DISEÑO DE PROTOTIPO PARA LA IMPLEMENTACIÓN DE UN SISTEMA BIG DATA

LA TESIS PRESENTADA AL DEPARTAMENTO DE INGENIERÍA Y CIENCIAS BÁSICAS Y LA COMISIÓN DE ESTUDIOS DE PREGRADO DE LA UNIVERSIDAD POLITÉCNICO GRANCOLOMBIANO EN CUMPLIMIENTO PARCIAL DE LOS REQUISITOS PARA EL GRADO DE INGENIERO DE SISTEMAS

DANIEL ROMERO, CARLOS VARGAS

Diciembre 2015

DISEÑO DE PROTOTIPO PARA LA IMPLEMENTACIÓN DE UN SISTEMA BIG DATA

DANIEL LEONARDO ROMERO ALBARRACIN
CARLOS ANDRES VARGAS LOPEZ

DIRECTOR
ALEXIS ROJAS CORDERO
CANDIDATO A DOCTOR EN COMPUTACIÓN DE ALTAS PRESTACIONES
UNIVERSIDAD AUTÓNOMA DE BARCELONA

UNIVERSIDAD POLITÉCNICO GRANCOLOMBIANO
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
BOGOTÁ
2015

Certifico que he leído esta tesis y que, en mi opinión, es totalmente adecuada en alcance y calidad como una tesis para el grado de ingeniero de sistemas.

_____

Certifico que he leído esta tesis y que, en mi opinión, es totalmente adecuada en alcance y calidad como una tesis para el grado de ingeniero de sistemas.

_____

Resumen

En el pasado se decía que la información es poder, pero en la actualidad el cómo se usa esta información es lo que hace la diferencia, por esta razón se da inicio a esta investigación.

En este documento se pretende que el lector conozca el concepto de Big Data y los conceptos tecnológicos que lo rodean, desde el punto de vista de software y hardware que puede llegar a variar dependiendo de la necesidad.

El almacenamiento de información y la velocidad al manipularla se convierte en un problema para una empresa encargada de almacenar transmisiones generadas por GPS. Se propone un análisis y diseño de un sistema Big Data el cual permite hacerle frente a estos problemas.

También se presenta, las bases de datos no relacionales, conocidas como bases de datos **NoSQL**; abordando el sistema ***MongoDB*** y el framework ***Hadoop***, además de una comparación entre las sentencias en SQL y en ***MongoDB***. Incluyendo algunos algoritmos para el análisis de datos. Adicionalmente patrones y tendencias para definir un modelo de minería de datos.

Agradecimientos

Al finalizar un trabajo tan arduo y lleno de dificultades es inevitable pensar en que esto no hubiera sido posible sin la participación de personas que han facilitado las cosas para que este trabajo llegue a un feliz término.

A Alexis Rojas Cordero desde un inicio nos puso los pies sobre la tierra y nos dio toda su experiencia y guía.

A Wilmar Jaimes Fernández que nos dio un aporte importante para el resultado final.

A Diego Rodríguez por su apoyo y poner toda su confianza en nosotros.

A la familia Romero Albarracín y Vargas López quienes nos apoyaron todo el tiempo.

A nuestros maestros quienes nunca desistieron al enseñarnos.

Dedico esta tesis a Angie quien fue un gran apoyo emocional durante el tiempo en que escribía esta tesis.

A todos los que nos apoyaron para escribir y concluir esta tesis.

Para ellos es esta dedicatoria de tesis, pues es a ellos a quienes se las debemos por su apoyo incondicional.

Contenido

Tabla de contenido

Lista de Tablas

Lista de Figuras

Lista de Anexos

Glosario

Terabyte: Un terabyte es una unidad de cantidad de información que equivale a 10^{12} bytes.

GPS: GPS son las siglas de (global positioning system) el cual es un sistema el cual permite conocer en todo el planeta tierra la posición actual de un objeto con una gran precisión.

MongoDB: Es un motor de base de datos NoSql el cual es de tipo documental y es muy conocido como almacén Big Data en la actualidad.

Framework: Es una estructura conceptual la cual sirve de soporte para facilitar el desarrollo de software.

Hadoop: Es un framework que soporta aplicaciones distribuidas el cual permite trabajar con un flujo de datos muy alto. También sirve como almacén Big Data.

SQL Server: Es un motor de base de datos tradicional SQL de modelo relacional de tecnología Microsoft.

Match Insights: Herramienta de análisis de datos creada por SAP.

Colombia 3.0: Es la cumbre de contenidos digitales más importante de américa latina que reúne a expertos nacionales e internacionales.

SAP-SE: Es una empresa multinacional alemana, que se dedica al diseño de productos informáticos

Streaming: Es una transmisión continua digital de multimedia a través de un red de computadoras.

House of Cards: Es una serie dramática estadounidense creada por Netflix, a partir de análisis del consumo de usuarios con Big Data.

Activos S.A.: Es una empresa colombiana encargada de administrar integralmente el talento humano.

Nutresa: Es una empresa colombiana de alimentos de todo tipo.

Colombina S.A.: Es una empresa colombiana de alimentos enfocada en dulces.

Verizon: Es una compañía de banda ancha y telecomunicaciones.

Call center: Es un área donde personal especialmente entrenado presta un servicio para realizar/recibir llamadas a clientes, socios comerciales o compañías con un fin específico.

DB2: Es un sistema de gestión de base de datos relacional creado por IBM, el cual integra XML nativo.

Informix: Es un sistema de gestión de base de datos relacional que fue adquirido por IBM, en los años 90, fue tan popular como Oracle.

Cognos: Es un software encargado de realizar inteligencia de negocios y administración del desempeño.

Acrónimos

API: (Application Programming Interface) conjunto de funciones y procedimientos utilizado por un software para sus aplicaciones.

BSON: Binary JSON.

GPL: (GNU General Public License) Licencia Pública General de GNU.

HDFS: (Hadoop Distributed File System) sistemas de archivos distribuido Hadoop.

HTTP: (Hypertext Transfer Protocol) protocolo de transferencia de hipertexto.

JSON: (JavaScript Object Notation) formato ligero para el intercambio de datos.

NOSQL: (No Structured Query Language) bases de datos no-relacionales.

RDBMS: (Relational Database Management System) sistema de Gestión de Base de Datos Relacional.

SQL: (Structured Query Language) lenguaje declarativo de acceso a bases de datos relacionales.

XML: (Extensible Markup Language): lenguaje de marcas extensibles, utilizado por algunas bases de datos

START-UPS: Es un término utilizado recientemente tiene como intención, montar un nuevo negocio con base en la tecnología, el cual necesita un apoyo generalmente monetario para empezar a trabajar en sus diferentes ideas.

SPIN-OFFS: Se refiere a un Proyecto que nace como continuación de un proyecto anterior, ya sea mejorándolo o utilizándolo como base.

SPSS: Es un software estadístico desarrollado por IBM caracterizado por trabajar con grandes bases de datos de una forma muy sencilla.

Capítulo 1

1 Introducción

En el año 2014 Brasil organizó el mundial de futbol, dando como ganador a Alemania. Por medio del evento Colombia 3.0, obtuvimos la información de que éste acontecimiento tan importante para el futbol, se había realizado con ayuda de 'Big Data'

SAP-SE, es una empresa multinacional alemana, que se dedica al diseño de productos informáticos, ésta empresa creó una herramienta de análisis de datos llamada "Match Insights", que analiza los datos generados por videos de cámaras instaladas en la cancha de futbol. SAP almacena esta información, y la examina, para que los entrenadores puedan tomar decisiones en tiempo real, calificar con más rapidez y veracidad el desempeño de un jugador. El futbol está cambiando a través del análisis de datos, dado que ofrece ventajas al que puede tener acceso a esta información.

Se espera que otros deportes empiecen a cambiar por medio del análisis de datos.

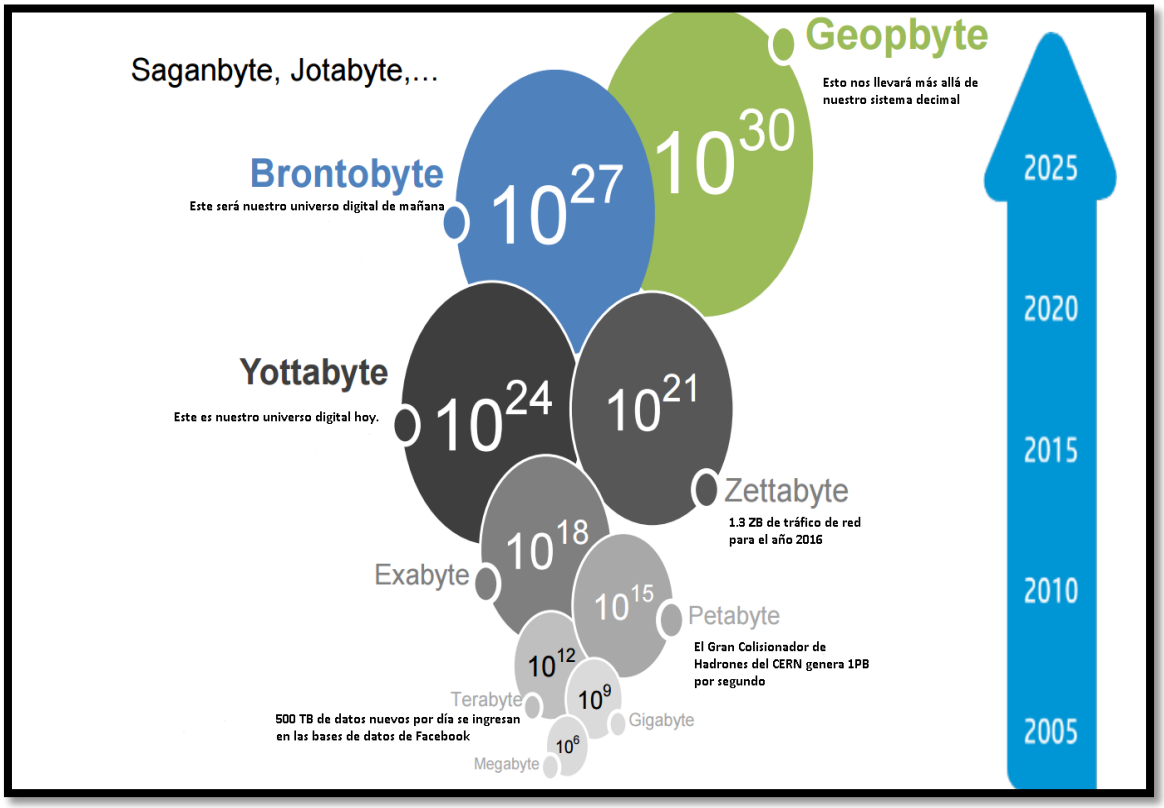


Figura 1 Escala de datos generados en el mundo. Fuente: [1], [2].

Analicemos la figura 1 en términos de bytes:

Megabyte	=	10^6	=	1,000,000
Gigabyte	=	10^9	=	1,000,000,000
Terabyte	=	10^{12}	=	1,000,000,000,000
Petabyte	=	10^{15}	=	1,000,000,000,000,000
Exabyte	=	10^{18}	=	1,000,000,000,000,000,000
Zettabyte	=	10^{21}	=	1,000,000,000,000,000,000,000
Yottabyte	=	10^{24}	=	1,000,000,000,000,000,000,000,000
Brontobyte	=	10^{27}	=	1,000,000,000,000,000,000,000,000,000
Geopbyte	=	10^{30}	=	1,000,000,000,000,000,000,000,000,000,000
Saganbyte	=	10^{33}	=	1,000,000,000,000,000,000,000,000,000,000,000
Jotanbyte	=	10^{36}	=	1,000,000,000,000,000,000,000,000,000,000,000,000

Tabla 1 Escala de equivalencias en bytes. Fuente: los autores.

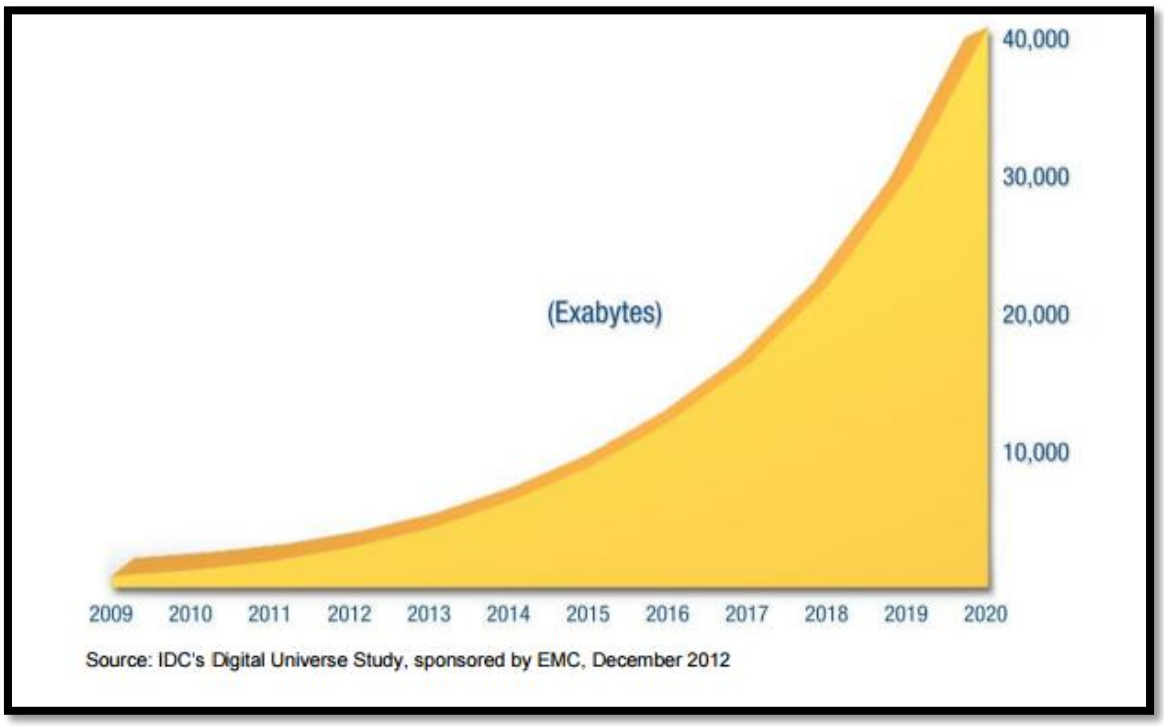


Figura 2. Los datos a través del tiempo. Fuente: [2].

Analizando la figura N°2, podemos evidenciar el crecimiento de los datos a través del tiempo. En la actualidad (año 2015), se están generando 10.000 Exabytes, eso quiere decir 1^{22} bytes y se pronostica que para el 2020 se produzcan 40.000 Exabytes, es decir 4^{22} bytes.

Como podemos apreciar en la figura N°1, se pronosticar que en el año 2030, el mundo superará la barrera del Geopbyte, razón por la cual, es momento de incursionar en el estudio, investigación y desarrollo de metodologías, prototipos y herramientas que permitan afrontar dicho reto.

Una de las razones que nos ha llevado a incursionar en estos ambientes, es debido a la necesidad de aprender a manejar este tipo de tecnologías que no son del todo conocidas en Colombia, pero que se requieren para el desarrollo de la Nación.

Capítulo 2

2 Generalidades

2.1 Antecedentes

Si ponemos la mirada en la actualidad, se puede percibir que cada día son más los datos generados por: los dispositivos móviles, las investigaciones científicas, las redes sociales, los dispositivos GPS, etc. En resumen, la humanidad genera cada día más y más datos que pueden ser almacenados y analizados, esto se convierte en un reto desde el punto de vista computacional.

Actualmente los datos son almacenados y analizados en sistemas computacionales. Esto nos conduce a enfrentar dos problemas principales:

- **El almacenamiento:** debido al espacio que se requiere para almacenar los datos, es necesario recurrir a las bases de datos distribuidas, esto permite almacenar grandes volúmenes de datos, y a modelos de bases de datos no convencionales como **NoSQL**.
- **La velocidad:** debido a que los datos almacenados deben ser procesados, analizados, depurados, etc. El rendimiento en términos de velocidad debe ser eficiente, una vez más se recurre a los sistemas distribuidos para dar solución a este reto, además la información tratada en muchos casos es en tiempo real y lo más importante en este caso es el acceso de forma instantánea.

Sin embargo existen más retos por cubrir como son la veracidad y la variedad.

Según un artículo publicado el 8 de Julio de 2015 por la revista Dinero de Colombia: "*El año pasado el operador móvil norteamericano Verizon redujo la fuga de clientes sometiendo a análisis el sentimiento detectado en las llamadas recibidas en su call center. Ningún analista humano podía procesar las llamadas que recibe el operador telefónico más grande de Estados Unidos, con más de 110 millones de abonados, pero una plataforma de analítica, dotada con un sistema de inteligencia artificial capaz de interpretar una llamada de voz o un trino en Twitter, sí lo hizo, y ayudó a tomar decisiones clave para el negocio*". [3]

En resumen lo anterior evidencia que con el uso de Big Data una empresa puede tener ventajas frente a sus competidores. Además es claro que Big Data se enlaza con otros avances tecnológicos como es la inteligencia artificial. Suena utópico que un sistema pueda analizar el sentimiento de las personas, pero eso se puede conseguir al día de hoy con el uso tecnología que en este caso particular permite entender y tomar mejores decisiones en lo que respecta al cliente.

En el mismo artículo y entrando en el tema de los GPS, se asegura que: "*La información emanada del GPS en cada auto vendido llega a manos de los fabricantes, que la utilizan para mejorar los diseños con base en los datos recibidos acerca del uso del automóvil. Varios operadores móviles en el mundo están aprovechando el conocimiento que tienen de la localización minuto a minuto de cada uno de sus clientes –porque el operador conoce en tiempo real en dónde está cada teléfono activo– para ofrecer a terceros la posibilidad de enviar publicidad contextual. Si estoy cerca de una tienda Starbucks*

podría recibir un SMS con la oferta del día. De ese tamaño es la batalla por los clientes." [3]

Telefónica Brasil, tiene listo un proyecto que permite mejorar el plan del transporte masivo en tiempo real, se pretende que por medio de las transmisiones del celular a las que tienen acceso, se ubiquen los lugares o estaciones donde más se concentra la gente a tomar el transporte y con esta información realizar un plan mucho más efectivo, esta información se obtuvo en el evento Colombia 3.0 que se realizó el año 2015.

Es acá donde queremos hacer un alto y le preguntamos al lector ¿La privacidad se está comprometiendo con el uso de la tecnología y en especial con el análisis de datos?

Según César Ayala, director de ventas de innovación para la región, de SAP, "*La era de internet de las cosas y de la social media ha generado un nuevo orden en el mundo de los negocios y las compañías no están utilizando todas las posibles fuentes de información para tener una foto profunda de la realidad*" Por este argumento tenemos el deber de presentar esta investigación y permitir que una empresa conozca las herramientas que necesita para convertir datos en información.

Ahora vamos a presentar un caso en donde la televisión orientada al Streaming y el procesamiento de información hicieron que a través de la analítica automatizada una empresa lograra tener éxito, siguiendo el artículo también se encontró lo siguiente: "*Netflix es uno de los ejemplos emblemáticos del provecho que puede sacarse del enfoque Big Data. Asesorados por la firma norteamericana Teradata, el popular servicio de video por demanda debe buena parte de su éxito a la analítica automatizada. El perfil de House of Cards se construyó con base en el análisis de las reacciones, hábitos y gustos detectados en la masa enorme de clientes.*

Netflix toma nota de qué escena transcurría cuando un cliente en cualquier país abandonó la película, y con la información de que dispone puede clasificar internamente sus filmes y series considerando las emociones que suscitan entre el público y ya no las viejas categorías de edad y sexo en que se basa la televisión tradicional. Y allí están los resultados. Netflix está redefiniendo el futuro de la televisión en el mundo." [3]

Empresas tan grandes y exitosas como las nombradas anteriormente se dieron cuenta de lo que representa la información, y que al usarla tienen ventaja de sus competidores y hacen que sus servicios tiendan a ser exitosos.

Los 15 grandes del Big Data:

1. **IBM:** *"IBM fue el proveedor más grande de Big Data en el 2012 con un ingreso de 1,3 mil millones de dólares, según un reporte reciente de Wikibon, gracias a la venta de productos y servicios relacionados con Big Data. Las ofertas incluyen hardware de servidor y de almacenamiento, software de base de datos, aplicaciones analíticas y servicios asociados. Los productos más conocidos son las plataformas de base de datos DB2, Informix e InfoSphere, y las aplicaciones analíticas Cognos y SPSS. IBM también apoya la plataforma de análisis de datos de código abierto Hadoop."* [4]

2. **HP:** "*HP fue el segundo proveedor más grande de Big Data en el 2012 por sus ingresos de 664 millones de dólares. Esta empresa también ofrece una mezcla de hardware, software y servicios, y es conocida por la plataforma de análisis Vertica*" [4]

3. **Teradata:** "*Teradata fue el tercer proveedor más grande de Big Data del 2012 con un ingreso de 435 millones de dólares. Esta es conocida por sus plataformas de hardware, de software analítico y de base de datos. También ofrece herramientas analíticas específicas para industrias de distribución y transporte*" [4]

4. **Oracle:** "*Aunque Oracle es conocido principalmente por su conocida base de datos, también es un gran jugador en el ámbito de Big Data. Su Oracle Big Data Appliance combina un servidor Intel, distribución Hadoop de Cloudera y la base de datos NoSQL de Oracle. Fue el cuarto gran proveedor en el 2012 con un ingreso de 415 millones de dólares*" [4]

5. **SAP:** "*SAP ofrece una variedad de herramientas analíticas, pero es más conocido por su base de datos en memoria, HANA. Fue el quinto gran proveedor de Big Data en el 2012 con un ingreso de 368 millones de dólares*" [4]

6. **EMC:** "*EMC ayuda a las compañías a almacenar y analizar Big Data y es también la sede del Marketing Science Lab, un think tnak en análisis de Big Data que se enfoca en analizar datos de marketing. Esta primavera ocupó los titulares con su spinf-off de Pivotal, también respaldado por VMware y General Electric. Pivotal combina el Hadoop con la base de datos Greenplum de EMC y herramientas de consulta HAWQ. EMC fue el sexto gran proveedor de Big Data en el 2012 con un ingreso de 336 millones de dólares*" [4]

7. **Amazon:** "*Amazon es conocido por su plataforma en la nube, pero también ofrece un número de productos de Big Data, incluyendo el Elastic MapReduce basado en Hadoop, la base de datos Big Data DynamoDB, el almacén de datos paralelamente masivo RedShift, y todos funcionan bien con Amazon Web Services.*" [4]

8. **Microsoft:** "*La estrategia de Big Data de Microsoft incluye una asociación con Hortonworks, una empresa nueva de Big Data, y la herramienta HDInsights basada en la plataforma de datos de Hortonworks. Microsoft también es conocida por su servidor de base de datos SQL y fue el noveno gran proveedor de Big Data en el 2012 con un ingreso de 196 millones de dólares*"

9. **Google:** "*Las ofertas de Big Data de Google incluyen BigQuery, una plataforma de análisis de Big Data basada en la nube. La compañía recibió 36 millones de dólares en ingresos relacionados con Big Data el año pasado.*" [4]

10. **VMware:** "*VMware es conocido por sus soluciones de virtualización y de nube, pero se está convirtiendo en un muy buen jugador de Big Data. En Junio publicó el anuncio de VMware vSphere Big Data Extensions, el cual permite que vSphere controle las implementaciones de Hadoop y hacer que lanzar proyectos de Big Data se vuelva mucho más sencillo para las empresas. VMware recibió, el año pasado, 32 millones de dólares en ingresos relacionados con Big Data, casi tanto como Google.*" [4]

11. **Cloudera:** "*Cloudera está en la lista de los principales proveedores de Big Data con más de 141 millones de dólares en fondos de capital de riesgo y ha atraído a varios fundadores conocidos y de gran nombre en Big Data que vienen de Google,*

Facebook, Oracle y Yahoo. La compañía lanzó por primera vez la plataforma Apache Hadoop para clientes empresariales en el 2008." [4]

12. **Hortonworks:** "*Hortonworks es otro proveedor de Hadoop y ha recibido más de 70 millones de dólares en inversiones de capital de riesgo luego de la escisión de Yahoo en el 2011. Está creciendo para ir directamente contra Cloudera, y es muy conocido por sus alianzas estratégicas con Microsoft, Rackspace, Red Hat, Teradata y otras compañías.*" [4]

13. **Splunk:** "*Splunk tuvo la mayor cuota de mercado de todos los vendedores de Big Data únicamente, con un ingreso de 186 millones de dólares, según Wikibon. La compañía se especializa en análisis de datos de máquinas."* [4]

14. **10Gen:** "*10Gen es conocida por su código abierto MongoDB que es la base de datos NoSQL líder. Entre los inversores estratégicos se encuentra Intel, Red Hat e In-Q-Tel. El año pasado, 10Gen quedó tercero entre los vendedores de Hadoop y NoSQL únicamente, con un ingreso de 36 millones de dólares*". [4]

15. **MapR:** "*Conocido por M7, su base de datos NoSQL, MapR funciona con la plataforma de Amazon en la nube y con Google Compute Engine. El año pasado quedó cuarto entre los vendedores de Hadoop y NoSQL únicamente, con un ingreso de 23 millones de dólares.*" [4]

Big Data en Colombia:

El articulo nos entrega la siguiente información: "*Colombia va a experimentar en breve un boom de Big Data en virtud de la próxima implementación de la factura electrónica, que le permitirá a la Dian registrar cada compra cuando esta tenga lugar.*

Hoy la Dian recibe esta información mediante los reportes de los sistemas ERP de las empresas, en los periodos de declaraciones. La factura electrónica permitirá obtener esa información minuto a minuto, lo cual permitirá un análisis gigantesco de patrones de evasión, según explica el experto Víctor Muñoz, vicepresidente comercial y de mercadeo de Carvajal Tecnología y Servicios, que ofrece consultoría y soluciones de Big Data" [3]

Recolectando información encontramos 3 casos de empresas que aplican Big Data:

- **Activos S.A.**, de la mano de IBM implemento: "u*n nuevo sistema experto que permite consolidar una plataforma para soportar los procesos críticos de negocios del cliente.*" [5]
- **Nutresa**, implemento un sistema que: "*Permite controlar los procesos financieros, logísticos, de distribución, de contacto con los proveedores y clientes, de ventas y de marketing.*" [5]
- **Colombina S.A.**, implemento un sistema que: *"Permitió mejorar la distribución de la información de las ventas de cada una de sus sucursales.*" [5]

La información recolectada en el texto anterior nos lleva a la conclusión que las empresas de Colombia y el mundo cada día están haciendo más uso de Big Data para la toma de decisiones, creando la necesidad de científicos de datos y científicos de computación.

En marzo de 2014, el gobierno de Colombia a través de Colciencias, pretendia construir una Herramienta de consecución de datos y análisis de los mismos, que sirvan para consumo nacional y creación de start-ups con base en spin-offs. Busca con esto; Crear El CEA - Centro de Excelencia y Apropiación, este debe tener:

- Talento Humano capacitado, disciplinario y poli-Valente.
- Capacidad de análisis en la ingeniería de los datos.
- Contar con grupos de Investigación de alto nivel para formar talento humano a nivel de.
  - Maestría
  - Doctorado
  - Post-doctorado en temáticas de los datos
- Se busca concebir un Big-Player (proveedor mundial de información como datos de desastres, pestes (ébola), epidemias, seguridad policial (interpol), datos para los físicos (ejemplo: para el CERN), datos para los químicos como Dinámica Molecular, secuenciación genómica, etc. Finanzas como los Datawarehose, bancos de estudiantes potenciales para las universidades, etc.
- Ofrecer servicios de análisis y procesamiento de datos.
- Desarrollar soluciones innovadoras con base en los datos, por ejemplo para móviles.
- Formar talento humano en manejo de grandes volúmenes de datos.

El 2 de septiembre de 2015 en la Institución Universitaria Politécnico Grancolombiano, se llevó a cabo una conferencia sobre Big Data, dada por el señor Steven Adler que es jefe de estrategia de Información de IBM y líder mundial en desarrollo tecnológico por cerca de 20 años. En esta conferencia se habló sobre Open Data y se mencionó un término llamado People Data.

Open Data: También llamados datos abiertos, se consideran una práctica que se centra en que ciertos datos (no importa los que sean), estén disponibles para cualquier persona del mundo, sin ninguna restricción de derechos de autor, de patentes u otras formas de control.

"Podemos considerar datos abiertos a todos aquellos datos accesibles y reutilizables, sin exigencia de permisos específicos. Cabe la posibilidad de que los tipos de reutilización puedan estar controlados mediante algún tipo de licencia." [6]

People Data: Según el señor Steven Adler, un nuevo concepto se avecina, People Data o Datos de las personas, este concepto se expone explicando la forma en que las personas del común pueden generar datos libres, por ejemplo si alguien del barrio se da cuenta que hay muchos habitantes de la calle alrededor, entonces puede crear una base de datos abierta y hacer que todos sus vecinos colaboren, puede ser con fotos o lugares donde suelen estar, de esta manera es posible que una ONG o alguien en el mundo que pueda ayudar tenga información precisa y pueda dar una mano. [7]

Steven conto al público sobre su experiencia en África, se llevó una sorpresa cuando noto que en Guinea, Liberia y Sierra Leona lo que faltaba cuando el ébola golpeó el África occidental a principios de enero 2014 fue información de calidad sobre la capacidad de salud en toda la región, la comunicación y la coordinación, la gobernanza, y los procedimientos prácticos sobre cómo manejar las epidemias y emergencias de salud. [8]

Encuentro Mundial Big Data, Bogotá 28 y 29 de Octubre de 2015:

CEA

El 29 de Octubre de 2015 en la ciudad de Bogotá se lanzó el CEA (Centro de excelencia y apropiación) Big Data y Data Analytics. De que se trata del CEA según Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia "*Se trata de impulsar, promocionar y desarrollar las Tecnologías de la Información y las Comunicaciones en la industria, la academia y los diferentes sectores que puedan intervenir en ella, con el fin de contribuir al incremento de la productividad y la competitividad tanto del sector TIC como de otros sectores de la economía donde la analítica, la ciencia y la ingeniería de los datos generan valor."*

WATSON - IBM

Este encuentro nos sirvió para conocer las tres fases por las que ha pasado ***Watson:***

1. En la primera fase se nos dijo que esta máquina logro vencer al mejor jugador de ajedrez del momento Gari Kaspárov en 1997.
2. En la segunda fase se nos dijo que esta máquina logro vencer a los mejores jugadores del concurso de televisión Jeopardy, la dificultad de este juego es que el moderador da la respuesta y el concursarte debe dar la pregunta, el que más este cerca de la pregunta es el que gana. Watson hizo el mayor puntaje posible en el programa.
3. En la tercera fase se nos dijo que esta máquina ahora está incursionando en la medicina. Se puso como ejemplo el uso de Watson en un diagnóstico para una señora que tenía cáncer de seno. El doctor había dado su diagnóstico, sin embargo dotaron a Watson de la información necesaria para que pudiera dar un veredicto y para la tranquilidad del doctor, su diagnóstico era muy parecido al que decía Watson. Este apoyo que presenta Watson, dando una mano a los médicos nos da un indicio de lo que se puede conseguir, usando los datos de manera correcta.

