A GUIDE TO CLIMATE DATASETS: SUMMARY AND RESEARCH CHALLENGES

T.DHIVYA[1] M.DHIVYAPRIYA[2] R.INDUJA[3] K.KARTHICK.[4].,

[1,2,3] *UG scholar,* [4] *Assistant professor, Department of computer science and engineering, Kongunadu College of Engg & Tech, Trichy,*
(dhivtraj@gmail.com,dhivyapriyamohan@gmail.com,indhujaramu.nkl@gmail.com,karthivel.me@gmail.com)

***ABSTRACT*: Nowadays the Earth system knowledge has reached the massive growth and so that creates the circumstances to know about the global's physical procedures. From the previous few decades the global weather datasets are collected in various formats by the variety of observations .The recent technological approach is being a big challenges to the ancient approaches of the datacentric. In this project, the familiarize the different modules of the Earth system dataset and explains the challenges of the datacentric in evaluating Earth system data. We present this evaluation to overcome the challenges in getting exact information about the climate datasets.**

KEYWORDS: Meteorology, Data models, Ocean temperature, Earth, Atmospheric measurements, sea measurements

1. INTRODUCTION

Earth system datasets that collects the various informations about the surface which are acquired using various acquirement procedures and with varying data features. The local sensor recording collects the observational data about the earth's surface and they can also be collected through instruments attached on the remote sensing satellites. These observations are available at uneven locations with rare coverage in space and time. The observations are converted into fixed spatial and temporal grids by using different interpolation, sampling and aggregation techniques

The number of data is collected by the surface variables such as temperature, precipitation and wind. Climate in a wider sense is the state, including a statistical description .For instance to understand the simple impacts of average temperature[1].Better understanding of aspects of the climate can be ignored from the social perspective .The climatic has extent as well as the computing and climate datasets are analysed through the inputs .The data collection in user interface platform are processed by the user and the climate services information system are get their observation and monitoring system with that research for modelling and prediction in that the capacity development[2].

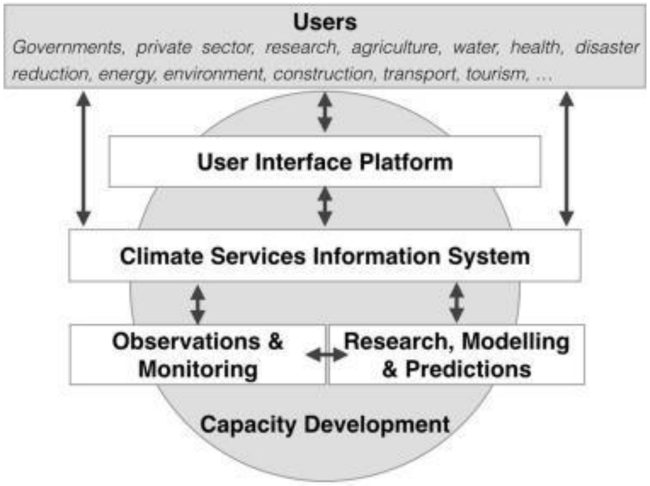

Users
Governments, private sector, research, agriculture, water, health, disaster reduction, energy, environment, construction, transport, tourism, ...
User Interface Platform
Climate Services Information System
Observations & Monitoring
Research, Modelling & Predictions
Capacity Development

Fig 1.1: Climate dataset analysis

2. RELATED WORK

MySQL was used as the backend in the existing system that has various drawbacks which is limitation of data where the processing large set of datas. If once the data is lost they can't be recovered .so the hadoop tool is used for the proposed system. In the existing system, it takes more time and maintenances cost is very high.

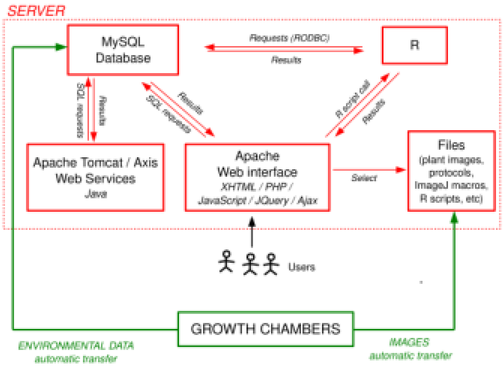

SERVER
MySQL Database
Requests (RODBC)
Results
R
SQL requests
Results
SQL requests
Results
R script call
Results
Apache Tomcat / Axis Web Services Java
Apache Web interface XHTML / PHP / JavaScript / JQuery / Ajax
Select
Files (plant images, protocols, ImageJ macros, R scripts, etc)
Users
GROWTH CHAMBERS
ENVIRONMENTAL DATA automatic transfer
IMAGES automatic transfer

Fig 2.1: MySQL database analysis

This structure represents the evaluation of the MYSQL databases.

