University at Buffalo
The State University of New York

Department of Computer Science and Engineering
School of Engineering and Applied Science

CSE 587- Data Intensive Computing
Project II- Problem 2

provided by:
Alireza Farasat (Person Number: **50060827**)
Rahul Gopalsamy (Person Number: **50163719**)

Spring 2016

Abstract:

The report is aimed at developing an intuition in classroom scheduling scenario at University at Buffalo. Dataset provided in ftp://ftp.cse.buffalo.edu/users/bina/bina_classschedule2.csv were used for our analysis. MapReduce programs were used to answer questions related to our data. In this phase, 23 sample questions addressing some problems in class scheduling have been created and answered using MapReduce (MR). On one hand, 20 of the questions are explored with five 1-step MapReduce jobs. On the other hand, 3 more sophisticated queries have been proposed which need 2-step MR jobs to answer. We developed three 2-step MR programs to study **Anomaly Detection**, **Statistical Inference** and **Regression Analysis**.

1- Introduction:

Classroom scheduling in an Optimization problem, subjected to various constraints. Efficient allocation is important for better utilization of building resources. The following are the major factors that influence this scheduling problem

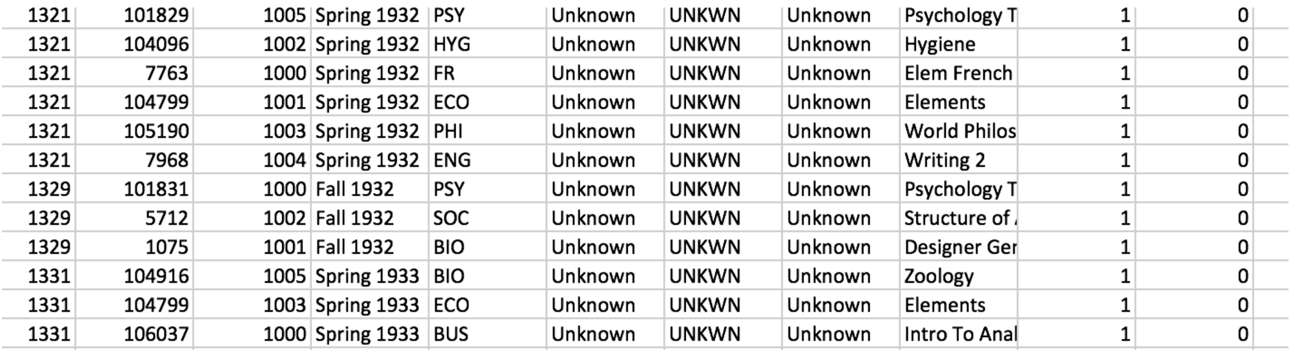
1) Number & capacity of classrooms.
2) Number of students.
3) Number of courses offered per semester.

A set of 23 questions were framed to get an in-depth view on the previous course-class assignment.

2- Dataset Overview:

The data used for this analysis had the following information from the year 1931 to 2017.

1) Semester ID
2) Course ID
3) ID
4) Semester and year
5) Department
6) Location
7) Day
8) Time
9) Course name
10) Number of enrolled students
11) Maximum capacity of the building



1321	101829	1005	Spring 1932	PSY	Unknown	UNKWN	Unknown	Psychology T	1	0
1321	104096	1002	Spring 1932	HYG	Unknown	UNKWN	Unknown	Hygiene	1	0
1321	7763	1000	Spring 1932	FR	Unknown	UNKWN	Unknown	Elem French	1	0
1321	104799	1001	Spring 1932	ECO	Unknown	UNKWN	Unknown	Elements	1	0
1321	105190	1003	Spring 1932	PHI	Unknown	UNKWN	Unknown	World Philos	1	0
1321	7968	1004	Spring 1932	ENG	Unknown	UNKWN	Unknown	Writing 2	1	0
1329	101831	1000	Fall 1932	PSY	Unknown	UNKWN	Unknown	Psychology T	1	0
1329	5712	1002	Fall 1932	SOC	Unknown	UNKWN	Unknown	Structure of ,	1	0
1329	1075	1001	Fall 1932	BIO	Unknown	UNKWN	Unknown	Designer Ger	1	0
1331	104916	1005	Spring 1933	BIO	Unknown	UNKWN	Unknown	Zoology	1	0
1331	104799	1003	Spring 1933	ECO	Unknown	UNKWN	Unknown	Elements	1	0
1331	106037	1000	Spring 1933	BUS	Unknown	UNKWN	Unknown	Intro To Anal	1	0

