Name of Student : Pushkar Sane			
Roll Number : 45		**LAB Practical Number:** 5	
Title of LAB Practical : To create a Pig Data Model, Read and Store Data and Perform following Pig Operations, 1. Pig Latin Basic 2. Pig Data Types, 3. Download the data 4. Create your Script 5. Save and Execute the Script 6. Pig Operations : Diagnostic Operators, Grouping and Joining, Combining & Splitting, Filtering, Sorting.			
DOP :03/09/2024		**DOS :** 06/09/2024	
CO Mapped : CO5	**PO Mapped:** PO1, PO2, PO3, PO4, PO5, PO7, PSO1, PSO2	**Faculty Signature:**	**Marks :**

Aim:

To create a Pig Data Model, Read and Store Data and Perform following Pig Operations,

1. Pig Latin Basic
2. Pig Data Types,
3. Download the data
4. Create your Script
5. Save and Execute the Script
6. Pig Operations : Diagnostic Operators, Grouping and Joining, Combining & Splitting, Filtering, Sorting.

Theory

1. Pig Latin Basics

Pig Latin is the language used to analyze large data sets in Apache Pig. It provides a way to define a sequence of data transformations using various operators like LOAD, FILTER, GROUP, and STORE. It abstracts away the complexity of writing MapReduce jobs directly, simplifying the process of working with big data.

- **High-level Language**: Pig Latin allows users to write concise, human-readable scripts that define data flows.
- **Data Flow**: The Pig engine converts these scripts into a series of MapReduce jobs, which run on a Hadoop cluster, making it easy to process terabytes of data.
- **Data Structure Support**: Pig handles structured and unstructured data in formats like JSON, text, and binary, making it versatile for handling various data types.
- **Operations**: The language allows the user to perform common operations like joining datasets, filtering data, and performing aggregations.

---

2. Pig Data Types

Pig supports a wide range of data types, which are crucial for working with different datasets. These types fall into two categories: **simple** and **complex**.

Simple Data Types:

- **int**: 32-bit signed integer (e.g., 10, 200).
- **long**: 64-bit signed integer (e.g., 100000L).
- **float**: 32-bit floating-point number (e.g., 3.14f).
- **double**: 64-bit floating-point number (e.g., 3.14145).
- **chararray**: A sequence of characters, representing strings (e.g., "hello").
- **bytearray**: A generic type for any unknown data format.

Complex Data Types:

- **Tuple**: An ordered set of fields. It is similar to a row in a table (e.g., (1, 'John', 25)).
- **Bag**: A collection of tuples. Bags are unordered and allow duplicate tuples (e.g., {(1, 'John', 25), (2, 'Jane', 30)}).
- **Map**: A key-value pair structure. The key must be a chararray, but the value can be of any data type (e.g., ['name' -> 'John', 'age' -> 25]).

These data types make Pig powerful for manipulating datasets, allowing complex data modeling and transformation.

---

3. Download the Data

In order to execute a Pig script, you must have the relevant data. For this assignment, you can either download a sample dataset or create your own. Here's a small example dataset representing interns' information:

1,Ashok,22,Mumbai
2,Sudha,23,Delhi
3,Maya,23,Tokyo
4,Sara,25,NewYork
5,Suraj,23,Mumbai
6,Sandy,22,Chennai

This dataset contains fields like id, name, age, and city, which you will use to perform operations such as filtering, grouping, and joining in your Pig script.

You can save this data in a file, say interns_data.txt, on your local file system or HDFS.

---

4. Create your Script

Once you have the data, you can create a Pig Latin script to process it. Here's an example script based on the dataset mentioned:

```
-- Load the data
interns_data = LOAD '/path_to_data/interns_data.txt' USING PigStorage(',')
    AS (id: int, name: chararray, age: int, city: chararray);

-- Filter data to get interns aged 23
interns_age_23 = FILTER interns_data BY age == 23;

-- Group data by city
grouped_by_city = GROUP interns_data BY city;

-- Display results
DUMP interns_age_23;
DUMP grouped_by_city;
```

This script performs three major steps:

1. **Loading the data**: The data is loaded from the file using PigStorage, which is a built-in function to handle CSV-like files.
2. **Filtering**: The FILTER operation is applied to only retrieve interns aged 23.
3. **Grouping**: The GROUP operation is used to group interns by city.

---

5. Save and Execute the Script

After writing your script, save it as a .pig file (e.g., interns.pig). You can then execute it on your system.

Local Mode:

If you are running Pig in local mode (without a Hadoop cluster), you can execute the script using the following command:

pig -x local interns.pig

MapReduce Mode:

If you have access to a Hadoop cluster, you can run the script in MapReduce mode:

pig interns.pig

Pig will automatically convert your script into a series of MapReduce jobs and execute them on the Hadoop cluster.

---

6. Pig Operations

Pig provides several useful operations for processing and analyzing data. These include:

Diagnostic Operators:

- **DUMP**: Displays the results of a relation (data) on the console. This is useful for debugging purposes.
    - Example: DUMP interns_age_23;
- **DESCRIBE**: Provides the schema of a relation, showing the field names and their data types.
    - Example: DESCRIBE interns_data;
- **EXPLAIN**: Gives an execution plan for a Pig script or relation. This helps in understanding how the script will be processed.
    - Example: EXPLAIN interns_data;

- **ILLUSTRATE**: Runs the script on a small sample of data to illustrate the transformations that will occur.
    - Example: ILLUSTRATE interns_data;

Grouping and Joining:

- **GROUP**: Groups data by a specified field, such as city, so that operations can be applied on each group.
    - Example: grouped_by_city = GROUP interns_data BY city;
- **JOIN**: Combines two or more datasets based on a common field, similar to SQL joins.
    - Example: joined_data = JOIN interns_data BY id, other_data BY id;

Combining & Splitting:

- **UNION**: Combines two or more datasets into one.
    - Example: combined_data = UNION data1, data2;
- **SPLIT**: Divides a dataset into two or more parts based on conditions.
    - Example: SPLIT interns_data INTO young_interns IF age < 23, older_interns IF age >= 23;

Filtering:

- **FILTER**: Extracts records that meet a specific condition.
    - Example: filtered_data = FILTER interns_data BY age == 23;

Sorting:

- **ORDER**: Sorts the data based on one or more fields.
    - Example: sorted_data = ORDER interns_data BY age ASC;

Commands-

1. Running Pig in Local Mode

```
pig -x local
```

This command runs Pig in **local mode**, meaning it will not use Hadoop and will work with your local filesystem.

2. Loading and Dumping Data

Load Data from data_model:

```
DataModels = LOAD 'data_model'
   USING PigStorage(';')
   AS (name: chararray,
      address: tuple(city: chararray, pincode: chararray),
      result: bag{info: tuple(sub: chararray, marks: int)},
      m: map[int]);
```

Dump the Loaded Data:

```
DUMP DataModels;
```

Load Another Dataset (Student Information):

```
student_info = LOAD '/home/cloudera/Desktop/sample_student_data.txt'
   USING PigStorage(',')
      AS (id: chararray, fname: chararray, lname: chararray, age: int, phone: chararray, city: chararray);
```

Dump the Student Data:

```
DUMP student_info;
```

3. Store Data

Storing the student data using a custom delimiter (|):

```
STORE student_info INTO 'sampleoutput' USING PigStorage('|');
```

4. Operators

Describe the Schema:

```
DESCRIBE student_info;
```

5. Filtering Data

Filter Students by Age:

```
filterstudent = FILTER student_info BY age > 22;
DUMP filterstudent;
```

Explain the Filter Operation:

```
EXPLAIN filterstudent;
```

Select Specific Fields:

```
foreachstudent = FOREACH filterstudent GENERATE id, fname, age, city;
DUMP foreachstudent;
```

6. Group Data

Group by City:

```
groupstudent = GROUP student_info BY city;
DUMP groupstudent;
```

Illustrate the Foreach Operation:
pig

```
ILLUSTRATE foreachstudent;
```

Group by Multiple Columns (City, Age):

```
groupstudent2 = GROUP student_info BY (city, age);
DUMP groupstudent2;
```

Group All:

```
groupstudent3 = GROUP student_info ALL;
DUMP groupstudent3;
```

7. COGROUP Operation

Load Intern Data:

```
I = LOAD '/home/cloudera/Desktop/interns_data.txt'
   USING PigStorage(',')
   AS (id: chararray, fname: chararray, age: int, city: chararray);
DUMP I;
```

COGROUP by Age:

```
CG = COGROUP student_info BY age, I BY age;
DUMP CG;
```

8. Self Join

Load the Same Data Again:

```
student_infoB = LOAD '/home/cloudera/Desktop/sample_student_data.txt'
   USING PigStorage(',')
      AS (id: chararray, fname: chararray, lname: chararray, age: int, phone: chararray, city: chararray);
```

Perform a Self Join:

pig

```
SelfJoin = JOIN student_info BY age, student_infoB BY age;
DUMP SelfJoin;
```

9. Join Operations

Inner Join by City:

```
InnerJoin = JOIN student_info BY city, I BY city;
DUMP InnerJoin;
```

Left Join by City:

```
LeftJoin = JOIN student_info BY city LEFT, I BY city;
DUMP LeftJoin;
```

Right Join by City:

```
RightJoin = JOIN student_info BY city RIGHT OUTER, I BY city;
DUMP RightJoin;
```

Full Outer Join by City:

```
FullJoin = JOIN student_info BY city FULL OUTER, I BY city;
DUMP FullJoin;
```

10. Sorting Data

Sort in Descending Order by ID:

```
B5 = ORDER student_info BY id DESC;
DUMP B5;
```

Sort in Ascending Order by ID:

```
B = ORDER student_info BY id ASC;
DUMP B;
```

11. Limit

Limit the Results to 3 Records:

pig

```
C = LIMIT B 3;
```

```
DUMP C;
```

12. Union

Perform Union on Limited Data:

```
c = LIMIT B 3;
c1 = LIMIT B 5;
UnionData = UNION c, c1;
DUMP UnionData;
```

13. Split Data

Split Data into Two Sets:

```
SPLIT student_info INTO younger_students IF age < 23, older_students IF age >= 23;
DUMP younger_students;
DUMP older_students;
```

14. Filter by City

Filter Students from Mumbai:

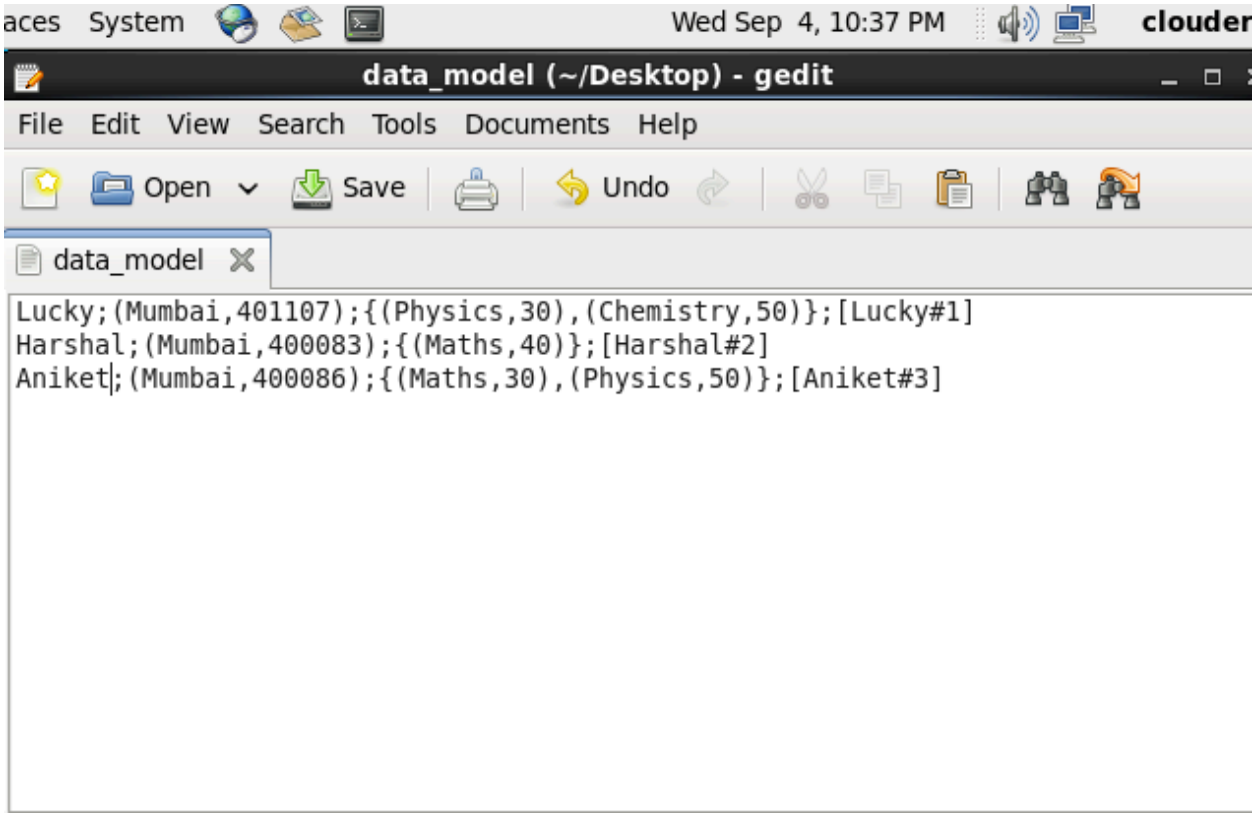
```
filter_city = FILTER student_info BY city == 'Mumbai';
DUMP filter_city;
```

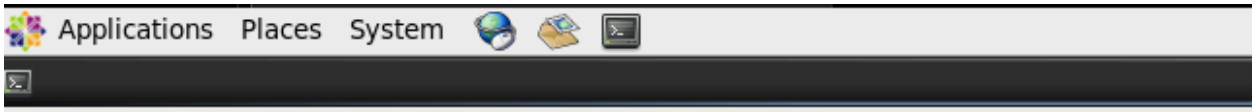
15. Distinct

Find Distinct Cities:

```
all_city = FOREACH student_info GENERATE city;
distinct_cities = DISTINCT all_city;

        DUMP distinct_cities;
```

