

**WEATHER ANALYTICS**

**USING**

**HADOOP ECOSYSTEM**

**WEATHER ANALYTICS**

**USING**

**HADOOP ECOSYSTEM**

***in partial fulfillment for the award of the degree***

***of***

**Bachelor of Technology**

**in**

**Information Technology**

**Dr. Mahalingam College of Engineering and Technology**

**Pollachi – 642 003**

**An Autonomous Institution**

**Affiliated to Anna University, Chennai - 600 025**

**MAY 2017**

**MINI PROJECT REPORT**

**Submitted by**

**K.HEMACHANDRU.(14BIT072)**

**R.SEKAR.(14BIT098)**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr.S.Ramakrishnan,

Prof. & Head,

Information Technology,

Dr. Mahalingam College of Engineering and Technology,

NPT-MCET Campus,

Pollachi – 642 003, India.

Certified that this mini project report,

**“WEATHER ANALYTICS USING HADOOP ECOSYSTEM”**

is the bonafide work of

**K.HEMACHANDRU.(14BIT072)**

**R.SEKAR.(14BIT098)**

**INTERNAL EXAMINER**

Submitted for the Autonomous End Semester Examination Mini Project

viva-voce held on \_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ms.K.Kanagasabapathi.

Supervisor,

Assistant Professor,

Information Technology,

Dr. Mahalingam College of Engineering and Technology,

NPT-MCET Campus,

Pollachi – 642 003, India.

Pollachi – 642003 India

**Dr. Mahalingam College of Engineering and Technology**

**Pollachi – 642 003**

**An Autonomous Institution**

**Affiliated to Anna University, Chennai - 600 025**

**EXTERNAL EXAMINER**

who carried out the project work under my supervision.

BONAFIDE CERTIFICATE

**“WEATHER ANALYTICS USING**

**HADOOP ECOSYSTEM”**

ABSTRACT

Weather-data analysis is about lakhs and lakhs of information’s of climate, temperature, atmospheric pressure and some more details about regions, conditions. This data is transformed into understandable format using java programs. These huge amount of data is the loaded onto a Hadoop Distributed File System. This file system consists of cluster. Once the data is loaded onto the HDFS file system the data is balanced across the cluster system are Fault tolerant ,as there exists replication of data among the clusters and if any one of the node fails the whole setup doesn't crash, HDFS obtains the replicated file system. After the map reduce program implementation the Minimized key, value pair of data is created. The output data contains the raw data for our predictions. We can found the way of output by customizing our Map Reduce program, which is written in Java using eclipse. Then finally we can judge the weather conditions by the Output predicted using Splunk tool for Visualization.

**ACKNOWLEDGEMENT**

First we would like to praise the almighty for his grace to bless us with this project and for giving us the confidence to complete it successfully.

We wish to acknowledge with thanks for the excellent encouragement given by the management of our college and **Dr.C.Ramaswamy., M.E., F.I.V., Ph.D.,** Secretary, NIA Educational Institutions.

We sincerely thank our Director **Dr.S.Vijayarangan., M.E., Ph.D., FIE.,** for his constant support for our project.

We wish to express our hearty thanks to **Dr.M.Ramakrishnan, B.E., M.E., Ph.D.,** Principal of our college, for his constant motivation regarding our Project work.

We express our extreme gratefulness to **Dr.S.Ramakrishnan, M.E., Ph.D.,** Professor & Head of the Department, Information Technology for his direction delivered at all times required.

We express our warm and sincere thanks to our guide **Ms.S.Rubika, M.E.,** Assistant Professor, Information Technology, for her tireless and meticulous efforts in bringing out this project to its logical conclusion.

We are committed to place our heartfelt thanks to our project coordinator and all our teaching and non-teaching staff members, lab technicians and friends, and all the noble hearts that gave us immense encouragement towards the completion of our Project.**TABLE OF CONTENTS**

[1. INTRODUCTION 1](#_Toc385279220)

[1.1 Overview 1](#_Toc385279221)

[1.2 Existing System 1](#_Toc385279222)

[1.3 Problem Definition 2](#_Toc385279228)

[1.4 Proposed System 2](#_Toc385279228)

[1.5 Architecture 2](#_Toc385279228)

[2. REQUIREMENT SPECIFICATIONS 4](#_Toc385279229)

[2.1 Software Requirements 4](#_Toc385279228)

[2.2 Hardware Requirements 4](#_Toc385279228)

[3. IMPLEMENATION 5](#_Toc385279220)

[3.1 Modules 5](#_Toc385279221)

[3.2 Module Descriptions 6](#_Toc385279222)

[3.2.1 Data Preparation 6](#_Toc385279222)

[3.2.2 Hadoop Setups 6](#_Toc385279222)

[3.2.3 MapReduce 6](#_Toc385279222)

[3.2.4 Algorithm Implementation 7](#_Toc385279222)

[4. RESULTS AND DISCUSSIONS 8](#_Toc385279229)

[5. CONCLUSION AND FUTURE WORK 9](#_Toc385279229)

[6. REFERENCES 10](#_Toc385279229)

[APPENDIX A : Coding A.1](#_Toc385279231)

[APPENDIX B : Screen Shots B.1](#_Toc385279236)

LIST OF ABBREVIATIONS (in alphabetic order)

HDFS: Hadoop Distributed File System.

MR: Map Reduce.

