CSE 587 DATA INTENSIVE COMPUTING

PROJECT 2 PART B

CHAINED MAPPER-REDUCER

VAIBHAV LELLA, 50169859, vaibhal

KARTHIK CHAGANTI, 50169441, kchagant

Introduction:

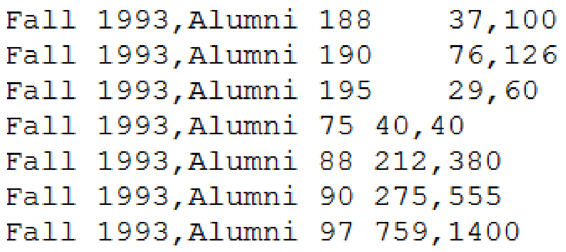
- We have chosen to answer five questions on the given data using chained MR.
- Each of the five MR job chains will have one temporary intermediate output of reducer 1 and a final output of the reducer 2.
- VM used: “Cloudera QuickStart”
- IDE: Eclipse

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

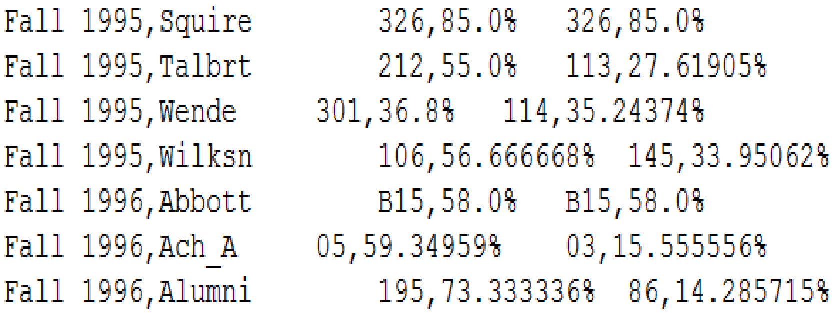
Question 1:

Q1. Give the maximum and minimum utilized (Enrollment/Capacity) rooms in every building Semester-Year wise

- ***1st MR job in the chain:*** Gives the enrollment and capacity of all the rooms in every building semester-year wise.

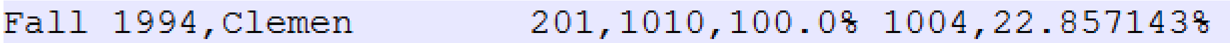
Input	*bina_classschedule.csv*
Output	*part-r-00000* in *Temp1* folder
Example screenshots *(Semester Year, Building Room Numbers Enrollment and Capacity respectively)*	 Fall 1993,Alumni 188 37,100 Fall 1993,Alumni 190 76,126 Fall 1993,Alumni 195 29,60 Fall 1993,Alumni 75 40,40 Fall 1993,Alumni 88 212,380 Fall 1993,Alumni 90 275,555 Fall 1993,Alumni 97 759,1400
Java File Name	*MaxMinbuildingUtilization.java*

- ***Final MR job in the chain:*** Calculate utilization factors of all the rooms in a building and output the room/s with highest and minimum most factors per building semester-year wise.

Input	*part-r-00000* in *Temp1* folder
Output	*part-r-00000* in *output1* folder
Example screenshots *(Semester Year, Building, Room Number/s, Max.Utilization factor (%), Room Number/s, Min.Utilization factor (%) respectively)*	 Fall 1995,Squire 326,85.0% 326,85.0% Fall 1995,Talbrt 212,55.0% 113,27.61905% Fall 1995,Wende 301,36.8% 114,35.24374% Fall 1995,Wilksn 106,56.666668% 145,33.95062% Fall 1996,Abbott B15,58.0% B15,58.0% Fall 1996,Ach_A 05,59.34959% 03,15.555556% Fall 1996,Alumni 195,73.333336% 86,14.285715%
Java File Name	*MaxMinbuildingUtilization.java*

