ISSN No:-2456-2165

Creating a Health Data Management Platform using Hadoop

Atul Bengeri[#1], Dr. Amol C Goje[^1]
[#1]Chairman, Board of Studies, Healthcare Management, Savitribhai Phule Pune University, Pune, India
[^1]Chairman, Board of Studies, Computer Management, Savitribhai Phule Pune University, Pune, India

Abstract:- Customary, conventional healthcare Database Management Systems are used as a repository of data and to process structured data efficiently, but in case of diverse variety and huge volumes of data it becomes arduous to handle such mammoth volumes. The question arises of what and how to process such data from various sources which could be structured as well as unstructured and in a distributed manner? Hadoop is open source framework, based on distributed computing, which is capable of storing and processing Big Data, which may comprise of structured, unstructured as well as semi-structured data. In this paper, we summarize the basic operations performed on healthcare data in a Data Management Lifecycle.

Keywords:- *Big Data, Data Analysis, Distributed Computing, ETL Hadoop, Healthcare, MapReduce.*

I. INTRODUCTION

Huge amounts of data get generated by the healthcare providers from record keeping of patient related data, health and medical device data, data regarding drug research, health insurance data, images with graphic, audio and video data and of late, patient generated data as well. This data thus generated could be structured or unstructured. Big Data includes handling both structured (RDBMS) and unstructured (multimedia, flat files) data. Big data is assuring a tremendous revolution in healthcare with important advancements from management activities of chronic diseases to prediction of disease in prior stages through the observation of the symptoms. Many healthcare organizations are bringing big data into practice which is of prime focus for researchers. The main focus remains at leveraging healthcare data and obtaining insights from it and make right decisions in appropriate time. Hadoop, due to its distributive nature helps in making information available to all its stakeholders instantly.

The management and analytics activities are performed using systemized collection of patient and population, electronically stored health information in digital format known as an Electronic Medical Record (EMR) or Electronic Health Record (EHR). These records can be shared across various healthcare enterprises via a secured network. These records generally contain broad range and diverse set of facts and figures such as medical history of patient, demographical information, medication and any specific allergy of patient, immunization status, laboratory results, personal information like age and weight, vital signs, and billing information.

II. METHODOLOGY

There are some open source software like OpenEMR and DHIS2.org that aid healthcare data management and analysis by supporting various data visualization features such as Tables, Charts, Pivot Table. The basic idea to these software was to get a brief of the schema used in real world healthcare organizations. There are also various random data generating sites like mockaroo.com. The interesting part is to obtain data from diverse, disparate and different sources to load it in the Hadoop Distributed File System (HDFS) through an ETL (Extract Transform Load) process and then populate the data using sql scripts, sql command or a sql procedure.

In our paper, we present a brief idea about how management and analysis of Healthcare data can be done using Hadoop framework.

Section III provides a brief on the Literature Survey and the related work done by various researchers across the world
Section IV gives a brief overview of the ETL process.
Section V throws light on the System Architecture and the Hadoop Ecosystem components used in our system.
Section VI provides the details of implementation.
Section VII is about the technologies.
Section VIII provide inputs on the future scope and conclusion.

III. LITERATURE SURVEY

In the paper "*Designing A Health Data Management System Based Hadoop-Agent*" by Fadoua Khennou, et al., they have presented an e-healthcare framework for health data management that links several numerous Electronic Medical Records (EMRs) implemented for all the health organizations and an Electronic Health Record (EHR) date warehouse kind of as a centralized system. They also attempt to provide a solution to the technical glitches of storing, loading and managing the health data using the Hadoop ecosystem framework. Furthermore, they have offered the theory of intermediary agents with a purpose that could play a vital role in smooth sharing of the medical data across distinct establishments.[1]

Yang Jin, et al., in their paper "*A Distributed Storage Model for EHR Based on HBase*" have suggested a distributed stowage for electronic healthcare record (EHR) which is based on HBase. This model consists of an electronic healthcare record storage that is used to organize the data and two Namenodes such that they respond to the list of relevant DataNodes.[2]

ISSN No:-2456-2165

In the paper, "*Research and Implementation of Massive Healthcare Data Management and Analysis Based on Hadoop*" by Hongyong Yu and Deshuai Wang, they deliberate around the Big data management, its handling and analysis solution based on Hadoop to achieve better scalability without compromising on the performance outcomes and taking into account the fault tolerance. They also elucidate on the 2 different data analysis methods constructed upon MapReduce and Hive.[3].

