UNIT – I

What is Big Data?

According to Gartner, the definition of Big Data –

"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

This definition clearly answers the "What is Big Data?" question – Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

The History of Big Data

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

Benefits of Big Data and Data Analytics

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Types of Big Data

Now that we are on track with what is big data, let’s have a look at the types of big data:

a) Structured

Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. **For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc.,** will be present in an organized manner.

b) Unstructured

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

c) Semi-structured

Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus we come to the end of types of data.

Characteristics of Big Data

Back in 2001, Gartner analyst Doug Laney listed the **3 ‘V’s of Big Data – Variety, Velocity, and Volume.** Let’s discuss the characteristics of big data.
These characteristics, isolated, are enough to know what big data is. Let’s look at them in depth:

a) Variety

Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

b) Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

c) Volume

Volume is one of the characteristics of big data. We already know that Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses. Thus comes to the end of characteristics of big data.

Why is Big Data Important?

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

1. **Cost Savings**: Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.
2. **Time Reductions: The** high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning.
3. **Understand the market conditions**: By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.
4. **Control online reputation:** Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.
5. **Using Big Data Analytics to Boost Customer Acquisition and Retention**
   The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.
6. **Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights**

Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.

7. **Big Data Analytics As a Driver of Innovations and Product Development**
   Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

Business Intelligence vs Big Data

Although Big Data and Business Intelligence are two technologies used to analyze data to help companies in the decision-making process, there are differences between both of them. They differ in the way they work as much as in the type of data they analyze.

Traditional BI methodology is based on the principle of grouping all business data into a central server. Typically, this data is analyzed in offline mode, after storing the information in an environment called Data Warehouse. The data is structured in a conventional relational database with an additional set of indexes and forms of access to the tables (multidimensional cubes).

A Big Data solution differs in many aspects to BI to use. These are the main differences between Big Data and Business Intelligence:

1. In a Big Data environment, information is stored on a distributed file system, rather than on a central server. It is a much safer and more flexible space.
2. Big Data solutions carry the processing functions to the data, rather than the data to the functions. As the analysis is centered on the information, it´s easier to handle larger amounts of information in a more agile way.
3. Big Data can analyze data in different formats, both structured and unstructured. The volume of unstructured data (those not stored in a traditional database) is growing at levels much higher than the structured data. Nevertheless, its analysis carries different challenges. Big Data solutions solve them by allowing a global analysis of various sources of information.
4. Data processed by Big Data solutions can be historical or come from real-time sources. Thus, companies can make decisions that affect their business in an agile and efficient way.
5. Big Data technology uses parallel mass processing (MPP) concepts, which improves the speed of analysis. With MPP many instructions are executed simultaneously, and since the various jobs are divided into several parallel execution parts, at the end the overall results are reunited and presented. This allows you to analyze large volumes of information quickly.

Big Data vs Data Warehouse

Big Data has become the reality of doing business for organizations today. There is a boom in the amount of structured as well as raw data that floods every organization daily. If this data is managed well, it can lead to powerful insights and quality decision making.

Big data analytics is the process of examining large data sets containing a variety of data types to discover some knowledge in databases, to identify interesting patterns and establish relationships to solve problems, market trends, customer preferences, and other useful information. Companies and businesses that implement Big Data Analytics often reap several business benefits. Companies implement Big Data Analytics because they want to make more informed business decisions.

A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the Data Warehouse through the processes of extraction, transformation and loading (ETL tools). Data analysis tools, such as business intelligence software, access the data within the warehouse.

Hadoop Environment Big Data Analytics

Hadoop is changing the perception of handling Big Data especially the unstructured data. Let's know how Apache Hadoop software library, which is a framework, plays a vital role in handling Big Data. Apache Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers, as both versions might be vulnerable to failures.

Hadoop Community Package Consists of

- File system and OS level abstractions
- A MapReduce engine (either MapReduce or YARN)
- The Hadoop Distributed File System (HDFS)
- Java ARchive (JAR) files
- Scripts needed to start Hadoop
- Source code, documentation and a contribution section

Activities performed on Big Data

- **Store –** Big data need to be collected in a seamless repository, and it is not necessary to store in a single physical database.
- **Process –** The process becomes more tedious than traditional one in terms of cleansing, enriching, calculating, transforming, and running algorithms.
- **Access –** There is no business sense of it at all when the data cannot be searched, retrieved easily, and can be virtually showcased along the business lines.

Classification of analytics

Descriptive analytics

Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.

Data aggregation and **data mining** are two techniques used in descriptive analytics to discover historical data. Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts.

Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning. Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment.

Advantages:

- Quickly and easily report on the Return on Investment (ROI) by showing how performance achieved business or target goals.
- Identify gaps and performance issues early - before they become problems.
- Identify specific learners who require additional support, regardless of how many students or employees there are.
- Identify successful learners in order to offer positive feedback or additional resources.
- Analyze the value and impact of course design and learning resources.

Predictive analytics

Predictive Analytics is a statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors

The software for predictive analytics has moved beyond the realm of statisticians and is becoming more affordable and accessible for different markets and industries, including the field of learning & development.

For online learning specifically, predictive analytics is often found incorporated in the Learning Management System (LMS), but can also be purchased separately as specialized software.

For the learner, predictive forecasting could be as simple as a dashboard located on the main screen after logging in to access a course. Analyzing data from past and current progress, visual indicators in the dashboard could be provided to signal whether the employee was on track with training requirements.

Advantages:

- **Personalize the training needs** of employees by identifying their gaps, strengths, and weaknesses; specific learning resources and training can be offered to support individual needs.
- **Retain Talent** by tracking and understanding employee career progression and forecasting what skills and learning resources would best benefit their career paths. Knowing what skills employees need also benefits the design of future training.
- **Support employees** who may be falling behind or not reaching their potential by offering intervention support before their performance puts them at risk.
- **Simplified reporting** and visuals that keep everyone updated when predictive forecasting is required.

Prescriptive analytics

Prescriptive analytics is a statistical method used to generate recommendations and make decisions based on the computational findings of algorithmic models.

Generating automated decisions or recommendations requires specific and unique algorithmic models and clear direction from those utilizing the analytical technique. A recommendation cannot be generated without knowing what to look for or what problem is desired to be solved. In this way, prescriptive analytics begins with a problem.

Example

A Training Manager uses predictive analysis to discover that most learners without a particular skill will not complete the newly launched course. What could be done? Now prescriptive analytics can be of assistance on the matter and help determine options for action. Perhaps an algorithm can detect the learners who require that new course, but lack that particular skill, and send an automated recommendation that they take an additional training resource to acquire the missing skill.

The accuracy of a generated decision or recommendation, however, is only as good as the quality of data and the algorithmic models developed. What may work for one company's training needs may not make sense when put into practice in another company's training department. Models are generally recommended to be tailored for each unique situation and need.

Descriptive vs Predictive vs Prescriptive Analytics

Descriptive Analytics is focused solely on historical data.

You can think of Predictive Analytics as then using this historical data to develop statistical models that will then forecast about future possibilities.

Prescriptive Analytics takes Predictive Analytics a step further and takes the possible forecasted outcomes and predicts consequences for these outcomes.

What Big Data Analytics Challenges

1. Need For Synchronization Across Disparate Data Sources

As data sets are becoming bigger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is overlooked, it will create gaps and lead to wrong messages and insights.

2. Acute Shortage Of Professionals Who Understand Big Data Analysis

The analysis of data is important to make this voluminous amount of data being produced in every minute, useful. With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been created in the market. It is important for business organizations to hire a data scientist having skills that are varied as the job of a data scientist is multidisciplinary. Another major challenge faced by businesses is the shortage of professionals who understand Big Data analysis. There is a sharp shortage of data scientists in comparison to the massive amount of data being produced.

3. Getting Meaningful Insights Through The Use Of Big Data Analytics

It is imperative for business organizations to gain important insights from Big Data analytics, and also it is important that only the relevant department has access to this information. A big challenge faced by the companies in the Big Data analytics is mending this wide gap in an effective manner.

4. Getting Voluminous Data Into The Big Data Platform

It is hardly surprising that data is growing with every passing day. This simply indicates that business organizations need to handle a large amount of data on daily basis. The amount and variety of data available these days can overwhelm any data engineer and that is why it is considered vital to make data accessibility easy and convenient for brand owners and managers.

5. Uncertainty Of Data Management Landscape

With the rise of Big Data, new technologies and companies are being developed every day. However, a big challenge faced by the companies in the Big Data analytics is to find out which technology will be best suited to them without the introduction of new problems and potential risks.

6. Data Storage And Quality

Business organizations are growing at a rapid pace. With the tremendous growth of the companies and large business organizations, increases the amount of data produced. The storage of this massive amount of data is becoming a real challenge for everyone. Popular data storage options like data lakes/ warehouses are commonly used to gather and store large quantities of unstructured and structured data in its native format. The real problem arises when a data lakes/ warehouse try to combine unstructured and inconsistent data from diverse sources, it encounters errors. Missing data, inconsistent data, logic conflicts, and duplicates data all result in data quality challenges.

