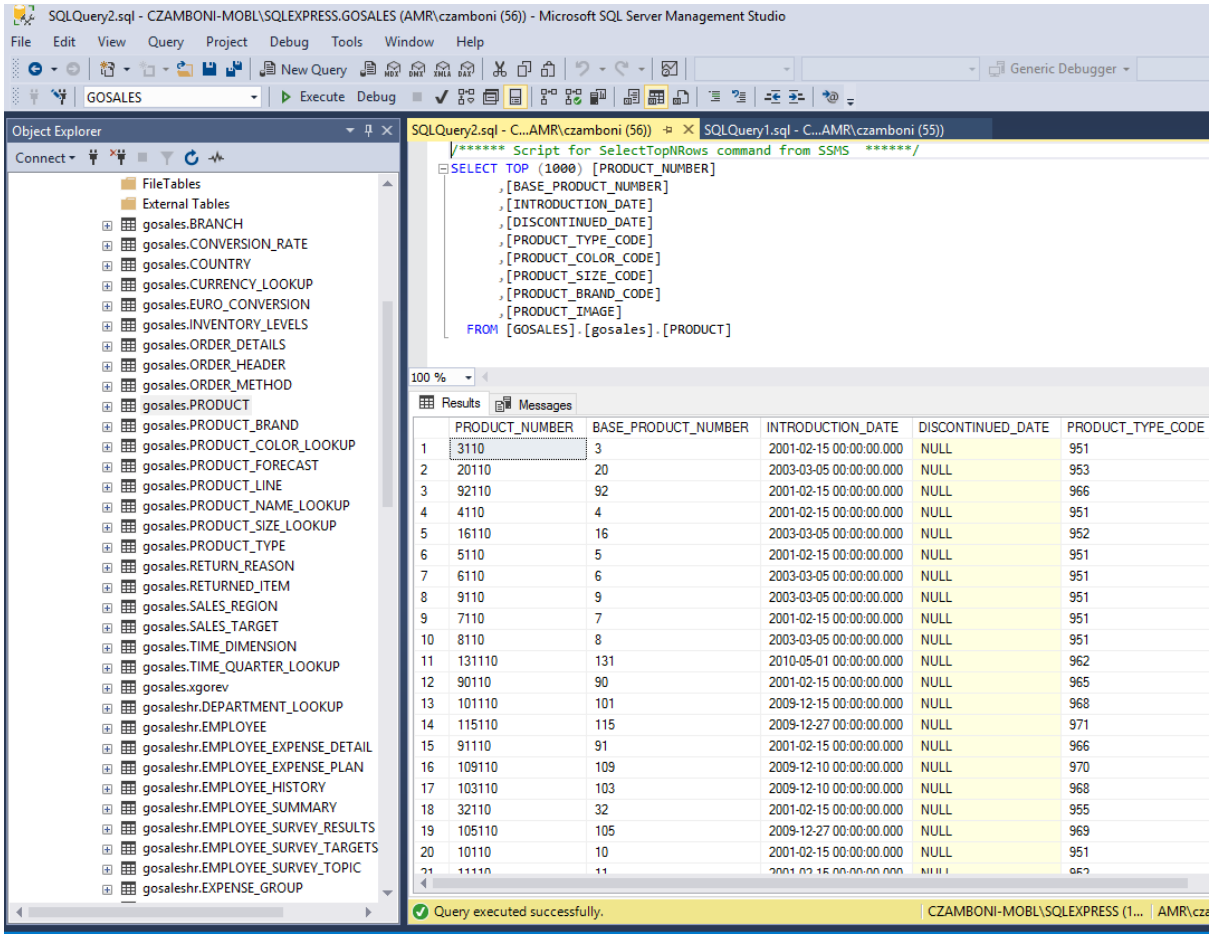
Acciones Prevención: Descripción de las Acciones o Medidas a Adoptar para evitar (mitigar) la aparición final del Riesgo.

Acciones Corrección: Descripción de las Acciones o Medidas a Adoptar en el caso en el que el Riesgo finalmente se haya producido.

14. Implementación (prototipo)

Inicialmente el cliente ya cuenta con un motor de base de datos SQL Server con un base de datos operacional:

Imagen 19 – Sql Server (EDW)

