22UDA501 – Introduction to Data Analytics

Introductory Background

Data has become the most critical factor in business. Different technologies, methodologies, and systems have been invented to process, transform, analyze, and store data in this data-driven world.

- **Big data** refers to any large and complex collection of data.
- **Data analytics** is the process of extracting meaningful information from data.
- **Data science** is a multidisciplinary field that aims to produce broader insights.

Each of these technologies complements one another yet can be used as separate entities. For instance, big data can be used to store large sets of data, and data analytics techniques can extract information from simpler datasets.

Big data

Big data simply refers to extremely large data sets. This size, combined with the complexity and evolving nature of these data sets, has enabled them to surpass the capabilities of traditional data management tools. This way, data warehouses and data lakes have emerged as the go-to solutions to handle big data, far surpassing the power of traditional databases.

Some data sets that we can consider truly big data include:

- Stock market data
- Social media
- Sporting events and games
- Scientific and research data

Characteristics of big data

- **Volume.** Big data is enormous, far surpassing the capabilities of normal data storage and processing methods. The volume of data determines **if it can be categorized as big data.**

- **Variety.** Large data sets are not limited to a single kind of data—instead, they consist of various kinds of data. Big data consists of different kinds of data, from **tabular databases to images and audio** data regardless of data structure.

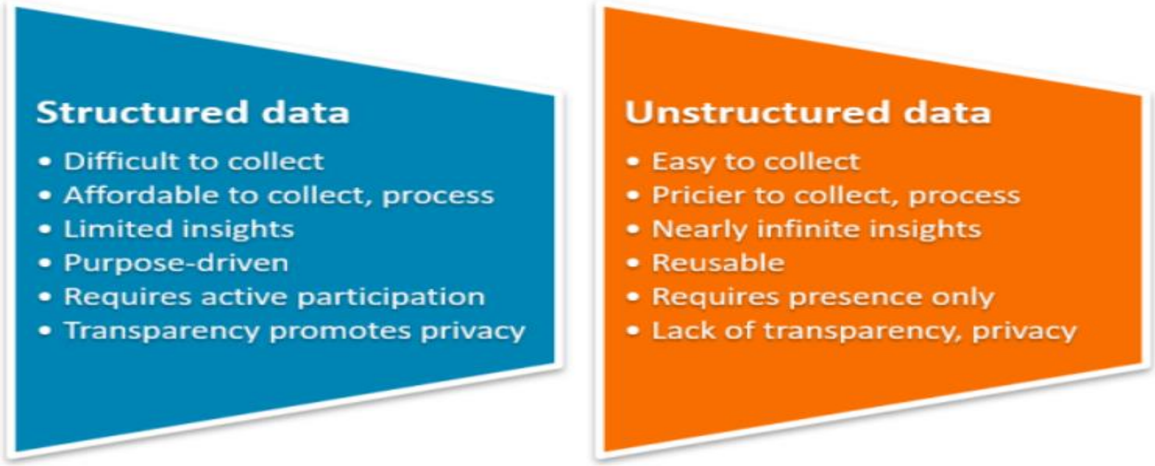
- **Velocity. The speed at which data is generated**. In Big Data, **new data is constantly generated and added to the data sets frequently**. This is highly prevalent when dealing with continuously evolving data such as social media, IoT devices, and monitoring services.

- **Veracity or variability.** There will inevitably be some inconsistencies in the data sets due to the enormity and complexity of big data. Therefore, you must account for variability to **properly manage and process big data.**

- **Value.** The usefulness of Big Data assets. The worthiness of the output of big data analysis can be **subjective and is evaluated based on unique business objectives.**

Types of big data

- **Structured data.** Any data set that **adheres to a specific structure** can be called structured data. These structured data sets can be processed relatively easily compared to other data types as users can exactly identify the structure of the data. A good example for structured data will be a distributed **RDBMS** which contains data in organized table structures.
- **Semi-structured data.** This type of data **does not adhere to a specific structure** yet retains some kind of observable structure such as a grouping or an organized hierarchy. Some examples of semi-structured data will **be markup languages (XML), web pages, emails, etc.**
- **Unstructured data.** This type of data consists of data that **does not adhere to a schema or a preset structure.** It is the most common type of data when dealing with **big data—things like text, pictures, video, and audio** all come up under this type.



