Introduction to Data Science

Data science always deals with real-world complexity and it is also a kind of emerging technology of this era. Nowadays unimaginable data are produced. What can be done with all this data? Proper analysis of all this data improves business and will also increase profit.

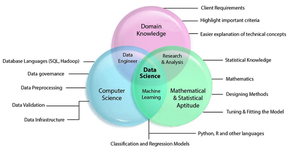
We all know that the internet of things makes each and everyone's life easier. An immeasurable amount of data is generated from different social media to the IoT devices.

Example: **Uber provides cab service according to the customer requirements.**

Nowadays most people are using Uber for transportation. Why uber is convenient and what makes them so rich in this era?

As we all know uber is a very successful company. it is a very successful company not only because of their availability of cabs or their service. Then what makes them so successful. The answer to this question is so simple "**Data make them very successful.**" that doesn't mean that data is not only enough to grow a business. But use data efficiently which means we should know how to utilize data to draw useful insights in order to solve problems and this is where the role of data science comes.

What do you mean by data science?



Data science is the process that is used to extract meaningful insights from data or find out the hidden patterns within a data or to identify the problems from the data which is used to improve the business.

- Data science is the process that extracts useful information from the data by using different scientific methods. **Scientific methods** are
  - Programming
  - Statistics
  - Business

- Data science will combine different fields such as artificial intelligence (AI), statistics, and data analysis in order to extract value from data.
- The data which is required to solve the problem are obtained from different data sources like the web, social media, etc.
- To solve complex problems which exist in the real world, useful insights are derived from data.

Why Data Science?

Data science is all about using data to create an impact for the company in order to grow its business. Patterns within the data find out using data science. Different types of statistical techniques are used in order to analyze and draw insights from the data.

Each and every company needs very skilled data scientists in order to analyze and process data. Their aim is not only to analyze data but also to improve the quality of data. Data scientists mainly solve real company problems using data. Many companies like apple, google, Facebook, and Amazon uses data science to improve their products and it helps the company to grow their business. These companies will appoint skilled data scientists who will analyze the data and uses those data to improve their business.

If we want to bring a change in a business, first we have to analyze the data which is generated. As we all know immeasurable data is generated each and every second. These generated data may contain use full insights, values, knowledge, etc. These useful insights, values, and patterns are extracted from the data using different techniques. Hence it helps the data scientist to create a change in the business and help them to take care of full decisions.

Need for data science

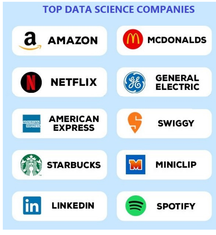
We need data science because of the big data explosion . for example different social media is producing a big amount of data each second. Millions and millions of emails are sent, people are watching and viewing more than 20 million videos, photo views are more than that, millions of search queries, and many downloads these all are happening at the same time over the internet.

All these data are useful but we should make use of these data efficiently. We can't even imagine the size of data and here is the need for data science. There will be a lot of useful insights and these insights are extracted from the data using data science techniques by the data scientists.

Many organizations are using data science to improve their business.

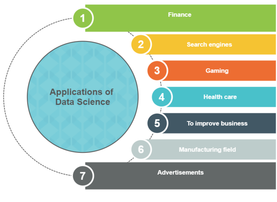
For example, Google and Netflix

1. **Google:** Google is a search engine that is used by people to get appropriate answers to their queries . Here data science algorithms are used to get better results within seconds.
2. **Netflix:** Most of us are using Netflix to watch movies, shows, sports, series, etc. We have seen Netflix producing a recommended list of movies for us. How is it generating the recommended list? The answer to this question is very simple, it is with the help of data science



Application of data science

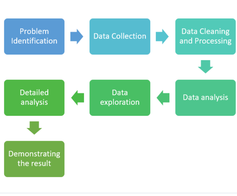
There are plenty of applications for data science. Some important applications of data science are given below.



1. **Risks are detected by the banking companies using data science.**
   Bad debts and losses create a lot of problems within the banking companies. Some customers may not able to repay their loans on time due to their unstable financial stability. This will create a big loss for the company. The solution to the problem is applying data science. Before sanctioning loans their personal details, their income details everything is collected, using these data skilled data scientists can dig out the useful insights which help the company to prevent losses.
2. **Use of data science by the online search engines.**
   There are many online search engines like yahoo, google, AOL, etc. These search engines provide answers to our queries within seconds. Data science algorithms are used by search engines in order to give results in seconds.
3. **Gaming**
   In most online motion games, we are competing with the computer. We may always wonder how we lose and computer wins all the time. It is because in computer games they are applying data science where our opponent that means the computer always knows our previous moves. a skilled data scientists always analyze that and design the game according to that.
4. **Data science is used in medical fields.**
   Genetic issues and reactions to particular drugs can be identified. It can be identified using data science techniques. If a person's personal genome data is obtained a detailed study on human DNA can be done which helps to predict the risks and individual attention can be given.it also helps in medical image analysis.
5. **Advertisements**
   Targeted advertisements are given for particular users. It is completely based on their past behavior. A detailed study is done on each and every user's past behavior.
6. **To improve business data science is used.**
   Different types of business are there in our society. So data science can be used to identify the business patterns, to analyze and to extract useful information's which will help to improve business.
7. **Manufacturing field**
   For monitoring the systems, detecting anomalies, for predicting the problems data science can be used.

Steps involved in the data science process

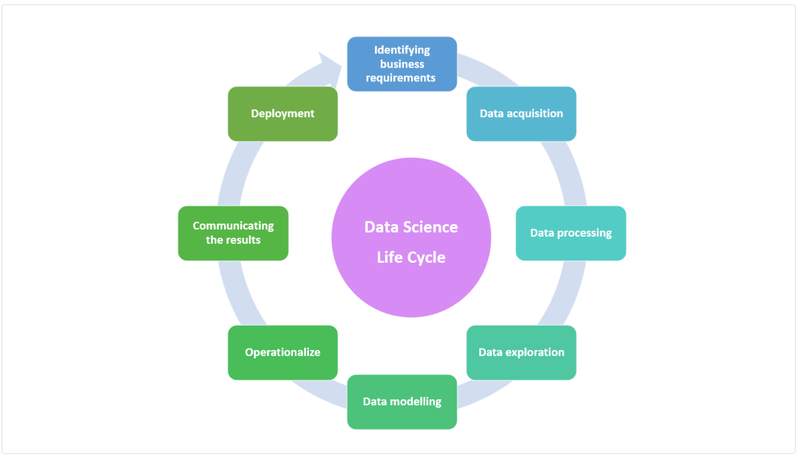
- Step 1: Identifying the problem
- Step 2: Raw data collection for the desired problem
- Step 3: Data cleaning and processing
- Step 4: Data analysis
- Step 5:data exploration
- Step 6: Detailed analysis
- Step 7: Demonstrating the result



Process of Data Science Life Cycle

In this tutorial, we are going to discuss the entire life cycle of data science. It's all about how to execute the data or the assigned project.

The Data science life cycle is a kind of framework that provides some information or steps about how to develop a data science project. It mainly contains some steps that should be followed by the data scientist when they begin a project and continue until the end of the project. These steps are not fixed and they can be bypassed or can be repeated until we get a corrector satisfied results. According to the project sometimes it may take a few months to complete because this is a lengthy procedure.



1. Identifying business requirements
2. Data acquisition
3. Data processing
4. Data exploration
5. Data modelling
6. Operationalize
7. Communicating the results
8. Deployment

1. Identifying business requirements

This is the very first step in the data science life cycle. Gathering all the information from the available data sources Identifying the problem and finding out the objectives are the main two things done in this step.