3. PROPOSED SYSTEM

Proposed system deals with the hadoop tool for providing database [3]. This process has unlimited datas and add various machines to the cluster thus we obtain results with less time ,high throughput and the cost for maintenance is very less and use joins, partitions and bucketing methods in hadoop [4]. There is no data loss problem and it has efficient data processing.

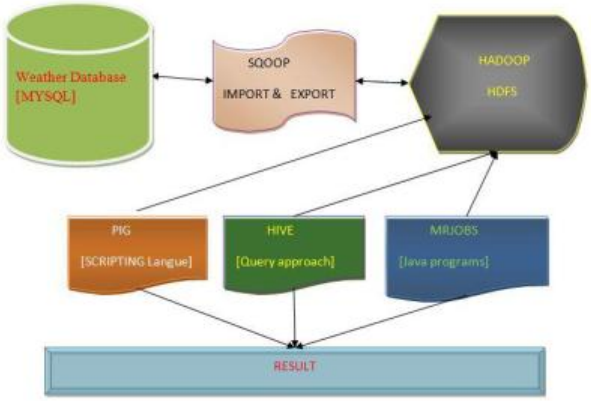


Fig3.1: Hadoop database model

MODULES

- Data Preprocessing Module
- Data Migration Module With Sqoop
- Data Analytic Module With Hive
- Data Analytic Module With Pig
- Data Analytic Module With Map Reduce

3.1.1 DATA PREPROCESSING MODULE

In this module to create Data set for Weather dataset it contains a table with twenty cities each day temperatures details for last 15 years and this data first provide in MySQL database with help of this dataset we analysis this project.

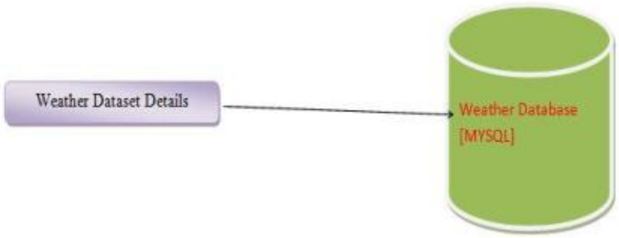


Fig3.1.1:Preprocessing Weather dataset

3.1.2 DATA MIGRATION MODULE WITH SQOOP

This Data migration module is ready with dataset. So that the aim is transfer the dataset into Hadoop (HDFS), that will be happen in this module. Sqoop is a command-line interface application for transferring data between relational databases and Hadoop In this module the fetch dataset into hadoop (HDFS) using sqoop Tool. Using sqoop to perform lot of the function, such that if the module want to fetch the particular column or if the modules want to fetch the dataset with specific condition that will be support by Sqoop Tool and data will be stored in Hadoop (HDFS).

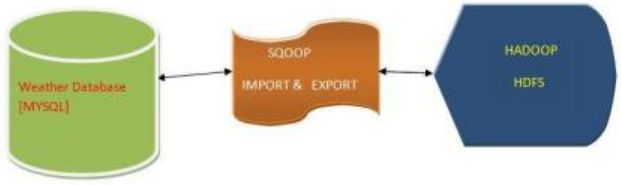


Fig3.1.2:processing dataset with sqoop

3.1.3 DATA ANALYTIC MODULE WITH HIVE

Hive is a data ware house system for Hadoop. It runs SQL like queries called HQL (Hive query language) that gets internally converted to map reduce jobs. Hive was developed by Facebook. Hive supports Data definition Language (DDL), Data Manipulation Language (DML) and user defined functions. In this module the analysis of dataset using HIVE tool which will be stored in hadoop (HDFS).For analysis dataset HIVE using HQL Language. Using hive to perform Tables creations, joins, Partition, Bucketing concept. Hive analysis the only Structure Language.

