International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

Data Analysis and Prediction System for Meteorological Data

Shrikshel Boralkar[1], Shivraj Jawalge[2], Vijay Jadhav[3], Sagar Ingale[4], Dr. J.S.Umale[5]

[1]*Shrikshel Boralkar, Dept. of Computer Engineering, PCCOE College, Maharashtra, India*
[2]*Shivraj Jawalge, Dept. of Computer Engineering, PCCOE College, Maharashtra, India*
[3]*Vijay Jadhav, Dept. of Computer Engineering, PCCOE College, Maharashtra, India*
[4]*Sagar Ingale, Dept. of Computer Engineering, PCCOE College, Maharashtra, India*
[5] *Dr. Prof. J. S. Umale, Dept. of Computer Engineering, PCCOE College, Maharashtra, India*

Abstract – *Weather prediction plays very important role in our daily life. Farmers are dependent on weather conditions. Builders, fisherman, all who work outdoor need weather prediction. In the earlier implementations multiple attempts are done for data analysis system for weather prediction. More efforts are expected to analyze the real time data to improve the prediction accuracy.*

Currently running systems are stand alone and having dependency on centralized storage. They need suitable data visualization of outputs. The analytic algorithms used are need to be more efficient or they need to produce reliable and correct output.

In order to handle such type of situations system should be modified. Better analyzing algorithms running on distributed systems can be used to enhance the accuracy of result, performance and reliability.

In this paper we are doing the comparative study of Data processing techniques. The data files having the data collected from various sensors including various parameters are processed and graphs are generated for analysis and prediction purpose. These graphs are generated by three ways i.e. Sequential, In-parallel by making use of threads and in Hadoop (distributed approach).

The distributed or parallel model will overcome the drawbacks of existing system i.e. time consumption.

Key Words: **Weather data, Analysis and Prediction, Hadoop, Parallel Computing, Graph plotting.**

1. INTRODUCTION

The need of weather forecasting began with early civilizations using reoccurring meteorological events to help them monitor the changes in the weather. Throughout the centuries, attempts have been made to produce forecasts based on personal observations. In ancient times people predicted the weather from cloud patterns and by observing patterns of events. Although these ancient methods of weather forecasting were used for centuries, they were not always reliable. Another limitation was that information about the current state of the weather could not be communicated to places far away, i.e. No real time prediction.

Meteorologists use a variety of tools to help them gather information about weather and climate. Some more familiar ones are thermometers which measure air temperature, anemometers which gauge wind speeds, and barometers which provide information on air pressure. These instruments allow users to gather data about what is happening near Earth's surface. Using these data, crude weather maps were drawn and surface wind patterns and storm systems could be identified and studied. While the meteorological instruments were being refined during the nineteenth centuries, but they were not able to process the data in real time.

Over the past few centuries, physical laws governing aspects of the atmosphere have been expressed and refined through mathematical equations. Despite the advances made by scientists, it took them, several months to produce a wildly inaccurate six-hour forecast. During the data assimilation process, information gained from the observations is used in conjunction with a numerical model's most recent forecast for the time that observations were made to produce the meteorological analysis.

By using various processing methods like Map-reduce pre-processing etc. We can perform analysis. Graph is plotted on the pre-processed data which is used for further processing.

2. RELATED WORK

Timely rainfall forecast is a big challenge and requirement for a country like India. Various rainfall forecast models have been created in past. Numerical Weather forecasting model is used to generate Short range and Medium range forecasts which uses partial differential equations to conduct multiple numerical predictions. Ensemble forecasting is a numerical weather prediction model that uses different forecasts model with slightly different input to gather better forecasts.

Bayesian approach based model for rainfall prediction is created where posterior probabilities are used to calculate likelihood of each class label for input data instance and the one with maximum likelihood is considered resulting output. Five years data is taken as input to compute rainfall using Karl person coefficient which is then compared with rainfall statistics for future years predicted using multiple regression technique. C. E. Decision tree classification algorithm to classify land cover for different vegetation's using remotely sensed data is also used. They have described three different types of decision trees. Univariate trees tests single attribute at any particular node of tree whereas multivariate tree uses

more than one attribute for testing condition at branch while splitting. Hybrid trees are heterogeneous trees as they use more than one algorithm to build the tree. Error complexity method by Breiman for CART trees gives a set of pruned trees and one with lowest misclassification rate is considered final tree. Pessimistic Pruning by Quinlan is used to prune trees build by C4.5 algorithm.

A comparative study of classification algorithms for forecasting rainfall is published by an author Deepti Gupta [1] in which a record level analysis is carried out. Following are the algorithms which are compared by Deepti Gupta [1].

