International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 194-199
e-ISSN:2278-621X

BIG DATA AGGREGATION USING HADOOP AND MAP REDUCE TECHNIQUE FOR WEATHER FORECASTING

Dr. Doreswamy[1], Ibrahim Gad[2] and B.R. Manjunatha[3]

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 thelarge 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 datasetis unpredictable and frequently change according to global climate changes. The data used is from thenational 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.

[1] *Department of Computer Science, Mangalore University,Mangalagangotri- 574 199,Karnataka, INDIA*
[2] *Departmentof Computer Science, Faculty of science, Tanta University, Tanta, Egypt*
[3] *Department of Marine Geology Mangalore University, Mangalagangotri- 574 199, Karnataka, INDIA*

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 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'spublications to have Nine V's characteristics (9V's characteristics)[2]. These 9V's characteristicsare: (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualizationand 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 afaster 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 thatsupports distributed processing of large chunks of data using simple programming models.The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in additionto other modules.

Hadoop is an open-source framework for processing a large amount of data acrossclusters of computers with the use of high-level languages. Its modulesprovide 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. Usercan submit jobs to Hadoop from his desktop machine in remote location from the Hadoop cluster[5].

Two main components of Hadoop are Hadoop Distributed File System (HDFS) and MapReduce[11]. 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 amounts of unstructured data in parallel based on the division of a big dataset into smaller independent chunks see Figure 1.

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 onsee Figure 2[11].

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.

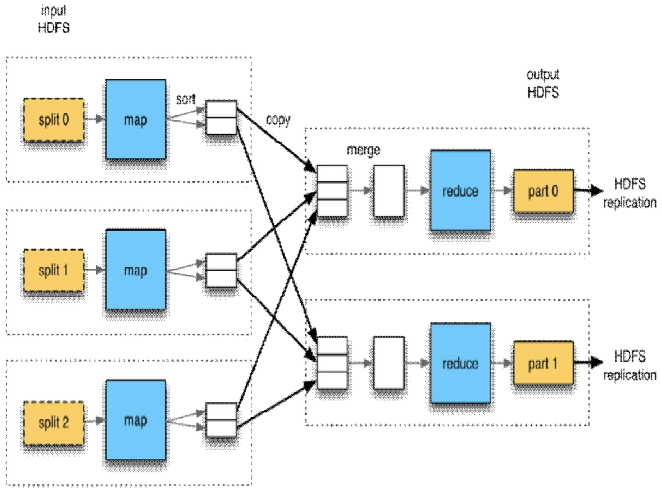

Figure 1. Map Reduce framework

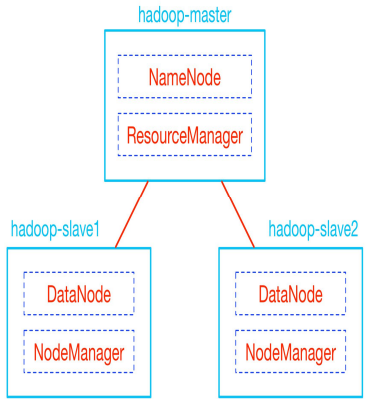

Figure 2. Hadoop cluster

IV. RELATED WORK

Riyaz P.A. et al.[4],describes the analysis of hugeamounts 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 extremelyflexible and scalable to be able to analyze Pentabytesof data across an extremely wideincreasing wealth of weather variables. Data processed by Apache Hadoop and Apache Spark. Experiments performed to select thebest tools among Hadoop using Pig and Hive Queries.

Ramya M. G. et al.[3],explains the meteorological data storage aswell as analysis platform based on Hadoop framework with the help of online logisticregression algorithm for prediction. This platform is based on distributed filesystemHDFS which incorporates distributed database HBase, data warehouse management andefficient query processing tool Hive, data migration tool Sqoop. The best data miningprediction algorithm regression also integrated into the system. This architecture hasan ability of mass storage of meteorological data, efficient query, and analysis, climatechange prediction.