Podemos evidenciar que en la actualidad ya existen avances tecnológicos de grandes dimensiones, estos avances permiten que la humanidad viva un poco mejor y tome mejores decisiones.

El científico de datos

El científico de datos es una nueva profesión que nació de la necesidad de tener profesionales en el área de Big Data. Es una profesión que a medida de que pase el tiempo se convertirá en una de las solicitadas, además de eso es una de las profesiones que tienen mejor remuneración. Esta información fue obtenida en el encuentro mundial Big Data.

Queremos presentar la red social para científicos de datos, en esta se presentan ofertas laborales y también es posible conectar con otros científicos de datos alrededor del mundo. Para acceder es necesario ingresar a este link https://www.kaggle.com/.

2.2 Planteamiento del problema

Big Data es un concepto que está tomando auge en los últimos años, sin embargo ya fue usado por empresas como Google con su buscador. Este concepto está llegando poco a poco a Colombia, y para las pequeñas empresas es muy difícil entender a que se enfrentan sin que exista un documento claro. No conocemos una guía paso a paso que permita implementar un sistema Big Data, tampoco que tecnologías se pueden usar para dicho fin.

Con base en lo expuesto anteriormente se plantean las siguientes preguntas:

- ¿Su empresa está preparada para Big Data?
- ¿Su empresa conoce las herramientas que debe usar para la implementación de un sistema Big Data?
- ¿Su empresa conoce las opciones que puede usar para analizar sus datos y sacar provecho de ello?

2.3 Objetivos

2.3.1 Objetivo general

Realizar el análisis y diseño de un modelo para la implementación de un sistema que requiera aplicar la tecnología del Big Data.

2.3.2 Objetivos específicos

- Realizar el estado del arte de “Big Data”
- Elaborar el análisis y diseño del prototipo del modelo con base en datos de prueba
- Plasmar la investigación en un documento de entrega final
- Elaborar un artículo de revisión para publicar en una revista indexada en Colciencias.

2.4 Justificación

Recopilamos los siguientes datos que nos permiten saber la relevancia este proyecto en el área de conocimiento:

- 2.5 quintillones de bytes de datos son creados cada día (quintillón = 1,000,000,000,000,000,000 bytes).
- “Big Data es la posibilidad de tratar con velocidad una cantidad ingente de datos de origen muy variado, las llamadas tres V: volumen, velocidad y variedad”, explicó Juan Luis Quincoces. “Siempre se dijo que la información es poder y nunca hemos estado tan cerca de analizar y usar tanta información como la que emana del llamado Big Data”, añadió.
- ‘Big Data’ aporta una información que podrá acelerar la investigación en el ámbito celular, evitar epidemias, prever incendios, ayudar en la investigación ambiental, participar de la seguridad física y virtual, luchar contra el crimen o personalizar los servicios.

- Una de las ventajas más importantes que posee el análisis de Big Data es su capacidad para ofrecer las mejores decisiones de negocio en una fracción del tiempo.

Entonces con esta información recopilada podemos asegurar que Big Data es relevante en la computación ya que se usan teorías matemáticas para que por ejemplo a través de la computación, un gerente de una empresa pueda tomar mejores decisiones.

En principio este proyecto es importante en el politécnico Grancolombiano, ya que se va a realizar un aporte a la investigación de la universidad y además se realizará la aplicación de los conocimientos adquiridos en sistemas distribuidos, sistemas operacionales, programación, bases de datos, etc., donde tanto el software como en el hardware y en general, las tecnologías están cambiando todo el tiempo.

Un ejemplo es el caso de las bases de datos relacionales las cuales brindan una excelente estructura, pero cuando su volumen es muy amplio o tiene mucha concurrencia de usuarios empiezan a tener problemas de rendimiento; en el caso de la Big Data se pretende que este problema se pueda solventar.

2.5 Delimitación

2.5.1 Tiempo

Se hará uso de la aplicación Trello [9] para llevar un control sobre lo que se está haciendo del proyecto, cada integrante del proyecto tendrá asignada tareas específicas:

Fase	**Mes 1**	**Mes 2**	**Mes 3**	**Mes 4**
Consulta Bibliográfica	X	X	X	
Elaboración del Documento y el artículo	X	X	X	X
Desarrollo del modelo	X	X	X	
Pruebas			X	X
Preparación de Sustentación				X

Tabla 2 Tiempos de ejecución de la tesis. Fuente: los autores.

2.5.2 Alcance

Este proyecto está dirigido a una empresa que ofrezca servicios de rastreo de vehículos a través de GPS, donde se generan cantidades ingentes de datos y se enfrentan al problema de almacenar y tratar esta información.

Dada la magnitud y complejidad del proyecto, se desarrollará solamente el análisis y diseño de un prototipo para la implementación de un sistema aplicando la tecnología del Big Data.

Capítulo 3

3 Marco teórico

3.1 ¿Qué es Big Data?

Según IBM, "*La tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos."* [10]

Según Wikipedia, "*El Big Data o Datos masivos es un concepto que hace referencia a la acumulación masiva de datos y a los procedimientos usados para identificar patrones recurrentes dentro de esos datos. Otras denominaciones para el mismo concepto son datos masivos o datos a gran escala.*" [10]

Según SAS, "*Es la expansión de datos estructurados y no estructurados que inunda su organización todos los días; y si se manejan bien, pueden proveer información poderosa.*

Imagine poder analizar datos para determinar la causa de origen de fallas, o detectar comportamiento fraudulento antes de que afecte los ingresos. Implementar las soluciones correctas para sacar el mayor provecho de Big Data (del manejo a la analítica de datos) puede ser clave para el éxito de su negocio." [10]

Es importante aceptar que Big Data significa diferentes cosas para muchas personas eso depende del enfoque que se le quiera dar por ejemplo, se dice que Big Data es el análisis de grandes cantidades de datos sociales, análisis de los medios de comunicación, datos en tiempo real, etc. Entonces se entiende que Big Data es toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales, estos datos pueden ser estructurados o no estructurados, y si se manejan bien se obtendrán ventajas alucinantes.

Las 4 V de Big Data:

En la mayoría de artículos se habla de las 3 V's, sin embargo IBM implementa una cuarta V que se reflejada en las siguientes definiciones, estas se hacen en base a una encuesta que el mismo IBM realizo:

- ***Volumen:*** *La cantidad de datos. Siendo quizá la característica que se asocia con mayor frecuencia a big data, el volumen hace referencia a las cantidades masivas de datos que las organizaciones intentan aprovechar para mejorar la toma de decisiones en toda la empresa. Los volúmenes de datos continúan aumentado a un ritmo sin precedentes. No obstante, lo que constituye un volumen verdaderamente "alto" varía en función del sector e incluso de la ubicación geográfica y es más pequeño que los petabytes y zetabytes a los que a menudo se*

hace referencia. Algo más de la mitad de los encuestados consideran que conjuntos de datos de entre un terabyte y un petabyte ya son big data, mientras que otro 30% simplemente no sabía cuantificar este parámetro para su empresa. Aun así, todos ellos estaban de acuerdo en que sea lo que fuere que se considere un "volumen alto" hoy en día, mañana lo será más. [10]

- ***Variedad:***
  Diferentes tipos y fuentes de datos. La variedad tiene que ver con gestionar la complejidad de múltiples tipos de datos, incluidos los datos estructurados, semiestructurados y no estructurados. Las organizaciones necesitan integrar y analizar datos de un complejo abanico de fuentes de información tanto tradicional como no tradicional procedentes tanto de dentro como de fuera de la empresa.

  Con la profusión de sensores, dispositivos inteligentes y tecnologías de colaboración social, los datos que se generan presentan innumerables formas entre las que se incluyen texto, datos web, tuits, datos de sensores, audio, vídeo, secuencias de clic, archivos de registro y mucho más. [10]

- ***Velocidad:***
  Los datos en movimiento. La velocidad a la que se crean, procesan y analizan los datos continúa aumentando.
  Contribuir a una mayor velocidad es la naturaleza en tiempo real de la creación de datos, así como la necesidad de incorporar datos en streaming a los procesos de negocio y la toma de decisiones. La velocidad afecta a la latencia: el tiempo de espera entre el momento en el que se crean los datos, el momento en el que se captan y el momento en el que están accesibles. Hoy en día, los datos se generan de forma continua a una velocidad a la que a los sistemas tradicionales les resulta imposible captarlos, almacenarlos y analizarlos. Para los procesos en los que el tiempo resulta fundamental, tales como la detección de fraude en tiempo real o el marketing "instantáneo" multicanal, ciertos tipos de datos deben analizarse en tiempo real para que resulten útiles para el negocio. [10]

- ***Veracidad:***
  La incertidumbre de los datos. La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. Esforzarse por conseguir unos datos de alta calidad es un requisito importante y un reto fundamental de big data, pero incluso los mejores métodos de limpieza de datos no pueden eliminar la imprevisibilidad inherente de algunos datos, como el tiempo, la economía o las futuras decisiones de compra de un cliente. La necesidad de reconocer y planificar la incertidumbre es una dimensión de big data que surge a medida que los directivos intentan comprender mejor el mundo incierto que les rodea (véase el recuadro "Veracidad, la cuarta V". [10]

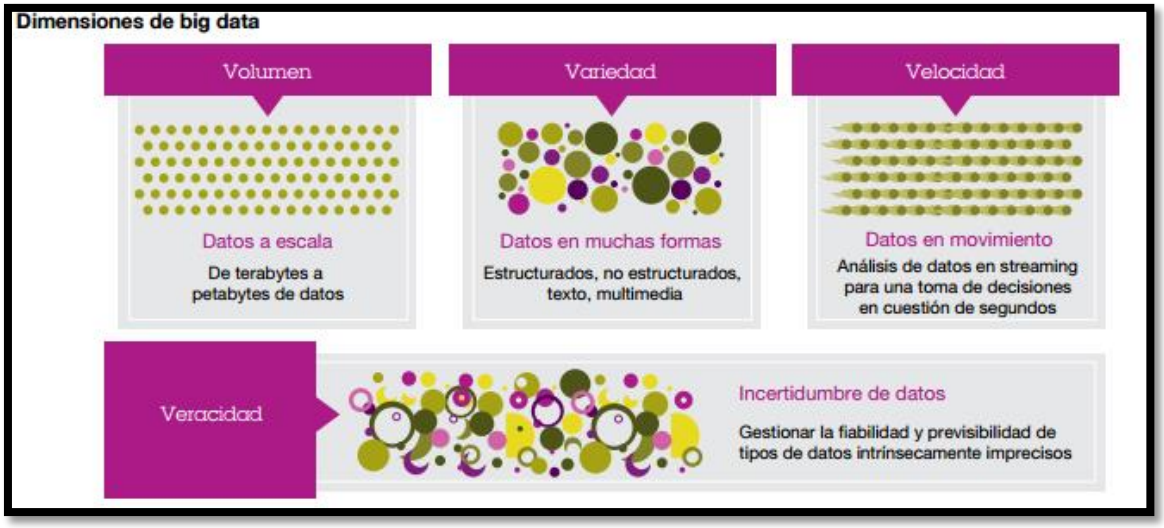


Figura 3 Cuatro dimensiones de Big Data. Fuente: [10].

Analizando la figura N°3 se puede apreciar que Big Data se divide en 4 grandes dimensiones el volumen, la variedad, la velocidad y por último la veracidad. Si lo que enfrenta tiene alguno de estos 4 retos entonces ya está hablando de un problema que puede solucionar aplicando tecnologías Big Data.

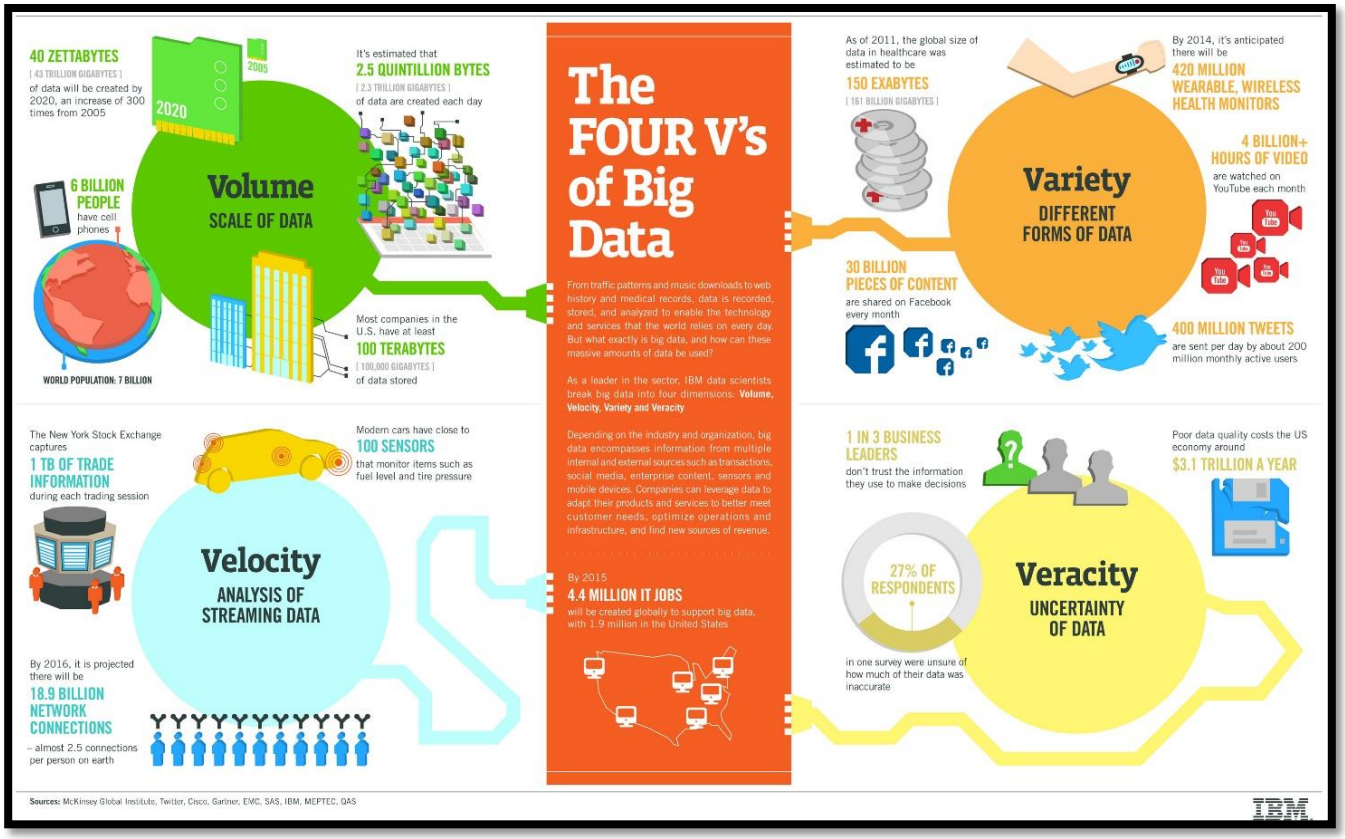


Figura 4. Las 4 V de Big Data - Fuentes, Mckinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS. Fuente: [10].

En la figura N°4, se pueden observar las dimensiones que componen Big Data, y se presentan unos datos interesantes sobre lo que representa en la actualidad cada una de las dimensiones, por ejemplo:

- Para el **volumen** se dice que 40 Zettabytes serán creados en 2020, es decir un aumento de 300 veces con respecto a 2005. 6 billones de personas tienen teléfonos celulares, de una población mundial de 7 billones. Se estima que 2.5 Quintillones de Bytes se crean cada día.
- Para la **velocidad** se dice que la Bolsa de Nueva York Captura 1 TB de información comercial durante cada sesión de negociación, y esta información debe estar disponible de manera instantánea. Los vehículos modernos tienen 100 sensores, que monitorean el vehículo satelitalmente. Para el 2016 se proyecta que habrá 18.9 billones de conexiones de red, eso equivale a 2.5 conexiones por cada persona en la tierra.
- Para la **variedad** se dice que a partir de 2011, el tamaño global de datos en la asistencia sanitaria se estimó en 150 exabytes. 400 millones de tweets son enviados por día en alrededor 200 millones de usuarios activos cada mes.
- Para la **veracidad** 1 de cada 3 líderes empresariales no confían en la información que utilizan para tomar decisiones. La mala calidad de los datos le cuesta a la economía de Estados Unidos alrededor de $ 3,1 trillones en un año. 27 % de los encuestados en una encuesta no estaban seguros de cómo gran parte de sus datos eran incorrectos.

3.2 ¿Por qué Big Data es importante?

La importancia de este método radica en que es una tecnología literalmente nueva, pero con un gran potencial el cual debe ser aprovechado en este momento.

Todos debemos estar cada vez más preparados para Big Data, por los retos tecnológicos que implica, y por los efectos, a menudo intangibles, que transformarán la manera en que trabajamos y vivimos. [11]

- ***NEW DATA**: "Los Big Data que se generan, se recopilan y se almacenan hoy en día contienen una enorme cantidad de información que antes simplemente no estaba disponible y era totalmente desconocida para nosotros.*

  La mayoría de los datos de Big Data tiene un contenido nuevo. Un ejemplo de esto, son aquellos relacionados al mundo del comercio electrónico. Atrás quedaron los días en que los sitios webs se dedicaban a la captura de datos transaccionales, como los detalles de las compras, para mejorar el cálculo de la demanda, la optimización del abastecimiento de existencias y los ajustes de precios.

  Hoy en día, los comercios electrónicos se centran en capturar el flujo de clics que los clientes realizan durante una transacción. Capturan la ruta que realizamos antes de finalizar la transacción o antes de abandonar nuestro carrito de compras. Estos datos incluyen información tangible del consumidor, como los estilos, los colores y los tamaños que busca.

  Pero ahora, más importante, también captura la información que el consumidor consultó durante el proceso; por ejemplo la opinión y la clasificación que otros

consumidores hicieron del producto, la sostenibilidad del proceso de fabricación y la disponibilidad de otros elementos complementarios.

El análisis del flujo de clics apunta a conocer mejor los procesos de compra humanos para adaptar nuestras futuras experiencias de compra y, por supuesto, para mejorar nuestra productividad de compra."

- ***UNLOCKING VALUE**:"El valor que contiene Big Data puede descubrirse a través de sus análisis automáticos, dado que Big Data son datos digitales. Los análisis de datos tienen la sorprendente capacidad de transformarlos en información nueva, la cual puede llevar a tomar medidas inteligentes.*

  Big Data por sí mismo no es la única razón por la cual es importante. Todos estos datos deben analizarse para develar el valor que encierran.

  El Business Intelligence (BI) realiza una búsqueda en todo el historial de transacciones de una base de datos para generar informes, modelar tendencias y proporcionar estadísticas sobre el rendimiento del sistema. Big Data no solo amplía la cantidad de datos actualmente disponibles para BI sino que también agrega un nivel de complejidad sin precedentes que amplía las fronteras de los análisis."

- ***SHAPING THE FUTURE**: "Es posible que el tipo, la profundidad y la sofisticación de los análisis que se pueden realizar hoy en día y en un futuro cercano nos permitan ser mucho más proactivos. Esto influye en cómo se perfila el futuro en oposición a solamente reaccionar ante las consecuencias imprevistas del pasado.*

  El tipo de información que encierra Big Data también nos ha permitido dar algunos saltos reemplazando el simple modelado del futuro en función del continuo de su historial, por la verdadera prevención e influencia en las acciones futuras.

  En el mundo de BI tradicional, la analítica realizaba análisis históricos para aprender del pasado y mejorar la eficiencia en el futuro, pero era trabajo del usuario identificar las partes pertinentes de la información y cómo utilizarla. En lugar de esto, los análisis de Big Data realizan análisis predictivos que buscan respuestas a preguntas como: ¿qué sucederá después? O bien, ¿qué sucedería si estas tendencias continúan? Y lo más importante, ¿por qué esté sucediendo esto y cómo se puede cambiar?

  Los análisis predictivos nos permiten modelar e influir en el futuro, al evitar que ocurran ciertos hechos y de esta forma cambiar el curso de las acciones. También nos permiten prever las preferencias de las personas y dar recomendaciones basándonos en ellas. Consideremos un negocio que ha prosperado gracias a este tipo de análisis: libros de Amazon.

  En esencia, la importancia se debe a que no solo captura el hecho sino también el comportamiento y los pensamientos de las personas, a un nivel muy fino de granularidad, casi en tiempo real.
  En realidad Big Data contiene información que puede revelarse mediante los análisis que a su vez se vuelven cada vez más predictivos.

Este tipo de información se incorporará a la estructura de nuestra vida diaria, transformando realmente la forma en que trabajamos y vivimos."

3.3 ¿Dónde aplicar Big Data?

Es evidente que Big Data se puede aplicar en todo aquello que necesite una solución que tenga que ver con las 4 V's. Big Data ya se aplicó como se explicó anteriormente en la selección Alemana de futbol, además en la campaña de reelección del presidente de estados unidos Barack Obama, también se aplicó para dar mayor rapidez al análisis del Genoma Humano, y el gran acelerador de hadrones en suiza es otro de los grandes ejemplos en donde se aplica Big Data.

Big Data es un universo en donde todo aquel que le gusten los retos tiene oportunidad. Desde un ingeniero, un economista, un matemático, un estadístico, un biólogo etc. Muchas ciencias están implicadas en este fenómeno llamado Big Data.

Se puede usar Big Data desde una pequeña empresa de ventas por internet hasta una empresa que maneje transmisiones de GPS, según su necesidad se puede dar una solución informática que use Big Data.

En bases de datos de un crecimiento horizontal desmesurado para realizar un correcto manejo del almacenamiento y análisis de los datos.

3.4 ¿Desde qué cantidad de información se considera Big Data?

El concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de *Petabytes* y *Exabytes* de datos.

Se puede considerar Big Data a todo aquel problema que presente un reto de:

- Almacenamiento
- Análisis
- Velocidad
- Veracidad
- Todas las anteriores en un mismo escenario,

Es decir Big Data no solo es almacenar información en grandes cantidades, también es analizar información o permitir que la información esté disponible en tiempo real.

3.5 Conceptos relacionados con Big Data

3.5.1 Almacén de datos o Data Warehouse

Es una colección de datos que permite almacenar información con diferente origen. Se caracteriza por su gran tamaño, y por ser el centro de toma de decisiones, es decir es acá donde se almacena y procesa la información. [12]

Se encontró que las principales diferencias entre una base de datos operacional y un almacén de datos son:

- Las bases de datos operacionales tratan datos operacionales, mientras que los almacenes de datos tratan datos del negocio para entregar información.
- En las bases de datos operacionales se almacena la información actual, mientras que en los almacenes de datos se permite almacenar históricos.
- En las bases de datos operacionales se puede acceder a la información detallada, mientras que los almacenes de datos se puede acceder a la detallada y la resumida.

El *Data Warehousing* es el proceso que facilita la creación y explotación de un Almacén de Datos.

Dentro de las funciones del *Data Warehousing* encontramos:

- Las bases de datos de toda clase se integran en una sola.
- La visualización de resultados en forma gráfica de los datos, en muchos casos totalizados y listos para la toma de decisiones.

Dentro de las características del almacén de datos encontramos que se encuentra que la información puede ser clasificada con respecto al interés de las empresas. Además la información permite que no se presente la volatilidad, es decir, se permite la manipulación de datos almacenados a modo e histórico.

Dentro de la arquitectura de un almacén de datos que la estructura básica incluye:

- Datos operacionales.
- Extracción de datos.
- Transformación de datos.
- Carga de datos.
- Almacén.
- Herramienta de acceso

Por último se menciona la estructura lógica de un almacén de datos, que está compuesta por:

- Metadatos
- Datos detallados actuales
- Datos detallados históricos.
- Datos ligeramente resumidos.
- Datos muy resumidos.

En su gran mayoría las empresas grandes prestan servicios de almacén de datos, esto debido a que pueden tener una arquitectura óptima para prestar este servicio. Estos servicios se prestan en lo que hoy se llama la nube.

3.5.2 Minería de datos o Data Mining

Es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

En la página oficial de Microsoft se recopilo la siguiente información: "*La minería de datos, utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante*

la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- ***Previsión:*** *calcular las ventas y predecir las cargas de servidor o el tiempo de inactividad del servidor.*
- ***Riesgo y probabilidad:*** *elegir los mejores clientes para la distribución de correo directo, determinar el punto de equilibrio probable para los escenarios de riesgo, y asignar probabilidades a diagnósticos u otros resultados.*
- ***Recomendaciones:*** *determinar los productos que se pueden vender juntos y generar recomendaciones.*
- ***Buscar secuencias:*** *analizar los artículos que los clientes han introducido en el carrito de compra y predecir los posibles eventos.*
- ***Agrupación:*** *separar los clientes o los eventos en clústeres de elementos relacionados, y analizar y predecir afinidades.*

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo." [13] Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. **Definir el problema:** Este paso es el más importante ya que el éxito del modelo y los demás pasos dependen de este punto. Se deben analizar los requisitos empresariales, definir el ámbito, definir las métricas por las que se evaluara el modelo y definir los objetivos concretos del modelo. Como recomendación Microsoft dice que debemos hacernos las siguientes preguntas:

   - *"¿Qué está buscando? ¿Qué tipos de relaciones intenta buscar?"*
   - *"¿Refleja el problema que está intentando resolver las directivas o procesos de la empresa?"*
   - *"¿Desea realizar predicciones a partir del modelo de minería de datos o solamente buscar asociaciones y patrones interesantes?"*
   - *"¿Qué resultado o atributo desea predecir?"*
   - *"¿Qué tipo de datos tiene y qué tipo de información hay en cada columna? En caso de que haya varias tablas, ¿cómo se relacionan? ¿Necesita limpiar, agregar o procesar los datos antes de poder usarlos?"*
   - *"¿Cómo se distribuyen los datos? ¿Los datos son estacionales? ¿Los datos representan con precisión los procesos de la empresa?"* [13]

2. **Preparar los datos:** En este paso se pretende limpiar los datos, se deben tratar de tal forma que no se presenten incoherencias, quitando los datos inválidos, además se deben consolidar en caso de que tengan diferentes orígenes, también se debe determinar los orígenes de datos, y determinar las columnas que son más adecuadas para el análisis.
3. **Explorar los datos:** En este paso es recomendable obtener los valores máximos y mínimos, calcular la media y las desviaciones estándar, y examinar la distribución de los datos. Esto con el fin de determinar si los datos son lo suficientemente confiables, este importante es de una importancia alta ya que

de esto dependerá la veracidad de la información. Por ejemplo una desviación estándar grande puede indicar que de agregar más datos se mejoraría el modelo.

4. **Generar modelos:** En este paso se debe usar los conocimientos adquiridos en el paso anterior y generar el modelo o modelos de minería de datos. Se debe definir las columnas de datos que se van a usar.
5. **Explorar y validar los modelos:** En este paso se busca comprobar la eficiencia del modelo. Este paso se le conoce como las pruebas del modelo, antes de instalar en un ambiente de producción es de suma importancia realizar las pruebas del modelo.
6. **Implementar y actualizar los modelos:** En este paso el objetivo es instalar el modelo en un ambiente de producción. Una vez se cumplan los pasos anteriores entonces es el momento de instalar.

En la figura N°5 se puede apreciar el diagrama con el fin de describir las relaciones existentes entre los pasos anteriores y SQL Server. Este presenta los pasos para realizar un proceso de minería de datos.

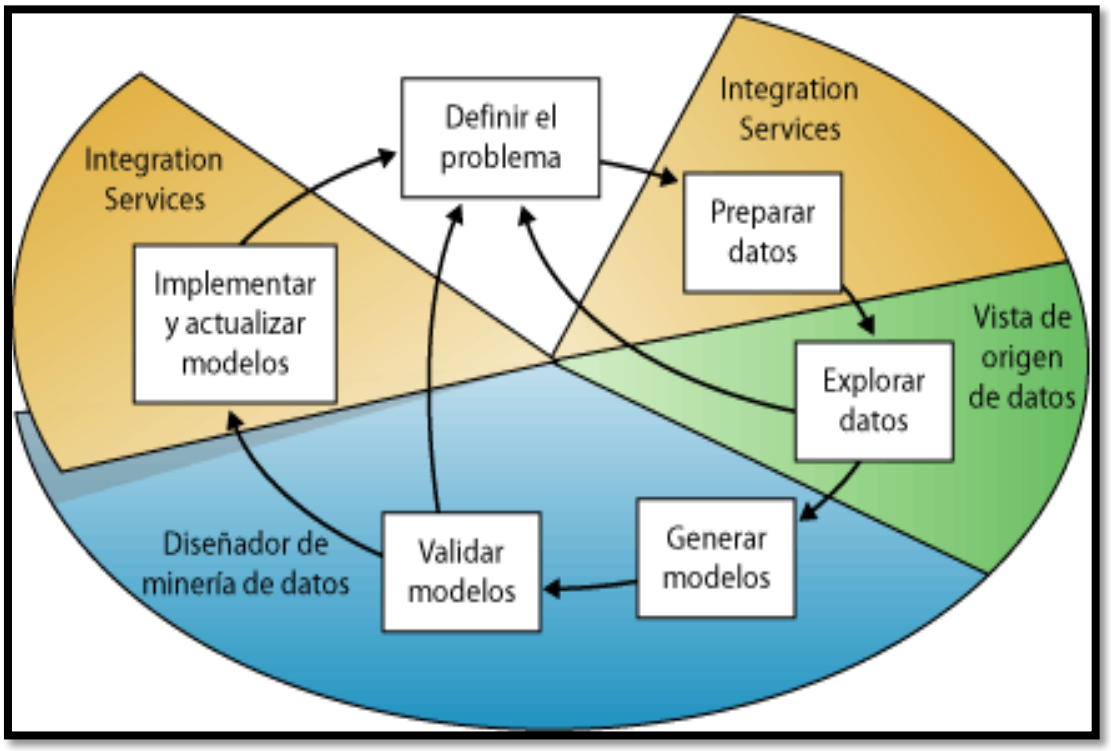


Figura 5 Relaciones existentes entre cada paso del proceso y SQL Server. Fuente: [13].

