Big Data Processing Course

Spring Semester 2016

Homework Assignment 2.

April 7, 2016

Contents

Chapter 1

Introduction

1. Homework submission is strictly in **groups of 4 students**, groups with less than 4 should ask for an explicit permition from the lecturer.
2. Homework deadline is due to: **01/05/2016 23:59**.
3. Please pay attention to the due date. I urge you to start working on the homework right away and not wait until the last minute.
4. Please remember that a significat part iof your homework solution is understanding of requirements, therefore I urge you to read through this document very careful and make sure you understand each single part of it. Solutions verified with automatic tooling, therefore in case of misunderstanding of requirements or in case of non compliance, points will not be returned back.

1.1 Goals

Main purpose of the current submission is to learn how to setup Hadoop cluster, since following excercises will heavily rely on that this is very important homework. You are required to provide a valid configuration for two nodes Hadoop cluster and in order to complete this task you will have to use Vagrant tool. It's assumed that you already fairly familiar with Vagrant from the homework 1, altghough it was very basic introduction and you didn't dealt directly with configuring. Now, in this assignment you will learn how to provision virtual machines with required configuration using Vagrant tool. There are many way to provision virtual machines with Vagrant however for now we will focus on how to leverage Linux shell environment to setup two nodes Hadoop cluster within virtual environment.

1.2 Get Started

Following the instructions, which allows you to get started with work on your assignment solution. In order to provide unified development and testing environment you're required to validate and compile your solution using virtual boxes. Since our check is automated you also required to work with **Vagrant** to manage virtual machine activity. Please read carefully tutorials following links below based on your preferences.

1.2.1 Install VirtualBox

VirtualBox is a cross-platform virtualization application. For one thing, it installs on your existing Intel or AMD-based computers, whether they are running Windows, Mac, Linux or Solaris operating systems. Secondly, it extends the capabilities of your existing computer so that it can run multiple operating systems (inside multiple virtual machines) at the same time. So, for example, you can run Windows and Linux on your Mac, run Windows Server 2008 on your Linux server, run Linux on your Windows PC, and so on, all alongside your existing applications. You can install and run as many virtual machines as you like – the only practical limits are disk space and memory.

VirtualBox is deceptively simple yet also very powerful. It can run everywhere from small embedded systems or desktop class machines all the way up to datacenter deployments and even Cloud environments. Bellow links to the tutorials which explains how to install Virtualbox on different environments, please make sure you download and install VirtualBox of version 5.0 or greater.

Windows

Virtualbox installation for Windows platform follow instructions from follow link:
http://www.htpcbeginner.com/install-virtualbox-on-windows/

Linux (Ubuntu)

Virtualbox installation for linux Ubuntu distribution platform follow instructions in follow link:
http://ubuntuhandbook.org/index.php/2015/07/install-virtualbox-5-0-ubuntu-15-04-14-04-12-04/

1.2.2 Install Vagrant

Vagrant is the application which aims to help in creating virtual machines in automatic manner, facilitating many reusable options and configurations. Since during the course you will learn how to work with clustered environment of Hadoop and Spark nodes. Vagrant will help you to create and setup such clustered environments on your laptops/private computers and validate your solutions. Moreover using Vagrant files will enable unified automatic check of your assignments. Bellow links to the tutorials which explains how to install Vagrant.

Vagrant binaries could be downloaded here: https://www.vagrantup.com/downloads.html.

Window

Vagrant installation instruction for Windows platform can be found here:
http://www.sitepoint.com/getting-started-vagrant-windows/

Linux(Ubuntu)

Following instructions to install Vagrant on Ubuntu:
https://www.godaddy.com/garage/tech/config/install-vagrant-ubuntu-14-04/

Chapter 2

Assignment Details

2.1 Deep dive into Vagrant

Your first task is to learn about Vagrant and how actually it could be used to manage your virtual machines, also how to setup and configure virtual machine parameters, such that you will be able to use it further to setup your Hadoop cluster on two virtual machines. In order to help you to get started and cover most complex parts of the Vagrant configuration you are provided with a Vagrantfile, which already includes initial configuration for your cluster.

