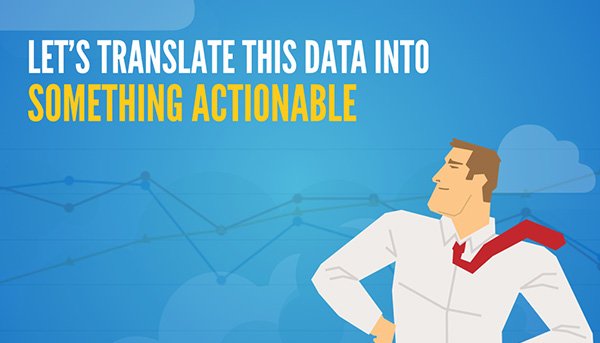
**Parallel processing of UB Classroom Scheduling data using Hadoop MapReduce**

Data Intensive Computing - CSE 587





Submitted by

Ramanpreet Singh Khinda (rkhinda | 5016-9622)

Elroy Preetham Alva (elroypre | 5016-8107)

**Question 1: UB Course Demand**

Every semester UB offers great courses to its students and provide the best world class faculty to teach those courses. However, we have observed that some of the courses are always in high demand either due to increasing technology demand of that field or due to extra ordinary teaching faculty.

UB student affairs want to globalize some of its courses which are highly demanding among students but we are not able to shortlist the courses. Can you utilize the “UB Course Scheduling data” to gather an insight on which courses have always been the first choice for the students?

**Solution:**

We are using number of students enrolled and class capacity to get an insight on the course demand**.** Since the problem is to find the most demanding courses we tackle it using 2 step MR Job. In the 1st step we calculating the total enrollment and capacity of the class for each course over the years and than in the 2nd step averaging over the number of semesters that course have been offered.

In some of the courses the enrolled student’s count is greater than the class capacity which shows that those courses are very high in demand. However, only this factor was not sufficient to derive the results so we are also calculating the %age of vacant seats to figure out that on an average how many seats remained vacant and this information will provide us course demand. We have categorized the course demand into 5 different categories High Demand (< 10%), Above Average Demand (10-25 %), Average Demand (25-40 %), Below Average Demand (40-70 %) and No Demand (> 70%)

Below are the input/output key-value pairs from the Mappers and Reducers used in this MR Job: -

**Input to Mapper 1**

bina\_classschedule.csv

**Output of Mapper 1 <Key, Value>**

<CourseName\_DepartmentName, StudentsEnrolled\_ClassCapacity>

**Input to Reducer 1 <Key, <List of Values>>**

<CourseName\_DepartmentName, <StudentsEnrolled\_ClassCapacity>>

**Output of Reducer 1 <Key, Value>**

<CourseName\_DepartmentName, TotalEnrollment\_TotalCapacity\_SemestersCourseOffered>

**Input to Mapper 2**

Temp file created with output of Reducer 1

**Output of Mapper 2 <Key, Value>**

<CourseName\_DepartmentName, TotalEnrollment\_TotalCapacity\_SemestersCourseOffered >

**Input to Reducer 1 <Key, <List of Values>>**

<CourseName\_DepartmentName, <TotalEnrollment\_TotalCapacity\_SemestersCourseOffered>>

**Output of Reducer 1 <Key, Value>**

<CourseName, DepartmentName\_AverageEnrollment\_AverageCapacity\_CourseDemand>

**Question 2: Lecture Time Analysis**

Class scheduling is very complex problem and its all the more difficult in a department where the enrollments are increasing. We have observed that some of the classes are small as compared to the number of students enrolled and vice versa. This not only make it difficult for the professor to deliver his/her lecture efficiently but also creates a bad image of us among international students.

Being a World Class University we do not want our students to fight for a seat during the class. Can you utilize the “UB Course Scheduling data” to provide an insight on what time of the day have remained most occupied with lectures during the years and what is the most suitable time to re-schedule these courses, so that we can allocate them class of proper size?

**Solution:**

We are using lecture timings and corresponding days of week to get an insight on which time the classes remain idle and can be utilized more efficiently**.** We tackle it using 2 step MR Job. In the 1st step we calculating the total number of lectures being held at a given time and the day than in the 2nd step finding the busiest time of the day for each year.

This information will not only help us analyze that at a particular time how many classes are being scheduled but also provide insight on busiest time of the day and the time when classes remain idle. Thus we can convince student affairs committee to re-schedule class timings of some of the courses and allocate a bigger classroom to them.

Below are the input/output key-value pairs from the Mappers and Reducers used in this MR Job: -

**Input to Mapper 1**

bina\_classschedule.csv

**Output of Mapper 1 <Key, Value>**

<Semester\_LectureTime\_DayOfWeek, LecturesCount>

**Input to Reducer 1 <Key, <List of Value>>**

<Semester\_LectureTime\_DayOfWeek, <LecturesCount >>

**Output of Reducer 1 <Key, Value>**

<Semester\_LectureTime\_DayOfWeek, TotalLecturesCount >

**Input to Mapper 2**

Temp file created with output of Reducer 1

**Output of Mapper 2 <Key, Value>**

<Semester\_DayOfWeek, LectureTime\_TotalLecturesCount >

**Input to Reducer 1 <Key, <List of Values>>**

<Semester\_DayOfWeek, <LectureTime\_TotalLecturesCount> >

**Output of Reducer 1 <Key, Value>**

<Semester\_DayOfWeek, MostBusyTime>