IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 6, June 2017

Big Data Challenges and its Tools

A.M. Chandrashekhar[1], Arpitha M G[2]

Assistant Professor, Department of Computer Science & Engineering, Sri Jayachamarajendra College of Engineering (SJCE), JSS S & T University Campus, Mysore, Karnataka, India[1]

M.Tech 2nd Semester, Department of Computer Science & Engineering, Sri Jayachamarajendra College of Engineering (SJCE), JSS S & T University Campus, Mysore, Karnataka, India[2]

Abstract: As number of user's increases data also increases. The phrase big data refers to bulk volume of data which is complicated in nature because it involves both structured and unstructured type of data which is difficult to analyze. Many different sources like social media postings, sensors which gives climate information, digital information etc contribute huge amount of data to the big data. To extract useful patterns and readable patterns from the big data it is necessary to use data mining technique. Big data stream refers to the data stream which is large in variety, veracity and velocity. Data processing is not an easy job unless of identifying, responding and locating the data hence it is more challenging and difficult it has to be done in well automated manner.

Keywords: Hadoop, no sql, pig, spark, 3v's.

I. INTRODUCTION

Big data mainly focus on data, not on borrowed data and that data is inflexible. Any data that crosses the fixed threshold can be referred to as 'big' in big data.
Lifecycle of data is composed of 4 A's

- Acquisition: Collection of scattered data
- Aggregatio: Integrated data is sent to analysis
- Analysis: Deals with extraction of knowledge
- Application: Log data is sent to acquisition.

At present we are facing a flood of data because we are getting billions and millions of data from many sources, big data is appearing in many vocabulary it varies from finance, business, genomics, meteorology, complex physical simulations etc. Examples of big data sources are NYSE [New York Stock Exchange] generates about 1TB, Social media impact like face book generates data in terms of photo, video, uploads which is more than 500TB, Single jet engine produces more than 10TB in just half an hour.
So the amount of data is exceeding the limitation of software tools hence it is facing many challenges such as sources, analysis, search, sharing, information security etc. Since the data is not in proper form then it will be not fruitful for advance use, To get convenient strings we need some tools/techniques to hold this type of data. Data warehouse cannot handle this semi structured and un structured type of data and this data warehouse cannot be able to process the appeal of big data that is modification, deletion or insertion since it is not in correct form.

Analyzing and extracting particular data is difficult to our brain to perceive from huge amount of data hence it need some advanced methods. Most important challenge is to inspect huge data and to get user expected data for decision making by using mining technique. Data mining is for some limit storing if it crosses that limit then all data in it lead to unachievable. Data evolved from Internet of Things will increase aggressively as the number of linked node increases. Key organizer for the advancement of big data is

- Development of storage quality
- Development of handle power
- Vacancy of data

Computational view of big data is meant for storage, lead to formatting, cleaning, data understanding accessing the data and last stage is data visualization.

II. BIG DATA THREATS AND SOLUTION

Threats in big data need to be concentrated otherwise it lead to downfall of the technology implementation and some unsatisfied output.

DOI10.17148/IJARCCE.2017.6608

The brief review of various issues is as follows:

Issues related to characteristics

Data volume as mentioned earlier it is boosting up for every second from various sources hence it is difficult to handle and maintain by this existing traditional system. Many website like E-commerce has increases their speed, E.g. website clicks. Hence data velocity management is much more than a bandwidth issue.

Storage and management issue

Amount of data has collapsed each time lead to invent a storage medium. Recent data collection is due to social media and data is creating from everyone and everywhere including journalist and writer.

Present disk technology only limited to 10^{12} per disk. But amount of data evolving is more compared to storage capacity. Assume that 1TB per second network has an effective transfer rate of 70%, bandwidth is about 100 MB thus transferring a data take too much of time. Hence it takes more time to transmit the data from a collection to a transform point than transform it. To solve this issue data should be transferred “in place” and send only concluded report.

Data administration issue

This is one the big problem related to big data. As mentioned earlier variety of data is from various sources and varies by size, format, type, and assemblage method. Till today there is no perfect administration explanation yet. This leads to a difference in the research literature on big data that needs to be filled.

