METEOROLOGICAL DATA ANALYSIS USING HDINSIGHT WITH RTVS

Mugdha Kulkarni[1], Priyusha Nair[2], Shruti Kulkarni[3], Swati Shekapure[4]

[1] *Student, Computer Engineering, Pune University, Maharashtra, India*
[2] *Student, Computer Engineering, Pune University, Maharashtra, India*
[3] *Student, Computer Engineering, Pune University, Maharashtra, India*
[4] *Professor, Computer Engineering, 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 the computer assisted process of digging through and analyzing 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 have focused on three main algorithms that are important for prediction on any kind of data. 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. There is also need of more accurate predictions based on any type of dataset. We propose a solution to this by using R programming language for analysis of weather data using Microsoft Azure HDInsight for good long term predictions. This solution can also be as a SMART CITY application.

Keywords: *Azure, HDInsight, Hive, RTVS, R, Time-Series, Naive forecast, Shiny, Markdown*

----------***----------

1. INTRODUCTION

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories.Weather data is an important source for testing the results of our analysis in an accurate way. Microsoft Azure is a cloud platform, which provides services like HDInsight for all Hadoop related technologies. In this paper, we plan to build a system that uses real time weather data stored on a Hadoop cluster and this available data is cleaned, summarized and queried using hive on azure. This data is then analyzed based on the trend and other parameters using R Programming language. It provides a system that is scalable, faster and more over gives accurate and good long term predictions.

2. MICROSOFT AZURE HDINSIGHT

Azure HDInsight is an Apache Hadoop distribution powered by the cloud. This means that it handles any amount of data, scaling from terabytes to petabytes on demand. Spin-up any number of nodes at any time.

- A data lake service
- Scale to petabytes on demand
- Crunch all data-structured, semi-structured, unstructured
- Developing Java, .NET, and more
- Skip buying and maintaining hardware
- Spin up Apache Hadoop, Spark and R clusters in the cloud
- Connect on-premises Hadoop clusters with the cloud

2.1 Hive

Apache Hive is a data warehouse system for Hadoop, which enables data summarization, querying, and analysis of data by using HiveQL. Hive can be used to interactively explore your data or to create reusable batch processing jobs.

Microsoft Azure provides Hive editor through cluster dashboard by providing cluster credentials. All Hive queries can be performed and submitted and results can be seen through job history. Table location and directories can be viewed though file browser.

The same can also be performed using Visual Studio 2015. By downloading Microsoft Azure SDK you can create Hive applications which allows you to query on the dataset loaded on HDInsight cluster. This is the done by configuring the name of the HDInsight cluster in the .hql file.

3. R TOOLS FOR VISUAL STUDIO (RTVS)

- R aware editor with IntelliSense
- R Interactive Window with multi-line editing
- Debugger with Locals and Stack views
- Plots and Shiny apps
- Support for CRAN R, Microsoft R open, Microsoft R Server
- Free and Open Source

3.1 Time Series

The function ts is used to create time-series objects. These are vector or matrices with class of"ts" (and additional attributes) which represent data which has been sampled at equispaced points in time. In the matrix case, each column of the matrix data is assumed to contain a single (univariate) time series. Time series must have at least one observation, and although they need not be numeric there is very limited support for non-numeric series

Time Series algorithm is used for trend analysis of the given data. It creates time series object that can be used as an input for the next forecasting algorithms.

3.2 K-Means

The data given by x are clustered by the *k*-means method, which aims to partition the points into *k*groups such that the sum of squares from points to the assigned cluster centers is minimized. At the minimum, all cluster centres are at the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre).

K-Means allows to cluster all items with similar properties which helps to study different behavior or patterns in data. In this case it helps us to understand the rainfall trend in data.

3.3 Naïve Forecast

naive() returns forecasts and prediction intervals for an ARIMA(0,1,0) random walk model applied to x. snaive() returns forecasts and prediction intervals from an ARIMA(0,0,0)(0,1,0)m model where m is the seasonal period.

Naïve method takes two inputs (tsobject of dataset, h)
Where tsobject is the time series object of the datset and h is the prediction period depending upon the value given by the user.

3.4 Classification

Classification allows us to classify the value predicted by the forecast function in user friendly format. Based on the cluster centers the predicted value is compared and classified. For example, consider weather dataset. The value obtained after naïve forecast method on this dataset is compared to its cluster centers and classified as sunny, rainy, cloudy etc.

3.5 Example

Consider the Air Passengers dataset available on R.
We can create a Time Series object of this dataset by using the following script.