LIST OF FIGURES

[Figure 1. Project Architecture](file:///C:\reghu\MCET\Courses\2014-Spring\2014MiniProjects\Guidelines\Final-Report\MIni_FinalReport_reghu-v01.docx#_Toc385279244) 2

[Figure 2. Project Flow Diagram 3](file:///C:\reghu\MCET\Courses\2014-Spring\2014MiniProjects\Guidelines\Final-Report\MIni_FinalReport_reghu-v01.docx#_Toc385279245)

[Figure 3. Map Reduce Flow 7](file:///C:\reghu\MCET\Courses\2014-Spring\2014MiniProjects\Guidelines\Final-Report\MIni_FinalReport_reghu-v01.docx#_Toc385279245)

LIST OF TABLES

[Table I. Map-Reduce Program Input-Output Key-Value Pair 7](#_Toc385241742)

**CHAPTER 1**

1.INTRODUCTION

Big Data is the process of examine large data sets containing variety of data types. The big data maintains the huge amount of data and process them. It is traditional data analysis; it is able to process the structured data, but not unstructured data. In big data it is able to process both structured and unstructured data. Weather analytics is the application of technology to predict the action of the atmosphere for a given location. It is important mainly for business agriculturist, farmers, disasters management etc. weather prediction is one of the most interesting and fascinating domain and plays significant role in meteorology.

* 1. **Overview:**

The Analysis of the climate change always has proven very important and useful. In Different stataes, there are regularly many events organized in different cities. These events might include the car racing, festivals, concerts, etc. As these are the outdoor concerts, they suffer a lot from the frequent weather changes, which is increasing due to global warming. To avoid these issues, they need to pre-plan and choose the data for their event in advance.This can work out only if they have any Analysis of the Climate data using the Hadoop and distributed system and map reduce. By using map reduce, we can also calculate the maximum and the minimum temperature for the hot days and cool days. So as the result we can discover useful information about event planning, such as location, time and statistical data,using the analysis and prediction using splunk tool for visualization.

* 1. **Existing System:**

Google weather analytics is an foremost system used in analytics but it is kept in protected way of using this phenomenon but as this project overview states that it simplifies the analytics which can be used as any normal person and we can offer as a private service for our mobile apps and web apps in future etc.to predict the climate changes obtained from the map reduce.

To provide the region temperature for the resultant years,months,dates based on the input data.

To provide visualization of the resultant data and compare the Min,Max,Avg temperature for analytics.

* 1. **Problem Definition:**

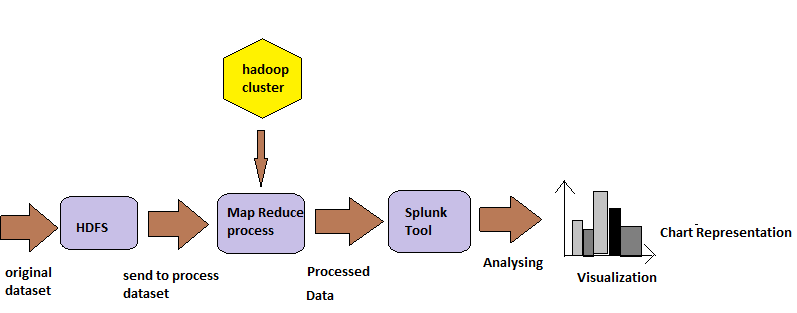
The over exploitation of natural resources has resulted in the serious environmental troubles. Additional problems have also come forward due to the increase in the world’s average temperature. Recent advances in the satellite technology and sensor data has improved in ground-based environmental observations.

* 1. **Proposed System:**

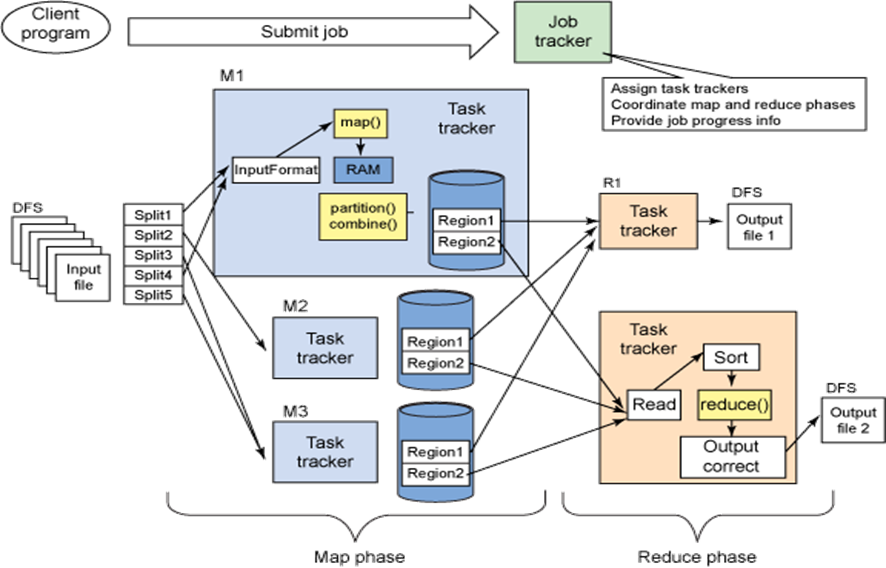
The Apache Hadoop project consists of the HDFS and Hadoop Map Reduce in addition to other modules, that provides high speed clustered processing for the analysis of large set of data smoothly and efficiently,the System which is also used for weather datasets and implement the mapreduce program for our specific data cluster analysis and the result dataset is processed for the future weather datas it seems to be a predicted value of weather forecast, In our project we use Splunk for the above specified analysis.

* 1. **Architecture:**

**WORK FLOW DIAGRAM**



**MAP REDUCE ARCHITECHTURE**



**HDFS ARCHITECTURE**



**Description:**

The resource manager works together with node manager and application manager. Node manager will take the instructions from the resource manager and will manage resource available on a single node. Application masters are responsible for negotiating resources with resource manager and for working with node manager to start the containers. The Node Manager (NM) is also YARN’s prenode agent, and takes care of the individual compute nodes in Hadoop cluster. The continuous features interaction such as air temperature, humidity would provide very valuable and efficient for most of the organization to work under any climatic condition.

