NoSQL Assignment 3

IMT2022001 Siddharth Menon
IMT2022006 Shreyas Arun Saggere
IMT2022082 Sathvik S. Rao
IMT2022507 Vishruth Vijay

April 14, 2025

Question 1:

The objective in this question was to define a Hive schema for the data files given, and create Hive tables accordingly. We initially created a staging table in order to load the data as it is. Then we loaded the data following our schema. We also defined an error schema for erroneous records. Our Hive schema looks as follows:

1. **Course_Attendance.csv** The **course_attendance_staging** table was used for staging.

Column Name	**Data Type**	**Description**
course	TEXT	Course identifier or name
name_hash	TEXT	Hashed student name
email_hash	TEXT	Hashed email address
member_id_hash	TEXT	Hashed member ID
classes_attended	INTEGER	Number of classes attended
classes_absent	INTEGER	Number of classes missed
instructors	ARRAY⟨*String*⟩	List of instructors (parsed from string)
avg_attendance_percent	DECIMAL(5,2)	Average attendance percentage

Table 1: Schema of the `course_attendance` Table

2. **Enrollment_Data.csv** The **enrollment_data_staging** table was used for staging.

Column Name	**Data Type**	**Description**
serial_no	INTEGER	Serial number
course	TEXT	Course name
status	TEXT	Enrollment status
course_type	TEXT	Type of course
course_variant	TEXT	Course variant
academia_lms	TEXT	NA
student_id	TEXT	Student ID
student_name	TEXT	Name of the student
program	TEXT	Program enrolled in
batch	TEXT	Batch or cohort
period	TEXT	Academic period or semester
enrollment_date	DATE	Date of enrollment
primary_faculty	ARRAY⟨TEXT⟩	List of faculty members for the course

Table 2: Schema of the `enrollment_data` Table

3. **grade_roster_report.csv**

Column Name	**Data Type**	**Description**
academy_location	STRING	Location of the academy or campus
student_id	STRING	Unique identifier for the student
student_status	STRING	Status of the student
admission_id	STRING	Admission ID associated with the student
admission_status	STRING	Status of the admission
student_name	STRING	Full name of the student
program_code_name	STRING	Program code and name
batch	STRING	Batch or cohort information
period	STRING	Academic period or term
subject_code_name	STRING	Subject code and name
course_type	STRING	Type of course
section	STRING	Section identifier
faculty_name	STRING	Name of the faculty/instructor
course_credit	INT	Credit value of the course
obtained_marks_grade	STRING	Marks or grade obtained by the student
out_of_marks_grade	STRING	Maximum marks or grade possible
exam_result	STRING	Exam result

Table 3: Schema of the `grade_roster_report` Table

4. **Erroneous record schema**

Column Name	**Data Type**	**Description**
source_table	STRING	Name of the source table where the error originated
original_row	STRING	Raw data row containing the error
column_name	STRING	Name of the column that caused the error
issue_type	STRING	Type or category of the issue (e.g., NULL, format error)
error_description	STRING	Detailed description of the error

Table 4: Schema of the `error_records` Table

Question 2

The **course_attendance table, enrollment_data table,grade_roster_report** table were joined on the fields **student_id and course field** to create a warehouse filtering out erroneous data. This warehouse was used in order to perform our analytical queries.
The analytical queries were:

1. **Query 1**: Comparing the mode grade and average attendance of different students of different programs for every course
2. **Query 2**: Percentage of students with a below 50% attendance record, per program, per course.
3. **Query 3**: Identification of top 3 courses by average attendance for each program.

Question 3

We set up a new data warehouse that is partitioned by **batch/program** and bucketed into 4 buckets by **course**. The goal was to evaluate the impact of partitioning on query performance. We chose to partition by **program** because our queries frequently include a 'GROUP BY' operation on the program field. For the same reason we chose to bucket by **course**.

Question 4: Converting Hive queries to Pig

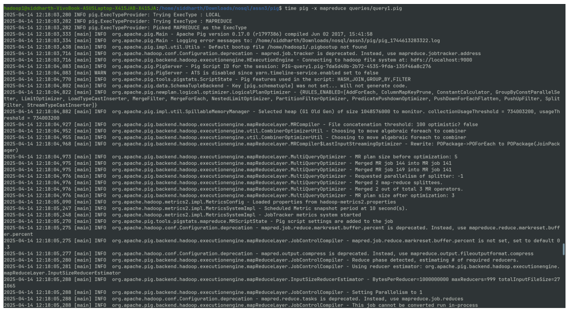
Important note: A difference in schema between Hive and Pig: in the Hive schema, we have two fields `batch` and `period` – which would contain values like (Master of Technology, 2024–2026 CSE), while these same fields are renamed to `program_name` and `batch`, where `batch` is analogous to the `period` field in the Hive schema. For this section, those two fields will be referred to by the names they are referred to in the Pig schema.

To analyse the runtime performance of pig queries, after performing transformations on the data, we had to then use the `dump` command, as Pig uses lazy evaluation, and the actual operations aren't done until some kind of output operation is done.

Furthermore, since the code for each of the queries was put in a different file and executed one after the other from the terminal, the load had to be done in all the three query files, causing an additional overhead for a load every time. If the 3 queries are executed together, and not timed separately, the time taken is lesser, as we load only once. That analysis, too, has been done.

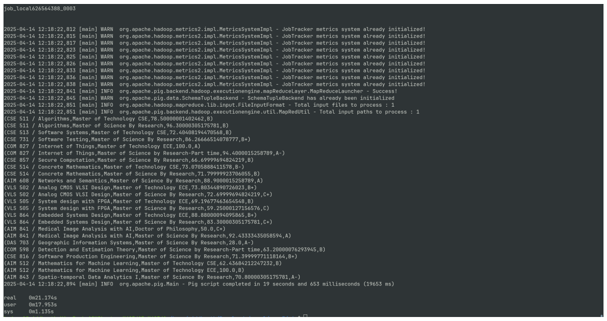
The three queries that were converted from Hive to Pig Latin were:

1. Comparing the mode grade and average attendance of different students of different programs for every course.



```
hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA:/home/siddharth/Downloads/nosql/assn3/pig$ time pig -x mapreduce queries/query1.pig
2025-04-14 12:18:03,280 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-04-14 12:18:03,282 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-04-14 12:18:03,282 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-04-14 12:18:03,333 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-04-14 12:18:03,334 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/siddharth/Downloads/nosql/assn3/pig/pig_1744613283322.log
2025-04-14 12:18:03,638 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop1/.pigbootup not found
2025-04-14 12:18:03,716 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-04-14 12:18:03,716 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2025-04-14 12:18:04,083 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-query1.pig-7da5d40b-2b72-4535-9fda-135f46a8c276
2025-04-14 12:18:04,083 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2025-04-14 12:18:04,770 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,FILTER
2025-04-14 12:18:04,802 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2025-04-14 12:18:04,822 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSe
tter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, Split
Filter, StreamTypeCastInserter]}
2025-04-14 12:18:04,882 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageTh
reshold = 734003200
2025-04-14 12:18:04,927 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2025-04-14 12:18:04,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2025-04-14 12:18:04,955 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2025-04-14 12:18:04,968 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPack
ager)
2025-04-14 12:18:04,973 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2025-04-14 12:18:04,975 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged MR job 144 into MR job 141
2025-04-14 12:18:04,975 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged MR job 149 into MR job 141
2025-04-14 12:18:04,976 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Requested parallelism of splitter: -1
2025-04-14 12:18:04,976 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 2 map-reduce splittees.
2025-04-14 12:18:04,976 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 2 out of total 3 MR operators.
2025-04-14 12:18:04,976 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2025-04-14 12:18:05,090 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2025-04-14 12:18:05,247 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2025-04-14 12:18:05,248 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2025-04-14 12:18:05,270 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2025-04-14 12:18:05,275 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buff
er.percent
2025-04-14 12:18:05,275 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0
.3
2025-04-14 12:18:05,277 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2025-04-14 12:18:05,280 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2025-04-14 12:18:05,281 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.
mapReduceLayer.InputSizeReducerEstimator
2025-04-14 12:18:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=27
1065
2025-04-14 12:18:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2025-04-14 12:18:05,288 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2025-04-14 12:18:05,288 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
```

