Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application

Ashish Juneja, Nripendra Narayan Das
Deaprtment of Computer Science & EngineeringFaculty of Engineering & Technology
Manav Rachna International Institute of Research and Studies
Faridabad, India
ash_juneja@hotmail.com, nndas.fet@mriu.edu.in

Abstract—**Big Data has become an imminent part of all industries and business sectors today. All organizations in any sector like energy, banking, retail, hardware, networking, etc all generate huge quantum of heterogenous data which if mined, processed and analyzed accurately can reveal immensely useful patterns for business heads to apply to generate and grow their businesses. Big Data helps in acquiring, processing and analyzing large amounts of heterogeneous data to derive valuable results. Quality of information is affected by size, speed and format in which data is generated. Hence, Quality of Big Data is of great relevance and importance. We propose addressing various aspects of the raw data to improve its quality in the pre-processing stage, as the raw data may not usable as-is. We are exploring process like Cleansing to fix as much data as feasible, Noise filters to remove bad data, as well sub-processes for Integration and Filtering along with Data Transformation/Normalization. We evaluate and profile the Big Data during acquisition stage, which is adapted to expectations to avoid cost overheads later while also improving and leading to accurate data analysis. 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.**

Keywords— Big Data, Big Data Quality, Data Quality, preprocessing, pre-processing.

I. INTRODUCTION

Big data is an evolving phase which means large volumes of both structured, semi-structured and unstructured data that pose a difficult task to be processed using traditional methods and databases. It is an approach for informed decision making using analytical techniques to describe any data set that is large enough that requires the use of high level programming skill and methodologies to make the data into an asset for an organization.

Such voluminous data can come from several different sources such as business transaction systems, customer databases, mobile applications, websites, machine-generated data and real-time data sensors used in internet of things (IoT) environments. This comes with complexities commonly known as 5Vs i.e. Volume, Variety, Velocity, Veracity, Value.

1) Volume: This is indicative of the huge data sets created at high frequency rates.

2) Variety: This deals with the different data types, i.e. structured, semi-structured, unstructured or all of these.

3) Velocity: This deals with the speed and frequency at which data may be generated by an application.

4) Veracity: This deals with the accuracy, truthfulness of the data, and if its authentic.

5) Value: This deals with the worthiness of data extracted from various raw data available. Just having data abundance doesn't essentially imply being able to extract usefulness from it.

In a Big Data system, data holds all essence to all the knowledge and possibilities of its applications. In fact, Data Quality is most often the reason for any business' data and information problems. The key data dimensions are:

- *Completeness:* Is data missing or not user friendly?
- *Timeliness:* Is data available for use in the time frame in which it is expected?
- *Conformity:* Is the data conforming to expected format?
- *Uniqueness:* Is the data duplicated within the available data set?
- *Integrity:* Ensure integrity of data and its relationships along with source or lineage of the data. Is the integrity ensured?
- Consistency: Is there a single source of truth or are different versions for the same data entity available across multiple environments?
- *Accuracy:* Is the data accurately representing the business data as expected?

The Data in a Big Data system traverses four phase within the Big Data Lifecycle, these are: 1) Data Origin Identification, 2) Data Acquisition and Cleansing, 3) Data Aggregation and Storage and 4) Data Analysis [1]-[5], as depicted in "Fig. 1".

1st Phase **Data Origin Identification** phase is concerned with the raw data being generated from a variety of sources and in abundance. The sources may include social sites, financial applications, customer relation applications, media web sites, images, etc. It is critical to understand the source(s) of the data, and identify it’s veracity or reliability level as the next phase uses this understanding to process the data.

2nd Phase **Data Acquisition and Cleansing** phase assimilates the data from many sources. This raw data may be messed up with anomalies including corrupted values, badly formatted and unsuitable for consumption by the Big Data application, a combination of structured, semi-structured and unstructured data. Such data needs to filtered and cleansed, reformatted and structured, deduped, remove illegal values and compressed. These pre-processing steps are crucial to transform the data to levels suitable or valuable for analysis.

3rd Phase **Data Aggregation and Storage** phase ensures the from many heterogenous sources is suitable aggregated with joins across source databases or stored within databases or file formats on which the analysis is planned.

4th Phase **Data Analysis** phase infuses relevance and sense into gathered data. This is a complex and evolving process that executes by comparing data characteristics to identify patterns using corrections as per domain knowledge or experience. The analysis results aim to help the users aware of the current state, make forecasts and informed decisions.

