Analyzing Weather Data impacts using Apache Hadoop

Dr S. Anitha[1] M. Kalaivani[2]

[1]Assistant professor, PG Department of Computer Science,
Dwaraka Doss Goverdhan Doss Vaishnav College,
Arumbakkam, email: anitasenthil@gmail.com

[2]Assistant professor, PG Department of Computer Science,
Dwaraka Doss Goverdhan Doss Vaishnav College,
Arumbakkam, email: kalaivani@dgvaishnavcollege.edu.in

Abstract: In data science era, the scale of weather data is enormous and rising rapidly. Apache Hadoop is a fast and efficient framework which has been used in many applications in big data field. However, for the large-scale weather dataset, the traditional algorithms are not capable enough to satisfy the genuine application requirements efficiently. Hadoop is a framework which deals with Big and Huge variety of datasets which supports processing components that collectively called Hadoop Ecosystem. This paper proposes efficient weather data analyses are carried out by Apache MapReduce and Apache Pig in Hadoop framework. Weather datasets are taken from NCDC Database for this proposed research. The impacts of Weather analysis are obtained from both Mapreduce and Apache pig and they were compared.

Keywords: Big data Analytics, Hadoop Ecosystem, Apache MapReduce, NCDC Datasets

1. Introduction:

In the atmospheric sciences, weather data is really rich and valued, which requires a mass of scientific computing, and provides services to the communities. Climate data are dramatically increasing in volume and complexity, since users of these data in the scientific community and the public are rapidly increasing [1]. For analyzing the large number of data, traditional data analysis techniques have failed to carry out analysis on larger data sets effectively. Newly, the dominant platform that has proved in processing hefty sets of data is Hadoop, which is considered to be operative for distributed file processing and distributed storage of wider range of data. The main component of Hadoop is HDFS and Mapreduce. MapReduce is a programming model for computing bigger data sets and HDFS is a Hadoop Distributed File System that stores data in the type of memory blocks and distributes the data across cluster of nodes. In this paper, Apache Mapreduce is used to analyze the NCDC weather dataset. This paper is organized as follows: Section 2 includes the review of literature and an outline of the paper. Section 3 undertakes the research methodology and Section 4 discusses the Results and Discussion of this study. Section 5 concludes the proposed method and scope for future work.

2. Review of Literature:

MapReduce is a key technology of using cloud computing to process a big amount of data. It is a parallel programming model and an associated implementation for processing and generating huge datasets in a broad variety of real world tasks proposed by Google. It is not only a programming model, but also a task scheduling model. It is compose of two essential functions: Map and Reduce, defined by users. A Map function is used to handle every Input and

convert it as an intermediate key/value pair. A Reduce function is specified to combine all of the intermediate value with the same middle key [2]. Google MapReduce is typically utilized to perform distributed computing on clusters of computers. Thereby, the effect originally achieved only by expensive high-performance computer can be achieved by low-cost computing services.

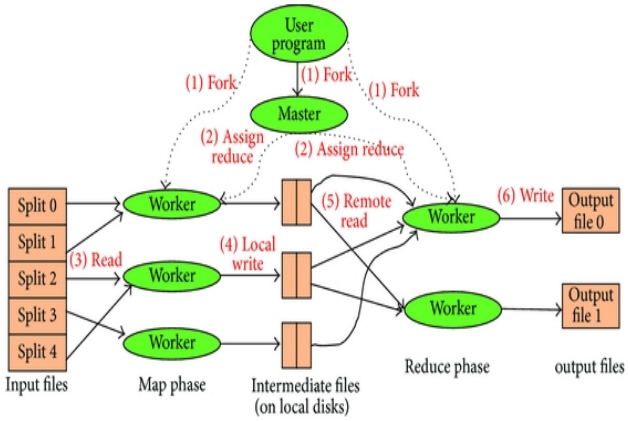

Fig-1.Mapreduce paradigm

2. Research Methodology

In this research, Analyzing weather data of Fairbanks, Alaska is utilized to find cold and hot days using MapReduce and Hadoop. The National Climatic Data Center (NCDC) is the world's largest active archive of weather data. In this paper, this research work consists of the following five phases,

1. Pre-processing the NCDC datasets
2. Feature selection from the datasets
3. Loading the preprocessed dataset to HDFS
4. Analysis of NCDC weather data set with Mapreduce and Pig
5. Comparing and Evaluating the results.

