UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

DISEÑO DE UN CURSO TEÓRICO Y PRÁCTICO SOBRE: BIG DATA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INGENIERÍA DE REDES DE COMUNICACIONES

POR

MARIA FERNANDA LAVERDE SALAZAR

PROFESOR GUÍA:

ALFONSO EHIJO BENBOW

MIEMBROS DE LA COMISIÓN:

SEBASTIÁN RÍOS PEREZ

JORGE SANDOVAL ARENAS

Santiago, Chile
2015

RESUMEN DE LA TESIS PARA OPTAR AL TITULO DE: Magíster en ingeniería de redes y comunicaciones.
POR: María Fernanda Laverde Salazar
FECHA: 10/01/2016
PROFESOR GUÍA: Alfonso Ehijo

DISEÑO DE UN CURSO TEÓRICO Y PRÁCTICO SOBRE: BIG DATA

La veloz expansión del uso de la tecnología, genera un conjunto de desafíos en cuanto al manejo y análisis de grandes cantidades de datos que se generan a una gran velocidad, ya que se debe lidiar con situaciones vinculadas tanto con los datos, el software & hardware y además la relaciones entre clientes y proveedores de servicios.

El Big Data es una etapa en la era digital, y no representa un concepto aislado, ya que para su correcto aprovechamiento es necesario establecer una integración con los métodos de análisis de datos que permitirán sacar provecho a la información recolectada. La posibilidad de tomar decisiones y luego llevar a cabo acciones útiles a través de los resultados obtenidos, mediante herramientas de análisis de datos, es lo que constituye el núcleo del Big Data Analytics.

Tal como lo expresa Michael Minelli (coautor del libro Big Data, Big Analytics: "Big Data no es sólo un proceso para almacenar enormes cantidades de data en un data warehouse (…) Es la habilidad de tomar mejores decisiones y tomar acciones útiles en el momento preciso".

El trabajo de grado que se desarrolla a continuación corresponde al "Diseño e implementación de un curso teórico y práctico sobre Big Data". Dicho curso está orientado a alumnos de pregrado de la Universidad de Chile, y se basa en un diseño curricular siguiendo una metodología docente específica la cual se estructura en bloques de planificación y desarrollo. El programa se divide en módulos de aprendizaje definidos mediante temas y objetivos, y tiene una duración de cuarenta (40) horas en total entre clases teóricas (20 horas divididas en 10 clases) y prácticas (20 horas divididas en 5 laboratorios).

El objetivo general del curso, es integrar conocimientos relacionados con la forma de almacenar, administrar y aprovechar mediante herramientas específicas, el incremento sustancial del volumen de datos que se manejan diariamente, e inclusive cada segundo, en las empresas de tecnología y comunicación de las cuales, en su mayoría, día a día somos los principales generadores de data.

AGRADECIMIENTOS

Agradezco,

A mis padres y mi familia, por ser los pilares fundamentales de mi presente. A pesar de que hoy nos separa la distancia, siempre los he sentido cercanos en mi corazón. Gracias por el apoyo incondicional y el amor que me han dado. Los amo infinitamente hoy y siempre.

A Roger, por ser un compañero de vida excepcional, por siempre aconsejarme, entenderme, soportarme y lo más importante gracias por darme tu amor sincero.

A mi profesor Guía Alfonso Ehijo por guiarme en la realización de este proyecto. A los profesores del Magíster, que durante estos dos años compartieron sus conocimientos y experiencias para formarme como profesional.

A mis compañeros de Magíster por los momentos compartidos, las salidas, los asados, los paseos en bici y los buenos ratos que pasamos durante este tiempo.

A Chile, gracias por tanto.

TABLA DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1

INTRODUCCIÓN

En el presente capítulo se describen los aspectos preliminares del trabajo, destacando la motivación por la cual se llevó a cabo la investigación, el objetivo general, objetivos específicos, y la metodología utilizada.

1.1 MOTIVACIÓN

El Big Data analytics es un elemento clave en el análisis de datos, y es de esperarse, pues en los últimos años ha existido un incremento importante de dispositivos y aplicaciones cuyo funcionamiento nace y depende de las redes, generando así un aumento exponencial en el tráfico de datos. He aquí el dilema, como sociedad estamos tan acostumbrados al cambio constante de tecnologías que ya no nos damos cuenta cuando está ocurriendo.

Los operadores móviles, son el ejemplo perfecto para demostrar el gran tráfico de información que se encuentra viajando a través de las redes y que diariamente se almacenan en diferentes modalidades. El concepto en el cual todos los dispositivos (desde Smartphones hasta sensores) eventualmente estarán conectados gracias a la tecnología móvil, genera la necesidad de almacenar y analizar la información proveniente de las transacciones entre personas y entre dispositivos directamente. Un ejemplo práctico podría ser el de las estaciones móviles: cada torre o estación base está constantemente comunicándose con todos los teléfonos móviles que se encuentren en su área de cobertura, sólo imaginemos el nivel de data que se intercambia entre los billones de teléfonos que existen a nivel mundial, o para acortar un poco ésta cifra, imaginemos la gran cantidad de teléfonos móviles que una compañía puede tener bajo suscripción. Las personas usan sus teléfonos para un sinfín de tareas, por ejemplo: comunicación con otras personas mediante texto, voz, redes sociales, conexión a internet, descarga de aplicaciones, compartir archivos, e inclusive hasta para realizar transacciones bancarias (y justo ahora que los smartphones tienen dominado el mercado, éstas tareas aumentan cada vez más).

La situación de las empresas que brindan convergencia de servicios, es ideal para la aplicación de métodos de análisis de Big Data y promover beneficios del uso de herramientas tales como Hadoop, ya que no sólo se enfoca en solucionar el problema del manejo de grandes cantidades de datos, sino que también da herramientas para enfrentarse a varios aspectos importantes, tales como el marketing, las relaciones con clientes y la tarificación.

La data que se almacena puede proporcionar a los proveedores de tecnología, información invaluable acerca del comportamiento de la red y de los usuarios que la

utilizan; desde saber con certeza la experiencia de llamadas hasta saber cuáles son las probabilidades de abandono de un cliente. Este punto representa una parte vital del Big Data Analytics, ya que además de recolectar y almacenar la data, es posible obtener un valor real de la información mediante herramientas de análisis. El Big Data Analytics es el núcleo del Big Data, mediante el cual se obtienen la importancia y los grandes beneficios que hoy en día brinda a la sociedad digital.

Como seres comunicativos por naturaleza, en un entorno cada vez más demandante en el ámbito tecnológico, muchos pueden preguntarse ¿De dónde provienen los millones de bytes a los cuales denominamos Big Data? ¿Cómo podemos almacenar y analizar ésta cantidad impensable de datos? ¿Es posible aprovechar la data como mecanismo de mejora para empresas proveedoras de servicios móviles? Entre éstas y otras interrogantes, se encuentra la motivación principal del presente trabajo de grado, en el cual se generan respuestas a dichas interrogantes y se plasman en un curso teórico con experiencias prácticas para orientar a los oyentes en este nuevo mundo de la revolución de los grandes volúmenes de datos.

1.2 OBJETIVO GENERAL

El objetivo general del presente trabajo de grado consiste en diseñar e implementar un curso teórico y práctico sobre el Big Data orientado a operadores convergentes, para estudiar cómo el fenómeno del Big Data se puede utilizar para mejorar el rendimiento de los proveedores de servicios móviles, los cuales hoy en día brindan un portafolio amplio de prestaciones a los usuarios.

1.3 OBJETIVOS ESPECÍFICOS

- Realizar el estudio de las bases teóricas y prácticas que se adapten al contenido del programa de curso.
- Desarrollar un programa con metodología de aprendizaje basado en competencias por módulos de instrucción.
- Cuantificar el programa en diez (10) clases teóricas de dos (2) horas cada una, y cinco (5) laboratorios de cuatro (4) horas cada uno, sumando un total de cuarenta (40) horas académicas.
- Plasmar los contenidos teóricos y prácticos en presentaciones que faciliten la transmisión de la información.
- Establecer el programa de evaluaciones.
- Realizar pruebas de las experiencias prácticas para validar la metodología utilizada.

1.4 METODOLOGÍA

La metodología y estrategia utilizada para el desarrollo del trabajo de grado, de forma generalizada, consideró las siguientes actividades:

- Investigación y recopilación de información relacionada con el Big Data.
- Análisis de procesos para el desarrollo eficiente de cursos y metodologías docentes.
- Aplicación de metodologías docentes para la definición de unidades programáticas, módulos de aprendizaje y régimen de evaluaciones.
- Desarrollo del contenido teórico y práctico del curso
- Elaboración de material de apoyo para la ejecución de las clases.
- Validación del contenido práctico materializado en actividades de laboratorio.
- Materialización de todos los contenidos anteriores en el trabajo de grado.

1.5 DESCRIPCIÓN DE CONTENIDOS

- Para un mejor seguimiento y comprensión del presente trabajo, se divide en ocho capítulos concisos donde se explican los diferentes tópicos relacionados con la investigación.
- El Capítulo I corresponde al actual contenido, donde se explicaron los objetivos y motivación del trabajo.
- El Capítulo II, coincide con los antecedentes de los tópicos relacionados al trabajo de grado, donde se establecen las bases teóricas necesarias para la comprensión de los temas.
- En el Capítulo III, se desarrolla la metodología general a seguir para el diseño e implementación del curso teórico-práctico, comenzando por la planificación de los módulos correspondiente a cada segmento de aprendizaje basado en las metodologías docentes y la planificación curricular que corresponde.
- El Capítulo IV contiene los resultados obtenidos según las experiencias prácticas realizadas y el análisis de impacto de las tecnologías aplicadas en el ámbito de los operadores convergentes.
- Los capítulos V, VI y VI corresponden a la discusión del trabajo, las conclusiones y las referencias bibliográficas respectivamente.

2

ANTECEDENTES

En el presente capítulo se establecerán las bases teóricas fundamentales para comprender el Big Data de una forma global. Se detalla su evolución y desarrollo, estructura, herramientas y lenguajes utilizados, Big data Analytics, así como también conceptos importantes que permiten una interpretación más completa de lo que en el siguiente trabajo se desarrolla.

2.1 FENÓMENO DEL BIG DATA

Big data es la nueva generación de almacenamiento, análisis de datos (data warehousing) y análisis de negocios. La mejor parte de este fenómeno es el paso adelantado en la innovación y el cambio, pero esta nueva parte de la era digital no surgió bruscamente, por el contrario, ha estado mostrándose durante algún tiempo. El manejo de grandes cantidades de datos ha sido utilizado por décadas en la industria, manejando toneladas de data transaccional a través de los años usando métodos de almacenamiento a papel y guardándolos en estantes de oficina. [1]

Hoy en día, en el fenómeno del Big Data no sólo se consideran fuentes de datos a los generados por las transacciones o documentos tales como se hacían en tiempos pasados, si no que las principales fuentes actuales de generación de data son las personas per se, sin dejar de lado a los equipos de la era digital, los cuales generan un gran cantidad de data con sólo encenderlos. Por ejemplo los smartphones, los servidores de datos generan continuamente mensajes de log para dar información acerca de las actividades y status del servicio; los equipos de ciencia generan data de medidas detalladas de proyectos y experimentos; las compañías recaudan información acerca de las ventas, operaciones, clientes, preferencias etc. (por ejemplo: Google, Amazon)

Otro ejemplo de rápido crecimiento, son los nodos de sensores conectados en red, los cuales están presentes hoy en día en los sectores industriales, de transporte, automotriz, entre otros. Y por último, no podemos dejar de lado uno de los mayores generadores de data a nivel mundial, la Internet.

En un estudio publicado recientemente por la IDC, se estima que la tecnología de Big Data y el mercado de servicios, crecerá en un 26.4% lo cual representa un cifra de 41.5 billones de dólares en el año 2018, o alrededor de seis veces la tasa de crecimiento del mercado de la tecnología de la información.

2.2 EVOLUCIÓN DE LA INFORMACIÓN – 5VS DEL BIG DATA

Para explicar las etapas evolutivas por la cual ha pasado la información que hoy en día generamos, organizamos y almacenamos, es necesario tener en cuenta ciertas características que definen el llamado fenómeno del Big Data.

Douglas Laney en su informe introducido en 2001, define tres de los parámetros más importantes en el Big Data, llamados popularmente "Las 3 Vs del Big Data" [2], las cuales corresponden a Volumen, Velocidad y Variedad. Hoy en día se han añadido dos parámetros adicionales que complementan la explicación del fenómeno del Big Data, éstas son la Veracidad y el Valor de la data.

2.2.1 Volumen

Se refiere a la gran cantidad de data generada cada segundo. Emails, Twitters, fotos, videos, Data de sensores, mensajería instantánea y otros muchos datos que se producen y comparten cada segundo.

Este incremento hace que la data sea muy extensa para almacenarla y analizarla usando tecnología de base de datos tradicional. El avance de la tecnología del Big Data, ha permitido almacenar y analizar la data con la ayuda de sistemas distribuidos, donde parte de la data es almacenada en diferentes ubicaciones y luego es reunida nuevamente mediante softwares especializados.

Hoy en día, a pesar de que se habla de cantidades enormes de data, no hay una idea concreta de cuantificación cuando se debe considerar un cierto volumen como "Grande". Por ejemplo, hace 10 años atrás se consideraba que 500 Terabytes era una cantidad enorme de data, hoy en día 1 Petabyte es considerado una cantidad considerable de data, y el promedio va aumentando más y más hacia términos como Exabyte e incluso Zettabyte.

Sin embargo, el volumen de datos no sólo viene de fuentes adicionales de data[1], si no también está relacionado con un cambio de mentalidad de empresas y particulares que lleva a incrementar el volumen de data. Muchas compañías y empresas consideran la data como un activo importante de su organización, aprovechando al máximo las ventajas que puedan obtener.

Un ejemplo claro de producción masiva de data, es el resultado de la interacción entre dispositivos, conocido como M2M[2]. Según estudios recientes realizados por el GSMA[3], establecen que las conexiones de M2M alcanzaron los 195 millones en 2013, creciendo a una tasa de 40% por año entre 2010 y 2013,

[1] Sensores, servidores, Smartphones, redes sociales, equipos electrónicos
[2] M2M: Machine to Machine
[3] GSMA: GSM Association

esperándose un crecimiento a 250 millones este año. En el 2013 M2M representó el 2.8% de las comunicaciones móviles globales, doblando el 1.4% del 2010.

Siguiendo con los datos de la investigación de GSMA, alrededor de 428 operadores móviles ofrecen servicios de M2M alrededor de 187 países, lo cual equivale a un 40% de los operadores móviles mundiales. [3]

Por otra parte, el número de dispositivos conectados a la red cada día se incrementa en una forma exponencial, lo cual está demostrado por muchos estudios e investigaciones realizadas que analizan la curva de crecimiento de ésta tecnología. Tal es el caso del estudio realizado en el año 2011 por la misma organización GSMA, en el cual se establecía que el en el 2011 existían un aproximado de nueve millones (9) de dispositivos y se esperaba que para el año 2020, esa cifra aumentara a veinticuatro (24) billones en 2020. Estos dispositivos generan gran cantidad de data no estructurada incluyendo: datos de temperatura por ejemplo en los sensores, ubicación, porcentaje de humedad, sonido, preferencias, y un sinfín de data que pueda ser útil dependiendo del área de estudio. [3]

2.2.2 Velocidad

La velocidad se refiere a dos cosas principalmente: la primera es el procesamiento de la data tomando en cuenta la velocidad a la cual fue creada; y la segunda se refiere a la necesidad de entregar el valor de dicha dentro de un cierto período de tiempo.

