AN EFFICIENT KNOWLEDGE OF BIG DATA PROCESSING &TECHNIQUES

Dr. (Mrs) T. Shanmugavadivu MCA., Ph.D.,
Assistant Professor, Department of CS (SF),
NGM College, Pollachi, Tamilnadu, India..
Ms.JeevaShanthini MCA.,
Assistant Professor, Department of CS (SF),
NGM College, Pollachi, Tamilnadu, India.

Abstact

A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing .Analysis of these massive at a requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. This high speed growing data is unstructured in the form of blogs, posts , tweets , news articles, video , audio etc. This all is termed as Big Data. Big data is said to bethe massive volume of information that is difficult to be processed using traditional database techniques. Big data can be of both types structured or unstructured. The growth of big data is not stoppable due to social network**.**

Keywords—Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data

I] Introduction:

Big Data are high volume, high velocity, or high-variety data that requires unique forms of processing to enable enhanced decision making, insight discovery, and process optimization

The term 'big data' is explain itself − a collection of large data sets that normal computing techniques cannot process. The term not only refers to the data, but also to the various frameworks, tools, and techniques involved. Technological advancement and the usage of new channels of communication like social networking and new, stronger devices have presented a challenge to industry that we have to find other ways to handle this large volume of data. All this big data is useful when processed. All this data analysis in meaning manner can provide a greater insight and can help in better decision making. The big data has features shown in Figure 1 which makes in different and complex for processing.

II] Types of Big Data Analysis

1. Descriptive Analytics

This summarizes past data into a form that people can easily read. This helps in creating reports, like a revenue, profit and loss statements, sales report, and so on. Also, it helps in the matrix formation of social media data.

2. Diagnostic Analytics

This type of analysis is done to understand the reason that caused a problem. Techniques like drill-down approach, deep data mining ,and data recovery are examples. Organizations use diagnostic analytics because they provide an in-depth insight into a particular problem.

An marketing and product based company report shows that their sales have gone down, although customers are adding products to their carts. This can be due to various reasons like the form didn't load correctly, the shipping fee is too high, or there are not enough payment options available. This is where you can use diagnostic analytics to find the reason.

3. Predictive Analytics

This type of analytics sees into the historical and current data to make predictions of the future. Predictive analytics uses data mining, AI, and

machine learning to analyze current data and make predictions about the future. It works on predicting customer needs, market trends etc.

Online payment service provider determines what kind of precautions they have to take to protect their clients against fraudulent transactions. Using predictive analytics, the company uses all the past payment data and user behavior data and builds an algorithm that predicts fraudulent activities.

4. Prescriptive Analytics

This type of analytics provides the solution to a particular problem specified .Perspective analytics works with both descriptive and predictive analytics. Most of the time, on Artificial Intelligence and machine learning.: Prescriptive analytics can be used to maximize

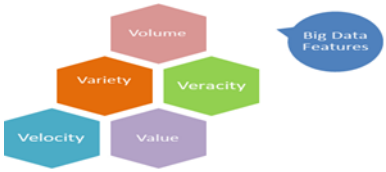


Figure2: Tools of Big

1. R Programming

R programming is one of the finest big data analytics tools. It is a important statistics programming language that can be used for statistical analysis, scientific computing, data visualization. The R programming language can also extend itself to perform various big data analytics operations and provides patterns.

- It includes a set of operations for working with arrays, and matrices.
- Effective and efficient storage facility and data handling.
- It offers a total, integrated set of big data tools for data analysis.
- It includes graphical data analysis tools that may be seen on screen or printed.

2. Apache Hadoop

Apache Hadoop is the top big data analytics tool which is open source. It is a software framework used to store data and run applications on clustering of commodity servers. It is the framework that consists ofa software ecosystem which is called as HADOOP ecosystem.

- **Distributed Processing & Storage:** The framework offers a lot of flexibility and manages distributed processing and storage on its own. Processing logic should be only written by the user.
- **Highly and easily scalable:** Vertical and horizontalscalabilityarebothavailableinHADOOP.Inhorizontalscalability it allows more no beaded to the system on the fly as data volume and processing demands rise without affecting existing systems or applications.
- **Cost-effective:** Hadoop delivers cost efficiency by introducing massively parallel computation to commodity servers. It always results in a significant drop in the cost per terabyte of storage. The commodity servers are increased and decreased as per need.

3. MongoDB

MongoDB is one of the popular document data store software in the world. It is based storage of unstructured data with higher volume of data than RDBMS-based database software has failed to do. MongoDB is robust, and it is one of the best big data analytics tools.

- **High Performance:** Due to distinctiveness such as scalability, replication, indexing ,and others, Mongo DB has a very high speed compared to other databases.
- **Replication:** Mongo DB enables high availability and redundancy by creating numerous copies of the data and sending these copies to a separate server. This ensures that in case one server fails, the data can be accessed from another server. This effectively assure availability and reliability of data.
- **Indexing:** Every field in the documents in the Mongo DB database is indexed with main and secondary indices. This makes it easier and faster to obtain or search data from the volume of data. If the data isn't indexed, the database will have to search each document individually for the query. Indexing makes the search of data faster which is necessary in big data processing.

4. Rapid Miner

Rapid Miner is one of the platforms for analysts to integrate data preparation,

machine learning, analytical model deployment, etc. It is the best big data analytics tools free that can be used for data analytics and text mining.

