International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Survey on Application of Big Data in weather monitoring system

Shivam Patil[1], Vanita S. Babanne[2]

[1]Student Third year Computer Engineering R.M.D. Sinhgad School of Engineering, Warje, Pune, Maharastra
[2]Asst. prof of Computer R.M.D. Sinhgad School of Engineering, Warje, Pune, Maharashtra

---

Abstract - *Weather forecasts are made by collecting as much data as possible about the current state of the atmosphere to determine how the atmosphere evolves in the future. To handle such humongous data – "Big Data" is introduced. Big Data has become an imminent part of all industries and business sectors today. we propose a Pre-Processing Framework to address quality of data in weather monitoring. Hence, it is imperative to improve Data quality even it is absorbed and utilized in an industry's Big Data system. In this paper, we propose a Pre-Processing Framework to address quality of data in a weather monitoring and forecasting application that also takes into account global warming parameters and raises alerts/notifications to warn users and scientists in advance.*

Key Words: **Big Data, Pre-Processing, Data Quality**

MOTIVATION*—We have conceptualized a Weather Monitoring and Forecasting Application to raises alerts/notifications to warn users and scientists in advance.*

1. INTRODUCTION

Big data is data that has greater variety arriving in increasing volumes and with higher velocity. This is called the three Vs. Big data is larger, more complicated data sets, especially from new data sources. These data sets are so humongous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you have been able to tackle before.

- Volume defines the large amount of data that is produced each day, for example. The generation of data is so large and complex that it can no longer be saved or analyzed using traditional data processing methods.
- Variety refers to the different of data types and data sources. 80 percent of the data in the world today is unstructured and at first glance does not show any indication of relationships. With the help of Big Data such algorithms, data is able to be sorted in a structured manner and examined for relationships. Data does not always comprise only traditional datasets, but also images, videos and audios.
- Velocity refers to the speed with which the data is produced, analyzed and reprocessed.

2. LITERATURE SURVEY

Year	Author	Objective	Methodology	Conclusion
IEEE/ 2015	Ikbal T. , Rachida D. & Mohamed A.S.	Solving the numerous data quality Issues that occur when attempting to apply data quality concepts at large data sets	Data cleansing: The process of finding and correcting errors	We demonstrated that using our data quality Selection framework helps gaining time and Resources

IEEE/ 2015	I.A.T. Hashem ,I. aqoob, N.B.Anuar, S. Mok-htar, A. Gani, & S. Ullah	The goal of this study is to implement a comprehensive investigation of the status of big data in cloud computing environments and provide the definition, characteristics, and classification of big data.	Cloud computing is a technology that uses the internet and servers to maintain large data and applications.	We discussed Hadoop architecture components such as MapReduce and HDFS.
IEEE/ 2014	Nang Tang	The aim is to report on an ongoing investigation into software productivity and its influencing factors	The data set contained a large number of cases with zero values and values which seemed to be the result of incorrect data.	It was recognized that the confidential nature of the data and concerns into the quality Of the data need to be addressed.
IEEE/ 2014	Divya Tomar and Sonali Agrawal	To obtain Data quality Before applying data mining techniques to get useful knowledge	Data mining is searching for hidden, valid, and potentially useful patterns in huge data sets.	It starts with Preprocessing techniques which includes detailed description of different data cleaning.
IEEE/ 2014	C-Y, Zhang and C.L Philip Chen	To demonstrate an immediate view about Big Data, including Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies.	Quantum computing is an area of computing focused on developing computer technology which is based on the principles of quantum theory, which explains the behavior of energy and material on the subatomic and atomic levels	we currently adopt to deal with the Big Data problems. We also discuss different underlying methodologies to handle the data flood, for example, granular computing, cloud computing, bio-inspired computing, and quantum computing

IEEE/ 2014	H. Hu, Y. Wen, T.-S Chua and X.Li	To provide an overall picture for amateur readers and in still a do-it-yourself spirit for advanced audiences to customize their own big-data solutions.	The chain of big data value consists of four phases: data generation, data acquisition, data storage, and data analysis.	we have presented the concept of big data and highlighted the big data value chain, which covers the entire big data lifecycle

3. LIVE SURVEY

1. Esri partner Weather Decision Technologies, Inc. (WDT), uses advanced GIS from Esri to better organize and analyze this big data. WDT provides weather monitoring and mapping services to different companies: energy corporations to help them predict electrical outages and keep offshore oil rigs safe; agriculture agencies for crop insurance; freight transportation industries to aid with route design; and concert and sporting event organizers for planning and safety.

