Chapter 3
Introduction to BigData

Introduction to BigData

- Big Data refers to extremely large datasets
- Cannot be processed efficiently using traditional data management tools.
- These datasets are often generated from diverse sources
- Such as social media, IoT devices, financial transactions, and enterprise applications.

Introduction to BigData

Key Characteristics of Big Data (The 5Vs)

1. **Volume** : Large amounts of data, often in terabytes or petabytes.
2. **Velocity** : The speed at which data is generated, collected, and processed.
3. **Variety** : Different types of data (structured, semi-structured, and unstructured).
4. **Veracity** : Ensuring the quality, accuracy, and reliability of data.
5. **Value** : Extracting meaningful insights to drive business decisions.

Introduction to BigData

Example Sources of Big Data

- **Social Media** : Facebook, Twitter, Instagram generate huge amounts of data.
- **E-commerce** : Websites like Amazon and eBay track customer behavior.
- **IoT (Internet of Things)** : Sensors, smart devices, and wearables generate real-time data.
- **Healthcare** :Medical records, genetic data, and imaging files..

Introduction to BigData

- **Sources of Big Data**

Introduction to BigData

- beyond the relational concept:
- Database: a data warehouse
- It is not just tables and relationships.
- Can contain documents, images, video…

Introduction to BigData

Databases are everywhere:

- Google searches
- Social networks: Twitter, Facebook
- Music / video: Spotify, YouTube, IMDb
- photo: Flickr, Picasa
- commerce: Amazon, eBay
- travel: Expedia, TripAdvisor, AirBnB
- encyclopedias: Wikipedia, Dbpedia
- medical and scientific databases
- data mining

Introduction to BigData

Importance of Big Data

➔enabling **data-driven decision-making**

➔ **real-time analytics**.

Introduction to BigData

Importance of Big Data

➔**Enhanced Customer Insights** : Companies analyze user behavior for personalized marketing.

➔ **Improved Operational Efficiency** :Businesses optimize supply chains and logistics.

➔ **Predictive Analytics** : Healthcare and finance industries use data to predict trends and risks.

➔**Artificial Intelligence & Machine Learning** : Large datasets improve AI models.

Introduction to BigData

- **Real-World Applications:**
- **Netflix** :Uses Big Data to recommend movies and shows.
- **Uber** : Analyzes ride demand and pricing in real-time.
- **Banks & Finance** : Detect fraudulent transactions instantly.
- **Smart Cities** :Traffic management, public safety, and energy consumption optimization.

Introduction to BigData

Challenges of Big Data

Storage : Managing petabytes of data efficiently.

Processing Speed : Analyzing data in real-time.

Data Integration : Combining structured and unstructured data from multiple sources.

Scalability :Handling the continuous growth of data.

