Quality Issues with Big Data analytics

Abhishek Jain[1],
[1]Student,
Department of Computer Science and Engineering
Ganga Institute of Technology and Management,
Jhajjar, India

Mahesh Kumar Malkani[2]
[2]Assistant Professor,
Department of Computer Science and Engineering
Ganga Institute of Technology and Management,
Jhajjar, India

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data.' 'In this paper techniques related with the big data pipelining or processing are discussed. There are number of issues related with the analytical results of the big data and the big data processing software. The major issues are included in this paper and will be helpful for the new beginners in this area of big data analytics.

Keywords— Big Data Reliability; Big Data Analytic.

I. INTRODUCTION

Big data involves the data sets that are having the large volume, velocity and variety and it is becoming difficult to process this dataset using traditional data management tools or processing applications. The size of data in 2011 is roughly 1.8 Zettabytes. There is supporting networking infrastructure and have to manage nearly 60 times more information by year 2020.Issues of efficiency, economics and privacy need to be carefully planned when adding the big data building blocks in the already developed data and networking infrastructure. Big data is defined in terms of three ways which includes Infrastructure Perspective, Analytics perspective and from the business perspectives. In terms of infrastructure perspective Big data has been defined as data with high volume, velocity, and variety, veracity, viability, validity, value and volatility(8V) and unpredictability. The data here is very much large that current, traditional methods can't be used to process it. In terms of the analytics perspective, big data can be defined in terms of the events which are having the low probability of happening and are different from the traditional small sampled data. From the business perspective, big data are giving the opportunities of gaining the actionable intelligence. Actionable intelligence is the output of the big data software that will be directly useful for the business and other perspectives without any modifications or any production process.

These definitions are providing the important context that must be considered by the big data analysis software.

The quality of analytics system is one of its most important attributes. Organizations are trying to ensure highest quality for the systems being developed, but ensuring the same is very difficult due to the following reasons-

- Increasing software size
- Budget constraints
- Time constraints
- Shortage of skilled manpower.

The present situation requires that the organizations be equipped with techniques that enable them to improve the quality of deliverables. There are number of factors that are used to measure the system development quality and each attribute can be used to measure the product performance. These includes Reliability, Security, Usability, Availability, Correctness, Efficiency, Integrity and the testability, flexibility, Reusability and the interoperability. Reliability is one of the most important quality attributes for developed systems. In modern society there is a great impact of big data software reliability because of its vast applications in different areas affecting many millions of people directly or indirectly. So there is requirement to predict the reliability in terms of consistency of the analytical results of the big data software. Big data are having the need of super computers, such machines can now be analyzed with the desktop computers with standard software and the modern fast developing techniques has the advantages for the development in the big data handling techniques.

The big data analytics is the procedure for examining the large amount of datasets which is having the variety, velocity and the volume of the data types that is the big data. This analysis will uncover the unknown patterns, unknown correlations and other useful information related with the business. The analytics results can be directly used for more effective marketing, new revenue opportunities and better improved customer services etc. without any processing to the analytics results. There is a need to check the quality factors associated with these analytical results.

II.LITERATURE REVIEW

Sharma in 2010 [5] proposed a new approach of Distance Based for the Optimal selection of SRGMs. This approach is useful for the optimal selection of parameters in the big data analytical results. C. E. and A. Peter gave "Research Directions for Engineering big data analytics software," in Jan/Feb 2015, this article gave major factors that are helpful in learning the factors that are related with the quality of the analytical results. A group at united state gave the ways for rhea challenges and opportunities with the big data. L. Cai and Y. Zhu[3] gave the Challenges of Data Quality and Data Quality Assessment in the Big Data Era. This challenges are providing the methods that can be helpful in the research in big data analytics. X. Wu, [4] X. Liu and S. Dai, The Reliability of Big Data gave the methods that are helpful in the reliability assessment of the big data. They used the random sampling method, which is the statistical technique for assessing the reliability of data, other such techniques can be used.

III.PHASES IN THE PROCESSING PIPELINE

There are five major steps in processing of the big data, and are giving the analytical results starting from the Acquisition/recording state then doing the Extraction/Cleaning following the Integration/Aggregation and at last doing the analysis and Interpretation of the datasets. These steps are more explored as follows-

Data Acquisition and Recording -Big Data cannot be obtained directly, data have to be recorded from the data sources and need to be filtered [3].

Information Extraction /Cleaning-The information which has been collected is not in a ready to use format for analysis .This needs to be extracted using different data extraction techniques. Data cleaning considers/assumes well-understood conditions on valid data error finding models [2].

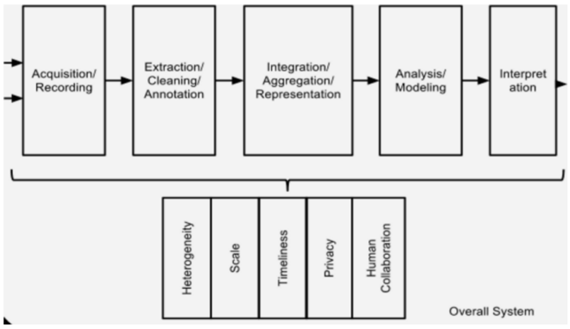

Fig-1 Major steps in the analysis of big data: big data pipeline

Data Integration, Aggregation, and Representation –There is a problem with today's Big Data analytics, which is the lack of correlation between database systems, which handles the data and provide queries for processing of the no SQL queries, such as data mining and statistical analyses. The big data needs to be properly integrated and presented.

Query Processing/Data Modeling, and Analysis -Querying and mining Big Data methods are functionally different from traditional statistical analysis which is done using the small samples. New optimal techniques and models for analysis are required to be used [6] and ranking of the techniques on particular methods is also required to be done so that optimal model and method can be selected [7].