Tomando en cuenta el masivo crecimiento de la data (Volumen), bien sea por el surgimiento de nuevas fuentes de data o por la evolución de la mentalidad de las personas hoy en día, surge un desafío que se refiere al análisis de data con la mayor rapidez posible.

El desafío yace en primer lugar, cuando se necesita procesar una gran cantidad de data mientras se mantiene un estado consistente del servicio. Una vía de manejar el problema, es filtrar la data, descartando lo innecesario y sólo almacenando las piezas importantes. Igualmente, el filtrado per se consumirá recursos y tiempo mientras procesa la data. En el caso que no sea posible realizar el filtrado de data (lo cual puede ocurrir), surge la necesidad de extraer y almacenar automáticamente la metadata[4] junto con la data en tiempo real.
En segundo lugar, hay que tomar en cuenta la puntualidad de la extracción de la información, al igual que el análisis de la misma. En muchas situaciones, el análisis en tiempo real es necesario para tomar acciones antes que la información pierda su valor. Sin embargo, en muchas oportunidades no es suficiente el análisis de los datos y la extracción de la información, sino que también es necesario tomar acciones inmediatas para aplicar la idea o plan sobre dicha información. [4]

[4] Metadata: Información descriptiva el contenido de un archivo, objeto o dato en particular.

2.2.3 Variedad

El gran impulsador de Big Data, es el potencial que tiene para usar diferentes fuentes de data, y combinar e integrar dichas fuentes como base para el Big Data Analytics. La data puede ser dividida en dos grandes grupos: data estructurada y data no estructurada. El análisis tradicional de data, se enfoca principalmente en la data estructurada.
Para almacenar y analizar data no estructurada, la necesidad de agregar nuevos atributos que soporten nuevos tipos de data sin cambiar la estructura de las bases de datos, está aumentando. [5]

Estas nuevas formas de data, se generan en fuentes tales como páginas web, redes sociales, mensajería instantánea, emails, datos de sensores, comunicación M2M y un sinfín de nuevas tecnologías que cada segundo generan data que no está contemplada dentro de tablas organizadas o que cumplen con patrones estructurados.

2.2.4 Veracidad

Para afrontar los desafíos en cuanto al desempeño y la capacidad que surgen de la falta de veracidad, es importante tener estrategias y herramientas de calidad como parte de una infraestructura de Big Data. Esto incluye evaluar el uso deseado del Big Data dentro de la organización y determinar qué tan precisa debe ser la data para así cumplir la meta en un caso determinado.

2.2.5 Valor

El valor de la data se refiere al procesamiento de la data y las ideas producidas durante el análisis, ya que la data está típicamente ligada con una meta o beneficio inmediato. Esto no quiere decir que el valor de la data está limitado sólo a un primer uso o al análisis inicial, por lo contrario el valor completo de la data está determinado por el posible análisis futuro de las tareas, la forma en cómo se realiza y cómo la data es usada a lo largo del tiempo.
La Data puede ser reusada, extendida y recombinada con otro conjunto de data, ésta representa una de las razones por las cuales hoy en día más y más empresas ven la data como un activo para su organización, y la tendencia es coleccionar la data potencial sin importar que en ese preciso momento no la necesiten inmediatamente, por lo contrario la guardan asumiendo que en algún futuro será de utilidad para ofrecer algún valor.

2.3 HADOOP

Hadoop es para muchos un sinónimo de Big Data debido a sus grandes capacidades de manejar grandes cantidades de data (no estructurada) en un período

de tiempo muy pequeño y de una forma económicamente responsable. Por ésta razón el ecosistema de Hadoop juega un rol mayor en el Big Data Analytics.

2.3.1 Introducción a Apache Hadoop

Hadoop fue un proyecto creado por Doug Cutting, el creador de Apache Lucene. Hadoop proporciona un conjunto de herramientas entendibles para construir sistemas distribuidos, incluyendo almacenamiento de data, análisis y coordinación de data. [6]

Hadoop se originó de un proyecto llamado Apache Nutch, un motor de búsqueda en la web de código abierto. Después de un tiempo, al darse cuenta que las estructuras existentes no escalarían a los billones de páginas en la web, los iniciadores del proyecto desarrollaron una implementación en código abierto basada en el sistema distribuido de Google, y lo llamaron Nutch Distributed File System (NDFS).

En 2004, Google lanzó un nuevo reléase que introducía a MapReduce, el cual consistía en modelo de programación paralelo y una implementación adicional para procesamiento, análisis y generación de grandes estructuras de data a través de un clúster realizado con equipos de bajo costo (llamados en inglés Commodity Hardware). Alrededor de un año después del lanzamiento del reléase de Google, todos los algoritmos Nutch comenzaron a usar MapReduce y NDFS. En 2006, Nutch se convirtió en un sub-proyecto separado bajo el nombre de Hadoop, y dos años después se convirtió en un proyecto de alto nivel de Apache, confirmando así su éxito.

Hadoop soporta varios subproyectos bajo la licencia Apache, y proporciona y soporta el desarrollo de softwares de licencia libre que brindan un framework para el desarrollo de aplicaciones computacionales distribuidas altamente escalables.

Los sub-proyectos que corren bajo Hadoop son:

- ***Hadoop Core:*** Proporciona un sistema de archivos distribuidos (HDFS) y MapReduce.
- ***HBase:*** Está construido en el Hadoop Core para otorgar una base de datos escalable y distribuida.
- ***Pig:*** Es un lenguaje de alto nivel de flujos de datos y un framework de ejecución para computación paralela. Este es configurado en el top de Hadoop core.
- ***ZooKepeer:*** Es un sistema de coordinación confiable y altamente disponible. Las aplicaciones distribuidas usan ZooKeeper para almacenar y mediar actualizaciones.
- ***Hive:*** Es una infraestructura de data warehouse construida en el Hadoop Core que proporciona agregación de data.
  (Ver Anexo I – Título 1)

2.3.2 Estructura de Apache Hadoop

La estructura de Hadoop consiste en un conjunto de máquinas que corren HDFS y MapReduce (YARN), las cuales conforman un Clúster. Cada máquina individual se conoce como Nodo.

Un clúster puede tener uno o miles de nodos, y tiene un escalamiento lineal, es decir, por cada nodo que se agrega se genera una capacidad y desempeño proporcional a la cantidad de nodos. Existen dos tipos particulares de Nodos que se diferencian por la clase de proceso que corre en ellos. Están los Master Nodes, que corren los procesos globales de administración; y los Worker Nodes, que corren la data y procesos de aplicaciones locales.

El proyecto Apache Hadoop, tal como se menciona en definiciones anteriores, fue desarrollado con el objetivo de crear un sistema distribuido de cómputo y programación que permitiera un desarrollo más fácil de aplicaciones, y cuya filosofía es proporcionar una gran escalabilidad sobre clusters construidos de hardware económico (llamados "commodity hardware").

La estructura de Hadoop está motivada y basada en gran parte en los papers publicados por Google en los cuales se mostraban, de forma libre, las estructuras usadas para el análisis de data. Tales son los ejemplos de Google File System y Google MapReduce. De acuerdo con lo dicho anteriormente, el core de Hadoop consiste en el HDFS (Hadoop Distributed File System) y el Hadoop MapReduce. El HDFS, se refiere al sistema distribuido de archivos de Hadoop, mientras que el MapReduce, es un algoritmo de reducción de grandes cantidades de data.

Típicamente, tanto HDFS como MapReduce, trabajan juntos en un clúster, ya que la data entrante proveniente de MapReduce, es almacenada en una instancia de HDFS en el mismo clúster. Sin embargo, esto no siempre se cumple, ya que HDFS puede ser usado sin MapReduce. Hadoop puede integrar distintos sistemas de archivos distribuidos, por ejemplo el llamado KFS[5].

2.3.3 Hadoop Distributed File System (HDFS)

El almacenamiento de data en el ecosistema de Hadoop, mayormente, es logrado mediante el uso de un sistema distribuido de archivos conocido como HDFS. Los archivos son divididos en bloques de menor tamaño los cuales son esparcidos a través de múltiples nodos en el sistema, lo cual hace posible procesar los bloques de archivos en paralelo [7]. Adicionalmente, este sistema añade dos funcionalidades:

- Escalabilidad: Permite añadir nuevos nodos al sistema
- Alta disponibilidad: Permite replicar los bloques a través de múltiples nodos.

[5] KFS – Kosmos File System

El HDFS se ha convertido hoy en día en un estándar para el procesamiento de datos a grandes escalas, y es usado por muchas grandes firmas en la industria. [8]

2.3.4 MapReduce

MapReduce es un modelo de programación usado ampliamente en ambientes donde se necesita procesar grandes cantidades de data en una vía altamente paralela. Las implementaciones de MapReduce están basadas en un modelo master-slave. La falla del esclavo está administrada mediante la reasignación de sus tareas hacia otro slave, mientras que las fallas que se produzcan en el master, no son administradas por implementaciones directas de MapReduce.

2.4 BIG DATA

El término Big Data ha surgido en los últimos años para hacer referencia a una gran cantidad de datos de diversa índole generados en los sistemas de comunicación actuales.

> "Big Data se refiere a la data cuyo tamaño está más allá de la capacidad de las herramientas de software de las típicas bases de datos para capturar, almacenar, administrar y analizar" [9]

Otra de las definiciones que engloban el significado de Big Data se refiere a la establecida por la IDC[6]en su estudio de "The Digital Universe" la cual cita:

> "IDC define las tecnologías del Big Data como una nueva generación de tecnologías y arquitecturas, designadas para económicamente extraer valor de grandes volúmenes de una extensa variedad de data, estableciendo capturas, análisis y descubrimientos de alta velocidad" [10]

Cuando se habla de Big Data, hay que tener presente que en su mayoría se trata de data no estructurada, a diferencia de las grandes base de datos basadas en análisis SQL. Esto es importante, ya que la data estructurada y no estructurada, difieren totalmente en la manera de estudiarlas y analizarlas. Existen tres características principales cuando se habla de Big Data, estas son: la data propiamente, el análisis de la data, y la presentación de los resultados del análisis.

La cantidad de data que es generada diariamente, ha aumentado significativamente especialmente desde el año 2000, año en el cual se considera que las tecnologías comenzaron a cambiar de análogo a digital.

[6] IDC International Data Corporation

2.4.1 Fuentes de data

Uno de los problemas al establecer una solución de Big Data, comienza cuando las capas de fuentes de datos de diferentes volúmenes, velocidades y variedades conviven conjuntamente y deben ser incluidas en el conjunto final de Big Data para ser analizados. En la Figura 1 se muestra algunas de las fuentes actuales de generación de data. [6] (*Ver Anexo I – Título 2*).

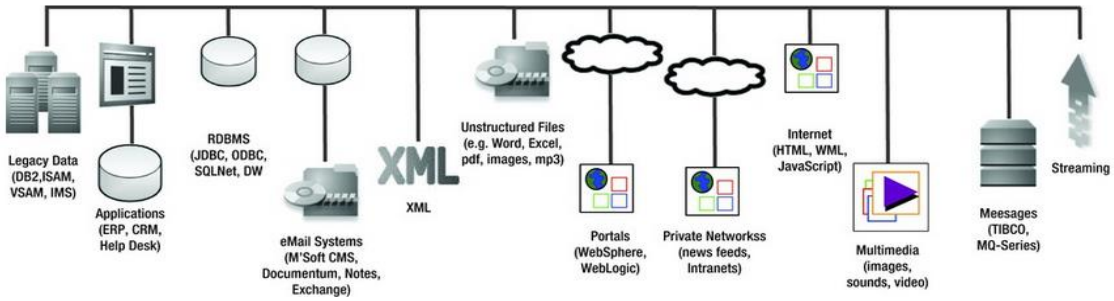


Figura 1 Fuentes de data

2.4.2 Arquitectura

Antes de incursionar en el mundo del Big Data, es necesario establecer una arquitectura con los componentes necesarios para formar el stack de la solución. Una arquitectura de administración para Big Data, debería ser capaz de consumir una infinidad de fuentes de data en una forma rápida y económica. En la figura 2.1 se muestra el diagrama arquitectural de los principales elementos que hacen posible el funcionamiento integral del stack. (*Ver Anexo I – Título 3*)

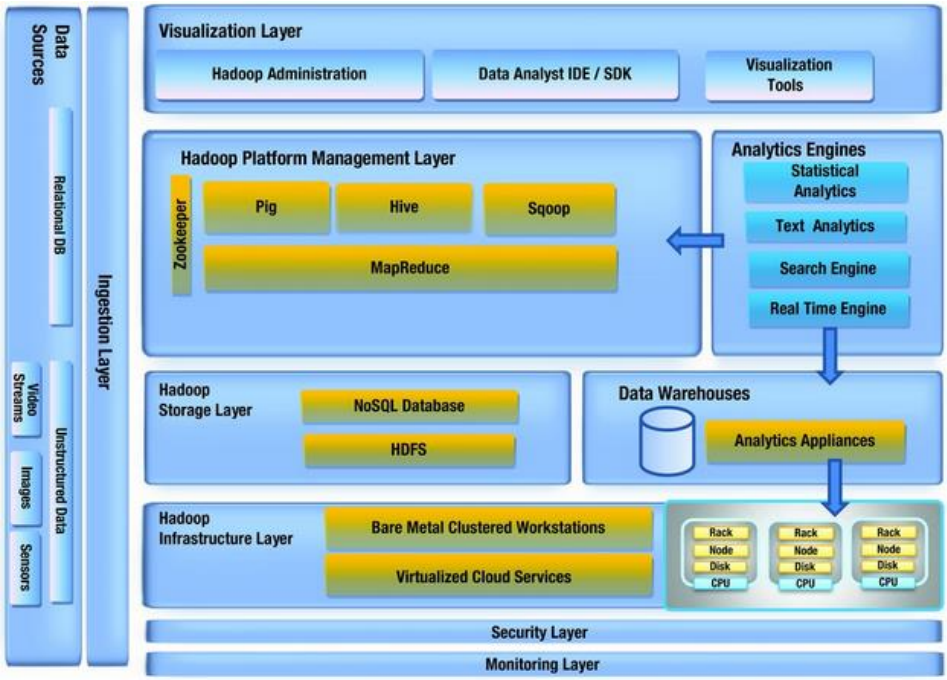


Figura 2 Modelo de arquitectura para Big Data

2.5 BIG DATA ANALYTICS

En el ambiente en el cual nos desarrollamos hoy en día, cada vez se generan mayores cantidades de data debido al crecimiento exponencial de equipos tecnológicos que de por sí con sólo encenderlos, generan tráfico de información. Por ejemplo, los usuarios de las redes móviles y fijas diariamente suben contenido a la red, bien sea mediante la publicación de fotos, videos en las redes sociales, blogs, almacenamiento en la nube, los datos generados por equipos médicos, educativos, empresas, usuarios de internet, encuestas, representan una cantidad bastante amplia de datos, la cual debe ser almacenada y analizada para obtener patrones de comportamiento e información relevante que puede ser usada para diversas situaciones. Éste proceso de análisis de data, forma parte del Big Data Analytics.