Before starting a project first we should have a clear idea about what are the project requirements, what is the need of the clients, and market trends. Everything should be identified clearly to get better results.

How do identify the business requirements?

- In this step, the data scientist will conduct a meeting with the client.
- Asks relevant questions to identify the problem.
- Understanding the problem.
- Note down the objectives of the problem.

2. Data acquisition

Collecting data from multiple sources means, here the data scientists gather information from different sources like databases, APIS, webpage, online repositories, etc. Needed information should be collected from the available sources. In order to read data from specific sources, some special packages like R or Python are available. Many precious data are also gathered from social media like youtube, Twitter, Facebook, etc.

Gathering data from files is known as the conventional way of data gathering. The main 5 methods to collect information are by conducting surveys as well as questionnaires, the 2nd method is by conducting interviews where data is collected directly from the respondents by asking questions to them, the 3rd one is data collection from group discussions, 4th method is direct observation and final method is gathering information from documents. These are the 5 main methods used to gather information.

3. Data processing

Cleaning of data and transformation of data is done in this step.

The obtained data will not be clean so before moving to the next step it is needed to process the data which means raw data which is obtained from different sources should be cleaned. The cleaning of raw data is done by scrubbing and filtering. While cleaning the data if we are noticing any missing data sets proper replacement should be done. This means replacing and withdrawing values are also done while cleaning the raw data.

- **Data cleaning**: Data cleaning is done because sometimes the raw data which is obtained from multiple sources may contain missing values, duplicate values, misspelled attributes, and so on.
- **Data transformation**: Data transformation is the process of converting or transforming data into the desired format.

4. Data exploration

Understand different patterns from the data which is cleaned and useful insights are retrieved from that. Before using the data it should be examined. Why this is done because data may contain different types of data like numerical data, ordinal data, nominal data, as well as categorical data. Different data types should be handled in a different manner so only proper examination of data will help to identify different data types.

In order to understand different patterns data scientists can use histograms, Microsoft Excel spreadsheets, etc.

5. Data modeling

In data modeling, a model is created which predicts the target most accurately and the model which is created is evaluated and tested in order to check the efficiency. Various tools used in model planning are R, python MATLAB, and SAS.

Mainly there are two types of data models and they are conceptual data models and physical data models. The concepts of database and the existing relationship between them are represented visually in the conceptual data model and this model contains entities/subtypes, attributes, relationships, and integrity rules. After the conceptual data modeling, the next step in action is the physical data model. That means each data entity's attributes are clearly defined here. The physical data model consists of tables, columns, keys, and triggers. The data modeling contains many diagrams, symbols or text these are mainly to represent the data based on how it interrelates.

6. Operationalize

The model which is chosen in the above step is checked in this stage which means it will run the model to understand how it works. After running the model findings based on its performance are documented and final reports are submitted. The final report consists of all the necessary briefings, code details, performance details, and technical information. If their findings are perfect then it is ready to use and the team will also deliver the final reports, code as well as technical documents in this stage.

7. Communicating the results

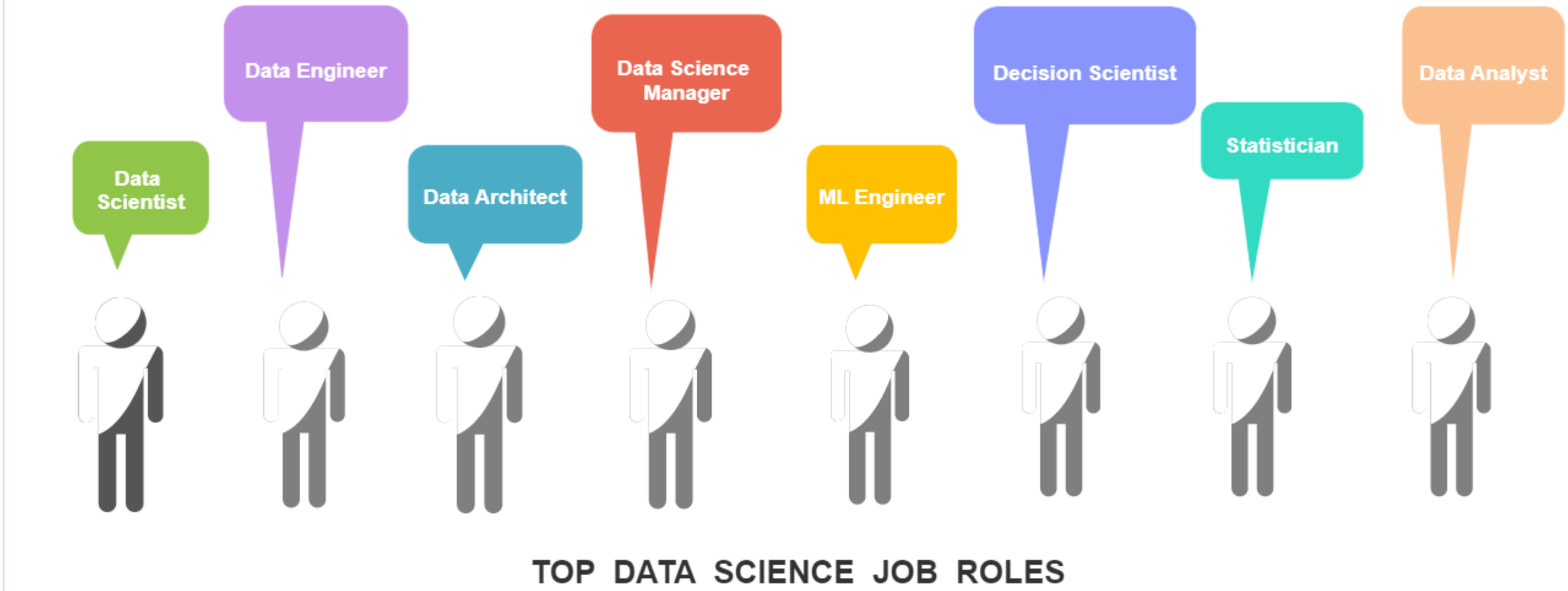
In this stage, the project results and valuable findings can be communicated to the responsible people. Here they can compare it with the initial stage where we identified the problem with our findings, in order to check whether useful insights are given by the data.

After the discussion, if the quality of the result needs an improvement or any fault is detected then again we have to start from the very first step. If no mistakes are found and the higher authority is happy with the data scientists and his team's work then we can move to the last step.

8. Deployment

Deployment is the final step of a data lifecycle. The main goal of this stage is to deploy the model into a production environment. Validating the performance of the models is done by the users. Monitoring the performance of the model is done very carefully and if any changes are needed that is also done in this stage. Continuous monitoring will be done because data trends are changing day by day so according to that we should make some changes to the model mainly to adjust to new evolving trends and to avoid performance regression.

Top Data science Jobs Roles



1 **Data Scientist**

Data scientists mainly find out hidden patterns and useful insights from the data. They mainly analyze the raw data and handle them very well. In order to extract useful insights, they use various statistical procedures. Both structured and unstructured information are handled very well by skilled data scientists.

2 **Data engineer**

Data scientists need big data pipelines and models to work upon. Data engineers are the ones who will help data scientists to build data pipelines and models. Managing, maintaining, and testing data models are also done by a data engineer.

Essential requirements for a data engineer 1. Knowledge of database models 2. Knowledge of ETL

3 **Data architect**

They will organize and manage both macro and microdata. The blueprints of a company's data platform are implemented by a data architect.

Tools used by a data architect XML SQL Hive Spark Pig

4 **Data science manager**

Data science projects are handled and managed by a data science manager. They are responsible for executing all of their plans and submitting the outcome before the deadline. In order to guide the teammates, a data science manager should have good communication skills and very good leadership qualities.

