ITDO6014
AI AND DS-1

Module 4: Introduction to DS

Introduction to Data Science

- **What is Data?**
- Data is the collection of facts and bits of information. In the real world, the data is either structured or unstructured.
- **Structured data** is data that has an order and a well-defined structure. As the structured data is consistent and well-defined, it is an easy task to store and access it. Also, searching for data is easy as we can use indexes to store structured data.

Introduction to Data Science

- Another type is unstructured data. It is an inconsistent type as it doesn't have any structure, format, or sequence. The unstructured data is error-prone when we perform indexing on it. Hence, it is a difficult task to understand and operate on unstructured data. Interestingly, in the real world, more than structured data, what we have always is inconsistent unstructured data. It can be in the form of audio, video, text, or any other format.

Introduction to Data Science

- **Why is data important?**
- Look at the statistics below to see what happens in the daily data life:
- Average daily –
- People across the world:
  - Send more than 300 billion emails and 500 million tweets
  - Send over 65 billion messages via WhatsApp
  - Perform 5.6 billion searches on Google
- Facebook creates nearly 4 petabytes of data
- By the year 2025, there will be 463 exabytes of data worldwide!

Introduction to Data Science

- Data is one of the biggest assets any company has in the present time. This, in fact, was long predicted by **Forbes** when it stated: '**The total data market is expected to nearly double in size. It will grow from US$69.6 billion in revenue in 2015 to US$132.3 billion in 2020.**' By these statistics, we can infer how important data is and the need to utilize it for businesses.

Introduction to Data Science

- **Use Case of Bank Payments**
- Suppose, some customers make payments to their respective merchants (such as Paytm, Amazon, Flipkart, etc.). The customers use the Citi bank debit card for the transactions. Now, the merchants collect the data related to transactions. This may include the mode of payment, data of the payment receivers, the time of the transaction, and the amount. The merchants analyze the data and build specific data products on top of these parameters. These data products exclude the confidential details of the customers. They consist of the following details of the transactions:

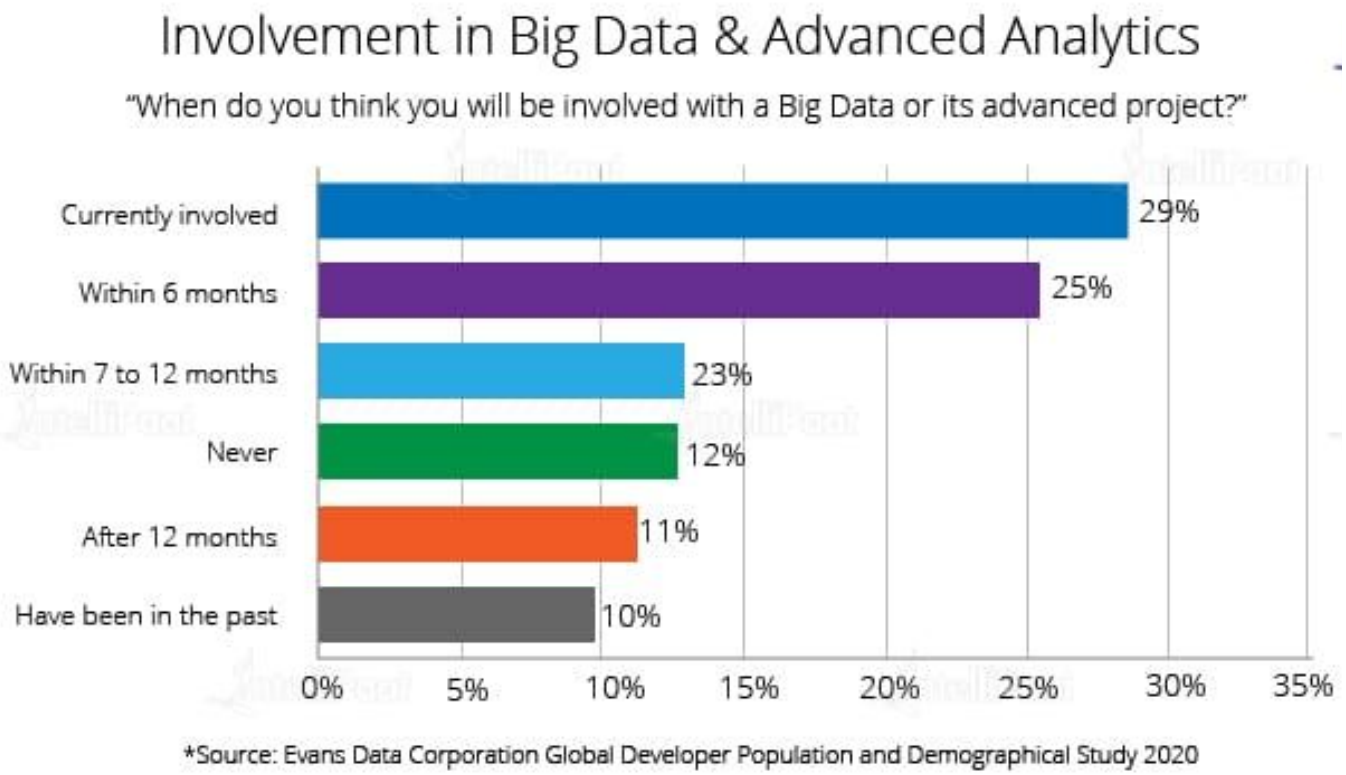
Introduction to Data Science

- The **banks utilize the data to target customers by providing them with exciting offers.** Due to this, the customers start making transactions through those banks that provide the greatest offer. These customer payments increase the revenue base of the banks. This is how data helps in increasing revenue generation for the banks, as well as for the merchants.

Introduction to Data Science

- **What is Big Data?**
- Big Data, Data Science, and Data Analytics are not just some technical jargon but are significant concepts contributing to the field of technology. While these terms are interlinked, there are fundamental differences among them.
- According to Forbes, today, **there are millions of developers (more than 25% of developers globally) who are working on projects of Big Data and Advanced Analytics.**

Introduction to Data Science



*Source: Evans Data Corporation Global Developer Population and Demographical Study 2020

Introduction to Data Science

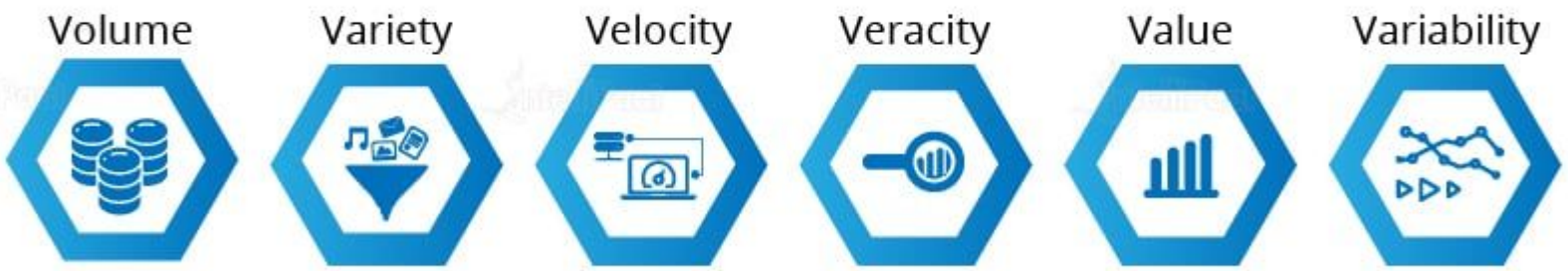
- Big Data refers to huge volumes of data. It deals with large and complex sets of data that a traditional data processing system cannot handle. Big Data consists of tools and techniques that extract data, store it systematically, and extract useful information out of the data. Here are various types of data that Big Data deals with:

Introduction to Data Science

- **Structured Data**: This type of data contains organized data. It has a fixed schema. Thus, it is easy to understand and analyze structured data.
- **Semi-structured Data**: The data in the form of various file formats like XML, JSON, and CSV is categorized as semi-structured data. It is partially organized data, which makes it difficult to understand.
- **Unstructured Data**: This type of data does not have a well-defined structure or a schema. The real-world data is always unstructured and hence challenging to understand. This data is generated through various digital channels including mobile phones, the Internet, social media, and e-commerce websites.

