Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

Big Data-Driven AI Models for Predictive Meteorological Analytics

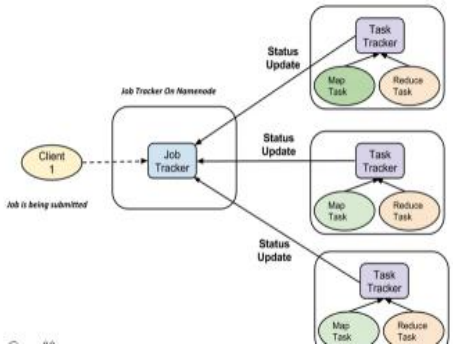
Ms. G. Vijayalakshmi.,
Assistant Professor, Department of Computer Science and Engineering ,
Sri Bharathi Engineering college for women ,Pudukkottai.
gvlyadav28@gmail.com

Abstract- Big data refers to a vast amount of data that requires new technology in order to potentially extract value from it through methods of analysis and capture Over 116 weather stations and over 1000 observation centers have contributed to the NCDC's (National Climatic Data Center) big data. Their data is unstructured, making it difficult to interpret. The massive volumes of data that have been deposited onto the Hadoop Distributed File System, Apache Pig, and Apache Hive are processed in this study using mappers and reducers. The aforementioned dataset has been explained using the methods provided, and the project's final product is the maximum, minimum, and average temperature for the specified time and date.

Keywords- Keywords: Big Data, Hadoop, HDFS, MapReduce, Mapper, Reducer, Min, Max, Average, NCDC

I.INTRODUCTION

The process of examining enormous data sets that include a variety of data kinds is known as "big data" [1]. Large volumes of data are processed and maintained via big data. Conventional data analysis is capable of handling structured data, but not unstructured data. It can handle both structured and unstructured data in large data. Big data refers to data sets that are usually too large for commonly used software tools to handle, collect, process, and curate. The size of big data can range from gigabytes to many petabytes. Weather forecasting is the use of technology to forecast how the environment will behave for a specific region. Weather prediction is one of the most fascinating and exciting fields, and it is crucial for farmers, disaster relief, business agriculturists, etc. It also plays a big part in aerograph. There are many factors that affect how well weather forecasting is implemented, such as data mining techniques, which are unable to accurately predict weather in the near future. The graph is constructed to show the minimum and maximum temperatures for each specific year. Weather data for the upcoming year is forecast using data from the prior year.



II. PROGRAMMING MAPREDUCE

The MapReduce application operates in three stages: reduce, shuffle, and map. **A map stage:** Processing the input data is the responsibility of the map or mapper. The input data is gathered in the Hadoop file system (HDFS) and takes the form of a file or directory [4] [5]. The Reduce job aggregates those data tuples (key-value pairs) into a smaller collection of tuples using the Map's result as input.

Date	Temp (°C)	Humidity (%)	Wind Speed (km/h)	Pressure (hPa)	Rainfall (mm)	Weather Condition
2024-01-01	30	65	12	1012	0	Sunny
2024-01-02	29	70	15	1010	2	Cloudy
2024-01-03	27	80	20	1008	10	Rainy
2024-01-04	31	60	10	1015	0	Sunny
2024-01-05	28	75	18	1009	5	Rainy

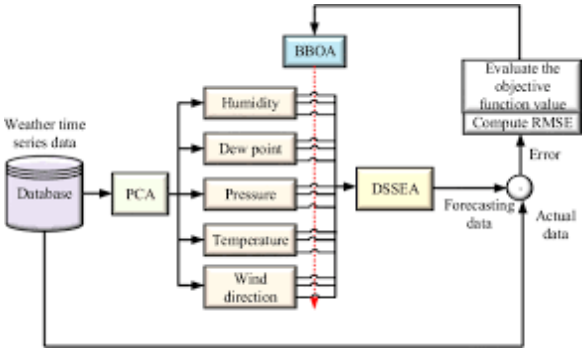
TABLE-1 Dataset Description

III.THE DATASET DESCRIPTION

For the benefit of programmers and other people who need to refer to them, a dataset is a collection of pictures of the things or data objects in a data model.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

The data dictionary utilized in this weather prediction is shown in table.

