e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

ANALYSIS OF WEATHER USING HADOOP

Akkshaya Sri J*[1], Mohindran S. R.*[2], Rohinth. S*[3], Sanchez Innocencia D*[4], Dr. M. Marimuthu*[5], Dr. P. Velvadivu*[6]

*[1,2,3,4]III M.Sc (Integrated), Decision and Computing Sciences,
Coimbatore Institute of Technology, Coimbatore, India

*[5,6]Assistant Professor, Department of Computing,
Coimbatore Institute of Technology, Coimbatore, India.

ABSTRACT

The Temperature is the one which is mainly considered for the changing weather conditions. This paper deals with the analysis of Temperature, where we use Big Data features like MapReduce to calculate the Average Temperature with respect to its year. We use HIVE, which is an open-source data warehousing software used for reading large datasets. Using HIVE, we write queries to fetch the data and we can also update it. Further, we use Random Forest Regression Technique, a Machine Learning Algorithm to classify and predict the temperature based on the classification. The dataset for this analysis is taken from the Data.gov website which contained data for about 85 years and we have taken the data for about 30 years from 1990 – 2020. The above analysis is done using Hadoop and Pyspark.

Keywords: MapReduce, HIVE, Random Forest Regression, Hadoop, PySpark

I. INTRODUCTION

Data has been growing since its existence in various domains like share market, social media etc. The data may be of any kind like structured, unstructured and semi-structured. BigData is one which provides a lot of techniques and tools for processing efficiently any kind of data and storing it in tools which helps to store large data. Big data can handle data sets with sizes beyond the ability of commonly used software tools to capture and process the data.

The change of Temperature plays a vital role to predict the Weather, where temperature of a day determines the Weather conditions. Forecasting of Weather is very much useful in many sectors through which they get affected like the Agricultural Sector, Water Resources, Air Traffic and also major causes in the field of Tourism. Forecasting of Temperature will be much useful for the people of these sectors so that they may make a calculation of the Weather Conditions and prepare accordingly. For our analysis, we first use HIVE queries for table creation, retrieval of data and then use the MapReduce programming framework to make the analysis on temperature and finally we build a Machine Learning Model using PYSPARK for prediction.

II. LITERATURE REVIEW

In the paper "Novel Weather Data Analysis Using Hadoop and MapReduce", the author has considered two important Big Data techniques which are Hadoop and Mapreduce to analyze the large dataset. They even developed a system that adopts the historical weather data by using the Big Data Techniques mentioned.

Chouksey P in his paper "A Review of Weather Data Analytics using Big Data" has explained the weather parameters which are to be used for analysis. It suggested that weather data is mostly analyzed using MapReduce. So he stated that comparison study of weather data using MapReduce and Spark is very much needed.

Shiva Naga Jyothi CH along with his colleagues in his paper "Weather Data Analysis Using Hadoop MapReduce and Spark" presents the analysis of weather data by calculating maximum, minimum, average temperatures and top coldest and hottest stations in a particular location or year using various weather parameters.

III. PRELIMINARIES

A. MapReduce

MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment. Then, the reducer aggregates those intermediate data tuples into a smaller set of tuples or key-value pairs which is the final output. Here using Mapreduce, we find the average temperature

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

across different years. The Mapper takes in the Weather input file as an Input and emits the Year and Temperature. The Reducer emits the data for Average Temperature.

B. Apache Hive

Now-a-days, the use of Hive Metastore is one of the big constants in big data implementation. Hive ships with the metastore service . This service allows you to manage your metastore as any other database. You can host this service on any of the popular RDBMS (e.g. MySQL, PostgreSQL etc.). This allows you to query your metastore with simple SQL queries, along with provisions of backup and disaster recovery. We have installed hive with mysql as a metastore and loaded the data. Here, we have used HIVE queries to Load and Create tables, to find the Maximum and Minimum values with respect to its years and also to find the second maximum temperature with respect to its year.

IV. MACHINE LEARNING ALGORITHM

A. Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.Since, there are high chances of having missing values in the dataset, unwanted attributes etc. We have cleaned it up using preprocessing techniques. After taking the dataset, we have dropped two variables of our dataset which won't affect the further analysis. We have also checked if there are any missing values in the dataset.

B. PySpark

Today, PySpark is being used by the majority of Data Scientists because of the library set it contains. PySpark is one which links the Python API to the spark core. We are implementing the machine learning algorithm using pyspark. For our dataset, we have used Random forest Regression to classify the Data and then predict the temperature.