3.5.3 Algoritmos de minería de datos

La herramienta Analysis Services de Microsoft incluye los siguientes tipos de algoritmos:

- **Algoritmos de clasificación:** *"Son los que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos."* [14]
- **Algoritmos de regresión:** *"que predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos."* [14]

- **Algoritmos de segmentación:** “que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.” [14]
- **Algoritmos de asociación:** “que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra.” [14]
- **Algoritmos de análisis de secuencias:** “que resumen secuencias o episodios frecuentes en los datos, como un flujo de rutas web.” [14]

Elegir un algoritmo por tarea:

La siguiente tabla se presenta con el fin de escoger un algoritmo según la necesidad que se presente:

Ejemplos de tareas	Algoritmos de Microsoft que se pueden usar
Predecir un atributo discreto • Marcar los clientes de una lista de posibles compradores como clientes con buenas o malas perspectivas. • Calcular la probabilidad de que un servidor genere un error en los próximos 6 meses. • Clasificar la evolución de los pacientes y explorar los factores relacionados.	• Algoritmo de árboles de decisión de Microsoft • Algoritmo Bayes naive de Microsoft • Algoritmo de clústeres de Microsoft • Algoritmo de red neuronal de Microsoft
Predecir un atributo continuo • Pronosticar las ventas del año próximo. • Predecir los visitantes del sitio a partir de tendencias históricas y estacionales proporcionadas. • Generar una puntuación de riesgo a partir de datos demográficos.	• Algoritmo de árboles de decisión de Microsoft • Algoritmo de serie temporal de Microsoft • Algoritmo de regresión lineal de Microsoft
Predecir una secuencia • Realizar un análisis clickstream del sitio web de una empresa. • Analizar los factores que dan como resultado errores en el servidor. • Capturar y analizar secuencias de actividades durante las visitas de pacientes externos, para formular las prácticas recomendadas en las actividades comunes.	• Algoritmo de clústeres de secuencia de Microsoft

Buscar grupos de elementos comunes en las transacciones • Usar el análisis de la cesta de la compra para determinar la posición del producto. • Sugerir a un cliente la compra de productos adicionales. • Analizar los datos de una encuesta a los visitantes a un evento, para descubrir qué actividades o stands estaban correlacionados con el fin de programar actividades futuras.	• Algoritmo de asociación de Microsoft • Algoritmo de árboles de decisión de Microsoft
Buscar grupos de elementos similares • Crear grupos de pacientes con perfiles de riesgo en función de atributos como datos demográficos y comportamientos. • Analizar usuarios mediante patrones de búsqueda y compra de productos. • Identificar servidores con características de uso similares.	• Algoritmo de clústeres de Microsoft • Algoritmo de clústeres de secuencia de Microsoft

Tabla 3 Elegir algoritmo por tarea. Fuente: [14].

3.6 Computación en la nube o Cloud Computing

Es la práctica de utilizar en una red servidores remotos alojados en internet para almacenar, gestionar y procesar los datos, en lugar de un servidor local o un ordenador personal.

La nube permite que no se necesite una gran infraestructura para lograr tener buenos resultados, esto quiere decir que una pequeña empresa puede montar sus servicios de tecnología en la nube. Cuando alguien decide usar la nube entonces decimos que se está subcontratando, de esta manera se le dan tareas complejas a quien ya tiene experiencia en hacerlo por ejemplo Azure. Esto generalmente se está vendiendo como un servicio por grandes empresas como Microsoft, Amazon, IBM, etc.

Hoy en día podemos ver que la nube se está convirtiendo en algo cotidiano, por ejemplo el caso de Google Drive u One Drive, que permiten tener siempre documentos disponibles, dese cualquier dispositivo.

La nube representa el gran avance de la computación y se hace llamar computación distribuida. En síntesis se ponen computadores de gran capacidad, interconectados a través de redes con otros computadores, dando como resultado un súper computador, que lograra que las tareas de gran capacidad de cómputo hoy en día sean una realidad.

Contextualizando Big Data y computación en la nube, entonces entendemos que para hablar de Big Data debemos hablar de computación distribuida y por supuesto de computación en la nube. Si se quiere una solución rápida sin tener contratar expertos en computación distribuida ni esperar largo tiempo para ver resultados, al primer lugar donde se debe dirigir a la nube, como lo mencionamos anteriormente, las más grandes empresas de tecnología venden servicios de Big Data.

3.7 Inteligencia de negocio o Business intelligence

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

El proceso de Business Intelligence incluye la comunicación de los descubrimientos y efectuar los cambios.

Las áreas incluyen clientes, proveedores, productos, servicios y competidores." (GARTNER, 2012)

La inteligencia de negocios es la unión de las tecnologías descritas en los numerales anteriores, se podría decir que esto hace parte de la definición del concepto Big Data. BI convierte los datos en información y la información en decisiones, esto es lo que al final le interesa a los gerentes o las personas encargadas de tomar decisiones.

Cuando una empresa decide aplicar la inteligencia de negocios, y esta se aplica de manera correcto, se ejecutan en gran parte los conceptos de Big Data.

3.8 Analysis de Big Data o Big Data Analytics

3.8.1 Proyecto R

Se trata de un proyecto de software libre, resultado de la implementación GNU del lenguaje S.

Es un lenguaje y entorno de programación para análisis estadístico y gráfico, el cual brinda una excelente estructura para relacionar con almacenes de Big Data para plasmar gráficamente los análisis realizados de un conjunto de datos.

"R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos.

Entre otras características dispone de:

- *Almacenamiento y manipulación efectiva de datos,*
- *operadores para cálculo sobre variables indexadas (Arrays), en particular matrices,*
- *una amplia, coherente e integrada colección de herramientas para análisis de datos,*
- *posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y*
- *un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas*

y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

El término "entorno" lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy específicas e inflexibles, como ocurre frecuentemente con otros programas de análisis de datos.

R es en gran parte un vehículo para el desarrollo de nuevos métodos de análisis interactivo de datos. Como tal es muy dinámico y las diferentes versiones no siempre son totalmente compatibles con las anteriores. Algunos usuarios prefieren los cambios debido a los nuevos métodos y tecnología que los acompañan, a otros sin embargo les molesta ya que algún código anterior deja de funcionar. Aunque R puede entenderse como un lenguaje de programación, los programas escritos en R deben considerarse esencialmente efímeros." [15]

Figura 6 Logo R. Fuente: [16].

En la figura N°6, se puede apreciar el logo con el cual se representa R en el mercado.

3.8.2 Métodos

3.8.2.1 Redes neuronales

3.8.2.1.1 Historia de las redes neuronales

"Alan Turing, en 1936, fue el primero en estudiar el cerebro como una forma de ver el mundo de la computación; sin embargo, los primeros teóricos que concibieron los fundamentos de la computación neuronal fueron Warren McCulloch, un neurofisiólogo,y Walter Pitts, un matemático, quienes, en 1943, lanzaron una teoría acerca de la forma de trabajar de las neuronas. Ellos modelaron una red neuronal simple mediante circuitos eléctricos. Otro importante libro en el inicio de las teorías de las redes neuronales fue escrito en 1949 por Donald Hebb, La organización del comportamiento, en el que establece una conexión entre psicología y fisiología.

En 1957, Frank Rosenblatt comenzó el desarrollo del Perceptrón, es el modelo más antiguo de red neuronal, y se usa hoy en día de varias formas para la aplicación de reconocedor de patrones. Este modelo era capaz de generalizar; es decir, después de haber aprendido una serie de patrones era capaz de reconocer otros similares, aunque no se le hubieran presentado anteriormente. Sin embargo, tenía una serie de limitaciones, quizás la más conocida era la incapacidad para resolver el problema de la función OR-

exclusiva y, en general, no era capaz de clasificar clases no separables linealmente.

En 1959, Bernard Widrow y Marcial Hoff, de Stanford, desarrollaron el modelo ADALINE (ADAptive LINear Elements). Esta fue la primera red neuronal aplicada a un problema real (filtros adaptativos para eliminar ecos en las líneas telefónicas) y se ha usado comercialmente durante varias décadas.

Uno de los mayores investigadores de las redes neuronales desde los años 60 hasta nuestros días es Stephen Grossberg (Universidad de Boston). A partir de su extenso conocimiento fisiológico, ha escrito numerosos libros y desarrollado modelos de redes neuronales. Estudió los mecanismos de la percepción y la memoria. Grossberg realizó en 1967 una red, Avalancha, que consistía en elementos discretos con actividad que varía con el tiempo que satisface ecuaciones diferenciales continuas, para resolver actividades tales como reconocimiento continuo del habla y aprendizaje del movimiento de los brazos de un robot.

En 1969 surgieron numerosas críticas que frenaron, hasta 1982, el crecimiento que estaban experimentando las investigaciones sobre redes neuronales. Marvin Minsky y Seymour Papert, del MIT, publicaron un libro, Perceptrons, que además de contener un análisis matemático detallado del Perceptrón, consideraba que la extensión a Perceptrones multinivel (el Perceptrón original solo poseía una capa) era completamente estéril. Las limitaciones del Perceptrón eran importantes, sobre todo su incapacidad para resolver muchos problemas interesantes. Esta fue una de las razones por la cual la investigación en redes neuronales quedó rezagada por más de 10 años.

A pesar del libro Perceptrons, algunos investigadores continuaron con su trabajo. Tal fue el caso de James Anderson, que desarrolló un modelo lineal llamado Asociador Lineal, que consistía en 4 elementos integradores lineales (neuronas) que sumaban sus entradas. También desarrolló una potente extensión del Asociador Lineal llamada Brain-State-in-a-Box (BSB) en 1977.

En Europa y Japón, las investigaciones también continuaron. Kunihiko Fukushima desarrolló en 1980 el Neocognitrón, un modelo de red neuronal para el reconocimiento de patrones visuales. Teuvo Kohonen, un ingeniero electrónico de la universidad de Helsinki, desarrolló un modelo similar al de Anderson pero independientemente.

En 1982 comenzó a resurgir el interés por las redes neuronales gracias a dos trabajos importantes. John Hopfield presentó su trabajo (basado en la física estadística), en el cual describe con claridad y rigor matemático una red a la que ha dado su nombre, que es una variación del Asociador Lineal, pero, además, mostró cómo tales redes pueden trabajar y qué pueden hacer. El otro trabajo pertenece a Teuvo Kohonen, con un modelo con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro, este modelo tenía dos variantes denominadas LVQ (Lerning Vector Cuantization) y SOM (Self-Organizing Map).

En 1986, el desarrollo del algoritmo back-propagation fue dado a conocer por Rumelhart, Hinton y Williams. Ese mismo año, el libro Parallel Distributed Processing, fue publicado por Rumelhart y McClelland, siendo este libro la

mayor influencia para la aplicación generalizada del algoritmo back propagation." [17]

3.8.2.1.1 Modelo biológico de las redes neuronales

Ahora se va a mostrar el modelo biológico de las neuronas:

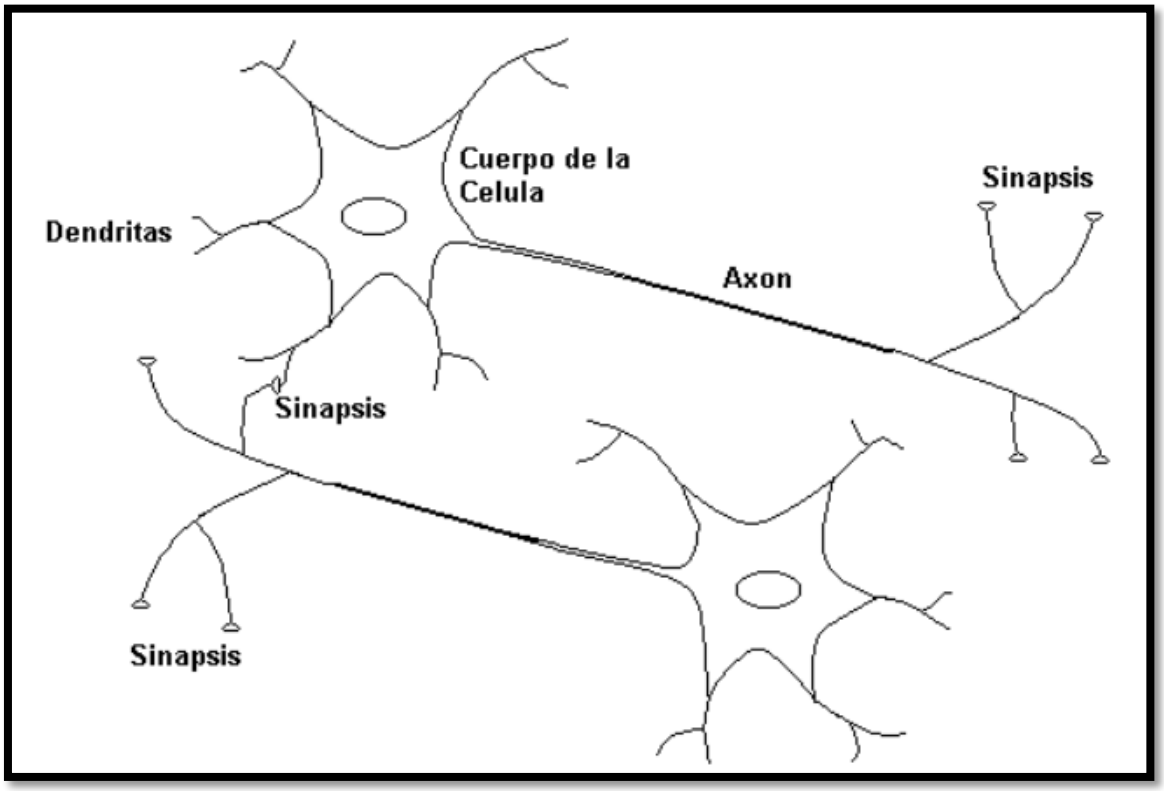


Figura 7 Modelo biológico neuronas. Fuente [17].

En la figura N°7 se puede apreciar que, *"el cerebro consta de un gran número de elementos altamente interconectados (aproximadamente 104 conexiones por elemento), llamados neuronas. Estas neuronas tienen tres componentes principales, las dendritas, el cuerpo de la célula o soma, y el axón. Las dendritas, son el árbol receptor de la red, son como fibras nerviosas que cargan de señales eléctricas el cuerpo de la célula. El cuerpo de la célula, realiza la suma de esas señales de entrada. El axón es una fibra larga que lleva la señal desde el cuerpo de la célula hacia otras neuronas. El punto de contacto entre un axón de una célula y una dendrita de otra célula es llamado sinápsis, la longitud de la sinápsis es determinada por la complejidad del proceso químico que estabiliza la función de la red neuronal.*

Todas las neuronas conducen la información de forma similar, esta viaja a lo largo de axones en breves impulsos eléctricos, denominados potenciales de acción; los potenciales de acción que alcanzan una amplitud máxima de unos 100 mV y duran un par de ms, son resultado del desplazamiento a través de la membrana celular de iones de sodio dotados de carga positiva, que pasan desde el fluido extracelular hasta el citoplasma intracelular; seguidos de un desplazamiento de iones de 5 potasio (carga negativa) que se desplazan desde el fluido intracelular al extracelular." [17]

3.8.2.1.1 Modelo artificial de las redes neuronales

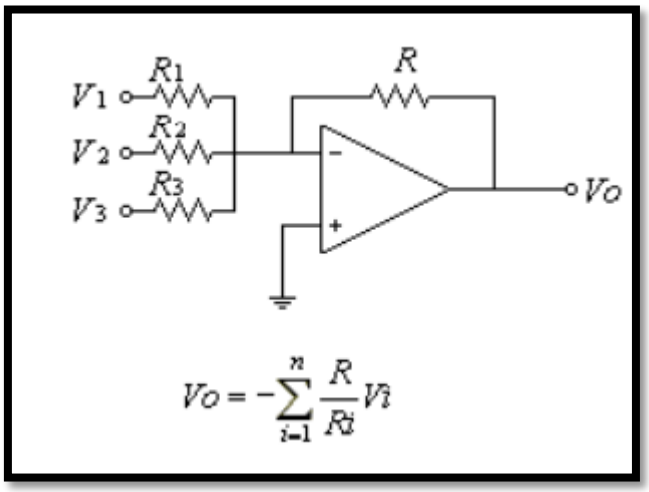


Figura 8 Modelo artificial de neurona I. Fuente: [17].

En la figura N°8 se observa que el modelo de una red neuronal artificial es un modelo matemático, en donde se tiene como objetivo sumar ciertas entradas para tomar una decisión y luego dar una salida. Esto nos lleva a pensar o asemejar este modelo a un sumador hecho con un amplificador operacional.

De esta forma se planteó la simulación de una neurona artificial frente a una neurona biológica.

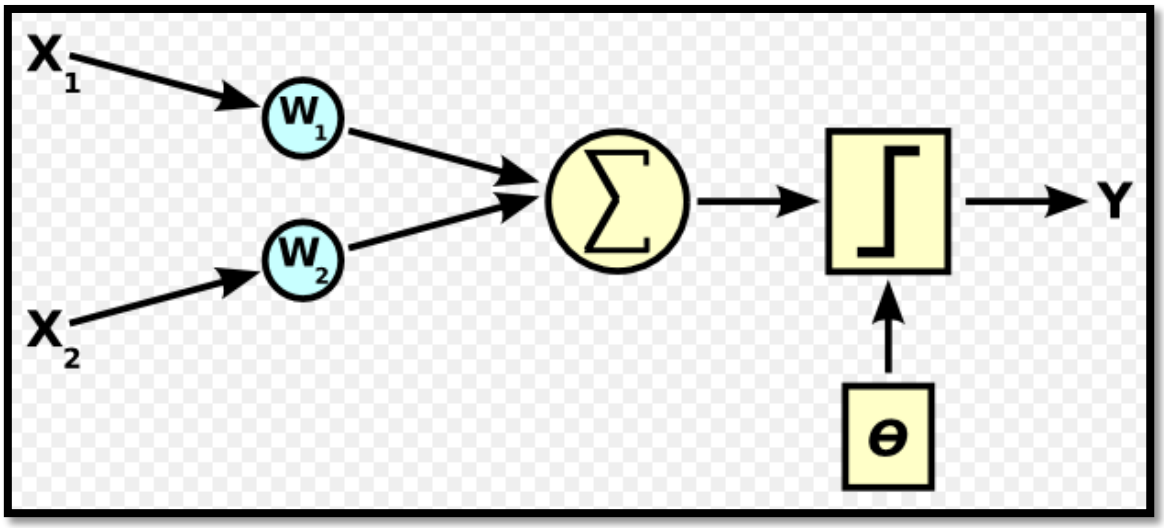


Figura 9 Modelo artificial de neurona II. Fuente: [18].

En la figura N°9 se puede apreciar que, existen dos entradas X1, X2 estas entradas tienen un peso llamada W1, W2, luego se realiza una sumatoria de esas entradas con respecto a su peso, para entregar una salida. Este es el modelo propuesto para el funcionamiento de una neurona artificial.

3.8.2.1.2 Similitudes de las redes neuronales biológicas y las artificiales:

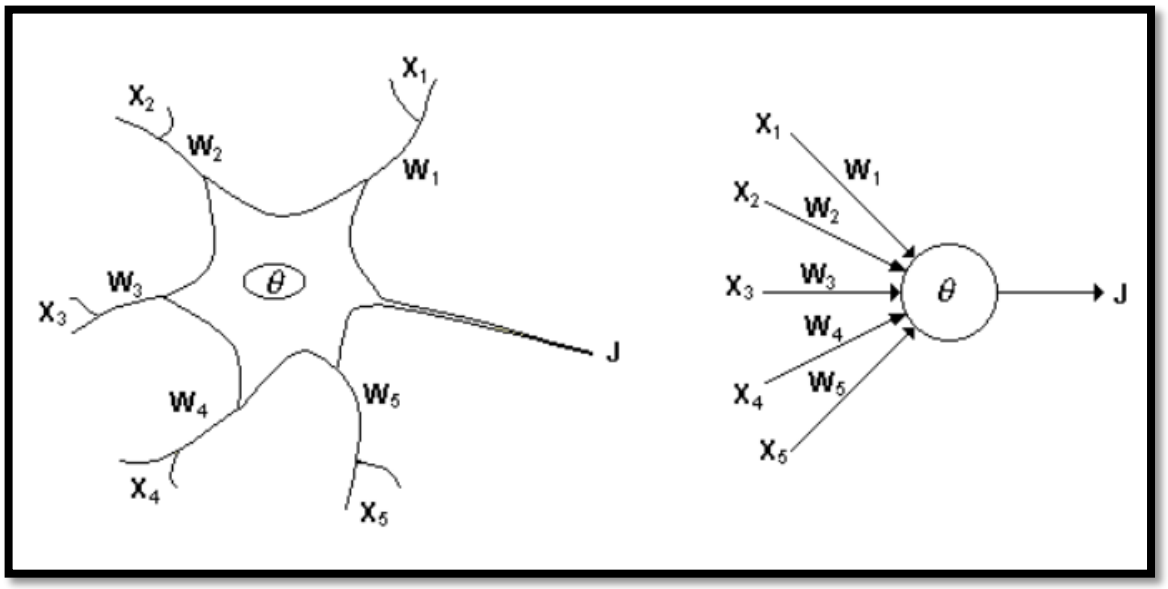


Figura 10 Similitudes entre red biológica y artificial. Fuente: [17].

Como se puede aprecia en la figura N°10:

- Las entradas Xn son las señales que son capturadas por las dentinas y generalmente provienen de otras neuronas.
- Wi representa la intensidad de las sinapsis, esto permite diferenciar las más fuertes de las débiles,
- *θ es la función umbral que la neurona debe sobrepasar para activarse; este proceso ocurre biológicamente en el cuerpo de la célula* [17]

3.8.2.1.3 Modelo red neuronal

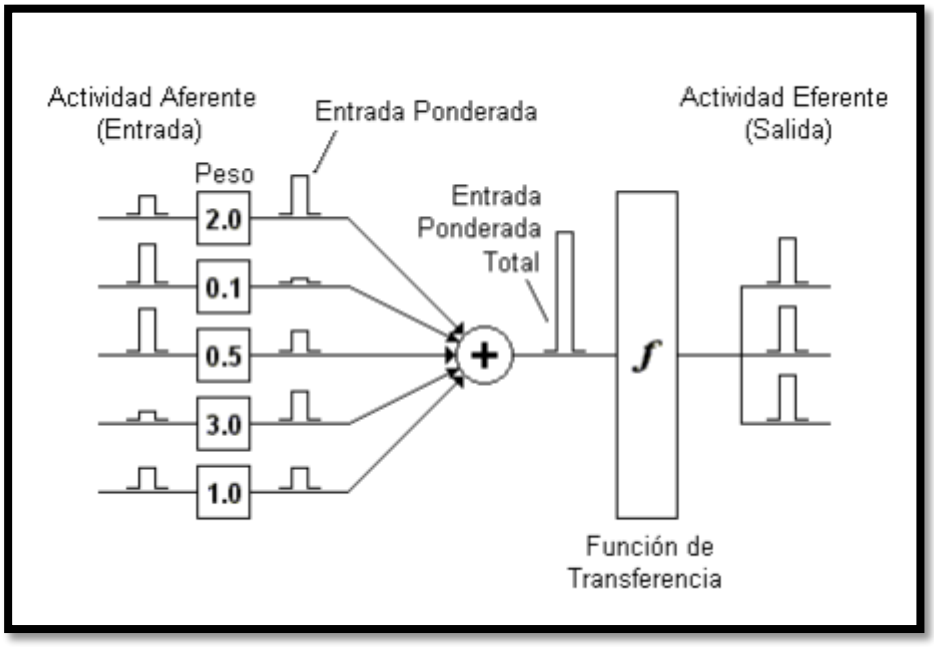


Figura 11 Recorrido de un conjunto de señales que entran a la red. Fuente: [17].

Se puede observar en la figura N°11, que el modelo presenta una entrada llamada *Actividad Aferente,* estas entradas pueden ser salidas de otras neuronas artificiales. Cada entrada tiene un peso, que afectara si se llega a sobrepasar el umbral, luego hace un proceso de ponderación y se genera una entrada ponderada, para hacer la sumatoria de las entradas ponderadas y generar una entrada ponderada Total. Se activa la verificación del umbral llamada función de transferencia, y por último se genera una salida o *Actividad Eferente.*

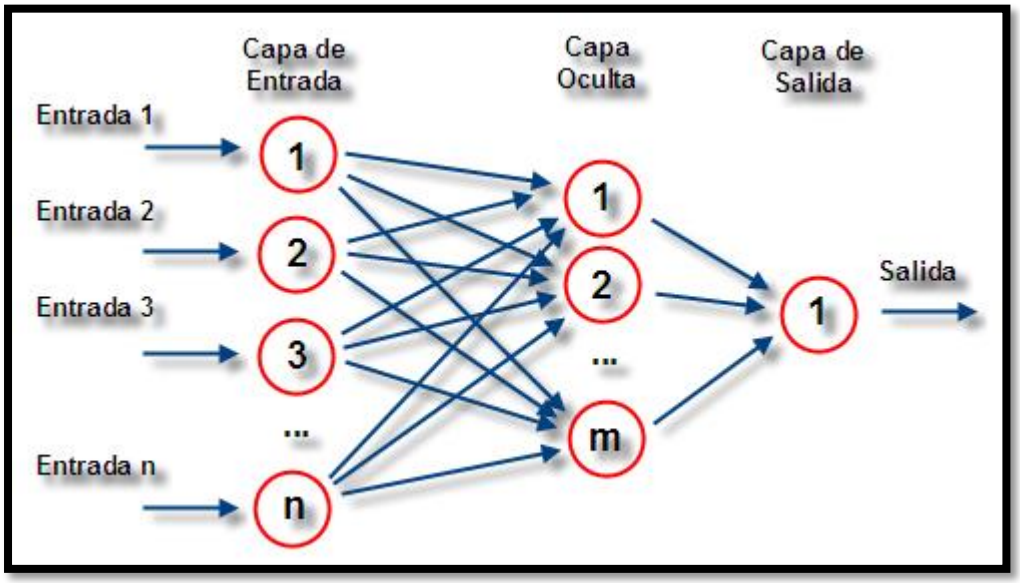


Figura 12 Red Neuronal Artificial Fuente: [19].

En la figura N°12 se puede aparecía una red neuronal de tipo Perceptrón simple con n neuronas de entrada, m neuronas en su capa oculta y una neurona de salida. [19]

3.8.2.1.1 Algoritmo Microsoft, redes neuronales

Microsoft para este fin ofrece la herramienta SQL Server Analysis Services, el algoritmo funciona de la siguiente forma; primero "*combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente, puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada."* [20]

El algoritmo crea una red formada por hasta tres niveles de neuronas. Que son:

- **Nivel de entrada:** *"las neuronas de entrada definen todos los valores de atributos de entrada para el modelo de minería de datos, así como sus probabilidades."* [20]
- **Nivel oculto:** *"las neuronas ocultas reciben entradas de las neuronas de entrada y proporcionan salidas a las neuronas de salida. El nivel oculto es donde se asignan pesos a las distintas probabilidades de las entradas. Un peso describe la relevancia o importancia de una entrada determinada para*

la neurona oculta. Cuanto mayor sea el peso asignado a una entrada, más importante será el valor de dicha entrada. Los pesos pueden ser negativos, lo que significa que la entrada puede desactivar, en lugar de activar, un resultado concreto." [20]

- **Nivel de salida:** *"las neuronas de salida representan valores de atributo de predicción para el modelo de minería de datos."* [20]

Una vez este procesado el modelo, entonces se puede usar la red y los pesos dentro de cada nodo para realizar predicciones.

Un ejemplo de lo anterior es:

"Suponga que conoce estos hechos sobre una clase de clientes potenciales:

- *Mediana edad (40 a 50 años).*
- *Tiene casa en propiedad.*
- *Tiene dos hijos que todavía viven en casa.*

¿Cómo puede correlacionar estos atributos con la probabilidad de que el cliente haga una compra?

Mediante la construcción de un modelo de red neuronal que use los hábitos de compra como resultado de destino, puede explorar diversas combinaciones de atributos del cliente, tales como ingresos altos, y descubrir qué combinación de atributos es más probable que influya en los hábitos de compra. Por ejemplo, puede que descubra que el factor determinante es la distancia entre su domicilio." [20]

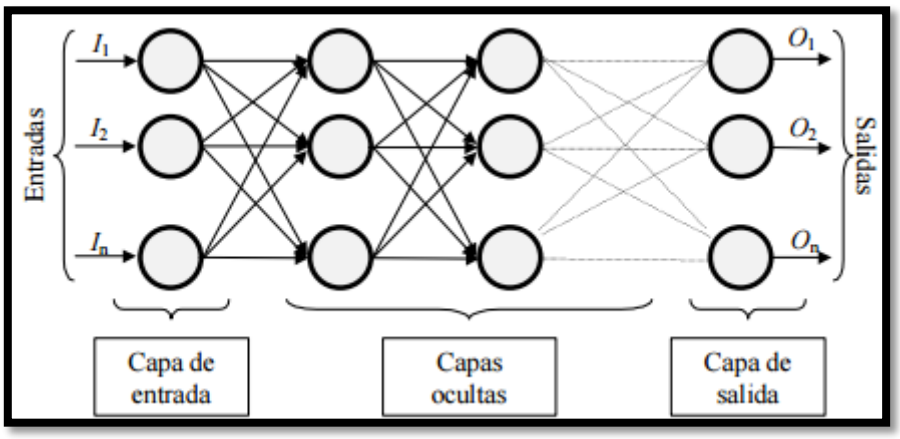


Figura 13 ejemplo de una red neuronal totalmente conectada [21].