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

- ***Note*:** Incase if there are multiple rooms in a single building with same values of either max or min utilization factors, then all those room numbers are shown followed by a comma and the factor!
    - **Eg:** In the below output example, rooms 201 and 1010 both have same maximum utilization factors of 100%. Hence both the ties are shown beside each other!



```
Fall 1994,Clemen        201,1010,100.0% 1004,22.857143%
```

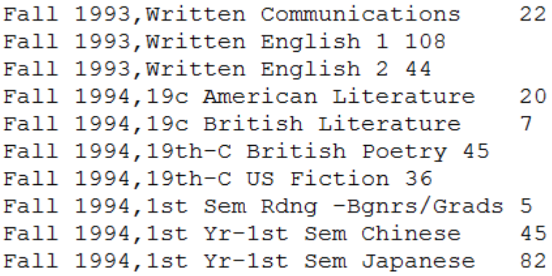
- **Questions/Inference from the result:**
    - *Maximum Utilized Rooms?*
    - *Most under-utilized rooms?*
- **Run Command:**
  hadoop jar /home/cloudera/MaxMinbuildingUtilization.jar MapReduce.MaxMinbuildingUtilization input/bina_classschedule.csv output1

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

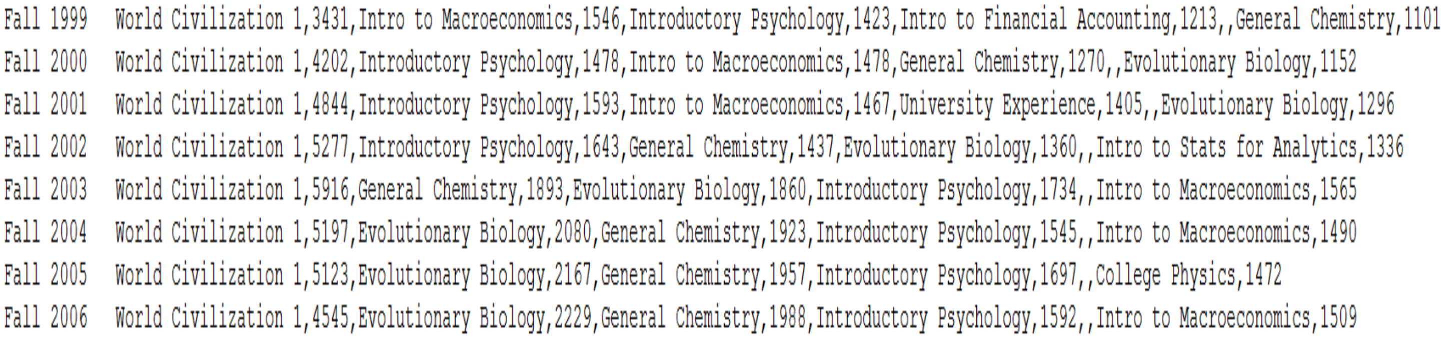
Question 2:

Q2. Give the top 5 highest enrolled courses every Semester-Year wise

- ***1st MR job in the chain:*** Calculate the total enrollment of each of all courses in a semester, semester-year wise.

Input	*bina_classschedule.csv*
Output	*part-r-00000* in *Temp2* folder
Example screenshots *(Semester Year, Course Name, Enrollment respectively)*	 Fall 1993,Written Communications 22 Fall 1993,Written English 1 108 Fall 1993,Written English 2 44 Fall 1994,19c American Literature 20 Fall 1994,19c British Literature 7 Fall 1994,19th-C British Poetry 45 Fall 1994,19th-C US Fiction 36 Fall 1994,1st Sem Rdng -Bgnrs/Grads 5 Fall 1994,1st Yr-1st Sem Chinese 45 Fall 1994,1st Yr-1st Sem Japanese 82
Java File Name	*top5EnrolledCoursesBySem.java*

- ***Final MR job in the chain:*** Arrange the top five highest enrolled courses in a semester in descending order and show them semester-year wise.

