Unit 3
Data and Knowledge Management

Introduction

Managing Data

Database Approach

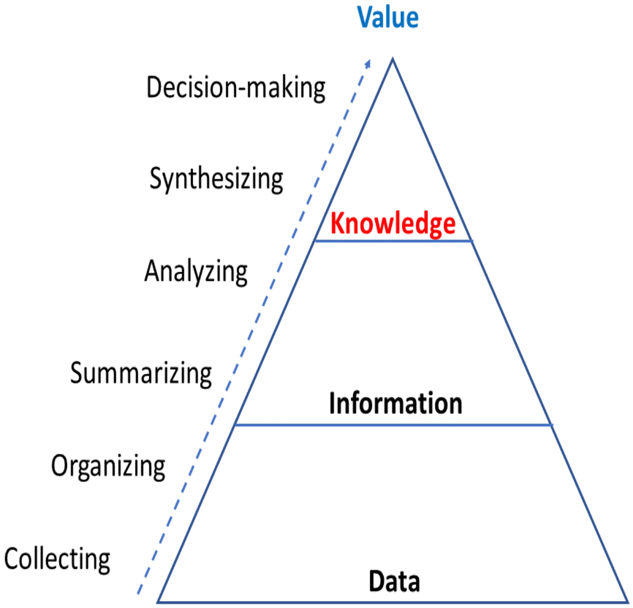
Big Data

Data Warehouses and Data Mart

Knowledge Management.

Data and Knowledge Management

- Data and knowledge management involve the processes and systems used to organize, store, and retrieve information, enabling organizations to make data-driven decisions and leverage knowledge effectively.
- Data and knowledge management (DKM) systems collect, manage, and provide controlled access to data and knowledge resources.
- These systems may also provide critical analytical and visualization capabilities to support research and decision processes. Data within the DKM may be at any stage of its lifecycle.



- **Data**: Raw facts and figures without context (e.g., 100, "John").
- **Information**: Processed data with meaning (e.g., "John scored 100 in mathematics").
- **Knowledge**: Insights derived from information to guide decisions.

Example:

- Data: "500 units sold."
- Information: "500 units of product X were sold in region Y last quarter."
- Knowledge: "Product X has high demand in region Y during the summer."

- ➢ Data and Knowledge Management is a critical area within business information systems, focusing on how data and information are collected, stored, processed, and utilized to support organizational decision-making and operations.
- ➢ It encompasses various topics, including:

1. Data Management

- **Definition**: The process of collecting, organizing, storing, and maintaining data for efficient and secure access and analysis.

- **Components**:
  - **Database Management Systems (DBMS)**: Software for storing and retrieving users' data while considering security, accuracy, and consistency.
  - **Data Warehousing**: Centralized repositories designed to store integrated data from multiple sources for analytical purposes.
  - **Data Integration**: Combining data from different sources into a unified view.
  - **Data Governance**: Establishing policies and procedures to ensure data quality, security, and compliance.

2. Knowledge Management (KM)

- **Definition**: A systematic process of capturing, distributing, and effectively using knowledge to enhance organizational learning and decision-making.
- **Key Aspects**:
  - **Knowledge Capture**: Identifying and documenting explicit (easily transferable) and tacit (intuitive, experience-based) knowledge.
  - **Knowledge Sharing**: Facilitating communication and collaboration through tools like intranets, knowledge bases, and forums.
  - **Knowledge Utilization**: Applying knowledge in problem-solving and innovation.

- **Tools**:
  - Knowledge Repositories
  - Expert Systems
  - Decision Support Systems

3. Big Data and Analytics

- Utilizing vast amounts of structured and unstructured data to generate insights.
- **Technologies**:
  - Hadoop
  - Spark

- **Applications**: Predictive analytics, business intelligence, and data mining.

