Big Data Analytics

Kundan Kumar Jha(22MSM40052)

Course Name:	Big Data
Course Code:	22SDT-654
Program:	M. Sc (Data Science)
Department:	Mathematics

Abstract:

Big data is getting a lot of attention in the IT world today. The rapid rise of the internet and digital commerce has led to an increased need for data storage and analysis, and IT departments are facing serious challenges in security, and check this growing knowledge. The reason organizations collect and store more data than ever before is because their business depends on it. The type of data generated is not always data-driven data, but data also that includes documents, images, audio, video and content from social media. Big data analytics is a way to extract valuable data from this huge data that can trigger new business opportunities and maximize customer retention.

In this paper, we will discuss about big data analytics in detail i.e., big data and analytics and why both together. We study about various steps that are required for big data analytics and also data type of big data analytics i.e. structured, semi-structured and unstructured data.

Here we will discuss about the different types of big data analytics with the benefits and barriers of big data analytics. You will also study about the various tools and technology that’s help in big data analytics.

Keywords: Big Data Analytics, analysis, structured and unstructured data.

Introduction

Big data analytics is where advanced analytics techniques work on big data. So big data analytics is really about two things - big data and analytics - and the two of them combining to create one of the most important Business Intelligence (BI) topics today. Let's start by defining Advanced Analytics, then move on to Big Data and the combination of the two.

Advanced Analytics

According to a survey, 38 percent of research organizations reported that they have implemented analytics, and 85 percent will implement it soon. Why run to use of high-level analytics? First, as can be seen in many of the "jobs" we've experienced in recent years, the economic changes are huge. Analytics helps us discover what has changed

and what we need to do about it. To that end, advanced analytics are the best way to discover new customers, identify top suppliers, engage with brands, understand seasonal sales, and more.

We can call it not as "advanced analytics" as it is "discovery analytics" because that's what users are trying to achieve. (Some call it “exploratory analytics.”) In other words, for big data analytics, users are mostly business analysts trying to discover the latest business trends which are not discovered before. To do this, the analyst needs a lot of data with lot of information. This is often information that the business has not yet used to review.

For example, in the middle of the recent economic recession, companies have recently been affected by new form of customer's churn. To find the root causes of this new form of customer churn, business analysts will collect the terabytes of detailed data from operational applications to get a view of end customers. Analysts can combine this new data with historical data from the data warehouse. After dozens of questions, the analyst found a new churn behaviour in a group of customers.

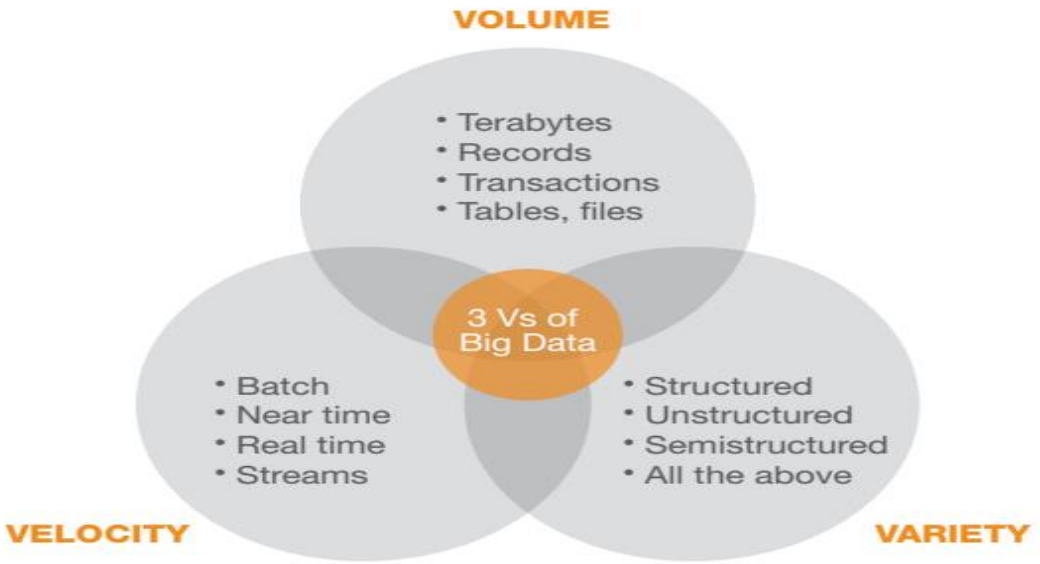
Luckily, this discovery will lead to metrics, reports, analytical models, or other BI products that companies can use to monitor and forecast data. Scientific research for big data can be done with a variety of analysis tools, including analysis tools based on SQL query, data mining, fact analysis, integration, data visualization, natural language processing, text analysis, artificial intelligence, and more.

