Big Data: Big Challenges and Opportunities

P.Punitha Ponmalar[1], G.Sujatha[2]

Department of Computer Science, Sri Meenakshi Govt Arts College,Madurai

[1]p_punithanadraj74@yahoo.co.in, [2]sujisekar05@gmail.com

***Abstract*—** Big Data has drawn great awareness from researchers of information sciences, policy and decision makers of different organizations. As the speed of information growth, excessive data are generated that makes great troubles to humans. In this huge volume of data, so much potential and highly useful values are hidden. On the other hand, Big Data also has many challenges like capturing data, storing data, analysis of data and visualization of data. This paper is an attempt to review the challenges and opportunities of big data by introducing the general background of big data followed by the review of related technologies.

***Ke ywords*—** **Bigdata , Internet of things, Hadoop, Big data Analysis ,Data collection**

I. INTRODUCTION

Big Data is a term which is used to describe massive amount of data generating from digital sources or the internet. From the past few years data is exponentially growing due to the use of connected devices such as smart phone's, tablets, laptops and desktop computer. This generated data volume is so vast and overwhelming which makes complex to process and analyze using traditional software systems consuming more time. Doug Laney [3] discussed about 3 V's in Big Data management are shown in figure 1.

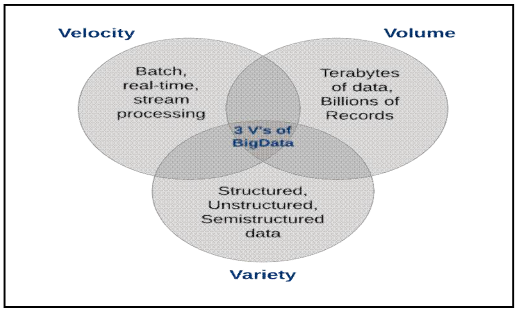
Volume : Volume is the amount of data generated by organizations or individuals. Today the size of data is increased to Exa bytes. The grand scale and rise of data outstrips traditional store and analysis techniques [4].

Variety : Variety makes health care data really big. Big data comes from a great variety of sources such as internal, external, social and behavioural and generally has in three types: structured, semi structured and unstructured. Structured data inserts a data warehouse already tagged and easily sorted but unstructured data is random and difficult to analyze. Semi-structured data does not conform to fixed fields but contains tags to separate data elements [4][13].

Velocity: Velocity describes the rate at which data is generated, captured and shared [4][16]. Healthcare Data can be generated from two sources: humans, or sensors. With a few exceptions like diagnostic imaging and intensive care monitoring, most of the data used in healthcare is entered by people, which effectively limits the rate at which healthcare organizations can generate data.
Nowadays, there are 2 more V's:

Variability: There are changes in the structure of the data and how users want to interpret that data.

Value: Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach [6].


Velocity
Volume
Batch, real-time, stream processing
Terabytes of data, Billions of Records
3 V's of BigData
Structured, Unstructured, Semistructured data
Variety

This big data needs to be processed and analyzed to extract knowledge for decision-making and cost-saving.

II. RELATED WORK

Fahad et al. [5] present a survey of existing clustering algorithms of different categories. In [2] the authors focus on the most popular and most used algorithms in the literature like k-means, they presents some comparative work of these algorithms. Another recent research [11] presents a general view of data mining algorithms and platforms that can be used in the field of Big Data by discussing different challenges and characteristics. Paper [4] discusses some of Big Data mining algorithms to find the most appropriate among them using a comprehensive comparison. Others in [14] are discussed classification algorithms and how it is used in statistics and apply them to specific databases. Researchers in [7] present a review of some old algorithms that can handle large data set as Nearest Neighbor Search, Decision Tree and Neural Network. In [8], Laney et al. discuss different data management techniques. They present an overview of different categories of data mining. The scalability of the parallel k-means algorithm has also been demonstrated [10]. In [12] proposes a classification algorithm for Big Data based on feature selection. Cui [3] addresses the Big Data processing problem using the K-means algorithm that proposes a new model of treatment with Map Reduce to eliminate iteration dependency and achieve high performance.

III. FLOW MODEL FOR BIG DATA

The Big Data Analytics review here consists of four important levels that carry out Data Sources, Information Ingestion, Data Analytics, Information Analysis and Information Consumption.

