IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique

Basvanth Reddy[1], Prof. B.A Patil[2]

Department of CSE, KLE Dr. MSS College of Engg., & Tech, Belgaum[1]

Professor, Computer Science & Engg, KLE DR M S Sheshgiri College of Engg & Tech., Belgaum[2]

Abstract: The assists of big data collects large volume of data, it is great computational challenge for the big data Hadoop which uses map reduce to maintain and process this data and also helps to extract useful information in an efficient manner . Keeping these things in mind there is need for designing system architecture that helps to predict weather forecast for future. It helps people to take decision in advance for their any outdoor events. In our proposed architecture there are 3 main units, such as: Data acquisition unit, Data processing unit, Data analysis and decision unit. DAU collects data from the satellite and sends to the different base station and finally stores in the national climatic data centre (NCDC).DPU it plays key role in the architecture, it takes the data from the NCDC and HDFS takes the data from this NCDC and store in it and later processed by MapReduce framework and produce required output .DADU it performs the compilation based on the results produced by the DPU and makes decision to produce required results. Hence in our proposed architecture we are focusing on the offline data that is stored in the NCDC to predict the weather analysis by Hadoop map reduce framework the output of result consists of, minimum temperature, maximum temperature, number of hot days and cold days and also predict future weather forecast, which brings the great significance of our work.

Keywords: Big data, HDFS, Hadoop, YARN, MAPREDUCE.

I. INTRODUCTION

Big Data is the process of examine large data sets containing variety of data types. The big data maintains the huge amount of data and process them. It is traditional data analysis; it is able to process the structured data, but not unstructured data. In big data it is able to process both structured and unstructured data. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage and process the data. Big data size ranges from terabytes to many petabytes of data.

Weather prediction is the application of technology to predict the action of the atmosphere for a given location. It is important mainly for business agriculturist, farmers, disasters management etc. weather prediction is one of the most interesting and fascinating domain and plays significant role in meteorology. There are several limitations in better implementation of weather forecasting for example in data mining techniques; it cannot predict weather short term efficiently. They used small limited areas for weather forecasting. It is difficult task to predict weather due to dynamic changes in the atmosphere.

Climate change has been seeking a lot of attention since long time. The antagonistic effect of this climate is being felt in every part of the earth. There are many examples for these, such as sea levels are rising, less rainfall, increase in humidity. The propose system overcomes the some issues that occurred by using other techniques. In this project we use the concept of Big data Hadoop. In the proposed architecture we are able to process offline data, which is stored in the National Climatic Data Centre (NCDC). Through this we are able to find out the maximum temperature and minimum temperature of year, and able to predict the future weather forecast. Finally, we plot the graph for the obtained MAX and MIN temperature for each moth of the particular year to visualize the temperature. Based on the previous year data weather data of coming year is predicted.

A. Overview of the Project

The prediction of the climate change always has proven very important and useful. In the United States of America there are regularly many events organized in different cities. These events might include the car racing, festivals, concerts, etc. As these are the outdoor concerts, they suffer a lot from the frequent weather changes, which is increasing due to global warming.

To avoid these issues, they need to pre-plan and choose the data for their event in advance.This can work out only if they have any predictions of the climate data using the Hadoop and distributed system and map reduce. By using map reduce, we can also calculate the maximum and the minimum temperature for the hot days and cool days. So as the result we can discover useful information about event planning, such as location, time and statistical data.

B. Aim and Objective of the Project

To provide

- The maximum and minimum temperature of city for year.
- To predict the climate changes obtained from the map reduce.

DOI 10.17148/IJARCCE.2016.56142

- To be able to provide schedule the events based on this climate data.
- To be prepared for the different natural calamities like humidity and cold.
- To provide visualization of the obtained data and compare the increase and decrease in global warming.

C. Problem Statement

The over exploitation of natural resources has resulted in the serious environmental troubles. Additional problems have also come forward due to the increase in the world's average temperature. Recent advances in the satellite technology and sensor data has improved in ground-based environmental observations.

