ANNAMACHARYA UNIVERSITY

New Boyanapalli, RAJAMPET, 516126, Andhra Pradesl1

Department of Computer science Engineering &(DS)

LECTURE NOTES

II B.Tech - II Semester

Introduction To Data science

(24AADI41T)

Acade1nic Year 2025-26

Prepared By:

Mr.P.Harikrishna
Assistant Professor,
Department of AI & DS,
Annamacharya University.

Title of the Course : Introduction to Data Science
Category : Professional Core
Year : II
Semester : II Semester
Course Code : 24AAID41T
Branches : AI&DS , CSE(DS)

Lecture Hours	Tutorial Hours	Practice Hours	Credits
3	-	-	3

Course Objectives: This course will be able to
1. Knowledge and expertise to become a data scientist.
2. Essential concepts of statistics and machine learning that are vital for data science.
3. Significance of exploratory data analysis (EDA) in data science.
4. Critically evaluate data visualizations presented on the dashboards
5. Suitability and limitations of tools and techniques related to data science process

Course Outcomes:
At the end of the course, the student will be able to
1. Understand the data science process and its components, including data cleansing and exploratory analysis.
2. Apply machine learning techniques for feature engineering and model selection using Python tools.
3. Analyze the NoSQL movement and the principles guiding NoSQL databases in handling big data.
4. Utilize graph databases and Python libraries for text mining and analytics in data science applications.
5. Develop data visualizations and interactive dashboards to present findings from data science projects.

Unit 1 Introduction to Data Science **10**

Introduction to Data science, benefits and uses, facets of data, data science process in brief, big data ecosystem, and data science

Data Science process**:** Overview, defining goals and creating project charter, retrieving data, cleansing, integrating, and transforming data

Unit 2 Handling large Data **10**

Applications of machine learning in Data science, role of ML in DS, Python tools like sklearn, modelling process for feature engineering, model selection, validation and prediction, types of ML, semi-supervised learning

Handling large data**:** problems and general techniques for handling large data, programming tips for dealing large data, case studies on DS projects for predicting malicious URLs, for building recommender systems.

Unit 3 NoSQL Movement for handling Bigdata **10**

NoSQL movement for handling Bigdata**:** Distributing data storage and processing with Hadoop framework, case study on risk assessment for loan sanctioning, ACID principle of relational databases, CAP theorem, base principle of NoSQL databases, types of NoSQL databases, case study on disease diagnosis and profiling.

Unit 4 Applications of Data Science 10

Tools and Applications of Data Science: Introducing Neo4jfor dealing with graph databases, graph query language Cypher, Applications graph databases, Python libraries like nltk and SQLite for handling Text mining and analytics, case study on classifying Reddit posts.

Unit 5 Data Visualization and Prototype Application Development 10

Data Visualization and Prototype Application Development**:** Data Visualization options, Crossfilter, the JavaScript MapReduce library, Creating an interactive dashboard with dc.js, Dashboard development tools. Applying the Data Science process for real-world problem-solving scenarios as a detailed case study.

Prescribed Textbooks:

1. Davy Cielen, Arno D.B.Meysman, and Mohamed Ali, “Introducing to Data Science using Python tools”, Manning Publications Co, Dreamtech press, 2016

2. Prateek Gupta, “Data Science with Jupyter” BPB publishers, 2019 for basics

Reference Books:

1. Joel Grus, “Data Science From Scratch”, OReilly, 2019
2. Doing Data Science: Straight Talk From The Frontline, 1 st Edition, Cathy O’Neil and Rachel Schutt, O’Reilly, 2013

CO-PO Mapping:

Course outcomes	Engineering Knowledge	Problem Analysis	Design/Development t of solutions	Conduct investigations of Complex Problems	Engineering tools usage	The Engineer and World	Ethics	Individual and collaborative teamwork	Communication	Project Ma engagement and Finance	Life-long Learning
24AAID41T-1	3	3				3			3	2	
24AAID41T-2	3	3	3	3		3	2		3		
24AAID41T-3	3	3	3	3		3			3	2	
24AAID41T-4	3	3	3	3	1	3	2		3		
24AAID41T-5	3	3	3	3		3			3		

UNIT-1

1)INTRODUCTION TO DATA SCIENCE

Data is a collection of raw facts. It is a set of characters used to collect, store and transmit information for a specific purpose. Data can be in any form, i.e., text, image, audio, etc. Data comes from the Latin word 'Datum,' which means 'something given'. Data can be structured as well as unstructured. Processed data is termed as Information.

Before the Internet era began, handling data was easier as there was no concept of Big Data, when people started using Internet widely especially with the arrival of Facebook and YouTube in the early 2000s, almost everybody started using the Internet. There began the generation of a large amount of data. With the generation of such large amounts of data, its storage and process became difficult, So the concept of Big Data came. The size of Big Data is expansive (in terabytes or petabytes) and grows exponentially with time.

Handling of such huge amount of data is a challenging task for every organization. So, to handle, process, and analysis of large data we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science.

Data science is a deep study of the massive/Large amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI) and machine learning(ML). Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

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

2)BENEFITS AND USES

Benefits & uses of Data Science:

1. **Improved Decision-Making**

   By using data science to address problems and inform viewpoints, data scientists play a critical role in allowing better decision-making. To analyze and process large datasets and to extract insightful data, they use different methodologies, that can enable companies and organizations to make wise decisions.

   For suppose A data scientist examine patient data in a healthcare organization, for instance, to find trends and patterns that can improve patient outcomes. In the retail sector, data analyst may be used to develop new goods and services and to have a better understanding of customer behavior.

2. **Increased Efficiency**

   Business operations can be made more efficient and costs can be cut with the use of data science. Businesses can find inefficiencies and potential improvement by analyzing data. Then, modifications that boost efficiency and eliminate inefficiency.

   For suppose Data analytics helps business identify patterns and trends, enabling them to increase efficiency and eliminate inefficiencies.

3.Competitive Advantage

   By empowering to make better decisions and discover new opportunities, data science provide a competitive advantage. Businesses may remain competitive by utilizing data to obtain insights into their processes and customers.

   For suppose A store could use data science to examine sales data and find new trends. Based on this knowledge, the merchant can create new products or change their marketing plan to benefit from these trends before their rivals.

4.Predictive Analytics

Based on past data, data science can be used to forecast future results. Businesses can find trends and forecast future occurrences by using machine learning algorithms to analyze massive datasets.

5.Efficient Resource Allocation

Utilizing data on resource allocation, demand trends, and supply chain dynamics. Data science helps organizations in maximizing resource allocation. So waste is reduced and operational efficiency is increased while resources like inventory, people, and equipment are appropriately allocated.

6.Continuous Improvement

Organizations with a culture of continuous development benefit from data science. Organizations can assess performance, monitor advancement, and efficiency fields for development by analyzing data. This data-driven strategy encourages an attitude of constant improvement and innovation.

7.Innovation and New Opportunities

Data science may help companies innovate and create new opportunities. Data science is playing a key role behind innovation, allowing companies to find new perspectives. Additionally, data science can find new business ideas by examining data, market dynamics, and customer behavior.

8.Personalized Marketing and Customer Segmentation

Organizations can segment their consumer bases and develop individualized marketing efforts using data science. This allows them to better understand individual preferences and needs.

For Suppose, a retail business can utilize data science approaches to recognize high-value clients and develop marketing campaigns or loyalty schemes to improve clients. Similar to this, an e-commerce platform can make relevant product recommendations based on a user's browsing history and buying products by using customer segmentation.

9.Enhanced Customer Experience

Discovering customer preferences and behavior can be accomplished through data analysis. The customer experience can be improved by using this information to create goods and services that are fulfil to the need of the user.

Using data science for example, analyze prior customer purchases and make customized product recommendations. The probability of repeat business might rise as a result.

10.Better Healthcare Outcomes

The healthcare sector becoming a good transformation because of data science. Data scientists can gain insights to increase diagnosis accuracy, optimize treatment strategies, and improve patient care, resulting in better healthcare outcomes, by analyzing patient data, medical records, and clinical studies.

Additionally, by taking a patient's unique characteristics, such as genetics, lifestyle, and previous treatment outcomes, data science enables the optimization of treatment programmes. Data scientists can find patterns and connections in large-scale clinical data that help them choose the best treatments for certain patient profiles.

3)FACETS OF DATA

Faceting is a way to get an overview of a specific data. Facets correspond to properties of the information elements.

There are many facets of data science, including: Identifying the structure of data. Accessing and importing data. Cleaning, filtering, reorganizing, augmenting, aggregating data and Visualizing data.

The main facets of data science, including:

a) Structured
b) Unstructured
c) Natural language
d) Machine-generated
e) Graph-based
f) Audio, video and images

g) Streaming data

A) Structured:

• The term structured data refers to data that is identifiable, because it is organized in a structure. The most common form of structured data is a database where specific information is stored based on particular way of columns and rows.

• so, the Structured data is arranged in rows and columns format. It helps for application to retrieve and process data easily. Database management system is used for storing structured data. Structured data is understood by computers and is also efficiently organized for human readers.

• Excel table is an example of structured data.

B) Unstructured Data

• Unstructured data is data that does not follow a specific format. So Unstructured data has no identifiable structure. Because Rows and columns are not used for unstructured data. Therefore, it is difficult to retrieve required information.

• The unstructured data can be in the form of Text: (Documents, email messages, customer feedbacks), audio, video, images. Email is an example of unstructured data.

Even today in most of the organizations more than 80 % of the data are in unstructured form. So extracting information from these various sources is a very big challenge.

Characteristics of unstructured data:

1. There is no structural format for the data.

2. Data can be of any type.

3. Unstructured data does not follow any structural rules.

4. There are no predefined formats.

5. Since there is no structural format for unstructured data, it is unpredictable in nature.

C) Natural Language

• Natural language is a special type of unstructured data.

• Natural language processing enables machines to recognize characters, words and sentences, then apply meaning and understanding to that information. This helps machines to understand language as humans do.

•For natural language processing to help machines understand human language, it must go through speech recognition, natural language understanding and machine translation.

D) Machine - Generated Data

• Machine-generated data is an information that is produced by mechanical or digital devices without human interaction. That means the data entered manually by an end-user is not recognized to be machine-generated.

• Examples of machine data are web server logs, call detail records, network event logs and telemetry.

• Both Machine-to-Machine (M2M) and Human-to-Machine (H2M) interactions generate machine data. Machine data is generated continuously by every processor-based system, as well as many consumer-oriented systems.

It can be either structured or unstructured. In recent years, the increase of machine data has surged(rised). The expansion of mobile devices, virtual servers and desktops, as well as cloud- based services and RFID (Radio Freequency Identification) technologies.

E) Graph-based or Network Data

•Graphs are data structures to describe relationships and interactions between entities in complex systems. Generally, a graph contains a collection of entities called nodes and edges.

• Nodes represent entities, which can be of any object type that is relevant to our problem domain. By connecting nodes with edges, we will end up with a graph (network) of nodes.

• A graph database stores nodes and relationships instead of tables or documents. Graph databases are used to store graph-based data and are queried with specialized query languages such as SPARQL (pronounced as sparkle).

• Graph databases are capable of sophisticated **fraud prevention.** With graph databases, we can use relationships to process financial and purchase transactions in real time. With graph queries, we are able to detect that, for example, a purchaser is using the same email address and credit card as included is a known fraud case.

• Graph databases can also easily detect relationship patterns such as multiple people associated with a personal email address or multiple people sharing the same IP address but residing in different physical addresses.

• Graph theory is the main method in social network analysis in the early history of the social network concept. The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links (for example influencers and the followers).

• Influencers on social network have been identified as users that have impact on the activities or opinion of other users by way of followership or influence on decision made by other users on the network as shown in following diagram.

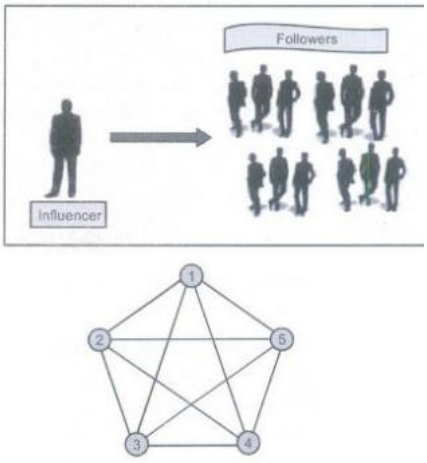


Fig. 1.2.2 : Graph on 5 vertices

f) Audio, Image and Video

•The terms audio and video commonly refers to the time-based media storage format for sound/music and moving pictures information. Audio and video digital recording, also referred as audio and video.

. It is important to remark that multimedia data is one of the most important sources of information and knowledge; the integration, transformation and indexing of multimedia data bring significant challenges in data management and analysis. Many challenges have to be addressed including big data for the nature of Data Science.

• Data Science is playing an important role to address these challenges in multimedia data. Multimedia data usually contains various forms of media, such as text, image, video.

G) Streaming Data

. Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously and in small sizes (order of Kilobytes).

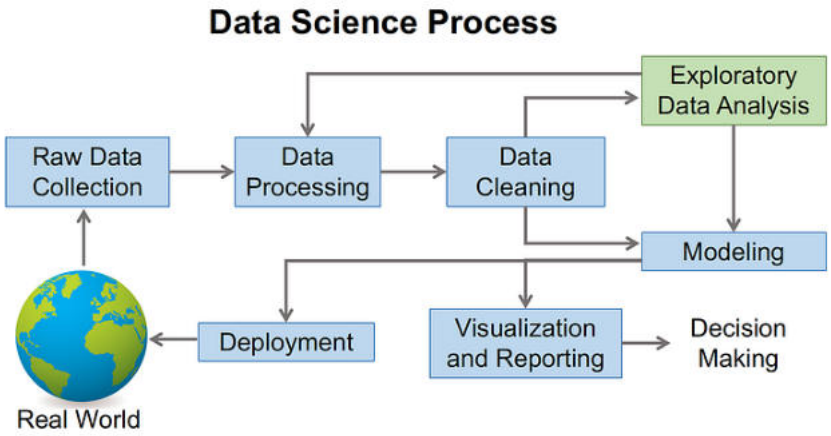
• Streaming data includes a wide variety of data such as log files generated by customers using your mobile or web applications, ecommerce purchases, in-game player activity, information from social networks.