Over a 660000 courses were provided by the university at this time frame. Several data were unknown, and for our analysis we have filtered out missing information as and when needed.

3- Simple Queries using 1-step MapReduce:

We categorized the data set in terms of

- courses
- Buildings
- Time and day
- Department

A set of questions were addressed based on this categorization to get the required insight.

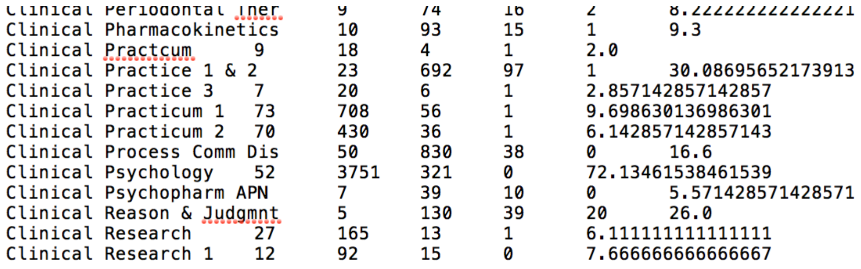
<u>Program 1: (Problem2Q6.java , p6.jar, part-r-00000) singlestageMR</u>

This program was aimed at answering a set of questions related to the courses that were offered over time.

Mapper-Output	
Key	**Value**
course	Number of students enrolled
Reducer-Output	
Key	**Value**
course	Number of times course offered , Total number of students who took the course, Highest class strength, minimum class strength, average number students for the course.

Q1) Number of times each course were offered?
Q2) What is the Overall number of students that have done a particular course?
Q3) What is the maximum strength of a particular course?
Q4) What is the minimum strength of a particular course?
Q5) What is the average strength of the course?

<u>Sample output:</u>



```
Clinical Periodontal Ther       9       74      16      2       8.222222222222221
Clinical Pharmacokinetics       10      93      15      1       9.3
Clinical Practcum       9       18      4       1       2.0
Clinical Practice 1 & 2         23      692     97      1       30.08695652173913
Clinical Practice 3     7       20      6       1       2.857142857142857
Clinical Practicum 1    73      708     56      1       9.698630136986301
Clinical Practicum 2    70      430     36      1       6.142857142857143
Clinical Process Comm Dis       50      830     38      0       16.6
Clinical Psychology     52      3751    321     0       72.13461538461539
Clinical Psychopharm APN        7       39      10      0       5.571428571428571
Clinical Reason & Judgmnt       5       130     39      20      26.0
Clinical Research       27      165     13      1       6.111111111111111
Clinical Research 1     12      92      15      0       7.666666666666667
```

<u>Discussion:</u>

The above questions help us to get an intuition on individual courses. Certain courses were offered only once. This helps us to identify the popular course with most number of students. These courses should be given bigger rooms when scheduling is done.

Program 2: (Problem2Q7.java , p7.jar,part-r-00001) singlestageMR

This program was aimed at answering a set of questions related to all the Academic Departments in the university.

Mapper-Output	
Key	**Value**
Year, Department	number of students enrolled

Reducer-Output	
Key	**Value**
Year, Department	Number of course offered in that semester, Total number of students, Highest class strength, number of courses having students above 50, average number students for the course.

Q6) Number of courses offered under each department during fall, spring, summer and winter semester?