II.LITERATURE SURVEY

In [1] It also includes the decision support systems for the more deep analytics and descriptive, which is the most dangerous part of the cloud environment. There are many problems faced by the application developer and DBMS Designers. (1) Provides both the types of systems (a) support bulletin large application (b) for ad hoc analytics and decision support. This paper provides the deep analysis of the systems that supports the update and as well as the great degree of work in web applications and also provides the report of the state-of-the-art in this particular domain. Here the design choices are made by the particular selected successful system of large scale DBMS, which analyses the application demands and also the access patterns, and enumerate the required data for a cloud-bound DBMS.

In [2] As it is known the huge data acquisition and storage becomes increase single more costlier, and hence many enterprises are hiring the statisticians so that they can handle the complexity of the data analysis. In (2) provides the main focus on the upcoming practice of Magnetic, Agile Deep (MAD) data analysis as a radical departure from the traditional enterprise data warehouses and Business Intelligence. The design philosophy, techniques and experience provided in the MAD analytics is for one of the world's biggest advertising networks. In this paper, the algorithm that runs parallel for the complex statistical techniques focuses mainly on the density of methods.
The outcome in this paper reflects on the database system features that allows the agile design and flexible algorithm improvement using both the SQL and Map reduce.

In [3] The author in (2) provides the mechanism of MapReduce interface that allows the user to reduce the complexity of database. The (3) provides the process of working model of the MapReduce. In this paper, the author provides the complete description of MapReduce programming model. In paper [3] author says MapReduce is a programming model which is also associated with implementation for processing and generating the large amount of data set. Here the user consider the map as the function that will process a key/ value pair so that it can generate a set of intermediate key/ value pairs, and the reduce function will merges all the outcome value from the map side with same value, and same intermediate key. This paper also describes the implementation of MAP reduce, which runs on a large cluster of different machines and which are highly scalable. The outcome of this paper provides a programming model which has been successfully used at Google for many different purposes. As this programming model helps to work parallel in distributed systems and hides the details of the parallization. In this the large datasets are reduced and they are easily expressible.

In [4] As and off the days are moving the analytics over the "big data" has become the key to the success in many business, scientific and engineering disciplines and also include government endeavours. The Map reduce engine, the pluggable distributed storage engines are part in the Hadoop software. The Hadoop software also consists of the array of procedure to annotative interface, which has become the most popular choice in big data.

In [5] From the recent few years, it has been proven the growth in the volume and availability of the data. The results of these occurs from the usage of different types of the sources i.e., devices, computers, sensors or social networks, which produce every day a huge amount of data. This data may be either structure, semi-structured or unstructured data sets.Data modelling provides the depth look of the data models that are used to define and support operational database, and Big Data Technologies, whereas the data analytics provides the different types of operations that can be performed over the data model.

In [6] The big data is usually generated by the online transaction emails, clicks, social network data, remote sensing data and their many other application. These data are dumped into the database that grow very huge in size and thus which increases the complexity to stores, manage, process, analyse and visualize the typical database software tools. The updating in the big data sensing and computer technology has changed fundamentally in the way the data is collected, processed and analysed and managed [9]. The recently designed sensors that are used in earth observatory system are generating huge amount of data.

In [7] Decision making is one of the important thing in big data. Huge data is being generated by different users to store and process in powerful data centres. Hence, it has become the necessity of generating a network in structure to gather the data that are generated rapidly and geologically to express these networks it is necessary to extend and interconnect multiple data centres and even interconnect the server nodes within the data centre. [26] Provides each and every segment in the network highway such as providing an access to all the networks that connect to data sources, the internet backbone, which allows them to route to the remote data centres. This helps to build a proper network infrastructure to access data from different networks.