5 **Machine learning engineer(ML engineer)**

In order to extract meaningful information from the large amount of data that is given, the Ml engineers will write programs and develop Algorithms. They should be very familiar with machine learning algorithms.

6 **Decision scientists**

With the help of artificial intelligence and machine learning, a decision scientist helps the company to make appropriate business decisions.

7 **Statistician**

They will identify various trends in the market using a statistical model. Tools used by a statistician R, python, SQL SAS ,SPSS Matlab Statsa

8 **Data Analyst**

They are the people who collect as well as interpret data to solve a particular problem.

Data Scientist Job Description: Role, Responsibilities,Salary and Skills Required

In this tutorial, we are going to discuss who are data scientists, how to become a successful data scientist, their job roles, their responsibilities, and the required skills for them. Analysis of data, pre-processing of data, collection of data, and taking useful insights from the raw data all these things are done by a data scientist to improve the business of a particular sector.

Who are data scientists?

Data scientists are the people who practice data science. They will use their skills for analyzing data properly which is collected from the web, smartphones, social media, and so on. Combining statistics, computer science, and mathematics is the main role of a data scientist. They work on both structured and unstructured data, in order to get useful insights needed to improve the overall business of a company.

A data scientist should not only have technical knowledge but also should be able to communicate his ideas with each and everyone right from the customers to the higher authorities. So it is very necessary for a data scientist to become an effective communicator as well as a good leader.

How will data scientists get useful insights from data?

Data exploration

- When a data scientist gets a challenging question or a challenging situation then each and every data scientists act like a detective who will investigate and they will try to understand different patterns as well as different characteristics of the data.
- They utilize the data for the betterment of an organization's business.

Skills needed for a data scientists

- **Deep knowledge in mathematics**: to find a solution for a particular problem they have to build many predictive models and these predictive models are based on very hard maths.
- **Proper utilization of technology**: in order to work with complex algorithms data scientists should analyze the data properly and utilize the technology.
- **Be an expert in coding languages**: They should be efficient with the coding languages and with some core languages such as SQL, Python, R, and sass.
- **Proper understanding of a business**: suppose a data set of a particular business is given to a data scientist they should study and analyze it properly to find out where the things are going wrong and to find out a proper solution.

Responsibilities of a data scientist

- **Solving business problems**: one of the main responsibilities of a data scientist is to identify the business problem faced by a particular sector and take appropriate measures in order to solve and improve the business.
- **Discuss the findings and predictions**: using effective reports and effective visualizations of data the data scientists should communicate all the findings as well as the predictions to the management and the IT department.
- **Gathering useful information**: As we all know unimaginable data is generated each and every second from different sources. So data scientists should gather useful insights from different sources and use them for the improvement of the business.
- **Avoiding repetitive work**: To avoid repetitive work they should always find out new and appropriate algorithms in order to solve the existing problems.
- **Identify the trends and patterns**: from the data they should be able to find out the trends and patterns which will help them to improve their business activities.
- **Building appropriate models**: data scientists are responsible for building appropriate models mainly to address business problems. All the business problems need to be solved by the company to improve the business. In order to solve the business problems nowadays, each company is appointing data scientists and they will build appropriate models to address the problems which are faced by a particular sector.

Qualifications required for a data scientists

The minimum qualification for a data scientist is a bachelor's degree in data science or in computer science or any other course which is close to a field that is related to computer science. But nowadays most companies require a master's degree to get a role of a data scientist.

Programming knowledge, deep math skills, analyzing risks, statistical analysis, communication and software engineering skills, knowledge of using could tools, and visualizing the data all are very basic skills required for becoming a data scientist.

How a data scientist uses data science to improve the business?

Let us see how the data scientists appointed by Walmart use data science to improve their business.

Walmart Use Case

- Who is Walmart?
  - It is the world's biggest retailer and has more than 20,000 stores in 28 countries.
- The reason behind Walmart's success.
  - The data analysts and data scientists present at Walmart will conduct a detailed study about the customers in order to know every detail about their customers.

So they know for example if a customer buys pop tarts they might also buy cookies.

The data scientists will use the data which they get from customers and proper analysis is done in order to know what a particular customer is looking for.

Case 1 : (case study): Halloween and cookies sales

During the Halloween festival, data scientists at Walmart analyzed the customers .so they came to know that during the Halloween festival a specific type of cookie is very popular in all Walmart stores which means they found there is a connection between Halloween and sales of cookies. They also identified 2 stores at Walmart that were not at all selling the cookies due to simple stocking oversight. The data scientists identified the situation and solved the problem and which helps to grow the business.

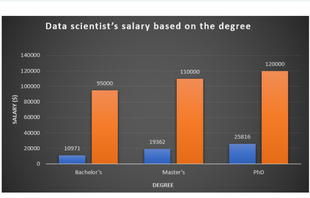
How did data scientists come to know that a particular cookie is very popular during the Halloween festival?

1. Analyzing the data and obtaining useful insights from the data to grow business
2. In order to grow business, they invest a lot of money, effort, and time.
3. To find the hidden patterns from the data they invest a lot of time.
4. When they find out the hidden patterns or relation between two products they will start giving offers or some sorts of discounts only to attract customers.
5. So Walmart uses data in a very effective manner and proper analysis of data and processing of data is done very well and useful insights are found to improve the business and to attract the business

Data scientists salary

The salary of data scientists depends upon many factors.

1. The salary package of a data scientists in India and the USA



In India and USA, there is a correlation between the salary provided for data scientists based on their degrees. When the degree is high then the salary is also high.

Ph.D. holders have the highest salary both in India as well as in the USA.

2. Salary based on experience

More experience in the relevant field will provide more salary both in USA and India.

3. Salary based on location

Salaries will be very high in big cities both in India and USA.

4. Salary based on companies

Different companies provide different salary packages. So many companies are looking for skilled data scientists. Among them, some companies provide high salaries for data scientists.

Data Scientists Vs Data Analyst

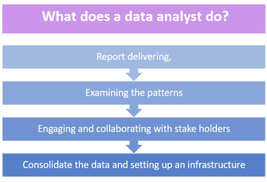
In this tutorial, we are going to discuss what are the main differences as well as similarities between data scientists and data analysts. We will also look into what a data scientist and data analyst do, their education and work experience, what are the required skills for both data analysts and data scientists, and a detailed comparison of their skills, work experience, salary, and education qualification.

Data scientists are the experts who will use their technical skills in order to solve problems that are very complex. They will also use algorithms, technical skills, and scientific methods in order to extract useful information or insights from both structured as well as unstructured data. Data scientists use data very carefully in order to make better and more suitable business decisions. Where a data analyst mainly deals with cleaning data, transforming data, generating inferences, and solving problems that means identifying the problem is not done by a data analyst they will only solve the existing problem. but a data scientist will first identify the problem then they will solve the problem. An important skill required for a data scientist is communication skills and it is not necessary for a data analyst.

What does a data analyst do?

A data analyst will collect data from different sources and organize the collected data. After organizing the data they will do the analysis.

1. **Define the problem**: They mainly determine the needs of a customer and generates plan according to the need of a customer. Finally, they communicate the plan with their team.
2. **Data collection**: data are collected from multiple sources such as database backups, flat files, and APIs. Data analysts work with the programmers in order to create the ETL process. The final step done in data collection is data aggregation.
3. **Data cleaning**: Raw data is always messy so in order to use data cleaning is very necessary. Cleaning data make it more useable. Normalization and standardisation are done on the data after data cleaning. Finally, data validation is done.
4. Report delivering, examining the patterns, engaging and collaborating with stakeholders, finally they will consolidate the data and infrastructure is made out that these are the main roles done by a data analyst.