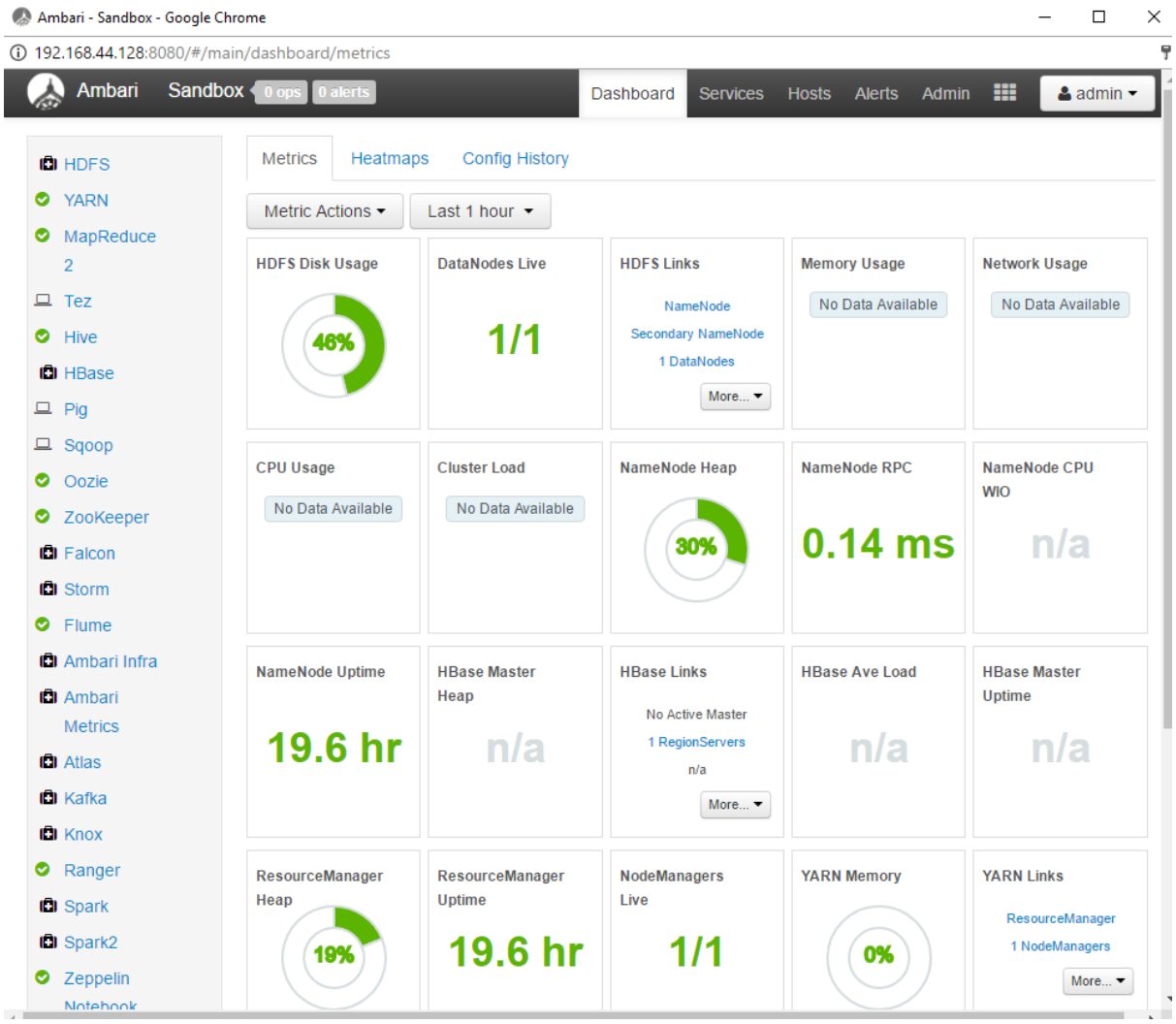


A su vez tiene su EDW con el modelado conceptual, tablas de dimensiones y hechos diseñados.

Por lo que el primer paso es instalar el sandbox de Hortonworks, que consiste en una máquina virtual (VM) con CentOS (Linux) como sistema operativo y con Hadoop y su ecosistema instalados.

Una vez que tengamos la VM corriendo podemos abrir desde un navegador web *Apache Ambari*, el sistema open source para la gestión del ciclo de vida, administración y monitoreo de clusters Apache Hadoop.

