International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 5 (2015) pp. 13431-13438

http://www.ripublication.com

A Survey Paper on Big Data Analytics

R.Wesley Daniel, Prof D.Venkata Subramanian[2], E.Jasper Samuel[3], B.Sambalaji[4]

[1]M.E, Department Of CSE, Saveetha University, Saveetha Nagar, Thandalam, Chennai-602105, India
[2]Assistant Professor, Department Of CSE, Saveetha University, Saveetha Nagar, Thandalam, Chennai-602105, India
[3]M.E, Department Of CSE, Saveetha Engineering College, Saveetha Nagar, Thandalam, Chennai-602105, India
[4]M.E, Department Of CSE, Saveetha University, Saveetha Nagar, Thandalam, Chennai-602105, India
leydaniel07@gmail.com
venkata.dayanandan@gmail.com
jasperisaac17@gmail.com
sambalajibj06@gmail.com

Abstract-

The objective of this paper is to review the recent developments in big data, their challenges issues and applications of big data [1]. A huge amount of data is stored in rapid rate that become exceeding the limitations. This rapid growth of data will result in data storage problems, transfer speed and security issues. Hence this survey focuses on recent technology big data analytics, to confront the problems of managing huge volumes of data and their challenging issues. Processing large number of datasets using traditional classification and management tools becomes more difficult [2]. Big data enables to discover the relationship among the stored datasets to process the data effortlessly. This paper also discusses available management tools for big data analytics.

Keywords- Big Data, Data Analytics, Challegnges and Security issues.

I. INTRODUCTION

With an increase in the volume of data warehousing, new methodologies need to be designed and implemented to improve the standard and quality of data in industries [5]. As the database of any organization is the most essential entity and all decisions rely on the same, the data should be of certain acceptable quality. The data volume

tends to grow day by day. Some organizations have mechanisms to purge the database periodically and also backups are taken regularly. These operations have an impact on the performance of the database servers. The data needs to be organized effectively so as to extract useful and relevant information from the same [7]. Many important domains of human endeavour such as meteorology, emergency services, logistics, and science are increasingly dependent upon remotely sensed information. In this context, it is widely believed that the Internet of Things (IoT) paradigm will revolutionize these domains by enabling professionals in these and other similar domains to share and access remotely sensed data over the Internet in real-time [9].

Cosmo is an early attempt in this direction. It provides Restful services to upload and download sensor data. If business and technology trends in other computing domains are any indicators, the next few years will very likely see a cloud-based 'virtualization' of domain sensing functionalities. Service providers will independently install and maintain domain sensing infrastructures, and the data feeds from various sensors will be exposed as services over the Internet. Service consumers can embed these data feeds into their applications [12]. It is noteworthy that this is a manifestation of the Data as a Service (DaaS) model, the viability of which has been validated by the recent emergence of several data markets including Azure Data Market and Info Chimps.

In consumer world, the amount of data about consumers, suppliers has been exploding through millions of networked sensors and consumer video surveillance systems. Multimedia and individuals on various social networks platforms will continue to increase the exponential growth. The unedited, user generated information of these social network platform include both structured and unstructured data which fall into a category of Big Data analysis. In general, Big Data is characterized on three dimensions of growth of data known as the 3V: Volume, Variety, and Velocity. On the other hand, consumer behaviour analysis is concerned with the study of inter actions among the consumers, products and operations such as purchasing, saving, brand choice and etc. Moreover, consumers are no longer what they used to be. Today's consumers have evolved beyond being merely 'buyers'. So, more insights information is necessary for analysing consumer behaviour. In this aspect, Big Data has become a central role for making data driven decision making processes [14]. The nature of data collected in the big data platform is significantly different. To analyse the Big Data, both in terms of quantity and the nature of the data, is a challenging task.

II. BIG DATA ANALYTICS

The following section represents the big data analytics stages to process the data. The first stage in big data analytics is to collect the raw data. The raw data is collected from all units of organizations. Then the value is added to raw data by some useful inputs. Large volumes of data can be collected from various sources like log files, internet, sensing sensors, satellites, mobile equipment, laboratories, chat messages, super computers [15]. The next stage involves filtering, in this stage noisy and irrelevant data are removed from the collected raw data through data preprocessing

methods. The third stage involves the classification of collected and filtered raw data. The classification is based on similarity between the data's and relevant to the data analysis stage. For example, if big data analysis is about finding total sales of an organization the grouping and classification of data is based on previous sales history.

After the raw data is collected, filtered and classified the next step is to analyze the collected data. The analysis of data is performed to handle the large amounts of data in an organizations as well as to analyze the existing relationship between collected data and to predict the future business developments. The analysis tools should be capable of processing both structured and unstructured data [16]. The various available for data analyzing are data mining, statistics analysis, visualizations, machine learning and cluster analysis. Among these available tools data mining is using more for analyzing data because of its ability to identify useful patterns. Big data analysis should be capable of analyzing both large volumes of data and data in different formats. So big data analysis tools must be able to adapt the architecture that will support all different analysis techniques like statistical analysis, machine learning, and visualizations [11]. Different storage systems should included for effectively storing different formats of data for big data analysis. The accessibility of data should be easy and fast for effective analysis of data. The different analysis methods are discussed below.