Screenshots:






```
cloudera@quickstart Desktop]$ pig -x local
og4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)

og4j:WARN Please initialize the log4j system properly.
og4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
o.
024-09-04 22:38:55,581 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
2.0-cdh5.12.0 (rexported) compiled Jun 29 2017, 04:34:31
024-09-04 22:38:55,582 [main] INFO  org.apache.pig.Main - Logging error message
 to: /home/cloudera/Desktop/pig_1725514735504.log
024-09-04 22:38:55,651 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
tup file /home/cloudera/.pigbootup not found
024-09-04 22:38:56,524 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - fs.default.name is deprecated. Instead, use fs.defaultFS
024-09-04 22:38:56,528 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ss
024-09-04 22:38:56,547 [main] INFO  org.apache.pig.backend.hadoop.executionengi
e.HExecutionEngine - Connecting to hadoop file system at: file:///
024-09-04 22:38:58,431 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
024-09-04 22:38:58,932 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - fs.default.name is deprecated. Instead, use fs.defaultFS
024-09-04 22:38:58,941 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ss
024-09-04 22:38:58,947 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
024-09-04 22:38:59,316 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - fs.default.name is deprecated. Instead, use fs.defaultFS
024-09-04 22:38:59,333 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ss
024-09-04 22:38:59,342 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
024-09-04 22:38:59,816 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - fs.default.name is deprecated. Instead, use fs.defaultFS
024-09-04 22:38:59,826 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ss
024-09-04 22:38:59,841 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
024-09-04 22:39:00,184 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - fs.default.name is deprecated. Instead, use fs.defaultFS
024-09-04 22:39:00,201 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
tion - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ss
```

```
grunt> DataModels = LOAD 'data_model'
    USING PigStorage(';')
    AS (name:chararray,
        address:tuple(city:chararray, pincode:chararray),
        result:bag{info:tuple(sub:chararray, marks:int)},
        m:map[int]);
```

```
grunt> DUMP DataModels;
2024-09-04 23:00:23,207 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig feature
s used in the script: UNKNOWN
2024-09-04 23:00:23,429 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOpt
imizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, Group
ByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFil
ter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitF
ilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFi
lterOptimizer]}
2024-09-04 23:00:23,888 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-04 23:00:24,010 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-04 23:00:24,011 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-04 23:00:24,212 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - sess
ion.id is deprecated. Instead, use dfs.metrics.session-id
2024-09-04 23:00:24,215 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Initializing
 JVM Metrics with processName=JobTracker, sessionId=
2024-09-04 23:00:24,335 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script
settings are added to the job
2024-09-04 23:00:24,598 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markrese
t.buffer.percent
2024-09-04 23:00:24,598 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to def
ault 0.3
2024-09-04 23:00:24,598 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2024-09-04 23:00:24,737 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.JobControlCompiler - Setting up single store job
2024-09-04 23:00:24,839 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.sche
matuple] is false, will not generate code.
2024-09-04 23:00:24,839 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting proc
ess to move generated code to distributed cache
2024-09-04 23:00:24,839 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed c
```

```
Successfully read records from: "file:///home/cloudera/Desktop/data_model"

Output(s):
Successfully stored records in: "file:/tmp/temp-770510020/tmp-882682946"

Job DAG:
job_local740547110_0001


2024-09-04 23:00:47,933 [main] INFO  org.apache.pig.backend.hadoop.executionengine.map
2024-09-04 23:00:47,949 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
2024-09-04 23:00:47,949 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
2024-09-04 23:00:47,951 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
2024-09-04 23:00:47,952 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTu
2024-09-04 23:00:48,031 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFo
2024-09-04 23:00:48,031 [main] INFO  org.apache.pig.backend.hadoop.executionengine.uti
(Lucky,(Mumbai,401107),{(Physics,30),(Chemistry,50)},[Lucky#1])
(Harshal,(Mumbai,400083),{(Maths,40)},[Harshal#2])
(Aniket,(Mumbai,400086),{(Maths,30),(Physics,50)},[Aniket#3])
```

001,Lucky,Patil,21,9848022337,Mumbai

002,Harshal,Battacharya,22,9848022338,Kolkata

003,Aniket,Khanna,22,9848022339,Mumbai

004,Atharva,Agarwal,21,9848022330,Pune

005,Raj,Jain,23,9848022336,Mumbai

006,Shubham,Mishra,23,9848022335,Chennai

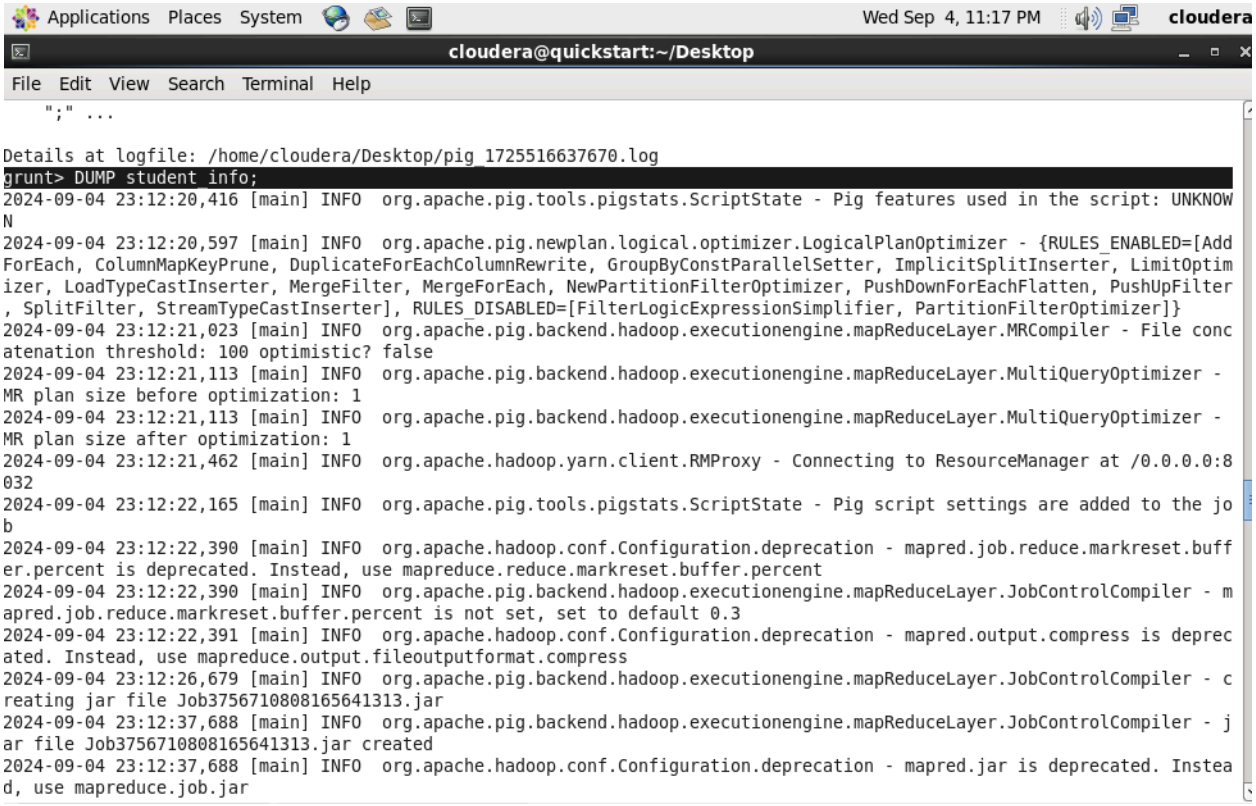
007,Kaushal,Nayak,24,9848022334,Kolkata

008,Shreemane,Nambiar,24,9848022333,Chennai

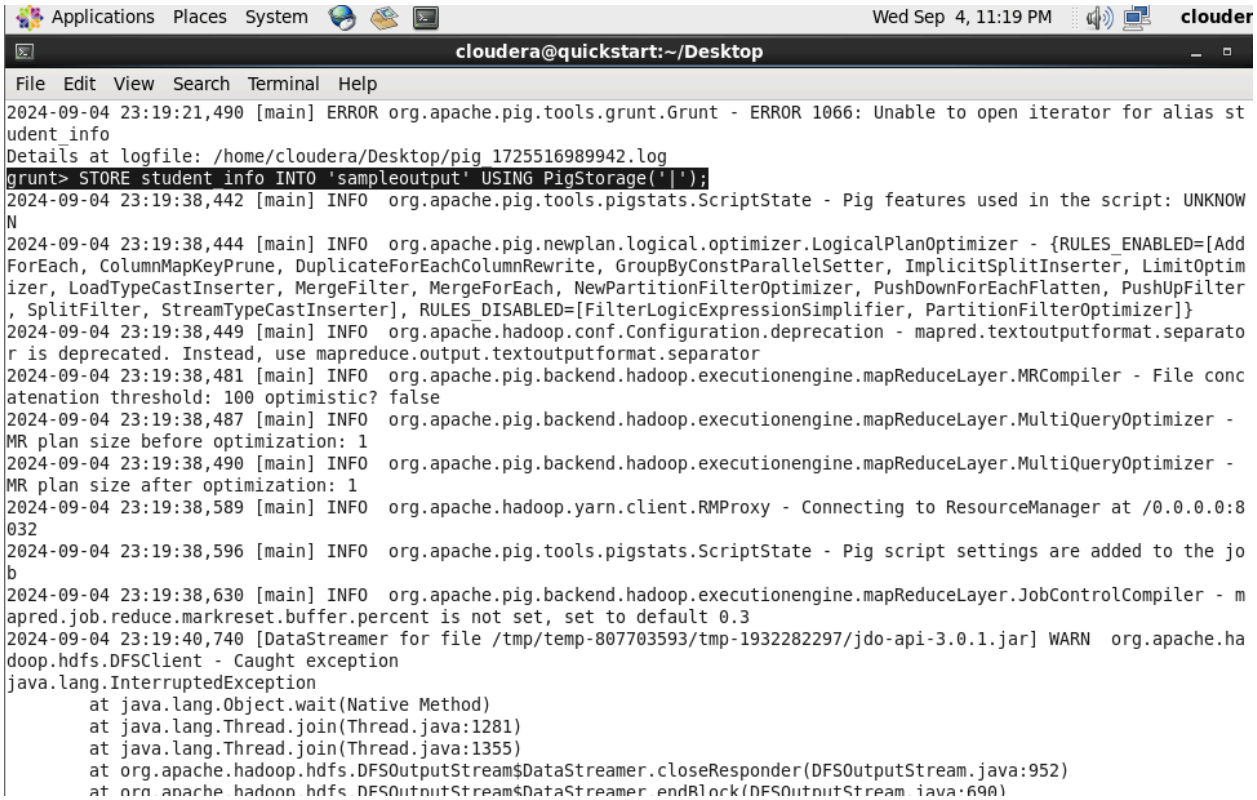
*sample_student_data

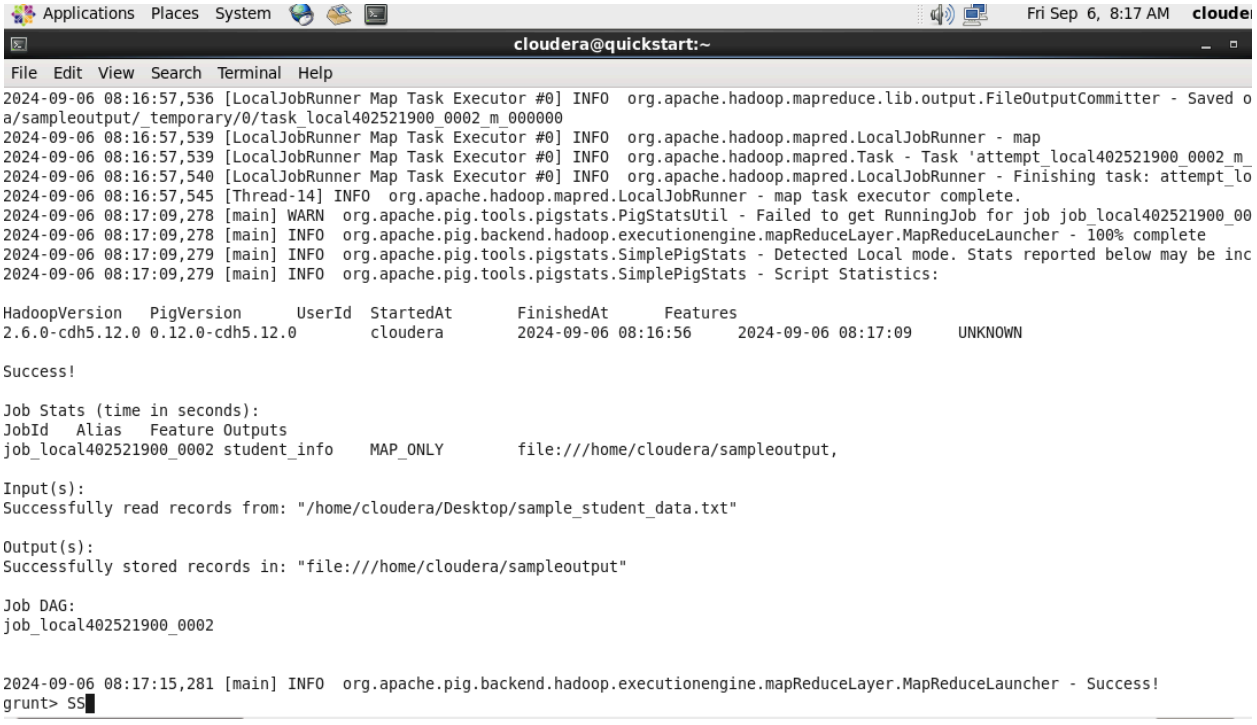
```
001,Lucky,Patil,21,9848022337,Mumbai
002,Harshal,Battacharya,22,9848022338,Kolkata
003,Aniket,Khanna,22,9848022339,Mumbai
004,Atharva,Agarwal,21,9848022330,Pune
005,Raj,Jain,23,9848022336,Mumbai
006,Shubham,Mishra,23,9848022335,Chennai
007,Kaushal,Nayak,24,9848022334,Kolkata
008,Shreemane,Nambiar,24,9848022333,Chennai
```

```
grunt> student_info = LOAD '/home/cloudera/Desktop/sample_student_data.txt'
    USING PigStorage(',')
    AS (id: chararray, fname: chararray, lname: chararray, age: int, phone: chararray, city: chararray);
```




```
2024-09-06 06:00:01,220 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - map
2024-09-06 06:00:01,220 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task 'attempt_local1640725062_0001_m_000000_0' done.
2024-09-06 06:00:01,220 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1640725062_0001_m_000000_0
2024-09-06 06:00:01,221 [Thread-8] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2024-09-06 06:00:06,414 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-09-06 06:00:06,414 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2024-09-06 06:00:12,416 [main] WARN  org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local1640725062_0001
2024-09-06 06:00:12,421 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-09-06 06:00:12,422 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2024-09-06 06:00:12,425 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2024-09-06 05:59:58     2024-09-06 06:00:12     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local1640725062_0001        student_info    MAP_ONLY        file:/tmp/temp142338119/tmp696824699,

Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp142338119/tmp696824699"

Job DAG:
job_local1640725062_0001


2024-09-06 06:00:18,427 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 06:00:18,431 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 06:00:18,431 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 06:00:18,431 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 06:00:18,431 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 06:00:18,459 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 06:00:18,459 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(001,Lucky,Patil,21,9848022337,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(003,Aniket,Khanna,22,9848022339,Mumbai)
(004,Atharva,Agarwal,21,9848022330,Pune)
(005,Raj,Jain,23,9848022336,Mumbai)
(006,Shubham,Mishra,23,9848022335,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata)
(008,Shreemane,Nambiar,24,9848022333,Chennai)
```

