MODELO DE EVALUACIÓN DE PLATAFORMAS TECNOLÓGICAS PARA EL ALMACENAMIENTO Y PROCESAMIENTO DE GRANDES DATOS CASO EN SALUD ASISTENCIAL

MODEL OF EVALUATION OF TECHNOLOGICAL PLATFORMS FOR THE STORING AND PROCESSING OF BIG DATA - HEALTH CARE CASE

Aspirante

ELKIN ANDRES VILLA SANCHEZ, PMP©

Asesores

EDWIN NELSON MONTOYA M, PHD

UNIVERSIDAD EAFIT

ESCUELA DE INGENIERÍA

DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS

MAESTRÍA EN INGENIERÍA

JUNIO 2018

TABLAS

FIGURAS

GLOSARIO

A

Alta disponibilidad: Es una característica del sistema que asegura una continuidad operacional durante un periodo de tiempo determinado.

API: Significa interfaz de programación de aplicaciones. Es una interfaz que define la manera de trabajar con distintos componentes de programación.

B

Base de datos: Es un conjunto de datos que pertenecen a un mismo contexto y están almacenados en un mismo sistema con la intención de ser usados repetidas veces.

Base de datos relacional: Es una base de datos que cumple con el modelo relacional.

Batch: Es una ejecución de una serie de trabajos o procesos informáticos que no requieren de la intervención manual para iniciarse.

Bit: Unidad básica de información en la informática y las comunicaciones digitales.

Byte: Unidad de información digital formada por una secuencia de ocho bits.

C

Clúster: Conjunto de computadores interconectados que actúan como si fueran uno solo.

Consola de comandos: También conocida como línea de comandos, es una herramienta que permite a los usuarios de una aplicación darle instrucciones a través de comandos de texto sencillos.

D

Dataflow o Flujo de información: Es el diseño de una ejecución de una serie de procesos que se comunican enviándose información a través de distintos canales.

Data warehouse: Es una colección de datos orientada a un ámbito empresarial u organizativo y almacenado en un sistema no volátil e integrado que ayuda a la toma de decisiones.

Distribuido: La computación distribuida es un modelo de computación en la que se usan una serie de computadores organizados en clústeres.

Double: Tipo básico de número racional en coma flotante de doble precisión.

E

Escalabilidad: Es una propiedad de un sistema que indica su capacidad de reacción y de adaptación a los cambios de envergadura, ya sean al crecer o disminuir.

ETL: Proceso de transformación de datos realizado para extraer datos de una fuente y almacenarlos en una base de datos o data warehouse.

F

Framework: Conjunto de conceptos y tecnologías que sirven de base para el desarrollo de aplicaciones. Acostumbra a incluir bibliotecas de software, lenguajes, soportes a aplicaciones de desarrollo.

H

Hash: Función que realiza una transformación normalmente de reducción a su entrada, dejando una salida de valor hash de tamaño finito.

Hardware almacenamiento: Unidades de almacenamiento secundario, principalmente discos duros, discos compactos, cintas magnéticas etc.

Hadoop: Es un sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos.

HDFS: Es un sistema de archivo distribuido que permite que los archivos de datos no se guarden en una única máquina, sino que la información sea distribuida.

L

Log File: Es un archivo que almacena un conjunto de entradas que informan de los distintos eventos, cambios en los estados de los procesos, comunicaciones, errores que se producen en un computador.

M

Modelo relacional: Es un modelo de datos basado en la lógica de predicados y en la teoría de conjuntos. Está compuesto fundamentalmente por tablas y se basa en el uso de las relaciones entre estas.

N

NoSQL: Es un sistema de gestión de bases de datos que ofrecen una alternativa al modelo relacional tradicional en aspectos muy importantes, destacando la no utilización de SQL.

O

Open Source: Hace referencia al código distribuido y desarrollado libremente, de manera abierta a todo el mundo, dando acceso al código fuente del proyecto.

Outsourcing: Es el proceso empresarial de destinar recursos a contratar una empresa externa para realizar ciertas tareas determinadas.

P

Phishing: Ataque informático que consiste en suplantar la identidad de una persona o una entidad para engañar al atacado.

R

Repositorio: Servidor centralizado donde se almacena y se mantiene la información digital como bases de datos, aplicaciones o código para poder ser accedidos remotamente.

RFID: Sistema de almacenamiento y recuperación de pequeñas cantidades de datos que se utiliza mediante etiquetas RFID y que generalmente se usan para la identificación.

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

S

Script: Archivo de texto plano que contiene una serie de comandos simples o códigos sencillos en lenguajes interpretados cuyo objetivo es generalmente el de realizar tareas de orquestación de procesos o monitorización.

Sensor: Dispositivo capaz de detectar eventos físicos cambios de luz, movimiento, temperatura y convertirlo en información analógica o digital.

Sistema gestor de bases de datos (SGBD) Conjunto de programas que permite el almacenamiento, modificación y extracción de la información en una base de datos. También incluyen métodos de administración y monitorización.

SQL: Lenguaje de consultas estructuradas para bases de datos relacionales.

Streaming: Distribución de datos de manera constante en forma de flujo continúo sin interrupción usada por ejemplo en la transmisión de contenido multimedia a través de Internet.

V

Virtualización: Simular un entorno físico de hardware mediante software dentro de otro entorno físico real.

W

Wordcount: Programa que cuenta el número de palabras de texto determinado.

CONTENIDO

RESUMEN

Big Data puede ser considerada como una tendencia en el avance de la tecnología del almacenamiento de grandes datos y procesamiento paralelo masivo que ha abierto la puerta a un nuevo enfoque para la comprensión y la toma de decisiones mediante la analítica avanzada de los datos, que utiliza grandes cantidades de datos (estructurados, no estructurados y semi-estructurados) que sería demasiado costoso o se tendrían limitaciones tecnológicas para cargar y procesar en una base de datos relacional para su análisis. Así, el concepto de Big Data se aplica a toda la información que no puede ser procesada o analizada utilizando herramientas o procesos tradicionales. En términos generales, Big Data y los procesos que dicha técnica representa tiene un amplio espectro de aplicaciones potenciales. El mayor desafío para la inversión en Big Data se produce con relación a los proyectos vinculados a la toma de decisiones sobre una gran cantidad de datos, definición de estrategias y la obtención de mejores experiencias sobre los actos de consumo de las personas. El desafío de Big Data consiste en capturar, almacenar, buscar, compartir y agregar valor a los datos poco utilizados o inaccesibles hasta la fecha. No es relevante el volumen de datos o su naturaleza; sino lo que importa es su valor potencial, que sólo las nuevas tecnologías especializadas en Big Data pueden explotar.

El objetivo de esta tecnología es aportar y descubrir un conocimiento oculto a partir de grandes volúmenes de datos.

En este trabajo se presenta un estudio aplicando a dos diferentes tipos de almacenamientos y procesamiento de altos volúmenes de datos, específicamente para datos clínicos estructurados. Un almacenamiento/procesamiento en infraestructura tradicional de base de datos SQL y el otro al almacenamiento/procesamiento en una infraestructura Big Data en Hadoop/Spark.

Los experimentos se han llevado a cabo utilizando el ecosistema Hadoop y Apache Spark y demuestran que esta combinación de capas permite realizar operaciones de analítica descriptiva sobre grandes volúmenes de datos y disponer de resultados realizados en tiempo real, con calidad comparable a trabajos similares sobre contextos no Big Data.

1. PROPUESTA DE INVESTIGACIÓN

1.1. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN

El sistema de salud en Colombia está regulado por el gobierno nacional, por intermedio del Ministerio de la Salud y Protección Social bajo mandato constitucional y delegado 100, expedida en 1993, la cual reglamenta el sistema general de seguridad social.

El porcentaje de población asegurada en Colombia ha aumentado significativamente en los últimos ocho años.

De otro lado, uno de los principales retos que deberá enfrentar el presidente, será el de buscar soluciones a la crisis financiera del sistema de salud. Esos ajustes serán costosos, ya que deberán liquidarse o fusionarse gran parte de las Empresas Promotoras de Salud, EPS, y de paso habrá que inyectarle millonarios recursos a la red hospitalaria para atender sus deudas en los próximos meses. Las quejas de los usuarios, el cierre de centros hospitalarios, la insostenibilidad que reclaman algunos aseguradores y el estancamiento de los recursos han desatado la controversia sobre si cambiar el actual sistema de salud en Colombia en su totalidad o realizar una reestructuración de fondo.

El sector salud colombiano vive un momento difícil. Son comunes las quejas por la mala calidad de la atención, por la no prestación de servicios, el no pago de los salarios al personal médico, el colapso de las salas de urgencias, la cantidad de trámites que hay que realizar para acceder a procedimientos de alto costo, las enormes deudas de las EPS con las IPS e incluso, los problemas de sanidad en las cárceles del país. Detrás de estas situaciones se esconde un profundo desequilibrio financiero del sistema, considerables inconvenientes con la gestión de los recursos y transparencia en su administración, falta de claridad en cuanto a las prestaciones que los recursos públicos realmente pueden cubrir y hábitos de vida que no contribuyen a prevenir la enfermedad y promover la salud. Desde el Gobierno Nacional se vienen implementando estrategias para mejorar el flujo de recursos dentro del sistema, controlar el gasto en salud mediante la regulación del precio de los medicamentos, diseñar una estructura adecuada para la prestación de los servicios y otras medidas que, sin embargo, dejan la
Sensación de que hace falta una transformación de fondo.

Actualmente las necesidades de manejo de grandes volúmenes de información en el área de la salud, especialmente en nivel asistencial es vital para tomar decisiones que se prestan en servicios farmacéuticos, urgencias y hospitalarios entre otros.

En síntesis, el problema de investigación de este proyecto es proponer un Modelo de Evaluación que ayude al almacenamiento y procesamiento de altos volúmenes de datos, específicamente datos clínicos estructurados.

Las entidades medianas en salud en Colombia, especialmente en Medellín no se encuentran preparada para adoptar una arquitectura para el almacenamiento y procesamiento para la gestión de grandes volúmenes de datos desde la disciplina Big Data; sin embargo: ¿Que tan preparadas están las entidades de salud tecnológicamente para el uso de esta información?

Consecuente con lo anterior este trabajo de Maestría **apunta a proponer una Arquitectura de tecnología de información para el almacenamiento y procesamiento de altos volúmenes de datos clínicos en nivel asistencial**, promoviendo el fortalecimiento del equipo de atención que se encarga de la prestación de servicios de salud, en torno a las principales características de morbilidad de la población y los lineamientos de APS (Atención primaria en salud), con el ánimo de mejorar la capacidad resolutiva del primer nivel de atención.
Dado que la principal experiencia tecnológica de los sistemas de información médicos o en salud, tanto para el mundo en línea (OLTP – On Line Transaction Processing) como en el contexto analítico BI/DWH (Business Intelligence – Data Warehouse) está dominado por la tecnología SQL, se plantea la problemática de cómo hacer una transición gradual, inicialmente manteniendo la misma tecnología SQL en el mundo analítico Big Data.

1.2. OBJETIVOS DEL PROYECTO

1.2.1. Objetivo General

Proponer un **modelo de evaluación** que oriente la toma de decisiones al almacenar y procesar grandes volúmenes de datos desde bases de datos tradicionales hacia plataformas Big Data basadas en Hadoop/Spark.

1.2.2. Objetivos específicos

a) Identificar el estado del arte alrededor de tecnologías Big Data y Big Data Analytics asociadas al almacenamiento y procesamiento de datos en asistencia clínica.

b) Caracterizar el conjunto de servicios para la gestión de datos grandes soportados por tecnología Big Data y sistemas gestores de bases de datos tradicionales que se puedan aplicar a las necesidades de procesamiento y análisis de información en salud.

c) Definir un esquema de evaluación de plataformas de gestión de grandes volúmenes de datos que orienten la toma de decisiones asociadas al almacenamiento y procesamiento de este tipo de datos desde bases de datos tradicionales basadas en SQL hacia plataformas Big Data basadas en el ecosistema hadoop/spark con soporte en SQL.

1.3. METODOLOGÍA DE INVESTIGACIÓN

Esta investigación se realizará bajo un paradigma cualitativo y se utilizará una metodología de investigación analítica-descriptiva para abordar todos los conceptos claves necesarios para establecer los modelos de comparación de tecnología convencional SQL hacia tecnologías Big Data para entidades del sector salud.

Adicionalmente se realizará un caso de estudio para verificar la hipótesis y contrastar la viabilidad del modelo tecnológico propuesto con los resultados conseguidos en la entidad de salud.

Para cumplir acabildad esta investigación inicialmente se plantearán todas las tecnologías y conceptos a utilizar durante el desarrollo del proyecto, realizando previamente un análisis frente a sus características, con el objetivo de determinar cuáles son las más viables para la implementación de esta arquitectura.

En la segunda fase, una vez definidas las herramientas a utilizar y teniendo claro los objetivos a cumplir, se iniciará con la construcción del ambiente Big Data ejecutando las pruebas en los equipos de la universidad EAFIT, siguiendo así la finalidad de esta metodología, donde se parte de lo abstracto para llegar a lo concreto.

Fuentes de información: Dentro de las fuentes de información que se van a utilizar para el desarrollo del proyecto de investigación, se encuentra:

Fuentes primarias: Libros especializados en el manejo de Big Data, que lograran identificar conceptos claves para conseguir una implementación optima a la propuesta; también artículos de Internet, los cuales se tomarán como guía para escoger e instalar las herramientas más adecuadas para la construcción del ambiente.

Fuentes secundarias: La participación en conferencia sobre Big Data, tanto de manera presencial como de manera virtual, adicionalmente la asesoría con personas que tengan altos conocimientos acerca del tema, logrando así un óptimo desarrollo del proyecto.

2. MARCO DE REFERENCIA

2.1. ARQUITECTURAS Y TECNOLOGÍAS BIG DATA

En el 2013 el volumen de datos creado por humanos diariamente fue de 2.5 quintillones de bytes, y el 90% de los datos del mundo había sido creados en los dos años previos a 2013 **(IBM, 2013).**

Las organizaciones están incrementando el volumen de datos transaccionales lo cual hace que se capturen billones de datos de sus clientes, proveedores y sus operaciones. Millones de sensores de red están siendo agregados en el mundo físico en dispositivos inteligentes tales como teléfonos, medidores de energía, automóviles y máquinas industriales, en ese sentido crean datos y se comunican generando la nueva era del Internet de las Cosas (IoT) (Paul zikopoulos, Chris Eaton., 2011).

Miles de millones de personas en el mundo utilizan las redes sociales, equipos inteligentes, PCs, portátiles, entre otros; contribuyendo a la gran cantidad de datos disponibles, por otro lado, el creciente volumen de contenidos multimedia aporta considerablemente al aumento exponencial de los datos, por ejemplo, hoy se genera más de **2000 bytes** extras por cada segundo de vídeo en alta resolución que la generada por una página de texto (McKinsey, 2014).

En la actualidad la cantidad de información digital que se genera diariamente en nuestro planeta crece exponencialmente, lo cual ha desencadenado que muchas empresas y organizaciones, deseen utilizar esta información con el objetivo de mejorar las prestaciones de sus servicios o negocios. Por lo tanto, el objetivo fundamental de la tecnología Big data es dotar de una infraestructura tecnológica a las empresas y organizaciones con la finalidad de poder almacenar, procesar y analizar de manera económica, rápida y flexible la gran cantidad de datos que se generan diariamente, para ello es necesario el desarrollo y la implantación tanto de hardware como de software específicos que gestionen esta explosión de datos con el objetivo de extraer valor para obtener información útil para nuestros objetivos o negocios (Cortés, 2014).

El término de Big Data, en general se define con 7 dimensiones llamada las 7V (Martínez-Prieto, 2014), Volumen, Variedad, Velocidad, Veracidad, Valor, Variabilidad y Visualización:

Volumen: Como su propio nombre indica, la tecnología Big Data (datos masivos) ha de ser capaz de gestionar un gran volumen de datos que se generan diariamente por las empresas y organizaciones de todo el mundo.

Variedad: Big data ha de tener la capacidad de combinar una gran variedad de información digital en los diferentes formatos en las que se puedan presentar ya sean en formato video, audio o texto, caracterizada en información estructurada, semi-estructurado y no estructurada.

Velocidad: La tecnología Big Data ha de ser capaz de almacenar y trabajar en tiempo real con las fuentes generadoras de información como sensores, cámaras de videos, redes sociales, blogs, páginas web; fuentes que generan millones y millones de datos al segundo, por otro lado, la capacidad de análisis de dichos datos ha de ser rápidos reduciendo los largos tiempos de procesamiento que presentaban las herramientas tradicionales de analítica.

Veracidad: Big Data ha de ser capaz de tratar y analizar inteligentemente este vasto volumen de datos con la finalidad de obtener una información verídica y útil que permita mejorar nuestra toma de decisiones.

Valor: Es la capacidad de convertir los datos en valor. Es importante que las empresas hagan un gran intento de recoger y aprovechar los datos grandes. Esto puede también interpretarse como la toma de decisiones que permitan la mejora de rendimiento organizacional en muchas dimensiones basada en la información que se obtiene de la analítica en Big Data.

Variabilidad: Son los datos cuyo significado está en constante cambio. Este es particularmente el caso cuando la recolección de datos se basa en el procesamiento del lenguaje.

Visualización: Esta es la parte dura de grandes volúmenes de datos, hacer que la gran cantidad de datos sea comprensible de manera que sea fácil de entender y leer. Con los análisis y visualizaciones adecuadas, los datos brutos se pueden utilizar para tomas de decisiones adecuadas. Las visualizaciones por supuesto no significan gráficas ordinarias o gráficos circulares. Significan gráficos complejos que pueden incluir muchas variables de datos sin dejar de ser comprensible y legible. La visualización puede no ser la parte más difícil con respecto al uso de la tecnología, pero lograr buenos resultados seguro que es la parte más difícil. Contar una historia compleja en un gráfico es muy difícil, pero también es extremadamente crucial.

