DATA INTESIVE COMPUTING

PROJECT 2 STAGE 2

OSHIN SANJAY PATWA

UB Id: oshinsan

Person #: 50169203

PRITHVI GOLLU INDRAKUMAR

UB Id: pgolluin

Person #: 50169089

Department of Computer Science

State University of New York at Buffalo

Problem statement:

Analyzing the data and performing exploratory data analysis on the data set related to State University of New York at Buffalo. This data set cons0ists of data for related to the number of students registered, courses offered, time scheduling of classes in various hall rooms in UB campus and their maximum capacities.

Goal:

Our main goal is to understand this data and come up with a list of relevant questions to realize important trends in this given data set. Also we came up with MR algorithms to extract "knowledge" from this data and provide answers to the set of questions designed.

Data Characteristics:

We referred to bina_classschedule.csv file provided by Professor Bina Ramamurthy. We worked on semi structured data in CSV format.

We designed a set of **9 complex questions each with 2 or more mappers** and came up with map reduce solutions to these problems in an attempt to give optimized solutions to these questions. This would help us comprehend the extensive data set in a constructive way which might reflect some important observations, thus helping to make better decisions in the future. This would surely help in the effective utilization of the classroom based on the courses offered in them, students registered etc.

Here is a brief explanation of each questions pertaining to its implementation details and conclusions derived and their usefulness.

PROBLEM 1: To find top 20 courses for which maximum number of students were registered.

Implementation details:

Data Cleaning:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like "Unknown" and "Arr" (for halls).

Map Reduce job: 3 MR Jobs

Job 1: Found the number of students registered for each course.
Key: Name of course value: Number of students registered

Job 2: To sort the records on values
In Mapper 2 we swap the Key, value pairs obtained from stage 2 and in Reducer 2 display the swapped key value pair i.e. key: Number of students registered Value: Course Name. This stage will output the records in sorted order based on the initial values as Hadoop sorts the keys by default.

Job 3: To display top 20 records
In Mapper 3 we set the Key as "1" so that in Reducer we can put all the values in one Array. Accessing the first element of an array will give us the maximum value i.e. course for which the maximum number of students are registered. Thus we have accessed the first 20 records to display the top 20 courses for which maximum courses are registered.

Output:

```
                Bytes Read=221288
        File Output Format Counters
                Bytes Written=535
hadoop@hadoop-VirtualBox:~/hadoop$ hdfs dfs -cat plo1/*
16/04/15 16:01:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
154409  General Chemistry
139710  World Civilization 1
121540  World Civilization 2
103194  Writing 1
70018   Introductory Psychology
63553   College Physics
62191   Intro to Macroeconomics
56938   Intro to Microeconomics
56469   Special Topics
55166   College Calculus 1
50689   Surv Calculus & Appl 1
48160   College Calculus 2
46543   Psychological Statistics
43908   Evolutionary Biology
43279   Intro to Financial Accounting
43117   Intro to Stats for Analytics
42912   Organic Chemistry I
40822   Cell Biology
39215   Human Nutrition
38541   Gen Chem for Engineers
hadoop@hadoop-VirtualBox:~/hadoop$
```

Conclusion:

Maximum number of students have registered for General Chemistry up till now. Besides World Civilization 1 and World Civilization 2 are also in demand. These courses are more in demand at UB according the data provided and hence they should be allocated comparatively greater lecture rooms or lecture rooms having greater capacity so as to avoid mismanagement. These courses should be allocated a greater capacity as they seem to be more demand.

PROBLEM 2: Special Case: To find in which year semester had maximum students registered for General Chemistry.

Implementation details:

Data Cleaning:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). Also in this case we filtered out the data on the basis of the field course_name and considered only those fields that hade “General Chemistry” in their course_name field as we are considering it as a special case.

Map Reduce job: 2 MR Jobs

Job 1: Found the number of students registered for General chemistry for each year semester respectively.

Key: Year semester field value: Number of students registered

Job 2: To sort the records on values
In Mapper 2 we swap the Key, value pairs obtained from stage 2 and in Reducer 2 display the swapped key value pair i.e. key: Number of students registered Value: Year semester. This stage will output the records in sorted order based on the initial values as Hadoop sorts the keys by default.