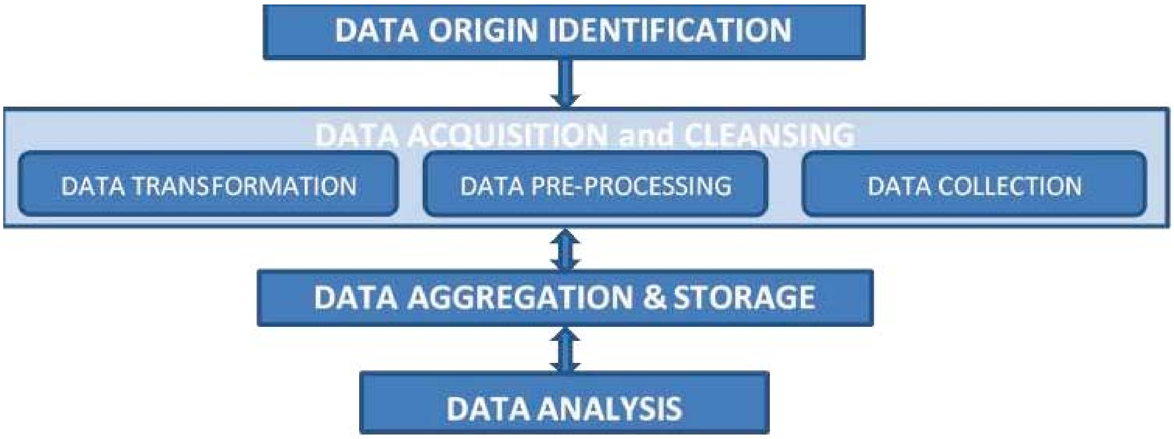
DATA ORIGIN IDENTIFICATION
DATA ACQUISITION and CLEANSING
DATA TRANSFORMATION
DATA PRE-PROCESSING
DATA COLLECTION
DATA AGGREGATION & STORAGE
DATA ANALYSIS

Fig. 1. Big Data Lifecycle

II. IMPORTANCE OF DATA QUALITY

Quality of Data has to be continually tracked, monitored and tuned to be best utilized and be as effective as possible while analyzing.

A lot of data is accumulated during the business lifecycle creating data that could be captured as unstructured, lacking quality of desired parameters. These businesses could include the power, energy, social, retail, e-commerce, and so many other existing or upcoming industrial sectors. Each sector has its data characteristics that are specific to its domain or nature of business. So, the data has to be useful for its users to be evaluated. A domain’s Data Quality needs to be appropriately conditioned to its specific domain making it quite a complex science to address. Hence, this paper focusses on improving Quality of the Data in early stages before it is consumed for further processing or analysis. It aims to resolve quality issues for large data sets using processes as manual efforts usually fail for large volumes of data.

Data, as we now understand, has to be domain conscious and suitable for consumptions by its users. Appropriateness of data characteristics like the format, structure, timing and data type variations need to be addressed as well from a quality perspective. The better the data source, better may be the output quality for the pre-processed data i.e. Data Quality is proportional to its source quality. Hence, this quality management and curing of data aberrations and complexities have an indirect bearing with effort and cost spend to produce quality data for analysis in Big Data systems or applications.

Building a Data Quality framework should consider multiple factor with critical ones being business domain, source(s) of data, structured/unstructured data.

i. *Data Quality Dimensions*

Data Quality dimensions are a means to assess the quality of data. These may be Intrinsic or/and Contextual [8][9]. Even though there is no standard or universal regulatory definition on these quality dimensions, the attempt here is to broadly agree on the commonly accepted ones. As such the Contextual may vary per business domain, application or relevance. The most popular Intrinsic dimensions include Accuracy, Consistency, Uniqueness, Timeliness, Validity and finally, Completeness.

ii. *Data Profiling*

Profiling data revolves around establishing a rule's framework to ensure efficient assessment of the quality of the data supported by a specific definition and characteristics on the quality of the data.

To ensure quality of data in a Big Data system data may assessed and transformed through numerous iterations in an effort to cleanse and also progress from an unstructured to a more structured state.

iii. *Data Quality Framework*

In [7] Data Quality framework establishes rules that aim to ensure data with enhanced quality. Processes to cleanse data, dedupe data, remove corrupted data instances and many more sub-process form part of the quality framework.

iv. *Big Data Quality*

Data Quality for the still evolving and growing field of Big Data is in itself a highly complex subject. Large organizations earlier believed having captured data from various business processes, multiple divisions, sales, profits, geographies, and locations parameters, etc. would empower them to magnify their business and spread in more areas strategically. However, the challenge remained how to tap and mine efficiently the huge volumes of data in a qualitative manner using standardized quality processes that also cleanse and improve the data quality as part of the Big Data life cycle. Ensuring and transforming the data to a quality one is crucial for any industry's Big Data platform to be able to analyze the data accurately and assess patterns that help in devising future strategies accurately or as best possible.

