ISSN: 2321-7782 (Online)
Volume 3, Issue 7, July 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Data Mining With Big Data: A Servey Paper

Manisha V.Kharat[1]
M.E Student (Computer Science & Engg),
Anuradha College of Engineering
Chikhli, Maharashtra, India

Dhiraj V. Bhise[2]
Assistant Professor
Department of Computer Science & Engg
Anuradha College of Engineering Chikhli, Maharashtra, India

Abstract: Big data is large volume, heterogeneous, distributed data. The primary sources for big data are from business applications, public web, social media, Weather Forecasting, and Electricity Demand Supply and so on. "Big data mining" involves knowledge discovery from these large data sets. For data processing Big data processing framework relay on cluster computers and parallel execution framework provided by Map-Reduce. This paper gives an overview of big data along with its challenges and characteristics.

Keywords: Big data, Data Mining.

I. INTRODUCTION

Today we are living in an era of digital world. With the rapid increase in digitization the amount of Structured, semi structured and unstructured data being generated and stored is exploding. In Big data the information comes from multiple, heterogeneous, autonomous sources with complex relationship and continuously growing. Every day, 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [1].for example Flicker, a public picture sharing site, where in an average 1.8 million photos per day are receive from February to march 2012[2]. Currently Big Data processing depends upon parallel programming models like Map Reduce, as well as providing computing platform of Big Data services. Thus making big data mining or knowledge discovery of large datasets a difficult process. Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter. The most fundamental challenge for big data applications is to explore the large volumes of data and extract useful information or knowledge for future actions.

There are different types of data such as relational, structural, textual, semi structured, graph data, streaming data etc can be included in big data.

II. BIG DATA

Big data refers to large data sets that are challenging to store, search, share, visualize and analyze. It is high volume, high velocity, high variety information assets that demand cost effective, innovative forms of information processing for enhance insight and decision making. Big data is coined to address massive volumes of data sets usually huge, sparse, incomplete, uncertain, complex or dynamic, which are mainly coming from multiple and autonomous sources. The 3Vs that define big data are Volume, Velocity and Variety [3].

» *Volume*

Volume means vast amount of data generated in every second [4]. It is a scale characteristic. The data is in rest state. Machine generated data are examples for these characteristics. Nowadays data volume is increasing exponentially.

» ***Velocity***

The second generated characteristics of big data are velocity or speed. Velocity is the speed at which data generated. The streaming data may not be massive and its state is in motion. It should have high speed data. Example is data created through social media. The data is begin generated fast and need to be processed fast. Online Data Analytics includes these types of big data. E-Promotions and health care monitoring are examples. In e-promotion, **b**ased on our current location and our purchase history, what we like will send promotions right now for store next to us. In Healthcare monitoring, sensors monitoring our activities and body. Any abnormal measurements require immediate reaction can be immediately identified through this.

» ***Variety***

Variety is another important characteristic of big data. Various data formats, types, and structures can be referred here. The type of data may include different verities such as Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…It also includes s static data and streaming data . A single application can be generating by collecting many types of data. To extract the knowledge all these types of data need to be linked together.

Now two more V's also contributed to big data. They are veracity and value of data [5].

» ***Veracity***

Veracity means data in doubt. The uncertainty of data can be found due to the inconsistency and incompleteness. The messiness of data (Abbreviation, colloquial speech etc) may result the veracity.

» ***Value***

Value gives importance to the profit gained by organizations who invest in Big Data technologies.

III. Challenges In Big Data Mining

Big Data has different characteristics such as it is large volume, heterogeneous, autonomous source with distributed and centralized control, seek to explore complex and evolving relationship among data [1].These different characteristics of Big Data make it challenge for discovering useful information or knowledge from it. After analyzing and research challenge form a three tier structure framework to mention different challenges at different tier, as shown in fig.1. The challenges at tier I focus on low-level data accessing and arithmetic computing procedures, Challenges on information sharing and Privacy. Big Data often stored on different location and it is continuously growing that's why an effective computing platform to take distributed large scale data storage into consideration for computing. Tier II concentrate on high-level semantics, application domain knowledge for different applications of big data and the user privacy issues. This information provides benefits to Big data access but also add a technical barriers to Big Data access (Tier I) and mining algorithms (Tier II). The Outmost tier is tier III which challenges the actual mining algorithms. At this tier III the mining challenges concentrate on algorithm designs in tacking the difficulties which is raised by the big data volumes, distributed data distribution, complex and dynamic characteristics. Tier III contains three stages. In first stage sparse, heterogeneous, uncertain, incomplete and multisource data is preprocessed by data fusion technique. In second stage after preprocessing stage complex and dynamic data are Mined. Third stage is for local learning and model fusion, where the global knowledge is obtained by local learning and model fusion is tested and the relevant information is feedback to preprocessing stage. Big Data is carry out computing on the PB (Pet byte) or even on EB (Exabyte) data with complex computing process, so parallel computing infrastructure, programming language support and software model utilizing to efficiently analyze and mine distributed data Map Reduce mechanism is suitable for large scale data mining task on clusters.