1. Requirement Specifications:
   1. Software Requirements

Software’s used in this projects are

1. **Ubuntu 16.04 –** Operating System
2. **Hadoop 2.7.3**– For Processing Large Data set.
3. **Eclipse** – For Creating Map Reduce Java Programs
4. **Java Jdk 1.8** – For Running Map Reduce Program
5. **Splunk –** Data Visualization and prediction.
   1. Hardware Requirements

The Minimum hardware Requirement for this Project are

* + 1. 4 GB of RAM
    2. 500 GB of Hard Disk
    3. Virtualization Enabled System
    4. Pentium 5 processor.

1. IMPLEMENTATION:

This Project is implemented using three modules. In first Module the raw data is preprocessed using MR program which run on Hadoop Cluster. In second module preprocessed data inserted in to Hive database for future usage and schema representation of data.In third module the visualized format of data is displayed using Splunk Tool.

* 1. Module
* Pre-Processing
  + - Input : Raw Log Data
    - Output : Preprocessed Weather Data
    - Tool Used : Map Reduce .jar file
* Data Initialization
  + - Input : Preprocessed Weather Data
    - Output : Stored for Backup.
    - It is fed into Splunk folder for Visualization.
* Analysis
  + - Input : Preprocessed Weather Data Output.
    - Output : Visualized data in the form of charts, graphs…
    - Prediction using Machine Learning Toolkit.
    - Tool Used : Splunk Tool.
  1. Module Description:
     1. Data Preparation

The Weather Data set which present in form text file for processing which is directly fed into the input HDFS file storage in Hadoop.The Weather Dataset contains multiple fields we use mapreduce to process the needed fields in the dataset and the Output Dataset is Saved for future Use.

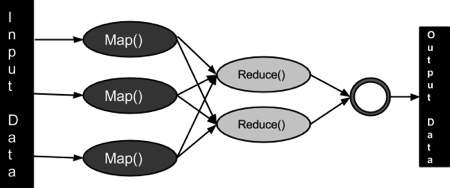
3.2.2 Hadoop Setup

* + - 1. Install Hadoop Latest Version by extracting tar file.
      2. Update $HOME/.bashrc.
      3. Setup Hadoop Distributed File System (HDFS)
      4. Configuration. hadoop-env.sh. conf/\*-site.xml.
      5. Formatting the HDFS filesystem via the NameNode.
      6. Starting the single-node cluster.
      7. Running a MapReduce job.

3.2.3 Map-Reduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

Figure 3. Map Reduce Flow



MapReduce Algorithm executes in two stages, namely map stage and reduce stage

.**Map stage** : The mapper’s job is to process the input data. The input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line.

**Reduce stage** : The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

3.2.4.Algorithm Implemention:

* + Installing the Eclipse Neon.
  + Create New project.
  + Create the Mapper class, Reducer Class,Driver class.
  + Add dependent Jar files.
  + Use the java class and mapper program to cut sort the columns from the dataset.
  + Using Reducer it reduces the common datasets in the fields.

A complex mapreduce program on weather dataset. Here I am using one of the dataset of year 2015 of  Austin, Texas . We will do analytics on the dataset and classify whether it was a hot day or a cold day depending on the temperature recorded by NCDC.

Steps for Running this SetUp :

* Create the project code for our convenience of output needed.
* Import the project in eclipse IDE in the same way it was told in earlier guide and change the jar paths with the jar files present in the lib directory of this project.
* When the project is not having any error, we will export it as a jar file. Right Click on the Project file and click on Export. Select jar file.
* Give the path where you want to save the file.
* Select the main file by clicking on browse.
* Click on Finish to export.
* Weather.txt is our Dataset for our project it size measures upto 1.2 mb . we can extent upto tb’s and more.
* Before running the mapreduce program to check what it does, see that your cluster is up and all the hadoop daemons are running.
* Send the weather dataset on to HDFS.
* **Command:** hdfs dfs -put Downloads/weather\_data.txt /  
  **Command:** hdfs dfs -ls /
* Run the jar file.
* **Command:** hadoop jar temp.jar /weather.txt /output

1. RESULT AND DESCRIPTION:

* The Result of the this project states are,
* The Output file of a Average -Temperature of the day from the Processed Data.
* This Temperature is used for the prediction of next day’s Average Temperature.
* It also Shows the Max and Min Temperature of the day in Dataset using Splunk tool for better Understanding.

As well as we can implement for any datasets which by changes in mapreduce program these are th essentials steps to follow in the project implementation.We can visualize the source Datasets and resultant Datases after Processing.The Tool Used to Vizuzlise the Weather Dataset is Splunk.The Mchine Learning ToolKit is used for the Prediction of Temperatures or any other field specified.

5.Conclusion and Future work:

In Weather Data analysis using big data environment, The method used in our project is Hadoop with map reduces to analyse the sensor data, which is stored in the National Climatic Data Centre (NCDC) is a efficient solution. Map reduce is frame work for highly parallel and distributed systems across huge dataset. It is used to analyse for the given data and predict required output to our project. By using map reduce with hadoop helps in removing scalability bottleneck.

This type of technology used to analyse large data sets has potential to great enhancement to weather forecast. Hence we predict the future weather forecast, minimum and maximum temperature, hot days and cold days based on the data obtained from the NCDC. This helps for the people to preplanning for outdoor events based on the weather conditions.In Future the predictions are simplified and the any person can use this idea for providing service to the mobile or any other platforms to predict before the events occurred.

6. REFERENCES:

[1]. P.Agarwal, S.Das and A.E.Abbadi, “Bigdata and cloud computing: Current state and future opportunities” in Proc Int Conf Extending Database Technol.

[2]. J.Cohen, B.Dolan, M.Dunlap, J.M.Hellenstein and C.Welton, “Mad Skills: New Analysis practices for Bigdata”.

[3]. J.Dean and S.Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters” Commun.

[4]. H.Herodotou et al, “Starfish: A Self-tuning System for Bigdata Analytics”.