- Rapid Miner can connect to various Hadoop clusters, which includes Cloud era

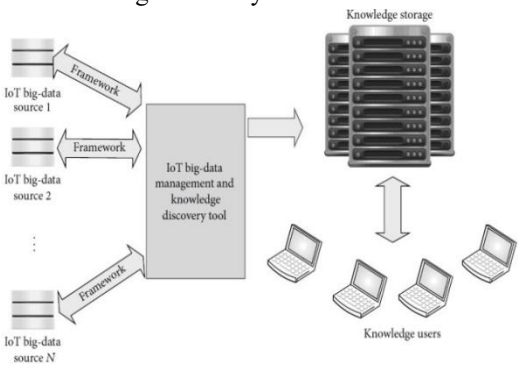
III] APPLICATION OF BIG DATA

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science 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. Main focus of this section is to discuss open research issues 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, bio inspired computing, and quantum computing.

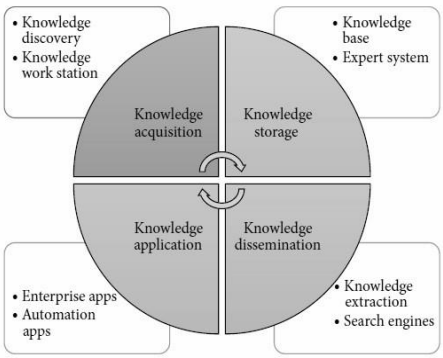
A. IoT for Big Data Analytics

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). Thus, 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 has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept to IoT is becoming more pertinent to the realistic world due to the development of mobile de- vices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. In a broader sense, just like the internet, Internet of Things enables the devices to exist in amyriadofplacesandfacilitatesapplicationsrangingfro mtrivialtothecrucial.Conversely,itis still mystifying to understand IoT well, including definitions, content and differences from other similar concepts. Several diversified technologies such as computational intelligence, and big-data

Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the re- searchers can develop tools to extract meaningful information from these data using machine learning techniques. Under- standing these 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.



Knowledge exploration system have originated from theories of human information processing such as frames, rules, tagging and semantic networks. In general, it consists of four segments such as knowledge acquisition, knowledgebase, knowledge dissemination, and knowledge application. In knowledge acquisition phase, knowledge is discovered by using various traditional and computational intelligence techniques. The discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge. Knowledge dissemination is important for obtaining meaningful information from the knowledge base. Knowledge extraction is a process that searches documents, knowledge within documents as well as knowledge bases



B. Cloud Computing for Big Data Analytics

The development of virtualization technologies have made super computing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been oneof the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. 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 there sources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big data and cloud computing are discussed in detail by many researchers which high lights 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.

Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business actively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results. This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques.

Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the market place and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software service (SaaS) from a whole crew of companies such as Net Suite, Cloud9,Jobscience etc. Another advantage of cloud computing is cloud storage which provides a possible way for storing big data. The obvious one is the time and cost that are needed to upload and download big data in the cloud environment. Else, it becomes difficult to control the distribution of computation and the underlying hardware.

C.Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired ny nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications.

D.Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously This exponential improvement in computer systems might be possible. If are that are exceptionally difficult on recent computers, of course today’s big data problems. The main technical difficulty in building quantum computer could soon be possible. Quantum computing provides away mechanics to process the information. In traditional computer, information is presented by long strings of bit which encode either a zero or a one. On the other hand a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a

qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states.

IV] CHALLENGE IN BIG DATA

To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data.

A. Data Storage and Analysis

The size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally becuase there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the top priority for the knowledge discovery and representation. The prime reason is being that, it must be accessed easily and promptly for further analysis.

B. Scalability and Visualization of Data

The most important challenge for big data analysis techniques is its scalability and security. In the last decades researchers 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.

v] CONCLUSION

In recent years data are generated at a dramatic pace. Analyzing these data is challenging for. To this end in this paper, we survey the various research issues, challenges, and tools used to analyze these big data .From this survey, it is understood that every big data platform individual focus. Some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing. We belive that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

VI] REFERENCES

[1] M.K.Kakhani,S.KakhaniandS.R.Biradar,Researchissues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.

[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] 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.

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

[6] 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.

[7] 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.

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

[9] MH.Kuo, T.Sahama, A.W. Kushniruk, E.M. Borycki and D.K.Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence,1 (2014), pp.114-126.

[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.

[11] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[12] N. Khan et. al, "Big Data: Survey, Technologies, Opportunities, and Challenges", The Scientific World Journal, vol.2014, Issue.4, pp.1-18, 2014.

[13]Online source, [Available] Top 7 Big Data Analytics Tools To Use In The Year 2022 (statanalytica.com)

[14] Online source, [Available] Big Data Analytics: Types, Tools and Applications [Updated] (simplilearn.com)

[15] S Kaushal, J.K. Bajwa, "Analytical Review of User Perceived Testing Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, issue 10, 2012.

[16] S. M. Ali et.al, "Big Data Visualization: Tools and Challenges", 2nd International Conference on Contemporary Computing and Informatics,2016.

[17] Firas D. Ahmed, Mazlina Binti Abdul Majid Age, Aws Naser Jaber, Mohd Sharifudd in Ahmad Agent Based Big Data Analytics .