Data: *Data* is a collection of raw facts and numbers used to analyze something or make decisions. Computer *data* is information in a form that can be processed by a computer.

Data Analysis: **Data analysis** is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making

Data Driven Decision Making: Data-driven decision-making (DDDM) is defined as using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives. When organizations realize the full value of their data, that means everyone—whether you're a business analyst, sales manager, or human resource specialist—is empowered to make better decisions with data, every day. However, this is not achieved by simply choosing the appropriate analytics technology to identify the next strategic opportunity.

Challenges in becoming a data driven organization

Due to technical disruption Data Science is becoming strong day by day an organization will have many benefits through data driven decision making in spite of the benefits there are many hurdles to become a n data driven organization

i. **Data Quality:**

Any data that is inaccurate, incomplete and out of date is of no use to the organization. Having poor quality data only increases your risk of making bad decisions, eventually leading to a loss in revenue.

Organizations should follow these 5 principles to cultivate good quality data.

1. Accuracy
2. Completeness
3. Consistency
4. Uniqueness
5. Timeliness

ii. **Tools for access,extraction,processing and analysis:**

Organizations need adequate storage and processing capabilities to access the data Another challenge is that the data may be available in human readable formats (for instance, in pdf files) rather than machine readable formats, So, it must be processed with the Optical text recognition softwares. So, even when the reliable data is available, the Organizations must possess the appropriate tools and technology for accessing the data and storing it and processing the data

iii Cost

Another challenge for becoming a data-driven organization is to find the balance between cost and access. Lot of human hours will be taken for data cleaning, analysis, management which is costing the organization. It incurs considerable amount in expanding the datasets. If the data is not first party data (data generated by the organization), then the data acquired from other parties might also incur cost

(iv) **Legacy Data**

The Organizations old information that is stored in an old /obsolete format is called as Legacy data function and to retain comprehensive business knowledge. Many Organizations sit on a big pile of old data (For example., Data stored in Excel 5.0 from 1990's) or use outdated systems to manage valuable information. Though these legacy data are still valuable to the Organization, its poor usability and inefficiency poses high maintenance costs. So, these legacy files need to be converted to new file formats to make them accessible, readable and usable in the future.

(v) **Organizational Politics**

Organization politics affects what data is shared with in the organization, within the departments or between the departments. This human element can create hurdles as it affects how and what data is shared to whom. The data formats are critical in sharing and matching across the datasets and easiness of processing this data depends on the decisions of humans on data formats.

(vi) **Data Silos**

A Data Silo is the collection of data or information held by one group that is isolated from and inaccessible to other groups of the enterprise. It can result from many factors including culture and competition between the departments that causes those employees to keep data from each other, rather than working together. Finance, administration, HR, and other departments need different information to do their work, and so do they collect. But the problem of data silos arises when the data is made inaccessible to other parts of the enterprise. For example, when accounting can't access current data from operations. Data Silos limit the view of data, threaten the data integrity, wastes resources and discourage collaborative work. As the quantity and diversity of data continue to grow too.

(vii) **Linking the Data**

To take appropriate decisions, the datasets from different sources must be combined, not an easy task. Data sets are owned and managed by various groups (Organizations, Internet service providers, vendors). Every group has their own interest in sharing the Linking data across data sets is a challenge.

(viii) **Strong Leadership**

The success of a data-driven organization critically depends on strong leadership. The leader to be aware of all the challenges and limitations of data sets, its cost and quality. The right resource teams have to be assigned in this task of moving towards data.

(ix) **Knowledge of Data Scientists**:

The challenge is that many organizations are not aware of how they could be using intelligent software to use data-driven insights to increase efficiency and revenue. Lack of training the intelligent software and usage of statistical techniques and inability to understand the quality data and its sources, and inability to draw the inferences poses challenges to the organizations become a data-driven organization.

ANALYZING DATA PRACTICES IN ORGANIZATIONS

Data is becoming the core corporate asset. Data Sources and the amount of data is growing, so ability to utilize this data and turn it into knowledge. The Companies which are seeing it as source of competitive advantage are transforming their culture to data driven.

Data Governance is the initiative that takes to create and enforce a set of Rules and policies regarding to the data management and plays an important role in building a data-driven culture. Effective data governance leads to improvement in data quality, decrease in data management cost as required. Data Governance ensures that the data inconsistent, trustworthy and Data.the process of Data Governance involves Data Steward.

Data Steward is a role within an organization responsible for utilizing an organization's data governance processes to make sure of fitness of data elements. They ensure high-quality data is easily accessible in a consistent manner. Data stewards share some responsibilities with data custodians.

Data Custodians are responsible for the safe custody, transport, storage of the data and implementation of the business rules.

Structured Data and Unstructured Data:

What is structured data?

So, structured data is the type of data that is well-organized and accurately formatted. This data exists in the format of relational databases (RDBMSs), meaning the information is stored in tables with rows and columns that are connected.

For analytical purposes, you can use data warehouses. DWs are central data storages used by companies for data analysis and reporting.

There is a special programming language used for handling relational databases and warehouses called SQL, which stands for Structured Query Language and was developed back in the 1970s by IBM.

What is unstructured data?

It makes sense that if the definition of structured data implies a neat organization of components in a predetermined manner, the definition of unstructured data will be the opposite. The pieces of such data aren't structured in a pre-defined way, meaning data is stored in its native formats.

One of the ways to manage unstructured data is to opt for non-relational databases, also known as NoSQL.

If there's a need to keep data in its raw native formats for further analysis, storage repositories called data lakes will be the way to go. A data lake is a storage repository or system meant to store huge volumes of data in its natural/raw formats.

Unstructured data examples. There is a wide array of forms that make up unstructured data such as email, text files, social media posts, video, images, audio, sensor data, and so on.

	Structured data	Unstructured data
What is it?	Data that fits in a predefined data model or schema.	Data without an underlying model to discern attributes.
Basic example	An Excel table.	A collection of video files.
Best for	An associated collection of discrete, short, non-continuous numerical and text values.	An associated collection of data, objects, or files where the attributes change or are unknown.

Storage types	Relational databases, graph databases, spatial databases, OLAP cubes, and more.	File systems, DAM systems, CMSs, version control systems, and more.
Biggest benefit	Easier to organize, clean, search, and analyze.	Can analyze data that can’t be easily shaped into structured data.
Biggest challenge	All data must fit in the prescribed data model.	Can be difficult to analyze.
Main analysis technique	SQL queries.	Varies.

How data benefits to organization:

Data provides us with actual facts and metrics which makes decision making trustworthy. Here are the benefits of Data to any Organization

1. **The Decisions are taken more confidently:**

Earlier Decision making made on launching a new product or give discount on product, train employee, acquiring new companies based on intuition or gut feeling but by basing on data collection and analysis of data the decision is taken more confidently

2. Competitive advantage:

As the decisions taken on proper analysis of data the company would be a bit ahead of competitors by detecting the opportunities. of competitors by detecting the opportunities, and the threats than our competitor companies.

3. Less organization politics:

AS decisions are supported with strong data, statistics and software, there will be less scope of politics based on ego levels of the employee involved in decision making therefore morale of employees improves.

4. Timely Decisions:

With proper tools to access, process and analyze the data, a lot of time will be savedand decisions are taken at the right time.

5. Proactive Decision Making:

Proactive data-driven decision-making can help businesses to grow in several ways. For example, it can help companies to identify new market opportunities, optimize their marketing campaigns, and improve their product development process, enabling them to seize first-mover advantages.

6.Increase in the Team spirit:

With data-driven insights, you can set realistic and relevant goals for your team, identify and address gaps and issues in their skills, knowledge, and behaviors, provide timely and constructive feedback and recognition, adjust your leadership style and communication methods to match their preferences and expectations

Data Science Use Cases that are Changing the World

Earlier we saw many **data science applications**. Today we will see the diverse data science use cases. We will take examples of social media, e-commerce, transportation, and healthcare to demonstrate some of the important data science use cases in contemporary industries.

1. Facebook – Using Data to Revolutionize Social Networking & Advertising

Facebook is a social-media leader of the world today. With millions of users around the world, Facebook utilizes a large scale quantitative research through data science to gain insights about the social interactions of the people.

Facebook has become a hub of innovation where it has been using advanced techniques in data science to study user behavior and gain insights to improve their product. Facebook makes use of advanced technology in data science called **deep learning**.

Using deep learning, Facebook makes use of facial recognition and text analysis. In facial recognition, Facebook uses powerful neural networks to classify faces in the photographs. It uses its own text understanding engine called “DeepText” to understand user sentences.

It also uses Deep Text to understand people’s interest and aligning photographs with texts.

However, more than being a social media platform, Facebook is more of an advertisement corporation. It uses deep learning for targeted advertising. Using this, it decides what kind of advertisements the users should view.

It uses the insights gained from the data to cluster users based on their preferences and provides them with the advertisements that appeal to them.

Now, let’s have a look at another data science use case – Amazon

2. Amazon – Transforming E-commerce with Data Science

Since its inception, Amazon has been working hard to make itself a customer-centric platform. Amazon heavily relies on **predictive analytics** to increase customer satisfaction. It does so through a personalized recommendation system.

This recommendation system is a hybrid type that also involves collaborative filtering which is comprehensive in nature. Amazon analyzes the historical purchases of the user to recommend more products.

This also comes through the suggestions that are drawn from the other users who use similar products or provide similar ratings.

Amazon has an anticipatory shipping model that uses big data for predicting the products that are most likely to be purchased by its users. It analyzes the pattern of your purchases and sends products to your nearest warehouse which you may utilize in the future.

Amazon also optimizes the prices on its websites by keeping in mind various parameters like the user activity, order history, prices offered by the competitors, product availability, etc. Using this method, Amazon provides discounts on popular items and earns profits on less popular items.

Another area where every e-commerce platform is addressing is **Fraud Detection**. Amazon has its own novel ways and algorithms to detect fraud sellers and fraudulent purchases.

Other than online platforms, Amazon has been optimizing the packaging of products in warehouses and increasing the efficiency of packaging lines through the data collected from the workers.

3. Uber – Using Data to Make Rides Better

Next in data science use cases is Uber. Uber is a popular smartphone application that allows you to book a cab. Uber makes extensive use of **Big Data**. After all, Uber has to maintain a large database of drivers, customers, and several other records.

It is therefore, rooted in Big Data and makes use of it to derive insights and provide the best services to its users. Uber shares the big data principle with crowdsourcing. That is, registered drivers in the area can help anyone who wants to go somewhere.

As mentioned above, Uber contains a database of drivers. Therefore, whenever you hail for a cab, Uber matches your profile with the most suitable driver. What differentiates Uber from other cab companies is that Uber charges you based on the time it takes to cover the distance and not the distance itself.

It calculates the time taken through various algorithms that also make use of data related to traffic density and weather conditions.

Uber makes the best use of data science to calculate its surge pricing. When there are less drivers available to more riders, the price of the ride goes up. This happens only during the scarcity of drivers in any given area.

However, if the demand for Uber rides is less, then Uber charges a lower rate. This dynamic pricing is rooted in Big Data and makes excellent usage of data science to calculate the fares based on the parameters.

4. Bank of America – Using Data to Leverage Customer Experience

10 years ago, Bank of America was one of the first financial companies to provide mobile banking to its customers. Recently, BoA launched Erica which is their first virtual financial assistant. It is considered as the world's finest innovation in finance domain.

Currently, Erica is serving as a customer advisor to more than 45 million users around the world. Erica also makes use of Speech Recognition to take customer inputs, which is a technological advancement in the field of Data Science.

Furthermore, several other banks like BoA are making use of **Data Science and predictive analytics**. Using data science, banking industries are able to detect frauds in payments and

customer information. It also prevents frauds regarding insurances, credit cards, and accounting.

In order to minimize the losses, a bank needs to detect fraud sooner. In order to carry this out, banks employ data scientists to use their quantitative knowledge where they apply algorithms like association, clustering, forecasting, and classification.

Risk modeling is another important area that is supervised by the banks to regulate financial activities. Using Machine Learning, banks are able to minimize risk modeling.

Through analytical solutions, banks can make data-driven decisions that are based on transparency and risk analysis. Furthermore, Bank of America detected the high-risk accounts using this technology of big data.

Various banks like BoA are understanding their customers through an intelligent customer segmentation approach. Through various data-mining techniques, banks are able to segment their customers in the high-value and low-value segments.

There are various techniques that a data scientist makes use of such as clustering, logistic regression, decision trees to help the banks to understand the Customer Lifetime Value (CLV) and take group them in the appropriate segments.

Data Science Case Study – How Netflix Used Data Science to Improve its Recommendation System?

Do you remember the last movie you watched on Netflix? I don’t want to know the name; just think about it- after watching the movie, were you recommended of similar movies? How does Netflix know what you’d like? The secret here is Data Science.

Netflix uses Data Science to cater relevant and interesting recommendations to you. So, today, in this article, we will discuss the same. Let’s start exploring Data Science at Netflix with a basic introduction to Netflix.

Data Science at Netflix

Netflix initially started as a DVD rental service in 1998. It mostly relied on a third party postal services to deliver its DVDs to the users. This resulted in heavy losses which they soon mitigated with the **introduction of their online streaming service** in 2007.

In order to make this happen, Netflix invested in a lot of algorithms to provide a flawless movie experience to its users. One of such algorithms is the **recommendation system** that is used by Netflix to provide suggestions to the users.

A recommendation system understands the needs of the users and provides suggestions of the various cinematographic products.

What is a Recommendation System?

A recommendation system is a platform that provides its users with various contents based on their preferences and likings. A recommendation system takes the information about the user as an input.

This information can be in the form of the past usage of product or the ratings that were provided to the product. It then processes this information to predict how much the user would rate or prefer the product. A recommendation system makes use of a variety of **machine learning algorithms**.

Another important role that a recommendation system plays today is to search for similarity between different products. In the case of Netflix, the recommendation system searches for movies that are similar to the ones you have watched or have liked previously.

This is an important method for scenarios that involve cold start. In cold start, the company does not have much of the user data available to generate recommendations.

Therefore, based on the movies that are watched, Netflix provides recommendations of the films that share a degree of similarity. There are two main types of Recommendation Systems –

1. Content-based recommendation systems

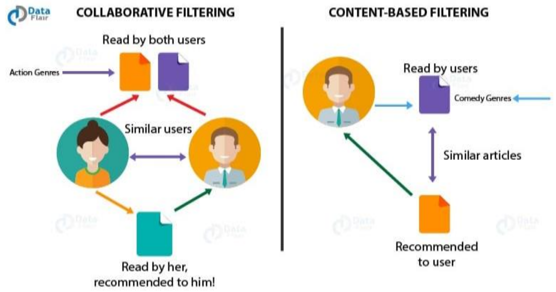
In a content-based recommendation system, the background knowledge of the products and customer information are taken into consideration. Based on the content that you have viewed on Netflix, it provides you with similar suggestions.

For example, if you have watched a film that has a sci-fi genre, the content-based recommendation system will provide you with suggestions for similar films that have the same genre.

2. Collaborative filtering recommendation systems

Unlike the content based filtering that provided recommendations of similar products, Collaborative Filtering provides recommendations based on the similar profiles of its users. One key advantage of collaborative filtering is that it is independent of the product knowledge.

Rather, it relies on the users with a basic assumption that what the users liked in the past will also like in the future. **For example**, if a person A watches crime, sci-fi and thriller genres and B watches sci-fi, thriller and action genres then A will also like action and B will like crime genre.