Input	*part-r-00000* in *Temp2* folder
Output	*part-r-00000* in *output2* folder
Screenshot: (*Semster-Year, CourseName, Enrollment, CourseName2, Enrollment……CourseName5, Enrollment)*  Fall 1999 World Civilization 1,3431,Intro to Macroeconomics,1546,Introductory Psychology,1423,Intro to Financial Accounting,1213,,General Chemistry,1101 Fall 2000 World Civilization 1,4202,Introductory Psychology,1478,Intro to Macroeconomics,1478,General Chemistry,1270,,Evolutionary Biology,1152 Fall 2001 World Civilization 1,4844,Introductory Psychology,1593,Intro to Macroeconomics,1467,University Experience,1405,,Evolutionary Biology,1296 Fall 2002 World Civilization 1,5277,Introductory Psychology,1643,General Chemistry,1437,Evolutionary Biology,1360,,Intro to Stats for Analytics,1336 Fall 2003 World Civilization 1,5916,General Chemistry,1893,Evolutionary Biology,1860,Introductory Psychology,1734,,Intro to Macroeconomics,1565 Fall 2004 World Civilization 1,5197,Evolutionary Biology,2080,General Chemistry,1923,Introductory Psychology,1545,,Intro to Macroeconomics,1490 Fall 2005 World Civilization 1,5123,Evolutionary Biology,2167,General Chemistry,1957,Introductory Psychology,1697,,College Physics,1472 Fall 2006 World Civilization 1,4545,Evolutionary Biology,2229,General Chemistry,1988,Introductory Psychology,1592,,Intro to Macroeconomics,1509	

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

- **Questions/Inference from the result:**
    - *Most enrolled course in a given semester?*
    - *Most enrolled course over a given list of semesters: From Fall 1999 to Fall 2006, World Civilization 1 was the most enrolled as shown*
- **Run Command:**

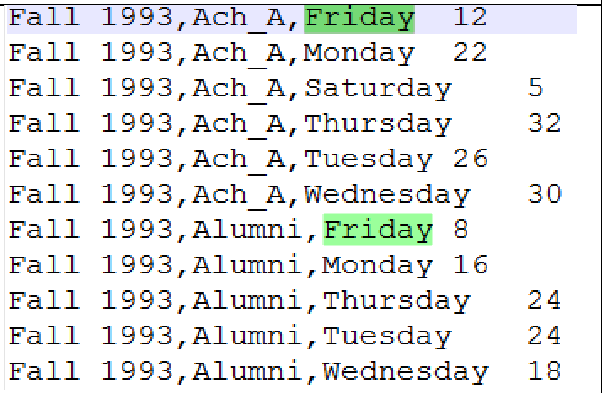
```
hadoop jar /home/cloudera/top5EnrolledCoursesBySem.jar MapReduce.top5EnrolledCoursesBySem
input/bina_classschedule.csv output2
```

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

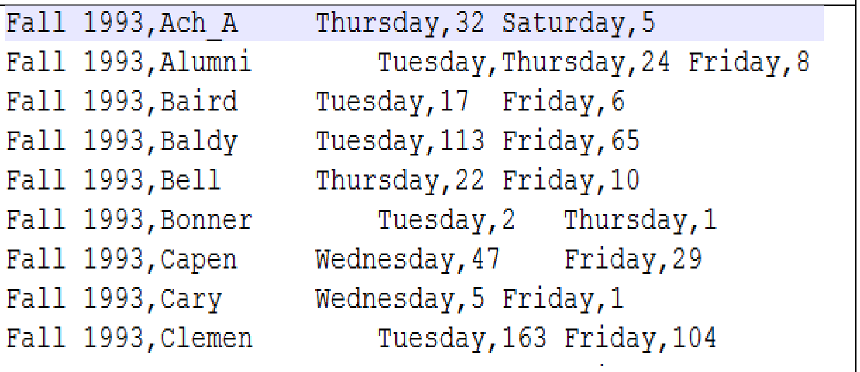
Question 3:

Q3. Give the busiest and the idlest day of the week, by number of reservations for every building, Semester-Year Wise

- ***1st MR job in the chain:*** Calculate and give the total number of reservations for classes on each day of the week in every building Semester-Year wise

Input	*bina_classschedule.csv*
Output	*part-r-00000* in *Temp3* folder
Example screenshots *(Semester Year, Building, Day, Reservations respectively)*	 Fall 1993,Ach_A,Friday 12 Fall 1993,Ach_A,Monday 22 Fall 1993,Ach_A,Saturday 5 Fall 1993,Ach_A,Thursday 32 Fall 1993,Ach_A,Tuesday 26 Fall 1993,Ach_A,Wednesday 30 Fall 1993,Alumni,Friday 8 Fall 1993,Alumni,Monday 16 Fall 1993,Alumni,Thursday 24 Fall 1993,Alumni,Tuesday 24 Fall 1993,Alumni,Wednesday 18
Java File Name	*busyIdleDayOfWeek.java*

- ***Final MR job in the chain:*** Calculate among all the days per building, the busiest and idlest days according to their reservations. Shown the result in a semester-year wise.
-

Input	*part-r-00000* in *Temp3* folder
Output	*part-r-00000* in *output3* folder
Example screenshots *(Semester Year, Building, Busiest Day, Reservation, Idlest-day,Reservation respectively)*	 Fall 1993,Ach_A Thursday,32 Saturday,5 Fall 1993,Alumni Tuesday,Thursday,24 Friday,8 Fall 1993,Baird Tuesday,17 Friday,6 Fall 1993,Baldy Tuesday,113 Friday,65 Fall 1993,Bell Thursday,22 Friday,10 Fall 1993,Bonner Tuesday,2 Thursday,1 Fall 1993,Capen Wednesday,47 Friday,29 Fall 1993,Cary Wednesday,5 Friday,1 Fall 1993,Clemen Tuesday,163 Friday,104
Java File Name	*busyIdleDayOfWeek.java*

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

- **Questions/Inference from the result:**
    - *The busiest day in a given building in a given semester?*
    - *The idlest day in a given building in a given semester?*
    - *The busiest day among all buildings in a given semester? (can be inferred from the output with a further analysis)*
    - *………*
- **Run Command:**

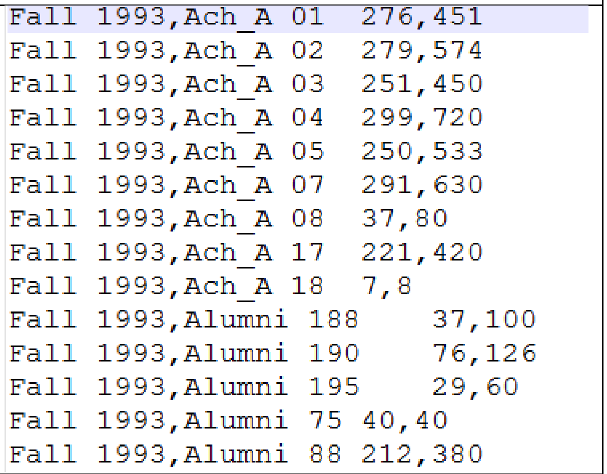
```
hadoop jar /home/cloudera/busyIdleDayOfWeek.jar MapReduce.busyIdleDayOfWeek
input/bina_classschedule.csv output3
```

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

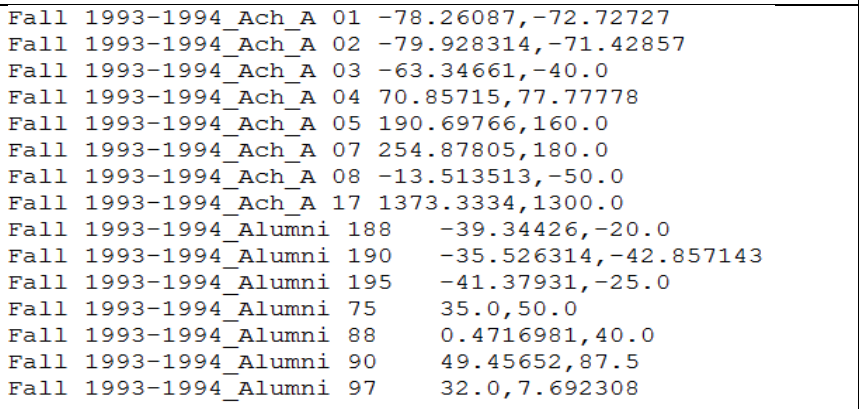
Question 4:

Q4. Give the percentage change in the enrollment and capacity across consecutive years for every room in every building for every semester

- ***1st MR job in the chain:*** Calculate and give the total enrollment and capacities for every room in every building semester-year wise

Input	*bina_classschedule.csv*
Output	*part-r-00000* in *Temp4* folder
Example screenshots *(Semester Year, Building-Room number, Enrollment, capacity respectively)*	 Fall 1993,Ach_A 01 276,451 Fall 1993,Ach_A 02 279,574 Fall 1993,Ach_A 03 251,450 Fall 1993,Ach_A 04 299,720 Fall 1993,Ach_A 05 250,533 Fall 1993,Ach_A 07 291,630 Fall 1993,Ach_A 08 37,80 Fall 1993,Ach_A 17 221,420 Fall 1993,Ach_A 18 7,8 Fall 1993,Alumni 188 37,100 Fall 1993,Alumni 190 76,126 Fall 1993,Alumni 195 29,60 Fall 1993,Alumni 75 40,40 Fall 1993,Alumni 88 212,380
Java File Name	*changeOverTheYears.java*

- ***Final MR job in the chain:*** Calculate the percentage change in both the enrollment and capacity of a given room across consecutive years for all semesters. Calculate those for all the rooms in all the buildings and show the result.
-

Input	*part-r-00000* in *Temp4* folder
Output	*part-r-00000* in *output4* folder
Example screenshots *(Semester, Consecutive-Years, Building Room number, percentage changes in enrollment, capacity respectively)*	 Fall 1993-1994_Ach_A 01 -78.26087,-72.72727 Fall 1993-1994_Ach_A 02 -79.928314,-71.42857 Fall 1993-1994_Ach_A 03 -63.34661,-40.0 Fall 1993-1994_Ach_A 04 70.85715,77.77778 Fall 1993-1994_Ach_A 05 190.69766,160.0 Fall 1993-1994_Ach_A 07 254.87805,180.0 Fall 1993-1994_Ach_A 08 -13.513513,-50.0 Fall 1993-1994_Ach_A 17 1373.3334,1300.0 Fall 1993-1994_Alumni 188 -39.34426,-20.0 Fall 1993-1994_Alumni 190 -35.526314,-42.857143 Fall 1993-1994_Alumni 195 -41.37931,-25.0 Fall 1993-1994_Alumni 75 35.0,50.0 Fall 1993-1994_Alumni 88 0.4716981,40.0 Fall 1993-1994_Alumni 90 49.45652,87.5 Fall 1993-1994_Alumni 97 32.0,7.692308

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

- **Questions/Inference from the result:**
    - *How did the enrollment vary among two years in a given room in a given building?*
    - *How did the capacity vary among two years? This is weird, as usually the capacity of a room tends to stay the same unless new chairs are added or some chairs are removed.*
- **Run Command:**

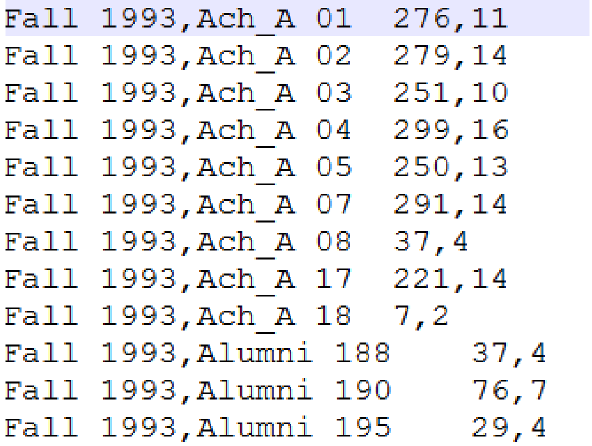
```
hadoop jar /home/cloudera/changeOverTheYears.jar MapReduce.changeOverTheYears
input/bina_classschedule.csv output4
```