4. The Role of IT in Knowledge Management

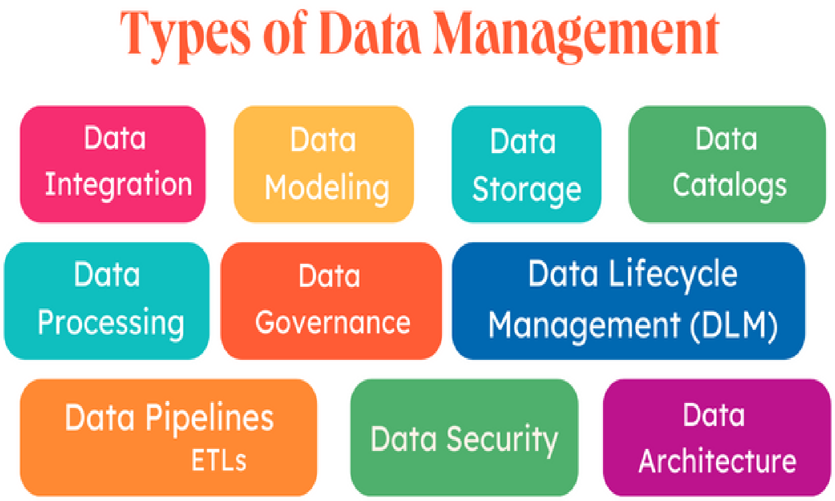
- Enhancing the accessibility and usability of knowledge through advanced technologies:
  - Cloud computing for data storage and collaboration.
  - Artificial intelligence for automating knowledge processing and decision-making.
  - Social computing to facilitate community-driven knowledge sharing.

Managing Data

Effective data management ensures accuracy, consistency, security, and accessibility of data within an organization.

Key Aspects:

1. **Data Quality**: Ensuring data is accurate, complete, and reliable.
2. **Data Governance**: Policies and practices to manage data responsibly.
3. **Data Security**: Protecting data from unauthorized access and breaches.



Case Analysis:

- **Scenario**: A retail company experiences inaccurate sales reporting due to inconsistent data entries.
- **Solution**: Implement a centralized data management system with validation rules to ensure accuracy.

➢ Managing data is a structured process of handling data efficiently and effectively throughout its lifecycle, ensuring its availability, integrity, and usability for organizational needs.

- Effective data management is essential for informed decision-making, operational efficiency, and strategic planning.

- Below are the key components and practices for managing data:

1. Data Lifecycle Management

The data lifecycle encompasses all stages from data creation to disposal:

- **Data Creation**: Generating data through various means, such as transactions, sensors, or user input.
- **Data Storage**: Securely storing data in databases, data warehouses, or cloud platforms.
- **Data Usage**: Utilizing data for analytics, reporting, and operational processes.
- **Data Archiving**: Retaining infrequently accessed data for historical or compliance purposes.
- **Data Disposal**: Securely deleting data when no longer needed.

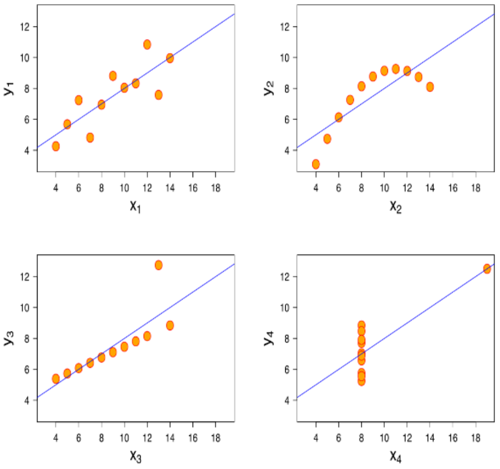
2. Data Storage and Organization

- **Relational Databases**: Structured data stored in tables using SQL (e.g., MySQL, PostgreSQL).
- **Non-Relational Databases**: For unstructured or semi-structured data (e.g., MongoDB, Cassandra).
- **Data Lakes**: Repositories storing raw data in its native format.
- **Data Warehouses**: Centralized systems designed for analytical queries and reporting.

