**UNIT-1**

**INTRODUCTION TO DATA SCIENCE**

1. **What is a data science? What is the importance of data science?**

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.

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that that you can find something new and meaningful data science uses the most powerful hardware,programming systems and more efficient algorithms to solve the data related problems. It is the future of artificial intelligence

**Example:** Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

But in today's world, data is becoming so vast, i.e., approximately ****2.5 quintals bytes**** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

* With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
* Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.
* Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
* Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

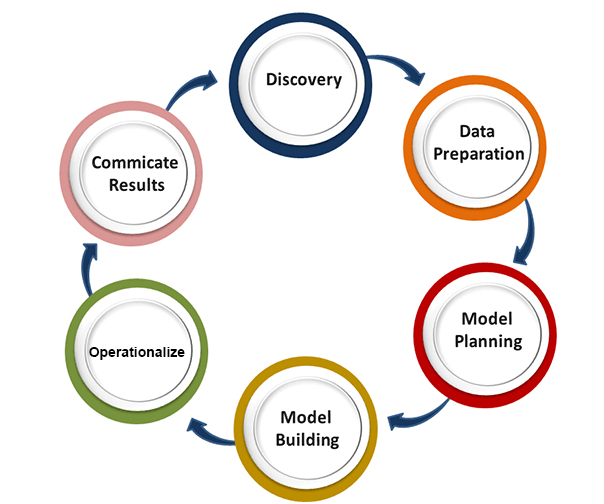
## Explain the Applications of Data Science:

* ****Image recognition and speech recognition:****  
  Data science is currently using for Image and speech recognition. When you upload an image on Facebook and start getting the suggestion to tag to your friends. This automatic tagging suggestion uses image recognition algorithm, which is part of data science.  
  When you say something using, "Ok Google, Siri, Cortana", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.
* ****Gaming world:****  
  In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.
* ****Internet search:****  
  When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.
* ****Transport:****  
  Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.
* ****Healthcare:****  
  In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.
* ****Recommendation systems:****  
  Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.
* ****Risk detection:****  
  Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued.  
  Most of the finance companies are looking for the data scientist to avoid risk and any type of losses with an increase in customer satisfaction.

**3.Explain about life cycle of the data science or process of data science?**

## Data Science Lifecycle

The life-cycle of data science is explained as below diagram.



The main phases of data science life cycle are given below:

****1. Discovery:**** The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget. In this phase, we need to determine all the requirements of the project such as the number of people, technology, time, data, an end goal, and then we can frame the business problem on first hypothesis level.

****2. Data preparation:**** Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:

* Data cleaning
* Data Reduction
* Data integration
* Data transformation,

After performing all the above tasks, we can easily use this data for our further processes.

****3. Model Planning:**** In this phase, we need to determine the various methods and techniques to establish the relation between input variables. We will apply Exploratory data analytics(EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:

* SQL Analysis Services
* R
* SAS
* Python

****4. Model-building:**** In this phase, the process of model building starts. We will create datasets for training and testing purpose. We will apply different techniques such as association, classification, and clustering, to build the model.

Following are some common Model building tools:

* SAS Enterprise Miner
* WEKA
* SPCS Modeler
* MATLAB

****5. Operationalize:**** In this phase, we will deliver the final reports of the project, along with briefings, code, and technical documents. This phase provides you a clear overview of complete project performance and other components on a small scale before the full deployment.

****6. Communicate results:**** In this phase, we will check if we reach the goal, which we have set on the initial phase. We will communicate the findings and final result with the business team.

**4.What is a Data Scientist? Explain the Responsibilities of Data Scientist?**

**Data scientist:** A data scientist is a professional responsible for collecting,analyzing and interpreting extremely large amount of data. The data scientist role in on offshoot of several traditional technical roles , including mathematician, scientist ,statistician and computer professional. This job requires the use of advanced analytic technologies including machine learning and predictive modeling

* Data mining or extracting usable data from valuable data sources.
* Using machine learning tools to select features create and optimize classifiers.
* Carrying out preprocessing of structured and unstructured data
* Enhancing data collection procedures to include all relevant information for developing analytic systems.
* Processing , cleaning and validating the integrity of data to be used for analysis.
* Analyzing large amount of information to fing patterns and solutions.
* Developing prediction systems and machine learning algorithms
* Presenting results in a clear manner
* Propose solutions and strategies to tackle business challenges
* Collaborate with business and IT teams

1. **What is the qualification of data scientist?**

* Programming skills-knowledge of statistical programming language like R,Python,and database query languages like SQL,Hive,Pig is desirable. Familiarity with java, or C++ is an added advantages
* Statistics-good applied statistical skills includes knowledge of statistical tests, distributions, regression,maximum likelihood estimators etc
* Machine learning-Good knowledge of machine learning methods like K-nearest neighbors, navie bayes , SVM , Decision Forests.
* Strong maths skills
* Data wrangling - proficiency in handling imperfections in data is an important aspect of a data scientist job description
* Stong software engineering background
* Hands on experience with data science tools
* Problem solving aptitude
* Analytical mind and great business sense
* Degree in computer science or any engineering or relevant field is preferred
* Experience as data analyst or Data scientist

### **6. Explain the different types of data science jobs**

If you learn data science, then you get the opportunity to find the various exciting job roles in this domain. The main job roles are given below:

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer
5. Data Architect
6. Data Administrator
7. Business Analyst
8. Business Intelligence Manager

Below is the explanation of some critical job titles of data science.