v. *Big Data Pre-Processing*

Pre-Processing of data or processing of data's quality at the early stage of any Big Data system's lifecycle enhance and refine the quality of the data. Data Pre-Processing lifecycle typically cover these sub-processes:

a) Data Consolidation and Integration

Data may be sourced from multiple locations that be may be in various forms, structured/semi-structured/unstructured, varying formats, junk, etc. Data from all these sources needs to be consolidated homogeneously to form a single and final source of truth for the data to be used in the Big Data system. Technologies like ETL [10] Extract, Transform and Load are popular and established mechanisms.

b) Data Enhancements and Enrichment

Data from various sources is consolidated, and during that data details are updated using additional information obtained from other supportive sources to create fused data that is Enriched with more information and possibly also enhanced qualitatively.

c) Data transformation

Data transformation involves many a step or sub-processed like capturing or pulling data from multiple sources, data may need to be reformatted, normalized, aggregated, even updated using regulatory standards.

d) Data reduction

Data reduction is the process of reducing the amount of data so that it becomes non-redundant. This helps in increasing the data storage efficiency and reducing costs by removing data that is not important and retaining only the meaningful parts for that particular work/task.

e) Data discretization

This process extracts and segregates data into intervals so that it can be efficiently utilized within available mining algorithm and techniques.

f) Data Cleansing

It is a process which helps improve the Quality of Data by removing the data which diminishes the usability of it. The steps involved in this process are removing the inaccurate, incomplete or irrelevant data from the data acquired so that it can be processed and analyzed upon to extract beneficial value from it [11][12].

III. Managing and Refining Quality of Big Data

Pre-Processing data and deciding a strategy on how the data from multiple sources will be formatted, stored and managed as a single source of truth is crucial to the success of corporates. Huge data volumes and multiple data sources may lead to data quality problems that require data cleansing solutions to resolve or minimize the issues.

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/external systems and may occur in varying formats (unstructured/structured) that may be non-homogeneous if attempted to store under a single storage system. All these variances around data makes the data useless unless it is dealt with and hence, poses quite a challenge.

The Big Data Pre-Processing lifecycle framework as shown in "Fig. 2" gives a 30000 feet view of the Big Data's journey through various lifecycle stages of the quality processing framework considered for a Weather Monitoring and Forecasting Application. Data from variety of sources that could be structured/semi-structured/unstructured enters the Pre-Processing data quality framework at the Pre-Processing Activity Selection stage to undergo Data Integration, Cleansing and Enrichment. This is followed by Technique Selection stage where Data Quality dimensions are improved by reducing data anomalies using techniques and rules of Auto-Discovery, Domain and User-Defined.

Thereafter, in the next stage the Data Quality profile is addressed based on many quality dimensions including the source of the data, duplication, detecting anomalies based on rules that have been set when Data Quality Rules were assumed. The output of this stage is an optimized data profile. In the subsequent stage, the Data Quality profiles are validated using samples of data from available resources i.e. data quality rules are tested. These steps are included in Pre-Processing automation, that is iterated recursively until the Data Quality achieves expected levels as per a pre-defined or agreed 'definition of done'. To realize the expected quality, a final Adapter stage may involve user interaction to repeatedly evaluate and improve the data quality profile as per domain.

Weather Monitoring and Forecasting Application

To understand the Data Pre-Processing lifecycle, we have conceptualized a Weather Monitoring and Forecasting Application. Such a solution involves receiving and capturing data from many sources like IOT Sensors for various weather parameters (including temperature, rain, humidity, dust/visibility levels, altitude, longitude, latitude, etc.), external APIs or micro-services for parameters from remote locations, travel guideline input data for locations, etc. The application also expects data on data point variations in temperature parameters as per global warming from various geographical locations.

The application will capture global temperature data average for last 5 decades. It will then capture fresh data for locations across the world to monitor any change in temperature above 1.5 degrees Celsius. Any change above that will trigger notification alarms to the scientific community. At the recent Paris Agreement, a goal has been set for temperature level across the globe to be constrained to not increase by 2 degrees beyond the pre-industrial one. This is since scientists have determined that more than 2 degrees of warming could have catastrophic effects for earth and mankind.

According to a report by the UN, even a jump of up to 1.5 degrees Celsius from the current 1 degree since 19th century would lead to shortages of water supply and flooding in coastal areas, coral reefs would start to vanish.

With the data gathered from several methods mentioned above and from several small locations that make up for a large area we can generate better weather models that can help us predict accurately the weather pattern in the near feature. We realize that with the increase in average temperature, it leads to varied affects depending on the type of regions and locations around the world.