DOI 10.17148/IJARCCE.2016.56142

In [8] The Smart Grid is used as it includes the different types of operational and every measure which also include smart meters, renewable energy sources and energy efficiency resources. Smart Grid also contains the next version of electric grid, which maintains the demand and supply of the customer in a balanced way. [27] deals with formation of the cloud based grid which can further help to analyse the Big-data and hence, helping into take decisions of the demand of customer needs. The data here deals with power usage patterns of customers, its weather data location and present demand and supply details. This grid will operate by the data being fetched from the cloud storage.

In [9] Airborne and Space borne platform the valuable data for mapping when obtained through remote sensing, also provide environmental monitoring, disaster management and civil and military intelligence. To survey the complete value of these data sets, correct and appropriate information is needed to be extracted and then present in a standard format.

[32] Provides the principal strategies of object oriented analysis and also allows the implementation of expert knowledge with fuzzy methods combination. The author also describes the strategies of object-oriented analysis, which can be used to present the outcome in a standard format.

III. PROBLEM DEFINITION AND OBJECTIVE:

A. Problem Definition:
The over exploitation of natural resources has resulted in the serious environmental troubles. Additional problems have also come forward due to the increase in the world's average temperature. Recent advances in the satellite technology and sensor data has improved in ground-based environmental observations.

B. Objective:
To provide

- The maximum and minimum temperature of city for year.
- To predict the climate changes obtained from the map reduce.
- To be able to provide schedule the events based on this climate data.
- To be prepared for the different natural calamities like humidity and cold.
- To provide visualization of the obtained data and compare the increase and decrease in global warming.

C. System Design:
As the name indicates i.e., Hadoop distributed file system, the large amount of data is distributed, stores and provides easier access. The files stored here are done in the redundant fashion so that they can reuse the system from the possible data loss and hence avoiding failure. As the data is distributed among many machines, the HDFS provides the parallel processing.

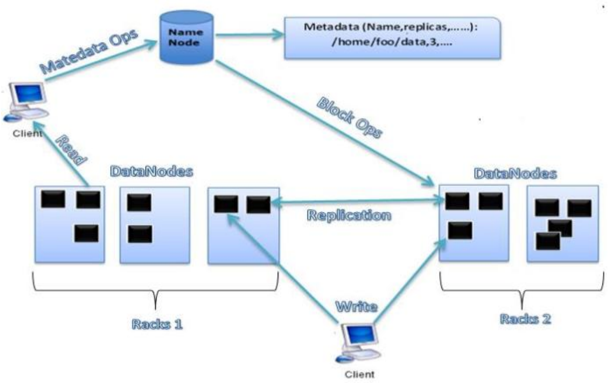

Figure 1: Architecture diagram of HDFS.

Name Node
The name node is the suitable hardware that contains the GNU/ LINUX OS and name node framework software. The name node in the file system acts as a master server and it does its tasks as follows:

- The execution operation of the file system such as renaming, closing and opening files and directories is done by name node.
- The name space of the file system is managed by name node.
- Regulates the client's access to files.

Data Node
As in the name node, the data node is a suitable hardware that resides the GNU/ LINUX operating system and data node software. The cluster contains many nodes and each node in that cluster will be a data node. The data storage of the system is maintained by these nodes. Tasks of data node are:

- Based upon the client request, it performs the react and writes operation.
- The name node will give certain instructions to data node such as creation, deletion and replication and it performs all these operations.

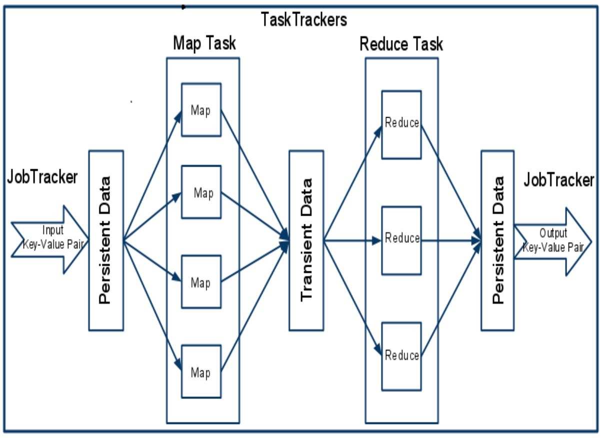

Figure 2: Architecture of MAP REDUCE

The basic definition of map reduce is defined in previous section. In Map Reduce, we are able to write application to process the huge amount of data in parallel.

