ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.6

Challenges and Opportunities of Big Data Analytics

Bare Bhakti C.[1], Dr.S.N.Kini[2]
Student[1], Professor[2]
Department of Computer Engineering
Jayawantrao Sawant College of Engineering, SPPU, Pune, India

Abstract:
This paper focuses on challenges and opportunities in Big Data Analytics. Big Data is a term in which there is no unique unit to specify size of data. Big data is defined as either large volume offline data or continuous streaming data with no control on its speed. There are some challenges to deal with offline large data and high speed continuous streaming of data. These challenges provide opportunities to overcome the problem. To provide consistent approach, we need to provide analytics techniques. To implement these techniques, Apache Sparks Technology is efficient. So in this paper, challenges and opportunities are discussed.

Keywords: Big Data, Big Data Analytics, Offline stored data, High Streaming Data , Apache Sparks , deep learning, machine learning

1.INTRODUCTION

An internet plays vital role which connects a world with one click. It made our lives easy so that one can send and receive data with no time. As data comes in scene, user can share, update, retrieve data efficiently. As number of user increase, size of data also increases. This data then become vast to manage and store. Here, the term Big Data comes. Big Data can be defined as high speed continuous streaming data or offline stored data with large volume. So there is source of data with tremendous volume. To get proper knowledge of data, we need to extract information. This information is helpful for the analytics to be used by other applications. Traditional Data Analytics Techniques are not sufficient to deal with large amount of data. In this paper, we will discuss challenges of traditional analytics techniques and opportunities for Big Data Analytics.

1.1 Big Data

A classification of data sets without their format can be done by Big Data. Big Data is having following characteristics.[1]

- Volume: There is no specific size of Big Data. Volume of Big data is tremendous which requires distributed system to process it.
- Velocity : As defined above, Big data is high speed continuous data which becomes difficult to handle in timely manner so that performance of system can be achieved.
- Variety: A Data is generated at every event done by user. A data can be any format like text, video, audio, structured data or even clicks made by user. Big Data contain multiple formats of data.
- Veracity: Big Data is collected from various sources with or without maintaining quality of data. Veracity is the reliability of data collected.
- Variability: The speed of continuous streaming varies from one source to another. Quality of data is also variable.
- Value: Big Data with given properties is not problem, providing proper analytics by using data is the challenge.

Finally, one can define Big Data by the given properties. Challenges of Big Data are the opportunities to provide solution over analytics.

1.2 Why Big Data Analytics

Traditional analytics techniques become insufficient to deal with Big Data. In Traditional Analytics, format and structure of data is known. So there is no real time decision making with these techniques when we work with Big Data. To overcome this, Big Data Analytics is used which focuses on predictive analytics to specify what might happen in future. These kind of analytics is useful in forecasting applications, prediction applications, health care systems etc. Moreover, Big Data complexity terms in data, computational and system complexity. Big data analytics aims to manage large amount of data despite of software, hardware, bandwidth constraints. Creating visualizations of Big Data helps to expose hidden patterns of data which leads to the deep learning. Building set of constraints makes the system efficient to analyse big data. By evaluating constraints on the use of data one can assess the data patterns and lifecycle of data. Another technique to analyse Big Data with high performance is to distribute the work over clusters of nodes. Dealing with Big Data on Local machine is not sufficient. But we can implement model on Local machine and after successful testing of functionality of module, these modules can be distributed over a clusters of node. This approach is known as Proof Of Concepts (POC). As we submit implantation module to clusters, each node in cluster gets copy of the module (i.e. code) for the execution of local data. After processing of individual node data get combined into single unit. This approach enhances the performance of system by providing high speed execution.

2. LITERATURE SURVEY

This section focuses on review of different papers on Big Data and challenges, implementation techniques, analytics , etc.

Sr.No.1
Paper Title: Big Data And Cloud Computing : Current State and Future Opportunities. [2]
Topic: Big Data and Cloud Computing

Description: It focuses on cloud computing and Big Data. Scalable Database Management supports heavy applications and ad-hoc analytics and decision support.
Advantages: scalability, elasticity, fault-tolerance, self-manageability, and ability to run on commodity hardware
Disadvantages: To provide feasibility of system to make effective use of available resources and minimize operation cost.

