Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Meteorological Data Analysis Using HDInsight

Mugdha Kulkarni, Priyusha Nair, Shruti Kulkarni
Pune University, Maharashtra,
India

Abstract— Weather Prediction is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems across the world. As per survey weather reports generated are huge in amount and in unstructured format. There is need for analysis of real time weather data for giving predictions. Data mining is computer assisted process of digging through and analysing enormous sets of data and then extracting the meaningful data. In today's world Big Data processing is the need of an hour. Ability to represent and query data with little and no apparent structure arises in several fields. We propose a solution to this by providing dynamic data mining technique which gives real time weather forecast and also providing a hadoop cluster to it.

Keywords— Dynamic, Hadoop, Pig, Hive, Clustering, Azure.

I. INTRODUCTION

Meteorological data analysis is a system which considers real time data while making predictions and giving out weather forecasts. It should be scalable, portable and should work on variety of client systems.it should be able to handle Big Data and give outputs according to visualization effect entered by the end user. The developer must have greatest privilege over all other users, including the weather forecasting personnel and authenticated users. Till date, various weather mobile applications have been developed using clustering and regression algorithms, but real time analysis is still a big challenge. We will be providing a part solution of dynamic data mining along with accurate long term prediction.

II. ENHANCED K-MEANS

In this algorithm that involves clustering process, the input remains in the same order in which data items are entered .The whole process is divided into two phases.

A. Phase I

In phase I, the cluster size is fixed and the output of the first phase forms initial clusters. Here, the input array of elements is scanned and split up into sub –arrays, which represent the initial clusters. This phase includes following steps

1) Find the size of cluster Si (1= i = k) by Floor (n/k).
2) Where n= number of data points D_p (a1, a2, a3, …… an)
3) K= number of clusters.
4) Create K number of Arrays Ak
5) Move data points (D_p) from Input Array to A_k until
6) Si =Floor (n/k).
7) Continue Step 3 until all D_p removed from input array
8) Exit with having k initial clusters

B. Phase-II

In phase II, the cluster sizes vary and the output of this phase are the finalized clusters.Intial clusters are inputs for the phase. The centroids of these initial clusters are computed first, on the basis of which distance from other data elements are calculated. Furthermore the data elements having less or equal distance remains in the same cluster otherwise they are moved to appropriate clusters. The entire process continues until no changes in the clusters are detected. Steps involved in this phase are

1) Compute the Arithmetic Mean M of all initial clusters C_I
2) Set 1 j k
3) Compute the distance D of all D_p to M of Initial Clusters C_j
4) If D of D_p and M is less than or equal to other
5) distances of M_i (1 i k) then D_p stay in same cluster
6) Else D_p having less D is assigned to Corresponding C_i
7) For each cluster C_j (1 j k), Re compute the M and
8) Move D_p until no change in clusters.

III. DYNAMIC DATA MINING

Big Data Analytics is an emerging field of analytics which requires algorithms and mining methods to be implemented on huge amount of raw data.

To work with a large volume of data, we need to have time-bound and effective methods of both mining as well as preprocessing. As the data is increasing day by day, there is an ardent need of data mining to be dynamic. Within this too, insert updations are easier to accommodate than the delete and modify updations on a huge dataset, the size of which often reaches gigabytes. In our paper, we offer a partial solution to this problem by using a weighted table which calculates a unique parameter for every record. This helps us in pinpointing to the record which has been changed (updated/deleted) since the last time the database was used to analyze data or gather information from. However, the data must be clean in order to be suitable for this algorithm to work it on. The purpose of dynamic data mining technique is to find solution (Item Summation) that is able to take into consideration all updates (insert, update and delete modifications) into account.

IV. PREDICTION

As we will be implementing enhanced k-means and dynamic solution for our datasets, use of Naive Bayesian model and algorithm will be more feasible and easier than any other weather prediction methods. Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem (or Bayes' rule) with strong independence (naive) assumptions.

$$\text{Bayes' rule: } P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

The basic idea of Bayes' rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed. From Bayes' rule, we have

1) A priori probability of H or P (H): This is the probability of an event before the evidence is observed.
2) A posterior probability of H or P (H | E): This is the probability of an event after the evidence is observed.