3. Data Quality Management

Maintaining high-quality data involves:

- **Accuracy**: Ensuring data is correct and error-free.



- **Completeness**: Avoiding missing or incomplete data entries.
- **Consistency**: Standardizing data formats and values across sources.
- **Timeliness**: Keeping data up-to-date for relevant use.

4. Data Integration

Combining data from multiple sources into a unified view:

- **ETL (Extract, Transform, Load)**: Extracting data, transforming it for analysis, and loading it into storage.
- **API Integration**: Using APIs to connect disparate systems.
- **Data Virtualization**: Accessing and querying data without physical movement.

5. Data Security and Privacy

Protecting sensitive data from unauthorized access or breaches:

- **Access Controls**: Implementing role-based permissions.
- **Encryption**: Securing data in transit and at rest.
- **Compliance**: Adhering to regulations like GDPR, HIPAA, or CCPA.

6. Data Governance

Establishing policies and frameworks to manage data assets:

- **Roles and Responsibilities**: Defining who manages and accesses data.
- **Data Stewardship**: Assigning individuals to ensure data quality and compliance.
- **Metadata Management**: Documenting data definitions, origins, and usage.

7. Data Analytics and Reporting

Using data to generate insights and reports:

- **Business Intelligence Tools**: Tools like Tableau, Power BI, or Qlik for visualization.

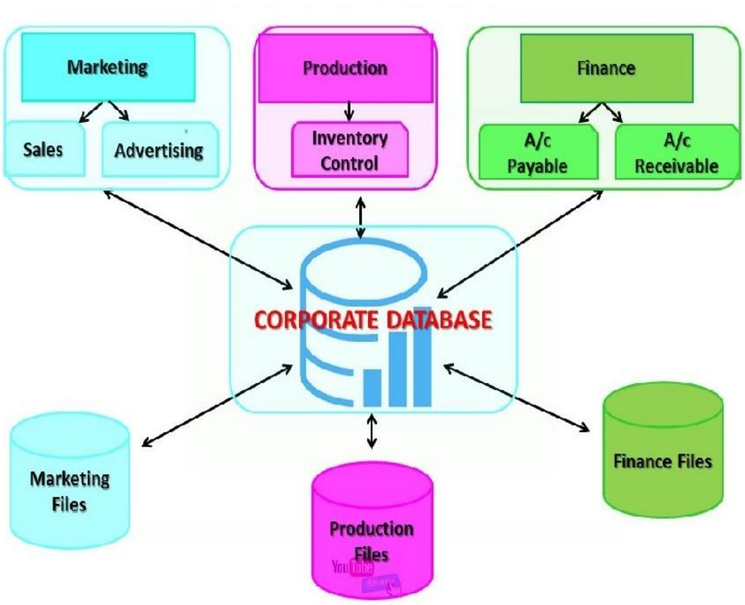
- **Predictive Analytics**: Forecasting trends and outcomes.
- **Real-Time Analytics**: Analyzing data streams instantly.

8. Emerging Technologies in Data Management

- **Cloud Data Management**: Storing and processing data on platforms like AWS, Azure, or Google Cloud.
- **Big Data Solutions**: Leveraging technologies like Hadoop and Spark.
- **Artificial Intelligence (AI)**: Enhancing data processing and insights through machine learning.

Database Approach

A database is an organized collection of data that can be accessed, managed, and updated efficiently.



Features:

- **Data Centralization**: Stores data in one location.
- **Minimized Redundancy**: Reduces duplication of data.
- **Improved Data Integrity**: Ensures data consistency.
- **Efficient Querying**: Enables quick retrieval of data using SQL.

Example:

- A library system uses a database to track books, members, and loan records.

Case Analysis:

- **Scenario**: A university uses spreadsheets for student records, leading to duplication and errors.
- **Solution**: Transition to a relational database system to maintain a single source of truth for all student data.

