

Big Data Technique for the Weather Prediction using Hadoop MapReduce

Khalid Adam¹, Izzeldin I. Mohamed^{2*}, Nidal A. Ahmed³, Younis M. Younis³, and Nazar Elfadil³

^{1,2}College of Engineering, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia

³Faculty of Computing, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia

³College of Internet of Things Engineering, Hohai University, Jiangsu Sheng, China

³College of Computing, Fahad Bin Sultan University, Tabuk, Kingdom of Saudi Arabia

Abstract—Currently analyzing large amounts of data has become a big challenge. This data could be medical, scientific, financial, climatological, or marketing. Several techniques are used to analysis meaningful information by use Big Data technologies. Weather one of area use big data technologies to support numerous important domain such as water resources, agriculture, air traffic, and tourism. Weather prediction is field of meteorology that is done by collecting data from the different stations related to the current state of the weather like Temperate, Humidity and Visibility. Thus, the most challenging problem for scientists to analysis this big amount of data. in this paper we focus on analyzing the weather dataset using Hadoop/MapReduce and we used the historical dataset from NOAA. The temperature, humidity and visibility attributes has been extracted from the dataset by the MapReduce Algorithm into structure data. Graphical analysis has been used to represent the result from the MapReduce Algorithm.

Keywords—Weather Prediction, Big Data, Hadoop, MapReduce

I. INTRODUCTION

Weather changes are one of the major problems in the world. For example, in 2014 Malaysia faced cool temperature resulted several diseases occurs, such as flu. In Thailand, due to this change, more than three peoples die due to cool temperature less than 10 Celsius and it is not a new thing. Based on upper quote, it could be concluded; the weather changes involve changes in climate condition in a series of data. The detection of this event could involve three major thresholds are velocity (how rapid the changes in weather data changes), variety (it is possible the changes in a series of data affected other data) and veracity (the effect of velocity and variety could change the nature of the data), and these called the 3V of Big Data characteristics [1].

Thus, the main idea about “Big Data” it refers to the 3Vs characteristics which are volume of the data that used to analyze the challenging of the weather data issues. Then, combined with statistical methods and computational intelligence approaches. Thus, Big Data revolution to various traditional research domains including weather, healthcare, genomics, biological and environmental among many others. The important of Big Data “let data talking” thus, when the data volume is big enough (Big Data), we will be discovered new value from the data via the big data tools (i. e., Hadoop, spark, statistical techniques). According to Bryson “Weather is the original Big Data problem” It has been discussed earlier though many approaches including weather forecasting. Nick Wakeman with reference to Hurricane Sandy stated in his paper about the importance of Big Data in weather forecasting and with the help of available data, in three-days out, forecasters predicted within 10 miles where landfall would occur [2].

As result Big Data has big impact in the weather prediction, which its application of science and technology to predict the conditions of the atmosphere for a given location and time by collecting quantitative data about the current state of the atmosphere at a given place [3]. The Big Data weather prediction of climate has proven to be very important and useful for it is always relating to the judgment of government to help protect the growth of agricultural crops, the warning of hurricanes, floods and so on. Nevertheless, the conditions' analysis is combined with Big Data, which made it a significant challenge and a good candidate for Big Data tools such as Hadoop.

Hadoop/MapReduce is one of the best-known Big Data tool for turning raw data into useful information. It is a method for taking Big Data sets and performance computations on it across cluster of computers in parallel way. It serves as a model for how to analysis Big Data and is often used to refer to the actual implementation of this model [4]. Thus, Hadoop with MapReduce has been the leading open source framework for many years. The aim of this paper is using Big Data Tool (i.e, Hadoop) for the weather predication [5]. In this paper, we used dataset from national oceanic and atmospheric administration (NOAA) (<http://www.noaa.gov/>) as it shown in Figure 1 to created big data model for weather predication (i, e., humidity, visibility, temp). Therefore, the paper organized big data concepts and characteristics in section II, Hadoop/MapReduce in section III. Then related works in section IV, the experimental results and discussion presented in section V and finally the conclusion presented in section VI.

```

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 Tendancy, 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
    
```

Figure1: NOAA Weather data

II. BIG DATA AND WEATHER DATA

The weather Big Data is collected by the weather station directly. The weather information in an area usually influences each other. Thus, we make a hypothesis that the weather conditions can be calculated for a site from weather data collected from the stations around the target site using a model constructed from the data analysis. Ramya M. G. et al. [6], provided platform for data storage and analysis based on Hadoop and machine learning algorithm (i.e., logistic regression) for weather prediction. This platform can storage massive amount of data thus has ability of analysis and weather changes predication.

