Weather Data Analysis Using Hadoop: Applications and Challenges

Weather Data Analysis Using Hadoop: Applications and Challenges

Mohammed Adam Ibrahim Fakherldin[3 *], Khalid Adam[2], Noor Akma Abu Bakar[1] and Mazlina Abdul Majid[1]

[1]Faculty of Computer System & Software Engineering, Universiti Malaysia Pahang
[2]Faculty of Electric and Electronic Engineering, Universiti Malaysia Pahang
[3]Faculty of Computer Science and Information Systems, Jazan University

Email: khalidwsn15@gmail.com

Abstract. Weather data is very crucial in every aspect of human daily life. It plays an important role in many sectors such as agriculture, tourism, government planning, industry and so on. Weather has a variety of parameters like temperature, pressure, humidity and wind speed. The meteorological department deployed sensors for each weather parameter at different geographical locations to collect data. This data is stored mostly in the unstructured format. Thus, a big amount of data has been collected and archived. Therefore, storage and processing of this big data for accurate weather prediction is a huge challenge. Hadoop an apache product it used to support big data sets in a distributed environment. Hadoop has greatest advantages over scalable and fault-tolerant distributed processing technologies. 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.

1. Introduction

Big Data has become one of the buzzwords in IT, during the last couple of years. Originally it was created by companies which had to manage fast increase rates of data such as web data, data resulting from scientific or business simulations or other data sources. Some of those business companies' models are basically based on indexing and using this large amount of data. The challenges to handle the fast growing of data amount on the web e.g. lead Google to develop the Google File System [1] and MapReduce [2].

Furthermore, most of the cities have become smart. Thus, many sensors devices utilized in smart city can be used to measure weather parameters [3]. Which led the weather department collect and analysis huge amount of data like temperature [4]. These different sensors value such as temperature, humidity to predict the rain fall etc. Consequently, when the number of sensors/devices increases, the data becomes high volume and velocity [5]. Therefore, there is an essential of a scalable analytics tool to process massive amount of data.

However, the conventional method of process the data is very slow. Compare to process the sensor data with Hadoop framework which remove the scalability bottleneck. Hadoop is a framework used for handling huge amount of data. Mainly the processing engine is MapReduce, which is currently one of the most common big data processing frameworks available. MapReduce is a framework for performing highly parallelizable and distributable algorithms across big data sets using commodity computers cluster [6]. Thus, using Hadoop/MapReduce the temperature can be analyses without scalability problems. The speed of processing data can improve rapidly when across multi cluster distributed network.

2. Related Works

[7] describe 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 datasets available for weather forecast. Its total size is around 20 GB. Results show that temperature analyzed effectively by Using MapReduce with Hadoop.

[6] gives a detailed description of build a platform that is extremely flexible and scalable to be able to analyse petabytes 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.

[8] 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 file system HDFS which includes distributed database HBase, data warehouse management and useful 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.

3. Experimental Setup and Results

The experiments were carried out in a physical cluster environment, the researcher used three computers. Hadoop cluster on Linux Ubuntu 14.04 where one computer ran a NameNode and ResourceManager and the remaining ran Datanode and DataManager. Each of the computers has the following configuration: Core i7 processor, 4 GB main memory, and 1 TB disk space as shown in figure 1. The researcher used a Hadoop-2.7.1. The max replication factor “dfs.replication.max” is used to set the replication limit of blocks.

Figure 1. Hadoop cluster

The proposed System use dataset of NCDC contains the following parameters: station number, station name, date, country, Precipitation, Temperature, and Wind and so on as shown in figure 2. The data files are stored in HDFS as shown in figure 3. 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 stores the final results in HDFS as show in figure 4.

This stage incorporates the goals of data stored in HDFS. This is finished utilizing MapReduce framework. The fundamental procedure of MapReduce is shown in figure 2. MapReduce structure takes a split-apply-reduce combine strategy. The last output is the value which has more key-value occurrences.

