01 Introduction to Big Data

What is Big Data?

- Big data refers to datasets that are too large or complex to be captured, managed, and dealt with by traditional data-processing application software

The 4V's of big data

1- **Volume [amount of data]**
   - Refers to the **massive size of data being generated every second**. Data now is come in terabyte, petabyte, or even zetabyte
2- **Velocity [speed of data]**
   - Data is being generated and needs to be processed at **high speed**. Some data must be handled in real-time for decision making
3- **Variety [types of data]**
   - Refers to the types of data. The data can be structured, un-structured, semi structured
4- **Value [create insights]**
   - The data in useless unless it can create insights or benefits
   - The main goal of big data is to turn raw data into value

Difference Between Big Data and Traditional processing

	Big Data Processing	Traditional Data Processing
Data sale	Large in GB, TB, PB	Small in MB
Data type	Multitype. Can be structured, semi-structure, unstructured	Single type, mainly structured data type
Mode-data relationship	Modes are set after data is generated. Modes evolve as data increase	Modes are set before data is generated
Tool	No size fit all	One size fit all

Main Computing Modes of Big Data Applications

1- **Batch computing**
   - Process massive data in batches
   - Major technologies include Mapreduce, and Spark
2- **Stream computing**
   - Calculates and processes streaming data in real-time
   - Major technologies include Spark, Storm, Flink, Flime, Dstream, ...
3- **Graph computing**
   - Processes large scale graph structure data.
   - Major technologies like GraphX, Gelly, PowerGraph

4- Query and Analysis computing

- Storage management and query analytics of massive data
- Major technologies include Hive, impala, Dremel, and Cassendra

Hadoop Big Data Ecosystem