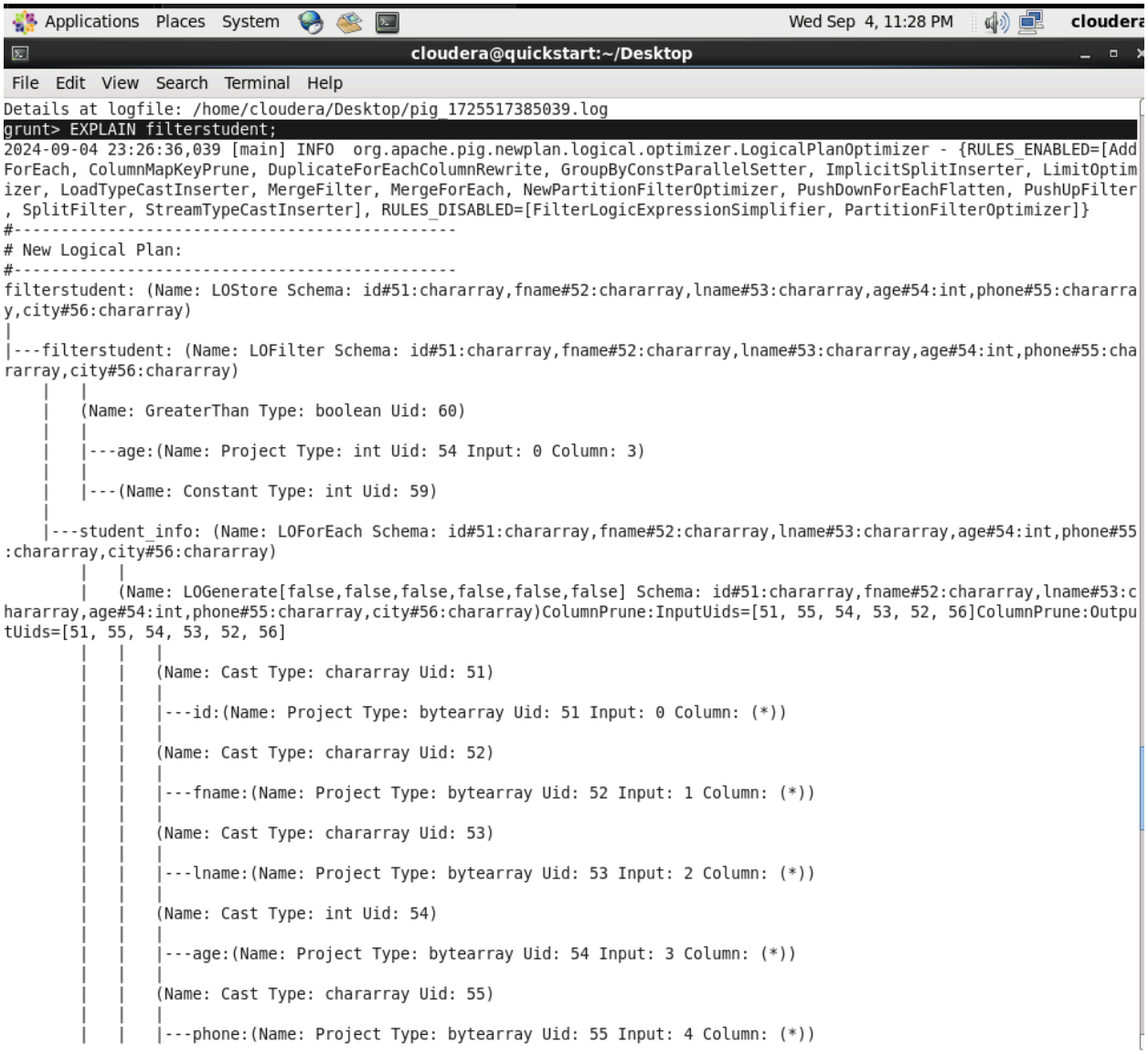

Applications Places System
Wed Sep 4, 11:19 PM
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
2024-09-04 23:19:21,490 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1066: Unable to open iterator for alias student_info
Details at logfile: /home/cloudera/Desktop/pig_1725516989942.log
grunt> STORE student_info INTO 'sampleoutput' USING PigStorage('|');
2024-09-04 23:19:38,442 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-04 23:19:38,444 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-04 23:19:38,449 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2024-09-04 23:19:38,481 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-04 23:19:38,487 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-04 23:19:38,490 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-04 23:19:38,589 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-09-04 23:19:38,596 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-04 23:19:38,630 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-04 23:19:40,740 [DataStreamer for file /tmp/temp-807703593/tmp-1932282297/jdo-api-3.0.1.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
at java.lang.Object.wait(Native Method)
at java.lang.Thread.join(Thread.java:1281)
at java.lang.Thread.join(Thread.java:1355)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)


Applications Places System
Fri Sep 6, 8:17 AM
cloudera@quickstart:~
File Edit View Search Terminal Help
2024-09-06 08:16:57,536 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved o
a/sampleoutput/_temporary/0/task_local402521900_0002_m_000000
2024-09-06 08:16:57,539 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - map
2024-09-06 08:16:57,539 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local402521900_0002_m_
2024-09-06 08:16:57,540 [LocalJobRunner Map Task Executor #0] INFO org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_lo
2024-09-06 08:16:57,545 [Thread-14] INFO org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2024-09-06 08:17:09,278 [main] WARN org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local402521900_00
2024-09-06 08:17:09,278 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-09-06 08:17:09,279 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be inc
2024-09-06 08:17:09,279 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0 cloudera 2024-09-06 08:16:56 2024-09-06 08:17:09 UNKNOWN
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local402521900_0002 student_info MAP_ONLY file:///home/cloudera/sampleoutput,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:///home/cloudera/sampleoutput"
Job DAG:
job_local402521900_0002
2024-09-06 08:17:15,281 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> SS

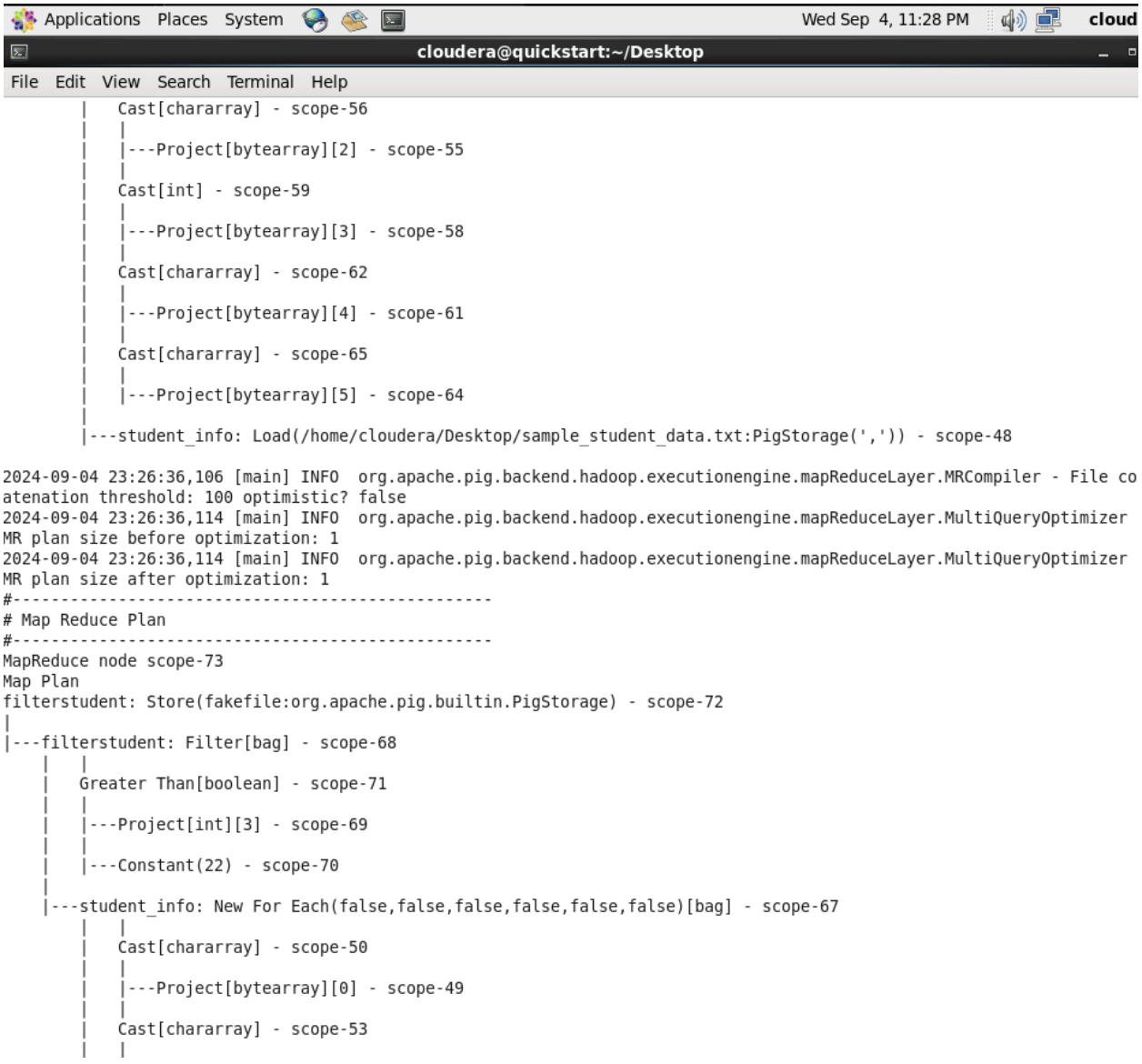
```
2024-09-06 08:17:15,281 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> DESCRIBE student_info;
2024-09-06 08:17:58,245 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2024-09-06 08:17:58,245 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use m
apreduce.jobtracker.address
2024-09-06 08:17:58,245 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
e dfs.bytes-per-checksum
student_info: {id: chararray,fname: chararray,lname: chararray,age: int,phone: chararray,city: chararray}
grunt>
```

```
Applications  Places  System                                   Wed Sep 4, 11:25 PM   cloude
cloudera@quickstart:~/Desktop
File  Edit  View  Search  Terminal  Help

Details at logfile: /home/cloudera/Desktop/pig_1725517385039.log
grunt> filterstudent = FILTER student_info BY age > 22;
grunt> DUMP filterstudent;
2024-09-04 23:24:48,487 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2024-09-04 23:24:48,493 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[Add
ForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter
, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-04 23:24:48,590 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File conc
atenation threshold: 100 optimistic? false
2024-09-04 23:24:48,611 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
MR plan size before optimization: 1
2024-09-04 23:24:48,611 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
MR plan size after optimization: 1
2024-09-04 23:24:48,753 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8
032
2024-09-04 23:24:48,759 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the jo
b
2024-09-04 23:24:48,802 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - m
apred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-04 23:24:49,252 [DataStreamer for file /tmp/temp1279935065/tmp2121047525/libthrift-0.9.3.jar] WARN  org.apache.h
adoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
2024-09-04 23:24:51,609 [DataStreamer for file /tmp/temp1279935065/tmp913367045/jdo-api-3.0.1.jar] WARN  org.apache.hado
op.hdfs.DFSClient - Caught exception
```

```

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local1539446827_0003        filterstudent,student_info      MAP_ONLY        file:/tmp/temp142338119/tmp-308534649,

Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp142338119/tmp-308534649"

Job DAG:
job_local1539446827_0003


2024-09-06 08:19:07,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:19:07,908 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
efaultFS
2024-09-06 08:19:07,908 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
apreduce.jobtracker.address
2024-09-06 08:19:07,908 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
e dfs.bytes-per-checksum
2024-09-06 08:19:07,909 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:19:07,953 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:19:07,953 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(005,Raj,Jain,23,9848022336,Mumbai)
(006,Shubham,Mishra,23,9848022335,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata)
```