Output:

```
3303    :Fall 2003_General Chemistry
3420    :Fall 2004_General Chemistry
3433    :Fall 1995_General Chemistry
3505    :Fall 1993_General Chemistry
3666    :Spring 2004_General Chemistry
3769    :Fall 2006_General Chemistry
3796    :Fall 2005_General Chemistry
3910    :Spring 2005_General Chemistry
4023    :Spring 2006_General Chemistry
4040    :Spring 2011_General Chemistry
4077    :Spring 2007_General Chemistry
4090    :Fall 2010_General Chemistry
4111    :Fall 2007_General Chemistry
4127    :Fall 2008_General Chemistry
4134    :Spring 2015_General Chemistry
4238    :Spring 2010_General Chemistry
4307    :Fall 2009_General Chemistry
4314    :Spring 2008_General Chemistry
4380    :Spring 2016_General Chemistry
4495    :Spring 2009_General Chemistry
4644    :Spring 2014_General Chemistry
5061    :Fall 2014_General Chemistry
5397    :Fall 2015_General Chemistry
hadoop@hadoop-VirtualBox:~/hadoop$
hadoop@hadoop-VirtualBox:~/hadoop$
```

Conclusion:

Maximum students had registered for General Chemistry last year in Fall semester i.e. Fall 2015. Also as we can observe from the output greater number of students have registered for this subject in Fall and Spring semesters. Hence this subject should be allocated larger or greater number of classrooms in Fall and Spring semesters. The course capacity should be probably increased for these semesters so that students can get the benefits of the course.

PROBLEM 3: Which hall utilized the maximum time since the year 2000.

Implementation details:

Data Cleaning:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). We have also ignored those tuples that have their time field as “Unknown” or “Before 8: 00 AM”. We have considered only those years since 2000.

Map Reduce job: 3 MR Jobs

Job 1: Found the time utilized by each hall.

Key: Name of hall value: Time utilized (set to 1) as all the classes have a duration of one hour.

Job 2: To sort the records on values
In Mapper 2 we swap the Key, value pairs obtained from stage 2 and in Reducer 2 display the swapped key value pair i.e. key: Name of hall Value: time. This stage will output the records in sorted order based on the initial values as Hadoop sorts the keys by default.

Job 3: To display topmost record
In Mapper 3 we set the Key as “1” so that in Reducer we can put all the values in one Array. Accessing the first element of an array will give us the maximum value i.e. the hall which utilized maximum time.

Output:

```
hadoop@hadoop-VirtualBox:~/hadoop$ hdfs dfs -cat plo3/*
16/04/15 17:15:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19934   Baldy
hadoop@hadoop-VirtualBox:~/hadoop$
hadoop@hadoop-VirtualBox:~/hadoop$
```

Conclusion:

Baldy hall seems to be more occupied as compared to other halls. The schedule for this hall should be decided well in advance as the hall seems to be occupied.

PROBLEM 4: To find the trends of time utilized by Baldy hall for consecutive years since 2000.

Implementation details:

Data Cleaning:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). We have also ignored those tuples that have their time field as “Unknown” or “Before 8: 00 AM”. We have considered only those years since 2000.

Map Reduce job: 2 MR Jobs

JOB 1: To find the time utilized by Baldy hall in each year.
Key: Year Value: Time utilized (set to 1 as all the classes take one hour)

JOB2: To find the change in time utilization for consecutive years:

In Mapper2 we have set the key twice and the reducer 2 displays the change in time utilized for those consecutive years for which the key was set.

Output:

```
hadoop@hadoop-VirtualBox:~/hadoop$ hdfs dfs -cat plo4/*
16/04/15 18:24:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2000-2001       310
2001-2002       -267
2002-2003       -229
2003-2004       51
2004-2005       -61
2005-2006       -60
2006-2007       75
2007-2008       64
2008-2009       22
2009-2010       -66
2010-2011       111
2011-2012       -36
2012-2013       -99
2013-2014       107
2014-2015       -10
2015-2016       88
2016-2017       1762
hadoop@hadoop-VirtualBox:~/hadoop$
```

Conclusion:

There was a remarkable increase in utilization of Baldy hall from year 2016 to 2017 by 1762 hours. Roughly the time utilized by Baldy has increased over the years. Even if there was a decrease in the time utilization that was comparatively less.

PROBLEM 5: What was the peak time for which most number of classes were held for each year.

Implementation details:

Data Cleaning:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). We have also ignored those tuples that

have their time field as “Unknown” or “Before 8: 00 AM”. We have considered only those years since 2000. Besides we have also filtered out the day’s field (field 3), we have ignored those tuples that have “Uknwn” or “Arr” in the days field.