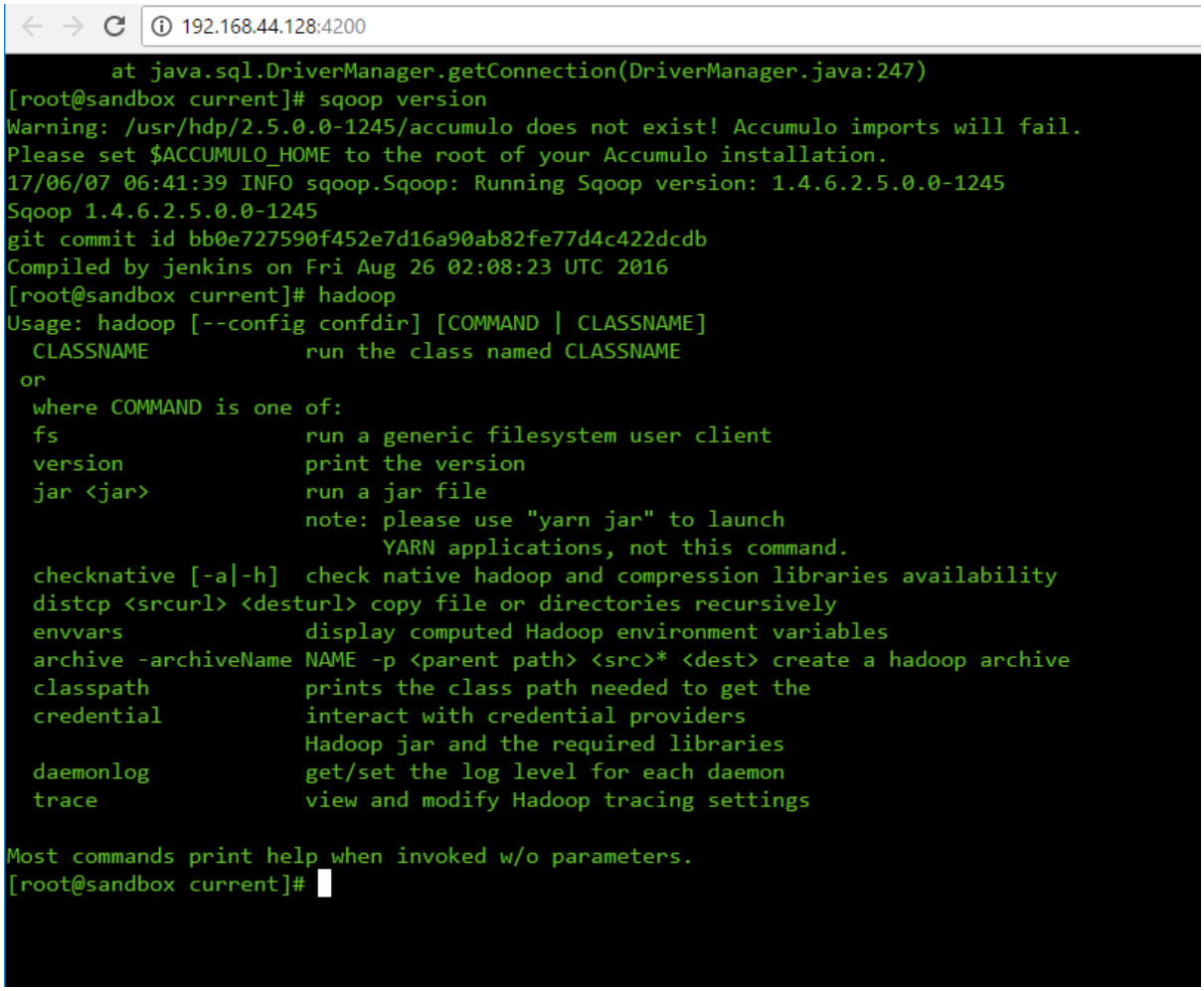
Imagen 20 – Apache Ambari



Desde aquí verificamos que los servicios de HDFS, Flume, Sqoop, Hive y Pig estan instalados y corriendo.

Desde la consola SSH web que proporciona el sandbox vamos a proceder a configurar y correr los comandos necesarios de Sqoop para traer a hadoop las tablas de la base de datos operacional y del EDW necesarias:

Imagen 21 – Consola SSH web.



Es necesario asegurarse que se cuenta con el conector JDBC para Sql Server, en este caso sqljdbc_6.0.8112.100_enu.tar, en la carpeta:

/usr/hdp/current/sqoop-client/lib

Con el siguiente comando podemos ver en consola el listado de todas las tablas disponibles en la base de datos del cliente:

```
sqoop list-databases --connect 'jdbc:sqlserver://192.168.0.10;database=GOSALES' --
username sqoop --password sqoop
```

Mientras que con la herramienta Import traemos las tablas y generamos la mata-data necesaria en H-Catalog para que Hive tenga acceso a esos datos:

```
sqoop import --connect <jdbc-url> -table <table-name> --hcatalog-table txn <other sqoop options>
```

Sqoop nos permite especificar exactamente qué datos queremos importar desde la fuente seleccionada:

Importamos una tabla completa de esta forma:

```
sqoop import --connect jdbc:sqlserver://db.foo.com/bar --table EMPLOYEES
```

Importamos un subset de columans de una table así:

```
sqoop import --connect jdbc:sqlserver://db.foo.com/bar --table EMPLOYEES --columns "employee_id,first_name,last_name,job_title"
```

Importamos solo los últimos registros de una table con una clausula WHERE y entonces agregamos estos datos a una tabla existente:

```
sqoop import --connect jdbc:sqlserver://db.foo.com/bar --table EMPLOYEES  --where "start_date > '2010-01-01'"
```

```
sqoop import --connect jdbc:sqlserver://db.foo.com/bar --table EMPLOYEES --where "id > 100000" --target-dir /incremental_dataset --append
```

Incluso se puede usar sentencias SQL en los comandos:

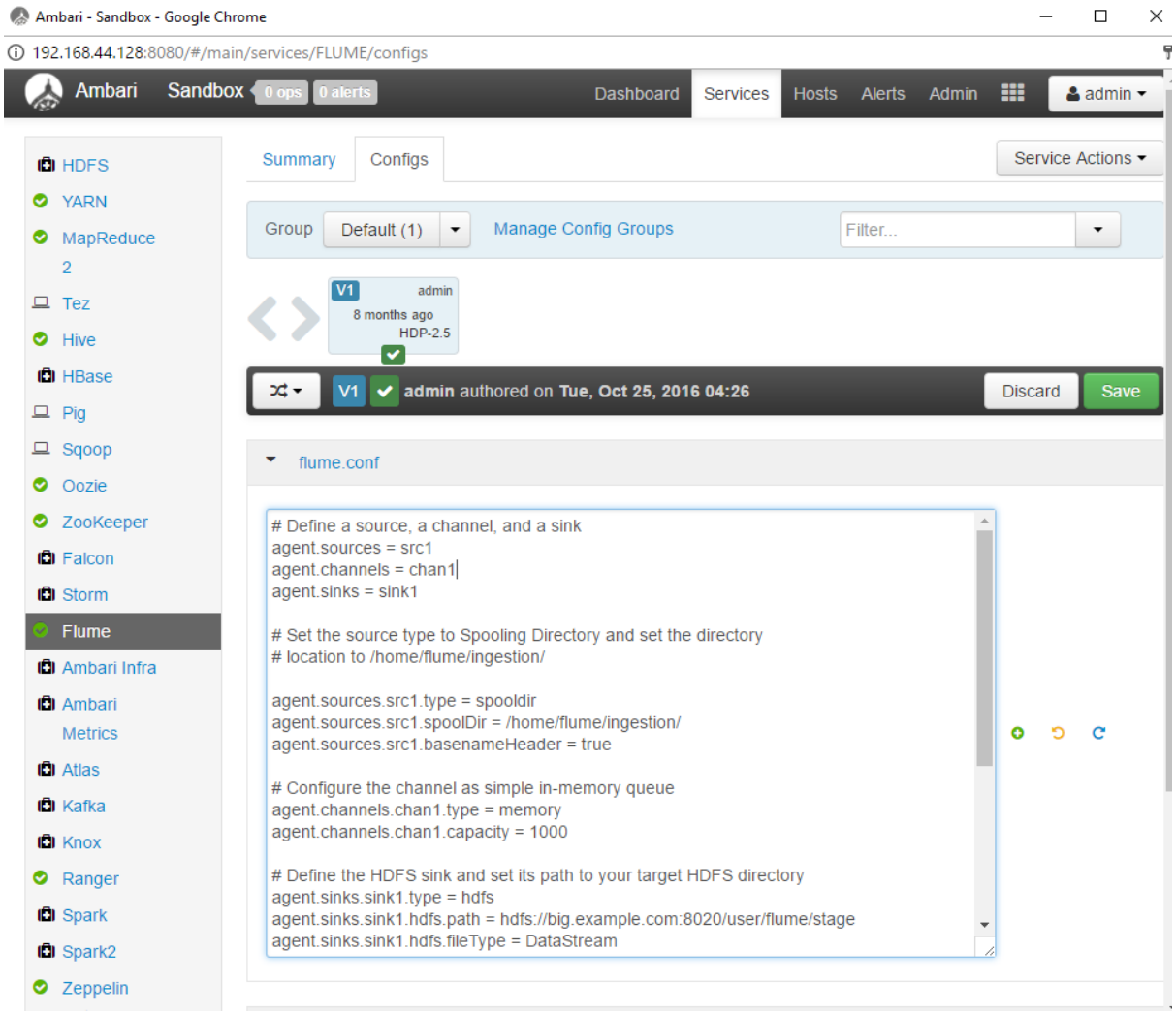
```
sqoop import --query 'SELECT a.*, b.* FROM a JOIN b on (a.id == b.id) WHERE $CONDITIONS' --split-by a.id --target-dir /user/foo/joinresults
```

Los datos siempre son importados en HDFS, pero como decíamos antes si le indicamos --**hive-import** en el comando logramos que se genera la metadata necesaria en H-Catalog para que las demás aplicaciones puedan accedar a los datos en forma de tabla, independientemente de su estructura original (.cvs, json, base de datos, ORC, etc.)

Una vez que ya tenemos los datos provenientes de bases de datos relacionales en Hadoop procedemos a importar nuevas fuentes de datos. En este caso incorporaremos archivos logs de un servidor web del cliente donde tiene funcionando su portal de comercio electrónico.

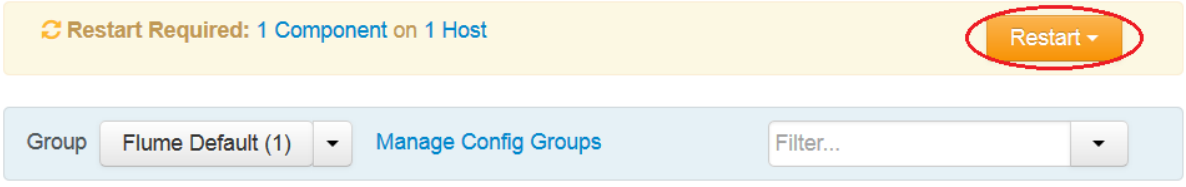
Para esto como definimos en la etapa de análisis y diseño usamos Flume, por lo que vamos a configurar el agente necesario desde Ambari:

Imagen 22 – Configuración de Flume



Allí especificamos la fuente de los datos, el tipo de canal (in-memory) y el destino en HDFS donde serán almacenados.