Para ampliar el alcance y definición de Big Data, se adicionan las siguientes definiciones:

Analítica Big data: Actualmente, las empresas que requieren conocer lo que está pasando en su negocio necesitan analizar a detalle grandes volúmenes de datos para ver algo que nunca antes habían visto, aprovechando datos antes descartados por la Inteligencia de Negocios (BI). Algunos de los datos sin explotar resultan ajenos a la naturaleza de las empresas ya que proceden de sensores, dispositivos, terceros, aplicaciones Web y medios sociales. Agrupar todos estos datos permite apreciar que Big Data no sólo involucra grandes volúmenes de datos también contempla una gran variedad de tipos de datos, entregados a varias velocidades y frecuencias.

Al inicio Big Data fue usado para resolver un problema técnico, en la actualidad se considera como una oportunidad de negocios para el manejo de grandes cantidades de información detallada.

Analítica Big Data trata de la unión de dos técnicas: Big Data y -Analytics-, y de cómo se han unido para crear hoy en día una de las más profundas tendencias en BI; en sí Analítica Big Data es la aplicación de técnicas analíticas avanzadas para conjuntos muy grandes de datos (Hurwitz, J., Nugent, A., Halper, F. y Kaufman, M, 2013).

Analítica Avanzada: Es una colección de diferentes tipos de técnicas, modelos y algoritmos avanzados que incluyen los basados en Minería de Datos, Aprendizaje de Máquina, Inteligencia Artificial, Análisis Predictivo, Procesamiento de Lenguaje Natural, entre otros (TMFORUM, 2012).

The Association for Data-driven Marketing and Advertisting ha generado varias definiciones de Big Data desde diferentes perspectivas (McKinsey, 2014) y son:

Perspectiva Comercial: Big Data es el término actual que se denomina a la amplia utilización de los datos recogidos de fuentes digitales, tecnológicas y analógicas. Big Data es usado para mejorar el conocimiento del negocio de los mercados, lo que permite mejorar en la experiencia del cliente y el desempeño organizacional.

Perspectiva Marketing: Big Data, es el uso exitoso de los datos en el liderazgo del marketing para mejorar la experiencia del cliente, lograr un mejor tercambio de valor entre clientes y organizaciones, y mejorar el desempeño de los negocios.

Perspectiva Global: Big Data es la colección de volúmenes grandes de información variada, usados para extender nuestro entendimiento del ambiente, medicina, ciencia, negocios y experiencia humana.

En base a las definiciones revisadas se concluye que Big Data no es simplemente una técnica o herramienta, es una tecnología que permite almacenar y procesar altos volúmenes de información a velocidades altas, complejos y variables que necesitan modernas técnicas de analítica y tecnologías para extraer, guardar, distribuir, interpretar y gestionar los datos.

Muchas de las tecnologías que conforman Big Data, tales como virtualización, procesamiento paralelo, sistemas de archivos distribuidos y base de datos en memoria han existido por décadas; y otras como MapReduce y Hadoop son nuevas, pero todas estas tecnologías combinadas entre si permiten abordar significativamente los nuevos problemas del negocio.

2.2. DOMINIOS DE APLICACIÓN DEL BIG DATA

El manejo de grandes volúmenes de datos es utilizado y se puede utilizar en múltiples áreas, por ejemplo:

- Seguridad: Su potencial reside en la capacidad de análisis de volúmenes de datos antes impensable de una manera óptima y ágil. Existen, por ejemplo, modelos de análisis del comportamiento humano para prevenir atentados terroristas obtenidos mediante un análisis permanente de las cámaras, sensores y accesos secuenciales a un sistema.
- Investigación médica: La investigación médica puede mejorar muchísimo si es capaz de asimilar una enorme cantidad de datos (monitorización, historiales, tratamientos, etc.) y estructurarlos para el establecimiento de diagnósticos o las síntesis de medicamentos.
- Gobierno y toma de decisiones: Big Data ofrece una mejora y optimización en los procesos de toma de decisiones de empresas y gobiernos, permitiendo entre muchas otras, el soporte a la toma decisiones, siendo complementario a las plataformas de Business Intelligence (BI).
- Internet 2.0: Genera una gran multitud de datos que difícilmente se podrán gestionar sin un Big Data. Motores de búsqueda, sistemas de recomendación, redes sociales cada vez se extienden a más ámbitos de nuestra sociedad

- CRM: La gestión de la relación de una empresa con sus clientes suele implicar la gestión de Almacén de Datos y la interrelación de diversidad de datos (comercial, operaciones, marketing), diversos canales (web, redes sociales, correo) y formatos. Big Data facilita las operaciones de análisis y seguimiento, favoreciendo la fidelidad y descubrimiento de nuevo mercado
- Logística: El sector logístico mejora notablemente gracias a las posibilidades analíticas de un Big Data y su potencial para el despliegue de servicios específicos (movilidad, tracking, seguridad, etc.). El ejemplo más popular se encuentra en el control de otras (la ruta optima permite a los vehículos circular con la máxima capacidad de carga, pudiendo recorrer rutas mejorando tiempos, consumos y contaminación).

2.3. TECNOLOGÍAS BIG DATA

2.3.1. Ecosistema Apache Hadoop para aplicaciones distribuidas

Apache hadoop es un conjunto de aplicaciones de software open source para el almacenamiento y procesamiento a gran escala de conjuntos de datos en un gran número de máquinas o servidores "commodity". Licenciado bajo la Apache License 2.0.

Dentro del contexto de Data Mining (DM) y Machine Learning (ML), el procesamiento y ejecución distribuida de los algoritmos es algo relativamente nuevo. Muchos de las implementaciones actuales de estos algoritmos sofisticas de DM y ML suponen que los datos están en la misma máquina y que el procesamiento lo realiza la misma máquina (incluyendo las características de múltiple de procesadores y GPU), pero Big Data rompe este esquema precisamente, ya que ante el alto volumen de datos y la necesidad de procesarlos como un conjunto, requiere una tecnología transparente para datos y procesamiento distribuido, aspecto que ofrece Hadoop así como otros frameworks como Spark.

Con la llegada y proliferación de la tecnología y la capacidad de computación se elaboraron algoritmos automatizados de procesamiento de los datos hasta la década de los 90.

En dicha época el diseño de los algoritmos no estaba enfocado al manejo de grandes volúmenes de datos (almacenamiento y recuperación de altos volúmenes de datos por ejemplo de sistemas de archivos distribuidos altamente escalables), sino más bien al aprovechamiento de la potencia de cálculo, lo que provoca que actualmente muchos de esos algoritmos sean inviables con las cantidades de datos que se pueden llegar a manejar hoy en día.

Aquí es donde Hadoop y las herramientas de su ecosistema ofrecen una alternativa, otorgándonos las herramientas necesarias para distribuir y paralelizar, de forma rápida y eficaz, los algoritmos que se adapten al paradigma MapReduce, Spark entre otros. De esta forma Hadoop ofrece una capa de abstracción a la problemática de paralización y gestión de tareas. Sobre esta plataforma existen herramientas adicionales que re-implementan algunos de los algoritmos de data mining ya conocidos adaptándolos a las necesidades y recursos de este momento.

Hadoop nos ofrece la base, más adelante se detallarán las herramientas que nos ofrecen la posibilidad de trabajar con algoritmos de data mining de forma distribuida (hadoop. apache, 2016).

A continuación se detallan los componentes principales del ecosistema Hadoop que hacen posible el análisis de grandes volúmenes de datos en tiempos razonables hasta tiempo real inclusive:

Commons: Es un conjunto de proyectos de Apache Software Foundation que originalmente formaron parte de Jakarta Project. El propósito de estos proyectos consiste en proveer componentes de software Java reutilizables, en código abierto. (Commons, 2016). El proyecto Apache Commons se compone de tres partes:

- Commons Proper: Un repositorio de componentes Java reutilizables.
- Commons Sandbox: Un espacio de trabajo para el desarrollo de componentes de Java.
- Commons Dormant: Un repositorio de componentes que se encuentran actualmente inactivas.

HDFS (Hadoop Distributed File System): Es el sistema de archivo distribuido utilizado por Hadoop. Está especialmente diseñado para cumplir con las necesidades propias de Hadoop. Las dos ideas principales de HDFS es por un lado que sea un sistema de archivo que facilite una alta escalabilidad tolerante a fallos. Por otro lado, Hadoop necesita que los problemas que se estén intentando solucionar involucren un gran número de datos. HDFS debe garantizar un alto rendimiento de datos para que Hadoop sea capaz de procesar

Características de HDFS

Es adecuado para el almacenamiento y procesamiento distribuido

- Hadoop proporciona una interfaz de comandos para interactuar con HDFS
- Streaming o flujo de ingesta y egesta de datos hacia y desde el sistema de archivo HDFS
- HDFS proporciona seguridad de archivos y autenticación.

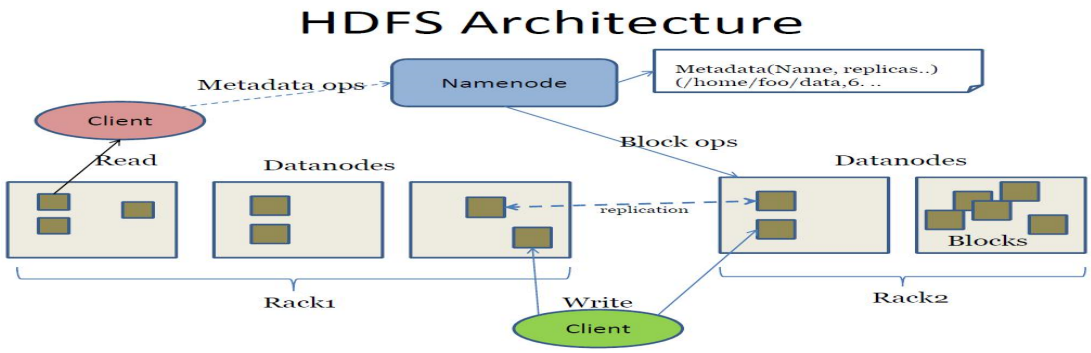

Figura 1. Arquitectura HDFS

En un clúster de HDFS encontramos dos tipos de nodos diferentes

Namenodes: Son los encargados de gestionar el espacio de nombres del sistema de archivo. Se encarga de la administración del sistema de archivos, mantiene el sistema de archivos y la metadata asociada a éstos y a los directorios. Almacena los datos relacionados con la posición de un bloque para un archivo en específico durante un periodo de tiempo, estos datos son actualizados al iniciar el sistema y cada cierto tiempo.

Datanodes: Son los que almacenan los bloques de información y los recuperan bajo demanda. Se encargan de almacenar y distribuir los bloques entre los distintos nodos. La distribución de los bloques es realizada cuando reciben un aviso desde un Namenode o algún cliente solicita una distribución. Una vez realizada la distribución, estos realizan un reporte al Namenode, este reporte se realiza periódicamente y contiene los bloques los cuales están siendo almacenados en ese nodo en específico.

Map/Reduce: El modelo de programación MapReduce se basa en dos funciones llamadas Map y Reduce. La entrada a dicho modelo es un conjunto de pares <clave, valor>y la salida es otro conjunto de pares < clave, valor> (MapReduce, 2016).

- Función Map. A partir del conjunto de pares clave/valor de entrada se genera un conjunto de datos intermedios. La función Map asocia claves idénticas al mismo grupo de datos intermedios. Cada grupo de datos intermedios estará formado por una clave y un conjunto de valores, por lo tanto, estos datos intermedios van a ser a su vez la entrada de la función de Reduce, después de pasar por un proceso de ordenamiento y distribución hacia los respectivos

Reduce, proceso que se realiza de forma determinística y provista por el Hadoop sin intervención del usuario o programador.

- Función Reduce. La fase de Reduce se encargará de manipular y combinar los datos provenientes de la fase anterior para producir a su vez un resultado formado por otro conjunto de claves/valores (MapReduce, 2016).

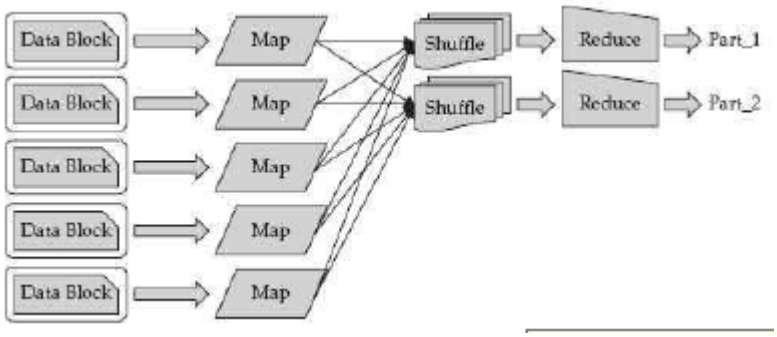

Figura 2. Proceso Map/Reduce
Muestra un modelo de cómo diferentes bloques de datos pueden ser "mapeados" por grupos de acuerdo a la <clave, valor>.

Pig: Es un Framework y lenguaje scripting para el análisis de grandes conjuntos de datos que ejecuta sobre Map/Reduce y puede tomar los datos de entrada desde múltiples fuentes como HDFS. Consiste en un lenguaje de alto nivel (Pig Latin) para la expresión de los programas de análisis de datos, junto con la infraestructura necesaria para la evaluación de estos programas. La característica más importante es que su estructura es susceptible de paralelismo que a su vez le permite manejar grandes conjuntos de datos

En la actualidad, la capa de procesamiento de Pig se compone de un compilador que produce secuencias de programas de Map/Reduce, para el que ya existen implementaciones paralelas a gran escala.

Pig posee las siguientes características

- Facilidad de programación. Es trivial para lograr la ejecución paralela de tareas de análisis simples. Las tareas complejas compuestas de múltiples transformaciones de datos relacionados entre sí, están codificados explícitamente como secuencias de flujo de datos, lo que hace que sean fáciles de escribir, entender y mantener.

- Oportunidades de optimización. La forma en que se codifican las tareas permite que el sistema pueda optimizar su ejecución de forma automática, lo que permite al usuario centrarse en la semántica en lugar de la eficiencia

- Extensibilidad: Los usuarios pueden crear sus propias funciones para realizar el procesamiento de propósito especial.

Hive: Es la infraestructura de procesamiento de datos estructurados mediante de interfaz SQL construida sobre Apache Hadoop para proporcionar integración con numerosas herramientas de analítica basada en SQL (Plataformas OLAP, Data Warehouse, herramientas de visualización, entre otras), igualmente soporta analíticas complejas y distribuidas basadas en datos estructurados gestionados vía SQL.

Proporciona un mecanismo lenguaje similar a SQL llamado HiveQL (HQL). Hive facilita la integración entre Hadoop y herramientas para la inteligencia de negocios y la visualización. Hive permite al usuario explorar y estructurar los datos, analizarlos y luego convertirla en conocimiento del negocio.

Estas son algunas de las características de Hive:

- Cientos de usuarios pueden consultar simultáneamente los datos utilizando un lenguaje familiar y ampliamente utilizado con Sql.

- Los tiempos de respuesta son típicamente mucho más rápido que otros tipos de consultas sobre el mismo tipo de conjuntos de datos.

- Controladores JDBC y ODBC, para integrarlos con otras aplicaciones para extraer datos. Hive permite a los usuarios leer los datos en formatos arbitrarios.

Sqoop: Es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales. Sqoop importar los datos de un sistema de gestión de bases de datos relacionales (RDBMS) como MySQL u Oracle al sistema de archivos distribuido Hadoop (HDFS), donde transforma los datos y luego los exporta de nuevo a un Rdbms.

Sqoop automatiza la mayor parte de este proceso, basándose en la base de datos para describir el esquema de importación de los datos. Sqoop utiliza MapReduce para importar y exportar los datos, lo que proporciona el funcionamiento en paralelo, así como tolerancia a fallos

Chukwa Es un sistema de recopilación de datos de código abierto para el seguimiento de grandes sistemas distribuidos. Se construye en la parte superior del sistema de archivos distribuido Hadoop (HDFS) y Map/Reduce. Chukwa también incluye un conjunto de herramientas flexibles para la visualización, seguimiento y análisis de resultados de los datos recogidos.

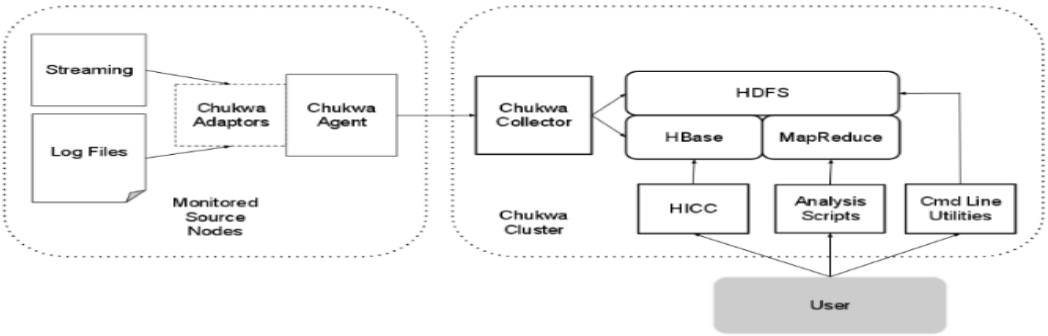

Figura 3 Diagrama de chukwa

Muestra la interacción de Chukwa con las otras tecnologías y la función que tiene dentro de los procesos de datos.

Flume Es un servicio distribuido, confiable y disponible para recolectar, agregar y mover grandes cantidades de datos de registro eficientemente. Cuenta con una arquitectura simple y flexible basada en transmisión de flujos de datos. Es robusto y tolerante a fallos con los mecanismos de fiabilidad, conmutación por error y los mecanismos de recuperación. Se utiliza un modelo de datos extensible simple que permite una aplicación analítica en línea, el cual se puede ver en la siguiente figura.