How data analysts do collaborate with stakeholders?

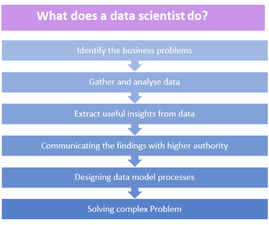
One of the important roles, as well as responsibilities of a data analyst, mainly includes collaborating and engaging with salespeople and marketers who are there within the organization. This means data analysts should engage as well as collaborate with the different departments of a company. A data analyst should cooperate with peers who work in other departments.

What do data scientists do?

The data model process of designing, and creating suitable algorithms and predictive models is mainly done by a data scientist. Hence each and every data scientists spend more time designing tools, finding out useful insights from the data, and identifying the business problems and data frameworks.

When we compare a data scientist with a data analyst they will always find out new tools or will develop new tools as well as methods in order to solve complex problems faced by a particular company in their business sector. They will always identify and solve business problems by extracting useful insights from the data with the help of predictive models. Identifying the problem is the very first thing done by a data scientist and after that gathering and data analysis is done in order to extract useful insights from the data.

Hence data scientists mainly identify the business problems, gather and analyze data, extract useful insights from data, communicate the findings with higher authority as well as with their colleagues, and design data model processes. They will use their technical knowledge, mathematical skills, and coding skills in order to solve as well as extract useful insights from data.



What are the differences as well as similarities between data analysts and data scientists?

- Programming is done by both data scientist and data analyst but a data analyst will only do basic programming and a data scientist will always do advanced programming.
- When data scientists do predictive modelling which means predictions are made about the future events which are unknown, the data analyst will do static modelling. Static modelling is mainly used to specify the object's structure which is there within the problem domain.
- A data scientist should be familiar with database systems for example MSQL, HIVE, python, java etc and a data analyst should be very familiar with the concepts like data warehousing and business intelligence and they should also have a strong base of Hadoop based analytics such as HBase, Hive, MapReduce jobs, cascading etc.
- A data scientist should have strong fundamentals of maths whereas a data analyst should have strong statistical skills.
- Programming is used by both data scientists and data analysts mainly for cleaning, transforming and analysing data.
- If someone wants to start their career in analytics then the data analyst role is more suitable for them and if someone wants to ease the tasks of humans by creating advanced machine learning models and deep learning techniques then data scientist is the better option for them.
- Based on the past patterns data scientists can predict the future of the business whereas data analysts cant do that. For both at least a bachelor's degree in the quantitative field (mathematics, computer science or statistics) is necessary.
- Data visualization, statistics, math, data mining and data warehousing are the common skills used by both data scientists as well as a data analysts.
- Both are working with the data sets. But the only difference is both are using different tools and using different skills in order to solve the business problems.

Educational requirements needed: data scientists and data analysts.

Educational requirements which are needed for a person to become a data analyst is a bachelor's degree mainly in the fields like statistics, mathematics, computer science or finance.

If one has a master's degree in data science, information technology, mathematics or statistics they can surely work as a data scientist.

Apart from this if one can earn a professional certificate from IBM or Google in data analytics and if you are skilled enough to acquire all the basic needs required for a data analyst you will be able to work as a data analyst for certain companies. A data analyst can become a data scientist if he/ she is ready to study more or ready to gain the required skills which are needed to become a data scientist.

Skills needed for data scientists as well as for data analysts

Both should have great knowledge of mathematics. The data scientists should be experts in advanced statistics and predictive analytics whereas a data analyst should have great knowledge in foundational mathematics and statistics. Software and tools used by a data analyst are SAS, Excel, and business intelligence software. Hadoop, MySQL, TensorFlow, and Spark are the tools used by data scientists. Analytical thinking and data visualizations are the other skills required for a data analyst. Machine learning, as well as data modelling, are the other skills required for a data scientist. Programming skills required for a data analyst are basic fluency in R, Python, and SQL. Advanced object-oriented programming is the main programming skill required for data scientists.

Responsibilities of a data scientist and data analyst

Data Scientist	Data Analyst
• To discover new opportunities they will use current data. • Machine learning models, as well as analytical methods, are developed. • Detailed data cleaning is done by a data scientist • A/B testing is conducted	• In order to solve a problem, they will use pre-existing data. • Create reports and dashboards. • Very basic data cleaning is done by a data analyst. • A data analyst will help to collect incremental data from new sources.

Qualification and skills

Data Scientist	Data Analyst

• Masters degree or higher is required
• For some positions PhD is must
• Degrees mainly in computer science, statistics, mathematics, economics, physics, and machine learning is required.
• Skills required: Great knowledge in
  • SQL
  • R/Phython, pandas,Numpy,tensorflow
  • NLP
  • Apache spark
  • SAS/SPSS
	• Bachelor's degree or higher is preferred
	• For some positions master's required
	• Degrees mainly in computer science, statistics, mathematics, economics, and physics is required.
	• Skills required- Great knowledge in
	  • SQL
	  • R/Python
	  • Data modelling
	  • ExcelAWS/Azure

Business analyst Vs Data engineer

As we all know data science has a great demand in the coming years. Data science is a very vast topic that also consists of many small and very well-defined topics. There are many job roles that come under data science. In this module let us mainly discuss the role of a business analyst as well as a data engineer.

Who is a business analyst?

A business analyst will first identify the area of a business that needs improvement in order to strengthen the business processes. They will also cooperate with their teammates by sharing their ideas as well as their findings. The business of a particular company is improved by a business analyst by analyzing the business processes, and services that are provided by the company and also by analyzing data and products to provide better services. They also opt for a disciplined approach while introducing or when they try to make changes in the business process for better results.

Business trends will change over time so these changes are continuously analyzed by a business analyst. They mainly focus on the goal of a company by identifying the past, as well as current business and proper steps, that are taken to achieve the business goals.

Role of a business analyst

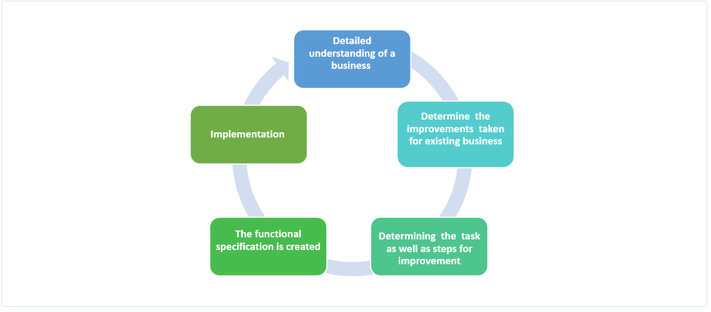
1. **Detailed understanding of a business**
   Understanding the business is an important role of a business analyst. A business analyst will first understand the business of a particular company with the past as well as current business. By comparing and analysing the past and current business they will get a clear idea about how does the business works.

2. **Determine the improvements taken for existing business**
   As we all know business trends are changing day by day so it is very necessary to analyse the changes and take further steps to improve the business. The business analyst should clearly determine what all improvements can be done to the existing business process to achieve the business goals.

3. **Determining the task as well as steps for improvement**
   If the business analyst got a clear idea about the business and how to improve the business then the next role is to determine the task and steps for the improvements. Here they will determine what are their requirements in order to improve the existing business process.

4. **The functional specification is created**
   The business analyst will mainly design the features which is required for the IT system to improve the existing business process. To make changes they will create functional specification and it is also one of the main role of a business analyst.