All of these technologies have been around for years, most of them in the 1990s. What makes different today is that many organizations use it. This is because many of these methods scale well with very little data for very large, multi-terabyte datasets.

Big Data

Big data refers to the concept of big data with three V’s (3Vs). The first is the volume based on the amount of data on the storage medium. The second is Variety, which refers to various heterogeneous and complex types of information. Data can be structured, unstructured or semi-structured data created by humans or machines. The third is Velocity, which indicates the speed of data that needs to be processed for big data.

Most definitions of big data focus on the size of the data stored. Size matters, but big data also has other important characteristics, such as data velocity and data variety. The Three Vs of Big Data (Volume, Variety, and velocity) create a comprehensive definition that dispels the myth that Big Data is only about volume. Also, each of the three Vs has its own branches of analysis.



Three main characteristics of big data: volume, variety and velocity or 3Vs. The volume of the data is how big it is. Velocity refers to how quickly data changes or how much data is created. Finally, variety includes different types of data and different uses and methods of data analysis.

Why Put Big Data and Analytics Together Now?

Big data provides a large sample size that improves the results of analytical tools: Most tools designed for data mining or statistical analysis tend to be optimized for big data. In fact, the general rule is that, if our data sample is larger then statistics and other analysed will be more accurate.

Much can be learned from messy data as long as it is Big: Most modern tools and techniques for advanced analytics and big data are tolerant of raw data and their transformation patterns, data is not structure and is weak data. This is good because

discovery and predictive analytics depend on many details, even the data in question. For example, fraud detection applications often depend on outliers and non-standard data as fraud's indicators.

Big Data Analytics

It is a process used to extract meaningful hidden patterns and customers preferences from large amount of raw data which can be used for better decision making. Big data analytics can be used in multiple way to improve business such as:

- To understand customers behaviour.
- To predict future trends in business.
- To improve market complain.
- To detect fraud and misuse.

Steps of Performing Big Data Analysis

1. Data Collection:

An organisation can gather structured, unstructured data from different sources such as cloud storage, mobile application or IOT devices.

2. Data Processing:

Once the data is collected and stored, it must be organised properly to get accurate result. This data can be processed in large blocks overtime known as **batch processing** and small processed at one time known as **stream processing**.

Stream processing is more expensive and complex.

3. Data Cleaning:

Data should be formatted correctly and irrelevant or duplicate entries should be deleted.

4. Data Analysis:

Data analysis is used to extract meaningful pattern, add influence, business decisions. This method includes data mining, predictive analysis and deep learning.

Some Big Data Analytic method are following:

- **Data mining** analyzes large datasets to identify patterns and relationships by identifying anomalies and creating datasets.
- **Predictive analytics** uses an organization's historical data to predict the future and identify future risks and opportunities.
- **Deep learning** leverages human learning models by applying artificial intelligence and machine learning to layer algorithms and find patterns in complex and abstract data.

Datatypes for Big Data Analytics

Like in social media, we are dealing with unstructured data like audio or video recording and images. Big data is the solution to maintain unstructured as well as semi-structured data.

1. Structured Data:

The data which has a proper structured like table in rdbms is known as structured data. We can easily process it.

Example: tables and records.

2. Semi-Structured Data:

The data which is not fully formatted and it is not stored in any table is known as semi-structure data.

Example: CSV file.

3. Unstructured Data:

The data which does not have any structured is known as unstructured data. It is complex to process.

Example: Audio, Video and Images.

Types of Big Data Analytics

In big data analytics, there are mainly four types. These are

- Descriptive
- Diagnostics
- Predictive
- Prescriptive

1. Descriptive:

It describes the past data in the form of reports such as a company sells and profit report.