DOI 10.17148/IJARCCE.2016.56142

The Map Reduce mainly contains two main tasks, namely “MAP” and “REDUCE”. The task of the map is to take a set of data and converting this data to another set of data, where the individual data elements are broken down into tuples i.e., key/ value pairs. The work of reduce task is to take the O/O of map task as input and reduce those data tuples into smaller tuples. As the name indicates the reduce task is always done after map task.

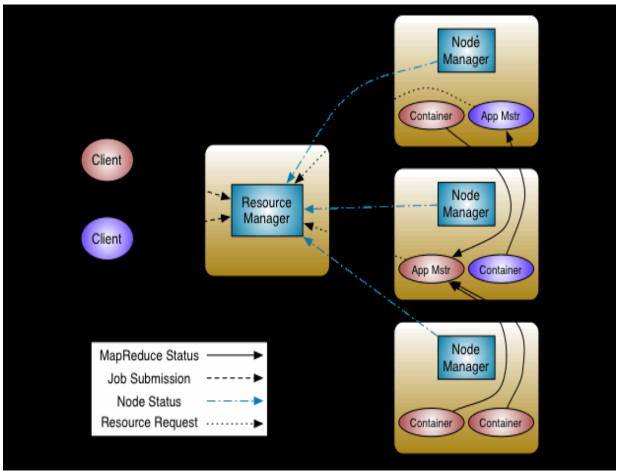

Figure 3: Architecture of YARN

Yarn consist of central resource manager which arbitrates all available cluster resources area and as per node manager which takes direction from the Resource Manager and responsible for managing resources available on a single node. The resource manager is the master that judges all the available cluster resources and thus helps manage the distributed application running on the YARN system.

The resource manager works together with node manager and application manager. Node manager will take the instructions from the resource manager and will manage resource available on a single node. Application masters are responsible for negotiating resources with resource manager and for working with node manager to start the containers. The Node Manager (NM) is also YARN’s pre-node agent, and takes care of the individual compute nodes in Hadoop cluster. The continuous features interaction such as air temperature, humidity would provide very valuable and efficient for most of the organization to work under any climatic condition.

IV. RESULTS:

The proposed system uses the temperature datasets of 2013, 2014, 2015 from NCDC. These records are stored in the HDFS and perform map reduce function. Map reduce execution is shown in fig below “the results shows adding more number of systems to the network will speed up the entire data processing”. This is one of the major advantage of the map reduce with hadoop frame work.

The following is the graphical outcome result of weather for the month November and September 2014

