BIG DATA ANALYTICS – TOOLS, TECHNIQUES AND CHALLENGES

V. Niranjana

Guest Lecturer, Department of Computer Science,
St. Joseph's College of Arts and Science for Women, Hosur, Tamil Nadu, (India)

ABSTRACT

Data that is complex in terms of volume, variety, velocity and/or its relation to other data makes it hard to handle using traditional database management or tools. Big data analyticsrefers to analysis techniques operated on data sets classified as "big data". Today, every organization across the globe is faced with an unprecedented growth in data. The digital universe of data was expected to expand to 2.7 zettabytes (ZB) by the end of 2012. Then it's predicted to double every two years, reaching 8 ZB of data by 2015. That's really big data. In short, Big Data is about quickly deriving business value from a range of new and emerging data sources, including social media data, location data generated by smartphones and other roaming devices, public information available online and data from sensors embedded in cars, buildings and other objects. I present here the basics of Big Data Analytics, big data analytics tools, techniques and technical challenges of big data.

Keywords : Analytics, Big data, Hadoop, Mapreduce, NoSQL.

I. INTRODUCTION

According to the Oxford dictionary, analytics is defined as "the systematic computational analysis of data or statistics" or, "information resulting from the systematic analysis of data or statistics". Since information is the source for knowledge and even wisdom, analytics is very important in many different fields, both scientific and organizational, especially for decision making. For example, without analytics, the procurement department of a supermarket chain would have a hard time deciding what to buy and in which numbers.

The capability to store data quickly isn't new. What's new is the capability to do something meaningful with that data, quickly and cost-effectively. Businesses and governments have been storing huge amounts of data for decades. What we are witnessing now, however, is an explosion of new techniques for analyzing those large data sets. In addition to new capabilities for handling large amounts of data, we're also seeing a proliferation of new technologies designed to handle complex, non-traditional data — precisely the kinds of unstructured or semi-structured data generated by social media, mobile communications, customer service records, warranties, census reports, sensors, and web logs.

The journey often begins with traditional enterprise data and tools, which yield insights about everything from sales forecasts to inventory levels. The data typically resides in a data warehouse and is analyzed with SQL-based business intelligence (BI) tools. Much of the data in the warehouse comes from business transactions originally captured in an OLTP database. While reports and dashboards account for the majority of BI use, more and more organizations are performing “what-if” analysis on multi-dimensional databases, especially within the context of financial planning and forecasting [1].

These planning and forecasting applications can benefit from big data but organizations need advanced analytics to make this goal a reality. For more advanced data analysis such as statistical analysis, data mining, predictive analytics, and text mining, companies have traditionally moved the data to dedicated servers for analysis. Exporting the data out of the data warehouse, creating copies of it in external analytical servers, and deriving insights and predictions is time consuming. It also requires duplicate data storage environments and specialized data analysis skills.

New types of data are supplementing traditional data sources and familiar BI activities. For example, weblog files track the movement of visitors to a website, revealing who clicked where and when. This data can reveal how people interact with your site. Social media helps you understanding what people are thinking or how they feel about something. It can be derived from web pages, social media sites, tweets, blog entries, email exchanges, search indexes, click streams, equipment sensors, and all types of multimedia files including audio, video, and photographic.This data can be collected not only from computers, but also from billions of mobile phones, tens of billions of social media posts, and an ever-expanding array of networked sensors from cars, utility meters, shipping containers, shop floor equipment, point of sale terminals and many other sources. Most of this data is less dense and more information poor, and doesn’t fit immediately into your data warehouse. As we will see, some of it is better placed in Hadoop Distributed File System (HDFS) or in non-relational databases, commonly called NoSQL databases. In many cases, this is the starting point for big data analysis.

II.CHARACTERIZATION OF BIG DATA

Big data is often described using 5 V’s: Volume, Velocity, Variety, Veracity and Value. Fig. 1 shows the characterization of big data.

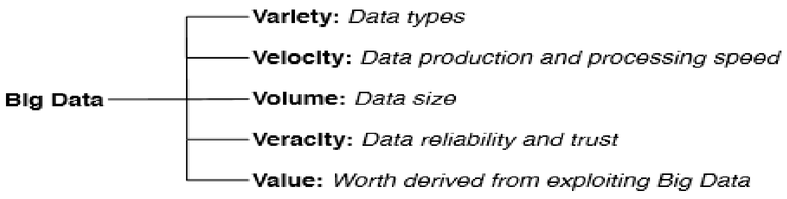
Big Data
Variety: Data types
Velocity: Data production and processing speed
Volume: Data size
Veracity: Data reliability and trust
Value: Worth derived from exploiting Big Data

Figure 1: Characterization of Big Data.

2.1. Variety

Variety [2] describes the fact that Big Data can come from many different sources, in various formats and structures. For example, social media sites and networks of sensors generate a stream of ever-changing data. As well as text, this might include, for example, geographical information, images, videos and audio.

2.2. Velocity

Velocity reflects the sheer speed at which this data is generated and changes. For example, the data associated with a particular hashtag on Twitter often has a high velocity. Tweets fly by in a blur. In some instances they move so fast that the information they contain can't easily be stored, yet it still needs to be analyzed.