2. Diagnostics:

It is used to understand the cause of a problem in an organisation.

For example,

An e-commerce shows in the report that sells has gone down although customers are adding product to their cart. This can be due to various reasons such as shipping fees is high, payment option are not enough. So, diagnostics analytics can be used to find the reason.

3. Predictive Analytics:

It compares historical data and present data to make future predication such as market trends, customer preference.

4. Prescriptive:

It is used to prescribe the solution to a particular problem. It is combination of descriptive and predictive analysis.

Big Data Analytics Tools and Technology

Big data analytics cannot be reduced to a single tool or process. Instead, different types of tools work together to help you collect, process, clean, and analyze big data. Listed below are some key players in the big data ecosystem:

- **Hadoop** is an open-source framework for storing and processing big data on a cluster of hardware devices. The framework is free and can handle a lot of structured and unstructured data, making it an essential part of big data work.

- **NoSQL databases** are non-relational data management systems that do not need a fixed schema, making them an excellent choice for large, raw, unstructured data. NoSQL stands for "Not Only SQL" and these databases can handle different types of data.

- **MapReduce** is an essential part of the Hadoop framework and has two functions. The first is the mapping, which filters the data to different nodes within the cluster. The second is reduction, which improves and reduces the results of each node to answer the question.

- **YARN** stands for “Yet Another Resource Negotiator.” It is another component of the second-generation Hadoop. Cluster management systems facilitate task scheduling and cluster resource management.

- **Spark** is an open-source computing framework that uses data asymmetry and error-based algorithms to provide a functional interface for the entire cluster. Spark can perform batch and stream operations for fast computation.

- **Tableau** is an end-to-end data analytics platform that enables you to organize, analyze, collaborate and share your big data. Tableau specializes in self-service visual analytics, allowing people to ask new questions about managing big data and easily share insights across the organization.

Benefits of Big Data Analytics

1.Risk Management:

It can identify various fraud activities. A business can use it to produce a list of suspicious activity or root cause of any problem>

2. Product Development Innovation:

It can be use to analyse existing product design and suggest any need for improvement.

3. Better Decision Making:

A business can use it to decide weather a particular location is suitable for a new outlet or not; by analysing different factor such as accessibility of that location.

4. To Improve Customer Experience:

An organisation can improve the customer experience through data analysis inbuilt good customer relation.

Difficulties in Big Data Analytics

Above we saw benefits of big data analytics but also have barriers in big data analytics. Some of the barriers in big data analytics are following:

1. Lack of staff and skills is a big problem for big data analytics:

After all, big data analytics is new to many organizations. The intelligence process is not the same as Business Research and Information Systems, where most organizations develop intelligence. Other skills related issues include the challenge of building big data analytics systems and the issues handling big data for end users.

2. Lack of business support can hinder big data analytics initiatives:

There are lack of economic support and lack of compelling business case, overall cost is issue.

3. Problems with database software can affect big data analytics:

Problems arise when existing database software lacks in-database analytics, has scalability issues with large datasets, cannot handle queries fast enough, or load information fast enough.

In a related question, managing big data in a database is difficult when the database is modelled only for reporting.

Conclusion:

In recent years, data has been produced by the unexpected. Analyzing this data is difficult for the average person. For the end of this article, we explore various research questions, challenges, and tools for analyzing this huge amount of data. From these studies, we learned that every big data platform has its own personal concerns. Some are designed for batch processing, while others are good at real-time analysis. Each big data platform also has specific functions. Different technologies used for analysis include statistical analysis, machine learning, data mining, smart analytics, cloud computing, quantum computing, and data flow processing. We believe that future researchers will pay more attention to these techniques to solve big data problems.

References:

[1] Philip Russom, “Big Data Analytics – TDWI Best Practices Report”, Fourth Quarter 2011.

[2] Tableau site, “Big Data Analytics”, “https://www.tableau.com/learn/articles/big-data-analytics”.

[3] Yousef A. Aburawai, Abdulbaset Albaour, “Big Data: Review Paper”, International Journal of Advance Research and Innovative Ideas in Education, February 2021.