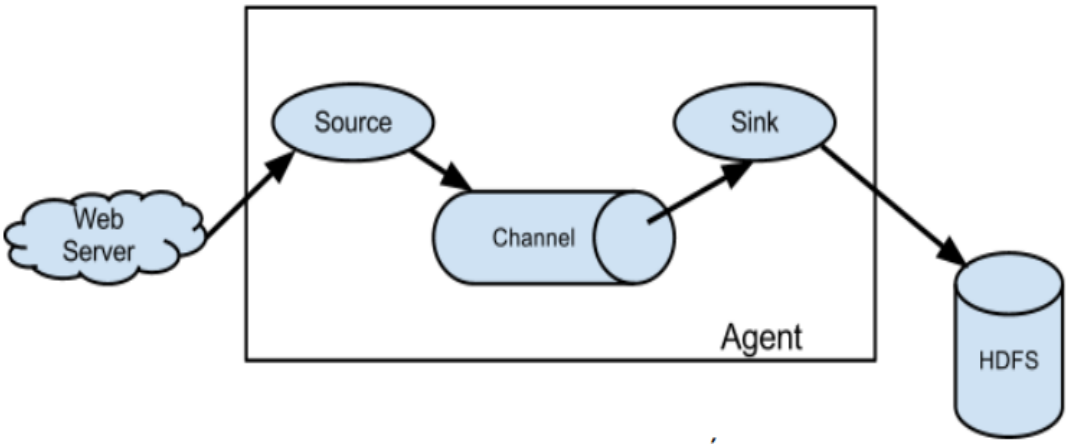

Figura 4 Diagrama flume

Sistema distribuido para capturar de forma eficiente, agregar y mover grandes cantidades de datos log de diferentes orígenes

Avro: Es un sistema de serialización de datos; contiene lo siguiente

- Estructuras de datos.
- Formato de datos binario.
- Un archivo contenedor para almacenar datos persistentes.
- Llamada a procedimiento remoto (RPC).

- Fácil integración con lenguajes dinámicos. La generación de código no está obligado a leer y escribir archivos de datos ni a utilizar o implementar protocolos RPC.

Cuando los datos de Avro se almacenan en un archivo, su esquema se almacena con él, para que los archivos se puedan procesar posteriormente con cualquier programa.
Si el programa de lectura de los datos espera un esquema diferente esto puede ser fácilmente resuelto, ya que ambos esquemas están presentes.

2.3.2 Apache spark

Apache Hadoop es un framework open source que ofrece soluciones en Big Data para procesar grandes conjuntos de datos en clusters de computadoras usando modelos de programación sencillos. Hadoop se compone principalmente de dos componentes: HDFS (Hadoop Distributed File System) (Hurwitz, J., Nugent, A., Halper, F. y Kaufman, M, 2013) y el framework de Map-Reduce (Dean, 2004).

Hadoop está diseñado para ser simple y fácil de usar, flexible en cuanto al procesamiento y almacenamiento de datos, tolerante a fallos y con elevada escalabilidad.

No obstante, presenta una serie de inconvenientes (Gopalani, S., & Arora, R. 2015).

- Flujo de datos fijo: Provee facilidad de uso con una abstracción simple que también es un flujo de datos fijo. Es decir, varios algoritmos complejos son difíciles de implementar en un solo trabajo de procesamiento (Job). Además, algunos algoritmos que requieren múltiples entradas no son compatibles.

- Baja eficiencia: Con la tolerancia a fallos y la escalabilidad como objetivos primarios, las operaciones no siempre son las más eficientes.

- Latencia: Por su inherente procesamiento en batch, sufre el problema de la latencia.

Sin embargo, recientemente, con la introducción de Apache Spark, se dispone de un nuevo modelo de computación en el contexto de Big Data que ofrece una interfaz de programación que permite disminuir los esfuerzos de programación y ofrece mejor rendimiento en los tipos de problemas relacionados con Big Data (Gopalani, S., & Arora, R., 2015).

Mientras que Hadoop ofrece poca flexibilidad a la hora de crear flujos de datos ya que sigue un esquema de ejecución fijo, Spark ofrece un esquema de computación más flexible, debido a que se basa en un flujo de ejecución de grafo a cíclico dirigido, y es posible modificar el flujo con distintas transformaciones y acciones.

Spark es un motor de computación en clúster, rápido y de propósito general para procesamiento de datos a gran escala. Comenzó como un proyecto de investigación en UC Berkeley en el AMPLab, con el objetivo de diseñar un modelo de programación compatible con una clase de aplicaciones más amplia que Map-Reduce mientras mantenían la tolerancia fallos.

Además, Spark dispone de una librería de aprendizaje automático (MLlib), consulta de Datos con SQL (Spark SQL) y procesamiento de flujos (Spark Streaming).

Spark expone un interfaz de programación funcional (Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I, 2010) y puede utilizarse con varios lenguajes de programación (Java, Scala, Python, R). La principal fortaleza de la programación funcional, al evitar el estado (y los efectos colaterales), es que el sistema entero obtiene transparencia referencial, que implica que cuando se pasan una serie de argumentos a una función siempre devuelve el mismo resultado, es decir, siempre se comporta de la misma forma, lo cual facilita su ejecución distribuida (Moseley, B., & Marks, P., 2006).

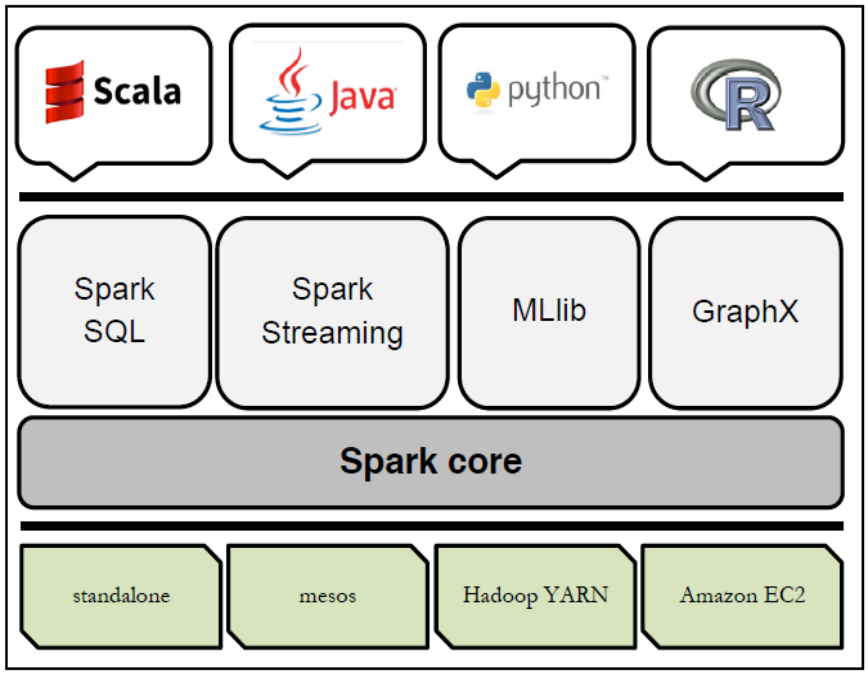

Figura 5. Módulos disponibles en spark,

Este diagrama es utilizado para el análisis y almacenamiento de enormes volúmenes de datos

Spark ofrece una abstracción llamada Resilient Distributed Datasets (RDD) para soportar las aplicaciones que trabajan con datos masivos de forma eficiente. Los datos en RDD se pueden almacenar en memoria entre consultas sin necesidad de replicación lo cual ofrece un mejor rendimiento. Los RDD permiten mejorar los modelos existentes hasta 100 veces en análisis que requieran procesamiento

iterativo. También permiten minería de datos interactiva, consultas SQL altamente eficientes, procesamiento de flujos y computación en paralelo sobre grafos (Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I., 2010), (Da Silva Morais, T. 2015).

2.4. BIG DATA EN SALUD

El modelo sanitario y en general el sector de la salud, es uno de los sectores donde Big data está teniendo mayor impacto en la actualidad y donde sus aplicaciones crecerán de un modo espectacular, tanto para el área médica, como también para las áreas de análisis de datos (historias médicas, análisis clínicos etc.), la gestión de centros de salud, la administración hospitalaria y la documentación científica.

De acuerdo al estudio de 2011 del Mckinsey Global Institute sobre Big data, sin lugar a dudas uno de los más referenciados en la Web, calcula que la aplicación de Big data en el campo de la salud podrían suponer un beneficio de 250.000 millones de euros al sector público europeo y unos 300.000 millones de dólares al sector de sanidad de los EEUU. Otro estudio posterior de octubre de 2012 también relativos al impacto de Big data en el sector de la salud confirman los mismos datos de presupuestos. (McKinsey, 2014)

Se define un dato estructurado como un dato que puede ser almacenado, consultado, analizado y manipulado por máquinas, normalmente, en modo tabla de datos. Un dato no estructurado o desestructurado **es todo lo contrario**. Datos estructurados, son los datos clásicos de los pacientes (nombre, edad, sexo etc.) y datos no estructurados son las fórmulas en papel, los registros médicos, las notas manuscritas de médicos y enfermeras, las grabaciones de voz, las radiografías, escáneres, resonancias magnéticas, TAC y otras imágenes médicas. A estos datos y pertenecientes a ambas categorías también se pueden considerar los archivos electrónica, la gestión administrativa y los datos clínicos, etc.

Los avances tecnológicos están generando una nueva gestión de datos de todo tipo, que provienen de los más variados dispositivos, sensores, aparatos médicos diversos, datos hospitalarios etc. y a ellos se suman los datos procedentes de los medios sociales (redes sociales, blogs, wikis, podcast) de los teléfonos inteligentes, de áreas tan voluminosas como importantes tales como genérica y genómica.

Otro informe (Salud., 2012) también de gran impacto es, "Big Data in digital Health" de la Fundación Rock Health analiza la situación actual y el potencial del Big data en el mundo de la salud. El informe utiliza los datos y la información obtenidos en entrevistas con emprendedores e inversionistas y calcula como se puede producir importantes ahorros en el sector sanitario. Según las conclusiones del informe hay cinco vías mediante las cuales Big Data puede cambiar la atención sanitaria y apoyar la investigación:

- Transformación de datos en información.
- Apoyo al autocuidado de las personas.
- Apoyo a los proveedores de cuidados médicos.
- Aumento del conocimiento y concienciación del estado de salud.

- Agrupamiento de los datos para expandir el ecosistema.

El informe concluye recomendando tres tendencias que emergen en el uso de los datos y que considera de gran trascendencia:

- Trabajar con conjuntos de datos limitados.
- Combinar una gran variedad de datos.
- Agrupamiento de datos para mejorar resultados.

La investigación médica puede mejorar muchísimo si es capaz de asimilar una enorme cantidad de datos (monitorización, historiales, tratamientos, etc.), especialmente no estructurados, y organizarlos o estructurarlos para definir las causas de enfermedades y establecer mejores tratamientos. Big data en sanidad se utilizará para predecir, prevenir y personalizar enfermedades, tratamientos y con ello los pacientes afectados. Los campos serán prácticamente casi todos los sectores de la sanidad, pero en particular podemos citar ya algunos en los que se están encontrados los mayores desafíos:

- La investigación genómica y la secuenciación del genoma.
- Operativa clínica.
- Autoayuda y colaboración ciudadana.
- Mejora en la atención personalizada al paciente.
- Monitorización remota de pacientes.
- Medicina personalizada para todos.
- Autopsias virtuales.
- Seguimiento de pacientes crónicos.
- Mejoras en los procesos médicos.

Las aplicaciones de Big Data en el sector salud y sanitario son numerosas y en aumento. Por ejemplo, los profesionales sanitarios pueden utilizar la analítica de Big data en tiempo real para saber dónde se está extendiendo un virus de la gripe y a qué ritmo, pueden adaptar la respuesta y garantizar el stock de vacunas suficiente para los sitios que lo necesiten.

2.4.1 Casos de Big data en salud

- **Análisis de los registros electrónicos en salud:** Este caso está dirigido al análisis de todos los pacientes en salud a partir de los registros electrónicos médicos, debido a que se están analizando sus datos. Su objetivo es optimizar en varios frentes, la prestación de asistencia sanitaria mediante el intercambio de información de los pacientes entre los proveedores para reducir los ensayos de serie de pedidos y reducir el tiempo necesario para prestar atención médica (Jason S. Mathias, 2012).
- **La red de Big data en el hospital**: Actualmente el hospital registra continuamente los datos de todos los instrumentos médicos en una sala de pediatría, mediante la captura

de los datos y el análisis, se pudo ayudar a los médicos a detectar una infección de 12 a 24 horas antes de lo que pudieron haber visto los médicos permitiendo iniciar un tratamiento que permiten a salvar la vida (Big data, 2014).

- **Control de los datos para investigar, informar y mejorar la salud:** Las organizaciones de la salud ya no tienen que buscar y recopilar los datos; ahora, el reto consiste en gestionar e informar los datos para el suministro y control de este proceso, además debe cumplir con los requisitos reglamentarios. Esta investigación incluye la información recopilada, así como las pruebas relacionadas con tratamientos específicos. Esto elimina la cantidad de registros que las entidades de salud a menudo reciben y que congestionan las líneas telefónicas (Big data, 2014).
- **Analítica de la telemedicina:** Mediante una plataforma de telemedicina puede ofrecerse servicios asistenciales al paciente cuando es difícil que este venga al hospital. Una plataforma de telemedicina puede capturar varios signos vitales del paciente, como la temperatura, la frecuencia cardíaca, la presión arterial y el ECG (electrocardiograma) que puede ser enviado a un repositorio central en tiempo real a través de una red de datos como internet. Una vez recopilado una serie de factores desencadenantes pueden ser colocados en los datos para detectar y responder a las condiciones de salud en el mundo real (Analytics, 2014).
- **Identificación delos pacientes:** Los aseguradores están utilizando análisis de grandes volúmenes de datos con el fin de detectar el fraude y robo de identidad médica. La compañía utiliza la analítica en los registros de voz a texto de las llamadas al centro de atención para identificar posibles fraudes, con el fin de predecir qué tipos de planes de tratamiento tienen más probabilidades de tener éxito (Frank Irving, 2014).
- **Analítica Big data y Hadoop en el desarrollo de servicios de salud en la india**: En este trabajo, se analizan y revelan los beneficios de Big Data analítica y Hadoop en las aplicaciones de la Salud, donde los datos fluyen y el volumen es masivo. En los países en desarrollo como la India, con población masiva y diversos problemas en el campo de la salud con respecto a los gastos, la satisfacción de las necesidades de la economía deprime a las personas, el acceso a los hospitales, la investigación en el campo de la medicina y especialmente en el tiempo de propagación de epidemias.

  En este trabajo se da la participación de Big Data Analítica y Hadoop, donde se revela el impacto de los servicios de asistencia sanitaria para que todos sus costos sean óptimos (D. Peter Augustine, 2014).
- **Tecnología Big data para el sector salud en el estado de guerrero**: Este trabajo presenta la problemática que tiene el Sector Salud en el estado de Guerrero en su manejo, en su almacenamiento y en el análisis de información generada por las unidades sanitarias, por lo que se propone aplicar la tecnología de Big Data utilizando Hadoop, Hbase, MapReduce, Hcatalog y así reducir los costos, con lo que puede llamar salud digital, permitiendo mejorar en los tratamientos a los pacientes, apoyar en las decisiones médicas cada que se requiera, el análisis de mayor número de variables de análisis comparativo de múltiples opciones de tratamiento y de esta

manera proponer la mejor, se pueden monitorear pacientes a través de dispositivos remotos, recolectar datos individuales.

Al utilizar esta tecnología se estará apoyando al sector sanitario estatal en la automatización de los sistemas de monitorización de pacientes, sistemas de soporte a las decisiones médicas, análisis comparativo de tratamientos y que los médicos puedan aplicar la medicina más personalidad de acuerdo a cada paciente.

Posteriormente se podrán incorporar más unidades médicas que continúen alimentando la base de datos con información de pacientes, enfermedades, sobre todo las denominadas del nuevo siglo (Escobar Ayona Elías Marino, 2014).

- **Apoyo de Big data a los proveedores de salud y mejora en la atención al paciente:** Big Data se emplea también como herramienta para la inteligencia de negocio en la gestión sanitaria. Las plataformas actuales permiten explorar y medir en segundos una enorme cantidad de datos clínicos, de gestión financiera, redes sociales, etc., para medir el rendimiento y generar nuevos modelos de pago (pago por rendimiento). Servirán para mejorar la atención al paciente y reducir costos (Onehealth, 2015).
- **Aumentar el conocimiento con Big data:** Existen diversas herramientas que se utilizan para aumentar el conocimiento y resolver una gran variedad de problemas basándose en los datos. Sproxil utiliza Big Data para identificar medicamentos falsificados, protegiendo así la salud del paciente, y para que las empresas farmacéuticas puedan rastrear la distribución de medicamentos y prevenir el fraude. (Sproxil, 2014) Asthmapolis recoge datos de los pacientes con asma y les proporciona retroalimentación que les ayuda a controlar mejor su enfermedad. Un dispositivo móvil mediante una aplicación para IOS / Android se conecta con inhaladores para el asma que disponen de un sensor que controla la hora y el lugar de los hechos en el momento de tomar la dosis y recoge datos de los posibles factores ambientales desencadenantes, así como los síntomas para rastrear brotes de enfermedades y ofrece previsiones para los usuarios de forma similar a la predicción del tiempo.
- **Apoyo al auto-cuidado y colaboración ciudadana con Big data:** Una nueva herramienta que nos ayuda al cuidado de nosotros mismos. Mediante la recopilación de los datos procedentes de sensores instalados en dispositivos "wearables" y smartphones, Big Data puede devolver información a los usuarios sobre su estado de salud. Aplicaciones para móviles como IBlueButton, de Humetrix, permiten el intercambio rápido y seguro de información entre pacientes y registros médicos. La plataforma Ginger.io, basada en la nube, recopila datos en tiempo real desde los smartphones acerca del comportamiento, activo y pasivo, de los pacientes para ayudar a los médicos, enfermeras, familiares y pacientes a gestionar su salud, comenzando por las enfermedades crónicas y con el consentimiento del paciente, la recogida de datos y análisis quedarán a disposición de los proveedores de salud y los investigadores a través de un panel de control. (Ginger.io, 2015)

2.4.2. Aplicaciones Map/Reduce en la asistencia médica (Healthcare)

Programación Map/Reduce para análisis de datos grandes clínicos

En este informe iniciamos enumerando los diferentes tipos de grandes conjuntos de datos clínicos, seguido por los esfuerzos que se desarrollan para aprovechar los datos y las ventajas analíticas. Estas ventajas se centran principalmente en análisis descriptivo y predictivo. La razón principal para usar el marco de programación MapReduce en los esfuerzos de velocidad de este tipo de análisis. Esto es debido al hecho de que este tipo de algoritmos analíticos es muy bien desarrollado y probado para el marco de MapReduce y la plataforma Hadoop, que puede manejar una gran cantidad de datos en una pequeña cantidad de tiempo. Los análisis prescriptivos requieren datos compartidos entre los nodos de computación, que lamentablemente no pueden ser alcanzado fácilmente (es decir, programas sofisticados con una gran cantidad de gestión de datos) utilizando MapReduce y, por lo tanto, no todos los problemas de optimización (es decir, análisis prescriptivo) puede implementarse en el marco de MapReduce (Emad A Mohammed, 2014).