Figure 1: Query 1 execution



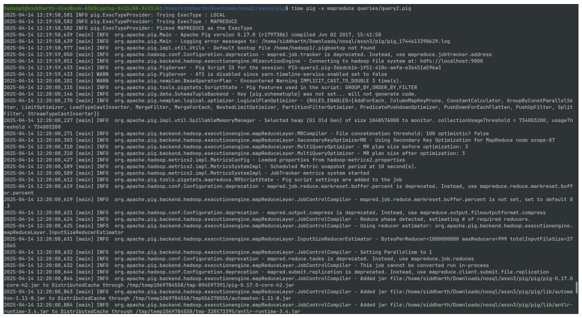
```
job_local626564388_0003


2025-04-14 12:18:22,812 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,815 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,817 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,823 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,825 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,826 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,833 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,836 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,838 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:18:22,841 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-04-14 12:18:22,845 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-14 12:18:22,851 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-14 12:18:22,851 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(CSE 511 / Algorithms,Master of Technology CSE,78.50000001402462,B)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781,B)
(CSE 513 / Software Systems,Master of Technology CSE,72.40408194470568,B)
(CSE 731 / Software Testing,Master of Science By Research,86.26666514078777,B+)
(COM 827 / Internet of Things,Master of Technology ECE,100.0,A)
(COM 827 / Internet of Things,Master of Science by Research-Part time,94.4000015258789,A-)
(CSE 857 / Secure Computation,Master of Science By Research,66.69999694824219,B)
(CSE 514 / Concrete Mathematics,Master of Technology CSE,73.0705888411578,B-)
(CSE 514 / Concrete Mathematics,Master of Science By Research,71.79999923706055,B)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789,A)
(VLS 502 / Analog CMOS VLSI Design,Master of Technology ECE,73.80344890726023,B+)
(VLS 502 / Analog CMOS VLSI Design,Master of Science By Research,72.69999694824219,C+)
(VLS 505 / System design with FPGA,Master of Technology ECE,69.19677463654548,B)
(VLS 505 / System design with FPGA,Master of Science By Research,59.25000127156576,C)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,88.88000094095865,B+)
(VLS 864 / Embedded Systems Design,Master of Science By Research,83.30000305175781,C+)
(AIM 841 / Medical Image Analysis with AI,Doctor of Philosophy,50.0,C+)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594,A)
(DAS 703 / Geographic Information Systems,Master of Science By Research,28.0,A-)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,63.20000076293945,B)
(CSE 816 / Software Production Engineering,Master of Science By Research,71.39999771118164,B+)
(AIM 512 / Mathematics for Machine Learning,Master of Technology CSE,62.43684212247232,B)
(AIM 512 / Mathematics for Machine Learning,Master of Technology ECE,100.0,B)
(AIM 843 / Spatio-temporal Data Analytics I,Master of Science By Research,70.80000305175781,A-)
2025-04-14 12:18:22,894 [main] INFO  org.apache.pig.Main - Pig script completed in 19 seconds and 653 milliseconds (19653 ms)

real    0m21.174s
user    0m17.953s
sys     0m1.135s
```

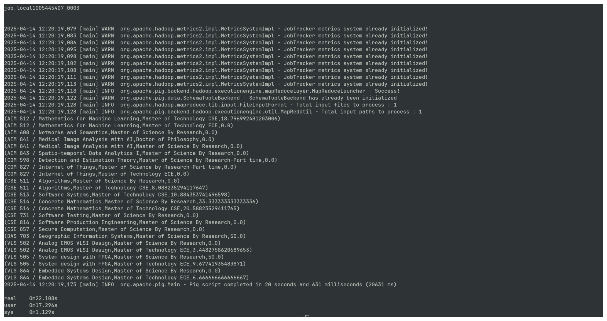
Figure 2: Time taken for query 1

2. Percentage of students with a below 50% attendance record, per program, per course.



```
hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA:/home/siddharth/Downloads/nosql/assn3/pig$ time pig -x mapreduce queries/query2.pig
2025-04-14 12:19:58,581 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-04-14 12:19:58,582 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-04-14 12:19:58,582 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-04-14 12:19:58,639 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-04-14 12:19:58,639 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/siddharth/Downloads/nosql/assn3/pig/pig_1744613398629.log
2025-04-14 12:19:58,977 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop1/.pigbootup not found
2025-04-14 12:19:59,050 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-04-14 12:19:59,051 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2025-04-14 12:19:59,433 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-query2.pig-36bdcb16-1f51-418c-aefa-e36451a596a3
2025-04-14 12:19:59,433 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2025-04-14 12:20:00,101 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 3 time(s).
2025-04-14 12:20:00,115 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER
2025-04-14 12:20:00,146 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2025-04-14 12:20:00,170 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSe
tter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, Split
Filter, StreamTypeCastInserter]}
2025-04-14 12:20:00,227 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageTh
reshold = 734003200
2025-04-14 12:20:00,271 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2025-04-14 12:20:00,303 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-87
2025-04-14 12:20:00,310 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2025-04-14 12:20:00,310 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2025-04-14 12:20:00,427 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2025-04-14 12:20:00,589 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2025-04-14 12:20:00,589 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2025-04-14 12:20:00,612 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2025-04-14 12:20:00,619 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buff
er.percent
2025-04-14 12:20:00,619 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0
.3
2025-04-14 12:20:00,621 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2025-04-14 12:20:00,624 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2025-04-14 12:20:00,625 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.
mapReduceLayer.InputSizeReducerEstimator
2025-04-14 12:20:00,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=27
1065
2025-04-14 12:20:00,632 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2025-04-14 12:20:00,632 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2025-04-14 12:20:00,632 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2025-04-14 12:20:00,644 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2025-04-14 12:20:00,846 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/pig-0.17.0
-core-h2.jar to DistributedCache through /tmp/temp1069784558/tmp-894597391/pig-0.17.0-core-h2.jar
2025-04-14 12:20:00,863 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/lib/automa
ton-1.11-8.jar to DistributedCache through /tmp/temp1069784558/tmp556370555/automaton-1.11-8.jar
2025-04-14 12:20:00,884 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/lib/antlr-
runtime-3.4.jar to DistributedCache through /tmp/temp1069784558/tmp-328573395/antlr-runtime-3.4.jar
```

