PUNE INSTITUTE OF COMPUTER TECHNOLOGY

Department of Computer Engineering

(2021- 2022)

DSBDAL

Batch: - N3

Health care systems with Hadoop Ecosystem components

Group members:

31367 – Gausiya Sayyad

31389 – Yash Rajput

Guided by: - Prof. M. S. Wakode

Table of Contents: -

1. Problem Statement

Write a case study to process data driven for Health care systems with Hadoop Ecosystem components as shown.

- HDFS: Hadoop Distributed File System
- YARN: Yet Another Resource Negotiator
- MapReduce: Programming based Data Processing
- Spark: In-Memory data processing
- PIG, HIVE: Query based processing of data services
- HBase: NoSQL Database (Provides real-time reads and writes)
- Mahout, Spark MLLib: (Provides analytical tools)
  Machine Learning algorithm libraries
  Solar, Lucene: Searching and Indexing

2. Motivation

- The analysis and prediction of future health conditions are still in developing stage. The data which is exerted in a little amount has risen greatly from a few bytes to terabytes, not only has the storage increased but also the dataset maintenance

- The traditional method of using data mining and diagnosis tools is difficult, therefore the need for big data tools and techniques arises.

3. Scope

Everyone likes that their work must be completed in less amount of time and the calculator on the salesforce cloud will help users/customers to do that.

Hence, many users will use such applications which help them to achieve their goals in less time.

4. Objective

By performing this case study, we shall be able to:

- HDFS: Hadoop Distributed File System
- YARN: Yet Another Resource Negotiator
- MapReduce: Programming based Data Processing
- Spark: In-Memory data processing
- PIG, HIVE: Query based processing of data services
- HBase: NoSQL Database (Provides real-time reads and writes) ● Mahout, Spark MLLib: (Provides analytical tools) Machine Learning algorithm libraries
- Solar, Lucene: Searching and Indexing

above components of Hadoop ecosystem in Health care system.

5. Outcomes

By performing this case study, student will be able to understand following components of Hadoop Ecosystem:

- HDFS
- YARN

- MapReduce
- Spark
- PIG, HIVE
- HBase
- Mahout, Spark MLLib
- Solar, Lucene: Searching and Indexing

6. Software and Hardware Requirements:

Software:

- Windows 10 OS, 64 bits
- Hadoop

Hardware:

- Processor: Intel i-5 8th gen
- Manufacturer: Acer Nitro 7
- Ram: 8 GB/ 16GB Optane memory

7. Theory

Introduction:

Every day, data is generated by a range of different applications, devices, and geographical research activities for the purposes of weather forecasting, weather prediction, disaster evaluation, crime detection, and the heath industry, to name a few. In current scenarios, big data is associated with core technologies and various enterprises including Google, Facebook, and IBM, which extract valuable information from the huge volumes of data collected. Big data is being generated rapidly in every field including healthcare, with respect to patient care, compliance, and various regulatory requirements. As the global population continues to increase along with the human lifespan, treatment delivery models are evolving quickly, and some of the decisions underlying these fast changes must be based on data. Healthcare shareholders are promised new knowledge from big data, so called both for its volume as well as its complexity and range. Pharmaceutical-industry experts and shareholders have begun to routinely analyze big data to obtain insight, but these activities are still in the early stages and must be coordinated to address healthcare delivery problems

and improve healthcare quality. Health informatics involves data acquisition, storage, and retrieval to provide better results by healthcare providers. In the healthcare system, data is characterized by its heterogeneity andvariety as a result of the linking of a diverse range of biomedical data sources including, for example, sensor data, imagery, gene arrays, laboratory tests, free text, and demographics

Four Vs of Big Data in Healthcare

Four primary attributes that are associated with big data: volume, velocity, variety, and veracity.

1. Volume: Big data is a term to referring to huge volumes of collected data. There is no fixed threshold for the volume of this data. Typically, the term is used with respect to massive-scale data which must be managed, stored, and analyzed using traditional databases and data processing architecture. The volume of data generated by modern IT and the healthcare system has been growing and is driven by the reduced costs of data storage and processing architectures and the need to extract valuable insights from data to improve business processes, efficiencies, and services to consumers.