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

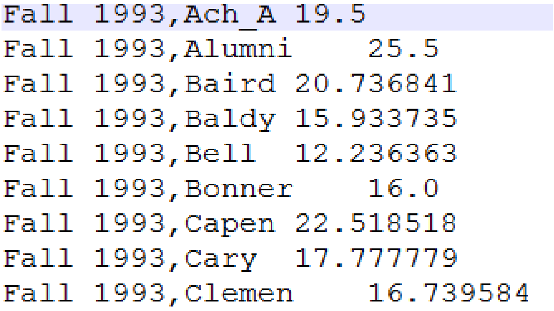
Question 5:

Q4. **Give the Average Enrollment for every building, Semester-Year Wise, computed by dividing the total enrollment of the all the rooms by total numbers rooms in the building.**

- ***1st MR job in the chain:*** Calculate and give the total enrollment of all the rooms in a building and the respective number of courses the room is registered for. The courses count basically gives the number of times the room is used for, for enrollment. Total enrollment = Enrollment per course multiplied by total courses that the room is used for.

Input	*bina_classschedule.csv*
Output	*part-r-00000* in *Temp5* folder
Example screenshots *(Semester Year, Building-Room number, Total Enrollment, Courses count)*	 Fall 1993,Ach_A 01 276,11 Fall 1993,Ach_A 02 279,14 Fall 1993,Ach_A 03 251,10 Fall 1993,Ach_A 04 299,16 Fall 1993,Ach_A 05 250,13 Fall 1993,Ach_A 07 291,14 Fall 1993,Ach_A 08 37,4 Fall 1993,Ach_A 17 221,14 Fall 1993,Ach_A 18 7,2 Fall 1993,Alumni 188 37,4 Fall 1993,Alumni 190 76,7 Fall 1993,Alumni 195 29,4
Java File Name	*AvgEnrol.java*

- ***Final MR job in the chain:*** Calculate the average enrollment of each of the rooms by dividing the total enrollment and the respective number of courses registered on that room. Show the result Semester year wise and for all the rooms in all the buildings!
-

Input	*part-r-00000* in *Temp5* folder
Output	*part-r-00000* in *output5* folder
Example screenshots *(Semester, Building room number, Average Enrollment respectively)*	 Fall 1993,Ach_A 19.5 Fall 1993,Alumni 25.5 Fall 1993,Baird 20.736841 Fall 1993,Baldy 15.933735 Fall 1993,Bell 12.236363 Fall 1993,Bonner 16.0 Fall 1993,Capen 22.518518 Fall 1993,Cary 17.777779 Fall 1993,Clemen 16.739584