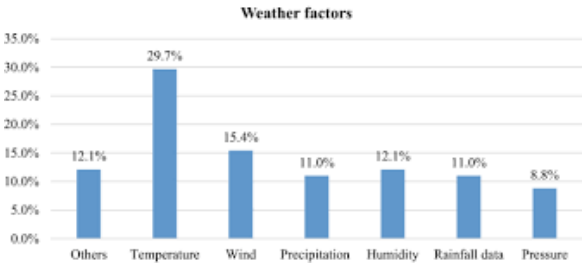


IV.PROPOSED METHODOLOGY FOR WEATHER FORECASTING BY USING BIG DATA ANALYTICS

The forecast of the climate variance perpetually has shown very usefully and essential. In the United States of America (USA) there are typically many effects designed in different cities. These issues might involve the concerts, car racing, festivals, etc. As these are the open-air concerts, they experience a lot from the daily weather variations, which is rising because of global warming. To avoid these issues, they need to pre-plan and choose the data for their event in advance. It can work out only if they had any predictions of the climate data using the Hadoop and distributed system and map reduce.

Calculate the highest and lowest temperatures on hot and cold days using map reduce. As a result, we may find important details about event preparation, including time, place, and statistical information.

Maximum, Minimum, and



Average: This stage determines the year's highest, lowest, and average temperatures in order to forecast future weather. Lastly, to display the temperature, plot the graph for the acquired MAX, MIN, and AVG temperature for each month of the specificyear.

Comparisons: The one to three years out accuracy is used to get the total accuracy percentage. percentages for text forecast precipitation, icon forecast precipitation, high temperature, and low temperature. The percentage of estimates that fall within three degrees is known as temperature accuracy. The percentage of accurate forecasts is known as precipitation accuracy. In the evening, the forecasts were gathered.

Seasons: This stage of seasonal forecasting aims to provide useful data regarding the "weather climate" that has been necessary in the following months. A weather forecast is not what the periodic forecast is. While climate has been defined as a statistical record of the weather events occurring within a designated season, weather can be analyzed as a picture of constantly changing atmospheric conditions.

Prediction: The long-term climate change projection has shown to be extremely significant and helpful. It can only function if it has estimations of the climatic data using distributed systems, Hadoop, and map reduction. Utilizing a map, determine the highest and lowest temperatures for hot and cold days. As a result, we may get important details about event preparation, including time, place, and statistical statistics.

Weather Reports: The list of places on the weather reports is displayed by this module. It has been depicted as a picture that may be used to determine the temperature in the previous and current years.

Weather Format: The list of places' weather forecast details is displayed by this module. It will be forecasted using the previous year's minimum, maximum, and average temperatures. By entering the place name in this module, a user can search for a certain location's weather forecast.

Reports: Based on user needs, this program generates reports such as the total number of required (Min, Max, and Average Temperature) weather forecast reports.

Proposed Algorithm for Weather Forecasting using MapReduce Programming

Input: Cleans Dataset for particular region/City, Prediction Dates, Prediction Attribute

Output: Prediction for a specific range and specified attribute.

Step1: Select all data from noisy data source, and verify each. While(i ≠ Ø) If(verified(i)) Then weight(i) = 1 Else Weight(i) = 0 End While Step1 Traverses the entire database and verifies the validity of each parameter if the parameter value is found noisy, zero weight has given to that record, and that record will not participate in the prediction process.

Step2: PRED_DATE = sequence to be predicted BASE_SEQ = (PRED_DATE) – (NO_OF_DAYS) The algorithm divides the whole data into equal chunks called sequences where every sequence is equal to the prediction time span, i.e., if the prediction is for 1 Month, the 12-year dataset has divided into monthly chunks. It has expected for the distance calculation in the dataset.