Una definición acertada del Big Data Analytics, la expresa Michael Mellini:

> "Big Data Analytics es el resultado natural de cuatro grandes tendencias globales: **Ley de Moore** (la cual básicamente dice que la tecnología siempre será más barata); **Computación móvil** (Smartphones, Tablets); **Redes Sociales** (Facebook, Twitter, Pinterest, etc.); y **Cloud Computing** (ya no necesitas tener tu propio hardware o software, puedes rentar o usar el de alguien más)"

El Big Data no es sólo una descripción de volumen de data, el foco real es el uso que se le da a dicha data. Según David Smith, el verdadero desafío es identificar

o desarrollar métodos confiables y efectivos desde el punto de vista económico para extraer el valor de todos los terabytes y petabytes de la data disponible hoy en día. Ahí es donde Big Data Analytics se convierte en algo necesario. Comparar los métodos de análisis tradicionales de datos con Big Data Analytics, es como comparar un carro llevado por caballos con un tractor de remolque. Las diferencias en velocidad, escala y complejidad son enormes. [1]

2.5.1 Tipos de Analytics

Para obtener valor de la data recolectada, es necesario que las organizaciones apliquen una estructura sistemática de análisis para obtener el mejor resultado y crear ideas que ayuden a tomar las mejores decisiones.

La estructura describe diferentes tipos de métodos de análisis:

Analytics Descriptivo: Consiste en el nivel más simple de análisis. Recolecta y divide la data en partes más pequeñas para facilitar el análisis. Describe el estado actual de la organización de acuerdo a tendencias y patrones. El análisis descriptivo fue el primer método utilizado antes del surgimiento del Big Data Analytics.

Analytics Diagnóstico: Consiste en el estudio de la data para validar/rechazar las hipótesis establecidas. Entender el porqué de los eventos.

Analytics Predictivo: Consiste en el modelamiento de la data para determinar posibilidades futuras. Utiliza técnicas de estadísticas, modelos, machine learning.

Analytics Prescriptivo: Consiste en establecer mecanismos de mejoras de acuerdo al análisis realizado en el análisis predictivo. Responde a las preguntas ¿Qué debería pasar ahora? ¿Cómo puedo mejorar los niveles de la organización? Etc.

2.5.2 Software usado en el análisis de data

Para analizar grandes volúmenes de data se necesitan herramientas que sean capaces de estructurar la data de forma que se puedan estudiar y generar algoritmos y patrones de estudio, usando base de datos especializadas y softwares orientados al análisis de datos con crecimiento masivo.

Algunos de los softwares comúnmente usados para el análisis de la data son: SAS, SPSS, y R. [11]

(*Ver Anexo I – Título 4*)

2.5.3 Big Data Analytics – Enfoque económico

El uso del Big Data Analytics, se está convirtiendo en un punto clave para la competencia y crecimiento en la industria, ya que la mayor parte de los sectores-incluyendo el sector público, salud, retail, marketing, ingeniería etc – pueden beneficiarse del resultado proveniente del análisis de las grandes cantidades de data que se generan en su funcionamiento. La recolección y el análisis de data, otorga a las organizaciones mayores ideas en cuanto a las preferencias y comportamientos de sus clientes o usuarios, de esta forma dicha data puede ser usada como base para la creación de productos y servicios. Esto permite a las organizaciones solucionar problemas emergentes de una forma más competitiva y en menor tiempo. [12]

Big Data representa un concepto innovador para muchas organizaciones, lo cual implica que al introducirlo dentro de sus procesos naturales, se deben tomar en cuenta una serie de parámetros tales como almacenamiento, escalabilidad, diseño de centro de datos mejorados entre otros. Este proceso, mediante el cual se rompen paradigmas antiguos para dar paso a nuevas tecnologías, involucra costos asociados a inversión de hardware, software, personal calificado, y soporte necesario; elementos que afectan el balance económico de la organización. Esto significa que el Retorno de Inversión (ROI[7]) y el Costo Total de Propiedad (TCO[8]) son elementos claves para elaborar un plan de negocio con Big Data. [12]

Para desarrollar un caso de negocios ligado al uso del Big Data Analytics, no hay una fórmula establecida, pero hay ciertos elementos que pueden ser usados para definir como debería estructurarse un caso de negocios para asegurar el éxito. Un caso sólido de negocio para Big Data Analytics debería incluir los siguientes ítems:

- Antecedentes completos del proyecto.
- Beneficios del análisis.
- Opciones posibles
- Alcance y costos
- Análisis de riesgos [12]

La mejor forma para entender el enfoque económico que el Big Data Analytics tiene hoy en día, es con ejemplos prácticos de la industria.

El Marketing es un ejemplo palpable de cómo el Big Data Analytics puede generar grandes beneficios. Hoy en día, la mayoría de las empresas tienen la capacidad de almacenar la información que sus usuarios generan a través de la red. Combinando dicha data con métodos específicos de análisis, las empresas pueden predecir, de una forma bastante precisa, el comportamiento de cada usuario, lo cual genera un importante activo a la hora de realizar campañas publicitarias orientadas

[7] ROI: Return of Investment

[8] TCO: Total Cost of Ownership

directamente a las necesidades ya estudiadas de los usuarios, obteniendo una mayor probabilidad de éxito e ingresos.

2.5.4 Factores que impulsan el Big Data

Los factores impulsadores de Big Data están directamente relacionados con la agilidad en la utilización y análisis de colecciones de data y flujos para obtener valor: incrementar ganancias, bajar costos, mejorar la experiencia del usuario, reducir riesgos, e incrementar la productividad. La explosión de la data choca contra el requerimiento de capturar, administrar y analizar información. [5]
Algunas tendencias claves que manejan la necesidad de plataformas de Big Data incluyen los siguientes factores:

- **Incremento de volúmenes de data que necesitan ser capturados y almacenados**: la escala de crecimiento supera la capacidad razonable de las base de datos tradicionales.
- **Rápida aceleración del crecimiento de la data:** Según la IDC, "Desde 2005 al 2020, el universo digital crecerá en un factor de 300, desde 130 exabytes hasta 40000 exabytes, o 40 trillones de gigabytes (más de 5200 gigabytes por cada hombre, mujer y niño en 2020). Desde ahora hasta el 2020, el universo digital será aproximadamente el doble cada dos años".
- **Incremento de volúmenes de data introducidos dentro de la red:** De acuerdo el Visual Networking Index Forecast de Cisco, para el 2016, el tráfico IP global está estimado que sea de 1.3 zettabytes[9]. Este incremento del tráfico de la red es atribuido al crecimiento del número de smartphones, tablets y otros dispositivos que hacen uso de internet , el crecimiento de la comunidad de usuarios de internet, el incremento del ancho de banda ofrecido por los carriers de telecomunicaciones, y la proliferación de la disponibilidad y conectividad de WiFi.
- **Crecimiento en la variación de tipos data para el análisis:** Alguna de las fuentes de data pueden reflejar elementos mínimos de estructura, mientras que otros pueden estar completamente no estructurados o incluso limitado a formatos específicos. Es por esto que para extraer información útil de toda esa data, las empresas deben mejorar sus alcances de administración de la data.
- **Métodos alternos y no sincronizados para facilitar la entrega de la data:** En un ambiente estructurado, hay lineamientos claros de las tareas discretas para la adquisición de la data o intercambio, tales como transferencias de archivos vía discos o mediante el protocolo de internet. Hoy en día, la publicación y el intercambio de data funciona de una forma impredecible con picos y valles, con datos provenientes de un amplio espectro de fuentes conectadas tales como páginas web, sistemas de procesamiento, e incluso "data abierta" y flujos provenientes de redes sociales. Esto crea presión para la rápida

[9] Zettabytes: 10^21 bytes

adquisición, absorción, y análisis manteniendo la consistencia a través de los diferentes data sets.

- C**reciente demanda por resultados en tiempo real del análisis de data:** Entregando información a diferentes áreas de negocios para análisis simultáneos, da nueva información y capacidades que nunca existieron en el pasado, permitiendo a los compradores revisar patrones de compra y tomar decisiones más precisas de acuerdo al producto.

3

METODOLOGÍA

En este capítulo se describe la metodología utilizada para alcanzar el objetivo general planteado en capítulos anteriores.

La metodología general se divide en tres grandes componentes; la primera corresponde a las metodologías docentes la cual explica los métodos y recursos utilizados para concretar el marco de enseñanza, la segunda corresponde a las metodologías relacionadas con el desarrollo e implementación de los módulos prácticos que se desarrollan en el programa y por último la metodología aplicada para el aprendizaje basado en competencias.

3.1 METODOLOGÍAS DOCENTES

Para el desarrollo y aplicación del curso, fue necesario aplicar un método de planificación que se basara en el aprendizaje basado en competencias (ABC) por módulos de objetivos, y así de ésta forma facilitar la comprensión y análisis global de los distintos tópicos relacionados con el curso.

El CERI (*Centre for Educational Research and Innovation*), establece una definición general de la metodología docente, la cual define como:

> "(..) Una búsqueda sistemática y original, asociada con el desarrollo de actividades con la finalidad de incrementar el caudal de conocimientos sobre la educación y el aprendizaje, y la utilización de ese conocimiento acumulado para promover nuevas aplicaciones o para mejorar el esfuerzo deliberado y sistemático en aras de transmitir, evocar o adquirir conocimiento, actitudes, habilidades y sensibilidades, y cualquier tipo de aprendizaje que resulte de este esfuerzo"

Figura 3 Metodología docente

La metodología docente utilizada en este proyecto, está basada en el aprendizaje por módulos de instrucción, y se podría esquematizar en cuatro etapas:

- Esquema de aprendizaje basado en competencias
- Planificación curricular
- Planificación y desarrollo de módulos y experiencias prácticas
- Revisión y validación

3.1.1 Esquema de aprendizaje basado en competencias

El aprendizaje basado en competencias incluye, el *saber* los conocimientos teóricos propios de cada área; el *saber aplicar* dichos conocimientos en situaciones determinadas, el *saber convivir* con los demás mediante actitudes y habilidades personales e interpersonales y, el *saber ser* y estar en el mundo que nos rodea. Las competencias agregan un valor añadido al proceso de enseñanza, generando así una relación entre los conocimientos, las habilidades y el comportamiento. [13]

En el sistema de evaluación por competencias, el plan curricular se formula y se expresa en competencias generales y específicas. Los cuatro elementos fundamentales del proceso de enseñanza-aprendizaje para lograr dichas competencias son:

1. Estrategia y metodología de enseñanza-aprendizaje
2. Modalidades
3. Seguimiento
4. Evaluación

En la Tabla 1, se muestran con más detalle los elementos fundamentales del proceso [13]

APRENDIZAJE BASADO EN COMPETENCIAS		
Elementos	**Definición**	**Tópicos**
Estrategia enseñanza-aprendizaje	Diseño de un proceso compuesto por procedimientos y normas que aseguran una decisión en función de los objetivos perseguidos.	Métodos: Exposición, estudio de casos, proyectos, resolución de problemas, laboratorios. Recursos: Presentaciones, charlas, material audiovisual. Tiempos de duración
Modalidades	Formas de organizar el proceso de enseñanza-aprendizaje.	Modalidad presencial Modalidad Semipresencial Modalidad OnLine
Seguimiento	Feedback del progreso del estudiante, o autoevaluación sobre cómo está desarrollando su estudio.	Tutoría individual y/o grupal, revisión de trabajos/proyectos, feedback de ejercicios y resolución de los mismos.
Evaluación	Consiste en la apreciación de los aspectos tanto académicos como formativos del estudiante.	**Qué** se va a evaluar: competencias generales y específicas trabajadas. **Cómo** se van a evaluar: Técnicas/ instrumentos que se van a emplear, ej.: examen, análisis de tareas, presentaciones orales, prueba de ejecución. **Criterios de evaluación:** La evaluación deberá reflejar un equilibrio entre las competencias trabajadas y las técnicas empleadas.

Tabla 1 Elementos del Aprendizaje Basado en Competencias

3.1.2 Planificación curricular

La planificación del currículo ha de entenderse como un proceso a través del cual se toman las decisiones respecto al qué, para qué, cómo, cuándo, dónde, en cuánto tiempo se pretende enseñar la materia. Es la toma de decisiones curriculares donde también está comprendida la forma cómo se evaluará (MINEDUC, Orientaciones, 2004).

Para el desarrollo del programa de curso, se desarrolló un esquema compuesto por dos niveles: el primer nivel corresponde a la etapa de planificación, y el segundo nivel consta de la etapa de desarrollo. En la figura 4 se muestra cada etapa con sus procesos correspondientes.

Figura 4 Esquema de planificación curricular

3.1.2.1 Búsqueda y Recolección de Información

En primera instancia, se establece la búsqueda y recolección de información que servirá de base para fundamentar los contenidos del curso. Igualmente los datos referentes al valor y misión del curso en general, tales como a quiénes estará dirigido el curso y orientarse hacia el tópico en el cual se quiere desarrollar el programa.

3.1.2.2 Formulación de Objetivos

Se formulan objetivos finales y específicos directamente relacionados con el aprendizaje que el alumno deberá tener al culminar el curso (objetivo final) y las tareas que le ayudarán a alcanzar el aprendizaje esperado (objetivos específicos).

3.1.2.3 Diseño del programa de curso

En este nivel se diseña el programa de curso basado en unidades programáticas que guardan estrecha relación con los objetivos planteados en el punto anterior. Comprende los contenidos, estrategias y recursos para el logro de objetivos. En el esquema de la Figura 4 se muestra cómo estará estructurado el programa de curso y las unidades programáticas, siguiendo el método de aprendizaje basado en competencias.

3.1.2.4 Diseño de módulos de instrucción

En este nivel se diseña detalladamente los módulos de instrucción correspondientes a cada unidad de enseñanza, con los contenidos orientados a los objetivos específicos planteados para cada unidad. Se escoge la estrategia de enseñanza según cada tópico y se establecen los recursos necesarios para cada módulo. En la Figura 5 se muestra la estructura del programa de curso que se usará para la definición de cada unidad programática.

PROGRAMA DE CURSO

Código	Nombre			
Nombre en inglés				
SCT	**Unidades Docentes**	**Horas de cátedra**	**Horas Docencia Auxiliar**	**Horas de trabajo personal**
Requisitos				**Carácter del curso**
Resultados de aprendizaje del curso				

Metodología Docente	Evaluación General

UNIDADES PROGRAMÁTICAS

Número	Nombre de la unidad	Duración en semanas
Contenidos	**Resultados de aprendizaje de la unidad**	**Referencias a la bibliografía**

Figura 5 Formato de definición de curso

3.1.2.5 Diseño del programa de evaluación

Como nivel final en la etapa de planificación, se diseña el programa de evaluación el cual servirá como indicador de la efectividad del aprendizaje y control de cumplimiento de objetivos, medido a través de evaluaciones teóricas y prácticas.

3.1.2.6 Diagnóstico preliminar

En ésta etapa se establecen los pasos a seguir para el desarrollo de clases. En primer lugar, la evaluación preliminar de la situación de los alumnos en cuanto a los contenidos de cada módulo, para establecer medidas iniciales antes de abordar los temas previstos, con el fin de adaptar el programa a las necesidades reales del curso.

3.1.2.7 Clase y evaluación formativa

La evaluación formativa puede ser utilizada para mejorar el programa antes de implementarlo formalmente, ya que en permite obtener respuestas a interrogantes tales como:

¿Es el enfoque correcto de objetivos el que se está implementando?

¿Los criterios están siendo medidos correctamente?

¿Es útil el diseño planteado?

¿El desarrollo del programa está alineado con la intención del diseño?