III. HADOOP AND MAPREDUCE

Doug Cutting, the maker of Apache Lucene, Hadoop gives a complete toolset for the distributed processing of Big Data sets across clusters of computers using simple programming models, including data analysis, data storage and coordination [7]. It is an open source framework for storing data and running application on clusters of commodity hardware. After realizing the traditional enterprise data analysis cannot handle big volumes of structured and unstructured data more efficiently. Thus, Hadoop, its initial cost savings are dramatic and continue to grow as your organizational data grows. Additionally, Hadoop has a strong Apache community behind it that

continues to contribute to its advancement, to general society. Almost a year later all Nutch algorithms had been ported to utilize MapReduce and HDFS as shown in Figure 2.

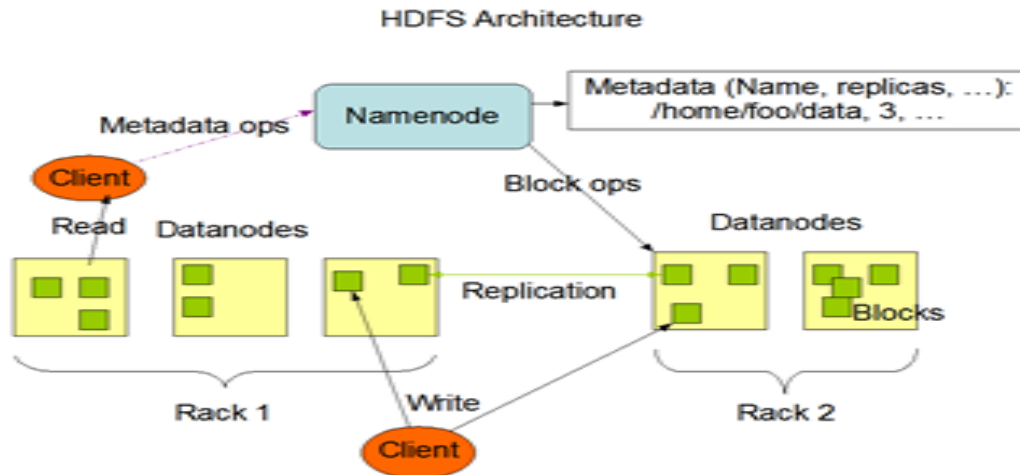


Figure2: MapReduce framework

IV. RELATED WORKS

Weather prediction is a hot topic. Much related research has been done with different models and methods. This section covers a number of papers in this field, which are published recently. In [8], the author introduces a novel Train Delay Prediction System whose function is to provide an integrated and holistic view of operational performance and enable high levels of rail operations efficiency of the Italian railway network. The weather data is working as the exogenous sources combined with historical data of train movements to build this reliable and robust data-driven model. Four different methods such as RFI systems, Random Forest algorithm, kerneled version of RLS and Extreme Learning Machine are supplied to make the comparisons to prove the novel RFI with weather data perform up to twice better than the current state-of-the-art-methodologies. Meanwhile, more data sources would be adding to the system as the conditions to improve the new system in the future. [9] describe a data mining study of agricultural meteorological patterns collected from the meteorological center of Bengaluru district. K-means and Hierarchical clustering techniques are used to exact patterns and obtain results, which play a crucial role in the decision making for sustainable agriculture. The result analysis is clustered by the types of crops such as mango, grapes, potatoes, and so on for each one has different conditions for analysis. The results show that the cluster techniques are effective to predict the information of weather details and the Hierarchical algorithm performs better than K-means.

In [10], the authors introduce a project that aims to forecast the chances of rainfall by using predictive analysis in Hadoop. This model captures relationships among many factors in the data to assign a score or weight pattern for future rainfall prediction by using historical data. The process is in an efficient manner for the large volume of data can be well processed by the big data techniques. The main method for the analysis is classify the weather type by using Naive Bayes with the weather data attribute of humidity. The system design and the plot of mean, maximum and precipitation parameter of humidity are supplied to improve that the more weather information can be efficiently predicted by using Naive Bayes in Hadoop Framework.