Issues related to confidentiality and guarantee

Person information log in a database need to be maintained confidentially and the worker don’t want to reveal this information to others or to owner.

Solution to big data problems

Huge blocks of data can be handled by hadoop where hadoop is a system used to save and process big data today many organizations are coming forward to handle the big data and providing many technologies and tools yam ,spool ,spark, pig and no sql database. This no sql database does not require any kind of relationship, usual way to store data in this method is key: value pair, implementation of this very similar to hadoop.

Hadoop is important because of its store and process of any kind of data, computing power, flexibility, low cost, and scalability. The most powerful mining of big data is to implement the competitive advantage and to compute value for many regions. Figure 1 shows big data processing framework in which data approach and distribution of information is said to be used to make decisions in time, only when data is authenticate, completes in timely manner

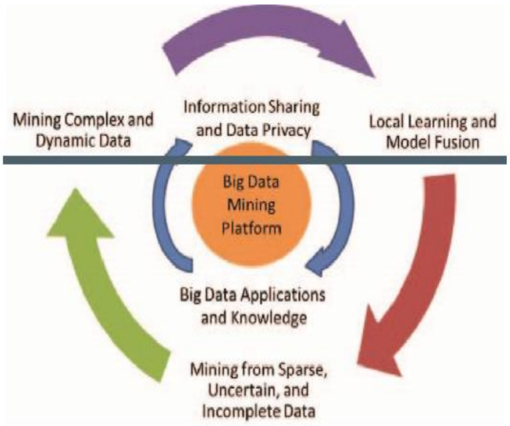
Mining Complex and Dynamic Data
Information Sharing and Data Privacy
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

Fig 1: Big data processing framework

Big data is mainly specified by its 3 v's

Velocity: It defines rate at which data is growing, nothing but speed of data from heterogeneous sources, speed not only restricted to coming data but also the flow of data and combined data.

Variety: This shows the richness of data that is type of data either text, image, audio, video this not only contains structured type of data but also semi structured type of data E.g. web files

Volume: Vast amount of data available in organization. Big data practical responses are, it provides platform for the data before data stored in data warehouse. Many organizations are still trying to establish convenient mechanics for big data.

DOI10.17148/IJARCCE.2017.6608

Old methods are not adapted to shared environment and big data complication. Enterprise need to execute queries through large volumes of unstructured data groups. This advanced to improvement of scalable browser based on particular searching technologies. It needs more advanced methods to achieve reliability and scalability

III. LITERATURE SURVEY

Web crawlers [1] are needed to various Web applications, such as Web browsers, Web archives, and Web notes, which preserve Web pages in their local store house. In this paper, we study the complication of crawl organizing that biases crawl ordering toward important pages. We advances a well set of crawling algorithms for sufficient and economical crawl ordering by precedence important pages with the well-known Page Rank as the relevant metric. In order to score URLs, the proposed algorithms uses variety features, including partial link structure, inter-host links, page titles, and theme relevance. We conduct a large-scale experiment using publicly available data sets to examine the effect of each feature on crawl ordering and evaluate the performance of many algorithms. The experimental results verify the efficiency of our schemes.

A nature-motivated theory to model aggregate behavior from the inspected data on blogs using swarm bright power, where the intent is to exactly model and judge the future observance of a large society after penetrating their communications during a training phase [2]. Clearly, an ant colony optimization model is trained with behavioral trend from the blog data and is tested over real-world blogs. Rising results were accomplished in trend prediction using ant colony based phenomenon classifier and CHI statistical measure.

Hadoop is necessary because of many useful properties like computing power, flexibility, low cost and scalability over the previous decade, there has been an ignition of passion in network experimentation beyond the physical and social sciences[3]. Here, we check-up the kinds of things that social scientists have tried to justify using social network analysis and present nutshell explanation of the basic assumptions intent, and explanatory mechanisms prevalent in the field a assembled advance to training a Bayesian network from shared different types of data. Bayesian network is learnt at the midway site using the data broadcast from the local site[4]. The entire data is modeled by merging both central and local Bayesian network to get a group of Bayesian network; preparatory outputs and theoretical notification that dispose the feasibility of our access are granted.