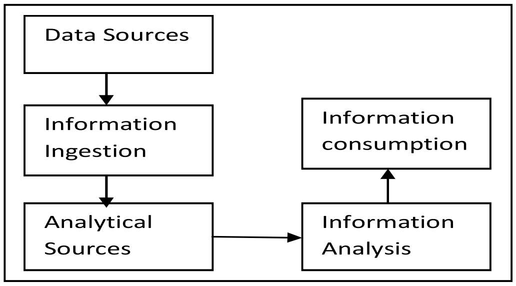


Fig 2 : Flow model of Big data

A. DATA SOURCES

Big data is different from the data being stored in traditional warehouses. The data stored there first needs to be cleansed, documented and even trusted. Moreover it should fit the basic structure of that warehouse to be stored but this is not the case with Big data it not only handles the data being stored in traditional warehouses but also the data not suitable to be stored in those warehouses. Thus there comes the point of access to mountains of data and better business strategies and decisions as analysis of more data is always better.

1) Log Storage in IT Industries

IT industries store large amount of data as Logs to deal with the problems which seem to be occurring rarely in order to solve them. But the storage of this data is done for few weeks or so though these logs need to be stored for longer duration because of their value. The Traditional Systems are not able to handle these logs because of their volume, raw and semi structured nature. Moreover these logs go on changing with the s/w and H/w updates occurring. Big data analytics not only does analysis on the whole /large data available to pinpoint the point of failures but also would increase the longevity of the log storage.

2) Sensor Data

Massive amount of sensor data is also a big challenge for Big data. All the industries at present dealing with this large amount of data make use of small portion of it for analysis because of the lack of the storage infrastructure and the analysis techniques. Moreover sensor data is characterized by both data in motion and data at rest. Thus safety, profit and efficiency all require large amount of data to be analyzed for better business insights.

3) Risk Analysis

It becomes important for financial institutions to model data in order to calculate the risk so that it falls under their acceptable thresholds. A lot amount of data is potentially underutilized and should be integrated within the model to determine the risk patterns more accurately.

4) Social Media

The most use of Big data is for the social media and customer sentiments. Keeping an eye on what the customers are saying about their products helps business organizations to get a kind of customer feedback. This feedback is then used to modify decisions and get more value out of their business.

B. INFORMATION INGESTION

As the second phase of the big data system, big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once we collect the raw data, we shall utilize an efficient transmission mechanism to send it to a proper storage management system to support different analytical applications. The collected datasets may sometimes include much redundant or useless data, which unnecessarily increases storage space and affects the subsequent data analysis. For example, high redundancy is very common among datasets collected by sensors for environment monitoring. Data compression technology can be applied to reduce the redundancy. Therefore, data pre-processing operations are indispensable to ensure efficient data storage and exploitation.

1) DATA CAPTURE

Data sets grow in size because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies, remote sensing, software logs, cameras, microphones, radio-frequency identification readers, wireless sensor networks, and so on.

2) DATA TRANSMISSION

Cloud data storage is popularly used as the development of cloud technologies. The network bandwidth capacity is the bottleneck in cloud and distributed systems, especially when the volume of communication is large.

3) DATA PREPROCESSING

Data pre-processing is aimed at data discovery and retrieval, data quality assurance, value addition, reuse and preservation over time. This field specifically involves a number of sub-fields including authentication, archiving, management, preservation, retrieval, and representation. The existing database management tools are unable to process Big Data that grow so large and complex. This situation will continue as the benefits of exploiting Big Data allowing researchers to analyse business trends, prevent diseases, and combat crime.

Though the size of Big Data keeps increasing exponentially, current capability to work with is only in the relatively lower levels of petabytes, exabytes and zettabytes of data. The classical approach of managing structured data includes two parts, one is a schema to storage the data set, and another is a relational database for data retrieval. For managing large-scale datasets in a structured way, data warehouses and data marts are two popular approaches. A data warehouse is a relational database system that is used to store and analyse data, also report the results to users. The data mart is based on a data warehouse and facilitates the access and analysis of the data warehouse. A data warehouse is mainly responsible to store data that is sourced from the operational systems. The pre-processing of the data is necessary before it is stored, such as data cleaning, transformation and cataloguing. After these pre-processing, the data is available for higher level online data mining functions.

C. DATA ANALYSIS

The first impression of Big Data is its volume, so the biggest and most important challenge is scalability when we deal with the Big Data analysis tasks. In the last few decades, researchers paid more attentions to accelerate analysis algorithms to cope with increasing volumes of data and speed up processors following the Moore's Law. For the former, it is necessary to develop sampling, on-line, and multi resolution analysis methods. In the aspect of Big Data analytical techniques, increment algorithms have good scalability property, not for all machine learning algorithms. Some researchers devote into this area. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology—although the clock cycle frequency of processors is doubling following Moore's Law, the clock speeds still highly lag behind. Alternatively, processors are being embedded with increasing numbers of cores. This shift in processors leads to the development of parallel computing. For those real-time Big Data applications, like navigation, social networks, finance, biomedicine, astronomy, intelligent transport systems, and internet of thing, timeliness is at the top priority. It is still a big challenge for stream processing involved by Big Data.