Fig-2.Mapreduce Function

Table-1 Steps to Mapreduce jobs run:

Step1: for Compiled the Java File: javac -classpath /home/student3/hadoop-common-2.6.1.jar:/home/student3/hadoop-mapreduce-client-core-2.6.1.jar:/home/cloudera/commons-cli-2.0.jar -d . MaxTemperature.java MaxTemperatureMapper.java MaxTemperatureReducer.java **Step 2**:Created the JAR file: jar -cvf hadoop-project.jar *class **Step 3**:Executed the jar file: hadoop jar hadoop-project.jar MaxTemperature /home/student3/Project/ /home/student3/Project_output111 **Step 4:**Copy the output file to local hdfs dfs -copyToLocal /home/student3/Project_output111/part-r-00000

3. Dataset Description

In this paper, Apache Map-Reduce weather analysis algorithm is applied in NCDC dataset for analyzing weather datasets to predict the weather condition for a particular year. The NCDC weather dataset is downloaded for year 1930 and loaded it in HDFS system. MapReduce and Pig algorithm is implemented in dataset to find the Min, Max, avg temparature for diffrent stations. Maximum and Minimum temperature are retrieved and are used to find the cold and hot data respectively. NCDC (National Climatic Data Center) is collected across more than 116 weather stations and more than 1000 observations centers. The data is unstructured, which becomes a challenging task to analyze it. Weather sensors are gathering weather statistics throughout the world in a huge volume of log data. NCDC weather data is unstructured and record-oriented. In this dataset, each row has lots of fields like longitude, latitude, daily max-min temperature, daily average temperature, etc. temperature is taken as the main element.

4. Results and Discussion

The data were cleaned, preprocessed, and then fed into mapreduce algorithm and Apache Pig. Hadoop is installed in pseudo distributed mode. The performance evaluation between the pig and Mapreduce is depicted in Fig-3 and below are the commands for the performance of hadoop. And according to the data analyzing speed and efficiency Apache Pig proves to be better than Mapreduce.

Table-2 output of weather data analysis.

In hdfs environment: Create the temporary content file in the input directory:

[cloudera@quickstart ~]$ hdfs dfs -mkdir weather_dir

Put the file.txt into hdfs:

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Desktop/wd.txt weather_dir/

[cloudera@quickstart ~]$ hdfs dfs -ls weather_dir/

Found 1 items

-rw-r--r-- 1 cloudera cloudera 41881 2019-10-09 22:16 weather_dir/wd.txt

To see the content of the file:

[cloudera@quickstart ~]$ hdfs dfs -cat weather_dir/wd.txt

23907 20150101 2.423 -98.08 30.62 2.2 -0.6 0.8 0.9 6.2 1.47 C 3.7 1.1 2.5 99.9 85.4 97.2 0.369 0.308 -99.000 -99.000 -99.000 7.0 8.1 -9999.0 -9999.0 - 9999.0 23907 20150102 2.423 -98.08 30.62 3.5 1.3 2.4 2.2 9.0 1.43 C 4.9 2.3 3.1 100.0 98.8 99.8 0.391 0.327 -99.000 -99.000 -99.000 7.1 7.9 -9999.0 -9999.0 - 9999.0 23907 20150103 2.423 -98.08 30.62 15.9 2.3 9.1 7.5 2.9 11.00 C 16.4 2.9 7.3 100.0 34.8 73.7 0.450 0.397 -99.000 -99.000 -99.000 7.6 7.9 -9999.0 -9999.0 - 9999.0 23907 20150104 2.423 -98.08 30.62 9.2 -1.3 3.9 4.2 0.0 13.24 C 12.4 -0.5 4.9 82.0 40.6 61.7 0.414 0.352 -99.000 -99.000 -99.000 7.3