En la figura N°13 se puede ver un esquema de red neuronal, con sus respectivas capas, es decir la capa de entrada, luego la capa oculta y por último la capa de salida, todas estas capas están interconectadas.

3.8.2.2 Reconocimiento de imágenes

Actualmente se adelantan trabajos de investigación en reconocimiento facial con la aplicación de diferentes técnicas de reconocimiento, como los métodos basados en imágenes 3D, basados en video, combinación de técnicas, o técnicas basadas en imágenes fijas.

"El reconocimiento facial, hoy por hoy, es uno de los campos de investigación más amplios que existen, con diferentes ramas como la biometría, reconocimiento de patrónes, reconocimiento por medio del mapa de las venas del rostro con luz infrarroja. etc. Con aplicación en áreas como la medicina, telefónica celular, vigilancia y control de acceso, procesos de investigación y criminalística, software comercial, entre otros." [22]

*"**Elementos para el procesamiento de las imágenes:***

Para trabajos en reconocimiento de rostros, es necesario proporcionar un procesamiento de las imágenes ya que estas deben llegar de forma adecuada a la solución. Ejemplo: el ángulo de enfoque, la iluminación, profundidad etc. Por lo cual es necesario retirar aquella información de la imagen que no nos interesa y entregarla en forma precisa.

Si bien existen muchas herramientas y lenguajes de programación que facilitan el procesamiento de las imágenes, todos estos contienen las implementaciones matemáticas correspondientes para realizar su trabajo. Aquí se pretende exponer con un enfoque técnico revisando cómo sucede el procesamiento de imágenes, claro está de forma general.

A continuación, algunas áreas en las que existen no solo trabajos finalizados sino una amplia gama de investigaciones en procesamiento de imágenes:

- *Procesamiento de vídeo.*
- *Creación de herramientas para la post-producción de cine digital*
- *Análisis de imágenes médicas.*
- *Fotografía digital.*
- *Visión estéreo.*
- *Reconstrucción tridimensional a partir de secuencias de vídeo.*
- *Restauración e interpretación de las imágenes tomadas por satélites.*
- *Reconocimiento de formas.*
- *Búsqueda de imágenes en la web.*
- *Compresión de imágenes.*
- *Procesamiento de superficies.*
- *Síntesis de imágenes.*
- *Simulación para videojuegos."* [22]

3.8.2.2.1 Técnicas basadas en imágenes fijas

Linear Discriminant Analysis (LDA): [22]

Es un método supervisado, que utiliza la información obtenida de un conjunto de imágenes de la misma persona, es decir, imágenes de la misma clase, desarrollando un conjunto de vectores y permitiendo maximizar la varianza entre clases diferentes, garantizando la discriminación entre las mismas.

LDA convierte un problema de alta dimensionalidad en uno de baja dimensionalidad, por medio de una matriz de proyección, cuyas columnas son llamadas FisherFaces.

Una de las limitaciones de esta técnica es que se genera singularidad o singularity Prob-lem, entre el número de imágenes y la dimensión de las mismas, lo cual se puede controlar la pseudo-inversa de la matriz de covarianza, con el fin de realizar previamente la reducción necesaria de los datos.

3.8.2.2.2 Técnicas basadas en video

Las técnicas basadas en video proponen la asociación de cada uno de los recuadros del video, e implica que el video analizado sea de baja resolución, esto, dado su alto coste computacional, esta técnica actualmente es objeto de investigación de Intel.

"Los resultados de las investigaciones y trabajos propuestos como el de Gorodnichy, donde propone una combinación entre la forma de pensar del ser humano y la técnica implementada, tal que el sistema reconozca un rostro que tenga mínimo 12 pixeles de separación entre los ojos del individuo, no evolucionaron ya que estudios posteriores a este, demuestran el alto coste computacional versus los resultados obtenidos, demostrando las ventajas de las técnicas basadas en imágenes fijas." [22]

3.9 Datos estructurados

3.9.1 Bases de datos relacionales

Es un modelo de base de datos el cual consta de tablas las cuales almacenan datos y tiene conexiones para relacionar los datos de las tablas.

Actualmente es uno de los modelos que más se utiliza en el mercado.

3.9.2 SQL Server

Es un sistema de base de datos relación RDBMS. Su propietario Microsoft se ha encargado de que muchas empresas usen este sistema, es decir se ha convertido en un sistema empresarial. Este sistema maneja su propio lenguaje de desarrollo el cual se llama TSQL (Transact-SQL).

El primer nombre que tuvo SQL Server en su etapa de desarrollo fue Yukon, y en el año 2005 fue lanzado.

SQL Server ha sido durante muchos años el sistema de base de datos preferido por las empresas, sin embargo hoy en día muchas tienen necesidades que no se cubren. Por este motivo Microsoft ha decidido ofrecer a las empresas soluciones en la nube. Además en los últimos años se han realizado alianzas con organizaciones como apache.org, dando como resultado que hoy en día el software libre y el pago vallan de la mano trabajando por un mismo objetivo.

Podemos encontrar integraciones con Hadoop y con sistemas como R, que permiten realizar análisis estadístico y cosas como aprendizaje de máquina.

En la siguiente imagen se puede apreciar la arquitectura que usa SQL Server en una implementación clásica:

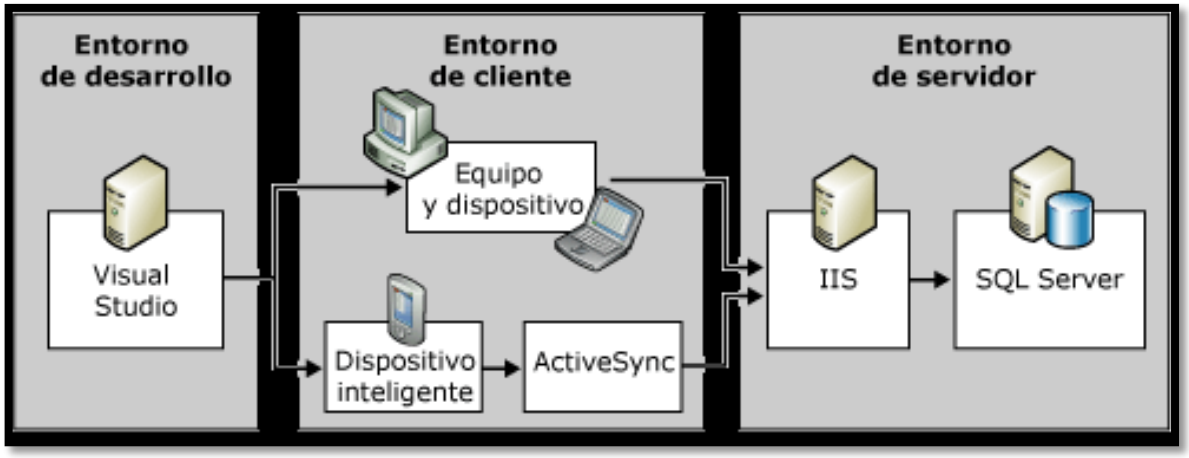


Figura 14 Arquitectura SQL Server en sistema clásico. Fuente: [23].

En la figura N°14 se puede apreciar la arquitectura de una empresa promedio en el mercado de hoy en día. De esta forma funcionan muchas organizaciones.

Para acceder a la base de datos SQL Server, comúnmente se usa el servidor IIS (Internet Information Service), que permite la comunicación entre los dispositivos las aplicaciones, los servicios web con la base de datos.

Figura 15 Arquitectura Base de Datos Centralizada. Fuente: [24].

En la figura N°15 se muestra como se centralizan los datos en un DataCenter, esta es la arquitectura con la que funcionan muchas organizaciones.

3.10 Bases de datos NoSQL

Son bases de datos que difieren del modelo clásico de bases de datos relacionales, por lo cual los datos que almacenan no necesitan estructuras fijas como son las tablas. El NoSQL hace referencia a “Not Only SQL” para aclarar que no solo funciona con consultas del tipo SQL.

Existen varios tipos de clasificaciones según la forma en la que estos motores almacenan sus datos como son: “BigTable”, “Bases de datos Documentales”, “Bases de datos orientadas a grafos”.

En los anexos del 1 al 7 se puede ver una comparación entre el lenguaje SQL que se trabaja actualmente y el lenguaje NoSQL.

3.10.1 MongoDB

Es un sistema de base de datos multiplataforma (Linux, Windows OS X, Solaris) orientado a documentos de código abierto.

Este motor de base de datos en vez de guardar los datos en tablas, guarda estructuras de datos en documentos tipos JSON con un esquema dinámico.

En los anexos del 1 al 7 se puede ver el lenguaje NoSQL que se propone para el manejo de MongoDB.

MongoDB puede ser integrado con ***Hadoop*** de las siguientes formas:

1. **Batch Aggregation:**

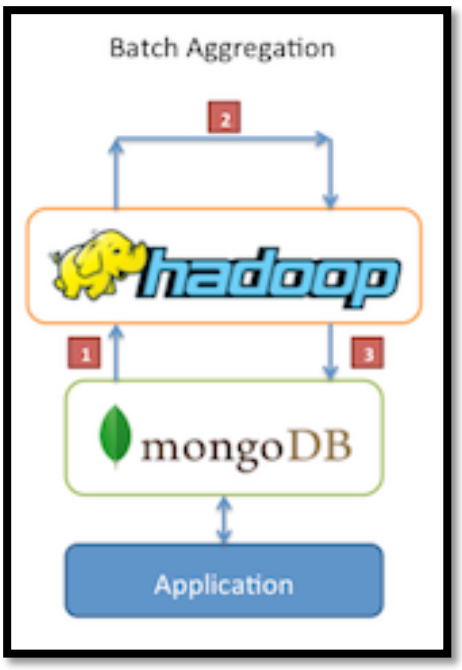


Figura 16 MongoDB Batch Aggregation. Fuente: [25].

Como se puede apreciar en la figura N°16, *“en este escenario los datos se extraen de MongoDB y procesados dentro de Hadoop a través de uno o más puestos de trabajo MapReduce. Los datos también pueden ser traídos de fuentes adicionales dentro de estos puestos de trabajo MapReduce para desarrollar una solución multi-fuente de datos. La salida de estos puestos de trabajo MapReduce entonces se puede escribir de nuevo a MongoDB para su posterior consulta y análisis ad-hoc.”* [25]

2. Data Warehouse:

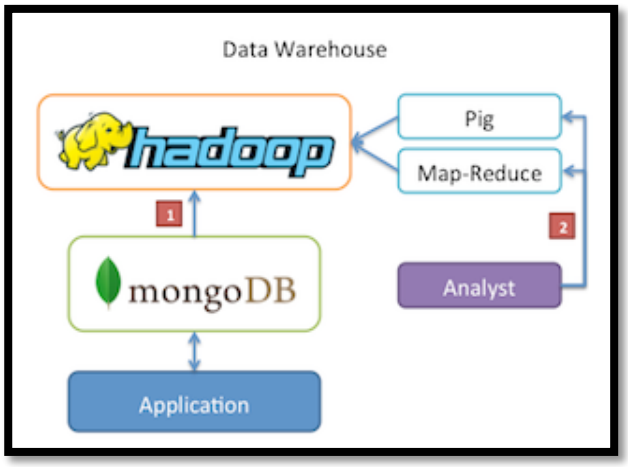


Figura 17 MongoDB DataWarehouse. Fuente: [25].

Como se puede apreciar en la figura N°17, "*en un escenario típico de producción, los datos de su aplicación pueden vivir en varios almacenes de datos, cada uno con su propio lenguaje de consulta y funcionalidad. Para reducir la complejidad en estos escenarios, Hadoop puede ser utilizado como un almacén de datos y actuar como un repositorio centralizado para los datos de las diversas fuentes.*

En esta situación, usted podría tener Jobs periódicos de MapReduce que permiten la carga de datos a MongoDB en Hadoop. Esto podría ser en forma de "diario" o cargas de datos "semanales" sacados de MongoDB través MapReduce. Una vez que los datos de MongoDB estan disponibles en Hadoop, y datos de otras fuentes también están disponibles, los datos del conjunto de datos más grandes se pueden consultar en contra. Los analistas de datos ahora tienen la opción de utilizar MapReduce o pig para crear Jobs que consultan las bases de datos más grandes que incorporan datos en MongoDB." [25].

3. ETL Data:

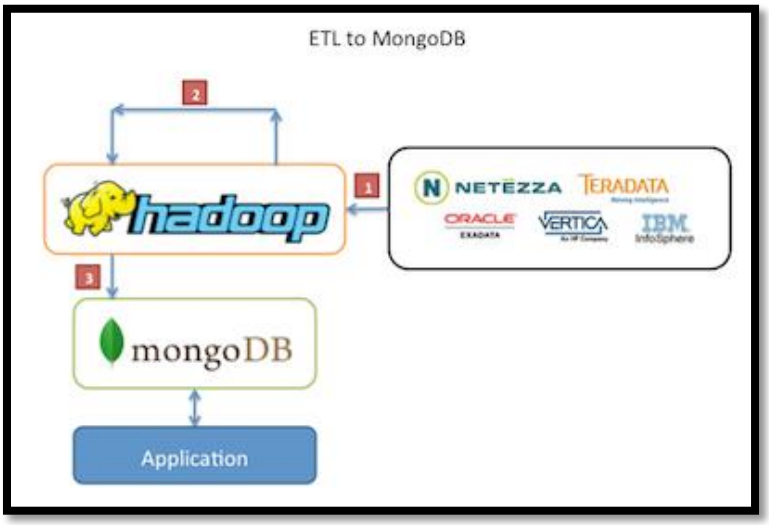


Figura 18 MongoDB ETL Data. Fuente: [25].

Como se aprecia en la figura 18*, "MongoDB puede ser el almacén de datos operativa para su aplicación, pero también puede haber otros almacenes de datos que sujetan los datos de su organización. En este escenario, es útil ser capaz de mover datos de un almacén de datos a otro, ya sea a partir de datos de su aplicación a otra base de datos o viceversa. Mover los datos es mucho más complejo que una simple tubería desde un mecanismo a otro, que es donde se puede utilizar Hadoop.*

En este escenario, se utilizan Jobs de Map-Reduce para extraer, transformar y cargar datos de un lugar de almacenamiento a otro. Hadoop puede actuar como un mecanismo de ETL complejo para migrar datos en diversas formas a través de uno o más Jobs MapReduce que jalan de los datos de un lugar donde se almacenan los datos, aplicando varias transformaciones (la aplicación de nuevos diseños de datos u otra agregación) y cargando los datos a otra lugar de almacenamiento. Este enfoque se puede utilizar para mover datos desde o hacia MongoDB, dependiendo del resultado deseado." [25]

Como se puede apreciar mongoDB permite la integración con Hadoop y este a su vez con otras arquitecturas, es decir se pueden hacer procesos de ***Batch Aggregation***, también se puede hacer modelos de ***Data Warehouse***, y adicionalmente se observa la arquitectura de una ***ETL*** con mongoDB, Hadoop y otro lugar de almacenamiento.

3.10.2 Cloudant

En el año 2014 IBM anunció un acuerdo definitivo para la adquisición de la empresa ubicada en Boston, Massachusetts, EEUU., **Cloudant, Inc**., proveedor privado de base de datos -as- a-service (DBaaS) que permite a los desarrolladores crear fácil y rápidamente, la siguiente generación de aplicaciones web y móviles.

"IBM está liderando la responsabilidad para ayudar a sus clientes a aprovechar los grandes volúmenes de datos, el cómputo de nube y las tecnologías móviles. *Cloudant se sitúa de lleno en el nexo de estas tres áreas clave de transformación y permite a los clientes entregar aplicaciones rápidamente con un nivel totalmente nuevo de innovación y ricas en datos al mercado.*" [26] Dijo Sean Poulley, vicepresidente de Bases de Datos y Almacenamiento de datos de IBM.

Cloudant complementa el portafolio de Big Data y Analytics de IBM más allá de la gestión tradicional de datos, proporcionando base de datos-como-servicio que permite a los clientes simplificar y acelerar el desarrollo de la participación y aplicaciones móviles y web escalables. Cloudant también es parte integral de las soluciones MobileFirst de IBM. Permite a los desarrolladores que utilizan Worklight, software de IBM para desarrollo de aplicaciones móviles, para crear rápidamente aplicaciones flexibles, confiables y escalables, que incluyen una variedad de datos estructurados y no estructurados.

Los servicios administrados de nube Cloudent logran:

- Almacenar los datos de cualquier estructura como documentos auto-descritos de JSON
- Aprovechar un sistema de replicación multi-master y principios de diseño distribuidos avanzados para lograr agrupaciones de bases de datos

elásticas que pueden abarcar varios bastidores, los centros de datos o proveedores de la nube.

- Permitir la distribución global de datos así como cargas geo-balanceadas a fin de proporcionar alta disponibilidad y un rendimiento mejorado para aplicaciones que requieren que la información se encuentran cerca de los usuarios.
- Proporcionar búsqueda de texto completo, consulta geo-espacial y temporal avanzada y flexible, así como la indexación en tiempo real.
- Integrarse a través de una interfaz de programación de aplicaciones REST (API).
- Permitir la replicación fácil de datos así como la sincronización para las aplicaciones móviles, con código abierto, librerías de software del dispositivo nativo.
- Ofrecer monitoreo 24x7 y la gestión realizada por sus expertos en Big Data.

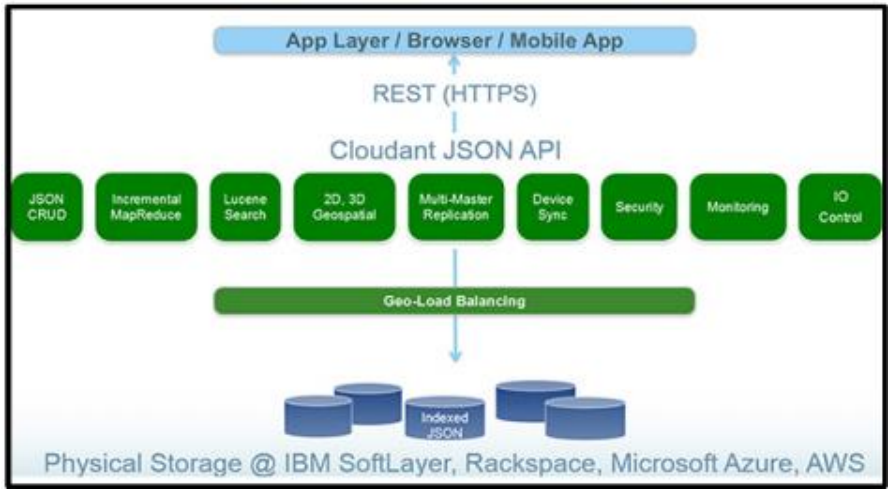


Figura 19 La base de datos de IBM Cloudant NoSQL, servicios y API de la capa vista. Fuente: [27].

En la figura 19 se ve puede observar la arquitectura de la base de datos Cloudant NoSQL. Los componentes del API JSON que se encuentra disponibles y la arquitectura REST usada en las aplicaciones o lo que se puede llamar el cliente.

3.10.3 Hadoop

3.10.3.1 Componentes principales de Hadoop.

Figura 20. Componentes principales de Hadoop. Fuente: Adapta de [28].

La figura N°20 representa los componentes básicos para poner en funcionamiento un ambiente Big Data, el conjunto de estos componentes reciben el nombre de Hadoop.

3.10.3.2 HDFS

Si usted desea más de 4000 computadoras trabajando en sus datos, entonces más vale que distribuya sus datos a lo largo de más de 4000 computadoras. HDFS hace esto para usted. HDFS tiene pocas partes movibles. Datanodes almacena sus datos, y Namenode da seguimiento al lugar donde se almacenan las cosas. Hay otras cuestiones, pero usted ya tiene lo suficiente para iniciar. [29]

3.10.3.3 MapReduce

Este es el modelo de programación para Hadoop. Existen dos fases, no es de sorprender que se llamen Map y Reduce. Para impresionar a sus amigos dígales que hay un tipo de mezcla entre la fase Map y la fase Reduce. JobTracker gestiona los más de 4000 componentes de su trabajo MapReduce. TaskTrackers toma órdenes de JobTracker. Si le gusta Java entonces codifíquelo en Java. Si a usted le gusta SQL u otro lenguaje que no sea Java tiene suerte, usted puede usar una utilidad llamada Hadoop Streaming. [29]

3.10.3.4 Hadoop Streaming

Una utilidad para permitir a MapReduce codificar en cualquier lenguaje: C, Perl, Python, C++, Bash, etc. Los ejemplos incluyen un correlacionador Python y un reductor AWK. [29]

3.10.3.5 Hive and Hue

Si a usted le gusta SQL, estará encantado de escuchar que usted puede escribir SQL y hacer que Hive lo convierta a un trabajo de MapReduce. No, usted no obtiene un entorno ANSI-SQL completo, pero usted obtiene 4000 notas y escalabilidad multi-Petabyte. Hue le brinda una intefaz gráfica basada en navegador para realizar su trabajo Hive. [29]

3.10.3.6 Pig

Un entorno de programación de nivel alto para realizar codificación MapReduce. El lenguaje Pig es llamado Pig Latin. A usted puede parecerle el nombre poco convencional, pero obtiene rentabilidad y alta disponibilidad increíbles. [29]

3.10.3.7 Sqoop

Proporciona transferencia de datos bidireccional entre Hadoop y su base de datos relacional favorita. [29]

3.10.3.8 Oozie

Gestiona flujo de trabajo Hadoop. Esto no reemplaza a su planificador o herramienta BPM, pero proporciona ramificación de "if-then-else" y control dentro de sus trabajos Hadoop. [29]

3.10.3.9 HBase

Un almacenamiento de valor de clave súper escalable. Funciona similarmente a un hash-map persistente (para los aficionados de python piensen en diccionario). No es una base de datos relacional pese al nombre HBase. [29]

3.10.3.10 FlumeNG

Un cargador en tiempo real para transmitir sus datos hacia Hadoop. Almacena datos en HDFS y HBase. Usted deseará iniciar con FlumeNG, que mejora el canal original. [29]

3.10.3.11 Whirr

Suministro de nube para Hadoop. Usted puede arrancar un clúster en unos cuantos minutos con un archivo de configuración muy corto. [29]

3.10.3.12 Mahout

Aprendizaje de máquina para Hadoop. Usado para análisis predictivos y otros análisis avanzados. [29]

3.10.3.13 Fuse

Hace que el sistema HDFS parezca como un sistema de archivos normal para que usted pueda usar ls, rm, cd, y otros en datos HDFS. [29]

3.10.3.14 Zookeeper

Usado para gestionar sincronización para el clúster. Usted no estará trabajando mucho con Zookeeper, pero trabaja mucho por usted. Si usted piensa que necesita escribir un programa que use Zookeeper usted es ya sea muy, muy inteligente y podría formar un comité para un proyecto Apache, o usted está a punto de tener un día terrible. [29]

3.10.3.15 Arquitectura Hadoop

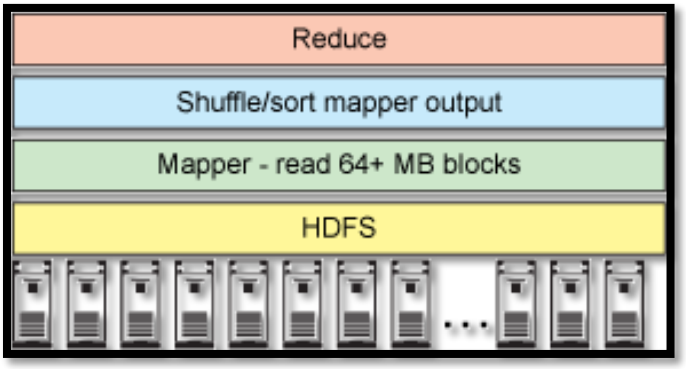


Figura 21 Arquitectura Hadoop. Fuentes: [30].

En la figura N°21 se aprecia que esta arquitectura se base en una granja de servidores que tienen en una escala el componente HDFS, luego el Mapper, seguido de la salida del mapper de forma ordenada y por último el reduce. De esta manera se presenta la arquitectura Hadoop.

3.10.4 Seguridad en Hadoop

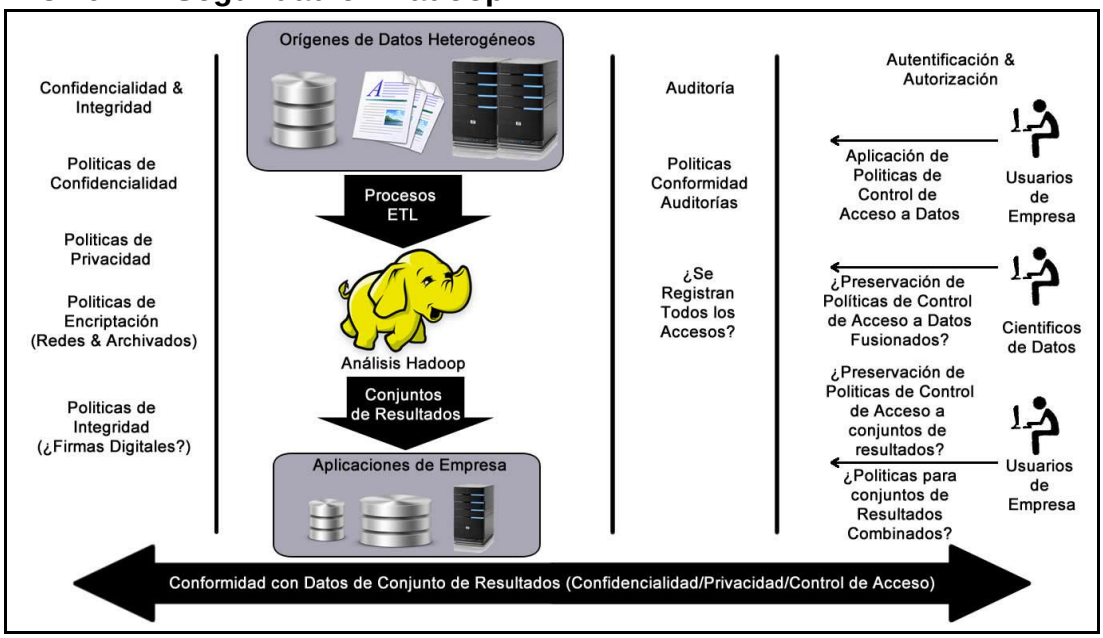


Figura 22 Preocupaciones de seguridad en el ciclo vital de datos Hadoop .Fuente: [28].

En la figura N°22. Se puede apreciar los principales problemas de seguridad y sus características. En los siguientes numerales se explica cada uno de estos problemas al detalle.

3.10.4.1 Autenticación

Autenticación significa validar la identidad del sujeto. Un Sujeto puede ser un usuario, una aplicación, una tarea u otro "Personaje" en un sistema. Hadoop se puede configurar para utilizar Kerberos en la autenticación de usuarios, servicios y servidores en un cluster Hadoop. La autenticación proporciona una

cierta seguridad de que los usuarios y servicios son quienes dicen ser y frustra la suplantación de usuarios, tareas y servicios por parte de sistemas maliciosos. [28]

3.10.4.2 Autorización

La autorización significa determinar qué tiene permiso para hacer un sujeto. Después de que se valide la identidad del sujeto en la autentificación, los sistemas tienen que determinar las credenciales de autorización del sujeto y las comparan con una política de autorización expresada para dar acceso a los recursos solicitados. En la actualidad Hadoop proporciona un determinado nivel de control de acceso, utilizando las ACL para expresar la política de control de acceso para aspectos determinados de Hadoop y permisos de archivos de manera similar a UNIX para el propietario y usuarios de grupos. [28]

Además de lo que proporciona Hadoop, la mayoría de las organizaciones empresariales tienen controles adicionales para la autorización. Por ejemplo, una organización puede tener uno o más de los siguientes: [28]

- *Directorios LDAP o instancias AD que almacenan grupos, funciones y permisos para los sujetos.*
- *Servicios de atributos que utilizan atributos como credenciales de autorización para los sujetos.*
- *STS (Security Token Service, Componente de servicio que compila, firma y emite tokens de seguridad) se utiliza para emitir tokens relacionadas con las credenciales de autorización de un sujeto y para emitir decisiones de autorización en las transacciones.*
- *Servicios de políticas que utilizan estándares como XACML y SAML para expresar la política de control de acceso para recursos y proporcionan decisiones de control de acceso para sujetos.*

3.10.4.3 Confidencialidad

La confidencialidad es el objeto de la seguridad para restringir información sensible y que únicamente las partes autorizadas puedan verla. Cuando la información sensible se transmite por la red, puede ser un requisito que esta información no sea vista pos fisgones durante el tránsito. Esto se consigue mediante la encriptación de la red. Algunas organizaciones requieren la encriptación de los datos en el disco o la encriptación de los datos almacenados, done la criptografía se utiliza en el lugar de almacenamiento de los datos, reduciendo el riesgo de robo de datos sin proteger. [28]

Hadoop proporciona la capacidad y los mecanismos para ofrecer encriptación de red. Sin embargo, no nos dota de las capacidades para encriptar datos almacenados. [28]

3.10.4.4 Integridad

Integridad significa garantizar que los datos no han sido alterados mientras estaban en tránsito o almacenados. Habitualmente, esto se consigue mediante la criptografía y el uso de resúmenes del mensaje, códigos hash o como un efecto secundario de una firma digital. Cunado Hadoop está configurado para implementar la encriptación de red aplica la integridad de datos en tránsito. [28]

3.10.4.5 Auditoria

La mayoría de las compañías dependen de la auditoria de seguridad para proporcionar seguridad sobre las cuestiones de conformidad e identificar brechas de seguridad en potencia. Desde luego, Hadoop se puede configurar para registrar todos los accesos. NameNode almacena un registro local y se puede configurar un registro de auditorías para que escriba en un volumen seguro y garantice la integridad del registro. Las organizaciones pueden tener requisitos de auditoría más exigentes relacionados con la autentificación y la autorización. [28]

3.10.4.6 Ejemplos de aislamiento de red

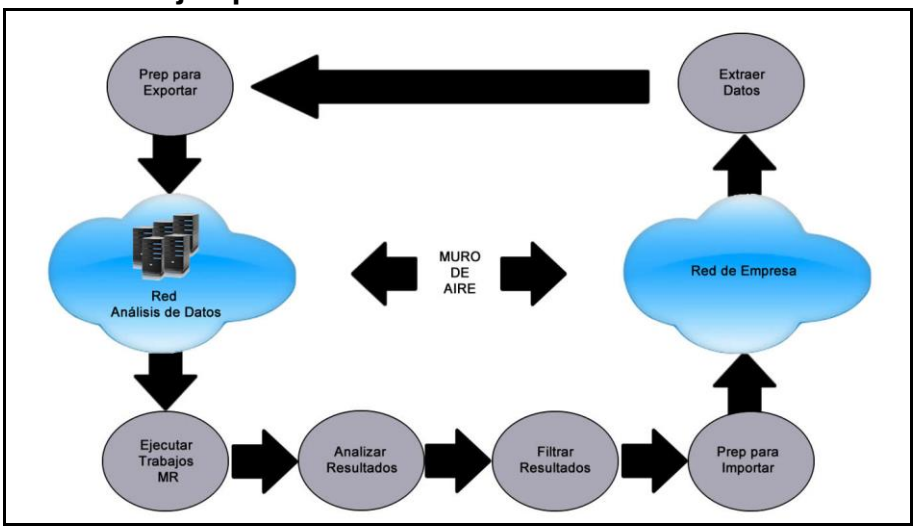


Figura 23 Aislamiento de red "muro de aire". Fuente: [28].

