International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

Exploiting and Gaining New Insights for Big Data Analysis

K.Vishnu Vandana
Assistant Professor, Dept. of CSE
Brindavan Institute of Technology and Science, Kurnool, Andhra Pradesh.

S. Yunus Basha
Assistant Professor, Dept.of CSE
Brindavan Institute of Technology and Sciences, Kurnool, Andhra Pradesh.

G. Pratiba Priyadarshini
Assistant Professor, Dept.of CSE
Brindavan Institute of Technology and Science, Kurnool, Andhra Pradesh.

***Abstract*— Apart from any domain, big data is having a tremendous impact on the enterprise. The amount of business data that is generated is getting increased every day and most types of information are being stored in digital formats. One of the challenges is to learn how to deal with all of these new data types and determine which information is potentially useful. It is not just access to new data sources, but the patterns and inter-relationships among these elements that are of interest. You need analytics to uncover insights that will help your business. Pressing needs to process greater amounts of unstructured text data in less time to assist predictive modeling in a data mining environment are met by high-performance text mining techniques. These techniques enable users to continue to work in the environments that they are familiar with and, at the same time, to benefit from the computation power that is provided by High-Performance Analytics.**

This paper not only studies Big data but also bring new data types and storage mechanisms, and analysis as well.

Keywords—Business Data, Events, Transactions, Big Data, Unstructured Data, Text Mining.

I. INTRODUCTION

Most of the days, the data we need is in digitization form. The amount of data being created and stored now a day is in both structured and unstructured form. Mean while business and organizations, individuals contribute to the data volume which became huge. It became almost impossible for organizations to search and manage this data by using conventional relational databases and data warehousing technologies. Proper analysis of this data can give better advantages to the organizations for decision making.

Big data is a new concept adopting new benefits for a better business. Big data refers to huge data sets organized by larger volumes, greater variety and complexity, generated by various sources. Big data is an area exploring new ideas and opportunities in the area of Information Technology. Big data analytics requires a new approach to capture, store, and analyze data to serve our needs. This Data analysis enables the executives to get the relevant data for making decisions in a short period of time. Analyzing such data manually is a challenging task and time consuming process. Better decision making can be done when we can mine this data in less time. The only solution to this problem is by clustering the huge available unstructured and structured data in the form of data sets and implement new innovative technologies.

II. BIG DATA

Big Data refers to data represented in datasets whose size is more than the ability of conventional database software technologies to capture, store, manage and analyse. There is no proper concept to organize a dataset where is a need of new technologies. According to IDC, Big Data technologies defines as a new generation of technologies and architectures designed to extract data from very large volumes of a wide variety of data by using high velocity capture, discovery and analysis. There must be an alternative way to gain access needed value and process it from these data.

Big Data is not only its size but also includes its categories and quantity. Some organisations generated mere gigabytes or terabytes of data storage. Data volume will continue to increase in spite of organization type and its size. There is a need for the companies to store all sorts of data.

Data can come from different sources and in different types. Data in an organization has become critical because it includes not only structured traditional relational data, but also semi-structured and unstructured data.

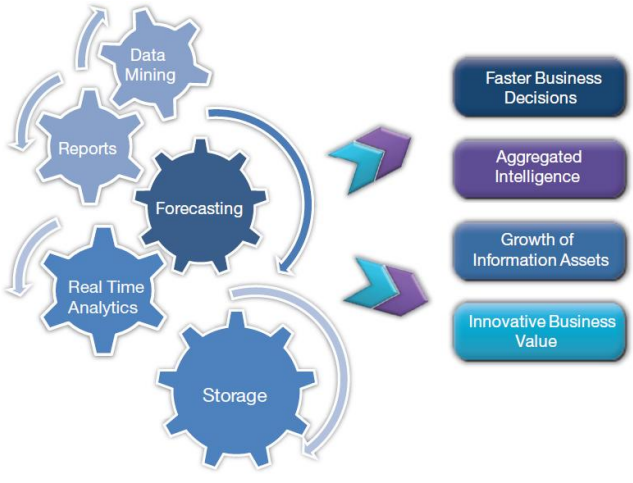

Fig 1: Big Data Sources

A. Existing Big Data Technologies

Today there are no existing specific Big Data technologies exists because analytics projects are typically complex and diverse in nature. There are no proven comprehensive Big Data certification and standards in

place. According to Hadoop the term "Big Data" is more popular for handling huge volumes of unstructured data generating continuously from various organizational sources. These Distributed File System is highly scalable, redundant and processed environment data that can be used to process different types of high range large-scale computing projects.

Big Data technology can be broken down into two major components such as the hardware component refers to the component and infrastructure layer and the software component, which is further divided into data organisation and management software, analytics and discovery software, and decision support and automation software.

