Working with

Hadoop and SPARK

Lab Session

COMP 330: Tools and Models for Data Science

Chris Jermaine, Kia Teymourian

Fall, 2015

Rice University
Computer Science Departments

Usage of Hadoop and Spark

1) **Hadoop**
    - 1.1. Local Installation and Running
    - 1.2. Cluster usage on AWS

2) **SPARK**
    - 2.1. Local Installation and Running
    - 2.1. Cluster on AWS

1. Hadoop – Local

- This is about Java programming
- ***Installation steps:***
    1. **Java JDK** 1.6 or higher (Open JDK or Oracle Version)
    2. **Maven** (https://maven.apache.org/)
        - Find packages for your OS
        - Maven can download required libraries and compile your project
    3. You might use an **IDE like Eclipse, Netbeans, IntelliJ**
    4. Install **maven plugin** for your IDE like http://www.eclipse.org/m2e/ for eclipse
    5. download **“WordCount.zip”** project from piazza resources and create a Java/Maven project in your IDE
    6. Compile/run and debug your project locally

Deployment of your project to AWS Elastic MapReduce

- Steps:

  1. Create a **Cluster** on AWS "**Elastic MapReduce**"
     - It will automatically start a set of EC2 machines and install hadoop on them.

  2. **Create a Jar file** from your project with all of the dependency libraries

  3. Deploy and run your **Jar file on your cluster**
     - **Two Methods:**
       - Method-1: Use the Elastic MapReduce GUI to add a "step" on your cluster
       - Method-2: Login to the Master Node by using a SSH connection

         On the command line run

         `hadoop -jar  YOURJARFILE.jar  INPUT OUTPUT`

- ***`INPUT and OUTPUT are directory paths like S3, HDFS, ...`***

SPARK - Local

- This is about **Python** programming
- Steps:

    1. Install python (Version 3) (It requires Java JDK as well)
    2. Install Scala - http://www.scala-lang.org
    2. You may want to use an IDE (Eclipse)
    3. Install python plug-in for your IDE (like http://www.pydev.org/ PyDev is a Python IDE for Eclipse)
    4. Download Spark Package from http://spark.apache.org/downloads.html Choose a package type: pre-built package for Hadoop 2.6 or later
    5. Unpack the file spark-1.5.0-bin-hadoop2.6.tgz
    6. Add library files (zip files) from spark-1.5.0-bin-hadoop2.6/python/lib to your project
    7. Now you can compile and run SPARK programs locally

SPARK – Cluster

- Deployment on AWS Amazon Elastic MapReduce

  1. Create a Cluster on Amazon Elastic MapReduce including SPARK

  2. SSH login to the master Node (you can see which one is the master node on cluster page. You can not ssh to slave nodes from outside Amazon network)

  3. After login to the Master Node you can deploy your application by using this command:

  # ***spark-submit YOURAPP.py INPUT OUTPUT***

  4. Check the SPARK GUI to see if your application is running (you need to enable a SSH tunnel to the Master Node)

AWS Tips

- USE Amazon EC2 Spot Instances to save money
  - https://aws.amazon.com/ec2/spot/
- Shutdown instances when you do not need them, for example overnight running instances. You can run a new cluster whenever you want.
- Debug your code before you run it on AWS
- In Case of Errors on EC2/EMapReduce check the log files and debug