V. EXPERIMENTAL RESULTS

National Climatic Data Center (NCDC) have provideda 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 NCDCcontains the following parameters: station number, station name, date,country, Precipitation, Temperature, and Windas 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 thereducers. 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, andwindin different stations. Each reducer store the final results in HDFSsee Figure 3. Due to apractical limit, the analysis is executed in Hadoop standalone mode. Figure 4 shows MapReduce Framework execution. Figures(6, 7, 8) show the different final resultsof our proposed method.

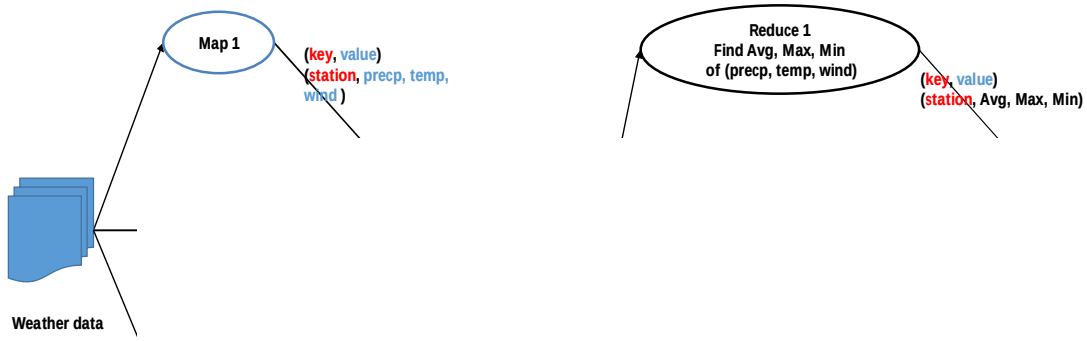

Figure 3. Proposed MapReduce Framework

```
16/10/28 22:06:29 INFO reduce.MergeManagerImpl: Merged 3 segments, 55984 bytes t
o disk to satisfy reduce memory limit
16/10/28 22:06:29 INFO reduce.MergeManagerImpl: Merging 1 files, 55984 bytes fro
m disk
16/10/28 22:06:29 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from
 memory into reduce
16/10/28 22:06:29 INFO mapred.Merger: Merging 1 sorted segments
16/10/28 22:06:29 INFO mapred.Merger: Down to the last merge-pass, with 1 segmen
ts left of total size: 55959 bytes
16/10/28 22:06:29 INFO mapred.LocalJobRunner: 3 / 3 copied.
16/10/28 22:06:29 INFO Configuration.deprecation: mapred.skip.on is deprecated.
Instead, use mapreduce.job.skiprecords
16/10/28 22:06:29 INFO mapred.Task: Task:attempt_local1522631095_0001_r_000000_0
 is done. And is in the process of committing
16/10/28 22:06:29 INFO mapred.LocalJobRunner: 3 / 3 copied.
16/10/28 22:06:29 INFO mapred.Task: Task attempt_local1522631095_0001_r_000000_0
 is allowed to commit now
16/10/28 22:06:29 INFO output.FileOutputCommitter: Saved output of task 'attempt
_local1522631095_0001_r_000000_0' to hdfs://localhost:9000/india/india-SEASON/_t
emporary/0/task_local1522631095_0001_r_000000
16/10/28 22:06:29 INFO mapred.LocalJobRunner: reduce > reduce
16/10/28 22:06:29 INFO mapred.Task: Task 'attempt_local1522631095_0001_r_000000_
0' done.
16/10/28 22:06:29 INFO mapred.LocalJobRunner: Finishing task: attempt_local15226
31095_0001_r_000000_0
16/10/28 22:06:29 INFO mapred.LocalJobRunner: reduce task executor complete.
16/10/28 22:06:29 INFO mapreduce.Job: Job job_local1522631095_0001 running in ub
er mode : false
16/10/28 22:06:29 INFO mapreduce.Job:  map 100% reduce 100%
16/10/28 22:06:29 INFO mapreduce.Job: Job job_local1522631095_0001 completed suc
cessfully
16/10/28 22:06:29 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=147699
                FILE: Number of bytes written=1146561
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=148646
                HDFS: Number of bytes written=2800
```