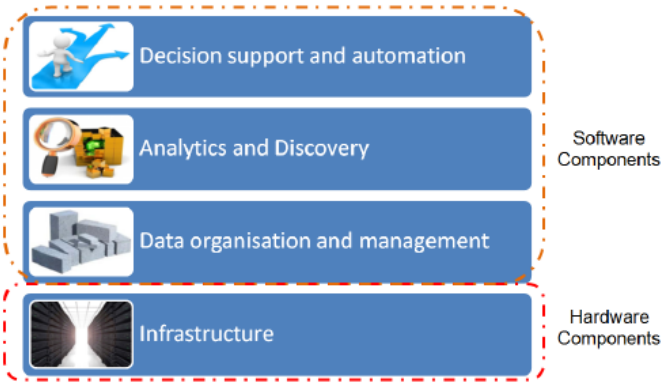

Fig 2: Big Data Technology Stack

B. *Big Data Paradigms*

Big data analytics is the process of using analytical algorithms running on powerful platforms to uncover potentials hidden big data. According to the requirement and processing time, big data analytics can be categorized into two alternative paradigms. *Streaming Processing* is an assumption that the potential value of data depends on data generated time and can analyse as soon as its derived results. In this paradigm, data arrives in the form of continuous stream arrival carries large volume, and only a small portion of the entire stream is stored in memory. For over decades we have been studied streaming processing theory and technologies and such streaming processing paradigm is used for online applications. *Batch Processing* is a paradigm; data are first stored and then analyzed. MapReduce is a dominant batch-processing model with an idea of dividing data into small chunks. Next, process these chunks in parallel and in a distributed manner to generate intermediate results. The final result is derived by consolidating all the intermediate results. This model schedules the resources close to data location and avoids data transfer communication overhead. There are many differences between these two processing paradigms, as summarized in Table 1.

Table 1: Streaming Processing Vs Batch Processing

	streaming processing	**batch processing**
Input	stream of new data or updates	data chunks
Data size	infinite or unknown in advance	known & finite
Storage	not store or store non-trival portion in memory	store
Hardware	typical single limited amount of memory	multiple CPUs, memories
Processing	a single or few pass(es) over data	processed in multiple rounds
Time	a few seconds or even millisec-onds	much longer
Applications	web mining, sensor networks, traffic monitoring	widely adopted in almost every domain

In general, the streaming processing paradigm are narrow and is suitable for applications in which data are generated in the form of a stream and rapid processing is required to obtain approximation results. Recently, most applications have adopted the batch-processing paradigm to achieve a faster response. Moreover, some research has been made to integrate the advantages of these two paradigms. Big data platforms can use alternative processing paradigms which will cause architectural distinctions in the associated platforms. For example, batch-processing-based platforms typically encompass more complex data storage and management systems, whereas streaming-processing-based platforms do not. In practice, we can customize the platform according to the data characteristics and application requirements.

III. BIG DATA ANALYSIS

Advantages of big data involves a cultural and technical changes throughout the business, from exploring new opportunities to expanding your sphere of inquiry and to exploiting new insights as you merge traditional and big data analytics. The journey often begins with a conventional enterprise data and tools, which leads about everything from sales forecasts to inventory levels. The data mainly resides in a data warehouse and is analyzed with SQL-based business intelligence tools. Most of the data in the warehouse comes from business transactions originally captured in an online transactional processing database where reports and information account for the majority of organizations are performing "what-if" analysis on multi-dimensional databases, especially in the context of planning and forecasting. These applications can benefit from big data but organizations need to mine the data and to make this goal a reality.

Most advanced data analysis such as statistical analysis, data mining, predictive analytics, and text mining, companies have traditionally stored the data to various dedicated servers for analysis. Exporting the data out of the warehouse creating copies in external servers to make predictions based on time consuming manner. It also need duplicate data storage environments and specialized data analysis skills for better data mining. Once you built a predictive model, using that model with production data involves editing of existing model or the additional movement of large data volumes from a data warehouse to an external data analysis server. At this point the data is scored and the results are moved back to the data warehouse. This process of moving and re-purposing data to

create actionable information can take long time to complete.

While many organizations have achieved better performance in exploiting their data through data analysis, at the early stages of creating an analytic model and can deliver real business value from big data. Moreover, new technologies are becoming appropriate achievements in the area of information technology and data analysts gained benefits within the database itself.

At the same time, new data types are adding traditional data sources and familiar BI activities. This data can reveal how users interact with your site. People are thinking about social media which helps thinking or how they feel about something. It can be derived from web pages, social networking sites, search engines, email messengers, search indexes, data streaming, and all types of multimedia files. This data can be collected not only from computers over internet and many other sources. Most of this data is less dense, inaccurate and doesn't fit into your data warehouse. In many cases, this is the starting point for big data analysis.

