Parallel Processing of Big Data Using Hadoop MapReduce – CSE587

By

Bhargav Repuri (50133959) – brepuri

Pradyumna Reddy Kortha (50167960) - pkortha

Introduction:

Class room scheduling for courses is complex problem. It is all the more difficult in a department where the enrollments are increasing and number of courses and class sizes are increasing. In this project we will design a set of 20 questions that will provide insights into the situation of classroom scheduling at UB North Campus and provide answers by implementing MapReduce algorithms using Hadoop.

Data Source:

For this project we will use the Class room of University At Buffalo. The data is for courses and class rooms from 1931 to 2017. There are two .csv files which have the Class Room scheduling for Courses and One .tsv file for exam rooms scheduling. Following are the headers for Class Room Scheduling:

Semester ID, Semester Name, Location, Days of the Week, Class Time, Course Number, Course Name, Current Number of Students, Max Number of Students Allowed

Questions and Solutions

Question1:

What is the Utilization of different Courses over all the years?

Answer1:

For this task, we calculated the utilization by calculating the Key Value pair for Mapper: <Sem_Course ,(Enrollment)/Maximum Capacity of the Building)>

In the Reducer we output the Key Value pairs which we obtained from the mapper. There is no additional processing in reducer.

Output is as follows:

WordPad Document

Question:2

How many courses occurred for every week-day over all the years?

Answer:2

For this task, the Key Value pair for Mapper: <Sem_Day ,1>.

In the mapper we parse the string and determine whether it is a range of days(M-F) or a string of days (MWF) and process accordingly and emit the Key Value pairs accordingly for all the days on which the course is held

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Day,Sum of all the values >

Output is as follows:

WordPad Document

Question:3

What are the Top Ten Timings which are popular over all the years?

Answer:3

In this we have used two Mapper – Reducers – The First will do the following:

For this task, the Key Value pair for Mapper1: <Sem_Time ,1>

In the Reducer11 we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Time,Sum of all the values >Output is as follows:

The Second one will do the following:

We are using a TreeMap with Key and Value as <Integer,Text> . For Every Input which we got from the Reducer1 we will split on " \\t" and get the count and use that as a Key for Storing in the TreeMap and the Value is the entire Key Value Pair obtained from the Reducer1 . When the TreeMap size exceed 10 we are removing the first Key. We have Override the CleanUp function in the Mapper2 which will emit all the 10 values which are stored in the TreeMap with every Key as NULLWritable . We Have kept the Key as NullWritable because we want the ouput of all the Mappers to go the same reducer. At the Reducer2(TopTenReducer) we have inserted to the TreeMap. So for all the Key Value pairs which the Reducer2 is receiving we insert them into another TreeMap with the Key obtained by parsing the value.

Whenever the TreeMap size exceeds 10 we remove the first Key entry.

After processing all the values the Reducer2 emits all the 10 values stored in the TreeMap

Fall 2015_Before 8:00AM	6371
Fall 2016_Before 8:00AM	6349
Fall 2014_Before 8:00AM	6186
Fall 2013_Before 8:00AM	5369
Fall 2012_Before 8:00AM	5310

Spring 2016_Before 8:00AM 3537

Spring 2015_Before 8:00AM 3366

Spring 2014_Before 8:00AM 2594

Spring 2013_Before 8:00AM 1598

Summer 2016_Before 8:00AM 905

Question:4

What are Buildings which are Efficiently Utilized over all the years?

Answer:4

For this task, we calculated the utilization by calculating the Key Value pair for Mapper: <Sem_Building ,(Enrollment)/Maximum Capacity of the Building)>

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Combined Utilization of all Course which are held in that Building>

Output is as follows:

WordPad Document

Question:5

What are the Top Ten Buildings which are popular in terms of Efficient Utilization?

Answer:5

In this we have used two Mapper – Reducers – The First will do the following:

For this task, we calculated the utilization by calculating the Key Value pair for Mapper1: <Sem_Building , ,(Enrollment)/Maximum Capacity of the Building)>

In the Reducer11 we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Sum of all the values >Output is as follows:

The Second one will do the following:

We are using a TreeMap with Key and Value as <Integer,Text> . For Every Input which we got from the Reducer1 we will split on " \\t" and get the count and use that as a Key for Storing in the TreeMap and the Value is the entire Key Value Pair obtained from the Reducer1 . When the TreeMap size exceed 10 we are removing the first Key. We have Override the CleanUp function in the Mapper2 which will emit all the 10 values which are stored in the TreeMap with every Key as NULLWritable . We Have kept the Key as NullWritable because we want the ouput of all the Mappers to go the same reducer. At the Reducer2(TopTenReducer) we have inserted to the TreeMap. So for all the Key Value pairs which the Reducer2 is receiving we insert them into another TreeMap with the Key obtained by parsing the value.

Whenever the TreeMap size exceeds 10 we remove the first Key entry.

