# A Novel Approach for Analysing the Weather by using Big Data Tool

**Sheeba Ann Thomas[1] Anju Rachel Oommen[2] Smita C Thomas[3]**
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Mount Zion College of Engineering Kadammanitta, Pathanamthitta ,India

*Abstract*— In the last few decades, the generation of data has increased and it is expected to increase in future. So that it is necessary to process a large amount of data set in weather and analyse the same using the traditional methods. In the existing system, it aims to forecast the chances of rainfall by using predictive analysis in Hadoop. It helps to predict the rainfall in minimum and maximum by taking the data as input. It uses the apache PIG. It have some disadvantages, so in proposed system we use a Naïve Bayse algorithm to predict earth quake, floods , etc., It focuses on meteorological data to predict the seasons to separate the weather data based on Longitudinal and Latitudinal which can be used to analyse the reliability factor of cyclone, earthquake, rainfall, temperature and humidity. It provides specific service to an assessment of pollution impacts from different organization and thermal power plants. It is easy and fast to predict class of test data set. It also performs well in multi class prediction.

***Key words:*** Weather, Hadoop, Big data, Naïve Bayse, Cyclone

## I. INTRODUCTION

Rainfall prediction modelling involves a combination of computer simulations, and findings the trends and patterns. The weather data used for the research include daily temperature, daily pressure and monthly rainfall. Data mining for meteorological Data and applied knowledge discovery process is used to extract knowledge from weather dataset. To predict the daily temperature value accurately a hybrid data mining technique is used. Using Naïve Bayes classification Technique, can be predict the level of meteorological weather condition of region in seasonally which based on cyclone form and level of humidity. The Big Data maintains the huge amount of data and processes them efficiently. Big data includes data sets with sizes beyond the ability of commonly used software tools to capture, manage and process the data. We will be using Map reduce and Pig commands in order to analyze the data sets and to perform various operations on the data set. Based on the previous year's historical weather data set we are able to predict the future weather. Weather prediction is the application of technology to predict the weather for a given location based on historical data or current data as applicable. Climate change has been seeking a lot of attention since a long time due to the unexpected changes that occur. There are several limitations in better implementation of weather forecasting as a result it becomes difficult to predict weather short term with efficiency [1]. The prediction of climate has always proven to be very important and useful. Big data collects large volume of data and it is a great challenge for Hadoop, a part of Big Data, which uses Map Reduce and Pig to maintain and process the data and helps to extract useful information in an efficient manner [2]. The Big Data maintains the huge amount of data and processes them efficiently. Big data includes data sets with sizes beyond the ability of commonly used software tools to capture, manage and process the data. We will be using Map reduce and Pig commands in order to analyze the data sets and to perform various operations on the data set. Based on the previous year's historical weather data set we are able to predict the future weather

## II. LITERATURE REVIEW

This chapter investigates some researches in the prediction domain we have done. It also has detail study of each paper in the same field. It covers six papers of prediction analyses. In [4] the author describes design of patient customized healthcare system. It consists of 4 modules. Medical Data Collection Module – It stores big data of patient's health and medical information in the Hbase. Text Mining Hadoop Module – It analyses the collected unstructured data into structured data like patient's information, family history and stores the structured data into Hbase with a map-reduce framework. Disease Rule Creation Module – It generates disease rules by using disease information stored in Hbase. Disease Management Prediction Module – This module informs the risk index or result of disease prediction. In [5] the author describes that storm can be predicted using the previous year's data set. It contains huge number of records therefore can be used as a research idea. This paper defines the solution to predict using Map Reduce Framework. The data is classified using Support Vector Machine (SVM). Using this it can predict maximum Rain Storm. Map Reduce Framework is use for the Rain Storm Prediction. In [6] author describes that predicting daily behaviour of stock market is a serious issue for stock holders. Nowadays the stock market has been called for research in many fields due to its effects on financial challenging. By using linear regression we predict S&P 500 index behaviour and at the end we compared and evaluated the result of our proposed method with other approaches. Our System has good performance in terms of huge volume of data and the stock holders can invest more with confidence. By using integrated collective data it can determine market policies and their orientation which finally lead to increases in productivity and income. In [7] author describes that current video streaming algorithms use various estimation approaches to infer the variable bandwidth in cellular networks. This variable bandwidth sometimes leads to reduced quality of experience. There is no accurate bandwidth present due to which achieving reliable video streaming over cellular networks has proven to be difficult. Nowadays most content providers use adaptive bitrate (ABR) streaming. Existing algorithms fail to fully utilize available band-width. Here we are using PBA (Prediction Based Adaptation) algorithm that combines short term predictions. Using PBA we achieve nearly 96% of optimal quality and it also improves the quality of experience by accurate prediction.