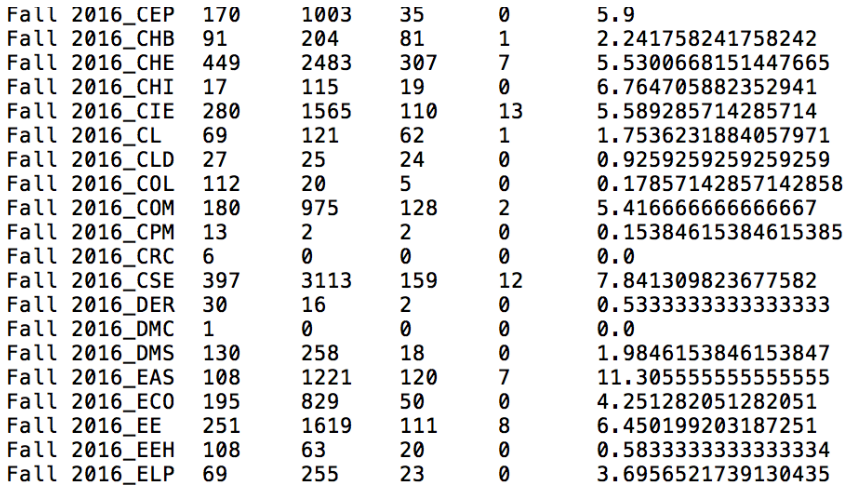
Q7) What is the Overall number of students that have done a particular course?

Q8) What is the maximum strength of a class in the department for given semester?

Q9) How many number of courses offered have more than 50 students for a given semester?

Q10) What is the average number of students in courses for each department for a given semester?

Sample Output:



```
Fall 2016_CEP  170     1003    35      0       5.9
Fall 2016_CHB  91      204     81      1       2.241758241758242
Fall 2016_CHE  449     2483    307     7       5.530066815144766 5
Fall 2016_CHI  17      115     19      0       6.764705882352941
Fall 2016_CIE  280     1565    110     13      5.589285714285714
Fall 2016_CL   69      121     62      1       1.7536231884057971
Fall 2016_CLD  27      25      24      0       0.9259259259259259
Fall 2016_COL  112     20      5       0       0.17857142857142858
Fall 2016_COM  180     975     128     2       5.416666666666667
Fall 2016_CPM  13      2       2       0       0.15384615384615385
Fall 2016_CRC  6       0       0       0       0.0
Fall 2016_CSE  397     3113    159     12      7.841309823677582
Fall 2016_DER  30      16      2       0       0.5333333333333333
Fall 2016_DMC  1       0       0       0       0.0
Fall 2016_DMS  130     258     18      0       1.9846153846153847
Fall 2016_EAS  108     1221    120     7       11.305555555555555
Fall 2016_ECO  195     829     50      0       4.251282051282051
Fall 2016_EE   251     1619    111     8       6.450199203187251
Fall 2016_EEH  108     63      20      0       0.5833333333333334
Fall 2016_ELP  69      255     23      0       3.6956521739130435
```

Discussion:

The above set of questions give us a in-depth view on each department of the university over the years. This can be further used to see the year over year growth in the department in terms of course as well the number of students enrolled.

Mapper-Output	
Key	**Value**
Year, Location	Number of students enrolled, Maximum capacity of the building.