5. **Implementation**
   Business analyst will also determine how to implement all the findings in order to improve the existing business. They will also determine the technical design which is needed for a business to improve and for goal achievement.



Who is a data engineer?

Basically, a data engineer is an IT worker who will prepare the data for analytical as well as operational uses. In order to gather information from different sources, data engineers will be building data pipelines. Data engineers always work with data scientists as well as with business analysts. The raw data is converted into a useable format and given to the data scientist by a data engineer. The data which is gathered from different sources are given to the business analyst by a data engineer for proper analysis.

A data engineer mainly builds data pipelines in order to gather data from different sources. They will help a data scientist in integrating, consolidating, and cleaning data that is gathered from different sources.

A data engineer deals with both structured as well unstructured data. Structured data consists of information that is well organized. Where unstructured data consist of text, images, audio, and video files. The main duties of a data engineer are developing software, and building as well as maintaining data pipelines. They will also maintain databases. Hadoop, NoSQL databases, and Spark are the main tools used by a data engineer.

Role of a data engineer

1. **Act as generalists**
   Data engineers are the ones who works with data analyst, data scientist as well as with small teams in order to collect data . After data collection they will also helps to process the data. Even though they have many skills their knowledge in system architecture is very poor. A data engineer is always a helping hand for a data scientist as well as for data analyst.

2. **Act as pipeline centric engineers**
   Here the data engineers mainly works with data analyst team who works with mid size data. They also help the data analyst with more complicated data science projects. These kind of data science engineers are mainly seen in mid size as well as with in the large companies.

3. **Act as database-centric engineers**
   The data engineers who act as database-centric engineers will implement the analytics database, maintain the analytics database as well as they will also populate the analytics database. This kind of data engineer is seen in large companies where their data has been distributed across several databases.

Responsibilities of a business analyst and data engineer

Business analyst	Data engineer
• Analysing future business requirements. • Analyzing various business possibilities • Tracking the requirements • Detailing of project • Team building • Documentation	• They will build data pipelines as well as warehouses • Scalability of data products are managed. • Develops, constructs and Maintains the architecture • Process raw data • Monitor and maintain systems • Prepare data for analysing

Qualification Need

Business analyst	Data engineer
• Bachelor's degree in IT/CS • Bachelor's or masters in IT/CS • Practical experience with Microsoft	• Bachelor's degree in computer science, math, statistics, and information systems. • Bachelor's degree in business/ IT • Master's degree • 2 to 5 years of experience

Skills Needed

Business analyst	Data engineer
• Communication • Problem-solving • Critical thinking • Analyzing techniques • Data modeling techniques • Decision-making abilities	• Knowledge in Databases • Communication skills • Programming • Knowledge in Bigdata and cloud

Tech Skills

Business analyst	Data engineer
• R • SAS • SPSS • STATA • Data mining	• SQL • Python • Cloud • Distributed computing

Salary

Business analyst	Data engineer
India Rs 7,60,000 per year(appox)	India Rs 9,50,000 per year(appox)

How do we solve a problem in data science?

Data science is the process that extracts useful information from the data which is obtained from different sources by using different scientific methods. As we all know there are many complex problems that exist in the real world. In data science, the useful insights which are derived from the data that are collected from different sources are used in order to solve the business problems that exist within a company. In this module let us discuss how to solve a problem in data science in detail.

First principle thinking approach in solving data science problems.

The first principle of thinking is an approach where new solutions are created by breaking down the problems after identifying the assumptions. Innovative solutions can be made using the first principle thinking approach.

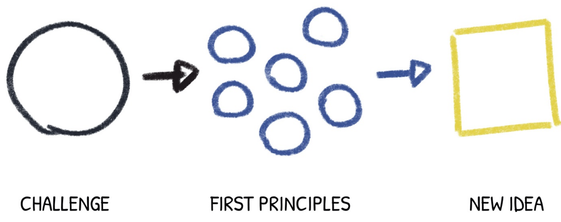
Suppose if a company is facing a problem then the very first step taken will be complex problem identification and it is broken down into smaller parts. Break down process continues until you can't break down any further. Finally, innovative solutions can be made which will help to solve the data science problem.

Traditional approach vs first principle approach

The traditional approach as well as the first principle approach is used to solve data science problems. Studies say that the first principle approach is the most appropriate and efficient method in order to solve data science problems. The traditional approach is also known as the Analog approach.

The traditional approach or Analog approach always begins with the existing ideas and some improvements are done to the options which are available. Finally, the best option is chosen to solve the problem. The main problem faced by a traditional approach is it won't solve the core problem.

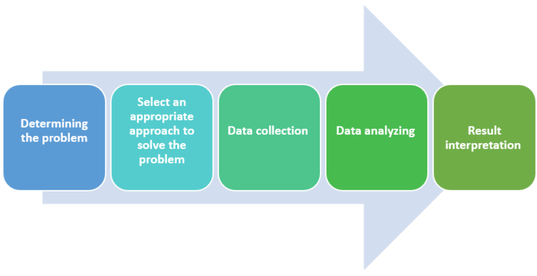
The first principle approach always identifies the assumptions and it will break down the problems into smaller components in such a way that they can't be divided any further. Finally, a new solution is created to solve the data science problem. In the first principle approach, most of the time is spent on identifying as well as understanding the problem because once the problem is identified clearly then the proper solution can be generated.



Traditional Approach	First principle Approach
• Begin with the ideas which are already existing. • Available options will be improved • The Best option will be selected.	• Start with identifying the assumptions. • Problems are divided into small components. • The New solution is created.

Steps were taken by the data scientists to approach a data science problem

Several steps are taken by a data scientist in order to solve a problem in data science.



1. **Determining the problem**

   Determining the problem is the very first step to solving a problem in data science. Problems should be defined properly in order to solve the problem. If the problems are not clear or if it is not defined properly it will be very difficult for each and every data scientist when they work it to find the solutions. So the identified problems should be defined clearly and properly.

2. **Select an appropriate approach to solve the problem**

   Mainly data scientist uses two types of approach

   1. Traditional approach(if needed link can be given to the above section)
   2. First principle approach

   Among these two approaches most commonly used approach is first principle approach. It is because the first principle approach always start with identifying the assumptions and the identified problems are divided into small components . Finally new solutions are created.

   Many data science algorithms are used in order to solve a problem in data science. linear regression, logistic regression, decision trees, naïve bayes, KNN, support vector machines , k mean clustering, PCA are some of the common data science algorithms mainly used to solve problems.

3. **Data Collection**

   When a data scientist identifies a problem they will define the problem properly and clearly then suitable approach is determined. After that the next thing is data collection. The data which is collected should be maintained properly along with the dates on which the data is collected.

   The collected data should be analysed properly and cleaning should be done . Data cleaning is a time taking process. Each and every data scientists spend more time for cleaning data. Cleaning data consist of removing the missing values, duplicate records identification and making some corrections if needed.

4. **Data Analysing**

   After data collection and data cleaning the next step is data analysing. In order to analyse the collected data from different sources so many data science libraries are available. If in this stage the selected data science approach is not working then suitable and appropriate approach is again selected.

5. **Result Interpretation**

   Once data analysing is done properly then the next step is interpreting the result. In this step the results are interpreted. The main four steps in result interpretation are assembling all the information properly, generate all the findings, conclusions are developed and finally all the recommendations are also developed.

