Data Science:
What is, Basics & Process

What is Data Science?

- Data Science is about data gathering, analysis and decision-making.
- Data Science is about finding patterns in data, through analysis, and make future predictions.

In generally, Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

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.

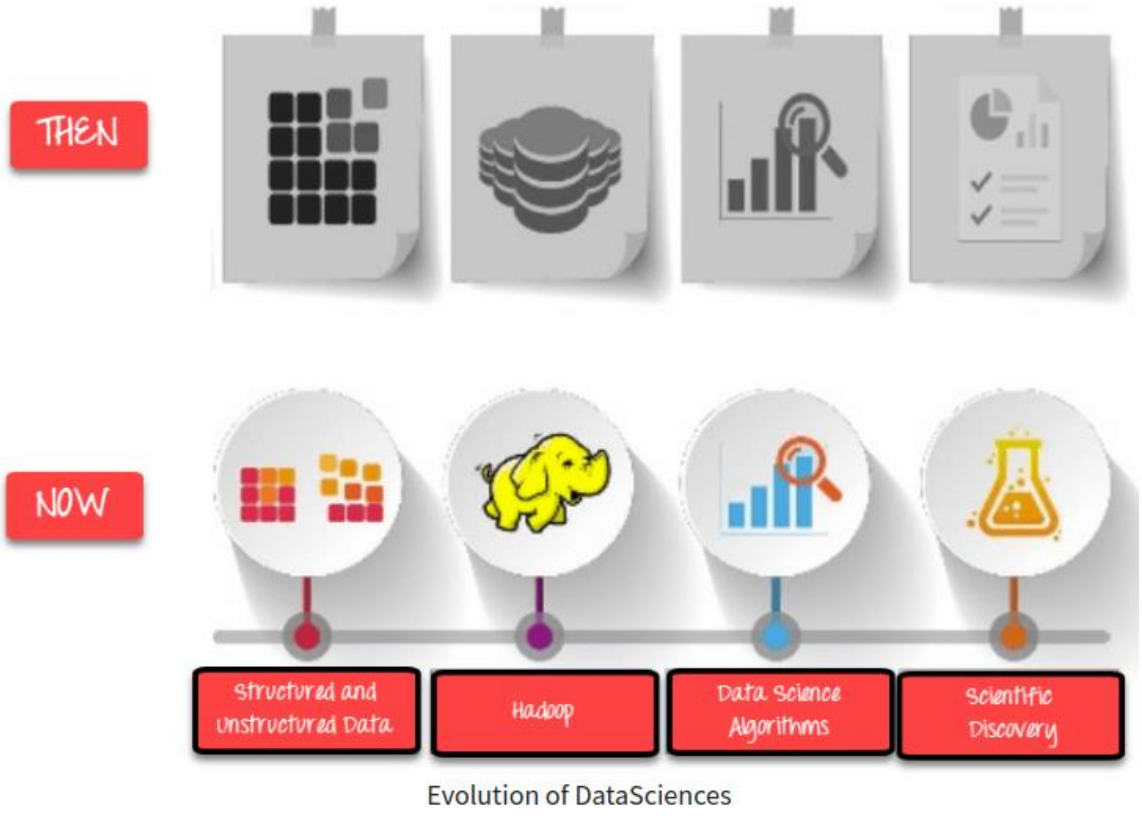
Example:

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.

Why Data Science?