- The **database approach** is a systematic method of managing data that replaces traditional file systems with a unified repository to store and manipulate data efficiently.
- This approach is centered around the use of a **Database Management System (DBMS)**, which enables users to store, retrieve, and update data securely and efficiently.

Key Characteristics of the Database Approach

1. **Centralized Data Management**:
   - All data is stored in a single, consistent repository.
   - Reduces data redundancy and inconsistencies compared to traditional file systems.
2. **Data Independence**:
   - **Logical Independence**: Changes in database structure do not affect application programs.
   - **Physical Independence**: Changes in storage devices or techniques do not affect how data is accessed.
3. **Data Sharing**:
   - Multiple users and applications can access the same data concurrently.
   - Ensures controlled and consistent sharing through transaction management.
4. **Minimized Redundancy**:
   - Avoids duplicate data entries by normalizing database design.
   - Saves storage space and ensures data integrity.
5. **Security and Integrity**:
   - Provides robust mechanisms to protect data from unauthorized access and corruption.
   - Maintains data integrity through constraints (e.g., primary keys, foreign keys).

Advantages of the Database Approach

1. **Improved Data Consistency**:
   - Centralized control ensures consistent data updates.
2. **Enhanced Security**:
   - Role-based access control ensures that only authorized users access sensitive data.
3. **Data Integrity**:
   - Built-in integrity constraints (e.g., NOT NULL, UNIQUE) maintain accurate and reliable data.
4. **Scalability**:
   - Supports growing data volumes and concurrent user access efficiently.
5. **Better Data Accessibility**:
   - Query languages (like SQL) allow users to retrieve and manipulate data easily.

Key Components of the Database Approach

1. **Database**:
   - A collection of organized data that can be easily accessed, managed, and updated.
2. **DBMS (Database Management System)**:
   - Software that provides an interface for interacting with the database.
   - Examples: MySQL, PostgreSQL, Oracle, MongoDB.
3. **Data Models**:
   - Logical frameworks to define the database structure and relationships.
   - Common models include:
     - Relational (e.g., SQL databases)
     - Hierarchical
     - Network
     - Object-Oriented
4. **Query Language**:

  - Used to communicate with the database (e.g., SQL, NoSQL).

5. **Users**:
    - **Database Administrators (DBAs)**: Responsible for managing the database environment.
    - **End Users**: Access the database for various operations.

Comparison: Database Approach vs. Traditional File Systems

Feature	**Database Approach**	**Traditional File Systems**
Data Redundancy	Minimal	High
Data Sharing	High	Limited
Data Security	Centralized and robust	Fragmented and weak
Data Integrity	Enforced via constraints	No built-in integrity checks
Scalability	Scalable with large datasets	Limited scalability

Database Approach in Practice

- **Applications**: Banking systems, e-commerce platforms, inventory management, ERP systems.
- **Popular DBMS Tools**:
    - Relational Databases: MySQL, PostgreSQL, Oracle DB.
    - NoSQL Databases: MongoDB, Cassandra, Redis.

Big Data

COLLECTION
STORAGE
NETWORK
RESEARCH
BIG DATA
CLOUD TECHNOLOGY
ANALYSIS
EXABITE
VISUALIZATION
VOLUME

Big Data refers to vast volumes of data generated at high velocity and in various formats.

Characteristics (3Vs):

1. **Volume**: Massive amounts of data.
2. **Velocity**: Speed of data generation and processing.
3. **Variety**: Different types of data (structured, unstructured, semi-structured).

Features:

- **Scalability**: Systems must scale to handle large datasets.
- **Real-Time Processing**: Enables immediate analysis.
- **Predictive Insights**: Extracts patterns for forecasting.

hadoop

Example:

- Social media platforms analyze user interactions to improve recommendations.

Case Analysis:

- **Scenario**: An e-commerce platform leverages big data to personalize user experiences and predict purchasing trends.

