ANNMACHARYA INSTITUTE OF TECHNOLOGY & SCIENCES,

RAJAMPET

(AUTONOMOUS)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATASCIENCE

LECTURE NOTES

NAME OF THE FACULTY: P.RENUKA

CLASS: II B.TECH II SEM

NAME OF THECOURSE: INTRODUCTION TO DATASCIENCE

SUBJECT CODE: 24AID41T

ACADEMIC YEAR:2025-2026

UNIT-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.

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

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 II – Handling Large Data & Machine Learning in Data Science

---

1. Handling Large Data

Large data refers to datasets that are **high in volume, velocity, and variety**, making them difficult to store, process, and analyze using traditional systems.

Challenges of Large Data

- High memory consumption
- Long processing time
- Scalability issues
- Data quality problems
- Data integration from multiple sources

2. Applications of Machine Learning in Data Science

Machine Learning (ML) enables systems to **learn from data and improve automatically** without explicit programming.

Key Applications

- Recommendation systems (Netflix, Amazon)
- Spam and fraud detection
- Image and speech recognition
- Predictive analytics
- Healthcare diagnosis
- Financial risk assessment
- Cybersecurity (malicious URL detection)

3. Role of Machine Learning in Data Science

Machine Learning is a **core component** of Data Science.

Role of ML

- Converts raw data into actionable insights
- Builds predictive and classification models
- Automates decision-making
- Identifies hidden patterns in large datasets
- Improves accuracy over time with more data

4. Python Tools for Data Science – Scikit-learn (sklearn)

`scikit-learn` is a popular Python library for ML.

Features

- Easy-to-use API
- Supports classification, regression, clustering
- Built-in datasets
- Model evaluation and validation tools

Common Modules

- `sklearn.linear_model`
- `sklearn.tree`
- `sklearn.ensemble`
- `sklearn.model_selection`
- `sklearn.metrics`

5. Model Process for Feature Engineering

Feature engineering is the process of **selecting, creating, and transforming variables** to improve model performance.

Steps

1. Feature selection
2. Feature extraction
3. Feature transformation
4. Handling missing values
5. Encoding categorical data
6. Feature scaling (normalization, standardization)

Importance

- Improves accuracy
- Reduces overfitting
- Reduces model complexity

6. Model Selection

Model selection involves choosing the **best ML algorithm** for a given problem.

Factors Affecting Model Selection

- Nature of data
- Size of dataset
- Problem type (classification/regression)
- Interpretability

- Accuracy requirements

Common Models

- Linear Regression
- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines
- K-Means Clustering

7. Validation and Prediction

Model Validation

Ensures the model performs well on unseen data.

Validation Techniques

- Train-test split
- Cross-validation
- K-fold validation

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- Mean Squared Error (MSE)

Prediction

Using a trained model to predict outcomes for new data.

8. Types of Machine Learning

1. Supervised Learning

- Uses labeled data
- Examples: Classification, Regression

2. Unsupervised Learning

- Uses unlabeled data
- Examples: Clustering, Dimensionality reduction

3. Semi-Supervised Learning

- Uses small labeled data + large unlabeled data

4. Reinforcement Learning

- Learns via rewards and penalties

9. Semi-Supervised Learning

Semi-supervised learning lies between supervised and unsupervised learning.

Characteristics

- Reduces labeling cost
- Improves performance over unsupervised learning
- Useful when labeled data is limited

Applications

- Text classification
- Web content categorization
- Speech recognition

10. Problems in Handling Large Data

- Memory limitations
- Slow I/O operations
- Data imbalance
- High dimensionality
- Noise and redundancy

General Techniques for Handling Large Data

- Data sampling
- Dimensionality reduction (PCA)
- Distributed computing
- Parallel processing
- Incremental learning
- Batch processing

Programming Tips for Dealing with Large Data

- Use generators instead of lists
- Process data in chunks
- Use efficient data structures
- Avoid unnecessary loops
- Optimize memory usage
- Use vectorized operations
- Prefer libraries like NumPy and Pandas

Case Study: Predicting Malicious URLs

Objective

Identify whether a URL is malicious or safe.

Steps

1. Data collection (URLs dataset)
2. Feature extraction (URL length, special characters)
3. Data preprocessing
4. Model training (Logistic Regression, Random Forest)
5. Validation and evaluation
6. Deployment for real-time prediction

Applications

- Cybersecurity
- Spam detection
- Phishing prevention

14. Case Study: Building Recommendation Systems

Objective

Recommend products or content to users.

Types

- Content-based filtering
- Collaborative filtering
- Hybrid recommendation systems

Steps

1. User behavior data collection
2. Feature extraction
3. Model selection (KNN, Matrix Factorization)
4. Training and validation
5. Recommendation generation

Applications

- E-commerce
- Streaming platforms
- Online learning systems

UNIT III – Distributed Data Storage & Processing

---

1. Distributed Data Storage

Distributed data storage refers to storing data across **multiple machines (nodes)** instead of a single system.