2.3. Volume

Volume refers to the fact that, Big Data involves analyzing comparatively huge amounts of information, typically starting at tens of terabytes.

2.4. Veracity

Veracity refers to the reliability or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable. But big data and analytics technology now allows us to work with these types of data. The volumes often make up for the lack of quality or accuracy.

2.2. Value

Value refers to the worth derived from big data. It is the most important aspect of big data. Through effective data mining and analytics, the massive amount of data that we collect throughout the normal course of doing business can be put to good use and yield value and business opportunities.

III. IMPLEMENTING BIG DATA: 7 TECHNIQUES TO CONSIDER

Big data analysis involves making "sense" out of large volumes of varied data that in its raw form lacks a data model to define what each element means in the context of the others. There are 7 widely used Big Data analysis techniques which are as follows:

3.1. Association Rule Learning

Association rule learning is a method for discovering interesting correlations between variables in large databases. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale (POS) systems. Association rule learning is being used to help place products in better proximity to each other in order to increase sales, extract information about visitors to websites from web server logs, analyze biological data to uncover new relationships, and monitor system logs to detect intruders and malicious activity

3. 2. Classification Tree Analysis

Statistical classification is a method of identifying categories that a new observation belongs to. It requires a training set of correctly identified observations – historical data in other words. Statistical classification is being used to automatically assign documents to categories, and categorize organisms into groupings.

3.3. Genetic Algorithms

Genetic algorithms are inspired by the way evolution works – that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to "evolve" useful solutions to problems that require optimization.

3.4. Machine Learning

Machine learning includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of "training data."

3.5. Regression Analysis

At a basic level, regression analysis involves manipulating some independent to see how it influences a dependent variable. It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.

3.6. Sentiment Analysis

Sentiment analysis helps researchers determine the sentiments of speakers or writers with respect to a topic. Sentiment analysis is being used to help improve service at a hotel chain by analyzing guest comments, customize incentives and services to address what customers are really asking for, and determine what consumers really think based on opinions from social media.

3.7. Social Network Analysis

Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals.

IV. APPROACHES FOR ANALYZING BIG DATA

There are five key approaches to analyzing big data and generating insight:

4.1. Discovery Tools

Discovery tools are useful throughout the information lifecycle for rapid, intuitive exploration and analysis of information from any combination of structured and unstructured sources. These tools permit analysis alongside traditional BI source systems[3]. Because there is no need for up-front modeling, users can draw new insights, come to meaningful conclusions, and make informed decisions quickly.

4.2. Business Intelligence (BI) Tools

BI tools are important for reporting, analysis and performance management, primarily with transactional data from data warehouses and production information systems. BI Tools provide comprehensive capabilities for business intelligence and performance management, including enterprise reporting, dashboards, ad-hoc analysis, scorecards, and what-if scenario analysis on an integrated, enterprise scale platform.

4.3. In-Database Analytics

In-Database Analytics include a variety of techniques for finding patterns and relationships in your data. Because these techniques are applied directly within the database, you eliminate data movement to and from other analytical servers, which accelerates information cycle times and reduces total cost of ownership.

4.4. Hadoop

Hadoop is useful for pre-processing data to identity macro trends or find nuggets of information, such as out-of-range values. It enables businesses to unlock potential value from new data using inexpensive commodity servers. Organizations primarily use Hadoop as a precursor to advanced forms of analytics.

4.5. Decision Management

Decision Management includes predictive modeling, business rules, and self-learning to take informed action based on the current context. This type of analysis enables individual recommendations across multiple channels, maximizing the value of every customer interaction.

V.TOOLS USED IN BIG DATA SCENARIO

- **NoSQL:** A NoSQL (often interpreted as **Not only SQL**) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling, and finer control over availability. The data structures used by NoSQL databases differ from those used in relational databases, making some operations faster in NoSQLand others faster in relational databases. The particular suitability of a given NoSQL database depends on the problem it must solve. NoSQL databases are increasingly used in big data and real-time web applications.

Some of the NoSQL databases are DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riakand ZooKeeper.

- **MapReduce** is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. A MapReduce [4] program is composed of a **Map()** procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a **Reduce()** procedure that performs a summary operation. Example :Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum.
- **Storage** the key requirements of big data storage are that it can handle very large amounts of data and keep scaling to keep up with growth, and that it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools. Exanple : S3, Hadoop Distributed File System.
- **Servers for big data:**Examples for big data servers are EC2, Google App Engine, Elastic, Beanstalk, Heroku.
- **Processing for big data**: The following tools are used to aggregate, manipulate, and mash up content from around .Example : R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets,Tinkerpop.

VI. TECHNICAL CHALLENGES

Many of Big Data’s technical challenges also apply to data it general. However, Big Data makes some of these more complex. The following are the technical challenges [5] faced by big data.

6.1. Data Integration

Since data is a key asset, it is increasingly important to have a clear understanding of how to ingest, understand and share that data in standard formats in order that business leaders can make better-informed decisions. Integration challenges arise when a business attempts to transfer external data to its system. Whether this is migrated as a batch or streamed, the infrastructure must be able to keep up with the speed or size of the incoming data.