IV.PROPOSED WORK

In proposed system to build a branch-based Big Data analytic skeleton for rapid reply and real-time ruling. The key demanding and investigation problem append designing big data examine workings to minimize big data content to a manageable size for preparing; - framework guess models from big data streams.

This type of models can flexibly unite to the dynamic alteration of the data. Figure 2 shows proposed architecture A awareness indexing skeleton to confirm real-time data governing and dividing for Big Data utilizations.

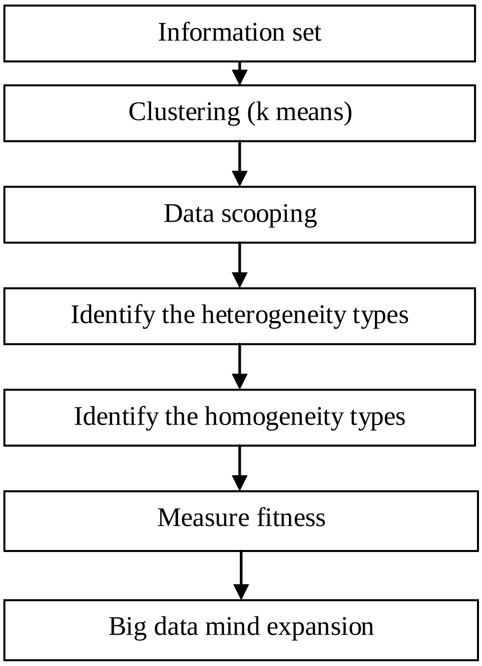

Fig 2: Proposed architecture

V. SCOPE OF THE WORK

Big Data look for complex, huge, enlarging data sets with various sources. Big Data are now vastly growing in all the fields, which includes biomedical, biological and physical sciences because of rapid advancement of networking, data clustering capacity. Benefits of big data handling are business can take desired decisions, developed and updated customer service, traditional comment has been recovered by big data automation this technology has been used to read and analyze the customer feedback and replies. Big data practical response is to provide platform for the data before data stored in data warehouse.

VI. METHODOLOGY

Many areas are there for research to improve the features and capabilities of big data. The proposed work mainly concentrates on platform for data acquisition, storage, transmission and large scale processing mechanism.
The work involves performance analysis at each step for the processing time, efficiency, scalability and flexibility. This proposed work mainly concentrates on the platform in big data for further improvement; it consists of mainly three phases as described below:

Data Acquisition: refers to the process of obtaining information and it contains the data collection, data transmission and data pre-processing units.

Data Processing: refers to integration and storage of data to obtain process data using different methods of transformations.

Data Services/Retrieval: provides the different services like access and use of data for the user. This acts like an interface for the user to collect and use the data from the different sources without much delay

VII.DISCUSSION

Even though big data has many advantages it has many challenges, To achieve big data mining platforms are required. At certain level many sources may results in missing some of the values. In some of the cases unwanted data, errors can be introduced into the original information so introducing a well good and safe protocol is a big challenge.

At the model level the main challenge is to combine discovered patterns to form a defined view. for this it needed an algorithm to combine decisions from variety of sources to produce a good model out of the big data. At some level big data mining main work is to frame a skeleton that needs to consider a difficult communication between models, various sources and samples along with some changes that is adopted with time and other aspect. a system that need to be developed such that unstructured data should be converted to useful patterns and this patterns should be useful in future.

Due to big data trend there is a demand for big data mining hence it is evolving in all the fields. This big data mining technologies helping users to better understand the useful pattern from vast data and to get a good feedback about the big data. From this many uses from big data mining technologies we can say that big data era has been occurred.

REFERENCES

[1] M.H. Alam, J.W. Ha, and S.K. Lee, “Novel Approaches to Crawling Important Pages Early, “Knowledge and Information Systems,vol. 33, no. 3, pp 707-734, Dec. 2012

[2] S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks,”Science,vol. 337, pp. 337-341, 2012.

[3] S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks,”Science,vol. 337, pp. 337-341, 2012

[4] S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems,vol. 33, no. 3, pp. 523-547, Dec. 2012.