Uso de Map/Reduce para análisis de imágenes médicas a escala

El crecimiento de la cantidad de datos de imágenes médicas producidas diariamente en los hospitales modernos, la adaptación del análisis de imagen médica tradicional y el enfoque hacia soluciones escalables de indexación.

El número de imágenes y su dimensionalidad aumentaron dramáticamente durante los últimos 20 años. Proponemos soluciones para análisis de imagen médica de gran escala basados en optimización de algoritmo y computación paralela. El concepto de Map/Reduce se utiliza para acelerar y hacer posibles procesamientos de imágenes médicas de gran escala. (Dimitiros Marconi, 2015)

2.4.3 Trabajos realizados con spark en healthcare

- **Flujo de trabajo basado en spark para el acoplamiento de registro probabilístico de datos sanitarios** Varias áreas, tales como la ciencia, la economía, las finanzas, la inteligencia de negocios, la salud y otros; están explorando grandes datos como una manera de producir información, y tomar mejor las decisiones, y mover hacia adelante sus sistemas y tecnologías.

  Específicamente en salud, los datos grandes representan un desafío ante los problemas de la mala calidad de los datos en algunas circunstancias y la necesidad de recuperar, agregar y procesar una gran cantidad de datos de diferentes bases de datos. Este trabajo se enfoca en un sistema de salud pública brasileño y en grandes bases de datos del Ministerio de salud y Ministerio de Desarrollo Social, Se presenta un análisis basado en spark para el procesamiento de datos y vinculación de registros probabilístico de esas bases de datos para producir almacenes de datos muy precisos (Robespierre Pita, 2014).

- **Resolución de problemas prácticos con el análisis profesional de la salud** en un mundo donde la atención es primordial, los proveedores de la salud deben

atender el cuadro completo de la salud del paciente, que puedan administrar los costos y riesgos al tiempo que ofrece atención personalizada de alta calidad. Para proveedores y pagadores por igual, con las condiciones de salud del paciente con precisión y completamente codificado, es crucial para el pago el cuidado entregado y riesgos los compartidos. Pacientes, que a menudo son responsables de un porcentaje creciente de los costos de su atención, también necesitan una comprensión clara y completa de su Salud en general. (Intel, 2014).

2.4.4. Datasets de salud

Para el desarrollo de este proyecto, se exploró y analizo varias fuentes de datos, que permitirán tener el volumen y oportunidades de análisis para realizar el caso de estudio de verificación de la hipótesis o pregunta de investigación. Después de un análisis de fuentes, y ante la imposibilidad de tener datasets en Colombia con volumen y tipo requerido, se seleccionó los datasets provistos por el sistema de salud del estado de Texas en Estados Unidos (http://www.dshs.texas.gov/ y http://www.healthdata.dshs.texas.gov/)

Dentro del análisis de otras fuentes, se consideró una base de datos abierta de información del sector de la salud de la página web de http://www.emrbots.org/, donde se ofrece una base de datos de pacientes ficticios. La base de datos contiene las mismas características que existen en la base de datos médica real, como los detalles del paciente de admisión, datos demográficos, datos socioeconómicos, laboratorios, medicamentos, etc.

Otras fuentes de datos generales en salud se listan a continuación:

DESCRIPCIÓN	LINK
Amazon Web Service - Datasets	http://aws.amazon.com/datasets
Stanford Large Network Dataset Collection	https://snap.stanford.edu/data/
Health Data - Datos del sector salud	http://www.healthdata.gov/dataset
ADNI (Alzheimer's Disease Neuro-Imaging Initiative)	http://www.adni-info.org/Scientists/ADNIScientistsHome/ADNIPublications.aspx
Texas Hospital Inpatient Discharge Public	http://www.dshs.state.tx.us/thcic/hospitals/HospitalData.shtm

The Dartmouth Atlas of Health Care	http://www.dartmouthatlas.org/
Data Sources from Second Workshop on Data Mining for Health care Management (DMHM 2011)	http://phpartners.org/health_stats.html#National%20Public%20Health%20Data%20Sets
NIH Data Sets	http://www.nlm.nih.gov/hsrinfo/datasites.html
Drug and Substance Abuse Data Sets	http://www.icpsr.umich.edu/icpsrweb/SAMHDA/download
Agency for Health Care Research and Quality	http://www.ahrq.gov/data/dataresources.htm
Framingham Health Care Data Set	http://www.framinghamheartstudy.org/share/index.html
Laboratory for Neuro-Imaging (UCLA)	https://ida.loni.ucla.edu/services/Menu/IdaData.jsp?page=DATA&subPage=AVAILABLE_DATA
Heritage Health care Prize	http://www.heritagehealthprize.com/
Behavioral Risk Factor Surveillance System (BRFSS)	http://www.cdc.gov/brfss/
Behavioral Risk Factor Surveillance System (BRFSS)	http://www.cdc.gov/brfss/
World Health Organization Data Sets	http://apps.who.int/ghodata/
World Bank Health Care Data Sets	http://data.worldbank.org/country
Data Science central (amplia lista):	http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free

2.5 ALMACENAMIENTO EN BIG DATA

Con el paso de los siglos el ser humano ha tenido la necesidad de guardar su conocimiento en diferentes soportes. El almacenamiento y recuperación de la información representan uno de los problemas a los que la humanidad se ha tenido

que enfrentar desde la invención de la escritura. Con la aparición de la computadora este problema se ha resuelto parcialmente con nuevos dispositivos de almacenamiento, diseños de conexión y estructuras de bases de datos.

Las tecnologías han evolucionado para atender las necesidades de almacenamiento, de ahí que nos encontramos en una etapa compleja donde es más fácil producir datos que guardarlos y administrarlos, en consecuencia, las capacidades de almacenamiento han tenido que crecer. Por lo tanto, se intuye que la cantidad de información digital que se produce en el mundo es inmensa; sin embargo, ignoramos su verdadera cantidad, asimismo su enorme dimensión. Existen Exabytes de datos almacenados en servidores de empresas que se han visto en la necesidad de ampliar su capacidad de espacio; en principio pareciera que el tamaño de almacenamiento es un problema, aunque se olvida que la naturaleza de los datos y la administración de la entrada y salida del sistema de información es otra cuestión de suma importancia.

El almacenamiento de datos puede verse desde dos perspectivas. La primera observación se puede hacer desde el punto de las estructuras de sistemas de almacenamiento con opciones como DAS (Direct Attached Storage o Almacenamiento de Conexión Directa), NAS (Network Attached Storage o Almacenamiento Conectado en Red), SAN (Storage Área Network o Red de área de Almacenamiento) y sistemas de almacenamiento en la nube, que incluye capacidades de espacio en unidades de discos duros tradicionales y sólidos, así como la tecnología de la Memoria de Cambio de Fase (PCM: Phase Change Memory). La segunda visión se enfoca en la naturaleza de los datos en una perspectiva más cercana a la administración de datos; probablemente se pueda tener la capacidad de espacio a través de los sistemas distribuidos de nube, pero surgen inconvenientes relacionados con la consistencia, disponibilidad y tolerancia de partición de los datos; es decir, se trata de una perspectiva más cercana a la administración de datos.

2.5.1 Sistemas de almacenamiento de datos digitales

Los datos almacenados a través de unidades de disco duro tradicionales tienen amplias capacidades de almacenamiento, aunque con poco rendimiento en el acceso, tomando en cuenta su naturaleza de funcionamiento, pues la velocidad del giro de los platos magnéticos implica un funcionamiento lento, razón por la cual son remplazados por discos de estado sólido (SSD) con chips que permiten una mayor velocidad en el acceso a la información guardada. Estos últimos tienen la desventaja de ser más caros que los discos duros tradicionales; no obstante, el costo del SSD sería menor al de las nuevas memorias emergentes (Su-Kyung Yoon, 2014).

El desarrollo de la tecnología en almacenamiento sigue presentando nuevas formas cambio de memoria, que de acuerdo con (K. Gopalakrishnan, 2010), tiene la bondad de facilidad de integración, escalabilidad, velocidad y resistencia; esta podría ser una opción en un sistema de almacenamiento de grandes capacidades de datos. Además, sus bondades prometen ser memorias no volátiles de próxima (Su-Kyung Yoon,

2014), considerando que muchas aplicaciones modernas exigen cada vez más para el manejo de grandes cantidades de datos.

Nos encontramos ante una etapa de transición, de modo que las unidades de disco duro tradicional, sólido y memorias de PCM en este tiempo nos ayudarán a resolver nuestros problemas de almacenamiento de grandes datos. Este desarrollo tecnológico ha permitido que los centros de datos dedicados al almacenamiento estén estructurados de una o varias combinaciones de las siguientes cuatro formas: DAS, SAN, NAS y almacenamiento en la nube. A continuación, describimos cada uno de ellos.

2.5.1.1. Almacenamiento de conexión directa (DAS)

Direct Attached Storage o DAS es una de las formas más sencillas y tradicionales del almacenamiento de conexión directa, donde las unidades de disco se encuentran conectadas directamente con los servidores o host a través de una interfaz de datos SCSI o IDE (Qunhui, 2010). Las conexiones en DAS tienen muchas ventajas, tales como: su instalación es fácil; el software es poco complejo; el costo en mantenimiento es bajo; la tecnología presenta madurez técnica, buena compatibilidad y, relativamente, es de menor gasto. Sin embrago, su deficiencia aparece en cuatro aspectos:

- ✓ La capacidad de almacenamiento está limitada por el servidor
- ✓ Su rendimiento de almacenamiento es directamente afectado por el servidor
- ✓ Los servidores dispersos geográficamente se limitan al intercambio de información y gestión cuando se tiene un servidor aislado
- ✓ La carga de almacenamiento de datos y el acceso en el servidor hará en general tener un pobre rendimiento.

El entorno de uso de este tipo de arquitectura de almacenamiento es ideal para el intercambio de archivos localizados en ambientes con un único servidor o unos cuantos servidores.

2.5.1.2. Almacenamiento conectado en red (NAS)

Para proveer el almacenamiento en red es necesaria una LAN o WAN, además de un dispositivo de almacenamiento dedicado y diseñado para esta infraestructura; su propósito es proporcionar a los usuarios un sistema de servicio de acceso e intercambio de información. El almacenamiento en red se caracteriza por el depósito masivo de datos, lo que incluye intercambio de datos limitados, fiabilidad y seguridad en los datos, y así como el simplificado y unificado en la gestión de datos. Aunque su principal bondad es la capacidad de expansión, donde se proporcionan tasas de transmisión de la información de acuerdo al volumen de datos. Las conexiones SAN y NAS son ejemplos claros del almacenamiento en red.

El Almacenamiento Conectado en Red o *NAS* (del *acrónimo inglés Network Attached Storage*) es un dispositivo que se conecta a la red y provee un almacén de datos que permite a varios hosts acceder al mismo lugar de almacenamiento a través de una red IP. El espacio de almacenamiento se presenta en la red con un nodo dedicado a través de un servidor de archivos, aunque en sistemas recientes este dispositivo puede ser un dispositivo inmerso en la red. NAS y LAN están en la misma red física; por lo tanto, NAS depende de ciertas características de LAN. Para ello necesita un gran ancho de banda en red y de muy alta potencia de procesamiento del CPU; cuando no se cumplen estas condiciones, la red se congestiona y su rendimiento se reduce (San-jun Liu, 2012).

Con el servidor de archivos se gestiona la entrada y salida de datos en el disco duro; además, se regula el acceso entre varios clientes de red. Para (Edelson, 2004), el almacenamiento en NAS tiene dos características. En primer lugar, es la conexión física, puesto que se conecta el servidor de archivos directamente al equipo de almacenamiento y otro punto a la red, evitando así la carga de entrada y salida de datos en el servidor; en segundo lugar, técnicamente, se reducen los movimientos del brazo de la unidad de disco duro y, por lo tanto, se reduce el desgaste. Sin embargo, en esencia la estructura de este tipo de almacenamiento muestra que todavía es un equipo de servidor tradicional.

Los principales beneficios de NAS son la facilidad de comunicación entre una computadora y el sistema de almacenamiento en comparación con una conexión de computadora a computadora. El intercambio y recuperación de datos mediante una sola fuente de almacenamiento genera menos errores, menos trabajo al tratar de mantener copias de seguridad, y mayor precisión en la búsqueda de información. Estos sistemas son más seguros, porque en lugar de almacenar los datos en un solo disco duro distribuyen copias de los datos entre distintos discos duros que actúan como uno solo. Cuando un disco duro falla, se alerta al administrador de redes, y la información continúa estando disponible para todos los usuarios (Edelson, 2004).

El sistema NAS tiene las ventajas tales como facilidad en la instalación, complementos o extensiones (Plugs), precio, flexibilidad de conexión, fácil mantenimiento, seguridad de autenticación, administración de espacio en disco y escalabilidad. Así las cosas, NAS es una opción ideal para organizaciones pequeñas y medianas que buscan, de una manera simple y rentable, lograr el acceso de datos rápido en nivel de archivo para varios clientes (Cunhe, 2002).

2.5.1.3. Red de área de almacenamiento (SAN)

La tecnología SAN, se orienta a la alta velocidad de procesamiento de datos masivos, lo que incluye alta velocidad en el acceso, almacenamiento seguro, intercambio de datos, respaldo de datos, migración de los datos, entre otras ventajas de los sistemas distribuidos. Se debe considerar que muchas organizaciones usan conexiones SAN con

cable UTP otras con fibra óptica; esta última se caracteriza por la alta velocidad de transmisión de información (Jia-Jun Zhu, 2006).

El canal de fibra es de gran fiabilidad, a causa de la tecnología de interconexión en gigabytes que provee la comunicación simultánea entre distintas estaciones de trabajo, mainframes, servidores, sistemas de almacenamiento y otros periféricos de entrada o salida. De ahí que el canal de fibra es ideal para mover grandes volúmenes de datos a través de largas distancias rápidamente y de forma fiable. La velocidad del puerto de fibra ahora ha alcanzado 4 GB para la transferencia de datos; aunque en ocasiones las limitaciones de SAN se deben al uso de cable UTP y del protocoló IP, esto se ha superado con el canal de fibra que da un mayor alcance y con un funcionamiento estable entre los dispositivos (Cunhe, 2002). El rasgo característico de esta arquitectura es el costo que sigue siendo demasiado alto para el uso general (Chao-Tung Yang, 2014).

Por su parte, el rendimiento es mayor con el uso del canal de fibra en comparación con las conexiones de cable par trenzado. En contraste, cuando se necesita gestionar el sistema de almacenamiento SAN será complejo por la gran cantidad de información, además del alto costo de la infraestructura para el uso de fibra óptica. Así SAN FC es adecuado para grandes unidades de información que tienen mayores presupuestos y requerimientos altos de transferencia y transmisión de datos.

2.5.1.4. Almacenamiento basado en la nube

El desarrollo del almacenamiento de datos en la nube, mejor conocido como cloud computing, se da gracias al uso de equipos virtuales (B. Furht, 2011); implica una infraestructura informática invisible para el usuario, pero al utilizarla parece que se tuviera un equipo físico real, permitiendo la gran ventaja de determinar el número de procesamiento, el sistema operativo, el tamaño de memoria RAM y de disco de almacenamiento. Esta elasticidad en la infraestructura es una de las técnicas usables en Big data (S. Sakr, 2011); (Eric E. Schadt, 2010) las tecnologías de virtualización han hecho que la computación sea accesible, asequible y rentable.

El nombre de *cloud computing* proviene de la utilización del símbolo con forma de nube o *cloud*, que es el diagrama usado en sistemas como una abstracción para determinar internet, mientras que computing implica la informática. Una buena definición de esto es la que ha formulado el NIST (Peter Mell, 2011), (Aguilar, 2012), que la considera como un modelo que permite el acceso bajo demanda a través de la red a un conjunto compartido de recursos de computación configurables (como por ejemplo red, servidores, almacenamiento, aplicaciones y servicios) que pueden ser rápidamente aprovisionados con el mínimo esfuerzo de gestión o interacción del proveedor del servicio.