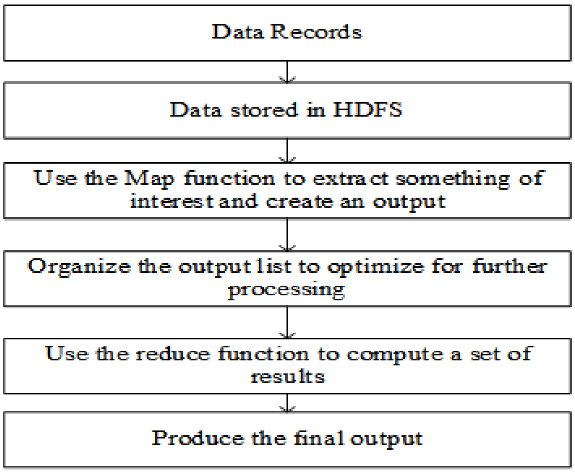
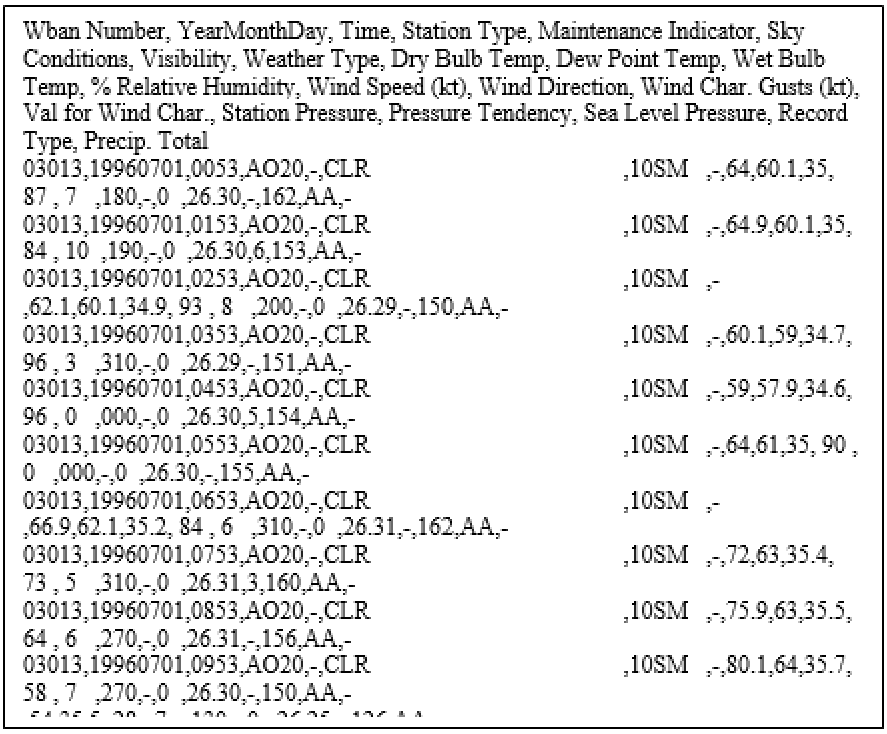
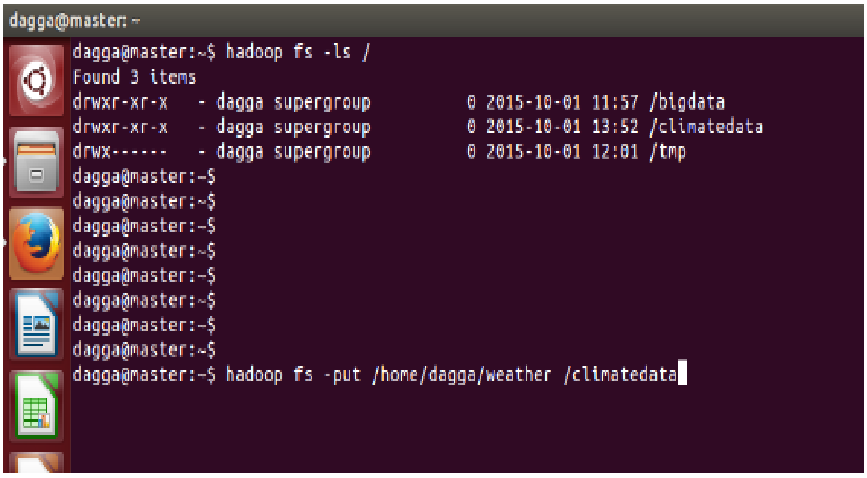

Data Records
Data stored in HDFS
Use the Map function to extract something of interest and create an output
Organize the output list to optimize for further processing
Use the reduce function to compute a set of results
Produce the final output