Introduction to Data Science

- **Characteristics of Big Data**
- There are certain characteristics of Big Data that define its structure and importance of it. The six characteristics of Big Data are described below:



Introduction to Data Science

- **Volume**: The amount of data generated per day from multiple sources is very high. Previously, it was a redundant task to store this big data. But, with the help of **Big Data Hadoop**, we can efficiently store these huge volumes of data.
- **Variety**: There are a variety of data collected from different sources. It can be an audio file, video, images, documents, or unstructured text. The tools in Big Data help in processing this variety of structured and unstructured data.
- **Velocity**: In this digital era, the number of Internet users is increasing rapidly day by day. Due to this, the speed of data generation gets enhanced. The term Velocity refers to how fast this data generation and its processing are happening. It is used to understand the trends in the data and meet the demands of the market.

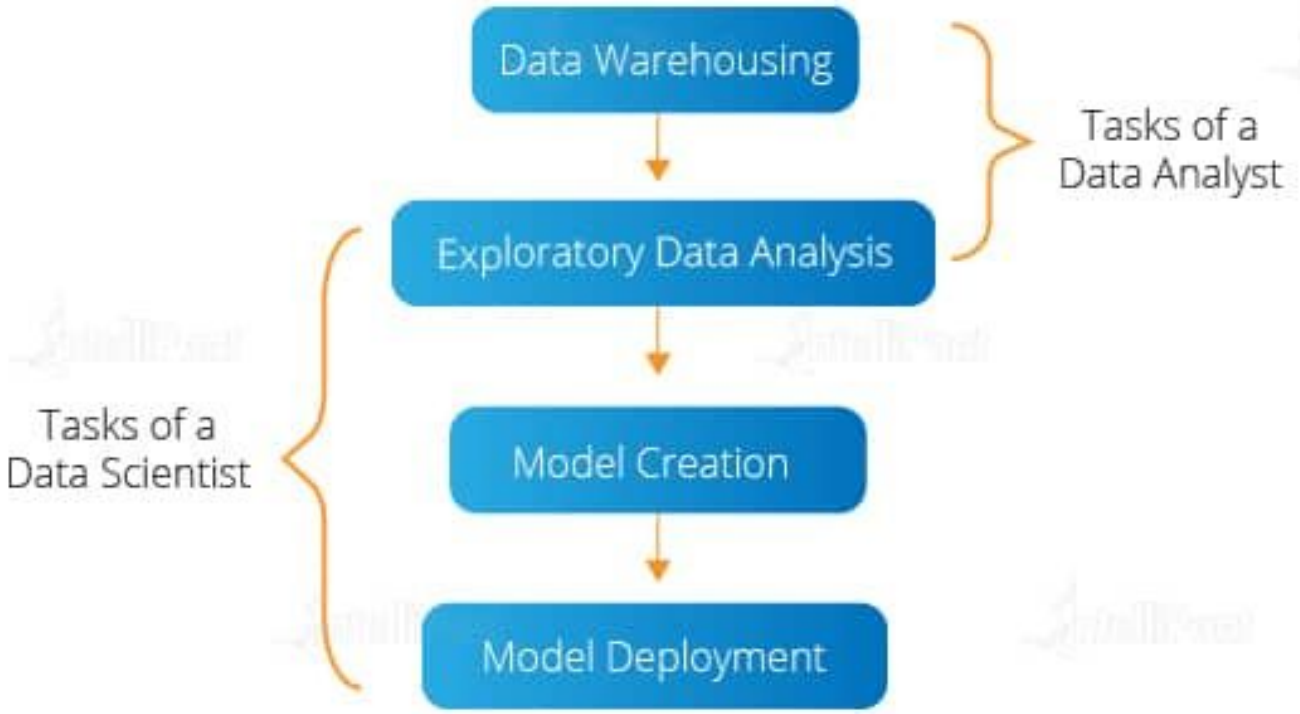
Introduction to Data Science

- **Veracity**: It relates to the quality of the data collected. Organizations need to take care of the quality of data while collecting it so that the data is relevant to them.
- **Value**: Big Data focuses on collecting data that creates some business value for the organizations. This helps them compete in the market and increase their profits.
- **Variability**: There is always a change in trends in the market. Variability refers to how often this change happens. Big Data helps in managing these drifts of data that benefit organizations to come up with the latest products.