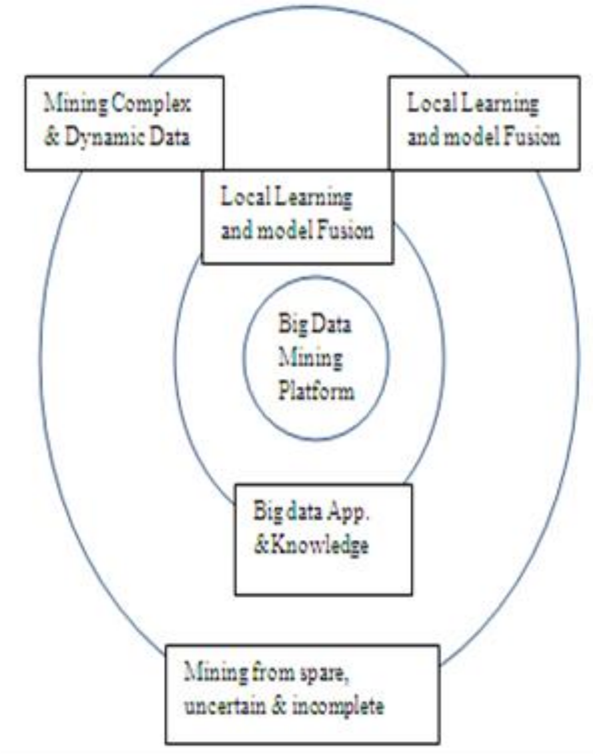
Mining Complex & Dynamic Data
Local Learning and model Fusion
Local Learning and model Fusion
Big Data Mining Platform
Big data App. &Knowledge
Mining from spare, uncertain & incomplete

Fig.1. A Big Data processing Framework

IV. BIG DATA SOURCES

The major sources of big data are from the following [3].

» ***Archives***

Archives are mainly maintained by organizations, to show the function of a particular person or organization functions. Accumulation of archives sometimes does not fit into the traditional storage systems and need systems with high processing capabilities. This voluminous archive contributes to big data.

» ***Media***

Users generate images, videos, audios, live streams, podcasts and so on contributes for big data.

C. Business applications

Huge volumes of data are generated from business applications as part of project management, marketing automation, productivity, customer relation management (CRM), enterprise resource planning (ERP) content management systems, procurement, human resource (HR), storage, talent management, Google Docs, intranets, portals and so on. These data contributes to big data

» Public web

Many organizations under government sector, weather, competitive, traffic regulatory compliance, healthcare services, economic, census, public finance, stock, open source intelligence (OSINT), the world bank, electronic data gathering analysis and retrieval (Edgar), Wikipedia and so on uses web services for communication. These data contributes to big data

» Social Media

Nowadays users rely on social media sites such as twitter, LinkedIn, facebook, tumblers, blog, slide share, YouTube, Google+, instagram, flicker, wordpress and so on for the creation and exchange of user generated contents. These social networking sites contribute to big data.