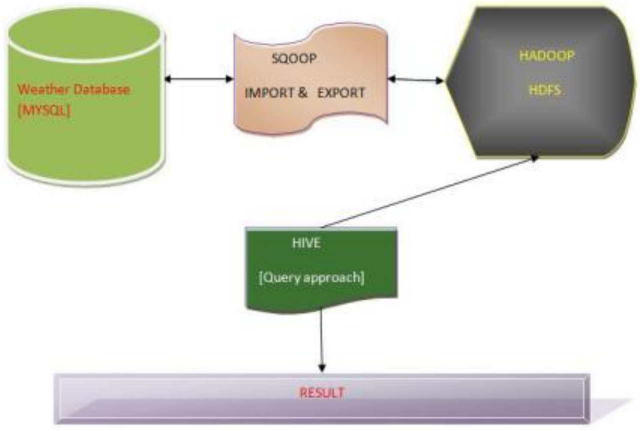


Fig3.1.3: processing dataset with hive

3.1.4 DATA ANALYTIC MODULE WITH PIG

Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop. The language for Pig is pig Latin. Pig handles both structure and unstructured language. It is also top of the map reduce process running background. In this module also used for analyzing the Data set through Pig using Latin Script data flow language in this all operators, functions and joins applying on the data.

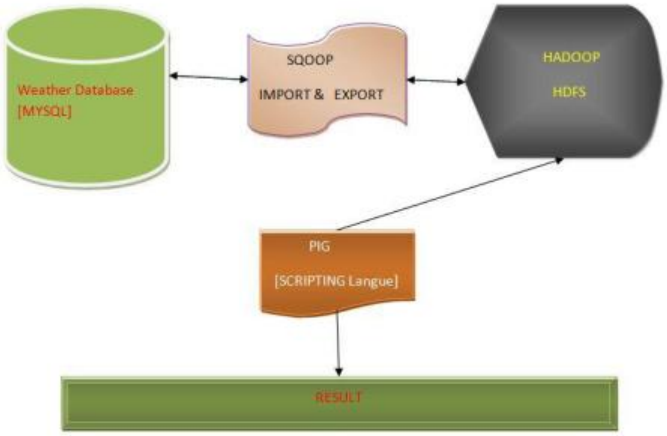


Fig3.1.4 processing dataset with pig

3.1.5 DATA ANALYTIC MODULE WITH MAPREDUCE

In this module the Map Reduce is dealing with the technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. It is used for analyzing the data set using MAP REDUCE. Map Reduce Run by Java Program.

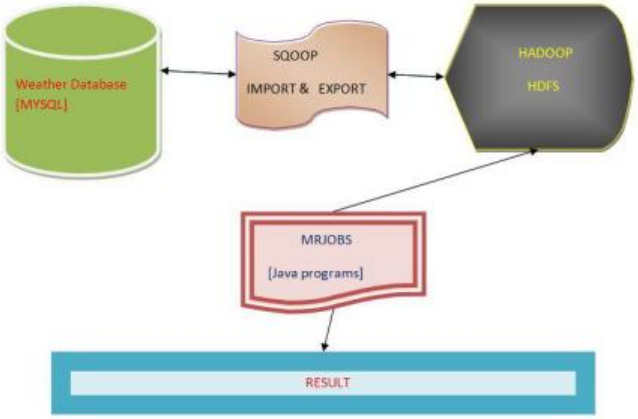


Fig3.1.5 processing dataset with mapreduce

4. MAPREDUCE ALGORITHM

Commonly Map Reduce paradigm is based on sending the computer to where the data resides.

Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

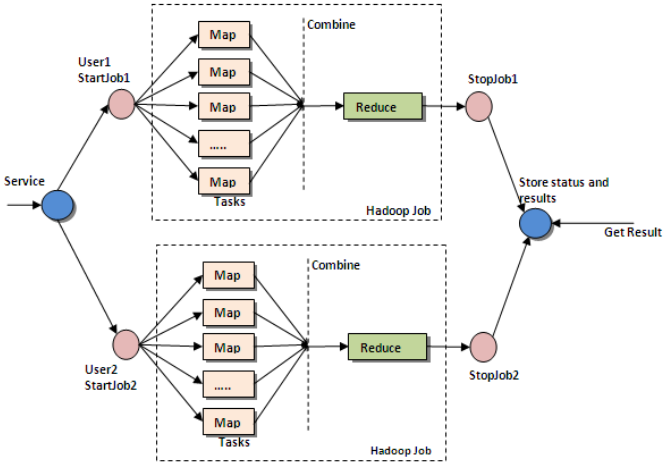


Fig 4.1 Mapreduce process

Map stage: The map or mapper's job is to process the input data. Generally 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 [5]. The mapper processes the data and creates several small chunks of data.

