# A – Class Discussion

## Big Data

### What are the characteristics of Big Data?

* Big Data is characterized by its volume, velocity, variety, veracity, value, and variability (often referred to as the Six V's).
* These traits describe the challenges and opportunities of processing and analyzing massive, complex datasets that traditional data management tools cannot handle efficiently.

### What are the (3 then 4 then) Six V's of Big Data?

* Original 3 V's:
  1. Volume – The enormous scale of data.
  2. Velocity – The speed at which data is generated and processed.
  3. Variety – The diverse types of data (structured, unstructured, semi-structured).
* Expanded to 4 V's:

4. Veracity – The uncertainty or trustworthiness of data.

* Now 6 V's:

5. Value – The usefulness of data in decision-making.

6. Variability – Inconsistencies in data flow or meaning.

### Who/What is generating these large Vs of data?

* **Humans**: Social media, online transactions, emails.
* **Machines/IoT**: Sensors, smart devices, industrial equipment.
* **Organizations**: Business transactions, healthcare records, scientific research.
* **Automated Systems**: AI, logs, surveillance cameras, satellites.

## Big Data and Cloud Computing.

### How are IoT, Cloud computing and Big Data linked?

* **IoT (Internet of Things)** generates massive amounts of data from sensors and devices.
* **Cloud Computing** provides the storage, processing power, and scalability needed to handle this data.
* **Big Data** technologies analyze and extract insights from the collected IoT data.
* **IoT** → Produces data (Volume, Velocity, Variety).
* **Cloud** → Stores & processes data (Scalability, On-demand resources).
* **Big Data** → Analyzes data (Machine Learning, Real-time analytics).

Together, they enable smart applications like predictive maintenance, smart cities, and real-time monitoring.

### What is NoSQL?

* NoSQL (Not Only SQL) is a type of database designed for unstructured, distributed, and large-scale data.
* Unlike traditional SQL databases (which use tables), NoSQL databases use flexible models like:
* Key-Value (e.g., Redis)
* Document (e.g., MongoDB)
* Column-Family (e.g., Cassandra)
* Graph (e.g., Neo4j)
* Handles Big Data efficiently.
* Scales horizontally (across servers).
* Supports high-speed read/write operations.

### What are the characteristics and advantages of using NoSQL for scalable IoT applications?

* **Characteristics of NoSQL for IoT:**
  + **Schema-less** → Adapts to varying IoT data formats (JSON, XML, sensor data).
  + **Distributed & Scalable** → Handles massive IoT device data across clusters.
  + **High Performance** → Low latency reads/writes for real-time analytics.
  + **Flexible Data Models** → Stores structured, semi-structured, and unstructured data.
* **Advantages for IoT:**
* **Handles High Velocity** – Billions of sensor readings per second.
* **Scalability** – Easily expands with growing IoT networks.
* **Fault Tolerance** – Replicates data across servers for reliability.
* **Cost-Effective** – Uses commodity hardware (unlike expensive SQL setups).

### Data Warehousing

### What is Data Warehousing?

* **Data Warehousing** is the process of collecting, storing, and managing structured data from multiple sources in a centralized repository optimized for **analytics and reporting** (rather than transaction processing).
* **Key Characteristics:**
* **Integrated** – Consolidates data from different sources (databases, IoT sensors, logs).
* **Time-variant** – Tracks historical data for trend analysis.
* **Non-volatile** – Data is read-only once stored (not frequently updated/deleted).
* **Subject-oriented** – Organized by business domains (e.g., sales, IoT sensor data).

### Why might we need Data Warehousing for building scalable IoT applications?

* **Challenges in IoT Data:**
  + **Massive Volume** – Billions of sensor readings daily.
  + **Multiple Sources** – Devices, APIs, legacy systems.
  + **Need for Historical Analysis** – Detecting patterns (e.g., predictive maintenance).
* **How Data Warehousing Helps IoT:**
* **Centralized Storage** – Aggregates dispersed IoT data for unified analysis.
* **Fast Query Performance** – Optimized for analytics (unlike raw databases).
* **Historical Insights** – Stores long-term data for trend detection.
* **Integration with BI Tools** – Powers dashboards (e.g., Tableau, Power BI).
* **Scalability** – Cloud-based warehouses (BigQuery, Redshift) handle IoT’s growing data.
* Use Case Example: A smart factory uses a data warehouse to:
  + Analyze sensor data for equipment failures.
  + Optimize energy usage over time.
  + Generate reports on operational efficiency.