SESSION 2

Types and Characteristics of Big Data

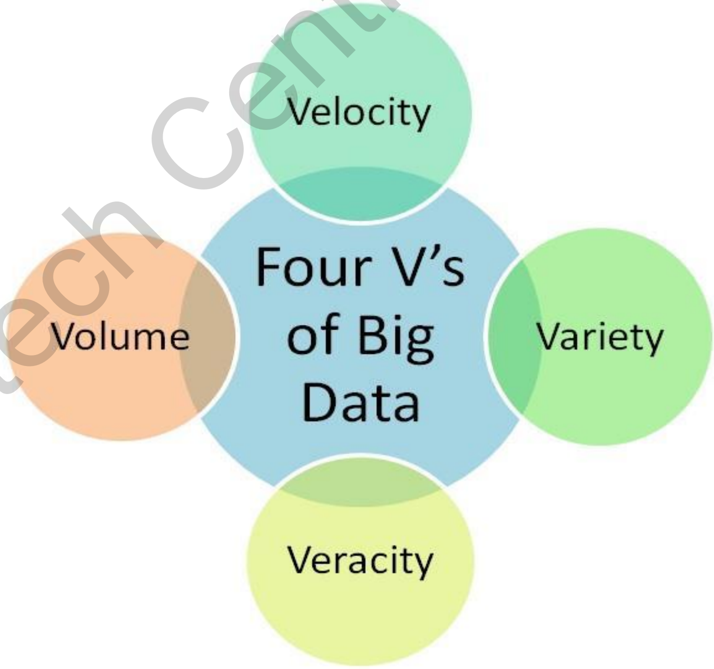
Objectives

- Describe the characteristics of Big Data
- Explain the types of data in Big Data

Characteristics of Big Data (1-2)

- Big Data is qualitative and cannot be quantified

- Four V's of Big Data are:



Characteristics of Big Data (2-2)

Volume

- Represents size of data, varies from Gigabytes to Terabytes or Petabytes