Sr.No. 2
Paper Title: MAD Skills: New Analysis Practices for Big Data[3]
Topic: Big Data Analytics
Description : Big Data Collection and analysis is a challenge. It provides Magnetic ,Agile and deep analysis.
Advantages: Quick Import and Frequent iterations.
Disadvantages: requires standardizing a vocabulary for objects like vectors, matrices, functions and functional.

Sr.No.3
Paper Title: Starfish: A self tuning system for Big Data Analytics[4]
Topic: Big Data Analytics.
Description: It uses MAD skills to express the features and overcomes challenges of MAD. Starfish is a MADDER and self-tuning system for analytics on big data.
Advantages: Data-lifecycle awareness, Elasticity, and robustness, minimizes cost and improves performance**.**
Disadvantage: leading us to different design choices.

Sr.No.4
Paper Title: MapReduce: Simplified Data Processing on Large Clusters[5].
Topic: Map-reduce.
Description: Map-Reduce is a programming model and an associated implementation for processing and generating large data sets.
Advantages: parallelize and distribute computations and to make such computations fault-tolerant, redundant execution can be used to reduce the impact of slow machines.
Disadvantages: -

Sr.No.5
Paper Title: Real Time Big Data Analytical Architecture for Remote Sensing Applications[6]
Topic: Big Data Analytics
Description : This uses a distributed approach by map reduce. Structural architecture is composed of Remote Sensing Data Acquisition unit , Data Processing Unit , Data Analysis and Decision-making Unit to provide analytics of remote sensing application data.
Advantages: Distributed Approach is used.
Disadvantages: Need to study format of satellite data.

Sr.No 6
Paper Title: Social Set Analysis: A Set Theoretical Approach to Big Data Analytics[7].
Topic: Big Data Analytics using set theory.
Description: It provides a new approach to big data analytics called social set analysis. Social set analysis consists of a generative framework for the philosophies of
Computational social science, theory of social data, conceptual and formal models of social data, and an analytical framework.
Advantages:
It helps to extract meaningful facts, actionable insights and valuable outcomes from Big Social Data analytics.

3. ANALYTICS OF SATELLITE DATA

As per definition of big data, a large volume of data stored offline is also big data. i.e. a data collected from remote sensing satellite applications. By which we can provide predictive analytics of data, forecasting , change detection etc. Data come at high speed and are tremendous in volume. So handling data with high volume is a challenge and it can be implemented by distributing process over systems with big data.

3.1 Scalable Data Management [2]

Scalable Data Management provides an efficient way to handle large amount of data with bigger applications to be used by many users. Figure 3.1 shows the scenario of scalable data management. Here, Data is processed through distributed approach. To speed processing load balancer distributes work over servers. It then aggregates results taken from servers and then performs interpretation. After which results of analytics are used by various applications.

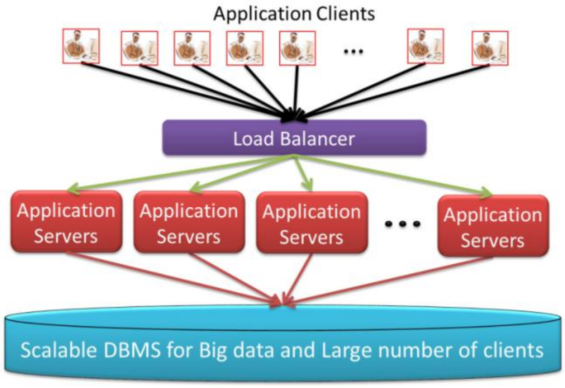

Figure.1. Scalable Data Management[2]

3.2 Analytics of Satellite Data

Fig 3.2 shows the structural architecture design to handle satellite data and to provide analytics which can be used to detect changes in an object observed at specific time intervals. It collects data across the world .We assume that data is preprocessed by satellite. For the further processing, data is stored offline and then secondary preprocessing is done. Useful data are filtered among large volume of data sets. Modeling of these filtered data helps to provide patterns of data by which one can analyze data and can uncover hidden knowledge. This knowledge is then used by various applications.

