WEATHER PREDICTION ANALYTICS USING MAPREDUCE-BASED LOGISTIC REGRESSION

SU HLAING MON THAN

M.C.Sc. **SEPTEMBER 2022**

WEATHER PREDICTION ANALYTICS USING MAPREDUCE-BASED LOGISTIC REGRESSION

By

SU HLAING MON THAN

B.C.Sc

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

September 2022

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to the following persons who supported and helped towards the success of the thesis.

Above all else, I would like to express my appreciation and sincere thanks to **Dr. Mie Mie Khin**, Rector, University of Computer Studies, Yangon, for giving the opportunity to develop this thesis.

I greatly appreciate and acknowledge to **Dr. Tun Myat Aung,** Principal and Pro-rector of the University of Computer Studies, Hinthada, for his kind permission and administrative support.

I would like to thank course coordinators, **Dr. Si Si Mar Win**, **Dr. Tin Zar Thaw and** Professors, Faculty of Computer Science, University of Computer Studies, Yangon, for their guidance, management and encouragement in the progress of the thesis.

I would like to express my deepest gratitude to my supervisor, **Dr. Hmway Hmway Tar**, Professor, Faculty of Computer Science, University of Information Technology, for her patient, supervision, tenderness, encouragement in the progress of the thesis. I will always remember her for being a mentor to me.

I am thankful to my thesis supervisors **Dr. May Phyo Thu** for her close supervision, proper guidance, valuable suggestions, guidance, advice and encouragement during the course of this work.

I am thankful to teacher **Daw Win Lai Lai Bo**, Lecturer, the Department of English, University of Computer Studies, Yangon, for modification of my thesis in checking grammatical errors, choice of words from the language point of view.

My sincere gratitude also goes to all my respectful teachers for teaching valuable lectures, guiding suggestions and sharing knowledge during the master course work and thesis work.

Finally yet importantly, I am deeply grateful to my beloved parents and my family for their mental and financial support, continuous encouragement, care and kindness, and endless love over the years.

I especially thank to my friends and colleagues for their true friendship, encouragement, support and assistance throughout my studies.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

---------- ----------

Date Su Hlaing Mon Than

ABSTRACT

In the modern world, the prediction of weather becomes a challenging task. The weather prediction system is very useful and important for our agriculture. Weather forecasting is important for investigating of many business and decreases crop damage. Agriculture is the vital role for our country's business and most of people are depend on developing activities. Regression is one of the main methods used in weather data prediction. Multinomial logistic regression is the important role of the system and forecasting of weather data based on temperature, humidity and wind. This system is applied to predict the weather data for a given location with multinomial logistic regression in order to obtain the desired prediction. The system of weather data prediction for a given location is Hinthada Region. It applies multinomial logistic regression and Map Reduce platform. In the system uses multinomial logistic regression to calculate the model. The various formats of weather datasets store in Hadoop Distributed File System and to obtain the optimum result the MapReduce Algorithm is used. This system predicts weather forecasting of Hinthada Region. Weather forecasting is one of the most important task for farmers in Hinthada Region and they decide their agriculture. Weather prediction helps agriculturist for decision making for their crop. The system helps farmers to use effective approach for weather prediction. Hinthada Region's economy is highly dependent on its agricultural products. The system helps the agriculturist to get the awareness of their business and income.

CONENTS

LIST OF FIGURES

LIST OF TABLES

Table **Page**

CHAPTER 1

INTRODUCTION

Nowadays, a massive amount of data, according to the development of technology in Internet of things, devices and social media, every bank transaction of data are big data. The huge amount of have a large number of features, attributes, characteristics, and many behaviors. These data are structured, unstructured, and semi-structured. The structured data is quantitative data, consists of numerical and text data. It is easy to analyze and process structured data, which is generally stored in a relational database and can be queried using structure query language (SQL). The unstructured data is qualitative data that lacks any predefined structure and can come in a variety of formats (images, mp3 files, WAV files, etc..), that is stored in a non-relational database and come be queried using NOSQL. Thus, big data means the datasets that cannot be recognized, obtained, managed, analyzed and processed by present tools.

Big data is high-volume, velocity, variety, variability, value, visualization, and validity.

- Volume: the amount of data to justify whether it should generated.
- Velocity: the data generation speed, that it cannot processed or analyzed using conventional data processing techniques.
- Variety: the data means different data source that provides information on that data sources.
- Veracity: the volumes of data comes for processing are valuable.
- Variability: the enormous amount of data are heterogeneity.
- Value: the data are great value but very low similarity.
- Visualization: the big data must be visualized with appropriate tools that perform different parameters to analyst.
- Validity: the correct processing of data should accurate results.

1.1 Big Data Technology

Big Data can be stored in numerous ways; it is often stored in a data lake. The data warehouse commonly built on relational databases. Private public and hybrid storing are the most common and effective ways to store data. Big data storage is compute and storage architecture, to collect and mange massive datasets and perform real-time data analyses. Big data storage are used in similar ways as traditional relational database management systems, online transactional processing (OLTP) solutions and data warehouse. The large amount of data are various types of storage, Massive Parallel Processing (MPP), New SQL database, Big Data Querying platforms, and distributed file systems. Massive Parallel Processing (MPPP) is the coordinated processing of a program by multiple processors working on different parts of the program. Each processor has its own operating system and memory. MPP speeds the performance of huge database that deal with massive amounts of data that support for high query performance and platform stability. NewSQL is the latest technology in the big data, class of relational database management system that seek to provide the scalability and availability of NoSQL with the consistency and usability of SQL. NoSQL database provides a mechanism for storage and retrieval of data. This data is modeled in means other than the tabular relations used in relational databases that supporting database consistency, scalability and availability.

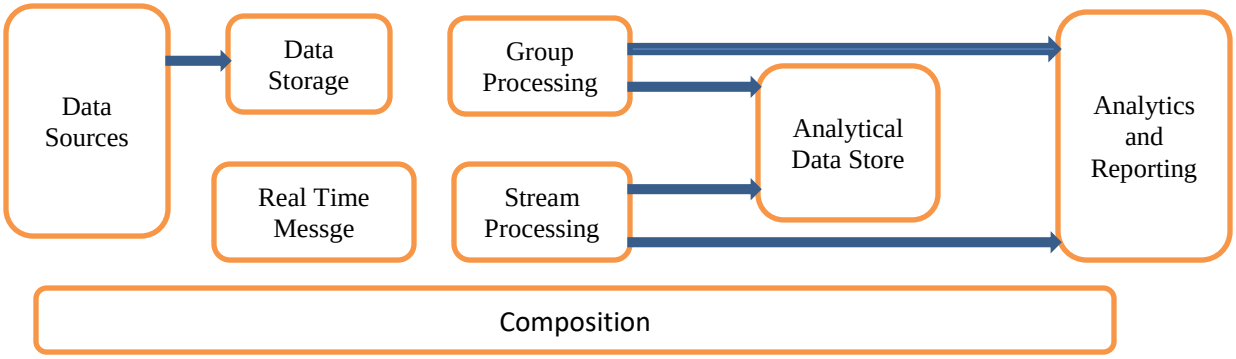


Figure 1.1. Big Data Architecture

The most famous of big data querying platforms are Hadoop, data warehouse and cloud storage. Hadoop is a software framework for distributed storage and processing of big data to handle large amounts of data and computation. Datawarehouse tools make it possible to manage data more efficiently as it enables to find access, visualize, and analyse

data to make achieve more desire results. The other method of storing massive amounts of data is cloud storage, that store data and information are stored electronically online where it can access from anywhere, negating the need for direct attachement of access to a hard drive or computer.

Hadoop distribute file system (HDFS) is an open source framework that is used to efficiently store and process large datasets. Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly. Hadoop is a distributed file system that runs on large clusters and provides high-throughput access the data. An open source framework, Hadoop was implement to support in performing big data analytics that provides scalability, reliability, and manageability, providing the implemented Map Reduce paradigm. Hadoop have two components: Hadoop distributed files system for the big data storage, and Map Reduce for big data analytics.

