2.3 Introduction to Big Data

Big data is defined as collections of datasets whose volume, velocity or variety is so large that it is difficult to store, manage, process and analyze the data using traditional databases and data processing tools.

2.3.1 Challenges

1. Meeting the need for speed
   - Storing data efficiently
   - Finding relevant data quickly
2. Visualization helps organizations perform analyses
3. Degree of granularity increases
   - user base of a company may grow explosively
   - companies release apps that produce even more real-time data for each customer
4. Privacy issues
   - data breach
   - ransomware
5. Fault tolerance
   - distributed systems are fault-tolerant
   - highly available architectures using data replication
6. Scalability and performance

7. Batch-processing and real-time processing

2.3.2 Characteristics

1. **Volume** of data involved is so large that it is difficult to store, process and analyze data on a single machine (MB, GB, TB, PB)
2. **Velocity** of data is very high and the data needs to be analyzed in real-time (moving target)

   Categories: batch >> period >> real-time
3. There is **variety** of data involved, which can be structured, unstructured or semi-structured, and is collected from multiple data sources

   Categories: databases >> photo/web/audio >> social/mobile
4. **Veracity** refers to quality of data. Cleaning the data can improve its quality.
5. **Value** contribution towards intended application requires various types of analytics to be performed such as descriptive, diagnostic, predictive and prescriptive analytics.
6. Big Data tools and frameworks have **distributed and parallel processing architectures** and can leverage the storage and computational resources of a large cluster of machines.