Subject: Fundamental of Statistics and AI

UNIT 3

3.1 Big Data

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.

1. Volume:

- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.
- *Example*: In the year 2020 we had almost 40000 ExaBytes of data.

2. Velocity:

- Velocity refers to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like 'velocity'.
- *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

3. Variety:

- It refers to nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

- **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
- **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.
- **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:

- After having the 4 V's into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V's

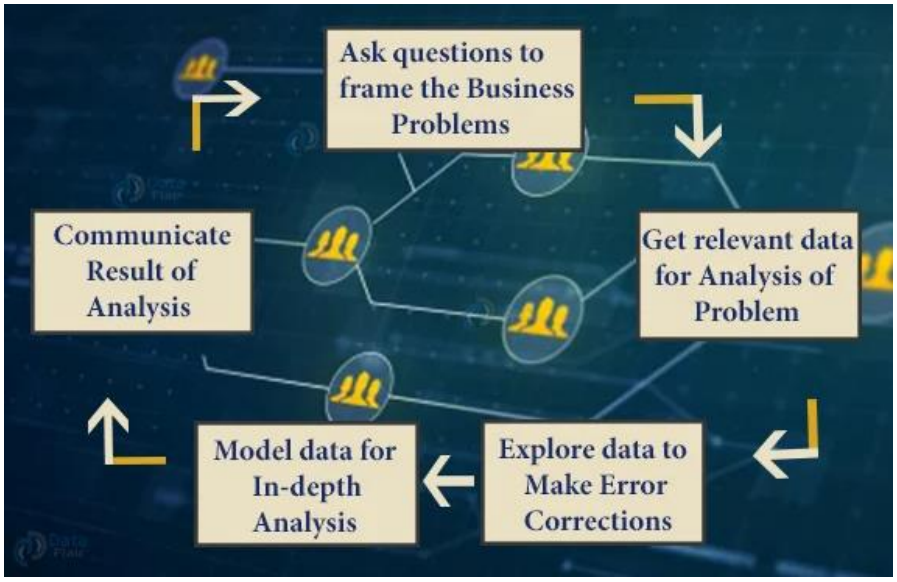
3.2 Data Science

- Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.
- Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems.
- It is the future of artificial intelligence.

In short, we can say that data science is all about:

- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.

3.3 Data Science Process



Step 1. Ask Questions to Frame the Business Problem

- In the first step, try to get an idea of what are the needs of a company and extract data based on it.
- The process of data science begin by asking the right questions to find what the problem is.
- Let's take a very common problem of a bag company – The sales problem.
    - Who are the target market and the customers?
    - How do you approach the target market?
    - How does the sales process look currently?
    - What information do you have about the target market?
    - How can we identify customers who are more likely to buy our product?
    -

Step 2. Get Relevant Data for Analysis of the Problem

- It is time to collect the data that will help you solve the problem. Before gathering the data, you should ask if the data required is already available with the company?
- Data related to following is required: *age, gender, previous customers transaction history*, etc.

Step 3. Explore the Data to Make Error Corrections

- Exploring the data is actually cleaning and organizing it.
- More than 70% of the data scientist's time is spent on this process.
- Various tools and techniques are put to use for this purpose like Python, R, SQL, etc.
- Are there missing values in the data i.e. are there customers without their contact numbers?
- Are there any invalid values? If there are, how can you fix it?
- Are there multiple datasets? Is merging datasets a good choice? If yes, then how should you merge them?

Step 4. Model the Data for In-depth Analysis

- Build a model of the data to answer the question.
- Validate the model against the data collected.
- Usage of various visualization tools to present data.
- Perform the necessary algorithms and statistical analysis.
- Compare results against other techniques and sources.

Step 5. Communicate the Results of the Analysis

- Graph or chart the information for presentation with tools – R, Python, Tableau, Excel.
- Use "storytelling" to fit the results.
- Answer the various follow-up questions.
- Present data in different formats- reports, websites.
- Answers will always spark more questions, and the process begins again.

3.4 Business Intelligence

- Business intelligence(BI) is basically a set of technologies, applications, and processes that are used by enterprises for business data analysis.
- Business intelligence tools enhance the chances of an enterprise to enter a new market as well as help in studying the impact of marketing efforts.

Difference between Business Intelligence and Data Science

Factor	Data Science	Business Intelligence
Concept	It is a field that uses mathematics, statistics and various other tools to discover the hidden patterns in the data.	It is basically a set of technologies, applications and processes that are used by the enterprises for business data analysis.
Focus	It focuses on the future.	It focuses the past and present.
Data	It deals with both structured as well as unstructured data.	It mainly deals only with structured data.
Flexibility	Data science is much more flexible as data sources can be added as per requirement.	It is less flexible as in case of business intelligence data sources need to be pre-planned.
Method	It makes the use of scientific method.	It makes the use of analytic method.
Complexity	It has a higher complexity in comparison to business intelligence.	It is much simpler when compared to data science.
Expertise	It's expertise is data scientist.	It's expertise is business user.
Questions	It deals with the questions what will happen and what if.	It deals with the question what happened.
Tools	It's tools are SAS, Apache Haddop & Spark, BigML,etc.	It's tools are InsightSquared Sales Analytics, Klipfolio, ThoughtSpot, Cyfe, TIBCO Spotfire etc.

3.5 Roles & Responsibilities of a Data Scientist

Management

- The Data Scientist plays an insignificant managerial role where he supports the construction of the base of futuristic and technical abilities within the Data and Analytics field in order to assist various planned and continuing data analytics projects.

Analytics

- The Data Scientist represents a scientific role where he plans, implements, and assesses high-level statistical models and strategies for application in the business's most complex issues.
- The Data Scientist develops econometric and statistical models for various problems including projections, classification, clustering, pattern analysis, sampling, simulations, and so forth.

Strategy/Design

- The Data Scientist performs a vital role in the advancement of innovative strategies to understand the business's consumer trends and management as well as ways to solve difficult business problems, for instance, the optimization of product fulfillment and entire profit.

Collaboration

- The role of the Data Scientist is not a solitary role and in this position, he collaborates with superior data scientists to communicate obstacles and findings to relevant stakeholders in an effort to enhance drive business performance and decision-making.

Knowledge

- The Data Scientist also takes leadership to explore different technologies and tools with the vision of creating innovative data-driven insights for the business at the most agile pace feasible. In this situation, the Data Scientist also uses initiative in assessing and utilizing new and enhanced data science methods for the business, which he delivers to senior management of approval.