Mimoh Ojha, Dr. Kirti Mathur in the paper, "*Proposed Application of Big Data Analytics in Healthcare at Maharaja Yeshwantrao Hospital*" mostly address the challenges faced by doctors and patients while they attempt to provide solution to these problems as well.[4]

In the paper of Thara D.K., Dr. Premasudha B.G., Ravi Ram V, Suma R "*Impact of Big Data in Healthcare: A Survey*", they examine various investigative attempts made in the healthcare domain using concepts and strategies of Big Data. Among these things, this paper likewise provides an acumen for the budding scholars to distinguish and recognize the impact of Big Data on healthcare and seek evidence around the meagre research efforts made in the field of healthcare using Big Data thus far.[5]

IV. OVRVIEW OF THE ETL PROCESS

ETL is a model in the data warehousing technology that deals with conjoining the data from various sources into data warehouse, data marts or relational database such that we can analyze the data for meaningful patterns and useful insights. Heterogeneous data from diverse and disparate sources forms the input for the ETL to transform the data into standardized, harmonized, homogeneous data. ETL process helps in analysing heterogeneous data through an automated, programmatic and structured manner to derive business analysis and intelligence from it.

The robust data warehousing process of ETL consists of the usual three stages which can be interchanged to Extract, Load, Transform (ELT) process as well:

1. Extract- This phase involves the mining or the extraction of data from disparate source systems. Common data-source formats include the flat files, relational databases, JSON and XML and they could possibly also include the unstructured non-relational database structures too. The data thus extracted from source systems can be used in multiple data warehouse, data lake and data lakehouse systems.

2. Transform- In this phase of data transformation, wherein the change or makeover of the data can be either constructive, destructive, aesthetic or structural in nature. This is accomplished by a set of functions that get applied on the extracted data (from the previous step) for the sake of preparing the data which becomes suitable for analyzing by loading onto the object target system. Some data may not not require any data manipulation or data transformation which is recognized as direct move or pass through data. The objective of this transformation step is to ensure that all the data conforms to the tidy data principles and constitutes a dataset that can be referred to as a uniform schema.

3. Load- This phase of the ETL process involves the stacking and loading of the altered, transformed data into the end system for analyzing by applying machine learning techniques. In numerous organizations, the ETL process is an iterative practice which is performed regularly for the sake of keeping the data warehouse updated with the tardiest data to ensure the authenticity and the veracity of the data in all aspects.

V. SYSTEM ARCHITECTURE

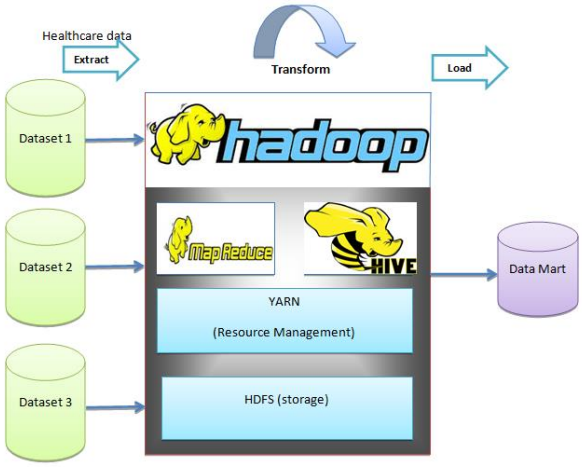

Fig 1

Handling diverse, disperate and voluminous data through conventional ETL or ELT datawarehouse process leads to were flaws related to handling Big data. A new distributed data storage and processing system known as Apache Hadoop has evolved over the years to not only provide a platform for storing and retrieving, but also for data evaluation and analysis purposes. The data from various disperate data sources are obtained and catalogued into different data repositories such as Dataset1, Dataset2 and Dataset3 respectively. The catalogued, combined data is then imported into the Hadoop framework using Flume or Sqoop. This approach of handling Big data ETL circumscribes around Hadoop which is cost effective and provides very efficient scalability and supple flexibility. The data in the Hadoop environment is processed much more effectively than RDBMS and that performs the transformations efficiently.

The Hadoop framework consists of the following components:

Map Reduce is the default processing framework used for writing applications to process in parallel and also used for batch data processing.

YARN is Yet Another Resource Negotiator that is used for resource allocation and management.

HDFS is the distributed file system that is redundant in

nature and used by Hadoop to store the data

Hive is a data warehouse software project that provides data summarization, query and analysis.

The transformed data from Hadoop is then loaded into the data mart. The data in the data mart can be used for prediction purposes.