 

## III. EXISTING SYSTEM

Weather prediction is the application of technology to predict the weather for a given location based on historical data or current data as applicable. Climate change has been seeking a lot of attention since a long time due to the unexpected changes that occur. There are several limitations in better implementation of weather forecasting as a result it becomes difficult to predict weather short term with efficiency [1]. The prediction of climate has always proven to be very important and useful. Big data collects large volume of data and it is a great challenge for Hadoop, a part of Big Data, which uses Map Reduce and Pig to maintain and process the data and helps to extract useful information in an efficient manner [2]. The Big Data maintains the huge amount of data and processes them efficiently. Big data includes data sets with sizes beyond the ability of commonly used software tools to capture, manage and process the data. We will be using Map reduce and Pig commands in order to analyze the data sets and to perform various operations on the data set. Based on the previous year's historical weather data set we are able to predict the future weather.

Big Data is that data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast and is unstructured and doesn't fit the structures of the architectures. To gain value from this data we need an alternative way to process it. Various fields for example that generate such large amounts of huge data are Facebook, Twitter ,Weather stations ,New York Stock Exchange , Worldwide electric transmissions etc. Thus in our project we are dealing with huge amount of unstructured weather data. Our paper focuses on the shifting of processes from single node data processing to Hadoop distributed file system for faster processing and the best technique to process the queries.

Weather forecasting is always a big challenge for the meteorologists to predict the state of the atmosphere at some future time and the weather conditions that may be expected. It is obvious that knowing the future of the weather can be important for individuals and organizations. Accurate weather forecasts can tell a farmer the best time to plant, an airport control tower what information to send to planes that are landing and taking off, and residents of a coastal region when a hurricane might strike

Humans have been looking for ways to forecast the weather for centuries. Scientifically-based weather forecasting was not possible until meteorologists were able to collect data about current weather conditions from a relatively widespread system of observing stations and organize that data in a timely fashion. Vilhelm and Jacob Bjerknes developed a weather station network in the 1920s that allowed for the collection of regional weather data. The weather data collected by the network could be transmitted nearly instantaneously by use of the telegraph, invented in the 1830s by Samuel F. B. Morse. The age of scientific forecasting, also referred to as synoptic forecasting, was under way. In the United States, weather forecasting is the responsibility of the National Weather Service (NWS). The future modernized structure of the NWS will include 116 weather forecast offices (WFO) and 13 river forecast centers, all collocated with WFOs. Thus Global weather data are collected at more than 1,000 observation points around the world and then sent to central stations maintained by the World Meteorological Organization, a division of the United Nations. Thus there is a need for a flexible platform for the maintenance of this Big Data and help Weather forecasting using that Big Data. Thus Apache open source Hadoop and Spark are the solutions for it that provides high speed clustered processing for the analysis of large set of data smoothly and efficiently.

Hadoop & Map Reduce is the most widely used models used today for Big Data processing. Hadoop is an open source large-scale data processing framework that supports distributed processing of large chunks of data using simple programming models. The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in addition to other modules. The software is modelled to harvest upon the processing power of clustered computing while managing failures at node level.

Apache Spark is the new competitor in the Big Data field. Spark design is not tied to MapReduce ,it has proved to be 100 times faster than Hadoop MapReduce in certain cases. Spark supports in-memory computing and performs much better on iterative algorithms, where the same code is executed multiple times and the output of one iteration is the input of the next one.

Apache Pig and Hive are two projects which are layered on top of Hadoop, and provide higher-level language to use Hadoop's MapReduce library. Pig provides the scripting language to describe operations like the reading, filtering and transforming, joining, and writing data which are exactly the same operations that MapReduce was originally designed for. And Hive offers even more specific and higher-level language, to query data by running Hadoop jobs, instead of directly scripting step-by-step all operation of several

Map Reduce jobs on Hadoop. The language is, very much SQL-like, by design. Apache Hive is still intended as a tool for long-running batch-oriented queries over a massive data and it's not "real-time" in any sense.