After processing all the values the Reducer2 emits all the 10 values stored in the TreeMap

Output is as follows:

Spring 2002_Intro to Microproc Lab 308.0

Spring 2003_Intro to Microproc Lab 282.0

Spring 2004_Intro to Microproc Lab 274.0

Spring 2007_Intro to Microproc Lab 268.0

Spring 2006_Intro to Microproc Lab 266.0

Spring 2001_Intro to Microproc Lab 262.0

Spring 2005_Intro to Microproc Lab 250.0

Fall 2000_World Civilization 1 248.0

Fall 1993_Introduction to Engng 222.0

Fall 1993_College Calculus 1 209.0

Question:6

What are all the Courses which have Efficient Seating Arrangement?

Answer:6

For this task, we calculated the utilization by calculating the Key Value pair for Mapper: <Sem_Course ,(Enrollment)/Maximum Capacity of the Building)>

In the Reducer we just emit only the rows which have values >=1

Output is as follows:

WordPad Document

Question:7

What are Top Ten Courses across all the departments which most of the Students Were Enrolled?

Answer:7

Mapper Key Value pairs are <NULLWritable,Text>

We are using a TreeMap with Key and Value as <Integer,Text> . When the TreeMap size exceed 10 we are removing the first Key. We have Override the CleanUp function in the Mapper1 which will emit all the 10 values which are stored in the TreeMap with every Key as NULLWritable . We Have kept the Key

as NullWritable because we want the ouput of all the Mappers to go the same reducer. At the Reducer(TopTenReducer) we have inserted to the TreeMap. So for all the Key Value pairs which the Reducer is receiving we insert them into another TreeMap with the Key obtained by parsing the value.

Whenever the TreeMap size exceeds 10 we remove the first Key entry.

After processing all the values the Reducer emits all the 10 values stored in the TreeMapOutput is as follows:

Fall 2014_Corporation Finance;574

Fall 2015_Corporation Finance;500

Fall 2014_Introductory Psychology;454

Spring 2015_Introductory Psychology;453

Spring 2008_Introduction to Sociology;452

Spring 2007_Introductory Psychology;451

Spring 2008_Introductory Psychology;450

Fall 2005_Evolutionary Biology;449

Fall 1995_Introductory Psychology;448

Spring 2005_Introductory Psychology;447

Question:8

What are the Different Departments and their Utilization of Classrooms?

Answer:8

For this task, we calculated the utilization by calculating the Key Value pair for Mapper: <Sem_Department ,(Enrollment)/Maximum Capacity of the Building)>

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Combined Utilization of all Course which are held in that Building>

Output is as follows:

WordPad Document

Question:9

What are the Top Ten Popular Departments in UB in terms of students Enrollment?

Answer:9

In this we have used two Mapper – Reducers – The First will do the following:

For this task, we calculated the utilization by calculating the Key Value pair for Mapper1:
<Sem_Department , ,Enrollment>

In the Reducer11 we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Department,Sum of all the values >Output is as follows:

The Second one will do the following:

We are using a TreeMap with Key and Value as <Integer,Text> . For Every Input which we got from the Reducer1 we will split on " \\t" and get the count and use that as a Key for Storing in the TreeMap and the Value is the entire Key Value Pair obtained from the Reducer1 . When the TreeMap size exceed 10 we are removing the first Key. We have Override the CleanUp function in the Mapper2 which will emit all the 10 values which are stored in the TreeMap with every Key as NULLWritable . We Have kept the Key as NullWritable because we want the ouput of all the Mappers to go the same reducer. At the Reducer2(TopTenReducer) we have inserted to the TreeMap. So for all the Key Value pairs which the Reducer2 is receiving we insert them into another TreeMap with the Key obtained by parsing the value.

Whenever the TreeMap size exceeds 10 we remove the first Key entry.

After processing all the values the Reducer2 emits all the 10 values stored in the TreeMap

Output is as follows:

Fall 2013_CHE 11156

Fall 2015_CHE 10853

Fall 2014_CHE 10836

Fall 2012_CHE 10418

Fall 2014_MTH 10107

Fall 2015_MTH 10023

Fall 2013_MTH 9917

Fall 2012_MTH 9737

Fall 2011_CHE 9605

Fall 2009_CHE 9431

Question:10

How many Courses each Building have served over all the years?

Answer:10

For this task, we calculated the utilization by calculating the Key Value pair for Mapper:
<Sem_Building,1>

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Total Courses >

Output is as follows:


WordPad Document

Question:11

Which Day of the Week is most populated in UB in terms of students over all the years combined – Top Ten?

Answer:11

In this we have used two Mapper – Reducers – The First will do the following:

For this task, we calculated the utilization by calculating the Key Value pair for Mapper1: <Sem_Day , ,Enrollment>

In the Reducer11 we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Day,Sum of all the values >Output is as follows:

The Second one will do the following:

We are using a TreeMap with Key and Value as <Integer,Text> . For Every Input which we got from the Reducer1 we will split on " \\t" and get the count and use that as a Key for Storing in the TreeMap and the Value is the entire Key Value Pair obtained from the Reducer1 . When the TreeMap size exceed 10 we are removing the first Key. We have Override the CleanUp function in the Mapper2 which will emit all the 10 values which are stored in the TreeMap with every Key as NULLWritable . We Have kept the Key as NullWritable because we want the ouput of all the Mappers to go the same reducer. At the Reducer2(TopTenReducer) we have inserted to the TreeMap. So for all the Key Value pairs which the Reducer2 is receiving we insert them into another TreeMap with the Key obtained by parsing the value.