Map Reduce job: 2 MR Jobs

JOB 1: To get how many classes are held respective to each time period.

Key: Year + time period e.g. 2000_3:00AM-3:59AM Value: Number of classes (We have calculated the number of classes) based on the length of the day’s field i.e. if the entry is M-F we have assigned the length 5, if MTW the length assigned is 3.

JOB2: To get the time for which maximum classes are held in each year respectively.

Mapper 2 passes the entire value to the reducer where the Reducer2 calculates the time for which maximum number of classes held keeping the year unique.

Output:

```
        File Output Format Counters
                Bytes Written=597
hadoop@hadoop-VirtualBox:~/hadoop$ hdfs dfs -cat plo/*
16/04/15 21:48:51 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
1993_9:00AM - 9:59AM    864
1994_9:00AM - 9:59AM    1984
1995_9:00AM - 9:59AM    1980
1996_9:00AM - 9:59AM    2068
1997_9:00AM - 9:59AM    2165
1998_9:00AM - 9:59AM    2145
1999_9:00AM - 9:59AM    2171
2000_9:00AM - 9:59AM    2184
2001_9:00AM - 9:59AM    2276
2002_9:00AM - 9:59AM    2291
2003_9:00AM - 9:59AM    2330
2004_9:00AM - 9:59AM    2359
2005_9:00AM - 9:59AM    2327
2006_9:00AM - 9:59AM    2234
2007_9:00AM - 9:59AM    2279
2008_9:00AM - 9:59AM    2332
2009_9:00AM - 9:59AM    2463
2010_9:00AM - 9:59AM    2325
2011_9:00AM - 9:59AM    2303
2012_9:00AM - 9:59AM    2321
2013_9:00AM - 9:59AM    2366
2014_9:00AM - 9:59AM    2395
2015_9:00AM - 9:59AM    2272
hadoop@hadoop-VirtualBox:~/hadoop$
```

Conclusion:

As we can see from the above output the peak time for almost all years is 9:00AM – 9:59 AM. This slot of the day seems to be very busy in all years and there seem to be many classes going on during this time period. Over all the classrooms in UB campus will be filled with students during this period as this is the peak lecture period. Probably more class rooms should be allocated to this time period to avoid mismanagement.

PROBLEM 6: To find the top 5 halls having the maximum capacity for the recent year.

Implementation details:

Data Cleaning:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). Also we have checked that if the number of students registered for the corresponding tuple is greater than zero.

Map Reduce job: 3 MR Jobs

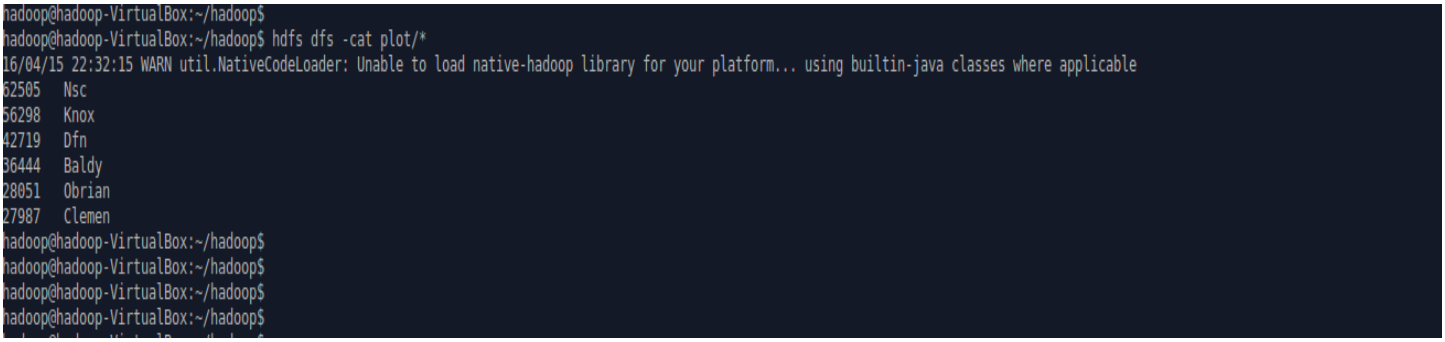
Job 1: To find the capacity for each hall.

Key: Name of hall value: Maximum capacity.

