IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

A Review of Weather Data Analytics using Big Data

Priyanka Chouksey[1], Abhishek Singh Chauhan[2]

M.Tech Research Scholar, NIIST, Bhopal (M.P), India[1]

Assistant Professor, NIIST, Bhopal (M.P), India[2]

Abstract: Weather plays an important role in every aspect of human life. It has direct impact on every section of human society. Weather forecast has lot of importance in agriculture sector, tourism sector and government agencies. Prior knowledge of weather can be very helpful for human to prepare themselves for any undesirable condition of climate. Various weather parameters like temperature, pressure, humidity, wind speed etc. plays an important role in the analysis of whether condition. Weather data from various sensors for various weather parameters are being generated at huge scale and in future this data will continue to grow. Big Data Analytics technology like Hadoop MapReduce and Spark are playing great role in handling huge amount of data. MapReduce has proven to be useful for batch processing while Spark has proven to be highly efficient in In-Memory Computing. This project aims to study the analytics of weather data using MapReduce and Spark.

Keywords: Hadoop, MapReduce, HDFS, Spark, Weather Data Analytics.

I. INTRODUCTION

Weather data analytics has lot of importance in human life. Accurate prediction of weather is very helpful to agriculture sector, tourism, and also planning for any natural calamities like flood, drought etc. Weather Prediction has lot of commercial value in news agency, government sector and industrial farming. Weather has enormous effect on psyche of a human being. Human mood can change as positive, negative or tiredness based on some changes in weather [1].

Prediction of climatic condition is very important challenge for every living being to sustain. To study the climate there is need to study meteorology. Meteorology is the interdisciplinary scientific study of atmosphere i.e. temperature, pressure, humidity, wind, etc. Usually, temperature, pressure, wind and humidity are the variables that are measured by a thermometer, barometer, anemometer, and hygrometer respectively [2]. Observations of these parameters are collected from various sensors deployed at different geographical location. This data is accumulated at meteorological department of various countries. This data is known as weather data.

At each location the values of various weather parameters is collected at a frequency of 3-4 times per hour. This data is stored in the unstructured format along with location, date and time. The structure of these formats is flat file which is separated by comma or tab or may be semicolons. It is difficult to process this unstructured data directly. The collective data is becoming very huge considering various parameters, their frequency of recording and number of locations. Day by day this data is growing and accumulated at enormous speed. Hence, to process this data using conventional methods and tools is becoming a challenge. Hadoop platform with MapReduce programming paradigm has proven to be very useful in processing huge unstructured data. Spark with in memory computing also gives very good performance for analysis of unstructured data.

As various other Big Data Technologies like Storm, NoSQL are also claiming their usefulness in storage and processing of huge data it is important to study their relative performance and usefulness in various domains. In the current project, the study of Big Data technology MapReduce and Spark is being studied and compared for Weather Data Analytics.

II. LITERATURE REVIEW

Shraddha et al. [2] proposed a method for weather forecasting using Adaptive technique in data mining. Method for identification of the occurrence of rare patterns in weather is proposed. Various steps of data mining like data collection, data pre-processing, data cleaning, data transformation and smoothing are explained. For knowledge discovery various methods of mining like Classification, Prediction, Clustering and Outlier Analysis are discussed. K-means clustering algorithm was discussed in detail for weather data.

Veershetty et al. [3] worked on building a platform using Hadoop to analyse the weather data. Temperature and yearly precipitation were chosen as weather parameter for extraction and analysis. The performance comparison of weather data using Pig and Hive is shown. The performance of HIVE is proven to be better in results. The

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

proposed analytical engine has capability to scale better in Hadoop cluster.

J Denissen at el. [1] has studied the effect of weather on daily mood using multilevel approach. Six weather parameters (temperature, wind power, sunlight, precipitation, air pressure, and photoperiod) were examined to predict the daily mood (positive effect, negative effect, and tiredness). This study showed the importance of weather to a daily life of human being.

Basvanath Reddy and Prof B.A.Patil worked on prediction of maximum and minimum temperature of a particular city for a particular year [4]. Basic detail about Hadoop, HDFS, YARN and MapReduce is explained for the implementation of their methods. The weather data was taken from NCDC analysis.

