An Analysis on Big Data Interrogation, Explore Problems and Tools

Dr.S.Brindha[1] , Dr.S.Sukumaran[2]

[1](Assistant Professor, Department of Computer Applications,
Vellalar College for Women, Thindal, Tamilnadu, India)
[2](Associate Professor, Department of Computer Science,
Erode Arts and Science College, Erode, Tamilnadu, India)

ABSTRACT: **The Analysis of Big Data datasets is an interdisciplinary endeavor that blends mathematics, statistics, computer science, and subject matter expertise. This mixture of skill sets and prospective has led to some confusion as to what comprises the field of Big Data and its analysis. The boundaries of what constitutes a Big Data problem are also changing due to the ever shifting and advancing landscape of software and hardware technology. This is due to the fact that the definition of Big Data takes into account that impact of the data's characteristics on the design of the solution environment. Most of the scientific departments like astronomy, energy effect physics, medical effectiveness, produce petatypes of data that must be shared with collaborators all over the world. Based on the research of the National Science Foundation supported International Research Network Connection links have been most important to enable the data, but as data transformation has increased so has the amount of information being collected to understand network to be maintained. New capabilities to measure and analyze the performance of international wide-area networks are essential to ensure end-users are able to take full advantage of such infrastructure for their big data applications. This paper supports to analysis of Different Big Data Methods and Algorithms to find the Big Data Interrogation, Explore problems and Tools.**

Keywords: Big Data, Datasets, Data Acquisition and Filtering, Data Extraction.

I. INTRODUCTION

Data within Big Data environments generally accumulates from being amassed within the enterprise via applications sensors and external sources. Data processed by a Big Data Solution can be used by enterprise applications directly. Big Data analytics helps organizations harness their data and use it to identify new opportunities. That in turn leads to smarter business moves more efficient operations, higher profits and happier customers [1]. Big Data might be petabytes of data consisting of billions to trillion of records of millions of people all from different sources, example web, sales, customer, contact centers, social media, mobile data so on. The Analysis of Big Data datasets is an interdisciplinary endeavor that blends mathematics, statistics, Computer Science, and subject matter expertise. This mixture of skill sets and perspectives has led to some confusion as to what comprises the field of big data and its analysis [2] [3]. The boundaries of what constitutes a Big Data problem are also changing due to the ever-shifting and advancing landscape of software and hardware technology. This is due to the fact that the definition of Big Data takes into account that impact of the data's characteristics on the design of the solution environment. Data within Big Data environments generally accumulates from being amassed within the enterprise via applications sensors and external sources. Data processed by a Big Data solution can be used by enterprise applications directly.

II. RELATED WORK

Collections or groups of related data are generally referred to as datasets. Each group or dataset member shares the same set of attributes or properties of other in the same dataset. Data Analysis is the process of examining data to find facts, relationships, pattern insights and or trends [4][2][6]. The overall goals of data analysis are to support better decision-making.

Big Data is important because it analytics helps organizations harness their data and use it to identify new opportunities. That in turn leads to smarter business moves and more efficient operations, higher profits and happier customers. Big data is a pharse used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques.

Challenges and their viable solutions:

- Handling Voluminous Data
- Big Data is generated is outpacing the development
- Computing and storage systems
- Shortage of Data Scientists
- Getting real time insights
- Data Governance and security
- Organizational Resistance
- Inaccurate Data

Inadequate knowledge about the technologies involved data privacy and inadequate analytical capabilities of organizations.

Three major challenges to implementing Big Data are, First Managing Hadoop. The foundational technology supporting every big data initiative is the Hadoop Analytics platform. The second one is the scalability challenge. Big data projects can grow an evolve rapidly. The final one is Big Data volume, velocity, and variety the big data security issues that come with them. The various challenge unstructured data accessibility, real time analytics fault tolerance.

III. BIG DATA ANALYTICS LIFE CYCLE

Big Data Analysis differs from traditional data analysis primarily due to the volume and variety characteristics of the data being processes. To address the distinct requirements for performing analysis on Big Data a step by step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing Data [7]. A Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.