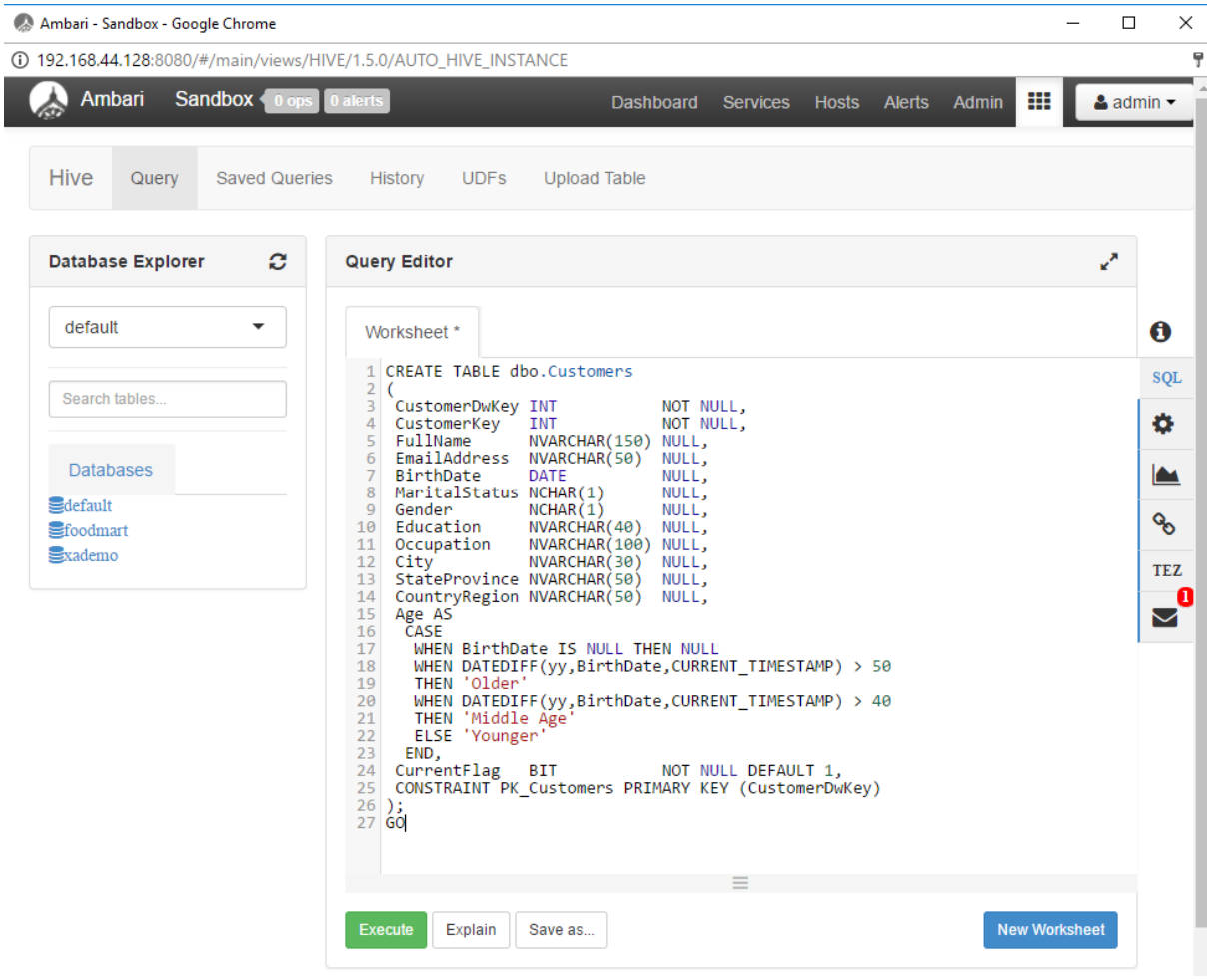
Tras configurar el agente reiniciamos el servicio:



Para implementar el requrimiento candidato podemos configurar otro agente de Flume para recibir información de nuestros productos directamente desde Twitter para luego ser analizada con el resto de los datos importados.

Una vez que tenemos todos los datos necesarios procederemos a crear nuestras tablas de hechos y dimensiones con los mismos procedimientos que nuestro cliente usa habitalmente para su EDW pero usando en este caso Hive.

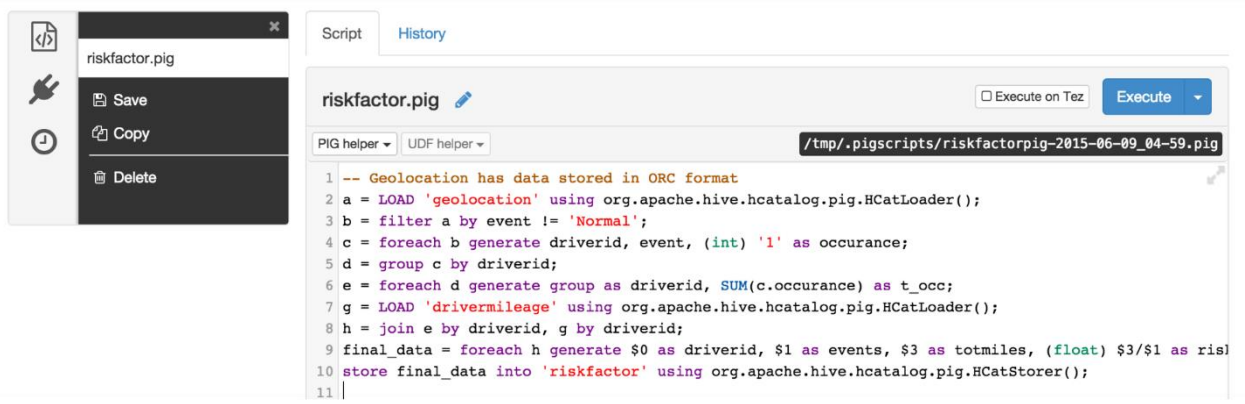
Imagen 23 – Modelado con Hive



Tras crear las tablas necesarias podemos usarlas pata hacer transformaciones de los datos mas complejas con Apache Pig y crear otras tablas que luego nos permitan hacer análisis mas

profundos, como cálculos de riesgo en ventas, patrones de comportamiento de usuarios según la navegación que hacen en el sitio de comercio electrónico de la empresa, análisis de sentimientos según las publicaciones en las redes sociales, entre otros.