A. Gayathri et al. [6] worked on the study of weather forecasting using data mining. Different types of forecasting like Now casting, Short range, Medium range and Long range forecasting is explained. Various weather parameters and different data mining techniques is explained along with different classification algorithms like Decision tree, Bayesian classification, Back propagation in the context of weather forecasting.

Riyaz and Surekha [7] worked on the temperature based weather data analysis of NCDC data. The details about MapReduce program execution including results are mentioned. They claimed that Hadoop is good for weather data analysis and has lot of industrial importance.

A Zaslavsky et al. [8] explains the Sensing as a Service and Big Data. Billions of sensing devices are connected to a computer networks and leading to generate huge amount of data on daily basis. Storage and processing of this enormous data is becoming a challenge. To process this data Hadoop, Spark and NoSQL [10, 11] technology can be used.

A Katal, M Wazid and R Goudar [9] talked about the issues, challenges, various tools and good practices about handling a Big Data. Various technical challenges to a computer scientist like fault tolerance, scalability, quality of data and processing of heterogeneous data are mentioned. Parallel programming model like MapReduce, Spark and Distributed File System are proposed as a good tool for Big Data

III. DISCUSSION

Enormous amount of data is being generated in various domains like Social Network, Weather, Scientific Experiment, Stock Market, Bioinformatics etc. This huge data is growing at an exponential speed. Out of this available data, mostly it is in unstructured format. Figure 1 show that more than 80% of data which is available today is in unstructured format while rest 20% comes under the category of structured and semi structured format. Processing unstructured data is challenge because all the available tools require data to be in structured format. Hence, we need technology and solutions which can process this huge unstructured data.

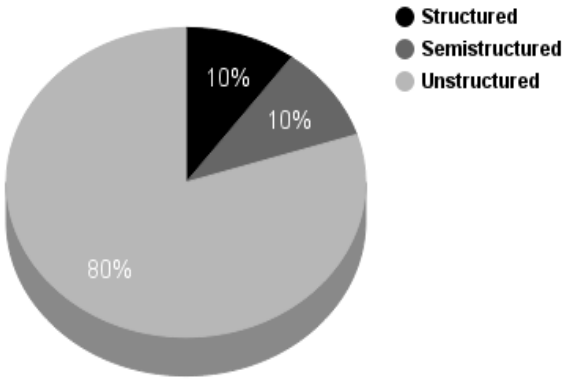

Fig. 1 Distribution of Data

A. Big Data

Big Data can be defined as a large amount of data which requires advanced tools and technologies so that it becomes possible to extract value from it by capturing and analysis process. Big data is more than just size. There are three characteristics of Big Data

a. Volume - It refers to the huge size of data which is being accumulated over the period of time. In case of weather data thousands of sensors are generating data for various weather parameters many times per hour from last many years. It poses the challenge to the storage and processing of huge data.

b. Velocity - It refers to the speed at which data is arriving. If the tool cannot process the data in a time equal to or less than at which it is arriving then it will lead to huge accumulation of unprocessed data.

c. Variety - It refers to various formats in which data is being generated. Like textual data, binary data, images and video data. In ASCII format XML, JSON, CSV are the different format in which data may come. Separate parser is a need to process each type of format.

B. Hadoop

Hadoop is the open source software library framework implemented in java for processing huge data on a cluster of commodity computers in a parallel and distributed manner. Hadoop is good for batch processing job. Hadoop has three major components viz. Hadoop Distributed FileSystem (HDFS), MapReduce Processing Engine and YARN.

1. HDFS - It is the open source implementation of the Google file system published in white paper[12]. It is block oriented, distributed, fault tolerant, reliable, scalable and robust file system supporting huge amount of storage which can give streaming access to a data. The

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

architecture of HDFS is shown in figure 2. It has master slave architecture and it has following main components.

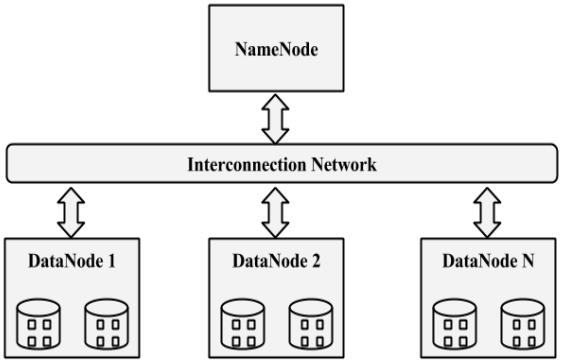

Fig. 2 HDFS Architecture

Name Node - It is the master node of HDFS. It manages the metadata about the file system. File operation take place through the name node only. It is the most important part of HDFS; its failure leads to crash of whole HDFS.

