Apache Spark (Big data)

by Hua Zhang

Agenda

- Introduction
- Spark Core (in Java)
  - Spark Session
  - Spark Read
  - Spark API + SQL
  - Spark Write
- Assignment
- Java vs Scala, based on Spark
- Spark Cluster, from small to big

Introduction

From Wikipedia

Apache Spark is an open-source unified analytics engine for **large-scale data** processing. Spark provides an interface for programming clusters with **implicit data parallelism** and **fault tolerance**.

- **Spark Core (API + SQL)**
- Spark Streaming
- Machine Learning (MLLib)
- Graph Processing (Graph X)

Spark Core

1. Checkout git project: https://github.com/happyhua/spark
2. Create SparkSession, entry point to all of Spark's functionality
3. Spark Read (DataFrameReader), load data files into DataFrame
4. Spark API + SQL: Scala, Java, Python and R
5. Spark Write

Assignment

1. Make sure that you can run Example main
2. Open Assignment.java file
3. Assignment 1: Load a CSV file into Dataset<Row>
4. Assignment 2: Aggregate data in Spark
5. Assignment 3: Join data in Spark

Java vs Scala, based on Spark

- The gap is smaller since Java 8, but still lots of shining stuff in Scala
- Comparison based on Spark framework
  - Case class in Scala
  - Implicit usage in Scala
  - Math sign in Scala

Spark Cluster

1. From https://spark.apache.org/downloads.html download the binary file
2. Unpack it and place it somewhere on your disk, for example, your home directory
3. Run the following command: ./start-worker.sh spark://hp-solus:7077

Questions?