Una bondad de los entornos de la nube es que, sin duda, proporciona una posible herramienta para el almacenamiento de grandes volúmenes de datos. El almacenamiento en la nube o cloud storage es el espacio para acopiar datos, información, objetos digitales, y otros, que se acceden por internet a través de un servicio web, mediante un navegador como Explorer, Firefox, Chrome o Safari. Además de un aprovisionamiento de recursos informáticos bajo demanda, con control variable para el usuario y neutrales ante sistemas operativos (Sosinsky, 2011), estas características hacen único al almacenamiento en la nube. Hay que tener en cuenta que el almacenamiento puede ser brindado por un proveedor de servicios (nube pública) o una versión privada (nube privada); esta última es creada por una organización particular para su uso interno, con un completo control de los recursos en tecnologías de información.

El servicio de almacenamiento en la nube significa que un proveedor renta espacio en su centro de almacenamiento a usuarios finales que carecen de almacenamiento propio o no desean adquirirlo. También, se usa cuando no se dispone de personal técnico especializado en la administración de sistemas informáticos, o cuando se adolece de conocimiento para implementar y mantener infraestructura en almacenamiento.

Las tecnologías del cloud computing ofrecen principalmente tres modelos de servicio, de acuerdo al NIST (Peter Mell, 2011):

- ✓ La infraestructura como servicio (laaS)
- ✓ La plataforma como servicio (Paas)
- ✓ El software como servicio (SaaS).

No obstante, para fines de almacenamiento en la nube el más adecuado es la modalidad laaS, a causa de que el proveedor ofrece al usuario recursos como capacidad de procesamiento, de almacenamiento, o comunicaciones, que el usuario puede utilizar para ejecutar cualquier tipo de software, desde sistemas operativos hasta aplicaciones.

Una de las grandes ventajas del almacenamiento en la nube es el ahorro de recursos económicos. El almacenamiento se alquila a un proveedor utilizando el modelo de pago por gigabyte almacenado o pago por unidades de datos transferidos. Pues el usuario únicamente paga por la cantidad de datos que transfiere y aloja en los servidores del proveedor, además se tiene que el almacenamiento en la nube permite buena extensibilidad y escalabilidad en el almacenamiento de la información, necesario cuando se manejan grandes cantidades de datos (Aguilar, 2012) .La nube no tiene delimitaciones geográficas como los países, implicando que nuestra información puede acabar siendo deslocalizada en una o varias regiones del mundo, en uno de los centros de procesos de datos del proveedor de servicios.

2.6. SISTEMAS DE PROCESAMIENTO EN BIG DATA

El objetivo de Big Data es procesar de forma integrada los datos dinámicos y estáticos, así como estructurados y no estructurados, un reto que requiere desarrollar arquitecturas de procesamiento con escalabilidad y extensibilidad extrema, capaz de integrar datos almacenados y datos en movimiento y que soporten la aplicación de técnicas de análisis sobre ambos tipos de datos.

El proceso de extraer información y descubrir conocimiento de flujos de datos cuya generación ocurre de forma rápida e ininterrumpida. Tales flujos pueden ser los producidos por el tráfico de la red, bitácoras del sistema, llamadas telefónicas, transacciones bancarias y datos provenientes de sensores (cámaras de video, RFID, entre otros). Las arquitecturas o plataformas diseñadas para el Sistema de análisis de datos tienen como principal requerimiento el de garantizar altas velocidades de desempeño, es decir, los datos deben ser procesados durante períodos pequeños de tiempo (cercano al tiempo real). También deben satisfacer exigencias de análisis distribuido donde los operadores procesarían los datos en línea, es decir, sin previo almacenamiento de estos. Mantener los valores de latencias bajos sin comprometer la escalabilidad es uno de los principales desafíos en dichas arquitecturas

Se definen tres tipos de arquitecturas generales orientadas al sistema de análisis de datos (Michael Stonebrake, 2005):

- **Sistema de gestión de bases de datos (DBMS, por sus siglas en inglés).** Es un sistema que permite el almacenamiento, modificación y extracción de la información registrada en una base de datos; además, proporcionan las herramientas necesarias para analizarlos. Permitiendo un amplio número de funciones relacionadas con la **seguridad de las bases de datos**. Por un lado, controlar el acceso a las mismas, asegurar su integridad. Y, por último, recuperar los datos tras un fallo del sistema y hacer copias de seguridad.

- **Motores de reglas** Capas de trabajar eficientemente con grandes volúmenes de información (Big Data), proporciona una interface para la elaboración de las reglas de análisis y detección en lenguaje natural. Esta facilidad permite a los propietarios analistas generar sin intermediarios sus estrategias de detección con eficiencia, agilidad y reducción de costos

- **Motores de procesamiento de flujo (spes, por sus siglas en inglés)** Es el que permite llevar a cabo tareas mediante el análisis de un flujo continuo y potencialmente infinito de datos, brindando respuesta en tiempos muy cercanos al tiempo real. La principal característica de procesamiento es que los datos de. flujo llegan a una velocidad tal que no es posible almacenarlos en su totalidad y si se pueden almacenar, el volumen de datos es tan grande que presenta la dificultad de analizarlos en tiempo de respuesta corta.

3. EVALUACIÓN DE CAPACIDADES DE ALMACENAMIENTO Y PROCESAMIENTO DE GRANDES VOLÚMENES DE INFORMACIÓN

En este capítulo se presenta un método de prueba y evaluación de capacidades tecnológicas, el cual puede orientar el uso de plataformas Big Data cuando se requiera trabajar con grandes volúmenes de datos, en lo que respecta al almacenamiento y procesamiento de información para el caso particular del sector salud.

3.1 Definición del Método

Para lograr el objetivo general se utilizó un ciclo de vida de analítica de Big Data (Paul Buhler, 2016), Este ciclo de vida consta de nueve fases, de las cuales solo algunas de ellas serán utilizadas para hacer énfasis en el objetivo de este proyecto referente a la prueba y evaluación de dos alternativas de almacenamiento y procesamiento a considerar (DBMS vs Hadoop/Hive).

El ciclo de vida contiene las siguientes fases:

1. Caso de negocio a ser evaluado (Business Case Evaluation)
2. Identificación de datos (Data Identification)
3. Adquisición de datos y filtrado (Data adquisition & Filtering)
4. Extracción de datos (Data extraction)
5. Validación de datos y limpieza (Data validation & Cleansing)
6. Agregación de datos y representación (Data aggregation & representation)
7. Análisis de datos (Data analysis)
8. Visualización de datos (Data visualization)
9. Utilización de análisis de resultados (Utilization of analysis results)

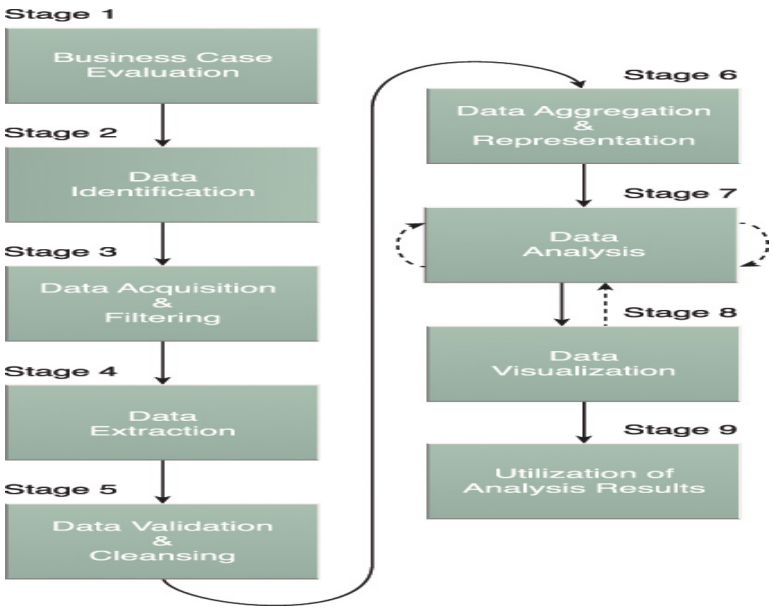

Figura 6. Ciclo de vida de analítica big data (Big Data Fundamentals)

3.2 Fases

Fase1: Caso de negocio a ser evaluado

Actualmente las empresas se en dilema si realmente necesitan entrar en el contexto de Big Data y analítica avanzada. Si se analiza este problema desde las 5V's, dos de ellas requieren de atención. Estas son Volumen y Velocidad, ya que se requeriría que la empresa hiciera una evaluación de capacidades para analizar si se tiene el volumen de datos requerido para realizar las analíticas que requieren grandes volúmenes de información en tiempos razonables. Actualmente la mayoría de las empresas realizan inteligencia de negocios desde hace mucho tiempo con volúmenes de datos controlado y en motores de bases de datos tradicionales.

Para el caso concreto de este proyecto y en el contexto del sistema de salud de Texas en Estados Unidos, cuenta con los datos de ingreso a salas de emergencia (pacientes externos), se registran muchos datos entre ellos, los más importantes son los códigos de diagnósticos de la atención, y cuales son atendidos por una real urgencia, quienes son devueltos, quienes pasan a hospitalización (pacientes internos) entre otra información.

Esta información permite medir la eficiencia de la atención, diagnósticos más recurrentes entre otros. Si bien esta analítica es descriptiva, permite evaluar las alternativas de almacenamiento y procesamiento de interés en este proyecto.

Este caso de negocio, esta principalmente centrado en medir el uso de las salas de emergencia en el estado de Texas en los Estados idos y paso a hospitalización para pacientes internos y externos[1]

Fase 2: Identificación y entendimiento de los datos

Para lograr identificar las características de volumen y velocidad de plataformas de gestión de datos tradicionales (estructurados – DBMS) vs. Big Data basado en HDFS y Hive o SparkSQL, se definieron los datos de los pacientes internos y externos del hospital de Texas Health Care Information Council (THCIC) del año 2010 y 2011.

Estos datos fueron clasificados en cuatro periodos correspondientes a cada trimestre del año.

Se utilizarán los datos Texas Health Care Information Council (THCIC) para pacientes internos (https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm) y para pacientes externos (https://www.dshs.texas.gov/thcic/OutpatientFacilities/OutpatientPudf.shtm) que es una Institución recolectora de datos hospitalarios de todos los hospitales con licencia

[1] Hospital Emergency Room Data Collection (Departament of State Health Services december 2010-2011)

estatal. Desafortunadamente en Colombia no se pudo aplicar este proyecto ya que no se cuenta con datos abiertos sobre los hospitales y centros de atención de **pacientes internos y externos.**

THCIC pasó a formar parte del Departamento de Servicios de Salud del Estado de Texas (DSHS) a partir del 1 de septiembre de 2004 y el Centro de Estadísticas de Salud del DSHS es ahora el responsable de la recopilación y publicación de los datos hospitalarios.

El Departamento de Servicios de Salud del Estado de Texas recopila y divulga información sobre las altas hospitalarias en hospitales en Texas y las visitas ambulatorias trimestralmente. Actualmente hay registros de más de 3 millones de paciente y 12 millones de visitas anuales de los pacientes. Los datasets de DSHS contienen información de nivel de paciente para estancias hospitalarias internas y ambulatorias individuales e incluyen datos sobre diagnósticos, procedimientos, cargos y datos demográficos de los pacientes. **RPC** (Consultores de Investigación y planificación) tiene estos archivos desde 1999 hasta el último trimestre publicado.

El Departamento Estatal de Servicios de Salud de Texas recopila y publica información trimestral sobre las altas hospitalarias y las visitas ambulatorias de pacientes internos y externos hospitalizados en Texas. Estos datos incluyen información sobre datos de diagnósticos, procedimientos, cargos y demografía del paciente. Algunos datos útiles de las tablas de paciente interno y externo incluyen: Género del paciente, Grupo de edad del paciente, Origen de la admisión condado del paciente y código postal, dos fuentes de pago, 25 códigos de diagnóstico (IC-D-9), y 25 códigos de procedimientos quirúrgicos.

Para proteger la identidad del paciente, los datos de alta del Hospital de Texas suprimen ciertos campos en los casos en que su inclusión puede poner en peligro el anonimato del paciente, y también suprimir las identidades del médico.

Fase 3: Adquisición de datos y Filtrado

Esta fase inicia con la descarga de los datos del funcionamiento del servicio del sitio web, permitiendo reportar la actividad del hospital, identificando las distintas variables que conforman el servicio, luego estos datos se someterán a un filtrado para descartar aquellos datos que no tienen ningún valor agregado.

Los datasets originales del Hospital de Texas Health Care Information Council (THCIC) del año 2010-2011, son archivos de datos de uso público (PUDF) del departamento de emergencia que contiene datos de los pacientes internos y pacientes externos. Estos archivos se almacenarán en el servidor “frontend” en la plataforma de experimentación.

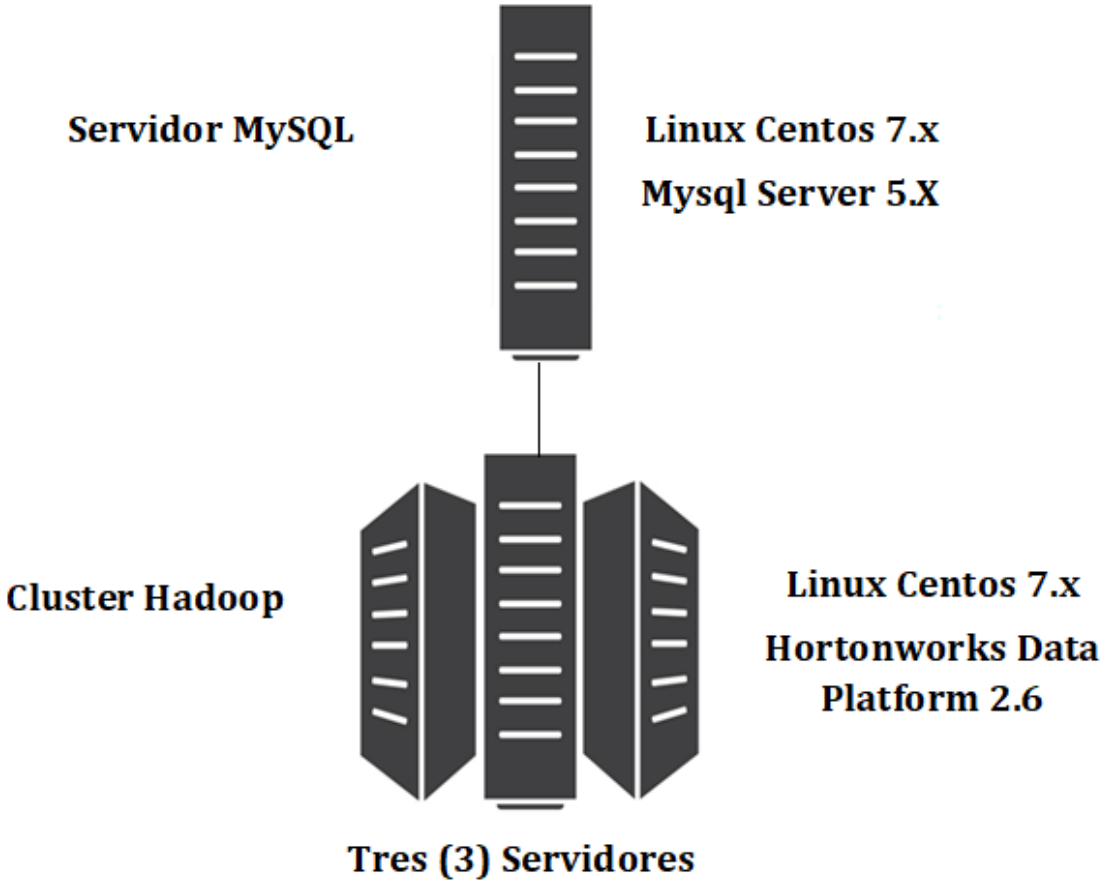

Figura 7. Plataforma de implementación

Los archivos del dataset información de los pacientes hospitalizados. El archivo de datos base PUDF se divide en dos:

1. **Dataset # 1:** El archivo contiene los datos necesarios del paciente (Con este se realizan las pruebas).
2. **Dataset # 2:** Contiene los datos esencialmente primordiales del paciente y algunos campos establecidos.

- El ID de registro permite vincular los archivos juntos.
- Los proveedores tienen, por ley, hasta el próximo trimestre (Tras la aprobación de la gestión) para presentar sus datos. Esto significa que los datos PUDF son instantáneos en el tiempo y cada trimestre puede contener algunas descargas fechadas en el trimestre anterior (es decir, para el año calendario los datos se aseguran que corresponden al primer trimestre del año).
- También se incluyen datos estadísticos, el cual contiene 10 variables que incluyen el ID THCIC, y el proveedor. Estas variables indican si la institución es un centro de enseñanza, un centro hospitalario pediátrico u otra especialidad.

Además, este archivo incluye comentarios del remitente. Contiene cualquier comentario cuando los datos respectivos que fueron presentados y certificados por una institución.

El PUDF está disponible en cuatro archivos de texto de formato de longitud fija o separado por tabs (TSV), El Dataset # 1 (registro lógico, su longitud es de 950 bytes), Dataset #2 (registro lógico, su longitud es de 700 bytes), cargos (registro lógico, su longitud es de 80 bytes), y datos del tipo de instalación (registro lógico, su longitud es de 69 bytes). Los datasets publicados después de 2004 incluyen 255 campos de datos en un archivo de datos base y 13 campos de datos en un archivo de cargos detallados. Los archivos de datos de años anteriores a 2004 incluyen solo 205 campos de datos. Los archivos también están disponibles en formato delimitado por tabuladores.

Después de analizar los datos, se filtraron los siguientes atributos:

- **Género del paciente**
- **Grupo de edad del paciente**
- **Fuente de admisión**
- **Condado del paciente y código postal**
- **Pagador principal**
- **Códigos de diagnóstico**
- **Códigos de procedimiento quirúrgico**
- **Estado de Descarga**

Fase 4: Extracción de los datos

Esta fase se inicia extrayendo los datasets del servidor del Hospital de Texas Health Care Information Council (THCIC) de los años 2010 - 2011 e almacenándolos en el servidor "frontend", estos datasets se encuentran en formato tipo texto para realizar la migración hacia la base de datos de Mysql y cluster Hadoop.