Data Node - It is the node where actual data in the form of block is stored. Two or three or may be more data node can store the same block to allow the redundancy. It leads to data availability even if one of the datanode fails. Data node periodically updates the namenode about their availability so that namenode is aware about the health of the HDFS cluster.

Block Size - HDFS is for storing file of big size in distributed fashion. HDFS stores file in the form of blocks. It is the minimum amount of data that HDFS can read or write. The default block size of file is 64MB in Hadoop 1.x while it is 128MB in Hadoop 2.x.

2. MapReduce - It is the processing engine of Hadoop [13]. MapReduce is the programming paradigm for implementing business logic in the key value pair format. Key value pair is the core of MapReduce programming.

Typically it has Mapper and Reducer but it has intermediate steps also which user may rewrite to override their default behaviour.

Mapper - User data is divided into chunks or blocks of size 128MB each. For each block, a separate Mapper is executed to process that block. Mapper transforms and or filters the input key value into another key value pair(s).

Partitioner- It categorizes the output key value of Mapper and pushes it to a particular Reducer. Programmer can implement the custom partitioner to override the default HashPartitioner which work on the hash value of a key.

Combiner - If the operation in the reducer is associative and commutative then at local Mapper level combiner can run which does exactly the same job as the reducer would do but at smaller data (data from only that Mapper) set so that load on final Reducer is reduced. It is optional but if used appropriately can help in great performance boost and saving of network traffic.

Shuffle and Sort - Before key value is given to the Reducer, the pair is shuffled and sorted so that data goes to a proper reducer in a sorted format with respect to a key which simplifies the task of Reducer. Shuffle and sort are done internally by the Hadoop framework.

Reducer - Reducer does the reduction or aggregation job from the key value generated from the Mapper to produce the final result.

3. YARN - It is Yet Another Resource Negotiator. It takes the responsibility of resource management from Map Reduce engine which was there in Hadoop 1.x. It gives the separation between job execution and resource management. It also enables the platform to run another type of programming model like MPI to be executed on top of it. It has following main components.

Resource Manager - It is responsible for the overall management of the cluster.

Node Manager - It is responsible for management of the particular node. Based on the request of Application Master it reserves and gives the resources to the Application master. Periodically it updates its information to the Resource Manager.

Application Master - When any job runs on the YARN cluster the Resource Manager allocates one Node Manager where Application Master can be scheduled to run. Application Master is responsible for whole resource management for that particular job with the help of Resource Manager.

C. Spark

Apache spark is fast and general purpose engine for large scale data processing [14-15]. Architecture of spark has spark core at it bottom and on top of which Spark SQL, MLlib, Spark streaming and GraphX libraries are provided for data processing. Apache Spark is very good for in memory computing. Spark has its own cluster management but it can work with Hadoop also.

There are three core building blocks of Spark programming. Resilient Distributed Datasets (RDD), Transformations and Action.RDD is an immutable data structure on which various transformations can be applied. After transformation any action on RDD can lead to complete lineage execution of transformation before result is produced.

IV. OBSERVATIONS

1. It is observed that the weather analytics is very useful to every sector of human society.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

2. Temperature, pressure, humidity and wind speed are some of the most important parameter for weather data analysis.
3. For weather data analysis people has mostly used the temperature as weather parameter fortheir analysis.
4. Looking at the tremendous speed at which of weather data is being generated, it is a huge challenge in terms of storage and processing.
5. Big Data technology like Hadoop and Spark are being used as a solution to address the challenges caused by Big Data generation.
6. HDFS has provided the solution for the storage challenge of Big Data
7. Hadoop MapReduce is good for batch processing and is giving good result on Hadoop cluster.
8. Intermediate data between Mapper and Reducer is stored on disks which resultsin the latency of the data access causing slower performance.
9. Spark is providing in memory computing which can be used for streaming data processing, graph data processing, machine learning and iterative computing.
10. Intermediate data between spark operations is stored in the memory itself which avoids the latency which is present in the MapReduce job execution which results in big performance boost up.
11. Hence, there is a need to have a comparison of a performance study between Hadoop MapReduce and for weather data analytics to conclude which Big Data technology is better suited for the analysis.