A. Example

To predict the chance or the probability of raining, we usually use some evidences such as the amount of dark cloud in the area. Let H be the event of raining and E be the evidence of dark cloud, then we have

$$P(\text{raining} \mid \text{thundering}) = \frac{P(\text{thundering} \mid \text{raining}) \times P(\text{raining})}{P(\text{thundering})}$$

1) P (thundering | raining) is the probability that there is dark cloud when it rains. Of course, "thundering" could occur in many other events such as overcast day or forest fire, but we only consider "thundering" in the context of event "raining". This probability can be obtained from historical data recorded by some meteorologists.
2) P (raining) is the *priori* probability of raining. This probability can be obtained from statistical record, for example, the number of rainy days throughout a year.
3) P (thundering) is the probability of the evidence "thundering" occurring. Again, this can be obtained from the statistical records, but the evidence is not usually well recorded compared to the main event. Therefore, sometimes the full evidence, i.e., P (thundering), is hard to obtain.

V. DATA FLOW

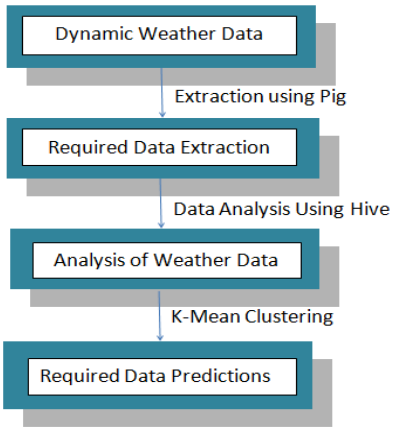

Fig.1 Data flow of proposed system.

Above figure represents data flow of the proposed system. As first step we will gather weather data with feasible attributes. Weather data has Synoptic data or climate data are the two classifications. Climate data is the official data record, usually provided after some quality control is performed on it. Synoptic data is the real-time data provided for use in aviation safety and forecast modelling. We will be using rainfall as parameter for further process. We will be using 62 years of rainfall data for around 28 regions. We cannot directly use any data without it being cleaned and pre-processed. We will be using Pig as a tool for ETL (extract, transform and load). Finally the extracted and transformed data is loaded onto the Hive for further processing. We will use clustering technique for data partition. After applying dynamic solution we will be implementing k-means algorithm for cluster formation and prediction.

VI. ADVANTAGES AND DISADVANTAGES

A. Advantages of Proposed System

1) Provide a more accurate weather forecast system.
2) To provide analysis of real time data.
3) System that provides good long term prediction.
4) It may be used as service in Azure.
5) Use of part solution will alter only one entry in the database instead of altering the whole database.

B. Advantages of Azure HDInsight

1) It offers all the advantages of Hadoop, plus the ability to integrate with Microsoft ecosystem.
2) Data upload/download speed is very high.
3) Azure also allows developers to produce applications using a variety of other languages like ASP, ASP.net, PHP, Python etc.

C. Disadvantages

1) Since Azure is not an open source tool, there will be addition of cost
2) Dataset may contain only a single attribute and size may vary that may affect accuracy for the overall system.
3) Need of internet access and also Azure account for the user.

VII. FUTURE SCOPE

A. Making Use of Proposed System for "SMART CITY" Applications

1) Can be used as an application for local trains in Mumbai to know weather conditions in rainy season.
2) For airport to manage daily flights.
3) Can be used by various meteorological departments in cities.
4) Wide use for agricultural growth of different variety of crops.
5) Can be widely used in harvesting different energy resources (e.g. Windmills).
6) Useful for construction of buildings depending upon weather conditions.

VIII. CONCLUSION

This paper presents a survey that involves use of data mining techniques and algorithms for weather predictions. Use of Enhanced k-means algorithm improves iterations and time complexity. Dynamic data mining algorithm uses a mathematical equation, special devised for the process of calculating a summation value. The summation value is unique for each record in the database. It is assumed that if the record is changed this summation value will change accordingly. Market needs accurate data which is possible by applying correct data mining technique. Still prediction will be a challenge as there is no accurate algorithm for it.Future work includes building of dynamic methods for faster and accurate data predictions.

ACKNOWLEDGMENT

Authors would like to take this opportunity to thank Mr. Pradeep Deshmukh and Prof.Swati Shekhapure for giving us all the help and guidance we needed. We are really grateful to them for their kind support and valuable suggestions.

REFERENCES

[1] Meghali A. Kalyankar, S. J. Alaspurkar, "Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), 114-118, February – 2013.
[2] Sarah N. Kohail, Alaa M. El-Hales, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJICT Journal Volume 1 No. 3, 2011
[3] Divya Chauhan, Jawahar Thakur, "Data Mining Techniques for Weather Prediction: A Review ", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 8, August 2014.
[4] University of Alberta, Osmar R. Zaïane, "Chapter I: Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1990.
[5] Badhiye S. S., Dr.Chatur P. N., Wakode B. V., "Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach", International Journal of Emerging Technology and Advanced Engineering, 2250-2459, Volume 2, Issue 1, January 2012.
[6] J. Han, M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