Hadoop distribute file system (HDFS) for administration of extensive data index sizes of gigabytes and petabytes. Hadoop allows the data in clusters of different commodity volumes of data on node systems. Map Reduce gives expository capacities to analyze enormous volumes of complex data. Hadoop cluster consists of one or several Master Nodes and many more Slave Nodes. Hadoop and Map Reduce form a flexible foundation that can linearly scale out by adding nodes. The processing of Hadoop that analyze and process datasets coming into the cluster. The structured and unstructured datasets are mapped, shuffled, sorted, merged and reduced into smaller manageable data blocks.

Map Reduce programs is composed of a map and reduce which is suitable for parallel processing of massive data stored in Hadoop. Map Reduce separates a task into smaller parts and assigns the input data to different systems called nodes. Map Reduce framework is parallel processing for problems across large datasets of a large numbers of nodes and collectively referred to as a cluster or map. The functions of Map Reduce works maps input values to output as a combination of (key, value) pairs. The output of each mapper is a set of pairs (key, value). A (key, value) consists of two related data elements, key is a constant that contains data set and value is a variable that belongs to the parameters.

Analytics is a process of discovering and communicating to extract the meaning patterns, which can find in data. Data analytics is that the process by applying to investigate the datasets and extract relationships which might be invisible patterns and information.

Big data analytics use business, organization, weather prediction that make better decision. Big data analytics is the process examining large data sets containing a massive amount of data types to examine in database, to identify interesting patterns and establish predictive analytics regression. Big Data analysis includes analytical methods of big data, systematic architecture of big data and data mining software for analysis. Big data analytic is the process to find the useful patterns and relationships in a large amount of datasets that uncovered with regular data management techniques and tools.

As the large amounts of big data are becoming growth from different sources, therefore they are trying to get their measurement and speculations towards its tracks. For the growing large amount of big data, to explore meaningful values and to evaluate the relationship of the statistical analysis.

Data analytics is required to find the hidden from high voluminous data, regression analysis may be more suitable to use. Regression analysis is a set of statistical process for estimating the relationships between outcome variable and one or more independent variable. Logistic regression uses to describe data and to explain the relationship between one dependent variable and one or more independent variable. Logistic regression uses for a different class problem known as classification problems. Regression analysis use to predict the values of a response variable as a predictive modellings. The most common form of regression analysis is simple linear regression, multinomial regression and etc.

Simple linear regression use to model the relationship between two continuous variables correlation provides a measure of the linear association between pairs of variables. , the objective is to predict the value of an output variable (or response) based on the value of an input (or predictor) variable. Simple linear regression use to estimate the relationship between two quantities variables by fitting a line to the observed data.

Multiple linear regression is a statistical method that uses two or more independent variables to predict the outcome of a dependent variable and analysts to determine the variation of the model and the relative contribution of each independent variable. Multiple linear regression works by considering the values of the available multiple independent variables and predicting the value of one dependent variable. The goal of multiple linear regression is to predict the output of the result of the process.

Multinomial logistic regression use to predict categorical membership on a dependent variable based on multiple independent variables. Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems with more than two possible discrete outcomes. The goal of multinomial logistic regression is to construct a model that explains the relationship between the explanatory variables and the outcome, so that the outcome of a new experiment can be correctly predicted for a new data point for which the explanatory variables.

1.2 Motivation of the Thesis

Hadoop and Map Reduce are the most widely used platform for Big Data processing. Weather prediction is big challenge for the world that knowing the future of the weather condition can be important for organizations and business. Weather forecasting system is very useful in the age of technology.

- Hadoop is an opensource framework for massive amount of data processing that supports distributed processing of large chunks of data.
- .The difficulties of weather forecasting problems to find and an enormous volume amount of datasets to easy analysis use Hadoop and Mapreduce framework.
- The evaluation result speed, shows the present state of weather condition that helps the agricultural economy base on crop productivity.
- Weather prediction is important to determine the farmers for effective use of water resources, crop productivity and pre planning of their crop.

Due to these facts, weather prediction is technological challenge, an exactly prediction gives accurate result. The forecasting model for the system uses multinomial logistic regression can predict the estimate of weather condition.

1.3 Objective of the Thesis

The main objectives of the thesis are as follow:

- To provide information people and organizations can use to reduce weather related looses.

- .To predict the condition of the weather for a given location and time.
- To reduce the devastation which may result from natural disasters.
- To increase the quality of current weather prediction by establishing weather forecasting system.

1.4 Contribution of the Thesis

The contribution of the thesis are as follows:

- The system includes Hadoop Distributed File System to store the large amounts of dataset.
- The system uses Map Reduce method that gives expository capacities to analyzing enormous volumes of complex data.
- MapReduce separates a task into smaller parts and assigns the input data to different systems called nodes.
- The system uses to formulate the equation for predicting the weather data with Multinomial Logistic Regression.
- Multinomial Logistic Regression approach, it can predict weather prediction in anyone of the future's year by using climate factors.
- Multinomial Logistic Regression predict future weather data efficiency and computational time for the system.

1.5 Organization of the Thesis

This thesis describes five chapters, abstract, acknowledgement and references.

Big Data is introduced for the weather prediction system in chapter one. This chapter consists of the motivation, contribution, aim, and objectives of the research process.

The chapter two is presented in Hadoop distribute file system (HDFS) and MapReduce method. This chapter is briefly explained big data's features, Map Reduce programming model and regression.

The chapter three describes the processing of Map Reduce method. The designing of weather prediction system, the detail explanation about MapReduce programming

model and regression analysis of Multinomial logistic regression for classification of calculation. informational advantage.

The chapter four includes weather prediction system of design and implementation. The overview of system design, the architecture of the system and the structure of weather prediction system describes in this chapter. The implementation of processing of the weather prediction system describes the chapter. Finally experimental results show with tables.

The conclusion of the system is chapter five. In this chapter, further extensions that the system's development that can be make described.

CHAPTER 2

BACKGROUND THEORY

This chapter describes the important of weather forecasting, the features of big data, the processing of Map Reduce and about multinomial logistic regression. Weather prediction is the use of science and technology to forecast the condition of the weather for a given area. The important of big data features and performance of Map Reduce describe this chapter. Multinomial logistic regression provides high speed clustered processing for the analysis of large set of data smoothly and efficiently.

2.1 Weather Forecasting

Weather prediction is always a big challenge for the meteorologist to forecast the state weather conditions at future time that may be expected. The knowing weather prediction is important for individuals and organizations. Weather forecasting can tell the farmers the best time to plant, the air transformation. Agricultural is the vital role for our country's business and most of people are depend on farming activities. The extremely growled and valued meteorological data has become a challenging task. The main purpose of weather prediction is to know the exactly current state of weather, to support the agriculturist and early to know the storm information.

To get the weather condition, use many ways. a dramatic pace. . Data mining, chi square test, neural network, machine learning and regression analysis and so on. At the system, according to the data use MapReduce method and exactly to know weather result use multinomial logistic regression.

2.1.1 Common Features of Big Data

Big Data is a collection large amount of complex data. The large size and complexity of data cannot use the traditional management tools, store and process.

Structured: The data that standardized formats can be stored, accessed and processed in the formed of fixed format means structured data. It is easy to use, access, process and it is easy the outcomes of accurate.

Unstructured: The data is not arrange different formats and these data does not fit formats. It is a challenge for conventional software to process and analyze. Traditional analytics tools are not easy to analyze unstructured data. The unstructured data use to manage non-relational database and data lakes.

Semistructured: Semistructured data is data that is not match to a data model but has some structured. Semistructured data is combination of structured and unstructured data. The semistructured data cannot organized in relational database. The semistructured model is database and enables to integrate data from various sources or exchange data between different systems.