En paralelo con la evaluación formativa, mediante clases expositivas se transmitirán los contenidos de cada módulo, tomando en cuenta los resultados del diagnóstico preliminar, con el fin de orientar y evaluar el curso de acuerdo al proceso de planificación establecido en un principio, aplicando los contenidos y recursos de cada módulo. [14]

3.1.2.8 Evaluación acumulativa

La evaluación acumulativa ocurre luego de la implementación de un programa o solución, y usualmente requiere la cantidad de tiempo de asentamiento apropiada, para que el objeto de evaluación tenga la oportunidad de tener por completo el impacto requerido en el desempeño que se quiere lograr [14]. Mediante este proceso se evalúa el cumplimiento de los objetivos de enseñanza y aprendizaje planteados al inicio del curso, así como también evaluar la efectividad y eficiencia del programa.

3.1.3 Planificación/desarrollo de módulos y experiencias prácticas

Para la planificación de cada módulo, se toman en cuenta los procesos de planificación curricular expuestos anteriormente, para luego ser definidos en una componente teórica o a través de una experiencia práctica según sea el caso. La estructura general de cada módulo estará compuesto según se muestra en la Figura 6.

Figura 6 Estructura de módulo de instrucción

Los módulos con experiencias prácticas, de igual forma estarán estructurados de acuerdo al diagrama de la figura 5, con la característica adicional de inclusión de guías preliminares, en las cuales se especificarán los pasos de la actividad a ejecutar, mostrando resultados según cada proceso.

3.1.4 Validación de resultados

La validación de resultados corresponde a la última fase de la metodología docente, y básicamente consiste en ejecutar un prototipo de prueba del programa desarrollado y orientarlo a un grupo de personas que se adecúen a los perfiles y requisitos de la potencial audiencia que se estableció para el curso. Esto con la finalidad de obtener información real de la aceptación y resultados del curso para establecer mejoras en el procedimiento, contenidos, duración y otros parámetros que sean relevantes para aumentar la eficiencia del programa.

Mediante la aplicación real del programa, será posible confirmar si el plan curricular establecido se adecúa a los intereses de la posible audiencia y si se debe profundizar en otros temas de interés, o por el contrario obviar ciertos temas que no se adapten al interés de la audiencia.

Por otra parte, también será posible evaluar la distribución de contenidos y requisitos según cada módulo y adicionalmente, en la práctica ayudará a determinar con mayor precisión, la duración de los módulos incluyendo posibles discusiones en clase según se incrementen las dudas o interrogantes de la audiencia frente a ciertos temas.

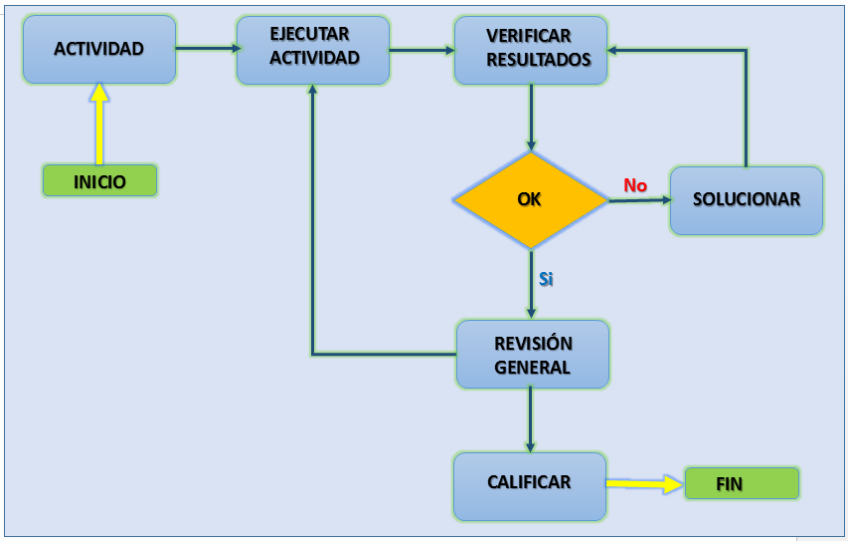


Figura 7 Proceso de validación de resultados

3.2 METODOLOGÍAS DE EXPERIENCIAS PRÁCTICAS

El programa consta de un módulo específico orientado a experiencias de laboratorio en las cuales se concreta gran parte de los conocimientos teóricos del curso. En dicho módulo, se establecen tres prácticas que deben hacerse en un ambiente acorde el cual cumpla los requisitos mínimos para obtener los resultados esperados.

Para el desarrollo de las actividades prácticas se estableció el método que se esquematiza bajo el concepto de la metodología docente pero que se enfoca hacia la elaboración de guías de laboratorio específicas siguiendo el objetivo establecido en cada módulo.

La metodología que se utilizó para desarrollar las actividades prácticas se basa en cuatro etapas principales:

Recopilación de información
Definición de tema
Establecimiento de objetivos
Desarrollo de laboratorio
Evaluación

La estructura general de las actividades prácticas desarrolladas en el laboratorio, será la que se muestra en la Figura 8.

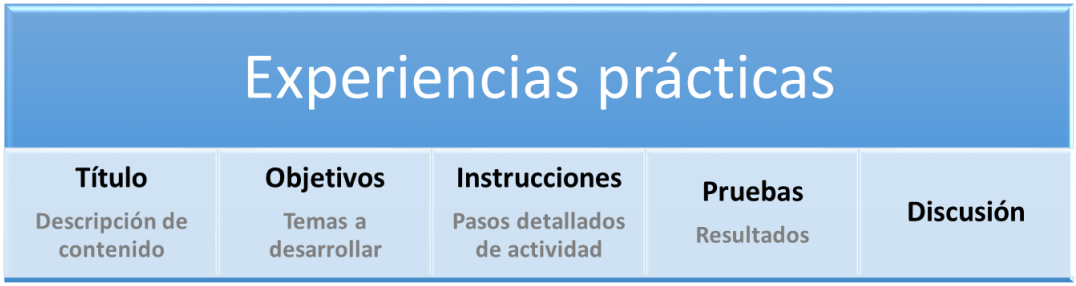


Figura 8 Estructura de las experiencias prácticas

4

RESULTADOS

En éste capítulo se presentan los resultados obtenidos mediante la aplicación de la metodología y objetivos planteados en capítulos anteriores. Debido a que el presente trabajo de grado se trata de un curso teórico/práctico, los resultados se dividirán en dos fracciones:

- Fracción I – Resultados Docentes: La primera fracción corresponde a los resultados en el ámbito docente, en el cual se especifican los módulos de instrucción y unidades programáticas correspondientes según la duración total del curso, así como también los contenidos y recursos utilizados para cada unidad de aprendizaje.
- Fracción II – Resultados prácticos: Los resultados de las experiencias prácticas y pruebas de laboratorio, corresponden a la segunda fracción, los cuales incluyen la instalación de software especializado para el manejo de la data.

4.1RESULTADOS DOCENTES

4.1.1. Requisitos básicos del curso

Para tomar el curso no es necesario tener conocimientos avanzados sobre Big Data, ya que para muchos es un tema nuevo en el cual quieren incurrir. Como requisitos básicos, para facilitar el entendimiento de algunos temas y realización de prácticas de laboratorio, es necesario contar con ciertas aptitudes (no excluyentes) tales como:

- Conocimiento intermedio de networking.
- Conocimiento intermedio de virtualización.
- Conocimiento básico de base de datos (lenguaje SQL)
- Manejo intermedio de Linux.

4.1.2. Potencial audiencia

El curso está dirigido a los interesados en obtener conocimientos generales sobre Big Data. Dentro de la audiencia, que se estima será de provecho el curso, se destacan los siguientes grupos:

- Alumnos de pre-grado y post-grado del Departamento de Ingeniería Eléctrica de la Universidad de Chile.

- Profesionales independientes que deseen tener conocimientos sobre el mundo del Big Data y tecnologías convergentes.

4.1.3. Período de duración del curso

En primer lugar, por ser un curso orientado a la docencia, las horas de clase se contabilizarán como horas académicas, es decir de cuarenta y cinco (45) minutos cada una.

El período de duración del curso, en promedio, se estima que sea de diez (10) semanas, realizando dos (2) clases por semana de dos (2) horas académicas cada clase, dando un total de cuarenta (40) horas académicas, período que considera la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile para la duración de un semestre académico.

El tiempo de duración del curso puede ser adaptado según las necesidades de los alumnos y de las condiciones que lo ameriten.

4.1.4. Objetivo general del curso

El Objetivo general que se busca con la realización de este curso, radica en que durante las clases, el alumno sea capaz de analizar, entender y describir los diferentes módulos de aprendizaje que se ofrecen según la unidad programática, y que al término del curso se tengan conocimientos sólidos de cómo se define, se compone y se desarrolla un ambiente óptimo para la aplicación del Big Data Analytics y cómo puede ser de provecho para ser utilizado en la industria de los operadores convergentes.

4.1.5. Objetivos específicos del curso

El curso se divide en tres unidades programáticas que engloban el objetivo general, y dentro de cada unidad se establecen módulos de aprendizajes orientados a profundizar cada uno de los tópicos establecidos.

4.1.6. Definición del programa de curso

En la Tabla 2 se muestra la definición del curso, según el formato definido por la FCFM de la Universidad de Chile.

Código	**Nombre**			
	Big Data			
Nombre en inglés				
Big Data				
SCT	**Unidades Docentes**	**Horas de cátedra**	**Horas Docencia Auxiliar**	**Horas de trabajo personal**
		40	0	60
Requisitos				**Carácter del curso**
Ninguno				Electivo
Resultados de aprendizaje del curso				
Al término del curso sobre Big Data, el estudiante demuestra en forma general, que: • Aplica adecuadamente los conceptos generales y las arquitecturas de Big Data estudiadas para servicios, aplicaciones y soluciones del Big Data. • Identifica los principales usos y beneficios del Big Data como tecnología y construye proposiciones de valor, destacando el racional económico del Big Data. • Caracteriza las plataformas estudiadas para el desarrollo de soluciones de Big Data, como por ejemplo: Hadoop, Cloudera, Spark.				

Tabla 2 Definición de programa de curso para la FCFM

4.1.6.1. Unidad programática I: Fenómeno del Big Data

Módulo 1: Introducción al Big Data

En este módulo introductorio se busca que el alumno se relacione con los términos básicos, los antecedentes y las características propias de lo que llamamos Big Data. Igualmente que maneje conceptos que ayudarán a un mejor entendimiento en capítulos posteriores.

Módulo 2: Desafíos del Big Data

En éste segundo módulo, se establecen conceptos y análisis de cómo esta nueva filosofía se ha vuelto tan importante y útil en nuestro día a día, indicando las oportunidades que se ofrecen dando una breve introducción al Big Data Analytics y la relación que existe hoy en día con las plataformas de cloud computing.

Módulo 3: Arquitectura del Big Data

Corresponde al último módulo de la primera unidad programática, y está orientado a explicar detalladamente la arquitectura propia de un ambiente para el Big Data, explicando sus componentes y capas de una forma detallada, así como también las estructuras utilizadas para el funcionamiento base del análisis de data.

4.1.6.2. Unidad programática II: Herramientas y técnicas de aplicación para Big Data.

Módulo 1: Almacenamiento de data

En este módulo se busca orientar al alumno en el proceso de almacenamiento de las grandes cantidades de data y especificar en los recursos claves que se deben tomar en consideración a la hora de aplicar un entorno basado en Big Data Analytics.

Módulo 2: Ecosistemas del Big Data

En este módulo se especifican los elementos que componen la arquitectura del Big Data, estableciendo las funciones de cada una de las distribuciones de Hadoop que hacen posible el proceso de recolección, análisis y almacenamiento de la diversidad de data que ingresa a una arquitectura de Big Data.

4.1.6.3. Unidad programática III: Big Data y los negocios

Módulo 1: Impulsores de mercado

Este módulo tiene como objetivo dar a conocer los puntos clave que impulsan la necesidad de adoptar la tecnología inherente al Big Data y hacer comparaciones con los escenarios en los que se aplica Big Data y aquellos en los que no.

Módulo 2: Estrategias de aplicación

En este módulo se busca orientar al alumno en las estrategias principales que se deben tomar en cuenta a la hora de aplicar el modelo de Big Data dentro de una empresa, siempre orientado hacia operadores que usen la convergencia de sus servicios al cliente, por ejemplo: operadores móviles.

4.1.6.4. Unidad programática IV: Desarrollo de un entorno práctico de Big Data

Módulo 1: Instalación de Apache Hadoop

Este módulo consiste en la primera experiencia práctica del curso, y tiene como objetivo establecer el primer contacto con la herramienta básica del Big Data: Hadoop.

Módulo 2: Aplicaciones prácticas MapReduce

En este módulo se busca crear una experiencia en la cual el alumno sea capaz de ver un resultado en el mundo real de la aplicación de MapReduce, que se considera una de las más importantes dentro del mundo del Big Data. Esta experiencia consiste en introducir un texto y correr un programa de contador de palabras.

4.1.7. Duración de unidades programáticas

El desarrollo del curso contempla cuatro unidades programáticas de las cuales 3 de ellas están compuestas por contenido teórico y la última contiene actividades prácticas las cuales se desarrollarán en el laboratorio. Se dividen en veinte clases (20), entre teóricas y prácticas, las cuales suman en total cuarenta (40) horas académicas. En la Tabla 3 se detalla la duración de cada unidad basadas en horas académicas y las clases que corresponden según el contenido.

Unidad programática	Módulo	Clase Tipo	N° Clase	Slides	Horas
I. Fenómeno del Big Data	Introducción al Big Data	Teórica	1	**17**	**2**
	Desafíos del Big Data	Teórica	2	**19**	**3**
	Arquitectura del Big Data	Teórica	3	**23**	**3**
II. Herramientas y técnicas de aplicación para Big Data.	Almacenamiento del Big Data	Teórica	4	**9**	**2**
	Ecosistemas del Big Data	Teórica	5-6	**9-8**	**3**
III. Big Data y los negocios	Impulsores de mercado para Big Data Analytics	Teórica	7	**13**	**2**
	Big Data Analytics	Teórica	8-9	**15-11**	**3**
	Estrategias de aplicación	Teórica	10	**11**	**2**
IV. Desarrollo de un entorno práctico de Big Data	Instalación de Hadoop (VMWare) Apache Hadoop V1.1.2	Práctica	11-12	**Guía Lab**	**4**
	Instalación de Sqoop	Práctica	13-14	**Guía Lab**	**4**
	Desarrollo de un programa de conteo de palabras. Permite observar cómo funciona el Map-Reduce de Hadoop.	Práctica	15-16	**Guía Lab**	**4**

	Instalación de Cloudera Manager Parte I	Práctica	17-18	**Guía Lab**	**4**
	Instalación de Cloudera Manager Parte II	Práctica	19-20	**Guía Lab**	**4**
		TOTAL			**40**

Tabla 3 Duración de unidades programáticas

4.1.8. Contenidos y Recursos de las Unidades Programáticas

Se establecen los contenidos que se expondrán en cada unidad, así como también las competencias adquiridas por el estudiante al finalizar cada unidad programática compuesta por módulos de aprendizaje.

Ver Anexo II – Título 1 para los contenidos más detallados de cada unidad programática según sus módulos de aprendizaje.