Figure 3: Query 2 execution



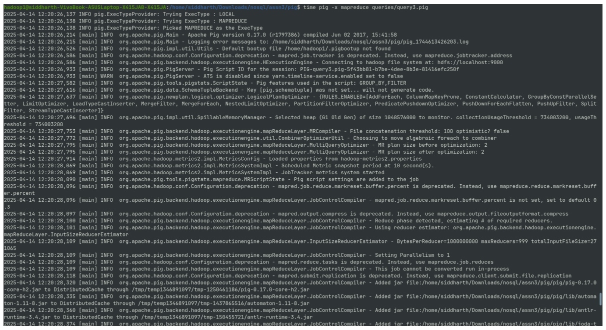
```
job_local1005445407_0003


2025-04-14 12:20:19,079 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,083 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,086 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,095 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,098 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,102 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,108 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,111 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,113 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2025-04-14 12:20:19,118 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2025-04-14 12:20:19,122 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2025-04-14 12:20:19,128 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2025-04-14 12:20:19,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(AIM 512 / Mathematics for Machine Learning,Master of Technology CSE,18.796992481203006)
(AIM 512 / Mathematics for Machine Learning,Master of Technology ECE,0.0)
(AIM 608 / Networks and Semantics,Master of Science By Research,0.0)
(AIM 841 / Medical Image Analysis with AI,Doctor of Philosophy,0.0)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,0.0)
(AIM 843 / Spatio-temporal Data Analytics I,Master of Science By Research,0.0)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,0.0)
(COM 827 / Internet of Things,Master of Science by Research-Part time,0.0)
(COM 827 / Internet of Things,Master of Technology ECE,0.0)
(CSE 511 / Algorithms,Master of Science By Research,0.0)
(CSE 511 / Algorithms,Master of Technology CSE,8.088235294117647)
(CSE 513 / Software Systems,Master of Technology CSE,10.884353741496598)
(CSE 514 / Concrete Mathematics,Master of Science By Research,33.333333333333336)
(CSE 514 / Concrete Mathematics,Master of Technology CSE,20.58823529411765)
(CSE 731 / Software Testing,Master of Science By Research,0.0)
(CSE 816 / Software Production Engineering,Master of Science By Research,0.0)
(CSE 857 / Secure Computation,Master of Science By Research,0.0)
(DAS 703 / Geographic Information Systems,Master of Science By Research,50.0)
(VLS 502 / Analog CMOS VLSI Design,Master of Science By Research,0.0)
(VLS 502 / Analog CMOS VLSI Design,Master of Technology ECE,3.4482758620689653)
(VLS 505 / System design with FPGA,Master of Science By Research,50.0)
(VLS 505 / System design with FPGA,Master of Technology ECE,9.67741935483871)
(VLS 864 / Embedded Systems Design,Master of Science By Research,0.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,6.666666666666667)
2025-04-14 12:20:19,173 [main] INFO  org.apache.pig.Main - Pig script completed in 20 seconds and 631 milliseconds (20631 ms)

real    0m22.108s
user    0m17.296s
sys     0m1.129s
```

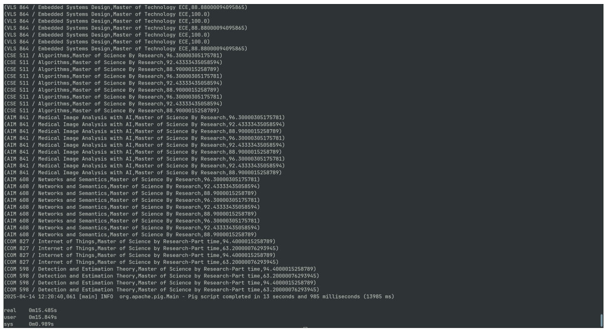
Figure 4: Time taken for query 2

3. Identification of top 3 courses by average attendance for each program.



```
hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA:/home/siddharth/Downloads/nosql/assn3/pig$ time pig -x mapreduce queries/query3.pig
2025-04-14 12:20:26,137 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-04-14 12:20:26,138 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-04-14 12:20:26,138 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-04-14 12:20:26,214 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-04-14 12:20:26,215 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/siddharth/Downloads/nosql/assn3/pig/pig_1744613426203.log
2025-04-14 12:20:26,526 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop1/.pigbootup not found
2025-04-14 12:20:26,586 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-04-14 12:20:26,586 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2025-04-14 12:20:26,933 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-query3.pig-5f43bb81-b7be-4dee-8b3e-81416efc250f
2025-04-14 12:20:26,933 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2025-04-14 12:20:27,582 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,FILTER
2025-04-14 12:20:27,616 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2025-04-14 12:20:27,637 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSe
tter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, Split
Filter, StreamTypeCastInserter]}
2025-04-14 12:20:27,696 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageTh
reshold = 734003200
2025-04-14 12:20:27,753 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2025-04-14 12:20:27,772 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2025-04-14 12:20:27,795 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2025-04-14 12:20:27,795 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 2
2025-04-14 12:20:27,914 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2025-04-14 12:20:28,069 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2025-04-14 12:20:28,069 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2025-04-14 12:20:28,090 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2025-04-14 12:20:28,096 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buff
er.percent
2025-04-14 12:20:28,096 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0
.3
2025-04-14 12:20:28,097 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2025-04-14 12:20:28,100 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2025-04-14 12:20:28,101 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.
mapReduceLayer.InputSizeReducerEstimator
2025-04-14 12:20:28,109 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=27
1065
2025-04-14 12:20:28,109 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2025-04-14 12:20:28,109 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2025-04-14 12:20:28,109 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2025-04-14 12:20:28,118 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
2025-04-14 12:20:28,320 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/pig-0.17.0
-core-h2.jar to DistributedCache through /tmp/temp1346891097/tmp-1250461186/pig-0.17.0-core-h2.jar
2025-04-14 12:20:28,335 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/lib/automa
ton-1.11-8.jar to DistributedCache through /tmp/temp1346891097/tmp-1437865516/automaton-1.11-8.jar
2025-04-14 12:20:28,360 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/lib/antlr-
runtime-3.4.jar to DistributedCache through /tmp/temp1346891097/tmp-150455721/antlr-runtime-3.4.jar
2025-04-14 12:20:28,374 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/siddharth/Downloads/nosql/assn3/pig/pig/lib/joda-t
```