7. Security And Privacy Of Data

Once business enterprises discover how to use Big Data, it brings them a wide range of possibilities and opportunities. However, it also involves the potential risks associated with big data when it comes to the privacy and the security of the data. The Big Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, the rise of voluminous amount of data increases privacy and security concerns.

Terminologies Used In Big Data Environments

- **As-a-service infrastructure**

Data-as-a-service, software-as-a-service, platform-as-a-service – all refer to the idea that rather than selling data, licences to use data, or platforms for running Big Data technology, it can be provided "as a service", rather than as a product. This reduces the upfront capital investment necessary for customers to begin putting their data, or platforms, to work for them, as the provider bears all of the costs of setting up and hosting the infrastructure. As a customer, as-a-service infrastructure can greatly reduce the initial cost and setup time of getting Big Data initiatives up and running.

- **Data science**

Data science is the professional field that deals with turning data into value such as new insights or predictive models. It brings together expertise from fields including statistics, mathematics, computer science, communication as well as domain expertise such as business knowledge. Data scientist has recently been voted the No 1 job in the U.S., based on current demand and salary and career opportunities.

- **Data mining**

Data mining is the process of discovering insights from data. In terms of Big Data, because it is so large, this is generally done by computational methods in an automated way using methods such as decision trees, clustering analysis and, most recently, machine learning. This can be thought of as using the brute mathematical power of computers to spot patterns in data which would not be visible to the human eye due to the complexity of the dataset.

- **Hadoop**

Hadoop is a framework for Big Data computing which has been released into the public domain as open source software, and so can freely be used by anyone. It consists of a number of modules all tailored for a different vital step of the Big Data process – from file storage (Hadoop File System – HDFS) to database (HBase) to carrying out data operations (Hadoop MapReduce – see below). It has become so popular due to its power and flexibility that it has developed its own industry of retailers (selling tailored versions), support service providers and consultants.

- **Predictive modelling**

At its simplest, this is predicting what will happen next based on data about what has happened previously. In the Big Data age, because there is more data around than ever before, predictions are becoming more and more accurate. Predictive modelling is a core component of most Big Data initiatives, which are formulated to help us choose the course of action which will lead to the most desirable outcome. The speed of modern computers and the volume of data available means that predictions can be made based on a huge number of variables, allowing an ever-increasing number of variables to be assessed for the probability that it will lead to success.

- **MapReduce**

MapReduce is a computing procedure for working with large datasets, which was devised due to difficulty of reading and analysing really Big Data using conventional computing methodologies. As its name suggest, it consists of two procedures – mapping (sorting information into the format needed for analysis – i.e. sorting a list of people according to their age) and reducing (performing an operation, such checking the age of everyone in the dataset to see who is over 21).

- **NoSQL**

NoSQL refers to a database format designed to hold more than data which is simply arranged into tables, rows, and columns, as is the case in a conventional relational database. This database format has proven very popular in Big Data applications because Big Data is often messy, unstructured and does not easily fit into traditional database frameworks.

- **Python**

Python is a programming language which has become very popular in the Big Data space due to its ability to work very well with large, unstructured datasets (see Part II for the difference between structured and unstructured data). It is considered to be easier to learn for a data science beginner than other languages such as R (see also Part II) and more flexible.

- **R Programming**

R is another programming language commonly used in Big Data, and can be thought of as more specialised than Python, being geared towards statistics. Its strength lies in its powerful handling of structured data. Like Python, it has an active community of users who are constantly expanding and adding to its capabilities by creating new libraries and extensions.

- **Recommendation engine**

A recommendation engine is basically an algorithm, or collection of algorithms, designed to match an entity (for example, a customer) with something they are looking for. Recommendation engines used by the likes of Netflix or Amazon heavily rely on Big Data technology to gain an overview of their customers and, using predictive modelling, match them with products to buy or content to consume. The economic incentives offered by recommendation engines has been a driving force behind a lot of commercial Big Data initiatives and developments over the last decade.

- **Real-time**

Real-time means “as it happens” and in Big Data refers to a system or process which is able to give data-driven insights based on what is happening at the present moment. Recent years have seen a large push for the development of systems capable of processing and offering insights in real-time (or near-real-time), and advances in computing power as well as development of techniques such as machine learning have made it a reality in many applications today.