- **Big Data** refers to extremely large datasets that cannot be effectively managed, processed, or analyzed using traditional data management tools and techniques.

- These datasets are characterized by their massive volume, high velocity, and wide variety, commonly referred to as the **3Vs of Big Data**.
- Big Data is essential for deriving insights, making predictions, and supporting decision-making in various industries.

Characteristics of Big Data (The 3Vs)

1. **Volume**:
    - Refers to the enormous size of data generated from various sources such as social media, IoT devices, transactions, and sensors.
    - Example: Social media platforms generate petabytes of data daily.
2. **Velocity**:
    - The speed at which data is generated, collected, and processed.
    - Example: Stock market data or streaming data from IoT devices require real-time processing.
3. **Variety**:
    - Refers to the diverse formats and types of data:
        - Structured: Tables, databases.
        - Unstructured: Videos, images, social media posts.
        - Semi-structured: JSON, XML.

Other Important Big Data Characteristics

- **Veracity**: Ensuring data accuracy and reliability.
- **Value**: Deriving meaningful insights and actionable intelligence from data.

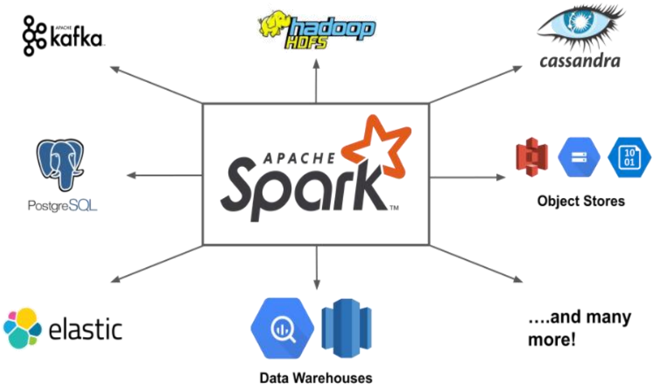
Sources of Big Data

1. **Social Media**: Platforms like Facebook, Twitter, and Instagram generate vast amounts of user-generated content.
2. **IoT Devices**: Sensors and smart devices continuously collect and transmit data.
3. **Transactional Data**: Data generated from online shopping, banking, and point-of-sale systems.

4. **Healthcare**: Electronic health records, medical imaging, and genomics data.
5. **Telecommunication**: Call records, network logs, and customer data.

Big Data Technologies

1. **Storage**:
   - **Hadoop Distributed File System (HDFS)**: Distributed storage for large datasets.
   - **Cloud Storage**: AWS S3, Google Cloud Storage, Azure Blob Storage.
2. **Processing**:
   - **Hadoop**: Framework for distributed storage and processing.
   - **Apache Spark**: Fast, in-memory data processing engine.
   - **Storm and Flink**: Real-time stream processing tools.
3. **Databases**:
   - **NoSQL Databases**: MongoDB, Cassandra, HBase.
   - **NewSQL Databases**: Google Spanner, CockroachDB.

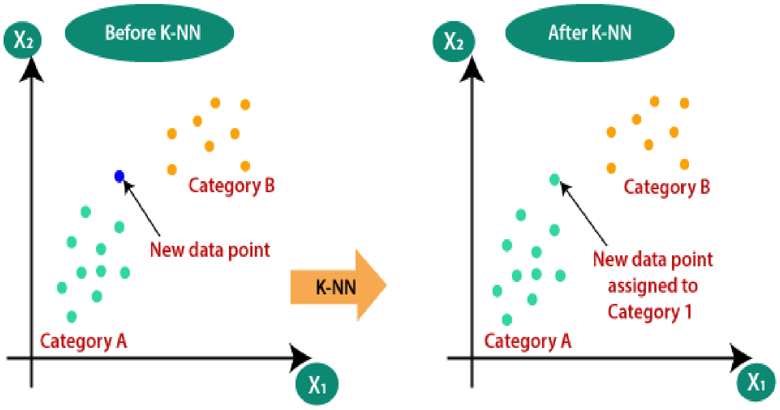


4. **Data Visualization**:
   - Tableau, Power BI, QlikView.