Figure 4. The output of the MapReduce program

	A	B	C	D	E	F	G
1	420270	SRINAGAR	20150101	IN	0	0.7	0.2
2	420270	SRINAGAR	20150102	IN	0	1	0.3
3	420270	SRINAGAR	20150103	IN	0	3.3	0.6
4	420270	SRINAGAR	20150104	IN	0	3.6	0.3
5	420270	SRINAGAR	20150105	IN	0	4.1	0.2
6	420270	SRINAGAR	20150106	IN	0	3.5	0.2
7	420270	SRINAGAR	20150107	IN	0	3.6	0.3
8	420270	SRINAGAR	20150108	IN	0	2.8	0.2
9	420270	SRINAGAR	20150109	IN	0	2.5	0.2
10	420270	SRINAGAR	20150110	IN	0	2.3	0.2
11	420270	SRINAGAR	20150111	IN	0	2.9	0.2
12	420270	SRINAGAR	20150112	IN	0	3.6	0.3
13	420270	SRINAGAR	20150113	IN	0.3	4.8	0.5
14	420270	SRINAGAR	20150114	IN	0	5.1	0.5
15	420270	SRINAGAR	20150115	IN	0	5.1	0.2
16	420270	SRINAGAR	20150116	IN	0	3.5	0.2
17	420270	SRINAGAR	20150117	IN	0	3.3	0.2
18	420270	SRINAGAR	20150118	IN	0	3.3	0.2
19	420270	SRINAGAR	20150119	IN	0	3.5	0.2
20	420270	SRINAGAR	20150120	IN	0	4.7	0.3
21	420270	SRINAGAR	20150121	IN	0	4.4	0.4
22	420270	SRINAGAR	20150122	IN	3	3.9	0.2
23	420270	SRINAGAR	20150123	IN	0	4.1	0.3
24	420270	SRINAGAR	20150124	IN	0	3.4	0.2
25	420270	SRINAGAR	20150125	IN	0	4.6	0.3
26	420270	SRINAGAR	20150126	IN	0	3.9	0.4
27	420270	SRINAGAR	20150127	IN	0	2.8	0.5
28	420270	SRINAGAR	20150128	IN	0	3.6	0.3
29	420270	SRINAGAR	20150129	IN	0.3	2.6	0.4

Figure 5. Weather data set

```
SRINAGAR-01 AVG PRCP=0.13, AVG TEMP=0.03, AVG WIND=0.00
 ,MAX PRCP=3, MAX TEMP=1,MAX WIND=0
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

SRINAGAR-02 AVG PRCP=2.29, AVG TEMP=0.14, AVG WIND=0.00
 ,MAX PRCP=45, MAX TEMP=2,MAX WIND=0
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

SRINAGAR-03 AVG PRCP=0.16, AVG TEMP=0.55, AVG WIND=0.00
 ,MAX PRCP=2, MAX TEMP=14,MAX WIND=0
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

SRINAGAR-04 AVG PRCP=2.03, AVG TEMP=1.43, AVG WIND=0.00
 ,MAX PRCP=45, MAX TEMP=18,MAX WIND=0
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

SRINAGAR-05 AVG PRCP=0.68, AVG TEMP=2.45, AVG WIND=0.00
 ,MAX PRCP=17, MAX TEMP=21,MAX WIND=0
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------
```

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

```
AMRITSAR-2014   AVG PRCP=0.89, AVG TEMP=1.58, AVG WIND=0.07
 ,MAX PRCP=146, MAX TEMP=33,MAX WIND=3
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

SRINAGAR-2015   AVG PRCP=0.95, AVG TEMP=0.95, AVG WIND=0.00
 ,MAX PRCP=49, MAX TEMP=26,MAX WIND=1
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------
```

