Big Data Analytics with Apache Spark

Debzani Deb

What is “Big Data”?

- Flickr has >6 billion photos
- Facebook has 1.7 billion active users
- Google is serving >1.2 billion queries/day on more than 27 billion items
- >2 billion videos/day watched on YouTube

How much data?

- Modern applications use massive data:
  - Rendering 'Avatar' movie required >1 petabyte of storage (1 million GB or a thousand TB)
  - eBay has >6.5 petabytes of user data
  - CERN's LHC will produce about 15 petabytes of data per year
  - In 2008, Google processed 20 petabytes per day
  - German Climate computing center dimensioned for 60 petabytes of climate data
  - Google now designing for 1 Exabyte (1B GB) of storage
  - NSA Utah Data Center is said to have 5 zettabyte (thousand Exabyte))
- How much is a zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
  - A stack of 1TB hard disks that is 25,400 km high

How much computation?

- No single computer can process that much data
  - Need to distribute data in many computers
  - Need to process this distributed data in a parallel fashion!
  - That results in faster computation!
- How many computers do modern services need?
  - Facebook is thought to have more than 60,000 servers
  - 1&1 Internet has over 70,000 servers
  - Akamai has 95,000 servers in 71 countries
  - Intel has ~100,000 servers in 97 data centers
  - Microsoft reportedly had at least 200,000 servers in 2008
  - Google is thought to have more than 1 million servers, is planning for 10 million.

Sequential vs. Distributed/Parallel

- Problem: Find the average of 1000 numbers.
- Sequential program:
  - Read the array that contains 1000 numbers
  - Use a for loop to find the **sum** of them.
  - Find average = sum/1000

- Parallel Program (when you have 4 Node parallel computer)
  - Master node reads the array.
  - Master node divides the array in 4 chunks
  - Master node assign each chunk (250 elements) to a node.
    - Each node calculates the **sum** of 250 numbers.
    - Each node send their **sums** to Master node.
  - Master node receive all sum and add them all to computer the final_sum.
  - Master node find average = final_sum/1000.

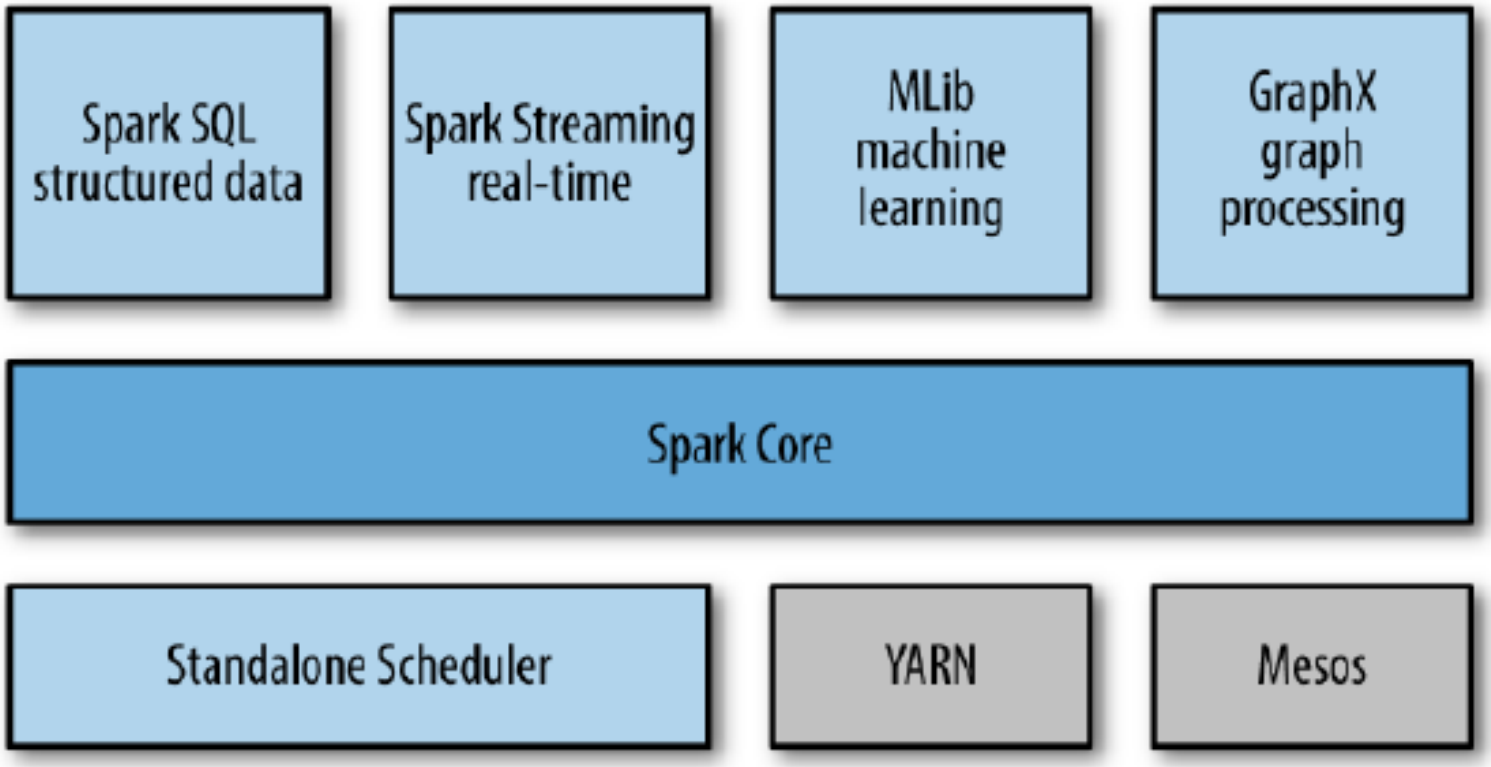
Distributed Processing

- Advantage
  - Faster computation
  - Can process real big data (Petabyte or even more)
- Disadvantage
  - Programming is hard.
  - Need to be concerned with
    - How do I distribute an algorithm?
    - How do I partition my dataset?
    - How do I manage communication between nodes?
    - How to I allocate tasks/jobs to nodes?
    - How to I scale?
    - How do I recover from machine failures?

What is Hadoop/Spark?

- Apache Hadoop or Spark are distributed Frameworks that allows us to write program that can distribute and process a task that works over petabytes of data while taking care of scalability and fault tolerance.
- Apache Spark is a memory-based data processing framework that is much faster than traditional Hadoop and supports SQL like querying on NO-SQL databases.
  - Flexible – can use Java, Python, Scala
  - SQL is integrated
  - Support varieties of File types such as text, CSV, JSON, SequenceFiles etc.
  - No schema required

Spark


Spark SQL
structured data
Spark Streaming
real-time
MLib
machine
learning
GraphX
graph
processing
Spark Core
Standalone Scheduler
YARN
Mesos

Spark SQL & DataFrame

- Spark SQL - A Spark module for structured data processing
- Provide Spark with more information about the structure of data and computation being performed
- DataFrame is one way to interact with Spark SQL
  - A distributed collection of data organized into named columns.
  - Conceptually equivalent to a **table** in relational database

Tutorial and Assignment

- Some of the computers in this lab (4116) and 2nd floor lab is equipped with necessary software. They all have "SPARK" written in a yellow paper on top of the monitor.
- We will be learning data processing using SPARK in this lab.
- You will code in python (using jupyter notebook).
- You will work as a pair
  - One submission from each pair
- The complete assignment is due on 4/26.