4.1.8.1. Unidad programática I

Número	**Nombre de la unidad**	**Duración en semanas**
1	**Fenómeno del Big Data**	**1,5**

Contenidos	**Resultados de aprendizaje de la unidad**	**Referencias a la bibliografía**
1. Definición, puntos clave, fuentes, conceptos básicos. 2. Diferencia entre Big Data vs métodos tradicionales, Oportunidad del Big Data, Cloud Computing y Big Data, BDA, Hadoop 3. Arquitectura de referencia para el Big Data, Fuentes de data, Capas, Virtualización.	El estudiante demuestra que es capaz de: - Comprender la definición y características generales del Big Data como proceso tecnológico. - Analizar las diferencias con otros modelos. - Analizar las arquitecturas para Big Data	[1] , [2] , [9]

Tabla 4. Contenidos y recursos de unidad programática I

4.1.8.2. Unidad programática II

Número	Nombre de la unidad	Duración en semanas
2	**Herramientas y técnicas de aplicación para Big Data**	**1,5**
Contenidos	**Resultados de aprendizaje de la unidad**	**Referencias a la bibliografía**
1. Arquitecturas de alto desempeño: HDFS, MapReduce. 2. Ecosistemas de Big Data: Hadoop, Zookeper, HBase, Hive, Pig, Mahout, Funcionamiento de Hadoop, Consideraciones y requerimientos técnicos. Base de datos para manejo de Big Data, Graph Analytics para Big Data	El estudiante demuestra que es capaz de: - Comprender las arquitecturas propuestas para Big Data. - Clasificar los ecosistemas para Big Data - Explicar los requerimientos y consideraciones técnicas.	[4] , [6], [7]

Tabla 5 Contenidos y recursos de unidad programática II

4.1.8.3. Unidad programática III

Número	Nombre de la unidad	Duración en semanas
3	**Big Data y los negocios**	**2**
Contenidos	**Resultados de aprendizaje de la unidad**	**Referencias a la bibliografía**
1. Impulsores de mercado para Big Data Analytics 2. Estrategias de aplicación Big Data Analytics Proceso para la integración de la tecnología	El estudiante demuestra que es capaz de: Identificar los principales impulsores de mercado para Big Data. Analizar y aplicar estrategias de Big Data Analytics.	[5], [8], [10], [11]

Tabla 6. Contenidos y recursos unidad programática III

4.1.8.4. Unidad programática IV

Número	Nombre de la unidad	Duración en semanas
4	**Desarrollo de un entorno práctico de Big Data**	**5**

Contenidos	Resultados de aprendizaje de la unidad	Referencias a la bibliografía
1.Instalación de Apache Hadoop 2.Uso de MapReduce 3.Instalación de Sqoop 4.Instalación de Cloudera Manager Parte I 5.Instalación de Cloudera Manager Parte II	El estudiante demuestra que es capaz de: - Instalar y comprender el funcionamiento y aplicación de	[15], [16], [17] [18]

Tabla 7. Contenidos y recursos unidad programática IV

4.1.9. Evaluaciones

Tomando en cuenta que la escala de evaluación que se utilizará será la de siete (7) puntos como la nota máxima. Las evaluaciones serán divididas en dos (2) controles cortos teóricos correspondientes a las tres unidades programáticas que contienen la información teórica. Las cinco (5) experiencias prácticas también serán evaluadas sobre la misma escala. La aprobación del curso requiere que el promedio de los controles teóricos y prácticos sea igual o mayor a 4 puntos.

En el curso no se aplicará examen final, de esta forma la nota acumulada será la correspondiente a la nota definitiva del curso. La idea principal de las evaluaciones es estimar el conocimiento adquirido de los estudiantes al terminar cada unidad programática. En la Tabla 8 se muestra el esquema evaluativo propuesto según la duración del curso.

Evaluaciones	Contenidos	Semana	Porcentaje
Control I	Unidad I	2	25%
Control II	Unidad II y III	4	25%
Práctica I	Laboratorio I	6	10%
Práctica II	Laboratorio II	7	10%
Práctica III	Laboratorio III	8	10%
Práctica IV	Laboratorio IV	9	10%
Práctica V	Laboratorio V	10	10%

Tabla 8. Esquema de evaluaciones

4.1.10. Material desarrollado

Se preparó un material de apoyo para las clases teóricas y experiencias prácticas, el cual incluye presentaciones con diapositivas para la fase teórica del curso, y guías de laboratorio para las experiencias prácticas, donde se especifican las actividades a seguir para concretar el fin de la actividad de prueba.

4.1.10.1. Clases expositivas

Para las clases presenciales, se utilizaron diapositivas las cuales abarcan el contenido de cada unidad programática, y de ésta forma usarlas como material de apoyo en la clase tanto como para el profesor como para el alumno, y de ésta forma tener un orden secuencial con los módulos establecidos en el programa.

El formato de las presentaciones es el que se muestra en la Figura 9:

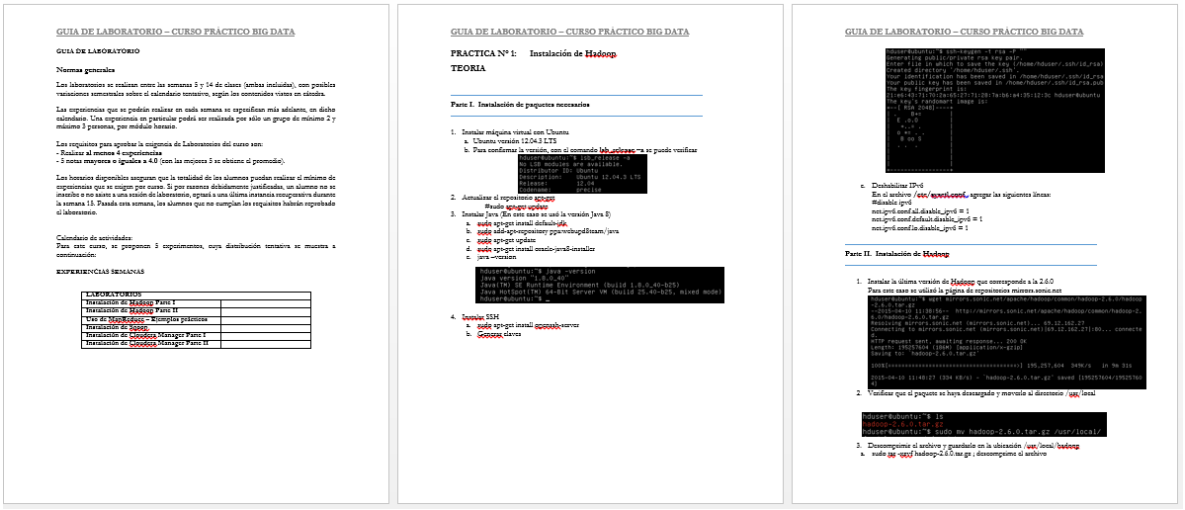
Figura 9 Modelo de presentaciones

4.1.10.2. Guías de laboratorio

Para las experiencias prácticas, típicamente se estila a seguir una guía la cual tenga especificados los pasos para llevar a cabo la actividad. Es por esto que para las prácticas de laboratorio contempladas para este curso, se desarrollaron guías cortas

para guiar al alumno en la instalación de las aplicaciones que facilitaran el entendimiento del tema.

Las guías están estructuradas según el esquema metodológico expuesto en el capítulo 2. En la Figura 10 se muestra la forma de estructuración de las guías prácticas.


GUIA DE LABORATORIO – CURSO PRÁCTICO BIG DATA

GUIA DE LABORATORIO

Normas generales

Las laboratorios se realizan entre las semanas 5 y 14 de clases (ambas incluidas), con posibles variaciones semanales sobre el calendario tentativo, según las necesidades vistas en cátedra.

Las experiencias que se podrán realizar en cada semana se expresarán más adelante, en dicho calendario. Una experiencia en particular podrá ser realizada por sólo un grupo de mínimo 2 y máximo 3 personas, por módulo horario.

Los requisitos para aprobar la exigencia de Laboratorios del curso son:
- Realizar al menos 4 experiencias
- 3 notas mayores o iguales a 4.0 (con las mejores 3 se obtiene el promedio).

Los horarios disponibles aseguran que la totalidad de los alumnos puedan realizar el mínimo de experiencias que se exigen por curso. Si por razones debidamente justificadas, un alumno no se inscribe o no asiste a una sesión de laboratorio, optará a una última instancia recuperativa durante la semana 15. Pasada esta semana, los alumnos que no cumplan los requisitos habrán reprobado el laboratorio.

Calendario de actividades
Para este curso, se propondrán 5 experimentos, cuya distribución tentativa se muestra a continuación:

EXPERIENCIAS SEMANAS

LABORATORIOS	
Instalación de Hadoop Parte I	
Instalación de Hadoop Parte II	
Uso de MapReduce – Ejemplos prácticos	
Instalación de Impala	
Instalación de Cloudera Manager Parte I	
Instalación de Cloudera Manager Parte II	

GUIA DE LABORATORIO – CURSO PRÁCTICO BIG DATA

PRACTICA N° 1: Instalación de Hadoop

TEORIA

Parte I. Instalación de paquetes necesarios

1. Instalar máquina virtual con Ubuntu
   a. Ubuntu versión 12.04.3 LTS
   b. Para confirmar la versión, con el comando lsb_release -a se puede verificar
2. Actualizar el repositorio apt-get
   Hacer apt-get update
3. Instalar Java (En este caso se usó la versión Java 8)
   a. sudo apt-get install default-jdk
   b. sudo add-apt-repository ppa:webupd8team/java
   c. sudo apt-get update
   d. sudo apt-get install oracle-java8-installer
   e. java -version
4. Instalar SSH
   a. sudo apt-get install openssh-server
   b. Generar claves

GUIA DE LABORATORIO – CURSO PRÁCTICO BIG DATA

c. Deshabilitar IPv6
   En el archivo /etc/sysctl.conf, agregar las siguientes líneas:
   #disable ipv6
   net.ipv6.conf.all.disable_ipv6 = 1
   net.ipv6.conf.default.disable_ipv6 = 1
   net.ipv6.conf.lo.disable_ipv6 = 1

Parte II. Instalación de Hadoop

1. Instalar la última versión de Hadoop que corresponde a la 2.6.0
   Para este caso se utilizó la página de repositorios mirrors.sonic.net
2. Verificar que el paquete se haya descargado y moverlo al directorio /usr local
3. Descomprimir el archivo y guardarlo en la ubicación /usr/local/hadoop
   a. sudo tar -xzvf hadoop-2.6.0.tar.gz ; descomprimo el archivo

Figura 10 Ejemplo de Guía de laboratorio.

4.2 RESULTADOS DE EXPERIENCIAS PRÁCTICAS

Los resultados fueron obtenidos luego de la realización de las prácticas contempladas para el curso. Para las pruebas se utilizaron máquinas virtuales donde se instaló Hadoop como plataforma de análisis.

4.2.1.1 Práctica I. Instalación de Hadoop

Para establecer una visión práctica de cómo funciona un ambiente desarrollado con Hadoop, se ejecutó la experiencia práctica basada en la instalación de ésta herramienta.
Para ello, se usaron los siguientes recursos:

- Máquina virtual (VMWare/VirtualBox) con Ubuntu V 12.04
- Java 8
- Apache Hadoop 2.6.0 (Versión actualizada)

El resultado obtenido de ésta experiencia fue la instalación de Hadoop como herramienta base para el estudio del Big Data, y en ese sentido orientar al alumno en las características y funcionamiento básico de la herramienta. [15]

En la Figura 11 se observa el resultado final de la instalación. Mediante la aplicación del comando ***jps***, se verifica que los daemons de Hadoop estén ejecutándose de manera correcta en el equipo.

```
hduser@ubuntu:/usr/local/hadoop$ jps
1746 SecondaryNameNode
1467 NameNode
1564 DataNode
1885 ResourceManager
2174 NodeManager
2270 Jps
```

Figura 11 Resultado de Instalación de Hadoop - Procesos

4.2.1.2 Práctica II. Uso de MapReduce – Ejemplos prácticos

En esta práctica se establecen ejemplos cuantificables del uso de MapReduce como herramienta de análisis de la data. La experiencia de laboratorio consiste en correr un programa que permita el ingreso de un texto y genere como salida la cantidad de repeticiones que tiene cada palabra; en general es un contador de palabras (WordCount). En la Figura 12 se muestra el resultado de una parte del conteo de palabras que se utilizó en la práctica. *[16]*

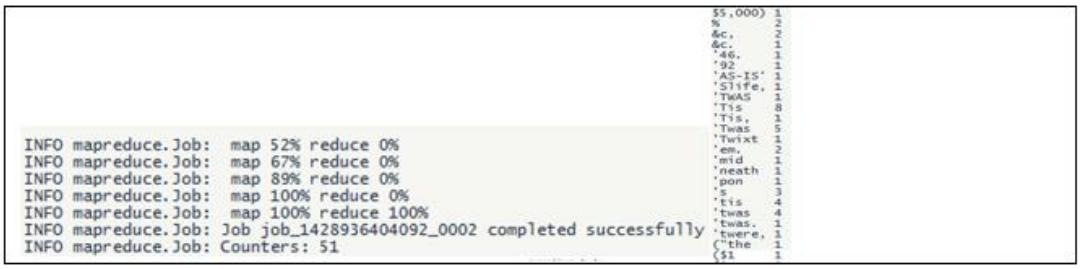


Figura 12 Resultado de programa corriendo en MapReduce

4.2.1.3 Práctica III. Instalación de Apache Sqoop

En ésta práctica se desarrolló la instalación de Apache Sqoop como herramienta de ingestión de data. La idea principal de instalar Sqoop es porque normalmente se realizan prácticas de cómo la data es analizada, pero muchas veces surge la duda de cuál es la herramienta que se utiliza para ingresar la data al ecosistema de Hadoop.

En general la experiencia consistió en instalar Sqoop y analizar los procesos y opciones que están involucrados en etapa de ingestión de la data, como se muestra en la Figura 13.

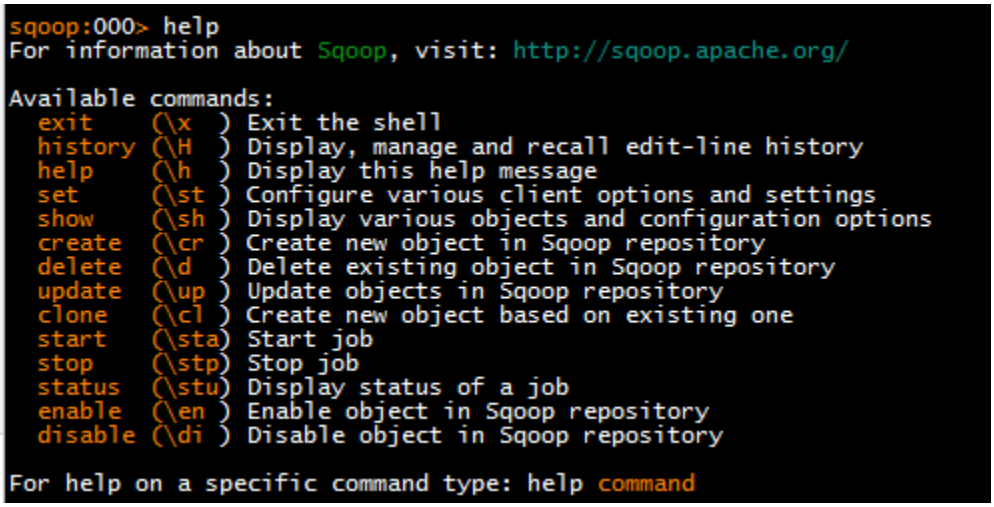


Figura 13 Resultado de Instalación de Apache Sqoop

4.2.1.4 Práctica IV. Instalación de Cloudera Manager Parte I

En ésta experiencia práctica se configura el Cloudera Manager, el cual consiste en una aplicación end-to-end para la administración de clusters CDH. La aplicación automatiza el proceso de instalación, reduciendo el tiempo de desarrollo, permitiendo la utilización de vistas en tiempo real de hosts y servicios corriendo.