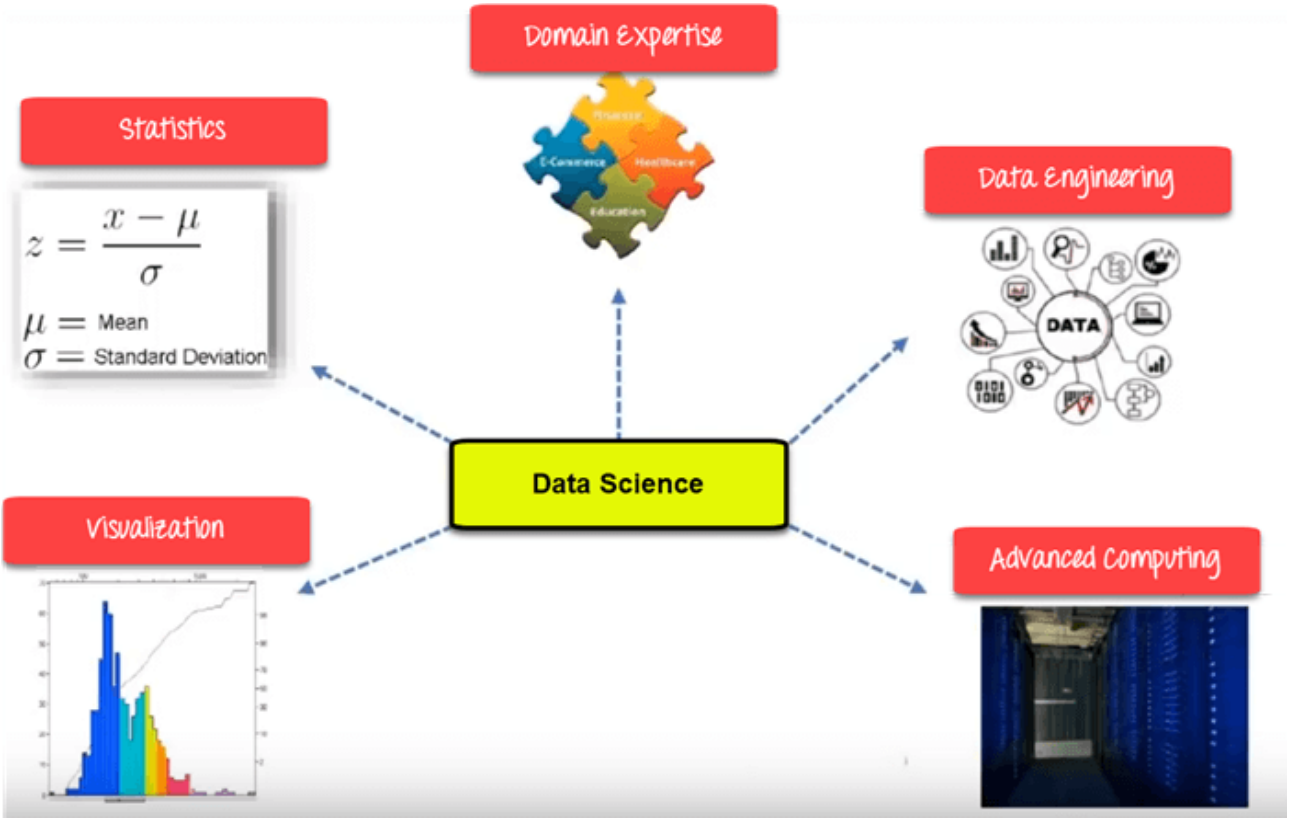
Here, are significant advantages of using Data Analytics Technology:

1. Data is the oil for today's world. With the right tools, technologies, algoithms, we can use data and convert it into a distinctive business advantage.
2. Data Science can help you to detect fraud using advanced machine learning algorithms
3. It helps you to prevent any significant monetary losses
4. Allows to build intelligence ability in machines
5. You can perform sentiment analysis to gauge customer brand loyalty
6. It enables you to take better and faster decisions
7. Helps you to recommend the right product to the right customer to enhance your business



Evolution of DataSciences

Data Science Components



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.

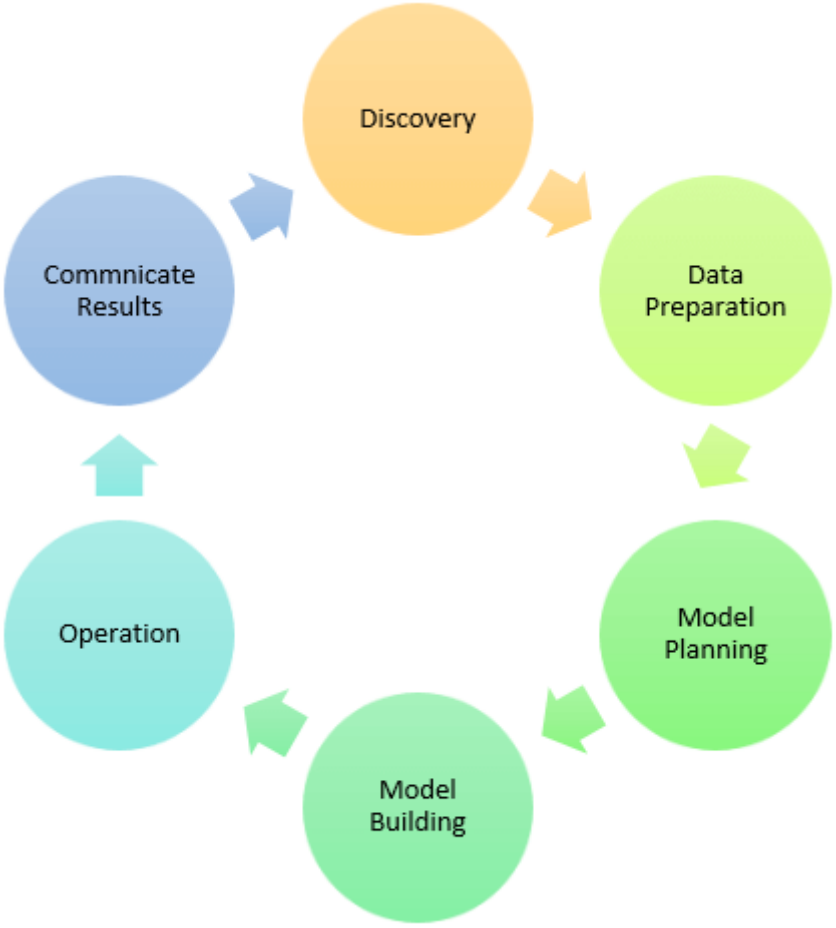
5. Advanced computing: Heavy lifting of data science is advanced computing. Advanced computing involves designing, writing, debugging, and maintaining the source code of computer programs.