Applications Places System
Wed Sep 4, 11:28 PM
cloudera
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Details at logfile: /home/cloudera/Desktop/pig_1725517385039.log
grunt> EXPLAIN filterstudent;
2024-09-04 23:26:36,039 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
#---
New Logical Plan:
#---
filterstudent: (Name: LOStore Schema: id#51:chararray,fname#52:chararray,lname#53:chararray,age#54:int,phone#55:chararray,city#56:chararray)
|
|---filterstudent: (Name: LOFilter Schema: id#51:chararray,fname#52:chararray,lname#53:chararray,age#54:int,phone#55:chararray,city#56:chararray)
| |
| (Name: GreaterThan Type: boolean Uid: 60)
| |
| |---age:(Name: Project Type: int Uid: 54 Input: 0 Column: 3)
| |
| |---(Name: Constant Type: int Uid: 59)
|
|---student_info: (Name: LOForEach Schema: id#51:chararray,fname#52:chararray,lname#53:chararray,age#54:int,phone#55:chararray,city#56:chararray)
| |
| (Name: LOGenerate[false,false,false,false,false,false] Schema: id#51:chararray,fname#52:chararray,lname#53:chararray,age#54:int,phone#55:chararray,city#56:chararray)ColumnPrune:InputUids=[51, 55, 54, 53, 52, 56]ColumnPrune:OutputUids=[51, 55, 54, 53, 52, 56]
| | |
| | (Name: Cast Type: chararray Uid: 51)
| | |
| | |---id:(Name: Project Type: bytearray Uid: 51 Input: 0 Column: (*))
| | |
| | (Name: Cast Type: chararray Uid: 52)
| | |
| | |---fname:(Name: Project Type: bytearray Uid: 52 Input: 1 Column: (*))
| | |
| | (Name: Cast Type: chararray Uid: 53)
| | |
| | |---lname:(Name: Project Type: bytearray Uid: 53 Input: 2 Column: (*))
| | |
| | (Name: Cast Type: int Uid: 54)
| | |
| | |---age:(Name: Project Type: bytearray Uid: 54 Input: 3 Column: (*))
| | |
| | (Name: Cast Type: chararray Uid: 55)
| | |
| | |---phone:(Name: Project Type: bytearray Uid: 55 Input: 4 Column: (*))

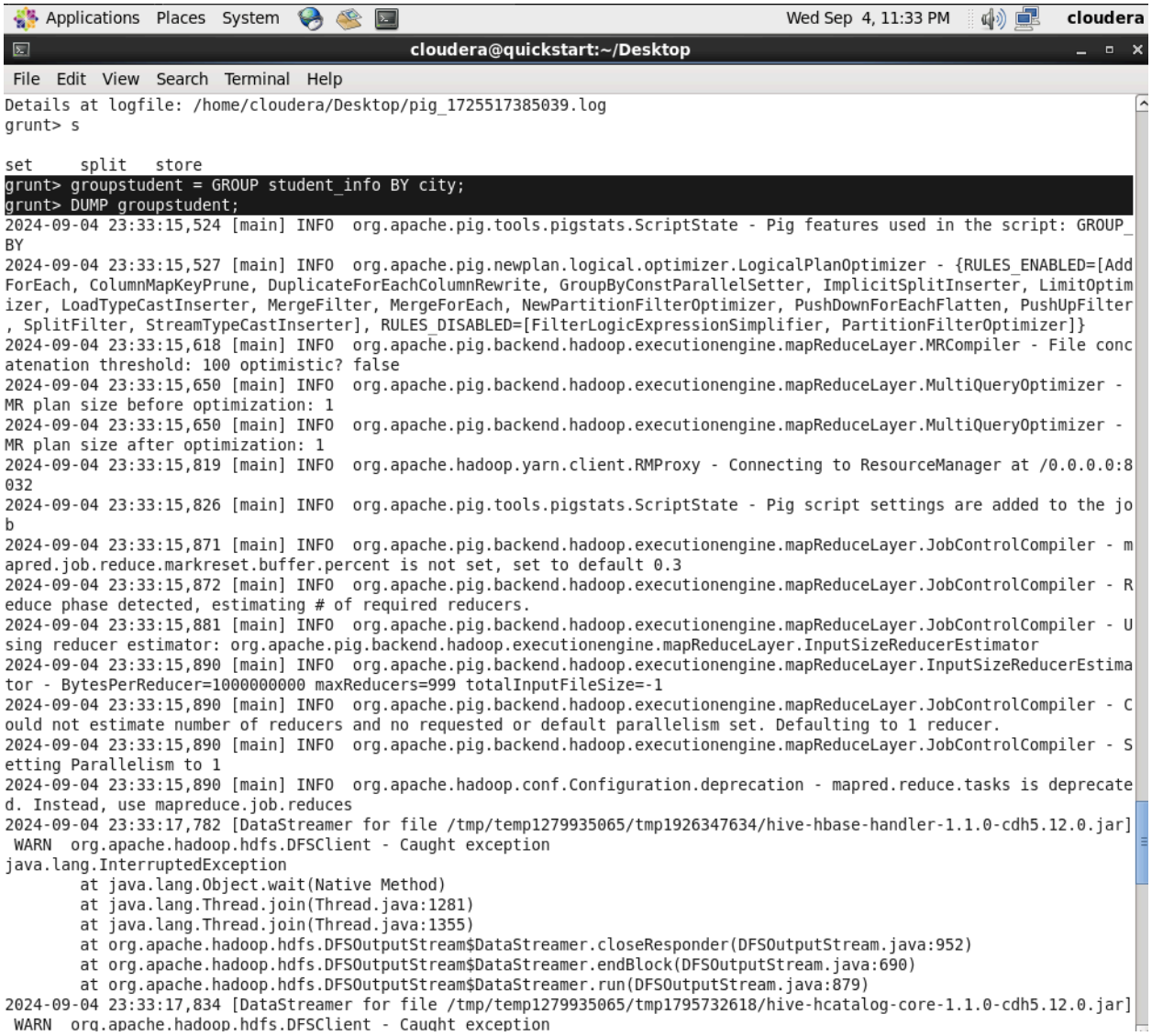
```
        |   |   |
        |   |   |---phone:(Name: Project Type: bytearray Uid: 55 Input: 4 Column: (*))
        |   |   |
        |   |   (Name: Cast Type: chararray Uid: 56)
        |   |   |
        |   |   |---city:(Name: Project Type: bytearray Uid: 56 Input: 5 Column: (*))
        |   |
        |   |---(Name: LOInnerLoad[0] Schema: id#51:bytearray)
        |   |
        |   |---(Name: LOInnerLoad[1] Schema: fname#52:bytearray)
        |   |
        |   |---(Name: LOInnerLoad[2] Schema: lname#53:bytearray)
        |   |
        |   |---(Name: LOInnerLoad[3] Schema: age#54:bytearray)
        |   |
        |   |---(Name: LOInnerLoad[4] Schema: phone#55:bytearray)
        |   |
        |   |---(Name: LOInnerLoad[5] Schema: city#56:bytearray)
        |
        |---student_info: (Name: LOLoad Schema: id#51:bytearray,fname#52:bytearray,lname#53:bytearray,age#54:bytearray,p
hone#55:bytearray,city#56:bytearray)RequiredFields:null

#-----------------------------------------------
# Physical Plan:
#-----------------------------------------------
filterstudent: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-72
|
|---filterstudent: Filter[bag] - scope-68
    |   |
    |   Greater Than[boolean] - scope-71
    |   |
    |   |---Project[int][3] - scope-69
    |   |
    |   |---Constant(22) - scope-70
    |
    |---student_info: New For Each(false,false,false,false,false,false)[bag] - scope-67
        |   |
        |   Cast[chararray] - scope-50
        |   |
        |   |---Project[bytearray][0] - scope-49
        |   |
        |   Cast[chararray] - scope-53
        |   |
        |   |---Project[bytearray][1] - scope-52
        |   |
        |   Cast[chararray] - scope-56
        |   |
```

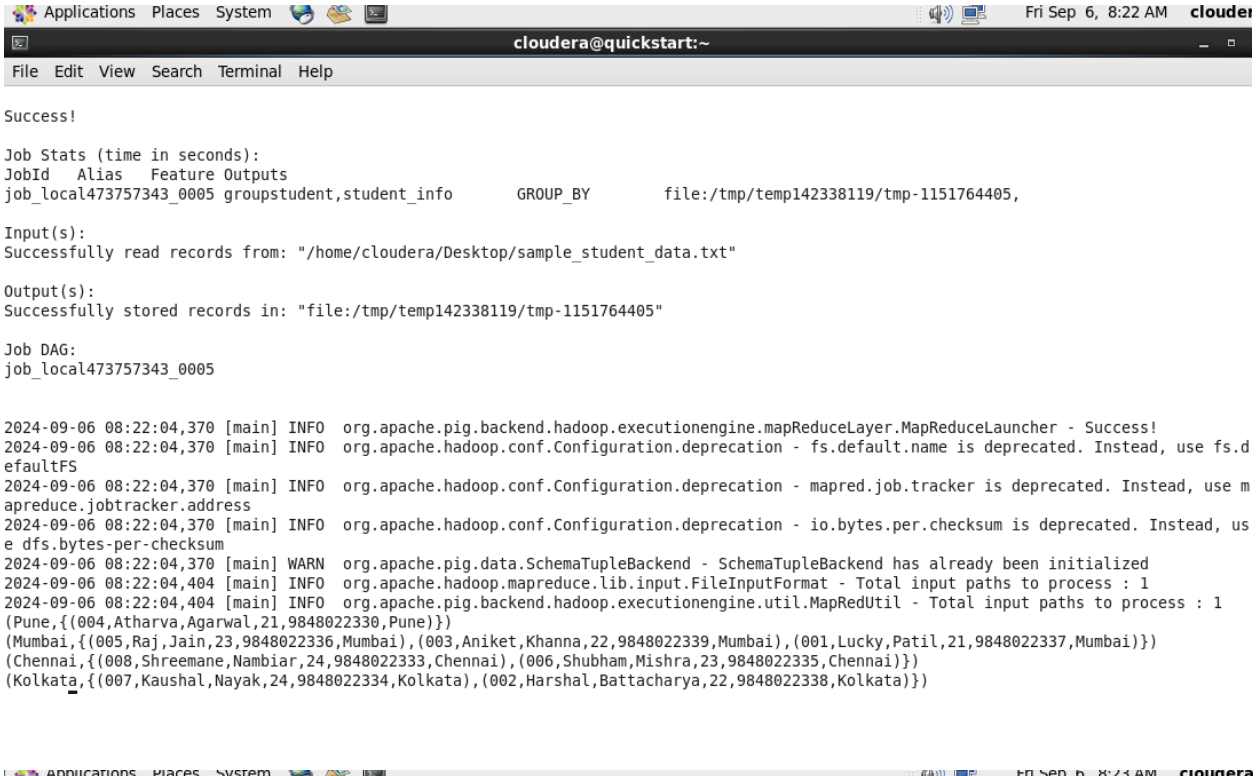

Applications Places System
Wed Sep 4, 11:28 PM
cloud
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
| Cast[chararray] - scope-56
| |
| |---Project[bytearray][2] - scope-55
| |
| Cast[int] - scope-59
| |
| |---Project[bytearray][3] - scope-58
| |
| Cast[chararray] - scope-62
| |
| |---Project[bytearray][4] - scope-61
| |
| Cast[chararray] - scope-65
| |
| |---Project[bytearray][5] - scope-64
|
|---student_info: Load(/home/cloudera/Desktop/sample_student_data.txt:PigStorage(',')) - scope-48
2024-09-04 23:26:36,106 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File co
atenation threshold: 100 optimistic? false
2024-09-04 23:26:36,114 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
MR plan size before optimization: 1
2024-09-04 23:26:36,114 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
MR plan size after optimization: 1
#---
Map Reduce Plan
#---
MapReduce node scope-73
Map Plan
filterstudent: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-72
|
|---filterstudent: Filter[bag] - scope-68
| |
| Greater Than[boolean] - scope-71
| |
| |---Project[int][3] - scope-69
| |
| |---Constant(22) - scope-70
|
|---student_info: New For Each(false,false,false,false,false,false)[bag] - scope-67
| |
| Cast[chararray] - scope-50
| |
| |---Project[bytearray][0] - scope-49
| |
| Cast[chararray] - scope-53
| |

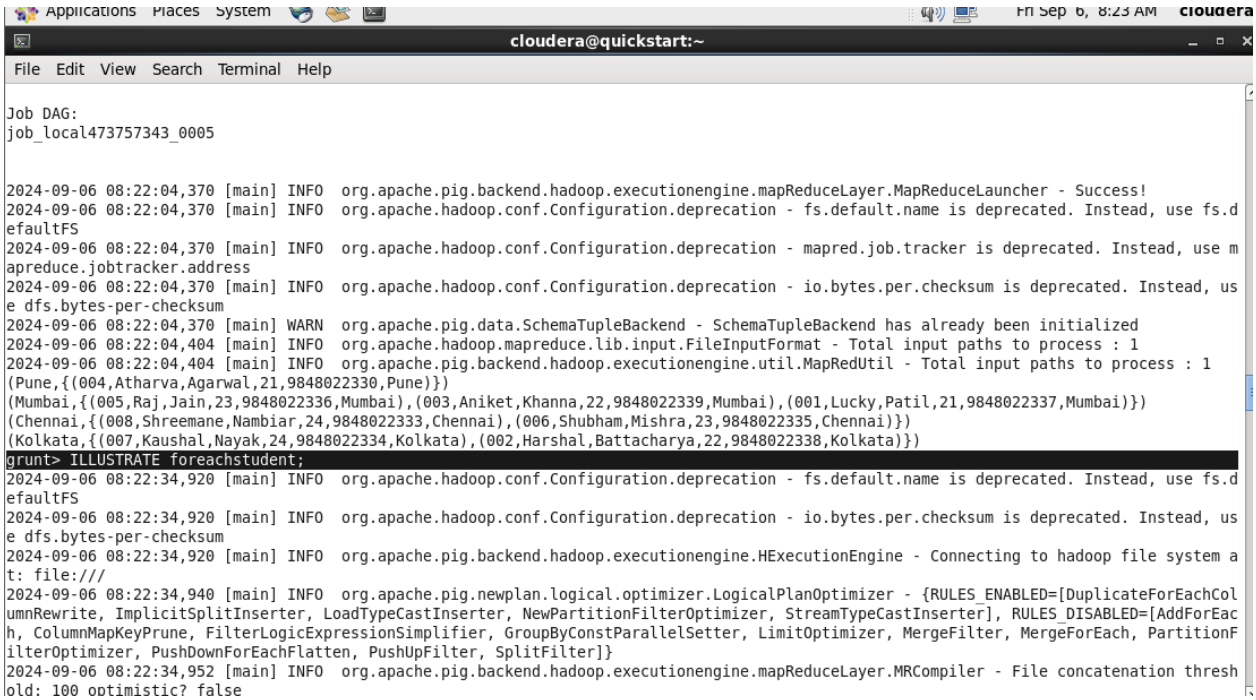
```
        |   |
        |   Cast[chararray] - scope-53
        |   |
        |   |---Project[bytearray][1] - scope-52
        |   |
        |   Cast[chararray] - scope-56
        |   |
        |   |---Project[bytearray][2] - scope-55
        |   |
        |   Cast[int] - scope-59
        |   |
        |   |---Project[bytearray][3] - scope-58
        |   |
        |   Cast[chararray] - scope-62
        |   |
        |   |---Project[bytearray][4] - scope-61
        |   |
        |   Cast[chararray] - scope-65
        |   |
        |   |---Project[bytearray][5] - scope-64
        |
        |---student_info: Load(/home/cloudera/Desktop/sample_student_data.txt:PigStorage(',')) - scope-48--------
Global sort: false
----------------
```