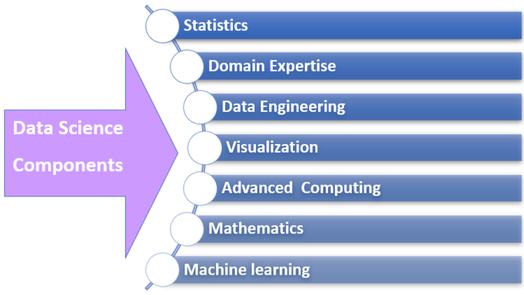
Some real-world problems with data science and how to solve them

Netflix: It is a subscription-based online platform that is used to watch movies, tv shows, and series with a strong internet connection. So Netflix uses data science in order to solve the problems. Netflix mainly uses collaborating filtering algorithm is mainly used for recommending movies to the Netflix users. The recommendation of movies is based on the movies that the users watched previously. Not only Netflix but also many other social media like YouTube, Hotstar, Facebook, etc use the same method to satisfy customer needs.

Uber: They use the user data which is available to them mainly to improve their customer services. Enter into the uber app and just one click will make the cab reach your destination where you are standing. All this is happing very smoothly because of the hard work which is done by each and every data scientist who works behind that particular company. This application interacts with the customers on one hand and with the drivers on the other hand. Here the data scientists mainly use deep learning, AI, and many other mechanisms to run the business smoothly and efficiently.

Data science components

Demand for data science jobs is increasing day by day. Most companies in the world are looking for employees who have a piece of great knowledge of data science. In this tutorial, let us discuss each and every component of data science.



Data science mainly consists of 7 components. Statistics, domain expertise, data engineering, visualization, advanced computing, mathematics, and machine learning are the main data science components.

1. **Statistics**

Statistics is the main and most important component of data science. As we all know data science is the process of taking useful insights from the data which are collected from different sources. These useful insights are used by data scientists to improve the business of a particular company.

In order to explore data statistical features are used by the data scientists. Statistical features mainly include data organizing which is mainly done to find out the minimum as well as maximum values, it is also used to find out the mean, mode, median values, and many more.

Some data consists of numerical data, collecting and analyzing numerical data is also very important. Statistics is a kind of tool or a way that is mainly used by each and every data scientist in order to collect as well as analyze a large amount of numerical data. Once the collection and analysis of numerical data are done then the useful insights are extracted from the data. Computer algorithms as well as some statistical formulas are used by the data scientists in order to dig out the useful insights from the raw data that is collected from multiple sources.

2. **Domain Expertise**

Domain Expertise is one of the other important components of data science. Domain expertise is a component that will mainly help for binding the data science together.

Domain expertise is defined as the deep core knowledge in a particular field or in a particular area. It will also play a very important role in decision-making and always binds the data together. Domain experts are very must required in various areas for improvement as well as take appropriate decisions in data science.

A domain expert will always help to identify the best data from the available sources and they will also be able to analyze how good a data is for use. All these are done very easily by them because of their deep knowledge as well as their experience in a particular field.

3. **Data engineering**

Data engineering is the process of acquiring, storing, retrieving, and data transforming. It includes metadata which means data about data.

Data is increasing day by day from different sources. A vast amount of data is produced each and every second whereas data engineering mainly deals with a large amount of data with the tools which is developed on their own. The main aim is to provide software solutions for the problems which are related to data.

The solution for the problems which are related to data is generated simply by creating a data pipeline and endpoints that are done within the system itself. Proper understanding of data technologies and frameworks are the main requirements which are needed for data science engineering. These are combined and used in order to create proper solutions which will surely enable the business processes.

4. **Visualization**

Visualization is the process of representing data in a visual context. While we are representing the data in a visual context it will make or help the people to understand the significance of data very clearly.

As we all know visualization is the process of representing data where the data representation is done using common graphs such as plots, charts, animations, etc. When data scientists use these types of graphs for visualization it will make the common people understand the complex data and their relationships very easily.

5. **Advanced Computing**

Advanced computing is nothing but an extended version of data science. It is the technology that mainly deals with designing as well as developing computing hardware and software. Advanced computing also defines a PC which is high end and different types of skills which is used on the PCs. Word processing, graphics as well as multimedia, spreadsheets, databases, computers, etc are the skills of advanced computing.

6. **Mathematics**

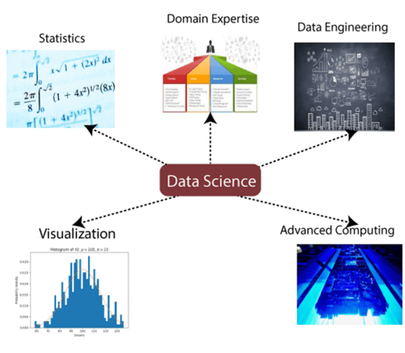
In order to find a solution for a particular problem data scientists who works under a particular company have to build many predictive models and these predictive models are based on very hard maths. So a core knowledge of mathematics is very necessary for a data scientist.

Data scientists, data analysts, and many more employees who work for the company will use their technical knowledge, mathematical skills, and coding skills in order to solve as well as to extract useful insights from data that is obtained from multiple sources. To become a good data scientist should have great knowledge of mathematics. The study of quantity, the structure, and the occurring changes in business are the main things involved in mathematics.

7. **Machine learning**

Machine learning is another component of data science. Do you know which is the backbone of data science? The answer to this question is very clear, the backbone of data science is machine learning. So what do you mean by machine learning? Machine learning is nothing but it is a process of providing training to the computers in order to make them act as human brains.

To solve business problems various machine learning algorithms are used in data science. Regression and supervised clustering are some of the techniques which are used in machine learning to solve problems. In order to identify the business trends and patterns, machine learning algorithms are used. For predicting the qualities machine learning plays an important role. Some important algorithms which are used in machine learning are linear regression algorithm, k means clustering, decision tree, etc



Statistics, domain expertise, data engineering, advanced computing, visualization, mathematics, and machine learning are the important components of data science. All these components are equally important in data science in order to improve the business by extracting useful insights and patters as well trends from the collected data.

Tools for data science

Data science is all about extracting useful insights from the data which are collected from different sources. These useful insights are extracted by a data scientist using different statistical tools and by using some programming languages.

So, in this module let us discuss various data science tools, their features as well as their benefits which are used by data scientists in order to extract useful information and insights from the collected data.

As we all know data scientists are the ones who will extract useful insights from the data for improving the business of a particular company. They are will do a lot of things such as data analysis, data cleaning, extracting useful insights, and communicating the findings with their team members as well as with the higher authority to improve the business. In order to do all these things, data science tools and some programming languages are used by data scientists to make the process easier. What are the tools used by data scientists, and what are their key features and their benefits all these things will be discussed more clearly in this module? Mainly these tools are used for data analysis and for providing predictions about the business to the higher authority.

Most important and highly used data science tools

So many data science tools are used by each and every data scientist while working in a company for improving their business. But some tools are used by the data scientists most frequently.

Frequently used data science tools by a data scientists

SAS, ApacheSpark, BigML,D3.js, MATLAB, Excel,ggplot2, Tableau, Jupyter, Matplotlib, NLTK, Scikit-learn, TensorFlow, Weka are the most frequently used data science tools which is used by data scientists.

1. SAS

SAS is a data science tool that is mainly and perfectly designed for statistical operations. A vast number of organizations are using SAS for analyzing data. For statistical modeling data, scientists mainly use the base SAS programming language.

Many statistical libraries are provided by the SAS for data scientists, not only libraries but also a lot of tools for analyzing the data, data modeling, and also for data organizing. SAS is a very powerful and strong tool and provides great support for the companies that use SAS. But it is seen that only large and multi-national companies will use SAS because this tool is very expensive so only large companies can afford it. The up-gradation which is available in this tool is also very expensive.