2.1.1 Vagrantfile

```
# -*- mode: ruby -*-
# vi: set ft=ruby :

Vagrant.configure(2) do |config|
    config.vm.define "nodeA" do |nodeA|
        nodeA.vm.box = "ubuntu/trusty64"
        nodeA.vm.network "forwarded_port", guest: 8088, host: 8088
        nodeA.vm.provision "shell", inline: <<-SHELL

            # !!! YOU NEED TO REPLACE HERE CORRECT IP ADDRESS !!!
            sudo echo "IP_ADDRESS slave" >> /etc/hosts

            # Update VM to the latest binaries from distribution
            # package.
            sudo apt-get update && sudo apt-get upgrade -y
            sudo apt-get install -y vim telnet wget curl htop nmon

            # Installing and configuring java.
            cp /vagrant/install_java.sh .
            ./install_java.sh

            # Passwordless ssh communication between two virtual nodes.
            su vagrant -c "ssh-keygen -t rsa -P '' -f /home/vagrant/.ssh/id_rsa"
            mkdir -p /vagrant/files/ssh/
            cp /home/vagrant/.ssh/id_rsa.pub /vagrant/files/ssh/master.pub
            cp /vagrant/after_startup.sh /home/vagrant/.
```

```
            # TODO: In order to complete Hadoop configuration you have to
            # provide here set of Linux shell commands which completes
            # instalation and configuration of Hadoop cluster.

            # !!! Fill your commands here !!!
        SHELL
    end

    config.vm.define "nodeB" do |nodeB|
        nodeB.vm.box = "ubuntu/trusty64"
        nodeA.vm.network "forwarded_port", guest: 8088, host: 9088
        nodeB.vm.provision "shell", inline: <<-SHELL
            # !!! YOU NEED TO REPLACE HERE CORRECT IP ADDRESS !!!
            sudo echo "IP_ADDRESS master" >> /etc/hosts

            # Update VM to the latest binaries from distribution
            # package.
            sudo apt-get update && sudo apt-get upgrade -y
            sudo apt-get install -y vim telnet wget curl htop nmon

            # Installing and configuring java.
            cp /vagrant/install_java.sh .
            ./install_java.sh

            # Passwordless ssh communication between two virtual nodes.
            su vagrant -c "ssh-keygen -t rsa -P '' -f /home/vagrant/.ssh/id_rsa"
            mkdir -p /vagrant/files/ssh/
            cp /home/vagrant/.ssh/id_rsa.pub /vagrant/files/ssh/slave.pub
            cp /vagrant/after_startup.sh /home/vagrant/.

            # TODO: In order to complete Hadoop configuration you have to
            # provide here set of Linux shell commands which completes
            # instalation and configuration of Hadoop cluster.

            # !!! Fill your commands here !!!
        SHELL
    end
end
```

Your goal is to learn Linux commands which will help you to complete installation on Hadoop cluster also you will have to understand which configuration parameter of Vagrant file you have to add in order to fullfill requirements.

More details about vagrant configuration you can find here: https://www.vagrantup.com/docs/

2.2 YARN cluster

1. You are required to create two virtual machines one which will serve as a master of Hadoop cluster and second one is a slave. Your first task will be to change Vagrant configuration file such that master will receive IP address **A.B.C.D** and slave node will receive IP address of **A.B.C.(D+1)** (you can add lines, however cannot delete or change already existing lines inside Vagrantfile). Please see how to get IP addresses:

   (a) **For groups of 3 people**. Master node IP address A.B.C.D wil be constructed from ID's of the team members in the following way: A is the **ID1 mod 100**, B is **ID2 mod 100**, C is **ID3 mod 100** and finally D is **(ID1+ID2+ID3) mod 100**, then slave IP is simply A.B.C.(D+1).

   (b) **For groups of 4 people**. Master node IP address A.B.C.D wil be constructed from ID's of the team members in the following way: A is the **ID1 mod 100**, B is **ID2 mod 100**, C is **ID3 mod 100** and finally D is **(ID4) mod 100**, then slave IP is simply A.B.C.(D+1).

   (c) **For groups of 5 people**. Master node IP address A.B.C.D wil be constructed from ID's of the team members in the following way: A is the **(ID1 + ID5) mod 100**, B is **(ID2 + ID5) mod 100**, C is **(ID3 + ID5) mod 100** and finally D is **(ID5+ID5) mod 100**, then slave IP is simply A.B.C.(D+1).

   Where ID1, ID2, ID3, ID4 and ID5 is students ID numbers according to the order in the README.txt file.