"The amount of data that we collect is humongous, about one terabyte per day," said Matt Gaffner, GIS solutions expert at WDT. "Over the years, we have assembled an archive of almost half a petabyte of weather data."

2. The dataset used for this analysis is obtained from Central Research Institute for Dryland Agriculture (CRIDA) which is a National Research Institute under the Indian Council of Agricultural Research (ICAR). This dataset is large enough that contain weather data from the year 1958–2014, nearly 56 years of data with 20,820 rows (days) and attributes like temperature, humidity, wind speed, sunshine, etc. The temperature attribute is taken for examination, and it was undergone for predicting the forecast value and compared with the available real-time data, i.e., the average temperature from the year 1958–2013 is used as input for the forecasting algorithms and forecasts the temperature of the year 2014. Real-time temperature (in Celsius) of the year 2014 is already available in the dataset and hence it was compared with the value predicted by the algorithms and the percentage of accuracy is determined. The data is collected from sensors so it is not purely structural data, and hence it was converted to comma-separated file (CSV) for easy access. The implementation is done using R programming which contains various packages for statistical analysis and techniques. The dataset is undergone for few steps of preprocessing to apply the data in R functions. The preprocessing details are given below:

i) The average temperature was read in a variable as a dataframe. This dataframe is converted to DATE type to make time series conversion easier.

ii) Convert this dataframe to zoo object by using zoo() function in zoo package (to be downloaded and installed).

iii) The time series object can be acquired using ts() in zoo package. The reason to convert the dataframe to zoo object followed by time series object without converting directly to time series is that compatibility errors may occur in direct conversion because of the frequency exceeding beyond 12. This time series object can be undergone with forecast predictions using forecast functions like rollmean() for moving average, HoltWinters(), etc. The forecast functions are built-in functions in R which contains several parameters that can be set up as per the need of the research. In our implementation, every 18 years data are separately analyzed as individual phase to study the behavior of the data pattern at each stage and the entire dataset temperature value is taken for prediction.

4. ALGORITHMIC SURVEY

1. Support Vector Machine is machine learning algorithm that can be employed for both classification and regression purposes. SVMs are based on the idea of finding a hyper plane that best divides a dataset into two classes. SVMs accuracy level is high as compared to other machine learning algorithms. They are more efficient because it uses a subset of training points. We are using SVM algorithm to classify the weather parameters to predict the rainfall percentage. The basic idea of SVM is applied to binary classification. Most of the previous approaches are the method that decomposes a multiclass problem into

multiple independent binary classification tasks. In practice, these methods usually bring about the inseparable cases which will reduce the accuracy of classifications. After obtaining the input data and applying the preprocessing method, the classification or prediction system can be used. The core of the forecasting system is based on support vector machine (SVM). Each instance in the training phase contains one "target value" and several "attributes". The objective of SVM is to generate a sculpt (based on the training data) which predicts the aim values of the test data given only the test data attributes.

2. IoT cloud platform is designed to store and process Internet of Things (IoT) data. This platform is built to take massive volumes of data generated by devices, sensors, applications, websites and initiate actions for real time responses.

5. CONCLUSION

This paper evaluated the case Study of the Weather Application and attempted to use the Big Data gathered from multiple sources to design a system capable of forecasting weather based on recent global warming concerns. Big Data is a science and process depending on many technologies and is still in an evolving phase. However, it emphasizes the importance of addressing Data in a Big Data system within the early stages to magnify its relevance. This can indeed enable and prepare organizations to take a leap forward in their growth and future strategies.

REFERENCES

[1] Ikbal Taleb, Rachida Dssouli and Mohamed Adel Serhani, "Big Data Pre-processing: A Quality Framework" in 2015 IEEE International Congress on Big Data (Bigdata Congress),2015.

[2] I.A.T. Hasem,I. Yaqoob, N.B Anuar, S.Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing : Review and open research issues,"inf.Syst,vol. 47, pp. 98-115,2015

[3] N. Tang, "Big Data Cleaning," in Web Technologies and Applications, L. Chen, Y. Jia, T. Sellis, and G. Liu, Eds. Springer International Publishing, 2014, pp. 13–24.

[4] Tomar, Divya and Sonali Agarwal. "A survey on pre-processing and postprocessing techniques in data mining." International Journal of Database Theory & Application 7.4(2014)

[5] C.L Philip Chen and C-Y, Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Inf. Sci.,vol. 275, pp.314-347,2014

[6] H. Hu, Y. Wen, T.-S Chua and X.Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," IEEE Access, vol. 2,pp.652687,2014