Need for Distributed Storage

- Handles very large datasets
- Improves scalability
- Provides fault tolerance
- Enhances availability and performance

Characteristics

- Data is partitioned across nodes
- Data replication for reliability
- Supports parallel access

Examples

- HDFS (Hadoop Distributed File System)
- Amazon S3
- Google File System

2. Processing with Hadoop Framework

Apache Hadoop is an **open-source framework** used for **distributed storage and processing** of big data.

Core Components of Hadoop

1. **HDFS (Storage Layer)**
2. **MapReduce (Processing Layer)**
3. **YARN (Resource Management Layer)**

2.1 Hadoop Distributed File System (HDFS)

- Stores large files by splitting them into blocks
- Blocks are replicated across nodes
- Master–slave architecture

Components

- NameNode – metadata management
- DataNode – stores actual data
- Secondary NameNode – backup support

2.2 MapReduce Processing Model

MapReduce processes data in **parallel**.

Map Phase

- Input data is split
- Key–value pairs are generated

Reduce Phase

- Aggregates and processes output from Map phase

Advantages

- High scalability
- Fault tolerance
- Cost-effective

3. Case Study: Risk Assessment for Loan Sanctioning

Objective

To assess the risk level of loan applicants using Data Science techniques.

Data Used

- Applicant income
- Credit history
- Employment details
- Loan amount
- Previous defaults

Process

1. Data collection from banks and financial records
2. Data preprocessing and cleansing
3. Feature selection

4. Model training using ML algorithms
5. Risk classification (Low, Medium, High)

Benefits

- Faster loan approval
- Reduced financial risk
- Improved decision-making

4. ACID Principles of Relational Databases

ACID ensures **reliable database transactions**.

ACID Properties

1. **Atomicity** – Transaction is all or nothing
2. **Consistency** – Database remains valid after transaction
3. **Isolation** – Concurrent transactions do not interfere
4. **Durability** – Changes persist even after failures

Importance

- Maintains data integrity
- Ensures accurate financial and business operations

5. CAP Theorem

CAP theorem defines limitations of distributed systems.

CAP Properties

- **Consistency (C)** – All nodes see same data
- **Availability (A)** – System responds to every request
- **Partition Tolerance (P)** – System continues during network failures

Key Rule

A distributed system can satisfy **only two of the three** at the same time.

Examples

- CP systems: HBase
- AP systems: Cassandra
- CA systems: Traditional RDBMS (not distributed)

6. BASE Principles of NoSQL Databases

BASE is an alternative to ACID for distributed systems.

BASE Properties

- **Basically Available** – System remains available
- **Soft State** – Data may change over time
- **Eventual Consistency** – Data becomes consistent eventually

Advantages

- Better scalability
- High availability
- Suitable for big data applications

7. Types of NoSQL Databases

1. Key-Value Stores

- Store data as key–value pairs
- Examples: Redis, DynamoDB

2. Document Stores

- Store semi-structured documents
- Examples: MongoDB, CouchDB

3. Column-Family Stores

- Store data in columns
- Examples: Cassandra, HBase

4. Graph Databases

- Store data as nodes and relationships
- Examples: Neo4j

---

8. Case Study: Disease Diagnosis and Profiling

Objective

To predict diseases and profile patients using large healthcare datasets.

Data Used

- Patient demographics
- Medical history
- Symptoms
- Lab test results

Process

1. Data acquisition from hospitals
2. Data integration and preprocessing
3. Feature extraction
4. ML model development
5. Disease prediction and patient profiling

Benefits

- Early disease detection
- Personalized treatment
- Improved healthcare outcomes

---

9. Comparison: ACID vs BASE

Feature	ACID	BASE
Consistency	Strong	Eventual
Availability	Lower	High
Scalability	Limited	High
Use Case	Banking systems	Big data apps

UNIT IV – Tools and Applications of Data Science

---

1. Tools and Applications of Data Science

Data Science uses a combination of **tools, techniques, and platforms** to extract knowledge from data.

Common Applications

- Business intelligence
- Healthcare analytics
- Social media analysis
- Recommendation systems
- Fraud detection
- Natural Language Processing (NLP)

Popular Data Science Tools

- Python, R
- Jupyter Notebook
- SQL databases
- NoSQL databases
- Big data tools

2. Graph Databases

Graph databases store data using **nodes, relationships, and properties**, making them ideal for highly connected data.

Advantages

- Efficient relationship traversal
- Flexible schema
- High performance for connected data
- Easy representation of real-world networks

Use Cases

- Social networks
- Fraud detection
- Recommendation engines
- Knowledge graphs

3. Introducing Neo4j

Neo4j is a **popular graph database** used to manage connected data.