## IV. PROPOSED SYSTEM

In the next version of our project we will use Apache Hadoop Framework and Map Reduce Framework and predict the rain using Naïve Bayes Algorithm. Naïve Bayes Algorithm is a classification technique based on Bayes Theorem. Naïve Bayes is easy to build and very much useful for large datasets. By using the Naïve Bayes equation we can find the future probability [12]. A system Architecture defines the behaviour, Structure and views of the system. An architecture description is a formal description and representation of a system; it supports structures and behaviour of the system. A system Architecture can develop system components, the expand systems developed, that will work together to implement the overall system.

### *A. Weather Data*

This Module contains Weather data which will be used for predicting the Rain. It contains various parameters that mean various columns. Data set of our Project is shown below:

 

### *B. Hadoop*

Hadoop is open source software and it is used to storing large data set in a distributed computing environment, Hadoop makes it possible to run applications on system with hundreds of hardware nodes. Hadoop supports range of related projects that can extend Hadoop performance. Complimentary software project includes Apache Pig, Apache Hive and Apache Spark etc. Apache Pig is a high level platform for creating programs that runs on Hadoop. Hadoop Distributed file system provides rapid data transfer rates among nodes and in case of node failure it allows the system to continue operating

### *C. HDFS (Hadoop Distributed file System)*

The Hadoop Distributed File System (HDFS) is similar to the Google File System (GFS) and it uses large cluster of data and it provides distributed file system, fault-tolerant manner. HDFS follows two architecture which is master and slave. The master node includes a single Name Node that handles the metadata

### *D. Data Gathering*

Data gathering is the process of measuring information on targeted variables in an established systematic fashion which then enable one to evaluate outcome. The goal of gathering data is to capture the data that then translated to structuring data. For predicting the weather the data is collected from the National Climate Data Centre. The gathered weather data includes any facts or numbers about the state of the atmosphere, including temperature, wind speed, rain or snow, humidity and pressure. There are some amazing ways to collect this kind of data. We have high-tech equipment that can measure everything with amazing accuracy and we can measure it from all sorts of places: the ground, the air and even more space. Some of the equipment we use to take these measurements includes thermometers, radar systems and even satellites.

### *E. Data Structuring*

Structuring of data is the process of organizing the data based on certain parameters. The parameters are like longitude, latitude, temperature- max, min, avg, mean, solar radiation, surface temperature-min, max, avg, soil moisture and soil temperature. It provides a means to manage large amount of data efficiently and accurately. General data structure types include the array, the file, the record, the table, the table, the tree, and so on. Any data structure is designed to organize data to suit a specific purpose so that it can be accessed and worked with in appropriate ways. Data structure for large datasets, which stores abstract of data, which is less in magnitude and capable of satisfying most of the queries of user as well as original data. The data structure had been explained under four headings, subdivision scheme, location codes, transformation function and sub division criteria algorithms for ATree. Two types of build algorithms are used with ATree. The role of build algorithm, which accepts raw data in standard formats. This builds the ATree according to the parameters which are shape, location code scheme, transformation function, subdivision criteria and metadata function. The ATree is build using two different approaches: Top down approach and Bottom up approach.

## V. CONCLUSION

Thus we have successfully found of the chances of rain from given dataset using Apache PIG. This was the first version of our project, in next version we will use Naïve Bayes algorithm in Hadoop Framework Apache PIG has some Disadvantages will be overcome in next version of this project. The prediction of earthquake, flood can also be done using Naïve Bayes Algorithm this is the future scope of our project.

## ACKNOWLEDGEMENT

We would like to thank, first and foremost, Almighty God, without his support this work would not have been possible. We would also like to thank all the faculty members of Mount Zion College of engineering, for their immense support.

## REFERENCES

[1] Gautam and P. Bedi, "MR-VSM: Map Reduce based vector Space Model for user profiling-an empirical study on News data," 2015 International Conference on Advances in Computing Communications and Informatics (ICACCI), Kochi, 2015, pp. 355-360.
[2] Anjali Gautam, Tulika , Radhika Dhingra, and Punam Bedi, "Use of NoSQL Database for Handling Semi Structured Data: An Empirical Study of News RSS Feeds," in Emerging Research in Computing, Information, Communication and Applications, 2015, in press.
[3] Viktor Mayer-Schoenberger & Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think.
[4] D Byung,Kwan Lee, EunHee Jeong, , " A Design of a Patientcustomized Healthcare System based on the Hadoop with Text Mining (PHSHT) for an efficient Disease Management and Prediction", Vol.8, No.8 (2014), pp. 131-150, "International Journal of Software Engineering and Its Applications",ISSN:1738-9984 IJSEIA.

 