Reduce stage: This stage is the arrangement of **Shuffle** stage and **Reduce** stage[6]. In that Reducer's job is to process the data that comes from the mapper. After the handing it produces a new form of output, which will be stored in the HDFS.

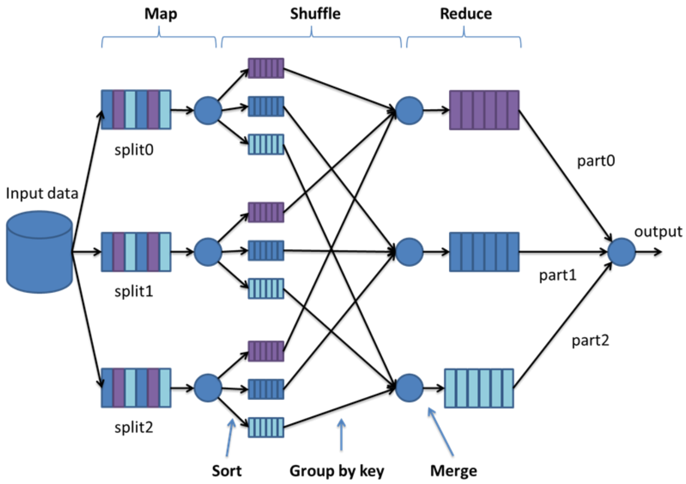


Fig 4.2 Representation of mapreduce

```
Class Mapper
      Method Map (nid n, node N)
           Emit (nid n,NODE_MSG N)
               For all nodeid m $\in$ N.AdjacencyList do
                    Emit (nid m,NBR_MSG N)
                        End for
Class Reducer
      Method Reduce(nid m, [MSG1 N1, MSG2 N2,…])
       M$\leftarrow\emptyset$
      for all $MSG_i$  $N_i \in$ [MSG1 N1, MSG2 N2,…] do
           if IsNode($MSG_i$) then
              M$\leftarrow$ $N_i$
           else if IsForwardNeighbor($MSG_i$)
             add $N_i$ to forward position weight matrix(FPWM)
           else if IsReverseNeighbor($MSG_i$)
             add $N_i$ to Reverse position weight matrix(FPWM)
      end if
 end for
  FCS $\leftarrow$ compute consensus sequence(FPWM)
      if  FCS != $\emptyset$ then
         for all forward neighbors u$\in$ M.AdjacencyList do
      if !consistent(u.sequence , FCS) then
          remove u from M.Adjacencylist
       end if
 end for
      end if
  RCS$\leftarrow$ compute consensus sequence(RPWM)
     if  RCS != $\emptyset$ then
       for all reverse neighbors w $\in$ M.AdjacencyList do
     if !consistent(w.sequence , RCS) then
        remove w from M.Adjacencylist
       end if
 end for
      end if
  Emit(nid m,node M)
```

4.3 Map Reduce algorithm

5.USECASEDIAGRAM

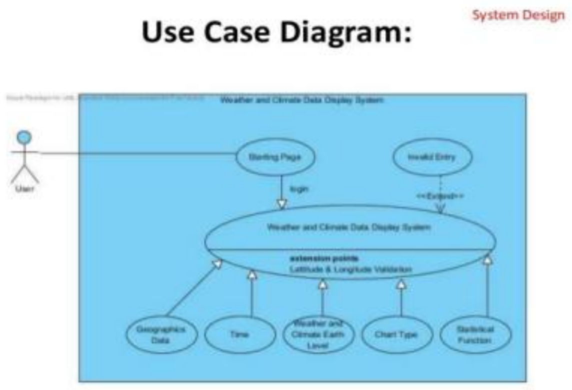


Fig 5.1:Usecase model

The user has to login the page initially.After the login user will reach the page which consists of datas of weather and climate.The user can get the details of geographics data,time,weather and climate earth level,chart type and statistical function.