[5] Marco Viceconti,Peter Hunter and Rod Hose, "Big Data, bi knowledge: big data for personalized health care", IEEE Journal of Biomedical and Health Informatics, Vol. PP, Issue 99, February 2015.Fong, S. Simon Fong, Wong R, Vasilakos A. V, "Accelerated PSO

[6] Hao Zhang, Gang Chen, Beng Chin Ooi, Kian-Lee Tan and Meihui Zhang, "In-Memory Big Data Management and Processing: A Survey", IEEE Transactions on Knowledge and Data Engineering,Vol.27, No.7, July 2015

[7] A. M. Chandrashekhar and K. Raghuveer, “Confederation of FCM Clustering, ANN and SVM Techniques of Data mining to Implement Hybrid NIDS Using Corrected KDD Cup Dataset”, Communication and Signal Processing (ICCSP) IEEE International Conference,2014, Page 672-676.

[8] AM. Chandrashekhar and K. Raghuveer , “Improvising Intrusion detection precision of ANN based NIDS by incorporating various data Normalization Technique – A Performance Appraisal”, IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014.

[9] A. M Chandrashekhar A M and K. Raghuveer, “Diverse and Conglomerate modi-operandi for Anomaly Intrusion Detection Systems”, International Journal of Computer Application (IJCA) Special Issue on “Network Security and Cryptography (NSC)”, 2011.

DOI10.17148/IJARCCE.2017.6608

[10] A. M Chandrashekhar A M and K. Raghuveer, “Hard Clustering Vs. Soft Clustering: A Close Contest for Attaining Supremacy in Hybrid NIDS Development”, Proceedings of International Conference on Communication and Computing (ICCC - 2014), Elsevier science and Technology Publications.

[11] A. M. Chandrashekhar and K. Raghuveer, “Amalgamation of K-means clustering algorithem with standard MLP and SVM based neural networks to implement network intrusion detection system”, Advanced Computing, Networking, and Informatics –Volume 2(June 2014), Volume 28 of the series Smart Inovation, Systems and Technologies pp 273-283.

[12] A. M. Chandrashekhar and K. Raghuveer, “Fusion of Multiple Data Mining Techniques for Effective Network Intrusion Detection – A Contemporary Approach”, Proceedings of Fifth International Conference on Security of Information and Networks (SIN 2012), 2012, Page 178-182.

[13] A. M. Chandrashekhar, Jagadish Revapgol, Vinayaka Pattanashetti, “Big Data Security Issues in Networking”, International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Volume 2, Issue 1, JAN-2016.

[14] Puneeth L Sankadal, A. M Chandrashekhar, Prashanth Chillabatte, “Network Security situation awareness system” International Journal of Advanced Research in Information and Communication Engineering(IJARICE), Volume 3, Issue 5, May 2015.

[15] P. Koushik, A.M.Chandrashekhar, Jagadeesh Takkalakaki, “Information security threats, awareness and cognizance” International Journal for Technical research in Engineering(IJTRE), Volume 2, Issue 9, May 2015.

[16] A.M.Chandrashekhar, Yadunandan Huded, H S Sachin Kumar, “Advances in Information security risk practices” International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5, May 2015.

[17] A. M. Chandrashekhar,Muktha G, Anjana D, “Cyberstalking and Cyberbullying: Effects and prevention measures”, Imperial Journal of Interdisciplinary Research (IJIR), Volume 2, Issue 2, JAN-2016.

[18] A.M.Chandrashekhar, Syed Tahseen Ahmed, Rahul N, “Analysis of Security Threats to Database Storage Systems” International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5, May 2015.

[19] A.M.Chandrashekhar, K.K. Sowmyashree, RS Sheethal, “Pyramidal aggregation on Communication security” International Journal of Advanced Research in Computer Science and Applications (IJARCSA), Volume 3, Issue 5, May 2015.

[20] A.M.Chandrashekhar, Rahil kumar Gupta, Shivaraj H. P, “Role of information security awareness in success of an organization” International Journal of Research(IJR), Volume 2, Issue 6, May 2015.

[21] A.M.Chandrashekhar, Huda Mirza Saifuddin, Spoorthi B.S, “Exploration of the ingredients of original security” International Journal of Advanced Research in Computer Science and Applications(IJARCSA), Volume 3, Issue 5, May 2015.

DOI10.17148/IJARCCE.2017.6608