Figure 2. MapReduce flow chart process

Table 1. Weather Dataset into ASCII Format

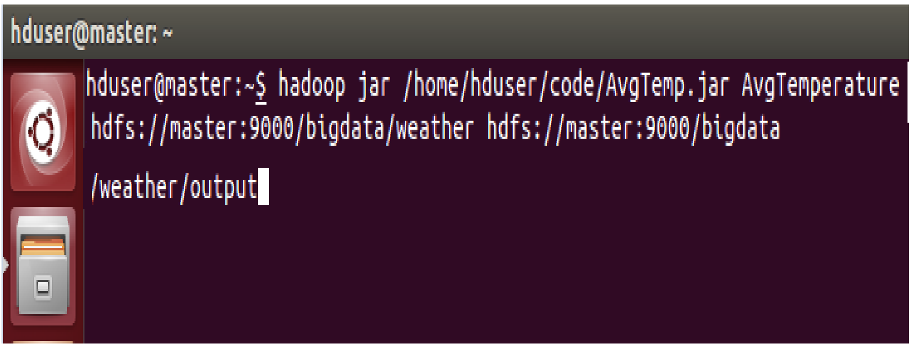

Wban Number, YearMonthDay, Time, Station Type, Maintenance Indicator, Sky Conditions, Visibility, Weather Type, Dry Bulb Temp, Dew Point Temp, Wet Bulb Temp, % Relative Humidity, Wind Speed (kt), Wind Direction, Wind Char. Gusts (kt), Val for Wind Char., Station Pressure, Pressure Tendency, Sea Level Pressure, Record Type, Precip. Total

```
03013,19960701,0053,AO20,-,CLR                    ,10SM  ,-,64,60.1,35,
87 , 7  ,180,-,0  ,26.30,-,162,AA,-
03013,19960701,0153,AO20,-,CLR                    ,10SM  ,-,64.9,60.1,35,
84 , 10  ,190,-,0  ,26.30,6,153,AA,-
03013,19960701,0253,AO20,-,CLR                    ,10SM  ,-
,62.1,60.1,34.9, 93 , 8  ,200,-,0  ,26.29,-,150,AA,-
03013,19960701,0353,AO20,-,CLR                    ,10SM  ,-,60.1,59,34.7,
96 , 3  ,310,-,0  ,26.29,-,151,AA,-
03013,19960701,0453,AO20,-,CLR                    ,10SM  ,-,59,57.9,34.6,
96 , 0  ,000,-,0  ,26.30,5,154,AA,-
03013,19960701,0553,AO20,-,CLR                    ,10SM  ,-,64,61,35, 90 ,
0  ,000,-,0  ,26.30,-,155,AA,-
03013,19960701,0653,AO20,-,CLR                    ,10SM  ,-
,66.9,62.1,35.2, 84 , 6  ,310,-,0  ,26.31,-,162,AA,-
03013,19960701,0753,AO20,-,CLR                    ,10SM  ,-,72,63,35.4,
73 , 5  ,310,-,0  ,26.31,3,160,AA,-
03013,19960701,0853,AO20,-,CLR                    ,10SM  ,-,75.9,63,35.5,
64 , 6  ,270,-,0  ,26.31,-,156,AA,-
03013,19960701,0953,AO20,-,CLR                    ,10SM  ,-,80.1,64,35.7,
58 , 7  ,270,-,0  ,26.30,-,150,AA,-
```



```
dagga@master: ~
dagga@master:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dagga supergroup          0 2015-10-01 11:57 /bigdata
drwxr-xr-x   - dagga supergroup          0 2015-10-01 13:52 /climatedata
drwx------   - dagga supergroup          0 2015-10-01 12:01 /tmp
dagga@master:~$
dagga@master:~$
dagga@master:~$
dagga@master:~$
dagga@master:~$
dagga@master:~$
dagga@master:~$
dagga@master:~$
dagga@master:~$ hadoop fs -put /home/dagga/weather /climatedata
```

Figure 3. Push the dataset into climatedata folder



```
hduser@master: ~
hduser@master:~$ hadoop jar /home/hduser/code/AvgTemp.jar AvgTemperature
 hdfs://master:9000/bigdata/weather hdfs://master:9000/bigdata
 /weather/output
```

Figure 4. Run the weather dataset

From the above figures 3 and 4, the last output will be climate information as it has increasingly key events. MapReduce is an effective way of data resolution of large amount of data in a cluster. The MapReduce program is made of a map () procedure that performs percolating and sorting the required data into the DataNode. This module follows a split-apply combine strategy.

4. Conclusion

In case of using traditional systems, to process millions of sensors data it is time consuming. Today IoT and the meteorological department uses various of sensors devices to collect data (e.g. temperature, humidity etc). Hadoop/MapReduce is a framework for processing huge amount in distributable way across large number of computers cluster. Using this framework, the sensors data can be analyzed efficiently. The major benefit of Hadoop framework speeds up the processing of huge data. Where the volume of data is increasing every day.

References

[1] Ghemawat S, Gobioff H, and Leung S T 2003 *The Google file system* **37** 5

[2] Dean J and Ghemawat S 2008 MapReduce: simplified data processing on large clusters *Commun. ACM*, **51** 1 107–113.

[3] Ge M, Bangui H, and Buhnova B 2018 Big Data for Internet of Things: A Survey *Futur. Gener. Comput. Syst.*

[4] Ismail K A, Majid M A, Zain J M, and Bakar N A A 2016 Big Data prediction framework for weather temperature based on MapReduce algorithm *Open Systems (ICOS), 2016 IEEE Conference* 13–17.

[5] Hammad K, Fakharaldien M, Zain J, and Majid M 2015 Big data analysis and storage *International Conference on Operations Excellence and Service Engineering* 10–11.

[6] Dagade V, Lagali M, Avadhani S, and Kalekar P 2015 Big data weather analytics using Hadoop *Int. J. Emerg. Technol. Comput. Sci. Electron* **14** 2.

[7] Riyaz P and Varghese S M 2015 Leveraging map reduce with hadoop for weather data analytics *IOSR J. Comput. Eng* **17** 3.

[8] Ramya M, Balaji C, and Girish L 2016 Environment Change Prediction to Adapt Climate-Smart Agriculture Using Big Data Analytics *Int. J. Adv. Res. Comput. Eng. Technol.* **4** 201.