- Data that is produced by machines

Velocity

- Represents speed at which data is produced
- Data is generated at different speeds, fields, and areas of technology

Variety

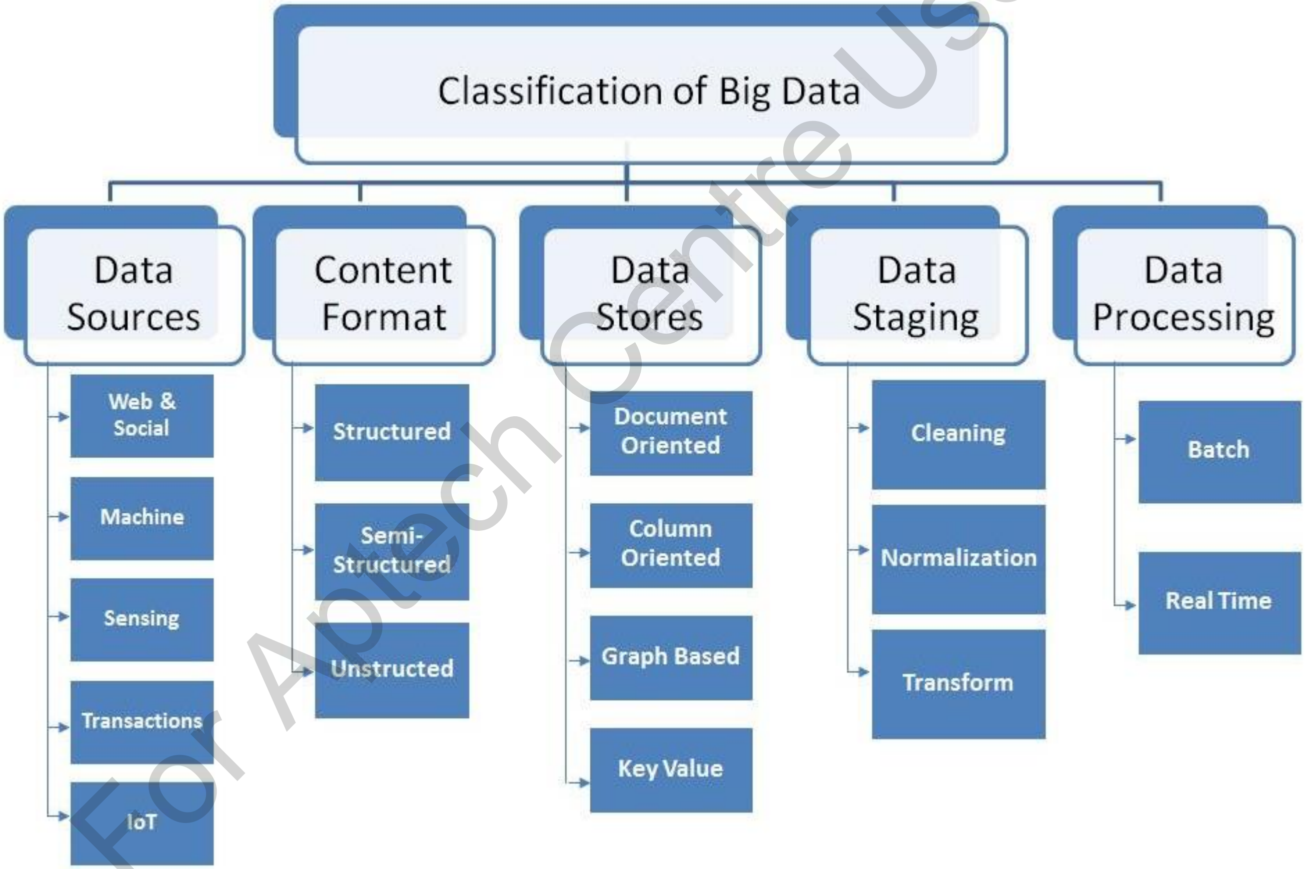
- Represents formats of data being produced or stored
- Unstructured data is produced in large volumes and stored in formats such as images, audio/video files, and sensor data.