The Big Data Analytics lifecycle can be divided into the 9 following nine stages:

1. Business case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

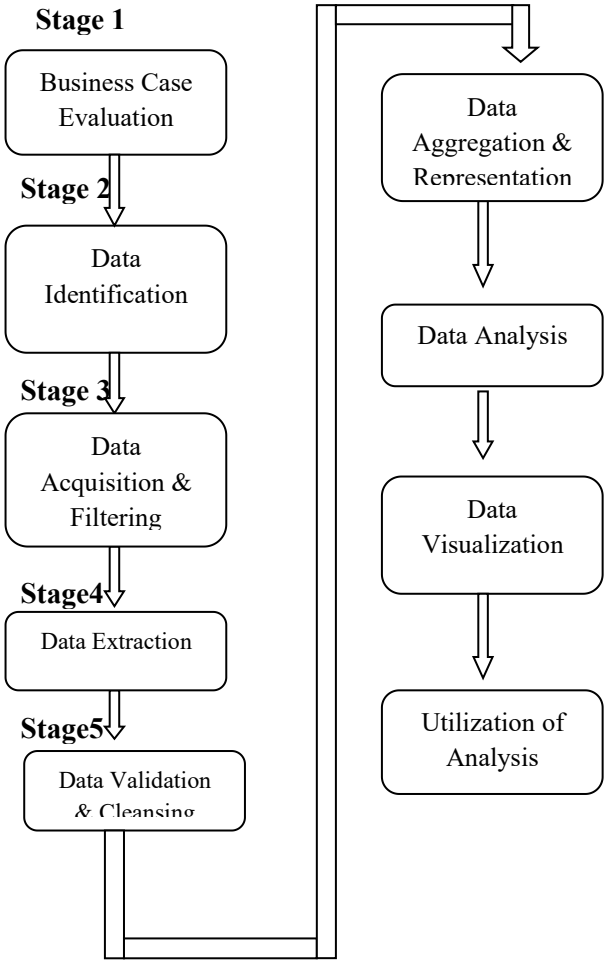

Fig.1 Big Data Life Cycle Analytics

3.1 BUSINESS CASE EVALUATION

Business Case Evaluation lifecycle must begin with a well defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the Analysis [5]. Business case be created, assessed and approved prior to proceeding with the actual hands on analysis tasks. An evaluation of a Big Data Analytics business case helps decision –makers understand the business resources that will need to be utilized and which business resources that will need to be utilized and which business challenges the analysis will tackle [4]. The efforts should be made to make the goals of the analysis project SMART, which stands for specific, measurable, attainable, relevant and timely. Based on business requirements that are documented in the business case, it can be determined whether the business problems being addressed are really Big Data problems. In order to qualify as a Big Data problem, a business problem needs to be directly related to one or more of the big data characteristics of volume or variety. Another outcome of this stage is the determination of the underlying budget required to carry out the analysis project. Any required purchase such as tools hardware and training must be understood in advance so that the anticipated investment can be weighed against the expected benefits of achieving the goals.

3.2 DATA ACQUITSTION AND FILTERING

The Data is gathered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives. Depending on the type of data source data may come as a collection of files, such as data purchased from a third

party data provider such as with twitter. Data classified as "corrupt" can include records with missing or non sensical values or invalid data types. Data that is filtered out for one analysis may possibly be valuable for a different type of analysis [8]. It is advisable to store a verbatim copy of the original dataset before proceeding with the filtering. To minimize the required storage space the verbatim copy can be compressed. The internal and external data needs to be persisted once it gets generated or enters the enterprise boundary. For batch analytics, this data is persisted to disk prior to analysis. In the case of real time analytics, the data is analyzed first and then persisted to disk Meta data can be added via automation to data from both internal and external data sources to improve the classification and querying.

3.3 DATA IDENTIFICATION

DI is dedicated to identifying the datasets required for the analysis project and their sources. Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal / external to the enterprise [9]. In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and matched against a pre defined dataset specification. A list of possible third party data providers such as data markets and publicly available datasets are compiled. Some forms of external data may be embedded within blogs or other types of content based web sites, in which case they may need to be harvested via automated tools. A number of internal and external datasets are identified. Internal data includes policy data, insurance application documents, claim data, claim adjuster notes, incident photographs, call center agent notes and emails. External data includes social media data weather reports, geographical data and census claim data consisting of multiple fields where one of the fields specifies if the claim was fraudulent or legitimate.

3.4 DATA EXTRACTION

Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. The underlying Big Data solution can use for the purpose of the data analysis. The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution [10]. Extracting the text analytics, which requires scans of whole documents, is simplified if the underlying Big Data solution can directly read the document in its native format. Further transformation is needed in order to separate the data into two separate fields as required by the Big Data solution.

3.5 DATA VALIDATION AND CLEANING

Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is predefined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity[11]. Its complexity can further make it difficult to arrive at a set of suitable validation constraints. The Data validation and cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data. Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data. For batch analytics, data validation and cleansing can be achieved via an offline ETL operation [12]. Provenance can play an important role in determining the accuracy and qualify of questionable data. Data that appears to be invalid may still be valuable in that it may possess hidden patterns and trends.