C. Random Forest Regression

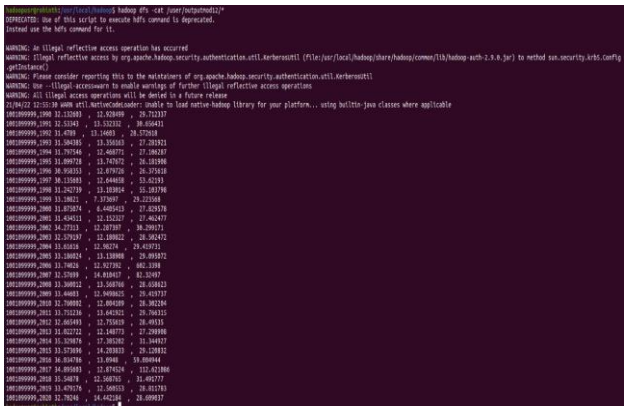
The Random forest is also sometimes called a random decision forest. It is an ensemble learning technique used for solving supervised learning tasks such as classification and regression. An advantageous feature of RF is that it can overcome the overfitting problem across its training dataset. Also an additional advantage is that it does not need any preprocessing initially and accepts even the missing values and predicts the output. Here, using Random Forest, we have first classified and calculated the accuracy score.

V. RESULTS AND DISCUSSION

A. MapReduce Program

Using MapReduce, we have taken the full dataset and then summed up the dataset year wise for the Temperature, Dew Point and the wind Speed taking into consideration the Month. And further we have calculated the Average of Temperature, Dew Point and the Wind Speed with respect to its years from 1990 - 2020 along with its station ID.

The Output is shown below.

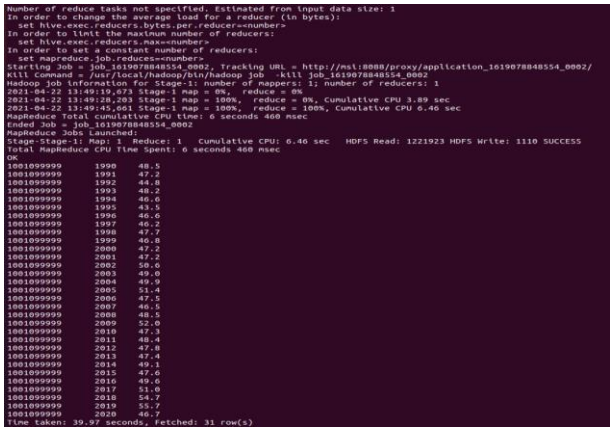


e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

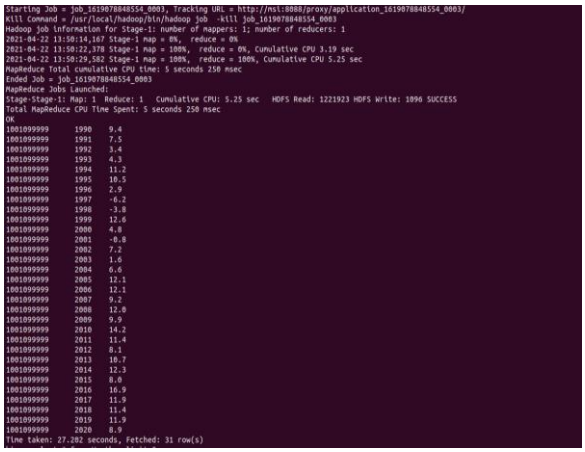
B. Apache Hive

Using HIVE, we have loaded and stored the data in the database and then created the table using the HIVE query and also created 3 more queries in HIVE for retrieving the data we want. Through which we find the Year, where we have the Maximum and Minimum Temperature.

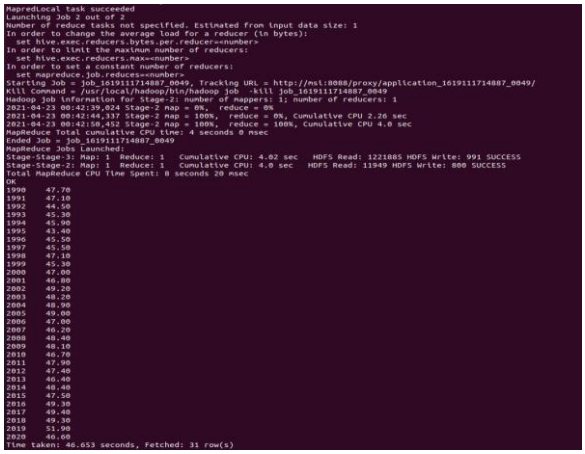
This is the output of the Maximum Temperature with respect to its years using HIVE SQL.