Interpretation-Big data analyzing capability is of no use if anyone cannot understand the analysis. There should be a decision-maker, that provides the result of analysis, and he has to interpret these results.

IV.PROBLEM ENCOUNTERD WITH BIGDATA ANALYTICS

There are number of problems encountered when big data analytics is done. These are related with the Requirements specification, Security, design including the reliability and testing of the big data analysis [1].

Requirements

Specifying the details of the requirements is important for the successful construction of software. The problem of requirement can be understood if one can develop a model for functional requirements specification and verification. To show this feature of big data software, there is need to develop a model which can include the attributes like time and functional details and this can be better understood as-

R_i fi (x_i t_ip $)_i$

where the system provides a functional feature
with some time constraint t ,and verifiable
during some period p.$_{ii}$

There is the necessity for the big data software to include all the above three parameters to completely define the requirement specification for such software .But all of the models are not including the constrained time and verifiable time and specify the requirement only in terms of the functional features.

Reliability

Reliability is the most difficult to achieve quality attribute. There is no direct way to check the consistency of the results of the data analytics. There are number of ways by which we can check the reliability of the data analytics result.

Random sampling is the main method for the analysis of the big data and accuracy of the random sampling completely depends on the randomness of the data. It is essential to make the analysis of the reliability for the big data. Big data provides the messiness of the data .The data samples taken will be more random and size of the samples does not matter. This messiness of the data can have the more reliable results. There is the need to analyze the reliability of these datasets [5].

In case of small datasets, dataset is not random, limited amount of data is there and have the fixed point in the graph. But in case of big data, data is very random and large and is showing the curve in the graph.

V. QUALITY ISSUES WITH BIG DATA ANALYTICS

Consistency of the data analytics results cannot be provided directly rather some human decisions and domain knowledge can be used to detect the inaccurate or extreme results. The other way is to use design for reliability method in which architectural tactics are used to enhance the reliability of the big data software analytics results. As the big data software are complex there construction is complex enough as these are requiring the multiprocessing technologies as CUDA, GPUs, Map Reduce and Hadoop etc. These are requiring the high skill development during the development phase. New environments and frame-works for improving the usability of these technologies are enhancing day by day to reduce the learning curve for the big data software and involve the Spark, Mahout and Storm etc.

TABLE-SHOWING THE BEHAVIOR OF SMALL DATA SET VERSUS LARGE DATA SET

Features	Small data	Big data
1.Entropy of data	Low	High
2.Modeling Techniques	Parametric	Non-Parametric
3Effect of data size	Efficiency, Easy	Accuracy
4.Method of problem solution	Direct rules are available	Non-linear models are available

In this table comparison of the small data and big data is done showing that there are different ways by which the two datasets are handled. In both of them Entropy of data, Modeling Techniques, Effect of data size on result and Method of problem solution are different. So different methods are required to handle them.

Security-
Security is also the major concern in big data analytics. There are two types of attacks these are Execution phase attack and attacks which occur during the training. Execution phase attacks the data input streams which are added to influence the actionable intelligence and has been generated by the big data software. In case of training phase attacks malicious attackers can create data generators that affect the reliability of the big data software analytics results. Denial of service is also the problem in big data soft wares and may cause the problem of denying access to system.

Testability
Testability of the big data is also the major issue in big data analytical results. Verification techniques should be there to test the validity of the results.

Availability
Data must have the accessibility and timeliness. There is a need to check whether a data access is available or not. Availability requires that the 1) data is easily made public or easy to purchase 2) data should arrive on a given time 3)data are required to be regularly updated 4)the data interval from data collection and processing to release meets requirements.
Availability incurs all these above features.

Scalability
Size of big data is a major challenge today. First Big data is having scalability in storage, as there are the increases in data density on the secondary storage devices. The current Redundant Array of Independent disks method that is in common use does not provide the level of performance and durability of the data those enterprises requires for dealing with large volumes of data. Second Data volume is increasing at high speed than computing resources and processor speeds that are available in the market areas. Techniques for the Parallel data processing that were used in the traditional approaches are not directly applied for the internal and intra node parallelism, because the architecture looks very different. As the power considerations in the future are likely to inhibit us from using all of the hardware in the system continuously, new data processing systems and techniques will be required to actively manage the power consumption of the processor. All these changes force us to rethink how data processing components are designed, built, and operated.

VI.CONCLUSION

In this paper basics of big data analytics are discussed with the techniques for big data processing .The major issues related with the big data analytics software and their results are discussed in this paper. This will be helpful for the researchers having an interest in the big data analytics. Everyone must support and encourage all the fundamental research towards addressing these issues if we have to achieve the promised use of Big Data.

REFERENCES

[1] C. E. a. A. Peter, "Research Directions for Engineering big data analytics software," Jan/feb 2015.

[2] B. Li, "Survey of Recent Research Progress and Issues in Big data," 2013.

[3] A. c. w. p. d. b. l. r. a. t. U. States, "Challenges and Opportunities with Big Data".

[4] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era".

[5] X. Wu, X. Liu and S. Dai, "The Reliability of Big Data".

[6] K. sharma, R. Garg, C. Nagpal and R.K.Garg, "Selection of Optimal Software Reliability Growth Models using a distance based Approach," *IEEE Transaction on Reliability*, pp. 266276, 2010.

[7] R. Garg, K. Sharma, C. Nagpal, R. Garg, R. Garg, R. Kumar and Sandhaya, "Ranking of Software engineering metrics by fuzzy-based matrix methodology," *Software Testing , Verification and Reliability*, vol. 23, pp. 149-168, 2011.