[5]. K.Michael and K.W.Millen, “Bigdata: New Opportunities and New Challenges”.

[6] R.D.Schneider, “Handoop for Dummies”, Special Edition. L.Ramaswamy, V.Lawson and S.V.Gogieni, “Towards a quality Centeric Bigdata Architecture for Fractured Sensor services”.

[7]. X.Y, F. Liu, J.Liu and H.Jin, “Building a network highway for Big data: Architecture and challenges.

[8]. M.Mayilvaganan and M.Sabitha, “A cloud-based architecture for Big data analytics in smart grid: A Proposal”.

[9]. V.C.Benz et al, “Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS ready information”.

[10]. M.Olson, “Hadoop: Scaleable, Flexible Data Storage and Analysis”.

[11]. Andue Riberio, Afonso Silva, Alberto Rodrigues De Silva, “Data Modelling and Data Analytics: A Survey from a big data perspective”.

WEB REFERENCE:

* + - 1. [**https://bigdatauniversity.com/learn/hadoop**/](https://bigdatauniversity.com/learn/hadoop/)
      2. <https://www.youtube.com/watch?v=pg3f1ftPlZU>
      3. <https://www.ncdc.noaa.gov/cdo-web/datasets>
      4. Apache Hadoop : http://hadoop.apache.org

APPENDIX:

A: Dataset Structure.

Field#  Name                           Units

———————————————

1    WBANNO                         XXXXX

2    LST\_DATE                       YYYYMMDD

3    CRX\_VN                         XXXXXX

4    LONGITUDE                      Decimal\_degrees

5    LATITUDE                       Decimal\_degrees

6    T\_DAILY\_MAX                    Celsius

7    T\_DAILY\_MIN                    Celsius

8    T\_DAILY\_MEAN                   Celsius

9    T\_DAILY\_AVG                    Celsius

10   P\_DAILY\_CALC                   mm

11   SOLARAD\_DAILY                  MJ/m^2

12   SUR\_TEMP\_DAILY\_TYPE            X

13   SUR\_TEMP\_DAILY\_MAX             Celsius

14   SUR\_TEMP\_DAILY\_MIN             Celsius

15   SUR\_TEMP\_DAILY\_AVG             Celsius

16   RH\_DAILY\_MAX                   %

17   RH\_DAILY\_MIN                   %

18   RH\_DAILY\_AVG                   %

19   SOIL\_MOISTURE\_5\_DAILY          m^3/m^3

20   SOIL\_MOISTURE\_10\_DAILY         m^3/m^3

21   SOIL\_MOISTURE\_20\_DAILY         m^3/m^3

22   SOIL\_MOISTURE\_50\_DAILY         m^3/m^3

23   SOIL\_MOISTURE\_100\_DAILY        m^3/m^3

24   SOIL\_TEMP\_5\_DAILY              Celsius

25   SOIL\_TEMP\_10\_DAILY             Celsius

26   SOIL\_TEMP\_20\_DAILY             Celsius

27   SOIL\_TEMP\_50\_DAILY             Celsius

28   SOIL\_TEMP\_100\_DAILY            Celsius

B: coding of Map Reduce program:

import java.io.IOException;

import java.util.Iterator;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.conf.Configuration;

public class Mapper {

MaxTemperatureMapper class is static and extends Mapper abstract class

having four hadoop generics type LongWritable, Text, Text, Text.

public static class MaxTemperatureMapper extends

Mapper<LongWritable, Text, Text, Text> {

@Override

public void map(LongWritable arg0, Text Value, Context context)

throws IOException, InterruptedException {

String line = Value.toString();

if (!(line.length() == 0)) {

String date = line.substring(6, 14);

float temp\_Max = Float

.parseFloat(line.substring(39, 45).trim());

float temp\_Min = Float

.parseFloat(line.substring(47, 53).trim());

if (temp\_Max > 35.0) {

// Hot day

context.write(new Text("Hot Day " + date),

new Text(String.valueOf(temp\_Max)));

}

if (temp\_Min < 10) {

// Cold day

context.write(new Text("Cold Day " + date),

new Text(String.valueOf(temp\_Min)));

}

}

}

}

public static class Reducer extends

Reducer<Text, Text, Text, Text> {

public void reduce(Text Key, Iterator<Text> Values, Context context)

throws IOException, InterruptedException {

String temperature = Values.next().toString();

context.write(Key, new Text(temperature));

}

}

public static void main(String[] args) throws Exception {

//reads the default configuration of cluster from the configuration xml files

Configuration conf = new Configuration();

Job job = new Job(conf, "weather example");

job.setJarByClass(Mapper.class);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

job.setMapperClass(MaxTemperatureMapper.class);

job.setReducerClass(Reducer.class);

job.setInputFormatClass(TextInputFormat.class);

job.setOutputFormatClass(TextOutputFormat.class);

Path OutputPath = new Path(args[1]);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

OutputPath.getFileSystem(conf).delete(OutputPath);

System.exit(job.waitForCompletion(true) ? 0 : 1);