Difference between Structured and Unstructured Data

Sr. No.	Parameters	Structured data	Unstructured data
1.	Representation	It is in discrete form i.e. stored in row and column format	Unstructured data is data that does not follow a specified format
2.	Meta data	Syntax	Semantics
3.	Storage	Database management system	Unmanaged file structure
4.	Standard	SQL, ADO.net, ODBC	Open XML, SMTO, SMS
5.	Integration tool	ETL	Batch processing or manual data entry
6.	Characteristics	With a structure document, certain information always appears in the same location on the page.	In unstructured document information can appear in unexpected places on the document.
7.	Used by organizations	Low volume operations	High volume operations

4)DATA SCIENCE PROCESS IN BRIEF

Another term for the data science process is “the data science life cycle”. The data science process is a systematic approach to solving a data problem. It provides a structured framework for our problem, and deciding how to solve it, then presenting the solution. It involves various stages, including problem definition, data collection, preprocessing, exploratory analysis, model building, and deployment.



A) Data collection:

Data Collection refers to the systematic process of gathering and analyzing information from various sources to get a complete idea of interest. Different sources of data collection include Primary Sources and Secondary Sources.

During the data collection phase, data scientists acquire the necessary data to address the defined problem. This involves identifying data sources, both internal and external, that contain relevant information. It may include structured data from databases, spreadsheets, or APIs, as well as unstructured data such as text documents, images, or social media.

Data scientists engage in exploratory data analysis (EDA) to comprehend the dataset's structure, size, and variables. By verifying the data, they ensure its integrity and determine if any additional data is required to enhance the analysis.

B) Data Preprocessing and Cleaning

Raw data is not in a particular format and does not suitable for analysis. Data preprocessing involves cleaning, transforming, and organizing the data to make it usable. This step includes handling missing values, dealing with outliers, addressing data inconsistencies, and performing feature engineering to create new variables or

modify existing ones. The goal is to ensure data quality and prepare the data for analysis.

Most of the data you collect during the collection phase will be unstructured, irrelevant, and unfiltered. Bad data produces bad results, so the accuracy of your analysis will depend on the quality of your data.

Cleaning data : This step is the most time-intensive/taken process, but finding and resolving flaws in your data is essential to building effective models.It Eliminates duplicate and null values, corrupt data, inconsistent data types, invalid entries, missing data, and improper formatting.

C)Exploratory Data Analysis (EDA):

Exploratory Data Analysis is an important step that involves summary, visualizing, missing values, outlier detection, correlation analysis….

- **Data Summary**: Generate descriptive statistics to summarize the main characteristics of the data, such as mean, median, standard deviation, minimum, and maximum values.
- **Data Visualization**: Create visual representations, including histograms, scatter plots, box plots, and bar charts, to gain overview to the distribution, patterns, and relationships within the data.
- **Identify Missing Values**: Identify and handle missing data by exploring the presence of null values or incomplete records.
- **Outlier Detection**: Detects outliers, which are extreme values that deviate significantly from the majority of the data points. Assess their impact and decide whether to store, remove, or transform them based on the analysis goals.
- **Correlation Analysis**: Explore the relationships between variables by calculating correlation coefficients, such as Pearson's correlation, to determine the strength and direction of linear associations.
- **Feature Importance**: Assess the importance of input features or variables using techniques such as feature ranking, importance scores, or permutation importance to understand their impact on the target variable.
- **Data Distribution:** Examine the distribution of variables and assess whether they follow a particular distribution, such as normal distribution, skewed distribution.
- **Data Exploration**: Data Exploration refers to the initial step in data analysis in which data analyst use data visualization and statistical techniques to describe dataset characterization such as quantity, accuracy.

- **Hypothesis Generation**: Formulate initial hypotheses about relationships, patterns, or potential causality in the data based on observations and initial analysis, which can guide further investigation.

D)Model Building and Machine Learning

With a solid understanding of the data, it's time to build predictive models or apply machine learning algorithms to extract valuable insights. Select the appropriate algorithms based on the problem type (classification, regression, clustering, etc.) and the nature of the data. Train the models using the prepared data and evaluate their performance using suitable methods.

Different types of machine learning algorithms and techniques have been developed which can easily identify complex patterns in the data which will be a very difficult task to be done by a human.

E) Interpretation and Insights

Once models are built, it's important to interpret their results and extract meaningful insights. Understand the factors driving the models' predictions or outcomes and assess their significance in the context of the problem. Communicate the insights in a clear and actionable manner to stakeholders, enabling informed decision-making.

F) Deployment and Monitoring

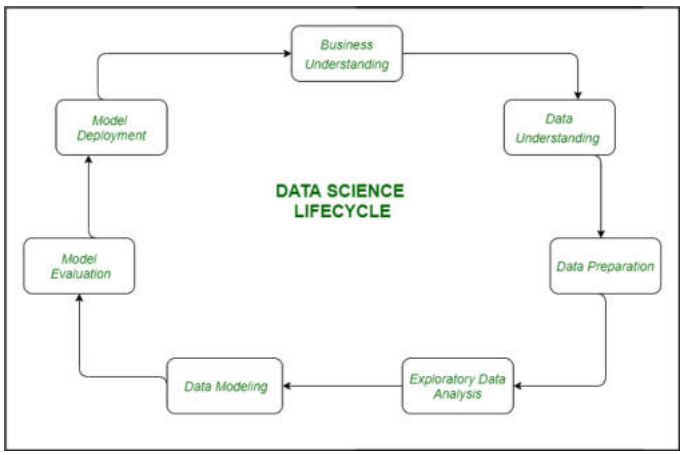
The data science process doesn't end with insights. To realize the full value of data science, it's important to deploy the models into production. Integrate the models into the business workflow or decision-making systems. Continuously monitor the performance of the models, updating and retraining them as new data becomes available. This ensures the models remain accurate and relevant over time.

Deployment:

- Integrate the developed models into the target production environment or systems.
- Develop APIs or interfaces that allow external systems or applications to interact with the deployed model, enabling easy integration and data exchange.
- Conduct testings to ensure the deployed model functions as expected, producing accurate predictions or outcomes in real-world scenarios.
- Implement appropriate security measures to protect the deployed model.

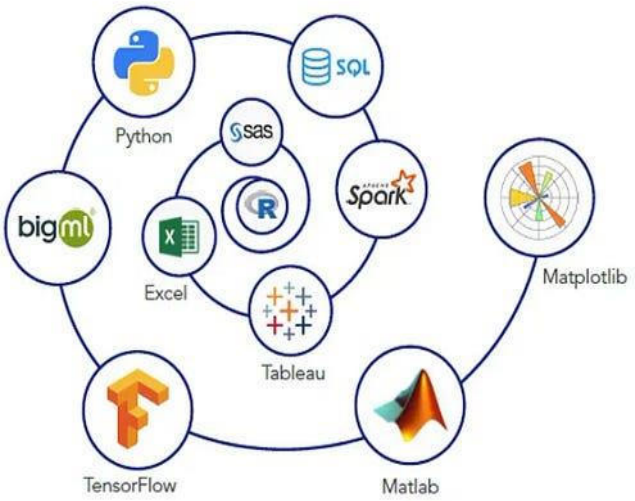
Monitoring:

- Define and monitor performance methods specific to the deployed model, such as prediction accuracy, response time, or resource utilization.
- Continuously monitor the quality and consistency of incoming data to ensure it meets the requirements of the deployed model, detecting and handling anomalies/errors.
- Monitor the performance of the deployed model over time, assessing its accuracy, stability, and any degradation in performance.
- Capture and analyze prediction errors or unexpected outcomes to identify potential issues for improvement. Use techniques like error logs, confusion matrices, or anomaly detection.
- Gather feedback from end-users or stakeholders to understand their experience with the deployed model, addressing any usability issues, and incorporating necessary improvements or updates.
- Document the deployment process, monitoring strategies, and any changes made to the model or its environment. This documentation ensures transparency, reproducibility, and for future maintenance or updates.



Tools for Data Science Process:

There are various tools and programming languages used in the Data Science process, such as MATLAB, Tableau/Power BI, Python, and R. These tools provide utility features for different tasks in Data Science, making the process more efficient and effective.



5)BIG DATA ECO SYSTEM AND DATA SCIENCE

The term Ecosystem is defined in scientific as a complex network or interconnected systems. The big data ecosystem refers to the interconnected network of organizations, technology platforms and applications that support big data. The ecosystem includes companies that develop and deploy big data solutions, as well as those who use big data to make business decisions.

The big data ecosystem is growing at a very fast, and it will require significant investment in order to keep up. As the industry continues to increase, businesses will need to find ways to work with larger data sets and create efficiencies through collaboration. To do this, they will need to understand the basics of the big data ecosystem and its components.

The big data ecosystem has five key components:

1. Data sources: Every business needs access to reliable and large data sets in order to make informed decisions. In order to find these sources, businesses need to identify where their data comes from and how it can be accessed. This can be done through a variety of methods, such as market research or surveys.

2. Platforms: Businesses use a number of different platforms to store, process and analyze their data. These platforms can come from traditional technology companies such as Microsoft or Amazon, or new entrants such as google Cloud platform or Apples iCloud.

3. Applications: Businesses use a wide range of applications in order to process their data. These applications can be used for everything from analyzing customer behavior to manufacturing products.

4. Data management: All businesses require effective ways to manage their data sets so that they are organized, effective and accessible. This can be done through a number of methods, including manual process or automatic processes such as imilating cubes from various source datasets into a single report or exporting all your tables into an Excel file for analysis.

5. Collaboration: All businesses need effective ways to collaborate with other organizations in order to share information and make better decisions. This can be done through a variety of methods, including online surveys or collaborations with outside experts (such as developers who can help improve the efficiency of your existing solutions).

Big data ecosystem with the advances in technology and the rapid evolution of computing technology, it is becoming a very important to process and manage huge amount of information without the use of supercomputers. There are some tools and techniques that are available for data management like Google BigTable, Data Stream Management System (DSMS), NoSQL amongst others.

However, there is an urgent need for companies to deploy special tools and technologies that can be used to store, access, analyse and large amounts of data in near-real time. Big Data cannot be stored in a single machine so several machines are required. Common tools that are used to manipulate Big Data are Hadoop, MapReduce, and BigTable. These tools are able to process large amount of data efficiently.

UNIT-2

CHAPTER-1

1)APPLICATIONS OF MACHINE LEARNING IN DATA SCIENCE

Data Science: Data science is a deep study of the massive/Large amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

Machine Learning : Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focus on the using data and algorithms to enable AI to imitate the way that humans learn. Arthur Samuel first used the term "machine learning" in 1959. Machine Learning (ML) involves training algorithms to make predictions or decisions based on data.

Applications of machine learning in data science:
some of the most popular applications of Machine Learning in Data Science are.
1) Image Recognition
2) Speech Recognition
3) Stock-market Trading
4) Traffic prediction
5) Email Spam and Malware Filtering
6) Product Recommendation
7) Real-Time Navigation
8)On-line Fraud Detection
9)Medical Diagnosis
10)Virtual Personal Assistant

1)Image Recognition: Image Recognition is one of the most common applications of Machine Learning in Data Science. Image Recognition is used to identify objects, persons, places, digital images … The most popular use cases of this application are Face Recognition in Smartphones, Face Recognition devices in offices…

Facebook provides use a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm. It is based on the Facebook project named "Deep Face," which is responsible for face recognition and person identification in the picture.

2) Speech Recognition: Speech Recognition is a process of translating spoken words into text. (Or) Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text" Some of the examples are Siri, Google Assistant, Alexa, YouTube Closed Captioning, etc.

While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

3) Stock Market trading

Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's "Neural Network" is used for the prediction of stock market trends.

4) Traffic prediction

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as traffic is cleared, traffic is busy, or heavily congested with the help of two ways

- **Real Time location**
- **Average time has taken**

5)Email Spam and Malware Filtering:

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail

- **Header filter**
- **Content Filter**
- **General blacklists filter**
- **Permission filters**

Some machine learning algorithms such as **Multi-Layer Perceptron**, **Decision tree,** and **Naïve Bayes classifier are used for email spam filtering and malware detection.**

6) Product Recommendation:

Product Recommendation is mainly used by ECommerce and Entertainment companies like Amazon, Flipkart, Netflix etc. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product and this is because of machine learning. Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

7) Real-Time Navigation:

Google Maps is one of the most commonly used Real-Time Navigation applications. When you open the application, it sends the data back to Google, providing information about the route being traveled and traffic patterns at any given time of the day.

All the information given by the number of users using the application on regular basis has given Google a huge database of traffic data which allows Google Maps.

8) Online Fraud Detection

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways such as fake accounts, fake ids, and steal money in the middle of a transaction. So, to detect this, "Feed Forward Neural network "helps us by checking whether it is a genuine transaction or a fraud transaction.

9) Medical Diagnosis:

In medical science, machine learning is used for diseases diagnoses. Medical technology is growing very fast. For suppose it build 3D models that can predict the exact position of problems in the brain. It helps in finding brain tumors and other brain-related diseases easily.

10) Virtual Personal Assistant:

We have various virtual personal assistants such as Google assistant, Alexa, Siri. they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email…

2)ROLE OF ML IN DS

Data science definition:

ML definition:

Role of Machine Learning in Data Science:

Machine learning significantly boosts data science by improving analysis efficiency, patterns, predicting outcomes, and identifying anomalies in datasets, and improved decision-making.

1.Enabling predictive modeling:

Machine learning is like having a superpower. Because It can look at old data and find New patterns. Those patterns help guess what will happen next. It's accurate.

They can use it to make plans and good ideas. machine learning techniques are able to identify useful insights and make predictions that guide critical decisions in many different fields. For suppose in Finance Machine learning looks at old stock market information and guess what prices will do. It can help investors know when to buy or sell.

In Healthcare. It can look at patient information and guess if they might get sick. If they might recover sooner. That can make patients healthier.

2.Facilitating classification

Machine learning algorithms work like tools. They sort data into set of groups. This makes it easier to handle and understand information. By grouping items based on their qualities.