Whenever the TreeMap size exceeds 10 we remove the first Key entry.

After processing all the values the Reducer2 emits all the 10 values stored in the TreeMap

Output is as follows:

Fall 2016_Thursday 14168

Fall 2015_Thursday 14163

Fall 2014_Thursday 13831

Fall 2013_Thursday 12193

Fall 2012_Thursday 12051

Spring 2016_Thursday 8452

Spring 2015_Thursday 8151

Spring 2014_Thursday 6590

Spring 2013_Thursday 4568

Spring 2012_Thursday 2662

Question:12

How many students each building have served over all the years combined?

Answer:12

For this task, the Key Value pair for Mapper: <Sem_Building,(Enrollment) >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Total Enrollment>

Output is as follows:

WordPad Document

Question:13

How many Courses each department have offered over all the years in total?

Answer:13

For this task, the Key Value pair for Mapper: <Sem_Department ,1>

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Sum of values >

Output is as follows:

WordPad Document

Question:14

How many Online courses were offered in total?

Answer:14

For this task, Key Value pair for Mapper: <Sem_Online ,1 >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Online,Total Values >

Output is as follows:

Fall 2007	1
Fall 2008	1
Fall 2009	1
Fall 2010	1
Fall 2011	1
Fall 2012	55
Fall 2013	77
Fall 2014	98
Fall 2015	105
Fall 2016	128
Spring 2006	1
Spring 2010	3
Spring 2011	1
Spring 2012	14
Spring 2013	34
Spring 2014	49
Spring 2015	62
Spring 2016	83
Summer 2010	3
Summer 2011	1
Summer 2012	10
Summer 2013	92
Summer 2014	116
Summer 2015	111
Summer 2016	149
Winter 2014	6
Winter 2015	28
Winter 2016	39
Winter 2017	32

Question:15

How many number of Seminars are offered in each Semester over all the years?

Answer:15

For this task, Key Value pair for Mapper: <Sem_Seminar ,1 >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Seminar,Total Values >

Output is as follows:


WordPad Document

Question:16

How many students were enrolled in each department over all the years ?

Answer:16

For this task, Key Value pair for Mapper: <Sem_Department ,Enrollment >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Department,Total Values >

Output is as follows:


WordPad Document

Question:17

How many Exams Each Building have server over all the years?

Answer:17

For this task, Key Value pair for Mapper: <Sem_Building,1 >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Total Values >

Output is as follows:


WordPad Document

Question:18

What are the total number of exams happened in each Semester over all the years?

Answer:18

For this task, Key Value pair for Mapper: <Sem,1 >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem,Total Values >

Output is as follows:

WordPad Document

Question:19

What are the total number of exams happened for each department over all the years?

Answer:19

For this task, Key Value pair for Mapper: <Sem_Department,1 >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Department,Total Values >

Output is as follows:

WordPad Document

Question:20

What are the Top Ten Buildings which are more populated over all the years?

Answer:20

In this we have used two Mapper – Reducers – The First will do the following:

For this task, we calculated the utilization by calculating the Key Value pair for Mapper1: <Sem_Building , ,Enrollment>

In the Reducer11 we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Building,Sum of all the values >Output is as follows:

The Second one will do the following:

We are using a TreeMap with Key and Value as <Integer,Text> . For Every Input which we got from the Reducer1 we will split on " \\t" and get the count and use that as a Key for Storing in the TreeMap and the Value is the entire Key Value Pair obtained from the Reducer1 . When the TreeMap size exceed 10 we are removing the first Key. We have Override the CleanUp function in the Mapper2 which will emit all the 10 values which are stored in the TreeMap with every Key as NULLWritable . We Have kept the Key as NullWritable because we want the ouput of all the Mappers to go the same reducer. At the Reducer2(TopTenReducer) we have inserted to the TreeMap. So for all the Key Value pairs which the Reducer2 is receiving we insert them into another TreeMap with the Key obtained by parsing the value.

Whenever the TreeMap size exceeds 10 we remove the first Key entry.

After processing all the values the Reducer2 emits all the 10 values stored in the TreeMap

Output is as follows:

Fall 2015_Nsc 22261

Fall 2014_Nsc 21963

Fall 2013_Nsc 21708

Fall 2011_Nsc 20555

Fall 2010_Nsc 20389

Fall 2007_Nsc 19884

Fall 2009_Nsc 19808

Fall 2012_Nsc 19792

Fall 2006_Nsc 19725

Spring 2011_Nsc 19608

Question:21

What are the total classes happened at different timings in each Semester over all the years?

Answer:21

For this task, Key Value pair for Mapper: <Sem_Timing,1 >

In the Reducer we calculate the sum of all the Key Value pairs obtained from the Mappers and the output of the Reducer is <Sem_Timing,Total Values >

Output is as follows:

WordPad Document