International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Pre-Processing of Data to achieve Data Quality in Weather Monitoring App

Sonali Thosar, Bhagyashree Bhoyar, Tejaswini Patil

Sonali Thosar, Assistant Professor, DIT Pimpri, Pune, India
Bhagyashree Bhoyar, Assistant Professor, DIT Pimpri, Pune, India
Tejaswini Patil, Assistant Professor, DIT Pimpri, Pune, India

---***---

Abstract – *Big data plays crucial role in business sectors and industries today. Sectors like banking, retail, hardware, software, networking produce tremendous amount of data. This produced data is of heterogeneous type. To grow business this produced data can be used. This quantum of heterogeneous data can be used by acquiring patterns. i.e. Large amount of data is mined, processed, and then analyzed accurately to get useful patterns. Big data is used to acquire process and analyze heterogeneous data to achieve useful results. Quality of information depends on size of data, and type of data generated. Quality of data generated has crucial importance. Pre-processing stage is used to improve quality of data. We are investigating process like Cleansing to fix as much information as achievable, Noise channels to evacuate terrible information, too sub-forms for Integration and Filtering alongside Data Transformation or normalization. We assess and profile the Big Data during procurement stage, which is adjusted to desires to maintain a strategic distance from cost overheads later while additionally improving and prompting exact information investigation. Consequently, it is basic to improve Data quality even it is consumed and used in an industry's Big Data framework. In this paper, we propose a Pre-Processing Framework to address nature of information in a climate observing and estimating application that likewise considers a dangerous atmospheric global warming parameter and raises warnings to caution clients and researchers ahead of time*

Key Words: Big data quality, Pre-processing techniques, Big Data

1.INTRODUCTION

Big data is an advancing stage which implies huge volumes of both organized, semi-organized and unstructured information that represent a troublesome assignment to be prepared utilizing conventional strategies and databases. It is a methodology for educated dynamic utilizing investigative procedures to portray any informational index that is sufficiently enormous that requires the utilization of elevated level programming ability and philosophies to make the information into an advantage for an association.[1,2,]

Key dimensions of big data are Completeness, Timeliness, Conformity, Integrity, Consistency and Accuracy.

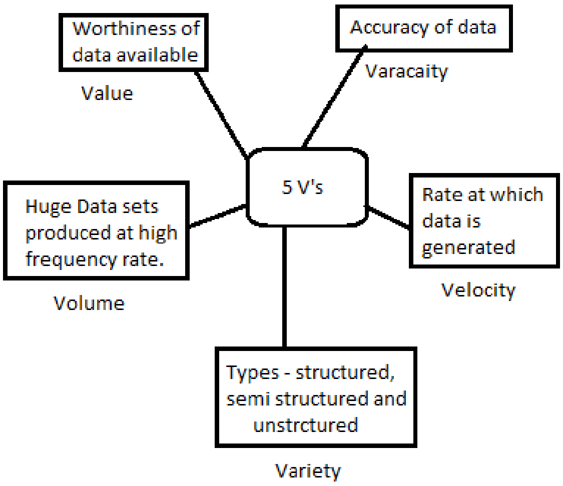
Worthiness of data available
Value
Accuracy of data
Varacaity
5 V's
Huge Data sets produced at high frequency rate.
Volume
Rate at which data is generated
Velocity
Types - structured, semi structured and unstrctured
Variety

Fig -1: 5V's in Big Data

Data traverse through 4 phases in Big Data System which are -1) Data Origin Identification, 2) Data Acquisition and Cleansing, 3) Data Aggregation and Storage and 4) Data Analysis. These phases are called as Big Data Lifecycle.

Data Origin Identification phase is related with the raw data being generated from a variety of sources. The sources are from different platforms like social sites, financial applications, customer relation applications, media web sites, images, etc. It is very difficult to understand the sources of the data.[3,4]

Information Acquisition and Cleansing stage acclimatizes the information from numerous sources. This crude information might be wrecked with irregularities.

These pre-preparing steps are urgent to change the information to levels reasonable or important for examination. Data should be filtered and reformed as it is from structured, semi structured, and unstructured sources. Pre-processing steps have prime importance to transform the data to levels suitable or valuable for analysis. Heterogeneous data should be aggregated with joins across source databases is responsibility of Data aggregation and storage phase.Last phase is Data Analysis phase. It infuses sense and relevance into combined data. This process is executed by comparing data characteristics to identify patterns.

1.1 DATA QUALITY

Data should be continuously monitored and tracked to verify its quality for analysis purpose. A great deal of information is aggregated during the business lifecycle making information that could be caught as unstructured, lacking nature of wanted parameters. Business can be industrial sectors, energy, power, retail, or e-commerce.

Each sector is specific to its domain and has its own data characteristics. Data should be useful to the users. Specific domains data must be checked for its domain making. This paper focuses on the Data Quality in early stages and analysis in early stages. Data quality can be failed using manual process. Maintain quality data is challenging. Timing, data type, format, structure is considered while considering quality of data. [8]

Data Quality depends upon its source. To build data quality framework several dimensions should be considered.