There is also a third type of recommendation system that combines both Content and Collaborative techniques. This form of recommendation system is known as **Hybrid Recommendation System**. Netflix makes the primary of use Hybrid Recommendation System for suggesting content to its users.

Most recommended – Marks & Spencer using Big Data to Analyze Customer Behaviour

Follow DataFlair on WhatsApp & Stay updated with latest technology trends.

How Netflix Solved its Recommendation Problem with Data Science

Back in 2006 when Netflix wanted to tap into the streaming market, it started off with a competition for movie rating prediction. It provided a prize of $ 1 million to whoever increased the accuracy of their then existing platform ‘Cinematch’ by 10%.

At the end of competition, the BellKor team presented their solution that increased the accuracy of prediction by 10.06%. With over 200 work hours and an ensemble of 107 algorithms provided them with this result.

Their final model gave an RMSE of 0.8712. For their solution, they made use of K-nearest neighbor algorithm for post-processing of the data.

Then they implemented a factorization model which is popularly known as **Singular Value Decomposition (SVD)** for providing an optimal dimensional embedding to its users.

They also made use of **Restricted Boltzmann Machines (RBM)** for enhancing the capability of the collaborative filtering model. These two algorithms in the ensemble, SVD and RBM provided them with the best results. A linear combination of these two algorithms reduced the RMSE to 0.88.

However, even after reduction of RMSE and increase in accuracy, Netflix suffered from two major challenges – Firstly, the data that provided during the competition comprised of 100 million movie ratings, as opposed to more than 5 billion ratings that Netflix constituted of.

Furthermore, the algorithms were static, meaning that they only dealt with historical data and did not take into account the dynamicity of users adding reviews in real-time. After Netflix overcame these challenges, it made the winning algorithms a part of its recommendation system.

Using Interleaving to Improve Personalization

Netflix uses Ranking Algorithms to provide a ranked list of movies and TV Shows that appeal the most to its users. However, with the presence of various ranking algorithms, it is often difficult to accommodate all of them and test their performance simultaneously.

While the traditional A/B testing on a reduced set of algorithms could not identify the best algorithms with smaller sample size and also consumed a lot of time, Netflix decided to innovate its algorithmic process.

In order to speed up its experimentation process of its ranking algorithms, Netflix implemented the interleaving technique that allowed it to identify best algorithms. This technique is applied in two stages to provide the best page ranking algorithm to provide personalized recommendations to its users.

In the first stage, experimentations to determine the member preference between the two ranking algorithms is carried out. Unlike the A/B testing where the two groups of viewers are exposed to the two ranking algorithms, Netflix makes use of interleaving to blend the rankings of algorithm A and B.

Netflix provides its users with enriched content based on this interleaving technique that is highly sensitive towards ranking the algorithm quality.

Importance of Context Awareness in Recommendations

Contextual Awareness is one of the key elements in personalizing recommendations for its users.

This not only improves the performance of the recommendation system but also prompts users to provide better feedback that would result in a quality recommendation. There are two categories of contextual classes:

Explicit

- Location
- Language
- Time of the Day
- Device

Inferred

- Binging Patterns
- Companion

In order to predict contexts, we make use of representation learning. It is a **deep learning technique** that performs feature engineering that discovers features without explicit programming.

Based on the time and periods of watching, Netflix bases its data on various parameters like Day, Week, Season and even longer periods like Olympics, FIFA, and elections.

It is the Right Time to Upgrade your Data Science Skills

For performing contextual predictions, Netflix treats recommendations as a sequence classification problem. It takes the input as a sequence of user-actions and performs predictions that output the next set of actions.

An example of a sequence problem is **Gru4Rec**. And in the case of contextual sequence prediction, the input consists of the contextual user actions as well as the current context of the user. This helps the recommendation engine solve the question:

“Based on all the historical actions that are taken by the user what is the most probable video that they will play right now?”

Proactive data Practioner:

Being proactive means structuring your data analytics to serve ongoing business objectives. It creates tangible, real actions that arise from data-driven insights. For example, you can use data analytics tools to create a business dashboard. The dashboard might display, on a dynamic, ever-changing basis, how your results are matching up to your KPIs.

A sales dashboard, for instance, could show sales results versus budget by product, region or sales person. You could create a dashboard that shows financial and operations data in real time so people can stay on top of cashflow.

Data: *Data* is a collection of raw facts and numbers used to analyze something or make decisions. Computer *data* is information in a form that can be processed by a computer.

Data Analysis: **Data analysis** is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making

Data Driven Decision Making: Data-driven decision-making (DDDM) is defined as using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives. When organizations realize the full value of their data, that means everyone—whether you're a business analyst, sales manager, or human resource specialist—is empowered to make better decisions with data, every day. However, this is not achieved by simply choosing the appropriate analytics technology to identify the next strategic opportunity.

Challenges in becoming a data driven organization

Due to technical disruption Data Science is becoming strong day by day an organization will have many benefits through data driven decision making in spite of the benefits there are many hurdles to become a n data driven organization

i. **Data Quality:**

Any data that is inaccurate, incomplete and out of date is of no use to the organization. Having poor quality data only increases your risk of making bad decisions, eventually leading to a loss in revenue.

Organizations should follow these 5 principles to cultivate good quality data.

1. Accuracy
2. Completeness
3. Consistency
4. Uniqueness
5. Timeliness

ii. **Tools for access,extraction,processing and analysis:**

Organizations need adequate storage and processing capabilities to access the data Another challenge is that the data may be available in human readable formats (for instance, in pdf files) rather than machine readable formats, So, it must be processed with the Optical text recognition softwares. So, even when the reliable data is available, the Organizations must possess the appropriate tools and technology for accessing the data and storing it and processing the data

iii Cost

Another challenge for becoming a data-driven organization is to find the balance between cost and access. Lot of human hours will be taken for data cleaning, analysis, management which is costing the organization. It incurs considerable amount in expanding the datasets. If the data is not first party data (data generated by the organization), then the data acquired from other parties might also incur cost

(iv) **Legacy Data**

The Organizations old information that is stored in an old /obsolete format is called as Legacy data function and to retain comprehensive business knowledge. Many Organizations sit on a big pile of old data (For example., Data stored in Excel 5.0 from 1990's) or use outdated systems to manage valuable information. Though these legacy data are still valuable to the Organization, its poor usability and inefficiency poses high maintenance costs. So, these legacy files need to be converted to new file formats to make them accessible, readable and usable in the future.

(v) **Organizational Politics**

Organization politics affects what data is shared with in the organization, within the departments or between the departments. This human element can create hurdles as it affects how and what data is shared to whom. The data formats are critical in sharing and matching across the datasets and easiness of processing this data depends on the decisions of humans on data formats.

(vi) **Data Silos**

A Data Silo is the collection of data or information held by one group that is isolated from and inaccessible to other groups of the enterprise. It can result from many factors including culture and competition between the departments that causes those employees to keep data from each other, rather than working together. Finance, administration, HR, and other departments need different information to do their work, and so do they collect. But the problem of data silos arises when the data is made inaccessible to other parts of the enterprise. For example, when accounting can't access current data from operations. Data Silos limit the view of data, threaten the data integrity, wastes resources and discourage collaborative work. As the quantity and diversity of data continue to grow too.

(vii) **Linking the Data**

To take appropriate decisions, the datasets from different sources must be combined, not an easy task. Data sets are owned and managed by various groups (Organizations, Internet service providers, vendors). Every group has their own interest in sharing the Linking data across data sets is a challenge.

(viii) **Strong Leadership**

The success of a data-driven organization critically depends on strong leadership. The leader to be aware of all the challenges and limitations of data sets, its cost and quality. The right resource teams have to be assigned in this task of moving towards data.

(ix) **Knowledge of Data Scientists**:

The challenge is that many organizations are not aware of how they could be using intelligent software to use data-driven insights to increase efficiency and revenue. Lack of training the intelligent software and usage of statistical techniques and inability to understand the quality data and its sources, and inability to draw the inferences poses challenges to the organizations become a data-driven organization.

ANALYZING DATA PRACTICES IN ORGANIZATIONS

Data is becoming the core corporate asset. Data Sources and the amount of data is growing, so ability to utilize this data and turn it into knowledge. The Companies which are seeing it as source of competitive advantage are transforming their culture to data driven.

Data Governance is the initiative that takes to create and enforce a set of Rules and policies regarding to the data management and plays an important role in building a data-driven culture. Effective data governance leads to improvement in data quality, decrease in data management cost as required. Data Governance ensures that the data inconsistent, trustworthy and Data.the process of Data Governance involves Data Steward.

Data Steward is a role within an organization responsible for utilizing an organization's data governance processes to make sure of fitness of data elements. They ensure high-quality data is easily accessible in a consistent manner. Data stewards share some responsibilities with data custodians.

Data Custodians are responsible for the safe custody, transport, storage of the data and implementation of the business rules.

Structured Data and Unstructured Data:

What is structured data?

So, structured data is the type of data that is well-organized and accurately formatted. This data exists in the format of relational databases (RDBMSs), meaning the information is stored in tables with rows and columns that are connected.

For analytical purposes, you can use data warehouses. DWs are central data storages used by companies for data analysis and reporting.

There is a special programming language used for handling relational databases and warehouses called SQL, which stands for Structured Query Language and was developed back in the 1970s by IBM.

What is unstructured data?

It makes sense that if the definition of structured data implies a neat organization of components in a predetermined manner, the definition of unstructured data will be the opposite. The pieces of such data aren't structured in a pre-defined way, meaning data is stored in its native formats.

One of the ways to manage unstructured data is to opt for non-relational databases, also known as NoSQL.

If there's a need to keep data in its raw native formats for further analysis, storage repositories called data lakes will be the way to go. A data lake is a storage repository or system meant to store huge volumes of data in its natural/raw formats.

Unstructured data examples. There is a wide array of forms that make up unstructured data such as email, text files, social media posts, video, images, audio, sensor data, and so on.

	Structured data	Unstructured data
What is it?	Data that fits in a predefined data model or schema.	Data without an underlying model to discern attributes.
Basic example	An Excel table.	A collection of video files.
Best for	An associated collection of discrete, short, non-continuous numerical and text values.	An associated collection of data, objects, or files where the attributes change or are unknown.

Storage types	Relational databases, graph databases, spatial databases, OLAP cubes, and more.	File systems, DAM systems, CMSs, version control systems, and more.
Biggest benefit	Easier to organize, clean, search, and analyze.	Can analyze data that can’t be easily shaped into structured data.
Biggest challenge	All data must fit in the prescribed data model.	Can be difficult to analyze.
Main analysis technique	SQL queries.	Varies.

How data benefits to organization:

Data provides us with actual facts and metrics which makes decision making trustworthy. Here are the benefits of Data to any Organization

1. **The Decisions are taken more confidently:**

Earlier Decision making made on launching a new product or give discount on product, train employee, acquiring new companies based on intuition or gut feeling but by basing on data collection and analysis of data the decision is taken more confidently

2. Competitive advantage:

As the decisions taken on proper analysis of data the company would be a bit ahead of competitors by detecting the opportunities. of competitors by detecting the opportunities, and the threats than our competitor companies.

3. Less organization politics:

AS decisions are supported with strong data, statistics and software, there will be less scope of politics based on ego levels of the employee involved in decision making therefore morale of employees improves.

4. Timely Decisions:

With proper tools to access, process and analyze the data, a lot of time will be savedand decisions are taken at the right time.

5. Proactive Decision Making:

Proactive data-driven decision-making can help businesses to grow in several ways. For example, it can help companies to identify new market opportunities, optimize their marketing campaigns, and improve their product development process, enabling them to seize first-mover advantages.

6.Increase in the Team spirit:

With data-driven insights, you can set realistic and relevant goals for your team, identify and address gaps and issues in their skills, knowledge, and behaviors, provide timely and constructive feedback and recognition, adjust your leadership style and communication methods to match their preferences and expectations

Data Science Use Cases that are Changing the World

Earlier we saw many **data science applications**. Today we will see the diverse data science use cases. We will take examples of social media, e-commerce, transportation, and healthcare to demonstrate some of the important data science use cases in contemporary industries.

1. Facebook – Using Data to Revolutionize Social Networking & Advertising

Facebook is a social-media leader of the world today. With millions of users around the world, Facebook utilizes a large scale quantitative research through data science to gain insights about the social interactions of the people.

Facebook has become a hub of innovation where it has been using advanced techniques in data science to study user behavior and gain insights to improve their product. Facebook makes use of advanced technology in data science called **deep learning**.

Using deep learning, Facebook makes use of facial recognition and text analysis. In facial recognition, Facebook uses powerful neural networks to classify faces in the photographs. It uses its own text understanding engine called “DeepText” to understand user sentences.

It also uses Deep Text to understand people’s interest and aligning photographs with texts.

However, more than being a social media platform, Facebook is more of an advertisement corporation. It uses deep learning for targeted advertising. Using this, it decides what kind of advertisements the users should view.

It uses the insights gained from the data to cluster users based on their preferences and provides them with the advertisements that appeal to them.

Now, let’s have a look at another data science use case – Amazon

2. Amazon – Transforming E-commerce with Data Science

Since its inception, Amazon has been working hard to make itself a customer-centric platform. Amazon heavily relies on **predictive analytics** to increase customer satisfaction. It does so through a personalized recommendation system.

This recommendation system is a hybrid type that also involves collaborative filtering which is comprehensive in nature. Amazon analyzes the historical purchases of the user to recommend more products.

This also comes through the suggestions that are drawn from the other users who use similar products or provide similar ratings.

Amazon has an anticipatory shipping model that uses big data for predicting the products that are most likely to be purchased by its users. It analyzes the pattern of your purchases and sends products to your nearest warehouse which you may utilize in the future.

Amazon also optimizes the prices on its websites by keeping in mind various parameters like the user activity, order history, prices offered by the competitors, product availability, etc. Using this method, Amazon provides discounts on popular items and earns profits on less popular items.

Another area where every e-commerce platform is addressing is **Fraud Detection**. Amazon has its own novel ways and algorithms to detect fraud sellers and fraudulent purchases.

Other than online platforms, Amazon has been optimizing the packaging of products in warehouses and increasing the efficiency of packaging lines through the data collected from the workers.

3. Uber – Using Data to Make Rides Better

Next in data science use cases is Uber. Uber is a popular smartphone application that allows you to book a cab. Uber makes extensive use of **Big Data**. After all, Uber has to maintain a large database of drivers, customers, and several other records.

It is therefore, rooted in Big Data and makes use of it to derive insights and provide the best services to its users. Uber shares the big data principle with crowdsourcing. That is, registered drivers in the area can help anyone who wants to go somewhere.

As mentioned above, Uber contains a database of drivers. Therefore, whenever you hail for a cab, Uber matches your profile with the most suitable driver. What differentiates Uber from other cab companies is that Uber charges you based on the time it takes to cover the distance and not the distance itself.

It calculates the time taken through various algorithms that also make use of data related to traffic density and weather conditions.

Uber makes the best use of data science to calculate its surge pricing. When there are less drivers available to more riders, the price of the ride goes up. This happens only during the scarcity of drivers in any given area.

However, if the demand for Uber rides is less, then Uber charges a lower rate. This dynamic pricing is rooted in Big Data and makes excellent usage of data science to calculate the fares based on the parameters.

4. Bank of America – Using Data to Leverage Customer Experience

10 years ago, Bank of America was one of the first financial companies to provide mobile banking to its customers. Recently, BoA launched Erica which is their first virtual financial assistant. It is considered as the world's finest innovation in finance domain.

