BIG DATA TECHNIQUES: HADOOP AND MAP REDUCE FOR WEATHER FORECASTING

Dr. Doreswamy¹ and Ibrahim Gad²

Abstract— Nowadays analyzing large amounts of data has become a big challenge. Data could be scientific, medical, meteorological, climatically, financial or marketing. Data mining techniques is used to extract meaningful information from large data set. Weather forecasting can be used to support many important sectors that are affected by climate like agriculture, water resources, air traffic, and tourism. Weather forecasting is an area of meteorology that is done by collecting data from the different stations related to the current state of the weather like temperate, rainfall, wind, and fog. Weather forecasting is the most challenging problem for scientists. Hadoop and MapReduce are the most models used to analysis a huge data set. This paper explains a system that uses the historical weather data of a region and apply the MapReduce and Hadoop techniques to analysis these historical data.

Keywords – Weather forecasting, Meteorology, Big data, Hadoop.

I. INTRODUCTION

Data mining techniques are the process of extracting meaningful information from the large data set. The process of extract meaningful information described as knowledge discovery that can be applied on any large data set. The main Data mining techniques are Classification, Clustering, Association rules, and Regression [13]. The different Data mining techniques used for solving weather forecasting problem.

Weather forecasting problem include prediction of temperature, rain, fog, winds, clouds, storm etc. [6]. Weather sensors collect data every hour at many locations and gather a huge data. Weather forecasting is always a big challenge because it is hard to predict the state of the atmosphere for the upcoming future because climate dataset is unpredictable and frequently change according to global climate changes. The data used is from the national climatic data center (NCDC), the format of dataset support a rich set of meteorological elements, which are good candidate for analysis with big data because it is semi-structured and record oriented [12].

The paper is organized as follows: The concepts of big data and its characteristics are given in section II. In section III, Hadoop and MapReduce are presented. In section IV, present the previous work of the other researchers. Finally, section V includes practical work of Hadoop and MapReduce.

II. BIG DATA

The term Big Data came around 2005, which refers to datasets that are big, also high in variety and velocity, which makes them difficult to process using traditional tools and techniques [1]. Big data created huge business and social opportunities in each field, enabling the discovery of

¹ Department of Computer Science, Mangalore University, Mangalagangotri- 574 199, Karnataka, INDIA
² Department of Computer Science, Faculty of Science, Tanta University, Tanta, Egypt
previously hidden patterns and the development of new insights to make decisions, ranging from web search to content recommendation and computational advertising. The term Big Data is now used almost everywhere in our daily life and it is a current technology and also going to rule the world in future and has emerged because people and different companies makes increasing use of data-intensive technologies[5,7].

Big data sizes are currently ranging from a Terabyte (TB or \(10^{12}\) or \(2^{40}\)) to Zettabyte ( ZB or \(10^{21}\) or \(2^{70}\)) in a single data set [8]. Like the physical universe, the digital universe is large. According to research conducted by IDC, from 2005 to 2020, the digital universe will grow from 130 Exabytes to 40,000 Exabyte’s, or 40 trillion gigabytes. From now, the digital universe will about double every two years until 2020 [9]. As stated by IBM, with machine-to-machine(M2M) communications, online/mobile social networks and pervasive handheld devices it creates 2.5 quintillion bytes of data in each day — so much that 90 percentage of the data in the world today has been produced in the last two years alone [10].

A. Characteristics of Big data—

Big Data has many characteristics or properties mentioned by n V’s characteristics. Set of V’s characteristics of the Big Data were collected from different researcher’s publications to have Nine V’s characteristics (9V’s characteristics)[2]. These 9V’s characteristics are: (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization and Value).

- **Veracity**: Big Data veracity refers to the biases, noise, and abnormality in data.
- **Variety**: Structured, semi-structured, and unstructured data besides text and more data types have emerged, such as record, log, audio, and hybrid data.
- **Velocity**: The created information at a faster pace than before, in which the different channels of Big Data increase the output content.
- **Volume**: The amount of data is known as volume of data, where the amount of data continues to explode.
- **Validity**: The data is correct and accurate for the intended use. Clearly, valid data is the key to making the right decisions.
- **Variability**: The data flows may be highly inconsistent with periodic peaks, daily, seasonal, and event-triggered peak data loads can be challenging to manage, especially with unstructured data involved.
- **Volatility**: Once retention period expires, we can easily destroy it.
- **Visualization**: means complex graphs that can include several variables of data while still remaining understandable and readable.
- **Value**: It has a low-value density as a result of extracting value from massive data. Useful data needs to be extracted from any data type and from a huge amount of data.

III.HADOOP

Hadoop and Map Reduce are the most widely used models used today for BigData processing. Hadoop is an open source large-scale data processing framework that supports distributed processing of large chunks of data using simple programming models. The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in addition to other modules.
Hadoop is an open-source framework for processing a large amount of data across clusters of computers with the use of high-level languages. Its modules provide easy to use languages, graphical interfaces, and administration tools for managing data on thousands of computers. Hadoop cluster is a set of machines networked together in one location. Data storage and processing all occur within this "cloud" of machines. Users can submit jobs to Hadoop from his desktop machine in remote locations from the Hadoop cluster.