The voluminous amount of data in which meaningful and useful values are hide to extracted in systematic analysis means big data analytics. Big data describes as large amount of data in complex structures increasing with high volume to gain values out of a big data deal. The characteristics of big data are

- Volume: The volume means data continuously increasing amount of data in terms of size.
- Varity: The varity means different forms uses Map Reduce method that gives expository capacities to analyzing enormous and types of data like structure, unstructured volumes of complex data.
- Velocity: The data is growing and streaming how fast the challenge of speed.
- Veracity: The massive velocity data which uncertain format, these data comes dimension of big data is veracity.
- Variability: The data that comes from different sources, it contains different data types, which needs to know the meaningful data from enormous amount of data. Variability is one of the characteristics of big data.
- Value: The data value is a little gentler of a concept that is weak and sometimes without proper application; high valuable data exists at datawarehouse without any value.

- Visualization: The data presents almost any type of graphical format that makes it easy to understand. The data visualization is not only the decision making to analyze data but also choosing the most effective way to visualize the data.

Validity: Validity uses accurate and correct data for the intended usage. The validity of big data use to get the results for decision-making and correct processing.

2.2 Hadoop

Hadoop is an open-source framework for processing huge amount of data that supports the processing of large chunks of data with using high-level languages. Hadoop is a large-scale data for processing a large amount of data across clusters of computers with the use of high-level languages. Two main components of Hadoop are Hadoop Distribute file system (HDFS) and Map Reduce. Hadoop clusters have many parallel paradigms that store and process large amount of datasets. These paradigms applied easy to use languages and a set of consumer product machines one location. Hadoop uses to efficiently store and process the data, which allows clustering multiple computer to analyze massive datasets in parallel more quickly. Hadoop is a hadoop distributed file system (HDFS) to manage for large datasets. In HDFS large datasets separates into smaller task and then stores in multi locations called Data nodes. These data nodes are connect to a Name node. HDFS process parallel processing and the storage of data is in informal avoiding error.

2.3 MapReduce

MapReduce is a programming paradigm that was develop to handle very large dataset and distribute the files across thousands of nodes. Map Reduce is a parallel programming model and a task-scheduling model. Map Reduce is not only a simple programming model but also an efficient distributed scheduling model. . Large amount of data cut into unrelated blocks by Map program and task to lots of computers to process, receiving distributed computing. The map operation takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce operation combines those data tuples based on the key and accordingly modifies the value of the key.

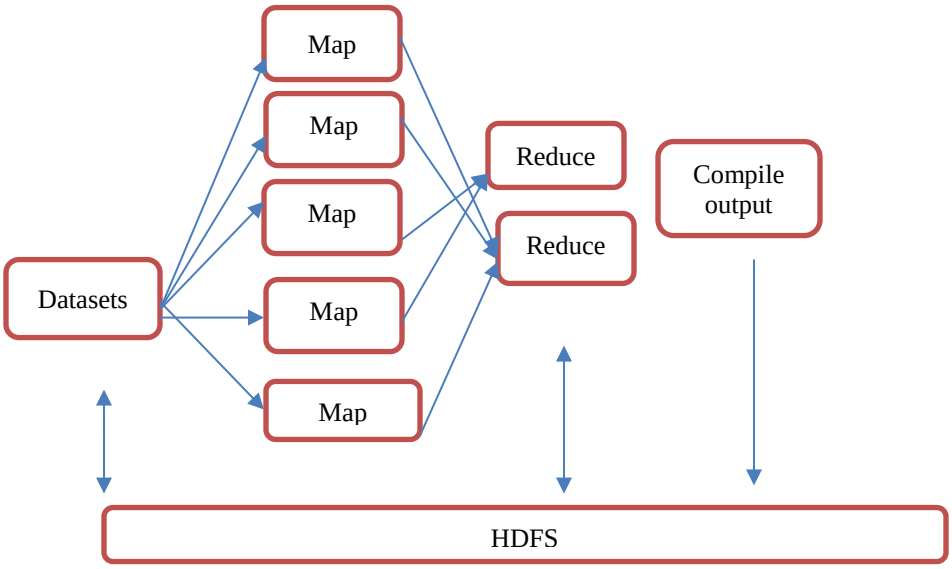


Fig: 2.1 The processing of MapReduce

2.4 Regression Analysis

Today, analysis of a large amount of data is difficult and conventional methods are not get the estimate accuracy. Regression analysis is popular method adjustment method. Regression analysis is widely used for prediction and forecasting, some regression used for classification problems that is based on data modelling and determining the best way to examine the relationships between dependent variables and independent variables. Regression use to solve the major prediction problems that deals with in data mining and machine learning. Regression is the process to find the extract formula or model distinguish data into continuous real values instead of using classes or discrete values. Regression analysis is a set of statistical methods used for estimation of relationships between a dependent variable and one or more independent variables. Regression method can applied in finance, investing, and other fields. The objectives of regression analysis is to estimate the appropriate model for the datasets. The process that is appropriate to perform regression

analysis helps to understand which points are important, which points can ignored, and how they are influencing each other. In statistical analysis, regression used to identify the associations between variables occurring in some data. There are numerous regression analysis approaches available for making predictions. The choice of predictive analytics technique determined by various parameters, including the number of independent variables, the form of the regression line, and the type of dependent variable. The most common of regression analysis approaches are

- Linear Regression
- Logistic Regression
- Polynomial Regression

2.4.1 Linear Regression

In static analytic regression is a linear approach for modelling the relationship between dependent variables and independent variables. Linear Regression is a simple and powerful model for predictive analytics for various fields. Linear regression is an extremely all-around technique that uses to address a variety of research questions and study aims. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change. Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. Linear regression not only tests for relationships but also magnitude their direction and strength. Regression analysis use to develop a more formal understanding of relationships between variables. The results of regression modeling, to determine the variables have an effect on the response known as explanatory modeling. The objective of regression modeling to design the system that will refine our process knowledge and drive further improvement. The goal of linear regression is prediction can use to fit a predictive model to the datasets of values of dependent variables and independent variables.

2.4.2 Logistic Regression

In regression analysis, logistic regression is estimating the parameters of a logistic model. Statistical model of logistic regression often use for classification and predictive analytics. Logistic regression is a process of modeling the probability of a discrete outcome

given an input variable. Logistic regression is a simple and more efficient method for binary and linear classification problems. Classification model of logistic regression is very easy to realize and achieves very good performance with linearly separable classes that is an extensively employed algorithm for classification problems. The model of logistic regression is a statistical method for binary classification that generalized to multiclass classification. Logistic regression is a flexible method for dichotomous classification. There are three types of logistic regression:

- Binary logistic regression only two possible outcomes for binary classification. There are two or more independent variables or predictors for logistic regression. . Logistic regression measures the relationship between the categorical target variable and useful for situations in which the outcome for a target variable can have only two possible outcomes.
- Multinomial is a types of logistic regression model the dependent variable has three or more possible outcomes. A multinomial logistic regression can help to determine the extremely influence for their levels of dependent variables and independent variables.
- Ordinal logistic regression model when the dependent variable has three or more possible outcome these values have defined order. An ordinal variable is a categorical variable for which there is a clear ordering of the category levels. The explanatory variables may be either continuous and estimating ordinal logistic regression models with statistical method is not difficult, but the interpretation of the model output can be heavy.

2.5 Chapter Summary

In this chapter, there are many sections to summarize the shape of system to implement for the system. The main objective of the chapter describes the reviews of related the system's process to motivate the structure of the system. The discussions of current trend of big data, the features of big data, Hadoop, Map Reduce, and the importance of predictive analytic for Regression Analysis. The statistical regression analysis for big

data using multinomial logistic regression model describes and intends to implement the proposed system.

CHAPTER 3

BIG DATA ANALYTICS

This chapter describes general background and history on big data, mapreduce platform and multinomial logistic regression. This section discuss more specifically the details and history of big data analytics. Section 3.1 describes big data analytics. Hadoop Architecture describes in Section 3.2. Section 3.3 explains about mapreduce platform and their properties. Section 3.4 briefly describes the multinomial logistic regression.