}

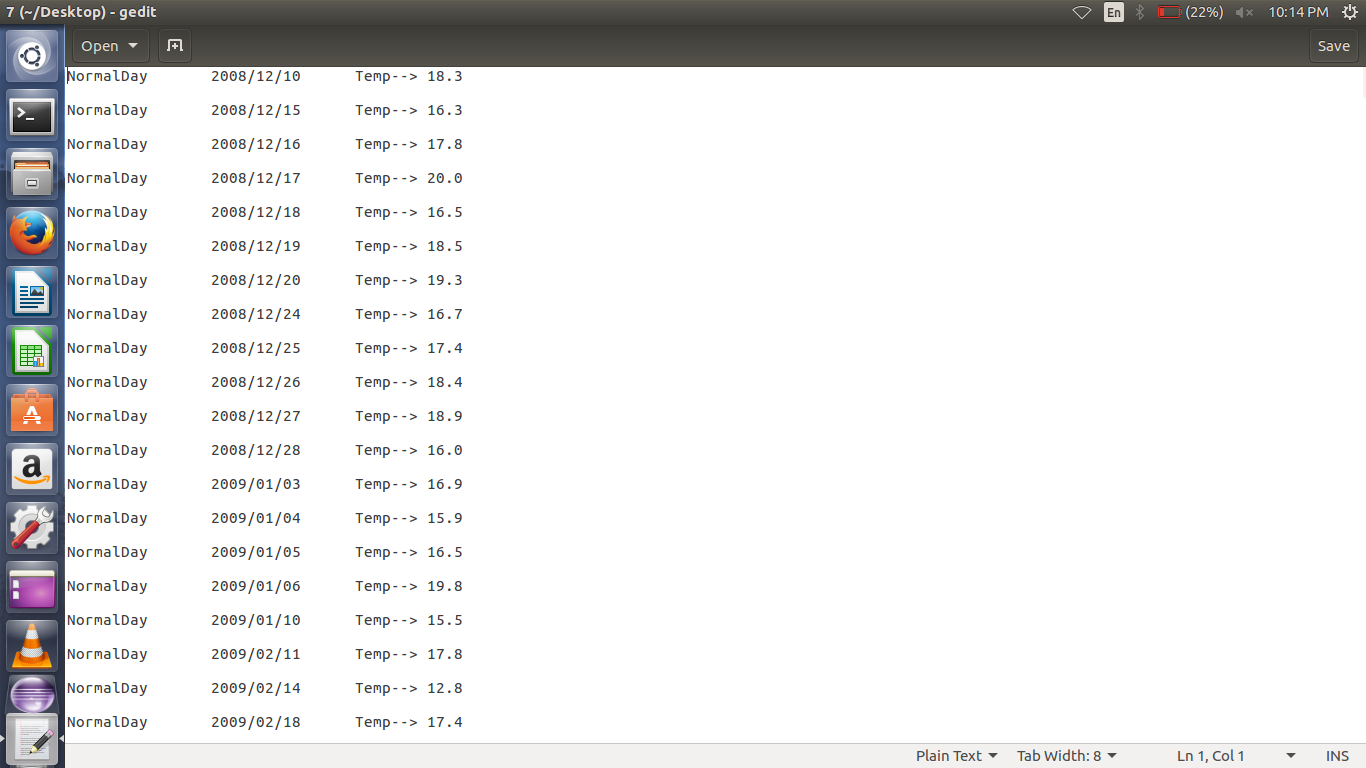
}

This is complete cod eto create the temp jar file to run our mapreduce program to done done a job on hadoop ecosystem.

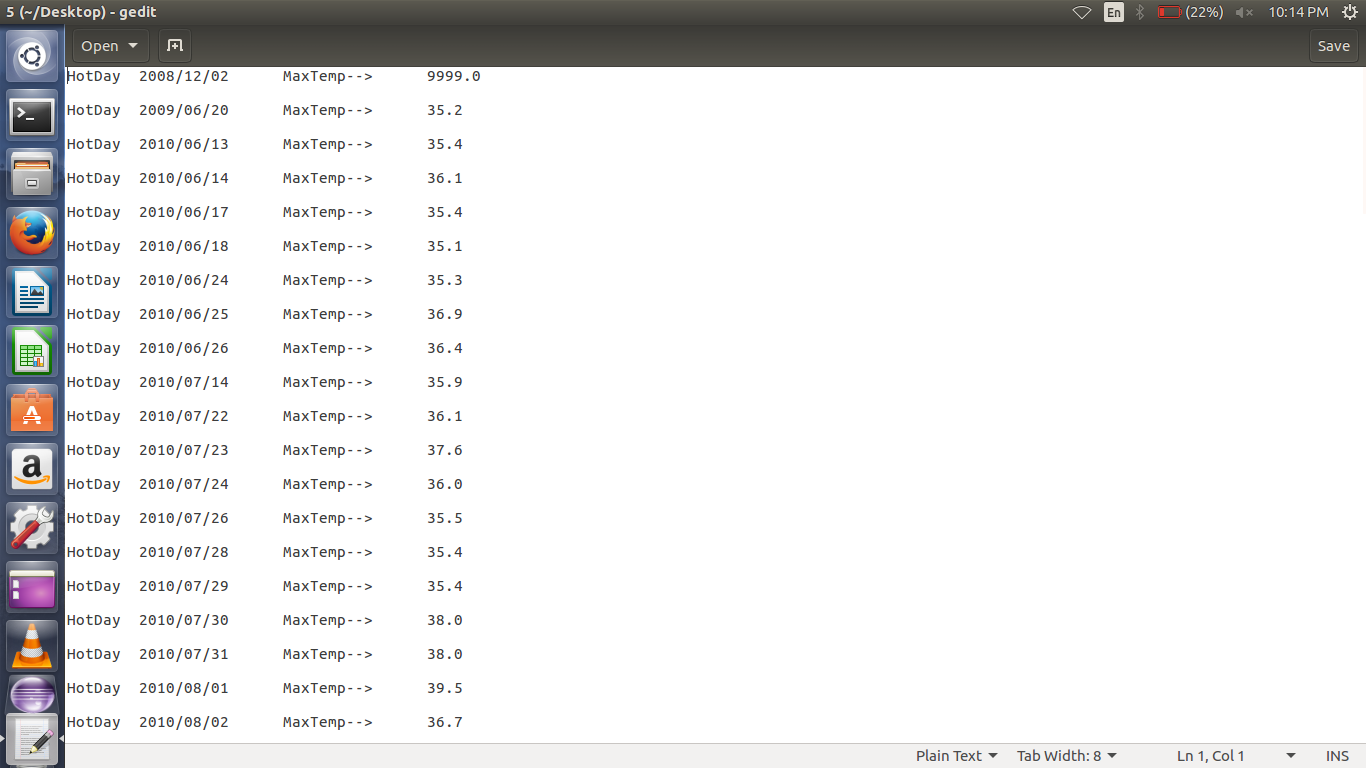
APPENDIX :