Data so obtained from various sources must aim to possess the characteristics of big data, as elucidated below –
1] Volume: Enormous amounts of data is generated by the healthcare organizations through numerous diverse sources. Data generated by patient records, past medical history, diagnosis information and medical devices can be humongous in size which can transcend into terabytes or petabytes or even exabytes or zettabytes.
2] Velocity: Different constituent divisions in the healthcare enterprises generate data simultaneously, which takes into account the speed of data generation and frequency of delivery. The flow of data is massive and can be continuous and is valuable for health analysts for making predictions of a possible illness, a disorder or a health condition.
3] Variety: As the name suggests, data so collected and collated can be either structured data or unstructured data or a mixed bag which may include flat files, spreadsheets, text, emails, photos, videos or even EMR and / or EHR too.
4] Veracity: All the data so collated and collected has to be genuine and needs to be handled by the system.
5] Validity: In healthcare, data is mostly time dependent. Correct and accurate patient details should be made available when needed.
6] Volatility: The record of the timeline on the validity of the data and the extent of time to which it is required to be stored in the system.

VI. IMPLEMENTATION

For setting up of entire Hadoop multinode system, firstly installation of software packages is required from www.apache.org website. After the installation of packages, configuration of the xml files like coresite, mapred, hdfs-site, yarn is required along with setting of environment variables. The default openjdk path is to be modified and is to be changed to default Hadoop's path with correct specified version of jdk. After single node cluster formation on different systems, we need to link them in a way that forms master-slave architecture. There are two types of nodes in cluster formation i.e. Namenode and Datanode. Namenode acts as master node and there can be *n* number of datanodes. Namenode stores all the metadata of HDFS while all the actual data is stored in datanodes. Proper mapping of ip address and node in slave file is to be done along with respective entries made in host file.

Sites like Dhis2.org and OpenEMR are some examples from which sample data can be collected and which give a brief overview of the schema that is suggestive in a healthcare management software. The examples can be used as reference for building a prototypical schema design. Dhis2.org provides sample data in PostgreSQL database while OpenEMR provides sample data in SQL database. The main aim is to show how healthcare related big data can be handled effectively in Hadoop, for that the data collected from various sources needs to be populated. The volume of populated data needs to be at least in the range of terabytes. Various techniques can be implemented to populate data such as by applying scripts, SQL queries or procedure. The populated data is to be imported in HDFS using SQOOP, which is a tool for importing and exporting RDBMS data in HDFS.

VII. TECHNOLOGIES

The healthcare data can be analysed using Hive. In Hive the data is stored in the form of tables. Various queries can be processed on Hive using HiveQL, which is much similar to SQL. Hive queries rely on Map Reduce jobs internally thereby reducing workload on developers to actually perform map reduce program. Although Hive performs efficiently, it is seen that its performance degrades for medium sized (10 to 200GB) data and it lacks resume capability.

It is an established fact that Big Data will help to revolutionize the way in which the healthcare organizations operate their clinical data in a more sophisticated way, obtain meaningful insights from it and make good decisions as compared to now. In near future we will see a lot of involvement of big data and Hadoop in healthcare organizations. This paper throws light on how Hadoop framework can be used in a cost effective manner for data management and thus creating a platform ready for analysis on healthcare data, utilizing the distributive computing nature and its capability to handle heterogeneous data.

VIII. CONCLUSIONS

It is an established fact that Big Data will help to revolutionize the way in which the healthcare organizations operate their clinical data in a more sophisticated way, obtain meaningful insights from it and make good decisions as compared to now. In near future we will see a lot of involvement of big data and Hadoop in healthcare organizations. This paper throws light on how Hadoop framework can be used in a cost effective manner for data management and thus creating a platform ready for analysis on healthcare data, utilizing the distributive computing nature and its capability to handle heterogeneous data.

REFERENCES

[1]. Fadoua Khennou, Youness Idrissi Khamlichi, and Nour El Houda Chaoui, "Designing A Health Data Management System Based Hadoop-Agent", IEEE, 2016, pp. 71-76.
[2]. Yang Jin, Tang Deyu, Zhou Yi, "A Distributed Storage Model for EHR Based on HBase", IEEE, 2011, pp. 369-372
[3]. Hongyong Yu, Deshuai Wang, "Research and Implementation of Massive Healthcare Data Management and Analysis Based on Hadoop", IEEE, 2012, pp. 514-517
[4]. Mimoh Ojha, Dr. Kirti Mathur, "Proposed Application of Big Data Analytics in Healthcare at Maharaja

Yeshwantrao Hospital", IEEE, 2016
[5]. Thara D.K., Dr. Premasudha B.G., Ravi Ram V, Suma R, "Impact of Big Data in Healthcare: A Survey", IEEE, 2016, pp. 729-735
[6]. Mukesh Borana, Manish Giri, Sarang Kamble, Kiran Deshpande, Shubhangi Edake, "Healthcare Data Analysis using Hadoop", IRJET, Vol-2 Issue-7, Oct-2015, pp. 583-586