3.1 Big Data

Big data is the procedure of extensive informational collections containing various information types. The enormous information keeps the humongous measure of information and procedure them. Big data analytics is the process of analysis the massive large amount of data. Analyzing of big data is a challenging task to the data. The enormous information keeps the humongous measure of information and procedure them. Big data analytics is the process of analysis the massive large amount of data. Analyzing of big data is a challenging task to the data analytics, because it is not fit to store the huge data on a traditional method. Traditional data warehouse for big data which makes more expensive, which involves large distributed file system. Hadoop is an open source java based programming framework that use in companies for internet user application like Yahoo, Facebook and Twitter etc. The large amount of data sets to store and process by using the goods hardware. Hadoop create as flexible alternative to the traditional data warehouse. In Traditional way, big data takes a lot of time to process the large data, by using Hadoop can process huge data very fast. These Big data is not possible to analysis by the traditional data analytics. This propose system present a study of store the big data and analysis by using Hadoop, MapReduce and Apache Spark.

Fig: 3.1 Big Data Analytics

3.2 Big Data Analysis

The Hadoop architecture is a package of the distributed file system, MapReduce engine and the HDFS (Hadoop Distributed File System). Hadoop is a java framework that utilizes a large cluster of commodity hardware to maintain and store big size data. Hadoop works on MapReduce Programming Algorithm. In Hadoop, data exists in a distributed file system that is as a Hadoop Distributed File system. The processing model is based on computational logic is sent to cluster nodes (server) containing data. The computational logic is a compiled version of a program written in a high-level language such as Java. Such a program, processes data stored in Hadoop HDFS. Apache Hadoop consists of two sub-projects.

- Hadoop MapReduce: MapReduce is a computational model and software framework for writing applications that are run on Hadoop. These MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.
- HDFS (Hadoop Distributed File System): HDFS takes care of the storage part of Hadoop applications. MapReduce applications process data from HDFS. HDFS creates multiple same task of data blocks and distributes them on compute nodes in a cluster. This distribution enables reliable and extremely rapid computations.

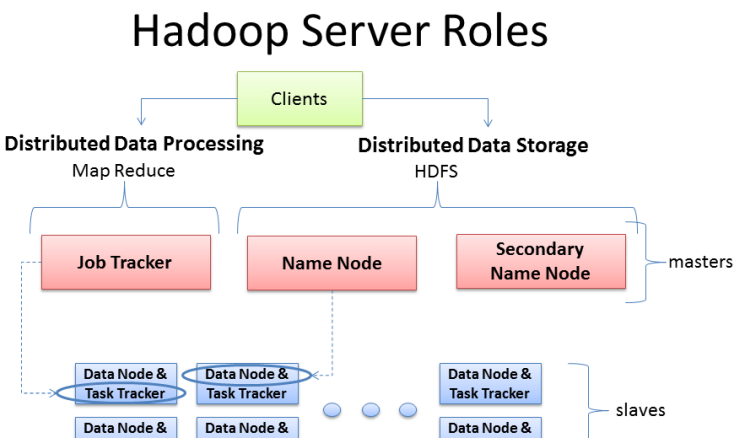


Fig: 3.2 Hadoop Architecture

3.3 Properties of MapReduce

Finding a function to divide the dataset into classes based on several parameters is the process of classification. MapReduce is a software framework and programming paradigm used for processing large amounts of data. MapReduce programming paradigm that process massive scalability across hundreds or thousands of servers in a Hadoop cluster. MapReduce is the heart of Apache Hadoop. MapReduce refers to two separate and distinct tasks that Hadoop programs perform. Map tasks deal with splitting and mapping of data while Reduce tasks shuffle and reduce the data

3.3.1 Components of Map Task

The map function is to extract large amount of data where individual elements are broken into tuples that are key-value pair. The detail of map task describes the following. Hadoop MapReduce: MapReduce is a computational model and software framework for writing applications that are run on Hadoop. These MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.

- RecordReader: The purpose of recordreader is data from its source and converts the data into key-value pair appropriate reading by the mapper. RecordReader connects with the input split (record) until it does not read the complete file.
- Map: The map or mapper's job is to process the input data that in the form of file or directory and stores in the Hadoop file system (HDFS). The mapper processes the data and creates several small chunks of data. The output of mapper is merged end sort by key.

- Combiners: Combiner uses for grouping the data in the Map workflow. Its task is similar to a local reducer. The intermediate key-value that generate in the Map combines with the help of this combiner. Using a combiner is not need as it is optional.
- PartitionarPartitional: PartitionarPartitional is responsible for fetching key-value pairs generated in the Mapper Phases. The partitioner generates the shards corresponding to each reducer. Then partitioner performs it's modulus with the number of reducers.

3.3.2 Components of Reduce Task

The reduce task is operates and reduces the results from each node into a cohesive answer to a query, referred to as the reducer. The detail of reduce task describes the following.

- Shuffle and Sort : The task of reducer process in which the Mapper organize the intermediate key-value and transfers them to the Reducer task known as shuffling. The system of shuffling process can sort the data using its key value. Mapping tasks are done shuffling begins that is why it is a faster process and does not wait for the completion of the task performed by mapper.
- Reduce: The main function of the reduce task is to gather the tuple generate from map and then perform some sorting and aggregation sort of process on that key-value depending on its key element.
- OutputFormat: The performance of operation in key-value pairs writes into the file with the help of record writer, each record in a new line, and the key and value in a space-separated manner.

3.4 Regression Analysis of Multinomial Logistic

Regression analysis is a powerful statistical tool that allows the process to examine the relationship between two or more variables of interest. The purpose of regression analysis is to predict the value of the dependent variable for individuals for some

information concerning the explanatory variables is available, in order to estimate the effect of some explanatory variable on the dependent variable. The proposed system use multinomial logistic regression for two or more dependent variables. The multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems. It uses to predict the probabilities of the different possible outcomes of a distributed dependent variable that gives a set of independent variables.

CHAPTER 4

SYSTEM DESIGN AND IMPLICATION

This chapter briefly describes processing of mapreduce platform and multinomial logistic regression. The proposed system designs for processing of weather forecasting with big data on mapreduce platform. The architecture of the analytics need to redesign so that it could handle historical data to forecasting. For detail clustering, the system use multinomial logistic.

4.1 Overview of the System

The overview of the system is illustrated in figure 4.1.

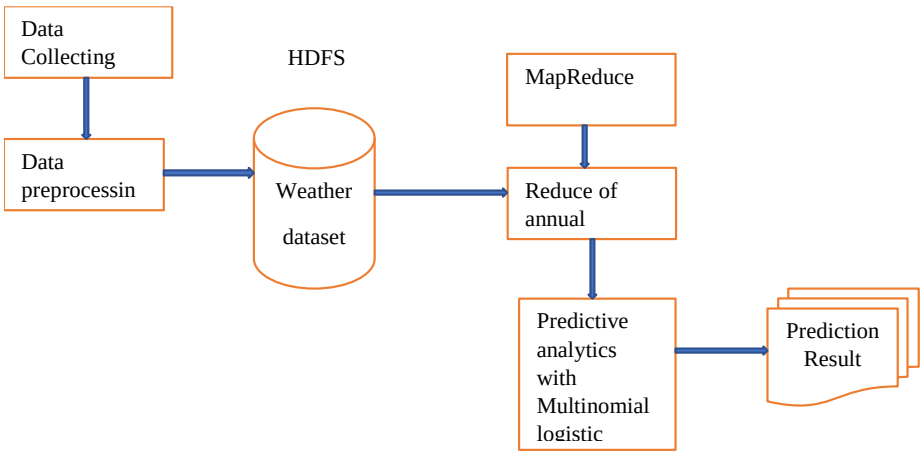


Figure 4.1 The system flow diagram

The overview design diagram of the system shows in Figure 4.1. The proposed system is predictive analytics for weather data using multinomial logistic regression. In the proposed system, weather data files are stored in HDFS and then data files split and goes to different mappers on the MapReduce. This data uses to formulate the equation for weather forecasting with multinomial logistic regression. Multinomial logistic regression implemented for forecasting of weather and computational time for the process that find to be better than traditional methods. Multinomial logistic regression use for the categorizing

the data to predict the weather forecasting. Weather forecasting problem include prediction of the weather condition of the state at real time analytics. The objective of this project is to analyze the meteorologists' data using Hadoop and MapReduce for extracting knowledge about the weather for the purpose of better decision-making. Hinthada historical weather data enables to work as well as real world data where the Hadoop distributed file system uses for faster processing and compared to the latest technique to know the processing speed. Linear Regression

