ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

A Multilevel Procedural Big Data analysis of Weather Prediction

Garg Avanti [1], Saurabh Singh [2]

M. Tech Scholar, Jaipur Engineering College, Jaipur, India[1]

Assistant Professor, Jaipur Engineering College, Jaipur, India [2]

ABSTRACT: One of the difficulties involves figuring out how to manage this new information composes and deciding which data can possibly give value to your business. It is not simply access to new information sources, chosen occasions or exchanges or blog entries, however the examples and entomb - connections among these components that are of intrigue. Gathering heaps of differing kinds of information rapidly does not make esteem. You require examination to reveal experiences that will help your business. That is the thing that this paper is about.

KEYWORDS: dataset, weather dataset

I. Introduction

Data mining is experiencing a critical move in the volume, assortment, esteem and speed of data expanding essentially every year. The volume of data made is outpacing the measure of as of now Data to such an extent, to the point that most associations don't recognize what esteem is in their data. At a similar that time data mining is changing, equipment abilities have likewise experienced sensational changes. Similarly as data mining isn't a certain something yet an accumulation of numerous means, speculations, and calculations, equipment can be dismembered into various parts. The comparing part changes are not generally in a state of harmony with this expanded request in data mining, machine learning, and huge logical issues. The four parts of circle, memory, focal preparing unit, and system can be thought of as four legs of the equipment stage stool. To have a helpful stool, every one of the legs must be of a similar length or clients will be disappointed, stand up, and leave to locate a superior stool; so excessively should the equipment framework for data mining be in adjust concerning the segments to give clients the best understanding for their expository Data Quality Control.

Information Warehouse Management Tools are customizing applications that focus and change information from operational structures and weights it into the information distribution center.

The area of information distribution center organization is particularly mind boggling as information got from operational sources, for instance, those information beginning from esteem based business programming courses of action like Supply Chain Management (SCM), Point of Sale, Customer Serving Software and Enterprise Resource Planning (ERP) and organization programming to encounter the ETL (remove, change, stack) process.

To encourage data around the data warehouse, proficient ETL devices ought to be utilized. Organizations may either need to purchase outsider instruments or build up their own particular ETL devices by allocating their in-house software engineers to carry out the activity. As a rule, the general guideline is that the more mind boggling the data change necessities are, the more worthwhile it is to simply buy outsider ETL devices.

II. Foundation and need for big data analytics

Capacity and recovery of huge measure of organized and also unstructured data at an alluring time slack is a test. Some of these confinements to deal with and process tremendous measure of data with the customary stockpiling strategies prompted the rise of the term Big Data. In spite of the fact that big data has picked up consideration because of the development of the Internet, however it can't be contrasted and it. It is past the Internet, however, Web makes it simpler to gather and offer learning also data in crude shape. Big Data is about how these data can be put away,

DOI: 10.15680/IJIRCCE.2018.0602018

prepared, and grasped to such an extent that it can be utilized for foreseeing the future game-plan with an awesome exactness and worthy time delay.

Advertisers center on target promoting, protection suppliers center around giving customized protections to their clients, and medicinal services suppliers center around giving quality and ease treatment to patients. In spite of the headways in data stockpiling, accumulation, examination and calculations identified with anticipating human conduct; it is essential to comprehend the hidden driving and additionally the directing variables (advertise, law, social standards and design), which can help in creating hearty models that can deal with big data but then yield high forecast precision (Boyd and Crawford, 2011).

The present and rising focal point of big data examination is to investigate customary strategies, for example, control based frameworks, design mining, choice trees and other data mining procedures to create business administers even on the extensive data sets proficiently. It can be accomplished by either creating calculations that utilizations circulated data stockpiling, in-memory calculation or by utilizing group registering for parallel calculation. Prior these procedures were done utilizing framework figuring, which was surpassed by distributed computing as of late.