Como se aprecia en la figura N°23*. Una organización crea una red "Análisis de datos" separada de la red empresarial de las organizaciones con un "muro de aire" que impide la transferencia de datos entre las dos redes. Los científicos de datos con los controles de acceso adecuado ejecutan consultas y operaciones MapReduce sobre los clústeres de Hadoop en la red de análisis de datos y el acceso a esta red está controlado por seguridad física y/o autenticación ante las maquinas del cliente utilizadas para ejecutar consultas.* [28]

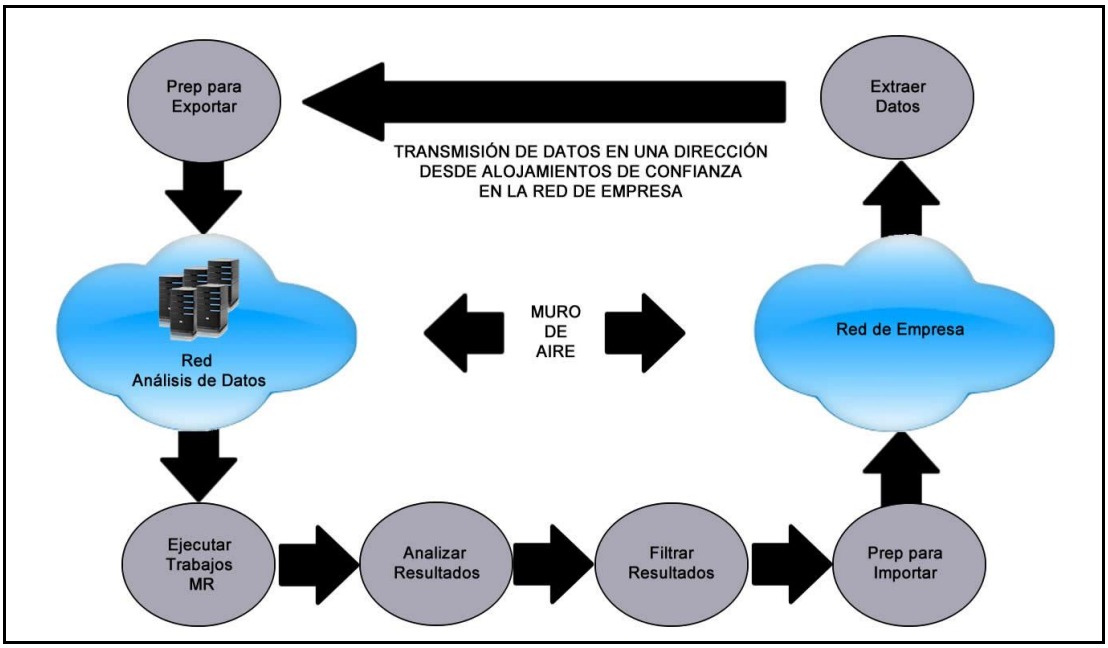


Figura 24. Aislamiento de red con transmisiones en una dirección. Fuente: [28].

> *El proceso ETL se puede realizar sobre la red, eliminando el primer paso del proceso anterior.* [28]

3.11 Infraestructura en la Nube

3.11.1 Amazon Web Services (AWS)

Es una colección de servicios en la nube que en conjunto forman una plataforma de computación en la nube, es una solución ofrecida por Amazon para que las empresas no se enfoquen en la infraestructura de un sistema si no en el negocio.

3.11.2 Componentes Principales AWS

3.11.2.1 Amazon S3 (Simple Storage Service)

Amazon Simple Storage Service (Amazon S3) ofrece a los desarrolladores y los profesionales de TI un almacenamiento de objetos seguro, duradero y altamente escalable. Amazon S3 es fácil de utilizar e incorpora una sencilla interfaz de servicios web para almacenar y recuperar cualquier cantidad de datos desde cualquier ubicación de la web. Con Amazon S3 pagará únicamente por el almacenamiento que realmente use. Sin cuota mínima ni coste de contratación.

Ofrece varios tipos de almacenamiento diseñados para distintos casos de uso, como Amazon S3 Estándar para el almacenamiento de uso general de datos a los que se accede con frecuencia, Amazon S3 Estándar – Acceso poco frecuente (Estándar – IA) para datos de larga duración a los que se accede con menos frecuencia y Amazon Glacier para el archivado a largo plazo.

3.11.2.2 Amazon EC2 (Elastic Compute Cloud)

Es un servicio web que proporciona capacidad informática con tamaño modificable en la nube. Está diseñado para facilitar a los desarrolladores la informática en la nube escalable basada en web.

La sencilla interfaz de servicios web de Amazon EC2 permite obtener y configurar su capacidad con una fracción mínima. Proporciona un control completo sobre sus recursos informáticos y permite ejecutarse en el entorno informático acreditado de Amazon.

3.11.2.3 Amazon Redshift (Data Warehouse)

Amazon Redshift es un almacén de datos rápido y totalmente gestionado a escala de petabytes que permite analizar todos los datos empleando de forma sencilla y rentable las herramientas de inteligencia empresarial existentes.

3.11.2.4 Amazon EMR (Elastic Map Reduce)

Amazon Elastic MapReduce (Amazon EMR) es un servicio web que facilita el procesamiento rápido y rentable de grandes cantidades de datos.

Capítulo 4

4 Marco conceptual

4.1 Diseño

4.1.1 Levantamiento de información

- SCI es una compañía de tecnología para la gestión de flotas y de combustible.
  Estos son algunos de los productos que ofrece:

Gestión de compras de combustible
Gestión automática del reaprovisionamiento
Administración de almacenamiento de combustible. Fuente: [31]

Soluciones telemáticas
Mantenimiento preventivo.
Fuente: [31].

Seguridad de flota y estación por medio de cámaras. Fuente: [31].

Gestión de flotas por medio de GPS Velocidad y Gestión ralentí. Fuente: [31]

Figura 25. Productos ofrecidos por SCI. Fuente: [31].

Se puede ver claramente que en cada uno de estos productos ofrecidos por esta empresa se puede hacer uso de Big Data. Sin embargo decidimos centrarnos en dar una solución para el producto GPS. Con base en la experiencia que nos deje GPS Tracking continuaremos con los demás productos.

SCI se enfrenta a problemas como, el almacenamiento, la velocidad y para nuestra opinión se requiere de un análisis de datos más sofisticado. Entonces es necesario realizar un diseño para el almacenamiento, velocidad y análisis de información.

4.1.2 Problemas actuales

- Se presentan un problema de almacenamiento, cada 6 meses tienen que liberar la tabla de transmisiones de GPS para que las consultas sean óptimas.
- En la aplicación web solo se puede ver un rango específico de registros en la tabla de eventos del GPS.
- Al intentar realizar consultas de gran tamaño, por ejemplo: consultar el último año de un vehículo no es posible, primero porque no se cuenta con la información de un año de un vehículo y segundo porque la consulta se puede quedar más de 10 minutos.

4.1.3 Tipos de datos

- Datos: Estos son los recogidos por las transmisiones de GPS, además existen dispositivos que permiten el control del combustible, que generan transacciones y datos que pueden ser analizados.
- Video: Existe un producto llamado Vision, que permite instalarle una cámara a un vehículo y grabar su recorrido, además se puede instalar en una estación de combustible y monitorear la operación.
- Imágenes: El producto Vision, permite tomar fotografías de algún instante que se requiera.
- Voz: No tienen ningún producto que maneje voz.

4.1.4 Reportes de base de datos

4.1.4.1 Reporte Uso de disco

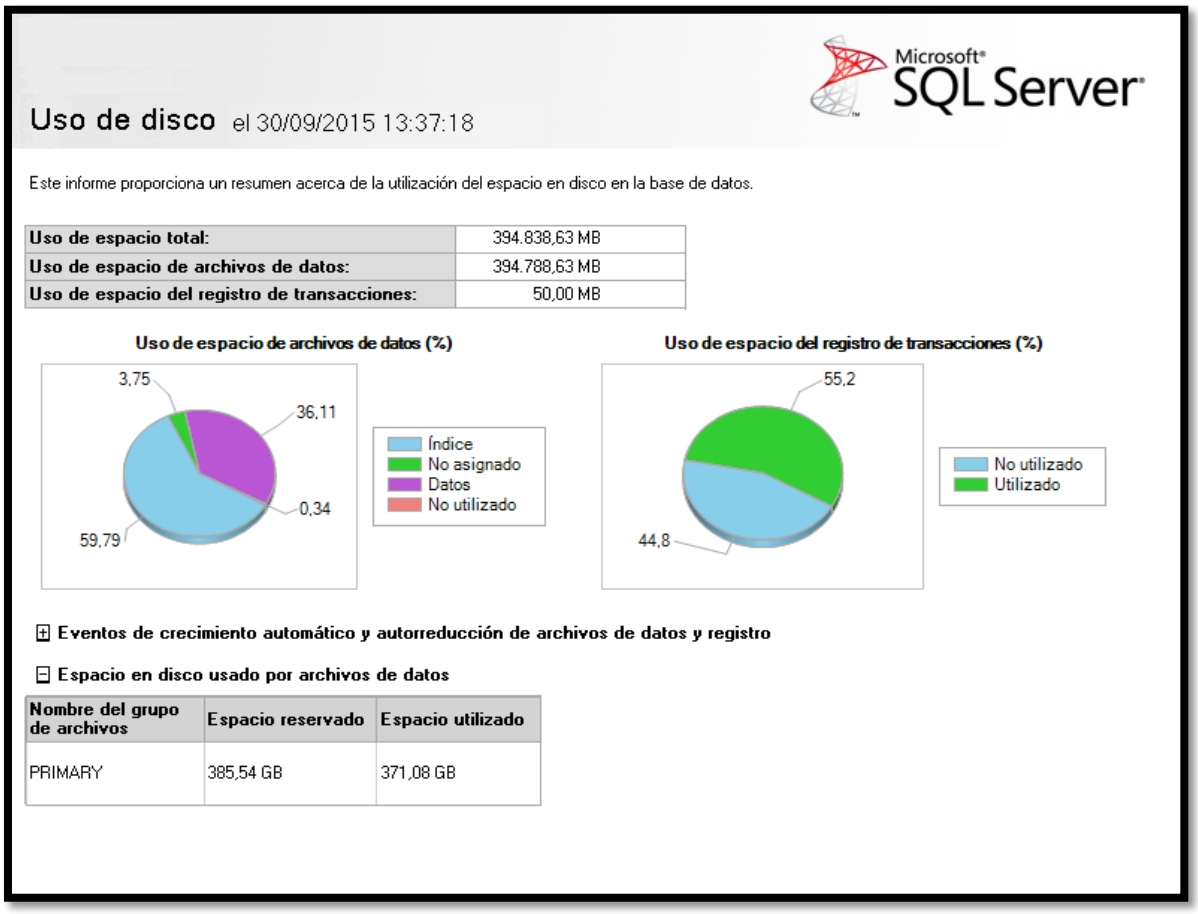


Figura 26 Uso de disco base de datos. Fuente: los autores.

Como se puede apreciar en la figura N°26, en los últimos 4 meses se almacenaron 371.08 GB eso quiere decir que cada mes la base de datos aproximadamente almacena 93 GB y cada día aproximadamente se almacenan 3,1 GB.

Con los datos anteriores se calcula que dentro de un año la base de datos pesara aproximadamente 1.116 GB es decir 1,116 TB, y dentro de 5 años la base de datos puede llegar a pesar aproximadamente 5.580 GB es decir 5,58 TB

4 Meses (GB)	1 Mes (GB)	Día (GB)
372,0	93,0	3,1

Tabla 4 Información tamaño base de datos. Fuente: los autores.

	Tamaño (GB)	Tamaño (TB)
1 Año	1.116,0	1,1
2 Años	2.232,0	2,2
3 Años	3.348,0	3,3
4 Años	4.464,0	4,4
5 Años	5.580,0	5,4
10 Años	11.160,0	10,9

Tabla 5 Estimado tamaño de base de datos. Fuente: los autores.

- El almacenamiento estimado en los próximos 5 años es de 5.4 TB aproximadamente.

4.1.5 Reporte tráfico de red

Para medir el tráfico de red de la empresa SCI, se instaló el software PRTG Network Monitor que permite realizar seguimiento al tráfico.

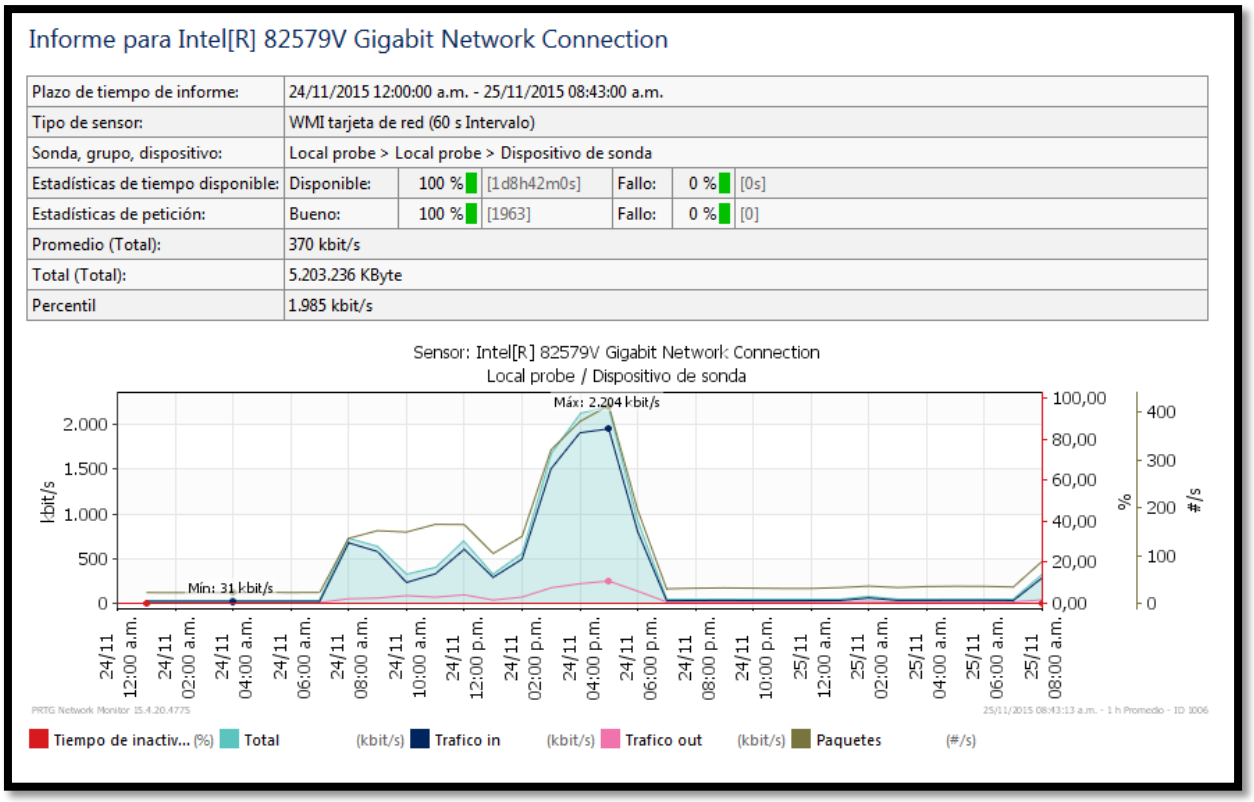


Figura 27. Tráfico de red SCI. Fuente: Los autores

En la figura N°27 se evidencia el tráfico de red para 1 servidor de la empresa.

Teniendo en cuenta la totalidad de servidores, el tráfico de red máximo es de 22.040 Kbit/s, y el mínimo es de 310 Kbit/s.

El 10% es la salida de la red, es decir 2.204 Kbit/s, mientras que 19.836 Kbit/s representaría la entrada.

4.1.6 Recomendación de servidores

Con la recomendación de un experto en el tema presentamos las siguientes maquinas con el fin de dar solución al almacenamiento y el procesamiento:

4.1.6.1 Recomendación de servidor de proceso y de base de datos.

1. **PowerEdge R630 Rack Server**

Figura 28. PowerEdge R630 Rack Server DELL. Fuente: [32].

PowerEdge R630 Rack Server	
Procesador	Intel® Xeon® E5 2600 v3 processors
Sistema operativo	• Microsoft® Windows Server® 2008 R2 • Microsoft Windows Server 2012 • Microsoft Windows Server 2012 R2 • Novell® SUSE® Linux Enterprise Server • Red Hat® Enterprise Linux • VMware® ESX®
Chipset	Intel C610 series chipset
Memoria	24 DIMM slots, DDR4 memory Architecture: Up to 2133MT/s DDR4 DIMMs Memory Type: RDIMM, LRDIMM Memory Module Sockets: 24 Minimum RAM: 4GB (one module) Maximum RAM: Up to 768GB (24 DIMM slots): 4GB/8GB/16GB/32GB
Almacenamiento	HDD: SAS, SATA, nearline SAS SSD: SAS, SATA, PowerEdge Express Flash NVMe PCIe SSD 24 x 1.8” SSD – up to 23TB via 0.96TB hot-plug SATA SSD 10 x 2.5” – up to 18TB via 1.8TB hot-plug SAS HDD 8 x 2.5” – up to 14TB via 1.8TB hot-plug SAS HDD

Slots	2 CPUs, 3 slots Slot 1: Half length, half height - PCIe 3.0 x16 (x16 connector) Slot 2: Half length, half height - PCIe 3.0 x8 (x16 connector) Slot 3: Half length, half height - PCIe 3.0 x16 (x16 connector) 2CPUs, 2 slots Slot 1: Half length, half height - PCIe 3.0 x16 (x16 connector) Slot 2: 3/4 length, full height - PCIe 3.0 x16 (x16 connector) 1CPU, 2 slots Slot 1: Half length, half height - PCIe 3.0 x8 (x16 connector) Slot 2: 3/4 length, full height - PCIe 3.0 x16 (x16 connector Dedicated RAID card slot
Controladores RAID	Internal: PERC S130 (SW RAID), PERC H330 PERC H730 PERC H730P External: PERC H830 External HBAs (non-RAID): 12Gbps SAS HBA
Controladores de red	4 x 1Gb, 2 x 1Gb + 2 x 10Gb, 4 x 10Gb
Comunicaciones	Broadcom® 5719 quad-port 1Gb NIC Broadcom 5720 dual-port 1Gb NIC Broadcom 57810 dual-port 10Gb DA/SFP+ CNA Broadcom 57810 dual-port 10Gb Base-T network adapter Intel® Ethernet I350 dual-port 1Gb server adapter Intel Ethernet I350 quad-port 1Gb server adapter Intel Ethernet X540 dual-port 10GBASE-T server adapter Mellanox® ConnectX®-3 dual-port 10Gb Direct Attach/SFP+ server network adapter Mellanox ConnectX-3 dual-port 40Gb Direct Attach/QSFP server network adapter Emulex® LPE 12000, single-port 8Gb Fibre Channel HBA Emulex LPE 12002, dual-port 8Gb Fibre Channel HBA Emulex LPe16000B, single-port 16Gb Fibre Channel HBA Emulex LPe16002B, dual-port 16Gb Fibre Channel HBA Emulex OneConnect OCe14102-U1-D 2-port PCIe 10GbE CNA

	QLogic® 2560, single-port 8Gb Optical Fibre Channel HBA QLogic 2562, dual-port 8Gb Optical Fibre Channel HBA Qlogic 2660, single-port 16GB, Fibre Channel HBA, full height Qlogic 2662, dual-port 16GB, Fibre Channel HBA, full height
Energía	1100W AC, 86 mm (Platinum) 1100W DC, 86 mm 750W AC, 86 mm (Platinum) 750W AC, 86 mm (Titanium) 495W AC, 86 mm (Platinum)

Tabla 6. Recomendación 1, PowerEdge R630 Rack Server DELL [32]

Dell PowerEdge R710

Figura 29 Dell PowerEdge R710. Fuente: [33].

En la figura N°29 se puede apreciar una de las opciones recomendadas para el modelo.

Dell PowerEdge R710	
Procesador	Quad-core or six-core Intel® Xeon® processor 5500 and 5600 series
Sistema operativo	• Microsoft Windows® Small Business Server 2011 • Microsoft Windows Small Business Server 2008 • Microsoft Windows Server® 2008 SP2, x86/x64 (x64 • includes Hyper-V®) • Microsoft Windows Server 2008 R2 SP1, x64 (includes • Hyper-V v2) • Microsoft Windows HPC Server 2008 R2 • Novell® SUSE® Linux® Enterprise Server • Red Hat® Enterprise Linux • Oracle® Solaris™
Chipset	Intel 5520

Memoria	Up to 288GB (18 DIMM slots): 1GB/2GB/4GB/8GB/16GB DDR3 up to 1333MHz
Almacenamiento	Up to 18TB • **Hot-plug hard drive options:** 2.5" SAS SSD, SATA SSD, SAS (15K, 10K), nearline SAS (7.2K), SATA (7.2K) 3.5" SAS (15K, 10K), nearline SAS (7.2K), SATA (7.2K) • **Solid state storage cards:** Fusion-io® 160GB ioDrive PCIe solid state storage card Fusion-io 640GB ioDrive Duo PCIe solid state storage card Fusion-io 320GB ioDrive Mono PCIe solid state storage card Fusion-io 640GB ioDrive Mono PCIe solid state storage card Fusion-io 1.28TB ioDrive Mono PCIe solid state storage card
Slots	4 PCIe G2 slots + 1 storage slot: two x8 slots, two x4 slots, one x4 storage slot
Controladores RAID	Internal: PERC H200 (6Gb/s) PERC H700 (6Gb/s) with 512MB battery-backed cache; 512MB, 1GB Non-Volatile battery-backed cache SAS 6/iR PERC 6/i with 256MB battery-backed cache External: PERC H800 (6Gb/s) with 512MB of battery-backed cache; 512MB, 1GB Non-Volatile battery-backed cache PERC 6/E with 256MB or 512MB of battery-backed cache External HBAs (non-RAID): 6Gbps SAS HBA SAS 5/E HBA LSI2032 PCIe SCSI HBA

Comunicaciones	Four embedded Broadcom® NetXtreme® II 5709c Gigabit Ethernet NIC with failover and load balancing; TOE (TCPIP Offload Engine) supported on Microsoft® Windows Server® 2003 SP1 or higher with Scalable Networking Pack; Optional 1GBe and 10GBe add-in NICs Broadcom NetXtreme II 57711 Dual Port Direct Attach 10Gb Ethernet PCI-Express Network Interface Card with TOE and iSCSI Offload Intel Gigabit ET Dual Port Server Adapter and Intel Gigabit ET Quad Port Server Adapter Dual Port 10GB Enhanced Intel Ethernet Server Adapter X520-DA2 (FcoE Ready for Future Enablement)
Energía	Energy Smart: Two hot-plug, high-efficient 570W power supplies or High Output: Two hot-plug 870W power supplies Uninterruptible power supplies: 1000W–5600W 2700W–5600W High-Efficiency Online Extended Battery Module (EBM) Network Management Card

Tabla 7. Recomendación 2, Dell PowerEdge R710. Fuente: [33].

Cualquiera de estos dos equipos mencionados anteriormente puede usarse para la implementación del Sistema Big Data.

4.1.6.2 Recomendación de arreglo de discos.

Al realizar la investigación de cuál es el arreglo de discos que más se adapta a Hadoop y a la necesidad de la empresa, se encontró el siguiente servidor:

Sistema HP Apollo 4500/4510

Para almacenamiento de objetos, el Apollo 4510 ultra denso incluye 1 servidor y hasta 68 factor de forma grande (LFF) en un chasis de 4U para un máximo de 544 TB por cada sistema

Para el análisis de Hadoop y otras soluciones de Big Data, el Apollo 4530 ofrece exclusivamente 3 servidores por chasis, ideal para albergar 3 copias de datos en un único sistema. Apollo 4510 y Apollo 4530 le permiten obtener todo el valor de su Big Data al costo adecuado y en una cantidad de espacio mínima.

Especificaciones de arreglo de discos:

- **Procesadores:** Procesadores Intel Xeon E5-2600v3
- **Número de procesadores:** 1 o 2 por nodo

- **Número de nodos de computación:** 1 o 3
- **Núcleo de procesador disponible:** 6, 8, 10, 12, 14, 16
- **Velocidad del procesador:** 1,6-2,6 GHz
- **Ranuras de memoria:** 16 ranuras DIMM máximo por nodo (8 ranuras DIMM por procesador)
- **Memoria máxima:** 512 GB por nodo
- **Tipo de memoria:** HP SmartMemory - DDR4 R y L DIMM
- **Compatibilidad con unidades:** (68) LFF SAS/SATA/SSD para Apollo 4510, (15) LFF SAS/SATA/SSD por servidor para Apollo 4530-
- **Tipo de almacenamiento:** SAS LFF de 3,5 pulgadas con conexión en caliente, SATA LFF de 3,5 pulgadas con conexión en caliente, SSD SAS LFF de 3,5 pulgadas con conexión en caliente, SSD SATA LFF de 3,5 pulgadas con conexión en caliente, SSD SATA SFF de 2,5 pulgadas con conexión en caliente, SDD SATA de 6 Gb M.2
- **Controladores de red:** HP Ethernet 1 Gb 2 puertos 361T integrado
- **Controladores de almacenamiento:** HP Smart Array P440/P441, HP Smart Array P840/841, HP H240/241 Smart Host Bus Adapter, HP H244br Smart Host Bus Adapter
- **Ranuras de expansión:** Hasta 4 ranuras Gen3 PCIe
- **Gestión:** HP iLO4 Management, HP Insight Control
- **Características de los ventiladores del sistema:** 5 módulos de ventiladores duales estándar de serie
- **Tipo de fuente de alimentación:** Hasta 4 fuentes de alimentación, 800 W y 1400 W Fuente de alimentación redundante de conexión en caliente (hasta el 94 % de eficacia)
- **Factor de forma:** 4U

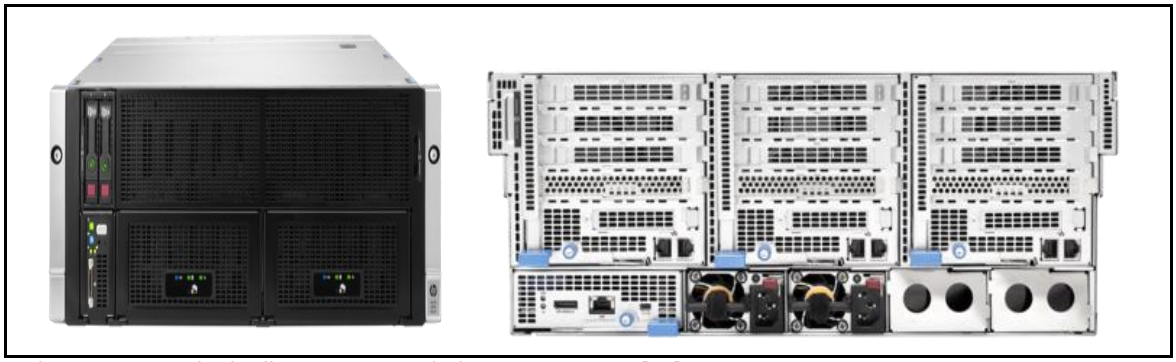


Figura 30. Arreglo de discos recomendado I Y II. Fuente: [34].

Figura 31. Arreglo de discos recomendado III. Fuente: [34].

La figura N°30 y N°31, representan una imagen física del arreglo de discos recomendado para el modelo.

Selección de Switch: Según las especificaciones del arreglo de discos se recomiendan el uso de alguno de estos dos switch, es decir que para sacar el máximo de este arreglo se debería usar alguno de estos dos:

1. En la figura N°32 se puede apreciar el, El HP 5900AF-48XGT-4QSFP+, 10GbE que es un switch TdR ideal, con cuarenta y ocho puertos de 10 GbE y cuatro enlaces ascendentes 40GbE proporcionando elasticidad, alta disponibilidad y escalabilidad de apoyo. Además, este modelo viene con soporte para cables CAT6 (cables de cobre) y Software Defined Networking (SDN). Un interruptor de administración dedicada para el tráfico de la OIT no se requiere que el ProLiant DL360 Gen9 y Apolo 4530 son capaces de compartir el tráfico iLO sobre NIC.

Figura 32. HP 5900AF-48G-4XG-2QSFP, Recomendación 1. Fuente: [35].