Figure 5: Query 3 execution

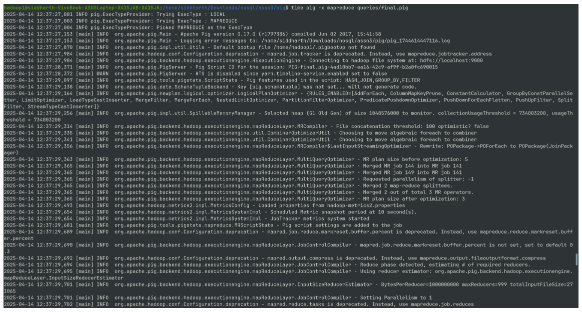


```
(VLS 864 / Embedded Systems Design,Master of Technology ECE,88.88000094095865)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,88.88000094095865)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,88.88000094095865)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781)
(CSE 511 / Algorithms,Master of Science By Research,92.43333435058594)
(CSE 511 / Algorithms,Master of Science By Research,88.9000015258789)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781)
(CSE 511 / Algorithms,Master of Science By Research,92.43333435058594)
(CSE 511 / Algorithms,Master of Science By Research,88.9000015258789)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781)
(CSE 511 / Algorithms,Master of Science By Research,92.43333435058594)
(CSE 511 / Algorithms,Master of Science By Research,88.9000015258789)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,96.30000305175781)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,88.9000015258789)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,96.30000305175781)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,88.9000015258789)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,96.30000305175781)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,88.9000015258789)
(AIM 608 / Networks and Semantics,Master of Science By Research,96.30000305175781)
(AIM 608 / Networks and Semantics,Master of Science By Research,92.43333435058594)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789)
(AIM 608 / Networks and Semantics,Master of Science By Research,96.30000305175781)
(AIM 608 / Networks and Semantics,Master of Science By Research,92.43333435058594)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789)
(AIM 608 / Networks and Semantics,Master of Science By Research,96.30000305175781)
(AIM 608 / Networks and Semantics,Master of Science By Research,92.43333435058594)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789)
(COM 827 / Internet of Things,Master of Science by Research-Part time,94.4000015258789)
(COM 827 / Internet of Things,Master of Science by Research-Part time,63.20000076293945)
(COM 827 / Internet of Things,Master of Science by Research-Part time,94.4000015258789)
(COM 827 / Internet of Things,Master of Science by Research-Part time,63.20000076293945)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,94.4000015258789)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,63.20000076293945)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,94.4000015258789)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,63.20000076293945)
2025-04-14 12:20:40,061 [main] INFO  org.apache.pig.Main - Pig script completed in 13 seconds and 985 milliseconds (13985 ms)

real    0m15.485s
user    0m15.849s
sys     0m0.989s
```

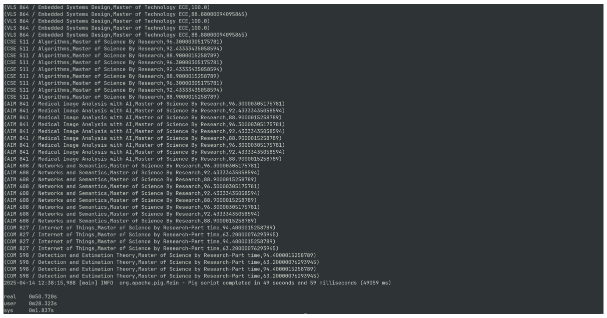
Figure 6: Time taken for query 3

Executing all the 3 queries together and then timing it:



```
hadoop1@siddharth-VivoBook-ASUSLaptop-X415JAB-X415JA:/home/siddharth/Downloads/nosql/assn3/pig$ time pig -x mapreduce queries/final.pig
2025-04-14 12:37:27,001 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2025-04-14 12:37:27,003 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2025-04-14 12:37:27,004 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-04-14 12:37:27,153 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-04-14 12:37:27,153 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/siddharth/Downloads/nosql/assn3/pig/pig_1744614447116.log
2025-04-14 12:37:27,870 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop1/.pigbootup not found
2025-04-14 12:37:27,984 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-04-14 12:37:27,984 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2025-04-14 12:37:28,371 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-final.pig-4ad10bb7-ea16-42c9-af9f-b2a0fc690015
2025-04-14 12:37:28,372 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2025-04-14 12:37:29,097 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN,GROUP_BY,FILTER
2025-04-14 12:37:29,138 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2025-04-14 12:37:29,164 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSe
tter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, Split
Filter, StreamTypeCastInserter]}
2025-04-14 12:37:29,256 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageTh
reshold = 734003200
2025-04-14 12:37:29,314 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2025-04-14 12:37:29,335 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2025-04-14 12:37:29,341 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
2025-04-14 12:37:29,356 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPack
ager)
2025-04-14 12:37:29,363 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 5
2025-04-14 12:37:29,365 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged MR job 144 into MR job 141
2025-04-14 12:37:29,365 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged MR job 149 into MR job 141
2025-04-14 12:37:29,365 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Requested parallelism of splitter: -1
2025-04-14 12:37:29,365 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 2 map-reduce splittees.
2025-04-14 12:37:29,365 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 2 out of total 3 MR operators.
2025-04-14 12:37:29,365 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2025-04-14 12:37:29,493 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties
2025-04-14 12:37:29,654 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2025-04-14 12:37:29,654 [main] INFO  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2025-04-14 12:37:29,681 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2025-04-14 12:37:29,689 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buff
er.percent
2025-04-14 12:37:29,690 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0
.3
2025-04-14 12:37:29,692 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2025-04-14 12:37:29,694 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2025-04-14 12:37:29,695 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.
mapReduceLayer.InputSizeReducerEstimator
2025-04-14 12:37:29,701 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=27
1065
2025-04-14 12:37:29,701 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2025-04-14 12:37:29,702 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
```

Figure 7: All queries together - execution



```
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,88.88000094095865)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,100.0)
(VLS 864 / Embedded Systems Design,Master of Technology ECE,88.88000094095865)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781)
(CSE 511 / Algorithms,Master of Science By Research,92.43333435058594)
(CSE 511 / Algorithms,Master of Science By Research,88.9000015258789)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781)
(CSE 511 / Algorithms,Master of Science By Research,92.43333435058594)
(CSE 511 / Algorithms,Master of Science By Research,88.9000015258789)
(CSE 511 / Algorithms,Master of Science By Research,96.30000305175781)
(CSE 511 / Algorithms,Master of Science By Research,92.43333435058594)
(CSE 511 / Algorithms,Master of Science By Research,88.9000015258789)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,96.30000305175781)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,88.9000015258789)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,96.30000305175781)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,88.9000015258789)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,96.30000305175781)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,92.43333435058594)
(AIM 841 / Medical Image Analysis with AI,Master of Science By Research,88.9000015258789)
(AIM 608 / Networks and Semantics,Master of Science By Research,96.30000305175781)
(AIM 608 / Networks and Semantics,Master of Science By Research,92.43333435058594)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789)
(AIM 608 / Networks and Semantics,Master of Science By Research,96.30000305175781)
(AIM 608 / Networks and Semantics,Master of Science By Research,92.43333435058594)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789)
(AIM 608 / Networks and Semantics,Master of Science By Research,96.30000305175781)
(AIM 608 / Networks and Semantics,Master of Science By Research,92.43333435058594)
(AIM 608 / Networks and Semantics,Master of Science By Research,88.9000015258789)
(COM 827 / Internet of Things,Master of Science by Research-Part time,94.4000015258789)
(COM 827 / Internet of Things,Master of Science by Research-Part time,63.20000076293945)
(COM 827 / Internet of Things,Master of Science by Research-Part time,94.4000015258789)
(COM 827 / Internet of Things,Master of Science by Research-Part time,63.20000076293945)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,94.4000015258789)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,63.20000076293945)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,94.4000015258789)
(COM 598 / Detection and Estimation Theory,Master of Science by Research-Part time,63.20000076293945)
2025-04-14 12:38:15,988 [main] INFO  org.apache.pig.Main - Pig script completed in 49 seconds and 59 milliseconds (49059 ms)

real    0m50.720s
user    0m28.323s
sys     0m1.837s
```

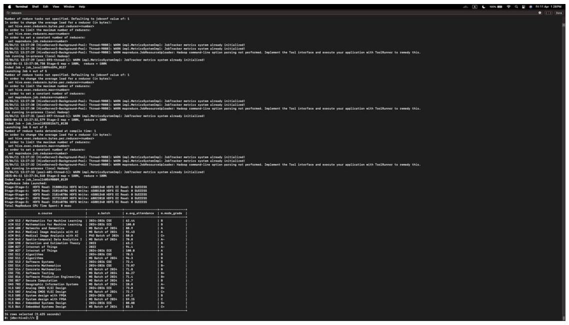
Figure 8: All queries together - time taken