6. HDFS:Hadoop File System was developed using distributed file system design. It runs on commodity hardware. The other distributed system, HDFS is highly fault tolerant and it was designed using low-cost hardware .HDFS holds very excess of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to set free the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

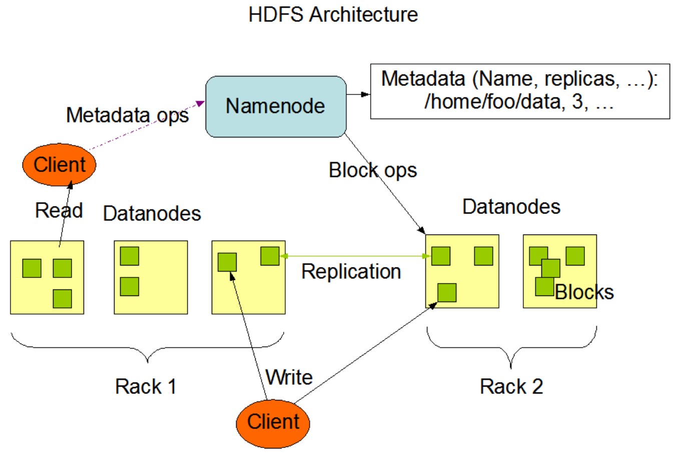


Fig 6.1 Representation of HDFS

7. RESULT

7.1 Data preprocessing Module:

The Weather dataset collecting the temperatures of various cities are represented in the module data preprocessing. That three collected datasets are accessed in form of Preprocessing with datasets of cities. In the module Data preprocessing handles in the gathered datasets of the process.

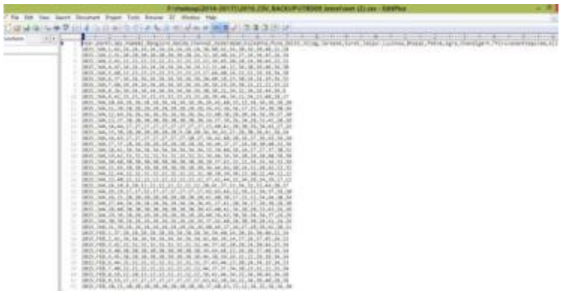
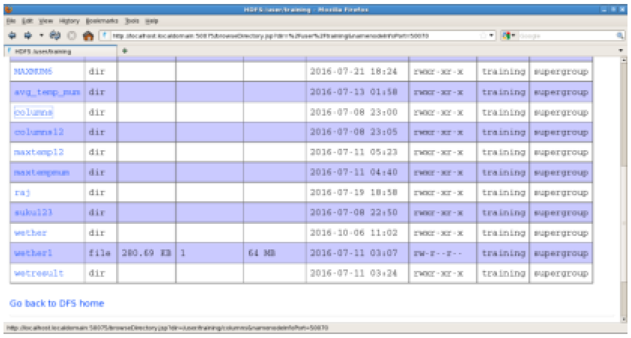


Fig 7.1.1 Data processing module

7.2.TO RETRIEVE THE PARTICULAR COLUMNS

The gathered datasets are retrieved from Hadoop distributed file system Database. As per the needs of user the data's are retrieved [7]. The large amount datas are accessed and get the retrieval of weather data.



MAXMUM6	dir				2016-07-21 18:24	rwxr-xr-x	training	supergroup
avg_temp_mum	dir				2016-07-13 01:58	rwxr-xr-x	training	supergroup
columns	dir				2016-07-08 23:00	rwxr-xr-x	training	supergroup
columns12	dir				2016-07-08 23:05	rwxr-xr-x	training	supergroup
maxtemp12	dir				2016-07-11 05:23	rwxr-xr-x	training	supergroup
maxtempmum	dir				2016-07-11 04:40	rwxr-xr-x	training	supergroup
raj	dir				2016-07-19 18:58	rwxr-xr-x	training	supergroup
suku123	dir				2016-07-08 22:50	rwxr-xr-x	training	supergroup
wether	dir				2016-10-06 11:02	rwxr-xr-x	training	supergroup
wether1	file	280.69 KB	1	64 MB	2016-07-11 03:07	rw-r--r--	training	supergroup
wetresult	dir				2016-07-11 03:24	rwxr-xr-x	training	supergroup

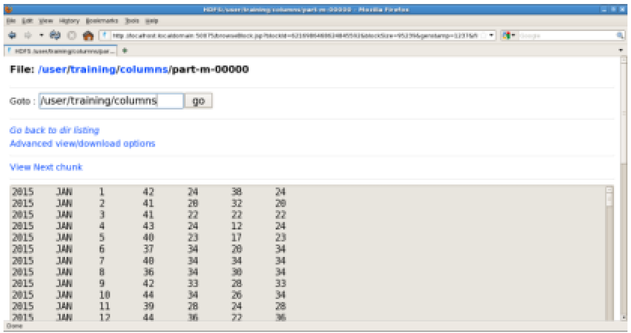
Go back to DFS home

Fig7.2.1 Retrieval of data

These data are framed and gives tabular form of output data.By using their hadoop framework tool the large set of weather data[8].