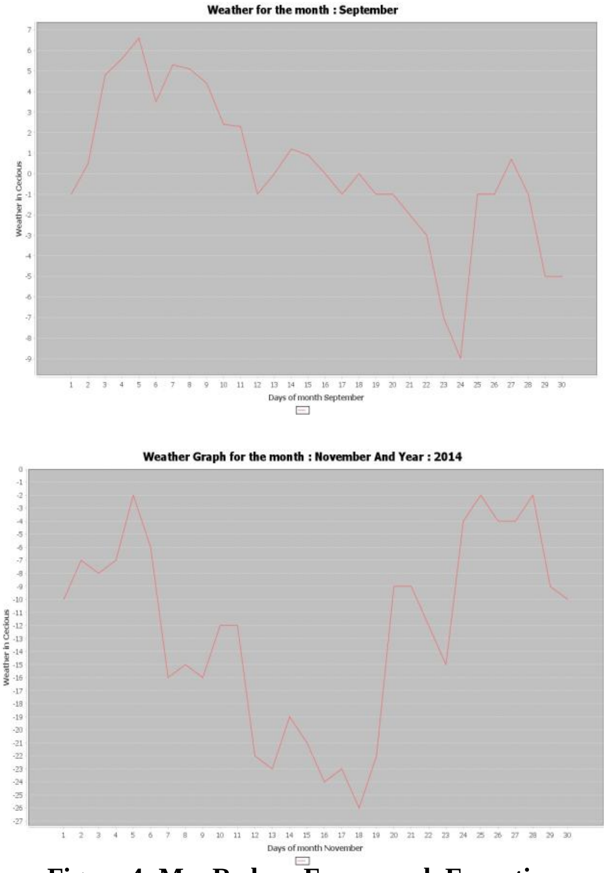

Figure 4: MapReduce Framework Execution

V. CONCLUSION

In this paper have proposed weather prediction using big data environment. The method used in our project is Hadoop with map reduces to analyse the sensor data, which is stored in the National Climatic Data Centre (NCDC) is a efficient solution.

Map reduce is frame work for highly parallel and distributed systems across huge dataset.
It is used to analyse for the given data and predict required output to our project. By using map reduce with hadoop helps in removing scalability bottleneck. This type of technology used to analyse large data sets has potential to great enhancement to weather forecast.

Hence we predict the future weather forecast, minimum and maximum temperature, hot days and cold days based on the data obtained from the NCDC.

This helps for the people to preplanning for outdoor events based on the weather conditions.

REFERENCES

[1]. P.Agarwal, S.Das and A.E.Abbadi, “Bigdata and cloud computing: Current state and future opportunities” in Proc Int Conf Extending Database Technol.

[2]. J.Cohen, B.Dolan, M.Dunlap, J.M.Hellenstein and C.Welton, “Mad Skills: New Analysis practices for Bigdata”.

[3]. J.Dean and S.Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters” Commun.

DOI 10.17148/IJARCCE.2016.56142

[4]. H.Herodotou et al, “Starfish: A Self-tuning System for Bigdata Analytics”.
[5]. K.Michael and K.W.Millen, “Bigdata: New Opportunities and New Challenges”.
[6] R.D.Schneider, “Handoop for Dummies”, Special Edition. L.Ramaswamy, V.Lawson and S.V.Gogieni, “Towards a quality Centeric Bigdata Architecture for Fractured Sensor services”.
[7]. X.Y, F. Liu, J.Liu and H.Jin, “Building a network highway for Big data: Architecture and challenges.
[8]. M.Mayilvaganan and M.Sabitha, “A cloud-based architecture for Big data analytics in smart grid: A Proposal”.
[9]. V.C.Benz et al, “Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS ready information”.
[10]. M.Olson, “Hadoop: Scaleable, Flexible Data Storage and Analysis”.
[11]. Andue Riberio, Afonso Silva, Alberto Rodrigues De Silva, “Data Modelling and Data Analytics: A Survey from a big data perspective”.
[12]. Nilamdhab Mishra, Chung-Chin, Lin Hsieu-Tsung Chang, “A cognitive adopted framework for IOT Bigdata Management and Knowledge Discovery Prospective”.

BIOGRAPHIES

Basvanth Reddy is a M. Tech student in the Department of Computer Science & Engineering, KLE Dr. M S Sheshgiri College of Engineering, Belgaum. He completed Bachelor of Engineering in Computer science & Engineering from P.D.A college of engineering gulbarga in the year 2014.

Prof. B.A Patil is currently head of the Department of Computer Science & Engg, KLE DR M S Sheshgiri College of Engg.& Tech, belgavi. He did his M.E degree in sangli wal chand college of engg in year 2003.

DOI 10.17148/IJARCCE.2016.56142