Two main components of Hadoop are Hadoop Distributed File System (HDFS) and MapReduce. HDFS is a distributed file system management for large datasets of sizes of gigabytes and petabytes. And MapReduce is a programming framework for managing and processing a huge amount of unstructured data in parallel based on the division of a big dataset into smaller independent chunks.

A. HDFS Architecture

HDFS has a master/slave architecture. The main components of an HDFS cluster are a single NameNode, a master server that manages the file system and control access to files by clients. In addition, there are a number of DataNodes, each node usually contain one DataNode in the cluster, which manage storage associated with these nodes that they run on see Figure 2.

B. MapReduce

The Map Reduce software framework which was originally introduced by Google in 2004 is a programming model, which now adopted by Apache Hadoop, consists of splitting the large chunks of data and 'Map' and 'Reduce' phases. MapReduce is a processing large datasets in parallel using lots of computer running in a cluster. We can extend the mapper class with our own instruction for handling various input in a specific manner. During map master node instructs worker nodes to process local input data and Hadoop performs shuffle process. Thus master node collects the results from all reducers and compilers to answer overall query see Figure 1.

IV. RELATED WORK

Riyaz P.A. et al.[4], describes the analysis of huge amounts of climatic data by using MapReduce with Hadoop. Huge amounts of climatic data collected, stored, and processed for accurate prediction of weather. Climatic data collected by using different types of sensors to store the following parameters: temperature, humidity, etc. Weather datasets collected from National Climatic Data Center (NCDC). Daily Global Weather Measurements 1929-2009 (NCDC, GSOD)
dataset is one of the biggest dataset available for weather forecast. Its total size is around 20 GB. Results show that temperature analyzed effectively by Using MapReduce with Hadoop.

VeershettyDagade et al.[6], gives a detailed description of build a platform that is extremely flexible and scalable to be able to analyze Pentabytes of data across an extremely wide increasing wealth of weather variables. Data processed by Apache Hadoop and Apache Spark. Experiments performed to select the best tools among Hadoop using Pig and Hive Queries.

Ramya M. G. et al.[3], explains the meteorological data storage as well as analysis platform based on Hadoop framework with the help of online logistic regression algorithm for prediction. This platform is based on distributed filesystem HDFS which incorporates distributed database HBase, data warehouse management and efficient query processing tool Hive, data migration tool Sqoop. The best data mining prediction algorithm regression also integrated into the system. This architecture has an ability of mass storage of meteorological data, efficient query, and analysis, climate change prediction.

V. EXPERIMENTAL RESULTS

National Climatic Data Center (NCDC) have provided a huge historical weather datasets. Daily Global Weather Measurements 1929-2016 (NCDC, GSOD) dataset is one of the biggest historical weather dataset available for weather forecasting. Its total size is around 20 GB. It is available on National Climatic Data Center (NCDC) web services [12]. The United States National Climatic Data Center (NCDC), previously known as the National Weather Records Center (NWRC), in Asheville, North Carolina is the world’s largest active archive of weather data. The Center has more than 150 years of data on hand with 224 gigabytes of new information added each day. NCDC store 99 percent of all NOAA data, including over 320 million paper records; 2.5 million microfiche records; over 1.2 petabytes of digital data residing in a mass storage environment. NCDC has satellite weather images back to 1960.

The proposed System use dataset of NCDC contains the following parameters: station number, station name, date, country, Precipitation, Temperature, and Wind as shown in Figure 5. The data files are stored in HDFS. Then, weather files are split and goes to different mappers. The output of each mapper is a set of pairs (key, value) where key is consists of station name, date and value is contains the parameters: Precipitation, Temperature, and Wind. Then the output of mappers is merged and sort by key. Finally, all results sent to the reducers. For each reducer calculate Average (monthly, yearly, and seasonal), Maximum (monthly, yearly, and seasonal), and Minimum (monthly, yearly, and seasonal), for each parameter precipitation, Temperature, and wind in different stations. Each reducer store the final results in HDF See Figure 3. Due to a practical limit, the analysis is executed in Hadoop standalone mode. Figure 4 shows MapReduce Framework execution. Figures(6, 7, 8) show the different final result of our proposed method.
Figure 3. Proposed MapReduce Framework

Figure 4. The output of the MapReduce program

Figure 5. Weather data set

Figure 6. The final result for Months of each station.
VI. CONCLUSION

In case of using traditional systems, to process millions of records in the era of Internet of things, the meteorological department faces time-consuming processes. Using MapReduce with Hadoop, the weather data can be analyzed effectively. The major advantage of MapReduce with Hadoop frameworks speeds up the processing of data. Where the volume of data is increasing every day.

REFERENCES