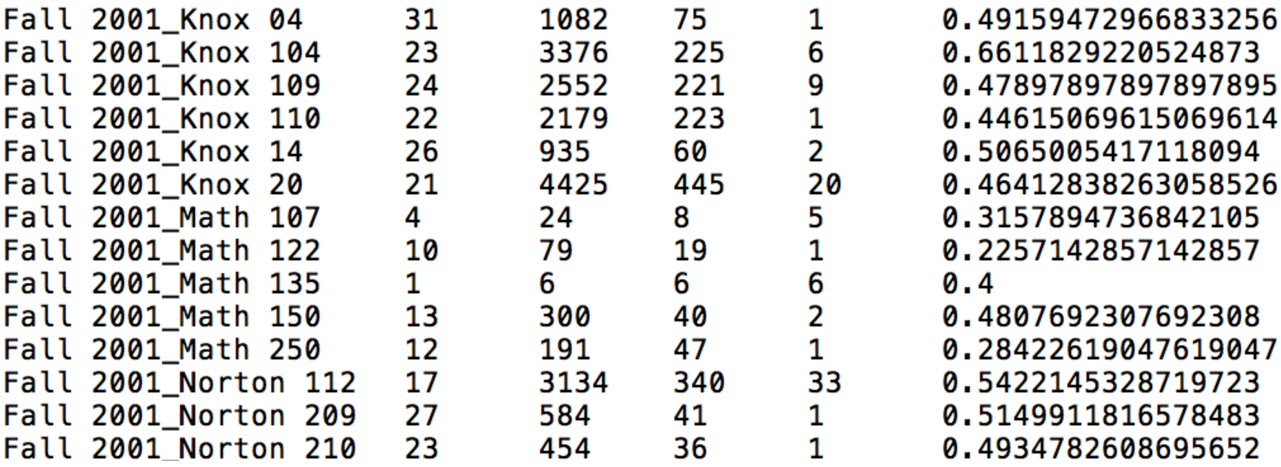
Program 3: (Problem2Q10.java , p10.jar, part-r-00004) singlestageMR

This program was aimed at answering a set of questions related to all the Buildings in the university for each year.

Reducer-Output	
Key	**Value**
Year, Location	Number of course offered, Total number of students, Highest class strength, minimum class strength, building utility ratio.

Q11) Number of courses offered in each building during a particular semester?
Q12) What is the Overall number of students using the building for a given semester?
Q13) What is the maximum strength of a class in the building for a given year?
Q14) What is the minimum strength of a class in the building for a given year?
Q15) What is the utility rate of the building?

Sample Output:



```
Fall 2001_Knox 04      31      1082    75      1       0.49159472966833256
Fall 2001_Knox 104     23      3376    225     6       0.6611829220524873
Fall 2001_Knox 109     24      2552    221     9       0.47897897897897895
Fall 2001_Knox 110     22      2179    223     1       0.44615069615069614
Fall 2001_Knox 14      26      935     60      2       0.5065005417118094
Fall 2001_Knox 20      21      4425    445     20      0.46412838263058526
Fall 2001_Math 107     4       24      8       5       0.3157894736842105
Fall 2001_Math 122     10      79      19      1       0.2257142857142857
Fall 2001_Math 135     1       6       6       6       0.4
Fall 2001_Math 150     13      300     40      2       0.4807692307692308
Fall 2001_Math 250     12      191     47      1       0.28422619047619047
Fall 2001_Norton 112   17      3134    340     33      0.5422145328719723
Fall 2001_Norton 209   27      584     41      1       0.5149911816578483
Fall 2001_Norton 210   23      454     36      1       0.4934782608695652
```

Discussion:

The goal of scheduling is to maximize the utilization of building. The above set of questions are important for efficient scheduling. The building utilization rate is calculated by taking the ratio between strength of students of the class to the maximum capacity of the class. A good course class assignment should maximize this ratio.

Program 4: (Problem2Q9.java , p9.jar, part-r-00003) singlestageMR

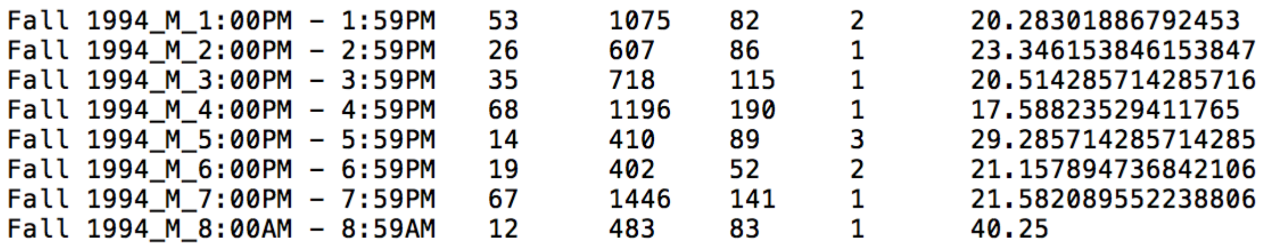
This program is aimed at answering a set of questions related to time and day for a given semester.