Currently, Erica is serving as a customer advisor to more than 45 million users around the world. Erica also makes use of Speech Recognition to take customer inputs, which is a technological advancement in the field of Data Science.

Furthermore, several other banks like BoA are making use of **Data Science and predictive analytics**. Using data science, banking industries are able to detect frauds in payments and

customer information. It also prevents frauds regarding insurances, credit cards, and accounting.

In order to minimize the losses, a bank needs to detect fraud sooner. In order to carry this out, banks employ data scientists to use their quantitative knowledge where they apply algorithms like association, clustering, forecasting, and classification.

Risk modeling is another important area that is supervised by the banks to regulate financial activities. Using Machine Learning, banks are able to minimize risk modeling.

Through analytical solutions, banks can make data-driven decisions that are based on transparency and risk analysis. Furthermore, Bank of America detected the high-risk accounts using this technology of big data.

Various banks like BoA are understanding their customers through an intelligent customer segmentation approach. Through various data-mining techniques, banks are able to segment their customers in the high-value and low-value segments.

There are various techniques that a data scientist makes use of such as clustering, logistic regression, decision trees to help the banks to understand the Customer Lifetime Value (CLV) and take group them in the appropriate segments.

Data Science Case Study – How Netflix Used Data Science to Improve its Recommendation System?

Do you remember the last movie you watched on Netflix? I don't want to know the name; just think about it- after watching the movie, were you recommended of similar movies? How does Netflix know what you'd like? The secret here is Data Science.

Netflix uses Data Science to cater relevant and interesting recommendations to you. So, today, in this article, we will discuss the same. Let's start exploring Data Science at Netflix with a basic introduction to Netflix.

Data Science at Netflix

Netflix initially started as a DVD rental service in 1998. It mostly relied on a third party postal services to deliver its DVDs to the users. This resulted in heavy losses which they soon mitigated with the **introduction of their online streaming service** in 2007.

In order to make this happen, Netflix invested in a lot of algorithms to provide a flawless movie experience to its users. One of such algorithms is the **recommendation system** that is used by Netflix to provide suggestions to the users.

A recommendation system understands the needs of the users and provides suggestions of the various cinematographic products.

What is a Recommendation System?

A recommendation system is a platform that provides its users with various contents based on their preferences and likings. A recommendation system takes the information about the user as an input.

This information can be in the form of the past usage of product or the ratings that were provided to the product. It then processes this information to predict how much the user would rate or prefer the product. A recommendation system makes use of a variety of **machine learning algorithms**.

Another important role that a recommendation system plays today is to search for similarity between different products. In the case of Netflix, the recommendation system searches for movies that are similar to the ones you have watched or have liked previously.

This is an important method for scenarios that involve cold start. In cold start, the company does not have much of the user data available to generate recommendations.

Therefore, based on the movies that are watched, Netflix provides recommendations of the films that share a degree of similarity. There are two main types of Recommendation Systems –

1. Content-based recommendation systems

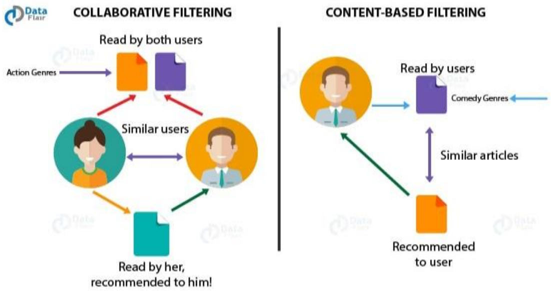
In a content-based recommendation system, the background knowledge of the products and customer information are taken into consideration. Based on the content that you have viewed on Netflix, it provides you with similar suggestions.

For example, if you have watched a film that has a sci-fi genre, the content-based recommendation system will provide you with suggestions for similar films that have the same genre.

2. Collaborative filtering recommendation systems

Unlike the content based filtering that provided recommendations of similar products, Collaborative Filtering provides recommendations based on the similar profiles of its users. One key advantage of collaborative filtering is that it is independent of the product knowledge.

Rather, it relies on the users with a basic assumption that what the users liked in the past will also like in the future. **For example**, if a person A watches crime, sci-fi and thriller genres and B watches sci-fi, thriller and action genres then A will also like action and B will like crime genre.



There is also a third type of recommendation system that combines both Content and Collaborative techniques. This form of recommendation system is known as **Hybrid Recommendation System**. Netflix makes the primary of use Hybrid Recommendation System for suggesting content to its users.

Most recommended – Marks & Spencer using Big Data to Analyze Customer Behaviour

Follow DataFlair on WhatsApp & Stay updated with latest technology trends.

How Netflix Solved its Recommendation Problem with Data Science

Back in 2006 when Netflix wanted to tap into the streaming market, it started off with a competition for movie rating prediction. It provided a prize of $ 1 million to whoever increased the accuracy of their then existing platform 'Cinematch' by 10%.

At the end of competition, the BellKor team presented their solution that increased the accuracy of prediction by 10.06%. With over 200 work hours and an ensemble of 107 algorithms provided them with this result.

Their final model gave an RMSE of 0.8712. For their solution, they made use of K-nearest neighbor algorithm for post-processing of the data.

Then they implemented a factorization model which is popularly known as **Singular Value Decomposition (SVD)** for providing an optimal dimensional embedding to its users.

They also made use of **Restricted Boltzmann Machines (RBM)** for enhancing the capability of the collaborative filtering model. These two algorithms in the ensemble, SVD and RBM provided them with the best results. A linear combination of these two algorithms reduced the RMSE to 0.88.

However, even after reduction of RMSE and increase in accuracy, Netflix suffered from two major challenges – Firstly, the data that provided during the competition comprised of 100 million movie ratings, as opposed to more than 5 billion ratings that Netflix constituted of.

Furthermore, the algorithms were static, meaning that they only dealt with historical data and did not take into account the dynamicity of users adding reviews in real-time. After Netflix overcame these challenges, it made the winning algorithms a part of its recommendation system.

Using Interleaving to Improve Personalization

Netflix uses Ranking Algorithms to provide a ranked list of movies and TV Shows that appeal the most to its users. However, with the presence of various ranking algorithms, it is often difficult to accommodate all of them and test their performance simultaneously.

While the traditional A/B testing on a reduced set of algorithms could not identify the best algorithms with smaller sample size and also consumed a lot of time, Netflix decided to innovate its algorithmic process.

In order to speed up its experimentation process of its ranking algorithms, Netflix implemented the interleaving technique that allowed it to identify best algorithms. This technique is applied in two stages to provide the best page ranking algorithm to provide personalized recommendations to its users.

In the first stage, experimentations to determine the member preference between the two ranking algorithms is carried out. Unlike the A/B testing where the two groups of viewers are exposed to the two ranking algorithms, Netflix makes use of interleaving to blend the rankings of algorithm A and B.

Netflix provides its users with enriched content based on this interleaving technique that is highly sensitive towards ranking the algorithm quality.

Importance of Context Awareness in Recommendations

Contextual Awareness is one of the key elements in personalizing recommendations for its users.

This not only improves the performance of the recommendation system but also prompts users to provide better feedback that would result in a quality recommendation. There are two categories of contextual classes:

Explicit

- Location
- Language
- Time of the Day
- Device

Inferred

- Binging Patterns
- Companion

In order to predict contexts, we make use of representation learning. It is a **deep learning technique** that performs feature engineering that discovers features without explicit programming.

Based on the time and periods of watching, Netflix bases its data on various parameters like Day, Week, Season and even longer periods like Olympics, FIFA, and elections.

It is the Right Time to Upgrade your Data Science Skills

For performing contextual predictions, Netflix treats recommendations as a sequence classification problem. It takes the input as a sequence of user-actions and performs predictions that output the next set of actions.

An example of a sequence problem is **Gru4Rec**. And in the case of contextual sequence prediction, the input consists of the contextual user actions as well as the current context of the user. This helps the recommendation engine solve the question:

“Based on all the historical actions that are taken by the user what is the most probable video that they will play right now?”

Proactive data Practioner:

Being proactive means structuring your data analytics to serve ongoing business objectives. It creates tangible, real actions that arise from data-driven insights. For example, you can use data analytics tools to create a business dashboard. The dashboard might display, on a dynamic, ever-changing basis, how your results are matching up to your KPIs.

A sales dashboard, for instance, could show sales results versus budget by product, region or sales person. You could create a dashboard that shows financial and operations data in real time so people can stay on top of cashflow.

Data: *Data* is a collection of raw facts and numbers used to analyze something or make decisions. Computer *data* is information in a form that can be processed by a computer.

Data Analysis: **Data analysis** is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making

Data Driven Decision Making: Data-driven decision-making (DDDM) is defined as using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives. When organizations realize the full value of their data, that means everyone—whether you're a business analyst, sales manager, or human resource specialist—is empowered to make better decisions with data, every day. However, this is not achieved by simply choosing the appropriate analytics technology to identify the next strategic opportunity.

Challenges in becoming a data driven organization

Due to technical disruption Data Science is becoming strong day by day an organization will have many benefits through data driven decision making in spite of the benefits there are many hurdles to become a n data driven organization

i. **Data Quality:**

Any data that is inaccurate, incomplete and out of date is of no use to the organization. Having poor quality data only increases your risk of making bad decisions, eventually leading to a loss in revenue.

Organizations should follow these 5 principles to cultivate good quality data.

1. Accuracy
2. Completeness
3. Consistency
4. Uniqueness
5. Timeliness

ii. **Tools for access,extraction,processing and analysis:**

Organizations need adequate storage and processing capabilities to access the data Another challenge is that the data may be available in human readable formats (for instance, in pdf files) rather than machine readable formats, So, it must be processed with the Optical text recognition softwares. So, even when the reliable data is available, the Organizations must possess the appropriate tools and technology for accessing the data and storing it and processing the data

iii Cost

Another challenge for becoming a data-driven organization is to find the balance between cost and access. Lot of human hours will be taken for data cleaning, analysis, management which is costing the organization. It incurs considerable amount in expanding the datasets. If the data is not first party data (data generated by the organization), then the data acquired from other parties might also incur cost

(iv) **Legacy Data**

The Organizations old information that is stored in an old /obsolete format is called as Legacy data function and to retain comprehensive business knowledge. Many Organizations sit on a big pile of old data (For example., Data stored in Excel 5.0 from 1990's) or use outdated systems to manage valuable information. Though these legacy data are still valuable to the Organization, its poor usability and inefficiency poses high maintenance costs. So, these legacy files need to be converted to new file formats to make them accessible, readable and usable in the future.

(v) **Organizational Politics**

Organization politics affects what data is shared with in the organization, within the departments or between the departments. This human element can create hurdles as it affects how and what data is shared to whom. The data formats are critical in sharing and matching across the datasets and easiness of processing this data depends on the decisions of humans on data formats.

(vi) **Data Silos**

A Data Silo is the collection of data or information held by one group that is isolated from and inaccessible to other groups of the enterprise. It can result from many factors including culture and competition between the departments that causes those employees to keep data from each other, rather than working together. Finance, administration, HR, and other departments need different information to do their work, and so do they collect. But the problem of data silos arises when the data is made inaccessible to other parts of the enterprise. For example, when accounting can't access current data from operations. Data Silos limit the view of data, threaten the data integrity, wastes resources and discourage collaborative work. As the quantity and diversity of data continue to grow too.

(vii) **Linking the Data**

To take appropriate decisions, the datasets from different sources must be combined, not an easy task. Data sets are owned and managed by various groups (Organizations, Internet service providers, vendors). Every group has their own interest in sharing the Linking data across data sets is a challenge.

(viii) **Strong Leadership**

The success of a data-driven organization critically depends on strong leadership. The leader to be aware of all the challenges and limitations of data sets, its cost and quality. The right resource teams have to be assigned in this task of moving towards data.

(ix) **Knowledge of Data Scientists**:

The challenge is that many organizations are not aware of how they could be using intelligent software to use data-driven insights to increase efficiency and revenue. Lack of training the intelligent software and usage of statistical techniques and inability to understand the quality data and its sources, and inability to draw the inferences poses challenges to the organizations become a data-driven organization.

ANALYZING DATA PRACTICES IN ORGANIZATIONS

Data is becoming the core corporate asset. Data Sources and the amount of data is growing, so ability to utilize this data and turn it into knowledge. The Companies which are seeing it as source of competitive advantage are transforming their culture to data driven.

Data Governance is the initiative that takes to create and enforce a set of Rules and policies regarding to the data management and plays an important role in building a data-driven culture. Effective data governance leads to improvement in data quality, decrease in data management cost as required. Data Governance ensures that the data inconsistent, trustworthy and Data.the process of Data Governance involves Data Steward.

Data Steward is a role within an organization responsible for utilizing an organization's data governance processes to make sure of fitness of data elements. They ensure high-quality data is easily accessible in a consistent manner. Data stewards share some responsibilities with data custodians.

Data Custodians are responsible for the safe custody, transport, storage of the data and implementation of the business rules.

Structured Data and Unstructured Data:

What is structured data?

So, structured data is the type of data that is well-organized and accurately formatted. This data exists in the format of relational databases (RDBMSs), meaning the information is stored in tables with rows and columns that are connected.

For analytical purposes, you can use data warehouses. DWs are central data storages used by companies for data analysis and reporting.

There is a special programming language used for handling relational databases and warehouses called SQL, which stands for Structured Query Language and was developed back in the 1970s by IBM.

What is unstructured data?

It makes sense that if the definition of structured data implies a neat organization of components in a predetermined manner, the definition of unstructured data will be the opposite. The pieces of such data aren't structured in a pre-defined way, meaning data is stored in its native formats.

One of the ways to manage unstructured data is to opt for non-relational databases, also known as NoSQL.

If there's a need to keep data in its raw native formats for further analysis, storage repositories called data lakes will be the way to go. A data lake is a storage repository or system meant to store huge volumes of data in its natural/raw formats.

Unstructured data examples. There is a wide array of forms that make up unstructured data such as email, text files, social media posts, video, images, audio, sensor data, and so on.

	Structured data	Unstructured data
What is it?	Data that fits in a predefined data model or schema.	Data without an underlying model to discern attributes.
Basic example	An Excel table.	A collection of video files.
Best for	An associated collection of discrete, short, non-continuous numerical and text values.	An associated collection of data, objects, or files where the attributes change or are unknown.

Storage types	Relational databases, graph databases, spatial databases, OLAP cubes, and more.	File systems, DAM systems, CMSs, version control systems, and more.
Biggest benefit	Easier to organize, clean, search, and analyze.	Can analyze data that can't be easily shaped into structured data.
Biggest challenge	All data must fit in the prescribed data model.	Can be difficult to analyze.
Main analysis technique	SQL queries.	Varies.

How data benefits to organization:

Data provides us with actual facts and metrics which makes decision making trustworthy. Here are the benefits of Data to any Organization

1. **The Decisions are taken more confidently:**

Earlier Decision making made on launching a new product or give discount on product, train employee, acquiring new companies based on intuition or gut feeling but by basing on data collection and analysis of data the decision is taken more confidently

2. Competitive advantage:

As the decisions taken on proper analysis of data the company would be a bit ahead of competitors by detecting the opportunities. of competitors by detecting the opportunities, and the threats than our competitor companies.

3. Less organization politics:

AS decisions are supported with strong data, statistics and software, there will be less scope of politics based on ego levels of the employee involved in decision making therefore morale of employees improves.

