A Survey on Challenges In Bigdata

Santhosh Pawar
Dept of CS
KSWU, Vijayapura

Abstract- *Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. There are five dimensions to big data known as Volume, Variety, Velocity and the recently added Veracity and Value.*

Big data refers to volume of data which is complicated in nature because it involves both structured and unstructured type of data which is complex to analyze. Many different sources like social media postings like facebook , sensors which gives climate information, digital information etc contribute huge amount of data to the big data.We use data mining to extract useful patterns and readable patterns from the big data it is necessary to use data mining technique.

Keywords- Bigdata, Volume, Velocity, Variety, Hadoop, analysis.

I. INTRODUCTION

Big data is a term that describes the large volume of data with both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves .

Volume. Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

Velocity. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

we consider two additional dimensions when it comes to big data:

Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data.

Complexity. Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

Lifecycle of data is composed of 4 A's

Acquisition : Collection of scattered data
Aggregation : 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 billions 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

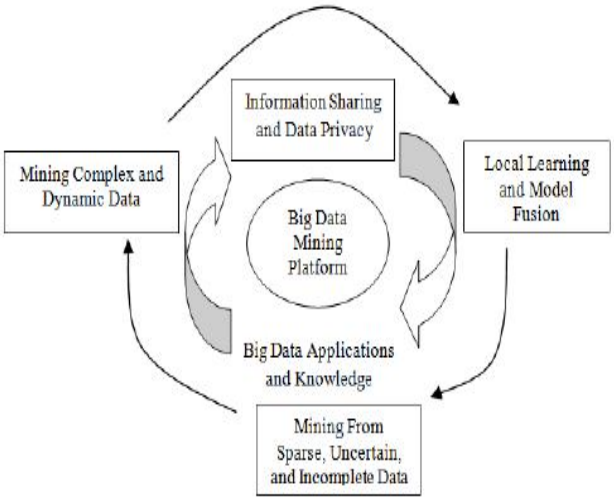
Computational view of big data is meant for storage, lead to formatting, cleaning, data understanding accessing the data and last stage is data visualization.

II. BIG DATA THREATS

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

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 1012per 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



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.

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 towards 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 classier 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. CONCLUSION

The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making. When you combine big data with high-powered analytics, you can accomplish business-related tasks such as:

- Determining root causes of failures, issues and defects in near-real time.
- Generating coupons at the point of sale based on the customer's buying habits.
- Recalculating entire risk portfolios in minutes.
- Detecting fraudulent behavior before it affects your organization.

Because of more demand for Bigdata and Applications to recognize the useful pattern we use data mining in bigdata.

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.