DADS 6002 / CI 7301

Big Data Analytics

Hadoop Lab

Quickstart Virtual Machine Installation

1. Install Oracle VirtualBox (Virtual Machine) on Window (Requires RAM at least 8 GB)
   1.1 Download VirtualBox from https://www.virtualbox.org/wiki/Downloads
   1.2 Install VirtualBox by executing the downloaded file.
2. Download the zip file of the Cloudera Quickstart Machine Image from the MS Teams. Then unzip the file (using 7-zip).
3. Start the VirtualBox and click File->Import Appliance then choose the unzip file (.ovf) from 2. to import the cloudera virtual machine.
4. When the Quickstart completely starts, open a terminal window (click on a terminal icon) to work with the machine.

Share A Folder

- To share a folder of Cloudera virtual machine with Window host machine, do the follows:

1. Click Devices menu of the above Virtualbox menu bar, then select Shared Folders Setting
2. Click on Machine Folders and select add folder icon (one with + sign on the right)
3. Enter the pre-created folder path of the Window host to be shared e.g. d:/vbshare. The folder name can be used as the shared folder name e.g. vbshare. Then select Auto-mount and Make Permanent options.

Share A Foloder

4. On a terminal window of the virtual machine, create a shared directory, e.g.

```
# mkdir /home/cloudera/vbshare
```

5. Login as Super user or root

```
# sudo su –
```

6. Mount the shared Window folder using shared folder vbshare to the pre-created directory /home/cloudera/vbshare

```
# mount –t vboxsf vbshare /home/cloudera/vbshare
# exit
```

Hadoop Lab 1

1. Under Quickstart
2. # cd /home/cloudera
3. Edit a file, “test.txt”, using nano command as follows

```
# nano
```

Enter multiple words and multiple lines, then press ctrl o to save to the file named test.txt, and then ctrl x to exit from the nano command. The file will be used for word count problem later.

Hadoop Lab 1

5. Execute the following hadoop file system commands.

```
# hadoop fs –help | more
# hadoop fs -put test.txt /user/cloudera/test.txt
# hadoop fs -ls /user/cloudera
# hadoop fs -cat /user/cloudera/test.txt
# hadoop fs -get /user/cloudera/test.txt test1.txt
# hadoop fs -mkdir /user/cloudera/temp
# hadoop fs -cp /user/cloudera/test.txt /user/cloudera/temp/test.txt
# hadoop fs -rm /user/cloudera/temp/test.txt
# hadoop fs -rmdir /user/cloudera/temp
# hadoop fs -ls
```

Hadoop Lab 1

Submit a MapReduce job for Word Count problem

- Download two files, mapper.py and reducer.py, from MS Teams into the shared folder.

- Submit a MapReduce job through Hadoop Streaming

  ```
  # hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
  –input /user/cloudera/test.txt –output /user/cloudera/wc –
  mapper “python mapper.py” –reducer “python reducer.py” –
  file  mapper.py –file reducer.py
  ```

- Output will be written in multiple files under /user/cloudera/wc e.g. part-00000, ….

Edit a mapper.py as follows:

```
#!/usr/bin/env python

from operator import itemgetter
import sys

for line in sys.stdin:
    for word in line.split():
        print(word+”\t1”)
```

Edit a reducer.py as follows:

```
#!/usr/bin/env python

from operator import itemgetter
import sys

curkey = None
total = 0
for line in sys.stdin:
    key, val = line.split(“\t”,1)
    val = int(val)

    if key == curkey:
        total += val
    else:
        if curkey is not None:
                print(curkey+”\t”+str(total))
        curkey = key
        total = val

print(curkey+”\t”+str(total))
```