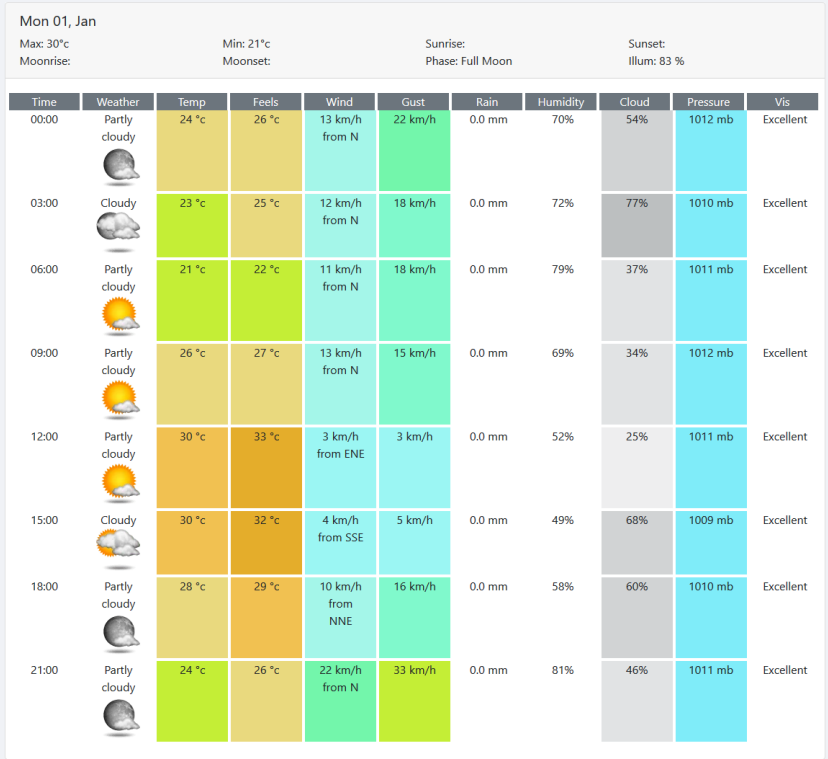
- Data collecting of Hinthada historical weather data source.
- Data pre-processing and transforming into standard for data source.
- Data store on HDFS.
- Data extract using mapreduce.
- To provide the efficient processing of multinomial logistic regression
- Evaluation performance of the system

4.2 Weather Data Analytics

Weather prediction is one of the important applications for the big challenge given location. The big challenge for meteorologists to predict the status of the atmosphere and climatic conditions that may be expected. The weather forecasts can help farmers to know the best time to plant, an airport transformation, storm warning and river overflow. The data center, here we can be able to work on historical as well as real world data where the Hadoop distributed file system use for faster processing and compared that with the modern technique like Spark to know the processing speed.

The data generated by them is unstructured, which becomes a challenging task to analyze it. The dataset has various parameters like temperature, pressure, humidity, speed of the wind and so on. The meteorological department of Hinthada collects these data from the online internet weather service that have deployed for every parameter at various geographical locations. The large amount of data collected, archived in unstructured format. The storage and processing this data for weather condition prediction is a big challenge. Most of the Big Data technologies like Hadoop MapReduce develop to solve these challenges using distributed computing. This project presents the analysis of weather

data by excellent, average, poor and good stations in a particular location or year using various weather parameters. The weather data is growing at a large amounts increases with high speed by various domains like social media, share market etc; Big Data apply various tools and techniques for efficient storing and processing of the massive amount data. The weather data has big amount that it does not match traditional data analysis. Big data can handle data sets with sizes beyond the ability of commonly used software tools to capture and process the data.



Mon 01, Jan

Max: 30°c | Min: 21°c | Sunrise: | Sunset:
Moonrise: | Moonset: | Phase: Full Moon | Illum: 83 %

Time	Weather	Temp	Feels	Wind	Gust	Rain	Humidity	Cloud	Pressure	Vis
00:00	Partly cloudy	24 °c	26 °c	13 km/h from N	22 km/h	0.0 mm	70%	54%	1012 mb	Excellent
03:00	Cloudy	23 °c	25 °c	12 km/h from N	18 km/h	0.0 mm	72%	77%	1010 mb	Excellent
06:00	Partly cloudy	21 °c	22 °c	11 km/h from N	18 km/h	0.0 mm	79%	37%	1011 mb	Excellent
09:00	Partly cloudy	26 °c	27 °c	13 km/h from N	15 km/h	0.0 mm	69%	34%	1012 mb	Excellent
12:00	Partly cloudy	30 °c	33 °c	3 km/h from ENE	3 km/h	0.0 mm	52%	25%	1011 mb	Excellent
15:00	Cloudy	30 °c	32 °c	4 km/h from SSE	5 km/h	0.0 mm	49%	68%	1009 mb	Excellent
18:00	Partly cloudy	28 °c	29 °c	10 km/h from NNE	16 km/h	0.0 mm	58%	60%	1010 mb	Excellent
21:00	Partly cloudy	24 °c	26 °c	22 km/h from N	33 km/h	0.0 mm	81%	46%	1011 mb	Excellent

Figure 4.2 The original weather dataset

4.2.1 Data Pre-processing

Data pre-processing is important step of the proposed system. This step, data transforms into appropriate format. In the propose system, big data sets are input datasets for predictive data analysis of the system. The weather data collected from Hinthada Wikipedia. The data generated by them is unstructured, which becomes a challenging task to analyze it. The original dataset have eleven attributes. The proposed system use eight attributes. This data transforms into understandable format using java programs.

Table 4.1 The pre-processing of weather data

Weather	Temp _C	Feels_C	Gust _kmph	Rain _mm	Humidity _percent	Cloud Prec	Pressure	Vis
Light rain shower	26.C	29.C	24km/h	1.1mm	90%	75%	1004mb	Excellent
Light rain shower	26.C	29.C	27km/h	2.4mm	88%	81%	1003mb	Excellent
Moderate or heavy rain shower	26.C	29.C	27km/h	6.6mm	87%	70%	1004mb	Good
Patchy light rain	23.C	26.C	22km/h	0.1mm	92%	60%	1004mb	Average
Patchy light rain	23.C	26.C	22km/h	0.1mm	91%	60%	1006mb	Excellent
Torrential rain shower	27.C	30.C	25km/h	13.9mm	85%	87%	1006mb	Average

4.3 Data Storage

The proposed system of dataset has 2 gigabyte. The huge amount of data load onto Hadoop distributed file system and file system consists of number of clusters. The data load onto HDFS file system that system appropriate across the cluster. The HDFS file systems are fault tolerant and it contains the replicated file system. The pre-processing data upload onto HDFS file system. The data use mappers and reducers to the process of data. HDFS design to store large files are broken into smaller chunks or blocks. HDFS replicates the data blocks to multiple machines in a cluster that makes the system reliable and fault-tolerant.

.

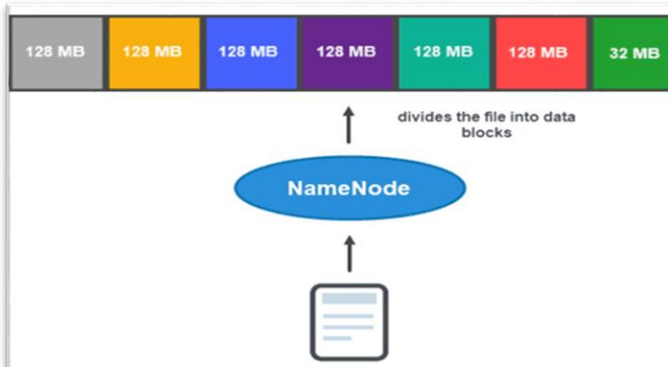


Figure 4.3 The weather data file store in HDFS

4.3.1 Hadoop Architecture

- Namenode : Namenode manages the file system namespace. Namenode is responsible for executing operations such as opening and closing of files, no data actually flows through the Namenode. Namenode executes the read and write operations while the data transforms into Datanodes.
- Secondary Namenode : The processing file keeps growing size the development updates are stored. The Secondary Namenode may not have enough resources, available, as it is performing other operations.
- Datanode : The Namenode stores the filesystem meta-data, the Datanode stores the data blocks and serve the read and write requests. The block reports to the Namenode.