Job 2: To sort the records on values
In Mapper 2 we swap the Key, value pairs obtained from stage 2 and in Reducer 2 display the swapped key value pair i.e. key: Maximum capacity Value: Hall Name. This stage will output the records in sorted order based on the initial values as Hadoop sorts the keys by default.

Job 3: To display top 5 records
In Mapper 3 we set the Key as "1" so that in Reducer we can put all the values in one Array. Accessing the first element of an array will give us the maximum value i.e. hall which has maximum capacity. Thus we have accessed the first 5 records to display the top 5 halls having the maximum capacity.

Output:



Conclusion:

Nsc Hall has the maximum capacity in the UB campus. Hence comparatively greater classes must be being held in this hall or this hall must be utilized for conducting events or seminars attended by huge people. Besides Baldy,knox and Dfn Hall also have greater capacity, as we all know Baldy hall has a Gym which required comparatively more space. As we have seen from Problem 3 Baldy hall has utilized the maximum time overall, also it has a comparatively greater capacity which might mostly be the reason for it being used for a greater time period.

PROBLEM 7: To find the trends and compare the number of students registered for online course v/s those on campus for a particular course for each year.

Implementation details:

Clean up:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like "Unknown" and "Arr" (for halls). Also we have checked that if the number of students registered for the corresponding tuple is greater than zero. We filtered the data to consider the courses offered only after 2009 as online courses were offered only after 2009.

Map Reduce job: 2 MR Jobs

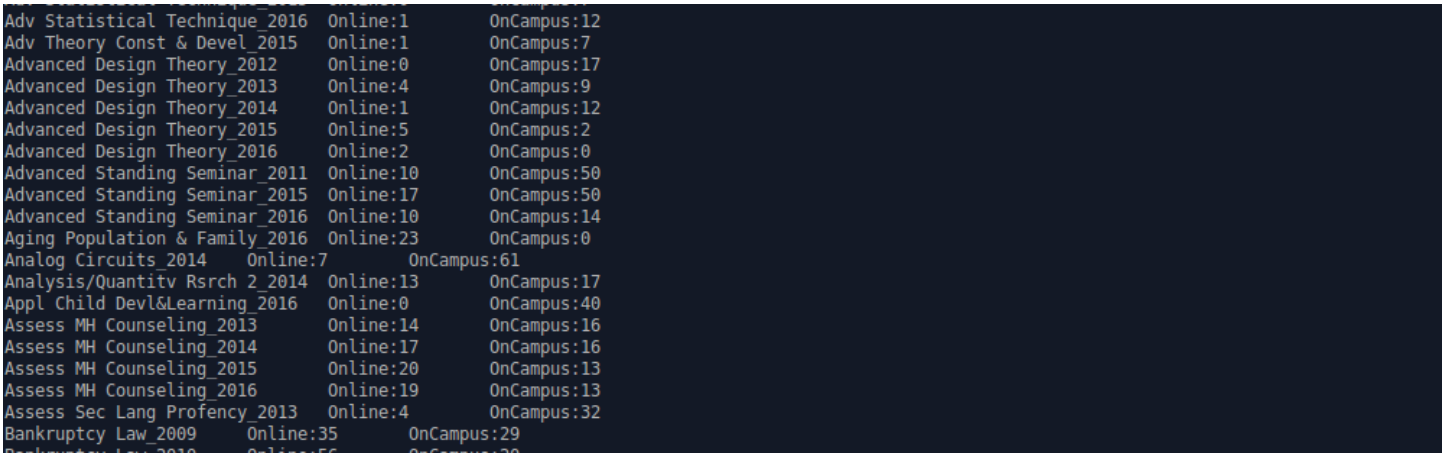
JOB 1: To find the number of students registered for online and on campus course (for each course respectively)

Key: Course_name+Year Value: Online or campus depending on the case

JOB2: Filtering those courses which offer both online and on campus course for that year.

Mapper 2 passes the entire value to the reducer where the Reducer2 calculates and makes the comparison between the number of students registered online and on campus for the same course in each year.

Output:



Conclusion:

This gives us a clear idea about which course had more demand online or on campus for each year. As we can see the output above for the subject Assess MH Counseling the demand for online courses is greater as compared to off campus and greater number of students are registering online. Thus allocating larger space classrooms such courses will probably not be wise decision. However, some courses like Analog Circuits have greater demand on campus hence they should positively be considered while allocating classes. This surely will give us a clear idea about the courses whose online demand is increasing, and probably increasing the capacity would help as people are considering taking up the online course option. Also a point worth noting is that online courses were offered in UB since the year 2009 but not before that.