C: SCREEN SHOTS OF THE RUNNING SYSTEM

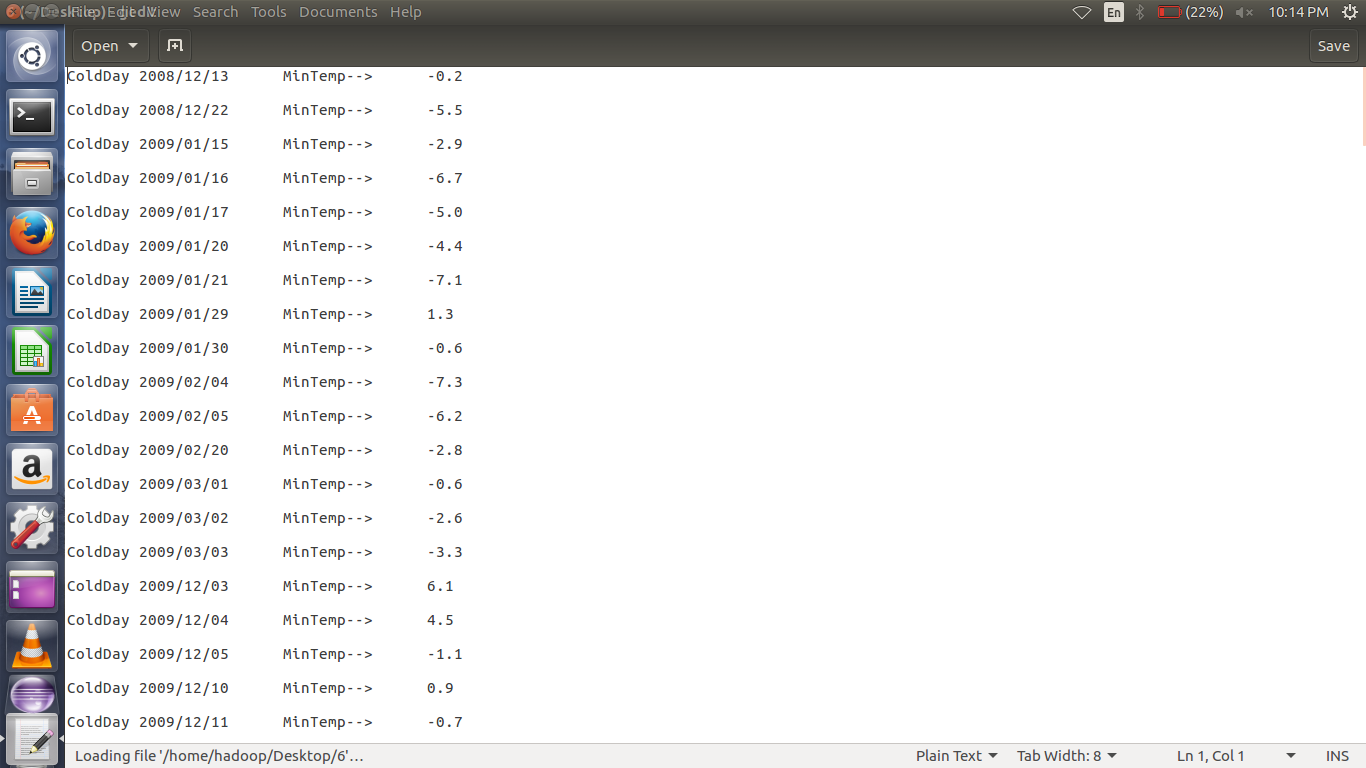
DATASETS:1 Normal Day Temperature.