4.4 Processing of MapReduce

Hadoop MapReduce is one of the most widely used models for BigData processing. MapReduce processing by splitting petabytes and gigabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application. The model is a specialization of the split-apply-combine strategy for data analysis. It process by the map and reduce functions commonly used in functional programming. MapReduce have many programming languages, with different levels of optimization. The popular

implementation of MapReduce that has support for distributed shuffles is part of Apache Hadoop. MapReduce framework is a parallel processing problem across large datasets using a large number of nodes, collectively referred to as a cluster that all nodes are on the same local network. Processing can occur on data stored either in a file system (unstructured) or in a database (structured). MapReduce can take advantage of the locality of data; processing it near the place it is stored in order to minimize communication overhead. MapReduce allows for the distributed processing of the map and reduction operations.

- Map: The function of map applies the node to the local data and writes the output to a temporary storage. A master node ensures that only one copy of the redundant input data is processed.
- Shuffle: The function of shuffle is nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node.
- Reduce: The function of reduce is nodes process each group of output data, per key, in parallel.

  MapReduce allows for the distributed processing of the map and reduction operations. Maps can be performed in parallel, provided that each mapping operation. The reducers can perform all outputs of the map operation that share the same key presents to the same reducer at the same time. MapReduce apply to significantly larger datasets than a single commodity server. MapReduce to sort a gigabyte of data in only a few hours.
- Prepare the Map() input :The Map processors, assigns the input key K1 that each processor would work on, and provides that processor with all the input data associated with that key.
- Run the user-provided Map () code : The Map() is run exactly once for each K1 key, generating output organized by key K2.
- Shuffle" the Map output to the Reduce processors : The MapReduce assigns the K2 key each processor should work on, and provides that processor with all the Map-generated data associated with that key Data extract using mapreduce.

- Run the user-provided Reduce () code : The Reduce () is run exactly once for each K2 key produced by the Map step.
- Produce the final output : The MapReduce system collects all the Reduce output, and sorts it by K2 to produce the final outcome.

MapReduce allows for the distributed processing of the map and reduction. Maps can be performed in parallel, provided that each mapping can be performed in parallel, provided that each mapping operation. The reducers can perform all outputs of the map operation that share the same key presents to the same reducer at the same time. MapReduce apply to significantly larger datasets than a single commodity server. MapReduce to sort a gigabyte of data in only a few hours.

- Prepare the Map() input –The Map processors, assigns the input key K1 that each processor would work on, and provides that processor with all the input data associated with that key.
- Run the user-provided Map () code – The Map() is run exactly once for each K1 key, generating output organized by key K2.
- "Shuffle" the Map output to the Reduce processors – The MapReduce assigns the K2 key each processor should work on, and provides that processor with all the Map-generated data associated with that key.
- Run the user-provided Reduce () code – The Reduce () is run exactly once for each K2 key produced by the Map step.
- Produce the final output – The MapReduce system collects all the Reduce output, and sorts it by K2 to produce the final outcome.

The five steps of MapReduce running in sequence – each step starts only after the previous step complete – although in practice they can be interleaved as long as the result is not affected.

The proposed system of dataset use MapReduce function. Weather, temperature, feels, gust, rain, humidity, cloud, pressure and vis are the proposed system of key value.

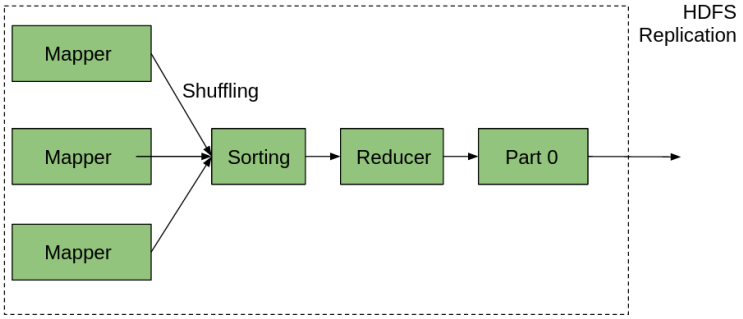


Figure 4.4 The proposed system of MapReduce processing

4.4.1 Regression Analysis for Classification Problems

Logistic Regression use for a different class of known as classification problems. The goal of logistic regression is correctly predict the category of outcome for individual cases using the most parsimonious model. Logistic regression is the appropriate regression analysis to communicate when the dependent variable is dichotomous (binary). Regression analyses, the logistic regression is a predictive analysis and use to describe data and to explain the relationship between one dependent binary variable and one or more nominal, independent variables. The logistic regression are widely used in statistics to model the probability of a certain class or event taking place, such as the probability of a team winning, of a patient being healthy, etc. The logistic regression use the most commonly model for binary regression. The logistic regression model is simple probability of output in terms of input and does not perform statistical classification but is not a classifier. The parameters of a logistic regression is the most commonly estimated by maximum-likelihood estimation (MLE).

4.4.2 Processing of Multinomial Logistic Regression

The proposed system of classification problem have four classes. Thus, to find the classification problem use multinomial logistic regression. Multinomial Logistic Regression is a classification method that generalizes logistic regression to multiclass with

more than two possible discrete outcomes. Multinomial logistic regression analysis use to find the best model to describe between the dependent variable and independent variables.

$$\pi_j = \frac{\exp(\sum_{k=1}^{k} \beta_{jk} x_{jk})}{1 + \sum_{j=1}^{j-1} (\sum_{k=1}^{k} \beta_{jk} x_{jk})} \quad j = 1,2,\ldots,j-1$$ Equation 1

$$\pi_j = \frac{1}{1 + \sum_{j=1}^{j-1} exp - (\sum_{k=1}^{k} \beta_{jk} x_{jk})}$$ Equation 2

Subscript k= coefficient of dependent variables

Subscript j= dependent variable category

The multinomial logistic regression has dependent variables and independents variables. The independents variables have their levels. The proposed system have independents variables and their levels.

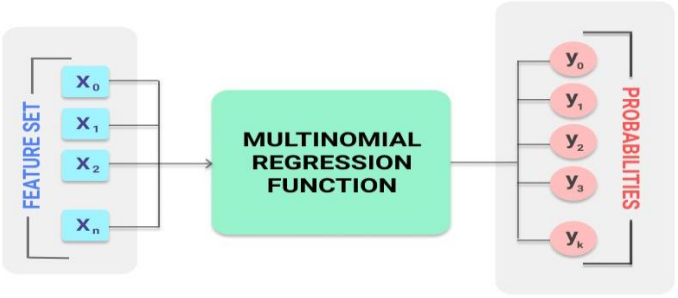


Figure 4.4 The architecture of multinomial logistic regression

Multinomial logistic regression models use for estimations where the dependent variable had more than two categories. Multinomial logistic regression analysis use to find the best model to describe between the dependent variable and independent variables. Multinomial logistic regression models. The dependent variable of which exhibit

multinomial distribution, while there are constraints over independent variables. The model calculate in the system, the odds ratios of the variables that compose the model obtain. The probability of a dependent variable to be in the nth category in a multinomial logistic regression model express.