2. Velocity: Velocity, which represents primary reason for the exponential growth of data, refers to how fast data is collected. Healthcare systems are generating data at increasingly higher speeds. In the volume and variety of the structured or unstructured data collected, the velocity of the generation of this data after processing requires a decision based on its output.

3. Variety: Variety refers to the form of the data, i.e., unstructured or structured, text, medical imagery, audio, video, and sensor data. Structured data information includes clinical data (patient record data), whichmust simply be collected, stored, and processed by a particular device. Structured data comprises just 5% to 10% of healthcare data. Unstructured or semi-structured data includes e-mails, photos, videos, audios, and other health related data such as hospital medical reports, physician's notes, paper prescriptions, and radiograph films.

4. Veracity: The veracity of data is the degree of assurance that the meaning of data is consistent. Different data sources vary in their levels of data credibility and reliability. The outcomes of bigdata analytics must be credible and error-free, but in healthcare, unsupervised machine learning algorithms make decisions that are used by automated machines based on data that may be worthless or misleading. Healthcare analytics are tasked with extracting useful insights from this data to treat patients and make the best possible decisions.

Impact of Big Data on the Health Care System:

The potential of big data is that it could revolutionize outcomes regarding the most suitable or accurate patient diagnosis and the accuracy information used in the health informatics system. As such, the investigation of huge amounts of information will have a powerful effect on medicinal services framework in five respects, or "pathways"

Improving outcomes for patients with respect to these pathways, as described below, will be the focus of thehealthcare system and will directly impact the patient.

1. Right Living:

Right living refers to the patient living a better and healthier life. By right living, patients could manage themselves by making the best decisions for themselves, based on the utilization of information mining better choices and enhancing their wellbeing. By choosing the right path for their daily health, regarding their diet, preventive care, exercise, and other activities of daily living, patients can play an active role in realizing a healthy life.

2. Right Care:

This pathway ensures that patients receive the most appropriate treatment available and that all providers obtain the same data and has the same objectives to avoid redundancy of planning and effort. This aspect has become more viable in the era of big data.

3. Right Provider:

Healthcare providers in this pathway can obtain an overall view of their patients by combining data from various sources such as medical equipment, public health statistics, and socioeconomic data. The accessibility of this information enables human service providers to conduct targeted investigations and develop the skills and abilities to identify and provide better treatment options to patients.

4. Right Innovation:

This pathway recognizes that new disease conditions, new treatments, and new medical will continue to evolve. Likewise, advancements in the provision of patient services, forexample, upgrading medications and the efficiency of research and development efforts, will enable new ways to promote wellbeing and patient health via

national social insurance system. The availability of early trial data is important for stakeholders. This data can be used to explore high-potential targets and identify techniques for improving traditional clinical treatment methods.

5. Right Value:

To improve the quality and value of health-related services, providers must pay careful and ongoing attention to their patients. Patients must obtain the most beneficial results identified by their social insurance system. Measures that could be taken to ensure the intelligent use of data includes, for example, identifying and destroying data misrepresentation, manipulations, and waste, and improving resources.

Big Data Analytics Architecture for Healthcare Informatics:

Currently, the main focus in big-data analytics isto gain an in-depth insight and understanding of big data rather than to collect it. Data analytics involves the development and application of algorithms for analyzing various complex data sets to extract meaningful knowledge, patterns, and information. In recent years, researchers have begun to consider the appropriate architectural framework for healthcare systems that utilize big-data analytics, one of which uses a four-layer architecture that comprises a transformation layer, data-source layer, big data platform layer, and analytical layer. In this layered system, data originates from different sources and has various formats and storage systems. Each layer has a specific data-processing functionality for performing specific tasks on the HDFS, using the MapReduce processing model. The other layers perform other tasks, i.e., report generation, query passing, data mining processing, and online analytical processing.

The main requirement in big-data analytical processing is to bundle the data at high speed to minimize the bundling time. The next priority in big-data analytical processing is to efficiently update and transform queries at a constant time. The third requirement in the big-data analytical processing is to utilize and efficiently manage the storage area space. The last specification of big-data analytics is to efficiently become familiar with the rapidly progressing workload notations. Big-data analytics frameworks differ from traditional healthcare processing systems with respect to how they process big data.