A. DATA MINING:

Data mining tools effectively identify the hidden valuable information from the large, fuzzy, incomplete, inaccurate and noisy data. The effective data mining algorithm includes SVM, apriori, CART, K-means, EM, naive bayes algorithms.

B. STATISTICAL ANALYSIS:

In statsitical analysis uncertatinity and randomness of data and results are analysed based on probability therory. There are two types of statistical analysis namely, inferred and described analysis methods. In inferred analysis methods the possible conclusion are drawn from the available datasets. In described analysis method, can be used to describe and summarize the data.

C. CLUSTER ANALYSIS:

Clustering is the process of grouping the similar datasets suchthat objects within the group are more similar than other data objects. The cluster analsysi follows unsupervised maching learning techinques where no training datasets are required.

D. CORRELATION ANALYSIS:

The correlation anlaysis is perfomed to analyze the relation between the data objects. The relation between two data objects is represented by numerical values. This correlation analysis is used to predict and control the data.

E. REGRESSION ANALYSIS:

Regression analysis is a mathematical method used to identify the hidden, complex and undetermined correlations between variables [17].

These are the various analysis techniques. The process of big data analytics is shown in figure 1 below.

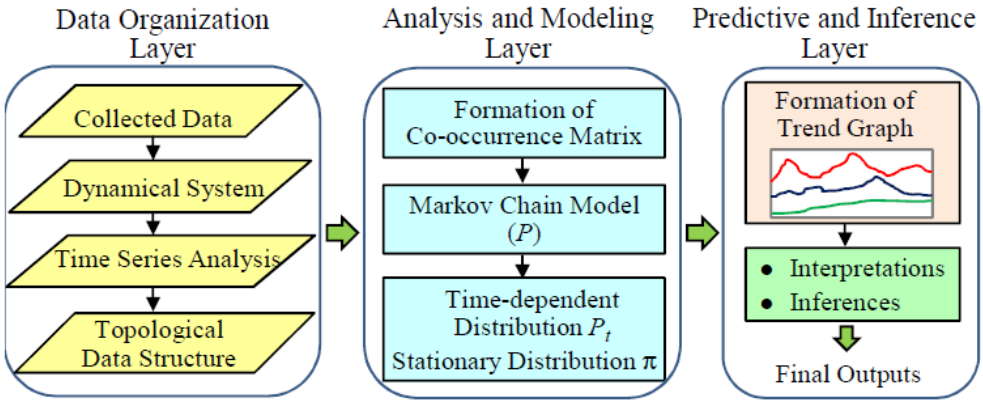

Figure 1:PROCESS OF BIG DATA ANALYTICS

III. CHALLENGES IN BIG DATA

There are so many challenges and security issues in big data analytics. The challenging issues are summarized in table 1 below.

TABLE 1 CHALLENGES AND SECURITY ISSUES IN BIG DATA

SNO	CHALLENGES	EXPLANATION
1	Scalability	Analyzing and managing large volumes of data requires big data to implement less time consuming easy navigating between large datasets is needed.
2	Accuracy	Analysis of data with increasing number of large datasets is extremely difficult. Finding the accurate and relevant data is difficult.
3	Complexity	Big data analysis tools finds difficult to analysis un-structured data's. An un-structured data includes videos and images.
4	Heterogeneity	Big data algorithms must be able to locate unknown and different patterns.
5	Privacy	The main problem in big data is data leakage. Where the confidential data may become publicly available. So maintaining the privacy of data is more important.
6	Availability	Data should be available to authorized users within the stipulate time. Providing analyzed data is difficult.
7	Integrity	In large scale operating organizations data may updated frequently. In such situation providing accurate data in a timely fashion is difficult.
8	Recovery	In case of any disasters like power failure or natural calamities may leads to the loss of data. So protecting data by implementing replication of data and recovering the loss data is more important in big data analytics.

IV. MANAGEMENT TOOLS IN BIG DATA

Because of popularity of internet increases rapidly, it is very difficult to manage large volumes of data without the need of sophisticated tools at a low cost. Big data requires multiple numbers of systems to store, process, analyses large volumes of data in real time [20]. Each of the organization should adapt their own management tools based on their functional activities and requirements. There are more numbers of tools available for big data analytics. These tools and their functions are summarized in table 2.