4. Timely Decisions:

With proper tools to access, process and analyze the data, a lot of time will be savedand decisions are taken at the right time.

5. Proactive Decision Making:

Proactive data-driven decision-making can help businesses to grow in several ways. For example, it can help companies to identify new market opportunities, optimize their marketing campaigns, and improve their product development process, enabling them to seize first-mover advantages.

6.Increase in the Team spirit:

With data-driven insights, you can set realistic and relevant goals for your team, identify and address gaps and issues in their skills, knowledge, and behaviors, provide timely and constructive feedback and recognition, adjust your leadership style and communication methods to match their preferences and expectations

Data Science Use Cases that are Changing the World

Earlier we saw many **data science applications**. Today we will see the diverse data science use cases. We will take examples of social media, e-commerce, transportation, and healthcare to demonstrate some of the important data science use cases in contemporary industries.

1. Facebook – Using Data to Revolutionize Social Networking & Advertising

Facebook is a social-media leader of the world today. With millions of users around the world, Facebook utilizes a large scale quantitative research through data science to gain insights about the social interactions of the people.

Facebook has become a hub of innovation where it has been using advanced techniques in data science to study user behavior and gain insights to improve their product. Facebook makes use of advanced technology in data science called **deep learning**.

Using deep learning, Facebook makes use of facial recognition and text analysis. In facial recognition, Facebook uses powerful neural networks to classify faces in the photographs. It uses its own text understanding engine called “DeepText” to understand user sentences.

It also uses Deep Text to understand people’s interest and aligning photographs with texts.

However, more than being a social media platform, Facebook is more of an advertisement corporation. It uses deep learning for targeted advertising. Using this, it decides what kind of advertisements the users should view.

It uses the insights gained from the data to cluster users based on their preferences and provides them with the advertisements that appeal to them.

Now, let's have a look at another data science use case – Amazon

2. Amazon – Transforming E-commerce with Data Science

Since its inception, Amazon has been working hard to make itself a customer-centric platform. Amazon heavily relies on **predictive analytics** to increase customer satisfaction. It does so through a personalized recommendation system.

This recommendation system is a hybrid type that also involves collaborative filtering which is comprehensive in nature. Amazon analyzes the historical purchases of the user to recommend more products.

This also comes through the suggestions that are drawn from the other users who use similar products or provide similar ratings.

Amazon has an anticipatory shipping model that uses big data for predicting the products that are most likely to be purchased by its users. It analyzes the pattern of your purchases and sends products to your nearest warehouse which you may utilize in the future.

Amazon also optimizes the prices on its websites by keeping in mind various parameters like the user activity, order history, prices offered by the competitors, product availability, etc. Using this method, Amazon provides discounts on popular items and earns profits on less popular items.

Another area where every e-commerce platform is addressing is **Fraud Detection**. Amazon has its own novel ways and algorithms to detect fraud sellers and fraudulent purchases.

Other than online platforms, Amazon has been optimizing the packaging of products in warehouses and increasing the efficiency of packaging lines through the data collected from the workers.

3. Uber – Using Data to Make Rides Better

Next in data science use cases is Uber. Uber is a popular smartphone application that allows you to book a cab. Uber makes extensive use of **Big Data**. After all, Uber has to maintain a large database of drivers, customers, and several other records.

It is therefore, rooted in Big Data and makes use of it to derive insights and provide the best services to its users. Uber shares the big data principle with crowdsourcing. That is, registered drivers in the area can help anyone who wants to go somewhere.

As mentioned above, Uber contains a database of drivers. Therefore, whenever you hail for a cab, Uber matches your profile with the most suitable driver. What differentiates Uber from other cab companies is that Uber charges you based on the time it takes to cover the distance and not the distance itself.

It calculates the time taken through various algorithms that also make use of data related to traffic density and weather conditions.

Uber makes the best use of data science to calculate its surge pricing. When there are less drivers available to more riders, the price of the ride goes up. This happens only during the scarcity of drivers in any given area.

However, if the demand for Uber rides is less, then Uber charges a lower rate. This dynamic pricing is rooted in Big Data and makes excellent usage of data science to calculate the fares based on the parameters.

4. Bank of America – Using Data to Leverage Customer Experience

10 years ago, Bank of America was one of the first financial companies to provide mobile banking to its customers. Recently, BoA launched Erica which is their first virtual financial assistant. It is considered as the world’s finest innovation in finance domain.

Currently, Erica is serving as a customer advisor to more than 45 million users around the world. Erica also makes use of Speech Recognition to take customer inputs, which is a technological advancement in the field of Data Science.

Furthermore, several other banks like BoA are making use of **Data Science and predictive analytics**. Using data science, banking industries are able to detect frauds in payments and

customer information. It also prevents frauds regarding insurances, credit cards, and accounting.

In order to minimize the losses, a bank needs to detect fraud sooner. In order to carry this out, banks employ data scientists to use their quantitative knowledge where they apply algorithms like association, clustering, forecasting, and classification.

Risk modeling is another important area that is supervised by the banks to regulate financial activities. Using Machine Learning, banks are able to minimize risk modeling.

Through analytical solutions, banks can make data-driven decisions that are based on transparency and risk analysis. Furthermore, Bank of America detected the high-risk accounts using this technology of big data.

Various banks like BoA are understanding their customers through an intelligent customer segmentation approach. Through various data-mining techniques, banks are able to segment their customers in the high-value and low-value segments.

There are various techniques that a data scientist makes use of such as clustering, logistic regression, decision trees to help the banks to understand the Customer Lifetime Value (CLV) and take group them in the appropriate segments.

Data Science Case Study – How Netflix Used Data Science to Improve its Recommendation System?

Do you remember the last movie you watched on Netflix? I don't want to know the name; just think about it- after watching the movie, were you recommended of similar movies? How does Netflix know what you'd like? The secret here is Data Science.

Netflix uses Data Science to cater relevant and interesting recommendations to you. So, today, in this article, we will discuss the same. Let's start exploring Data Science at Netflix with a basic introduction to Netflix.

Data Science at Netflix

Netflix initially started as a DVD rental service in 1998. It mostly relied on a third party postal services to deliver its DVDs to the users. This resulted in heavy losses which they soon mitigated with the **introduction of their online streaming service** in 2007.

In order to make this happen, Netflix invested in a lot of algorithms to provide a flawless movie experience to its users. One of such algorithms is the **recommendation system** that is used by Netflix to provide suggestions to the users.

A recommendation system understands the needs of the users and provides suggestions of the various cinematographic products.

What is a Recommendation System?

A recommendation system is a platform that provides its users with various contents based on their preferences and likings. A recommendation system takes the information about the user as an input.

This information can be in the form of the past usage of product or the ratings that were provided to the product. It then processes this information to predict how much the user would rate or prefer the product. A recommendation system makes use of a variety of **machine learning algorithms**.

Another important role that a recommendation system plays today is to search for similarity between different products. In the case of Netflix, the recommendation system searches for movies that are similar to the ones you have watched or have liked previously.

This is an important method for scenarios that involve cold start. In cold start, the company does not have much of the user data available to generate recommendations.

Therefore, based on the movies that are watched, Netflix provides recommendations of the films that share a degree of similarity. There are two main types of Recommendation Systems –

1. Content-based recommendation systems

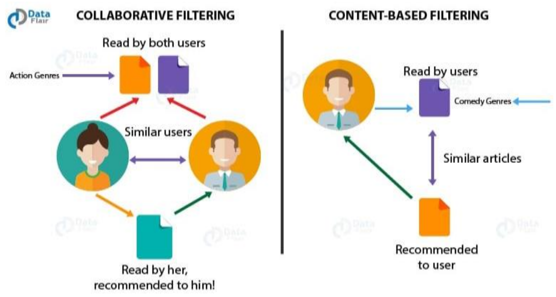
In a content-based recommendation system, the background knowledge of the products and customer information are taken into consideration. Based on the content that you have viewed on Netflix, it provides you with similar suggestions.

For example, if you have watched a film that has a sci-fi genre, the content-based recommendation system will provide you with suggestions for similar films that have the same genre.

2. Collaborative filtering recommendation systems

Unlike the content based filtering that provided recommendations of similar products, Collaborative Filtering provides recommendations based on the similar profiles of its users. One key advantage of collaborative filtering is that it is independent of the product knowledge.

Rather, it relies on the users with a basic assumption that what the users liked in the past will also like in the future. **For example**, if a person A watches crime, sci-fi and thriller genres and B watches sci-fi, thriller and action genres then A will also like action and B will like crime genre.



There is also a third type of recommendation system that combines both Content and Collaborative techniques. This form of recommendation system is known as **Hybrid Recommendation System**. Netflix makes the primary of use Hybrid Recommendation System for suggesting content to its users.

Most recommended – Marks & Spencer using Big Data to Analyze Customer Behaviour

Follow DataFlair on WhatsApp & Stay updated with latest technology trends.

How Netflix Solved its Recommendation Problem with Data Science

Back in 2006 when Netflix wanted to tap into the streaming market, it started off with a competition for movie rating prediction. It provided a prize of $ 1 million to whoever increased the accuracy of their then existing platform ‘Cinematch’ by 10%.

At the end of competition, the BellKor team presented their solution that increased the accuracy of prediction by 10.06%. With over 200 work hours and an ensemble of 107 algorithms provided them with this result.

Their final model gave an RMSE of 0.8712. For their solution, they made use of K-nearest neighbor algorithm for post-processing of the data.

Then they implemented a factorization model which is popularly known as **Singular Value Decomposition (SVD)** for providing an optimal dimensional embedding to its users.

They also made use of **Restricted Boltzmann Machines (RBM)** for enhancing the capability of the collaborative filtering model. These two algorithms in the ensemble, SVD and RBM provided them with the best results. A linear combination of these two algorithms reduced the RMSE to 0.88.

However, even after reduction of RMSE and increase in accuracy, Netflix suffered from two major challenges – Firstly, the data that provided during the competition comprised of 100 million movie ratings, as opposed to more than 5 billion ratings that Netflix constituted of.

Furthermore, the algorithms were static, meaning that they only dealt with historical data and did not take into account the dynamicity of users adding reviews in real-time. After Netflix overcame these challenges, it made the winning algorithms a part of its recommendation system.

Using Interleaving to Improve Personalization

Netflix uses Ranking Algorithms to provide a ranked list of movies and TV Shows that appeal the most to its users. However, with the presence of various ranking algorithms, it is often difficult to accommodate all of them and test their performance simultaneously.

While the traditional A/B testing on a reduced set of algorithms could not identify the best algorithms with smaller sample size and also consumed a lot of time, Netflix decided to innovate its algorithmic process.

In order to speed up its experimentation process of its ranking algorithms, Netflix implemented the interleaving technique that allowed it to identify best algorithms. This technique is applied in two stages to provide the best page ranking algorithm to provide personalized recommendations to its users.

In the first stage, experimentations to determine the member preference between the two ranking algorithms is carried out. Unlike the A/B testing where the two groups of viewers are exposed to the two ranking algorithms, Netflix makes use of interleaving to blend the rankings of algorithm A and B.

Netflix provides its users with enriched content based on this interleaving technique that is highly sensitive towards ranking the algorithm quality.

Importance of Context Awareness in Recommendations

Contextual Awareness is one of the key elements in personalizing recommendations for its users.

This not only improves the performance of the recommendation system but also prompts users to provide better feedback that would result in a quality recommendation. There are two categories of contextual classes:

Explicit

- Location
- Language
- Time of the Day
- Device

Inferred

- Binging Patterns
- Companion

In order to predict contexts, we make use of representation learning. It is a **deep learning technique** that performs feature engineering that discovers features without explicit programming.

Based on the time and periods of watching, Netflix bases its data on various parameters like Day, Week, Season and even longer periods like Olympics, FIFA, and elections.

It is the Right Time to Upgrade your Data Science Skills

For performing contextual predictions, Netflix treats recommendations as a sequence classification problem. It takes the input as a sequence of user-actions and performs predictions that output the next set of actions.

An example of a sequence problem is **Gru4Rec**. And in the case of contextual sequence prediction, the input consists of the contextual user actions as well as the current context of the user. This helps the recommendation engine solve the question:

“Based on all the historical actions that are taken by the user what is the most probable video that they will play right now?”

Proactive data Practioner:

Being proactive means structuring your data analytics to serve ongoing business objectives. It creates tangible, real actions that arise from data-driven insights. For example, you can use data analytics tools to create a business dashboard. The dashboard might display, on a dynamic, ever-changing basis, how your results are matching up to your KPIs.

A sales dashboard, for instance, could show sales results versus budget by product, region or sales person. You could create a dashboard that shows financial and operations data in real time so people can stay on top of cashflow.

Data: *Data* is a collection of raw facts and numbers used to analyze something or make decisions. Computer *data* is information in a form that can be processed by a computer.

Data Analysis: **Data analysis** is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making

Data Driven Decision Making: Data-driven decision-making (DDDM) is defined as using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives. When organizations realize the full value of their data, that means everyone—whether you're a business analyst, sales manager, or human resource specialist—is empowered to make better decisions with data, every day. However, this is not achieved by simply choosing the appropriate analytics technology to identify the next strategic opportunity.

Challenges in becoming a data driven organization

Due to technical disruption Data Science is becoming strong day by day an organization will have many benefits through data driven decision making in spite of the benefits there are many hurdles to become a n data driven organization

i. **Data Quality:**

Any data that is inaccurate, incomplete and out of date is of no use to the organization. Having poor quality data only increases your risk of making bad decisions, eventually leading to a loss in revenue.

Organizations should follow these 5 principles to cultivate good quality data.

1. Accuracy
2. Completeness
3. Consistency
4. Uniqueness
5. Timeliness

ii. **Tools for access,extraction,processing and analysis:**

Organizations need adequate storage and processing capabilities to access the data Another challenge is that the data may be available in human readable formats (for instance, in pdf files) rather than machine readable formats, So, it must be processed with the Optical text recognition softwares. So, even when the reliable data is available, the Organizations must possess the appropriate tools and technology for accessing the data and storing it and processing the data

iii Cost

Another challenge for becoming a data-driven organization is to find the balance between cost and access. Lot of human hours will be taken for data cleaning, analysis, management which is costing the organization. It incurs considerable amount in expanding the datasets. If the data is not first party data (data generated by the organization), then the data acquired from other parties might also incur cost

(iv) **Legacy Data**

The Organizations old information that is stored in an old /obsolete format is called as Legacy data function and to retain comprehensive business knowledge. Many Organizations sit on a big pile of old data (For example., Data stored in Excel 5.0 from 1990's) or use outdated systems to manage valuable information. Though these legacy data are still valuable to the Organization, its poor usability and inefficiency poses high maintenance costs. So, these legacy files need to be converted to new file formats to make them accessible, readable and usable in the future.

(v) **Organizational Politics**

Organization politics affects what data is shared with in the organization, within the departments or between the departments. This human element can create hurdles as it affects how and what data is shared to whom. The data formats are critical in sharing and matching across the datasets and easiness of processing this data depends on the decisions of humans on data formats.

(vi) **Data Silos**

A Data Silo is the collection of data or information held by one group that is isolated from and inaccessible to other groups of the enterprise. It can result from many factors including culture and competition between the departments that causes those employees to keep data from each other, rather than working together. Finance, administration, HR, and other departments need different information to do their work, and so do they collect. But the problem of data silos arises when the data is made inaccessible to other parts of the enterprise. For example, when accounting can't access current data from operations. Data Silos limit the view of data, threaten the data integrity, wastes resources and discourage collaborative work. As the quantity and diversity of data continue to grow too.