Introduction to Data Science

- **What is Data Analytics?**
- Data Analytics seeks to provide operational insights into complex business situations. The prime concern of a Data Analyst is looking into the historical data from a modern perspective and then, finding new and challenging business scenarios. After that, he/she applies methodologies to find better solutions. Not only this, but a Data Analyst also predicts the upcoming opportunities that the company can exploit.
- The responsibilities of a Data Analyst and a Data Scientist are similar to each other. However, they differ in the implementation part. The below diagram shows the difference between the responsibilities of a Data Analyst and a Data Scientist.

Introduction to Data Science



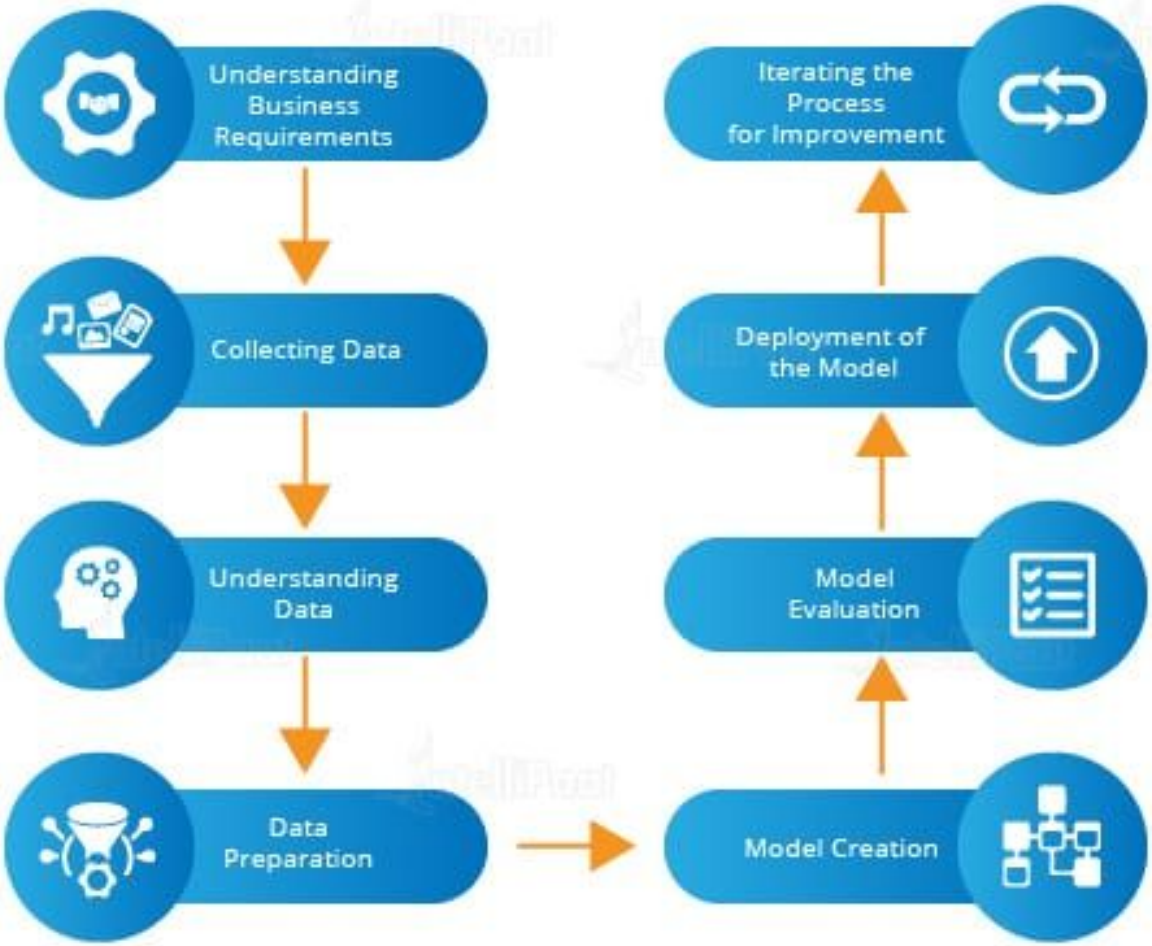
Introduction to Data Science

- Data Analysts collect data for their organizations from multiple sources. They perform exploratory data analysis to visualize the data. Then, they filter and clean the data by checking the reports generated with the help of the Data Analytics tools. After that, the data is analyzed with the help of a data visualization tool. Also, they build effective strategies to optimize the statistical analysis of the data. This helps organizations note down the growth or the market trend.
- Some of the tools used for Data Analytics are:
- R programming, Python, Tableau Public, SAS, RapidMiner, KNIME
- QlikView, Splunk

Introduction to Data Science

- **What is Data Science?**
- Data Science deals with the slicing and dicing of big chunks of data. It uses techniques to obtain insightful patterns and trends from the data. Data Scientists are responsible for uncovering the facts hidden in the complex web of unstructured data. This helps in making important business decisions in accordance with market trends. Data Science also involves the creation of Machine Learning models on top of the visualized data. To understand Data Science thoroughly, let's look at the **Data Science life cycle**:

Introduction to Data Science



Introduction to Data Science

- Understanding the Life Cycle of Data Science
- **Understanding business requirements**: Data Scientists perform a structural analysis of the business model. Then, they understand the market trends and customer needs. This helps to gather business requirements.
- **Collecting data**: The collection of valuable data is a necessary step in Data Science. The data is collected from multiple sources.
- **Data understanding**: The next step after data collection is understanding the data. For this, Data Scientists use data visualization tools and techniques.

Introduction to Data Science