Query	**Execution Time (seconds)**
Query 1	21.174
Query 2	22.108
Query 3	15.485
All Queries Executed Together	50.720

Table 5: Execution times of individual queries and combined execution

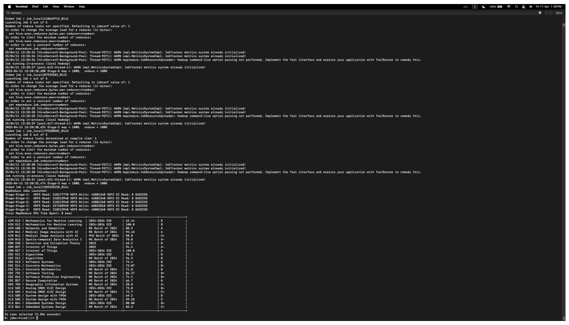
1 Comparing the performance of Hive(Partitioned and non-partitioned) and Pig

Query 1: Average attendance and mode grade per course per batch



```
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:27:28 [HiveServer2-Background-Pool: Thread-9080]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:27:28 [HiveServer2-Background-Pool: Thread-9080]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:27:29 [HiveServer2-Background-Pool: Thread-9080]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:27:29 [pool-593-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:27:30,738 Stage-5 map = 100%,  reduce = 100%
Ended Job = job_local180944594_0137
Launching Job 4 out of 5
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:27:30 [HiveServer2-Background-Pool: Thread-9080]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:27:30 [HiveServer2-Background-Pool: Thread-9080]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:27:30 [HiveServer2-Background-Pool: Thread-9080]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:27:31 [pool-597-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:27:32,579 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1833515671_0138
Launching Job 5 out of 5
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:27:32 [HiveServer2-Background-Pool: Thread-9080]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:27:32 [HiveServer2-Background-Pool: Thread-9080]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:27:32 [HiveServer2-Background-Pool: Thread-9080]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:27:33 [pool-601-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:27:34,540 Stage-3 map = 100%,  reduce = 100%
Ended Job = job_local1405490809_0139
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 218004316 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-4:  HDFS Read: 218140706 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-5:  HDFS Read: 218140706 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 327211059 HDFS Write: 68822010 HDFS EC Read: 0 SUCCESS
Stage-Stage-3:  HDFS Read: 218140706 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
+--------------------------------------------+-------------------+------------------+--------------+
|                  a.course                  |      a.batch      | a.avg_attendance | m.mode_grade |
+--------------------------------------------+-------------------+------------------+--------------+
| AIM 512 / Mathematics for Machine Learning | 2024-2026 CSE     | 62.44            | B            |
| AIM 512 / Mathematics for Machine Learning | 2024-2026 ECE     | 100.0            | B            |
| AIM 608 / Networks and Semantics           | MS Batch of 2024  | 88.9             | A            |
| AIM 841 / Medical Image Analysis with AI   | MS Batch of 2024  | 92.43            | A            |
| AIM 841 / Medical Image Analysis with AI   | PhD Batch of 2024 | 50.0             | C+           |
| AIM 843 / Spatio-temporal Data Analytics I | MS Batch of 2024  | 70.8             | A-           |
| COM 598 / Detection and Estimation Theory  | 2023              | 63.2             | B            |
| COM 827 / Internet of Things               | 2022              | 94.4             | A-           |
| COM 827 / Internet of Things               | 2024-2026 ECE     | 100.0            | A            |
| CSE 511 / Algorithms                       | 2024-2026 CSE     | 78.5             | B            |
| CSE 511 / Algorithms                       | MS Batch of 2024  | 96.3             | B            |
| CSE 513 / Software Systems                 | 2024-2026 CSE     | 72.4             | B            |
| CSE 514 / Concrete Mathematics             | 2024-2026 CSE     | 73.07            | B-           |
| CSE 514 / Concrete Mathematics             | MS Batch of 2024  | 71.8             | B            |
| CSE 731 / Software Testing                 | MS Batch of 2024  | 86.27            | B+           |
| CSE 816 / Software Production Engineering  | MS Batch of 2024  | 71.4             | B+           |
| CSE 857 / Secure Computation               | MS Batch of 2024  | 66.7             | B            |
| DAS 703 / Geographic Information Systems   | MS Batch of 2024  | 28.0             | A-           |
| VLS 502 / Analog CMOS VLSI Design          | 2024-2026 ECE     | 73.8             | B+           |
| VLS 502 / Analog CMOS VLSI Design          | MS Batch of 2024  | 72.7             | C+           |
| VLS 505 / System design with FPGA          | 2024-2026 ECE     | 69.2             | B            |
| VLS 505 / System design with FPGA          | MS Batch of 2024  | 59.25            | C            |
| VLS 864 / Embedded Systems Design          | 2024-2026 ECE     | 88.88            | B+           |
| VLS 864 / Embedded Systems Design          | MS Batch of 2024  | 83.3             | C+           |
+--------------------------------------------+-------------------+------------------+--------------+
24 rows selected (9.625 seconds)
0: jdbc:hive2://>
```