(vii) **Linking the Data**

To take appropriate decisions, the datasets from different sources must be combined, not an easy task. Data sets are owned and managed by various groups (Organizations, Internet service providers, vendors). Every group has their own interest in sharing the Linking data across data sets is a challenge.

(viii) **Strong Leadership**

The success of a data-driven organization critically depends on strong leadership. The leader to be aware of all the challenges and limitations of data sets, its cost and quality. The right resource teams have to be assigned in this task of moving towards data.

(ix) **Knowledge of Data Scientists**:

The challenge is that many organizations are not aware of how they could be using intelligent software to use data-driven insights to increase efficiency and revenue. Lack of training the intelligent software and usage of statistical techniques and inability to understand the quality data and its sources, and inability to draw the inferences poses challenges to the organizations become a data-driven organization.

ANALYZING DATA PRACTICES IN ORGANIZATIONS

Data is becoming the core corporate asset. Data Sources and the amount of data is growing, so ability to utilize this data and turn it into knowledge. The Companies which are seeing it as source of competitive advantage are transforming their culture to data driven.

Data Governance is the initiative that takes to create and enforce a set of Rules and policies regarding to the data management and plays an important role in building a data-driven culture. Effective data governance leads to improvement in data quality, decrease in data management cost as required. Data Governance ensures that the data inconsistent, trustworthy and Data.the process of Data Governance involves Data Steward.

Data Steward is a role within an organization responsible for utilizing an organization's data governance processes to make sure of fitness of data elements. They ensure high-quality data is easily accessible in a consistent manner. Data stewards share some responsibilities with data custodians.

Data Custodians are responsible for the safe custody, transport, storage of the data and implementation of the business rules.

Structured Data and Unstructured Data:

What is structured data?

So, structured data is the type of data that is well-organized and accurately formatted. This data exists in the format of relational databases (RDBMSs), meaning the information is stored in tables with rows and columns that are connected.

For analytical purposes, you can use data warehouses. DWs are central data storages used by companies for data analysis and reporting.

There is a special programming language used for handling relational databases and warehouses called SQL, which stands for Structured Query Language and was developed back in the 1970s by IBM.

What is unstructured data?

It makes sense that if the definition of structured data implies a neat organization of components in a predetermined manner, the definition of unstructured data will be the opposite. The pieces of such data aren't structured in a pre-defined way, meaning data is stored in its native formats.

One of the ways to manage unstructured data is to opt for non-relational databases, also known as NoSQL.

If there's a need to keep data in its raw native formats for further analysis, storage repositories called data lakes will be the way to go. A data lake is a storage repository or system meant to store huge volumes of data in its natural/raw formats.

Unstructured data examples. There is a wide array of forms that make up unstructured data such as email, text files, social media posts, video, images, audio, sensor data, and so on.

	Structured data	Unstructured data
What is it?	Data that fits in a predefined data model or schema.	Data without an underlying model to discern attributes.
Basic example	An Excel table.	A collection of video files.
Best for	An associated collection of discrete, short, non-continuous numerical and text values.	An associated collection of data, objects, or files where the attributes change or are unknown.

Storage types	Relational databases, graph databases, spatial databases, OLAP cubes, and more.	File systems, DAM systems, CMSs, version control systems, and more.
Biggest benefit	Easier to organize, clean, search, and analyze.	Can analyze data that can’t be easily shaped into structured data.
Biggest challenge	All data must fit in the prescribed data model.	Can be difficult to analyze.
Main analysis technique	SQL queries.	Varies.

How data benefits to organization:

Data provides us with actual facts and metrics which makes decision making trustworthy. Here are the benefits of Data to any Organization

1. **The Decisions are taken more confidently:**

Earlier Decision making made on launching a new product or give discount on product, train employee, acquiring new companies based on intuition or gut feeling but by basing on data collection and analysis of data the decision is taken more confidently

2. Competitive advantage:

As the decisions taken on proper analysis of data the company would be a bit ahead of competitors by detecting the opportunities. of competitors by detecting the opportunities, and the threats than our competitor companies.

3. Less organization politics:

AS decisions are supported with strong data, statistics and software, there will be less scope of politics based on ego levels of the employee involved in decision making therefore morale of employees improves.

4. Timely Decisions:

With proper tools to access, process and analyze the data, a lot of time will be savedand decisions are taken at the right time.

5. Proactive Decision Making:

Proactive data-driven decision-making can help businesses to grow in several ways. For example, it can help companies to identify new market opportunities, optimize their marketing campaigns, and improve their product development process, enabling them to seize first-mover advantages.

6.Increase in the Team spirit:

With data-driven insights, you can set realistic and relevant goals for your team, identify and address gaps and issues in their skills, knowledge, and behaviors, provide timely and constructive feedback and recognition, adjust your leadership style and communication methods to match their preferences and expectations

Data Science Use Cases that are Changing the World

Earlier we saw many **data science applications**. Today we will see the diverse data science use cases. We will take examples of social media, e-commerce, transportation, and healthcare to demonstrate some of the important data science use cases in contemporary industries.

1. Facebook – Using Data to Revolutionize Social Networking & Advertising

Facebook is a social-media leader of the world today. With millions of users around the world, Facebook utilizes a large scale quantitative research through data science to gain insights about the social interactions of the people.

Facebook has become a hub of innovation where it has been using advanced techniques in data science to study user behavior and gain insights to improve their product. Facebook makes use of advanced technology in data science called **deep learning**.

Using deep learning, Facebook makes use of facial recognition and text analysis. In facial recognition, Facebook uses powerful neural networks to classify faces in the photographs. It uses its own text understanding engine called “DeepText” to understand user sentences.

It also uses Deep Text to understand people’s interest and aligning photographs with texts.

However, more than being a social media platform, Facebook is more of an advertisement corporation. It uses deep learning for targeted advertising. Using this, it decides what kind of advertisements the users should view.

It uses the insights gained from the data to cluster users based on their preferences and provides them with the advertisements that appeal to them.

Now, let’s have a look at another data science use case – Amazon

2. Amazon – Transforming E-commerce with Data Science

Since its inception, Amazon has been working hard to make itself a customer-centric platform. Amazon heavily relies on **predictive analytics** to increase customer satisfaction. It does so through a personalized recommendation system.

This recommendation system is a hybrid type that also involves collaborative filtering which is comprehensive in nature. Amazon analyzes the historical purchases of the user to recommend more products.

This also comes through the suggestions that are drawn from the other users who use similar products or provide similar ratings.

Amazon has an anticipatory shipping model that uses big data for predicting the products that are most likely to be purchased by its users. It analyzes the pattern of your purchases and sends products to your nearest warehouse which you may utilize in the future.

Amazon also optimizes the prices on its websites by keeping in mind various parameters like the user activity, order history, prices offered by the competitors, product availability, etc. Using this method, Amazon provides discounts on popular items and earns profits on less popular items.

Another area where every e-commerce platform is addressing is **Fraud Detection**. Amazon has its own novel ways and algorithms to detect fraud sellers and fraudulent purchases.

Other than online platforms, Amazon has been optimizing the packaging of products in warehouses and increasing the efficiency of packaging lines through the data collected from the workers.

3. Uber – Using Data to Make Rides Better

Next in data science use cases is Uber. Uber is a popular smartphone application that allows you to book a cab. Uber makes extensive use of **Big Data**. After all, Uber has to maintain a large database of drivers, customers, and several other records.

It is therefore, rooted in Big Data and makes use of it to derive insights and provide the best services to its users. Uber shares the big data principle with crowdsourcing. That is, registered drivers in the area can help anyone who wants to go somewhere.

As mentioned above, Uber contains a database of drivers. Therefore, whenever you hail for a cab, Uber matches your profile with the most suitable driver. What differentiates Uber from other cab companies is that Uber charges you based on the time it takes to cover the distance and not the distance itself.

It calculates the time taken through various algorithms that also make use of data related to traffic density and weather conditions.

Uber makes the best use of data science to calculate its surge pricing. When there are less drivers available to more riders, the price of the ride goes up. This happens only during the scarcity of drivers in any given area.

However, if the demand for Uber rides is less, then Uber charges a lower rate. This dynamic pricing is rooted in Big Data and makes excellent usage of data science to calculate the fares based on the parameters.

4. Bank of America – Using Data to Leverage Customer Experience

10 years ago, Bank of America was one of the first financial companies to provide mobile banking to its customers. Recently, BoA launched Erica which is their first virtual financial assistant. It is considered as the world's finest innovation in finance domain.

Currently, Erica is serving as a customer advisor to more than 45 million users around the world. Erica also makes use of Speech Recognition to take customer inputs, which is a technological advancement in the field of Data Science.

Furthermore, several other banks like BoA are making use of **Data Science and predictive analytics**. Using data science, banking industries are able to detect frauds in payments and

customer information. It also prevents frauds regarding insurances, credit cards, and accounting.

In order to minimize the losses, a bank needs to detect fraud sooner. In order to carry this out, banks employ data scientists to use their quantitative knowledge where they apply algorithms like association, clustering, forecasting, and classification.

Risk modeling is another important area that is supervised by the banks to regulate financial activities. Using Machine Learning, banks are able to minimize risk modeling.

Through analytical solutions, banks can make data-driven decisions that are based on transparency and risk analysis. Furthermore, Bank of America detected the high-risk accounts using this technology of big data.

Various banks like BoA are understanding their customers through an intelligent customer segmentation approach. Through various data-mining techniques, banks are able to segment their customers in the high-value and low-value segments.

There are various techniques that a data scientist makes use of such as clustering, logistic regression, decision trees to help the banks to understand the Customer Lifetime Value (CLV) and take group them in the appropriate segments.

Data Science Case Study – How Netflix Used Data Science to Improve its Recommendation System?

Do you remember the last movie you watched on Netflix? I don’t want to know the name; just think about it- after watching the movie, were you recommended of similar movies? How does Netflix know what you’d like? The secret here is Data Science.

Netflix uses Data Science to cater relevant and interesting recommendations to you. So, today, in this article, we will discuss the same. Let’s start exploring Data Science at Netflix with a basic introduction to Netflix.

Data Science at Netflix

Netflix initially started as a DVD rental service in 1998. It mostly relied on a third party postal services to deliver its DVDs to the users. This resulted in heavy losses which they soon mitigated with the **introduction of their online streaming service** in 2007.

In order to make this happen, Netflix invested in a lot of algorithms to provide a flawless movie experience to its users. One of such algorithms is the **recommendation system** that is used by Netflix to provide suggestions to the users.

A recommendation system understands the needs of the users and provides suggestions of the various cinematographic products.

What is a Recommendation System?

A recommendation system is a platform that provides its users with various contents based on their preferences and likings. A recommendation system takes the information about the user as an input.

This information can be in the form of the past usage of product or the ratings that were provided to the product. It then processes this information to predict how much the user would rate or prefer the product. A recommendation system makes use of a variety of **machine learning algorithms**.

Another important role that a recommendation system plays today is to search for similarity between different products. In the case of Netflix, the recommendation system searches for movies that are similar to the ones you have watched or have liked previously.

This is an important method for scenarios that involve cold start. In cold start, the company does not have much of the user data available to generate recommendations.

Therefore, based on the movies that are watched, Netflix provides recommendations of the films that share a degree of similarity. There are two main types of Recommendation Systems –

1. Content-based recommendation systems

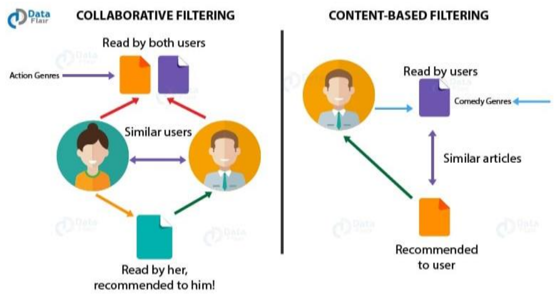
In a content-based recommendation system, the background knowledge of the products and customer information are taken into consideration. Based on the content that you have viewed on Netflix, it provides you with similar suggestions.

For example, if you have watched a film that has a sci-fi genre, the content-based recommendation system will provide you with suggestions for similar films that have the same genre.

2. Collaborative filtering recommendation systems

Unlike the content based filtering that provided recommendations of similar products, Collaborative Filtering provides recommendations based on the similar profiles of its users. One key advantage of collaborative filtering is that it is independent of the product knowledge.

Rather, it relies on the users with a basic assumption that what the users liked in the past will also like in the future. **For example**, if a person A watches crime, sci-fi and thriller genres and B watches sci-fi, thriller and action genres then A will also like action and B will like crime genre.



There is also a third type of recommendation system that combines both Content and Collaborative techniques. This form of recommendation system is known as **Hybrid Recommendation System**. Netflix makes the primary of use Hybrid Recommendation System for suggesting content to its users.

Most recommended – Marks & Spencer using Big Data to Analyze Customer Behaviour

Follow DataFlair on WhatsApp & Stay updated with latest technology trends.

How Netflix Solved its Recommendation Problem with Data Science

Back in 2006 when Netflix wanted to tap into the streaming market, it started off with a competition for movie rating prediction. It provided a prize of $ 1 million to whoever increased the accuracy of their then existing platform ‘Cinematch’ by 10%.

At the end of competition, the BellKor team presented their solution that increased the accuracy of prediction by 10.06%. With over 200 work hours and an ensemble of 107 algorithms provided them with this result.

Their final model gave an RMSE of 0.8712. For their solution, they made use of K-nearest neighbor algorithm for post-processing of the data.

Then they implemented a factorization model which is popularly known as **Singular Value Decomposition (SVD)** for providing an optimal dimensional embedding to its users.

They also made use of **Restricted Boltzmann Machines (RBM)** for enhancing the capability of the collaborative filtering model. These two algorithms in the ensemble, SVD and RBM provided them with the best results. A linear combination of these two algorithms reduced the RMSE to 0.88.

However, even after reduction of RMSE and increase in accuracy, Netflix suffered from two major challenges – Firstly, the data that provided during the competition comprised of 100 million movie ratings, as opposed to more than 5 billion ratings that Netflix constituted of.

Furthermore, the algorithms were static, meaning that they only dealt with historical data and did not take into account the dynamicity of users adding reviews in real-time. After Netflix overcame these challenges, it made the winning algorithms a part of its recommendation system.

Using Interleaving to Improve Personalization

Netflix uses Ranking Algorithms to provide a ranked list of movies and TV Shows that appeal the most to its users. However, with the presence of various ranking algorithms, it is often difficult to accommodate all of them and test their performance simultaneously.

While the traditional A/B testing on a reduced set of algorithms could not identify the best algorithms with smaller sample size and also consumed a lot of time, Netflix decided to innovate its algorithmic process.

In order to speed up its experimentation process of its ranking algorithms, Netflix implemented the interleaving technique that allowed it to identify best algorithms. This technique is applied in two stages to provide the best page ranking algorithm to provide personalized recommendations to its users.

In the first stage, experimentations to determine the member preference between the two ranking algorithms is carried out. Unlike the A/B testing where the two groups of viewers are exposed to the two ranking algorithms, Netflix makes use of interleaving to blend the rankings of algorithm A and B.

Netflix provides its users with enriched content based on this interleaving technique that is highly sensitive towards ranking the algorithm quality.

Importance of Context Awareness in Recommendations

Contextual Awareness is one of the key elements in personalizing recommendations for its users.