Imagen 24 – Procesamiento con Apache Pig



Tras el modelado vamos a usar nuevamente Sqoop para exportar estas tablas al EDW del cliente mediante la herramienta Export:

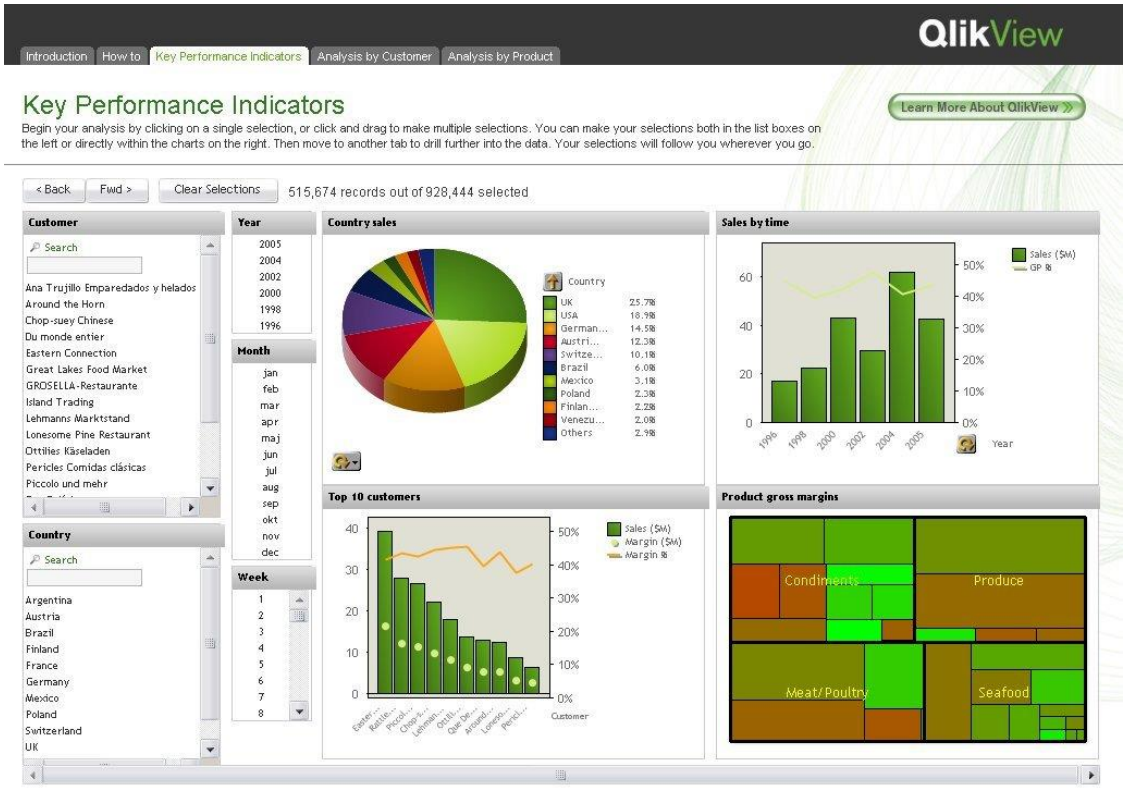
```
sqoop export --connect jdbc:sqlserver://db.example.com/foo --table bar  \
    --export-dir /results/bar_data
```

Aquí indicamos la tabla de la base de datos y el origen en HDFS.

Una vez procesada la información según los requerimientos podemos hacer los análisis que la nueva información permite mediante las herramientas de visualización que el cliente actualmente usa

En este caso se trata de QlikView para generar los reportes o gráficos necesarios que nos permitan hacer análisis y obtener conocimiento para la toma de decisiones. QlikView puede conectarse al EDW y pero tambien mediante el conector JDBC puede acceder directamente a los datos almacenados en Hadoop.

Imagen 25 – Análisis con QlikView



15. Conclusiones

Finalmente se le pudo demostrar la factibilidad técnica que le permite incorporar tecnología de Big Data a su EDW de una manera progresiva, sin generar demasiado impacto en lo económico ni en el esfuerzo requerido por su equipo de IT o el área de sistemas.

Se observó cómo una vez implementada la solución propuesta resulta más sencillo que antes incorporar nuevas fuentes de datos, almacenarlos y procesarlos sin utilizar hardware costoso, y también se reducen los tiempos en los procesos necesarios a la hora de implementar cambios en el modelo de datos y hacer nuevos análisis en función de las necesidaddes cambiantes de negocio y en el comportamiento de los clientes.

La solución propuesta, a través del análisis continuo de los logs de servidores, permitió rápidamente detectar potenciales defectos y problemas en las aplicaciones operacionales, y en el caso de los servidores web se logró un mejor conocimiento de sus clientes a la hora de realizar compras online, lo que permitió mejorar el portal de comercio electrónico en base a los análisis de los clics (clickstream) de los usuarios.

Tambien resultó interesante la propuesta de incorporar como fuente de datos a las redes sociales como Twitter, con el fin de capturar sensaciones y sentimientos en estas redes, por ejemplo, sobre nuevos lanzamientos de productos al mercado. Mediante técnicas de *sentiment analysis* se puede obtener rápidamente información muy valiosa a la hora de redefinir estrategias de marketing y re-diseño de productos y servicios.