For suppose Machine learning algorithms can sort products into groups like electronics, clothes, or home needs. Thus, customers can easily find what they want. Because this sorting is automated, machine learning algorithms save time and energy.so machine learning makes data management and understanding better.

3.Supporting anomaly detection

Machine learning plays an important role in detecting anomalies in datasets. This includes finance, cybersecurity, and healthcare. Here, finding anything unusual early on might stop big losses or risks. For example, in banks, machine learning algorithms can make transactions safe. This can stop fraud.

4.Real-world Applications

The influence of machine learning in data science industries, facilitating efficient analysis, predictive modeling, anomaly detection, and decision-making processes, enhancing overall productivity and effectiveness.

1.Business: Machine learning helps businesses improve service. It uses client data to suggestions, predict demands, and automate jobs, which elevates service and efficiency.it allows to gather valuable knowledge from large data.

2.Healthcare: Machine learning is changing the game in healthcare! It helps identify diseases, predicts how patients will do, and treatment plans to specific needs. This makes healthcare better. That means we can choose and diagnose conditions early.

Machine learning also forecasts how patients will recover based on factors like past medical issues, and how they respond to treatment.

3.Finance: Machine learning is most important in finance. It helps find fraud, check risks, and manage investments in the best way. It looks at lots of financial data to find regular patterns that might mean fraud.

4.Marketing: Machine Learning enables customer segmentation, campaign optimization, and personalized marketing strategies, improving targeting and

conversion rates.

5. Education: Machie learning supports personalized learning experience and performance prediction, enhancing student performance and academic outcomes.

Industry	Applications
E-Commerce	Recommendations, Demand prediction
Healthcare	Disease Diagnosis, Outcome Prediction
Finance	Fraud Detection
Marketing	Customer Segmentation
Transportation	Route Optimization
Manufacturing	Predictive Maintenance, Quality Control
Education	Performance Prediction

3)PYTHON TOOLS LIKE SKLEARN

A Python module called Scikit-learn offers a variety of supervised and unsupervised learning techniques. It is based on several technologies like NumPy, pandas, and Matplotlib.

Sklearn:

One of the most widely used machine learning packages on GitHub is Python's scikit-learn. French data scientist "David Cournapeau" developed scikits learn package. Its name refers to the idea that it's a modification to SciPy called "SciKit" (SciPy Toolkit), which was independently created and published.

Implementation of Sklearn:

Scikit-learn is mainly coded in Python and mainly utilizes the NumPy library for highly efficient array and linear algebra computations. Support vector machines, logistic regression, and linear SVMs are performed using coded in Cython(Cython is a static compiler for python) for LIBSVM and LIBLINEAR, respectively.

Scikit-learn works with other Python packages, including SciPy, Pandas data frames, NumPy for array vectorization, Matplotlib, seaborn and plotly for plotting graphs, and many more.

Key concepts and features

Algorithms for making decisions, such as Data are identified and categorized by classification as per the patterns. Regression is the process of forecasting or predicting data values using the historical and data average.

Clustering is the automatic collection of datasets with related data.

- Predictive analysis is supported by various algorithms, including neural networks for pattern recognition and straightforward linear regression.
- Compatibility with the libraries of NumPy, pandas, and matplotlib

A predictive model can be built or trained on input data by computers using machine learning (ML), eliminating the need for explicit programming. A subset of AI is machine learning (AI)

Scikit-learn's salient characteristics are:

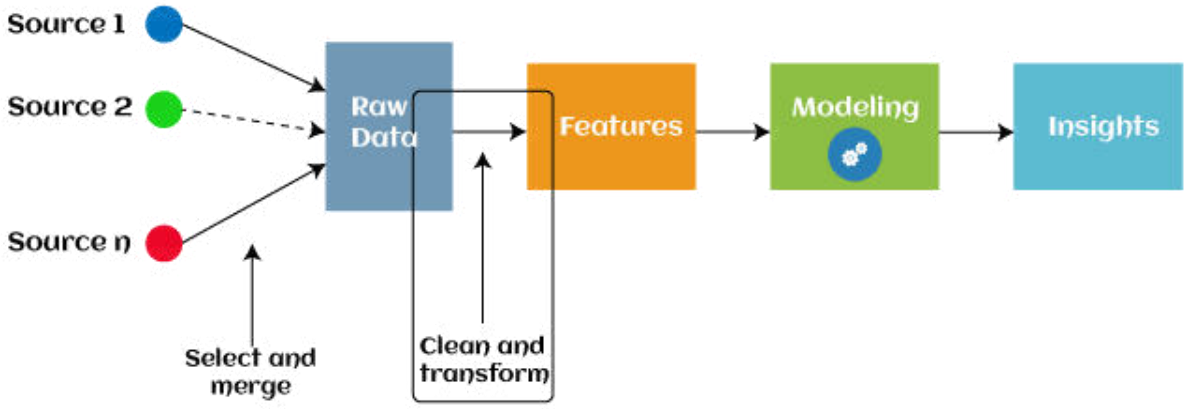
- **The package provides the functions for data mining and machine learning algorithms for data analysis that are easy to use and effective. Support vector machines, random forests, k-means, and other regression, classification, and clustering algorithms are included.**
- **The package is open source, accessible to everyone and reusable.**
- **It is built on SciPy, Matplotlib, and NumPy.**
- **The package has a commercially usable.**

4)MODELLING PROCESS FOR FEATURE ENGINEERING

Generally, all machine learning algorithms take input data to generate the output. The input data in a tabular form consisting of rows (instances or observations) and columns (variable or attributes), and these are generally known as features.

Feature engineering is the pre-processing step of machine learning, which extracts features from raw data.It is the process of creating new features or transforming existing features to improve the performance of a machine-learning model. It involves selecting relevant information from raw data and transforming it into a format that can be easily understood by a model. The goal is to improve model accuracy by providing more meaningful and relevant information.

In simple words Feature engineering is the process of transforming raw data into features that are suitable for machine learning models. And it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models.



The success of machine learning models mainly depends on the quality of the features used to train them. Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data.

It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.

Since 2016, automated feature engineering is also used in different machine learning software that helps in automatically extracting features from raw data. Feature engineering in ML contains mainly four processes:

1)**Feature Creation,.**

2)Transformations.

3) Feature Extraction.

4)Feature Selection.

1. **Feature Creation: Feature creation is finding the most useful variables to be used in a predictive model. The process is subjective, and it requires human creativity and interaction.** The new features are created by mixing existing features using addition, subtraction, and ration.
2. **Transformations: The transformation step of feature engineering involves adjusting the predictor variable to improve the accuracy and performance of the model. For example, it ensures that the model is flexible to take input of the variety**

of data; it ensures that all the variables are same. It improves the model's accuracy and ensures that all the features are within the acceptable range to avoid any error.

3. **Feature Extraction: Feature extraction is an automated feature engineering process that generates new variables by extracting them from the raw data. The main aim of this step is to reduce the volume of data so that it can be easily used and managed for data modelling. Feature extraction methods include cluster analysis, text analysis…**
4. **Feature Selection: While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the remaining features are either duplicate or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may reduce the overall performance and accuracy of the model.**

Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.

Steps in Feature Engineering

The steps of feature engineering may different as per different data scientists and ML engineers. However, there are mainly 3 steps involved in most machine learning algorithms.

- **Data Preparation: The first step is data preparation. In this step, raw data acquired from different resources are prepared to make it in a suitable format so that it can be used in the ML model**. The data preparation may contain cleaning of data, delivery, or loading.
- **Exploratory Analysis: Exploratory analysis or Exploratory data analysis (EDA) is an important step of features engineering, which is mainly used by data scientists. This step mainly involves analysis, summarization, Different data visualization techniques, and find the most appropriate statistical technique for data analysis, and to select the best features for the data.**
- **Benchmark: Benchmarking is a process of setting a standard baseline for accuracy to compare all the variables from this baseline. The benchmarking process is used to improve the predictability of the model and reduce the error rate.**

5)MODEL SELECTION

In machine learning, the process of selecting the top model or algorithm from a list of models to address a certain issue is referred to as model selection. That means Model selection is the process of selecting the best model from all the available models for a particular problem on the basis of different categories such as model accuracy, model efficiency and model complexity.

Model selection is an essential phase in the development of powerful and precise predictive models in the field of machine learning**. Model selection is the process of deciding which algorithm and model architecture is best suited for a particular task or dataset**. **The choice of an appropriate machine learning model is important because there are various levels of complexity, assumptions, and capabilities. During the selection process, it is important to take into account the assumptions, constraints, and important parameters that are unique to each model.**

The following steps are included in the model selection process:

- **Problem formulation: It includes the kind of predictions or task that you'd like the model to solve for example, classification, regression, or clustering**.
- **Candidate model selection: select a group of models that are appropriate for the issue. These models can include straight forward methods like decision trees or linear regression as well as more sophisticated ones like deep neural networks, random forests, or support vector machines.**
- **Performance evaluation: Establish measures for measuring how well each model performs. Common measurements include Area Under the Receiver's Operating Characteristic curve (AU-ROC), recall, F1-score, and accuracy.** The type of problem and the particular requirements will determine which metrics are used.
- **Training and evaluation: Each candidate model should be trained using a subset of the available data (the training set), and its performance should be assessed using a different subset (the validation set or via cross-validation)**.
- **Model comparison: Evaluate the performance of various models and determine which one performs best on the validation set. Consider elements like data handling capabilities, interpretability, computational difficulty, and accuracy.**
- **Hyper parameter tuning: Before training, many models require that certain hyper parameters, such as the learning rate, regularization strength, or the number of layers that are hidden in a neural network, use methods like grid search, random search, and Bayesian optimization to identify these hyperparameters' ideal values.**
- **Final model selection: After the models have been analyzed and fine-tuned, select the model that performs the best. Then, this model can be used to make predictions based on.**

Model Selection Techniques: Here are some methods/techniques for selecting models that are frequently used:

- **Train-Test Split: With this strategy, the available data is divided into two sets: a training set & a separate test set.** The models are evaluated using a predetermined evaluation metric on the test set after being trained on the training set. **This method offers a quick and easy way to evaluate a model's performance using hypothetical data**.
- **Cross-Validation: A resampling approach called cross-validation divides the data into various groups**. Lowering the variance in the evaluation makes it easier to generate an accurate assessment of the model's performance. **Cross-validation techniques that are frequently used include leave-one-out, stratified, and k-fold cross-validation.**
- **Grid Search: Hyper parameter tuning is done using the grid search technique. In order to do this, a grid containing hyperparameter values must be defined**. For each combination, the models are trained, assessed, and their performances are contrasted. **Finding the ideal hyperparameter settings to optimize the model's performance is made easier by grid search**.
- **Random Search: A set distribution for hyperparameter values is sampled at random as part of the random search hyperparameter tuning technique. In contrast to grid search, random search only investigates a portion of the hyperparameter field**. When a complete search is not possible due to the size of the search space, this strategy can be helpful.
- **Bayesian optimization: A more sophisticated method of hyperparameter is Bayesian optimization. It models the relationship between the performance of the model and the hyperparameters using a probabilistic model.** It intelligently chooses which set of hyperparameters to investigate next by updating the probabilistic model and iteratively assessing the model's performance. **When the search space is big and expensive to examine, Bayesian optimization is especially effective**.
- **Model averaging: This technique combines forecasts from various models to get a single prediction. For regression issues, this can be accomplished by averaging the predictions. Model averaging can increase overall prediction accuracy**.
- **Information Criteria: Information criteria offer a numerical assessment between model complexity and goodness. Example include the Bayesian Information Criterion (BIC). This method not useful if we use too complicated models.**
- **Domain Expertise & Prior Knowledge: The models that are more suitable given the specifics of the problem and the details of the data may be known by subject (Domain)experts.**
- **Model Performance Comparison: The best-performing model can be found by comparing many models Using the right assessment measures, it is complex task to evaluate the performance of various models. Depending on the issues.**

6)VALIDATION AND PREDICTION

Model validation is a core component of developing machine learning or artificial intelligence (ML/AI). It assesses the ability of an ML or statistical model to produce predictions with accuracy to be used to achieve business objectives.

Model validation is a set of processes and activities designed to ensure that an ML or an AI model performs as it should. This includes its design objectives and utility for the end user.

This can be done through testing, examining the construction of the model and the tools and data used to create it. Moreover, it is part of ML governance, the complete process of controlling access, implementing policies, and tracking model activity.

Why is model validation important?

Model validation is an important step in developing any machine learning or artificial intelligence **system. It helps ensure that the model performs as intended and can handle unseen data.**

Without proper model validation, the confidence in its ability to generalize well on unseen data can never be high. Furthermore, validation helps determine the best model, parameters, and accuracy metrics for the given task.

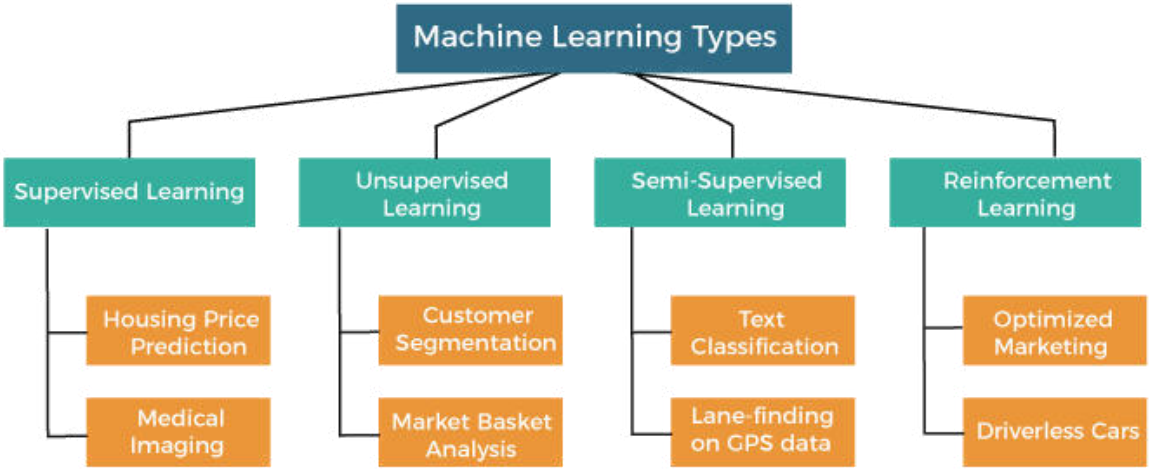
7)TYPES OF ML

Machine Learning Defintion:

Types of Machine Learning:

Machine learning is divided into mainly four types, which are:

1. **Supervised Machine Learning**
2. **Unsupervised Machine Learning**
3. **Semi-Supervised Machine Learning**
4. **Reinforcement Learning**



1. Supervised Machine Learning

As its name represents, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output.