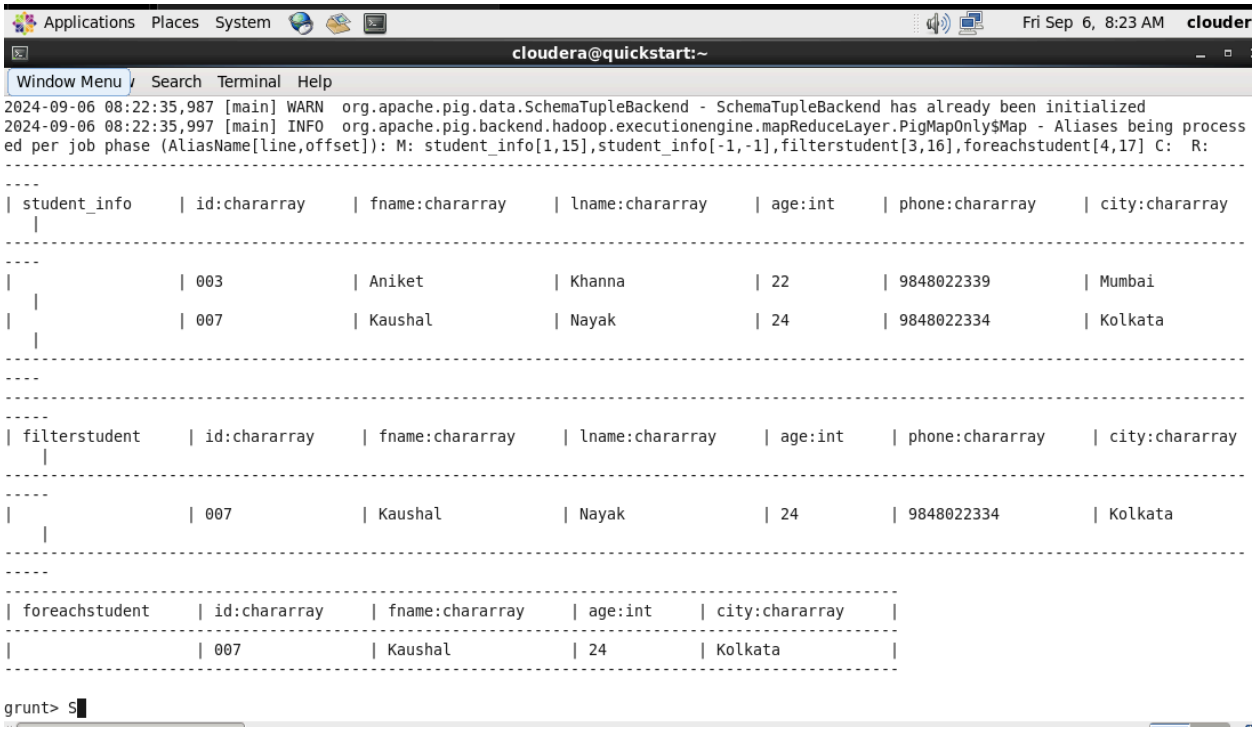
```
Applications  Places  System                                   Wed Sep 4, 11:31 PM   cloudera
cloudera@quickstart:~/Desktop
File  Edit  View  Search  Terminal  Help
grunt> foreachstudent = FOREACH filterstudent GENERATE id, fname, age, city;
grunt> DUMP foreachstudent;
2024-09-04 23:31:27,770 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2024-09-04 23:31:27,776 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[Add
ForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter
, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-04 23:31:27,818 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for studen
t_info: $2, $4
2024-09-04 23:31:27,889 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File conc
atenation threshold: 100 optimistic? false
2024-09-04 23:31:27,904 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
MR plan size before optimization: 1
2024-09-04 23:31:27,904 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer -
MR plan size after optimization: 1
2024-09-04 23:31:28,064 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8
032
2024-09-04 23:31:28,075 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the jo
b
2024-09-04 23:31:28,135 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - m
apred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-04 23:31:28,506 [DataStreamer for file /tmp/temp1279935065/tmp1766266818/libthrift-0.9.3.jar] WARN  org.apache.h
adoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
2024-09-04 23:31:29,861 [DataStreamer for file /tmp/temp1279935065/tmp366025451/hive-hbase-handler-1.1.0-cdh5.12.0.jar]
WARN  org.apache.hadoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
2024-09-04 23:31:29,969 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - c
reating jar file Job8363759557687636166.jar
2024-09-04 23:31:38,931 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - j
ar file Job8363759557687636166.jar created
2024-09-04 23:31:38,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - S
etting up single store job
2024-09-04 23:31:38,955 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not
generate code
```

```
Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local2042764522_0004        filterstudent,foreachstudent,student_info       MAP_ONLY        file:/tmp/temp142338119/tmp815996228,

Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp142338119/tmp815996228"

Job DAG:
job_local2042764522_0004


2024-09-06 08:21:07,977 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:21:07,977 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2024-09-06 08:21:07,977 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use m
apreduce.jobtracker.address
2024-09-06 08:21:07,978 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
e dfs.bytes-per-checksum
2024-09-06 08:21:07,981 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:21:08,027 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:21:08,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(005,Raj,23,Mumbai)
(006,Shubham,23,Chennai)
(007,Kaushal,24,Kolkata)
(008,Shreemane,24,Chennai)
```

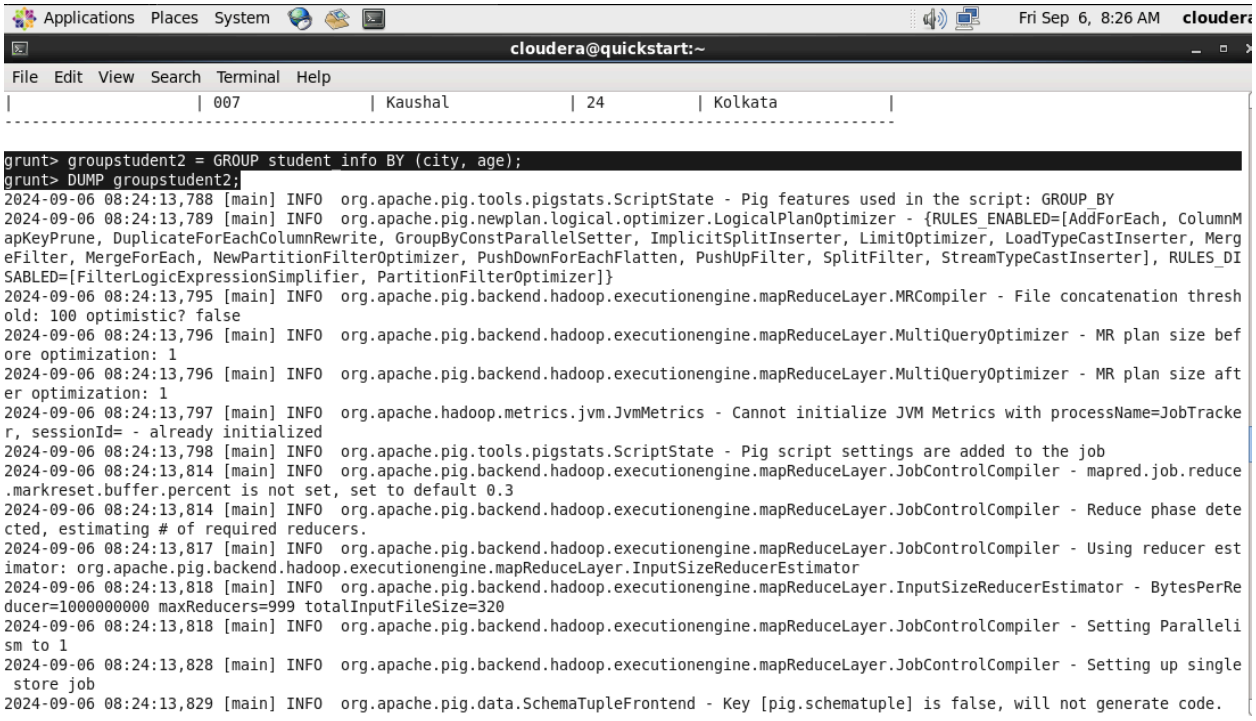

Applications Places System
Wed Sep 4, 11:33 PM
cloudera
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Details at logfile: /home/cloudera/Desktop/pig_1725517385039.log
grunt> s
set split store
grunt> groupstudent = GROUP student_info BY city;
grunt> DUMP groupstudent;
2024-09-04 23:33:15,524 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-09-04 23:33:15,527 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-04 23:33:15,618 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-04 23:33:15,650 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-04 23:33:15,650 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-04 23:33:15,819 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-09-04 23:33:15,826 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-04 23:33:15,871 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-04 23:33:15,872 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-04 23:33:15,881 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-04 23:33:15,890 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=-1
2024-09-04 23:33:15,890 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Could not estimate number of reducers and no requested or default parallelism set. Defaulting to 1 reducer.
2024-09-04 23:33:15,890 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-04 23:33:15,890 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2024-09-04 23:33:17,782 [DataStreamer for file /tmp/temp1279935065/tmp1926347634/hive-hbase-handler-1.1.0-cdh5.12.0.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
at java.lang.Object.wait(Native Method)
at java.lang.Thread.join(Thread.java:1281)
at java.lang.Thread.join(Thread.java:1355)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
2024-09-04 23:33:17,834 [DataStreamer for file /tmp/temp1279935065/tmp1795732618/hive-hcatalog-core-1.1.0-cdh5.12.0.jar] WARN org.apache.hadoop.hdfs.DFSClient - Caught exception


Applications Places System
Fri Sep 6, 8:22 AM
cloudera@quickstart:~
File Edit View Search Terminal Help
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local473757343_0005 groupstudent,student_info GROUP_BY file:/tmp/temp142338119/tmp-1151764405,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp142338119/tmp-1151764405"
Job DAG:
job_local473757343_0005
2024-09-06 08:22:04,370 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:22:04,370 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:22:04,370 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:22:04,370 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:22:04,370 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:22:04,404 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:22:04,404 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Pune,{(004,Atharva,Agarwal,21,9848022330,Pune)})
(Mumbai,{(005,Raj,Jain,23,9848022336,Mumbai),(003,Aniket,Khanna,22,9848022339,Mumbai),(001,Lucky,Patil,21,9848022337,Mumbai)})
(Chennai,{(008,Shreemane,Nambiar,24,9848022333,Chennai),(006,Shubham,Mishra,23,9848022335,Chennai)})
(Kolkata,{(007,Kaushal,Nayak,24,9848022334,Kolkata),(002,Harshal,Battacharya,22,9848022338,Kolkata)})


Applications Places System
Fri Sep 6, 8:23 AM
cloudera@quickstart:~
File Edit View Search Terminal Help
Job DAG:
job_local473757343_0005
2024-09-06 08:22:04,370 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:22:04,370 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:22:04,370 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:22:04,370 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:22:04,370 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:22:04,404 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:22:04,404 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Pune,{(004,Atharva,Agarwal,21,9848022330,Pune)})
(Mumbai,{(005,Raj,Jain,23,9848022336,Mumbai),(003,Aniket,Khanna,22,9848022339,Mumbai),(001,Lucky,Patil,21,9848022337,Mumbai)})
(Chennai,{(008,Shreemane,Nambiar,24,9848022333,Chennai),(006,Shubham,Mishra,23,9848022335,Chennai)})
(Kolkata,{(007,Kaushal,Nayak,24,9848022334,Kolkata),(002,Harshal,Battacharya,22,9848022338,Kolkata)})
grunt> ILLUSTRATE foreachstudent;
2024-09-06 08:22:34,920 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:22:34,920 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:22:34,920 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-09-06 08:22:34,940 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[DuplicateForEachColumnRewrite, ImplicitSplitInserter, LoadTypeCastInserter, NewPartitionFilterOptimizer, StreamTypeCastInserter], RULES_DISABLED=[AddForEach, ColumnMapKeyPrune, FilterLogicExpressionSimplifier, GroupByConstParallelSetter, LimitOptimizer, MergeFilter, MergeForEach, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter]}
2024-09-06 08:22:34,952 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false