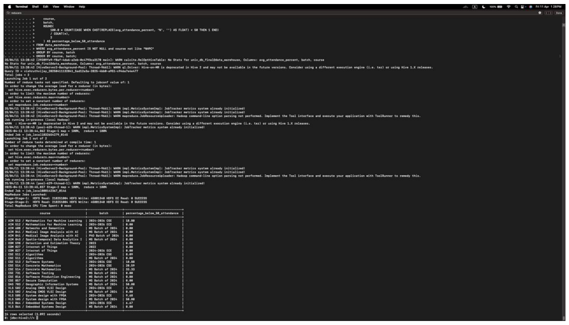
Figure 9: Query 1 - Non-Partitioned Execution



```
Ended Job = job_local1110669712_0141
Launching Job 3 out of 5
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:26 [HiveServer2-Background-Pool: Thread-9371]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:26 [HiveServer2-Background-Pool: Thread-9371]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:26 [HiveServer2-Background-Pool: Thread-9371]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:28:27 [pool-613-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:28:28,688 Stage-5 map = 100%,  reduce = 100%
Ended Job = job_local487925582_0142
Launching Job 4 out of 5
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:28 [HiveServer2-Background-Pool: Thread-9371]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:28 [HiveServer2-Background-Pool: Thread-9371]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:28 [HiveServer2-Background-Pool: Thread-9371]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:28:29 [pool-617-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:28:30,624 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1792588501_0143
Launching Job 5 out of 5
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:30 [HiveServer2-Background-Pool: Thread-9371]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:30 [HiveServer2-Background-Pool: Thread-9371]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:30 [HiveServer2-Background-Pool: Thread-9371]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:28:31 [pool-621-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:28:32,474 Stage-3 map = 100%,  reduce = 100%
Ended Job = job_local1282535228_0144
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 218177770 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-4:  HDFS Read: 218213940 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-5:  HDFS Read: 218213940 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 327320910 HDFS Write: 68822010 HDFS EC Read: 0 SUCCESS
Stage-Stage-3:  HDFS Read: 218213940 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
+--------------------------------------------+-------------------+------------------+--------------+
| AIM 512 / Mathematics for Machine Learning | 2024-2026 CSE     | 62.44            | B            |
| AIM 512 / Mathematics for Machine Learning | 2024-2026 ECE     | 100.0            | B            |
| AIM 608 / Networks and Semantics           | MS Batch of 2024  | 88.9             | A            |
| AIM 841 / Medical Image Analysis with AI   | MS Batch of 2024  | 92.43            | A            |
| AIM 841 / Medical Image Analysis with AI   | PhD Batch of 2024 | 50.0             | C+           |
| AIM 843 / Spatio-temporal Data Analytics I | MS Batch of 2024  | 70.8             | A-           |
| COM 598 / Detection and Estimation Theory  | 2023              | 63.2             | B            |
| COM 827 / Internet of Things               | 2022              | 94.4             | A-           |
| COM 827 / Internet of Things               | 2024-2026 ECE     | 100.0            | A            |
| CSE 511 / Algorithms                       | 2024-2026 CSE     | 78.5             | B            |
| CSE 511 / Algorithms                       | MS Batch of 2024  | 96.3             | B            |
| CSE 513 / Software Systems                 | 2024-2026 CSE     | 72.4             | B            |
| CSE 514 / Concrete Mathematics             | 2024-2026 CSE     | 73.07            | B-           |
| CSE 514 / Concrete Mathematics             | MS Batch of 2024  | 71.8             | B            |
| CSE 731 / Software Testing                 | MS Batch of 2024  | 86.27            | B+           |
| CSE 816 / Software Production Engineering  | MS Batch of 2024  | 71.4             | B+           |
| CSE 857 / Secure Computation               | MS Batch of 2024  | 66.7             | B            |
| DAS 703 / Geographic Information Systems   | MS Batch of 2024  | 28.0             | A-           |
| VLS 502 / Analog CMOS VLSI Design          | 2024-2026 ECE     | 73.8             | B+           |
| VLS 502 / Analog CMOS VLSI Design          | MS Batch of 2024  | 72.7             | C+           |
| VLS 505 / System design with FPGA          | 2024-2026 ECE     | 69.2             | B            |
| VLS 505 / System design with FPGA          | MS Batch of 2024  | 59.25            | C            |
| VLS 864 / Embedded Systems Design          | 2024-2026 ECE     | 88.88            | B+           |
| VLS 864 / Embedded Systems Design          | MS Batch of 2024  | 83.3             | C+           |
+--------------------------------------------+-------------------+------------------+--------------+
24 rows selected (9.356 seconds)
0: jdbc:hive2://>
```

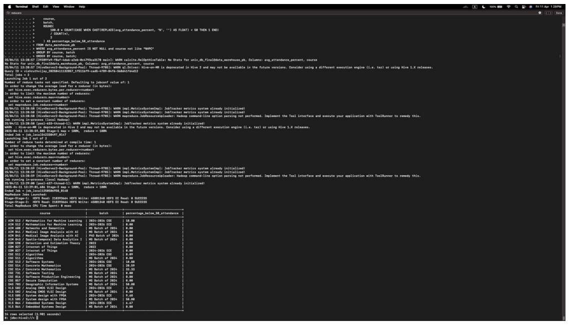
Figure 10: Query 1 - Partitioned Execution

Query 2: Percentage of students with below 50% attendance per course and program



```
. . . . . . . . . >     course,
. . . . . . . . . >     batch,
. . . . . . . . . >     ROUND(
. . . . . . . . . >         100.0 * COUNT(CASE WHEN CAST(REPLACE(avg_attendance_percent, '%', '') AS FLOAT) < 50 THEN 1 END)
. . . . . . . . . >         / COUNT(*),
. . . . . . . . . >         2
. . . . . . . . . >     ) AS percentage_below_50_attendance
. . . . . . . . . > FROM data_warehouse
. . . . . . . . . > WHERE avg_attendance_percent IS NOT NULL and course not like "%HPC"
. . . . . . . . . > GROUP BY course, batch
. . . . . . . . . > ORDER BY course, batch;
25/04/11 13:28:42 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN calcite.RelOptHiveTable: No Stats for univ_db_final@data_warehouse, Columns: avg_attendance_percent, batch, course
No Stats for univ_db_final@data_warehouse, Columns: avg_attendance_percent, batch, course
25/04/11 13:28:42 [HiveServer2-Background-Pool: Thread-9661]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
Query ID = vishruthvijay_20250411132841_5ad12a3a-2825-4bb0-a951-c946a7e4e477
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:42 [HiveServer2-Background-Pool: Thread-9661]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:42 [HiveServer2-Background-Pool: Thread-9661]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:42 [HiveServer2-Background-Pool: Thread-9661]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
25/04/11 13:28:43 [pool-625-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:28:44,062 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1832654279_0145
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:44 [HiveServer2-Background-Pool: Thread-9661]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:44 [HiveServer2-Background-Pool: Thread-9661]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:44 [HiveServer2-Background-Pool: Thread-9661]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:28:44 [pool-629-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:28:45,837 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local808142367_0146
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 218251004 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 218251004 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
+--------------------------------------------+-------------------+---------------------------------+
|                   course                   |       batch       | percentage_below_50_attendance  |
+--------------------------------------------+-------------------+---------------------------------+
| AIM 512 / Mathematics for Machine Learning | 2024-2026 CSE     | 18.80                           |
| AIM 512 / Mathematics for Machine Learning | 2024-2026 ECE     | 0.00                            |
| AIM 608 / Networks and Semantics           | MS Batch of 2024  | 0.00                            |
| AIM 841 / Medical Image Analysis with AI   | MS Batch of 2024  | 0.00                            |
| AIM 841 / Medical Image Analysis with AI   | PhD Batch of 2024 | 0.00                            |
| AIM 843 / Spatio-temporal Data Analytics I | MS Batch of 2024  | 0.00                            |
| COM 598 / Detection and Estimation Theory  | 2023              | 0.00                            |
| COM 827 / Internet of Things               | 2022              | 0.00                            |
| COM 827 / Internet of Things               | 2024-2026 ECE     | 0.00                            |
| CSE 511 / Algorithms                       | 2024-2026 CSE     | 8.09                            |
| CSE 511 / Algorithms                       | MS Batch of 2024  | 0.00                            |
| CSE 513 / Software Systems                 | 2024-2026 CSE     | 10.88                           |
| CSE 514 / Concrete Mathematics             | 2024-2026 CSE     | 20.59                           |
| CSE 514 / Concrete Mathematics             | MS Batch of 2024  | 33.33                           |
| CSE 731 / Software Testing                 | MS Batch of 2024  | 0.00                            |
| CSE 816 / Software Production Engineering  | MS Batch of 2024  | 0.00                            |
| CSE 857 / Secure Computation               | MS Batch of 2024  | 0.00                            |
| DAS 703 / Geographic Information Systems   | MS Batch of 2024  | 50.00                           |
| VLS 502 / Analog CMOS VLSI Design          | 2024-2026 ECE     | 3.45                            |
| VLS 502 / Analog CMOS VLSI Design          | MS Batch of 2024  | 0.00                            |
| VLS 505 / System design with FPGA          | 2024-2026 ECE     | 9.68                            |
| VLS 505 / System design with FPGA          | MS Batch of 2024  | 50.00                           |
| VLS 864 / Embedded Systems Design          | 2024-2026 ECE     | 6.67                            |
| VLS 864 / Embedded Systems Design          | MS Batch of 2024  | 0.00                            |
+--------------------------------------------+-------------------+---------------------------------+
24 rows selected (3.892 seconds)
0: jdbc:hive2://>
```