6.2. Data Transformation

Another challenge is data transformation — the need to define rules for handling data. Organizations also need to consider which data source is when records conflict, or whether to maintain multiple records. Handling duplicate records from disparate systems also requires a focus on data quality.

6.3. Complex Event Processing

Complex event processing (CEP) effectively means (near) real-time analytics. Matches are triggered from data based on either business or data management rules. For example, a rule might look for people with similar addresses in different types of data. But it is important to consider precisely how similar two records are before accepting a match.

6.4. Semantic Analysis

Semantic analysis is a way of extracting meaning from unstructured data. It can uncover people's sentiments towards, organizations and products, as well as unearthing trends, untapped customer needs, etc.

6.5. Historical Analysis

Historical analysis could be concerned with data from any point in the past. That is not necessarily last week or last month — it could equally be data from 10 seconds ago.

6.6. Search

Search is not always as simple as typing a word or phrase into a single text input box. Searching unstructured data might return a large number of irrelevant or unrelated results. Sometimes, users need to conduct more complicated searches containing multiple options and fields. Another consideration is how search results are presented

6.7. Data Storage

As data volumes increase storage systems are becoming ever more critical. Big Data requires reliable, fast-access storage.

6.8. Data Integrity

For any analysis to be truly meaningful it is important that the data being analyzed is as accurate, complete and up to date as possible. Erroneous data will produce misleading results and potentially incorrect insights.

6.9. Data Lifecycle Management

In order to manage the lifecycle of any data, organizations need to understand what that data is and its purpose. But, the potentially vast number of records involved with Big Data, and the speed at which the data changes, can give rise to the need for a new approach to data management.

6.10. Data Replication

Generally, data is stored in multiple locations in case one copy becomes corrupted or unavailable. This is known as data replication. The volumes involved in a Big Data solution raise questions about the scalability of such an approach.

6.11. Data Migration

When moving data in and out of a Big Data system, or migrating from one platform to another, organizations should consider the impact that the size of the data may have.

6.12. Visualization

While it is important to present data in a visually meaningful form, it is equally important to ensure presentation does not undermine the effectiveness of the system. Organizations need to consider the most appropriate way to display the results of Big Data analytics so that the data does not mislead.

6.13. Data Access

The final technical challenge relates to controlling who can access the data, what they can access, and when. Data security and access control is vital in order to ensure data is protected. Access controls should be fine-grained, allowing organizations not only to limit access, but also to limit knowledge of its existence.

VII. PRIVACY AND BIG DATA

The successful adoption of big data into an organization is by no means simple and can be a steep learning curve. It also presents some obvious challenges around privacy. By capturing and combining big data sets, an organization is able to create an extremely detailed profile of an individual. Regulations that balance productivity gains with privacy protection are yet to be developed. From a commercial or competitive perspective, organizations also face the challenge of sharing highly confidential big data; things like financial statements, patents, trade secrets and intellectual property.

This sensitive data must be securely stored and made available to those who need it. New innovations, like virtual data rooms, are making this task easier. Virtual data rooms allow authorized individuals to review confidential data and documents in a secure online space. Companies that implement a comprehensive data security system and have good privacy policies in place, which protect against the unauthorized disclosure of sensitive information, will be in the best position to adopt big data.

VIII. CONCLUSION

The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever increasing speed. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimize its operation, and reduce its costs.

Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in lightening these problems by providing resources on-demand with costs proportional to the actual usage. Further-more, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand.Like in many other technological areas, customs and ethics around big data take time to develop.

ACKNOWLEDGEMENT

I would like to thank God, Almighty for guiding and showering HIS blessings on me throughout my life and this paper in particular. Secondly I would like to thank my husband for his love and support. Next I would like to thank my parents and my daughters for their patience and care. I would also like to thank my colleagues and friends who helped me a lot in finalizing this paper within the limited time frame.

REFERENCES

[1]. “*Planning Guide Getting Started with Big Data*”, Intel IT Center.

[2]. http://en.wikipedia.org/wiki/Big_data#Technologies.

[3]. “*Big Data Analytics - Advanced Analytics in Oracle Database*”, Oracle Corporation, World Headquarters, U.S.A.

[4]. VigneshPrajapati, “*Big Data Analytics with R and Hadoop*”, Packt Publishing Ltd.

[5]. “*THE WHITE BOOK OF Big Data*”, Ian Mitchell, Mark Locke, Mark Wilson, and Andy Fuller, Published by Fujitsu Services Ltd.

AUTHOR ’S BIOGRAPHY:

Niranjana. V was born in Tirunelveli, TamilNadu (TN), India, in 1986. She received the Bachelor of Computer Science (B.Sc.[CS]) degree from the Madurai Kamaraj University (MKU), Madurai, TN, India, in 2007 and the Master of Computer Applications (M.C.A.) degree from Adhiyamaan college of Engineering (Autonomous), affiliated to Anna University, Coimbatore, TN, India, in 2010. Her research interests include Cloud computing and Big Data Analytics.