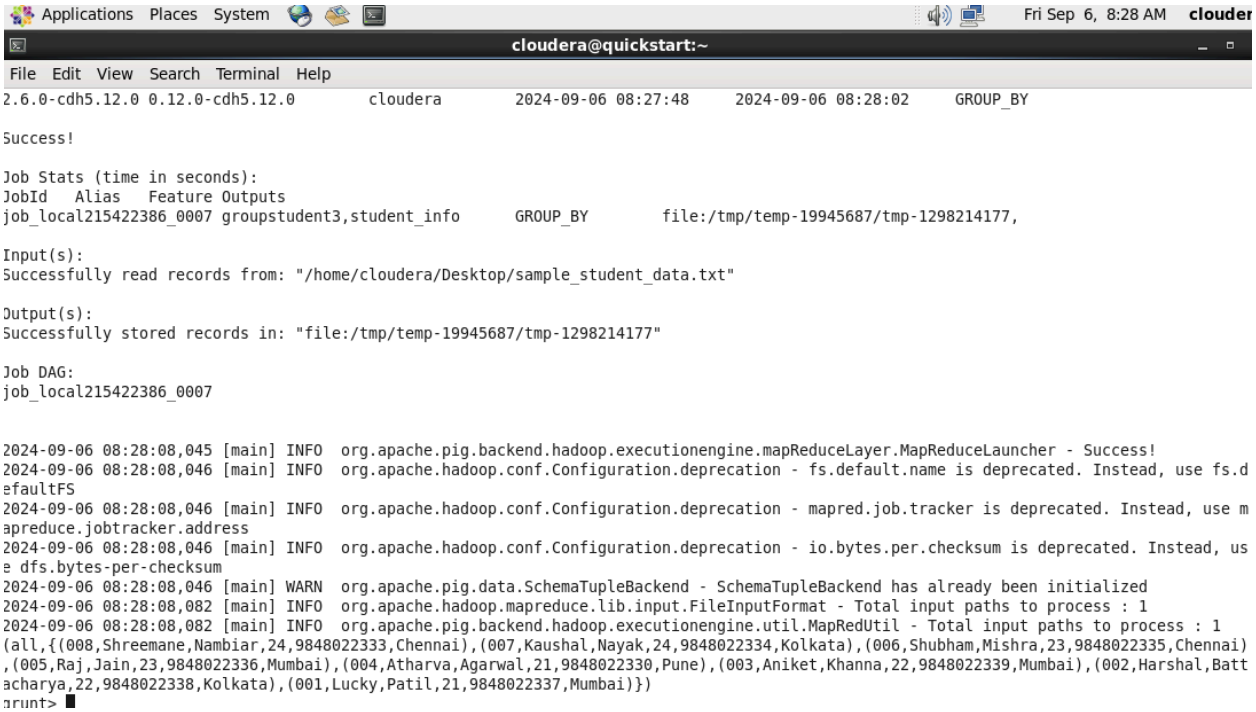
```
Applications  Places  System                                   Fri Sep 6, 8:23 AM  clouder
cloudera@quickstart:~
Window Menu  Search  Terminal  Help
2024-09-06 08:22:35,987 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:22:35,997 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being process
ed per job phase (AliasName[line,offset]): M: student_info[1,15],student_info[-1,-1],filterstudent[3,16],foreachstudent[4,17] C:  R:
-------------------------------------------------------------------------------------------------------------------------------------
----
| student_info     | id:chararray     | fname:chararray     | lname:chararray     | age:int     | phone:chararray     | city:chararray
   |
-------------------------------------------------------------------------------------------------------------------------------------
----
|                  | 003              | Aniket              | Khanna              | 22          | 9848022339          | Mumbai
   |
|                  | 007              | Kaushal             | Nayak               | 24          | 9848022334          | Kolkata
   |
-------------------------------------------------------------------------------------------------------------------------------------
----
-------------------------------------------------------------------------------------------------------------------------------------
-----
| filterstudent     | id:chararray     | fname:chararray     | lname:chararray     | age:int     | phone:chararray     | city:chararray
    |
-------------------------------------------------------------------------------------------------------------------------------------
-----
|                   | 007              | Kaushal             | Nayak               | 24          | 9848022334          | Kolkata
    |
-------------------------------------------------------------------------------------------------------------------------------------
-----
-----------------------------------------------------------------------------------------
| foreachstudent     | id:chararray     | fname:chararray     | age:int     | city:chararray     |
-----------------------------------------------------------------------------------------
|                    | 007              | Kaushal             | 24          | Kolkata            |
-----------------------------------------------------------------------------------------

grunt> S
```



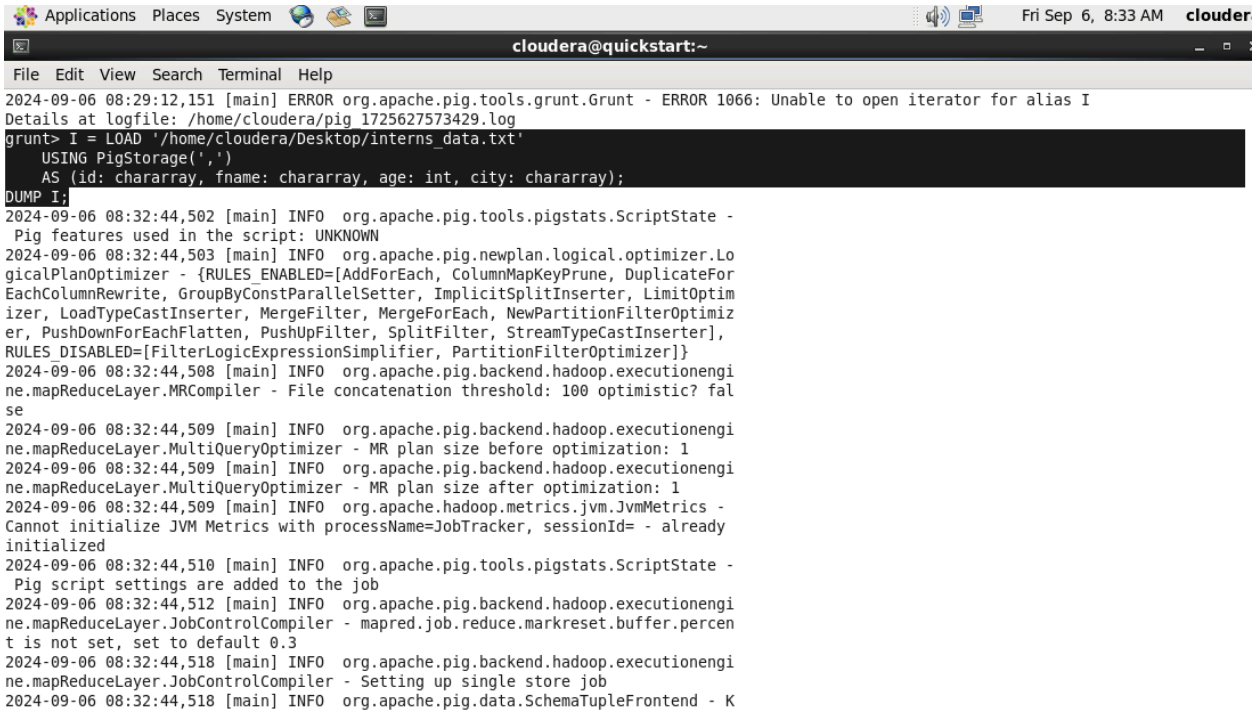
```
Applications  Places  System                                   Fri Sep 6, 8:26 AM  cloudera
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
|                   | 007              | Kaushal             | 24          | Kolkata            |
-----------------------------------------------------------------------------------------

grunt> groupstudent2 = GROUP student_info BY (city, age);
grunt> DUMP groupstudent2;
2024-09-06 08:24:13,788 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-09-06 08:24:13,789 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnM
apKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, Merg
eFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DI
SABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:24:13,795 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
old: 100 optimistic? false
2024-09-06 08:24:13,796 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
ore optimization: 1
2024-09-06 08:24:13,796 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
er optimization: 1
2024-09-06 08:24:13,797 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2024-09-06 08:24:13,798 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:24:13,814 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:24:13,814 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase dete
cted, estimating # of required reducers.
2024-09-06 08:24:13,817 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer est
imator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:24:13,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerRe
ducer=1000000000 maxReducers=999 totalInputFileSize=320
2024-09-06 08:24:13,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Paralleli
sm to 1
2024-09-06 08:24:13,828 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single
 store job
2024-09-06 08:24:13,829 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
```

```
JobId   Alias   Feature Outputs
job_local2021758745_0006        groupstudent2,student_info      GROUP_BY        file:/tmp/temp-19945687/tmp-206875807,

Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-206875807"

Job DAG:
job_local2021758745_0006


2024-09-06 08:24:32,366 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:24:32,367 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2024-09-06 08:24:32,367 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use m
apreduce.jobtracker.address
2024-09-06 08:24:32,368 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
e dfs.bytes-per-checksum
2024-09-06 08:24:32,368 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:24:32,418 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:24:32,418 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((Pune,21),{(004,Atharva,Agarwal,21,9848022330,Pune)})
((Mumbai,21),{(001,Lucky,Patil,21,9848022337,Mumbai)})
((Mumbai,22),{(003,Aniket,Khanna,22,9848022339,Mumbai)})
((Mumbai,23),{(005,Raj,Jain,23,9848022336,Mumbai)})
((Chennai,23),{(006,Shubham,Mishra,23,9848022335,Chennai)})
((Chennai,24),{(008,Shreemane,Nambiar,24,9848022333,Chennai)})
((Kolkata,22),{(002,Harshal,Battacharya,22,9848022338,Kolkata)})
((Kolkata,24),{(007,Kaushal,Nayak,24,9848022334,Kolkata)})
```

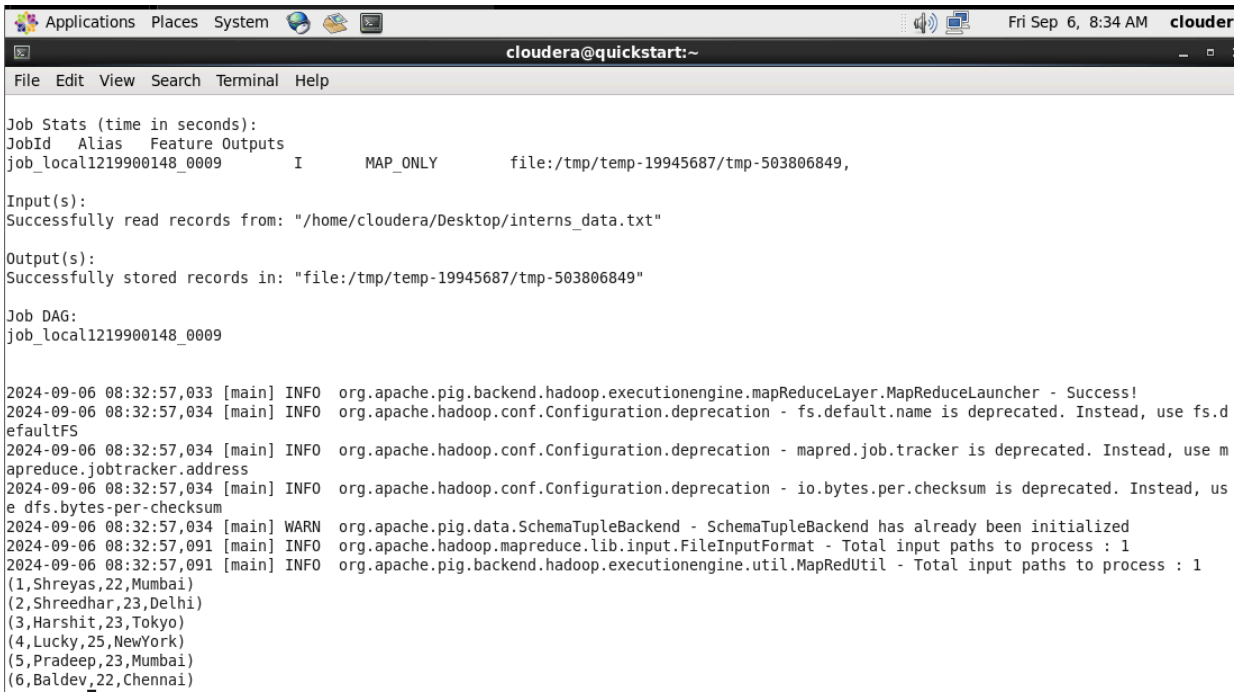
```
Applications  Places  System                                   Fri Sep 6, 8:28 AM  cloud
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
((Kolkata,24),{(007,Kaushal,Nayak,24,9848022334,Kolkata)})
grunt> groupstudent3 = GROUP student_info ALL;
grunt> DUMP groupstudent3;
2024-09-06 08:27:48,191 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2024-09-06 08:27:48,191 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Columm
apKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, Mer
eFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_D
SABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:27:48,261 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thres
old: 100 optimistic? false
2024-09-06 08:27:48,264 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size be
ore optimization: 1
2024-09-06 08:27:48,264 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size af
er optimization: 1
2024-09-06 08:27:48,267 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTrack
r, sessionId= - already initialized
2024-09-06 08:27:48,277 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:27:48,333 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduc
.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:27:48,339 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase det
cted, estimating # of required reducers.
2024-09-06 08:27:48,339 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallel
sm to 1
2024-09-06 08:27:48,460 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up singl
 store job
2024-09-06 08:27:48,468 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:27:48,468 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed c
che
2024-09-06 08:27:48,468 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode
Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725636468468-0
2024-09-06 08:27:48,625 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job
) waiting for submission.
```