Table 4.4.1 The pre-processing of weather data

Weather	Temp _C	Feels_C	Gust _kmph	Rain _mm	Humidity _percent	Cloud Prec	Pressure	Vis
Light_ rain_shower	26.C	29.C	24km/h	1.1mm	90%	75%	1004mb	Excellent
Light_ rain_shower	26.C	29.C	27km/h	2.4mm	88%	81%	1003mb	Excellent
Moderate_ or_heavy_ rain_shower	26.C	29.C	27km/h	6.6mm	87%	70%	1004mb	Good
Patchy_light _ rain	23.C	26.C	22km/h	0.1mm	92%	60%	1004mb	Average
Patchy_light _rain	23.C	26.C	22km/h	0.1mm	91%	60%	1006mb	Excellent
Torrential_rain_ shower	27.C	30.C	25km/h	13.9mm	85%	87%	1006mb	Average

4.4.3 System Implementation

The proposed system use the 2 gigabytes of weather datasets are store in Hadoop Distributed File System (HDFS). The raw data of annual weather data preprocessing for the proposed system. HDFS has a scalable storage for large files. HDFS stores each file as a sequence of block.

Figure 4.4.3 Datataset store in HDFS

The preprocessing of weather data file system has namespace. Thus, the weather file system store on Namenode. Namenode may not have enough resources available, as it is other operations. This process called checkpointing. Blocks replicate on the Datanode.

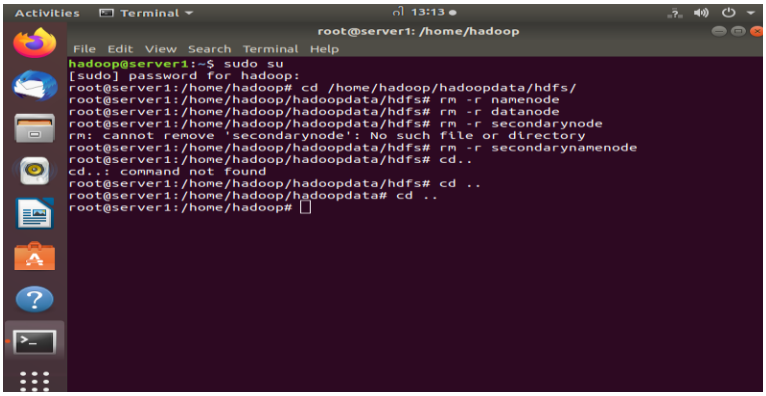


Figure 4.4.4 The weather data HDFS architecture

The preprocessing of weather data load on MapReduce platform. The MapReduce process the output of mapper sort by keys. The reducer store result in HDFS.

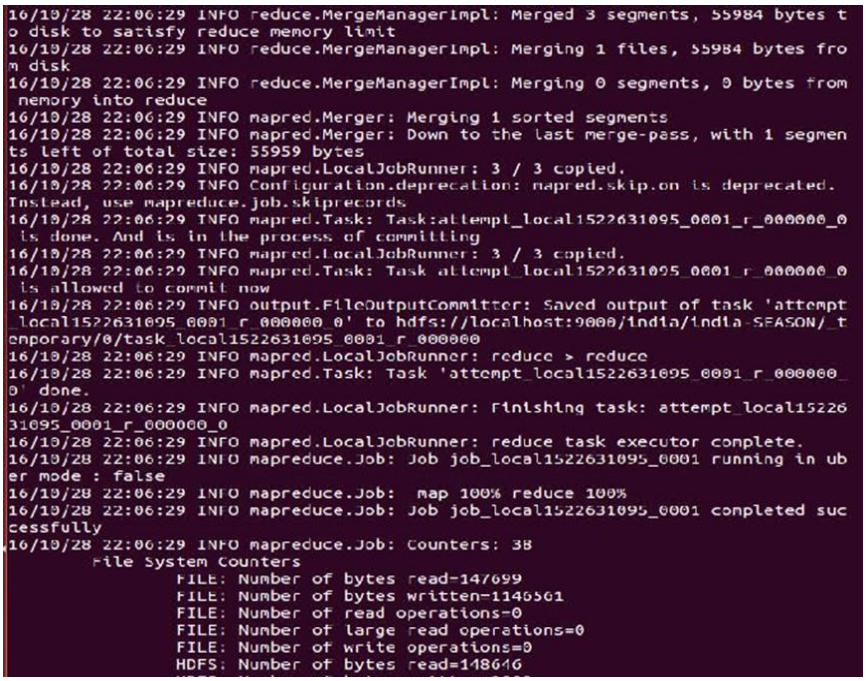


Figure 4.4.5 The processing of MapReduce result

The MapReduce results use to formulate the equation for weather forecasting with multinomial logistic regression. The multinomial logistic regression is the relationship of dependent variables and independent variables. The weather dataset of independent variables and levels of independent variables used in the proposed system. The multinomial logistic regression in which the number of dependent variable categories (Vis) have 3 levels, the sum of probabilities of each category is equal to "1".

$$P(E/P)+P(G/P)+P(A/P)+P(P/P)=1$$

P=Poor

E=Excellent

A=Average

G=Good

The multinomial logistic regression model have four levels that can identified with the overflow performance).For the dependent variable (Y) in the form of weather condition about the Excellent, Good, Average, Poor. For the independent variable, in

the form of characteristics of weather (X_1), Temperature (X_2), Feels (X_3), Gust (X_4), Rain (X_5), Humidity (X_6), Cloud (X_7), Pressure (X_8).

Data analysis technique using multinomial logistic regression performed estimating parameters estimation technique that aims to get a model that will use in classifying. The baseline category select 0 for a dependent variable that consist 0,1,2 and 3 categories. Three odd ratios calculate, each category compare with these ratio, and the model obtain the logistic model.

$$ln\frac{p(Y=0)}{p(Y=1)} = ln\frac{p(Y=0)}{p(Y=2)} - ln\frac{p(Y=1)}{p(Y=2)}$$

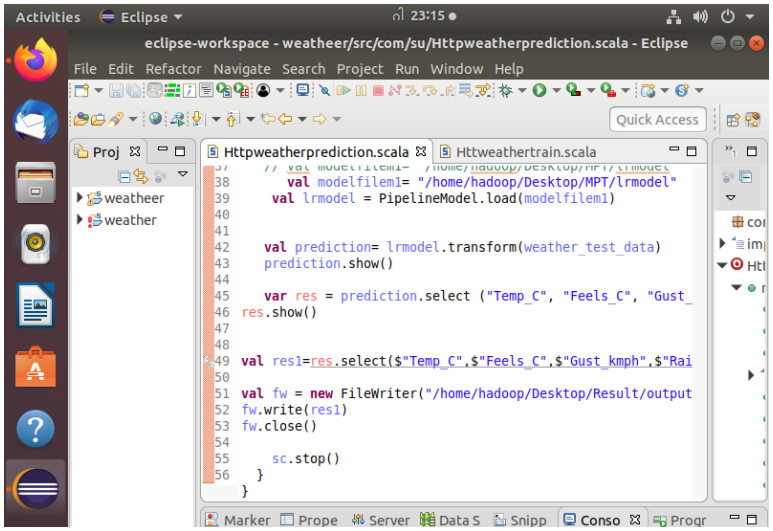


Figure 4.4.6 Working on terminal by weather prediction

The multinomial logistic regression model, load the data file to analyze and generates the independent variables of attributes.

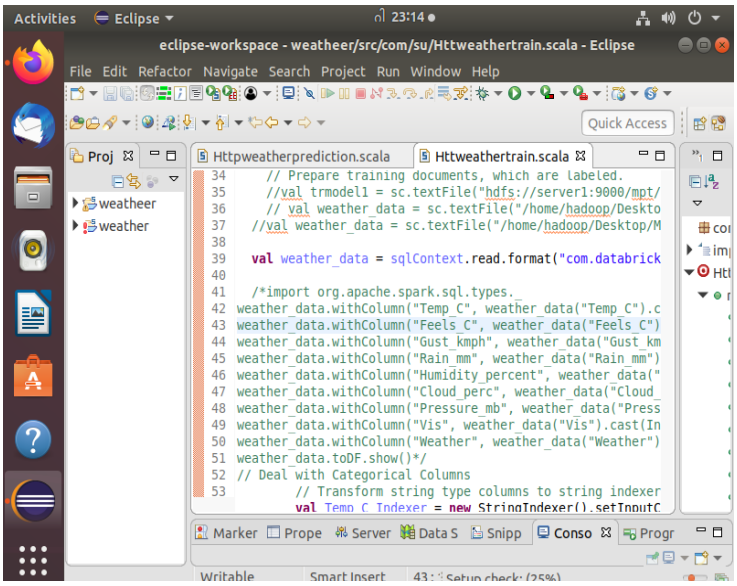


Figure 4.4.7 Working on terminal

The result file either store the data over HDFS or directly dump the data on the terminal. The proposed system process working with faster performance evaluation.