- tsObj<-ts(Air Passengers)
- plot(tsObj)

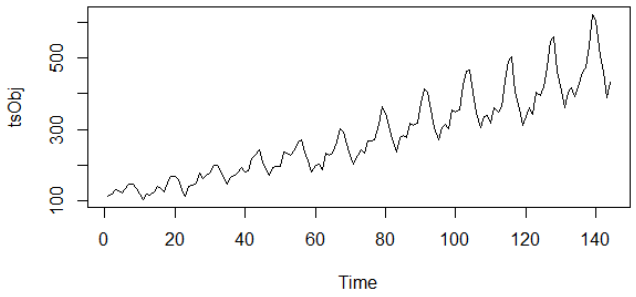

Fig-1: Time Series Plot

K-Means Clustering for this dataset can be done as follows.

- kmeasOut<-kmeans(tsObj,4)
  where 4 is number of clusters.

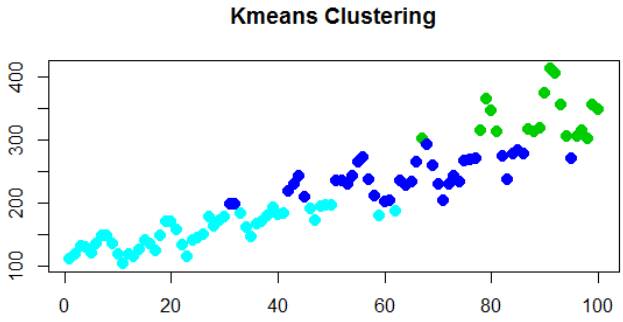

Fig-2: K-Means Plot

Naïve forecast method to predict future rainfall.

- NaiveOut<-rwf(tsObj, h=10)
- Plot(NaiveOut)

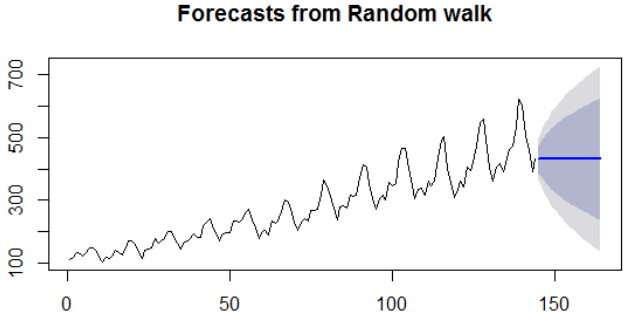

Fig-3: Naïve Forecast Plot

4. ARCHITECTURAL DESIGN

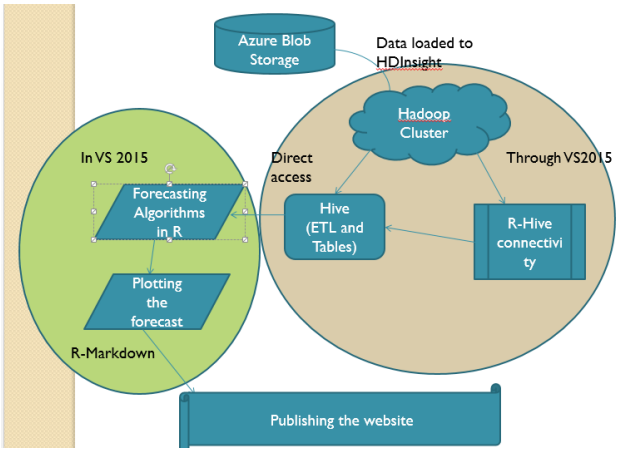

Fig-4: Architecture diagram

Brief about Architecture Design: As first step, we will upload dataset on Azure through Azure blob storage from local file System. Next, we will connect HDInsight cluster with Visual Studio 2015 for accessing hive tables and operations. Next we will make Hive- R connection through ROBC connections. Once connection has been established, we can read dataset from hive into R and then perform further operations and analysis on R.We will be using three algorithms in R for forecasting purpose. The very first and important algorithm is Time series that is used for analyzing the trend in the given data its object known as time series object is used an input for other algorithms. Next, is naïve method used for forecasting of weather based on the input of years you want the forecasting of! Then, after naïve we will be doing clustering and classification so the final output is understandable by the user.

5. FUTURE SCOPE

- Can be used for any dataset provided the API is available.
- API can be created using API Mgmt. Service in MS Azure.
- Important industry applications like customer feedback analysis, sales prediction; academic applications like determining the trend in education system etc. can use this service engine.
- We can pipeline different datasets like agricultural, share market data etc. along with this proposed system to make various SMART CITY applications.

6. CONCLUSION

- Market needs accurate data which is possible by applying correct data mining technique.
- As MS Azure is a recent platform, any new R&D is a surge to giving customers new services in a user-friendly way and on a software like Microsoft.
- With this proposed system, we can provide more accurate forecasting system that can used for various other prediction purposes.

ACKNOWLEDGEMENT

Authors would like to take this opportunity to thank Mr. Pradeep Deshmukh for giving us all the help and guidance we needed. We are really grateful to him for his 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]. 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.

[4]. J. Han, M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000