To see the output:

[cloudera@quickstart ~]$ hdfs dfs -ls out/

Found 2 items

-rw-r--r-- 1 cloudera cloudera 0 2019-10-09 22:20 out/_SUCCESS

-rw-r--r-- 1 cloudera cloudera 4632 2019-10-09 22:20 out/part-r-00000

[cloudera@quickstart ~]$ hdfs dfs **-cat out/part-r-00000**

```
1 The Day is Cold Day :20200101   -21.8
2 The Day is Cold Day :20200102   -23.4
3 The Day is Cold Day :20200103   -25.4
4 The Day is Cold Day :20200104   -26.8
5 The Day is Cold Day :20200105   -28.8
6 The Day is Cold Day :20200106   -30.0
7 The Day is Cold Day :20200107   -31.4
8 The Day is Cold Day :20200108   -33.6
9 The Day is Cold Day :20200109   -26.6
10 The Day is Cold Day :20200110   -24.3
```

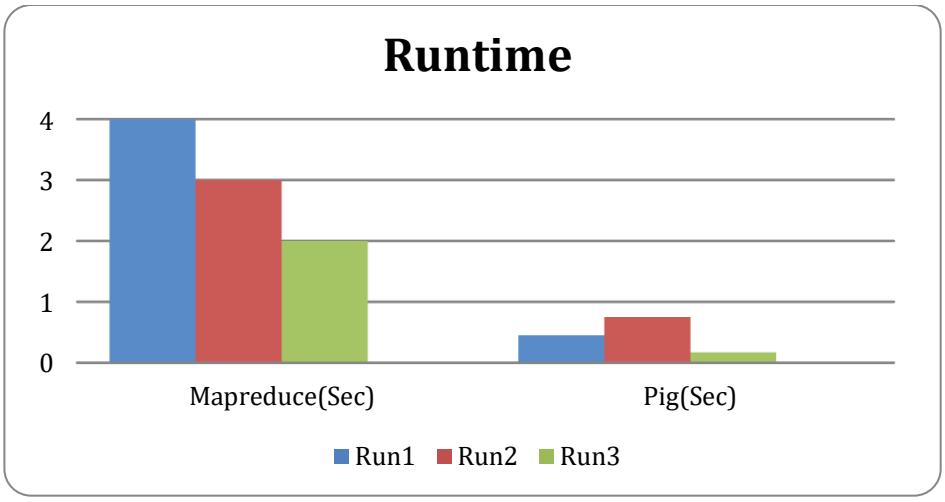

Fig-3 Comparison of Runtime –Mapreduce and Apche Pig

5 Conclusion

Weather data analysis algorithm is applied on the NCDC datasets. The analysis shows Cold and hot days along with the temperature. Mapper and reducer classes of Map reduce are used for analyzing the dataset and the results were compared in terms of execution time. Pig is outperformed and gained less elapsed time than Mapreduce. From this study, Apache Pig is the best tools for analyzing large number of instances in short execution time. Considering the merits and demerits of the proposed system.

Reference:

[1] Big Data Weather Analytics Using Hadoop" an International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume14 Issue 2.

[2] Novel Weather Data Analysis Using Hadoop and MapReduce" – A Case Study, 2019

[3] Arribas-Bel, Accidental, open and everywhere: Emerging data sources for the understanding of cities. Applied Geography, forthcoming.

[4] Chouksey P., Chauhan A., "A Review of Weather Data Analytics using Big Data", IJARCCE, ISSN: 2278- 1021 Volume-06, Issue01, Page No (365-368), January, 2017.

[5] 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.

[6] Riyaz P.A., Surekha M.V., "Leveraging MapReduce With Hadoop for Weather Data Analytics" IOSR Journal of Computer Engineering, Volume 17, Issue 03.

[7] 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.

[8] Riyaz P.A., Surekha M.V., "Leveraging MapReduce With Hadoop for Weather Data Analytics" IOSR Journal of Computer Engineering, Volume 17, Issue 03