The output data table is used to retrieve the specific climatic datasets such as temperature details,weather details,pressure details.

These datas can be viewed by the general users by their computer systems.



File: /user/training/columns/part-m-00000

Goto : /user/training/columns go

Go back to dir listing
Advanced view/download options

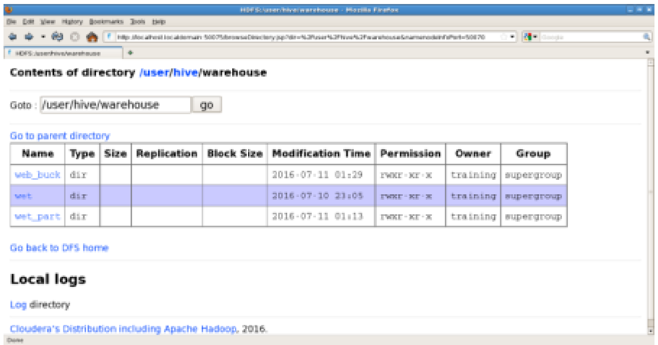
View Next chunk

```
2015    JAN    1     42    24    38    24
2015    JAN    2     41    20    32    20
2015    JAN    3     41    22    22    22
2015    JAN    4     43    24    12    24
2015    JAN    5     40    23    17    23
2015    JAN    6     37    34    20    34
2015    JAN    7     40    34    34    34
2015    JAN    8     36    34    30    34
2015    JAN    9     42    33    28    33
2015    JAN    10    44    34    26    34
2015    JAN    11    39    28    24    28
2015    JAN    12    44    36    22    36
```

Fig 7.2.2 Retrieval of column

7.3. TABLES CREATIONS

The Tabular form has created by the input set of data processed from the database. The output has retrieved from Tabular form .In the dataset created in the database system.



Contents of directory /user/hive/warehouse

Goto : /user/hive/warehouse go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
web_buck	dir				2016-07-11 01:29	rwxr-xr-x	training	supergroup
wet	dir				2016-07-10 23:05	rwxr-xr-x	training	supergroup
wet_part	dir				2016-07-11 01:13	rwxr-xr-x	training	supergroup

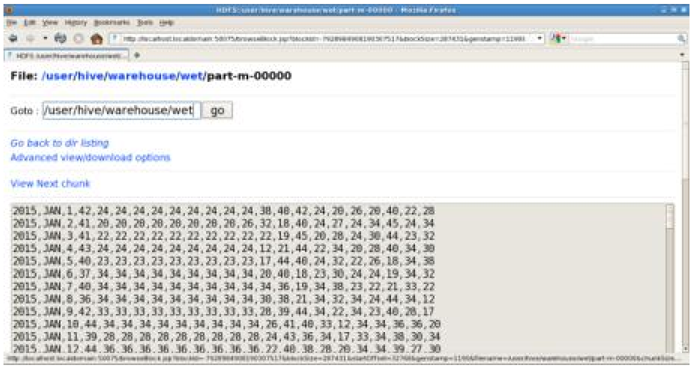
Go back to DFS home

Local logs

Log directory

Cloudera's Distribution including Apache Hadoop, 2016.