****1. Data Analyst:****

Data analyst is an individual, who performs mining of huge amount of data, models the data, looks for patterns, relationship, trends, and so on. At the end of the day, he comes up with visualization and reporting for analyzing the data for decision making and problem-solving process.

****Skill required:**** For becoming a data analyst, you must get a good background in ****mathematics, business intelligence, data mining****, and basic knowledge of ****statistics****. You should also be familiar with some computer languages and tools such as ****MATLAB, Python, SQL, Hive, Pig, Excel, SAS, R, JS, Spark****, etc.

****2. Machine Learning Expert:****

The machine learning expert is the one who works with various machine learning algorithms used in data science such as ****regression, clustering, classification, decision tree, random forest****, etc.

****Skill Required:**** Computer programming languages such as Python, C++, R, Java, and Hadoop. You should also have an understanding of various algorithms, problem-solving analytical skill, probability, and statistics.

****3. Data Engineer:****

A data engineer works with massive amount of data and responsible for building and maintaining the data architecture of a data science project. Data engineer also works for the creation of data set processes used in modeling, mining, acquisition, and verification.

****Skill required:**** Data engineer must have depth knowledge of ****SQL, MongoDB, Cassandra, HBase, Apache Spark, Hive, MapReduce****, with language knowledge of ****Python, C/C++, Java, Perl****, etc.

****4. Data Scientist:****

A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.

****Skill required:**** To become a data scientist, one should have technical language skills such as ****R, SAS, SQL, Python, Hive, Pig, Apache spark, MATLAB****. Data scientists must have an understanding of Statistics, Mathematics, visualization, and communication skills.

**7.Why to use phython for data science**

Python provides many useful features which make it popular and valuable from the other programming languages. Python is open source , interpreted and high level language It supports object-oriented programming, procedural programming approaches and provides dynamic memory allocation. We have listed below a few essential features.

### 1) Easy to Learn and Use

Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It is the recommended programming language for beginners.

### 2) Expressive Language

Python can perform complex tasks using a few lines of code. A simple example, the hello world program you simply type ****print("Hello World")****. It will take only one line to execute, while Java or C takes multiple lines.

### 3) Interpreted Language

Python is an interpreted language; it means the Python program is executed one line at a time. The advantage of being interpreted language, it makes debugging easy and portable.

### 4) Cross-platform Language

Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

### 5) Free and Open Source

Python is freely available for everyone. It is freely available on its official website [www.python.org](https://www.python.org/" \t "https://www.javatpoint.com/blank). It has a large community across the world that is dedicatedly working towards make new python modules and functions. Anyone can contribute to the Python community. The open-source means, "Anyone can download its source code without paying any penny."

### 6) Object-Oriented Language

Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code.

### 7) Extensible

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our Python code. It converts the program into byte code, and any platform can use that byte code.

### 8) Large Standard Library

It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting. There are various machine learning libraries, such as Tensor flow, Pandas, Numpy, Keras, and Pytorch, etc. Django, flask, pyramids are the popular framework for Python web development.

### 9) GUI Programming Support

Graphical User Interface is used for the developing Desktop application. PyQT5, Tkinter, Kivy are the libraries which are used for developing the web application.

### 10) Integrated

It can be easily integrated with languages like C, C++, and JAVA, etc. Python runs code line by line like C,C++ Java. It makes easy to debug the code.

### 11. Embeddable

The code of the other programming language can use in the Python source code. We can use Python source code in another programming language as well. It can embed other language into our code.

### 12. Dynamic Memory Allocation

In Python, we don't need to specify the data-type of the variable. When we assign some value to the variable, it automatically allocates the memory to the variable at run time. Suppose we are assigned integer value 15 to ****x,**** then we don't need to write ****int x = 15.**** Just write x = 15.

**8.Explain the components of the data science**

The main components of Data Science are given below:

****1. Statistics:**** Statistics is one of the most important components of data science. Statistics is a way to collect and analyze the numerical data in a large amount and finding meaningful insights from it.

****2. Domain Expertise:**** In data science, domain expertise binds data science together. Domain expertise means specialized knowledge or skills of a particular area. In data science, there are various areas for which we need domain experts.

****3. Data engineering:**** Data engineering is a part of data science, which involves acquiring, storing, retrieving, and transforming the data. Data engineering also includes metadata (data about data) to the data.

****4. Visualization:**** Data visualization is meant by representing data in a visual context so that people can easily understand the significance of data. Data visualization makes it easy to access the huge amount of data in visuals.

****5. Advanced computing:**** Heavy lifting of data science is advanced computing. Advanced computing involves designing, writing, debugging, and maintaining the source code of computer programs.

****6. Mathematics:**** Mathematics is the critical part of data science. Mathematics involves the study of quantity, structure, space, and changes. For a data scientist, knowledge of good mathematics is essential.

****7. Machine learning:**** Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain. In data science, we use various machine learning algorithms to solve the problems.