Veracity

- Refers to abnormality, biases, and noise in data
- Recommended to maintain clean data and prevent unwanted data

Big Data Classification

Big Data is classified into different categories:



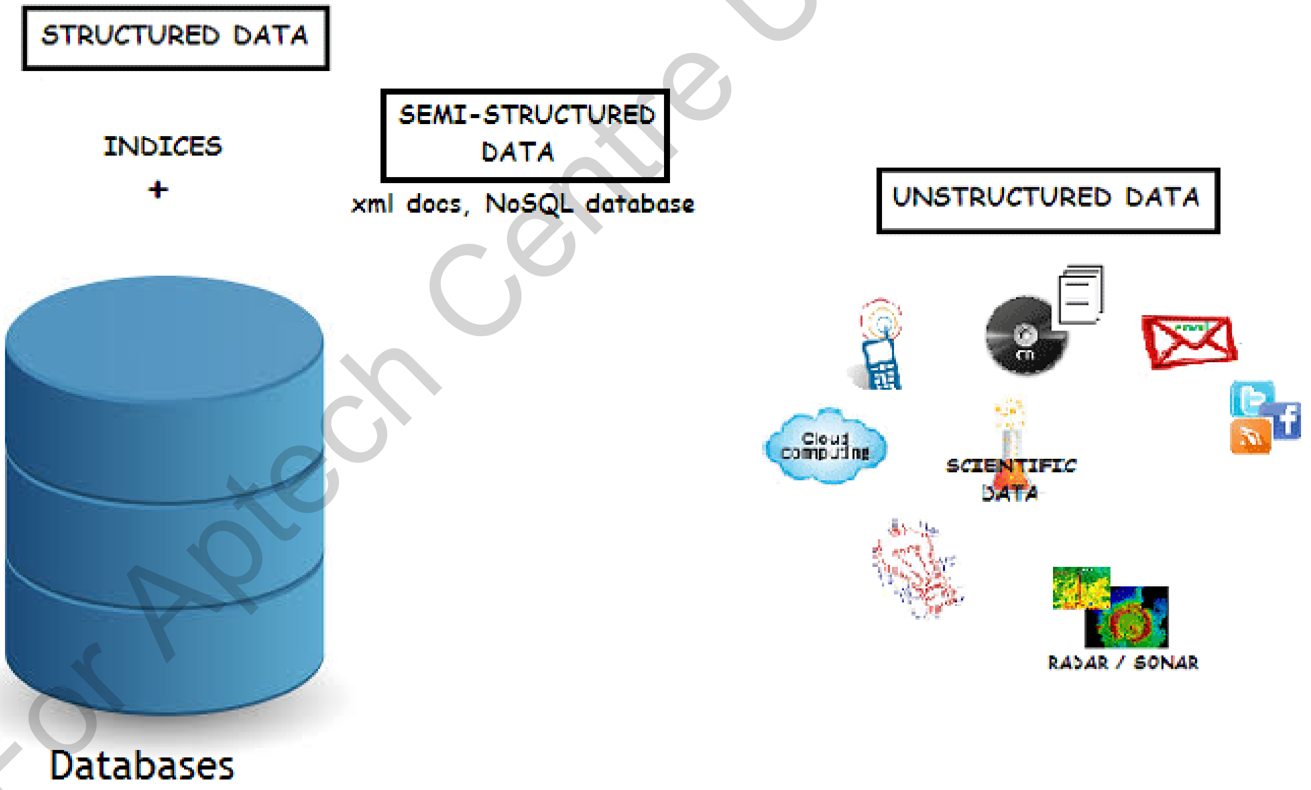
Data Sources

- It includes Internet data, sensing, and all stores of transactional data
- Data is collected from various sources, such as social media, sensing devices, machine generated data, transactional data, and uniquely identifiable Internet objects.
- Types of Data Sources are:

Data Source	Description
Social Media	The source of data or information that is generated through a URL for sharing or exchanging information and ideas to virtual communities and networks, such as Facebook, Twitter, Instagram, collaborative projects, blogs, and microblogs.
Machine-Generated Data	Data which is generated automatically from computers, medical devices, or other machines, with no human intervention.
Sensing	Data that is generated through signals.
Transactions	Data of an event that comprises time dimension for describing the data, such as finance and work data.
IoT	IoT represents a collection of objects that are uniquely identifiable devices as a part of the Internet. It includes smartphones, digital cameras, and tablets.

Content Format (1-2)

- It includes structured, unstructured, and semi-structured data.



STRUCTURED DATA

INDICES

+

Databases

SEMI-STRUCTURED DATA

xml docs, NoSQL database

UNSTRUCTURED DATA

Content Format (2-2)

Data Type	Description
Structured	Uses SQL, a database programming language, to manage and query data in RDBMS. Structured data is easy to manage, input, store, query, and analyze.
Semi-structured	Semi-structured data are structured data that are not organized in relational database models, for example, tables. It does not follow rules of conventional database systems.
Unstructured	Text messages, location information, videos, and social media data are formats of unstructured data. Such data does not have a specific format.