It is right to say that Big Data not only have produced many challenge and changed the directions of the development of the hardware, but also in software architectures. That is the swerve to cloud computing, which aggregates multiple disparate workloads into a large cluster of processors. In this direction, distributed computing is being developed at high speed recently.

D. DATA VISUALIZATION

The main objective of data visualization is to represent knowledge more intuitively and effectively by using different graphs. To convey information easily by providing knowledge hidden in the complex and large-scale data sets, both aesthetic form and functionality are necessary. Information that has been abstracted in some schematic forms, including attributes or variables for the units of information, is also valuable for data analysis. This way is much more intuitive than sophisticated approaches. Online marketplace eBay, have hundreds of millions active users and billions of goods sold each month, and they generate a lot of data. To make all that data understandable, eBay turned to Big Data visualization tool: Tableau, which has capability to transform large, complex data sets into intuitive pictures. The results are also interactive. Based on them, eBay employees can visualize search relevance and quality to monitor the latest customer feedback and conduct sentiment analysis.

For Big Data applications, it is particularly difficult to conduct data visualization because of the large size and high dimension of Big Data. However, current Big Data visualization tools mostly have poor performances in functionalities, scalability and response time. What we need to do is rethinking the way we visualize Big Data, not like the way we adopt before. For example, the history mechanisms for information visualization also are data-intensive and need more efficient approaches. Uncertainty can lead to a great challenge to effective uncertainty-aware visualization and arise in any stage of a visual analytics process. New framework for modelling uncertainty and characterizing the evolution of the uncertainty information are highly necessary through analytical processes. The shortage of talent will be a significant constraint to capture values from Big Data. In the United States, Big Data is expected to rapidly become a key determinant of competition across many sectors.

IV. CONCLUSIONS

This paper reviewed the big data challenges, its importance and opportunities. To accept and adapt to this new technology many challenges and issues exist which need to be brought up right in the beginning before it is too late. All those issues and challenges have been described with a flow model of Big data. These challenges and issues will help the information analysts who are moving towards this technology for increasing the value of the analysis.

REFERENCES

[1] *F. Bu, Z. Chen, Q. Zhang, and X. Wang, "Incomplete Big Data Clustering Algorithm Using Feature Selection and Partial Distance," In Digital Home (ICDH), 5th International Conference on. IEEE, p. 263- 266, 2014.*

[2] *A.BEN AYED, M.BEN HALIMA and M. ALIMI, "Survey on clustering methods: Towards fuzzy clustering for Big Data," In Soft Computing and Pattern Recognition (SoCPaR), 6th International Conference of. IEEE, p. 331-336, 2014.*

[3] *X. Cui, P. Zhu, X. Yang, K. Li, and C. Ji, "Optimized Big Data K-means clustering using MapReduce," The Journal of Supercomputing, vol. 70, no3,p.1249-1259,2014.*

[4] *C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012*

[5] *Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." Emerging Topics in Computing, IEEE Transactions on 2.3 (2014): 267-279.*

[6] *W. Fan, A Bifet, "Mining big data: current status, and forecast to the future", ACM SIGKDD Explorations Newsletter, Vol. 14, pp.1-5, 2013.*

[7] *D. Jianqiang, W. Fei and Y. Bo, "Accelerating BIRCH for clustering large scale streaming data using CUDA dynamic parallelism," In Intelligent Data Engineering and Automated Learning–IDEAL 2013. Springer Berlin Heidelberg, p. 409-416, 2013.*

[8] *D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001*

[9] *S. Madden, "From Databases to Big Data", IEEE Internet Computing, June 2012, v.16, pp.4-6*

[10] *H. S. Nagesh, S. Goil, and A. Choudhary, "A scalable parallel subspace clustering algorithm for massive data sets," In Parallel Processing, 2000. Proceedings. International Conference on. IEEE, p. 477-484, 2000*

[11] *A. Sherin, S. Uma, K.Saranya and M. Saranya Vani "Survey On Big Data Mining Platforms, Algorithms And Challenges". International Journal of Computer Science & Engineering Technology,Vol. 5 No, 2014.*

[12] *S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big Data Clustering: A Review," In Computational Science and Its Applications–ICCSA 2014. Springer International Publishing, p. 707- 720. 2014.*

[13] *S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011*

[14] *R. XU and D. WUNSCH, "Survey of clustering algorithms," Neural Networks, IEEE Transactions, vol. 16, no 3, p. 645-678, 2005.*

[15] *V.Bhat, H., Rao, P. G., Shenoy, P. D., Venugopal, K. R., Patnaik, L. M. :An Efficient Prediction Model for Diabetic Database using Soft Computing Techniques. LNCS, vol. 5908/2009, pp. 328-335. Springer Heilelberg (2009).*