There are no standards for dimensions of data quality. Some dimensions are commonly accepted like consistency, accuracy of data, completeness, and validity.

Table – 1 Big Data Pre-processing Phases in weather monitoring App

Big Data Preprocessing	
Phases	**Description**
Data Consolidation and Integration	Integration of data from multiple sources. Type of data can be structured, Semi Structured, Unstructured.
Data Enhancements and Enrichment	create fused data that is Enriched with more information and possibly also enhanced qualitatively.
Data transformation	Capturing data from multiple sources, data is reformatted, normalized, aggregated using regulatory standards.
Data reduction	reducing the amount of data so that it becomes non-redundant
Data discretization	extracts and segregates data into intervals
Data Cleansing	Process to improve Quality of data

2. PROPOSED SYSTEM

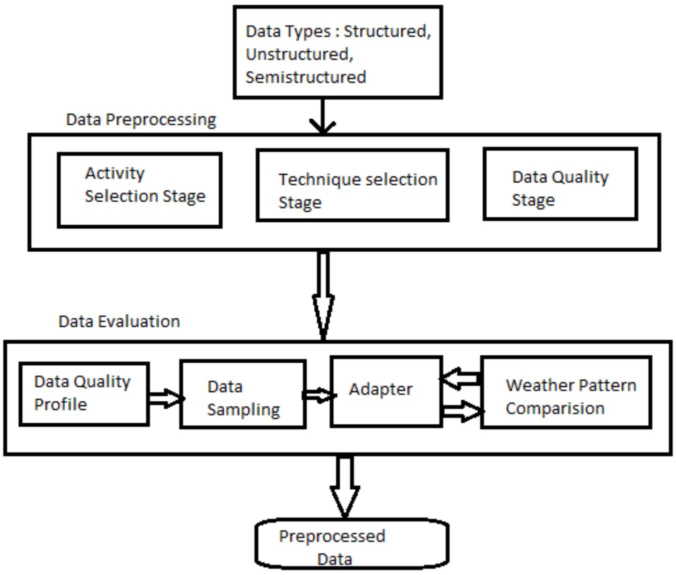

Fig -2 Preprocessing Framework for Weather Forecasting

In this application, data is captured from many sources like IoT sensors like weather parameters – rain, humidity, temperature humidity, longitude, and latitude. The application likewise expects information on information point varieties in temperature parameters according to a dangerous atmospheric deviation from different geological areas. It captures data for last decades which is average data. It also captures data for change in temperature above 1 degree Celsius. Change in the temperature triggers alarm.

Most organizations generate a huge quantum of data generated over time and within multiple projects. This huge data may help managers produce a variety of new reports that may initially impress the management, but do not help the corporate take any useful decisions as that data in up to 75% cases is found to be unstructured, complicated, duplicate and 33% believe it may be even inaccurate. This data may be coming from multiple applications, internal or external systems and may occur in varying formats All these variances around data makes the data useless. unless it is dealt with and hence, poses quite a challenge.

Data which is gathered from different methods and from small locations, we can generate useful weather models which helps us to predict weather pattern in the future. Increase in the temperature depends upon location in the world.

For this weather monitoring app, required data is not in the expected form and accurate data. In this, Pre-Processing stage Cleanse the input data. It applies rules to update or complete the same. Filters are applied as well to remove junk data that may also creep into the system if left unattended. Filters are useful to ensure quality data. Data quality is improved to achieve accuracy.

3. CONCLUSIONS

This paper is helpful in evaluation of Weather Application and attempts to use the Big Data which is being captured from multiple sources to design a system capable of forecasting weather based on recent global warming concerns.

Big Data is a process and science depend on many technologies and is still in an evolving phase. It is important to address big data in early stages which helps in the future strategies.

REFERENCES

[1] Tomar, Divya and Sonali Agarwal. “A survey on pre processing and postprocessing techniques in data mining.” International Journal of Database Theory & Application 7.4(2019)

[2] JayaramHariharakrishan, Mohanavalli.S, Srividya, Sundhara Kumar K.B, in IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP),2017. [4] 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.314-347,2018

[3] 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 and open research issues,”inf.Syst,vol. 47,pp 98-115

[4] C.Furber and M. Hepp,” Towards a Vcabular for Data Quality Management in Semantic Wb Architectures.”in Proceedings of the 1st International Workshop on Linked Web Data Management, New York,NY,USA,2015, pp.1-8.

[5] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, “Data quality: A survey of data quality dimensions,” in 2012 International Conference on Information Retrieval Knowledge Management (CAMP), 2012, pp. 300 –304.

[6] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, “Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model,” in 2014 47th Hawaii International Conference on System Sciences (HICSS), 2014, pp. 4700–4709.

[7] S. K. Bansal, “Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration,” in 2014 IEEE International Congress on Big Data (BigData Congress), 2014, pp. 522–529.

[8] G. A. Liebchen and M. Shepperd, “Software productivity analysis of a large data set and issues of confidentiality and data quality,” in Software Metrics, 2005. 11th IEEE International Symposium, 2005, p. 3 pp. –46.

[9] 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.