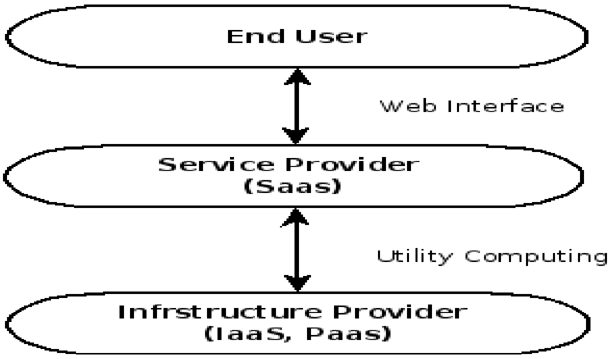


Fig 1: Infrastructure provider

III. TOWARDS DEVELOPING BIG DATA VALUE CHAIN

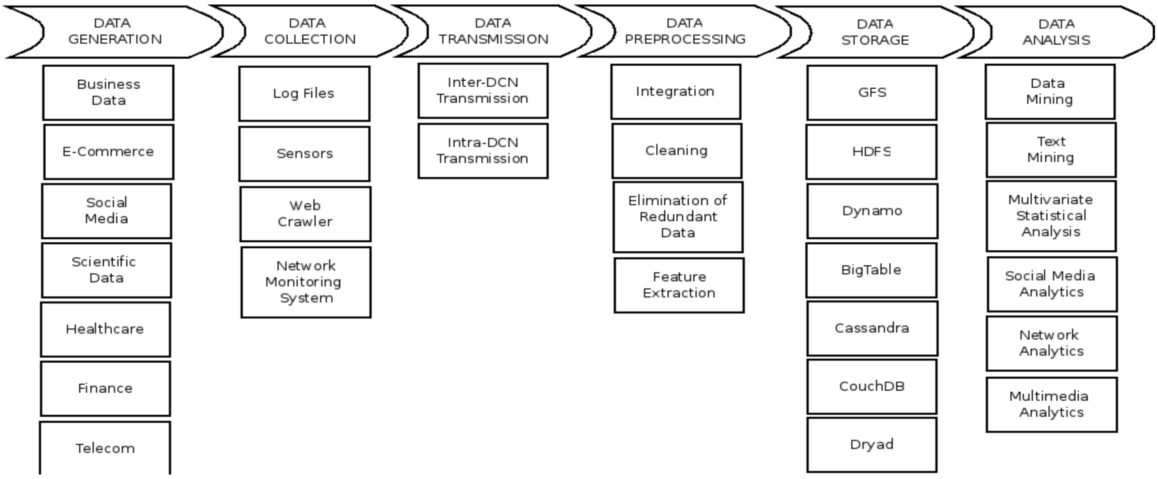


Fig 2 : Developing big data value chain

1) Data Generation: The above all progression the big data regard chain is the time of data. As inspected in the past territory, data is delivered from various sources that consolidate data from Call Detail Records (CDR), web diaries, Tweets and Facebook Page.

2) Data Collection: In this stage, the data is gotten from each and every possible datum sources). For instance, to suspect the customer beat in Telecom, data can be gotten from CDRs and slants/grievances of the customers on Social Networking Sites, for instance, Twitter (as tweets) and Facebook (appraisals shared on the association's Facebook page). The most regularly used procedures are log records, sensors, web crawlers and framework watching programming

3) Data Transmission: Once the data is accumulated, it is traded to a data storing and getting ready establishment for moreover dealing with and examination. It should be possible in two phases: Inter-Dynamic Circuit Network (DCN) transmission and Intra-DCN transmissions. Between DCN transmission deals with the trading of data from the data source to the data center while the last associates in the trade inside the data center. Beside limit of data, data center aides in social event, organizing and directing data.