TABLE 2 MANAGEMENT TOOLS OF BIG DATA

SNO.	TOOLS	DESCRIPTION
1	Hadoop	Hadoop is a java application, hadoop tool is used for spam filtering, and network finding and supports distributed processing on data.
2	HDFS	HDFS tools are useful when a system finds difficult to manage the data. All HDFS are replicated to perform parallel processing on data.
3	ZooKeeper	Zoo keeper tool is used to maintain, names and configuring large volumes of data using distributed processes.
4	HCatalog	Hcatalog is used to manage data's in HDFS and stores metadata and produce tables for stored data.
5	HBase	Hbase is open sourced, versioned management system used to manage the data. Hbase is depends on zoo keeper instances.
6	Pig	Pig tools is used to produces a scripting language and run time environment that enables users to execute map reduce on hadoop.
7	Hive	Hive tool is used to structure the data's stored in hadoop.
8	Avro	Avro serializes data and performs procedural calls to transfer data from one programming language to another. Data's are stored based on their own schema.
9	Mahout	Mahout tools acts as a library for data mining and machine learning. This tool is responsible for collection, tagging, grouping and mining of data.
10	Chukwa	Chukwa tool is responsible for pooling data from distributed system and analyzing it. Chukwa is related to HDFS and Map reduce tools.
11	Flume	Flume tool is used to cumulative and transfer of data to and from Hadoop.
12	Oozie	Oozie is used to manages, coordinates and executes the work flow of the process.

Fig 1: Architecture diagram for message passing based on clustering.

V. CONCLUSION

In this paper, we discuss the fundamental concepts of big data, big data analytics process, challenges and security issues in big data and available management tools in big data. Many scientific and research application depends on the analysis of large volumes of data for their progress. So attempts to be taken to overcome the challenges and security issues discussed in previous sections. Building a management tool for big data applications is a crucial one. Further research works to be conducted to explore more big data applications.

REFERENCE

[1] Antonio Criminisi, Jamie Shotton, Ender Konukoglu. "Decision Forests: A unified Framework for Classification, Regression, Density Estimation, Manifold learning and Semi Supervised Learning", Journal Foundation and Trends in Computer Graphics and Vision, Vol 7, Issue 2-3, Feb 2012

[2] Bettina Berendt, Soren Preibusch. " Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence", Journal Artificial Intelligence and Law, Vol 22, Issue 2, pp 175-209, ACM, Jun 2014

[3] Christos Doulkeridis, KjetilNorvag. "A survey of large-scale analytical query processing in MapReduce", Journal The VLDB journal – The International Journal on Very Large Databases, Vol 23, Issue 3, June 2014.

[4] EvgeniyYur'evichGorodov, VassilyVasil'evichGubarev. "Analytical review of data visualization methods in application to Big Data", Journal of Electrical and Computer Engineering, Article No. 22, Vol 2013, ACM, Jan 2013

[5] FloareaSerban, Joaquin Vanschoren, Jorg-Uwe Kietz, Abraham Bernstein. "A survey of intelligent assistants for Data Analysis", Journal ACM Computing Surveys (CSUR), Vol 45, Issue 3, Article No. 31, ACM, June 2013.

[6] JawwadShamsi, Muhammad Ali Khojaye, Mohammad Ali Qasmi. "Data Intensive Cloud Computing: Requirements, Expectations, Challenges and Solutions", Journal of Grid Computing, Vol 11, Issue 2, pp 281-310, ACM, Jun 2013.

[7] Liam Evans, NielsLohse, Mark Summers. "A fuzzy-decision-tree approach for Manufacturing technology selection exploiting experience-based information",

Journal Expert Systems with Applications: An International Journal, Vol 40, Issue 16, pp 6412-6426, ACM, Nov 2013.

[8] Madhushri Banerjee, Zhiyuan Chen, AryyaGangopadhyay. "A generic and distributed privacy preserving classification method with a worst-case privacy guarantee", Journal Distributed and Parallel Database, Vol 32, Issue 1, pp 5-35, ACM, March 2014

[9] MarcioP.Basgalupp, Rodrigo C. Barros, Andre C.P.L.F De Carvilo, Alex A.Freitas. "Evolving Decision Trees with beam search based initialization and lexicography multiobjective evaluation", Journal Information Sciences, Vol 258, pp 160-181, ACM, Feb 2014

[10] Ohbyung Kwon, Jae MunSIm. "Effects of data set features on the performances of classification algorithms", Journal Expert Systems with Applications: An International Journal, Vol 40, Issue 5, pp 1847-1857, ACM, Apr 2013.

[11] Stefan Strohmeier, Franca Piazza. "Domain Driven Data mining in Human Resource Management: A review of current research", Journal Expert Systems with Applications: An International Journal, Vol 40, Issue 7, pp 2410-2420, ACM, Jun 2013.

[12] Venkata Subramanian D, Angelina Geetha and Mohammed Hussain, "Measurement Process and Multi-dimensional Model For Evaluating Knowledge Management Systems", IEEE International Conference on Research and Innovation in Information Systems, 2011, pp 128-133, 2011.

[13] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy- Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[14] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[15] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems, vol. 30, no. 1, pp. 71- 88, 2005

[16] K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

[17] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[18] D. Howe et al., “Big Data: The Future of Biocuration,” Nature, vol. 455, pp. 47-50, Sept. 2008.

[19] A. Labrinidis and H. Jagadish, “Challenges and Opportunities with Big Data,” Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.

[20] Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining,” J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.