Step 3: While days ≠ Ø Selected_days[ ] =

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

DAY(day) of MONTH(month) (if Validated) End While Calculate Distance(Selected_days[]) SORT(Selected_days[], Distance) This step performs the key operation of the algorithm. It selects the similar record from the whole dataset, i.e., if we need to predict the weather for the 1st week of Jan 2003, then this step will select all records of the 1st week of January from the whole dataset. Further it calculates its distance, and finally, it sorts the results according to distance.

Step4: Find the K nearest neighbor and calculate mean. The last step extracts K nearest neighbors from the array and takes its mean as the predicted value for a specific day.
Step5: The process stopped when all data has examined.

V. RESULT AND DISCUSSION

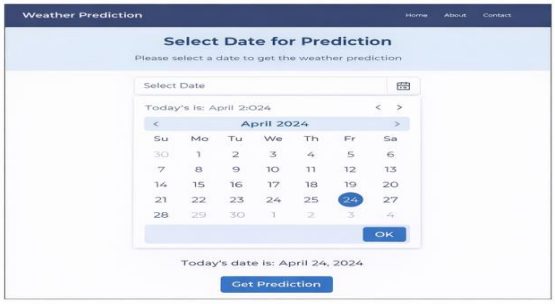
The performance study of fuzzy C-means and k-nearest neighbor in weather forecasting is shown in the following figure. Table 2 above shows that K-NN has a shorter execution time than Fuzzy C-Means. While the accuracy in Fuzzy C-means is only 57.14%, it is raised by 92.86% in KNN.

Parameters	Fuzzy C-Means Algorithm	K-Nearest Neighbor Algorithm
Accuracy	57.14%	92.86%
Execution Time	30 seconds	11 seconds

TABLE-2 Comparison of Performance Analysis of Fuzzy C-Means and K-Nearest Neighbor in Weather Forecasting

Figure 4a: Screenshot of the Home Page of the Weather Prediction

Figure 4b: Screenshot of the Select date for the weather prediction



VI.CONCLUSION

The suggested approach has examined previous data and sophisticated weather forecasting in a big data setting. An effective way to examine sensor data housed in the National Climatic Data Center (NCDC) is to use Hadoop with map reductions.

A framework for distributed, highly parallel systems spanning big datasets is called Map Reduce. The scalability bottleneck can be eliminated by utilizing map reduce with Hadoop. The weather forecast could be much improved by using this kind of technology to evaluate big databases. By giving random access to Big Data, the query tools greatly improve the ease of analytics. MapReduce is a framework that uses several computers to execute distributable algorithms across massive datasets. The weather data can be effectively evaluated using MapReduce with Hadoop, and the NCDC data may be used to anticipate future weather, minimum and maximum temperatures, hot and cold days, and more. It is beneficial for people to plan ahead for outdoor events according to the weather.

VII.REFERENCES

[1]. N.Padmaja, Prof. T.Sudha, "Big Data Analytics With Long Range Plan To Process Large Data Sets," International Journal of Advanced Scientific Technologies, Engineering and Management Sciences, pp.87-90.
[2]. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, "A Review Paper on Big Data and Hadoop," International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014, pp.1-7.
[3]. National Climatic Data Center Data Documentation for Data Set 3260 (DSI-3260). ftp://ftp.ncdc.noaa.gov/pub/data/noaa/dsi3260.pdf.
[4]. Pooja S.Honnutagi, "The Hadoop distributed file system," International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014, 6238-6243.
[5]. Jimmy Lin and Chris Dyer, "Data-Intensive Text Processing with MapReduce," This is the preproduction manuscript of a book in the Morgan & Claypool Synthesis Lectures on Human Language Technologies. Anticipated publication date is mid2010.