IV. Problem With The Existing Work

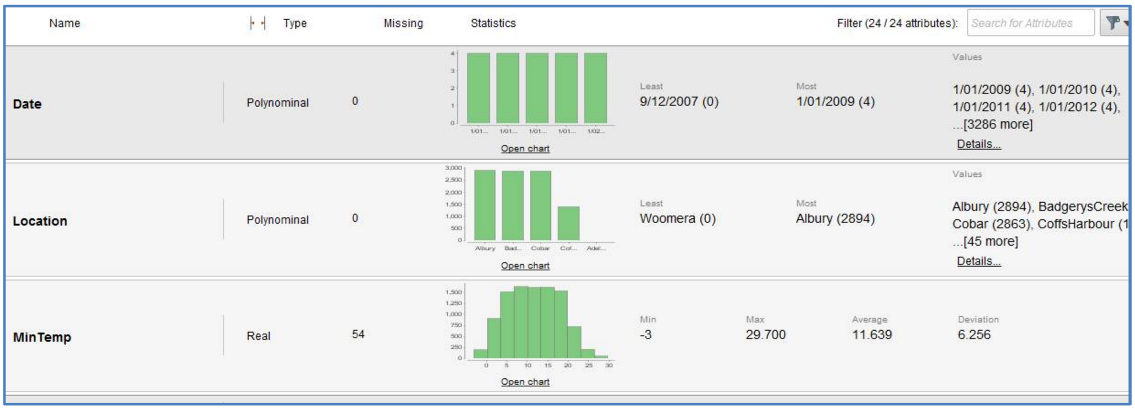
In previous work we saw that many tools use the basic statistics analysis to produce the result. They don't provide the wide range of variety of data analysis. In previous tools they worked to produce the result and compare the results from different tools or compare the performance or accuracy of the tools and methodologies or compare any factors among different algorithms. Their focus was on the result and its comparison. When data is not well understood, the result will not be more accurate. And to understand the result or produce more accurate result we must understand the data in depth. We must analyze the data fields and their dependencies factors on which result depends.

In our work we focused on the data field analysis. We analyzed the dependent fields or other fields in depth. We calculated the min values, maximum values of fields, their average and deviations. We generated the graph to understand the dataset graphically. It gives the visual understanding of data statistics. It would be very simple and easy to understand the real status of values in dataset.

V. Experimental Results

Here we can see the type of field (Polynominal/real/Integer), Missing values(if any) in the data field, statistics (using graphs) which contain Least value and most value. The least value denotes the frequency of the value which occurred less in the whole dataset and Most value denotes the frequency of the value which occurred most in the whole dataset. And Values show the other values of the field.

Some field which are real type, don't have the Least and Most value rather than they have Min and Max values. This type of field also have the average and deviation values.