Here, the labelled data specifies that some of the inputs are already mapped to the output. That means we can say that first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Suppose we have an input dataset of bike and car images. So, first, we will provide the training to the machine to understand the images, such as the structure & size, dimensions, height of bike and car.

After completion of training, we input the picture of a bike and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as the structure & size, dimensions and height, and find that it's a bike. So, it will put it in the bike category. This is the process of how the machine identifies the objects in Supervised Learning.

The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, **etc.**

Categories of Supervised Machine Learning

Supervised machine learning can be classified into two types.

- **Classification**

- **Regression**

a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "B. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are Spam Detection, Email filtering...

Some popular classification algorithms are…

- **Random Forest Algorithm**
- **Decision Tree Algorithm**
- **Logistic Regression Algorithm**
- **Support Vector Machine Algorithm**

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are..

- **Simple Linear Regression Algorithm**
- **Multivariate Regression Algorithm**
- **Decision Tree Algorithm**
- **Lasso Regression**

Advantages and Disadvantages of Supervised Learning

Advantages:

- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.

Disadvantages:

- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

- **ImageSegmentation:**

  Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
- **MedicalDiagnosis:**

  Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions.
- **Fraud Detection –**

  Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
- **Spam detection -**:

  In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
- **Speech Recognition –**

  Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

2. Unsupervised Machine Learning

Unsupervised learning is different from the Supervised learning technique, as its name represents, there is no need for supervision. It means, in unsupervised machine

learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences**.** Machines are instructed to find the hidden patterns from the input dataset.

suppose there is group of cars images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects based on size. Now the machine will discover its patterns and differences, such as colour difference, shape difference, and predict the output.

Categories of Unsupervised Machine Learning

Unsupervised Learning can be further classified into two types,

- **Clustering**
- **Association**

1) Clustering

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups.

Some of the popular clustering algorithms are given below:

- **K-Means Clustering algorithm**
- **Mean-shift algorithm**
- **DBSCAN Algorithm**
- **Principal Component Analysis**
- **Independent Component Analysis**

2) Association

Association learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables

accordingly. **This algorithm is mainly applied in Web usage mining, continuous production**, etc.

Some popular algorithms of Association rule learning are Apriori Algorithm, Eclat, FP-growth algorithm.

Advantages and Disadvantages of Unsupervised Learning Algorithm

Advantages:

- These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
- Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

Disadvantages:

- The output of an unsupervised algorithm can be less accurate as the dataset is not labelled.
- Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

Applications of Unsupervised Learning

- **Network Analysis: Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data**..
- **Anomaly Detection: Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraud transactions**.
- **Singular Value Decomposition: Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location**.

3. Semi-Supervised Learning

Semi-Supervised learning is a type of Machine Learning algorithm that is in between Supervised and Unsupervised machine learning. It represents the data between supervised and Unsupervised algorithms which uses the combination of labelled and unlabeled datasets.

To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data or unlabeled data.

For suppose Supervised learning is where a student is under the supervision of faculty college. Further, if that student is self-learning the same concept without any help from the faculty, then it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of faculty at college.

Advantages and disadvantages of Semi-supervised Learning

Advantages:

- It is simple and easy to understand the algorithm.
- It is highly efficient.
- It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages:

- Results may not be stable.
- We cannot apply these algorithms to network-level data.
- Accuracy is low.

4. Reinforcement Learning

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore taking action, learning from experiences, and improving its performance. Agent gets rewards for each good action and get punished for each bad action, hence the goal of reinforcement learning agent is to maximize the rewards.

The Reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life.

Due to its way of working, reinforcement learning is used in different fields such as Game theory, Operation Research, Information theory, multi-agent systems.

Categories of Reinforcement Learning

Reinforcement learning is categorized mainly into two types of methods/algorithms:

- **Positive Reinforcement Learning: Positive reinforcement learning specifies increasing the efficiency that the required behaviour** ,would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
- **Negative Reinforcement Learning**: **Negative reinforcement learning works exactly opposite to the positive RL. It decreases the Efficiency that the specific behaviour would occur**.

Real-world Use cases of Reinforcement Learning

- **VideoGames:**
  RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are AlphaGO and **AlphaGO Zero**.
- **Robotics:**
  RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning.
- **TextMining**:
  Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.

Advantages and Disadvantages of Reinforcement Learning

Advantages

- It helps in solving complex real-world problems which are difficult to be solved by general techniques.
- The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
- Helps in achieving long term results.

Disadvantage

- RL algorithms are not preferred for simple problems.
- RL algorithms require huge data and computations.

- Too much reinforcement learning can lead to an overload of states which can weaken the results.

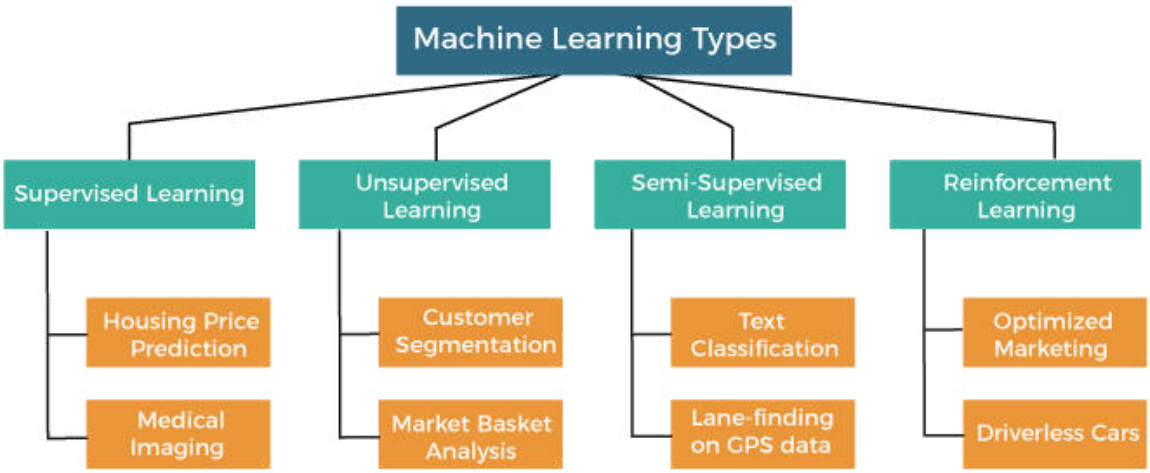
8)SEMI-SUPERVISED LEARNING

Machine Learning Defintion :

Types of Machine Learning:

Machine learning is divided into mainly four types, which are:

1. **Supervised Machine Learning**
2. **Unsupervised Machine Learning**
3. **Semi-Supervised Machine Learning**
4. **Reinforcement Learning**



Semi-Supervised learning is a type of Machine Learning algorithm that is in between Supervised and Unsupervised machine learning. It represents the data between supervised and Unsupervised algorithms which uses the combination of labelled and unlabeled datasets.

To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data or unlabeled data.

For suppose Supervised learning is where a student is under the supervision of faculty college. Further, if that student is self-learning the same concept without any help from the faculty, then it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of faculty at college.

Advantages and disadvantages of Semi-supervised Learning

Advantages:

- It is simple and easy to understand the algorithm.
- It is highly efficient.
- It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages:

- Results may not be stable.
- We cannot apply these algorithms to network-level data.
- Accuracy is low.

Working of Semi-Supervised Learning

Semi-supervised learning uses pseudo labeling to train the model with less labeled training data than supervised learning. The process can combine various neural network models and training ways:

- Firstly, it trains the model with less amount of training data similar to the supervised learning models. The training continues until the model gives accurate results.
- The algorithms use the unlabeled dataset with pseudo labels in the next step, and now the result may not be accurate.
- Now, the labels from labeled training data and pseudo labels data are linked together.
- The input data in labeled training data and unlabeled training data are also linked.
- In the end, again train the model with the new combined input as did in the first step. It will reduce errors and improve the accuracy of the model.

Real-world applications of Semi-supervised Learning-

Semi-supervised learning models are becoming more popular in the industries.

- Speech Analysis- **It is the most classic example of semi-supervised learning applications**. Since, labeling the audio data is the most task that requires many human resources, **this problem can be naturally overcome with the help of applying SSL in a Semi-supervised learning model**.

- **Web content classification- This is very critical and important to label each page on the internet because it needs human interaction. Still, this problem can be reduced through Semi-Supervised learning algorithms. Further, Google also uses semi-supervised learning algorithms to rank a webpage for a given query.**
- **Text document classifier- As we know, it would be very difficult to find a large amount of labeled text data, so semi-supervised learning is an ideal model to overcome this.**

Examples of Semi-Supervised Learning :

- **Text classification: In text classification, the goal is to classify a given text into one or more predefined categories. Semi-supervised learning can be used to train a text classification model using a small amount of labeled data and a large amount of unlabeled text data.**
- **Image classification: In image classification, the goal is to classify a given image into one or more predefined categories. Semi-supervised learning can be used to train an image classification model using a small amount of labeled data and a large amount of unlabeled image data.**
- **Anomaly detection: In anomaly detection, the goal is to detect patterns or observations that are unusual or different from the normal.**

UNIT-2

CHAPTER-2

HANDLING LARGE DATA

1)PROBLEMS AND GENERAL TECHNIQUES FOR HANDLING LARGE DATA

In general, a dataset is considered "large" when it exceeds the capacity of a computer's memory. It means the data cannot be loaded into memory all at once.

Characteristics of Large Data:

1. **Volume:** They contain a large number of records, might be 100s of gigabytes, terabytes, or even petabytes of data.
2. **Complexity:** They may involve different data, including structured and unstructured data.
3. **Velocity:** Data arrives and is generated at a high rate, and difficult to real-time processing.
4. **Variety:** Datasets may include text, images, videos, social media data, and more.

Challenges of Large Data:

- **Storage** - Large datasets require more storage capacity, and it can be expensive and difficult to manage and maintain to store the data.
- **Tools** –As traditional techniques may not be suitable for large datasets and can be difficult to use. So Specialized tools and techniques are needed to process such data
- **Access** - Collecting and accessing large datasets can be time-consuming…Transferring and communicating large datasets between systems or over networks can be slow.
- **Resources** - Designing and managing the infrastructure (Resource) to process and analyze large datasets is a difficult task.

Data Storage Strategies:

- **Distributed File Systems:** Systems like Hadoop Distributed File System (HDFS), Spark, and other cloud storage solutions are designed for storing and managing large datasets efficiently. They distribute data across multiple nodes, making it accessible in parallel.
- **Columnar Storage:** Utilizing columnar storage formats like Apache Parquet or Apache ORC(Optimized Row Columnar) can significantly reduce storage and improve query performance. These formats store data column-wise, allowing for efficient compression and selective column retrieval.
- **Data Partitioning:** Partitioning your data into smaller subsets can enhance query performance. It's particularly useful when dealing with time stamped data.
- **Data Compression:** By Using compression algorithms like Snappy or Gzip can reduce storage requirements without compromising data quality.

The specific data storage strategy utilized by your organization may bring its own challenges.

General Techniques:

1.Allocate More Memory

Some machine learning tools may be limited by a default memory configuration.Then you can re-configuration your tool to allocate more memory.A good example is Weka.(Waikato Environment For Knowledge Analysis.)

2. Work with a Smaller Sample

In this step make sure you need to work with all of the data? Take a random sample of your data, such as the first 1,000 rows. Use this smaller sample to work through your problem before going on all of your data.

3. Use a Computer with More Memory

In this technique we can get access to a larger computer with more memory. For example, a computer that work on a cloud service like Amazon Web Services that offers hundreds of gigabytes of RAM.

4. Change the Data Format

If your data stored in ASCII text, or CSV file then it can occupy more memory.in this case you can use less memory by using another data format. A good example is a binary format like GRIB(General Regularly-Distributed Information in Binary form), NetCDF(Network common Data Form), or HDF(Hierarchical Data Format).

There are many command line tools that you can use to transform one data format into another data format.

5. Stream Data

Does all of the data need to be in memory at the same time? you can use code or a library to stream data as-needed into memory..

6. Use a Big Data Platform

In some cases, you may need to work with a big data platform. That is, a platform designed for handling very large datasets, that allows you to use data transforms and machine learning algorithms .Two good examples are Hadoop and Spark.

2)PROGRAMMING TIPS FOR DEALING LARGE DATA

Introduction:

Characteristics:

Challenges:

Programming Tips for Dealing Large Data :

1)Use streaming or chunking

The simple way to handle large data is to use streaming or chunking techniques, which allow you to process data in smaller batches or units, rather than loading the whole data at once.

Streaming or chunking can reduce the memory and enable parallel or distributed processing. For example, you can use Python's pandas library to read and write data in chunks.

2)Compress or reduce data

Another way to handle large data is to compress or reduce it, which can save storage space and speed up processing. Compression or reduction can be done by applying various techniques, such as Filtering ,Encoding, Dimensionality reduction, Aggregating ,Sampling &Hashing .

For example, you can use Python's zlib methods to compress data files, or use scikit-learn's algorithms to reduce the dimensionality of data. But, compression or reduction may also cause some problems such as loss of information, quality.

3)Use external or cloud storage

A third way to handle large data is to use external or cloud storage, which can provide more capacity and flexibility than local memory. External or cloud storage can be accessed remotely through APIs or protocols, such as HTTP, FTP, S3(Simple Storage Services from Amazon), or HDFS.

For example, you can use Amazon S3 or Google Cloud Storage to store and retrieve large data files, or use Hadoop or Spark to process data on a distributed file system.

4)Use appropriate tools and frameworks

A fourth way to handle large data is to use appropriate tools and frameworks, which can offer specialized features and functionalities for dealing with large data. Tools and frameworks can be chosen based on the type, format, and purpose of the data, such as relational, non-relational, structured, unstructured, or streaming.