This not only improves the performance of the recommendation system but also prompts users to provide better feedback that would result in a quality recommendation. There are two categories of contextual classes:

Explicit

- Location
- Language
- Time of the Day
- Device

Inferred

- Binging Patterns
- Companion

In order to predict contexts, we make use of representation learning. It is a **deep learning technique** that performs feature engineering that discovers features without explicit programming.

Based on the time and periods of watching, Netflix bases its data on various parameters like Day, Week, Season and even longer periods like Olympics, FIFA, and elections.

It is the Right Time to Upgrade your Data Science Skills

For performing contextual predictions, Netflix treats recommendations as a sequence classification problem. It takes the input as a sequence of user-actions and performs predictions that output the next set of actions.

An example of a sequence problem is **Gru4Rec**. And in the case of contextual sequence prediction, the input consists of the contextual user actions as well as the current context of the user. This helps the recommendation engine solve the question:

“Based on all the historical actions that are taken by the user what is the most probable video that they will play right now?”

Proactive data Practioner:

Being proactive means structuring your data analytics to serve ongoing business objectives. It creates tangible, real actions that arise from data-driven insights. For example, you can use data analytics tools to create a business dashboard. The dashboard might display, on a dynamic, ever-changing basis, how your results are matching up to your KPIs.

A sales dashboard, for instance, could show sales results versus budget by product, region or sales person. You could create a dashboard that shows financial and operations data in real time so people can stay on top of cashflow.

UNIT III: BUSINESS ANALYTICS ECOSYSTEM: Relational Databases: Nature of relational databases - Purpose of the SQL language - Key aspects of ACID - Meaning of ETL - Not Only SQL: Big data and other data storage tools - Interacting with MongoDB - Document stores and graph stores - Big Data: Key functions of big data technologies - Utility of Hadoop - Purpose of MapReduce - Statistical Tool, Machine Learning, and Data Visualization: Tools for statistical analysis - Python and R - Purpose of machine learning - Visualization tools

A relational database management system (RDBMS): is a software program and capabilities that enable to create, update, administer and otherwise interact with a relational database. RDBMS store data in the form of tables, with most commercial relational database management systems using Structured Query Language (SQL) to access the database. However, since SQL was invented after the initial development of the relational model, it is not necessary for RDBMS use.

The RDBMS is the most popular database system among organizations across the world. It provides a dependable method of storing and retrieving large amounts of data while offering a combination of system performance and ease of implementation.

Nature of relational databases

The data tables used in a relational database store information about related objects. Each row holds a record with a unique id entifier -- known as a key -- and each column contains the attributes of the data. Each record assigns a value to each feature, making relationships between data points easy to identify.

Purpose of SQL:

SQL is a programming language used to communicate with relational databases which allows you to query the database in a variety of ways, using English-like statements. It's used on websites for back-end data storage and data processing solutions (for example, Facebook uses SQL).

Essentially, SQL provides CRUD functionality for databases. What does CRUD stand for?

- Create
- Read
- Update
- Delete.

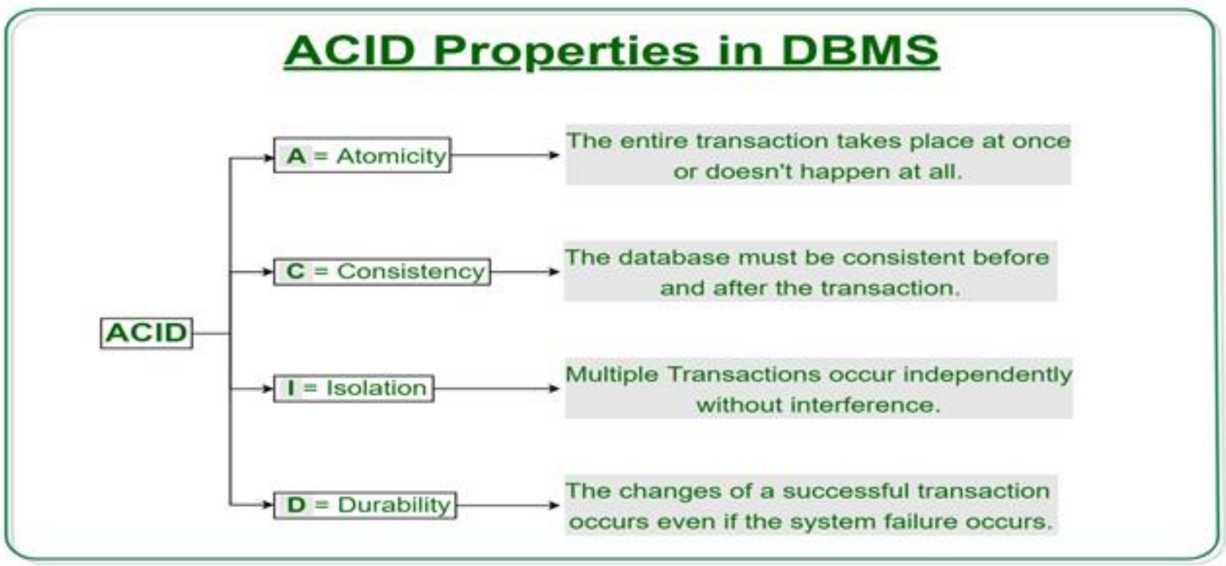
Types of SQL commands

The language can be broken down into four types of SQL commands – DDL, DML, DQL and DCL. Let's look at each of these sections.

- **DDL** (data definition language) – this is used to create and modify database objects like tables, users, and indices.
- **DML** (data manipulation language) – this is used to delete, add, and modify data within databases.
- **DCL** (data control language) – this is used to control access to any data within a database.
- **DQL** (data query language) – this is used to perform queries on the data and find information, and is composed of COMMAND statements only.

ACID properties

Transactions access data using read and write operations. In order to maintain consistency in a database, before and after the transaction, certain properties are followed. These are called **ACID** properties.


ACID Properties in DBMS
ACID
A = Atomicity
The entire transaction takes place at once or doesn't happen at all.
C = Consistency
The database must be consistent before and after the transaction.
I = Isolation
Multiple Transactions occur independently without interference.
D = Durability
The changes of a successful transaction occurs even if the system failure occurs.

Normalization:-Normalization is a process for evaluating and correcting table's structures to minimize data redundancy thereby avoiding the occurring of data anomalies.

NeedforNormalization

- To minimize data redundancy
- To avoid data anomalies resulting during insert, update or delete operations.

Normalform:-Normalization works through a series of stages called normal form.

Following are the different normal forms:-

- ➢ First normal form(1NF)
- ➢ Second normal form(2NF)
- ➢ Third normal form(3NF)

- **First Normal Form(1NF) :** A relation will be 1NF if it contains an atomic value.
- It states that an attribute of a table cannot hold multiple values. It must hold only single-valued attribute.
- First normal form disallows the multi-valued attribute, composite attribute, and their combinations.
- **Example:** Relation EMPLOYEE is not in 1NF because of multi-valued attribute EMP_PHONE.

- **EMPLOYEE table:**

EMP_ID	EMP_NAME	EMP_PHONE	EMP_STATE
14	John	7272826385, 9064738238	UP
20	Harry	8574783832	Bihar
12	Sam	7390372389, 8589830302	Punjab

- The decomposition of the EMPLOYEE table into 1NF has been shown below:

EMP_ID	EMP_NAME	EMP_PHONE	EMP_STATE
14	John	7272826385	UP
14	John	9064738238	UP
20	Harry	8574783832	Bihar
12	Sam	7390372389	Punjab
12	Sam	8589830302	Punjab

- **Second Normal Form(2NF):** In the 2NF, relational must be in 1NF.
- In the second normal form, all non-key attributes are fully functional dependent on the primary key
- **Example:** Let's assume, a school can store the data of teachers and the subjects they teach. In a school, a teacher can teach more than one subject.
- **TEACHER table**

TEACHER_ID	SUBJECT	TEACHER_AGE
25	Chemistry	30
25	Biology	30

47	English	35
83	Math	38
83	Computer	38

- In the given table, non-prime attribute TEACHER_AGE is dependent on TEACHER_ID which is a proper subset of a candidate key. That's why it violates the rule for 2NF.
- To convert the given table into 2NF, we decompose it into two tables:
- **TEACHER_DETAIL table:**

TEACHER_ID	TEACHER_AGE
25	30
47	35
83	38

- **TEACHER_SUBJECT table:**

TEACHER_ID	SUBJECT
25	Chemistry
25	Biology
47	English
83	Math
83	Computer

Third Normal Form(3NF):-

- A relation will be in 3NF if it is in 2NF and does not contain any transitive partial dependency.
- 3NF is used to reduce data duplication. It is also used to achieve the data integrity.
- If there is no transitive dependency for non-prime attributes, then the relation must be in third normal form.

A relation is in third normal form if it holds at least one of the following conditions for every non-trivial function dependency X → Y.

1. X is a super key.
2. Y is a prime attribute, i.e., each element of Y is part of some candidate key.
3. **Example:**
4. **EMPLOYEE_DETAIL table:**

EMP_ID	EMP_NAME	EMP_ZIP	EMP_STATE	EMP_CITY
222	Harry	201010	UP	Noida
333	Stephan	02228	US	Boston
444	Lan	60007	US	Chicago
555	Katharine	06389	UK	Norwich
666	John	462007	MP	Bhopal

Super key in the table above:

a. {EMP_ID}, {EMP_ID, EMP_NAME}, {EMP_ID, EMP_NAME, EMP_ZIP}....so on

5. **Candidate key:** {EMP_ID}
6. **Non-prime attributes:** In the given table, all attributes except EMP_ID are non-prime.
7. Here, EMP_STATE & EMP_CITY dependent on EMP_ZIP and EMP_ZIP dependent on EMP_ID. The non-prime attributes (EMP_STATE, EMP_CITY) transitively dependent on super key(EMP_ID). It violates the rule of the third normal form.
8. That's why we need to move the EMP_CITY and EMP_STATE to the new <EMPLOYEE_ZIP> table, with EMP_ZIP as a Primary key.
9. **EMPLOYEE table:**

EMP_ID	EMP_NAME	EMP_ZIP
222	Harry	201010

333	Stephan	02228
444	Lan	60007
555	Katharine	06389
666	John	462007

EMPLOYEE_ZIP table:

EMP_ZIP	EMP_STATE	EMP_CITY
201010	UP	Noida
02228	US	Boston
60007	US	Chicago
06389	UK	Norwich
462007	MP	Bhopal

ETL: ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.

ETL provides the foundation for data analytics and machine learning workstreams. Through a series of business rules, ETL cleanses and organizes data in a way which addresses specific business intelligence needs, like monthly reporting, but it can also tackle more advanced analytics, which can improve back-end processes or end user experiences.

How ETL works

The easiest way to understand how ETL works is to understand what happens in each step of the process.

Extract

During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of data sources, which can be structured or unstructured. Those sources include but are not limited to:

- SQL or NoSQL servers
- CRM and ERP systems
- Flat files
- Email
- Web pages

Transform

In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

- Filtering, cleansing, de-duplicating, validating, and authenticating the data.
- Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.
- Conducting audits to ensure data quality and compliance
- Removing, encrypting, or protecting data governed by industry or governmental regulators
- Formatting the data into tables or joined tables to match the schema of the target data warehouse.

Load

In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous and batch-driven. Typically, ETL takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest.

NOT ONLY SQL:

NoSQL is a type of database management system (DBMS) that is designed to handle and store large volumes of unstructured and semi-structured data. Unlike traditional relational databases that use tables with pre-defined schemas to store data, NoSQL databases use flexible data models that can adapt to changes in data structures and are capable of scaling horizontally to handle growing amounts of data.

Key Features of NoSQL:

1. **Dynamic schema:** NoSQL databases do not have a fixed schema and can accommodate changing data structures without the need for migrations or schema alterations.
2. **Horizontal scalability:** NoSQL databases are designed to scale out by adding more nodes to a database cluster, making them well-suited for handling large amounts of data and high levels of traffic.
3. **Document-based:** Some NoSQL databases, such as MongoDB, use a document-based data model, where data is stored in a **scales**semi-structured format, such as JSON or BSON.
4. **Key-value-based:** Other NoSQL databases, such as Redis, use a key-value data model, where data is stored as a collection of key-value pairs.
5. **Column-based:** Some NoSQL databases, such as Cassandra, use a column-based data model, where data is organized into columns instead of rows.
6. **Distributed and high availability:** NoSQL databases are often designed to be highly available and to automatically handle node failures and data replication across multiple nodes in a database cluster.
7. **Flexibility:** NoSQL databases allow developers to store and retrieve data in a flexible and dynamic manner, with support for multiple data types and changing data structures.
8. **Performance:** NoSQL databases are optimized for high performance and can handle a high volume of reads and writes, making them suitable for big data and real-time applications.

MongoDB is a document database. It stores data in a type of JSON format called BSON.

If you are unfamiliar with JSON, check out our JSON tutorial.

A record in MongoDB is a document, which is a data structure composed of key value pairs similar to the structure of JSON objects.

Start learning MongoDB now »

A MongoDB Document
Records in a MongoDB database are called documents, and the field values may include numbers, strings, booleans, arrays, or even nested documents.

Example Document

```
{
    title: "Post Title 1",
    body: "Body of post.",
    category: "News",
    likes: 1,
    tags: ["news", "events"],
    date: Date()
}
```

Example

Find all documents that have a category of "news".

```
db.posts.find( {category: "News"} )
```

Big data: Big data refers to data that is so large, fast or complex that it’s difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around for a long time. But the concept of big data gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three V’s:

Volume. Organizations collect data from a variety of sources, including transactions, smart (IoT) devices, industrial equipment, videos, images, audio, social media and more. In the past, storing all that data would have been too costly – but cheaper storage using data lakes, Hadoop and the cloud have eased the burden.

Velocity. With the growth in the Internet of Things, data streams into businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.

Apache Hadoop is an open source, Java-based software platform that manages data processing and storage for big data applications. The platform works by distributing Hadoop big data and analytics jobs across nodes in a computing cluster, breaking them down into smaller workloads that can be run in parallel. Some key benefits of Hadoop are

- **Scalability** - Unlike traditional systems that limit data storage, Hadoop is scalable as it operates in a distributed environment. This allowed data architects to build early data lakes on Hadoop.
- **Resilience** - The Hadoop Distributed File System (HDFS) is fundamentally resilient. Data stored on any node of a Hadoop cluster is also replicated on other nodes of the cluster to prepare for the possibility of hardware or software failures.
- **Flexibility** - Differing from relational database management systems, when working with Hadoop, you can store data in any format, including semi-structured or unstructured formats. Hadoop enables businesses to easily access new data sources and tap into different types of data.

challenges of Hadoop:

- **Complexity** - Hadoop is a low-level, Java-based framework that can be overly complex and difficult for end-users to work with. Hadoop architectures can also require significant expertise and resources to set up, maintain, and upgrade.
- **Performance** - Hadoop uses frequent reads and writes to disk to perform computations, which is time-consuming and inefficient compared to frameworks that aim to store and process data in memory as much as possible, like Apache Spark.

- **Long-term viability** - The impact of this decision on Hadoop users is still to be seen. This growing collection of concerns paired with the accelerated need to digitize has encouraged many companies to re-evaluate their relationship with Hadoop

What is MapReduce?

MapReduce is a Java-based, distributed execution framework within the Apache Hadoop Ecosystem. It takes away the complexity of distributed programming by exposing two processing steps that developers implement: 1) Map and 2) Reduce.