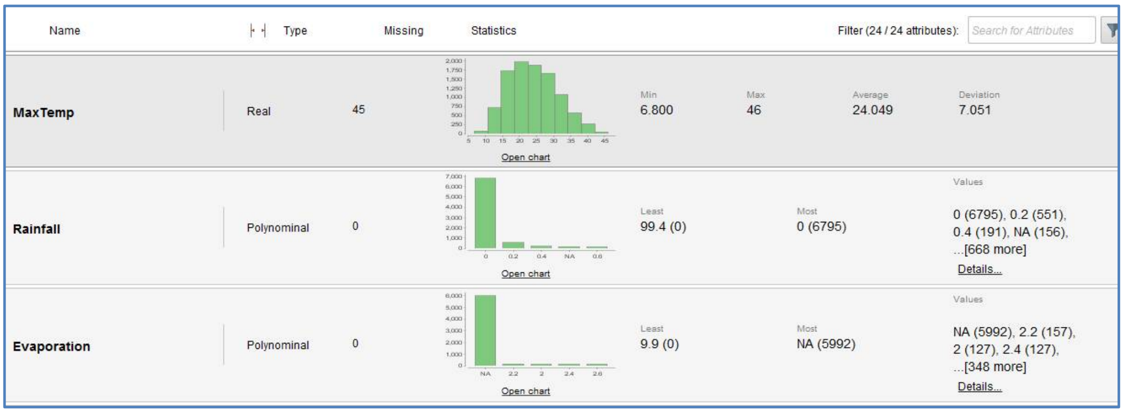


Filter (24 / 24 attributes): Search for Attributes

Name	Type	Missing	Statistics				
Date	Polynominal	0	Open chart	Least 9/12/2007 (0)	Most 1/01/2009 (4)		Values 1/01/2009 (4), 1/01/2010 (4), 1/01/2011 (4), 1/01/2012 (4), ...[3286 more] Details...
Location	Polynominal	0	Open chart	Least Woomera (0)	Most Albury (2894)		Values Albury (2894), BadgerysCreek Cobar (2863), CoffsHarbour (1 ...[45 more] Details...
MinTemp	Real	54	Open chart	Min -3	Max 29.700	Average 11.639	Deviation 6.256

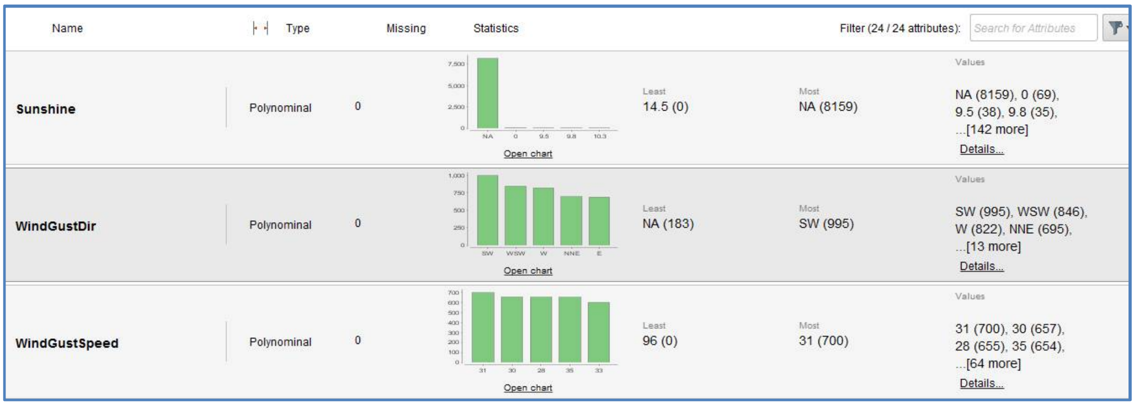
Figure 3 Date, Location and Min_Temp Statistics

DOI: 10.15680/IJIRCCE.2018.0602018



Name	Type	Missing	Statistics			
MaxTemp	Real	45	Min 6.800	Max 46	Average 24.049	Deviation 7.051
Rainfall	Polynominal	0	Least 99.4 (0)	Most 0 (6795)	Values 0 (6795), 0.2 (551), 0.4 (191), NA (156), ...[668 more]	
Evaporation	Polynominal	0	Least 9.9 (0)	Most NA (5992)	Values NA (5992), 2.2 (157), 2 (127), 2.4 (127), ...[348 more]	

Figure 4 Max_temp, Rainfall and Evaporation Statistics



Name	Type	Missing	Statistics		
Sunshine	Polynominal	0	Least 14.5 (0)	Most NA (8159)	Values NA (8159), 0 (69), 9.5 (38), 9.8 (35), ...[142 more]
WindGustDir	Polynominal	0	Least NA (183)	Most SW (995)	Values SW (995), WSW (846), W (822), NNE (695), ...[13 more]
WindGustSpeed	Polynominal	0	Least 96 (0)	Most 31 (700)	Values 31 (700), 30 (657), 28 (655), 35 (654), ...[64 more]

Figure 5 Sunshine, WindGustDir and WindGustSpeed Statistics

VI. Conclusion and Future Work

We use the Weather data as a dataset. In this dataset, there are various fields or factors on which weather depends. The class variable of fields or factor is rain tomorrow is to be determined. For "Rain Tomorrow" field all dependent fields are listed. To analyze the result and for statistics analysis, we use a tool named as Rapid Miner. This tool provides a wide range of data analysis or data statistics in graphical manner or text annotations.