PROBLEM 8: To find which hall was most efficiently utilized for each semester in each year since the year 2000.

Implementation details:

Clean up:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). Also we have checked that if the number of students registered for the corresponding tuple is greater than zero. Besides we have considered years since only 2000.

JOB 1: To find how efficiently utilized each class was used for each semester year. Here efficiency is a measure being calculated on the basis of difference between max capacity-registered students. Lesser the difference more efficiently has the class been used.

Key: Year Semester+Hall Name Value: Efficiency (Max capacity-Registered students).

JOB2: To find out how many classes were there in each hall which were efficiently utilized.To filter out those entries whose difference between Max capacity- registered students is less than or equal to 25.

Mapper 2 filters out those classes that have a difference of 25 or less for Max capacity-Registered students and passes on to the second Reducer

JOB3: To find a hall for each year with the maximum number of most efficiently utilized rooms.

Mapper 3 passes the entire value to the reducer where the Reducer2 calculates the hall which had the maximum classes efficiently utilized.

Output:

Conclusion:

On observing the output we can see that the most efficiently utilized hall was Capen in Winter 2016 over the years as there is difference of only one. This gives a clear idea about the how efficiency of the hall varied each year and in which year which hall was the most efficiently utilized Also we can see from the above output that Baldy is most frequently most efficiently utilized hall for each semester. From problem3 we know that Baldy also was utilized for maximum time besides it has comparatively greater capacity, from this we can probably conclude that Baldy seems to be always occupied thus making it efficiently utilized majority of the times. Also we can see that the halls are more efficiently utilized in the Fall semester comparatively, from problem 9 we will see that maximum students register during this semester thus being the causes for halls being mostly efficiently utilized. increasing its efficiency

PROBLEM 9: To find the semester for which generally greater students registered.

Implementation details:

Clean up:

We cleaned and filtered the field Students_registered (field 7). We ignored those tuples where there were zero number of students registered or this field had a non -numeric or invalid entry. Also if the corresponding field of Hall_Room has an invalid entry like “Unknown” and “Arr” (for halls). Also we have checked that if the number of students registered for the corresponding tuple is greater than zero.

Map Reduce job: 3 MR Jobs

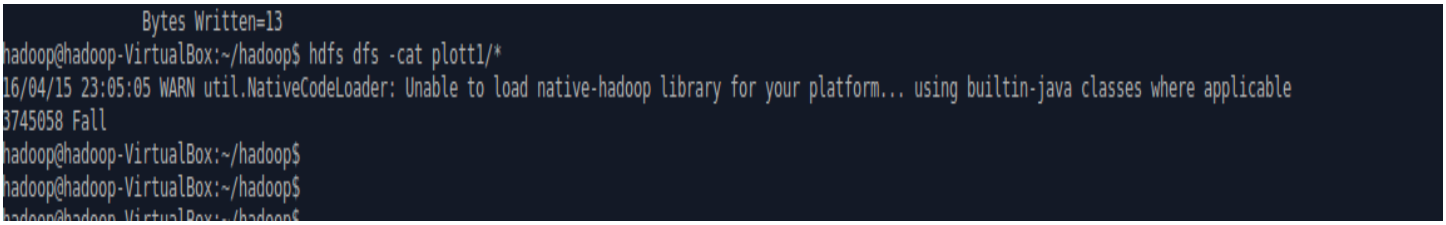
Job 1: Found the capacity for each hall.

Key: semester l value: Maximum capacity.

Job 2: To sort the records on values
In Mapper 2 we swap the Key, value pairs obtained from stage 2 and in Reducer 2 display the swapped key value pair i.e. key: Maximum registered Value: Semester. This stage will output the records in sorted order based on the initial values as Hadoop sorts the keys by default.

Job 3: To display topmost record
In Mapper 3 we set the Key as “1” so that in Reducer we can put all the values in one Array. Accessing the first element of an array will give us the maximum value i.e. the semester for which maximum students have registered.

Output:



Conclusion:

Greater number of students have registered for Fall semester. The Fall semester must be probably offering greater number of courses or having greater capacity. Hence many students must be registering for classes in Fall as compared to Spring, Winter and Summer.