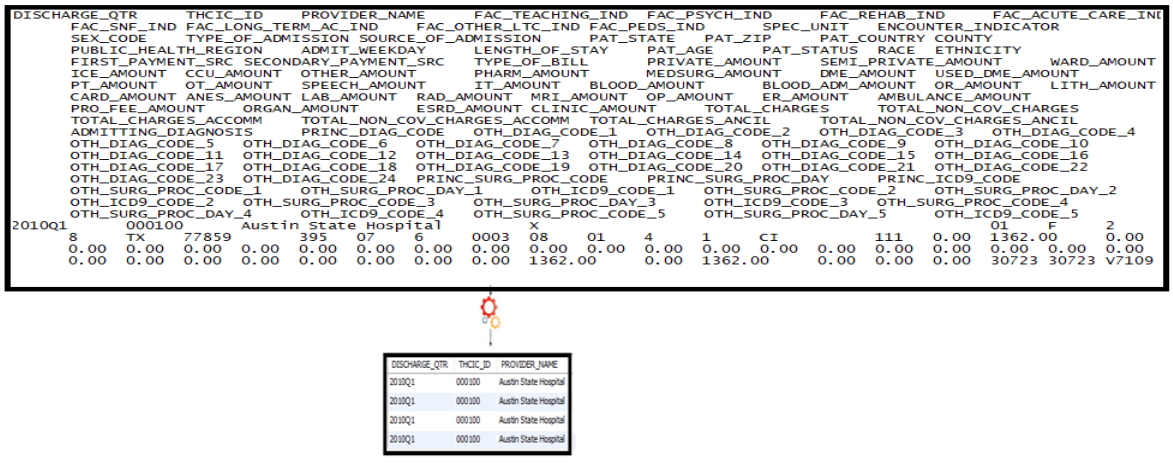

Figura 8. Discharge _QTR extraído de un documento tipo texto

- **Fase 5: Validación de los datos y limpieza**

  Una vez culminado el proceso anterior de la extracción de los datos y filtrado, ya para esta fase tenemos la base de datos consolidada con 255 campos de los cuales; 27 son numéricos, y 228 son alfanuméricos.

  Para realizar el proceso de validación y consistencia de los datos obtenidos se verifico que tanto el tamaño y la longitud del archivo fueran iguales en la extracción y carga.

  Se verifico con el diccionario de datos las condiciones de cada variable como son:
  - Longitud
  - Rango de valor
  - Tipo de dato
  - Obligatoriedad del campo
  - Restricciones entre otros

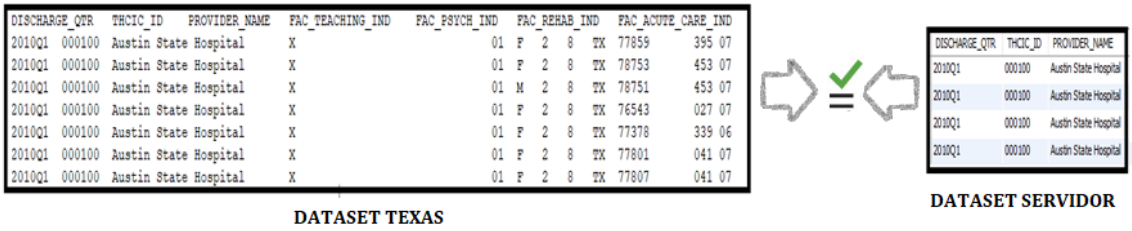

Figura 9. Validación de datos

Fase 6: Agregación de los datos y representación.

En esta fase se realizar el proceso de organización de datos donde los datos son buscados, recopilados y representados en un formato resumido, basado en informes para garantizar que solo se analice los datos correctos en la siguiente fase.

Por medio de la agregación se genera una nueve tabla temporal de información de la tabla origen no agregado, que respalda la posibilidad de agregar datos directamente desde la instrucción dada y de este modo es más directo y fácil visualizar esta información que es mostrada en la nueva tabla.

```
SELECT DISCHARGE, PRINC_DIAG_CODE, COUNT(*) AS CANTIDAD,@NUM := IF(@DISCHARGE = DISCHARGE, @NUM + 1,1) AS ROW_NUMBER, @DISCHARGE := DISCHARGE AS DUMMY FROM PACIENTE_INTERNOS GROUP BY DISCHARGE,
```

```
PRINC_DIAG_CODE HAVING ROW_NUMBER <= 5 ORDER BY DISCHARGE ASC, CANTIDAD DESC;
```

Fase 7. Análisis de los datos

El análisis de los datos es el proceso de examinar diferentes variables para descubrir patrones ocultos, correlaciones desconocidas y otra información útil.

Para la evaluación de las pruebas se utilizará dos plataformas de gestión de datos: una tradicional y la otra la plataforma Big Data basadas en el ecosistema Hadoop/Spark con soporte en SQL.

Debido a eso es necesario tomar en cuenta que para la preparación y la toma tiempos tendremos que procesar probablemente grandes cantidades de información.

La solución propuesta debe ser escalable en el tiempo y entre los diferentes tipos de aplicaciones, es por esta razón que es importante tomar en cuenta la utilización de herramientas para Big Data y DBMS.

Fase 8 y 9. NO APLICA PARA ESTE MODELO DE PLATAFORMA

4. DISEÑO DEL EXPERIMENTO

Para poder comparar el desempeño de los ambientes de Hive y MySQL se realizaron mediciones de tiempo de ejecución de procesamiento en los distintos ambientes, y de este modo poder evaluar y tomar una decisión orientada al almacenamiento y procesamiento de datos. Para medir el rendimiento de las bases de datos en ambos ambientes, se diseñaron varias consultas que analizan los datos desde el punto de vista descriptivo para tener una misma línea base. El factor más importante para una aplicación es el tiempo requerido para completar una tarea, y en el caso de las plataformas de gestión de datos tradicionales (estructurados – DBMS) vs. BigDataSql de los datos es el tiempo requerido para completar una consulta. Se realizan **tres** mediciones de tiempo en consultas y analíticas para obtener resultados más confiables del comportamiento de los dos ambientes. El tiempo de respuesta se medirá en segundos.

Las pruebas fueron ejecutadas de manera secuencial y exclusiva (todo el servidor y cluster dedicado a la consulta para asegurar el acceso y recursos exclusivos y evitar una sobrecarga de tareas en paralelo. Luego de cada ejecución se verificó el estado de los nodos y la integridad de los datos.

Las pruebas se ejecutan para dos tipos de pacientes:

- ✓ Paciente interno
- ✓ Paciente externo

Por otra parte, el análisis de los datos es para comparar y evaluar el tiempo de respuesta y la cantidad datos procesados, para evaluar la hipótesis ¨Una arquitectura de referencia para la gestión de altos volúmenes de información en el sector salud, puede mejorar la calidad de los servicios asistenciales en diferentes entidades de salud¨.

Para realizar las pruebas y evaluación de las consultas en las plataformas de gestión para las dos alternativas de almacenamiento y procesamiento se utiliza consultas de selección, agrupación y ordenación.

En las consultas realizamos un script seleccionando el **código principal** en nuestra base de datos, para nuestro caso fue Código de Diagnóstico más frecuente por visitas a urgencias durante el año 2010-2011 en paciente interno y externo.

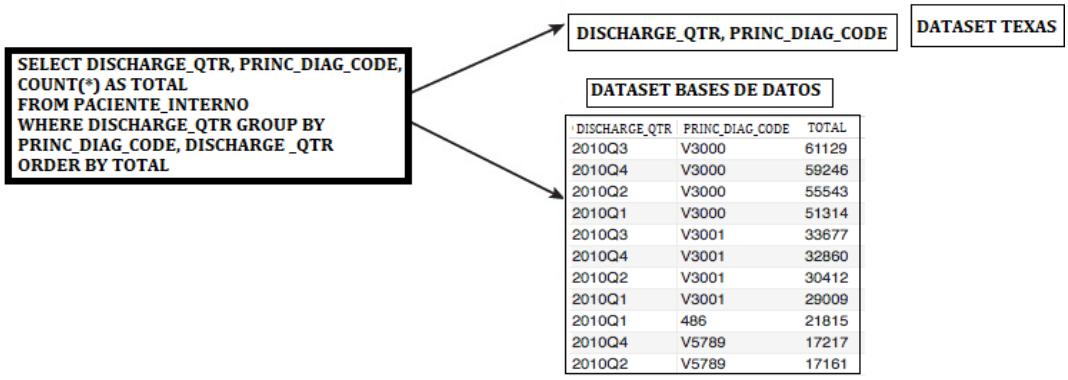

Figura 10. Estructura de datos con una solución Big Data

Consultas de agrupación y ordenación

Cada uno de los tiempos de consulta representa el promedio de tres ejecuciones. En el siguiente capítulo se presentan las tablas, gráficos e imágenes de los resultados obtenidos de la ejecución de las distintas pruebas divididas por tipo.

QUERY N°	MYSQL-HIVE (PACIENTE INTERNO Y EXTERNO)
1	SELECT count(*) FROM paciente_interno WHERE princ_diag_code = "4019"; SELECT count(*) FROM paciente_externo WHERE princ_diag_code = "4019"; Esta consulta realiza una agrupación por año y trimestre, informando el código de diagnóstico y la cantidad encontrada en forma descendiente para pacientes internos, Para realizar la consulta para paciente externo unicamente se cambia el nombre de la tabla y el trimestre a consultar en la base de datos
	SELECT count(*) FROM paciente_interno WHERE FIRST_PAYMENT_SRC = "MC"; SELECT count(*) FROM paciente_externo WHERE FIRST_PAYMENT_SRC = "MC"; Esta consulta se para averiguar la cantidad de pacientes internos y externos atendidos en la sala de emergencia de Texas por entidad pagadora.
3	SELECT COUNT(*) AS CANTIDAD FROM PACIENTE_INTERNOS GROUP BY DISCHARGE = "2010Q1"; SELECT COUNT(*) AS CANTIDAD FROM PACIENTE_EXTERNO GROUP BY SERVICE_QUARTER = "2010Q1"; **[En Los Query Unicamente Se Cambia El Periodo De Consulta 2010Q2,2010Q3,2010Q4]** Esta consulta nos informa la cantidad de pacientes internos que visitaron la sala de emergencia por el trimestre solicitado.

	Para realizar la consulta para pacientes externos unicamente se cambia el nombre de la tabla y el periodo a consultar en la base de datos.
4	SELECT MIN(TOTAL_CHARGES) MIN, MAX(TOTAL_CHARGES) MAX FROM PACIENTE_INTERNO; Esta consulta nos informa el valor Máximo y el valor mínimo recibidos por ingresos por pacientes internos. Para realizar la consulta para pacientes externos unicamente se cambia el nombre de la tabla en su base de datos
5	SELECT PRINC_DIAG_CODE FROM PACIENTE_INTERNO ORDER BY PRINC_DIAG_CODE DESC; SELECT PRINC_DIAG_CODE FROM PACIENTE_EXTERNO ORDER BY PRINC_DIAG_CODE DESC; Esta consulta ordena el código de diagnóstico en forma descendiente para paciente interno Para realizar la consulta para paciente externo unicamente se cambia el nombre de la tabla en la base de datos

Tabla 1. Consulta para el análisis de rendimiento de Apache Hive y MySQL

Plataforma Hardware y Software base

El equipo usado en las pruebas de rendimiento para MySQL y Hive posee las siguientes características:

	Hardware	Software base
Servidor MySQL	**Un (1) Servidor.** **16 vCPU** **64 GB RAM** **500 GB HDD**	**Linux Centos 7.x** **Mysql Server 5.X**
Cluster Hadoop	**Tres (3) Servidores** **16 vCPU c/serv.** **64 GB RAM c/serv.** **500 GB HDD c/serv.**	**Linux Centos 7.x** **Hortonworks Data Platform 2.6** **HDFS** **Hive** **Spark 2.x**

Direccion IP pública	192.168.10.80
Procesador:	Core 16(16)
RAM	62.72GB
Disco Duro	37.72 GB/433.91 GB(8.01% usado)
Versión	HDP-2.6.4.0

La siguientes tablas muestran el tiempo promedio de ejecución de las consultas en los dos plataformas de gestión de datos tradicionales, plataformas Big Data basadas en el ecosistema hadoop/spark con soporte en SQL.

Los cinco **mejores** códigos de diagnóstico para las visitas al servicio de urgencias hospitalarias en Texas en el año 2010**.**

El código de diagnóstico más frecuente informado para las visitas a Urgencias durante **el año 2010** para paciente interno y externo fue Hipertensión Esencial código (4019). El segundo código de diagnóstico de pacientes ingresados con mayor frecuencia fue la Hiperlipidemia (2724) para pacientes internos, mientras que para paciente externo fue la Diabetes Mellitus (25000). El tercer código para pacientes internos fue Diabetes mellitus código (25000) y para paciente externo fue es la Hiperlipidemia código (2724) y los dos últimos códigos de diagnóstico de visitas al servicio para los pacientes que ingresaron en el hospital fueron Insuficiencia renal no especificada (4280) y insuficiencia renal aguada no especificada (5849)

CINCO CODIGO DEL DIAGNOSTICO MAS FRECUENTES	CANTIDAD POR PACIENTE INTERNO	CANTIDAD POR PACIENTE EXTERNO
4019, Hipertensión esencial	777.696	939.840
2724, Hiperlipidemia	416.917	273.043
25000, diabetes mellitus tipo 2 sin complicaciones	358.562	432.312
4280, insuficiencia cardíaca congestiva, no especificada	259.076	86.182
5849, insuficiencia renal aguda, no especificada	161.483	8.362

a) ¿Cuál es el código de Diagnóstico más frecuente en las salas de emergencia en Texas para pacientes internos y externos para el año 2010-2011?

CODIGO DEL DIAGNOSTICO PACIENTE INTERNO (**HIVE)**	TIEMPO 1.	TIEMPO 2.	TIEMPO 3.	PROMEDIO
4019 Hipertension esencial	6,83	6,03	6,00	6,29
2724, Hiperlipidemia	6,44	6,43	4,70	5,86
25000, Diabetes mellitus tipo 2 sin complicaciones	6,37	6,58	6,03	6,33
4280, Insuficiencia cardíaca congestiva, no especificada	6,19	5,29	6,01	5,83
5849, Insuficiencia renal aguda, no especificada	5,76	5,97	5,90	5,88

Tabla 2. LOS 5 CODIGOS MÁS FRECUENTES PACIENTE INTERNO CON HIVE

CODIGO DEL DIAGNOSTICO PACIENTE INTERNO (MYSQL**)**	TIEMPO 1.	TIEMPO 2.	TIEMPO 3.	PROMEDIO
4019 Hipertension esencial	14,63	15,33	13,52	14,49
2724, Hiperlipidemia	13,52	14,00	13,71	13,74
25000, Diabetes mellitus tipo 2 sin complicaciones	13,52	13,54	13,49	13,52
4280, Insuficiencia cardíaca congestiva, no especificada	13,52	13,67	13,46	13,55
5849, Insuficiencia renal aguda, no especificada	14,41	13,57	14,44	14,14

Tabla 3. LOS 5 CODIGOS MÁS FRECUENTES PACIENTE INTERNO CON MYSQL

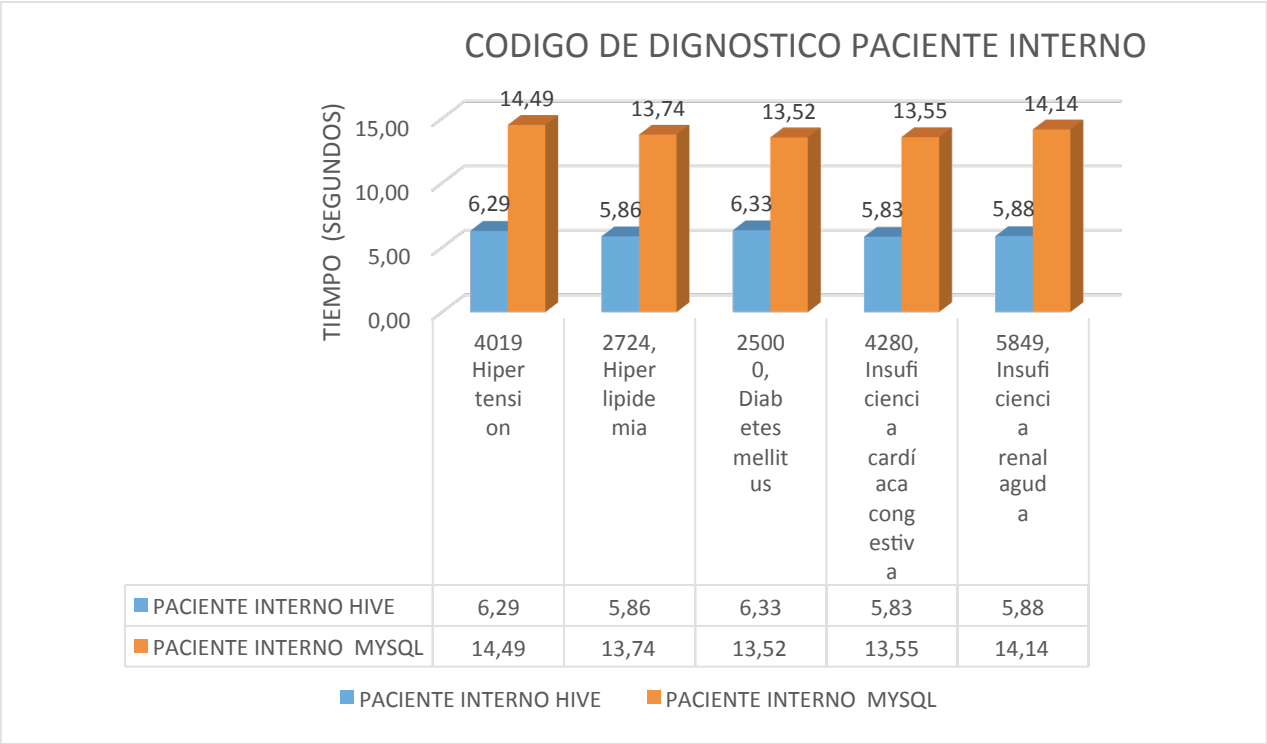

Figura 11. LOS 5 CODIGOS DE DIAGNOSTICOS MÁS FRECUENTES P-INTERNO

CODIGO DEL DIAGNOSTICO PACIENTE EXTERNO **(HIVE)**	TIEMPO 1.	TIEMPO 2.	TIEMPO 3.	PROMEDIO
4019 Hipertension esencial	6,19	6,78	5,85	6,27
2724, Hiperlipidemia	6,43	5,86	5,95	6,08
25000, Diabetes mellitus tipo 2 sin complicaciones	6,20	6,85	7,85	6,97
4280, Insuficiencia cardíaca congestiva, no especificada	6,56	6,54	4,93	6,01
5849, Insuficiencia renal aguda, no especificada	6,67	5,48	5,9	6,02

