Week -2 Continue Big Data

V's of Big Data

- Velocity (youtube uploads)
- Volume (Mobile, Desktop, Laptop, Wearable 2.5 quintillion per day, 10 million DVDs)
- Variety (text, picture, video)
- Veracity (Quality, Accuracy) (80% is unstructured)
- Value (Turn data into value)

Big Data Processing Tools

- **Hadoop** (a collection of tools that provides distributed storage and processing of big data)
  - Java Based Open Source
  - Across cluster of computers
  - HDFS → Hadoop Distributed File System
- **Hive** (Data Warehouse Query and analysis on top of Hadoop)
  - Open Source Warehouse Software
  - Very high latency →
  - Read based
  - Better suited for ETL
- **Spark** (Distributed analytics framework for complex, real-time data analytics)
  - Process large volume of data
  - ML, Live Analytics