3.6 DATA VISUALIZATION & SCALABILITY

The most important challenge for Big Data Analysis techniques is its scalability and security. In the last decades researches have paid attentions to accelerate data analysis and its speed up processors followed by Moore's Law. For the former, it is necessary to develop sampling, on-line, and multi resolution analysis techniques. Incremental techniques have good scalability property in the aspect of big data analysis. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology being embedded with increasing number of cores. The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpretation [13]. However, online marketplace like flipkart, amazon, e-bay have millions of users and billions of goods to sold each month. This generated a lot of data. Finally some company uses a tool Tableau for big data visualization. It has capability to transform large and complex data into intuitive pictures. This method helps employees of a company to visualize

search relevance, monitor latest customer feedback and their sentiment analysis. The current big data visualization tools mostly have poor performances in functionalities, scalability and response in time.

The result of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated. The same results may be presented in a number of different ways, which can influence the interpretation of the results. Consequently, it is important to use the most suitable visualization technique by keeping the business domain in context [14]. Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce "models" that encapsulate new insights and understandings about the nature of the patterns and relationships that exit within the data that was analyzed[15]. A model may look like a mathematical equation or a set of rules. Models can be used to improve business process logic and application system logic, and they can form the basis of a new system or software program.

3.7 DATA AGGREGATION AND REPRESENTATION

Data may be spread across multiple datasets, requiring that datasets be joined together via common fields like date and ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. The Data Aggregation and Representation stage, is dedicated to integrating multiple datasets together to arrive at a unified view[16]. The large volumes processed by Big Data solutions can make data aggregation a time and effort intensive operation. Reconciling these differences can require complex logic that is executed automatically without the need for human intervention. Future data analysis requirements need to be considered during this part to help foster data reusability. Whether data aggregation is required or not, it is important to understand that the same data can be stored in many different forms. One form may be better suited for a particular type of analysis than another. For meaningful analysis of data, it is decided to join together policy data, claim data and call center agent notes in a single dataset that is tabular in nature where each field can be referenced via a data query [17]. It is thought that this will not only help with the current data analysis task of detecting fraudulent claims but will also help with other data analysis task, such as risk evaluation and speedy settlement of claims. The resulting dataset is stored in a No SQL database.

3.8 DATA ANALYSIS

The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. The exploratory analysis approach will be explained shortly along with confirmatory analysis.

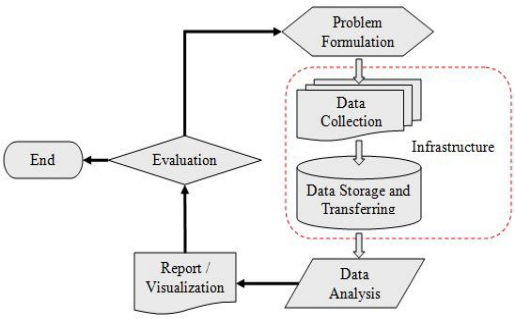

Fig.2 Workflow of Big Data

Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables. Confirmatory analysis is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand[18]. The proposed cause or assumption is called a hypothesis. The data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions. Data sampling techniques are typically used. Unexpected finding or anomalies are usually ignored since a predetermined cause was assumed.

3.9 DATA QUALITY

Data preprocessing comprises of Data Cleaning and Data Transformation. On observing the dataset, we see that the dataset is having certain missing values and noisy data too. So for further processing, we need to clean that data. This step is known as Data Cleaning. Dataset is having data in different-different formats. So, we need to reformat the data into some specific format. This step is termed as Data Transformation. This is the cleaned data having no missing values, no special symbols. On this cleaned dataset, we will perform the further mining algorithms.

IV. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are going to be the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. The research issues pertaining in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing. However, it is not limited to these issues. Big Data Analytics Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). The appliances are becoming the user of the internet, just like humans with the web browsers. Internet of things is attracting the attention of recent researchers for its most promising opportunities and challenges. It helps the future construction of information, network and communication technology. The new regulations of future will be eventually everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile devices, embedded and ubiquitous communication technologies, cloud computing and data analytics. IoT presents challenges in combinations of volume, velocity, and variety[19]. The Internet of things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Several variety technologies such as computational intelligence and big data can be incorporated together to improve the data management and knowledge of large scale automation applications.

Knowledge acquisition form IoT data is one of the biggest challenge that big data professional are facing. It is essential to develop infrastructure to analyze the internet of things data [20]. An IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information using the machine learning techniques. The streams of data generated from IoT devices and analyzing them to get meaningful information is a challenging issue and it leads to big data analytics. Key technologies that are associated with IoT. Knowledge exploration system have originated from theories of human information processing such as frames, rules, tagging and semantics networks. Generally it consists of four segments such as

- Knowledge Acquistion- knowledge is discovered by using various traditional and computational intelligence techniques.
- Knowledge Base – the discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge.
- Knowledge Dissemination – it is important for obtaining meaningful information form the knowledge base.
- Knowledge Extraction- it is a process that searches bases.