For the above Application, we can only imagine the amount of data being captured and stored from so many sources, systems and temperature monitoring sensors or instruments across the world from many countries, organizations, laboratories. All these will not be able to provide us with the required data in expected form, format, accuracy or ensure even completeness for all the parameters. Hence, having a Pre-Processing stage that Cleanse the input data, attempts to apply rules to update or complete the same, as well reformats using Adapters using various pre-processes is the crucial process to make Data with the Big Data lifecycle usable and very useful. Filters need to be applied as well to remove junk data that may also creep into the system if left unattended. Filters will also be useful to address partially correct or partially complete data that may be fixed using automatic rules (for known or assumed situations) or require user intervention to make the data useful.

With so many data sources and data types, the system will be subjected to huge data that may be structured or most likely unstructured or semi-structured. A continuous monitoring

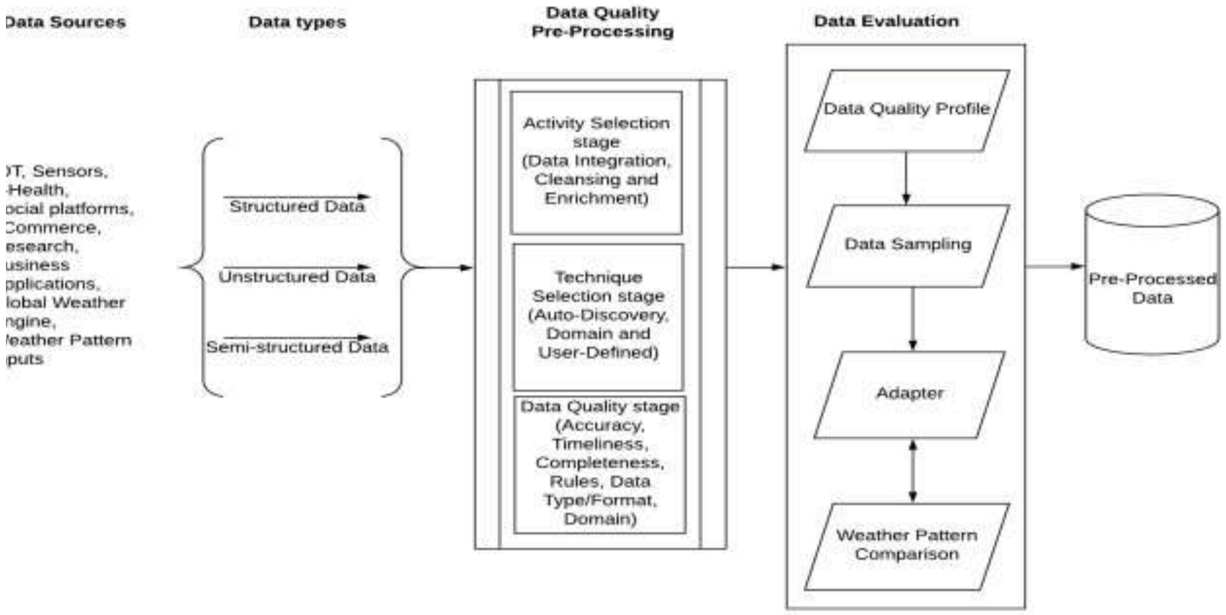
Data Sources
Data types
Data Quality Pre-Processing
Data Evaluation
)T, Sensors, -Health, ocial platforms, Commerce, esearch, usiness pplications, lobal Weather ngine, /eather Pattern puts
Structured Data
Unstructured Data
Semi-structured Data
Activity Selection stage (Data Integration, Cleansing and Enrichment)
Technique Selection stage (Auto-Discovery, Domain and User-Defined)
Data Quality stage (Accuracy, Timeliness, Completeness, Rules, Data Type/Format, Domain)
Data Quality Profile
Data Sampling
Adapter
Weather Pattern Comparison
Pre-Processed Data

Fig. 2. Pre-Processing Framework provides a bird's eye view of the different stages and the sequence of steps the Big Data undergoes right from the data's inception to be treated for various quality pre-processes resulting in Quality Pre-Processed Data.

system with sensors and external inputs is also bound to generate and capture huge volumes of data whose quality needs attention with cleansing or enriching as described in the lifecycle earlier. Applying domain specific rules and automated rules or user defined rules acquired by experience help in making data achieve better quality levels that the Forecasting system expects. The Data Quality is improved for accuracy, completeness, formatting and applying rules and knowledge as per the Weather Forecasting Domain. Further, the data evaluation stage helps evaluate the Data Quality as per the domain depicted in "Fig. 2".

The better the pre-processed Data Quality, better is the analysis or forecast the system will produce.

IV 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] 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)

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

[3] 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.314-347,2014

[5] 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

[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.652-687,2014

[7] 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,2011, pp.1-8.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.