V. BIG DATA MINING

Useful data can be retrieved from this large datasets with the aid of big data mining [4]. Here the data which are handled is big data, hence the term big data mining. Usually, data mining is the technique of analyzing data from different prospects and summarizing these data into interesting, understandable and useful models. For better decision making, the large repositories of data collected from different resources require a proper mechanism for extracting knowledge from the databases. Since big data scales far beyond the capacity of single PC, cluster computers, which have high computing powers and rely on parallel programming paradigms, are used. Thus a large attempt to exploit these huge parallel processing architectures was initiated. Big [6] data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactiveness that existing mining techniques and algorithms are incapable of. The need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM)has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop Map Reduce. NIMBLE [7] is a portable infrastructure that has been specifically designed to enable rapid implementation of parallel MLDM algorithms, running on top of Hadoop. Apache's Mahout [8] is a library of machine learning and data mining implementations. The library is also implemented on top of Hadoop using the Map Reduce programming model. Some important components of the library can run stand-alone. The main drawbacks of Mahout are that its learning cycle is too long and its lack of user-friendly interaction support. Besides, it does not implement all the needed data mining and machine learning algorithms. BC-PDM (Big Cloud-Parallel Data Mining) [9], as a cloud-based data mining platform, also based on Hadoop, provides access to large telecom data and business solutions for telecom operators; it supports parallel ETL process (extract, transform, and load), data mining, social network analysis, and text mining. BC-PDM tried to overcome the problem of single function of other approaches and to be more applicable for Business Intelligence. PEGASUS (Pet-scale Graph Mining System) and Graph both implement graph mining algorithms using parallel computing and they both run on top of Hadoop. Graph Lab is a graph-based, scalable framework, on which sever all graph-based machine learning and data mining algorithms are implemented. Distinctive algorithms used in data mining are as follows:[10]

» ***Classification trees:***

A famous data-mining system that is used to categorize a needy categorical variable based on size of one or many predictor variables. The outcome is a tree with links and nodes between the nodes that can be interpret to form if-then rules.

» ***Logistic regression:***

A algebraic technique that is a modification of standard regression but enlarges the idea to deal with sorting. It constructs a formula that predicts the possibility of the occurrence as a role of the independent variables.

» ***Neural networks:***

A software algorithm that is molded after the matching architecture of animal minds. The network includes of output nodes, hidden layers and input nodes. Each unit is allocated a weight. Data is specified to the input node, and by a method of trial and error, the algorithm correct the weights until it reaches a definite stopping criteria. Some groups have likened this to a black–box system.

» ***Clustering techniques like K-nearest neighbors:***

A procedure that identifies class of related records. The K-nearest neighbor technique evaluates the distances between the points and record in the historical data. It then allocates this record to the set of its nearest neighbor in a data group.

VI. CONCLUSION

In real-world applications managing and mining Big Data is Challenging task, as the data concern large in a volume, distributed and decentralized control and complex. There are several challenges at data, model and system level. We need computing platform to handle this Big Data. In this paper, an overview of big data along with it characteristics, challenges and sources are discussed. It is known that big data mining is an emerging trend in all science and engineering domains and also a promising research area. In spite of the limited work done on big data mining so far, it is believed that much work is required to overcome its challenges related to the above mentioned issues.

References

1. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding" Data Mining with Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014
2. F. Michel, "How Many Photos Are Uploaded to Flicker Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.
3. SHERIN A1, Dr S UMA2, SARANYA K3, SARANYA VANI M4" SURVEY ON BIG DATA MINING PLATFORMS, ALGORITHMS AND CHALLENGES" Sherin A et al. / International Journal of Computer Science & Engineering Technology (IJCSET), ISSN : 2229-3345 ,Vol. 5 No. 09 Sep 2014
4. SMITHA T, V. Suresh Kumar, "Application of Big Data in Data Mining" ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 7, July 2013
5. Anuradha, G., "Suggested techniques for clustering and mining of data streams", Published in: Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on Date of Conference:4-5 April 2014.
6. B R Prakash1*, Dr. M. Hanumanthappa 2," Issues and Challenges in the Era of Big Data "ISSN 2278-6856, Volume 3, Issue 4 July-August 2014 Mining,
7. NewVantage Partners: Big Data Executive Survey (2013) http://newvantage.com/wpcontent/ Uploads/2013/02/ NVP-Big- Data-Survey- 2013-Summary-Report.pdf
8. Xin, R.S., Rosen, J., Zaharia, M., Franklin, M., Shenker, S., Stoica, I.: Shark: SQL and Rich Analytics at Scale. In: ACM SIGMOD Conference (accepted, 2013)
9. Agrawal, D., Bernstein, P., Bertino, E., et al.: Challenges and Opportunities With big data Community White Paper Developed by Leading Researchers Across the United States (2012), http://cra.org/ccc/docs/init/bigdatawhitepaper. Pdf
10. Vitthal Yenkar, 2Prof.Mahip Bartere," Review on Data Mining with Big Data", Vitthal Yenkar *et al*, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 97-102 ,ISSN2320–088X.