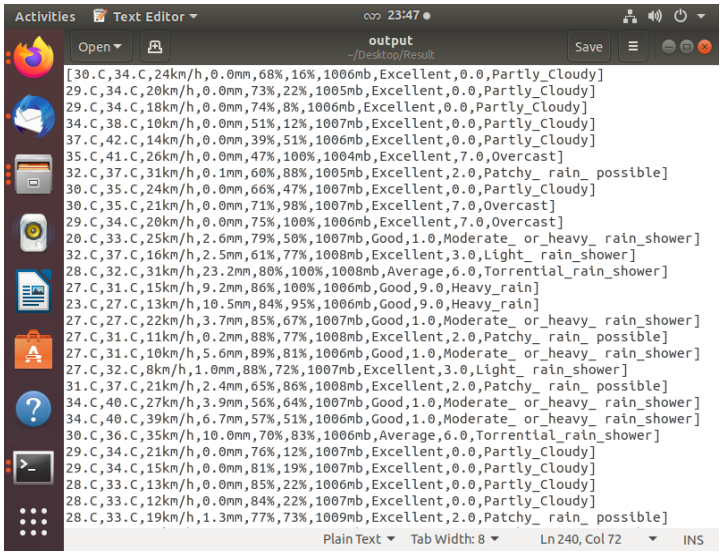


Figure 4.4.8 The output result file

4.5 Classification Accuracy Test

Table 4.5 Classification accuracy test

Observed			
Excellent	Good	Average	Poor
59%	34%	2%	4%

Predicted			
Excellent	Good	Average	Poor
87%	10%	2%	1%

The value of the accuracy of the classificationof weather condition using a multinomial logistic regression analysis In table shows that the percentage of Excellent of weather conditon is 87%, Good of weather conditon is 10%, Average of weather conditon is 2% and Poor of weather conditon is 1%.

CHAPTER 5

CONCLUSION AND FURTHER EXTENSION

Climate data is unique and clamorous in nature and has a huge volume of data. Precise weather prediction is essential for agriculture dependent countries like Myanmar. Using MapReduce can process a huge amount of data and can analyze effectively.

5.1 Conclusion

The proposed system describes the regression analysis of weather prediction using MapReduce framework. This system presents the important of weather prediction. For the growing amount of data to store that use HDFS file system. Complex and unstructured data use MapReduce platform. Multinomial logistic regression plays a main role in the evolution of weather forecasting classification problems. With the increasing amount of daily data is impossible to process and analyze data on traditional methods. The weather data can easily process using Hadoop Distributed File System in a very efficient manner.

5.2 Advantages and Limitations of the System

The propose system use Hadoop, is a platform that is highly scalable, ability to store as well as distribute large data sets. Hadoop's scale-out architecture with MapReduce Programming, allows the storage and processing of data in a very affordable manner. Hadoop offers support for numerous languages that can use for data processing and storage. The proposed system use Logistic regression is easier to implement, interpret, and very efficient to train.It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.

Hadoop is not suited for smaller files. It cannot handle firmly the live data and efficient iterative processing.

5.3 Further Extension

Hadoop and its MapReduce programming model is design to run on general-computers, and is not aim at custom hardware. The costs of operating Hadoop clusters and reduce processing is general-purpose processors to specialized processors. Graphics processing units (GPUs) is an example of such specialized processors that is currently being used in commercial cloud computing platforms.

AUTHOR'S PUBLICATION

Su Hlaing Mon Than, Hmway Hmway Tar, "Weather Prediction Analytics Using MapReduce-Based Logistic Regression", the Proceedings of the Conference on Parallel & Soft Computing (PSC 2022), University of Computer Studies, Yangon, Myanmar.

REFERENCES

[1] Anjana Joseph Joseph and M. Lakshmi "Storm Analysis with Raw Rainfall Dataset by using Artificial Neural Network and Min-Max Algorithms", Indian Journal of Science and Technology Vol 9(10), DOI March (2016).

[2] Basvanth Reddy', Prof. B.A Patil "Weather Prediction Based on Big Data using Hadoop Map Reduce Technique" (2018) International Journal vol. 5, Issue 6, June 2016

[3] B. Anurag, M. Prakash, V. Kanna, and P. Choudhary, "Weather Forecasting using MapReduce", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, No. 9, pp.1-8, 2017.

[4] C.P. Shabariram, K.E. Kannammal, and T. Manojpraphakar, "Rainfall Analysis and Rainstorm Prediction using MapReduce Framework", In: Proc. Of International Conference on Computer Communication and Informatics, pp.1-6, 2016.

[5] Doreswamy and G. Ibrahim, "Big Data Techniques: Hadoop And Map Reduce for Weather Forecasting", International Journal of Latest Trends in Engineering and Technology, Special Issue, pp.194-199, 2016.

[6] E.Ricciarddelli, A.Cersosimo, D. Cimini, and F. D Paola, "Analysis Of Heavy Rainfall Events Occurred in Italy By Using Numerical Weather Prediction, Microwave and Infrared Technique".

[7] Gwo-Fong Lin, Lu-Hsein Chen, "A non-linear rainfall-runoff model using radial basis function.

[8] Joko azhari Suyatno, Fhira Nhita and Aniq Atiqi Rohmaeati Rainfall forecasting in Bandaung Regency using C4.5 Algorithm", May (2018)

[9] K.A. Ismail and M. Abdulmajid, "Big Data Prediction Framework for Weather Temperature Based on MapReduce Algorithm", In: Proc. of International Conference on Open Systems, pp. 1-6,2016.

[10] M. Joshi, S. Shaikh, and P. Waghmode, “Farmer Buddy-Weather Prediction and Crop Suggestion using Artificial Neural Network on Map-Reduce Framework”, International Journal of Computer Applications, Vol. 159, No. 7, pp. 1-3,2017.

[11] M. Senthilkumar, N. Manikandan, U. Senthilkuma and R. Samy, “Weather Data Analysis Using Hadoop”, International Journal of Pharmacy and Technology, Vol.8, No.4, pp.21827-21834, 2016.

[12] P. ChandrashakerReddy and A. Sureshbabu, “Survey on Weather Prediction using Big Data Analystics”, In: Proc. of International Conf. On Electrical, Computer and Communication Technologies, pp.1-6, 2017.

[13] Q. Xiaoyun, K. Xiaoning, Z. Chao, J. Shuai and M. Xiuda, “Short-Term Prediction of Wind Power Based on Deep Long Short-Term Memory”, In: Proc. of International Conference on Asia-Pacific Power and Energy, pp.1148-1152, 2016.

[14] R. Basvanth and B.A. Patil, “Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique”, International Journal of Advanced Research in Computer & Communication Engineering, Vol.5, No.6, pp.1-6, 2016.

[15] S. Selvaragini, and E. Venkatesan, “Big Data Techniques For Weather Forecasting”, International Journal of Pure and Applied Mathematics, Vol.116, No.18, pp.195-201, 2017. [16] Gwo-Fong Lin, Lu-Hsein Chen, “A non-linear rainfall-runoff model using radial basis function.

[16] V. Dagade, L. Mahesh, A. Supriya. and K. Priya, “Big Data Weather Analytics Using Hadoop”, International Journal Technology in Computer Science & Electronics, Vol.14, No.2, pp.194-199,2015.

[17] Wei- Fang, Xuezhi, Wen, Victor Sheng, Wubin Pan Meteorological Data Analysis Using MapReduce, The Scientific World Journal, February 2014.