Po-Chen Chen and Mladen Kezunovic demonstrate a way to utilize historical weather data and climate change projections in a large (macro) geographical area to predict future electric load patterns in a relatively small (micro) geographical area in their paper [11]. The impact of temperature rising is based on the load while the deviations of the result is large depending on the changing data. Both the data and model are from Coupled Model Intercomparing Project. The future and historical peak load consumption are both supplied to show that the novel framework is proposed and its

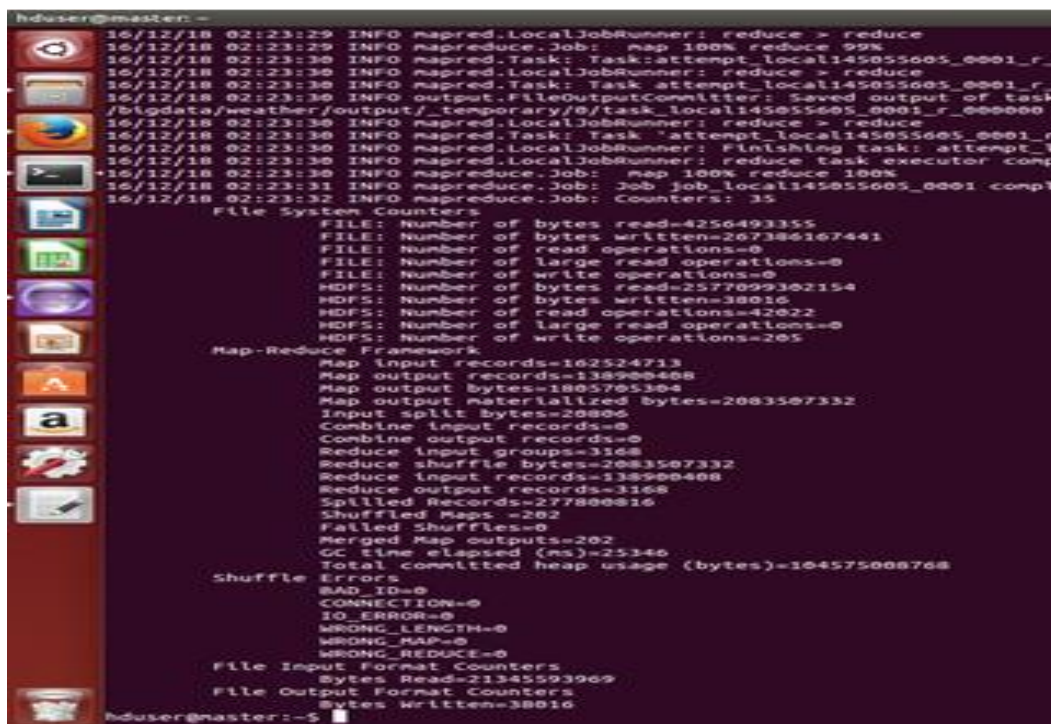
efficiency is higher than most models, which are also aimed at this research for it shows larger numbers for the temperature's increase.

Vincent and Katherine describe a method for loss compression of weather data [12] by representative the data as a sparse and adaptive subset, and the output used for solving an optimization problematic for the minimal loss of information. The series methods are combined with Numerical Weather Prediction (NWP) to support users who need substantially smaller datasets in exchange for some loss of data. Mean squared error (MSE) and peak-signal to noise ratio (PSNR) are used to judge the performance of various algorithms while the long running Genetic Algorithm (GA) gives the highest PSNR and the least loss of information. They enhance the result and reduced the amount of data and the loss of information can be minimized by use of adaptive sampling.