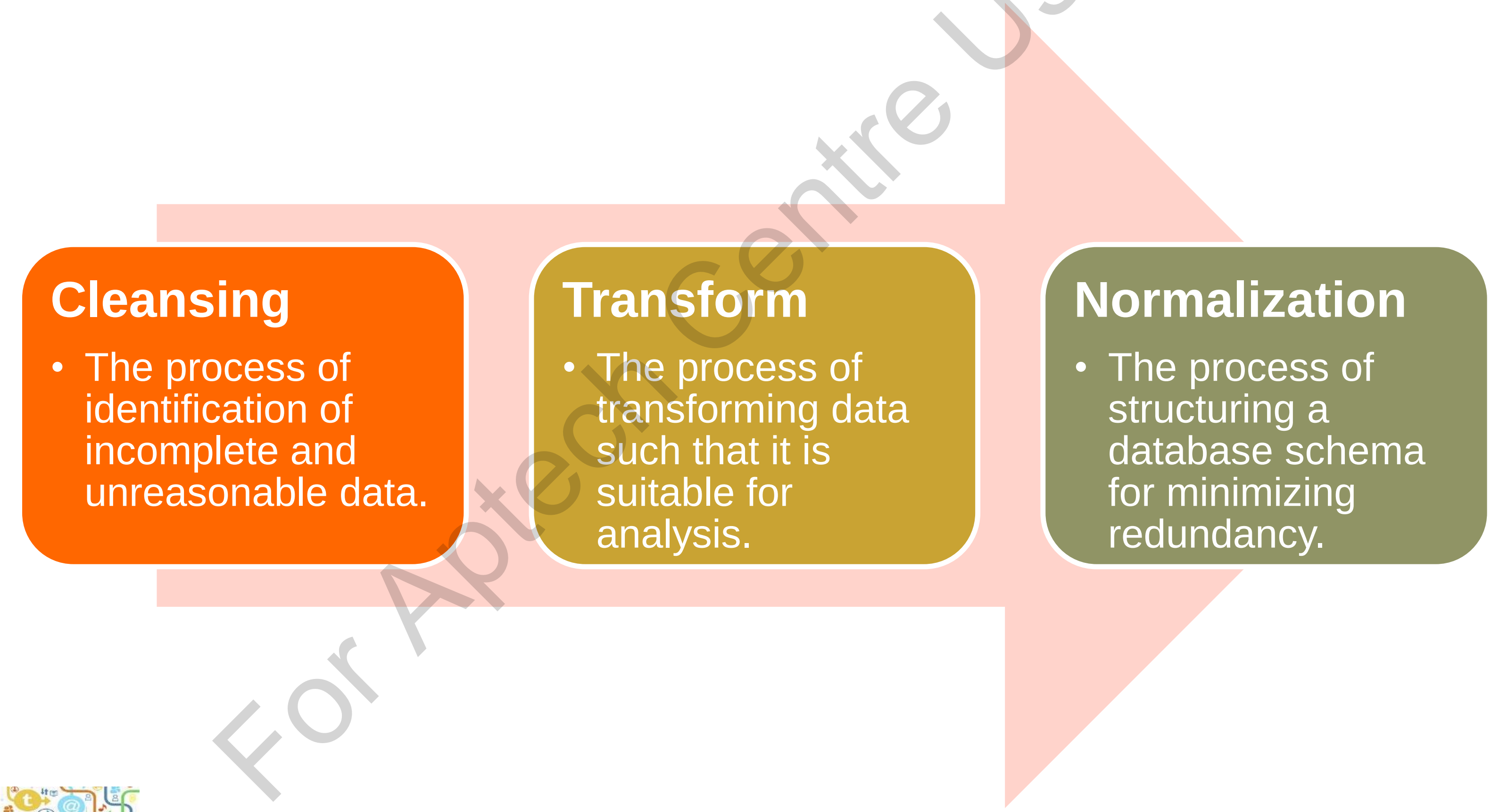
Data Stores

- It includes document-oriented, column-oriented, key-value, and graph database.
- Types of Data Stores are:

Data Store	Description
Document-oriented	Document-oriented data stores are mainly designed to store and retrieve collections of documents or information and support complex data forms in several standard formats, such as JSON, XML, and binary forms (for example, PDF and MS Word).
Column-oriented	A column-oriented database stores its content in columns aside from rows, with attribute values associated to the same column stored alongside. Column-oriented is different from classical database systems that store entire rows one after the other, such as BigTable.
Graph Database	A graph database, for example Neo4j, is designed to store and represent data that utilizes a graph model with nodes, edges, and properties related to one another through relations.
Key-value	Key-value is an alternative RDBMS that stores and accesses data created to scale to a very large volume.

Data Staging

- It includes three processes – Cleansing, Transform, and Normalization.



Cleansing

- The process of identification of incomplete and unreasonable data.