For example, you can use SQL or NoSQL databases to store and query large data, or use TensorFlow or PyTorch to build and train deep learning models on large data.

5)Optimize your code and algorithms

A fifth way to handle large data is to optimize your code and algorithms, which can improve the efficiency and performance of your data processing. Optimization can be done by applying various techniques, such as vectorization, caching, memorization & batching.

For example, you can use NumPy or Cython to vectorize your code.

6)Here's what else to consider

This is the step to share examples, new ideas, or insights that missed any of the previous sections.

3)CASE STUDIES ON DS PROJECTS FOR PREDICTING MALICIOUS URL's

The unique and specific address of each page on the Internet is called URL (Uniform Resource Locator).Malicious websites are well-known threats in cybersecurity. They act as an efficient tool for viruses, worms, and other types of malicious codes online and caused for most cyber attacks. Malicious URLs can be delivered through email links, text messages, browser pop-ups, page advertisements, etc.

These URLs may be links to Malicious/Harmful websites and downloaded to systems. These downloads can be viruses, worms, etc. Various techniques for malicious URL detectors have previously depends mainly on two Techniques 1) URL blacklisting 2) signature blacklisting.

1) URL blacklisting: Blacklisting involves maintaining a database of known malicious domains and comparing the hostname of a new URL to hostnames in that database. But main disadvantage of this method is It will be unable to detect new and unseen malicious URL, which will only be added to the blacklist after it has been observed as malicious url's.

2)Signature blacklisting : signature list provide a predictive approach that is generalizable across platforms and independent of prior knowledge of known signatures. Given a sample of malicious, ML techniques will extract features of known good and bad URLs and generalize these features to identify new and unseen good or bad URLs.

To improve the malicious URL detection methods,various machine learning techniques are developed.

1)Handling API Rate Limits and IP Blocking

To confirm that malicious URLs in the systems, we need to send multiple requests to VirusTotal. VirustTotal provides aggregated results from multiple virus scan engines. And also we send URLs through Shodan. Shodan is a search engine for all devices connected to the internet providing service-based features of the URL's server.

2)Fingerprinting URLS

In this Technique to extract URL characteristics that are separating malicious URLs from good URLs.

The URL fingerprinting process contains 3 types of URL features:

- URL String Characteristics: Features derived from the URL string itself.
- URL Domain Characteristics: Domain characteristics of the URLs domain. These include who provide information.
- Page Content Characteristics: Features extracted from the URL's page .

3)Removing Highly/multi Correlated Features

This method based on linear analysis.i.e pairs of features must not be correlated. The reason behind this assumption is that there is no additional information added to a model with multiple correlated features as the same information.

Multi-correlated features are also indication of redundant features in the data . By removing correlated features we can address the issues of feature redundancy .

4)Decision Trees

A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure.

Using a decision tree (DT) algorithm to select the best relevant subset of features. The effectiveness of the experimental outcomes is evaluated in terms of time, accuracy, and error reduction.

5)Feature Selection

Which variables are most useful in identifying a URL as 'malicious' or 'benign'? Computationally, we can automatically select what variables are most useful by testing which ones

‘improves’ or ‘fails to improve’ the overall performance of the prediction model. This process is called ‘Feature Selection’.

The goal of feature selection is to obtain a useful subset of the original data that is predictive of the target feature in such a way that useful information is not lost.

4)FOR BUILDING RECOMMENDER SYSTEMS

An AI-powered recommendation system is an intelligent technology used by businesses to suggest products, services, or content to users based on their preferences.This is the core concept behind recommender systems.

Recommendation engines use sophisticated algorithms and statistical models to predict and present users with items, services, or content that match their interests and preferences. Here system uses AI, which means it can analyze a lot of data quickly and understand patterns.

Types of Recommender Systems:

There are mainly 4 types of Recommender systems.

1) Collaborative filtering
2)Content Based Filtering
3)Hybrid Recommendations systems
4)Deep Learning Based Recommendations

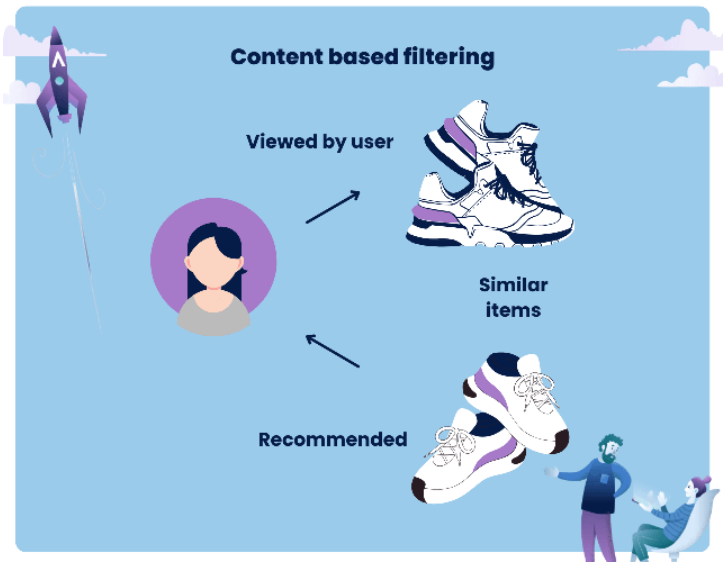
1)Collaborative filtering

The collaborative filtering model is a popular method used in AI-based recommendation systems. It works by collecting and analyzing information about users' behaviors, activities, or preferences, and predicting what they will like based on the preferences of other, similar users.

Collaborative filtering algorithms based on collecting and examining user data. For example, in case of movie recommendation system, collaborative filtering look at the movies you've watched and compare them to what others have watched. If you and another user liked many of the same movies, the system would recommend movies that the other user liked but you haven't seen.

The collaborative filtering approach can be very effective, but it has some limitations, such as requiring a large amount of user data to work well and sometimes recommending only popular Elements rather than our interest.

2)Content based filtering



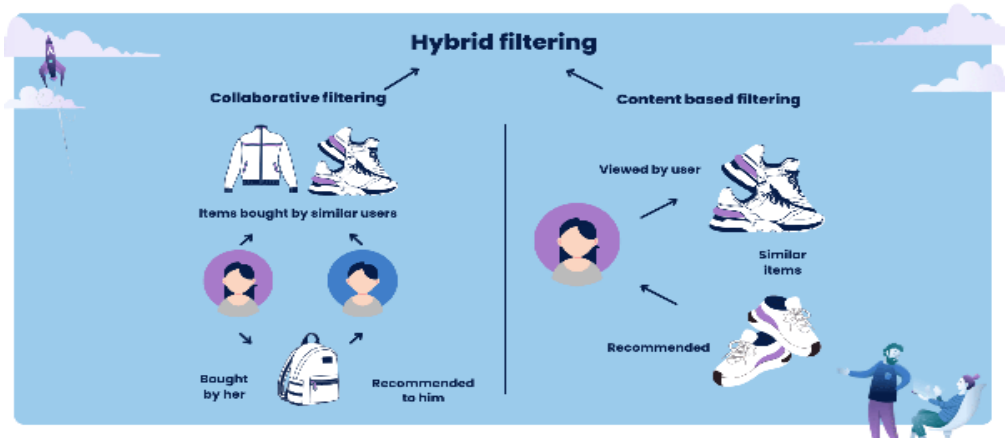
Content-based filtering is another key method used in AI-powered recommendation systems, focusing on the characteristics of the items themselves rather than on user behavior patterns**.**

This approach recommends products or content by understanding the characteristics of the items and matching them to a user's profile.

For example, in a Book recommendation system, content-based filtering looks at the attributes of books you searched in the past-such as author, user ratings and then finds other books with similar attributes.If you've searched several Data science books, the system will suggest other books in the Data science base.

Content-based filtering sometimes limit the variety of recommendations because it suggest items that are very similar to what you already know and like.

3)Hybrid recommendation systems



Hybrid recommendation systems combine the best of collaborative and content-based filtering to provide more accurate suggestions**.** That means This approach combines insights from user behavior (like collaborative filtering) and the characteristics of the content/product(like content-based filtering) to make recommendations.

For example, in a music streaming service, a hybrid system might use collaborative filtering to identify other users with music tastes similar to yours, and then use content-based filtering to suggest songs that match the specific artists, or music styles you prefer.

The hybrid approach is particularly effective because it can overcome the limitations of each method. It can provide more personalized recommendations than collaborative filtering alone,and it can provide a wider range of suggestions than content-based filtering, By combining these methods, hybrid systems can provide a more balanced and comprehensive recommendation experience.

4)Deep learning-based recommendations:

Deep learning and machine learning recommendation systems are advanced types of recommendation systems that use deep learning techniques, a subset of artificial intelligence (AI) that like the workings of the human brain in processing data.

These systems are capable of handling complex and large-scale data to provide highly personalized recommendations.In a deep learning recommendation system, layers of algorithms,

called neural networks, analyze large amounts of data, learning patterns and relationships within the data.

For example, in an e-commerce platform, such a system can analyze not just your past purchases and views but also more suitable factors like how long you look at an item, the sequence of your browsing, and your interactions with various product features.

These systems based on the machine learning model are particularly powerful because they can understand and utilize a wide range of data types, including text, images, and user interactions to create a comprehensive user profile and item descriptions.

UNIT-3
NOSQL MOVEMENT FOR HANDLING BIGDATA

INTRODUCTION:

Data is a collection of information. It can be different forms like text, numbers, media, bytes, etc. it can be stored in system memory, etc.

A database is an organized collection of data, so that it can be easily accessed and managed. The main purpose of the database is to operate a large amount of information by storing, retrieving, and managing data. There are many databases available like DBMS, MySQL, Oracle SQL Server, MongoDB(NOSQL)etc…

Big Data refers to large and complex datasets that cannot be effectively managed, processed, or analyzed using traditional data processing tools and methods.

1)DISTRIBUTING DATA STORAGE AND PROCESSING WITH HADOOP FRAMEWORK

Distributed Storage: Hadoop stores large data sets across multiple machines, allowing for the storage and processing of extremely large amounts of data.

Data Processing: Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It was created by Apache Software Foundation in 2006.It can be used to store large amount of structured and unstructured data.

The data processing with Hadoop frame work mainly consist 3 components.

1)HDFS (THE STORAGE LAYER)
2)MAP REDUCE (THE PROCESSING LAYER)
3)YARN (THE MANAGEMENT LAYER)

1. HDFS (The storage layer)

As the name represents, Hadoop Distributed File System is the storage layer of Hadoop and is responsible for storing the data in a distributed environment. It divides the data into several blocks of data and stores them across different data nodes. It has two main Nodes

A) **NameNode**: It is running on the master machine. It saves the locations of all the files stored in the clients file system and tracks where the data resides across the client.

When the client applications want to make certain operations on the data, it interacts with the NameNode. When the NameNode receives the request, it responds by returning a list of Data Node servers where the required data resides.

B) DataNode: This process runs on every client machine. One of its functionalities is to store each HDFS data block in a separate file in its local file system. That means, it contains the actual data in form of blocks. It sends the requests and waits for the responces from the NameNode to access the data.

2. MapReduce (The processing layer)

It is a programming technique based on Java that is used on top of the Hadoop framework for faster processing of large amount of data. It processes huge data in a distributed environment using many Data Nodes which enables parallel processing and faster execution.

MapReduce primarily consist of three phases namely the Map phase, the Sort phase, and the Reduce phase.

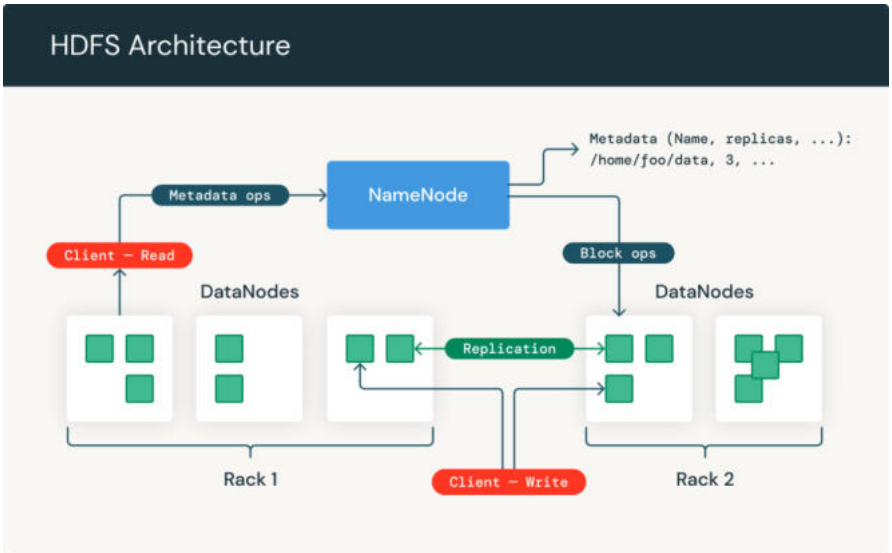
A) Map Phase: It is the first phase in the processing of the data. The main task in the map phase is to process each input and convert it into tuples (key-value pairs).

B) Sort Phase: This phase shuffle & sort the tuples (key-value pairs). The sorting is done on the basis of the keys in the key-value pairs.

C)Reduce Phase: The output of the map phase is an input to the reduce phase. It takes these key-value pairs and applies the reduce function on them to produce the result.

3. YARN (The management layer)

There are background processes running at each node (Node Manager on the client nodes and Resource Manager on the master node) that communicate with each other for the allocation of resources. The Resource Manager is the key element of the YARN layer which manages resources among all the applications and passes on the requests to the Node Manager.

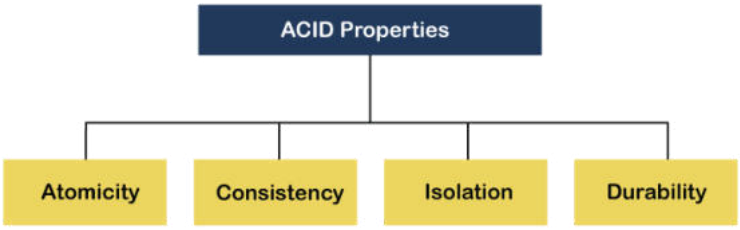


2)ACID PRINCIPLE OF RELATIONAL DATABASES

Database ACID is a set of database attributes that ensures that database transactions are completed efficiently. It consists of Atomicity, Consistency, Isolation, and Durability.