In [13], technique of a geostatistical interpolation called Kriging for short-term predications, and the weather data is used from surveillance data. Thus, this technique has high accuracy, which capture the spatiotemporal distribution of the temperature and wind data, and this allows providing a measure of the uncertainty associated with the prediction, short-term weather predictions and obtaining high-quality. Many approaches are supplied while the cross-validation is supplied to demonstrate that temperature and wind models generated using this technique can accurately capture the spatiotemporal distribution of these weather variables.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments are carried out in a physical cluster environment, the researcher used three PCs each of the computer has the following configuration: Core i7 processor, 4 GB main memory, and 1 TB disk space, and Hadoop cluster on Linux Ubuntu 14.04 where one PC (NameNode) and remaining ran as Datanode. We used dataset from NOAA dataset, which are the huge amount of weather dataset use for forecasting, and it considered as the world's largest active archive of weather data form the North Carolina. NOAA has more than 150 years of dataset on hand with 224 gigabytes of new information added each day. Figure 3 shown the output of the cluster ran, and Figure 4 demonstrate the output of data in DataNode in URL services <http://master:50070/cluster/nodes>.



```
hduser@master:~$ cat /tmp/hadoop-hduser/dfs/dfsutil/dfsutil-20181218022329.txt
16/12/18 02:23:29 INFO mapred.LocalJobRunner: reduce > reduce
16/12/18 02:23:29 INFO mapreduce.Job: map 100% reduce 92%
16/12/18 02:23:30 INFO mapred.Task: Task attempt_local145055005_0001_r_
16/12/18 02:23:30 INFO mapred.LocalJobRunner: reduce > reduce
16/12/18 02:23:30 INFO mapred.Task: Task attempt_local145055005_0001_r_
16/12/18 02:23:30 INFO output.FileOutputCommitter: Saved output of task
/bigdata/weather/output/temporary/0/task_local145055005_0001_r_000000
16/12/18 02:23:30 INFO mapred.LocalJobRunner: reduce > reduce
16/12/18 02:23:30 INFO mapred.Task: Task attempt_local145055005_0001_r_
16/12/18 02:23:30 INFO mapred.LocalJobRunner: Finishing task: attempt_l
16/12/18 02:23:30 INFO mapred.LocalJobRunner: reduce task executor comp
16/12/18 02:23:30 INFO mapreduce.Job: map 100% reduce 100%
16/12/18 02:23:31 INFO mapreduce.Job: Job job_local145055005_0001 compl
16/12/18 02:23:32 INFO mapreduce.Job: Counters: 35
File System Counters
FILE: Number of bytes read=4256403355
FILE: Number of bytes written=267386167441
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2378093302154
HDFS: Number of bytes written=38016
HDFS: Number of read operations=42022
HDFS: Number of large read operations=0
HDFS: Number of write operations=205
Map-Reduce Framework
Map input records=162524713
Map output records=138900408
Map output bytes=1805705304
Map output materialized bytes=2003507332
Input split bytes=20000
Combine input records=0
Combine output records=0
Reduce input groups=3168
Reduce shuffle bytes=2003507332
Reduce input records=138900408
Reduce output records=3168
Spilled Records=277800816
Shuffled Maps =202
Failed Shuffles=0
Merged Map outputs=202
GC time elapsed (ms)=25340
Total committed heap usage (bytes)=104575008700
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=21345593909
File Output Format Counters
Bytes Written=38016
hduser@master:~$
```

Figure3: The output of the MapReduce program

The summary of Hadoop cluster results in Figure 5 and Figure 6 the analysis base on MapReduce algorithm, which distributed the weather dataset in the cluster of three (nodes), as “map”