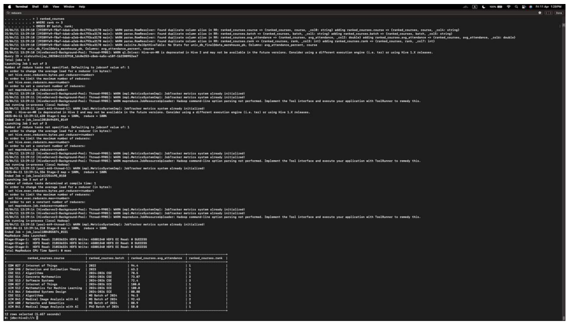
Figure 11: Query 2 - Non-Partitioned Execution



```
. . . . . . . . . >     course,
. . . . . . . . . >     batch,
. . . . . . . . . >     ROUND(
. . . . . . . . . >         100.0 * COUNT(CASE WHEN CAST(REPLACE(avg_attendance_percent, '%', '') AS FLOAT) < 50 THEN 1 END)
. . . . . . . . . >         / COUNT(*),
. . . . . . . . . >         2
. . . . . . . . . >     ) AS percentage_below_50_attendance
. . . . . . . . . > FROM data_warehouse_pb
. . . . . . . . . > WHERE avg_attendance_percent IS NOT NULL and course not like "%HPC"
. . . . . . . . . > GROUP BY course, batch
. . . . . . . . . > ORDER BY course, batch;
25/04/11 13:28:57 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN calcite.RelOptHiveTable: No Stats for univ_db_final@data_warehouse_pb, Columns: avg_attendance_percent, course
No Stats for univ_db_final@data_warehouse_pb, Columns: avg_attendance_percent, course
25/04/11 13:28:57 [HiveServer2-Background-Pool: Thread-9785]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
Query ID = vishruthvijay_20250411132857_1f511bf9-cad5-4f89-84f6-368d41f64d12
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:58 [HiveServer2-Background-Pool: Thread-9785]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:58 [HiveServer2-Background-Pool: Thread-9785]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:58 [HiveServer2-Background-Pool: Thread-9785]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:28:58 [pool-633-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
2025-04-11 13:28:59,805 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local542338497_0147
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:28:59 [HiveServer2-Background-Pool: Thread-9785]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:59 [HiveServer2-Background-Pool: Thread-9785]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:28:59 [HiveServer2-Background-Pool: Thread-9785]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:00 [pool-637-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:29:01,686 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1250506958_0148
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 218393664 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 218393664 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
+--------------------------------------------+-------------------+---------------------------------+
|                   course                   |       batch       | percentage_below_50_attendance  |
+--------------------------------------------+-------------------+---------------------------------+
| AIM 512 / Mathematics for Machine Learning | 2024-2026 CSE     | 18.80                           |
| AIM 512 / Mathematics for Machine Learning | 2024-2026 ECE     | 0.00                            |
| AIM 608 / Networks and Semantics           | MS Batch of 2024  | 0.00                            |
| AIM 841 / Medical Image Analysis with AI   | MS Batch of 2024  | 0.00                            |
| AIM 841 / Medical Image Analysis with AI   | PhD Batch of 2024 | 0.00                            |
| AIM 843 / Spatio-temporal Data Analytics I | MS Batch of 2024  | 0.00                            |
| COM 598 / Detection and Estimation Theory  | 2023              | 0.00                            |
| COM 827 / Internet of Things               | 2022              | 0.00                            |
| COM 827 / Internet of Things               | 2024-2026 ECE     | 0.00                            |
| CSE 511 / Algorithms                       | 2024-2026 CSE     | 8.09                            |
| CSE 511 / Algorithms                       | MS Batch of 2024  | 0.00                            |
| CSE 513 / Software Systems                 | 2024-2026 CSE     | 10.88                           |
| CSE 514 / Concrete Mathematics             | 2024-2026 CSE     | 20.59                           |
| CSE 514 / Concrete Mathematics             | MS Batch of 2024  | 33.33                           |
| CSE 731 / Software Testing                 | MS Batch of 2024  | 0.00                            |
| CSE 816 / Software Production Engineering  | MS Batch of 2024  | 0.00                            |
| CSE 857 / Secure Computation               | MS Batch of 2024  | 0.00                            |
| DAS 703 / Geographic Information Systems   | MS Batch of 2024  | 50.00                           |
| VLS 502 / Analog CMOS VLSI Design          | 2024-2026 ECE     | 3.45                            |
| VLS 502 / Analog CMOS VLSI Design          | MS Batch of 2024  | 0.00                            |
| VLS 505 / System design with FPGA          | 2024-2026 ECE     | 9.68                            |
| VLS 505 / System design with FPGA          | MS Batch of 2024  | 50.00                           |
| VLS 864 / Embedded Systems Design          | 2024-2026 ECE     | 6.67                            |
| VLS 864 / Embedded Systems Design          | MS Batch of 2024  | 0.00                            |
+--------------------------------------------+-------------------+---------------------------------+
24 rows selected (3.901 seconds)
0: jdbc:hive2://>
```

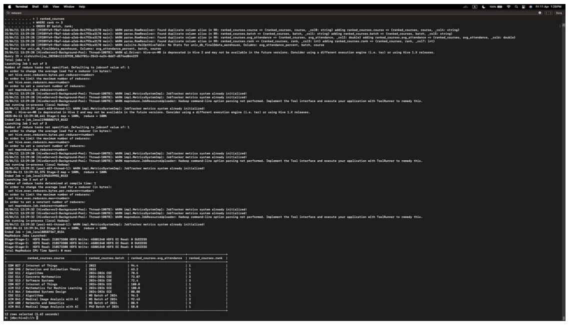
Figure 12: Query 2 - Partitioned Execution