ACID Compliance in Relational Databases:

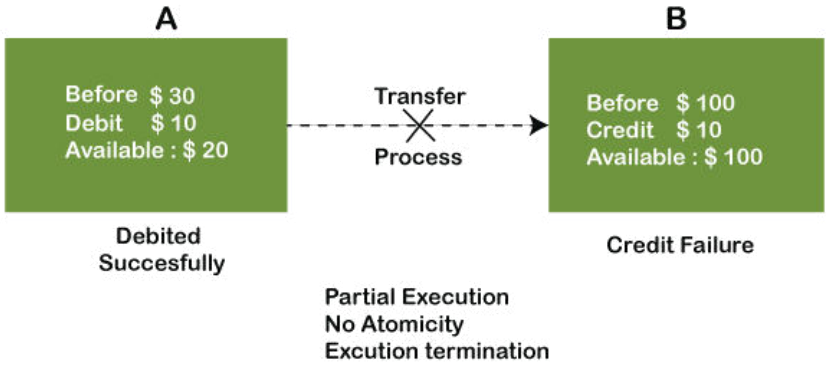
A transaction is a group of operations executed as a single unit of work.An example of a transaction is when money is transferred between bank accounts. Money must be debited from one account and credited to another.



1) Atomicity

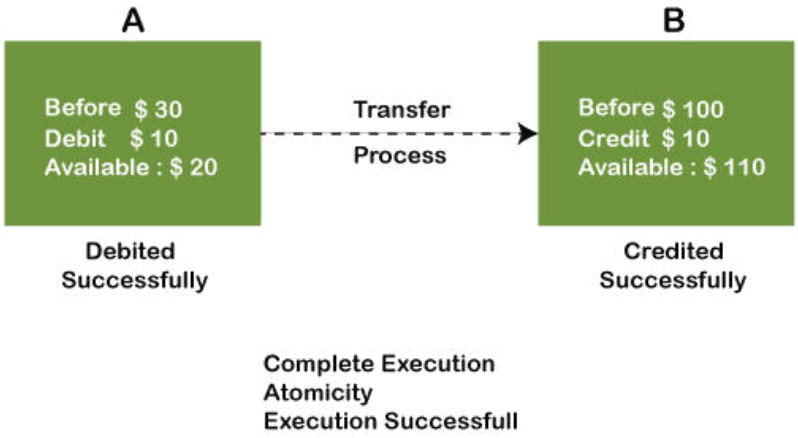
The term atomicity defines that the data remains atomic. It means if any operation is performed on the data, either it should be performed or executed completely or should not be executed at all. In Atomicity the operation should not break in between or execute partially.

Example: If A has account having $30 in his account from which he wishes to send $10 to B account. In account B, a sum of $ 100 is already present. When $10 will be transferred to account B, the sum will become $110. Now, there will be two operations that will take place. One is the amount of $10 that A wants to transfer will be debited from his account, and the same amount will get credited to account B.Now, suppose the first operation of debit executes successfully, but the credit operation fails. Thus, in account A, the value becomes $20, and to that of B account, it remains $100 as it was previously present.



In the above diagram,we can observe that after crediting $10, the amount is still $100 in account B. So, it is not an atomic transaction.

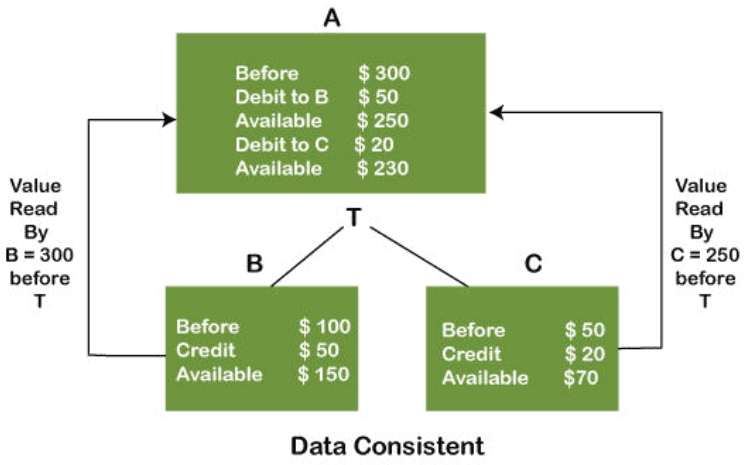
The below diagram shows that both debit and credit operations are done successfully. Thus the transaction is atomic.



Thus, when the amount loses atomicity, then in the bank systems, this becomes a huge issue, and so the atomicity is the main focus in the bank systems.

2) Consistency

The word **consistency** means that the value should remain preserved(saved/maintained) always. In DBMS, the integrity of the data should be maintained, which means if a change in the database is made, it should remain preserved always. In the case of transactions, the integrity of the data is very essential so that the database remains consistent before and after the transaction & the data should always be correct.



Data Consistent

Ex :

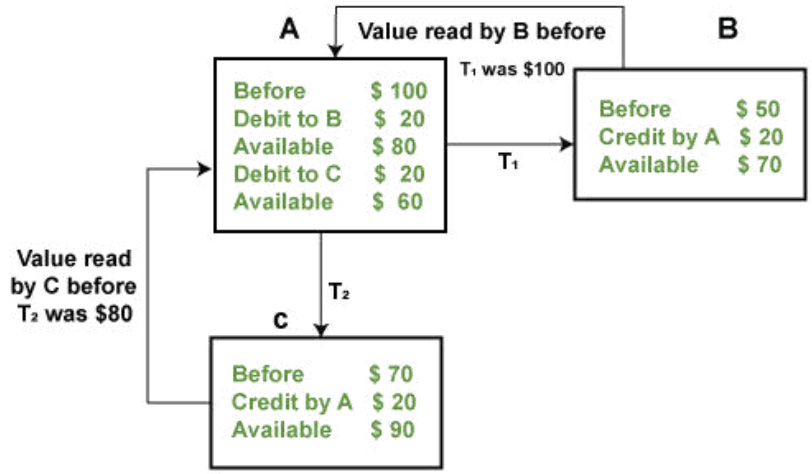
In the above diagram, there are three accounts, A, B, and C, where A is making a transaction T one by one to both B & C. There are two operations that take place, i.e., Debit and Credit. Account A firstly debits $50 to account B, and the amount in account A is read $300 by B before the transaction. After the successful transaction T, the available amount in B becomes $150. Now, A debits $20 to account C, and that time, the value read by C is $250 (that is correct as a debit of $50 has been successfully done to B). The debit and credit operation from account A to C has been done successfully. We can see that the transaction is done successfully, and the value is also read correctly. Thus, the data is consistent. In case the value read by B and C is $300, which means that data is inconsistent because when the debit operation executes, it will not be consistent.

3) Isolation

The term 'isolation' means separation. Isolation is the property of a database where no data should affect the other one and may occur concurrently. It means if two operations are being performed on two different databases, they may not affect the value of one another. In the case of transactions, when two or more transactions occur simultaneously, the consistency should remain maintained. Any changes that occur in any particular transaction will not be seen by other transactions until the change is not committed in the memory.

Example: If two operations are concurrently running on two different accounts, then the value of both accounts should not get affected. The value should remain persistent. As you can see in the below diagram, account A is making T1 and T2

transactions to account B and C, but both are executing independently without affecting each other. It is known as Isolation.



Isolation - Independent execution of T_1 & T_2 by A

4) Durability

Durability ensures the permanency of something. The term durability ensures that the data after the successful execution of the operation becomes permanent in the database. The durability of the data should be so perfect that even if the system fails but the database still survives. However, if gets lost, it becomes the responsibility of the recovery manager for ensuring the durability of the database. For storing the values, the COMMIT command must be used every time we make changes.

4)CAP THEOREM

The CAP theorem is also called Brewer's Theorem, because it was introduced by Professor Eric A. Brewer in 2000. CAP theorem states that in networked shared-data systems or distributed systems, we can only achieve at most two out of three guarantees for a database: **C**onsistency, **A**vailability and **P**artition Tolerance.It is very important to understand the CAP theorem as It makes the basics of choosing any NoSQL database based on the requirements.

1)Consistency: consistency means that all clients see the same data at the same time, doesn't matter which node they connect to in a distributed system. To achieve

consistency, whenever data is written to one node, it must be instantly forwarded to all the other nodes in the system before the write is successful.

2)Availability: Availability means that every non-failing node returns a response for all read and write requests in a reasonable amount of time, even if one or more nodes are down. That means all working nodes in the distributed system return a valid response for any request, without failing or exception.

3)Partition Tolerance: It means the system continues to operate if any loss or failure of the system.That means, even if there is a network issue and some of the computers are unreachable, still the system continues to perform.In this Partition tolerance, Partition represents to a communication break between nodes within a distributed system. Means if a node cannot receive any messages from another node in the system, because of network failure, server crash, or any other reason. Partition Tolerance is the ability of a system to continue to work even when errors may exist. The CAP theorem categorizes systems into three categories:

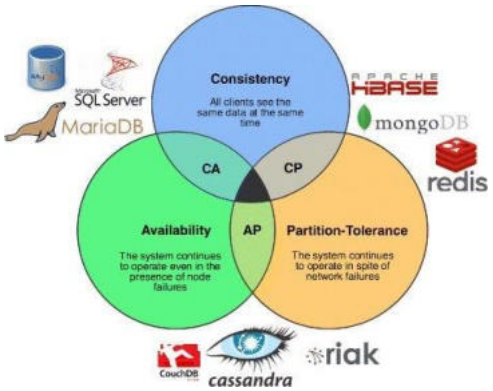
1)CP (Consistent and Partition Tolerant) database: A CP database delivers consistency and partition tolerance at the expense of availability. When a partition occurs between any two nodes, the system has to shut down the non-consistent node (i.e., make it unavailable) until the partition is resolved.

2)AP (Available and Partition Tolerant) database: An AP database delivers availability and partition tolerance at the expense of consistency. When a partition occurs, all nodes remain available but those at the wrong end of a partition might return an older version of data than others. When the partition is resolved, the AP databases typically resync the nodes to repair all inconsistencies in the system.

3)CA (Consistent and Available) database: A CA delivers consistency and availability in the absence of partition tolerance. As long as partition does not exist between any two nodes within the system, consistency and availability will remain

in place,because , without partition tolerance in place within the network, the entire system would be inoperable if an error occurs.

It is important to mention that **NoSQL databases** do not require a specific framework. They are known for being user friendly and having wide availability and scalability.The following diagram shows the classification of different databases based on the CAP theorem.



5)BASE PRINCIPLE OF NOSQL DATABASES

NoSQL are designed on the **BASE** principle:

- **B**asically **A**vailable: read and write operations are always available on all nodes, at the base of consistency. , BASE-modelled NoSQL databases will ensure availability of data by spreading and replicating it across the nodes of the database cluster.
- **S**oft state: without the guarantee of consistency, after some time, we can only predict the state with certain probability. Due to the lack of immediate consistency, data values may change over time.
- **E**ventually consistent: if the system is functional, the data will eventually get to consistent state, after enough time has passed.

ACID vs. BASE: What are the differences?

The fundamental difference between ACID and BASE database models is the way they deal with this limitation.

- The ACID model provides a consistent system.
- The BASE model provides high availability.

When compared to **ACID**-compliant databases. NoSQL does not guarantee atomicity of transations. NoSQL databases are not suited for some types of data. An

example is payment processing. That kind of data always needs to be saved to an **ACID** database.

Another difference between SQL and NoSQL is their behavior during network issues. NoSQL maintains service availability, while SQL favor data consistency.

Which Databases are Using the BASE Model?

MongoDB, Cassandra and Redis are among the most popular NoSQL solutions, together with Amazon DynamoDB and Couchbase.

6)TYPES OF NOSQL DATABASES

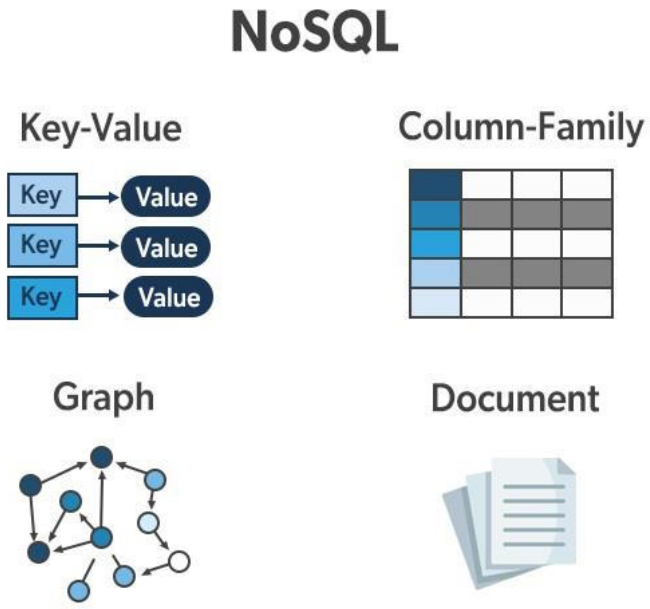
A database is a collection of structured data or information which is stored in a computer system and can be accessed easily. A database is usually managed by a Database Management System (DBMS).

NoSQL is a non-relational database that is used to store the data in the non-tabular form. NoSQL stands for Not only SQL. The main types are documents, key-value, wide-column, and graphs.

Types of NoSQL Database:

- Document-based databases
- Key-value stores
- Column-oriented databases

- Graph-based databases



1)Document-Based Database:

The document-based database is a non-relational database. Instead of storing the data in rows and columns (tables), it uses the documents to store the data in the database. A document database stores data in JSON, BSON, or XML documents.

In the Document database, the particular elements can be accessed by using the index value that is assigned for faster querying. Collections are the group of documents that store documents that have similar contents. Not all the documents are in any collection as they require a similar schema because document databases have a flexible schema.

Key features of documents database:

- Flexible schema: Documents in the database has a flexible schema. It means the documents in the database need not be the same schema.
- Faster creation and maintenance: the creation of documents is easy and minimal maintenance is required once we create the document.

