Lecture Notes 1

Big Data

- How to compute average(mean) of 10 integers
- What happens if size of all integers greater than storage on one machine?

Exponential Growth of Data

- Google Serach
- Facebook Likes
- Why companies need to store this data?

Unstructured and Structured Data?

- Unstructured - video, image, audio
- Structured - XML, JSON, sensor data
- The volume of unstructured data exploded in the past decade

Common BigData Use Cases

- ETL
- Text Mining
- Prediction Models and Analytics
- Graph creation and analytics

What do these workloads hav ein common?

- Huge Volume, variety, and Velocity

Traditional Distributed Systsms

- Data stored in central location
- Data copied to processors at runtime

- Fine for limited amounts of data

The Data Bottleneck

- Moden systms have much more data
- Solution - new style where we store/process huge amounts of data in parallel **big data systems**
    - Efficient storing, receiving, and processing large amounts of data
    - Distributed systems are the building blocks of this

Key Ideas

- Distribute data when the data is stored
- Bring the computation to the data and the data to the computation

Scalable and economical data storage, processing, and analysis

- Distributed and fault-tolerant
- Harness the power of industry standard hardware (virtualization)
- Heavily inspired by Open suorce tenchologies (HDFS, HBase, etc)
- Easy to develop applications (hides complexity)

A Typical BigData System

3 layers

- Processing - ie Batch processing, Analytical SQL, ML, etc.
- Resource Manager - Manages resources, schedules. The Operating system essentially
- Data Storage - Contains data ingestion systems

Popular BigData Systems

- Hadoop
- Spark