Structured data

- Difficult to collect
- Affordable to collect, process
- Limited insights
- Purpose-driven
- Requires active participation
- Transparency promotes privacy

Unstructured data

- Easy to collect
- Pricier to collect, process
- Nearly infinite insights
- Reusable
- Requires presence only
- Lack of transparency, privacy

Big data systems & tools

When it comes to managing big data, many solutions are available to store and process the data sets. Cloud providers like AWS, Azure, and GCP offer their own data warehousing and data lake implementations, such as:

- AWS Redshift
- GCP BigQuery
- Azure SQL Data Warehouse
- Azure Synapse Analytics
- Azure Data Lake

Data Analytics

Data Analytics is the process of analyzing data in order to extract meaningful data from a given data set. These analytics techniques and methods are carried out on big data in most cases, though they certainly can be applied to any data set.

In a scientific sense, a medical research organization can collect data from medical trials and evaluate the effectiveness of drugs or treatments accurately by analyzing those research data.

Combining these analytics with data visualization techniques will help you get a clearer picture of the underlying data and present them more flexibly and purposefully.

Types of analytics

While there are multiple analytics methods and techniques for data analytics, there are four types that apply to any data set.

- **Descriptive.** This refers **to understanding what has happened in the data set**. As the starting point in any analytics process, the descriptive analysis will help **users understand what has happened in the past.**

- **Diagnostic.** The next step of descriptive is diagnostic, which will consider the descriptive analysis and build on top of it **to understand why something happened.** It allows users to gain knowledge on the **exact information of root causes of past events, patterns, etc.**

- **Predictive.** As the name suggests**, predictive analytics will predict what will happen in the future.** This **will combine data from descriptive and diagnostic** analytics and use ML and AI techniques to predict future trends, patterns, problems, etc.

- **Prescriptive.** Prescriptive **analytics takes predictions from predictive analytics and takes it a step further by exploring *how* the predictions will happen**. This can be considered the most important type of analytics as it allows users to understand future events and tailor strategies to handle any predictions effectively.

Accuracy of data analytics

The most important thing to remember is that the accuracy of the analytics is based on the underlying data set. If there are inconsistencies or errors in the dataset, it will result in inefficiencies or outright incorrect analytics.

Any good analytical method will consider external factors like data purity, bias, and variance in the analytical methods. Normalization, purifying, and transforming raw data can significantly help in this aspect.

Data analytics tools & technologies

There are both open source and commercial products for data analytics. They will range from simple analytics tools such as **Microsoft Excel's** Analysis ToolPak that comes with Microsoft Office to SAP BusinessObjects suite and open source tools such as Apache Spark.

When considering cloud providers, **Azure is known as the best platform for data analytics** needs. It provides a complete toolset to cater to any need with its **Azure Synapse Analytics suite, Apache Spark-based Databricks, HDInsights, Machine Learning, etc**. AWS and GCP also provide tools such as Amazon QuickSight, Amazon Kinesis, GCP Stream Analytics to cater to analytics needs.

Additionally, specialized BI tools provide powerful analytics functionality with relatively simple configurations. Examples here include Microsoft PowerBI, SAS Business Intelligence, and Periscope Data Even programming languages like Python or R can be used to create custom analytics scripts and visualizations for more targeted and advanced analytics needs.

Finally, ML algorithms like TensorFlow and scikit-learn can be considered part of the data analytics toolbox—they are popular tools to use in the analytics process.

Data Science

Unlike the first two, data science cannot be limited to a single function or field. Data science is a multidisciplinary approach that extracts information from data by combining:

- Scientific methods
- Maths and statistics
- Programming
- Advanced analytics
- ML and AI
- Deep learning

In data analytics, the primary focus is to gain meaningful insights from the underlying data. The scope of Data Science far exceeds this purpose—data science will deal with everything, from analyzing complex data, creating new analytics algorithms and tools for data processing and purification, and even building powerful, useful visualizations.

Data science tools & technologies

This includes programming languages like R, Python, Julia, which can be used to create new algorithms, ML models, AI processes for big data platforms like Apache Spark and Apache Hadoop.

Data processing and purification tools such as Winpure, Data Ladder, and data visualization tools such as Microsoft Power Platform, Google Data Studio, Tableau to visualization frameworks like matplotlib and ploty can also be considered as data science tools.

As data science covers everything related to data, any tool or technology that is used in Big Data and Data Analytics can somehow be utilized in the Data Science process.