Tabla 4. LOS 5 CODIGOS MÁS FRECUENTES PACIENTE EXTERNO CON HIVE

CODIGO DEL DIAGNOSTICO PACIENTE EXTERNO (MYSQL)	TIEMPO 1.	TIEMPO 2.	TIEMPO 3.	PROMEDIO
4019 Hipertension esencial	31,31	32,25	31,54	31,70
2724, Hiperlipidemia	31,98	32,27	33,23	32,49
25000, Diabetes mellitus tipo 2 sin complicaciones	30,63	30,75	33,50	31,63
4280, Insuficiencia cardíaca congestiva, no especificada	31,40	30,45	32,11	31,32
5849, Insuficiencia renal aguda, no especificada	32,23	31,31	31,52	31,69

Tabla 5. LOS 5 CODIGOS MÁS FRECUENTES PACIENTE EXTERNO CON MYSQL

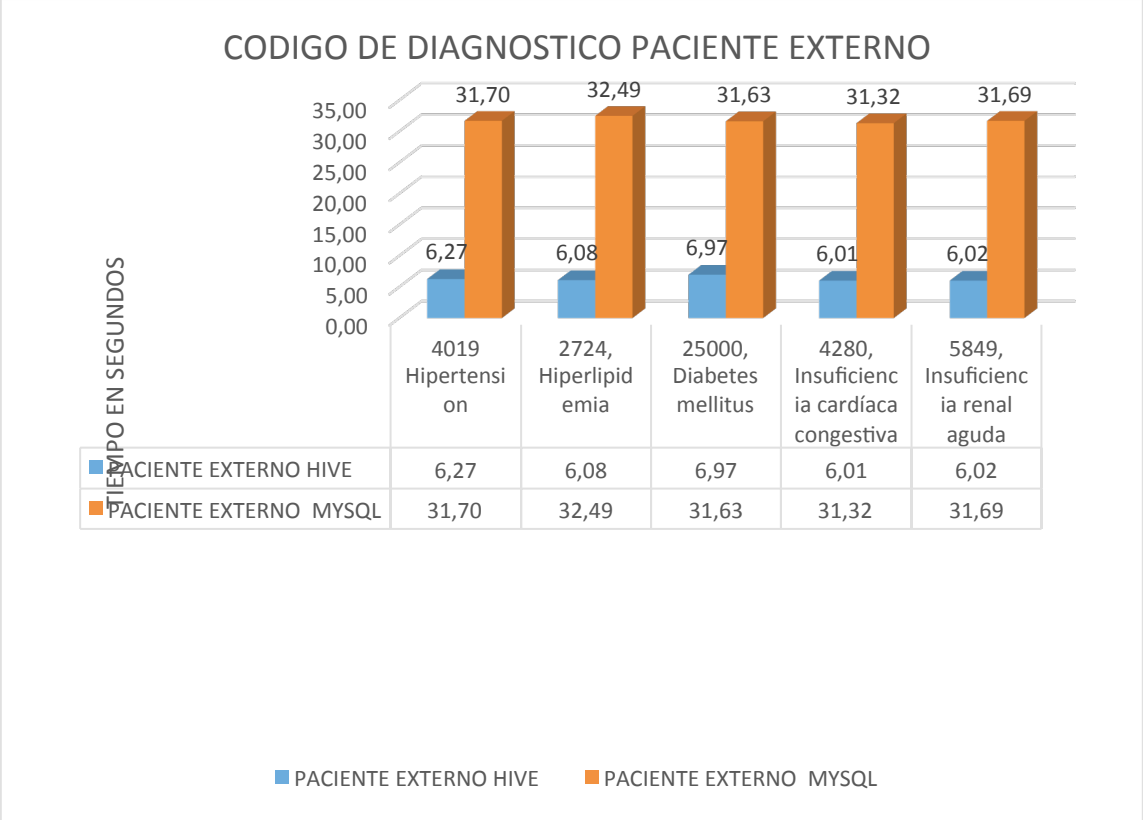

Figura 12. LOS 5 CODIGOS DE DIAGNOSTICOS MÁS FRECUENTES P-EXTERNO

b) ¿Cuál es la cantidad de pacientes internos y externos atendidos en la sala de emergencia de Texas por entidad pagadora?

HIVE					
PACIENTE INTERNO	CANTIDAD	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
MEDICARD (MC)	1.274.264	7,11	5,89	6,33	6,44
MEDICARE(MB)	23.460	5,83	5,54	6,21	5,86
OTROS (OF)	23.989	4,63	4,68	6,68	5,33
PRIVADO (09)	517.763	6,32	6,01	5,07	5,80
SIN SEGURO(15)	192.264	5,64	4,64	4,53	4,94

Tabla 6, PACIENTES INTERNOS POR ENTIDAD PAGDORA CON HIVE

MYSQL					
PACIENTE INTERNO	CANTIDAD	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
MEDICARD (MC)	1.274.264	14,89	15,12	15,17	15,06
MEDICARE(MB)	23.460	15,10	14,97	14,74	14,94
OTROS (OF)	23.989	14,95	14,03	14,97	14,65
PRIVADO (09)	517.763	15,01	16,54	15,09	15,55
SIN SEGURO(15)	192.264	14,92	15,01	14,94	14,96

Tabla 7. PACIENTES INTERNOS POR ENTIDAD PAGADORA CON MYSQL

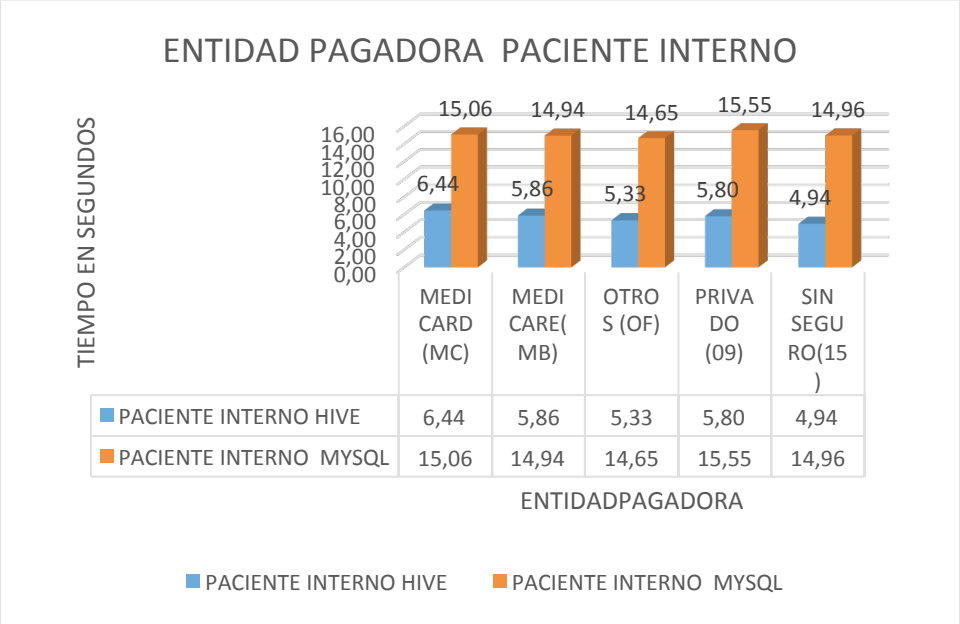

Figura 13. CANTIDAD PACIENTES POR ENTIDAD PAGADORA P. INTERNO

HIVE					
PACIENTE EXTERNO	CANTIDAD	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
MEDICARD (MC)	1.788.537	6,26	5,68	6,69	6,21
MEDICARE(MB)	1.772.241	6,33	5,05	6,75	6,04
OTROS (OF)	54.221	5,49	5,53	4,86	5,29
PRIVADO (09)	1.571.381	7,17	5,50	6,12	6,26
SIN SEGURO(15)	424.685	6,85	5,13	6,82	6,27

Tabla 8. . PACIENTES EXTERNOS POR ENTIDAD PAGADORA CON HIVE

MYSQL					
PACIENTE EXTERNO	CANTIDAD	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
MEDICARD (MC)	1.788.537	36,30	36,13	36,92	36,45
MEDICARE(MB)	1.772.241	35,32	35,26	35,67	35,42
OTROS (OF)	54.221	36,33	35,98	35,94	36,08
PRIVADO (09)	1.571.381	36,37	35,40	34,71	35,49
SIN SEGURO(15)	424.685	36,08	36,69	36,89	36,55

Tabla 9. PACIENTES EXTERNOS POR ENTIDAD PAGADORA CON MYSQL

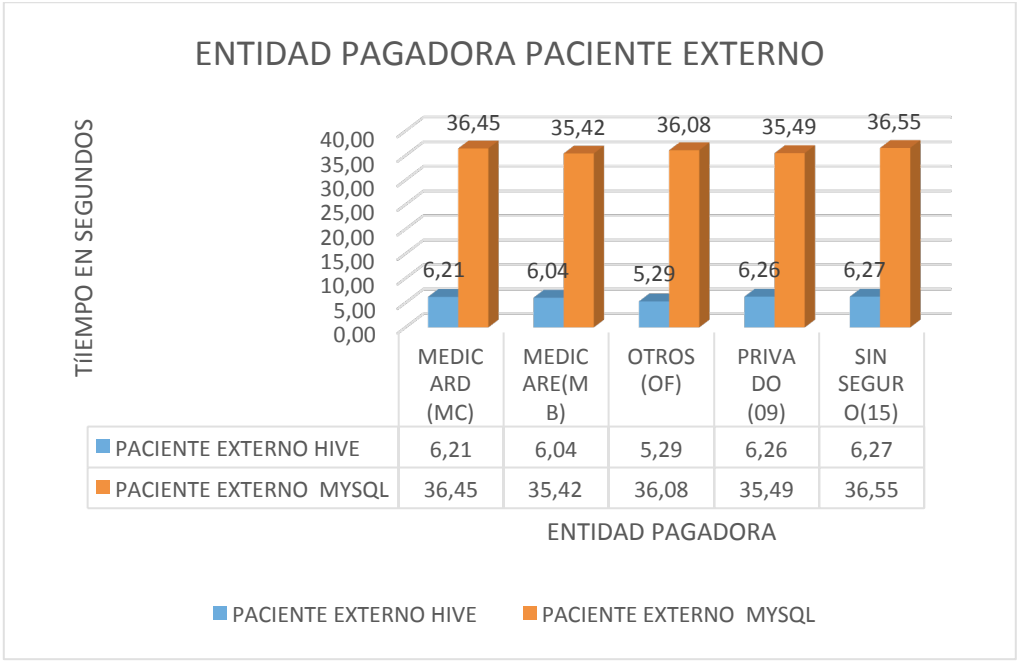

Figura 14. CANTIDAD PACIENTES POR ENTIDAD PAGADORA P.EXTERNO

c) Cual es la cantidad de pacientes internos y externos atendidos en la sala de urgencias en Texas durante los años 2010-2011?

HIVE				
PACIENTE INTERNO	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
5.874.517	5,64	4,90	5,22	5,25

Tabla 10. CANTIDAD DE PACIENTES INTERNOS ATENDIDOS CON HIVE

MYSQL				
PACIENTE INTERNO	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
5.874.517	12,53	11,26	11,33	11,71

Tabla 11. CANTIDAD DE PACIENTES INTENROS ATENDIDOS CON MYSQL

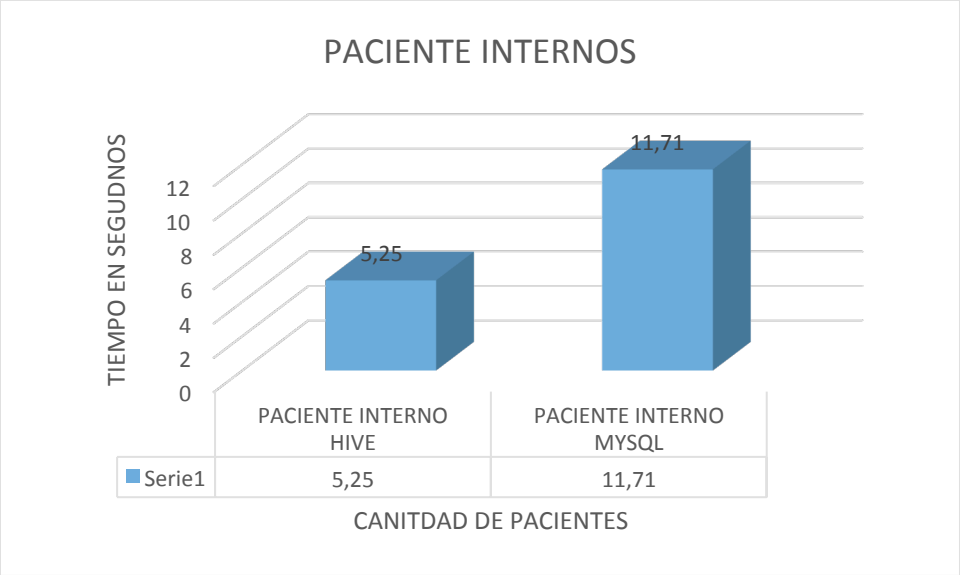

Figura 15. PACIENTES ATENDIDOS EN LA SALA DE EMERGENCIA

HIVE				
PACIENTE EXTERNO	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
17.048.893	5,43	5,18	5,53	5,38

Tabla 12, CANTIDAD DE PACIENTES EXTERNOS ATENDIDOS CON HIVE

MYSQL				
PACIENTE EXTERNO	TIEMPO 1	TIEMPO 2	TIEMPO 3	PROMEDIO
17.048.893	25,58	24,08	24,29	24,65

Tabla 13. CANTIDAD DE PACINES EXTERNOS ATENDIDOS CON MYSQL

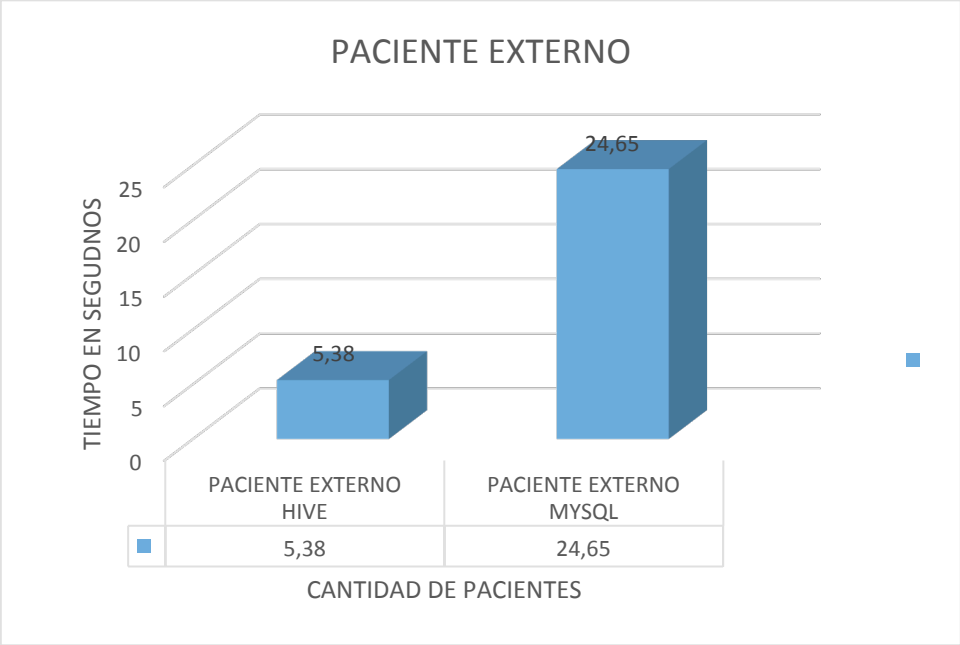

Figura 16. CANTIDAD DE PACIENTES ATENDIDOS EN LA SALA DE EMERGENCIA

d). ¿Cuál es el tiempo requerido para encontrar el valor máximo y mínimo de ingresos recibos por paciente durante los años 2010-2011?