2. En la figura N°33 se puede apreciar, el FlexFabric 5930-32QSFP+, 40GbE.Que es un switch que tiene una mejor conectividad con 32 puertos 40GbE, soportando hasta 104 puertos de 10 GbE a través del desglose cables y seis puertos de enlace ascendente 40GbE, redundancia switch de agregación y de alta disponibilidad (HA) de soporte con puertos de enlace IRF. SDN listo con OpenFlow 1.3.

Figura 33. FlexFabric 5930-32QSFP+, recomendación 2. Fuente: [35].

4.1.7 Diseños

De acuerdo con el estado del arte y la necesidad de la empresa se plantean los siguientes diseños:

4.1.7.1 Diseño de repositorio de imágenes y videos.

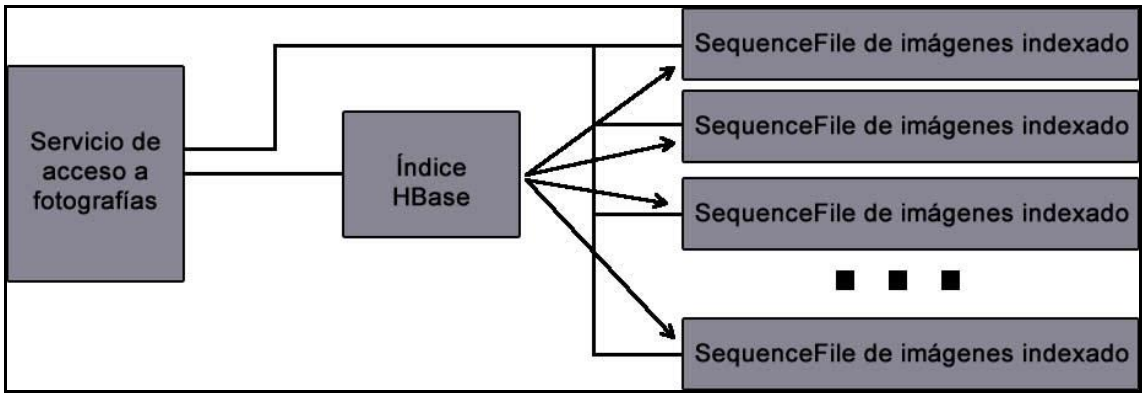


Figura 34 Diseño del sistema de gestión de imágenes con Hadoop. Fuente: [28].

En la figura N°34, se puede apreciar el enfoque utilizado para la implementación de la gestión de archivos. Se pretende que estos archivos sean fotografías. Se puede observar que existe un índice *HBase*, que se encarga de gestionar los *SequenceFile* de imágenes de indexado.

4.1.7.1 Diseño de repositorio de datos

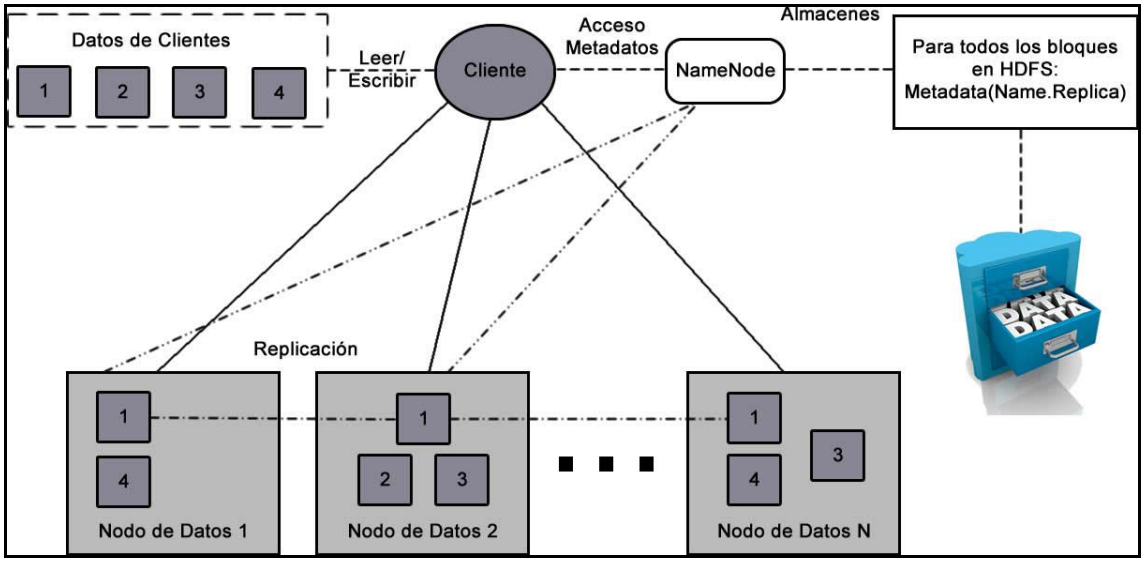


Figura 35 Diseño de repositorio de datos HDFS Hadoop. Fuente: [28].

En la figura N°35, se puede apreciar que, "*Los archivos individuales están divididos en bloques de n tamaño determinado que se almacena en el clúster de Hadoop. Un archivo puede estar compuesto de varios bloques, que se almacenan en distintos nodos de datos (maquinas individuales en el clúster) elegidos aleatoriamente sobre una base formada por bloques. Como resultado, el acceso a un archivo suele requerir acceder a varios nodos de datos, lo que significa que HDFS soporta los tamaños de archivos mucho mejor que la capacidad de disco de una única máquina".* [28]

4.1.7.1 *Diseño de servidores de procesamiento*

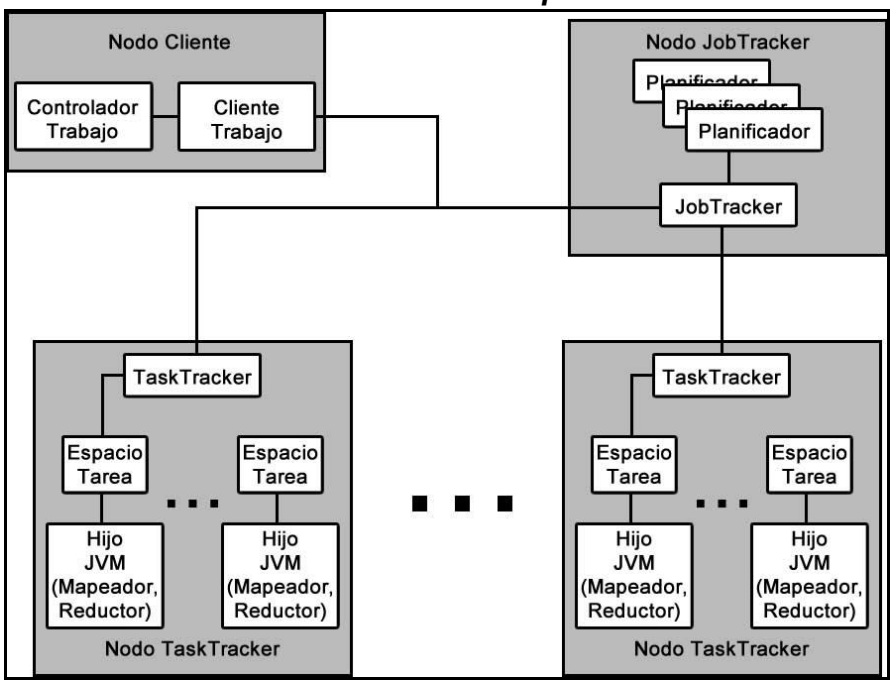


Figura 36 Procesamiento de datos con MapReduce Hadoop. Fuente: [28].

En la figura N°36, se puede apreciar que*, "la ejecución MapReduce de Hadoop utiliza un mecanismo de coordinación muy sencillo. Un controlador del trabajo utiliza InputFormat para particionar una ejecución de map basada en las divisiones de datos e inicia un cliente de trabajo, que comunica con JobTracker y envía el trabajo para su ejecucion. Una vez enviado el trabajo, el cliente del trabajo puede preguntar a JobTracker mientras espera que se complete el trabajo. JobTracker crea una tarea map para cada división y un conjunto de tareas reducer. La cantidad de tareas reducer a crear está determinada por la configuración del trabajo"* [28]

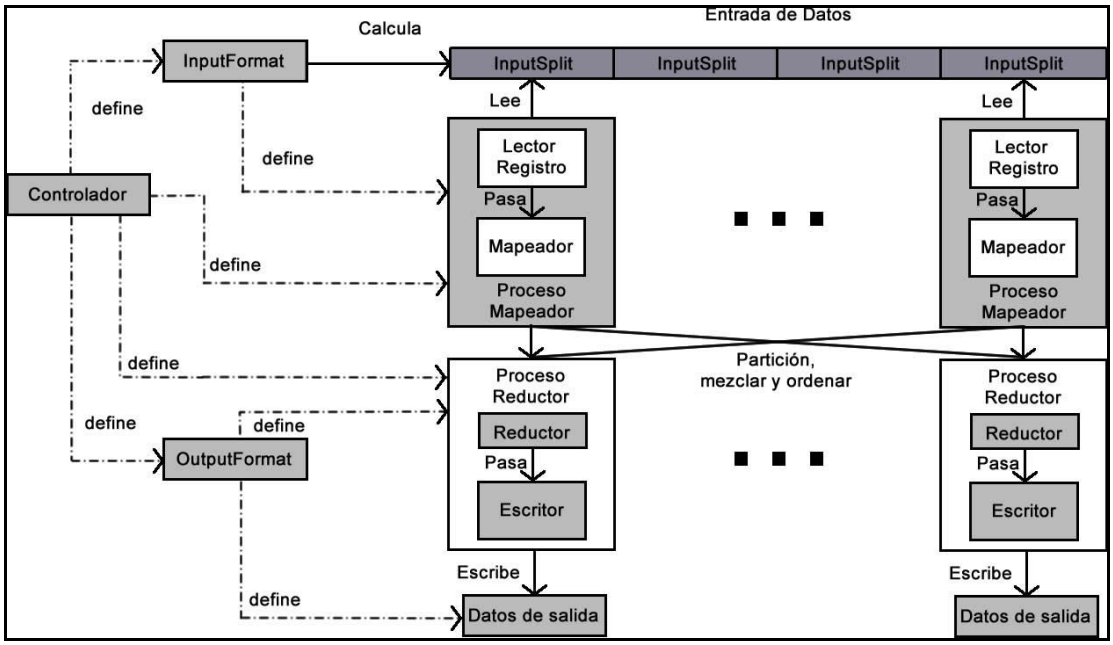


Figura 37. Arquitectura de Alto Nivel para el procesamiento. Fuente: [28].

Como se puede apreciar en la figura N°37*, "Cualquier dato almacenado en Hadoop (incluidos HDFS y HBase) o incluso fuera de Hadoop (una base de datos) puede ser utilizado como una entrada a la tarea MapReduce. De manera similar, la salida del trabajo puede ser almacenado en Hadoop (HDFS o HBase) o fuera de él. La estructura se ocupa de la programación y monitorización de las tareas también de repetir la ejecución de las tareas fallidas"* [28].

4.1.8 Diseño de modelo para la implementación de un sistema Big Data

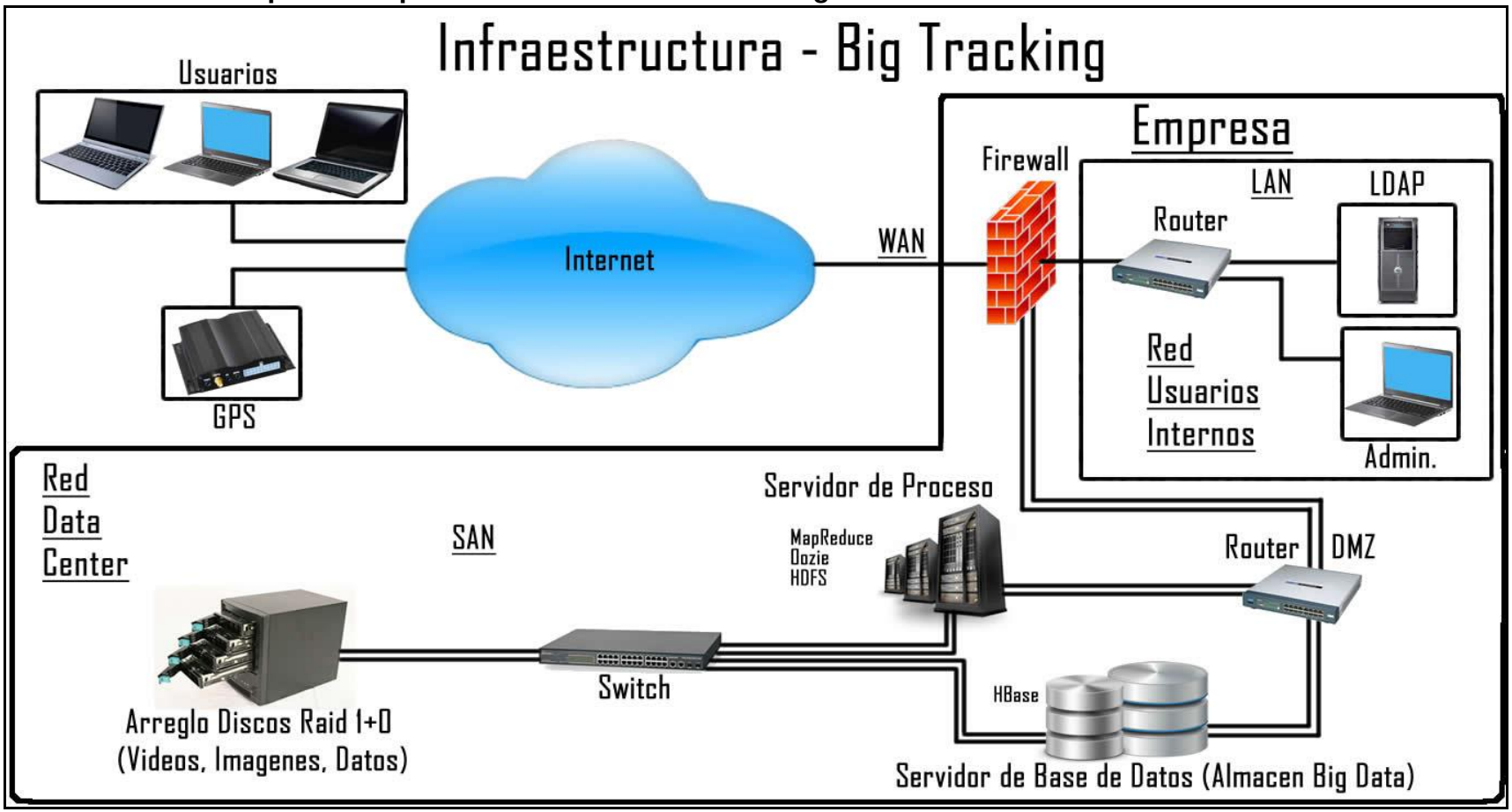


Figura 38, Diseño de arquitectura Big Tracking Fuente: los Autores.

En la figura N°38 se puede apreciar el diseño físico propuesto para dar solución al problema, en este se tiene en cuenta los usuarios de la empresa y además dos canales de comunicación entre el servidor de procesos, el servidor de base de datos y el arreglo de discos, dando una posible solución a la caída de uno de los canales, es decir si un canal de comunicación pierde comunicación entonces está disponible el otro y viceversa. Fuente: los autores.

4.1.9 Diseño de modelo para la implementación de un sistema Big Data en Amazon WebServices

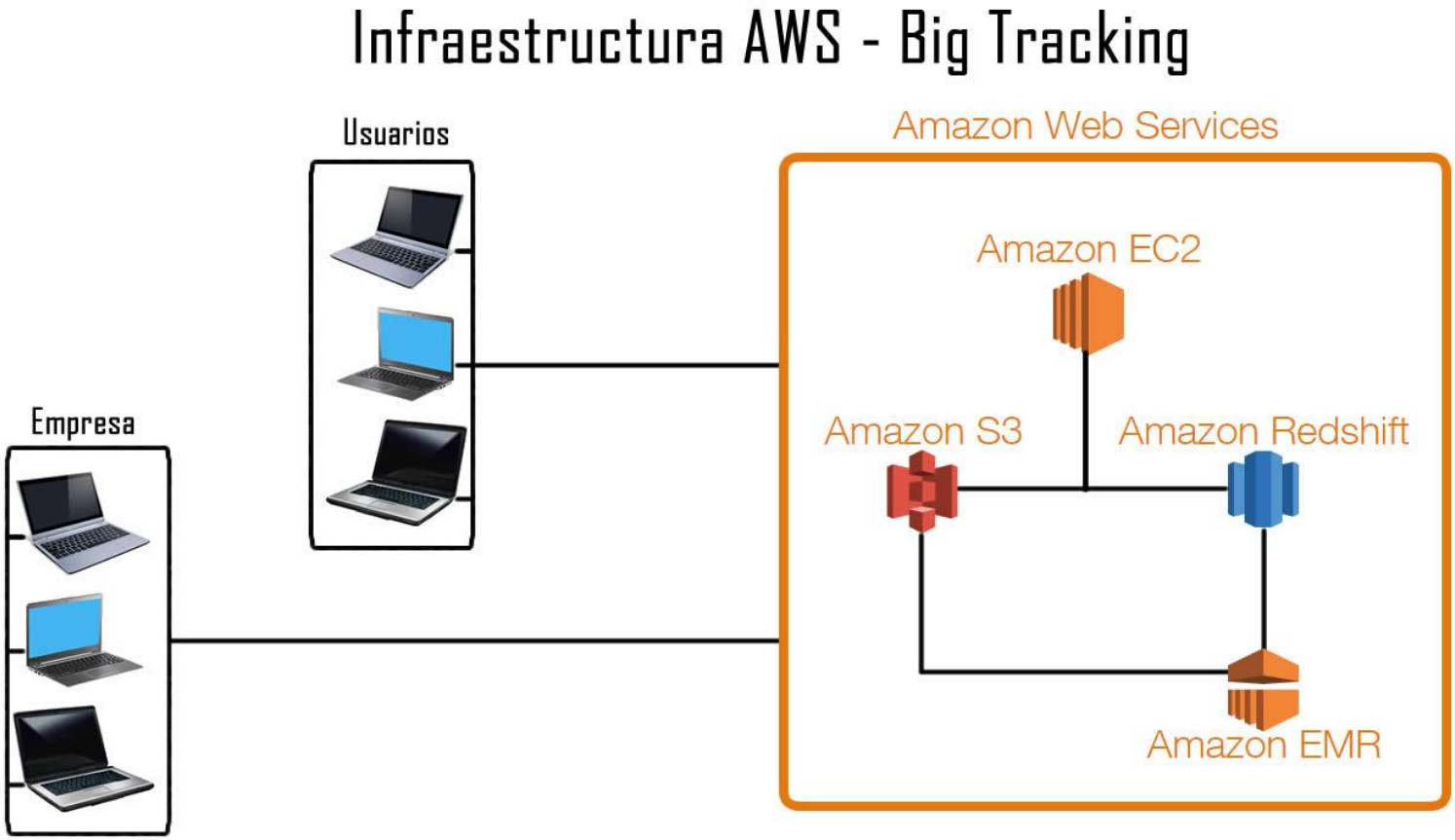


Figura 39, Diseño de arquitectura Big Tracking AWS. Fuente: los autores.

En la figura N°39 se puede apreciar el diseño propuesto para dar solución al problema, pero orientado a un conjunto de servicios que ofrece Amazon.

5 Conclusiones

1. Big Data son todos aquellos datos que no pueden ser procesados o analizados utilizando procesos o herramientas tradicionales, estos datos pueden ser estructurados o no estructurados, y si se manejan bien se obtendrán ventajas alucinantes.
2. Las empresas utilizan Big Data para obtener resultados centrados en el cliente, aprovechar los datos internos y crear un mejor ecosistema de información.
3. De nuestra investigación hemos podido extraer cinco recomendaciones clave para que las empresas puedan avanzar en sus iniciativas de Big Data y obtener el máximo valor de negocio:
    - Dedicar los esfuerzos iniciales a obtener resultados centrados en el cliente.
    - Desarrollar un plan de Big Data para toda la empresa.
    - Comenzar con datos ya existentes para lograr resultados a corto plazo.
    - Desarrollar capacidades analíticas sobre la base de prioridades de negocio.
    - Crear un caso de negocio sobre la base de resultados cuantificables.

4. Las empresas deben entender las opciones que tienen para poder implementar un sistema de Big Data dentro de su organización.
5. El conjunto de herramientas popular para dar solución a problemas de velocidad y almacenamiento es Hadoop. Para su uso se debe identificar muy rigurosamente los problemas.
6. En Hadoop la seguridad está centrada en la autentificación, la autorización, la confidencialidad, la integridad y la auditora, porque estos son aspectos clave de la seguridad para aplicaciones de empresa. La disponibilidad o garantizar el acceso a Hadoop es verdaderamente importante y ha sido una preocupación ocasionada por la manera en que fue diseñado.
7. Los clústeres de Hadoop son altamente fiables, tienen una trayectoria de disponibilidad excelente y se pueden complementar mediante otros mecanismos de seguridad en la empresa, como IDS (Instrusion Detection System, Sistema de detección de intrusos), que protege de ataques DoS.
8. Los análisis de Big Data se utilizan para encontrar patrones y tendencias que se sirven para aumentar la eficiencia y dotar la toma de decisiones de maneras que nunca antes habían sido posibles.
9. Big Data en un país como Colombia es una excelente oportunidad no solo laboral, también para que las personas se capaciten respecto al uso y el poder de los datos en la toma de decisiones.
10. Amazon Web Services es una muy buena opción para olvidarse de la infraestructura de una empresa y centrarse en el negocio.
11. Colombia necesita Científicos de Datos, es hora de que el país sobresalga en un área tecnológica e investigativa.
12. Un artículo de investigación sostuvo que “Hemos entrado en una era de grandes datos. A través de un mejor análisis de los grandes volúmenes de datos que se están haciendo disponibles, existe el potencial para hacer avances más rápidos en muchas disciplinas científicas y la mejora de la rentabilidad y el éxito de muchas empresas. Sin embargo, muchos desafíos técnicos deben ser abordados antes de que este potencial puede realizarse plenamente. Los desafíos incluyen no sólo las cuestiones obvias de escala, sino también la

heterogeneidad, falta de estructura, control de errores, la privacidad, la puntualidad, la procedencia y la visualización. Estos desafíos técnicos son comunes través de una gran variedad de dominios de aplicación, y por tanto no rentables para abordar en el contexto de un dominio solo. Además, estos desafíos requerirán soluciones transformadoras, y no serán abordados de forma natural por la próxima generación de productos industriales. Debemos apoyar y fomentar la investigación fundamental hacia la solución de estos problemas técnicos, si queremos alcanzar los beneficios prometidos de Big Data.” Tomado de [36].

6 Anexos

6.1 MongoDB vs SQL Server

Terminología y conceptos

La siguiente tabla presenta la terminología y los conceptos de SQL y la terminología y los conceptos MongoDB. [37]

Terminología SQL	Terminología MongoDB
database	database
table	collection
row	document o BSON document
column	field
Index	index
table joins	embedded documents and linking
primary key	primary key
aggregation (e.g. group by)	Ver tabla **Mapeo de agregación**

Anexo 1 Terminología SQL - MongoDB

Comparativo entre agregación en SQL y MongoDB

Terminología SQL	Terminología MongoDB
WHERE	$match
GROUP BY	$group
HAVING	$match
SELECT	$project
ORDER BY	$sort
LIMIT	$limit
SUM()	$sum
COUNT()	$sum
join	No existe una correspondencia directa, sin embargo $unwind realiza una tarea similar.

Anexo 2 Mapeo de agregación SQL – MongoDB

Comparativo entre sentencias básicas en SQL y MongoDB

SQL Schema Statements	**MongoDB Schema Statements**
CREATE TABLE users (id MEDIUMINT **NOT NULL** AUTO_INCREMENT, user_id Varchar(30), age Number, status char(1), **PRIMARY KEY** (id))	Creado implícitamente en la primera operación insert(). La llave primaria _id es automáticamente añadida si no se especifica campo acabado en _id. **db**.**users**.**insert**({ **user_id**: "abc123", **age**: 55, **status**: "A" }) Sin embargo, se puede crear también explícitamente una coleección: **db**.**createCollection**("users")
ALTER TABLE users **ADD** join_date DATETIME	**db**.**users**.**update**({ }, { **$set**: { **join_date**: **new** Date() } }, { **multi**: **true** })
ALTER TABLE users **DROP COLUMN** join_date	**db**.**users**.**update**({ }, { **$unset**: { **join_date**: "" } }, { **multi**: **true** })
CREATE INDEX idx_user_id_asc **ON** users(user_id)	**db**.**users**.**createIndex**({ **user_id**: 1 })
CREATE INDEX idx_user_id_asc_age_desc **ON** users(user_id, age **DESC**)	**db**.**users**.**createIndex**({ **user_id**: 1, **age**: -1 })
DROP TABLE users	**db**.**users**.**drop**()

Anexo 3 Sentencias Básicas

Comparativo entre Inserciones en SQL y MongoDB

SQL INSERT Statements	MongoDB insert() Statements
INSERT INTO users(user_id, age, status) **VALUES** ("bcd001", 45, "A")	**db**.**users**.**insert**({ user_id: "bcd001", age: 45, status: "A" })

Anexo 4 Inserciones SQL - MongoDB

Comparativo entre consultas SQL y MongoDB

SQL SELECT Statements	MongoDB find() Statements
SELECT * **FROM** users	**db**.**users**.**find**()
SELECT id, user_id, status **FROM** users	**db**.**users**.**find**({ }, { **user_id**: 1, **status**: 1 })
SELECT user_id, status **FROM** users	**db**.**users**.**find**({ }, { **user_id**: 1, **status**: 1, **_id**: 0 })
SELECT * **FROM** users **WHERE** status = "A"	**db**.**users**.**find**({ **status**: "A" })
SELECT user_id, status **FROM** users **WHERE** status = "A"	**db**.**users**.**find**({ **status**: "A" }, { **user_id**: 1, **status**: 1, **_id**: 0 })
SELECT * **FROM** users **WHERE** status != "A"	**db**.**users**.**find**({ **status**: { **$ne**: "A" } })
SELECT * **FROM** users **WHERE** status = "A" **AND** age = 50	**db**.**users**.**find**({ **status**: "A", **age**: 50 })

SELECT * **FROM** users **WHERE** status = "A" **OR** age = 50	**db**.**users**.**find**({ **$or**: [{ **status**: "A" } , { **age**: 50 }] })
SELECT * **FROM** users **WHERE** age > 25	**db**.**users**.**find**({ **age**: { **$gt**: 25 } })
SELECT * **FROM** users **WHERE** age < 25	**db**.**users**.**find**({ **age**: { **$lt**: 25 } })
SELECT * **FROM** users **WHERE** age > 25 **AND** age <= 50	**db**.**users**.**find**({ **age**: { **$gt**: 25, **$lte**: 50 } })
SELECT * **FROM** users **WHERE** user_id **like** "%bc%"	**db**.**users**.**find**({ **user_id**: /bc/ })
SELECT * **FROM** users **WHERE** user_id **like** "bc%"	**db**.**users**.**find**({ **user_id**: /^bc/ })
SELECT * **FROM** users **WHERE** status = "A" **ORDER BY** user_id **ASC**	**db**.**users**.**find**({ **status**: "A" }).**sort**({ **user_id**: 1 })
SELECT * **FROM** users **WHERE** status = "A" **ORDER BY** user_id **DESC**	**db**.**users**.**find**({ **status**: "A" }).**sort**({ **user_id**: -1 })
SELECT COUNT(*) **FROM** users	**db**.**users**.**count**() *or* **db**.**users**.**find**().**count**()

Anexo 5 Consultas SQL - MongoDB

Comparativo entre actualizaciones SQL - MongoDB

SQL Update Statements	**MongoDB update() Statements**
UPDATE users **SET** status = "C"	**db**.**users**.**update**({ **age**: { **$gt**: 25 } },

WHERE age > 25	{ **$set**: { **status**: "C" } }, { **multi**: **true** })
UPDATE users **SET** age = age + 3 **WHERE** status = "A"	**db.users.update**({ **status**: "A" } , { **$inc**: { **age**: 3 } }, { **multi**: **true** })

Anexo 6 Actualizaciones SQL - MongoDB

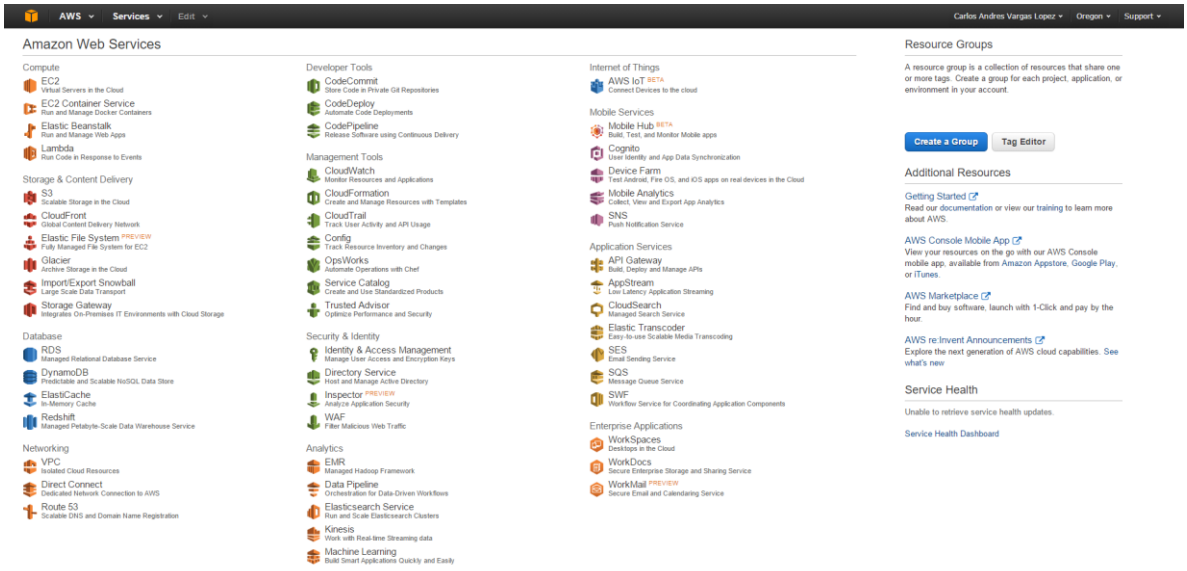
Comparativo entre eliminaciones SQL – MongoDB