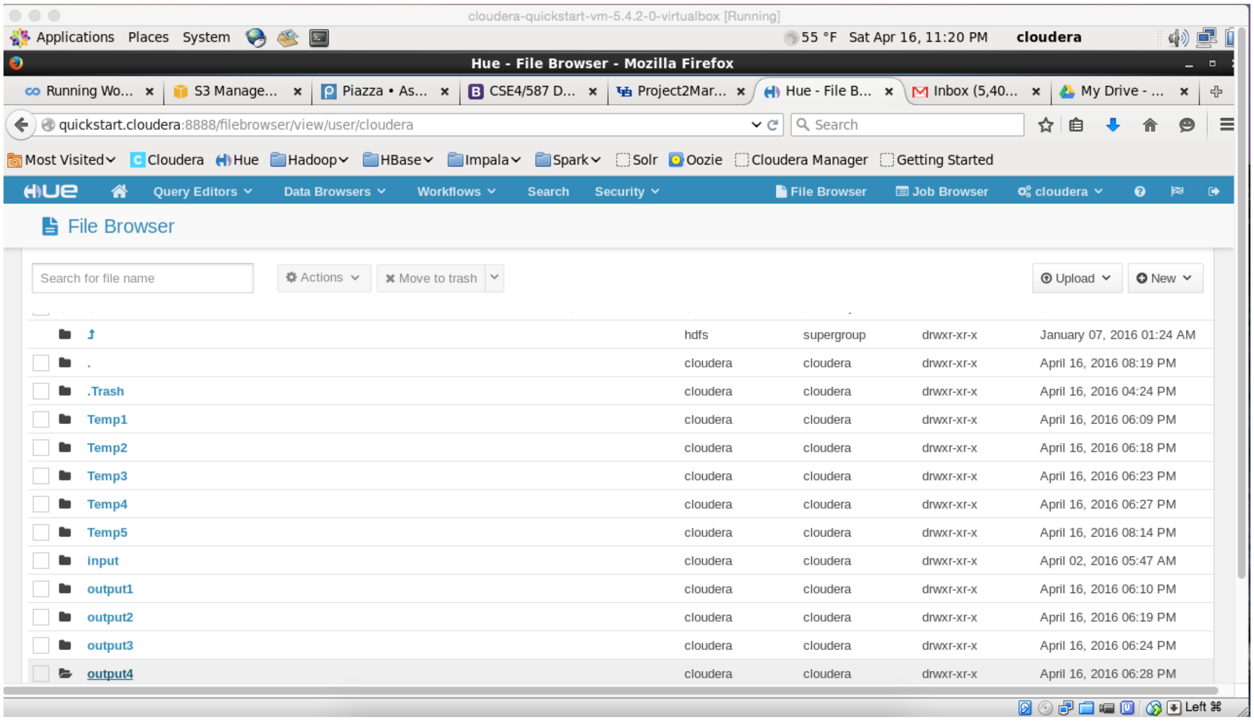
PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION

- **Questions/Inference from the result:**
  - *Average Enrollment for all the rooms in a building?*
  - *On an average, the most enrolled room in a building? – Can be inferred.*
  - *On an average, the most enrolled room in all the buidings – Can be inferred.*

- **Run Command:**

```
hadoop jar /home/cloudera/AvgEnrol.jar MapReduce.AvgEnrol input/bina_classschedule.csv output5
```

Cloudera QuickStart VM – HDFS Directory Snapshot:


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running]
Applications Places System
55 °F Sat Apr 16, 11:20 PM cloudera
Hue - File Browser - Mozilla Firefox
Running Wo... S3 Manage... Piazza • As... CSE4/587 D... Project2Mar... Hue - File B... Inbox (5,40... My Drive - ...
quickstart.cloudera:8888/filebrowser/view/user/cloudera
Search
Most Visited Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
File Browser
Search for file name
Actions
Move to trash
Upload
New
Left ⌘

↥	hdfs	supergroup	drwxr-xr-x	January 07, 2016 01:24 AM
.	cloudera	cloudera	drwxr-xr-x	April 16, 2016 08:19 PM
.Trash	cloudera	cloudera	drwxr-xr-x	April 16, 2016 04:24 PM
Temp1	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:09 PM
Temp2	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:18 PM
Temp3	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:23 PM
Temp4	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:27 PM
Temp5	cloudera	cloudera	drwxr-xr-x	April 16, 2016 08:14 PM
input	cloudera	cloudera	drwxr-xr-x	April 02, 2016 05:47 AM
output1	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:10 PM
output2	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:19 PM
output3	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:24 PM
output4	cloudera	cloudera	drwxr-xr-x	April 16, 2016 06:28 PM

- *The above snapshot has all the temp and output folders …*

PLEASE GOTO THE NEXT PAGE FOR NEXT QUESTION