Features of SAS

1. The ability for analysing data is very strong
2. For the 4th generation programming language this tool is very flexible.
3. Availability of SAS studio
4. Algorithms for data encryptions are available.
5. Different types of data formats are very well supported.
6. Availability of report output format
7. Management

2. Apache Spark

Apache Spark or simply spark is the most frequently used data science tool. It is an analytics engine that is very powerful and it is designed in such a way that it can handle all types of the batch as well as stream processing.

When we are comparing MapReduce with Spark it is very much clear that Spark is far better and faster than MapReduce. A lot of machine learning APIs are present in Apache Spark which leads to powerful predictions from the obtained data. Streaming data is handled very efficiently by this tool. Some tools will only handle historical data that too in batches but this tool is very powerful and it can handle real-time data very efficiently.

Features of Apache Spark

1. Advanced Analytics
2. Real-Time Stream Processing
3. False Tolerance
4. Lazy type of evaluation
5. Reusability
6. High Speed

3. BigML

BigML is another popular tool that is used in data science. In order to process machine learning algorithms, BigML will provide a completely interactable as well as cloud-based GUI environment.

A standardized software is provided by the BigML using cloud computing for meeting the company requirements. One of the main specialties of BigML is it specializes in predictive modeling. Many machine learning algorithms are used by BigML for example time series forecasting, clustering, classification, etc (link to ml)

4. D3.js

D3.js tool is completely based on Javascript. Animated transactions can be done using the D3.js tool. If a data scientist is working on a device that is IOT based where a client-side interaction, as well as visualization to process data, is required then the D3.js tool will be very useful.

Illustrations and transitory visualizations can be made by combining D3.js with CSS which will help a data scientist in order to implement the customized types of graphs on the web pages. In IoT, client-side interactions can be made with the help of the D3.js tool.

D3.js Features

1. Completly based on Javascript
2. Animated Transactions can be created
3. Used in IoT for client-side interactions
4. Interactive visualization can be made
5. Open source

5. MATLAB

This data science tool is mainly used for processing data that has more mathematical information. Statistical modeling, matrix functions, and algorithmic implementation of data which is collected from different sources are facilitated by the tool named MATLAB. Mostly this tool is used in scientific disciplines, for stimulating fuzzy logic as well as neural networks.

Powerful visualizations can be created by using the MATLAB graphics library. It is also used for processing images as well as signals. This is the most powerful tool because it can solve almost all sorts of problems right from data cleaning. It is very much useful for deep learning as well as for solving complex mathematical operations.

MATLAB features

1. Very easy to use
2. Errors and bugs can be fixed very easily
3. Fast
4. Display capabilities are excellent
5. It is Platform independent

6. EXCEL

It is the most commonly used data science tool. Excel tool is developed by Microsoft and it is mainly developed for spreadsheet calculations the amazing thing is nowadays it is also used for processing data, data visualization, and also for very complex calculations. One of the main disadvantages of this tool is a large amount of data calculation is not supported but powerful data visualizations and spreadsheets can be created.

Different types of formulas, tables, filters, and slicers are there within the EXCEL tool. It always provides an easy connection with the SQL and is always used for analyzing small-scale data.

EXCEL features

1. Analyse small-scale data
2. Easy connection with SQL
3. Spread sheet calculations and visualization is done
4. For complete data analysis excel tool is used

7. ggplot2

It is a special data science tool package that is mainly used in the R programming language for advanced data visualization. In order to create illustrious visualization powerful commands are used by the ggplot2 data science tool.

Customized visualization can be created using ggplot2 by a data scientist. There are many data visualizing tools in data science but this tool is very different from other data visualizing tools and it is very efficient also. When you are using the ggplot2 tool the intractability of the graphs is boosted, text labels can be added to the data points, and we can also easily annotate our data in visualizations.

8. Tableau

It is a data visualization software. This software package consists of powerful graphics which help to make visualizations more interactive. If the companies working field is business intelligence there the tool Tableau is used frequently. Geographical data visualization and the longitudes, as well as latitudes in the maps, can be easily plotted using Tableau. The main abilities of the tool Tableau are interfacing with databases, spreadsheets, Online Analytical Processing cubes and etc.

9. Jupyter

This tool is mainly used for helping developers who are involved in making open-source software as well as if they are experiencing interactive computing. It is an open-source tool completely based on IPython.

Julia, R, and python are the multiple languages that are supported by Jupiter. The requirements of data science are addressed mainly by Jupyter.

10. Matplotlib

This tool is mainly developed for python. From the analyzed data, this tool helps in generating graphs. Matplotlib is a library for plotting and also for visualization which has been strictly developed for Python. A simple line of code is used for plotting complex graphs.

Line plot, Scatter plot, histogram, bar chart, and pie chart are the plotting techniques used by matplotlib. Using this tool vast amount of data can be handled very easily and it can be represented using graphs, charts, etc very efficiently.

11. NLTK

Tokenization, stemming, tagging, parsing, and machine learning are the various language processing techniques and NLTK is the tool mostly used for this.

It contains more than 100 corpora, where corpora are nothing but a collection of data and it is used to construct machine learning models. This tool is specially used for text analytics and also for natural language processing tasks. Speech tagging, word segmentation, Machine Translation, and text-to-speech recognition are some of the applications of NLTK.

12. Scikit learn

In order to implement machine learning algorithms mostly sci-kit learn is used. It is used for analyzing data and it is simple and easy to use compared to other data science tools.
Data pre-processing, classification, regression, clustering, and dimensionality reduction are the various machine learning features that are supported by Scikit learn. When we are using this tool it will make the usage of complex machine algorithms easy.

13. TensorFlow

It is a tool with high processing ability because of that itself it has a variety of applications such as speech recognition, the discovery of drugs, language and image generation, classification of images,s, etc.

TensorFlow tool should be compulsorily known to all the data scientists who is specializing in Machine Learning. This tool can easily run on many platforms such as CPUs, GPUs, and also in TPU platforms.

14. Weka

The full form of this tool is Waikato Environment for Knowledge Analysis and shortly known as WEKA. This tool is a machine learning software and it is written in Java. It consists of many various types of machine learning algorithms and it is mainly used for data mining. It is also known as data mining or machine learning tools.

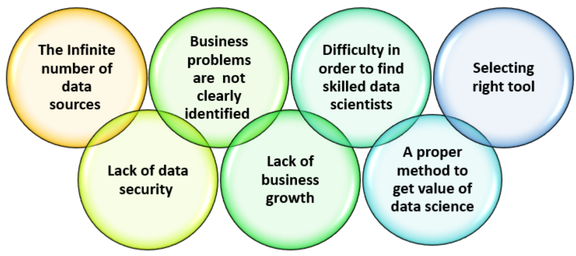
Compared to other data science tools Weka is less expensive. The main advantage of this tool is its availability this tool is free, and highly portable which means it can run on most modern platforms, it supports data pre-processing, clustering, regression, visualization, and selecting features.GUI and command line can be used very easily.

Challenges of data science technology

Nowadays data science plays a very important role in the growth of a company's business. Our world already changed into a digital world where data science becomes very significant. If an organization wanted to be successful there should be growth in its business and in this century it is possible mainly with the help of machines as well as data science.

Data science is very popular nowadays and there are many challenges faced by data science technology. There are many challenges such as problem identification, issues faced by data growth, lack of data security, sometimes it is very difficult to find out skilled data scientists and many more challenges exist. So this module lets us discuss some of the data science challenges and solutions.

Most important challenges of data science



1 The Infinite number of data sources

as we all know a vast amount of data is produced each second from multiple data sources. various software and many mobile applications such as ERPs and CRMs are used by most companies in order to collect as well as manage all the information which is mainly related to the customers, their sales, or employees associated with a particular company. Sometimes the collected data will be unstructured or semi-structured so data consolidation will become a very complex process.