REFERENCES

1. "The World's Technological Capacity to Store, Communicate, and Compute Information". *MartinHilbert.net*. Retrieved 13 April 2016.
2. *boyd, dana; Crawford, Kate (21 September 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.*doi:10.2139/ssrn.1926431.
3. "Data, data everywhere". *The Economist. 25 February 2010*. Retrieved 9 December 2012.
4. "Community cleverness required". *Nature. 455 (7209): 1. 4 September 2008.*doi:10.1038/455001a. PMID 18769385.
5. *Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". Science. 331 (6018): 703–5.*doi:10.1126/science.1197962. PMID 21311007.
6. *Hellerstein, Joe (9 November 2008).* "Parallel Programming in the Age of Big Data". *Gigaom Blog.*
7. *Segaran, Toby; Hammerbacher, Jeff (2009).* Beautiful Data: The Stories Behind Elegant Data Solutions. *O'Reilly Media. p. 257.* ISBN 978-0-596-15711-1.
8. *Hilbert, Martin; López, Priscila (2011).* "The World's Technological Capacity to Store, Communicate, and Compute Information". *Science. 332 (6025): 60–65.* doi:10.1126/science.1200970. PMID 21310967.
9. "IBM What is big data? – Bringing big data to the enterprise". *www.ibm.com.* Retrieved 26 August 2013.

10. *Reinsel, David; Gantz, John; Rydning, John (13 April 2017).* "Data Age 2025: The Evolution of Data to Life-Critical" *(PDF). seagate.com. Framingham, MA, US:* International Data Corporation. Retrieved 2 November 2017.
11. Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012
12. *Jacobs, A. (6 July 2009).* "The Pathologies of Big Data". *ACMQueue.*
13. *Magoulas, Roger; Lorica, Ben (February 2009).* "Introduction to Big Data".*Release 2.0. Sebastopol CA: O'Reilly Media (11).*
14. *John R. Mashey (25 April 1998).* "Big Data ... and the Next Wave of InfraStress" *(PDF). Slides from invited talk. Usenix.* Retrieved 28 September2016.
15. Amar Singh, "Spectre of Cyberterrorism: A Potential Threat to India's National Security," Indian Journal of Research, vol. 5, no. 3, 2016.
16. Neesha Jothia, Nur'Aini Abdul Rashidb, Wahidah Husainc, a*, ― Data Mining in Healthcare – A Review, Procedia Computer Science 72 (2015) 306 – 313
17. Blessing Ojemea*, Audrey Mboghob, ― Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders‖ Procedia Computer Science 96 (2016) 1294 – 1303
18. Lakshmi.B.Na*, Dr.Indumathi.T.Sb, Dr.Nandini Ravic, ― A study on C.5 Decision Tree Classification Algorithm for Risk Predictions during Pregnancy, Procedia Technology 24 (2016) 1542 – 1549
19. Btissam Zerhari1, Ayoub Ait Lahcen1,2, Salma Mouline1,‖ Big Data Clustering: Algorithms and Challenges‖, International Conference on Very large Databases, pp -487-499, 1994
20. R.Agrawal, R.Srikant, ―Mining Sequential Patterns‖, The 11th International conference on Data Engineering, pp-3-14, 1995
21. S.R.Jang, C-T.Sun, ―Neuro-Fuzzy and Soft Computing‖, ISBN-978-81-203-2243-1, PHI, 2011
22. T P Hong, K Y Lin and S L Wang, ―Fuzzy Data Mining for Interesting Generalized Association Rules‖, Fuzzy Sets & Symbols, Elsevier pp-255-269 2002
23. Cai, ―Mining Association Rules with Weighted Items‖ International Database Engineering and Applications Symposium 1998