V. CONCLUSION AND FUTURE WORK

Weather analytics has great influence on human society be it agriculture, tourism, sport event, government planning, news agency, industrial farming etc.Weather data is being generated from various sensors across many locations simultaneously at huge pace. This is producing the challenge for storage and processing. The Big Data technology like Hadoop, Spark can be used efficiently to handle this weather data.

Various studies have done on weather analysis especially for its temperature. There is a need to analyse all important weather parameters like temperature, pressure, humidity and wind speed. Also the technology benchmarking comparison for Hadoop and Spark is important to study which is better suited for weather data analysis.

ACKNOWLEDGMENT

I would like to acknowledge my Principal at NRI Institute of Information Science and Technology Bhopal, INDIA and the staff of college and my friends for supporting and motivating me for my research work. I would like to thank my family members for their support.

REFERENCES

[1] Denissen, Jaap JA, et al. "The effects of weather on daily mood: A multilevel approach" Emotion 8.5 (2008): 662.
[2] Miss. Shraddha V. Shingne, Prof. Anil D.Warbhe and Prof. Shyam Dubey, "Weather Forecasting using Adaptive technique in Data Mining", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, PP: 091 - 095
[3] Veershetty Dagade, Mahesh Lagali,Supriya Avadhani and Priya Kalekar, "Big Data Weather Analytics Using Hadoop",IJETCSE, ISSN: 0976-1353 Volume-14 Issue-02, April 2015
[4] Basvanth Reddy and Prof B. A. Patil, "Weather Prediction on Big Data Using Hadoop Map Reduce Technique", IJARCCE, ISSN: 2278-1021 Volume-05, Issue-06, Page No (643-647), June, 2016
[5] E Sreehari, J. Velmurugan and Dr. M. Venkatesan, "A Survey Paper on Climate Changes Prediction Using Data Mining", IJARCCE, ISSN: 2278-1021 Volume-05, Issue-02, Page No (294-296), February, 2016
[6] A Gayathri, M. Revathi, J. Velmurugan, "A Survey on Weather forecasting y Data Mining", IJARCCE, ISSN: 2278-1021 Volume-05, Issue-02, Page No (298-300), February, 2016
[7] Riyaz P.A., Surekha M.V., "Leveraging Map Reduce With Hadoop for Weather Data Analytics" IOSR Journal of Computer Engineering, Volume 17, Issue 03, May-June 2015
[8] Zaslavsky, Arkady, Charith Perera, and Dimitrios Georgakopoulos. "Sensing as a service and big data." arXiv preprint arXiv:1301.0159 (2013).
[9] Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: issues, challenges, tools and good practices." Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.
[10] Han, Jing, et al. "Survey on NoSQL database." Pervasive computing and applications (ICPCA), 2011 6th international conference on. IEEE, 2011.
[11] Tudorica, Bogdan George, and Cristian Bucur. "A comparison between several NoSQL databases with comments and notes." 2011 RoEduNet International Conference 10th Edition: Networking in Education and Research. IEEE, 2011.
[12] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.
[13] Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system." ACM SIGOPS operating systems review. Vol. 37. No. 5. ACM, 2003.
[14] Zaharia, Matei, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.
[15] Zaharia, Matei, et al. "Spark: cluster computing with working sets." HotCloud 10 (2010): 10-10.

BIOGRAPHIES

Priyanka Chouksey M.Tech Scholar in Computer Science and Engineering at NRI Institute of Information Science and Technology Bhopal, INDIA. She has done B.E. in CSE from Mittal Institute of Technology, Bhopal, INDIA. Her area of research in Algorithms, Data Mining, Big Data Analytics and Distributed Computing.

Abhishek Singh Chauhan An Assistant Professor in Computer Science and Engineering Department at NRI Institute of Information Science & Technology, Bhopal, INDIA. He has completed M.Tech in CSE from Samrat Ashok Technological Institutute and Pursuing PhD. from Bhagwant University, Ajmer, India. His main research interest includes Web Application Security & Network Security.