Fig7.3.1:Table creation



File: /user/hive/warehouse/wet/part-m-00000

Goto : /user/hive/warehouse/wet go

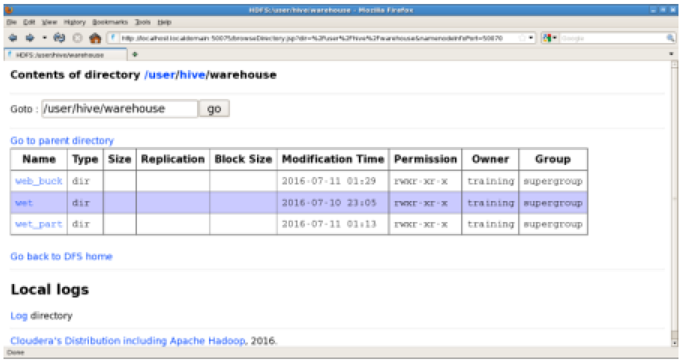
Go back to dir listing
Advanced view/download options

View Next chunk

```
2015,JAN,1,42,24,24,24,24,24,24,24,24,24,38,40,42,24,20,26,20,40,22,28
2015,JAN,2,41,20,20,20,20,20,20,20,20,26,32,18,40,24,27,24,34,45,24,34
2015,JAN,3,41,22,22,22,22,22,22,22,22,22,22,19,45,20,28,24,30,44,23,32
2015,JAN,4,43,24,24,24,24,24,24,24,24,24,12,21,44,22,34,20,28,40,34,30
2015,JAN,5,40,23,23,23,23,23,23,23,23,23,17,44,40,24,32,22,26,18,34,38
2015,JAN,6,37,34,34,34,34,34,34,34,34,34,20,40,18,23,30,24,24,19,34,32
2015,JAN,7,40,34,34,34,34,34,34,34,34,34,34,36,19,34,38,23,22,21,33,22
2015,JAN,8,36,34,34,34,34,34,34,34,34,34,30,38,21,34,32,34,24,44,34,12
2015,JAN,9,42,33,33,33,33,33,33,33,33,33,28,39,44,34,22,34,23,40,28,17
2015,JAN,10,44,34,34,34,34,34,34,34,34,34,26,41,40,33,12,34,34,36,36,20
2015,JAN,11,39,28,28,28,28,28,28,28,28,28,24,43,36,34,17,33,34,38,30,34
2015,JAN,12,44,36,36,36,36,36,36,36,36,36,22,40,38,28,28,34,34,39,27,30
```

Fig7.3.2:Retrieved dataset

The table has to be created and then the datas are inserted.After the datas are inserted they are processed and then they are retrieved.


HDFS:/user/hive/warehouse - Mozilla Firefox
File Edit View History Bookmarks Tools Help
HDFS:/user/hive/warehouse

Contents of directory /user/hive/warehouse

Goto : /user/hive/warehouse go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
web_buck	dir				2016-07-11 01:29	rwxr-xr-x	training	supergroup
wet	dir				2016-07-10 23:05	rwxr-xr-x	training	supergroup
wet_part	dir				2016-07-11 01:13	rwxr-xr-x	training	supergroup

Go back to DFS home

Local logs

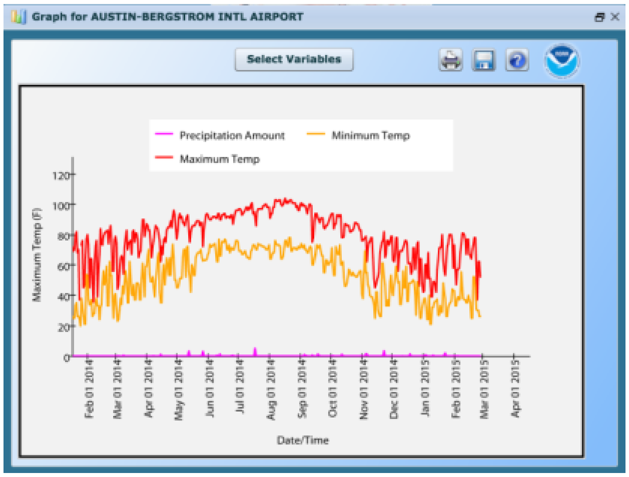
Log directory

Cloudera's Distribution including Apache Hadoop, 2016.

Fig7.3.3: Retrieval of data

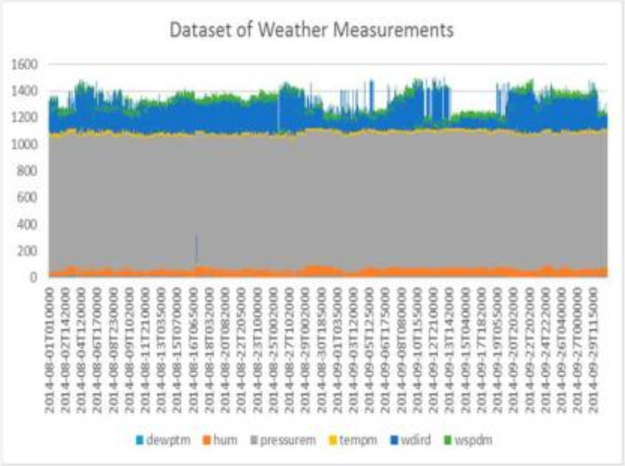
7.4 GRAPHICAL MODEL

The following graph represents the graphical model for weather dataset



7.4.1Monthly temperature representation

The following representation gives the details about the temperature, humidity and pressure.