Key Features

- Property graph model
- ACID compliant
- High scalability
- Fast graph traversal

Components

- Nodes – entities
- Relationships – connections
- Properties – attributes

4. Graph Query Language – Cypher

Cypher is a **declarative query language** used in Neo4j.

Characteristics

- Easy to read and write
- SQL-like syntax
- Designed for graph traversal

Basic Cypher Operations

- Create nodes and relationships
- Retrieve data using pattern matching
- Update and delete graph elements

Example Queries

- Create node
- Match relationships
- Return connected nodes

5. Applications of Graph Databases

Major Applications

- Social media platforms
- Network security
- Supply chain management
- Recommendation systems
- Knowledge representation

Benefits

- Faster queries for relationships

- Better insight into connected data
- Real-time graph analytics

6. Text Mining and Analytics

Text mining is the process of **extracting meaningful information from unstructured text data**.

Applications

- Sentiment analysis
- Topic modeling
- Spam detection
- Document classification

7. Python Libraries for Text Mining

7.1 NLTK (Natural Language Toolkit)

NLTK is a **popular Python library** for NLP.

Features

- Tokenization
- Stop-word removal
- Stemming and lemmatization
- POS tagging

Applications

- Text classification
- Language processing
- Sentiment analysis

7.2 SQLite for Text Analytics

SQLite is a **lightweight relational database**.

Features

- Server-less database
- Stores structured text data
- Fast data retrieval

Applications

- Storing processed text data
- Small-scale analytics
- Embedded systems

8. Text Preprocessing Steps

- Tokenization
- Lowercasing
- Stop-word removal
- Stemming / Lemmatization
- Removing punctuation

9. Case Study: Classifying Reddit Posts

Objective

Automatically classify Reddit posts into categories or sentiment types.

Data Source

- Reddit posts and comments

Process

1. Data collection using Reddit API
2. Text preprocessing using NLTK
3. Feature extraction (Bag of Words, TF-IDF)
4. Data storage using SQLite
5. Model training (Naive Bayes, SVM)
6. Model evaluation and classification

Benefits

- Community moderation
- Topic discovery
- Sentiment analysis

UNIT V – Data Visualization and PAD

---

1. Data Visualization

Data visualization is the **graphical representation of data** to help users understand patterns, trends, and insights.

Importance of Data Visualization

- Simplifies complex data
- Improves decision making
- Identifies trends and outliers
- Enhances communication of insights

2. PAD in Data Science

PAD represents **Descriptive, Analytical, and Prescriptive** approaches to data usage.

Descriptive Analytics

- Answers *What happened?*
- Uses charts, summaries, dashboards

Analytical (Diagnostic) Analytics

- Answers *Why did it happen?*
- Uses comparisons and correlations

Prescriptive Analytics

- Answers *What should be done?*
- Uses predictions and recommendations

3. Data Visualization Options

Common Visualization Types

- Bar charts
- Line charts
- Pie charts
- Scatter plots
- Histograms
- Heat maps

Advanced Visualization

- Interactive dashboards
- Geographic maps
- Network graphs

4. Crossfilter

Crossfilter is a **JavaScript library** for exploring large multivariate datasets in the browser.

Features

- Fast filtering of large datasets
- Works in real time
- Supports multiple dimensions

Applications

- Interactive dashboards
- Data exploration tools

5. JavaScript MapReduce Library

MapReduce is a programming model for **processing large datasets**.

JavaScript MapReduce

- Performs data aggregation
- Used for client-side data processing
- Supports parallel data computation

Benefits

- Efficient handling of large datasets
- Simplifies data transformation

6. Creating an Interactive Dashboard with dc.js

dc.js is a **JavaScript charting library** built on D3.js and Crossfilter.

Features

- Interactive charts
- Automatic filtering
- Real-time updates

Steps to Create Dashboard

1. Load dataset
2. Define dimensions and groups
3. Create charts (bar, pie, line)
4. Enable cross-filtering
5. Render dashboard

7. Dashboard Development Tools

Popular Tools

- Tableau
- Power BI
- QlikView
- Grafana
- Google Data Studio

Web-Based Tools

- D3.js
- dc.js
- Chart.js

Selection Criteria

- Data size
- Interactivity
- Ease of use
- Integration support

8. Best Practices for Dashboards

- Use clear labels and legends
- Avoid clutter
- Choose appropriate charts
- Maintain consistency
- Focus on key metrics

9. Applying Data Science Process – Real World Case Study

Case Study: Retail Sales Analysis Dashboard

Problem Statement

Analyze sales data to improve business decisions.

Data Science Process

1. **Problem Definition** – Understand business goals

2. **Data Collection** – Sales, customer, and product data
3. **Data Cleaning** – Remove missing and inconsistent data
4. **Data Analysis** – Identify trends and patterns
5. **Visualization** – Build interactive dashboards
6. **Insights & Decisions** – Improve pricing and inventory

Outcome

- Increased sales performance
- Improved customer satisfaction
- Better forecasting