This is the output of the Minimum Temperature with respect to its years using HIVE SQL



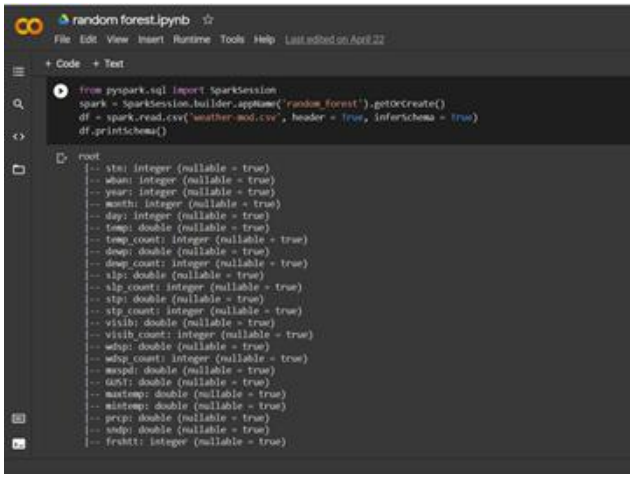
This is the output of the 2nd Maximum Temperature with respect to its years using HIVE SQL

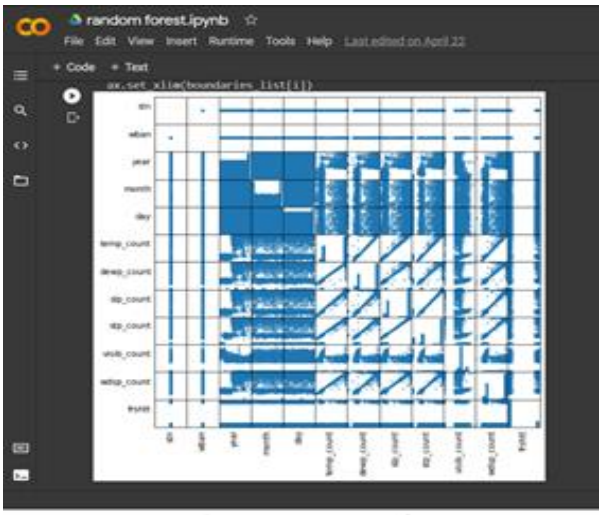


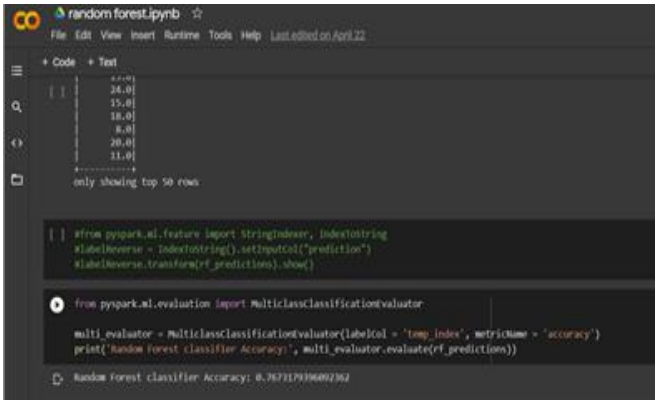
C. Machine Learning Algorithm

We have used the Random Forest Machine Learning Algorithm using the PySpark Environment by storing the data in PySpark. Here, we have found that the Accuracy of the Temperature is 0.77 and hence, we say that this is the best fit Model for the Analysis. The data has been read through PySpark and used for further analysis of Model fitting and prediction.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com







VI. CONCLUSION

Large amount of Weather Data has been collected every day which is being generated by the satellites, storing and processing of these large dataset seems to be very challenging. Hence, for storing these large data, we used an HIVE database and it's easy for retrieving the Maximum and Minimum Temperature and further more processes can also be done. For the Analysis, we have taken the main attributes as the Temperature, Dew Point and the Wind Speed, which is an essential attribute for Weather Forecasting. We used MapReduce and found the Average Temperature with respect to its Year. From the Analysis we have made, we conclude that for the Analysis of Weather, Temperature is one which is very important and also the Wind Speed and Dew Point are an important factor for the prediction of weather. And also we found that Random Forest Regression is one which is an appropriate Machine Learning Model to predict the Temperature.

VII. REFERENCES

[1] https://www.smartdatacollective.com/four-considerations-when-choosing-hadoop-distribution/
[2] https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/
[3] https://www.xajzkjdx.cn/gallery/311-april2020.pdf
[4] https://ieeexplore.ieee.org/document/8728444
[5] http://www.bookmetrix.com/detail/chapter/e5f8dfce-be6e-4d9b-a642-a2dc16d9bf5b
[6] ["Novel Weather Data Analysis Using Hadoop and MapReduce" – A Case Study, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).
[7] "Big Data Weather Analytics Using Hadoop" an International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume14 Issue 2.
[8] 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.