Figure 7. The final result of station with Year

```
AMRITSAR-2014-FALL  AVG PRCP=2.36, AVG TEMP=1.87, AVG WIND=0.04
 ,MAX PRCP=146, MAX TEMP=26,MAX WIND=3
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

AMRITSAR-2014-SPRING    AVG PRCP=0.49, AVG TEMP=1.63, AVG WIND=0.15
 ,MAX PRCP=18, MAX TEMP=26,MAX WIND=3
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

AMRITSAR-2014-SUMMER    AVG PRCP=0.63, AVG TEMP=1.30, AVG WIND=0.10
 ,MAX PRCP=32, MAX TEMP=33,MAX WIND=2
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
 ------------------------------------------------------------

AMRITSAR-2014-WINTER    AVG PRCP=0.07, AVG TEMP=1.52, AVG WIND=0.00
 ,MAX PRCP=2, MAX TEMP=16,MAX WIND=0
 ,MIN PRCP=0, MIN TEMP=0,MIN WIND=0
```

Figure 8. The result for Seasons of each station.

VI.CONCLUSION

In case of, using traditional systems to process millions of records is time consuming process. In the era of Internet of things, the meteorological department uses different sensors to get the temperature, humidity etc. MapReduce is a framework for executing distributable algorithms across huge datasets using a large number of computers. The major advantage of MapReduce with Hadoop frameworkspeeds up the processing of data. Using MapReduce with Hadoop, the weather data can be analyzed effectively

ACKNOWLEDGEMENT

The authors thank for financial assistance provided by the SERB-DST vide SB/EMEQ-137/2014, dated 21-03-2016, the ICCR fellowship by the Governments of India, and Egypt for carrying the project

REFERENCES

[1] Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, and Bongki Moon. Parallel data processing with mapreduce: a survey. *AcMsIGMoD Record*, 40(4):11–20, 2012.

[2] Suhail Sami Owais and Nada Sael Hussein. Extract five categories cpivw from the 9v's characteristics of the big data. *International Journal of Advanced Computer Science & Applications*,7 (3):254–258.*2016*

[3] M Ramya, ChetanBalaji, and L Girish. Environment change prediction to adapt climate-smart agriculture using big data analytics. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume*, 4, May 2015.

[4] Surekha Mariam Varghese Riyaz P.A. Leveraging map reduce with hadoop for weather data analytics. *IOSR Journal of Computer Engineering*, 17(3), May- Jun 2015.

[5] Caesar Wu, RajkumarBuyya, and KotagiriRamamohanarao. "Big data analytics = machine learning + cloud computing". *CoRR*, abs/1601.03115, 2016.

[6] SupriyaAvadhaniPriya KalekarVeershetty Dagade, Mahesh Lagali. Big data weather analytics using hadoop. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, 14(2), APRIL 2015.

[7] Toshniwal, Raghav, KanishkaGhoshDastidar, and AsokeNath. "Big Data Security Issues and Challenges." Complexity 2.2, 2015.

[8] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. "Big data: The next frontier for innovation, competition, and productivity", 2011.

[9] Gantz, John, and David Reinsel. "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", IDC iView: IDC Analyze the Future 2007, pp: 1-16, 2012.

[10] IBM, Big Data at the Speed of Business, "http://www-01.ibm.com/software/data/bigdata/", 2012.

[11] Kaur, Anureet. "Big Data: A Review of Challenges, Tools and Techniques. IJSRSET, 2(2),2016.

[12] National Climatic Data Centre, http://www.ncdc.noaa.gov

[13] S. Vijayarani,S. Maria Sylviaa ,A.Sakila, "Clustering Algorithms for Outlier DetectionPerformance Analysis" International Conference on Computing and Intelligence Systems, 04, 1213-1217, March 2015