2. You need to take care that hostname of master node will be **master** and of slave node **slave**.

3. Next you need to take care and download hadoop of version 2.7.2, you can follow the link and select required version to download: http://hadoop.apache.org/releases.html. In order to be able to automate the download process and use it inside vagrant tool consider to use linux tools such **wget** or **cURL**. For instance following command allows to download archive from the internet into current directory:

   ```
   wget http://somesite.com/files/archive.zip
   ```

   More examples could be found here:

   - **wget**: http://www.tecmint.com/10-wget-command-examples-in-linux/
   - **cURL**: http://www.thegeekstuff.com/2012/04/curl-examples/

4. Once you've downloaded hadoop distribution file you have to unpack it into virtual machine home folder, i.e. **/home/vagrant/hadoop-2.7.2**. Moreover you will have to setup environment variables required to run hadoop inside virtual machine (your need to discover them on your own).

5. Finally you need to change relevant Hadoop configuration files to setup cluster of two nodes, such that you will be able to run tasks on master node as well as on slaves, i.e. master should include ResourceManager, NodeManager, NameNode and DataNode entities.

   Hint!

   You can prepare required configuration files and place them in the folder nearby your vagrant file and then to add command to Vagrantfile to simply copy the file to the right location, for example you want to copy **"your_configuration.xml"**:

   ```
   cp /vagrant/your_configuration_file.xml /home/vagrant/hadoop-2.7.2/etc/
   ```

   Note: There are also many other valid ways of doing it, here we just proposing only one possible way of doing it.

2.3 Assignment folders structure

Listing 2.1: Assignment files

```
|---|
    |--- Vagrantfile
    |--- start.sh
    |--- install_java.sh
    |--- after_startup.sh
```

You should not change the folders structure.

- **Vagrantfile** - vagrant file with definitions of virtual machine you are going to work with
- **start.sh** - the shell script which will initialize Hadoop cluster.
- **install_java.sh** - shell script which helps to download and install Java 1.8 inside virtual machines
- **after_startup.sh** - auxilary shell script used to complete provisioning and configuration.

2.3.1 Testing distributed shell

Create, start and provision virtual machines with Hadoop cluster by running (it might take quite a while):

```
./start.sh
```

Once it will finish create your VM's and setup configuration of Hadoop cluster you should be able to access Hadoop cluster admin page at http://localhost:8088

- Login into master node

  ```
  vagrant ssh nodeA
  ```

- Enter hadoop bin folder

  ```
  cd hadoop-2.7.2/bin
  ```

- Execute distributed shell with the 'date' command and inspect output.

  ```
  hadoop jar \
  ../share/hadoop/yarn/hadoop-yarn-applications-distributedshell-2.2.0.jar \
  org.apache.hadoop.yarn.applications.distributedshell.Client \
  --jar ../share/hadoop/yarn/hadoop-yarn-applications-distributedshell-2.2.0.jar \
  --shell_command date --num_containers 2 --master_memory 1024
  ```

2.3.2 Testing map reduce configuration

- Login into master node

  ```
  vagrant ssh nodeA
  ```

- Enter hadoop bin folder

  ```
  cd hadoop-2.7.2/bin
  ```

- Execute example of random map-reduce writer

```
hadoop jar ../share/hadoop/mapreduce/hadoop-mapreduce-examples-2.2.0.jar \
randomwriter out
```

2.4 How to approach the solution

In order to solve this assingment and provide required configuration steps to automate Hadoop cluster with Vagrant, we encourage you first of all to start with manual configuration of the YARN cluster, once you will have clear understing of bits and bytes in the cluster setup process it will be much easier to provide automated solution and much clear the set of linux/shell command you have to add into Vagrantfile to complete you mission.

Chapter 3

Submission

This is an electronical submission via course portal on Moodle. To complete your submission you have to:

- Provide your students details at the top of each submitted file.
- Add **hw2_sol.pdf** file with short description of the implementation details and description which explains what was the responsibilities of each teammate during the assignment.
  - **NOTE:** Without this files 10pt will be reduced.
- Add **README.txt** file which will also include information in the format:

Listing 3.1: README.txt

```
First and Second Name;email1;ID1
First and Second Name;email2;ID2
First and Second Name;email3;ID3
First and Second Name;email4;ID4
```

- You need to submit your **Vagrantfile** only.
- All files should be archived with zip and final name for submission should be **homework2.zip**.
- **NOTE:** submissions which will be missed **README.txt** will not be graded.

GOOD LUCK!