Data science is mainly used to extract useful insights from the data which is collected from different sources. So it is very difficult for each data scientist to extract and understand insights from the data which is produced from heterogeneous sources. To solve this they may take more time because, in order to filter it more time is required and it becomes a time taking process, as a result, it always ends up with errors and improper decision making.

The one main solution to this is to standardize data so that accurate analysis can be done. The other solution to handle this problem is each company uses many sources to collect data so all the data sources which is used is listed and we will find out a centralized platform. This will mainly help to integrate data that are collected from those sources. Then a quality management plan and a data strategy are created this is because data collected from different sources are dynamic.

2 Lack of data security

Nowadays each and every company uses data science only to improve their business by extracting useful insights from the collected data. It also helps a company to identify the business opportunities, as well as it also improves the overall business performance. So lack of data security is the main challenge which is faced by data science. Virus attacks, theft, and attack faced by the data systems are some of the vulnerabilities. These all lead to a lack of data security. Among this information, theft is more vulnerable which means the information theft leads to the leakage of confidential data of the organization and sensitive data of the customers.

So only solution to this problem is for each and every company should strictly follow the three fundamentals of data security and they are confidentiality, integrity, and accessibility.

3 Business problems are not clearly identified

To solve a problem first we should identify the problem clearly. If it is not done clearly the root cause of the problem will be unclear and the solutions will be improper which will not give a solution to the problem. The business problems are not clearly identified because of many reasons such as the carelessness of data scientists, lack of skilled workers, getting into prediction before properly identifying the problem, and many more.

The solution to this problem is a proper method of strategizing a workflow. In order to create a workflow, a checklist should be created by collaborating all the information from all the departments. So this leads to proper identification of the problem.

4 Lack of business growth

The growth of the business completely depends on how well you are identifying the business problem and how properly you are solving the problem. In order to boost the business growth, we should identify key metrics such as we should be a very clear goal and vision, return on investment, number of production deployments, delivering actionable insights, and so on.

5 Difficulty in order to find skilled data scientists

There are many data scientists but the problem is, that it is very difficult to find out skilled data scientists. Skilled data scientists are the ones who can handle ML and AL algorithms very smoothly and they will also have a deep understanding in all these areas. This problem can be solved with proper monitoring and selecting data scientists with more experience.

6 A proper method should be used in order to get value out of data science

in order to get value out of data science, a proper method should be taken. The first thing is we should be able to understand the need of a company in order to improve their business. There should be proper communication between the team members which will help to make better decisions and healthy communication always build a bond between the teammates which will always help to make better decisions to improve the business. To get value out of data science proper understanding of the need of customers, the right customers should be targeted, and make the team more effective.

7 Selecting right tool

many tools are used by data scientists to solve many problems faced by a company to improve business. Each and every problem are different from one another and selecting the right tool will only help to solve the problem. Many tools(refer to the module of data science tools) are there, which one is appropriate that should be used then only data scientists can solve the problem. Selecting the right tool is most important. The solution to this is always to take help from experienced professionals and select the right tool.

8 Quality of data

The vast amount of data is collected from different sources and the quality of data may be different. To get a proper solution to a business problem quality of data is very important. If the data analysts select incorrect data it's too dangerous. If the input data quality is less then it affects the output badly. Data quality becomes less because of many reasons such as if errors occur at the time of data entry or the data disparity. the solution to this problem is, that monitoring should be done properly at the time of data entry and system integration can be done to solve the challenge faced by asymmetric data.

Prerequisites of Data Science

Today data science has become very important for every company. The role of data science in each company for improving their business is increasing day by day. In order to find out unseen patterns, extract meaningful information, and make perfect business decisions modern tools, as well as techniques, are used on the massive amount of data that is collected from different sources. In total, this is what data science deals with. To build predictive models data science uses machine learning algorithms.

If you want to start a career in data science, it is necessary to acquire a piece of very good knowledge in the course material data science, strong communication skills are required its only because then only the useful insights and the conclusions which are invented can be shared and discussed with the higher authorities as well as to the teammates. When working with real-time projects very good practical experience is also acquired.

So in this module let us discuss all the prerequisites that are required in order to learn data science. As we all data science is a technique that can be applied in any domain. Let's get into some of the prerequisites that needed to be known so that each and everyone can easily make a transition mainly toward data science.

Prerequisites of data science

Technical Data Science Prerequisites and Non-Technical Data Science Prerequisites are the main two categories of data science prerequisites.

Before starting to learn data science we should know some technical concepts and let us look what they are:

1. **Machine learning**

   Machine learning is known as the back bone of data science . In order to make quality predictions and estimations each and every data scientist should have a deep knowledge in machine learning. This will help the machines to take proper and right decisions mainly in real time with out the help of human beings intervention. The machine learning is the main branch of artificial intelligence and it is completely based on idea where the system will be able to learn from data, pattern identification and decisions are made with minimal human intervention.

2. **Modelling**

   Mathematical models are used to support data science. Quick calculations and predictions are made with the help of mathematical models all these are done on the data which is obtained from different sources. Modelling is mainly used to identify the most appropriate algorithms which is more suitable for problem solving and it also guide with how to train the models.

   The understanding of the problem, extracting the useful data, data cleaning ,Exploratory data analysis, features selection, incorporating the machine algorithms, testing the models and finally deploying the model are the various steps that are involved in data science modelling.

3. **Statistics**

   Statistics are known as the core of data science. In order to get meaningful insights from data first thing is to understand the data very well. To understand , to interpret and to evaluate the data in a detail manner statistics is the best tool.(link can given to basics of statistics)

   Mainly there are two types of statistics and they are descriptive statistics and inferential statistics. Descriptive statistics are again divided into measure of central tendency and measure of variability. Then measure of central tendency consists of mean, mode and median. The measure of variability consists of range, variance and dispersion. Data can be generated from different sources and these generated data are Collected and stored then it is Measured after that Analysing is done and finally it is visualised . All these are done successfully using statistical models and graphs.

4. **Programming**

   In order to execute a successful project completely based on data science high level programming is required. There are many programming language in that most common languages are python and R. Among these two languages python is the most common language because it is very easy to learn as well as it supports multiple libraries mainly for data science. Apache Hadoop, Tableau are the main programming tools in data science.

5. **Databases**

   How a data science works, how we should manage a database, and how we will extract the useful insights from data all these things should be known by a data scientist. Database plays very important role in each and every data science project its because we are obtaining data from different sources and initially it is stored in a database and data is retrieved from the database. A database is nothing but it is a structured set of data which will be there in the computer memory or it is stored in the cloud. There are various ways as well as methods in order to access the data. A data scientist should design, create and interact with the database which is there in the computer memory or cloud based on which project we are working. To handle structured data a data scientists needs SQL and the structured data is there in the relational database.

6. **Be good at mathematics**

   In the life cycle of data science project each and every modules consists of selecting the features, creating models, modelling every where mathematics is highly involved . Great knowledge in maths are required for each and every data scientist. Mathematical study is very important for a data scientist to reach somewhere in the data science career its all because to perform machine learning algorithms, to extract useful insights from data and for analysing the model for all these things mathematics is required. Statistics, probability, linear algebra and calculus are the main kind of maths used in data science.

7. **Data visualization**

   One of the important prerequisite for data science is data visualization. Representing the data with the help of graphs, pie charts, maps etc is known as data visualization.

   For better data visualisation there are multiple components such as data component, geometric component, mapping component, label component, scale component and ethical component. Data visualization is known as the subset of data science . The very effective data visualization techniques are scatter chart, bar charts, box plot, pair plot, kde charts, histogram, hexbin plots, line charts, heat maps, pie charts, area plot etc.