Mapper-Output	
Key	**Value**
Year, Day,Time	Number of students enrolled.

Reducer-Output	
Key	**Value**
Year, Day, Time	Number of course offered, Total number of students, Highest class strength, minimum class strength, average number of students per course.

Q16) Number of courses offered in a particular time on a particular day for a given semester?
Q17) What is the Overall number of students in a particular time on a particular day for a given semester?
Q18) What is the maximum strength of a class in a particular time on a particular day for a given semester?
Q19) What is the minimum strength of a class in a particular time on a particular day for a given semester?
Q20) What is the average number of students per course in a particular time on a particular day for a given semester?

Sample Output:



```
Fall 1994_M_1:00PM - 1:59PM    53     1075    82     2      20.28301886792453
Fall 1994_M_2:00PM - 2:59PM    26     607     86     1      23.346153846153847
Fall 1994_M_3:00PM - 3:59PM    35     718     115    1      20.514285714285716
Fall 1994_M_4:00PM - 4:59PM    68     1196    190    1      17.58823529411765
Fall 1994_M_5:00PM - 5:59PM    14     410     89     3      29.285714285714285
Fall 1994_M_6:00PM - 6:59PM    19     402     52     2      21.157894736842106
Fall 1994_M_7:00PM - 7:59PM    67     1446    141    1      21.582089552238806
Fall 1994_M_8:00AM - 8:59AM    12     483     83     1      40.25
```

Discussion:

Time is important factor in terms of scheduling. The number of courses allocated for given interval cannot exceed the overall number of classrooms. The above program will give us insight in the busiest time of the day with most number of courses. The time of the day with highest number of students using the classrooms. Most preferred time of the day for class.

4- Complex Queries using 2-steps MapReduce:

Q21) Anomaly Detection using a 2-step MR algorithm:
The question that we deal with in this section is how one can use an MR algorithm to detect anomalies (outliers) in huge datasets. This is an important step is data mining because many data mining and machine learning algorithms are sensitive respect to anomalies. In this case, an anomaly (outlier) is defined based on building's utilization rate. We are interested in finding UB's building which have been used significantly more that the capacity. Utilization rate is defined as the number of enrolments to the building capacity.

$$util = \frac{\#\ of\ Enrolment}{Capacity}$$

There exist different methods to detect outliers; however, one of the most frequently used technique is using average (AVG) and standard deviation (StD) of the data. In this method, outliers are defined as the observations which are not in the range of AVG±3StD.

$$StD = \sqrt{\frac{\sum_{i=1}^{n}(util_i - AvgUtil)^2}{n}}$$

If we assume that a Gaussian distribution can explain how the data is distributed, then the anomalies are the observations which are not in the tails of distribution and probability of such observations is less than 5%. The following table shows the structure of Mappers and Reducers.

Mapper 1-Output	
Key	**Value**
Building, Semester, Year	Utilization Rate.

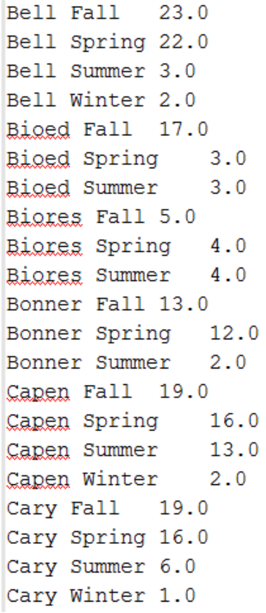
Reducer 1-Output	
Key	**Value**
Building, Semester, Year	Maximum, Average, Standard Deviation

Mapper 2-Output	
Key	**Value**
Building, Semester,	One (if this is an anomaly)

Reducer 2-Output	
Key	**Value**
Building, Semester,	Sum (# of years that the utilization rate is very high)

The final output of this program is which building have more students that the capacity. In other word, which buildings are more popular for overbooked courses?