- **Data preparation**: Since organizations need to create an effective strategy and model on the basis of data, Data Scientists prepare data accordingly. Suppose, if the need is for building a recommendation system on fashion trends, then Data Scientists have to prepare the data relevant to the trending fashion.
- **Model creation**: Data Science widely uses Machine Learning for building systems and models on top of the dataset prepared. Data Scientists use Machine Learning algorithms and techniques to build models. Organizations use these models to fulfill their business requirements.

Introduction to Data Science

- **Model evaluation**: Building a model is not enough. They have to assess the accuracy of the model. So, they use different data to train and evaluate the built model.
- **Deployment of the model**: After checking the performance of the model, it is deployed for implementation.
- **Iteration of the process**: The systems built with the help of Machine Learning learn from their experience. For this, Data Scientists expose them to a variety of real-time datasets. And the iteration of the learning process makes the models more accurate.

Introduction to Data Science

- Tools used by Data Scientist
- Tools used by Data Scientists for implementing the above steps are:
- Statistics and probability
- R and Python programming
- Tableau and Power BI for data visualization
- Machine Learning algorithms

Introduction to Data Science

	Impact on Various Sectors	
Big Data	**Data Science**	**Data Analytics**
• Retail • Banking and investment • Fraud detection and analyzing • Customer-centric applications • Operational analysis	• Web development • Digital advertisements • E-commerce • Internet search • Finance • Telecom • Utilities	• Travelling and transportation • Financial analysis • Retail • Research • Energy management • Healthcare
	Skills Required	
• Analytical skills • Mathematics and statistics • Java • Hadoop	• SAS • R/Python programming • Hadoop • SQL database • Analytical skills • Statistics • Mathematics • Visionary thinking	• Programming • Communication • Artificial Intelligence • Data wrangling skills

Introduction to Data Science

- **Skills for Becoming a Data Scientist**
- Data Science is a broad field of study. It requires knowledge of various fields such as programming, database, and Machine Learning. According to Forbes, 'Data Scientist jobs are among the best jobs in the IT industry.'
- To become a Data Scientist, you must acquire the below skillset:
- Good grasp of Python and R programming language
- Knowledge of mathematics especially statistics and probability
- Awareness of SQL database queries
- Knowledge of data mining
- Knowledge of how to work on data visualization tools

Introduction to Data Science

- Roles in Data Science Projects:
- https://www.projectpro.io/article/data-science-roles/647