HIVE				
PACIENTE INTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
MAXIMO Y MINIMO	8,81	8,91	8,61	8,78

Tabla 14. MAXIMO Y MINIMO VALOR DE INGRESO CON HIVE P. INTERNO

MYSQL				
PACIENTE INTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
MAXIMO Y MINIMO	14,19	13,68	13,92	13,93

Tabla 15. MAXIMO Y MINIMO VALOR DE INGRESO CON MYSQL P. INTERNO

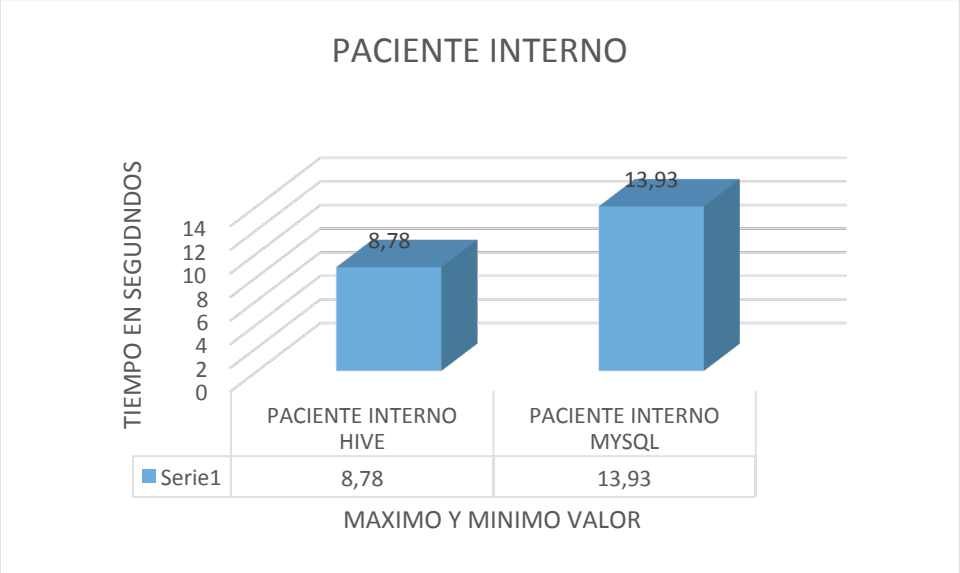

Figura 17. MAXIMO Y MINIMO VALOR DE INGRESOS RECIBIDOS P.INTERNO

HIVE				
PACIENTE EXTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
MAXIMO Y MINIMO	12,75	13,07	9,02	11,61

Tabla 16. MAXIMO Y MINIMO VALOR DE INGRESO CON HIVE P. EXTERNO

MYSQL				
PACIENTE EXTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
MAXIMO Y MINIMO	32,09	32,87	31,95	32,30

Tabla 17, MAXIMO Y MINIMO VALOR DE INGRESO CON MYSQL P. EXTERNO

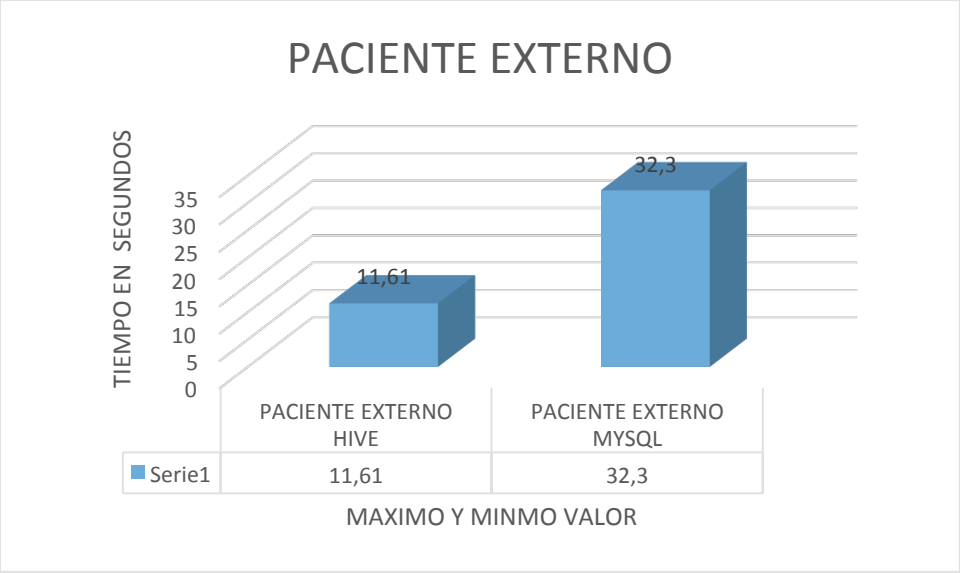

Figura 18. MAXIMO Y MINIMO VALOR DE INGRESOS RECIBIDOS P.EXTERNO

e). ¿Cuál es el código de diagnóstico más relevante en pacientes interno y externos para el periodo 2010-2011?

HIVE				
CODIGO DE DIAGNOSTICO PACIENTE INTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
V7612	9,55	9,52	9,53	6,36

Tabla 18. CODIGO MÁS RELEVANTE P.INTERNO CON HIVE

MYSQL				
CODIGO DE DIAGNOSTICO PACIENTE INTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
V7612	17,71	17,77	18,37	17,95

Tabla 19. CODIGO MÁS RELEVANTE P.INTERNO CON MYSQL

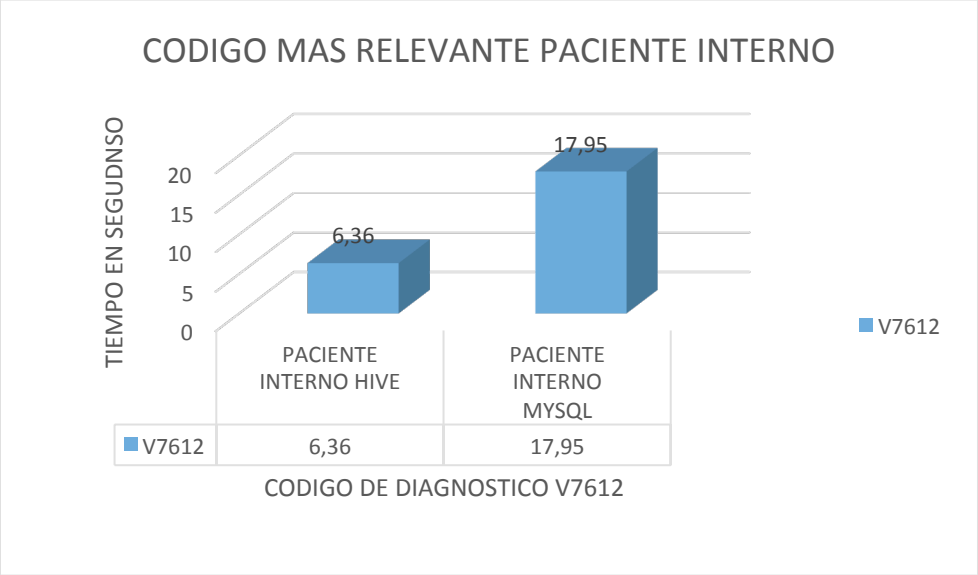

Figura 19. CODIGO DE DIGNOSTICO MÁS RELEVANTE P. INTERNO

HIVE				
CODIGO DE DIAGNOSTICO PACIENTE EXTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
V7612	9,49	9,51	9,56	9,52

Tabla 20. CODIGO MÁS RELEVANTE P EXTERNO CON HIVE

MYSQL				
CODIGO DE DIAGNOSTICO PACIENTE EXTERNO	**TIEMPO 1**	**TIEMPO 2**	**TIEMPO 3**	**PROMEDIO**
V7612	44,60	44,98	44,78	44,79

Tabla 21. CODIGO MÁS RELEVANTE P.IEXTERNO CON MYSQL

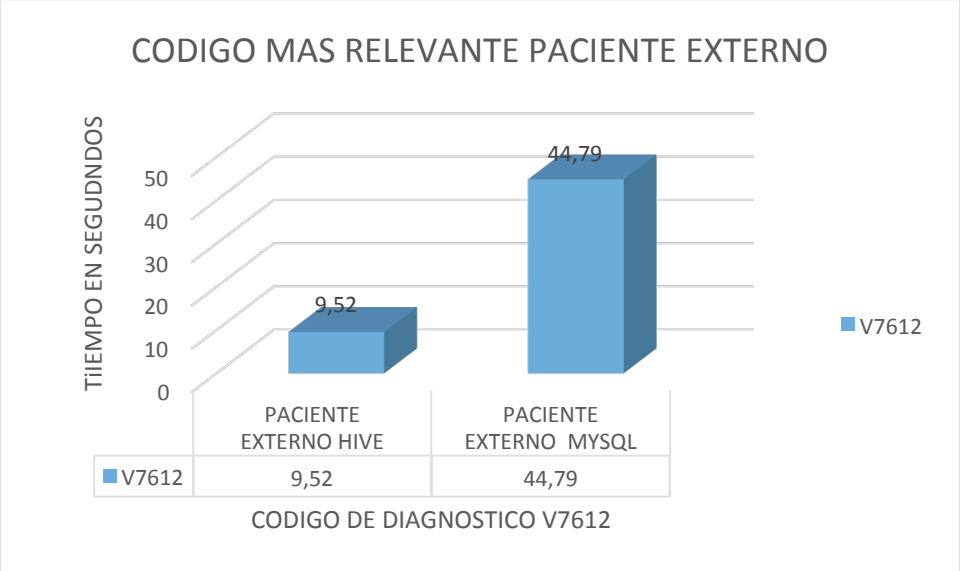

Figura 20. CODIGO DE DIGNOSTICO MÁS RELEVANTE P. EXTERNO

5. CONCLUSION Y TRABAJO FUTURO.

El cambio desde un ambiente DBMS vs BigDataSQL, es un desafío para los desarrolladores de sistemas. Esto se debe a que la historia y la evolución de la gestión de los sistemas de almacenamiento de datos, han estado estrechamente vinculados con las bases de datos relacionales y el lenguaje de consulta Sql, que es el lenguaje estándar para las bases relacionales.

En los últimos años se ha comenzado a desarrollar un sistema de software que demanda altos volúmenes de datos de distinto género, los cuales, afectaban la escalabilidad de todo el sistema porque difícilmente se podía modelar y manejar altos volúmenes de datos con el uso de las bases de datos relacionales. Tradicionalmente, la manera de resolver este problema fue aumentar las capacidades del hardware (llamada también escalabilidad vertical), sin embargo, esta solución llego a un punto donde el costo económico se volvió demasiado alto y la gestión del sistema demasiado compleja. Por lo tanto, el movimiento nosql ha presentado una solución a los desafíos recientes que enfrentan las bases de datos relacionales, porque proveen esquemas dinámicos, modelado de datos flexible, arquitectura escalable y almacenamiento eficiente de grandes datos, características que aumentan el rendimiento y escalabilidad.

De acuerdo con ese panorama, se realizó la comparación de rendimiento de los dos gestores de código abierto MySQL y Hive

Al realizar las pruebas de inserción, se ha demostrado que, en términos de tiempos de ejecución, Hive supera a MySQL, esto se debe a que Hive no impone un esquema a los documentos que son almacenados en la colección. Debido a que, cada documento puede tener su propio conjunto definido de campos, sin tener la necesidad de alterar la estructura o crear otra colección.

En la etapa de recuperación de registros, los índices son muy importantes, porque con ellos se pueden obtener los registros más rápidamente. En este caso todas las columnas del dataset están identificadas, estas características dan ciertas ventajas al momento de realizar una consulta en ambos ambientes.
En las pruebas realizadas, Hive fue superior al lograr ejecutar las tareas en menor tiempo, lo cual es muy importante cuando una aplicación debe soportar un uso intensivo de manipulación de datos, siempre y cuando no sean operaciones complejas. Los últimos estudios de investigación han concluido que hive es la plataforma que maneja de una manera más óptima los datos de gran volumen, analizando sus características, sus ventajas y desventajas y su uso en la inteligencia negocios.

6. BIBLIOGRAFIA

Analytics, H. (2014). Telemedicine Study. EU.

Antioquia, I. U. (2015). Ipsuniversitaria. Obtenido de ipsuniversitaria: http://www.ipsuniversitaria.com.co/es/quienes-somos

Antioquia, I. U. (s.f.). IPS UNIVERSIDAD DE ANTIOQUIA. Obtenido de http://www.ipsuniversitaria.com.co/es/quienes-somos/indicadores/indicadores-asistenciales#indicadores-asistenciales-2014

Apress. (2012). Enterprise Big Data Warehouse, BI Implemetations and analytics. White Paper, 84.

Bartłomiej Twardowski, Dominik Ry˙zko. (2016). Multi-agent architecture for real-time Big Data processing. Multi-agent architecture.

Big data. (2014). 10 Big Data Analytics Use Cases for Healthcare IT. Obtenido de bigdataanalyticsnews.com: http://bigdataanalyticsnews.com/10-big-data-analytics-use-cases-healthcare/

Chen, C. P. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347.

Commons. (2016). Commons. Obtenido de Commons: https://commons.apache.org/

Cortés, A. T. (2014). Big Data Technology to Exploit Climate Information/Consumption Models and to Predict Future Behaviours. International Technology Robotics Applications. Springer International Publishing, 25-36.

D. Peter Augustine. (2014). Leveraging Big Data Analytics and Hadoop in. EU.

Da Silva Morais, T. (2015). Survey on Frameworks for Distributed Computing. 95-105.

Dean, J. &. (2004). MapReduce: Simplified data processing on large clusters. Sixth Symposium on Operating System Design and Implementation.

Dimitrios Markonis, R. S. (2015). Using MapReduce for Large–scale Medical Image. arxiv, 10.

Emad A Mohammed, B. H. (2014). Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioMed Central, 8.

Escobar Ayona Elías Marino. (2014). Tecnología Big Data para el Sector Salud Del Estado de Guerrero. Propuesta, 1-10.

Fan, W. &. (2013). mining big data: current status, and forecast to the future. . ACM SIGKDD Explorations Newsletter, 1–5.

Frank Irving. (2014). 7 Big Data Use Cases for Healthcare. Obtenido de www.ingrammicroadvisor.com: http://www.ingrammicroadvisor.com/data-center/7-big-data-use-cases-for-healthcare

Ginger.io. (2015). www.ginger.io. Obtenido de www.ginger.io

Gopalani, S., & Arora, R. (2015). Comparing Apache Spark and Map Reduce. International Journal of Computer, 1-3.

hadoop.apache. (2016). hadoop.apache. Obtenido de hadoop.apache: https://hadoop.apache.org/.

Hcglobalgroup. (2010). Big Data y la Innovación Global en Servicios Actual y Futura. España.

Hurwitz, J., Nugent, A., Halper, F. y Kaufman, M. (2013). Big Data For Dummies.

Instituto Colombiano de Normas Técnicas y Certificación. (2015). Informe de auditoría de sistema de gestión. Medellín.

Instituto Colombiano de Vigilancia de Medicamentos y Alimentos INVIMA. (2013). Formato de certificado de buenas prácticas para banco de tejidos y médula ósea. Medellin.

Intel. (2014). Solving Practical Problems with Healthcare Analytics. White paper inel, 6.

Jason S. Mathias, M. D. (2012). Analyzing electronic health record data can help identify the overuse of cervical cancer screening. EU: U.S. Department of Health & Human Services.

Kumar, R., Gupta, N., Charu, S., & Jangir, S. K.. (2014). Manage Big Data through NewSQL. National Conference on Innovation in Wireless Communication and Networking Technology.

Laney, D. (2013).). Information Innovation: Innovation Key Initiative. Stanford. McKinsey.

Liu, X. I. (2014). Survey of real-time processing systems (X, Iftekhar, & & Xie, 2014) for big data. . Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS, 356–361.

MapReduce. (2016). MapReduce. Obtenido de MapReduce: https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce

Mariam Kiran Peter Murphy, Inder Monga, Jon Dugan Sartaj Singh Baveja. (2015). Lambda Architecture for Cost-effective Batch and Speed Big Data. EU.

Martínez-Prieto, M. A. (2014). The SOLID Architecture for Real-Time Management of Big Semantic Data. Future. Generation Computer Systems, 5.

Marz, Nathan; Warren, J.. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning Publications Co.

McKinsey. (25 de 02 de 2014). Big Data: The Next Frontier for Innovation, Competition, and Productivity. Obtenido de http://www.mckinsey.com: http://www.mckinsey.com/insights/business_technology/big_data_the_n

Mike Barlow. (2013). Big Data – Real Time Analytics: Emerging Architecture. 5-15.

Mobihealthnews. (s.f.). http://www.mobihealthnews.com. Obtenido de http://www.mobihealthnews.com

Moseley, B., & Marks, P.. (2006). Out of the tar pit. Software Practice Advancement (SPA). shaffner.us, 1-66.

Onehealth. (2015). http://www.onehealth.solutions/.

Paul zikopoulos, Chris Eaton. (2011). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media.
Paul zikopoulos, Chris Eaton. (2011). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. En Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Osborne Media. EU: McGraw-Hill.

Rizwan Patan and Rajasekhara babu. (2016). A NOVEL BIOMEDICAL DATA SOLUTIONS BY USING BIG DATA PLATFORMS FOR BETTER HEALTH CARE SERVICE. School of Computing Science and Engineering, VIT University Vellore, Tamil Nadu, India.

Robespierre Pita, C. P. (2014). A Spark-based workflow for probabilistic record. Distributed Systems Lab, 10.

Salud., F. s. (0ctubre de 2012). www.slideshare.net/RockHealth/rock-report-big-data. Obtenido de Fundación sin ánimo de lucro dedicada a la intersección de Tecnología y Salud.

Server, a. w. (2015). Lambda Architecture for Batch and Real - Time Processing on AWS with Spark Streaming and Spark SQL. Amazon web server, 12.

Shih, C. H. (2014). Building a CDR analytics platform for real-time services. Network Operations and Management Symposium (APNOMS), 1–5.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. Mass Storage Systems and Technologies (MSST). IEEE, 1-10.

Sproxil. (2014). http://www.sproxil.com/. Obtenido de http://www.sproxil.com/

TechAmerica Foundation. (2013). Demystifying Big Data. A practical guide to transforming the business of Government. White Paper, 11.

Tian, Y. A. (2014). DiNoDB: Efficient Large-Scale Raw Data Analytics Categories and Subject Descriptors. InProceedings of the First International Workshop on Bringing the Value of Big Data to Users (Data4U 2014), 1.

Tmforum. (2012). Insights Research: Big data: Big volumes, big payback and big challenge. Insights Research: Big data: Big volumes, big payback and big challenge.

Tmforum. (2012). Insights Research: Big data: Big volumes, big payback and big challenge. TMFORUM.