Big Data Analytics: Spark

Mohamed Ndaoud

Apache Spark

- Apache Spark is a popular framework in the field of Big Data.
- Running code on a single machine → using clusters.
- Spark is a distributed cluster-computing software framework:
    - It provides easy APIs
    - The end users barely need to know about the task and resource management across machines
- Spark is the most actively developed open source engine for parallel data processing on computer clusters

Apache Spark

- Spark:
  - supports widely used programming languages
  - includes libraries for diverse tasks
  - runs anywhere from a laptop to a cluster of thousands of servers

Apache Spark's Philosophy

- **Unified** platform for writing big data applications

Apache Spark's Philosophy

- **Unified** platform for writing big data applications
- **Computing Engine:** can be used with a wide variety of storage systems

Apache Spark's Philosophy

- **Unified** platform for writing big data applications
- **Computing Engine:** can be used with a wide variety of storage systems
- **Libraries**

Running Spark

- Spark can be used from Python, Java, or Scala, R, or SQL
- It is written in Scala, and runs on the Java Virtual Machine (JVM)
- Needs Java 6 or newer
- + Python interpreter (version 2.6 or newer)
- or a version of R on your machine.

Running Spark

- to get started with Spark:
  - download it (Homework 1) or
  - run it for free on
    - Databricks
    - Google Colab
    - etc

Spark's Basic Architecture

- A cluster puts the resources of many machines together
- A framework to coordinate work across them
- Spark: management and coordination of the execution of tasks across a cluster

Spark Applications

- Driver process:
  - runs your main() function
  - maintains information about the Spark Application
  - responds to user
  - distributes and schedules work across the executors
- Executor processes:
  - executes the work
  - reports the state of computation