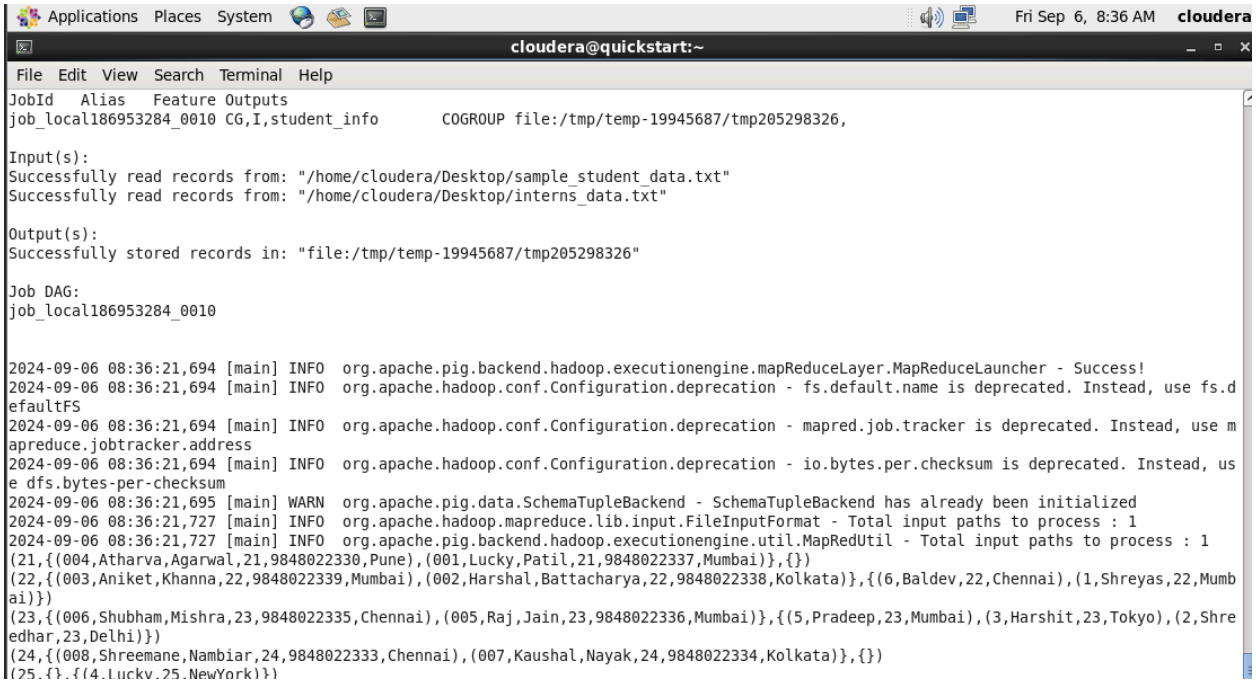
```
Applications  Places  System                                   Fri Sep 6, 8:28 AM  clouder
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2024-09-06 08:27:48     2024-09-06 08:28:02     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local215422386_0007 groupstudent3,student_info      GROUP_BY        file:/tmp/temp-19945687/tmp-1298214177,

Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"

Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-1298214177"

Job DAG:
job_local215422386_0007


2024-09-06 08:28:08,045 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:28:08,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2024-09-06 08:28:08,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use m
apreduce.jobtracker.address
2024-09-06 08:28:08,046 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
e dfs.bytes-per-checksum
2024-09-06 08:28:08,046 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:28:08,082 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:28:08,082 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(all,{(008,Shreemane,Nambiar,24,9848022333,Chennai),(007,Kaushal,Nayak,24,9848022334,Kolkata),(006,Shubham,Mishra,23,9848022335,Chennai)
,(005,Raj,Jain,23,9848022336,Mumbai),(004,Atharva,Agarwal,21,9848022330,Pune),(003,Aniket,Khanna,22,9848022339,Mumbai),(002,Harshal,Batt
acharya,22,9848022338,Kolkata),(001,Lucky,Patil,21,9848022337,Mumbai)})
grunt>
```



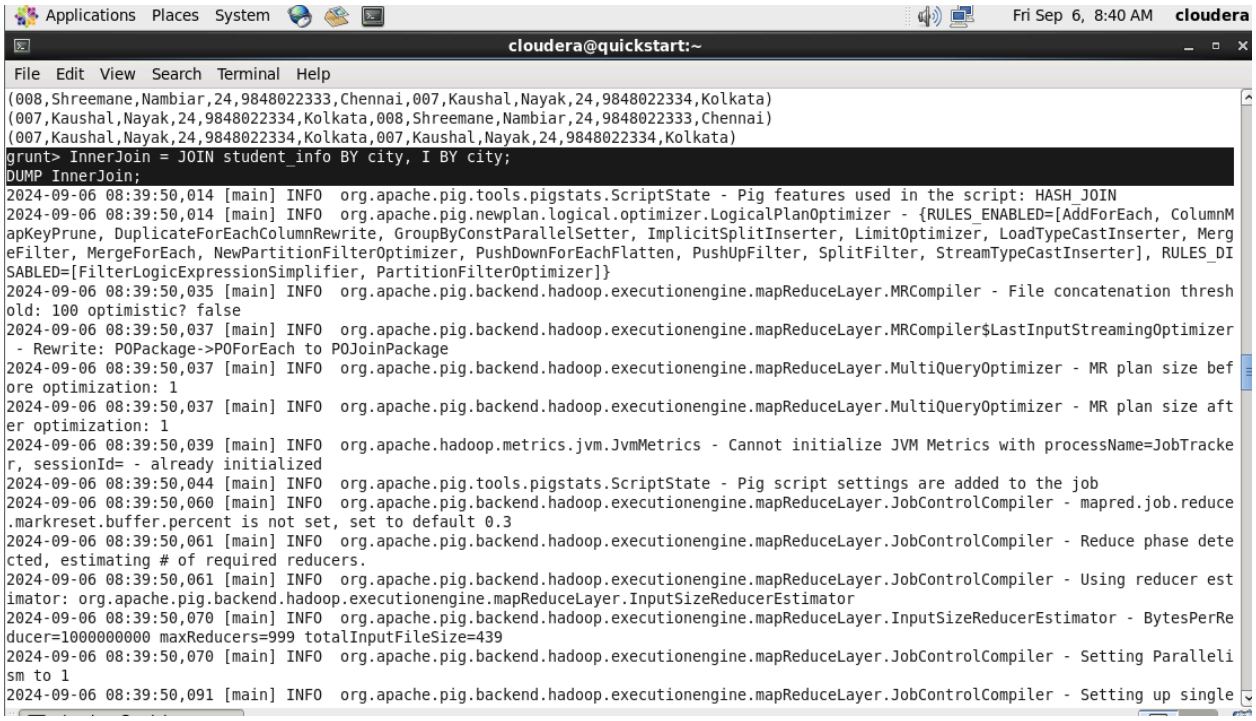
```
Applications  Places  System                                   Fri Sep 6, 8:33 AM  clouder
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
2024-09-06 08:29:12,151 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1066: Unable to open iterator for alias I
Details at logfile: /home/cloudera/pig_1725627573429.log
grunt> I = LOAD '/home/cloudera/Desktop/interns_data.txt'
    USING PigStorage(',')
    AS (id: chararray, fname: chararray, age: int, city: chararray);
DUMP I;
2024-09-06 08:32:44,502 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2024-09-06 08:32:44,503 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:32:44,508 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2024-09-06 08:32:44,509 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:32:44,509 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:32:44,509 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2024-09-06 08:32:44,510 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2024-09-06 08:32:44,512 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
t is not set, set to default 0.3
2024-09-06 08:32:44,518 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:32:44,518 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - K
```

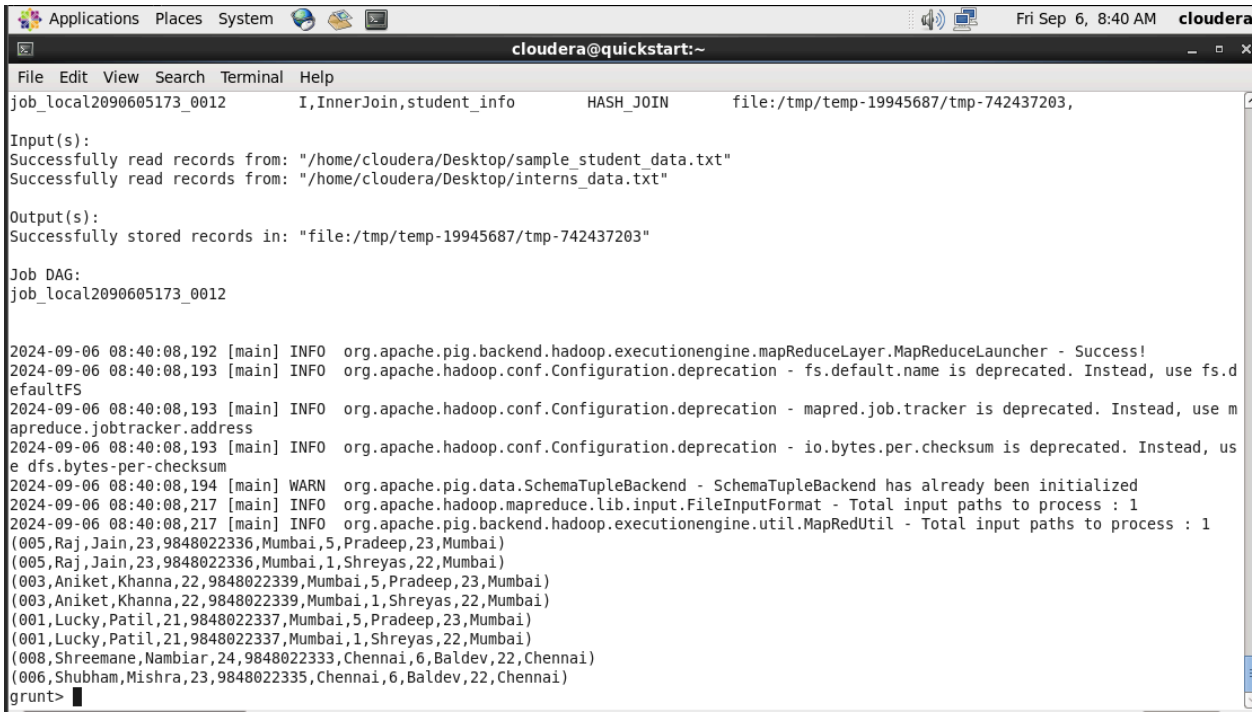

Applications Places System
Fri Sep 6, 8:34 AM clouder
cloudera@quickstart:~
File Edit View Search Terminal Help
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1219900148_0009 I MAP_ONLY file:/tmp/temp-19945687/tmp-503806849,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/interns_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-503806849"
Job DAG:
job_local1219900148_0009
2024-09-06 08:32:57,033 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:32:57,034 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:32:57,034 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:32:57,034 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:32:57,034 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:32:57,091 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:32:57,091 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Shreyas,22,Mumbai)
(2,Shreedhar,23,Delhi)
(3,Harshit,23,Tokyo)
(4,Lucky,25,NewYork)
(5,Pradeep,23,Mumbai)
(6,Baldev,22,Chennai)

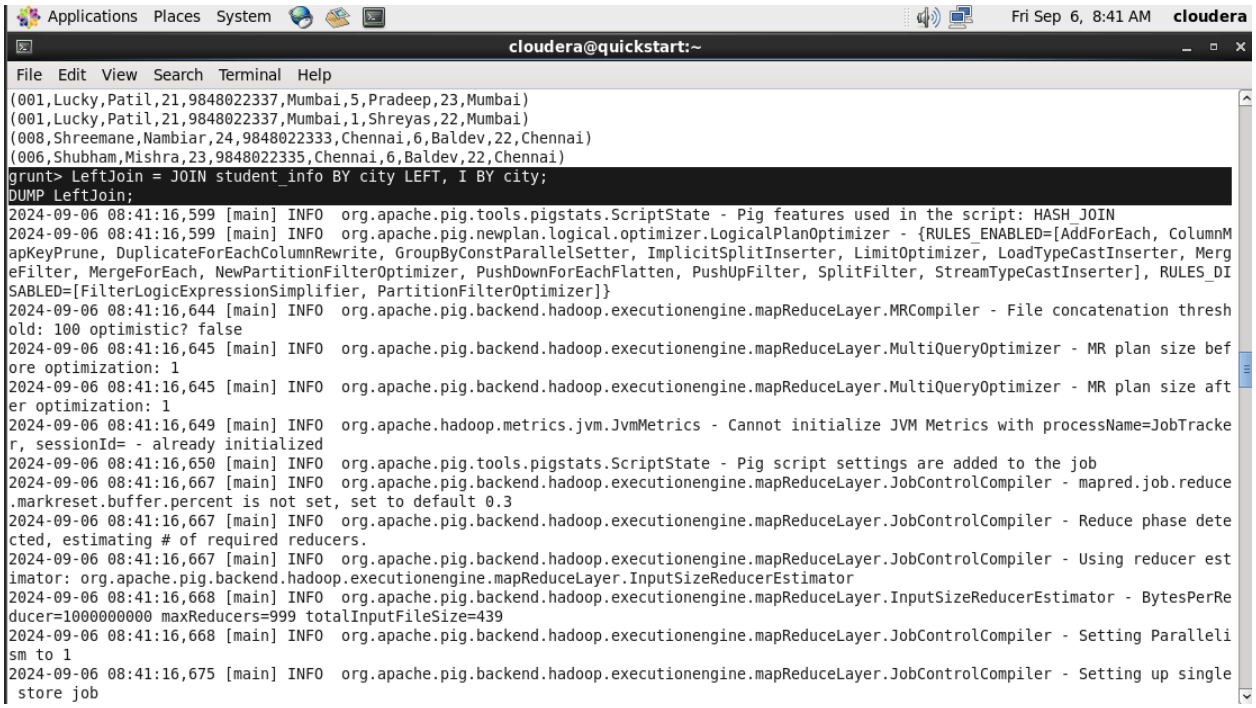
```
(5,Pradeep,23,Mumbai)
(6,Baldev,22,Chennai)
grunt> CG = COGROUP student_info BY age, I BY age;
grunt> DUMP CG;
2024-09-06 08:36:03,050 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: COGROUP
2024-09-06 08:36:03,056 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnM
apKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, Merg
eFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DI
SABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:36:03,106 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
old: 100 optimistic? false
2024-09-06 08:36:03,110 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
ore optimization: 1
2024-09-06 08:36:03,110 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
er optimization: 1
2024-09-06 08:36:03,113 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2024-09-06 08:36:03,114 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:36:03,127 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:36:03,127 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase dete
cted, estimating # of required reducers.
2024-09-06 08:36:03,127 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer est
imator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:36:03,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerRe
ducer=1000000000 maxReducers=999 totalInputFileSize=439
2024-09-06 08:36:03,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Paralleli
sm to 1
2024-09-06 08:36:03,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single
 store job
2024-09-06 08:36:03,137 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:36:03,137 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed ca
```