1. Decision tree classification is a nonparametric technique which uses a tree like structure to classify the input dataset instances. Each sample is assigned a class label by moving in a top down manner from root and testing the condition at the branch. It is based on greedy approach as best split is used at each step to build the tree. Decision trees are easy to work with and even with low domain knowledge. Decision trees can be drawn using ID3 algorithm, C4.5 and CART. ID3 and C4.5 were given by Quinlan where C4.5 is a successor of ID3. C4.5 adds some improvements to ID3 like pruning of tree to avoid overfitting of data, handling of missing values. ID3 uses Information gain to find best split while C4.5 uses Gain ratio. Classification and Regression trees (CART), an algorithm by Breiman uses Gini index as selection measure to find best attribute for splitting and constructing binary decision tree.[2] Classification trees are built when resulting class label is categorical like y or n whereas regression trees are used for numerical class labels like income of a person, rainfall in millimeters.

2. K nearest neighbors are lazy learners where learning is based on analogy. The algorithm cannot be used until the sample for which neighbors are to be computed is available. K nearest neighbors are computed for given instance t for which class label is to be predicted. The attributes or features are preferred to be numeric in nature as neighbors are computed using the distance metric.

3. Bayesian approach for classification is a statistical and linear classifier which predicts class label for data instance on the basis of distribution of attribute values. This is a parametric classification where the size of classifier remains fixed. Distribution can be normal (Gaussian), kernel, multivariate or multi nominal. Assuming normal distribution for weather data, Bayesian classifiers use Bayes theorem to find posterior probabilities of occurrence of input data instance in all classes. Class label having maximum conditional probability is assigned to data instance. Naive Bayes assumes that attributes have no effect on each other that is they have independent distribution of values.

4. An artificial neural network is a set of neurons arranged according to a specific architecture. Neurons are processing elements that transform input given to network into an output. Each of the connections between the neurons of one layer to another layer is assigned a particular weight which signifies its importance. A bias is available for each unit in the hidden layer and the output layer which acts as a threshold to vary the activity of neurons in the network. Pattern Recognition NN (PRNN) used here for analysis is feed forward two layer network which is used for classification to classify samples into one of the target class labels. The number of neurons in hidden layer depends on number of input and output. One neuron in output is used to represent two target class labels in training data in form of 0(n) and 1(y).

5. Apache Hadoop is an open-source software framework used for distributed storage and processing of very large data sets. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are a common occurrence and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called Map-Reduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality – nodes manipulating the data they have access to – to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

3. EXISTING SYSTEM

Indian Institute of Tropical and Meteorology uses graph based approach for prediction of weather. Various graphs are plotted which may be 2D or 3D which consists the information of various parameters recorded by sensors. These parameters include information of Temperature, pressure height, speed of wind etc. The data is recorded at different location and height for more fine grained information.

Based on these plotted graphs the current system recognizes the trend of the weather. By applying various methods like regression and other methods weather prediction is done.

The current system used for plotting the graphs based on the collected information is sequential. Sequential method of graph plotting takes time to plot. All the collected data is stored in the single file. But when we are plotting a graph we need the data of specific two or three parameters only. But then also we need to process the data file completely. This processing of complete data file sequentially takes more time.

So, this sequential processing time can be reduced by applying High performance techniques. These techniques include parallel processing, distributed processing etc.

In this paper we are doing the comparative study of sequential, parallel and distributed processing. Accordingly we can use for plotting the graph of required parameters.

4. PROPOSED SYSTEM

The distributed Hadoop system can be used for pre-processing of data file which decreases the size of data file and this results into less processing time for graph plotting.

The Hadoop architecture for pre-processing is as follows:

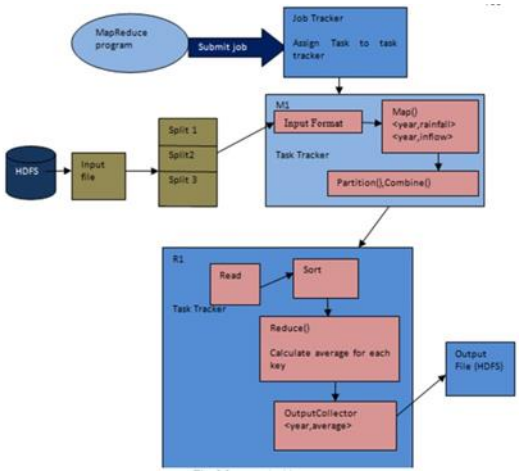


Fig – 1: System Architecture

After pre-processing the size of file is decreased and only required data is present in the data file. This is used for further processing so that the time requires for processing or graph plotting is very less.