SQL Delete Statements	MongoDB remove() Statements
DELETE FROM users **WHERE** status = "D"	**db.users.remove**({ **status**: "D" })
DELETE FROM users	**db.users.remove**({})

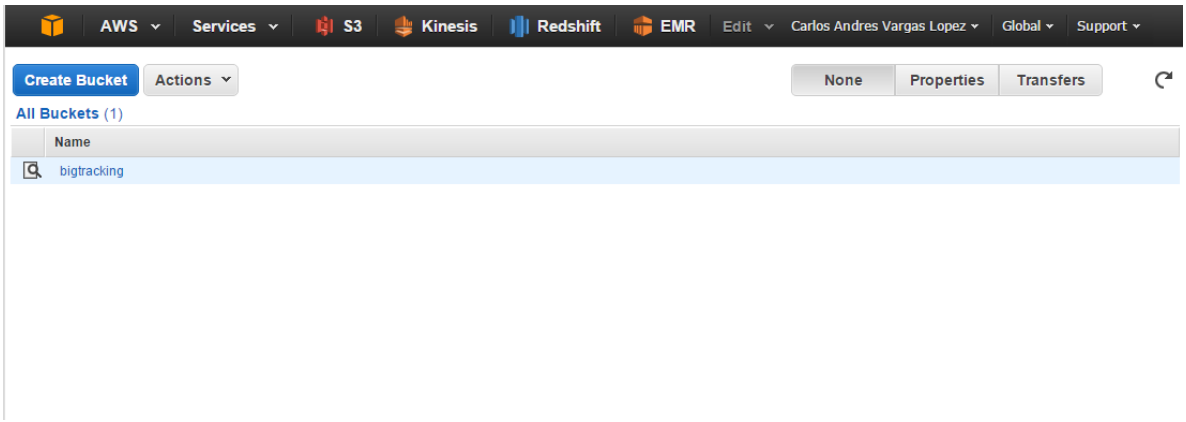
Anexo 7 Eliminaciones SQL – MongoDB

6.2 Configuración Ambiente AWS

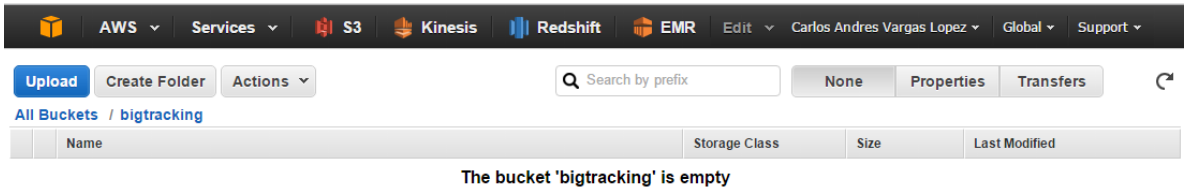
Consola AWS (Managment)



Anexo 8 Amazon S3 (Simple Storage Service) I

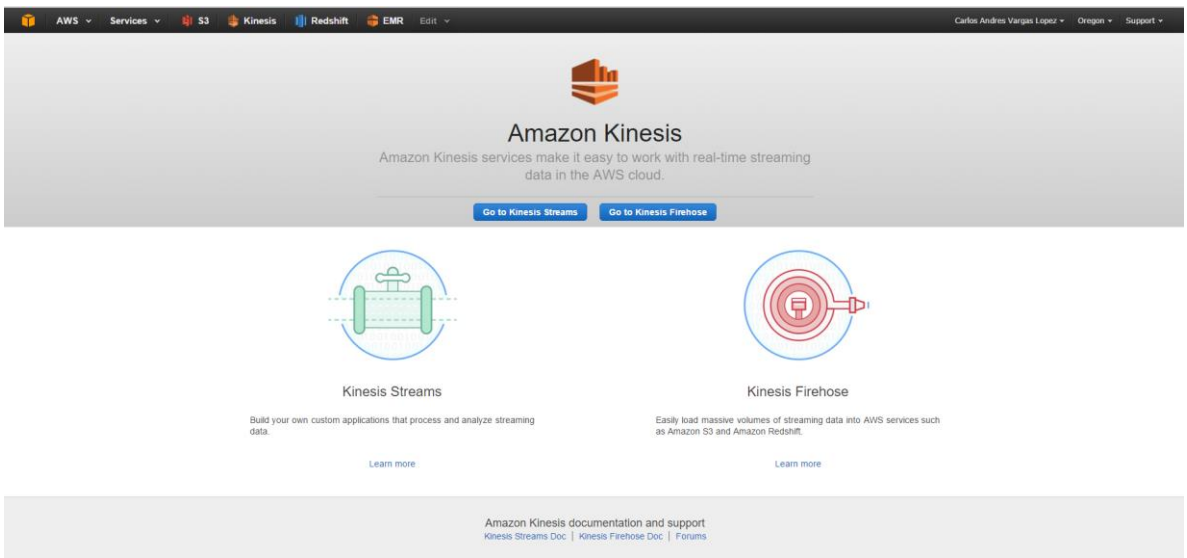


Anexo 9 Anexo 8 Amazon S3 (Simple Storage Service) II

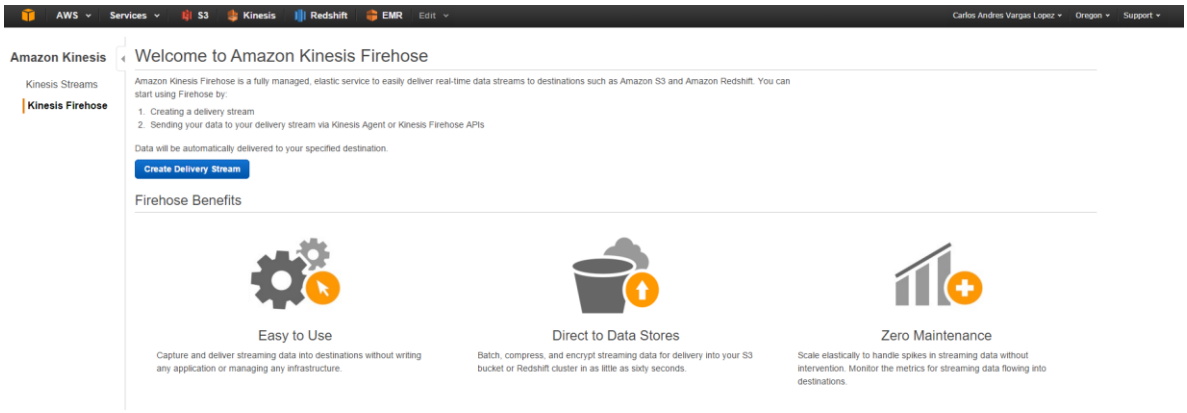


Anexo 10 Anexo 8 Amazon S3 (Simple Storage Service) III

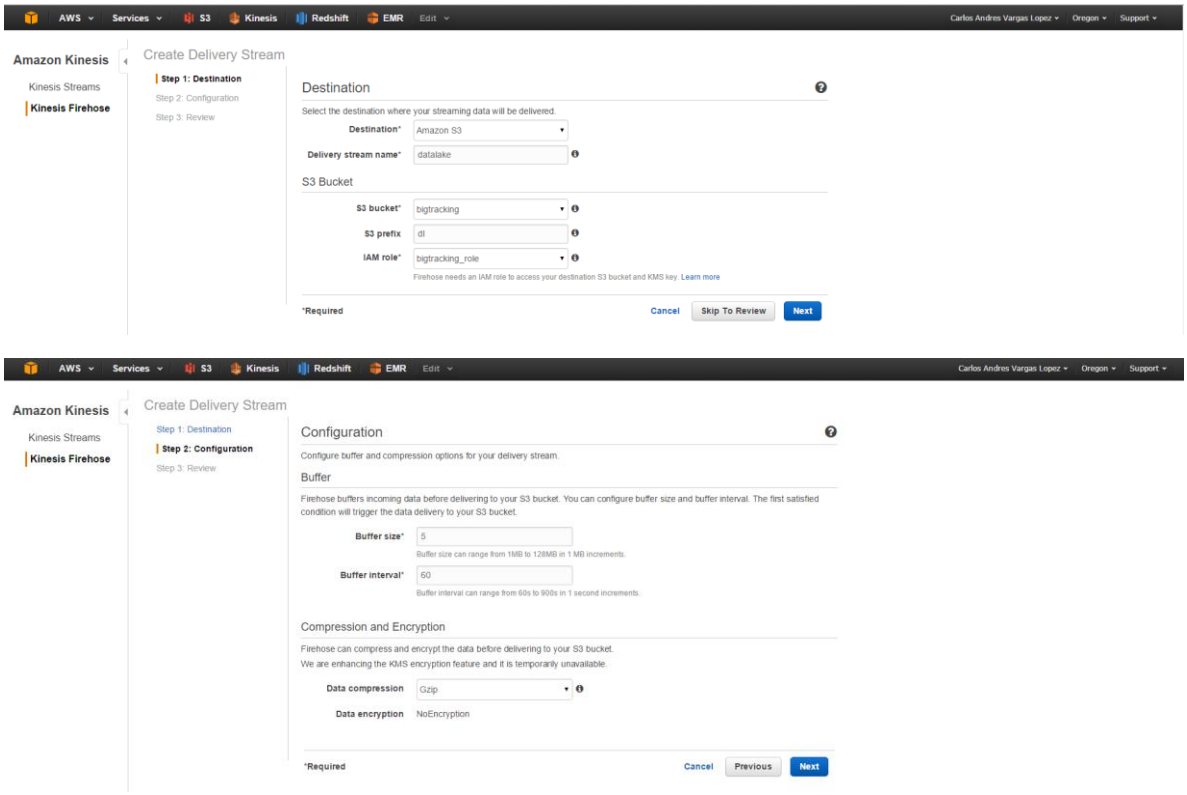
Amazon Kinesis (Facilitador del trabajo con transmisiones de datos en tiempo real en AWS)



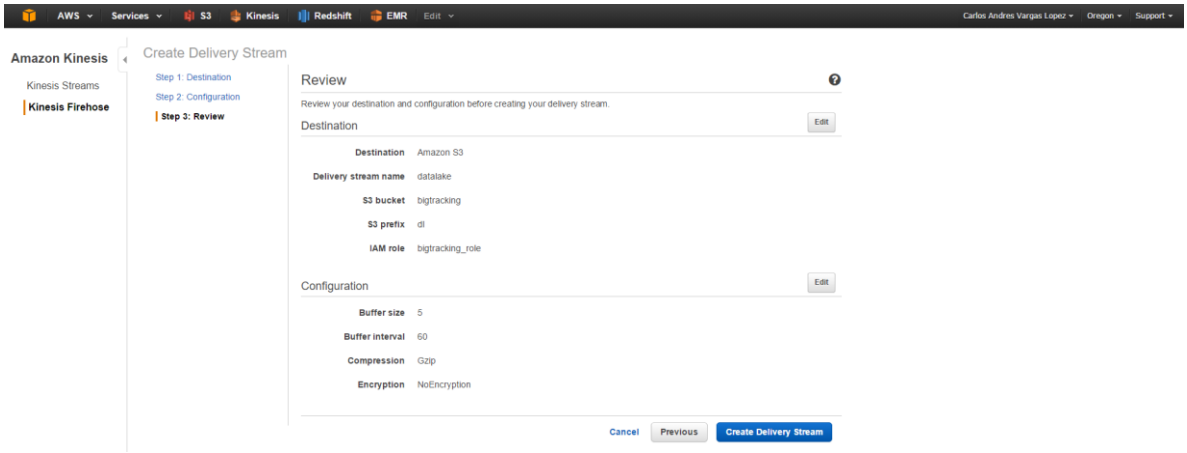
Anexo 11 Amazon Kinesis Firehose I



Anexo 12 Amazon Kinesis Firehose II

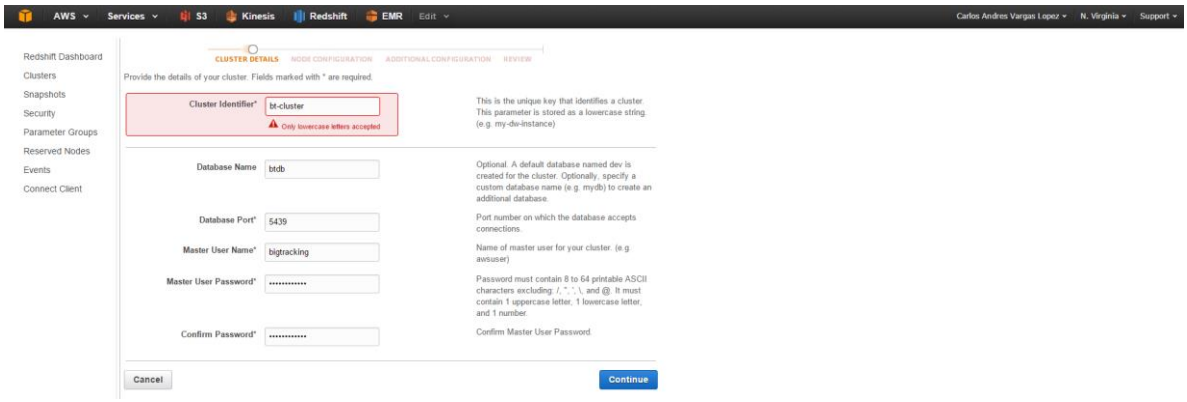


Anexo 13 Amazon Kinesis Firehose III

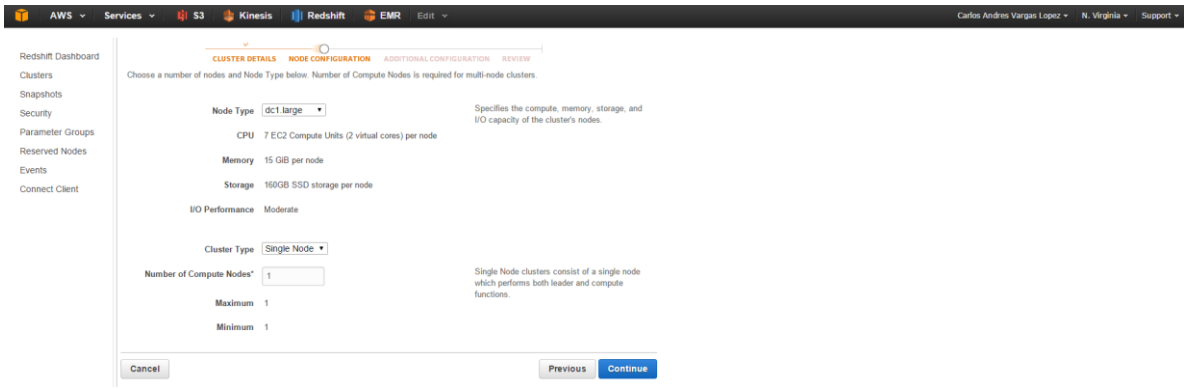


Anexo 14 Amazon Kinesis Firehose IV

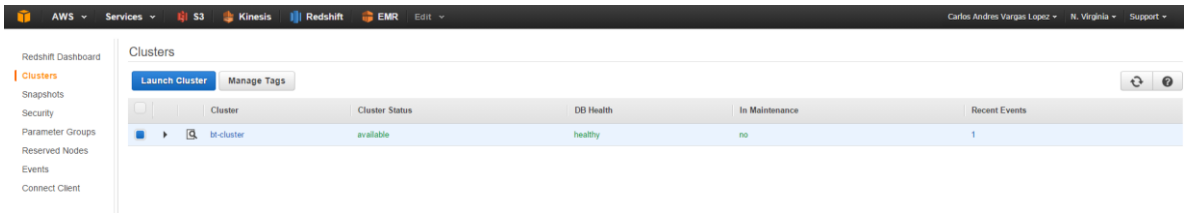
Amazon Redshift (Data Warehouse / Clusters)



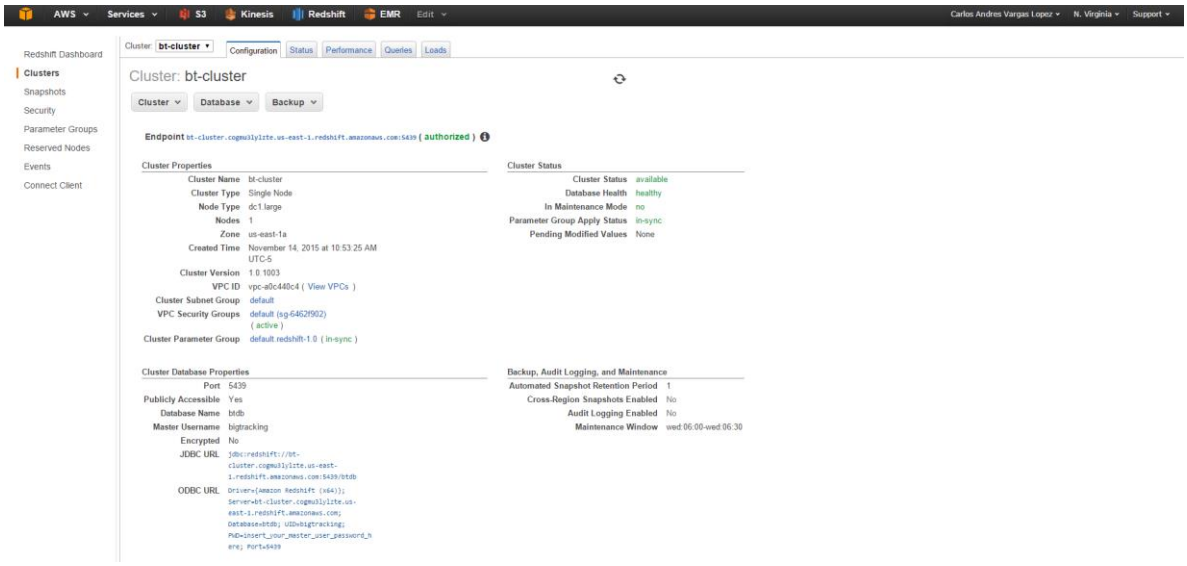
Anexo 15 Amazon Redshift (Data Warehouse / Clusters) I



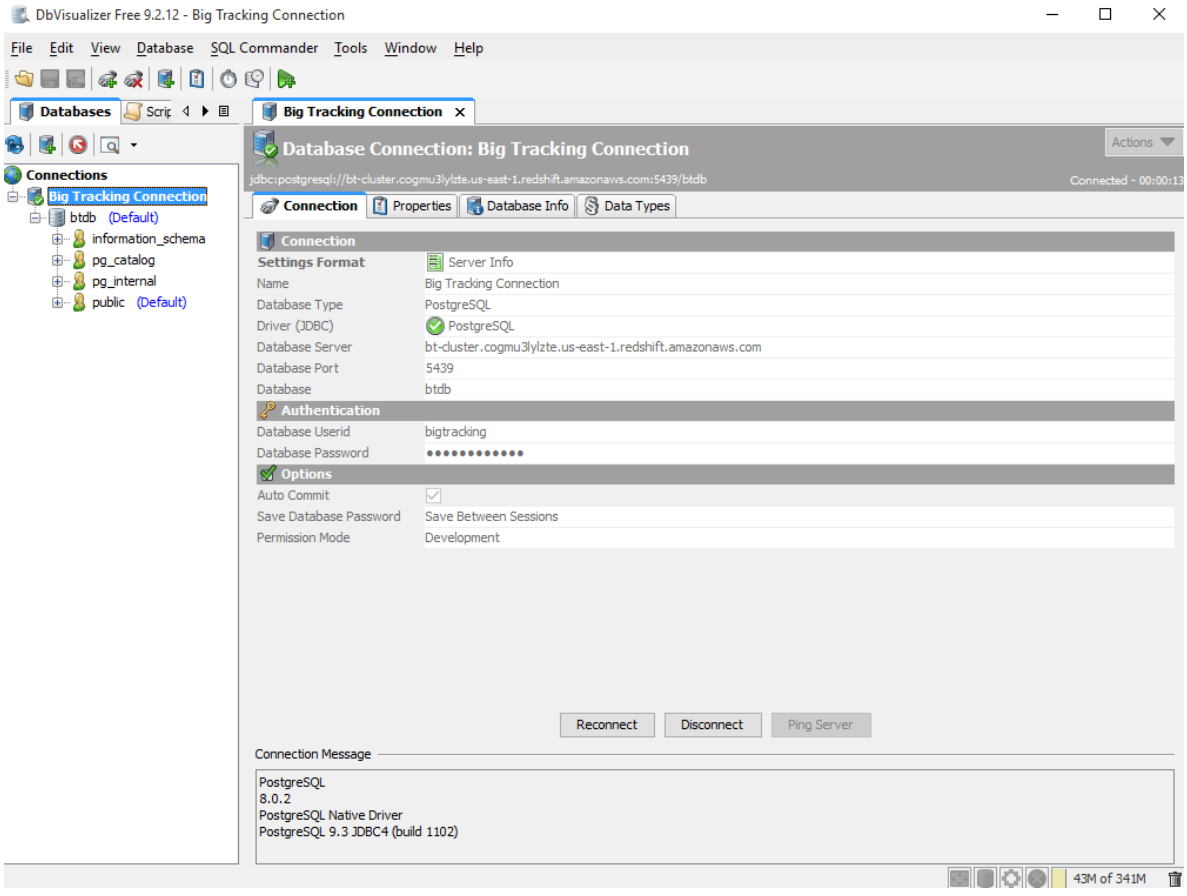
Anexo 16 Amazon Redshift (Data Warehouse / Clusters) II



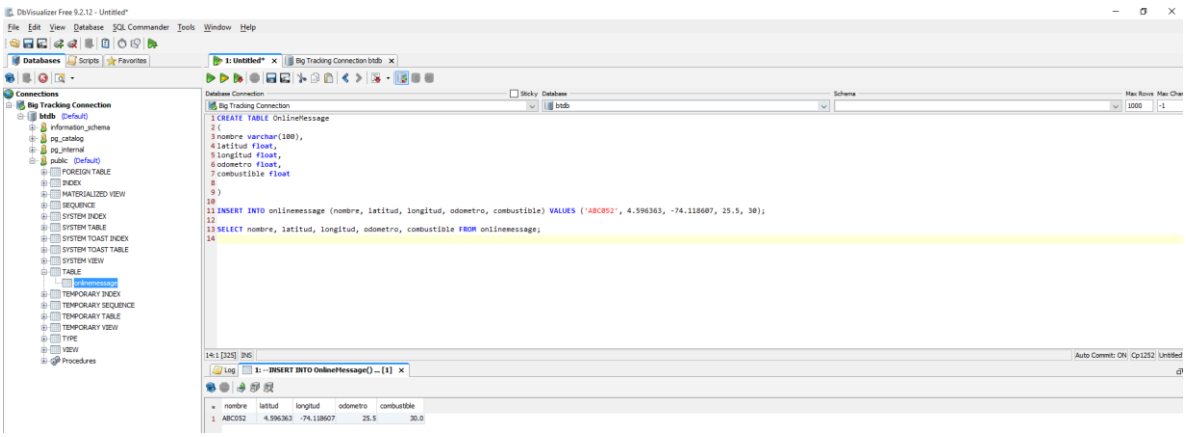
Anexo 17 Amazon Redshift (Data Warehouse / Clusters) III



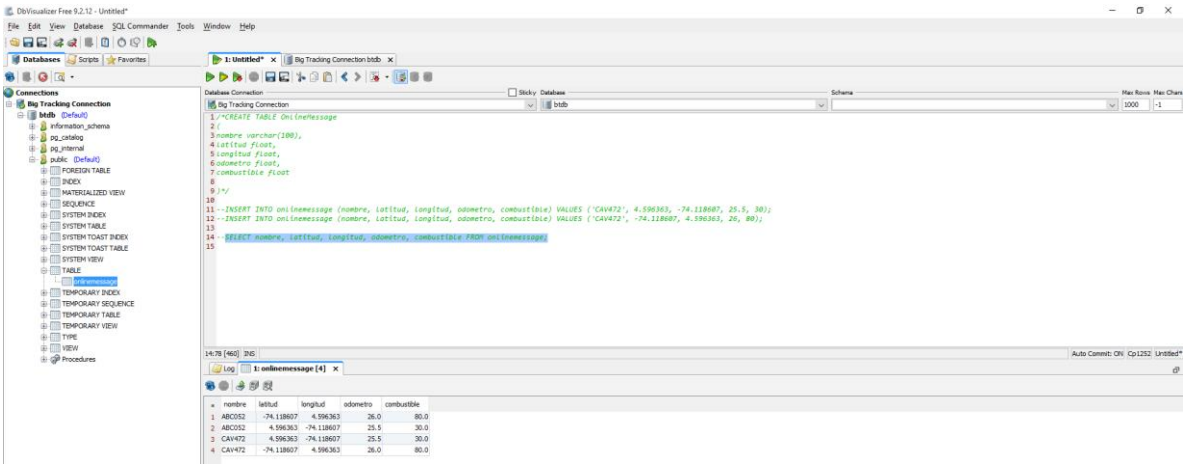
Anexo 18 Amazon Redshift (Data Warehouse / Clusters) IV



Anexo 19 DbVisualizer (Conexión a DB Amazon) I

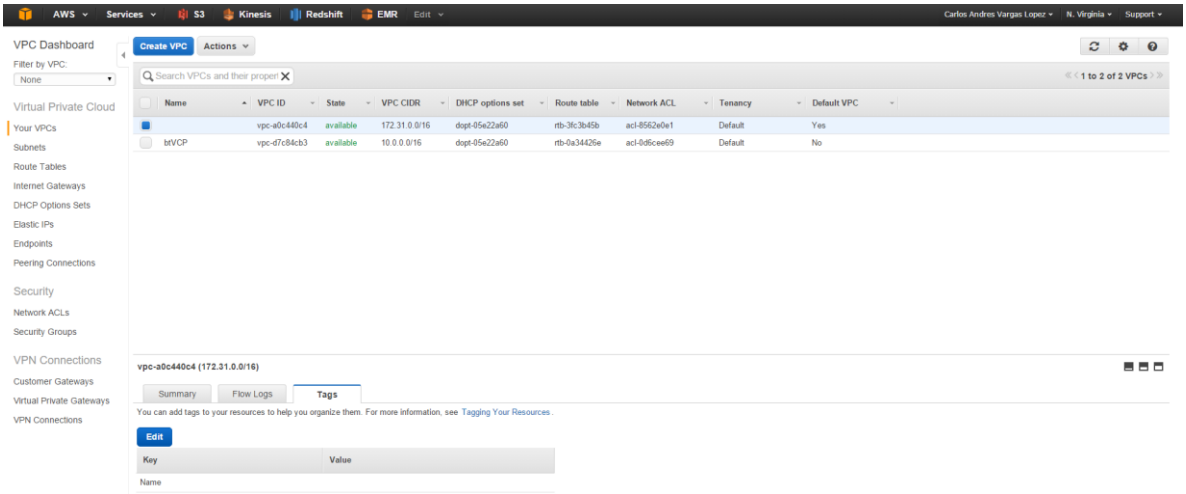


Anexo 20 DbVisualizer (Conexión a DB Amazon) II



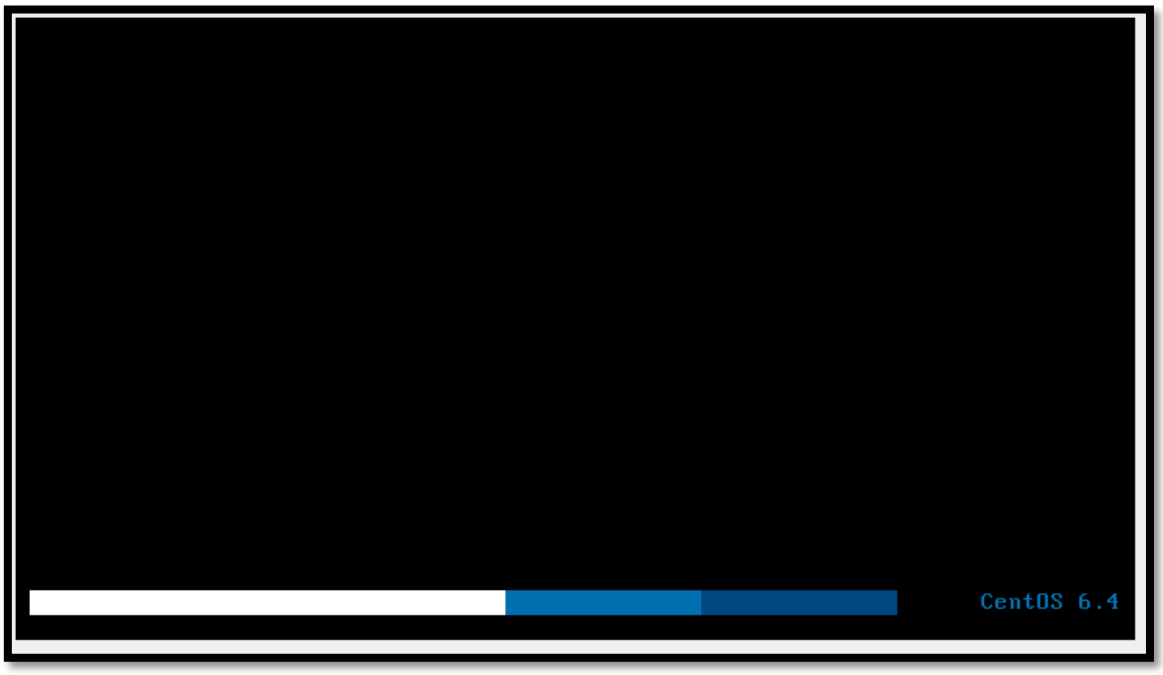
Anexo 21 DbVisualizer (Conexión a DB Amazon) III

Amazon VPC (SSL)