Transform

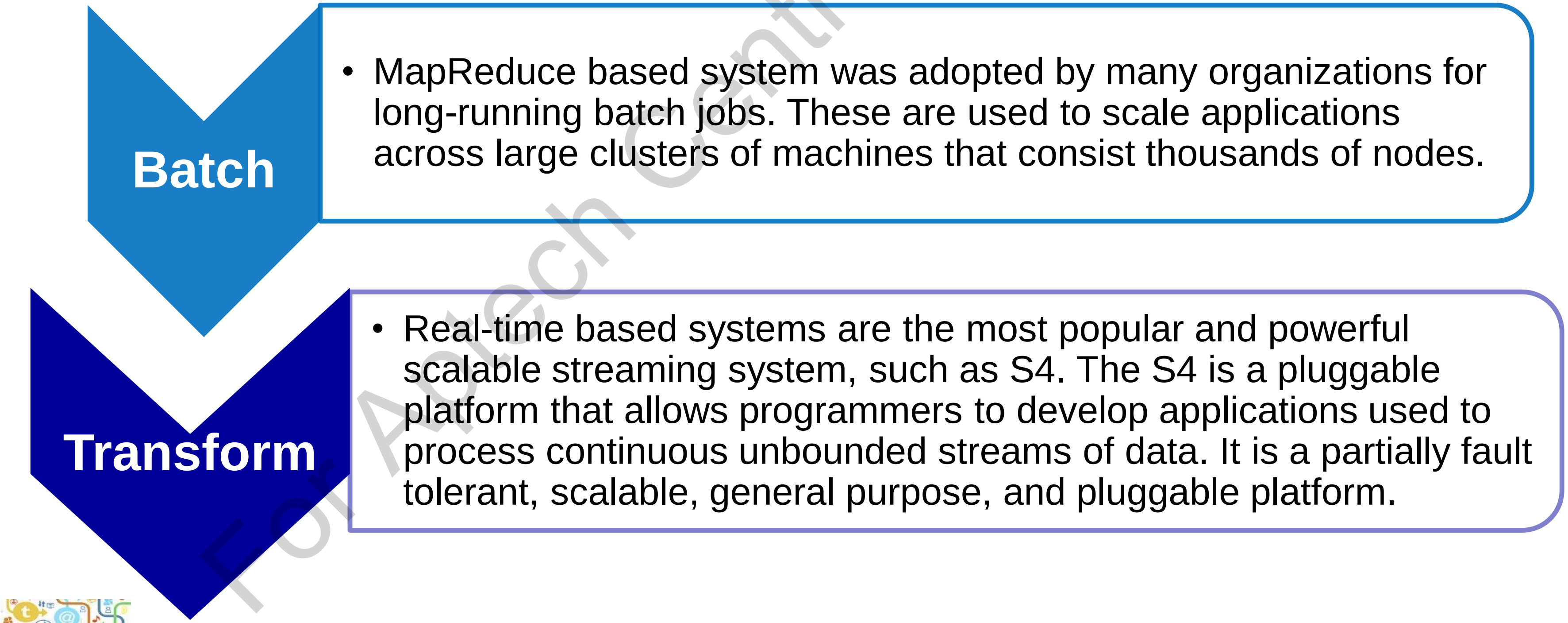
- The process of transforming data such that it is suitable for analysis.

Normalization

- The process of structuring a database schema for minimizing redundancy.

Data Processing

- It refers to processing of data in batches through batch jobs in MapReduce based system or Real-time based systems.
- Types of Data Processing are:



Batch

- MapReduce based system was adopted by many organizations for long-running batch jobs. These are used to scale applications across large clusters of machines that consist thousands of nodes.

Transform

- Real-time based systems are the most popular and powerful scalable streaming system, such as S4. The S4 is a pluggable platform that allows programmers to develop applications used to process continuous unbounded streams of data. It is a partially fault tolerant, scalable, general purpose, and pluggable platform.

Structured and Unstructured

- Structured Data:
  - □ Includes text or numerical data
  - □ Uses Structured Query Language (SQL) to manage and query data in RDBMS
  - □ Represents only 5% to 10% of informatics data
- Unstructured Data
  - □ Includes text messages, location information, videos, and social media data