The sample temperature versus pressure graph plotted sequentially based on data collected in one day 6th February 2013 is as follows:

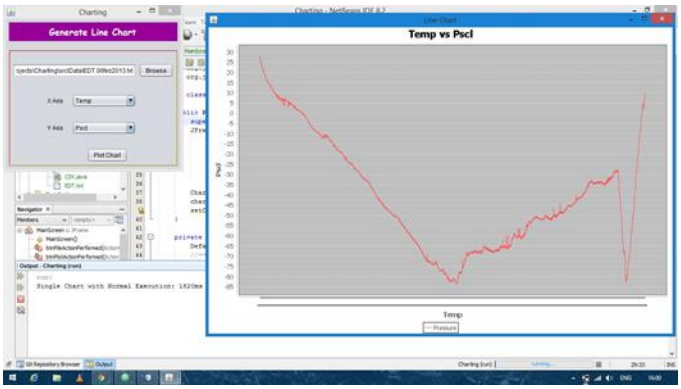
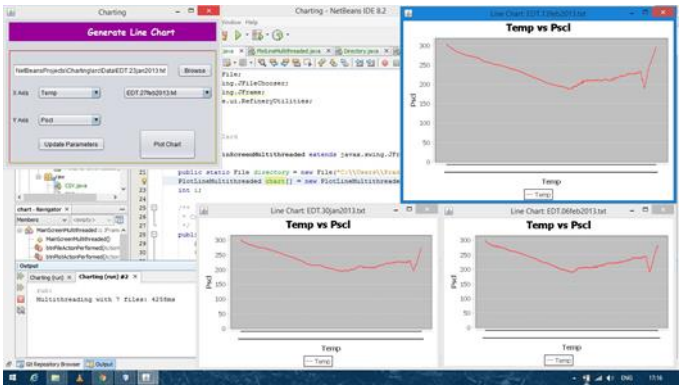


Fig -2: Sample temp Vs Pressure graph

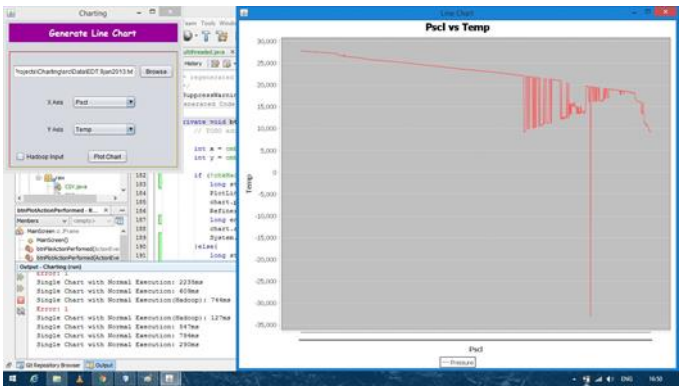
This sequential plotting requires 1820ms.

The parallel processed multithreaded program time for running and plotting the graph of 7 files simultaneously 4258ms.

These 7 files are of different days with same parameters. That is these files when processed in parallel they require less time as that of serial execution. The samples of some of these files and the execution time is Shown below:



When we pre-process the files using Hadoop the decreased file size decreases the processing time. The sample output screenshot is as follows:



From above screenshot we can say that the time required for sequential processing is maximum and parallel processing using thread is minimum. The Hadoop distributed architecture requires medium time, as file size were not large enough.

If file size is more we can prefer the Hadoop distributed approach over parallel or sequential.

4. CONCLUSIONS

From above comparison we can say that parallel or Distributed Hadoop approach is better than sequential approach at any time for data processing and graph plotting.

In between distributed and parallel processing, processing should be done depending upon the size of input dataset. If input data size is very large i.e. long term data analysis and prediction we should use Hadoop Distributed approach.

If file size is small i.e. short term analysis and prediction we can make use of threads (parallel processing) and can save even more time.

5. REFERENCES

[1] Depti Gupta and Udayan Ghose, A Comparative Study of Classification Algorithms for Forecasting Rainfall, IEEE 2015

[2] Lorenzo Luini and Carlo Capsoni, A Unified Model for the Prediction of Spatial and Temporal Rainfall Rate Statistics, IEEE Transactions on Antennas and Propagation, October 2013

[3] Wei Fang at al, Meteorological Data Analysis Using MapReduce, The Scientific World Journal February 2014.

[4] Brigitta Goger and Oliver Fuhrer, Current Challenges for Numerical Weather Prediction in Complex Terrain: Topography Representation and Parameterizations, IEEE 2016.

6. Authors

Shivraj Jawalge PCCOE Pune, Computer Engineering.

Shrikshel Boralkar PCCOE Pune, Computer Engineering.

Sagar Ingale PCCOE Pune, Computer Engineering.

Vijay Jadhav PCCOE Pune, Computer Engineering.