Introduction to BigData

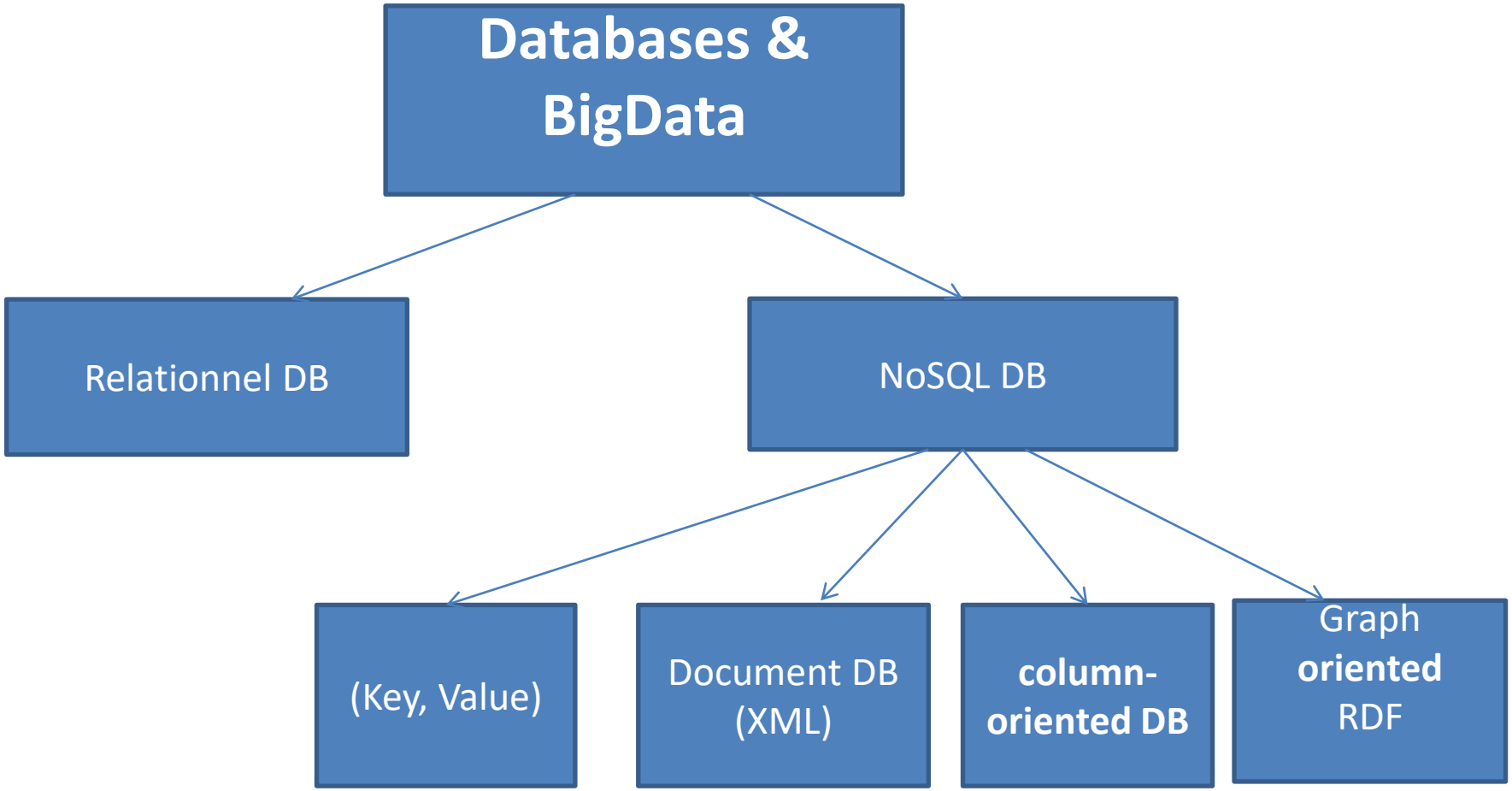
Tools & Technologies for Big Data

1. **Storage Technologies** : NoSQL Databases, HDFS, Amazon S3, Google Cloud Storage, Azure Blob Storage

2. **Processing Frameworks**: Apache Hadoop, Apache Spark, Apache Flink

3. **Data Analytics & Visualization** : Apache Hive, Presto, Google BigQuery – SQL-based analytics

Introduction to BigData



Introduction to BigData

NoSQL Databases :

- NoSQL, means "Not Only SQL",
- designate databases that do not follow the relational model based on SQL
- SQL (Structured Query Language).
- NoSQL are designed to handle large amounts of unstructured or semi-structured data.

Introduction to BigData

NoSQL Databases :

- With the arrival of the Internet and the exponential growth of data, traditional databases have shown their limits

- In terms of their ability to manage massive volumes of data, hence the emergence of NoSQL databases

Chapter 1
Introduction to BigData

Types of NoSQL Databases:

1. **Document-oriented Databases** : Store data in JSON or XML documents Examples: MongoDB, Couchbase.
2. **Key-Value Databases** : Store data as key-value pairs Examples: Redis, DynamoDB.

Types of NoSQL Databases:

3. **Column-oriented Databases** : Store data in columns rather than rows, Examples: HBase, Apache Cassandra.
4. **Graph Databases :** Stores and queries relationships between data in the form of graphs. Examples: Neo4j, Amazon Neptune