Data Intensive Computing

Project 2 Part 2

Using Hadoop-MapReduce to solve Complicated question on Class Schedule data

Guided By:

Prof. Bina

Junfei Wang (TA)

Luigi (TA)

Group Members:

Mayur Tale

Abhijeet Deshpande

Preface:

We prepared Five complicated question which would require at least two MR stages to solve. We used Eclipse to write Java code and Run the MR jobs on Hadoop provided by Junfei Wang during recitation.

Question 1:

Input: bina_classschedule.csv

What is change in rate of seats utilization (Total Enrollments/Total Capacity) over the years compared between consecutive years?

We used two MR stages to get the comparison between consecutive years. For the first stage we removed the rows with zero total capacity which could have resulted in divide by zero exception. We also removed the rows which has Unknown year.

The first MR stage generates reduced output as year mapped against the rate (total enrollments divided by total capacity).

The second MR stage generates reduced output showing the difference (Current year – Previous year). This difference can be negative, indicating that there has been decline in the rate whereas positive rate indicate increase in the rate.

The greater this difference, it indicates there was better seats utilization during the current year.

Question 2:

Input: bina_classschedule2.csv

What is the difference between in enrollments of students in Spring semester and fall semester for each department for each year?

We used two MR stages to get the difference between fall and spring semester for each years. For the first stage we removed the rows with Unknown year or department.

The first MR stage generates reduced output as department_semster_year as key and total enrollments for that year and semester.

The second MR stage generates reduced output as enrollments for spring semester minus enrollments for fall semester for each year.

The lesser the number (higher negative), indicates that there were more enrollments for that department for that year. The positive number indicates more enrollments in spring semester.

Question 3:

Input: bina_classschedule2.csv

What is the difference in enrollments for Engineering department for consecutive years?

As STEM (Engineering) is our and popular department in the university, we decided to do the findings on Enrollments for ENG department. We used two MR stages to get the difference between enrollments for consecutive years. For the first stage we removed the rows with Unknown year or department.

The first MR stage generates reduced output as year_department as key and total enrollments for that year as value.

The second MR stage generates reduced output as difference in enrollments for consecutive years for engineering department.

The greater difference indicates that there were more enrollments for the current compared to previous year for Engineering department and so the capacity for the engineering department should be increased accordingly. Whereas, the smaller difference (negative number) indicates that there was significant decrease in enrollments for Engineering and so capacity should be reduced accordingly.

Question 4:

Input: bina_classschedule.csv

Which building had the worst utilization of the capacity (capacity - enrollments) for each year?

Capacity utilization is an important part at our university because some classes overflow in enrollments such as CSE 4/587. We used two MR jobs to get the building with worst capacity utilization for each year. We removed the rows in data with unknown year or building and zero capacity rows.

The first MR stage is used to get the reduced output of year_building as key and the difference between capacity and enrollments as value. Whereas, the final MR stage gives the required reduced output of worst building for each year. This

information can be used to decide which buildings are not used properly and can be worked on their arrangements to maintain the highest utilization.

Question 5:

Input: bina_classchedule.csv

For each course, what is the difference in enrollments for Spring semester and Fall Semester for each year when the course was offered in the university?

We decided to check the popularity of each course in each year for Spring and Fall semester. We used two MR stages for this purpose and for cleaning the data removed rows with Unknown year or course.

In the first stage of MR, we generated the reduced out of yea, course and enrollments for Spring and Fall semester.

In the second stage, we took the difference between enrollments for each course for Spring semester and Fall semester for each year putting the difference as value.

The final output can be used to get the popularity for each course for Spring and Fall semester. Positive difference indicates that for the corresponding course, more students enroll in Spring compared to fall semester.

For Example, for Intro to Machine Learning course the difference in enrollments is 135 and 311 for 2015 and 2016 respectively. This indicates the course under prof. Chandola is very popular compared to that under prof. Srihari.