operation the NameNode takes the input data, partitioned into smaller sub-problems to distributes them to DataNodes. We executed Map and Reduce algorithm for the averaging Temperature, Visibility and Humidity of each year from 2004 to 2007. Each Map task extracts the temperature, Visibility and Humidity from the given year file. The output of the map phase is set of key value pairs. Set of keys are the years. Values are the temperature, Visibility and Humidity of each year. Then each reducer finds the average recorded temperature for each year.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxr-xr-x	hduser	supergroup	40.88 MB	28/12/2016 00:48:11	2	128 MB	200601hourly.txt
-rwxr-xr-x	hduser	supergroup	45.72 MB	28/12/2016 00:48:15	2	128 MB	200602hourly.txt
-rwxr-xr-x	hduser	supergroup	46.95 MB	28/12/2016 00:48:20	2	128 MB	200603hourly.txt
-rwxr-xr-x	hduser	supergroup	44.52 MB	28/12/2016 00:48:24	2	128 MB	200604hourly.txt
-rwxr-xr-x	hduser	supergroup	44.6 MB	28/12/2016 00:48:28	2	128 MB	200605hourly.txt
-rwxr-xr-x	hduser	supergroup	54.33 MB	28/12/2016 00:48:33	2	128 MB	200606hourly.txt
-rwxr-xr-x	hduser	supergroup	54.64 MB	28/12/2016 00:48:38	2	128 MB	200607hourly.txt
-rwxr-xr-x	hduser	supergroup	49.48 MB	28/12/2016 00:48:43	2	128 MB	200608hourly.txt
-rwxr-xr-x	hduser	supergroup	44.35 MB	28/12/2016 00:48:48	2	128 MB	200609hourly.txt
-rwxr-xr-x	hduser	supergroup	51.3 MB	28/12/2016 00:48:53	2	128 MB	200610hourly.txt
-rwxr-xr-x	hduser	supergroup	52.67 MB	28/12/2016 00:48:58	2	128 MB	200611hourly.txt
-rwxr-xr-x	hduser	supergroup	51.93 MB	28/12/2016 00:49:03	2	128 MB	200612hourly.txt
-rwxr-xr-x	hduser	supergroup	46.48 MB	28/12/2016 00:49:07	2	128 MB	200701hourly.txt
-rwxr-xr-x	hduser	supergroup	56.85 MB	28/12/2016 00:49:13	2	128 MB	200702hourly.txt
-rwxr-xr-x	hduser	supergroup	57.37 MB	28/12/2016 00:49:18	2	128 MB	200703hourly.txt
-rwxr-xr-x	hduser	supergroup	62.07 MB	28/12/2016 00:49:24	2	128 MB	200704hourly.txt