Big Data Analytics

Big Data analytics focuses on extracting meaningful insights from massive datasets through advanced analytical methods. Types include:

1. **Descriptive Analytics**: Understanding what happened.
   - Tools: Business Intelligence (BI) dashboards.



2. **Predictive Analytics**: Forecasting future trends.
    - Techniques: Machine learning, statistical modeling.
3. **Prescriptive Analytics**: Recommending actions based on data insights.
    - Tools: Optimization algorithms.

Applications of Big Data

1. **Healthcare**:
    - Predicting disease outbreaks.
    - Personalized medicine through genomic data analysis.
2. **Retail**:
    - Personalized recommendations (e.g., Amazon, Netflix).
    - Inventory management and demand forecasting.
3. **Finance**:
    - Fraud detection.
    - Real-time stock market analysis.
4. **Smart Cities**:
    - Traffic management using IoT data.
    - Energy usage optimization.
5. **Marketing**:
    - Customer sentiment analysis from social media data.
    - Targeted advertising.

Challenges in Big Data

1. **Data Privacy and Security**:
    - Protecting sensitive data from breaches and misuse.
2. **Data Integration**:
    - Combining data from diverse sources and formats.
3. **Scalability**:

- Managing growing data volumes effectively.

4. **Skill Gap**:
   - Lack of skilled professionals to manage and analyze Big Data.
5. **Infrastructure Costs**:
   - High costs of storage and processing infrastructure.

Future of Big Data

1. **AI and Machine Learning**:
   - Automating Big Data analysis for faster insights.
2. **Edge Computing**:
   - Processing data closer to the source (e.g., IoT devices) to reduce latency.
3. **Blockchain**:
   - Enhancing data security and integrity in distributed environments.
4. **Quantum Computing**:
   - Solving complex Big Data problems at unprecedented speeds.

5. Data Warehouses and Data Marts

Data Warehouse

A centralized repository for storing historical and current data from various sources for analysis and reporting.

Features:

- Subject-oriented (organized by business domain).
- Non-volatile (data remains unchanged).
- Time-variant (tracks data changes over time).

Data Mart

A subset of a data warehouse designed for a specific business function or department.

Features:

- Smaller in scope than a data warehouse.
- Easier to implement and maintain.

Example:

- A company's sales department uses a data mart for sales analytics.

Case Analysis:

- **Scenario**: A multinational company implements a data warehouse to consolidate global sales data for trend analysis and forecasting.

6. Knowledge Management (KM)

Knowledge Management refers to the processes of capturing, distributing, and effectively using organizational knowledge.

Key Processes:

1. **Knowledge Creation**: Developing new insights or solutions.
2. **Knowledge Storage**: Organizing knowledge in repositories.
3. **Knowledge Sharing**: Disseminating knowledge across the organization.
4. **Knowledge Application**: Using knowledge to make decisions or improve processes.

Features:

- Enhances collaboration and innovation.
- Reduces redundancy by reusing knowledge.
- Supports decision-making with shared expertise.

Example:

- A consulting firm creates a knowledge repository of best practices and project learnings accessible to all employees.

Case Analysis:

- **Scenario**: A software company implements KM tools to enable employees to access technical solutions and reduce support resolution times.

Summary of Key Features

Topic	Features
Managing Data	Quality, Governance, Security
Database Approach	Centralization, Integrity, Querying
Big Data	Scalability, Real-Time Processing
Data Warehouses/Marts	Subject-oriented, Time-variant
Knowledge Management	Collaboration, Decision Support

Q/A

Fill-in-the-Blanks Questions

Multiple-Choice Questions (MCQs)

Comprehensive Questions

Answers to Fill-in-the-Blanks

Answers to Multiple-Choice Questions (MCQs)