In the current health care system, data is processed using traditional tools installed in a single stand-alone system like a desktop computer. In contrast, big data is processed by clustering and scans multiple nodes of clusters in the network. This processing is based on the concept of parallelism to handle large medical data sets. Freely available frameworks, such as Hadoop, MapReduce, Pig, Sqoop, Hive, and HBase Avro, all have ability to process the health-related data sets for healthcare systems.

Big-data technologies broadly refer to scientific innovations that mimic those used for large datasets. In the first component is the requirement for big data sources for processing. In the second component clusters with a centralized big-data processing infrastructure are at the peak of high performance. It has been observed that the tools mainly available for big-data analytics processing provide data security, scalability, and manageability with the help of the MapReduce paradigm.In the third component, big data analytics applications have a storage domain to integrate accessed databases that use different applications. In the fourth component, are the most popular big-data analytics applications in healthcare systems, which include reports, Online Analytical Processing (OLAP), queries, and data mining.

1. HDFS (Hadoop Distributed File System):

The **Hadoop Distributed File System (HDFS)** is the essential information stockpiling framework utilized by Hadoop applications. It comprises NameNode, The Master and DataNodes. The Slave design to execute a disseminated record framework called Hadoop Distributed File System to get to information crosswise over exceedingly adaptable Hadoop Clusters in an effective way. Hadoop Framework in total consists of 5 daemon processes namely:

1. NameNode: NameNode is utilized to store the Metadata (data about the area, size of files/blocks) for HDFS. The Metadata could be put away on RAM or Hard-Disk. There will dependably be just a single NameNode in a cluster. The only way that the Hadoop framework can fail is when the NameNode will crash.

2. Secondary NameNode: It is used as a backup for NameNode. It holds practically same data as that of NameNode. On the off chance that NameNode falls flat, this one comes into picture.

3. DataNode: The actual user files or data is stored on DataNode. The number of DataNode depends on your data size and can be increased with the need.

The DataNode communicates to NameNode in definite interval of times.

4. Job Tracker: NameNode and DataNodes store points of interest and genuine information on HDFS. This information is likewise required to process according to users' prerequisites. A Developer writes a code to process the information.

5. Task Tracker: The Jobs taken by Job Trackers are in genuine performed by Task trackers. Each DataNode will have one task tracker. Task trackers communicate with Job trackers to send statuses of the undertaken job status.

HDFS bolsters the quick exchange of information between Master and Slaves as it is combined with MapReduce, an automatic system for information handling and to access information at a higher rate. When HDFS takes in information, it separates the data into partitioned squares and appropriates them to various nodes making the system effective via parallel processing.

2. YARN (Yet Another Resource Negotiator):

Hadoop YARN (Yet Another Resource Negotiator) is a Hadoop ecosystem component that provides the resource management. Yarn is also one the most important component of Hadoop Ecosystem. YARN is called as the operating system of Hadoop as it is responsible for managing and monitoring workloads. It allows
multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform. Contribute to society and human well-being.

YARN has been projected as a data operating system for Hadoop2.
Main features of YARN are:

- **Flexibility** – Enables other purpose-built data processing models beyond MapReduce (batch), such as interactive and streaming. Due to this

feature of YARN, other applications can also be run along with Map Reduce programs in Hadoop2.

- **Efficiency** – As many applications run on the same cluster, Hence, efficiency of Hadoop increases without much effect on quality of service.

- **Shared** – Provides a stable, reliable, secure foundation and shared operational services across multiple workloads. Additional programming models such as graph processing and iterative modeling are now possible for data processing.

3. MapReduce:

Map Reduction algorithm contains two important tasks, namely Map and

Reduce. • Mapping – Attained by Mapper Class

• Reduction – Attained by Reducer Class.

MapReduce utilizes different numerical calculations to separate an errand into little parts and dole out them to various frameworks. MapReduce calculation helps in sending the Map and Reduce errands to proper servers in a bunch. The tasks are executed in parallel in all the different nodes and finally the result is returned to the user.