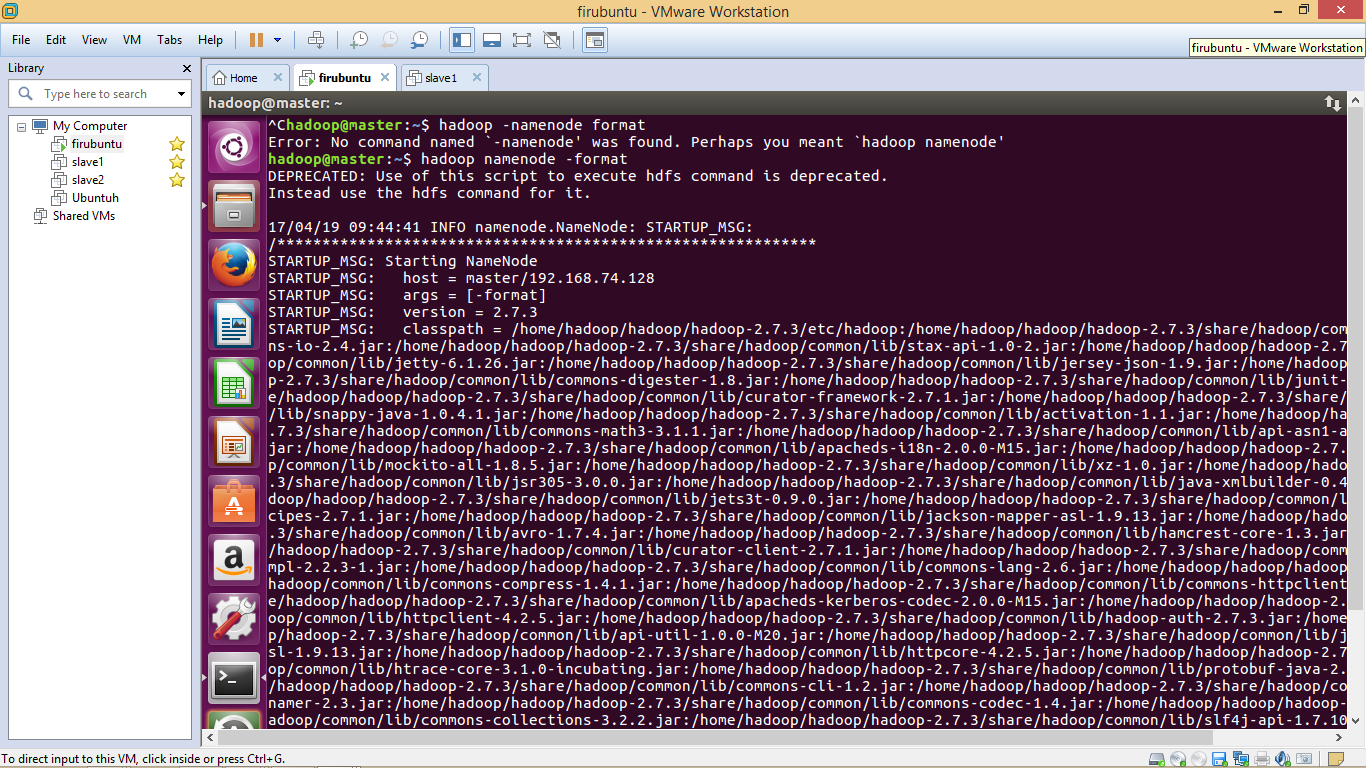
DATASET:2 Hot Day Temperature



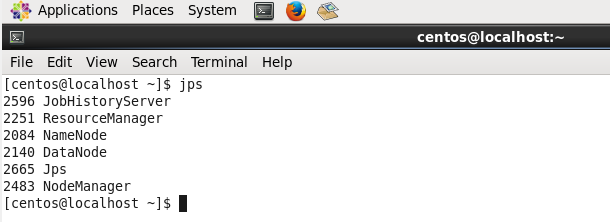
DATASET 3: Normal Day Temperature.



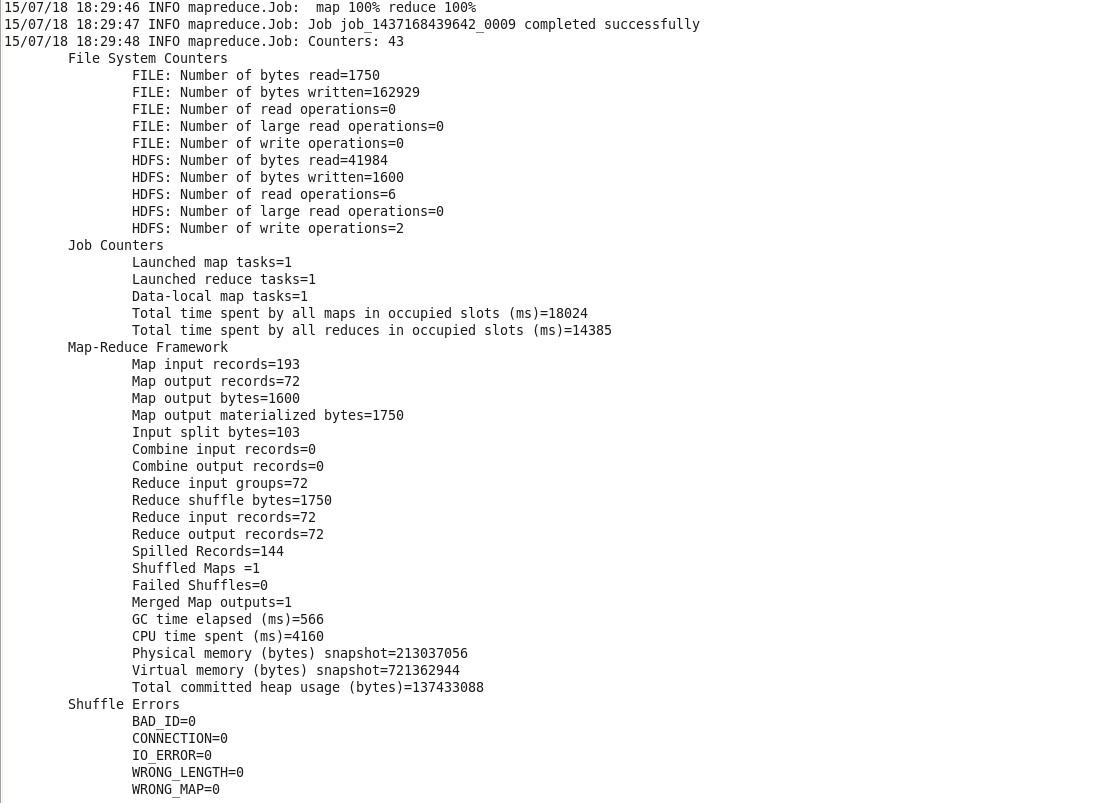
START: HADOOP SETUP.