Knowledge extraction is a process that searches documents, knowledge within documents as well as knowledge bases. It is the ultimate goal of knowledge discovery. The knowledge exploration system is necessarily iterative with the judgment of knowledge application. Beyond this survey paper there are many issues and applications are available in this area of knowledge exploration. There are many issues, discussions, and researches in this area of knowledge exploration. It is beyond scope of this survey paper. For better visualization, knowledge exploration system

V. CONCLUSION

Big Data analysis blends traditional statistical data analysis approaches with computational ones. Statistical sampling from a population is ideal when the entire dataset is available and this condition is typical of traditional batch processing scenarios. One challenge concerns the best way of balancing the accuracy of an analytic result against the runtime of the algorithm. In the long term, cost-effective persistent storage. An organization will operate its Big Data analysis engine at two speeds: processing streaming data as it arrives and performing batch analysis of this data as it accumulates to look for patterns and trends. Cloud Computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, this improves availability and cost reduction.

Open challenges and research issues of big data and cloud computing are discusses in detail by many researchers which highlights the challenges in data management data variety and velocity, data storage, data processing, and resource management. So cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools. We can observe that Big data have produced many challenges for the developments of the hardware and software which leads to parallel computing, cloud computing, distributed computing, visualization process, scalability. This paper supports to analysis of Different Big Data Methods and Algorithms to find the Big Data Interrogation, Explore problems and Tools.

References

[1] A. Jacobs, The pathologies of big data, Communications of the ACM, 52(8) (2009), Pp:36-44.

[2] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015),Pp:137-144.

[3] C. Lynch, Big data: How do your data grow?, Nature, 455 (2008), Pp:28-29.

[4] Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, International Neurourology Journal, 18 (2014), Pp:50-57.

[5] C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014), Pp:314-347.

[6] D.P.Acharjya, S.Dehuri and S. Sanyal Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.

[7] H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, Second International Conference on Cloud and Green Computing, 2012, Pp:675-683.

[8] H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, International Conference on Information Technology and Management Innovation, 2015, Pp:1041-1044.

[9] K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), Pp:2561-2573.

[10] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), Pp:228-232.

[11] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data: A review, Big Data Research, 2(3) (2015), Pp:87-93.

[12] P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), Pp:89-97.

[13] R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), Pp:1-12.

[14] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285 (2014),Pp:112-137.

[15] Dr.S.Brindha, Dr.S.Sukumaran Query Suggestion and Recommendation Using Bipartite Graph and K-Means clustering, IJARCCE, Volume 3, Issue 11, November 2014.

[16] Dr.S.Brindha, Dr.S.Sukumaran, Improve Air Quality Using Inner Pattern Data Plant Method For Deploying Taxonomy Method, International Journal of Science & Engineering Development Research (www.ijsdr.org) UGC Journal, ISSN:2455-2631, Vol.5, Issue 2, Pp: 56 - 61, February-2020.

[17] Dr.S.Brindha, Dr.S.Sukumaran, Comparative Study of Term Based Pattern Taxonomy Deploying Algorithms, International Journal for Innovative Research In Multidisciplinary Field (IJIRMF), ISSN: 2455-0620 Volume-6, Issue -4, Apr-2020.

[18] Dr.S.Brindha, Dr.S.Sukumaran, Concept Based Pattern Mining Text Using Clustering Techniques, International Journal of Research and Analytical Reviews (IJRAR) UGC Approved Journal, Volume 6, Issue 2, E-ISSN 2348-1269, P-ISSN 2349-5138, Pp: 701-705.

[19] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2(2) (2015), Pp:59-64.

[20] Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congresso da sociedada Brasileira de Computacao, 2014, Pp:1-6.

Authors

Dr.S.Brindha received B.Sc degree in Science from Bharathiyar University. She done her Master Degree in Information Science and Management in Periyar University and she awarded M.Phil Computer Science from the Bharathiyar University. She received the Ph.D degree in Computer Science from the Bharathiar University. She has 5 years of teaching experience and 6 years of Technical Experience in Hash Prompt Softwares Pvt. Ltd. At present she is working as Assistant Professor of Computer Applications in Vellalar College for Women, Erode, Tamilnadu, India. She published around 14 research papers in International Journals and Conferences. Her Research area includes Text Mining and Image Processing, Pattern Taxonomy Mining, Big Data, Artificial Intelligence.

Dr. S. Sukumaran graduated in 1985 with a Degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 28 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. He has guided for more than 55 M.Phil and 10 Ph.D Research Scholars in various fields. Currently he is Guiding 6 Ph.D Scholars. He is a member of Board studies of various Autonomous Colleges and Universities. He published around 75 research papers in National and International Journals and Conferences. His current research interests include Image Processing and Data Mining.