En esta primera parte de la experiencia, el alumno deberá identificar los componentes de la herramienta y descargar y configurar el instalador de Cloudera Manager para luego en la segunda parte configurar un clúster de ejemplo.

Al final de la instalación del asistente, la salida será la mostrada en la Figura 14.

Figura 14 Resultado de la configuración del Instalador Cloudera Manager

4.2.1.5 Práctica V. Instalación de Cloudera Manager Parte II

En ésta práctica se utilizó el Cloudera Manager para crear un clúster de Hadoop (CDH), el cual consiste en un tipo especial de clúster designado específicamente para almacenar y analizar grandes cantidades de data no estructurada en un ambiente distribuido. Cloudera Manager permite crear clusters y agregar tantos host como se requieran a cada clúster según las necesidades.

En la práctica de laboratorio, se configuró un clúster con un host para realizar las pruebas correspondientes y probar los servicios y aplicaciones que se pueden correr bajo el ambiente de Cloudera Manager.

Como resultado se obtuvo la inclusión del host en el clúster de Cloudera Manager.

5

DISCUSIÓN DE RESULTADOS

En este capítulo se discute la metodología aplicada para la realización de los resultados docentes, la validación de las experiencias prácticas y laboratorios así como también se analizan los beneficios e implicaciones que tienen estos temas para los operadores convergentes.

Para la validación de la metodología docente y práctica, se contó con el apoyo de los estudiantes del Magíster en Ingeniería de Redes de Comunicaciones (MIRC) de la Universidad de Chile, quienes actuaron como potencial audiencia ya que cumplían con los requerimientos y aptitudes que se establecieron en capítulos anteriores.

5.1 DISCUSIÓN Y VALIDACIÓN DE LA METODOLOGÍA DOCENTE

La metodología docente aplicada, se basó en los procesos de planificación curricular los cuales en un principio, comienzan con la búsqueda y recolección de información que permitió sentar las bases teóricas; la formulación de objetivos como segundo paso, permitió definir el alcance del programa de curso y cuáles eran los objetivos que se necesitaban alcanzar; como tercer nivel en la escala del proceso, se encuentra el diseño del programa de curso, donde se establecieron parámetros como la duración del curso, los tópicos que se abordarían y los temas más relevantes que se adaptaran a la formulación de objetivos en el paso dos.
Una vez realizado el diseño del curso, se procedió a diseñar los módulos de instrucción per se. Por otro lado, el diseño de programa de evaluación permitió establecer un sistema evaluativo correspondiente al nivel de la audiencia que se pretendía enseñar, en este caso fueron alumnos de la pregrado y postgrado de la Universidad de Chile. Por último las clases y las evaluaciones formativas forman el último escalón de la metodología docente utilizada, ya que mediante este proceso se logró transmitir la información y tópicos establecidos en pasos anteriores.

Para validar las clases y evaluaciones formativas de acuerdo a la metodología utilizada, se realizaron dos clases de prueba con los alumnos del MIRC. Por razones de tiempo no se dictó el programa en su totalidad pero con dos clases se observó una recepción positiva a la estructura del curso y sus contenidos, al igual que la duración

en semanas del mismo. Los alumnos mostraron interés en el contenido de las clases y se generaron dudas que permitieron la actualización de los contenidos del curso.

5.2 DISCUSIÓN Y VALIDACIÓN DE EXPERIENCIAS PRÁCTICAS

Las experiencias prácticas constan de cinco (5) laboratorios relacionados en su núcleo con la instalación de Hadoop como herramienta para el manejo y análisis del Big Data, así como también ciertos componentes que resultan de utilidad y fácil manejo para el nivel de curso que se está dictando.

Cada laboratorio se desarrolló de forma tal que se pudiese dar a conocer y aprender herramientas básicas de Big Data, sin necesidad de hacer uso de recursos especializados, como por ejemplo computadores con alto grado de procesamiento y memoria, ya que el curso está orientado a estudiantes y profesionales que quieran obtener conocimientos básicos-avanzados del mundo del Big Data de una forma práctica con herramientas de uso cotidiano.

Para la validación de las experiencias prácticas, se realizó un laboratorio de prueba con alumnos del MIRC. Por razones de tiempo no fue posible completar las cinco experiencias prácticas, y sólo se realizó una de ellas que fue la instalación de Hadoop por cónsola. Con dicha prueba se validó la técnica y metodología utilizada, la cual fue acogida de forma positiva por los alumnos resultando de interés y lográndose la culminación de la práctica en el tiempo estimado.

5.3 ALCANCE E IMPACTO DE LOS RESULTADOS OBTENIDOS

Se plantea una propuesta de curso electivo en el Departamento de Ingeniería Eléctrica de la Universidad de Chile, para aquellos que estén interesados en el aprendizaje de las tecnologías y herramientas que hoy en día hacen posible el constante crecimiento del Big Data Analytics y cómo se puede orientar a las aplicaciones de operadores que brindan diversos servicios (operadores convergentes).

El curso presenta una estructura de aprendizaje basado en competencias por módulos, con la idea de que el estudiante no sólo maneje conceptos teóricos y prácticos, si no que al mismo tiempo pueda obtener conocimiento sobre cómo aplicarlo a la vida diaria y laboral. De forma general, el curso se contempla en 10 clases teóricas de 2 horas académicas cada clase y 5 clases prácticas 5 horas cada una. Las clases teóricas se dictarán con material de apoyo basado en presentaciones de

aproximadamente 15 slides por clase, mientras que las clases prácticas se realizarán mediante el seguimiento de guías de laboratorio.

El programa se enfoca en conocimientos básicos-avanzados del análisis del Big Data, mediante un programa docente y práctico enfocado en transmitir conocimientos sólidos que permitan entender y manejar términos y herramientas que son básicas para comprender los usos avanzados de soluciones que manejan enormes cantidades de data.

El desarrollo del curso se plantea que se establezca en la sede de la Facultad, ya que se necesita un laboratorio dotado de equipos computacionales para que los alumnos puedan desarrollar las prácticas propuestas en un ambiente grato y adecuado.

Al final del curso, el alumno será capaz de reconocer los procesos principales de una solución para el análisis de Big Data basada en Apache Hadoop, y tener el conocimiento básico para desarrollar un entorno de análisis de Big Data según sus necesidades.

Dicho lo anterior, los contenidos expuestos en el curso, pueden ser actualizados y mejorados con la inclusión de nuevos temas que sean de interés para los alumnos y el profesorado, para así adaptar el aprendizaje a los nuevos requerimientos que día a día vamos demandando como sociedad tecnológica.

5.3.1 Potenciales aplicaciones del Big Data y Racional económico

En cuanto al racional económico del Big Data, en el curso se hace referencia a este tópico en la Unidad Programática III, la cual hace énfasis en los impulsores del mercado para Big Data Analytics y las estrategias de aplicación.

Hay que tomar en cuenta que el Big Data Analytics tiene un gran potencial en aplicaciones hoy y en el futuro. El Big Data Analytics es totalmente aplicable al ámbito de las telecomunicaciones, medicina, educación, seguridad, avances tecnológicos, etc. De hecho son muy pocos los campos en los que el Big Data no genera una utilidad. Hoy en día las empresas están usando el Big Data para reducir riesgos y costos, identificar oportunidades y mejorar su rendimiento.

Por lo dicho anteriormente, en el curso se hizo referencia a casos reales en diferentes ámbitos en los cuales el Big Data logró mejorar y actualizar el rendimiento de las empresas que lo implementaron.

En el ámbito de las telecomunicaciones, se habla de Telstra, una telco australiana, que está usando big data para desarrollar un sistema de análisis predictivo que usa su data operacional para ayudar a identificar problemas en la red

antes de que ocurran. La herramienta usada como base para desarrollar dicho sistema es Hadoop. De acuerdo a la entrevista realizada a la Telstra, indican que " el análisis predictivo y prescriptivo son tecnologías que requieren altas capacidades, cómputo y almacenamiento, así como también modelamiento de la data, coincidencia de patrones, procesamiento de eventos complejos y el análisis de base de datos multidimensionales. [17].

Otro ejemplo notable, es el caso de la compañía The Weather Channel, la cual se unió con IBM, desarrollando la aplicación Insights for weather, la cual permite integrar datos sobre el tiempo históricos y en tiempo real desde The Weather Company a la aplicación IBM Bluemix. [18]. Puede recuperar datos del tiempo correspondientes a un área especificada por una geolocalización, además los datos le permiten pronosticar, detectar y visualizar eventos del tiempo [18] Toda la data del clima es recolectada desde más de 100.000 sensores, así como también millones de Smartphones, edificios, y vehículos en movimiento. Dicha data se combina con datos de otras fuentes generando un promedio de más de 2.2 billones de predicciones en un día. Con esta data, otras industrias se verían beneficiadas, por ejemplo las compañías de seguro podrían advertir a sus clientes sobre un estado del tiempo no apto para circular en sus vehículos mediante notificaciones en tiempo real, y de esta forma evitar posibles accidentes.

6

CONCLUSIONES

Este trabajo de grado es el resultado de un proceso de investigación, y recolección para el desarrollo de un curso teórico y práctico sobre Big Data, el cual se contempla que forme parte de las materias electivas en el programa de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, para aquellos alumnos cuyo interés por el mundo del Big Data genere sugestión.

La metodología que se aplicó en el desarrollo del programa, se basó en el aprendizaje basado en competencias, la cual genera un rendimiento satisfactorio, ya que según las pruebas realizadas, se establecen conocimientos teóricos y prácticos que, además de otorgarle el conocimiento empírico a los alumnos, también permite desarrollar aptitudes y competencias que fueron de utilidad para el desempeño de actividades relacionadas con el entorno de aplicación del Big Data.

Se logró constituir un ambiente de laboratorios donde se establecieron situaciones que simulan un ambiente práctico, donde herramientas necesarias para el análisis de grandes cantidades de data fueron aplicadas y desarrolladas, como por ejemplo la instalación de Hadoop, con el cual el alumno pudo familiarizarse con el lenguaje y mecanismo usado para el modelamiento de los datos, así como también el análisis y estudio de los mismos mediante herramientas de Big Data Analytics tales como el entorno que brinda Cloudera. Los cambios en el entorno hacen del Big Data Analytics una tecnología atractiva para cualquier organización, mientras que las condiciones del mercado la hacen práctica. La combinación de modelos simplificados para el desarrollo, una amplia gama de herramientas de administración y un cómputo a bajo costo, efectivamente ha minimizado la barrera para entrar a este mundo, permitiendo que cada vez más organizaciones desarrollen y prueben las aplicaciones de alto rendimiento que pueden manejar grandes volúmenes de data y organizarlos según estructura y contenido.

Para el desarrollo de aplicaciones de Big Data, especialmente cuando se inclina hacia herramientas de fuente abierta, demanda una inversión de tiempo y recursos para asegurar que el procesamiento y análisis de datos estén listos para producir resultados, es por ello que para lograr rendimientos tangibles en las prácticas de laboratorio, lo ideal es mantener actualizados los equipos de laboratorio para que el alumno pueda hacer uso de la herramienta de estudio de la mejor forma posible y exista el mejor aprovechamiento del curso.

La estructura general del curso, está definida de forma que se puedan validar los contenidos de forma constante y de este modo, establecer un proceso de actualización continuo y totalmente adaptable a los requerimientos emergentes tanto de la sociedad como la industria. Tal como se indicó en capítulos anteriores, este trabajo tiene amplias proyecciones en el plano educativo e investigativo, el cual

está abierto a la incorporación de nuevos contenidos y experiencias que sean de provecho para el crecimiento del curso, permitiendo su continuidad.

Como idea final, se debe tomar en cuenta que no siempre es el tamaño de la data lo que hace el Big Data. La habilidad no sólo de capturar la data, sino además el análisis de dicha data de una forma efectiva económicamente, es lo que realmente hace poderoso al Big Data. Las aplicaciones de análisis para Big Data emplean una extensa variedad de herramientas y técnicas para su implementación. Cuando se organizan las ideas para implementar dichas aplicaciones, es importante pensar acerca de los parámetros que enmarcaran las necesidades para la evaluación y adquisición de la tecnología.

7

BIBLIOGRAFÍA

[1] M. Minelli, M. Chambers y D. Ambiga, Big Data, Big Analytics, Hoboken, New Jersey: John Wiley & Sons, Inc, 2013.

[2] D. Laney, «http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf,» Febrero 2001. [En línea].

[3] G. Intelligence, «www.gsma.com,» Febrero 2014. [En línea]. Available: http://www.gsma.com/connectedliving/wp-content/uploads/2014/02/M2M-report_GSMAi.pdf.

[4] N. Sawant y H. Shah, Big Data Application Architecture Q&A: A problem - solution Approach, Apress, 2013.

[5] D. Loshin, Big Data Analytics, Waltman, MA: Kauffman, Morgan, 2013.

[6] S. Teller, Hadoop Essentials, 2015.

[7] B. Verheij, «The process of big data solution adoption,» Septiembre 2013. [En línea]. Available: http://www.tbm.tudelft.nl/fileadmin/Faculteit/TBM/Over_de_Faculteit/Afdelingen/Afdeling_Infrastructure_Systems_and_Services/Sectie_Informatie_en_Communicatie_Technologie/medewerkers/jan_van_den_berg/news/doc/bverheij-big-data-adoption-process-final.pdf.

[8] V. Agneeswaran, «Big-Data–Theoretical, Engineering and Analytics Perspective,» 2012.

[9] Global Institute McKinsey, «Big data: The next frontier for innovation, competition, and productivity,» Junio 2011. [En línea]. Available: www.mckinsey.com/~/.../Big%20Data/MGI_big_data_full_report.ashx.

[10] J. Gantz y D. Reinsel, «"The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest growth in the Far East",» Diciembre 2012. [En línea]. Available: www.emc.com/leadership/digital-universe/index.htm.

[11] S. Kadre y V. Konasini, Practical Business Analytics Using SAS: A Hands-on Guide, Apress, 2015.

[12] F. J. Ohlhorst, Big Data Analytics: Turning Big Data into Big Money, New Jersey: John Wiley & Sons, Inc., 2012.

[13] A. VILLA SANCHEZ y M. POBLETE RUIZ, APRENDIZAJE BASADO EN COMPETENCIAS, BILBAO: EDICIONES MENSAJERO, 2007.

[14] I. Guerra-Lopez, Evaluación y mejora contínua, Bloomington: AuthorHouse, 2007.

[15] «Guía de Instalación de Hadoop,» [En línea]. Available: http://pingax.com/install-hadoop2-6-0-on-ubuntu/.

[16] M. Noll, «Applied Research. Big Data. Distributed Systems. Open Source,» [En línea]. Available: http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/. [Último acceso: Septiembre 2014].

[17] «http://www.itnews.com.au,» 2015. [En línea]. Available: http://www.itnews.com.au/news/telstra-deploys-hadoop-to-pre-empt-network-issues-404526. [Último acceso: Diciembre 2015].