Sample Output:



```
Bell Fall   23.0
Bell Spring 22.0
Bell Summer 3.0
Bell Winter 2.0
Bioed Fall  17.0
Bioed Spring    3.0
Bioed Summer    3.0
Biores Fall 5.0
Biores Spring   4.0
Biores Summer   4.0
Bonner Fall 13.0
Bonner Spring   12.0
Bonner Summer   2.0
Capen Fall  19.0
Capen Spring    16.0
Capen Summer    13.0
Capen Winter    2.0
Cary Fall   19.0
Cary Spring 16.0
Cary Summer 6.0
Cary Winter 1.0
```

Discussion:

As one can see in most cases, the buildings have more students that they should have in Fall semesters. This helps to plan the courses such that the courses which have more registrations are offered in Spring instead of Fall.

Q22) Statistical Inference about number of students in each course:

Another interesting question is to explore if there are statistically significant difference between the courses offered in Fall and Spring in terms of number of students. To answer this query one needs to compare the average of number of students who have registered in the same courses in Fall and Spring. However, comparing the average cannot answer the question precisely. Even though the averages are not equal, one cannot say that they are different because small changes are expected to the variations. In this MR algorithm, we compare the mean of number of students registered in courses in Fall and Spring Semester. To statistically analyze the situation, the following statistics is used:

$$z = \frac{\bar{X}_F - \bar{X}_S}{\sqrt{\frac{\sigma_F^2}{n} + \frac{\sigma_S^2}{n}}},$$

Where $\bar{X}_F - \bar{X}_S$ shows the difference between number of students registered in each course in Fall and Spring. σ_F^2 and σ_S^2 are variance of Fall and Spring respectively. Depending on the confidence interval, the difference for small values of z $(-1.96 \leq z \leq 1.96)$ is not significant.

Mapper 1-Output	
Key	**Value**
Course Name, Semester, Year	Number of students enrolled (after 1990).

Reducer 1-Output	
Key	**Value**
Course Name, Semester, Year	Sum, Average, Variance

Mapper 2-Output	
Key	**Value**
Course Name	Average, Variance, Semester,

Reducer 2-Output	
Key	**Value**
Course Name	Fall > Spring or Spring > Fall or Fall = Spring

Sample Output:

```
Data Anal Sys for Mkt Dec    Fall = Spring
Data Analysis    Spring = Fall
Data Communicatin/Network    Spring > Fall
Data Integration     Fall = Spring
Data Intensive Computing     Spring > Fall
Data Mining Fall > Spring
Data Models Query Lang   Fall = Spring
Data Structures Fall = Spring
Database Concepts    Fall = Spring
Database Management Syst     Fall > Spring
Database Mgt Systems     Fall > Spring
Database Systems     Spring > Fall
Death Pnlty Law & Pract Fall = Spring
Death in America     Fall = Spring
Definition of America    Fall = Spring
Democracy & Gender   Spring > Fall
Democracy and Gender     Spring > Fall
Democracy in America     Spring > Fall
Dent Mgt Special Needs PT    Spring > Fall
Dental Anatomy/Rdn Lab   Fall > Spring
Dental Biochemistry 1    Fall = Spring
Department Colloquium    Spring = Fall
```

Discussion:

This results also help the UB departments to plan to balance between more demanded courses such that sum of them are offered in Spring and others in Fall. This is very efficient in terms of computation since one can compute all of these statistics very quickly.