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.

Data Science Process

Now in this Data Science Tutorial, we will learn the Data Science Process:



1. Discovery:

The first phase is discovery, which involves asking the right questions. Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.

The data can be:

- Logs from webservers
- Data gathered from social media

- Census datasets
- Data streamed from online sources using APIs

2. Preparation:

Data preparation is also known as Data Munging. Data can have lots of inconsistencies like missing value, blank columns, incorrect data format.

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 stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

4. Model Building:

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.

5. Operationalize:

In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

6. Communicate Results

In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

Types of Data Science Job

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 algoithms, 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 varius 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.

Prerequisite for Data Science

1. Non-Technical Prerequisite:

- Curiosity: To learn data science, one must have curiosities. When you have curiosity and ask various questions, then you can understand the business problem easily.
- Critical Thinking: It is also required for a data scientist so that you can find multiple new ways to solve the problem with efficiency.
- Communication skills: Communication skills are most important for a data scientist because after solving a business problem, you need to communicate it with the team.

2. Technical Prerequisite:

- Machine learning: To understand data science, one needs to understand the concept of machine learning. Data science uses machine learning algorithms to solve various problems.
- Mathematical modeling: Mathematical modeling is required to make fast mathematical calculations and predictions from the available data.

- Statistics: Basic understanding of statistics is required, such as mean, median, or standard deviation. It is needed to extract knowledge and obtain better results from the data.
- Computer programming: For data science, knowledge of at least one programming language is required. R, Python, Spark are some required computer programming languages for data science.
- Databases: The depth understanding of Databases such as SQL, is essential for data science to get the data and to work with data.

Applications of Data Science

1. Internet Search:
Google search use Data science technology to search a specific result within a fraction of a second

2. Recommendation Systems:
To create a recommendation system. Example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

3. Image & Speech Recognition:
Speech recognizes system like Siri, Google assistant, Alexa runs on the technique of Data science. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

4. Gaming world:
EA Sports, Sony, Nintendo, are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning technique. It can update itself when you move to higher levels.

5. Online Price Comparison:
PriceRunner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

Challenges of Data science Technology

- ❿ High variety of information & data is required for accurate analysis
- ❿ Not adequate data science talent pool available
- ❿ Management does not provide financial support for a data science team
- ❿ Unavailability of/difficult access to data
- ❿ Data Science results not effectively used by business decision makers
- ❿ Explaining data science to others is difficult
- ❿ Privacy issues
- ❿ Lack of significant domain expert
- ❿ If an organization is very small, they can't have a Data Science team

Summary

1. Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes.

2. Statistics, Visualization, Deep Learning, Machine Learning, are important Data Science concepts.

3. Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results.

4.Important Data Scientist job roles are: 1) Data Scientist 2) Data Engineer 3) Data Analyst 4) Statistician 5) Data Architect 6) Data Admin 7) Business Analyst 8) Data/Analytics Manager

5. R, SQL, Python, SaS, are essential Data science tools

6. The predictions of Business Intelligence is looking backward while for Data Science it is looking forward.

7. Important applications of Data science are 1) Internet Search 2) Recommendation Systems 3) Image & Speech Recognition 4) Gaming world 5) Online Price Comparison.

8. High variety of information & data is the biggest challenge of Data Science technology.