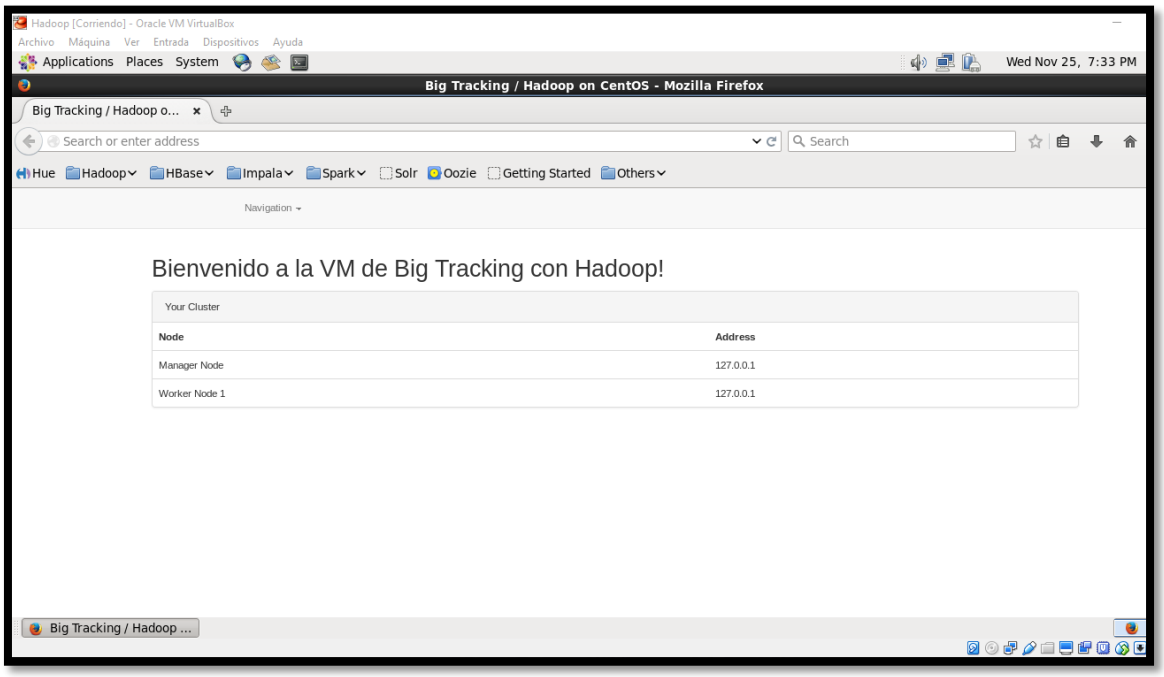


Anexo 22 Amazon VPC (SSL)

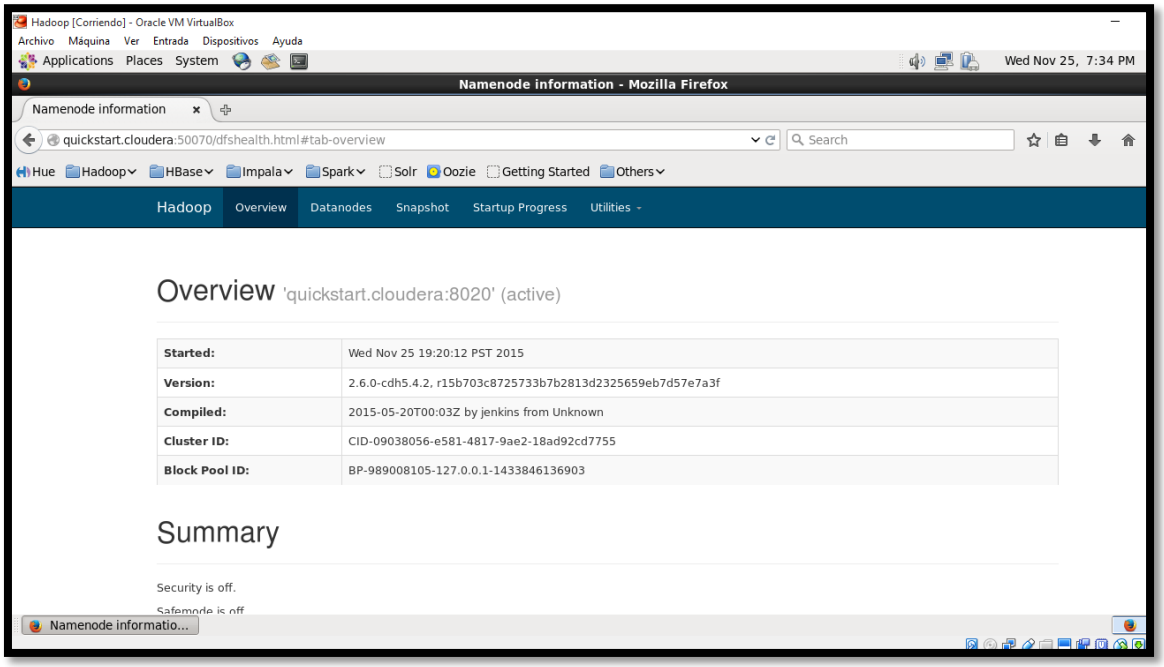
6.3 Ambiente Hadoop en VirtualBox



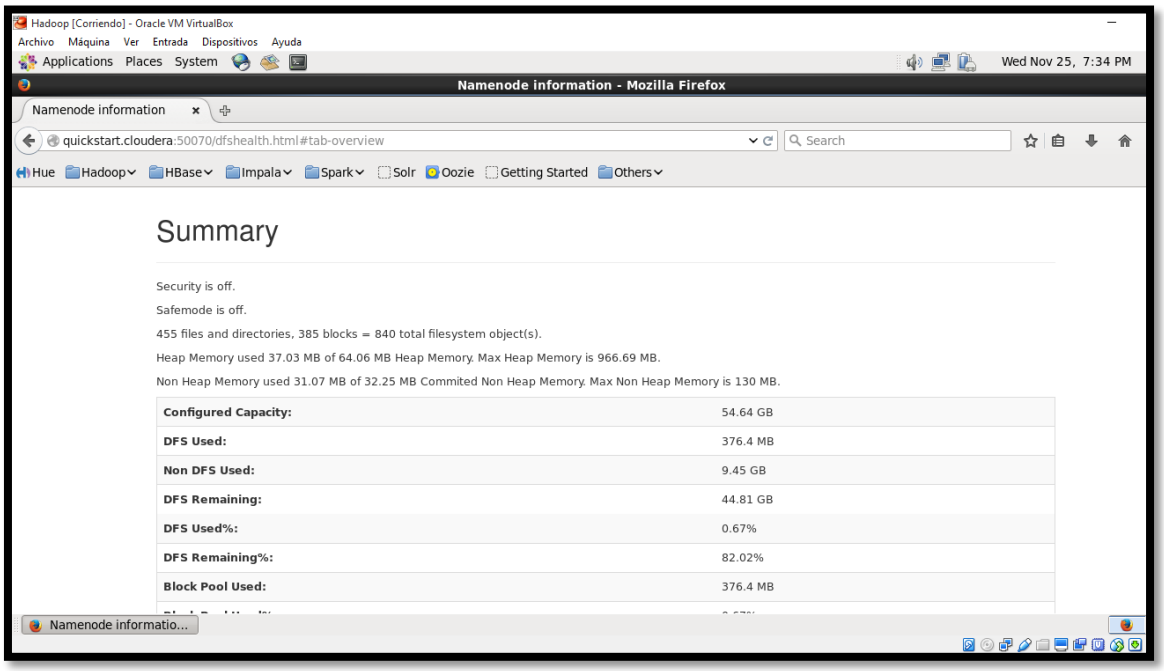
Anexo 23 Iniciando la máquina virtual en VirtualBox



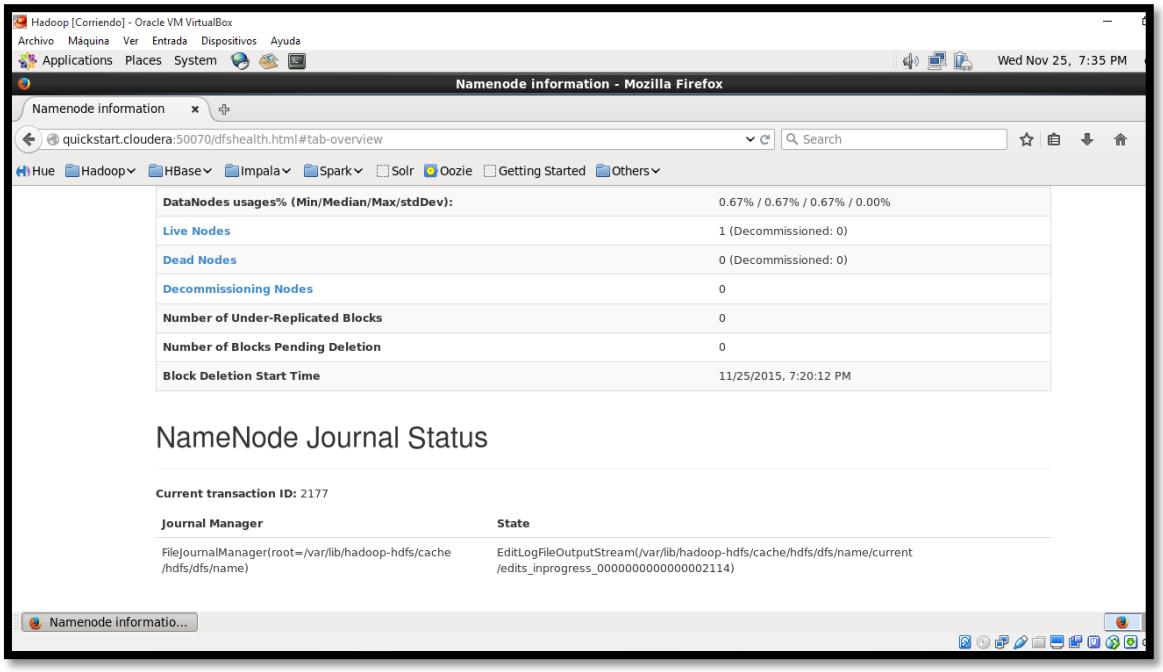
Anexo 24 Página principal de Hadoop



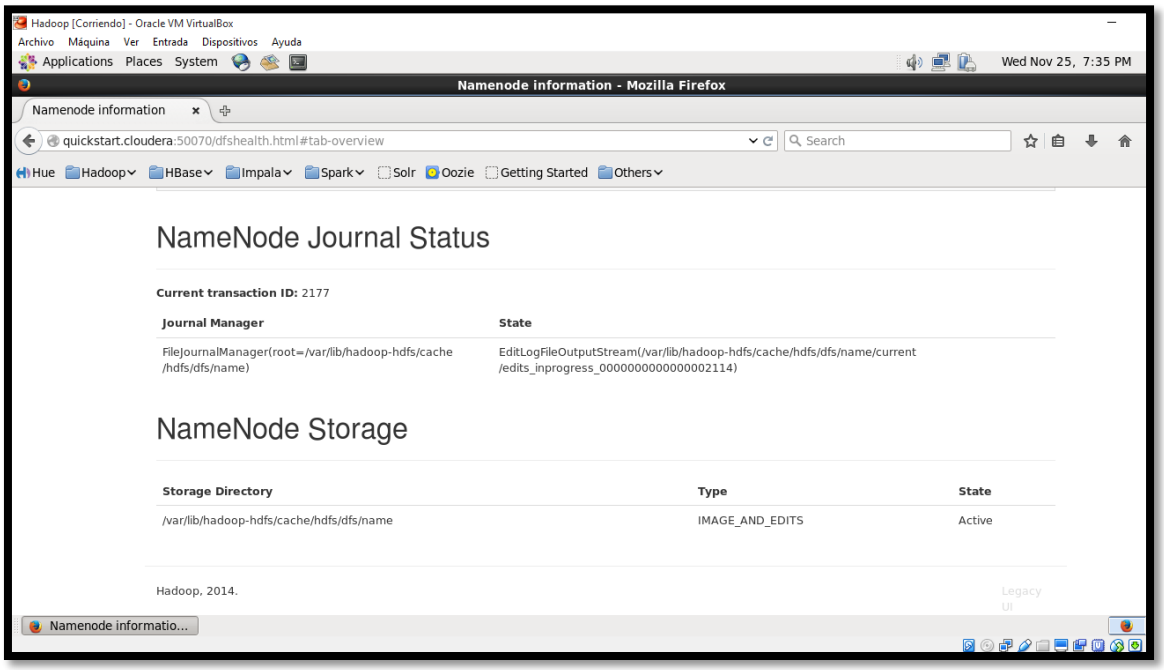
Anexo 25 Visión general de Hadoop



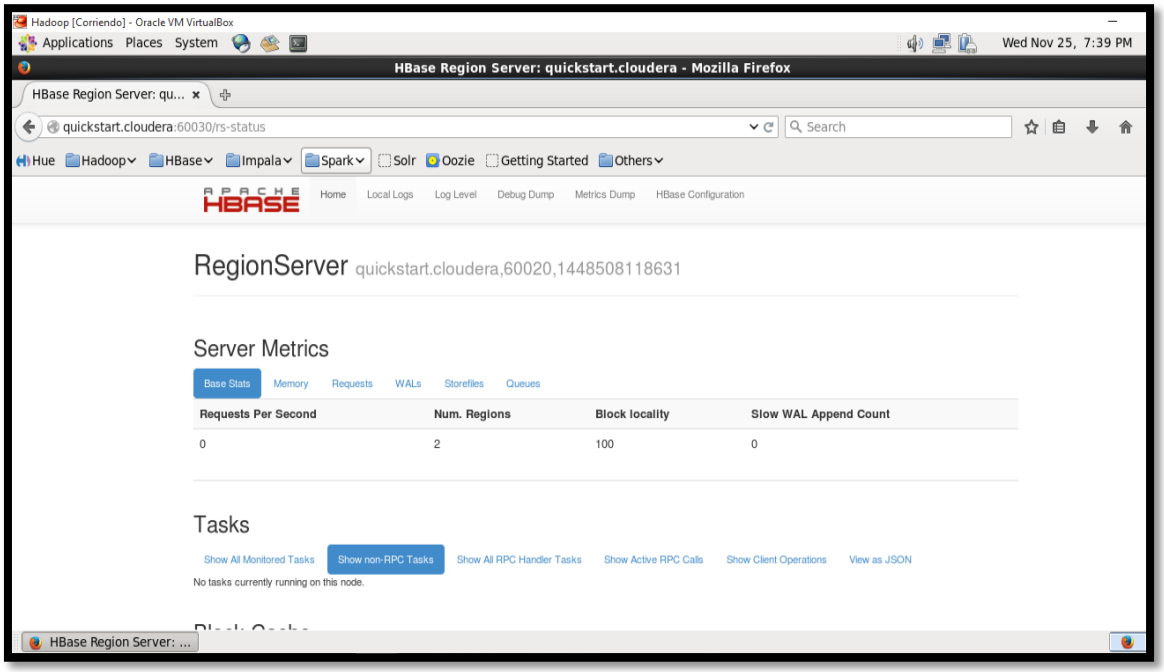
Anexo 26 Resumen de Hadoop I



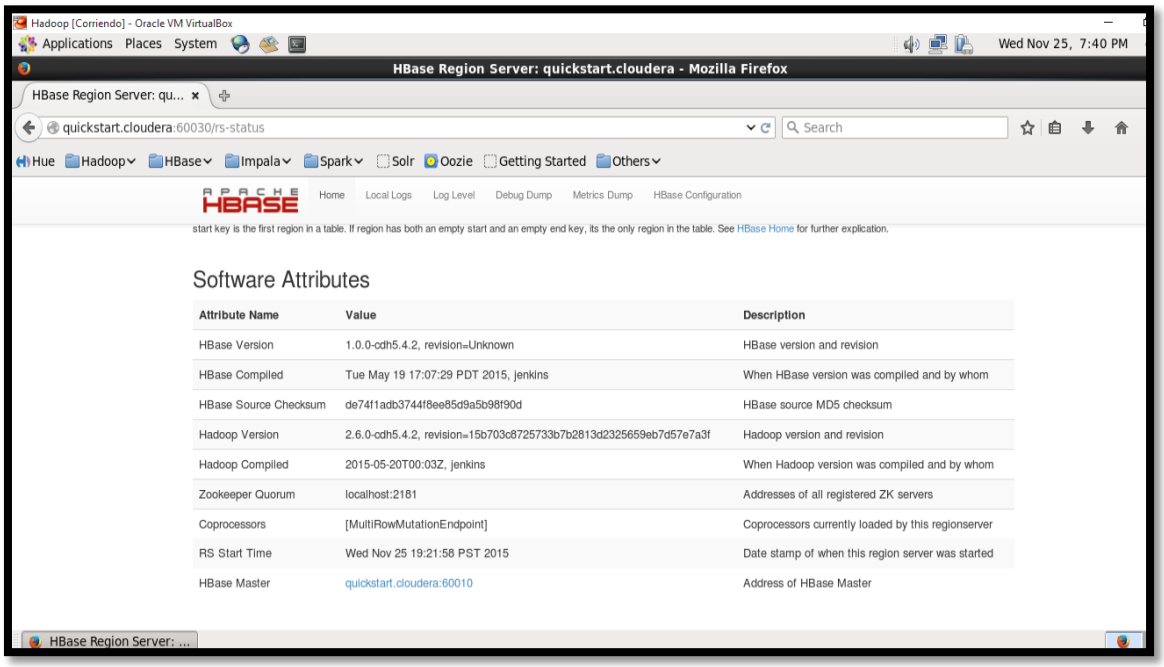
Anexo 27 Resumen de Hadoop II



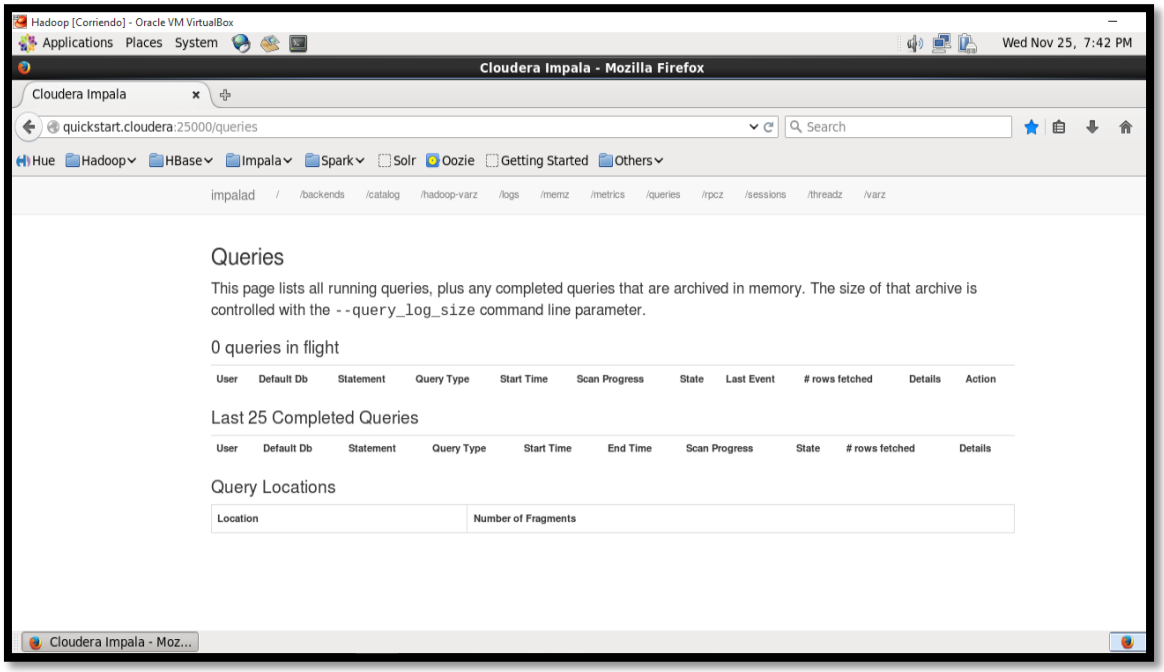
Anexo 28 Resumen de Hadoop III



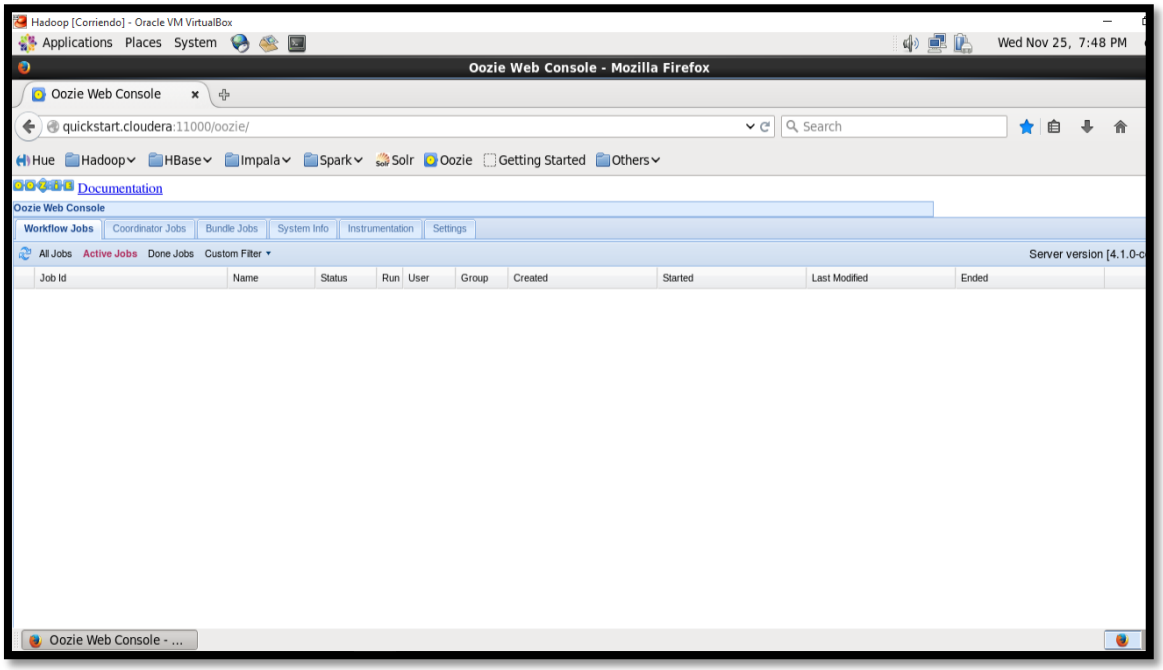
Anexo 29 HBase Hadoop, motor NoSQL.



Anexo 30 Atributos de HBase



Anexo 31 Lista de Querys en HBase



Anexo 32 Página principal de Oozie

Bibliografía Y Referencias

[1] «IDC,» 15 7 2015. [En línea]. Available: https://www.idc.com/.

[2] «EDC,» 15 7 2015. [En línea]. Available: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

[3] «Dinero - Big Data,» 20 9 2015 . [En línea]. Available: http://www.dinero.com/imprimir/210853.

[4] M. Korolov, «cioperu.pe,» 100 08 2015. [En línea]. Available: http://cioperu.pe/fotoreportaje/13833/los-15-grandes-del-big-data/.

[5] «Big Data Five Up,» 15 07 2015. [En línea]. Available: https://bigdatafiveup.wordpress.com/2014/06/01/big-data-en-colombia/.

[6] D. Abiertos, «Wikipedia,» 15 09 2015. [En línea]. Available: https://es.wikipedia.org/wiki/Datos_abiertos.

[7] S. Adler, «youtube,» 02 09 2015. [En línea]. Available: https://www.youtube.com/watch?v=CpLtceuujhs.

[8] S. Adler, «IBM - Open Data,» 25 2015 09. [En línea]. Available: http://www.ibmbigdatahub.com/blog/injecting-open-data-fight-against-ebola.

[9] «Trello,» 15 07 2015. [En línea]. Available: http://trello.com/.

[10] «IBM - Cuatro dimensiones de Big Data,» 15 10 2015. [En línea]. Available: http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf.

[11] «Data IQ,» 28 09 2015. [En línea]. Available: http://dataiq.com.ar/blog/por-que-es-importante-big-data/.

[12] «rhernado - Almacenes de Datos,» 17 10 2015. [En línea]. Available: http://www2.rhernando.net/modules/tutorials/doc/bd/dw.html.

[13] «msdn.microsoft - Mineria de Datos,» 25 10 2015. [En línea]. Available: https://msdn.microsoft.com/es-es/library/ms174949%28v=sql.120%29.aspx.

[14] «Msdn Microsoft - Algoritmos,» 30 10 2015. [En línea]. Available: https://msdn.microsoft.com/es-es/library/ms174949%28v=sql.120%29.aspx.

[15] «Cran -R,» 5 11 2015. [En línea]. Available: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf.

[16] «Sintetia - R,» 5 10 2015. [En línea]. Available: http://www.sintetia.com/wp-content/uploads/2012/03/r_project.png.

[17] F. Tanco, «dariolara,» 20 11 2013. [En línea]. Available: http://www.dariolara.com/tda/tds/RNA.pdf. [Último acceso: 20 11 2015].

[18] «commons.wikimedia.org,» Wikipedia, 1 6 2014. [En línea]. Available: https://commons.wikimedia.org/wiki/File:Computer.Science.AI.Neuron.svg. [Último acceso: 21 11 2015].

[19] «commons.wikimedia.org,» WIKIPEDIA, 14 12 2014. [En línea]. Available: https://commons.wikimedia.org/wiki/File:RedNeuronalArtificial.png. [Último acceso: 20 11 2014].

[20] «msdn.microsoft - Redes Neuronales,» 23 10 2015. [En línea]. Available: https://msdn.microsoft.com/es-es/library/ms174941(v=sql.120).aspx.

[21] «FRRO - Redes Neuronales,» 20 10 2015. [En línea]. Available: http://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf.

[22] «POLI - Reconocimineto Facial,» 10 11 2015. [En línea]. Available: http://repository.poligran.edu.co/bitstream/10823/600/1/Reconocimiento%20facial.pdf.

[23] «Technet Microsoft - SQL Server,» 30 10 2015. [En línea]. Available: https://technet.microsoft.com/es-es/library/ms172445(v=sql.105).aspx.

[24] «Photobucket - SQL Server,» 5 11 2015. [En línea]. Available: http://i368.photobucket.com/albums/oo127/analitico_bucket/Img%20Soft/diagram-sql2008-lg.gif.

[25] «Docs Mongodb - MongoDB,» 7 11 2015. [En línea]. Available: https://docs.mongodb.org/ecosystem/use-cases/Hadoop/.

[26] «IBMCloudant,» 15 10 2015. [En línea]. Available: https://www-03.ibm.com/press/mx/es/pressrelease/43317.wss.

[27] «RedBooks - Cloudant,» 21 10 2015. [En línea]. Available: http://www.redbooks.ibm.com/technotes/tips1187.pdf.

[28] K. T. A. Y. Boris Lublinsky, Hadoop Soluciones Big Data, España: ANAYA MULTIMEDIA, 2014.

[29] I. -. Hadoop, «IBM,» [En línea]. Available: http://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209Hadoopbigdata.

[30] M. Lurie, «IBM,» 06 10 2015. [En línea]. Available: http://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209Hadoopbigdata/.

[31] «SCI - Productos,» 14 10 2015. [En línea]. Available: http://www.sciufm.com/index.html.

[32] «DELL,» 1 11 2015. [En línea]. Available: http://www.dell.com/us/business/p/poweredge-r630/pd.

[33] «DEL,» 1 11 2015. [En línea]. Available: http://www.dell.com/downloads/global/products/pedge/r710-spec-sheet.pdf.

[34] HP, «hp,» hp, 1 1 2015. [En línea]. Available: http://www8.hp.com/pr/es/products/proliant-servers/product-detail.html?oid=8261835. [Último acceso: 25 11 2015].

[35] HP, «HP,» 1 10 2015. [En línea]. Available: http://h20195.www2.hp.com/v2/GetDocument.aspx?docname=4AA6-1292ENW&doctype=Technical%20white%20paper&doclang=EN_US&searchquery=&cc=pr&lc=es. [Último acceso: 25 11 2015].

[36] U. S. B. P. B. M. E. B. P. U. D. U. o. P. D. H. F. U. B. G. C. U. H. I. H. G. Divyakant Agrawal, *Challenges and Opportunities with Big Data,* United States: collaborative, 2012.

[37] MongoDB, «MongoDB Manual,» 29 09 2015. [En línea]. Available: http://docs.mongodb.org/manual/reference/sql-comparison/.

[38] «Hadoop,» 15 07 2015. [En línea]. Available: https://Hadoop.apache.org/docs/stable/.

[39] «Wikipedia,» 15 07 2015. [En línea]. Available: http://es.wikipedia.org/wiki/Big_data.

[40] «Vision Software,» 15 07 2015. [En línea]. Available: http://www.visionsoftware.com.co/big-data-relevante/.

[41] R. E. Walpole, Probabilidad & Estadística para ingeniería & ciencias, Pearson, 2006.

[42] «Nytimes,» 15 07 2015. [En línea]. Available: http://www.nytimes.com/2013/06/11/books/big-data-by-viktor-mayer-schonberger-and-kenneth-cukier.html?_r=1.

[43] «NoSql,» 15 07 2015. [En línea]. Available: http://www.nosql.es/blog/nosql/mapreduce.html.

[44] MongoDB , «MongoDB Manual,» 15 07 2015. [En línea]. Available: http://docs.mongodb.org/manual/.

[45] M. -. M. d. Datos, «MSDN Microsoft,» 28 09 2015. [En línea]. Available: https://msdn.microsoft.com/es-es/library/ms174949%28v=sql.120%29.aspx.

[46] I. Castillo, Estadística descriptiva y cálculo de probalidades, Madrid: Pearson, 2005.

[47] «IDC,» 15 07 2015. [En línea]. Available: https://www.idc.com/.

[48] «Hadoop,» 15 07 2015. [En línea]. Available: https://Hadoop.apache.org/docs/stable/.

[49] «EMC,» 22 07 2015. [En línea]. Available: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

[50] «Coursera,» 15 07 2015. [En línea]. Available: https://es.coursera.org/course/rprog.

[51] «comScore,» 15 07 2015. [En línea]. Available: http://www.comscore.com/lat/Insights/Presentations-and-Whitepapers/2014/What-is-Big-Data-and-why-is-it-important.

[52] «America Economia,» 15 07 2015. [En línea]. Available: http://tecno.americaeconomia.com/articulos/5-razones-de-por-que-big-data-y-la-analitica-son-importantes-para-tu-empresa.