A MapReduce system is usually composed of three steps (even though it's generalized as the combination of Map and Reduce operations/functions). The MapReduce operations are:

- **Map:** The input data is first split into smaller blocks. The Hadoop framework then decides how many mappers to use, based on the size of the data to be processed and the memory block available on each mapper server. Each block is then assigned to a mapper for processing. Each 'worker' node applies the map function to the local data, and writes the output to temporary storage. The primary (master) node ensures that only a single copy of the redundant input data is processed.
- **Shuffle, combine and partition:** worker nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node. As an optional process the combiner (a reducer) can run individually on each mapper server to reduce the data on each mapper even further making reducing the data footprint and shuffling and sorting easier. Partition (not optional) is the process that decides how the data has to be presented to the reducer and also assigns it to a particular reducer.
- **Reduce:** A reducer cannot start while a mapper is still in progress. Worker nodes process each group of <key,value> pairs output data, in parallel to produce <key,value> pairs as output. All the map output values that have the same key are assigned to a single reducer, which then aggregates the values for that key. Unlike the map function which is mandatory to filter and sort the initial data, the reduce function is optional.

What is Python?

What is Python?

Python is an interpreted, high-level object-oriented programming language. It comes with built-in data structures, dynamic typing(a process wherein type checks are done during the runtime), and binding(mapping of different objects with one another), which makes it a top language used for the development of applications. Python syntaxes are simple, easy to read, and easy to learn, hence **learning Python** is an excellent choice for beginners and experienced programmers alike.

The Python interpreter and libraries are free for distribution. . It includes libraries like Scikit, Keras, Tensorflow, Matplotlib, NumPy, Pandas, etc., that provide sophisticated functionalities. The addition of Jupyter Notebook, a web application to share the code live, makes the data science explanations smooth.

Advantages of Python

- **Versatility:** The language is one of the most versatile ones. It is neat, uncomplicated to use, and well-structured. Python's flexibility makes exploratory data analysis hassle-free. Python is object-oriented, but it makes a transition to functional features allowing itself into different paradigms of programming.
- **Open Source:** Python can be downloaded easily. It has one of the most active supporting forums, and anyone can contribute to improving the libraries and their functionalities.
- **Libraries:** Python has many libraries that are necessary to carry out major data science-related functions.
- **Productivity:** Its integration and control capabilities enhance and save a lot of time.
  Embeddable: Python codes are embeddable. Python codes can be integrated with other programming languages like C++.

Disadvantages of Python

- **Speed:** Python is an interpreted language and thus is relatively slower than other programming languages.
- **Mobile environment:** Python is not suitable for Android and iOS environments. Developers claim it to be weak language in such an environment. However, it can be used with additional efforts.
- **Memory consumption:** Python consumes a significant amount of RAM. The process gets slower when more objects need to be accessed.
- **Database Access Layers:** Python's database access layers are underdeveloped in comparison to Java Database Connectivity(JDBC), and Open Database Connectivity(ODBC), making it a less used database connectivity.
- **Threading:** Threading or the flow of multiple functions at the same time is a downside in Python due to its Global Interpreter Lock(GIL).

What is R?

R is a programming language for statistical analysis or computing and graphics. R comes with a wide range of statistical techniques such as linear modelling, non-linear modelling, statistical tests, clustering, etc. One of R's strengths is the ease at which a plot can be produced, including the mathematical notations and formulas.

R is available as free software. It compiles and works on UNIX, Windows, and macOS. R allows programmers to add additional functionality by defining user-specific functions. For intensive tasks, the user can link the C and C++ codes during the runtime. R can be extended with other languages like C++ using the packages.

Advantages of R Programming

- **Open Source:** R is an open-source language and is free to download and use. One can also contribute by optimizing its source code.
- **Platform independent:** R is platform-independent and can work on all operating systems like UNIX, Windows, and Mac.
- **Data Wrangling:** Through its packages like readr and dplyr, R has the capability of converting messy code into a structured one.

- **Plots and Graphs:** Through ggplot and plotly, R creates attractive graphs with notations and formulas.
- **Package Availability:** R has numerous packages dedicated to the development of machine learning, data analysis, and statistical projects.

Disadvantages of R

- **Memory:** R consumes more memory as all the objects get stored in the physical memory. Over time, as the program has more data, the process slows down.
- **Security:** R lacks basic security that makes it practically difficult to embed in web applications.
- **Difficult to learn:** Unlike Python, R is a complicated language and is difficult for a beginner to learn.
- **Slow Runtime:** R is a slow processing language. In comparison to other languages such as MATLAB and Python, it takes more time to give an output.
- **Data Handling:** Data handling in R is tedious as it requires all the data to be in one place. It is not ideal for Big Data. However, it does have an integration that makes handling slightly easier.

Python vs. R: Full Comparison

Python	**R Programming**
Python is a general-purpose language that is used for the deployment and development of various projects. Python has all the tools required to bring a project into the production environment.	R is a statistical language used for the analysis and visual representation of data.
Python is better suitable for machine learning, deep learning, and large-scale web applications.	R is suitable for statistical learning having powerful libraries for data experiment and exploration.
Python has a lot of libraries. However, it can be complex to understand all of them.	R has fewer libraries compared to Python and is easy to know.

Python can be used for various purposes like building a graphical user interface, develop games, etc., despite being an object-oriented language.	Along with object-oriented programming, R can also be used to develop music.
Python has a simple syntax and is easy to learn.	R has a relatively complex syntax and the learning curve is not straightforward.
Python’s statistical packages are less powerful.	R’s statistical packages are highly powerful.
Python is mainly used when the data analysis needs to be integrated with web applications.	R is generally used when the data analysis task requires standalone computation(analysis) and processing.
Python can be used to build applications from scratch.	R can be used to simplify complex mathematical problems.
There are many Python IDEs available to choose from, a few of them are Jupyter Notebook, Spyder, Pycharm, etc.	A few IDEs for the R language are RStudio, StatET, etc.
Python is more popular and has a vast user base. Primary users of Python include developers and programmers.	R is less popular among users. Its users include scientists and Research and development who frequently rely on data analysis.

UNIT IV: DATA LIFECYCLE MANAGEMENT: Data Life Cycle: Identify the stages in the data life cycle - Data in the organization: Distinguish between ways that data enters the organization - Identify the forms data takes as it is stored and used within the organization.

Data Life Cycle:

A **data lifecycle** illustrates how **data** (in all its various forms and derivatives, including data points, datasets, databases, data files, visualizations, and code) conceptually flows through its lifecycle of usefulness. While data lifecycles are helpful frameworks to discuss appropriate actions taken at different stages, it's important to remember that for most data the path is not linear and some actions may not occur at all.

Below are the 7 phases of Data Life Cycle every business must be informed:

1. Data Capture

The first experience that an item of data must have is to pass within the firewalls of the enterprise. This is Data Capture, which can be defined as the act of creating data values that do not yet exist and have never existed within the enterprise.

There are three main ways that data can be captured, and these are very important:

- **Data Acquisition:** the ingestion of already existing data that has been produced by an organization outside the enterprise
- **Data Entry:** the creation of new data values for the enterprise by human operators or devices that generate data for the enterprise
- **Signal Reception:** the capture of data created by devices, typically important in control systems, but becoming more important for information systems with the Internet of Things

2. Data Maintenance

Data Maintenance is about processing the data without deriving any value from it for the enterprise. It often involves tasks such as movement, integration, cleansing, enrichment, changed data capture, and familiar extract-transform-load processes.

3. Data Synthesis

This is comparatively new and perhaps still not a very common phase in the Data Life Cycle. It can be defined as the creation of data values via inductive logic, using other data as input.

It is the arena of analytics that uses modeling, such as is found in risk modeling, actuarial modeling, and modeling for investment decisions. Derivation by deductive logic is not part of this – that occurs in Data Maintenance. An example of deductive logic is Net Sales = Gross Sales – Taxes. If I know Gross Sales and Taxes, and I know the simple equation just outlined, then I can calculate Net Sales.

Inductive logic requires some expert experience, judgment, and opinion as a part of the logic, e.g., the way in which credit scores are created.

4. Data Usage

Data usage has special Data Governance challenges. One of them is whether it is legal to use the data in the ways that businesspeople want. This is referred to as "permitted use of data." There may be regulatory or contractual constraints on how data may actually be used, and part of the role of Data Governance is to ensure that these constraints are observed.

5. Data Publication

In being used, it is possible that our single data value may be sent outside of the enterprise. This is Data Publication, which can be defined as the sending of data to a location outside of the enterprise.

6. Data Archival

Data Archival is the copying of data to an environment where it is stored in case it is needed again in an active production environment and the removal of this data from all active production environments.

A data archive is simply a place where data is stored but where no maintenance, usage, or publication occurs. If necessary, the data can be restored to an environment where one or more of these occur.

7. Data Purging

We now come to the actual end of life of our single data value. Data Purging is the removal of every copy of a data item from the enterprise.

Ideally, this will be done from an archive. A Data Governance challenge in this phase of the data life cycle is proving that the purge has actually been done properly.

Data mining: *is* the process of analyzing enormous amounts of information and datasets, extracting (or “mining”) useful intelligence to help organizations solve problems, predict trends, mitigate risks, and find new opportunities. Data mining is like actual mining because, in both cases, the miners are sifting through mountains of material to find valuable resources and elements.

Data Mining Steps

When asking "what is data mining," let's break it down into the steps data scientists and analysts take when tackling a data mining project.

1. Understand Business

What is the company's current situation, the project's objectives, and what defines success?

2. Understand the Data

Figure out what kind of data is needed to solve the issue, and then collect it from the proper sources.

3. Prepare the Data

Resolve data quality problems like duplicate, missing, or corrupted data, then prepare the data in a format suitable to resolve the business problem.

4. Model the Data

Employ algorithms to ascertain data patterns. Data scientists create, test, and evaluate the model.

5. Evaluate the Data

Decide whether and how effective the results delivered by a particular model will help meet the business goal or remedy the problem.

6. Deploy the Solution

Give the results of the project to the people in charge of making decisions.

Data Mining Applications

Data mining is a useful and versatile tool for today's competitive businesses. Here are some data mining examples, showing a broad range of applications.

Banks

Data mining helps banks work with credit ratings and anti-fraud systems, analyzing customer financial data, purchasing transactions, and card transactions. Data mining also helps banks better understand their customers' online habits and preferences, which helps when designing a new marketing campaign.

Healthcare

Data mining helps doctors create more accurate diagnoses by bringing together every patient's medical history, physical examination results, medications, and treatment patterns. Mining also helps fight fraud and waste and bring about a more cost-effective health resource management strategy.

Marketing

Data mining helps bring together data on age, gender, tastes, income level, location, and spending habits to create more effective personalized loyalty campaigns. Data marketing can even predict which customers will more likely unsubscribe to a mailing list or other related service. Armed with that information, companies can take steps to retain those customers before they get the chance to leave!

Retail

Retail stores and supermarkets can use purchasing patterns to narrow down product associations and determine which items should be stocked in the store and where they should go.

Data in the organization: Data refers to the collection of facts in an organization that includes observations, experiences, events, experiments. Data that can enter into organization maybe in the format of letters, words, images, audio, video, tweets, likes, dislikes, numericals and so on. Data mining process. The data is still called data which is high level abstraction of data. Data can be classified as structured, semi structured and unstructured data. Unstructured or semi structured data consists of different combinations of text, pictures, audios, videos etc. On the other hand, structured data is one particular type of data,can be numeric or categorical data.

Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. A process called matching is done, to draw out the similarities or relations between the data and then they are grouped accordingly.

Example: sexuality is categorical data, as a person can be straight, homosexual, heterosexual, etc. and they are grouped together depending on the common characteristics possessed by them.

There are two subtypes of categorical data namely: Nominal data and Ordinal data.

- **Nominal data** – this is also called naming data. This is a type that names or labels the data and its characteristics are similar to a noun.

Example: person’s name, gender, school name.

- **Ordinal data** – this includes data or elements of data that is ranked, ordered or used on a rating scale. You can count and order ordinal data but it doesn’t allow you to measure it.

Example: seminar attendants are asked to rate their seminar experience on a scale of 1-5. Against each number, there will be options that will rate their satisfaction like “very good, good, average, bad, and very bad”.

Numerical data: refers to the data that is in the form of numbers, and not in any language or descriptive form. Often referred to as quantitative data, numerical data is collected in number form and stands different from any form of number data types due to its ability to be statistically and arithmetically calculated.

It doesn’t involve any natural language description and is quantitative in nature and it is used to measure quantities like a person’s height, age, IQ, etc.

It also has two subtypes known as Discrete data and Continuous data.

- **Discrete data** – Discrete data is used to represent countable items. It can take both numerical and categorical forms and group them into a list. This list can be finite or infinite too.

Discrete data basically takes countable numbers like 1, 2, 3, 4, 5, and so on. In the case of infinity, these numbers will keep going on.

Example: counting sugar cubes from a jar is finite countable. But counting sugar cubes from all over the world is infinite countable.

- **Continuous data** – As the name says, this form has data in the form of intervals. Or simply said ranges. Continuous numerical data represent measurements and their intervals fall on a number line. Hence, it doesn't involve taking counts of the items.

Example: in a school exam, students who scored 80%-100% come under distinction, 60%-80% have first-class and below 60% are second class.

Continuous data is further divided into two categories: Interval and Ratio.

- **Interval data** – interval data type refers to data that can be measured only along a scale at equal distances from each other. The numerical values in this data type can only undergo add and subtract operations. **Example**: body temperature can be measured in degree Celsius and degree Fahrenheit and neither of them can be 0.
- **Ratio data** – unlike interval data, ratio data has zero points. Being similar to interval data, zero point is the only difference they have. **Example**: in the body temperature, the zero point temperature can be measured in Kelvin.

Most of the unstructured data types like text messages, audios, videos, voice recordings have converted into some other categorical or numerical data before they can be processed. In an organization. So that data has to be converted into compatible form before taking it further. Analysis. If incompatible data or. Fitted to the data models, the predictions may not be accurate.

Types of data in an organization:

1. Transactional Data

Transactional data is the information that an agreement, exchange, or transfer that occurs between organizations or individuals. It is considered to be a special category of data as transactions have legal and commercial significance.

Examples of transactional data:

- Invoices: A bill of services or products.
- Trades: A trade that occurs in the market, eg. Stock market.
- Purchases: Customer purchases
- Returns: A record comprising of customer returned items and accepted by the seller.
- Payments: A payment towards debt or purchase.
- Credits: Fund added to an account, for example, an e-commerce site refunding the amount of a returned item.

- Debits: Funds removed from an account, for instance, a bank customer withdraws money from his/ her account.

The list goes on with payroll, asset sales, interest, contracts, etc.

2. Master Data

Information that an organization can agree upon is known as master data. Often organizations have, unlike information sources that may duplicate similar data with little agreement on standard definitions. Master data represents an opportunity to govern and manage data as a single source of reference.

Examples of Metadata:

- Product data: A catalog with product specifications and information.
- Transactions: Purchases and stock trades
- Tickets: To track problems and customer interactions.
- Analytical Data: Data that supports decision making.

3. Customer Data

All the information associated with a customer is known as customer data. Customer data is information that is essential for the core business processes and decision-making tools.

Examples of customer data:

- Customer Record: The record comprises the name of the customer and customer id.
- Accounts: A customer having various accounts, such as departments that make separate purchases.
- Contacts: Corporate customers ought to have several contacts that include various levels and functions within the customer organization, such as engineering staff or purchasing managers.
- Services: List of current services that the customer is using, such as configurations.