- No foreign keys: There is no dynamic relationship between two documents so documents can be independent of one another. So, there is no requirement for a foreign key in a document database.
- Open formats: To build a document we use XML, JSON, and others.

2)Key-Value Stores:

A key-value store is a non-relational database. The simplest form of a NoSQL database is a key-value store. Every data element in the database is stored in key-value pairs. The data can be retrieved by using a unique key allotted to each element in the database. The values can be simple data types like strings and numbers or complex objects.

A key-value store is like a relational database with only two columns which is the key and the value.

Key features of the key-value store:

- Simplicity.
- Scalability.
- Speed.

3)Column Oriented Databases:

A column-oriented database is a non-relational database that stores the data in columns instead of rows. That means when we want to run analytics on a small number of columns, you can read those columns directly without consuming memory with the unwanted data.

Columnar databases are designed to read data more efficiently and retrieve the data with greater speed. A columnar database is used to store a large amount of data. Key features of columnar oriented database:

- Scalability.
- Compression.
- Very responsive.

4)Graph-Based databases:

Graph-based databases focus on the relationship between the elements. It stores the data in the form of nodes in the database. The connections between the nodes are called links or relationships.

Key features of graph database:

- In a graph-based database, it is easy to identify the relationship between the data by using the links.
- The Query's output is real-time results.
- The speed depends upon the number of relationships among the database elements.
- Updating data is also easy, as adding a new node or edge to a graph database is a simple task that does not require total schema changes.

UNIT-4 (TOOLS & APPLICATIONS OF DATA SCIENCE)

1)INTRODUCTION TO NEO4J FOR DEALING WITH GRAPH DATABASES

Neo4j is the most famous database management system and it is also a NoSQL database system. It is a powerful, high-performance, open-source graph database that enables the efficient management and querying of highly connected data .

Neo4j is a powerful and flexible graph database management system, designed to efficiently store and query highly interconnected data. Unlike traditional relational databases, which store data in tables, Neo4j uses a graph structure to represent and navigate relationships between data entities.

Neo4j is different from Mysql or MongoDB it has its own features and it is designed to efficiently store and query highly interconnected data that's makes it special compared to other Database Management System.

Neo4j structure

Neo4j stores and present the data in the form of graph not in tabular format or not in a Json format. Here the whole data is represented by nodes and there you can create a relationship between nodes. That means the whole database collection will look like a graph, that's why it is making it unique from other database management system.

Graph Database

Graph database is a database used to model the data in the form of graph. It is a pictorial representation of a set of objects where some pairs of objects are connected by links. It is composed of two elements - nodes (vertices) and relationships (edges). Here, the nodes of a graph represent the entities while the relationships represent the association of these nodes.

- **Nodes** represent entities such as people, businesses, or any data item.
- **Edges** (or relationships) connect nodes and illustrate how entities are related.
- **Properties** provide additional information about nodes and relationships

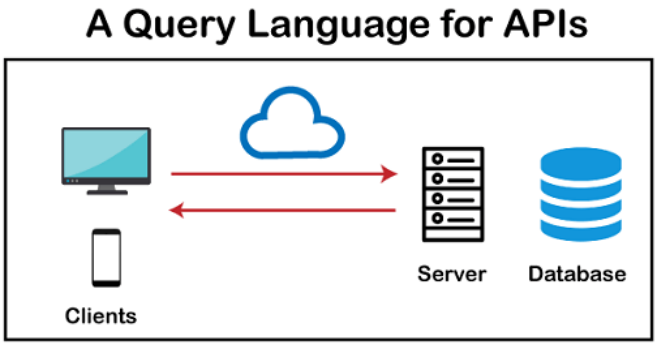
Popular Graph Databases

Neo4j is a popular Graph Database. Other Graph Databases are Oracle NoSQL Database, OrientDB, HypherGraphDB, GraphBase, InfiniteGraph, and AllegroGraph.

2)GRAPH QUERY LANGUAGE

GraphQL is a new API standard developed by Facebook**.** It is an open-source server-side technology, now maintained by a large companies and individuals of all over the world. It is also an execution engine that works as a data query language and used to fetch declarative data.

GraphQL is data query and manipulation language for your API and a server-side runtime for executing queries when you define a type system for your data. Unlike the REST APIs, a GraphQL server provides only a single endpoint and responds with the data for client request.



History of GraphQL

GraphQL was an internal project of Facebook. It was developed by Facebook in 2012 and was used in their mobile apps. After that, it was publicly released in 2015. A graph query language is a programming language used to query and manipulate graph databases. It allows developers to efficiently retrieve and update data stored in a graph structure.

Using a graph query language, you can navigate through these nodes and edges to find specific patterns or relationships. This makes it easier to handle complex queries that involve multiple levels of connections. For example, you can quickly find all the friends of a friend in a social network or identify the shortest path between two points in a transportation network.

Types of Graph Query Languages

1)Cypher

Cypher is a declarative query language developed by Neo4j, specifically designed for querying graph databases. When you write a Cypher query, you describe the structure of the data you want to retrieve, rather than detailing the steps to get there. This approach simplifies complex queries and allows you to focus on the relationships within the data.

2)Gremlin

Gremlin, part of the Apache Tinker Pop framework, is a graph traversal language that supports both imperative and declarative querying. Unlike Cypher, Gremlin allows you to write queries as a series of steps, which can be executed in a specific order. This flexibility makes Gremlin suitable for complex graph traversals and algorithms.

3)SPARQL

SPARQL is a query language for querying RDF (Resource Description Framework) data. It is a W3C recommendation and is widely used in the Web. SPARQL allows you to query and manipulate data stored in RDF format, which represents information as triples: subject, predicate, and object.

Graph Query Language vs. SQL

SQL is designed for querying tabular data, while graph query languages are optimized for graph structures. In SQL, data is organized into tables with rows and columns. This structure works well for many applications but can become difficult when dealing with large data. Graph query languages, on the other hand, represent data as nodes and edges. This makes easy to handle large amount of data.

3)CYPHER

Cypher is a declarative query language developed by Neo4j, specifically designed for querying graph databases. When you write a Cypher query, you describe the structure of the data you want to retrieve, rather than detailing the steps to get there. This approach simplifies complex queries and allows you to focus on the relationships within the data.

Cypher is unique because it provides a visual way of matching patterns and relationships. Cypher was inspired by an ASCII-art type of syntax where (nodes)-[:ARE_CONNECTED_TO]->(otherNodes) using rounded brackets for circular (nodes), and -[:ARROWS]-> for relationships.

Neo4j users use Cypher to construct expressive and efficient queries to do any kind of create, read, update, or delete (CRUD) on their graph..

The Neo4j has its own query language that called Cypher Language. It is similar to SQL.Neo4j does not work with tables, row or columns it deals with nodes. It is more satisfied to see the data in a graph format rather than in a table format.

Example: The Neo4j Cypher statement compare to SQL
MATCH (G:Company { name:"wipro" })

RETURN G

This Cypher statement will return the “Company” node where the “name” property is wipro. Here the “G” works like a variable to holds the data that your Cypher query requests after that it will return.

Below same query is written in SQL.

SELECT * FROM Company WHERE name = "wipro";

ASCII-Art Syntax: The Neo4j used **ASCII-Art** to create pattern.

(X)-[:wipro]->(Y)

- In the Neo4j the nodes are represented by “( )”.
- The relationship is represented by ” -> “.
- What kind of relationship is between the nodes are represented by ” [ ] ” like [:wipro]

So above description is helpful to decode the **ASCII-Art Syntax** given, (X)-[:wipro]->(Y). Here the X and Y are the nodes X to Y relation kind is “wipro”.

4)APPLICATIONS OF GRAPH DATABASES

……..graph data base defintion& intro…….

The following are the applications of Graph Data bases.

1)Greater Data complexity

2)Constant Evolution of data model

3)Simpler Relationship based querying

4)Fraud Detection

5)360-customer Views

6)Network Mapping

1)Greater Data Complexity: In a graph database, complex relationships between nodes can be added and removed in an easy-to-understand way, whereas within a tabular database this becomes difficult as additional data sources and use cases are introduced.

2)Constant Evolution of Data Model: In a graph, the data model can be modified continually without having to perform schema changes. For example, if a new property is required for a particular relation, this property can simply be added to nodes.

3)Simpler Relationship-based Querying: A graph database provides a query language that makes it easier to ask relationship-based questions of data. Whereas, performing complex queries in tables can be both difficult to understand and hard to optimise.

4)Fraud Detection

By using graph database, we can detect fraudulent behavior of a customer. Typically, in a fraud detection graph we will use entities from people such as names and dates of birth, as well as special entities such as IP addresses, device identifiers and access times. We can analyse the links between these entities and mark up those that have previously been marked as fraudulent. For example, use of multiple bank

accounts on a single device that has been previously used to access fraudulent accounts.

5)360-Degree Customer View

It is a typical for a business to have complete data about customers. To create a 360-degree customer view, we stream data into a graph, using a common data model to provide integration. For example, marketing in HubSpot, email tracking in MailChimp, website tracking in Mixpanel and so on.

6)Network Mapping

Network mapping is the best way for representation as a graph.It will use CMDBs (configuration management databases) and service catalogs to store networks of their systems.

A graph of the relationships between components not only enables interactive visualisations of the network but also use network tracing algorithms. For example

Dependency Management: Identify single points of failure and simulate the impact of their failure on services, to identify cascading failures before they happen.

Bottleneck Identification: Find weak links in network routing that could cause bottlenecks at times of high network utilisation.

Latency Evaluation: Estimate latency across paths in the network, and the impact on services accessed from various geographic regions.

5)PYTHON LIBRARIES LIKE NLTK & SQLite FOR HANDLING TEXT MINING & ANALYTICS

Python is a flexible programming language, particularly in the field of Natural Language Processing (NLP), With an extensive range of libraries and tools specifically designed for text analysis. Python has become suitable for researchers, developers, and data scientists working with textual data.

NLTK :

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on the interaction between computers and human language. It involves the

development of algorithms and models that enable computers to understand, interpret, and generate human language in a meaningful way. NLP has a wide range of applications, including machine translation, data analysis, text classification, and information extraction. NLTK stands for Natural Language Tool Kit, is a python package that we can use for NLP.

Python offers a rich ecosystem of libraries that facilitate various aspects of NLP. Some of the key libraries include:

1. NLTK (Natural Language Toolkit): NLTK is one of the most widely used libraries for NLP. It provides a comprehensive set of tools and resources for tasks such as tokenization, stemming, tagging, Named entity recognition, etc

2. spaCy: spaCy is a modern and efficient library for NLP. It offers fast and accurate tokenization, tagging, Named entity recognition, and dependency parsing. spaCy is known for its speed and scalability, making it suitable for processing large volumes of text data.

3. TextBlob: TextBlob is a simple library built on top of NLTK. It provides an easy-to-use interface for common NLP tasks such as sentiment analysis and language translation.

- Gensim: Gensim is a library for topic modeling and document similarity analysis. It provides efficient implementations of popular algorithms such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Word2Vec. Gensim allows users to extract meaningful topics from a collection of documents and measure the similarity between documents based on their content.

Data Preprocessing with Python

Before applying any NLP techniques, it is mandatory to preprocess the text. Python provides several libraries and techniques for data preprocessing, such as:

1)**Tokenization:** Tokenization is the process of splitting text into individual words or tokens. NLTK and SQlite offer efficient tokenization algorithms that handle complex cases such as contractions, punctuation, and special characters.

2)**Stopword Removal:** Stopwords are common words that do not carry much meaning and can be safely ignored in text analysis. NLTK provides a list of stopwords for various languages, which can be easily filtered from the text data.

3)**Lemmatization and Stemming:** Lemmatization and stemming are techniques used to reduce words to their base or root form. NLTK and SQlite offer lemmatization and stemming algorithms that normalize words and reduce their variations.

4)**Text Cleaning**: Text cleaning involves removing unwanted characters, numbers, and symbols from the text data. Regular expressions, available in the Python library, are commonly used for text cleaning tasks.

Text mining & Analytics Techniques with Python

Once the data is preprocessed, Python provides a wide range of techniques for text mining & analytics. Here are some commonly used techniques:

1. **Sentiment Analysis**: Sentiment analysis aims to determine the emotional tone of a piece of text, whether it is positive, negative, or neutral. TextBlob provides a

straightforward interface for sentiment analysis, allowing users to analyze the sentiment of text documents or individual sentences.

2. **Named Entity Recognition:** Named Entity Recognition (NER) identifies and classifies named entities in text, such as people, organizations, locations, and dates. NLTK and SQlite offer pre-trained models for NER, which can be used to extract valuable information from text.

3. **Topic Modeling:** Topic modeling is an unsupervised learning technique that discovers hidden thematic structures in a collection of documents. Gensim provides efficient implementations of popular topic modeling algorithms such as LSA and LDA, allowing users to extract meaningful topics from text data.

4. **Text Classification:** Text classification involves assigning predefined categories or labels to text documents based on their content. Python provides libraries such as scikit-learn, which offer various machine learning algorithms for text classification tasks.

6)CASE STUDY ON CLASSIFYING REDDIT POSTS

The process of data from Reddit's API is simple: it's just a basic request set up since they do not require a key to access the API. Generally Reddit posts classifiers are as follows…..

- Define the problem
- Gather & Clean the data
- Explore the data

- Model the data
- Evaluate the model
- Answer the problem

1)Define the Problem

In this step we need to creating a classification model that can distinguish which of two subreddits a post belongs to Reddit back-end developer defined every post and replaced the subreddit field with "(·L·)". As a result, none of the subreddit links will populate with posts until the subreddit fields of each post are re-assigned.

2)Gather & Clean the Data

Data gathering process involves using the requests library to loop through requests to recive data using Reddit's API. To get posts we need to add.json to the end of the url. Reddit also used the time.sleep()function at the end of loop to allow for a one second break in between requests.

After getting Reddit posts clean the data in their respective DataFrames. Here we checked for duplicate and null values, both of which occurred. For duplicate values we eliminate them by utilizing the drop_duplicates() function. Null values only occurred in Post Text column, this happens when a Reddit user decides to use only the title field.

3)Explore the data

In this process we generated & to get a visual understanding of the frequencies of words and how their commonality across subreddits.