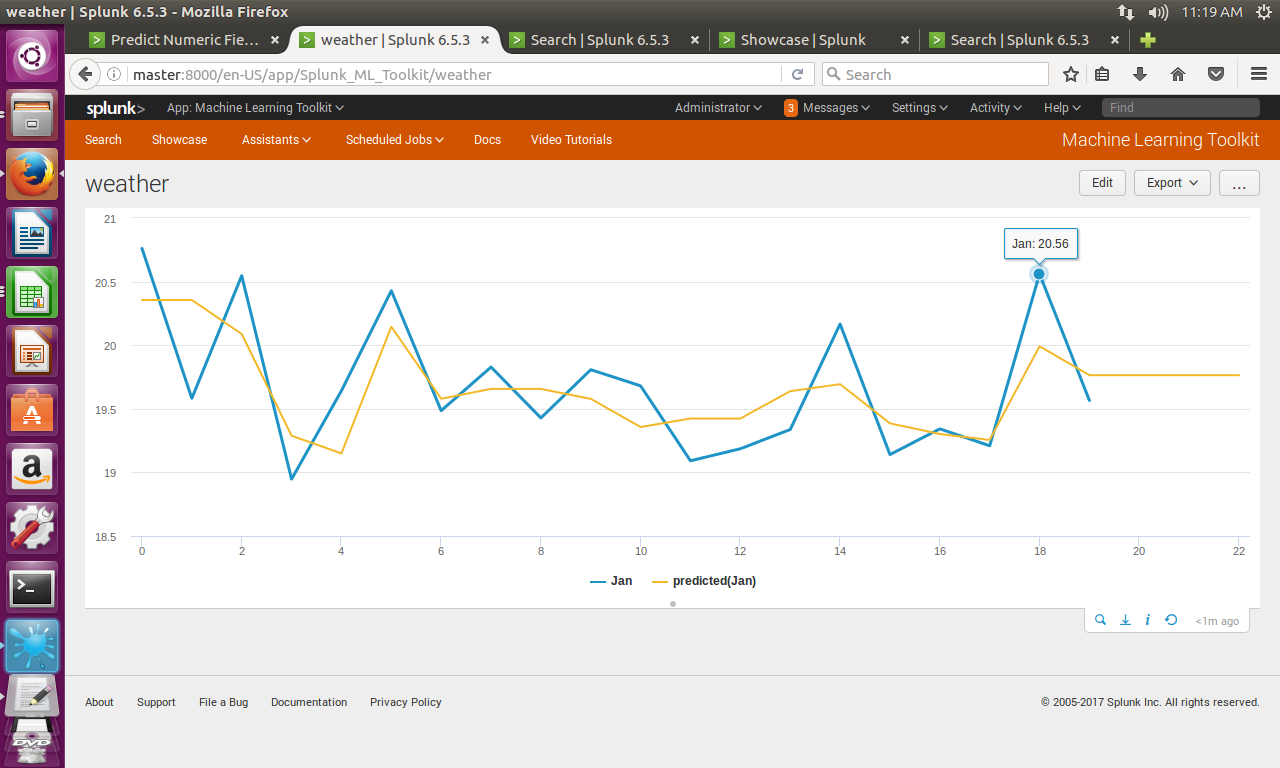
NODES OF HADOOP CLUUSTER:



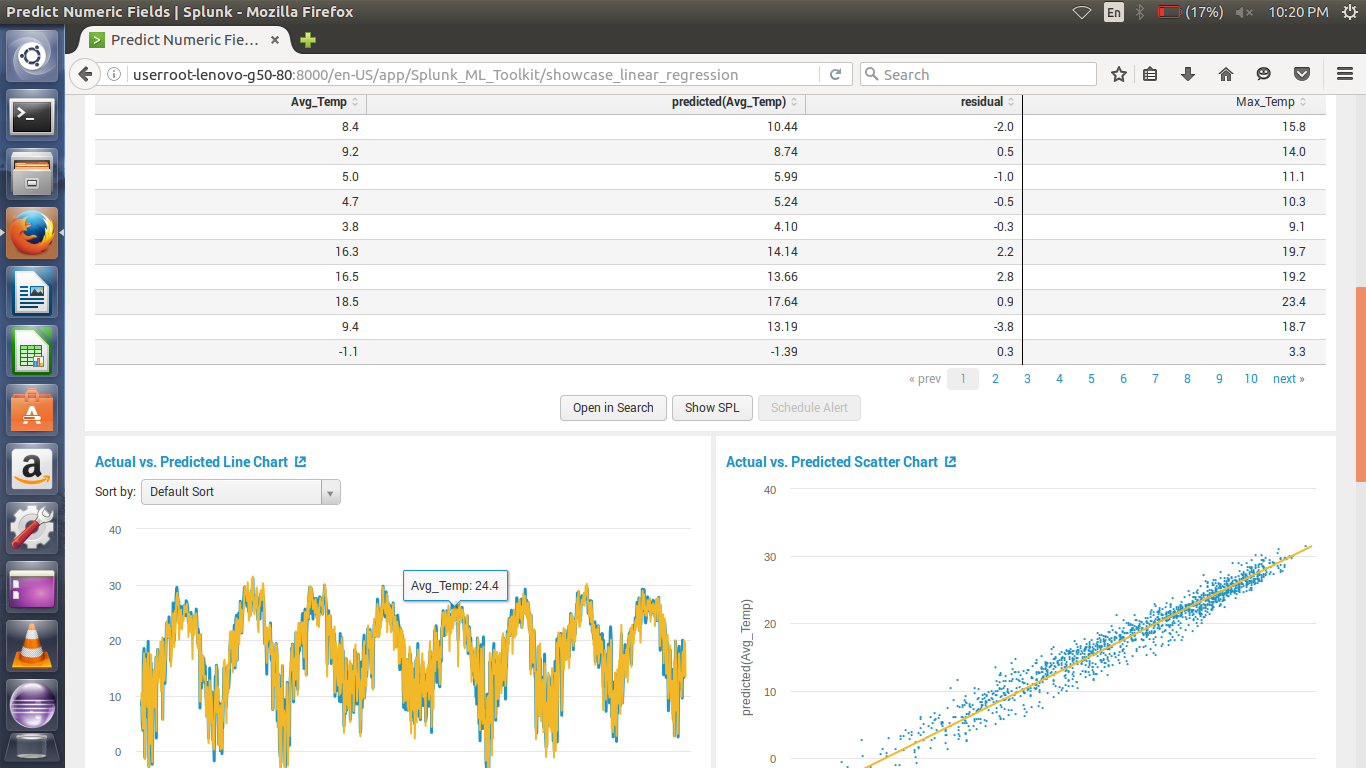
RUNNING MAPREDUCE:



PREDICTION:



NORMAL DAY: PREDICTION



OVERALL VIZUALIZATION OF DATASET:

