A SURVEY ON BIG DATA: IN REAL TIME, CHALLENGES AND OPPORTUNITIES

Shridevi Erayya Hombal
Department of Computer Science and Engineering
PESIT, Bangaluru South Campus
Bangaluru, India
shridevieh@gmail.com

Abstract- Big Data is the most challenging area to consider as it cannot be processed using traditional relational database management system. For some extent we can agree that big data leads to several benefits with the hope that it does not cause any harms meanwhile, it is essential to consider whether its changes are good or bad to society. This paper describes new challenges and opportunities of big data. Big data poses several challenges in terms of storage, management, analytics, networking and ethics. Several organizations use various analytics techniques to analyse the people sentiments and behaviour, those techniques ranges from crowdsourcing to genetic algorithms to neural networks to sentimental analysis. This data is collected from numerous sources including sensor networks [1], government data holdings, company market lead databases, and public profiles on social networking sites. The complexity level arises during data acquisition unit, when the real time or offline data requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to generate accurate metadata that describes the composition of data and the way it was collected and analysed.

Index Terms- Big Data, Genetic Algorithms, Sentimental Analysis, Neural Networks, Data Acquisition.

I. INTRODUCTION

Most of the data which is generated is streaming data [1] coming from sensors or web. Real time analytics is needed to analyse the data which is generated from applications such as sensor networks [2], manufacturing processes, call centre records, email, blogging, twitter posts and others. In real time processing the data arrives at high speed it degrades the system performance hence better algorithm need to be designed which utilizes the space and time in an efficient manner. As real time

data is a type of data stream, it poses several challenges while designing data mining algorithms [3]. The decision making is the most important part of the big data. Long back, decisions made were based on guesswork, now it is based on the data. The analysis of data is a major challenge as it involves many different phases each phase introduces challenges. The major steps of data analysis are shown in below figure Fig.1

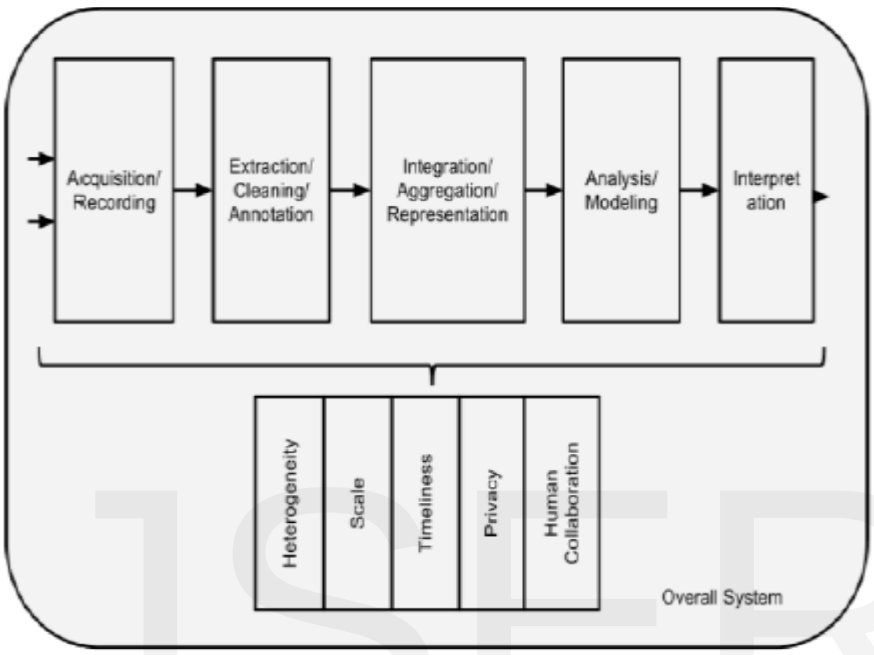

Figure 1. The Big Data Analysis Pipeline

The computational techniques can be applied to solve the big data problem. This paper, first gives the functionalities of each module and then considered the challenges [4] to exploit the big data and further gives the brief description about the opportunities exist in the big data domain.

II. LITERATURE SURVEY

Deploying an application leads to several challenges such as supporting applications which requires large scale of heavy updates and analytics of ad-hoc queries and decision support. Database management systems are a critical part in cloud infrastructure as it is using for decision support systems [5] and also application which requires heavy updations and also in deep analytics. So much research is happening in the field of scalable data management, to as it is required to analyse the data which is accumulated from different sources. This paper gives scalable analysis of the data using advanced technologies by making survey in this domain. The examination has carried out in the design which has carried out in the past in the domain of data management systems, and analysed access patterns and application demands of the application. Huge information is characterized basically by three parts which are volume (bulk datasets), velocity (real time information which is

gathered at the server farm) and variety (different classes of information from various sources) [6]. Many-sided Quality and variability are additionally two other characteristics [5]. Due to development in the information, there is need to take every necessary step in the area of information examination by various space people groups, for example, PC researchers, specialists and analysts Cost viable and convenient way enormous information investigation is understood in any field, for example, designing, business, experimental, government or instructive framework. Hadoop gives surely understood usefulness, for example, map diminish programming model which is executed by run time framework and it is pluggable capacity framework, is an understood choice to do information investigation. Hadoop gives the better execution and uses less assets, cash and time.

Cohen et al. introduced the term called MAD which means Magnetism, Agility and Depth to convey the characteristics that has been outputted once the analytics has done to the user [2].These terms are explained in detail below.

Magnetism: As the property of magnet which is pulling all sources of data without considering the missing qualities that keeps the useful information away or unknown pattern.

Agility: It keeps the information framework in a state of harmony/progressive with all information development.

Depth: All traditional database mechanism such as rollups and drilldown are replaced by new statistical techniques and a set of machine learning algorithms.

Hadoop gets the information and applying the map reduce programming model to interpret during the processing time not in the loading time respectively. Third Map Reduce calculations in Hadoop can be expressed directly in general purpose programming languages like Java or Python, domain-specific languages like R, or created consequently from SQL-like declarative languages like HiveQL and Pig Latin.

H. Herodotou et al., presents the Starfish which is a self-tuning framework for huge information analytics [7]. Starfish is intended to satisfy the client needs and to enhance the execution of information examination and it expand on the hadoop and it is a self-tuning framework for huge information investigation. Including MAD there are three more features are also important those are Data-lifecycle awareness, Elasticity, and Robustness. Thus the framework can be reclassified as MADDER[4]. The additional three properties are explained below.

Data-lifecycle awareness: As the information is originating from numerous assets, for example, key-esteem pair frameworks, databases or logging frameworks. This information is exchanged to the

examination framework for preparing. When it is done then it is exchanged back to the client confronting framework from nearer real time. Every day terabytes of information is going under this cycle[12]. Thus information life cycle awareness is required.

Elasticity: An Elastic framework modifies the expense of resources and time according to the necessities and workloads.

Robust: A powerful framework gives the administrations even if there should be an occurrence of programming bugs [8], corruption in the information or device failures As in the mid-2000, Google[9] faced the challenge of sorting the world's data subsequently it gives two key services to take care of the issue of effectively performing the crawling, indexing and replicating. The first is GFS (Google File System) which is scalable and fault tolerant and another is Map Reduce which isolates the task among numerous servers and performs the parallel processing.

Map Reduce is a programming model [12] used to process huge amount of datasets that has scattered over variety of real-world tasks. Clients determine map and reduce tasks, inner run time framework additionally handles the disappointments of machines and keeps schedule for inter-machine communication to make productive utilization of disks and system data transfer capacity. Over the past four years, Google implemented around ten thousand different map reduce programs and around ten thousand map reduce functions are running everyday on the Google's clusters which processes around twenty peta bytes of data Presently satellite imagery information is extended its area. Distinctive sort of databases which comprises of picture information needs a quick handling with a few fields. Remote sensing information is spreading its presence more in most recent couple of years.

Dhananjay G. Deshmukh presents a technique to perform unsupervised learning over this kind of symbolism data [10]. The execution is done utilizing C dialect and created utilizing AJAX, HTML, JavaScript and other web programming dialect. As constant information is getting produced massively it is important to guarantee about whether information is amassed and accumulated effectively with all components. Continuous data [11] comprises of numerous elements as of late it is testing errand to evacuate the components or qualities according to our advantage. There is a need of planning a design to save the components and extricating just required elements.

III. PHASES IN THE PROCESSING PIPELINE

- **Data Acquisition and Recording:**

As Big data is generated from some data generating source. Scientific experiments and simulations can easily produce petabytes of data today. Most of the data which is generated is of no

use hence, it needs to be filtered out. One challenge is to defining the filters which does not discard the useful information. As the data is online hence we need online analysis techniques [13] to process the streaming data on the fly.

One more important challenge is to generate the metadata in order to describe how it is recorded and measured and what type of data is recorded. One important issue with the metadata is data provenance. The recording information about the data at its birth is not useful unless it can be interpreted and carried through the data analysis pipeline. If any error happens during the processing at one step it may leads to subsequent analysis useless. With the suitable provenance, it is possible to identify the subsequent processing that depends on this step. Hence, research is needed in the field of generating the suitable metadata and the data systems that carry the data provenance and its metadata through the data analysis pipeline.

- **Information Extraction and Cleaning:**

As the data is generated this is not be in an understandable format. Hence information extraction process is needed that pulls out the required information and expressed in the form of structured format, which is suitable for future data analysis. The data extraction process is highly application dependent. Doing this correctly and completely is a technical challenge.

- **Data Integration, Aggregation, and Representation**

Data analysis is a challenging area to consider. For large scale data analysis, the data structure and semantics need to be expressed in the forms that are computer understandable. There is a strong research is going on in the field of data integration that can provide some of the answers. There is a need of work to be done to achieve error free difference solution.

To store the data there may be multiple database design patterns exist. Certain design techniques have advantages over the other for certain purposes and also the drawbacks as well for other purposes

- **Query Processing, Data Modeling, and Analysis**.

Big data mining and querying are different from the traditional statistical analysis on small samples. Compared to tiny particles the noisy big data is more valuable because statistics obtained from correlation analysis and frequent patterns usually overcome individual fluctuations often disclose more reliable hidden patterns and knowledge. The interconnected big data forms the large interconnected network; the data redundancy can be explored to compensate the missing data.

Data mining helps to improve the quality of the data and trustworthiness of the data, provide intelligent querying functions. Data mining requires scalable mining algorithms and big data computing environments. As real time data have errors, are heterogeneous and distributed across

multiple systems. The knowledge developed from the data can help to remove the errors and ambiguity. In future the big data queries can be generated for content creation on websites, to place the recommendations and also deciding whether to store the data set or to discard based on the ad hoc analysis. The problem with big data analysis is the lack of coordination between the data base systems, which provides the SQL querying and with help of analytics packages it also performs various non SQL processing such as statistical analysis and data mining. A declarative query language and the functions will improve the performance of analysis.

- **Interpretation.**

The ability to analyse the big data is of little use if the user cannot understand the analysis. Hence decision maker is needed to interpret these results provided the analysis results. This interpretation cannot happen in a vacuum. It involves the examination of list of assumptions made at each phases of the big data analysis process. It is of little use to provide only the results. Rather, it is a great work to provide how these results derived and how many inputs processed. This supportive information is known as provenance of data or result. It is needed to analyse how the data can be stored and captured in combination with the techniques to capture the metadata, we can build an infrastructure to interpret the analytical results which is obtained and to repeat the analysis for different data sets with different assumptions and parameters.

The system with visualization plays an important role for making the user to better understand the query results. A visualization tool presents the result in a powerful way which assists the interpretation, and also helps the user to collaborate with each other.

IV. CHALLENGES

The above lines give the description of different phases of the big data pipeline. Now we turn to describe some common challenges exist in many different phases of the analysis pipeline.

- **Data incompleteness:** Even after the data cleaning and error correction, some error and incompleteness in the data exist. These errors are handled during the data analysis phase. Doing this is a challenging area to consider.
- **Data Heterogeneity:** The machine learning algorithms always demands the homogeneous data as input. The first phase of the data analysis always requires the data to be in structured format. Many traditional data analysis systems require the data be in greater structure format. Computer systems work efficiently if they can store the multiple data items with a proper structure.
- **Scale:** The large and rapid growth of big data is a challenging issue. Data volume is increasing in size than compute resources.

- **Parallel processing of the data:** Problem with big data analytics is the lack of coordination between databases. Analysts are interested in impeding data from databases. Hadoop provides the basis of the solution for parallel processing of the data. In the past the, large data processing systems had to worry about the parallelism across the nodes in the cluster; now, one has to deal with the parallelism within a single node. The parallel data processing techniques that were applied in the past for processing the data across the nodes do not directly apply for the intra – node parallelism, since architecture looks different. There are many hardware resources such as processor caches and memory channels that are shared across the cores in a single node. Packing multiple sockets adds another level of complexity for intra-node parallelism.

There are many challenges exist in the field of cloud computing while applying it to network infrastructure and proposing this to many different big data applications. The k-means algorithm results in non-linear when the dataset increases in size. These problems can be solved using real time analytical architecture.

- **Timeliness:** If the data set is larger, then it will take longer amount of time to analyse. The system design makes lot of sense to process the large amount of data within a limited amount of time. Designing the better design not only includes the consideration of velocity but also data acquisition challenges as well.