Examples of Unstructured Data

Satellite Images

- It consists of weather data or satellite surveillance imagery data

Scientific Data

- It comprises seismic imagery, atmospheric information, and high energy physics

Photographs and Videos

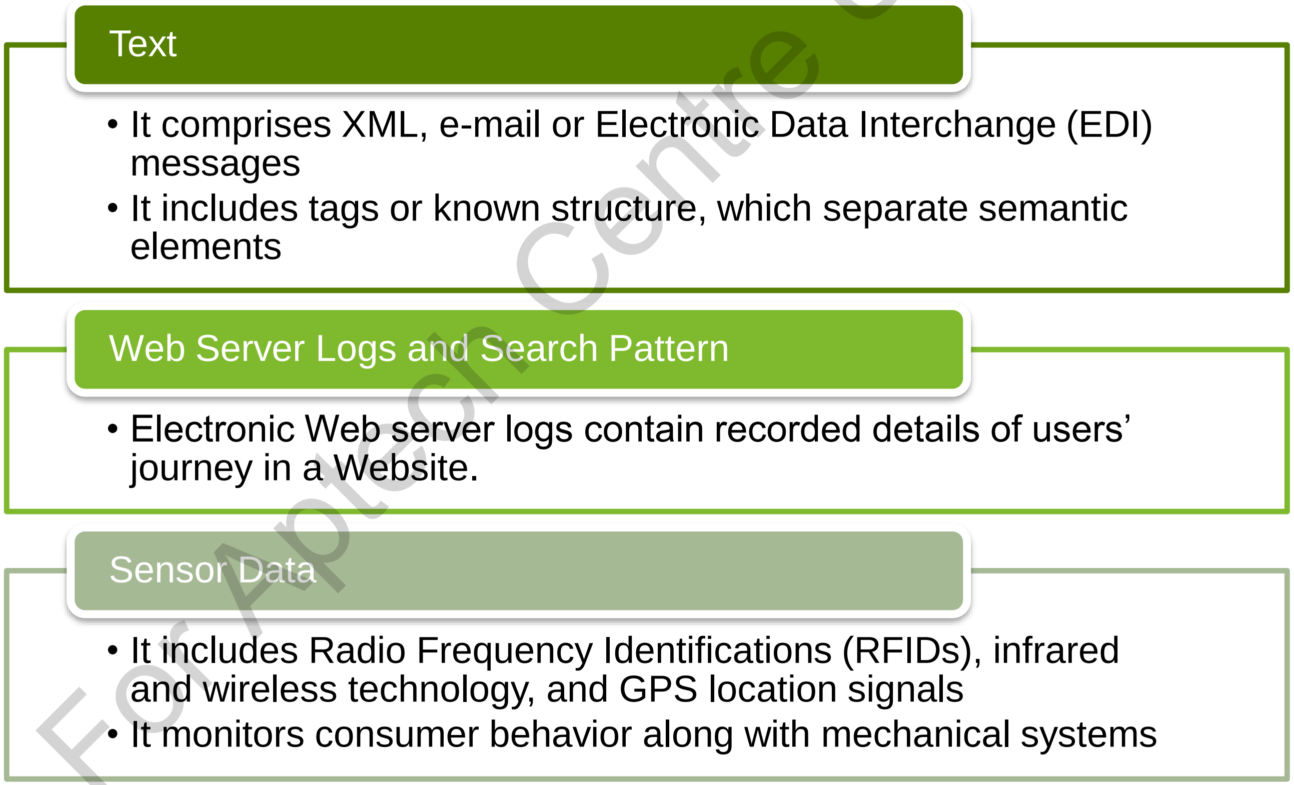
- This data comprises security, surveillance, and traffic video

Radar or Sonar Data

- It includes vehicular, meteorological, and oceanographic seismic profiles

Semi-Structured Data

- It does not follow rules of conventional database system.
- Types of semi-structured data are:



Text

- It comprises XML, e-mail or Electronic Data Interchange (EDI) messages
- It includes tags or known structure, which separate semantic elements

Web Server Logs and Search Pattern

- Electronic Web server logs contain recorded details of users' journey in a Website.

Sensor Data

- It includes Radio Frequency Identifications (RFIDs), infrared and wireless technology, and GPS location signals
- It monitors consumer behavior along with mechanical systems

Summary (1-2)

- The four Vs of Big Data are Volume, Velocity, Variety and Veracity.
- Volume refers to the size of data, velocity refers to the speed at which the data is being generated.
- Variety refers to the different formats in which the data is being stored and veracity refers to the abnormality in data.
- The classification of Big Data is based on five aspects: Data sources, Content format , Data stores, Data staging, and Data processing.

Summary (2-2)

- There are three types of data in Big Data: Structured, Unstructured, and Semi-Structured.
- Structured data includes text or numerical data.
- Unstructured data includes text messages, location information, videos, and social media data and it does not have any specific format.
- Semi-structured data includes text, Web server logs and search patterns, and sensor data.