Recomendaciones y líneas futuras de trabajo

Sin dudas la transformación digital en marcha es un proceso irreversible, no solo para las economías sino tambien para la cultura de las sociedades y las organizaciones. Y dentro de estas transformaciones Big Data es la tecnología central. Es por esto que se recomienda seguir invirtiendo en estas tecnologías, ampliando el cluster de servidores para lograr más capacidad de almacenamiento y procesamiento distribuido, y continuar de manera ininterrumpida con capacitaciones para que los ingenieros y analístas para que puedan desarrollar e incorporar en el corto y mediano plazo técnicas de analitica avanzada de datos, como Machine learning o Data mining, para extraer cada vez más información valiosa de la creciente cantidad de datos disponibles y mejorar el proceso de toma de decisiones de nuestro cliente.

La curva de aprendizaje de una tecnología como Hadoop y su ecosistema no es sencilla, es por eso que se recomienda un enfoque progresivo a medida que se potencia y fortalece el rol de data scientist en la organizacion, lo que aportará cada vez más inteligencia en el procesamiento y análisis de los datos que cada vez son más masivos, llegan a mayor velocidad y con los más variados formatos.

Logros

El principal logro de este trabajo es haber sido la puerta de ingreso para una pequeña o mediana organización en el proceso de transformación digital, poniendo en práctica las tecnologías de Big data hoy disponibles y que llegaron para quedarse, sin las cuales no les va a resultar fácil ser competitivos en un mercado global donde las empresas pioneras en adoptar los nuevos paradigmas ya estan haciendo uso de ellas desde hace años.

En lo personal este trabajo me permitió descubrir un futuro como profesional que me llevará a seguir investigando y aprendiendo sobre *Big data* y *Analítics* para poder deempeñar roles claves en las empresas de los próximos años.

16. Bibliografía

A Hortonworks White Paper. (March 2015). *Data Architecture Optimization with Hortonworks Data Platform.* Hortonworks. Recuperado el 01 de Junio de 2016

Accenture. (2014). *Big Success with Big data.* Recuperado el 2 de Diciembre de 2016, de https://www.accenture.com/us-en/insight-big-data-research

Accenture. (2014). *Las empresas consideran Big Data fundamental para su Transformación Digital.* Recuperado el 20 de Enero de 2017, de https://www.accenture.com/es-es/company-big-data-fundamental-transformacion-digital

Apache. (2013). *HDFS Architecture Guide.* Recuperado el 18 de Octubre de 2016, de https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Apache. (2013). *MapReduce Tutorial.* Recuperado el 18 de Octubre de 2016, de https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Apache. (2016). *Welcome to Apache Pig!* Recuperado el 19 de Octubre de 2016, de https://pig.apache.org/

Apache. (s.f.). *Ambari.* Recuperado el 17 de Octubre de 2016, de https://ambari.apache.org/

Apache. (s.f.). *APACHE HIVE TM.* Recuperado el 18 de Octubre de 2016, de https://hive.apache.org/

Apache. (s.f.). *Apache Sqoop.* Recuperado el 18 de Octubre de 2016, de http://sqoop.apache.org/

Apache. (s.f.). *Welcome to Apache Flume.* Recuperado el 18 de Octubre de 2016, de https://flume.apache.org/

Apache. (s.f.). *Welcome to Apache HBase.* Recuperado el 19 de Octubre de 2016, de https://hbase.apache.org/

Apache. (s.f.). *Welcome to Apache ZooKeeper.* Recuperado el 19 de Octubre de 2016, de https://zookeeper.apache.org/

Apache. (s.f.). *What is Apache Mahout?* Recuperado el 19 de Octubre de 2016, de http://mahout.apache.org/

Apache, H.-C. (2017). *H-Catalog*. Obtenido de https://cwiki.apache.org/confluence/display/Hive/HCatalog

Armah, N. A. (2013). *Big Data Analysis: The Next Frontier*. Bank of Canada Review. Recuperado el 03 de Agosto de 2016, de http://goo.gl/3yZMJ5

Azure, M. (2017). Obtenido de Linux Virtual Machines Pricing: https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/

Capgemini. (2013). The Principles of the Business Data Lake. Capgemini. Recuperado el 2 de Noviembre de 2016, de https://www.capgemini.com/resources/the-principles-of-the-business-data-lake

Castaño, M. (2016). *Big Data como pieza clave en el proceso de Transformación Digital.* Recuperado el 3 de Febrero de 2017, de Territorio Creativo: https://www.territoriocreativo.es/etc/2016/04/big-data-como-pieza-clave-en-el-proceso-de-transformacion-digital.html

CESSI. (2016). *Estado y Perspectivas de la Transformación Digital en las Empresas Argentinas*. Recuperado el 5 de Octubre de 2016, de http://cessi.org.ar/ver-noticias-estado-y-perspectivas-de-la-transformacion-digital-en-las-empresas-argentinas-2022

CESSI. (2016). *Mapa de la Transformación Digital.* Recuperado el 5 de Octubre de 2016, de http://cessi.org.ar/ver-noticias-mapa-de-la-transformacion-digital-2023

CPCIPC. (Mazo de 2017). *Tabla de honorarios*. Obtenido de Consejo Profesional de Ciencias Informáticas de la Provincia Córdoba: http://www.cpcipc.org.ar/content/honorarios