[18] «www.ibm.com,» [En línea]. Available: http://www.ibm.com/analytics/us/en/business/weather-insight.html.

[19] M. Y. Jimenez C, Base de datos relacionales y modelado de datos, Málaga: IC Editorial, 2014.

8

ANEXOS

En este capítulo se encuentra la información complementaria al trabajo expuesto, la cual incluye tablas, esquemas e información referente a los antecedentes teóricos así como también los resultados obtenidos de la elaboración del proyecto.

ANEXO I
Título 1 - ECOSISTEMAS DE BIG DATA (HADOOP)

DISTRIBUTED FILE-SYSTEM (SISTEMA DE ARCHIVOS DISTRIBUIDO)	
Apache HDFS	HDFS ofrece un método para almacenar grandes archivos a través de múltiples máquinas. Hadoop y HDFS se derivaron del informe de Google File System (GFS).
Red Hat GlusterFS	GlusterFS es un sistema de almacenamiento de archivos conectado a una red de escalamiento horizontal. Fue desarrollado por Gluster Inc. y luego por Red Hat Inc en el 2012.
Lustre File System	Lustre es un sistema de archivos distribuidos de alto rendimiento, usualmente configurado para administrar almacenamiento de data remotamente dentro de una SAN (Storage Area Network) que se comunican mediante protocolos SCSI (FCoE, SCSI, iSCSI)
DISTRIBUTED PROGRAMMING (PROGRAMACIÓN DISRIBUIDA)	
Apache MapReduce	MapReduce es un modelo de programación para procesar grandes cantidades de data con un alogritmo paralelo y distribuido en un clúster. La versión actual de MapReduce actúa está construida bajo el framework Apache YARN.
Apache Pig	Pig proporciona un motor para ejecutar flujo de data en paralelo en Hadoop, mediante un lenguaje (Pig Latin) para expresar dichos flujos de datos. Es mucho menos complejo que programar en MapReduce., a pesar de

	que la mayoría de los procesos ejecutados por Pig usa MapReduce.
Apache Spark	Spark provee una alternativa más fácil de usar Hadoop MapReduce y ofrece un rendimiento más rápido comparado con aplicaciones anteriores. Spark es un framework para escribir programas distribuidos rápidamente, mediante APIs en Scala, Java y Python.

NoSQL DATABASES	
Column Data Model (Modelo por columnas)	
Apache HBase	HBase está inspirado en Google BigTable. Realiza operaciones aleatorias de lectura/escritura en tiempo real en tablas muy grandes orientadas a columnas. Actúa como la base de datos de Hadoop y se usa para respaldar los trabajos de MapReduce.
Apache Cassandra	Cassandra actúa como una BDDB (Big Data Base) y se puede ejecutar sin HDFS o sobre HDFS. Cassandra es un proyecto de open source basado en un sistema de base de datos independiente inicialmente codificado por Facebook, quienes mientras implementaban el modelo de BigTable de Google, usaron un sistema inspirado por Amazon's Dynamo para almacenar data.
Apache Spark	Spark provee una alternativa más fácil de usar Hadoop MapReduce y ofrece un rendimiento más rápido comparado con aplicaciones anteriores. Spark es un framework para escribir programas distribuidos rápidamente, mediante APIs en Scala, Java y Python.
Files Data Model (Modelo por archivos)	
MongoDB	Sistema de base de datos orientado a archivos. Es parte de la familia de sistemas de base de datos NoSQL. En lugar de almacenar data en tablas como se haría en una base de datos clásica "relacional", MongoDB almacena data estructurada como documentos con extensión JSON (JavaScript Object Notation)
Stream Data Model (Modelo por flujo)	
EventStore	Base de datos funcional de open source con soporte para Procesamiento de Eventos

	Complejos. Almacena la data como una serie de eventos continuos en el tiempo. EventStore está escrito en C ,C++. Las aplicaciones que usan EventStore pueden estar escritas en JavaScript.
Key-Value Data Model (Modelo de clave única)	
Redis DataBase	Redis es un sistema de almacenamiento de clave única de código abierto. Actúa como un diccionario el cual mapea claves con valores. Soporta no sólo strings, si no también tipo de data abstracta.
Open TSDB	OPENTSDB, es una Base de Datos escalable en el tiempo (TSDB) escrita sobre HBase. Almacena, ordena y entrega métricas recolectadas desde los sistemas computacionales a grande escala.
Graph Data Model	
TitanDB	TitanDB es un base de datos gráfica altamente escalable optimizada para almacenar y encolar grandes gráficos con billones de vértices distribuidos a través de un clúster de varias máquinas.

SQL con HADOOP	
Apache Hive	Hive es una infraestructura de Data Warehouse desarrollada por Facebook. Permite la sumarización de la data, encolamiento, y análisis.
Cloudera Impala	Impala brinda una tecnología paralela escalable a Hadoop
DATA INGESTION (Inserción de data)	
Apache Flume	Flume es un servicio distribuido, confiable y disponible para la recolección, agregación y despliegue eficiente de grandes cantidades de logs de datos. Tiene una arquitectura simple y flexible basada en flujos de datos en streaming que permite aplicaciones de análisis online.
Apache Sqoop	Sqoop es un sistema para transferencia de masas de datos entre HDFS y base de datos estructuradas como HDFS. Parecido a Flume pero desde HDFS hacia RDBMS.

Apache Storm	Storm es un procesador de eventos complejos y un framework de computación distribuida, para el procesamiento rápido de grandes flujos de data. Un clúster de Storm, consiste en un master y nodos, con la coordinación de ZooKeeper.
Apache Kafka	Kafka es un message queueing desarrollado por Linkedin, que envía mensajes continuos al disco.
SERVICE PROGRAMMING	
Apache ZooKeeper	ZooKeeper es un servicio de coordinación que entrega herramientas para desarrollar correctamente aplicaciones distribuidas. ZooKeeper fue desarrollado en Yahoo! Research. Zookeeper es para construir sistemas distribuidos, simplificar los procesos de desarrollo.
SCHEDULING	
Apache Oozie	Oozie es un sistema de planificación para los trabajos de MapReduce usando DAGs (Direct Acyclical Graphs). Oozie puede gatillar trabajos por tiempo (frecuencia) y disponibilidad de data.
MACHINE LEARNING (Aprendizaje de máquina)	
Apache Mahout	Librería de aprendizaje de máquina y librería de matemática, funciona en el top de MapReduce.

[6]

Título 2 – Fuentes de data

Fuentes de alto volumen

1. Datos de dispositivos de conmutación
2. Datos de puntos de acceso

3. CDRs
4. Datos de las redes sociales

Variedad de fuentes

1. Imágenes, videos de redes sociales
2. Data de transacciones
3. Data de GPS
4. Data de call centers
5. E-mail
6. SMS

Fuentes de alta velocidad

1. CDR
2. Conversaciones de sitios de redes sociales
3. Data de GPS
4. Call center (Voz a texto) [6]

Título 3 – Arquitectura de un entorno de Big Data

Capa de ingestión

La capa de ingestión es la responsable de separar la información relevante de lo que se considera ruido. La capa de ingestión debe ser capaz de manejar un extenso volumen, alta velocidad y variedad de la data, así como también tener la capacidad de validar, depurar, transformar, reducir e integrar la data dentro del stack de Big Data para su posterior procesamiento. [4]

Los elementos que componen la capa de ingestión son los siguientes:

- Identificación: Identifica los diversos formatos conocidos o asigna formatos por defecto a la data no estructurada.
- Filtrado: Filtra la información entrante relevante para la empresa.
- Validación: Valida y analiza constantemente la data en contra de nueva metadata.
- Reducción: Reducción de ruido, incluye limpieza de data mediante la eliminación de ruido y minimizando las confusiones.
- Transformación: Involucra la división, convergencia, desnormalización[10] o sumarización de la data.

[10] Desnormalización (Base de datos): Es el proceso que persigue optimizar una base de datos por medio de agregar datos redundantes. [19]

- Compresión: Involucra la reducción del tamaño de la data pero no pierde la relevancia de la misma. No debería afectar los resultados del análisis luego de la compresión.
- Integración: Integra el conjunto de data dentro del almacenamiento de Hadoop (HDFS y NoSQL Databases).

En escenarios típicos de ingestión, se tiene data de múltiples fuentes para procesar. A medida que las fuentes aumentan, el procesamiento comienza a ser complicado. En el caso del Big Data muchas veces la estructura de la fuente de data no es conocida, por lo cual al seguir los enfoques tradicionales, crea una dificultad en el proceso de ingestión. Por ésta razón, se crearon patrones de ingestión que describen soluciones a problemas típicos, por ejemplo: patrón de extracción multi fuente, patrón convertidor de protocolo, patrón multi destino, patrones en tiempo real etc.

En la figura 3 se muestra el diagrama de bloques de la composición de la capa de ingestión y los patrones que se pueden utilizar.

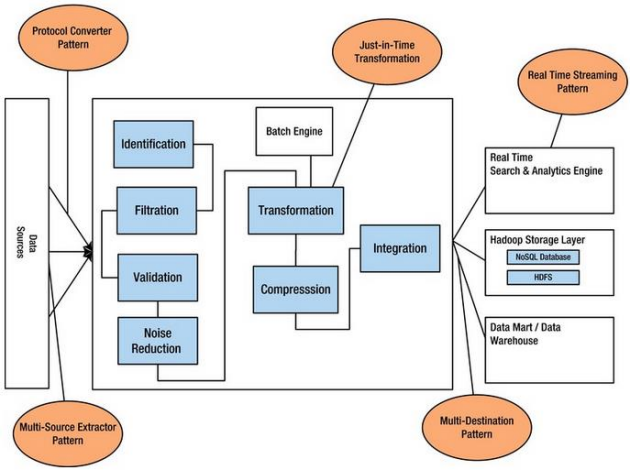


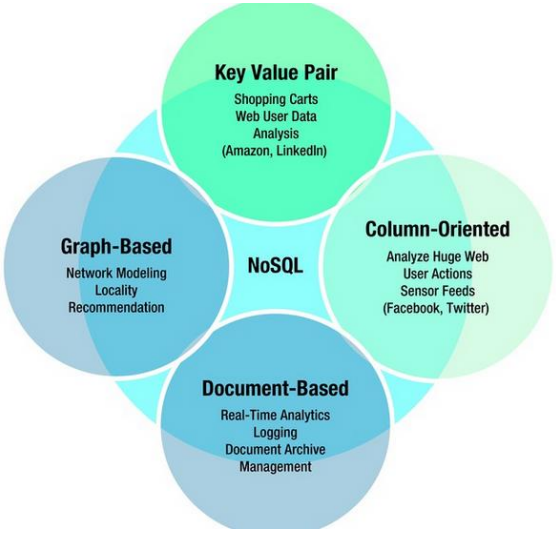
Diagrama de bloques de la capa de Ingestión

Capa de almacenamiento

Un cambio fundamental en la forma en que las empresas manejan Big Data, es el uso de almacenamiento y procesamiento distribuido. Un sistema de almacenamiento distribuido promete tolerancia a falla, y el paralelismo permite el procesamiento a altas velocidades de algoritmos distribuidos. HDFS representa la piedra angular de la capa de almacenamiento de la arquitectura de Big Data.

Hadoop permite la interacción con un clúster lógico de nodos de procesamiento y almacenamiento en vez de interactuar con sistemas operativos y CPU físicos. Los dos mayores componentes de Hadoop son: el sistema de archivos distribuidos (HDFS) que puede soportar petabytes de data y el motor Map Reduce que permite realizar operaciones en un batch.

HDFS requiere programas complejos de lectura/escritura que normalmente son escritos por desarrolladores. Para hacer la tarea más fácil, se necesitan almacenamiento de data no relacionales o las llamadas NoSQL. En la Figura 4 se muestra las aplicaciones de las base de datos NoSQL según cada necesidad.



Aplicaciones de las base de datos NoSQL

La capa de almacenamiento típicamente usa un proceso de batch para cargar data. El componente de integración de la capa de Ingestión, invoca varios mecanismos – tales como Sqoop, MapReduce, y otros – para subir data a la capa de almacenamiento, la cual proporciona patrones de almacenamiento que pueden ser implementados basados en los requerimientos de rendimiento, escalabilidad, y disponibilidad.

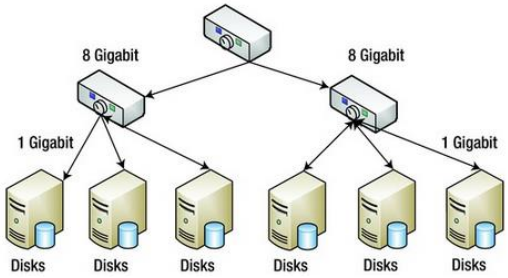
Capa de Infraestructura

La capa de infraestructura es aquella que sirve de soporte para la capa de almacenamiento, por lo cual es un nivel fundamental de la operación y escalabilidad de la arquitectura del Big Data. De hecho, la disponibilidad de una infraestructura física económica ha disparado el crecimiento del Big Data como una tendencia importante.

Hadoop utiliza una infraestructura física basada en un modelo de cómputo distribuido. Esto significa que la data puede ser almacenada físicamente en

diferentes ubicaciones y luego unirla a través de redes y sistemas de archivos distribuidos, formando una arquitectura "share-nothing", donde la data y las funciones requeridas para manipular la data, reside en un solo nodo.

Hadoop y HDFS puede administrar la capa de infraestructura en un ambiente de nube virtualizado (on-demand tal como nubes públicas) o en una red de commodity servers sobre una red de alto rendimiento (Gb – 10Gb). Típicamente para un ambiente de Big Data de alto rendimiento (20-40 nodos por rack), los requerimientos mínimos son procesadores de 8 núcleos, 24Gbs RAM, 4 a 12 TB de disco duro. En la Figura 5 se muestra una configuración simple de hardware para Big Data utilizando commodity servers.



Configuración simple de hardware para Big Data

Capa de Administración

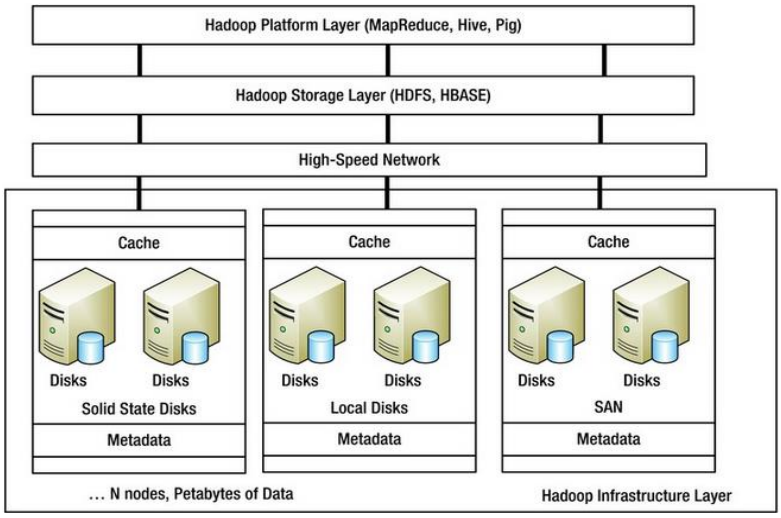
La capa de administración proporciona las herramientas y el lenguaje de query necesario para acceder a las base de datos NoSQL usando el sistema de almacenamiento de archivos HDFS. Con la evolución de las tecnologías computacionales, es posible administrar inmensos volúmenes de data los cuales previamente habrían sido manejados solo por supercomputadores a un gran costo.