Q22) Regression Analysis to Predict number of students in each course:

One of the most important aspect of planning is to have a rough estimation about the future. In this case, if one has some idea about the number of potential students, the planning is much easier. For each course and semester, one can create a time series of the number of enrolments. Lets Y and E denote year and number of enrolment respectively. We define:

$$\hat{E} = a_0 + a_1 Y$$

We can use the following formulas to calculate a_0 and a_1:

$$a_1 = \frac{\sum_{i=1}^{n} Y_i E_i - \frac{1}{n}\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i}$$

And

$$a_0 = \bar{E} - a_1 \bar{Y}$$

Where $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

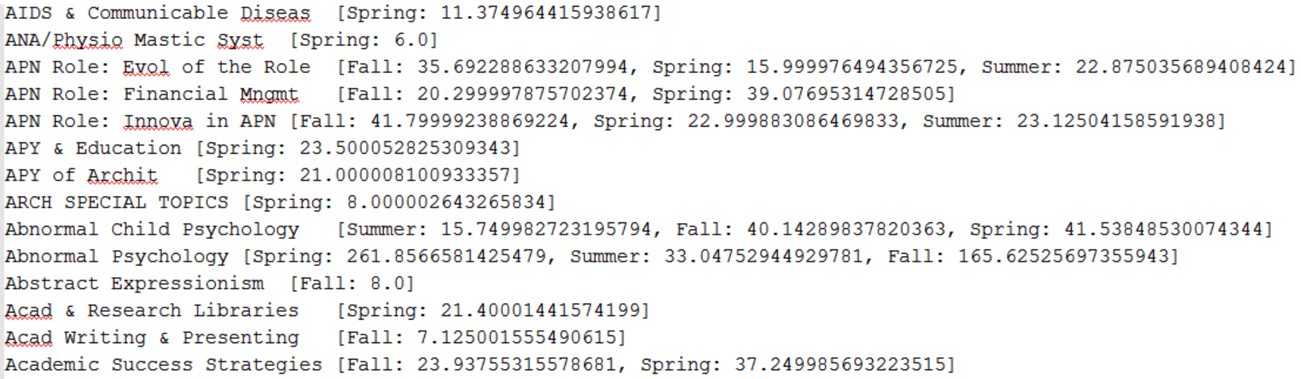
Mapper 1-Output	
Key	**Value**
Course Name, Semester	Year, Number of students enrolled (after 1990).

Reducer 1-Output	
Key	**Value**
Course Name, Semester	Regression coefficient (a0, a1)

Mapper 2-Output	
Key	**Value**
Course Name	Semester, predicted number of students

Reducer 2-Output	
Key	**Value**
Course Name	All Semesters, predicted value for each

Sample Output:



```
AIDS & Communicable Diseas  [Spring: 11.374964415938617]
ANA/Physio Mastic Syst  [Spring: 6.0]
APN Role: Evol of the Role  [Fall: 35.692288633207994, Spring: 15.999976494356725, Summer: 22.875035689408424]
APN Role: Financial Mngmt   [Fall: 20.299997875702374, Spring: 39.07695314728505]
APN Role: Innova in APN [Fall: 41.79999238869224, Spring: 22.999883086469833, Summer: 23.12504158591938]
APY & Education [Spring: 23.500052825309343]
APY of Archit   [Spring: 21.000008100933357]
ARCH SPECIAL TOPICS [Spring: 8.000002643265834]
Abnormal Child Psychology   [Summer: 15.749982723195794, Fall: 40.14289837820363, Spring: 41.53848530074344]
Abnormal Psychology [Spring: 261.8566581425479, Summer: 33.04752944929781, Fall: 165.62525697355943]
Abstract Expressionism  [Fall: 8.0]
Acad & Research Libraries   [Spring: 21.40001441574199]
Acad Writing & Presenting   [Fall: 7.125001555490615]
Academic Success Strategies [Fall: 23.93755315578681, Spring: 37.249985693223515]
```

Discussion:

Although the regression is not an effective method to predict values in time series, it provides the trend and smooths the data. In addition, it is very efficient to calculate regression for these many variables in parallel.