Dull, T. (2015). *Big Data Cheat Sheet on Hadoop.* Recuperado el 16 de Diciembre de 2016, de SAS: http://blogs.sas.com/content/customeranalytics/tag/big-data-cheat-sheet-on-hadoop/

Dull, T. (2016). *A Non-Geek's Big Data Playbook.* SAS. Recuperado el 16 de Diciembre de 2016

EDW. (2017). Obtenido de http://trisitsolutions.com/

Gartener IT Glossary. (2017). Obtenido de http://www.gartner.com/it-glossary/

Gartner. (2011). *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data.* Recuperado el 05 de Enero de 2017, de http://goo.gl/02llb9

Gartner. (2015). *Business Intelligence (BI).* Gartner glossary. Obtenido de http://www.gartner.com/it-glossary/business-intelligence-bi/

Gartner. (2015). *Data Warehouse.* Gartner glossary. Obtenido de http://www.gartner.com/it-glossary/data-warehouse

Global Center for Digital Business Transformation. (2015). *Digital Vortex: How Digital Disruption Is Redefining Industries.* Recuperado el 01 de Julio de 2016, de http://www.cisco.com/c/dam/en/us/solutions/collateral/industry-solutions/digital-vortex-report.pdf

Gruman, G. (14 de January de 2016). *What digital transformation really means.* Recuperado el 05 de Enero de 2017, de Infoworld: http://www.infoworld.com/article/3080644/it-management/what-digital-transformation-really-means.html

HBR. (2012). *Data Scientist: The Sexiest Job of the 21st Century.* Recuperado el 07 de Julio de 2016, de https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Hortonworks. (2016). *APACHE HADOOP*. Obtenido de http://hortonworks.com/apache/hadoop/

IBM. (2014). *Data warehouse augmentation.* Recuperado el 16 de Julio de 2016, de https://www.ibm.com/developerworks/library/ba-augment-data-warehouse1/index.html

Larissa T. Moss, S. A. (2003). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications.* Addison Wesley. Recuperado el 18 de Julio de 2016

McKinsey Global Institute. (2011). *Big data: The next frontier.* McKinsey Global Institute. Recuperado el 01 de Enero de 2017, de http://goo.gl/XbTLQs

Microsoft. (2009). *History of Business Intelligence.* Obtenido de Youtube: https://youtu.be/_1y5jBESLPE

Qlik. (2017). Obtenido de Meet QlikView: http://global.qlik.com/us

Sqoop, A. (2017). *Apache Sqoop.* Obtenido de http://sqoop.apache.org/

Tamara Dull, SAS. (2016). *SAS.* Recuperado el 28 de Febrero de 2017, de http://blogs.sas.com/content/customeranalytics/2015/08/14/marketers-ask-is-hadoop-enterprise-ready/

White paper, A. c. (2012). *Challenges and Opportunities with Big Data.* A community white paper developed by leading researchers across the United States. Recuperado el 18 de Junio de 2016, de http://goo.gl/t1HZSj

AUTORIZACIÓN PARA PUBLICAR Y DIFUNDIR TESIS DE POSGRADO O GRADO A LA UNIVERIDAD SIGLO 21

Por la presente, autorizo a la Universidad Siglo21 a difundir en su página web o bien a través de su campus virtual mi trabajo de Tesis según los datos que detallo a continuación, a los fines que la misma pueda ser leída por los visitantes de dicha página web y/o el cuerpo docente y/o alumnos de la Institución:

Autor-tesista *(apellido/s y nombre/s completos)*	César Zamboni
DNI *(del autor-tesista)*	18777068
Título y subtítulo *(completos de la Tesis)*	Integración de tecnologías Big Data en soluciones de Enterprise Data Warehouse
Correo electrónico *(del autor-tesista)*	cdzamboni@gmail.com
Unidad Académica *(donde se presentó la obra)*	Universidad Siglo 21

Datos de edición: *Lugar, editor, fecha e ISBN (para el caso de tesis ya publicadas), depósito en el Registro Nacional de Propiedad Intelectual y autorización de la Editorial (en el caso que corresponda).*	

Otorgo expreso consentimiento para que la copia electrónica de mi Tesis sea publicada en la página web y/o el campus virtual de la Universidad Siglo 21 según el siguiente detalle:

Texto completo de la Tesis *(Marcar SI/NO)*[1]	SI
Publicación parcial *(Informar que capítulos se publicarán)*	

[1] Advertencia: Se informa al autor/tesista que es conveniente publicar en la Biblioteca Digital las obras intelectuales editadas e inscriptas en el INPI para asegurar la plena protección de sus derechos intelectuales (Ley 11.723) y propiedad industrial (Ley 22.362 y Dec. 6673/63. Se recomienda la NO publicación de aquellas tesis que desarrollan un invento patentable, modelo de utilidad y diseño industrial que no ha sido registrado en el INPI, a los fines de preservar la novedad de la creación.

Otorgo expreso consentimiento para que la versión electrónica de este libro sea publicada en la en la página web y/o el campus virtual de la Universidad Siglo 21.

Lugar y fecha: _____

Firma autor-tesista	**Aclaración autor-tesista**

Esta Secretaría/Departamento de Grado/Posgrado de la Unidad Académica: _____certifica que la tesis adjunta es la aprobada y registrada en esta dependencia.

Firma Autoridad	Aclaración Autoridad

Sello de la Secretaría/Departamento de Posgrado