A. A New Approach to analyse Big Data

Big data analysis involves large volumes of varied data form lacks a data model to define what each element in the context of the others. There are several new issues on this new type of analysis:

- Discovery – In many cases it does not really know what you have and how different data sets are available which are related to each other. We need to identify them through a process of exploration and discovery.
- Iteration –Iteration is that it sometimes leads you down a path that turns out to be a dead end where experimentation is a part of the process. Many analysts and industry experts suggest starting with small, well-defined projects and gradually moving on to the next idea or field of inquiry.
- Flexible Capacity – Big data analysis is prepared to spend more time and utilize more resources to solve problems.
- Mining and Predicting – Big data analysis don't always know how the various data elements relate to each other. As we mine the data to discover patterns and relationships, predictive analytics.
- Decision Management –We need to consider how to automate and optimize how to automate and implement if you are using big data analytics of all those actions.

B. Big Data Analysis Requirements

We studied some of the techniques for Analyzing Big Data in the previous section and also some of methods to discover hidden relationships in big data. The three significant requirements for conducting these inquiries in an expedient way are Minimize data movement, use existing skills, and attend to data security.

Minimizing data movement is nothing but conserving computing resources. In conventional analysis scenarios, data is accessed to the computer, processed, and then sent back to the next destination. For example, production data might be extracted from e-business systems, transformed into a relational data type, and loaded into an operational data store organize them in the form of reporting. But as the data quantity grows, this type of ETL architecture becomes less efficient. There's just too much data to move around and makes sense to store and process.

We need to acquire new skills for the new data and data sources. Sometimes it is not possible to determine where analysis can be done where the skills are lacking, a combination of training, hiring and new tools will address the problem. Data security is essential for many corporate applications and data warehouse users are accustomed not only to define metrics, dimensions and attributes, but also to a reliable set of administration policies with security controls. Be awake of security and data management requirements of each analysis project and make sure that the tools can accommodate the needs.

C. Processing Big Data

In this section, we adopt text mining techniques to gain useful insights to improve performance.

Predictive analysis and opinion mining plays an important role in text analytics. They are also helpful to understand, identify the different aspects of services and measure overall satisfaction deemed important.

Free text analysis aspects are completely unexpected or unanticipated. Often, interpretation of text in isolation not even leads to the correct conclusion. Text is generated in an uncontrolled manner and is often grammatically incorrect, and may even contain non-textual elements such as icons, characters, symbols and so on which may be contextual. Text that is obtained can be noisy and even not relevant to the context of the discussion. In such situations, statistical analysis is adopted. Statistical analysis involves deriving information through patterns and trends through statistical techniques such as classification.

IV. FUTURE ENHANCEMENTS

Most of the data generated everyday all over the world is unstructured in nature and captures large volumes of information. Effective use of this information can help better know the past, understand the present and predict the future. Language, expressions and communication mechanisms are constantly evolving. While it would not be presumptuous to state that none of the systems will ever be perfect, with Big Data technologies. However, gleaning meaningful insights from unstructured text is an endeavour where humans will have to remain in the field to exploit the power of machines. A complete and experimental research is needed to use of Big Data and need to identify certain patterns to represent it.

V. CONCLUSION

The new era of big data is bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. The newest aspect of big data is generating new opportunities and new challenges for businesses across every industry. The challenge of data integration—incorporating data from social media and other unstructured data into a traditional BI environment—is one of the most urgent issues facing today. Organizations in every industry are trying to make sense of the massive influx of big data, as well as to develop analytic platforms that can synthesize traditional structured data with semi-structured and unstructured sources of information.

ACKNOWLEDGMENT

I want to take this opportunity to thank all the people who helped me during my conference sojourn. I understand that it is rather late to acknowledge their contributions, but as the saying goes, better late than never!

REFERENCES

[1] Savitha K, Vijaya MS – "Mining of Web Servers Logs in a Distributed Cluster Using Big Data Technologies", International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014.

[2] Sherin A, Dr S Uma, Saranya K, Saranya Vanim – "Survey On Big Data Mining Platforms, Algorithms and Challenges", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 09 Sep 2014,p.p.no854–861.

[3] Han Hu, Yonggang Wen, Tat-Seng Chua, Xuelong Li – "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", `Vol 2, May 12,2014,p.p.no652-681.

[4] Chanchal Yadav, Shuliang Wang , Manoj Kumar – Algorithm and approaches to handle large Data - A Survey", International Journal of Computer Science and Network, Vol 2, Issue 3, 2013.

[5] T.K.Das1 , P.Mohan Kumar – "BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013, p.p.no 153-157.

[6] Zheng Zhao, Russell Albright, James Cox, and Alicia Bieringer – "Big Data Meets Text Mining", Vol 4, 2013.