The Healthcare Industry uses it primarily for the following

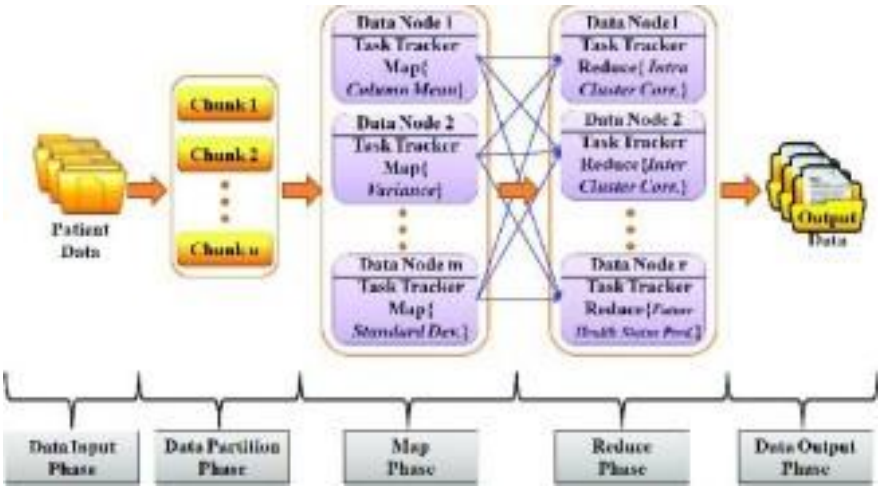
• Data Warehouse Optimization

• Patient Analysis

• Predictive Maintenance

Hadoop uses MapReduce algorithm to create tasks, called jobs which can be executed independently on different clusters (DataNodes) while the result is fetched

back to a single node (NameNode) for output.



As can be seen above, the system will group together the items having the same key. Finally, the system provides the requested output.

Patient's data is stored in a centralized repository which makes the system cost effective by reducing number of storage warehouses as well as eliminates any sort of data redundancy, which leads to the system being consistent as well.

4. HIVE:

The Hadoop ecosystem component, Apache Hive, is an open-source data warehouse system for querying and analysing large datasets stored in Hadoop files. Hive does three main functions: *data summarization, query, and analysis*. Hive use language called HiveQL (HQL), which is similar to SQL. HiveQL automatically translates SQL-like queries into MapReduce jobs which will execute on Hadoop.

Main parts of Hive are:

- **Metastore** – It stores the metadata.
- **Driver** – Manage the lifecycle of a HiveQL statement.
- **Query compiler** – Compiles HiveQL into Directed Acyclic Graph(DAG).
- **Hive server** – Provide a thrift interface and JDBC/ODBC server.

5. Pig:

Apache Pig is a high-level language platform for analyzing and querying huge dataset that are stored in HDFS. Pig as a component of Hadoop Ecosystem uses ***PigLatin*** language. It is very similar to SQL. It loads the data, applies the required filters and dumps the data in the required format. For Programs execution, pig requires Java runtime environment.

Features of Apache Pig:

- **Extensibility –** For carrying out special purpose processing, users can create their own function.
- **Optimization opportunities –** Pig allows the system to optimize automatic execution. This allows the user to pay attention to semantics instead of efficiency.
- **Handles all kinds of data –** Pig analyzes both structured as well as unstructured.

6. HBase:

Apache HBase is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of row and millions of columns. HBase is scalable, distributed, and NoSQL database that is built on top of HDFS. HBase, provide real-time access to read or write data in HDFS.

Components of Hbase

There are two HBase Components namely- HBase Master and RegionServer.

i. HBase Master

It is not part of the actual data storage but negotiates load balancing across all RegionServer.

ii. RegionServer

It is the worker node which handles read, writes, updates and

delete requests from clients. Region server process runs on every node in Hadoop cluster. Region server runs on HDFS DateNode.

7. Mahout:

Mahout is open source framework for creating scalable machine learning algorithm and data mining library. Once data is stored in Hadoop HDFS, mahout provides the data science tools to automatically find meaningful patterns in those big data sets.

Algorithms of Mahout are:

- **Clustering –** Here it takes the item in particular class and organizes them into naturally occurring groups, such that item belonging to the same group are similar to each other.
- **Collaborative filtering –** It mines user behavior and makes product recommendations (e.g. Amazon recommendations)
- **Classifications –** It learns from existing categorization and then assigns unclassified items to the best category.
- **Frequent pattern mining –** It analyzes items in a group (e.g. items in a shopping cart or terms in query session) and then identifies which items typically appear together.

7. Conclusion:

We have covered all the Hadoop Ecosystem Components in detail. Hence these Hadoop ecosystem components empower Hadoop functionality on the Healthcare system.