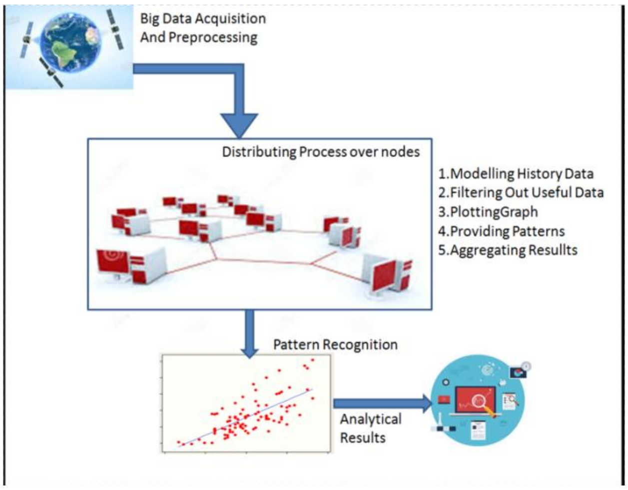

Figure.2. System Structure for Satellite Data

4. CONCLUSION

In this paper, we focused on remote sensing satellite data i.e. Big Data. To provide analytics of Big Data is challenging task. Handling large volume of data is crtical task. As we are interested in knowledge gain by information, mining of data from large volume data sets has limitations of designing filters in such way that it must not discard useful data from data set. So there is need to design filters in such way, so that it will include all required data without loss of useful data. Another problem is to provide high performance of systems. It can be achieved by distributing task over clusters of node so that work is done within less time with efficiency. To implement such techniques , Apache Hadoop and Apache Sparks can be used. Apache Hadoop uses Map-Reduce functionality to provide distribution of work. Apache Sparks is useful to implement iterative processing used in machine learning. Apache Sparks uses a functionality of Map-Reduce internally. One can extend the system to provide prediction of earthquake or tsunami, or can be used in health care systems.

5. ACKNOWLEDGEMENT

This is to acknowledge and thank all the individuals who played defining role in shaping this seminar report. Without their constant support, guidance and assistance this seminar report would not have been completed. Without their coordination, guidance and reviewing this task could not be completed alone. I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my guide Prof. S.N.Kini for giving his valuable guidance, inspiration and encouragement to embark this seminar. This seminar being conceptual one needed a lot of support from my guide so that I could achieve what I was set out to get.I would personally like to thank, Prof. M. D. Ingle, PG Seminar Co-ordinator, Computer Department, Prof. H. A. Hingoliwala, Head Of Computer Department and our Honorable Principal Dr. M. G. Jadhav Sir who creates a healthy environment for all of us to learn in best possible way.

6. REFERENCES

[1].Muhammad Habib ur Rehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah,Samee U. Khan, “Big Data Reduction Method : A Survey” Data Sci. Springer, DOI 10.1007/s41019-016-0022-0

[2].D. Agrawal, S. Das, and A. E. Abbadi, “Big Data and cloud computing:Current state and future opportunities,” in Proc. Int. Conf. Extending Database Technol. (EDBT), 2011, pp. 530–533.

[3].J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, “Madskills: New analysis practices for Big Data,” PVLDB, vol. 2, no. 2,pp. 1481–1492, 2009

[4].H. Herodotou et al., “Starfish: A self-tuning system for Big Data analytics, ”in Proc. 5th Int. Conf. Innovative Data Syst. Res. (CIDR), 2011,pp. 261–272.

[5].J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.

[6].Muhammad Mazhar Ullah Rathore, Anand Paul, Bo-Wei Chen, Bormin Huang, and Wen Ji, “Real-Time Big Data Analytical Architecture for Remote Sensing Application,” Ieee Journal Of Selected Topics In Applied Earth Observations And Remote Sensing 2015.

[7].Ravi Vatrapu,Raghava Rao Mukkamala, Abid Hussain,And Benjamin Flesch “ Social Set Analysis: A Set Theoretical Approach to Big Data Analytics” Digital Object Identifier 10.1109/ACCESS.2016.2559584 IEEE