Customer service tickets, feedback, locations, payment methods, offers, and quotes are a few more examples.

4. Machine Data

Data generated by machines without any human involvement is known as machine data. It is one of the important categories as machines create far more data when compared to humans.

Examples of machine data:

- Calculations: Data calculated from the other data. E.g., based on the market data, an algorithm calculates the risk estimate for an investment.

- Sensors: Devices that detect physical phenomena such as sound and light and turn into streams of data.
- Predictions: Artificial intelligence and algorithms that attempt to predict the future.
- Automation: Automated tasks that create data like controls, events, or commands.

5. Reference Data

Data used to structure and constrain other data is known as reference data. **It** is stable information with a known set of values that rarely change.

Examples of reference data:

- Science: List of known stars
- Geographical Locations: List of valid states for a country.
- Computing: List of standard computing values like HTTP status codes.
- Markets: List of currently valid stock tickets for a market.

Data storage in an organization: Data in an organization generally stores in different forms such as.

Data Repository is defined as a place to store data, develop data and organize data in a logical manner. In a. Repositories have specific reach, research domain, easy access and usage particular format of storing data. There are two natures of repositories are there

- Generalized: consists of two or more variety data sets.
- Specific: Focused on a particular research discipline, such as management.

Data Paper: Is the publication journal. Which describes the datasets and its types rather than publishing research investigation. Data papers Searchable meta documents published by scholars. Accessible online datasets. Data papers helps to share data. And reuse issues. Data paper consists of Interested in publications only? But others are mixed with published data papers along with multiple articles.

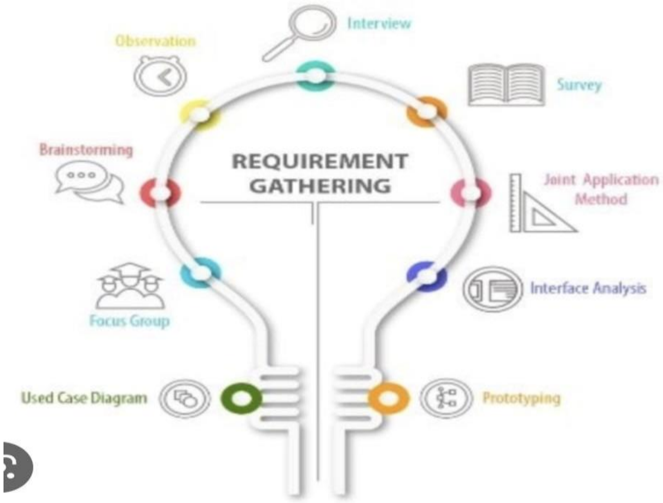
Database: Is an organization of synthesized data for easy access and maintenance. Database consists of data records

Data documents: at Sometimes index such as, Data Citation Index, Science Citation Index and traditional bibliography databases. Data can be gathered from primary or secondary sources. They're collecting the data. The data can be. Classify. And analyze with minimum three angles of analysis for better understanding of research objective. stop dictating

UNIT V: REQUIREMENTS GATHERING: Requirements gathering process: Analyze why requirement gathering process is critical to proper analysis - 3 V’s of data: Distinguish between the ways data is consumed (the three V’s of data) - Customer journey map: Understand how requirement gathering fits with the development of a customer journey map - Distinguish between the stages of the customer journey map

Requirements gathering techniques varies from Project to Project. Some requirements gathering techniques may prove highly beneficial for you in one project but may not be as productive in the other project or for some other company. Therefore the usefulness of a technique is determined by its need and the kind of advantages it offers in a particular project. There are 10 essential requirements gathering techniques that you must be aware of in order to manage projects in a better way and run your business successfully.

1. Brainstorming
2. Document Analysis
3. Focus Group
4. Interface Analysis
5. Interview
6. Observation
7. Prototyping
8. Requirements Workshop
9. Reverse Engineering
10. Survey



Brainstorming

Brainstorming can be utilised in requirements gathering to gather a good number of ideas from a group of people. Usually brainstorming is used in identifying all possible solutions to problems and simplifies the detail of opportunities. It casts a broad net, determining various discreet possibilities. Prioritisation of such possibilities is vital to locate needles in a haystack.

Document Analysis

Document Analysis is an important gathering technique. Evaluating the documentation of a present system can assist when making AS-IS process documents and also when driving the gap analysis for scoping of the migration projects. In today's world, you will also be determining the requirements that drove making of an existing system- a beginning point for documenting all current requirements. Chunks of information are mostly buried in present documents that assist you in putting questions as a part of validating the requirement completeness.

Focus Group

A focus group is a gathering of people who are customers or user representatives for a product to gain its feedback. The feedback can be collected about opportunities, needs, and problems to determine requirements or it can be collected to refine and validate the already elicited requirements. This type of market research is different from brainstorming in which it is a managed process with particular participants. There is a risk in following the crowd and some people think that focus groups are at best unproductive. One danger that we usually end up with is with least common denominator features.

Interface Analysis

Interface for any software product will either be human or machine. Integration with external devices and systems is another interface. The user-centric design approaches are quite effective to ensure that you make usable software. Interface analysis - analysing the touch points with another external system- is vital to ensure that you do not overlook requirements that are not instantly visible to the users.

Interview

Interviews of users and stakeholders are important in creating wonderful software. Without knowing the expectations and goal of the stakeholders and users, you are highly unlikely to satiate them. You also have to understand the perspective of every interviewee, in order to properly address and weigh their inputs. Like a good reporter, listening is a quality that assists an excellent analyst to gain better value through an interview as compared to an average analyst.

Observation

The observation covers the study of users in its natural habitat. By watching users, a process flow, pain points, awkward steps and opportunities can be determined by an analyst for improvement. Observation can either be passive or active. Passive observation provides better feedback to refine requirements on the same hand active observation works best for obtaining an understanding over an existing business process. You can use any of these approaches to uncover the implicit requirements that are often overlooked.

Prototyping

Prototyping can be very helpful at gathering feedback. Low fidelity prototypes make a good listening tool. Many a times, people are not able to articulate a specific need in the abstract. They can swiftly review whether a design approach would satisfy the need. Prototypes are very effectively done with fast sketches of storyboards and interfaces. Prototypes in some situations are also used as official requirements.

Requirements Workshop

Popularly known as JAD or joint application design, these workshops can be efficient for gathering requirements. The requirements workshops are more organised and structured than a brainstorming session where the involved parties get together to document requirements. Creation of domain model artifacts like activity programs or static diagrams is one of the ways to capture the collaboration. A workshop with two analysts is more effective than one in which one works as a facilitator and the other scribes the work together.

Reverse Engineering

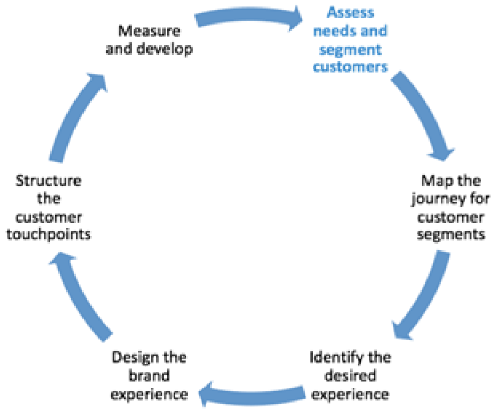
Is this a last resort or starting point? When a migration project does not have enough documentation of the current system, reverse engineering will determine what system does? It will not determine what thing went wrong with the system and what a system must do.

Survey

When gathering information from many people: too many to interview with time constraints and less budget: a questionnaire survey can be used. The survey insists the users to choose from the given options agree / disagree or rate something. Do not think that you can make a survey on your own but try to add meaningful insight in it. A well designed survey must give qualitative guidance for characterising the market. It should not be utilised for prioritising of requirements or features.

3 V's of data: covered in unit III

Customer Experience:


Assess needs and segment customers
Map the journey for customer segments
Identify the desired experience
Design the brand experience
Structure the customer touchpoints
Measure and develop

Customer Journey Map Definition

A customer journey map, also known as a customer experience map, is a visual representation that outlines the various steps and touchpoints a customer goes through when interacting with a company, product, or service.

Customer journey map is a tool used to understand and analyze the customer's experience, from the initial awareness or consideration of a product or service through the purchase and post-purchase stages.

It reveals customer actions, emotions, pain points and expectations along the customer journey. And it helps the business see things from the customer's perspective which in turn helps the business gain a deep understanding of the needs of the customer.

What are the Benefits of Using a Customer Journey Map?

There are many benefits to customer journey mapping. The customer journey map helps

- To enhance the customer experience. It helps businesses gain insights into customers' various touchpoints and interactions with the product or service.
- To reduce costs by identifying the areas the business should prioritize investing in and spending effort on. To innovate and differentiate by discovering the gaps between customer expectations and current customer experience, unmet customer needs, pain points, and opportunities.
- To improve customer satisfaction by identifying severe customer experience issues and eliminating them effectively.

- To increase customer loyalty by helping to build strong customer relationships by understanding their needs, preferences, and emotions.
- Data-driven decision-making based on gathered insights from customer research, feedback, and analytics.

What Are the Components of a Customer Journey Map?

A customer journey map typically includes the following components:

- Touchpoints: All the interactions and experiences a customer has with a company, including in-person, online, and mobile interactions.
- Customer personas: Representations of the target customer segments, including their demographics, behaviors, motivations, and pain points.
- Emotions: A visual representation of how the customer feels at different touchpoints during their journey.
- Channels: The ways in which a customer interacts with the company, such as website, phone, or in-person interactions.
- Data and insights: Customer behavior data and insights from surveys, analytics, or other sources.
- Pain points and opportunities: Identifications of areas where the customer experience can be improved, as well as opportunities for innovation and differentiation.
- Recommended actions: Specific recommendations for improving the customer experience, based on the journey map analysis.
- Alignment with company goals: A visual representation of how the customer journey aligns with the overall goals and strategy of the company.

Stages of customer journey map : Customer journey maps should include customer experiences, stakeholders, business goals and KPIs. Each customer journey map is based on five stages, which can be modified to suit any business model. Let’s dive into the details of each step:

Awareness

A customer’s journey with your company begins before visiting the facility. The customer becomes aware about the brand through word-of-mouth marketing and advertisements.

Business goal: Increase awareness

KPI: Maximum number of individuals reached

Evaluation and Consideration

After being introduced to your brand, the consumer will actively compare and evaluate alternate brands to choose the best available one. Next, the customer will plan to engage with the company based on the variety of products and services offered.

Business goal: Maximize customer engagement

KPI: Capture new customers

2. In-store Experience

Engagement

The customer makes a purchase or chooses a service with your company. They look for accessibility to products, easy return policies and short wait times in queues. and importantly, they want an impeccable customer service experience which answers queries at every stage.

Business goal: Increase sales

KPI: Shopping cart value

3. Post-sale

Retention

After using the service, the customer decides if the product satisfies their needs and meets expectations.

Business goal: Customer satisfaction services

KPI: Product ratings

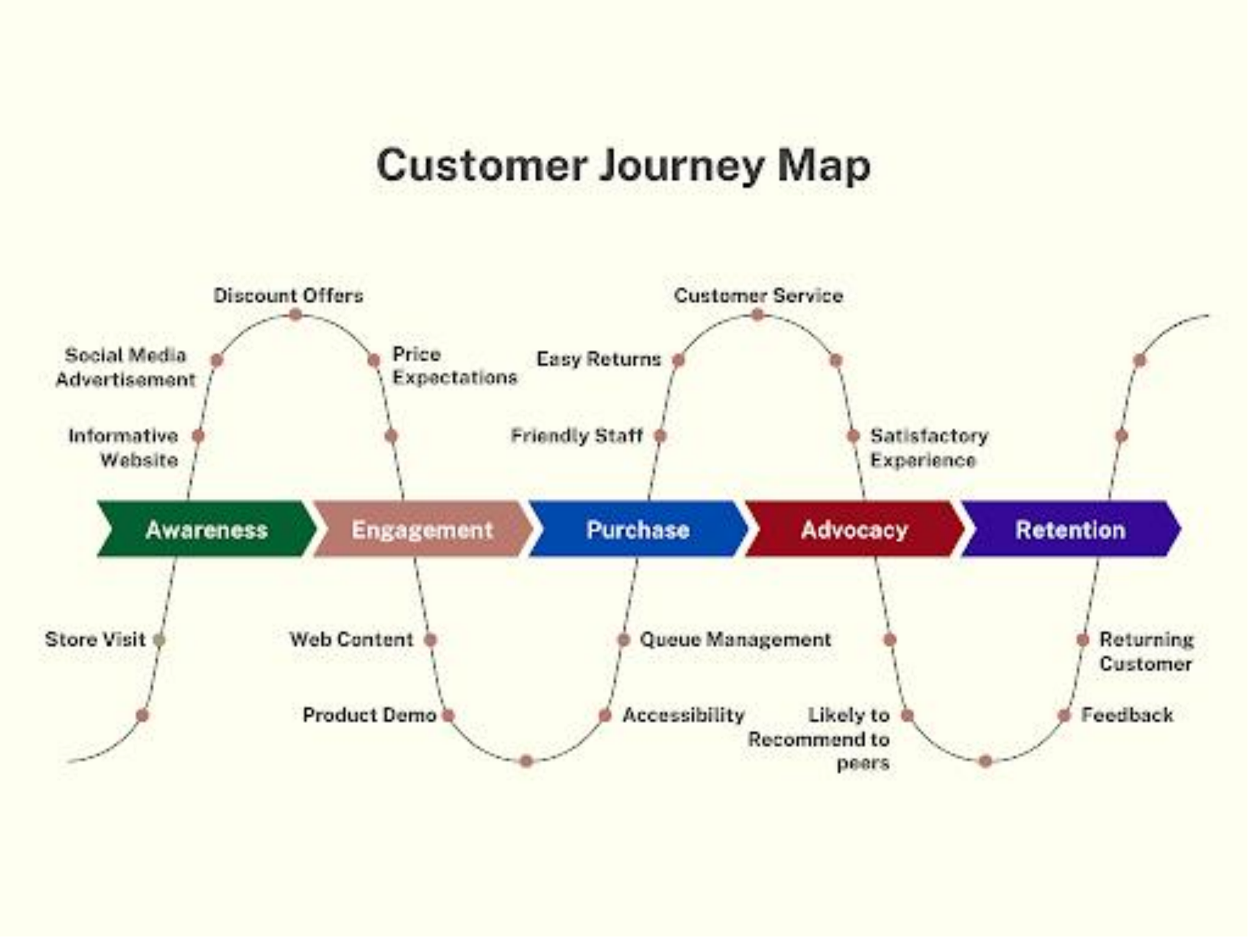
Advocacy

The last phase of the journey involves customers sharing experiences, be it negative or positive.

Business goal: Building long-term relations with customers

KPI: Cusomter retention rates

We have put together a sample of customer journey maps for a retail company. It highlights some interactions of consumers’ experiences with the business. Pair the stages of the journey maps with the template to design the best one for your company!


Customer Journey Map
Discount Offers
Customer Service
Social Media Advertisement
Price Expectations
Easy Returns
Informative Website
Friendly Staff
Satisfactory Experience
Awareness
Engagement
Purchase
Advocacy
Retention
Store Visit
Web Content
Queue Management
Returning Customer
Product Demo
Accessibility
Likely to Recommend to peers
Feedback