Hadoop y MapReduce son tecnologías que permiten a las empresas almacenar, acceder, y analizar grandes cantidades de data en tiempo real, para luego poder monetizas los beneficios de contar con tales cantidades de data. Dichas tecnologías llevan a uno de los problemas fundamentales: capacidad de procesar cantidades masivas de data de una forma eficiente, económica y en poco tiempo. Es por esto que la capa de administración de Hadoop utiliza queries, y administra las capas inferiores usando lenguajes de programación como Pig y Hive. Se usan varios patrones de acceso de data según cada escenario.

El patrón recomendado para el acceso a la data en una plataforma Hadoop es el mostrado en la Figura 6.

El Big Data se ha convertido en una funcionalidad muy popular en las compañías, por lo que la seguridad de la data se convierte en una preocupación

importante. Los hábitos de compra de los clientes, las historias médicas de los pacientes, tendencias de costos útiles, y descubrimientos demográficos para enfermedades genéticas –todas las áreas anteriores y otras muchos tipos de data necesitan ser protegidas, tanto para cumplir requerimientos de conformidad como para proteger la privacidad individual. Métodos adecuados de autorización y autenticación deben ser aplicados al análisis de Big Data. Estos requerimientos de seguridad deben formar parte del inicio de implementación y no ser una idea tardía o reflexión luego de la implementación.



Patrón para el acceso a la data en una plataforma Hadoop

Capa de Seguridad

¿Cuáles son las medidas de seguridad básicas que una arquitectura de Big Data debe tener?

En primer lugar, un *mapper* o *named node job tracker*, puede entregar resultados inesperados que generarán resultados incorrectos. Con grandes cantidades de data, tales violaciones de seguridad pueden pasar desapercibidas y causar daños significativos a las inferencias y cómputos.

NoSQL injection, es un target fácil para los hackers. Con grandes clusters utilizados de forma aleatoria para almacenar sets de Big Data, es muy fácil perder el rastro de dónde está almacenada la data e inclusive es fácil olvidarse de borrar la data que no se necesita. Tal data puede caer en manos erróneas y dejar a la empresa en una amenaza de seguridad.

Los proyectos de Big Data están sujetos inherentemente a problemas de seguridad debido a la arquitectura distribuida, el uso de un modelo de programación simple, y los servicios de framework opensource. Sin embargo, la seguridad debe ser

implementada en una vía que no dañe el rendimiento, escalabilidad, o funcionalidad, y debe ser relativamente simple de administrar y mantener.

Para implementar una base de seguridad, se deberían tomar en cuenta como mínimo los siguientes puntos de diseño:

Autenticación de nodos usando protocolos como Kerberos

1. Activar cifrado de archivos de capa (file-layer encryption)
2. Subscribirse a un servicio de administración de llaves para certificados y llaves confiables.
3. Usar herramientas como Chef o Puppet para validación durante el desarrollo de los set de datos o cuando se apliquen parches a los nodos virtuales.
4. Guardar en logs la comunicación entre nodos, y usar mecanismos de registro para trazar cualquier anomalía entre las capas.
5. Asegurar la comunicación entre nodos resulta una práctica segura –Por ejemplo: usar SSL (Secure Sockets Layer – SSL), TLS, etc.

Capa de Monitoreo

Con tantos clústers de almacenamiento de data y diversos puntos de ingestión, es importante tener el cuadro completo del stack de Big Data para asegurarse que los SLAs[11] se cumplan con el mínimo tiempo de falla.

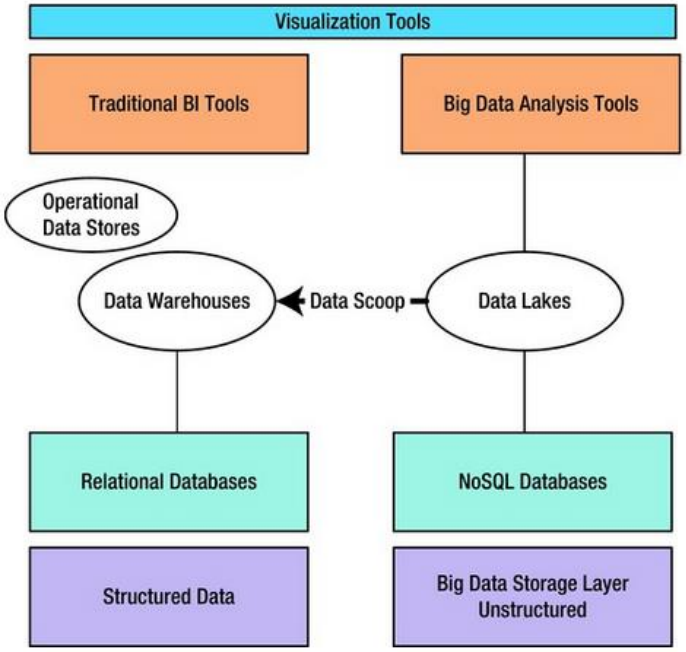
Los sistemas de monitoreo deben velar por los clúster de data que se desarrollan en un modo federado. El sistema de monitoreo debe estar al pendiente de los diferentes sistemas operativos y el hardware, así como también deberá proporcionar herramientas para el almacenamiento y visualización de la data. El rendimiento es un parámetro clave para monitorear, por lo que hay muy bajo costo operativo y alto paralelismo. Softwares de fuente abierta tales como Nagios son ampliamente usados para monitorear grandes stacks de Big Data.

Capa de Visualización

Un gran volumen de data puede generar sobrecarga de información. Sin embargo, si se incluyen métodos de visualización de forma temprana, como una parte integral del stack de Big Data, será de gran utilidad para los analistas de la data para obtener información rápidamente e incrementar su habilidad para observar diferentes aspectos de la data en varios modos visuales. Una vez que la salida del procesamiento de Hadoop sea recogido dentro de las tradicionales ODS (Operational Data Stores), data warehouse, y data marts, la capa de visualización puede trabajar

[11] SLAs: Service Level Agreements

en el top de esta agregación de data. Adicionalmente, si es requerido trabajar con información en tiempo real, los motores de tiempo real (real-time engines) pueden ser utilizadas. En la Figura 7 se muestran las interacciones entre las diferentes capas del stack de Big Data que permiten observar el poder de las herramientas de visualización.



Interacciones entre las diferentes capas del stack de Big Data

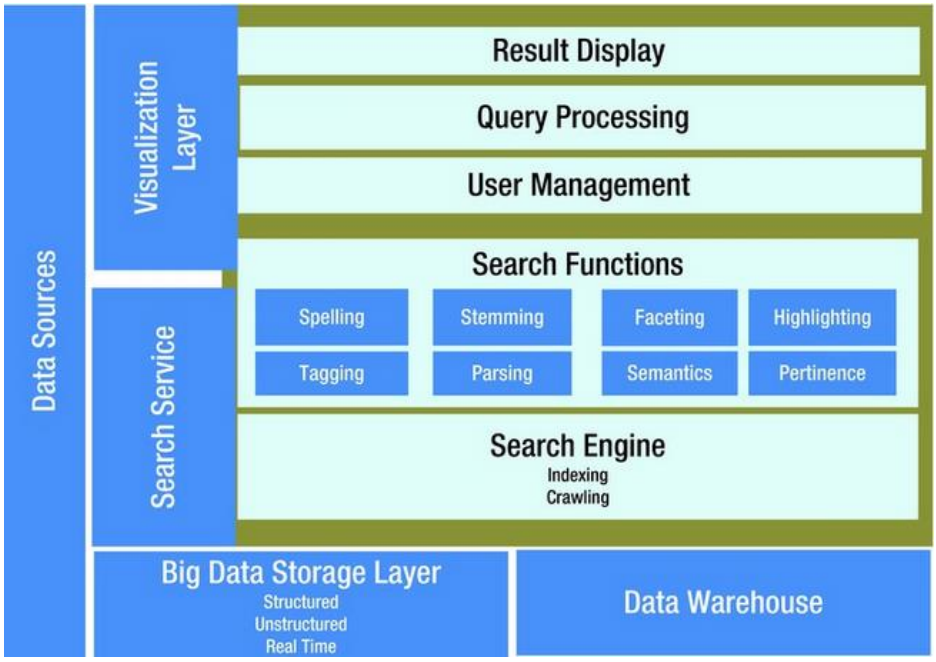
Motores de análisis

Las empresas necesitan adoptar diferentes enfoques para solucionar diferentes problemas usando Big Data; algunos análisis usan base de datos tradicionales, mientras otros usan Big Data combinado con métodos de BI[12] tradicionales. La mediación sucede cuando la data fluye entre las bases de datos tradicionales y los almacenes de Big Data (por ejemplo, a través de Hive/Hbase), en ambas direcciones, según los requerimientos, usando herramientas como Sqoop.

Search engine: Cuando se requiere el análisis de inmensos volúmenes de data, se necesitan motores de búsqueda con mecanismos de descubrimientos de data iterativos y cognitivos. La data cargada desde varias aplicaciones dentro del stack de Big Data debe ser indexada y buscada para el procesamiento del análisis de Big Data.

[12] BI: Business Intelligence

En la Figura 8 se muestra la arquitectura conceptual del nivel de búsqueda y como este interactúa con las otras capas del stack de Big Data.



Arquitectura conceptual del nivel de búsqueda

Real-time Engine: Para tomar ventaja de la información tan pronto como sea posible, las opciones en tiempo real están disponibles usando motores en tiempo real y almacenamiento de data NoSQL. En estos sistemas no es necesario esperar por una respuesta, ya que la velocidad de un almacenamiento NoSQL permitirá que los cálculos sean realizados a medida que la data esté disponible. Existen dos modos principales para el procesamiento en tiempo real: In-Memory Caching / In-Memory Database. En el modo de In-Memory Caching, la data es desplegada entre la aplicación y la base de datos para descongestionar la carga de la base de datos. En el modo In-Memory Database, la data es desplegada en la aplicación como una base de datos embebida.

Título 4 – Softwares usados en el análisis de data

SAS: Herramienta más ampliamente usada para análisis avanzado. Posee diversos algoritmos de modelamiento predictivo.

SPSS: Posee algoritmos eficientes de text mining y data mining.

R: Herramienta de fuente abierta para análisis de data, ampliamente usada. Tiene diversos paquetes para análisis de data.

MATLAB: Usada ampliamente para análisis numéricos y cómputo.

RapidMiner: Herramienta para la segmentación y clusterización; también puede ser usada para modelamiento convencional. Es de fuente abierta.

SAP: Herramienta para el manejo de operaciones de negocios y relaciones con clientes. Usada ampliamente como herramienta de rastreo de operaciones.

SAP HANA: Herramienta de base de datos para analytics en tiempo real.

Apache Mahout: Herramienta de análisis avanzado para Big Data. Es de fuente abierta.

Otras herramientas: Statistica, KXEN Modeler, GNU Octave, Knime.

Para elegir una herramienta de análisis se deben tener en cuenta varias consideraciones:

- La aplicación de negocio, su complejidad, y el nivel de experticia disponible en la organización.
- La información a largo plazo de los negocios, la información y la estrategia de análisis de la organización.
- Los procesos organizacionales existentes.
- Restricciones de presupuesto.
- Las inversiones propuestas o realizadas en el procesamiento de los sistemas de hardware, en las cuales se deben decidir factores tales como la potencia de procesamiento y la memoria disponible para que el software se pueda ejecutar de forma expedita.
- Estructura de la organización y dónde se llevará a cabo el trabajo de análisis.
- Ambiente de Proyecto y estructura de gobierno.
- Nivel de confort de la compañía al usar software de Fuente abierta, así como otras consideraciones legales.
- El tamaño de la data que será analizada.
- La sofisticación de gráficos y presentación requerida en el Proyecto.
- Cuáles serán las técnicas de análisis a ser utilizadas y que tan frecuentemente serán usadas.

- Cómo está organizada la data actual y cuan cómodo está el equipo que maneja la data.
- Consideraciones de sistemas Legacy. Tomar en cuenta si se está usando otra herramienta similar, y en caso de ser así, cuánto tiempo y recursos serán requeridos para un switch-over.
- Expectativas del ROI (Return of Investment) [11]

ANEXO II
Título I – Contenidos de cada Unidad Programática

Estructura de unidad programática

Unidad programática I

Módulos	Recursos	Contenidos
Introducción al Big Data	Clase expositiva Presentaciones	✓ Definición
		✓ Puntos clave del Big Data (5Vs)
		✓ Fuentes actuales de Big Data
		✓ Futuras fuentes generadoras de Big Data
		✓ Conceptos básicos en el ámbito del Big Data
		✓ Diferencia entre Big Data vs métodos tradicionales

Desafíos del Big Data	Clase expositiva Presentaciones	✓ Oportunidad del Big Data
		✓ Cloud Computing y Big Data
		✓ Big Data Analytics (Data estructurada y no estructurada)
		✓ IaaS, Daas, BDaaS
		✓ Introducción a Hadoop
Arquitectura del Big Data	Clase expositiva Presentaciones	✓ Definición de una arquitectura de referencia para el Big Data
		✓ Fuentes de data, Capa de Ingestión, Capa de Almacenamiento
		✓ Capa de Infraestructura, Capa de Administración de la plataforma,
		✓ Capa de Seguridad, Capa de Monitorización, Capa de Virtualización
		✓ Analytics Engines
		✓ Search Engines
		✓ Real-Time Engines

Unidad programática II

Módulos	Recursos	Contenidos
Almacenamiento del Big Data	Clase expositiva Presentaciones	✓ Recursos claves: Capacidad de procesamiento, memoria, almacenamiento, y red.
		✓ Patrones de almacenamiento
		✓ Arquitecturas de alto desempeño: HDFS, MapReduce.
Ecosistemas del Big Data	Clase expositiva Presentaciones	✓ Ecosistemas de Big Data: Hadoop, Zookeper, HBase, Hive, Pig, Mahout
		✓ Funcionamiento de Hadoop
		✓ Consideraciones y requerimientos técnicos
		✓ Base de datos para manejo de Big Data
		✓ Graph Analytics para Big Data
		✓ DaaS

Unidad Programática III

Módulos	Recursos	Contenidos
Impulsores de mercado para Big Data Analytics	Clase expositiva Presentaciones	✓ Puntos clave que impulsan la necesidad del uso de Big Data
		✓ Comparación de escenarios típicos y escenarios con Big Data
Estrategias de aplicación	Clase expositiva Presentaciones	✓ Big Data Analytics ✓ Preparar el ambiente para la escalabilidad
		✓ Plan estratégico para adoptar la tecnología
		✓ Keypoints
		✓ Estandarización de prácticas
		✓ Estudio del criterio de aceptación,
		✓ Proceso para la integración de la tecnología

Unidad Programática IV

Módulos	Recursos	Contenidos
Instalación de Apache Hadoop	Guía de laboratorio	✓ Instalación de paquetes necesarios
		✓ Instalación de Apache Hadoop ✓ Reconocimiento de los procesos de Hadoop
Uso de MapReduce	Guía de laboratorio	✓ Correr un programa de conteo de palabras con MapReduce
Instalación de Sqoop	Guía de laboratorio	✓ Descargar e instalar Sqoop ✓ Reconocer sus procesos
Instalación de Cloudera Manager Parte I	Guía de laboratorio	✓ Descargar y correr el instalador de Cloudera Manager Server
Instalación de Cloudera Manager Parte I	Guía de laboratorio	✓ Configurar un clúster de CDH