7.4.2Representation of tepm,hum,pressure

8. FUTURE ENHANCEMENTS

In this paper there is the usage of spark and it gives result hundred times faster than Hadoop. The secret is that it runs in-memory on the cluster, and that it isn't tied to Hardtop's Map Reduce two-stage paradigm. This makes repeated access to the same data much **faster**. **Spark** can run as a standalone or on top of Hadoop YARN, where it can read data directly from HDFS.

9. CONCLUSION

Map Reduce is a framework for executing highly parallelizable and distributable algorithms across huge data sets using a large number of commodity computers [9]. Using Map reduce with Hadoop, the temperature can be analyses effectively. The scalability bottleneck is removed by using Hadoop with Map Reduce. Addition of more systems to the distributed network gives faster processing of the data. The goal of this study was to analyze which city has highest temp, lowest temp recorded in year wise and month wise report generation of previous15year as the forecast for the following year. With the wide spread employment of these technologies throughout the commercial industry and the interests within the open-source communities, the capabilities of Map Reduce and Hadoop will continue to grow and mature [10]. The use of these types of technologies for large scale data analyses has the potential to greatly enhance the weather forecast too.

11. REFERENCES

1. K.E. Taylor, R.J. Stouer, and G.A. Meehl, “An Over- view of CMIP5 and the Experiment Design,” Bul- letin Am. Meteorological Soc., vol. 93, no. 4, 2016, pp. 485–498.

2. W. Tober, “A Computer Movie Simulating Urban Growth in the Detroit Region,” Economic Geogra- phy, vol. 46, no. 2, 2014, pp. 234–240.

3. A.R. Goncalves et al., “Multi-Task Sparse Structure Learning,” Proc. 23rd ACM Int’l Conf. Information and Knowledge Management, 2013, pp. 451–460.

4. K. Subbian and A. Banerjee, “Climate Multi-Model Regression Using Spatial Smoothing,” Proc. SIAM Int’l Conf. Data Mining, 2013, pp. 324–332

5. A. Karpatne et al., “Predictive Learning in the Presence of Heterogeneity and Limited Training Data,” Proc. SIAM Int’l Conf. Data Mining, 2012, pp. 253–261.

6. R.R. Vestavia, “Gaussian Multiple Instance Learning Approach for Mapping the Slums of the World Using Very High Resolution Imagery,” Proc. 19th ACM Int’l Conf. Knowledge Discovery and Data Mining, 2010, pp. 1419–1426

7 Goddard Earth Observing System Data Assimilation System Version 5 (GEOS-5). https://gmao.gsfc.nasa.gov/systems/geos5/.

8. Reanalysis Intercomparison and Observations. http://reanalyses.org.

9. Earth System Grid Federation. http://pcmdi9.llnl.gov.

10. CMIP3 Multi-Model Data. https://esg.llnl.gov:8443/about/registration.do.

11. K. E. Taylor, R. J. Stouffer, and G. A. Meehl, “An overview of cmip5 and the experiment design,” Bulletin of the American Meteorological Society, vol. 93, no. 4, pp. 485–498, 2008.

12.W. Tober, “A computer movie simulating urban growth in the Detroit region,” Economic Geography, vol. 46, no. 2, pp. 234–240, 2002.

13.Knutti, R., 2014: IPCC Working Group I AR5 snapshot: The rcp85 experiment. DKRZ World Data Center for Climate, accessed 14 October 2001,

14.Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2000: Global Historical Climatology Network–Monthly (GHCN-M), version 3. NOAA National Climatic Data Center, accessed 14 October 2000,

15.NOAA/NCDC, 2013: VIIRS Climate Raw Data Record (C-RDR) from Suomi NPP, version 1. NOAA/National Climatic Data Center. Subset used: October 1999–September 1999, accessed 14 October 1999.