- **Reporting**

The crucial “last step” of many Big Data initiative involves getting the right information to the people who need it to make decisions, at the right time. When this step is automated, analytics is applied to the insights themselves to ensure that they are communicated in a way that they will be understood and easy to act on. This will usually involve creating multiple reports based on the same data or insights but each intended for a different audience (for example, in-depth technical analysis for engineers, and an overview of the impact on the bottom line for c-level executives).

- **Spark**

Spark is another open source framework like Hadoop but more recently developed and more suited to handling cutting-edge Big Data tasks involving real time analytics and machine learning. Unlike Hadoop it does not include its own filesystem, though it is designed to work with Hadoop's HDFS or a number of other options. However, for certain data related processes it is able to calculate at over 100 times the speed of Hadoop, thanks to its in-memory processing capability. This means it is becoming an increasingly popular choice for projects involving deep learning, neural networks and other compute-intensive tasks.

- **Structured Data**

Structured data is simply data that can be arranged neatly into charts and tables consisting of rows, columns or multi-dimensioned matrixes. This is traditionally the way that computers have stored data, and information in this format can easily and simply be processed and mined for insights. Data gathered from machines is often a good example of structured data, where various data points – speed, temperature, rate of failure, RPM etc. – can be neatly recorded and tabulated for analysis.

- **Unstructured Data**

Unstructured data is any data which cannot easily be put into conventional charts and tables. This can include video data, pictures, recorded sounds, text written in human languages and a great deal more. This data has traditionally been far harder to draw insight from using computers which were generally designed to read and analyze structured information. However, since it has become apparent that a huge amount of value can be locked away in this unstructured data, great efforts have been made to create applications which are capable of understanding unstructured data – for example visual recognition and natural language processing.

- **Visualization**

Humans find it very hard to understand and draw insights from large amounts of text or numerical data – we can do it, but it takes time, and our concentration and attention is limited. For this reason effort has been made to develop computer applications capable of rendering information in a visual form – charts and graphics which highlight the most important insights which have resulted from our Big Data projects. A subfield of reporting (see above), visualizing is now often an automated process, with visualizations customized by algorithm to be understandable to the people who need to act or take decisions based on them.

Basic availability, Soft state and Eventual consistency

Basic availability implies continuous system availability despite network failures **and** tolerance to temporary in**consistency**.

Soft state refers to **state** change without input which is required for **eventual consistency**.

Eventual consistency means that if no further updates are made to a given updated data**base** item for long enough period of time , all users will see the same value for the updated item.

Top Analytics Tools

* **R** is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

- Effective data handling and storage facility,
- It provides a suite of operators for calculations on arrays, in particular, matrices,
- It provides coherent, integrated collection of big data tools for data analysis
- It provides graphical facilities for data analysis which display either on-screen or on hardcopy

* **Apache Spark** is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

Features:

- It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk
- It offers lighting Fast Processing
- Support for Sophisticated Analytics
- Ability to Integrate with Hadoop and Existing Hadoop Data

* **Plotly** is an analytics tool that lets users create charts and dashboards to share online.

Features:

- Easily turn any data into eye-catching and informative graphics
- It provides audited industries with fine-grained information on data provenance
- Plotly offers unlimited public file hosting through its free community plan

* **Lumify** is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

Features:

- It provides both 2D and 3D graph visualizations with a variety of automatic layouts

- It provides a variety of options for analyzing the links between entities on the graph
- It comes with specific ingest processing and interface elements for textual content, images, and videos
- It spaces feature allows you to organize work into a set of projects, or workspaces
- It is built on proven, scalable big data technologies

* **IBM SPSS Modeler** is a predictive big data analytics platform. It offers predictive models and delivers to individuals, groups, systems and the enterprise. It has a range of advanced algorithms and analysis techniques.

Features:

- Discover insights and solve problems faster by analyzing structured and unstructured data
- Use an intuitive interface for everyone to learn
- You can select from on-premises, cloud and hybrid deployment options
- Quickly choose the best performing algorithm based on model performance

* **MongoDB** is a NoSQL, document-oriented database written in C, C++, and JavaScript. It is free to use and is an open source tool that supports multiple operating systems including Windows Vista (and later versions), OS X (10.7 and later versions), Linux, Solaris, and FreeBSD.

Its main features include Aggregation, Adhoc-queries, Uses BSON format, Sharding, Indexing, Replication, Server-side execution of javascript, Schemaless, Capped collection, MongoDB management service (MMS), load balancing and file storage.

Features:

- Easy to learn.
- Provides support for multiple technologies and platforms.
- No hiccups in installation and maintenance.
- Reliable and low cost.