Query 3: Top 3 courses by average attendance per program



```
. . . . . . . . . > ) ranked_courses
. . . . . . . . . > WHERE rank <= 3
. . . . . . . . . > ORDER BY batch, rank;
25/04/11 13:29:10 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.course => {ranked_courses, course, _col0: string} adding ranked_courses.course => {ranked_courses, course, _col4: string}
25/04/11 13:29:10 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.batch => {ranked_courses, batch, _col1: string} adding ranked_courses.batch => {ranked_courses, batch, _col5: string}
25/04/11 13:29:10 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.avg_attendance => {ranked_courses, avg_attendance, _col2: double} adding ranked_courses.avg_attendance => {ranked_courses, avg_attendance, _col6: double}
25/04/11 13:29:10 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.rank => {ranked_courses, rank, _col3: int} adding ranked_courses.rank => {ranked_courses, rank, _col7: int}
25/04/11 13:29:10 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN calcite.RelOptHiveTable: No Stats for univ_db_final@data_warehouse_pb, Columns: avg_attendance_percent, course
No Stats for univ_db_final@data_warehouse_pb, Columns: avg_attendance_percent, course
25/04/11 13:29:10 [HiveServer2-Background-Pool: Thread-9905]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
Query ID = vishruthvijay_20250411132910_1dc0e253-c8e6-4a5c-a10f-1b2200962ee7
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:29:10 [HiveServer2-Background-Pool: Thread-9905]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:11 [HiveServer2-Background-Pool: Thread-9905]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:11 [HiveServer2-Background-Pool: Thread-9905]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:11 [pool-641-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
2025-04-11 13:29:12,620 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local2018494591_0149
Launching Job 2 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:29:12 [HiveServer2-Background-Pool: Thread-9905]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:12 [HiveServer2-Background-Pool: Thread-9905]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:12 [HiveServer2-Background-Pool: Thread-9905]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:13 [pool-645-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:29:14,336 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local612254495_0150
Launching Job 3 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:29:14 [HiveServer2-Background-Pool: Thread-9905]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:14 [HiveServer2-Background-Pool: Thread-9905]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:14 [HiveServer2-Background-Pool: Thread-9905]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:15 [pool-649-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:29:16,210 Stage-3 map = 100%,  reduce = 100%
Ended Job = job_local1804855874_0151
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 218536324 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 218536324 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-3:  HDFS Read: 218536324 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
+----------------------------------------------+------------------------+----------------------------------+----------------------+
|            ranked_courses.course             | ranked_courses.batch   | ranked_courses.avg_attendance    | ranked_courses.rank  |
+----------------------------------------------+------------------------+----------------------------------+----------------------+
| COM 827 / Internet of Things                 | 2022                   | 94.4                             | 1                    |
| COM 598 / Detection and Estimation Theory    | 2023                   | 63.2                             | 1                    |
| CSE 511 / Algorithms                         | 2024-2026 CSE          | 78.5                             | 1                    |
| CSE 514 / Concrete Mathematics               | 2024-2026 CSE          | 73.07                            | 2                    |
| CSE 513 / Software Systems                   | 2024-2026 CSE          | 72.4                             | 3                    |
| COM 827 / Internet of Things                 | 2024-2026 ECE          | 100.0                            | 1                    |
| AIM 512 / Mathematics for Machine Learning   | 2024-2026 ECE          | 100.0                            | 2                    |
| VLS 864 / Embedded Systems Design            | 2024-2026 ECE          | 88.88                            | 3                    |
| CSE 511 / Algorithms                         | MS Batch of 2024       | 96.3                             | 1                    |
| AIM 841 / Medical Image Analysis with AI     | MS Batch of 2024       | 92.43                            | 2                    |
| AIM 608 / Networks and Semantics             | MS Batch of 2024       | 88.9                             | 3                    |
| AIM 841 / Medical Image Analysis with AI     | PhD Batch of 2024      | 50.0                             | 1                    |
+----------------------------------------------+------------------------+----------------------------------+----------------------+
12 rows selected (5.657 seconds)
0: jdbc:hive2://>
```

Figure 13: Query 3 - Non-Partitioned Execution



```
. . . . . . . . . > ) ranked_courses
. . . . . . . . . > WHERE rank <= 3
. . . . . . . . . > ORDER BY batch, rank;
25/04/11 13:29:28 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.course => {ranked_courses, course, _col0: string} adding ranked_courses.course => {ranked_courses, course, _col4: string}
25/04/11 13:29:28 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.batch => {ranked_courses, batch, _col1: string} adding ranked_courses.batch => {ranked_courses, batch, _col5: string}
25/04/11 13:29:28 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.avg_attendance => {ranked_courses, avg_attendance, _col2: double} adding ranked_courses.avg_attendance => {ranked_courses, avg_attendance, _col6: double}
25/04/11 13:29:28 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN parse.RowResolver: Found duplicate column alias in RR: ranked_courses.rank => {ranked_courses, rank, _col3: int} adding ranked_courses.rank => {ranked_courses, rank, _col7: int}
25/04/11 13:29:28 [29289fe9-f8af-4da6-a2eb-0c4793ca3170 main]: WARN calcite.RelOptHiveTable: No Stats for univ_db_final@data_warehouse, Columns: avg_attendance_percent, batch, course
No Stats for univ_db_final@data_warehouse, Columns: avg_attendance_percent, batch, course
25/04/11 13:29:28 [HiveServer2-Background-Pool: Thread-10078]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
Query ID = vishruthvijay_20250411132928_58b2f01c-2543-4e24-8dd7-d574ed044229
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:29:28 [HiveServer2-Background-Pool: Thread-10078]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:28 [HiveServer2-Background-Pool: Thread-10078]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:28 [HiveServer2-Background-Pool: Thread-10078]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:29 [pool-653-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. tez) or using Hive 1.X releases.
2025-04-11 13:29:30,641 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1988885719_0152
Launching Job 2 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:29:30 [HiveServer2-Background-Pool: Thread-10078]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:30 [HiveServer2-Background-Pool: Thread-10078]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:30 [HiveServer2-Background-Pool: Thread-10078]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:31 [pool-657-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:29:32,375 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1396549955_0153
Launching Job 3 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/04/11 13:29:32 [HiveServer2-Background-Pool: Thread-10078]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:32 [HiveServer2-Background-Pool: Thread-10078]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/04/11 13:29:32 [HiveServer2-Background-Pool: Thread-10078]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/04/11 13:29:33 [pool-661-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-04-11 13:29:34,242 Stage-3 map = 100%,  reduce = 100%
Ended Job = job_local80587367_0154
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 218573388 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 218573388 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Stage-Stage-3:  HDFS Read: 218573388 HDFS Write: 45881340 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
+----------------------------------------------+------------------------+----------------------------------+----------------------+
|            ranked_courses.course             | ranked_courses.batch   | ranked_courses.avg_attendance    | ranked_courses.rank  |
+----------------------------------------------+------------------------+----------------------------------+----------------------+
| COM 827 / Internet of Things                 | 2022                   | 94.4                             | 1                    |
| COM 598 / Detection and Estimation Theory    | 2023                   | 63.2                             | 1                    |
| CSE 511 / Algorithms                         | 2024-2026 CSE          | 78.5                             | 1                    |
| CSE 514 / Concrete Mathematics               | 2024-2026 CSE          | 73.07                            | 2                    |
| CSE 513 / Software Systems                   | 2024-2026 CSE          | 72.4                             | 3                    |
| COM 827 / Internet of Things                 | 2024-2026 ECE          | 100.0                            | 1                    |
| AIM 512 / Mathematics for Machine Learning   | 2024-2026 ECE          | 100.0                            | 2                    |
| VLS 864 / Embedded Systems Design            | 2024-2026 ECE          | 88.88                            | 3                    |
| CSE 511 / Algorithms                         | MS Batch of 2024       | 96.3                             | 1                    |
| AIM 841 / Medical Image Analysis with AI     | MS Batch of 2024       | 92.43                            | 2                    |
| AIM 608 / Networks and Semantics             | MS Batch of 2024       | 88.9                             | 3                    |
| AIM 841 / Medical Image Analysis with AI     | PhD Batch of 2024      | 50.0                             | 1                    |
+----------------------------------------------+------------------------+----------------------------------+----------------------+
12 rows selected (5.62 seconds)
0: jdbc:hive2://>
```

Figure 14: Query 3 - Partitioned Execution

Execution Time Comparison Table

Query	Partitioned + Bucketed Time (s)	Non-Partitioned Time (s)	Pig (s)
Query 1	9.356	9.62	21.174
Query 2	3.892	3.901	22.108
Query 3	5.62	5.657	15.485

Table 6: Execution Time Comparison: Partitioned vs Non-Partitioned Data Warehouse