Lab 4 Report: AWS and Hadoop/MapReduce

Please **fill in the report** and submit the **pdf** to NYUClasses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name: |  | ID: |  | Date: |  |

# Objectives

* A peek of the Amazon AWS and managing VMs on cloud
* Understand the MapReduce concept.
* Get familiar with the Hadoop framework.
* Experience working a small Hadoop cluster with VMs.

# Experiments Tasks

## Basics

1. Go through the Apache Hadoop introduction to get the general idea about Hadoop:

<http://hadoop.apache.org/>

1. Go through the Apache Hadoop release notes to understand the evolution of Hadoop:

<http://hadoop.apache.org/releases.html>

1. Play with Amazon Web Services (AWS) EC2 and learn how to create instances on AWS

<https://aws.amazon.com/>

## Install Hadoop

1. Follow the instructions on   
   <http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html>  
   to set up Hadoop environment on your own Linux machine.
2. Compile the given WordCount java program and run it with MapReduce

## Build a Hadoop Cluster (Bonus)

1. Use two or more VMs to build a Hadoop Cluster with a master node and slave nodes.
2. Compile the given WordCount java program and run it with MapReduce

# Reports

1. When creating the AWS account, you can choose IAM user or Root user. What privilege does Root user have over IAM user? Which one will you choose and why?

|  |
| --- |
|  |

1. What are the major differences between Hadoop version 1, 2, and 3?

|  |
| --- |
|  |

1. What is YARN? What are advantages of YARN over MapReduce?

|  |
| --- |
|  |

1. What is Spark? What is the difference between Spark and Hadoop?

|  |
| --- |
|  |

1. What is Hadoop streaming? If I have a Mapper written in python, how can I use Hadoop Streaming to run the code with Hadoop?

|  |
| --- |
|  |

1. Screenshots of the practice of Hadoop (in 2.2) on your own computer. The screenshots should show the result of Hadoop execution ($ cat output/\*) as well as the files in HDFS ($ hdfs dfs -ls).

|  |
| --- |
|  |

1. Run WordCount on the Hadoop. Please attach the screenshots of your result, and what are the top 5 most frequent word in the provided txt file?

|  |
| --- |
|  |

1. In Hadoop Cluster Mode, what are the differences between Hadoop master and slave nodes? Also name what functionalities are performed on each node.

|  |
| --- |
|  |

1. Write a pseudo code to multiply large matrices using Hadoop MapReduce. **Explain** the function of your Mappers and Reducers.

|  |
| --- |
| Mapper:  Reducer: |

1. What is a “combiner” in MapReduce?

|  |
| --- |
|  |

1. Try to add a combiner to the question (i) and **explain its function**. Justify why adding the combiner can make your code faster.

|  |
| --- |
|  |

**(Bonus) Hadoop Cluster Mode**

1. Use jps commands on both VMs to show running Hadoop daemons and provide and screenshots.

|  |
| --- |
|  |

1. Screenshots of configuration files and IP addresses for Master node and Slave node of your small cluster as well as the MapReduce execution result. For each configuration file, please also briefly explain what it does.

|  |
| --- |
|  |

**We have zero tolerance to forged or fabricated data!!** A single piece of forged/fabricated data would bring the total score down to zero.