Figure4: The output of the MapReduce of website

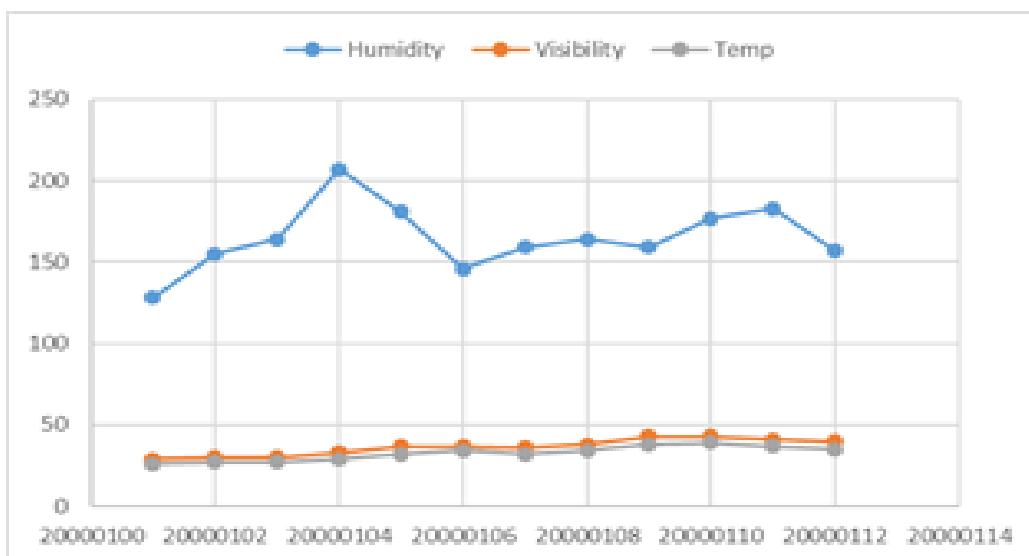


Figure5: The result for daily data

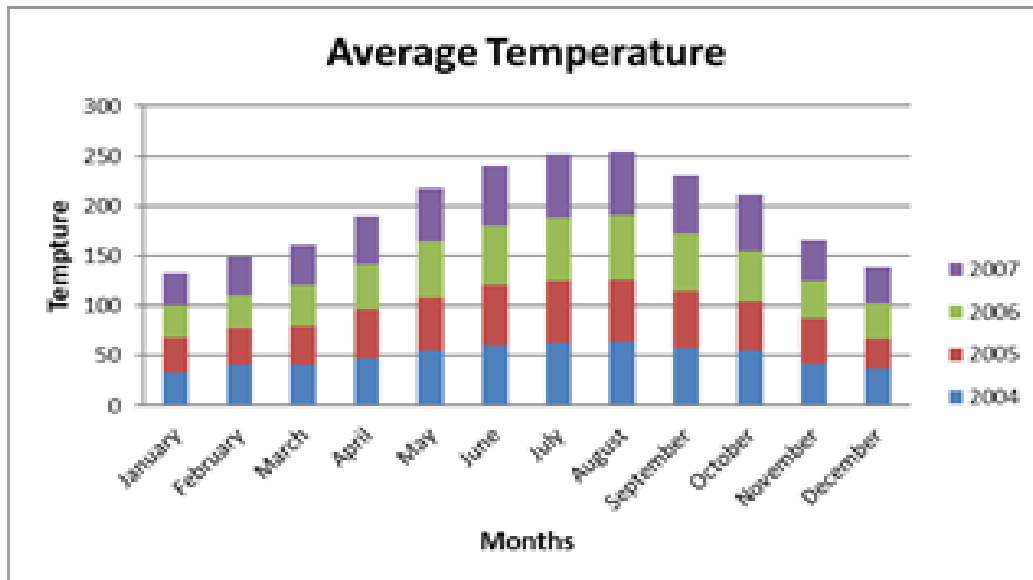


Figure6: The final result of station with Year

VI. CONCLUSION

In traditional analysis tools, when we processing millions of files of data is the time-consuming. Nevertheless, big data technologies use by the department of meteorological with variety of sensing devices to collecting weather data (i.e., temp, humidity). Hadoop/MapReduce used to analyze big data in an effective way, and it divided to two-part HDFS which it distributed the weather data in cluster nodes (i.e., commodity computers). MapReduce it uses to execute data in parallel and distribute Map and Reduce algorithms across the big dataset, thus the big data analyzed efficiently, and scalability bottleneck is removed by using Hadoop/MapReduce. Therefore, use big data technologies for big amount of data analyses has the potential to greatly enhance the weather prediction. This result has revealed the significant impact to the used of MapReduce Algorithm in weather prediction. In addition, the MapReduce results have discovered the significant pattern of temperature, humidity and visibility information, which is valuable for the weather prediction.

REFERENCES

- [1] R. Northcott, "Big data and prediction : Four case studies," no. February, 2019.
- [2] B. Anurag, M. Prakash, V. Kanna, and P. Choudhary, "Weather Forecasting using Map-Reduce," pp. 14945–14952, 2017.
- [3] V. Dagade, M. Lagali, S. Avadhani, and P. Kalekar, "Big Data Weather Analytics Using Hadoop," vol. 14, no. 2, pp. 847–851, 2015.
- [4] M. A. Majid, "Big Data Prediction Framework for Weather Temperature Based on MapReduce Algorithm," 2016 IEEE Conf. Open Syst., pp. 13–17, 2016.
- [5] A. K. Pandey, "A Hadoop based Weather Prediction Model for Classification of Weather Data," pp. 1–5, 2017.
- [6] P. C. Reddy, "Survey on Weather Prediction using Big Data Analytics."
- [7] V. Suryanarayana, B. S. Sathish, A. Ranganayakulu, and P. Ganesan, "Novel Weather Data Analysis Using Hadoop and MapReduce – A Case Study," 2019 5th Inf. Conf. Adv. Comput. Commun. Syst., pp. 204–207, 2019.
- [8] I. Gad, "BIG DATA TECHNIQUES : HADOOP AND MAP," pp. 194–199, 2016.
- [9] M. Adam, I. Fakherldin, K. Adam, N. Akma, and A. Bakar, "Weather Data Analysis Using Hadoop : Applications and Challenges Weather Data Analysis Using Hadoop : Applications and Challenges," 2019.
- [10] N. Yang, "Model with Big Data," pp. 1–7, 2018.
- [11] H. Hassani and E. S. Silva, "Forecasting with Big Data : A Review," Ann. Data Sci., vol. 2, no. February, pp. 5–19, 2015.
- [12] H. Hosni and A. Vulpiani, "Forecasting in Light of Big Data," pp. 557–569, 2018.
- [13] B. Reddy and P. B. A. Patil, "Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique," vol. 5, no. 6, pp. 643–647, 2016.