4)Model the data

In this step we can modeling process by creating vectors a and b then splitting data into training and test sets. Then move to feature engineering process by instantiating two Count Vectorizers for my Post Text and Title. Count Vectorizer converts a collection of text documents to a matrix of token counts. The hyper parameters (arguments) are

- stop_words='english' (Post Text & Title)
- strip_accents='ascii' (Post Text & Title)
- ngram_range=(1, 6), min_df=.03 (*Post Text*)
- ngram_range=(1, 3), min_df=.01 (*Title*)

Stop words removes words that commonly appear in the English language. Strip accents removes accents and performs other character normalization. Min_df ignores terms that have a document frequency strictly lower than the given threshold.An n-gram is just a string of n words in a row.

- Decision Trees & Random Forests

The difference in variations for these two models was the criterion parameter. One was set to 'gini' (Gini impurity) while the other was set to 'entropy' (information gain).

UNIT-5

DATA VISUALIZATION & PROTOTYPE APPLICATION DEVELOPMENT

1)DATA VISUALIZATION OPTIONS

Data visualization is the graphical representation of information and data. By using visual elements like tables, charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

- **Tables:** This consists of rows and columns used to compare variables.
- **Charts**: Bar charts, line charts, pie charts, etc.
- **Graphs**: Scatter plots, histograms, etc.
- **Maps**: Geographic maps, heat maps, etc.
- **Dashboards**: Interactive platforms that combine multiple visualizations..

The primary goal of data visualization is to make data more accessible and easier to understand, allowing users to identify patterns, trends, and outliers quickly. This is particularly important in the big data, where the large volume of information can be existed and it is difficult to understand without using effective visualization techniques.

Open-source visualization tools

Access to data visualization tools has never been easier. Open source libraries, such as D3.js, provide a way for analysts to present data in an interactive way. Some of the most popular open source visualization libraries include:

- **D3.js:** It is a front-end JavaScript library for producing dynamic, interactive data visualizations in web browsers. D3.js uses HTML, CSS, and SVG(scalar vector graphis) to create visual representations of data that can be viewed on any browser. It also provides features for interactions and animations.
- **ECharts:** A powerful charting and visualization library that offers an easy way to add interactive, and highly customizable charts to products, research papers, presentations, etc. Echarts are based on JavaScript and ZRender, a lightweight canvas library.

- **Vega:** Vega defines itself as “visualization grammar,” providing support to customize visualizations across large datasets which are accessible from the web.
- **deck.gl:** It is part of Uber's open source visualization framework suite. deck.gl is a framework, which is used for exploratory data analysis on big data. It helps to build high-performance GPU-powered (Graphics processing unit) visualization on the web.

2)CROSS FILTER

Data visualization is the graphical representation of information and data. By using visual elements like tables, charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

- **Tables:** This consists of rows and columns used to compare variables.
- **Charts**: Bar charts, line charts, pie charts, etc.
- **Graphs**: Scatter plots, histograms, etc.
- **Maps**: Geographic maps, heat maps, etc.
- **Dashboards**: Interactive platforms that combine multiple visualizations

Cross-filtering makes it easier for viewers to interact with the dashboard's data and see how reports interact with one another. Cross filter is a JavaScript library for exploring large datasets in the browser. Cross filter supports extremely fast interaction with coordinated views, even with datasets containing a million or more records.

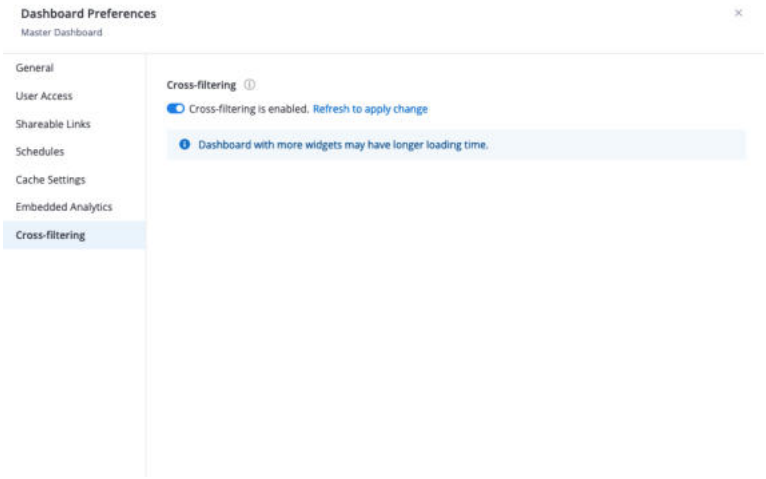
How to use cross-filtering :

1)Enable cross-filtering for each dashboard

You can enable/disable Cross-filtering for each individual dashboard.

- Go to Dashboard Preferences -> Cross-filtering and switch the toggle to ON.

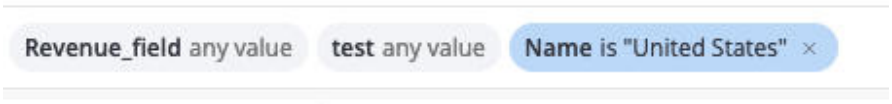
- Refresh the dashboard for the change to take effect.



2)Apply cross-filtering

To apply cross-filtering, click on any data point in chart. The following events will happen:

- Blue filters on the filter panel which indicates that all the values related to the clicked data point are being applied to all reports in dashboard.



- The reports are filtered by the blue filters. You might see a visual change in the charts,it appear in the right side of the report name, which indicates that the blue filters are applying to that report.

Examples:

Let's say you have a Sales Overview dashboard, which contains 3 reports: Product overview, Customer Profile, and Price. When you click on smart phones in Product overview (1), all the dimensions related to smartphone will be used as filter values for Customer Profile and Price.

3)THE JAVA SCRIPT MAPREDUCE LIBRARY

Map, reduce, and filter are array methods in JavaScript. Each one will iterate over an array and perform a transformation or computation. Each will return a new array based on the result of the function.

Map

The map() method is used for creating a new array from an existing one, applying a function to each one of the elements of the first array.

Syn : var new_array = arr.map(function callback( ) {

```
  // Return value for new_array

}[, thisArg])
```

In the callback,. Usually some action is performed on the value and then a new value is returned.

Example

In the following example, each number in an array is doubled.

```
const numbers = [1, 2, 3, 4];

const doubled = numbers.map(element =>element * 2);

console.log(doubled); // [2, 4, 6, 8]
```

Filter

The filter() method takes each element in an array and it applies a conditional statement. If this conditional returns true, the element gets pushed to the output array. If the condition returns false, the element does not get pushed to the output array.

Syn : var new_array = arr.filter(function callback( ) {

```
  // Return true or false
}[, thisArg])
```

The syntax for filter is similar to map, except the callback function should return true to keep the element, or false otherwise. In the callback, only the element is required.

In the following example, odd numbers are "filtered" out, leaving only even numbers.

```
 const numbers = [1, 2, 3, 4];
const evens = numbers.filter(element => element % 2 === 0);
console.log(evens); // [2, 4]
```

Reduce

The reduce() method reduces an array of values down to just one value. To get the output value, it runs a reducer function on each element of the array.
Syn : arr.reduce(callback[, initialValue]).

The callback argument is a function that will be called once for every element in the array. This function takes two arguments.

1) accumulator - the returned value of the previous iteration.
2) currentValue - the current element in the array

The initialValue argument is optional. If provided, it will be used as the initial accumulator value in the first call to the callback function.
The following example adds every number together in an array of numbers.

```
 const numbers = [1, 2, 3, 4];

const dec = numbers.reduce(element=>4) {

console.log(dec); // [3]
```

4)CREATING AN INTERACTIVE DASHBOARD WITH DC.JS

There are different ways to build an interactive dashboard.DC.js is a charting library built on top of D3.js and works natively with crossfilter, which is a popular JavaScript library used to explore millions of records in a short time.So DC.js is a JavaScript library used to make interactive dashboards in JavaScript.

By using Excel we can create a dynamic dashboard with pivot tables. But in various cases such as finance and insurance where the complexity was too big to handle.Another solution is to buy a specific tool like qlikview. This application allows you to build beautiful and interactive dashboards. But this tool is not free.

Finally, we use javascript there is crossfilter.js**.** This free javascript library allows you to manage a complex dataset really quickly**.** You can easily filter dataset . This will be really helpful to create our interactive dashboard.

The main advantage of dc.js is that the graphs created are dynamical, perfect for a dynamic dashboard. All your graphs are related to a crossfilter dataset. So when you click on a graph, it filters this dataset with the selected group and renders all your graph automatically.

How to work with dc.js:

In order to create a new dynamic graph you need four things:

1)The template of the graph

2)A crossfilter dataset

3)A crossfilter dimension

4)A crossfilter group.

The following pie chart is the most suited graph. Here is the template:

```
<div id="graph"></div>
<script>
var pieChart = dc.pieChart("#graph");
pie1
    .width(200)
    .height(200)
    .innerRadius(25)
    .label(function(d) {
    return d.key + ': ' + d.value;
  })
    .dimension(dimensionCategory)
    .group(quantityByCategory);
</script>
```

Advantages of dc.js :

dc.js is a great library in order to make a dynamic dashboard**.** For a dataset of 1MB, with 50 diffrent dimensions and groups, changing simultaneously the filter of all those dimensions take less than 0.1 seconds.

Disadvantage of dc.js:

The main disadvantage is we can't use easily, if you don't know how to deal with crossfilter.js.and also its difficult to combine dc.js with another graph library.

5)DASH BOARD DEVELOPMENT TOOLS

1. Microsoft Power BI

Microsoft Power BI is a web and cloud-based analytics and data visualisation platform.It is available as a desktop or mobile application with interactive reports, real-time dashboards, and datasets that can connect to data sources.

One of Power BI's unique features is its Q&A interface. Using the natural language you would use when asking a question it produce related answers , you can input any question to scout your data for specifics, and the Power BI technology will use suggestions, re-phrasings and autofill to present the answer.

2. Tableau Public

Tableau Public is part of the Tableau software that offers three different options: Tableau Public, Tableau Reader and Tableau Desktop. Tableau Public is the free package that offers data visualisation, analysis and business intelligence for companies.Tableau Public publishes your visualisations – maps, graphs, charts and other outputs – on the web through a simple user interface and live dashboard.

3. AddMaple

AddMaple provides flexible to data analysis and visualization, making it good choice for individuals and small teams looking to explore data without cost. The free version allows for interactive report writing and data exploration.Users can import up to 100 rows of data.

4. Databox

Databox is a cloud-based business analytics platform that is available on desktop and mobile, including iOS and Android. We can connect data directly – from Salesforce, Google Analytics, Hubspot, Facebook and others.

Databox has clean interface with a drag-and-drop editor for building custom dashboards,and also has also dashboard templates.It provides access including over 200 pre-built dashboard templates.

5.Vizzlo

Vizzlo is a simple charting and infographics platform with tools and applications for building different types of visual reports. It offers several charting options including classic bar and pie charts, Gantt charts and waterfall charts.

6.Piktochart

Piktochart is used by students, teachers, bloggers and marketers for telling stories with data, creating flyers or posters and publishing infographics and presentations.It has a drag and drop interface with useful design features including interactive maps,videos, hyperlinks, graphics,and templates.

PIKTOCHART

7. Canva

Canva is a popular graphic design platform than the other tools.It is widely used for data visualisation and infographic creation as well as presentations, and posters.

It includes access to a large library of graphics, fonts and photos, as well as design templates.And it is widely used for creating social media content. The free package includes over 100 design types, including posters and letters, and thousands of graphics and photos.

Canva

6)APPLYING THE DATA SCIENCE PROCESS FOR REAL WORLD PROBLEMS SOLVING SCENARIOS AS A DETAILE CASE STUDY

In this case study, we will explore how a data science process model can help to solve real world problems. This case study is a powerful example of how data science can transform a business, increasing efficiency, and improving decision-making.

1. Search Engines

Ever wonder how Google seems to know exactly what you're looking for? It's not magic; it's data science! Data science plays an important role in the functioning of search engines, enhancing their efficiency and accuracy in delivering relevant results.

2. Ecommerce

Data science is the key ingredient behind fast and efficient deliveries in the Ecommerce industry. Companies like Amazon and Flipkort depends on data-driven algorithms to optimize routes, considering factors like traffic patterns and weather conditions.

3. Finance

The financial sector is completely depends on data science, which is used to protect your money and make better investment decisions. Data science has revolutionized the financial sector, providing insights that drive decision-making, risk management, and customer service.

5. Healthcare

Data science has emerged as a pivotal element in the healthcare industry, enhancing patient care, improving operational efficiency, and driving medical research. Data science can analyze patient data to predict disease risk and promote preventive care. Based on medical history, and lifestyle factors, data science can help healthcare providers develop personalized treatment plans.

6. Image and Speech Recognition

Data science, particularly its subfields of machine learning and deep learning, plays an essential role in image and speech recognition. Machine learning algorithms can identify and locate objects within an image. This is used in various applications, from automated surveillance systems to self-driving cars.Deep learning techniques are used to identify and verify individuals based on their facial features. This is commonly used in security systems and social media applications.

Data science is used to convert spoken language into written text. This is used in applications like voice assistants, dictation software, and automated customer service systems.

7. Gaming

In the gaming industry, data science is used to create more effective and personalized experiences for players. Data science is making significant contributions to the gaming industry, enhancing game development, player experiences, and business performance.

8. Social Media

Data science plays a critical role in the operation and growth of social media platforms. It helps in understanding user behavior, content optimization, targeted advertising. Data science helps analyze user behavior to understand how users interact with the platform. This includes what they like, share, comment on, the time they spend on the platform, etc. Such insights can inform platform design decisions and content recommendation algorithms.

9. Sports Analytics

Data science is revolutionizing the sports world by providing teams and athletes with valuable insights to improve performance and gain a competitive edge. Data science has become a key player in the sports industry, significantly transforming how teams train, perform, and make strategic decisions.

10. Government

Data science is increasingly being utilized in government for various purposes, including policy-making, public service delivery, and improving operational efficiency.

11. Education

Data science has made significant contributions to the field of education, enhancing teaching methods, personalizing learning, and improving educational outcomes.