Finding the specified element in the large dataset which meet the criterion is always in need during the data analysis process. In the data analysis this type of searching the element happens repeatedly. Hence scanning the entire data set for a specific data element is an impractical process. The index structure provides the solution for finding the elements quickly.

- **Technical:** The challenges of the big data in the field of technical are related to requirement of real time analysis, new storage models, and parallel/distributed operators for data with new n-dimensional array-based data structures. These data need to be server or cloud-managed, compared and visualized with joint efforts from hardware/software engineers, computer scientists and statisticians.

Most of the challenges exist in the field of analysing the social media data such as tweets and blogs where in which most data is not in structured format. Images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge.

The challenges involves in the field of linking and integration of data. Because most of the data generated in digital format today, it is challenging aspect of creation to facilitate later linkage and to automatically link previously created data.

V. OPPORTUNITIES

Since internet introduction, the world is moving from text based communications to interactive communications which includes images, videos, maps and metadata information such as geo-location information, date and timestamps. By using telematics and telemetry devices in systems of systems, the user can steadily increasing the amount of data in bidirectional interactions such as machine to machine and people to machine. The even more important area is e-health networks that allow data sharing and data merging of images. Big data explores the hidden behavioural patterns. Basically big data is bridging the gap between what the people wants to do and how they actually do and how they interact with each other. This information is useful to companies as well as for government agencies to support decision making.

In the scientific domain, the secondary uses of the patient data could leads to the discovery of the cures for a wide range of diseases. The scientists embarking the research on two projects those are US brain initiative and human brain project, to construct the simulation of inner workings of human brain. To solve the scientific problems other big data types can also be examined that may ranges from geophysics to nanotechnology to climatology.

As part of future pattern in the data mining involves how to analyse the streaming data from social networks and smaller scale blogging applications. Social networks and smaller scale blogging applications data follows the data stream model. As the data is real time which is generating continuously, it is mandatory to use efficient algorithms to analyse this data set under very strict constraints such as space and time. Twitter search engine receives 600 million search queries per day and receives 3 billion requests a day via its API. Hence better streaming techniques are needed to deal with such huge amount of data. The classification problem related to this application domain is sentimental analysis, which involves the task of classifying the messages depending on whether they convey positive or negative feelings. Classification problem requires collecting the training data so that it is possible to apply appropriate training algorithms.

CONCLUSION

Today's era is big data, through the better analysis of this huge amount of data that are becoming available; there is the potential of making faster advances in many scientific disciplines and leads to the success of many enterprises. This paper provides new areas of research such as structured classification and its related application areas such as social networks. It also explores the scientific problems domain by examining the big data types that may ranges from geophysics to climatology to nanotechnology. However, many technical challenges also described in addition to scale, heterogeneity, incompleteness, timeliness, visualization and provenance at all stages of the pipeline from data acquisition to the result interpretation.

ACKNOWLEDGMENT

I wish to thank my parents for their support and encouragement throughout my study.

REFERENCES

[1] "C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.

[2] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.

[3] Frans C., "Data mining: past, present and future." The Knowledge Engineering Review (2011): Vol. 26:1, 25–29. Cambridge University Press. 30 November 2014.

[4] D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloud computing: Current state and future opportunities," in *Proc. Int. Conf. Extending Database Technol. (EDBT)*, 2011, pp. 530–533.

[5] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, "Mad skills: New analysis practices for Big Data," PVLDB, vol. 2, no. 2, pp. 1481–1492, 2009.

[6] K. Michael and K. W. Miller, "Big Data: New opportunities and new challenges guest editors' introduction]," IEEE Comput., vol. 46, no. 6, pp. 22–24, Jun. 2013.

[7] H. Herodotou et al., "Starfish: A self-tuning system for Big Data analytics," in Proc. 5th Int. Conf. Innovative Data Syst. Res. (CIDR), 2011, pp. 261–272.

[8] TeraByte-scale Data Cycle at LinkedIn. http://tinyurl.com/lukod6.

[9] HADOOP: Scalable,Flexible Data Storage and Analysis By Mike Olson

[10] Dhananjay G. Deshmukh ME CSE, Everest College of Engineering, Aurangabad: "A Web – Based System for Classification of Remote Sensing Data "

[11] Jianing Liu, Puming Li, Jin Zhong, Liang Liang Operation Indices for Smart Power Dispatch Center Design B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):

[12] Jeffrey Dean and Sanjay Ghemawat: " MapReduce: Simplified Data Processing on Large Clusters"

[13] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.