

DATABASE SYSTEMS

Assignment : 5

Deadline : 9:00PM, 10th November, 2017

Introduction to Hadoop and MapReduce

For this assignment, you would be running a Hadoop Virtual Machine on your system and write code for the following problems. It will roughly take you 2 hours to code.

Coding Language : Python

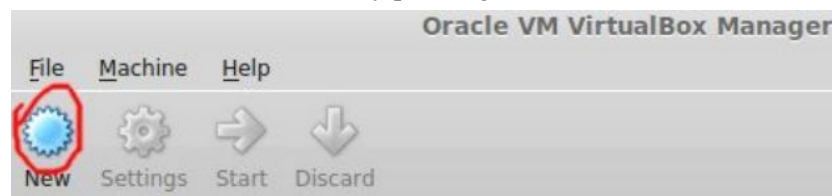
Virtual Machine Setup :

Downloading the VM

1. Download it from
<http://content.udacity-data.com/courses/ud617/Cloudera-Udacity-Training-VM-4.1.1.c.zip>
Warning the zipped file size is 1.7 GB. If you are on a Windows machine you will likely need to use WinRAR to open this .zip file because other methods fail to open the unzipped file (which exceeds the maximum specified 4GB for a .zip file).
2. MD5sum file can be found here
<http://content.udacity-data.com/courses/ud617/Cloudera-Udacity-Training-VM-4.1.1.c.zip.md5>
3. Unzip it. Warning the unzipped size is 4.2GB
4. MD5 hashes for files:
 - 8a610c151d4b1ebdce11542d13dd2a53 ClouderaTrainingVM4.1.1.c.log
 - 6b44c965c1c6062554bf4cc12d11e87e ClouderaTrainingVM4.1.1.c.plist
 - 46dedeba3e0affd8311431d7e370705e ClouderaTrainingVM4.1.1.c.vmdk
 - d41d8cd98f00b204e9800998ecf8427e ClouderaTrainingVM4.1.1.c.vmsd
 - 096956c1cbabeaa652ca63a2d5e14612 ClouderaTrainingVM4.1.1.c.vmx
 - c9f8a375e82ef1e9d96097850e237df9 ClouderaTrainingVM4.1.1.c.vmx
 - 0d7c8becb5a515068e81bb303c794e4f nvram

Using Oracle VirtualBox

1. Download and install VirtualBox from
<https://www.virtualbox.org/wiki/Downloads>
2. Create a new Virtual machine:
 - a. Create a new virtual machine by pressing the 'New' button:



- b. Choose a name, use 'Type': 'Linux':

Name and operating system

Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

Name:

Type:

Version:

- c. Press Next
- d. Select memory size for the VM.

Memory size

Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.

The recommended memory size is **512 MB**.

MB

4 MB 3584 MB

- e. Press Next
- f. Select 'Use an existing virtual hard drive file', click the button to browse to the directory you unzipped the provided VM image and press 'Create'.

Hard drive

If you wish you can add a virtual hard drive to the new machine. You can either create a new hard drive file or select one from the list or from another location using the folder icon.

If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard drive is **8,00 GB**.

☐ Do not add a virtual hard drive

☐ Create a virtual hard drive now

☒ Use an existing virtual hard drive file

- g. Start the VM!



Using VMWare

1. Download and install from https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/6_0
2. Create the Virtual Machine:
 - a. Click on 'Open a Virtual Machine' and, when prompted, navigate to the folder you unzipped the VM, choose the file and click 'Open'.

Welcome To VMware Player



Create a New Virtual Machine

Create a new virtual machine, which will then be added to the top of your library.



Open a Virtual Machine

Open an existing virtual machine, which will then be added to the top of your library.



Upgrade to VMware Workstation

Get advanced features like snapshots, developer tool integration, and more.

- b. Select the machine and click 'Play virtual machine'

Cloudera-Udacity-Training-VM-4.1.1.c


State: Powered Off

OS: Other

Version: Workstation 5.x virtual machine

RAM: 1 GB

 Play virtual machine

 Edit virtual machine settings

Dataset Download :

Dataset for the problem is a dataset on Airports which can be downloaded from moodle.

Problem 1:

Write Mapper and Reducer to get the number of Airports by :

1. Country
2. Type

Problem 2:

Write Mapper and Reducer to find the

1. Country
2. Region

having the highest number of airports

NOTE:

For both the problems and each part, write separate Mappers and Reducers and don't mix the problem.

Resources :

1. Unit 2 and Unit 3 from this online course (~ 12 hours). Unit 1 and 4 are not needed.
<https://in.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617>
2. Chapter 6 should suffice which is also free to download. <http://go.cloudera.com/udacity-lesson-2>

Deliverables/Upload Format :

RollNo / ProblemNumber

Mapper.py

Reducer.py

You can upload the code from the Virtual Machine itself.