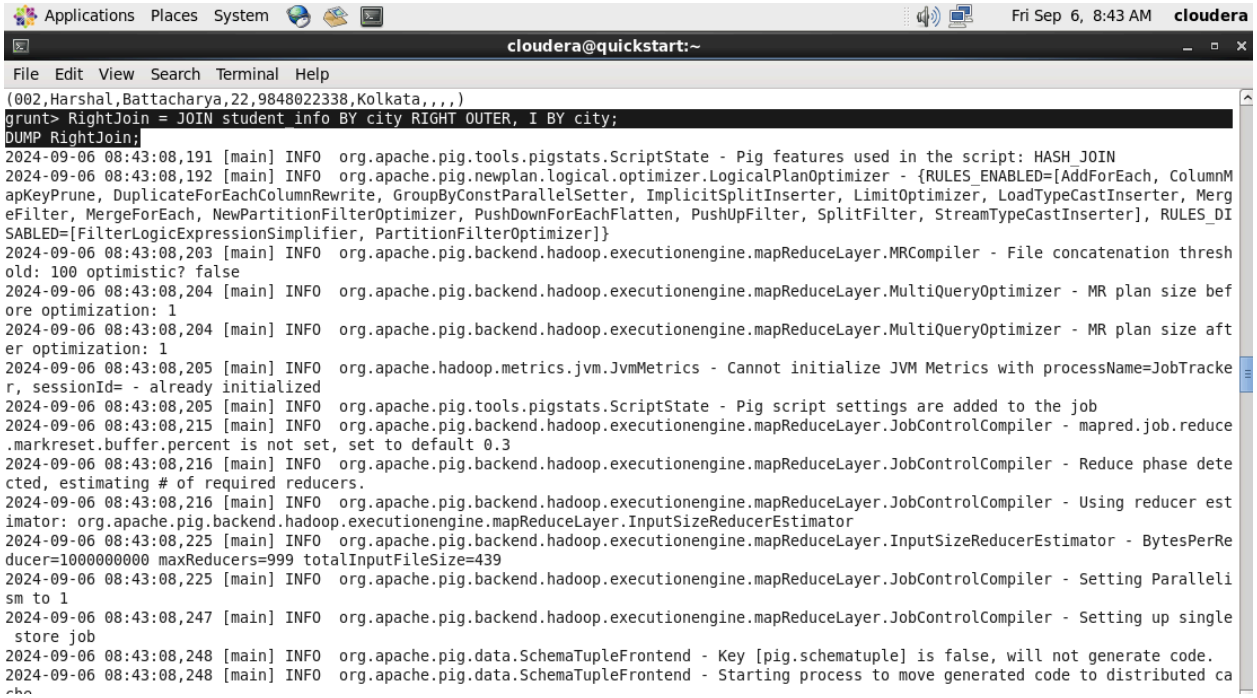

Applications Places System
Fri Sep 6, 8:38 AM
cloudera@quickstart:~
File Edit View Search Terminal Help
grunt> student_infoB = LOAD '/home/cloudera/Desktop/sample_student_data.txt'
USING PigStorage(',')
AS (id: chararray, fname: chararray, lname: chararray, age: int, phone: chararray, city: chararray);
grunt> SelfJoin = JOIN student_info BY age, student_infoB BY age;
grunt> DUMP SelfJoin;
2024-09-06 08:37:52,572 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2024-09-06 08:37:52,573 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:37:52,624 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:37:52,640 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POJoinPackage
2024-09-06 08:37:52,640 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:37:52,640 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:37:52,653 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:37:52,654 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:37:52,686 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:37:52,687 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-06 08:37:52,689 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:37:52,691 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=640
2024-09-06 08:37:52,691 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-06 08:37:52,704 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single
Job DAG:
job_local1020907157_0011
2024-09-06 08:38:11,593 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:38:11,594 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:38:11,594 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:38:11,594 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:38:11,594 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:38:11,644 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:38:11,644 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(004,Atharva,Agarwal,21,9848022330,Pune,004,Atharva,Agarwal,21,9848022330,Pune)
(004,Atharva,Agarwal,21,9848022330,Pune,001,Lucky,Patil,21,9848022337,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,004,Atharva,Agarwal,21,9848022330,Pune)
(001,Lucky,Patil,21,9848022337,Mumbai,001,Lucky,Patil,21,9848022337,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,003,Aniket,Khanna,22,9848022339,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,002,Harshal,Battacharya,22,9848022338,Kolkata)
(002,Harshal,Battacharya,22,9848022338,Kolkata,003,Aniket,Khanna,22,9848022339,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata,002,Harshal,Battacharya,22,9848022338,Kolkata)
(006,Shubham,Mishra,23,9848022335,Chennai,006,Shubham,Mishra,23,9848022335,Chennai)
(006,Shubham,Mishra,23,9848022335,Chennai,005,Raj,Jain,23,9848022336,Mumbai)
(005,Raj,Jain,23,9848022336,Mumbai,006,Shubham,Mishra,23,9848022335,Chennai)
(005,Raj,Jain,23,9848022336,Mumbai,005,Raj,Jain,23,9848022336,Mumbai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,008,Shreemane,Nambiar,24,9848022333,Chennai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,007,Kaushal,Nayak,24,9848022334,Kolkata)
(007,Kaushal,Nayak,24,9848022334,Kolkata,008,Shreemane,Nambiar,24,9848022333,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata,007,Kaushal,Nayak,24,9848022334,Kolkata)
grunt>


Applications Places System
Fri Sep 6, 8:40 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
(008,Shreemane,Nambiar,24,9848022333,Chennai,007,Kaushal,Nayak,24,9848022334,Kolkata)
(007,Kaushal,Nayak,24,9848022334,Kolkata,008,Shreemane,Nambiar,24,9848022333,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata,007,Kaushal,Nayak,24,9848022334,Kolkata)
grunt> InnerJoin = JOIN student_info BY city, I BY city;
DUMP InnerJoin;
2024-09-06 08:39:50,014 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2024-09-06 08:39:50,014 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:39:50,035 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:39:50,037 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POJoinPackage
2024-09-06 08:39:50,037 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:39:50,037 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:39:50,039 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:39:50,044 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:39:50,060 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:39:50,061 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-06 08:39:50,061 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:39:50,070 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=439
2024-09-06 08:39:50,070 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-06 08:39:50,091 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single

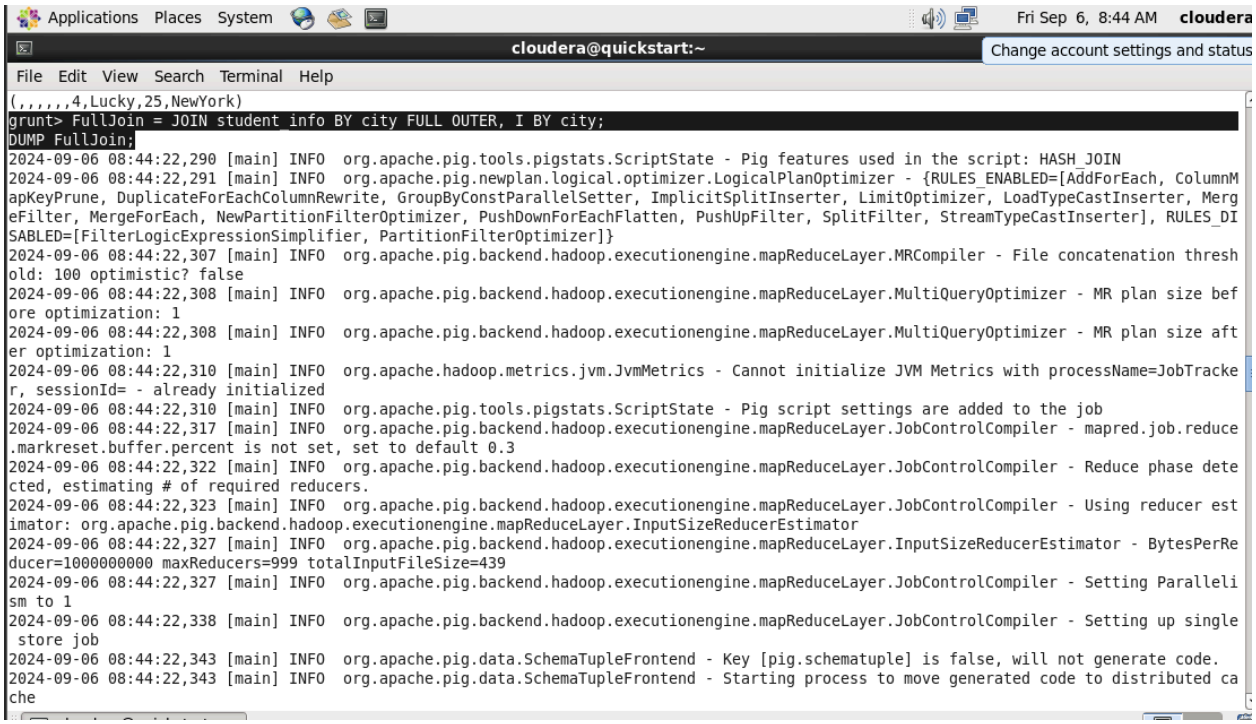

Applications Places System
Fri Sep 6, 8:40 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
job_local2090605173_0012 I,InnerJoin,student_info HASH_JOIN file:/tmp/temp-19945687/tmp-742437203,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Successfully read records from: "/home/cloudera/Desktop/interns_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-742437203"
Job DAG:
job_local2090605173_0012
2024-09-06 08:40:08,192 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:40:08,193 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:40:08,193 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:40:08,193 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:40:08,194 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:40:08,217 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:40:08,217 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(005,Raj,Jain,23,9848022336,Mumbai,5,Pradeep,23,Mumbai)
(005,Raj,Jain,23,9848022336,Mumbai,1,Shreyas,22,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,5,Pradeep,23,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,1,Shreyas,22,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,5,Pradeep,23,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,1,Shreyas,22,Mumbai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,6,Baldev,22,Chennai)
(006,Shubham,Mishra,23,9848022335,Chennai,6,Baldev,22,Chennai)
grunt>

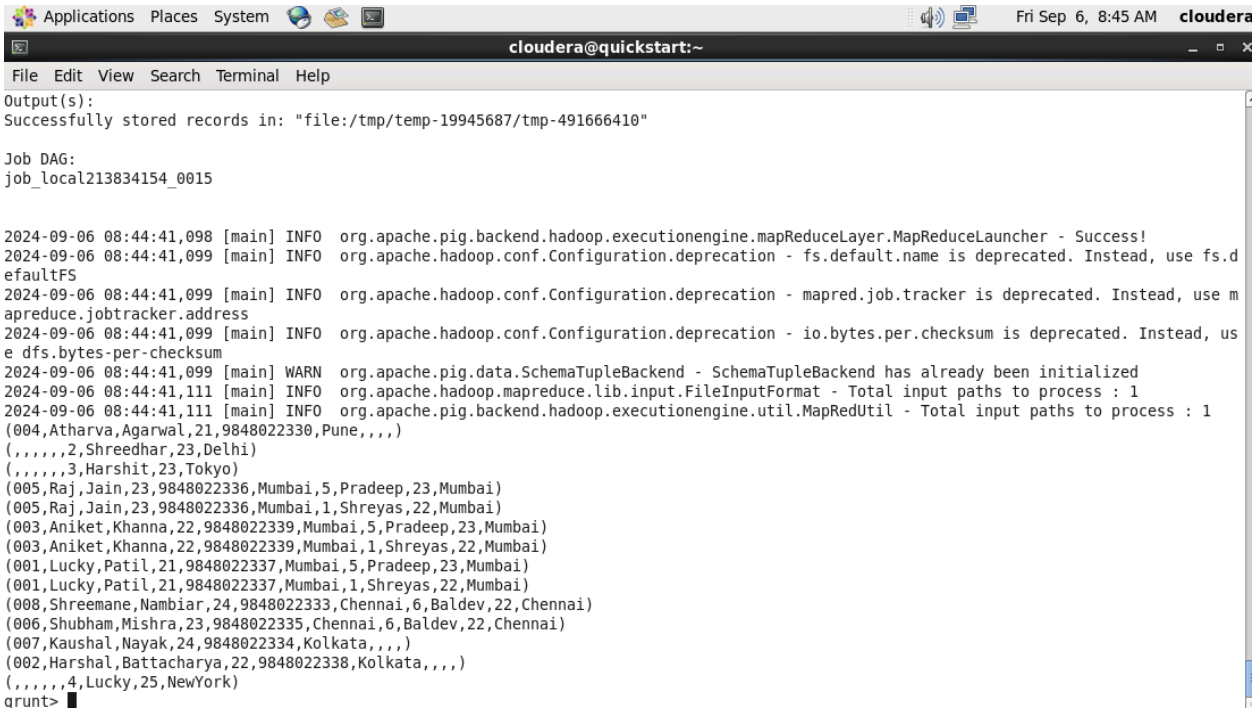

Applications Places System
Fri Sep 6, 8:41 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
(001,Lucky,Patil,21,9848022337,Mumbai,5,Pradeep,23,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,1,Shreyas,22,Mumbai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,6,Baldev,22,Chennai)
(006,Shubham,Mishra,23,9848022335,Chennai,6,Baldev,22,Chennai)
grunt> LeftJoin = JOIN student_info BY city LEFT, I BY city;
DUMP LeftJoin;
2024-09-06 08:41:16,599 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2024-09-06 08:41:16,599 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:41:16,644 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:41:16,645 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:41:16,645 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:41:16,649 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:41:16,650 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:41:16,667 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:41:16,667 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-06 08:41:16,667 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:41:16,668 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=439
2024-09-06 08:41:16,668 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-06 08:41:16,675 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job

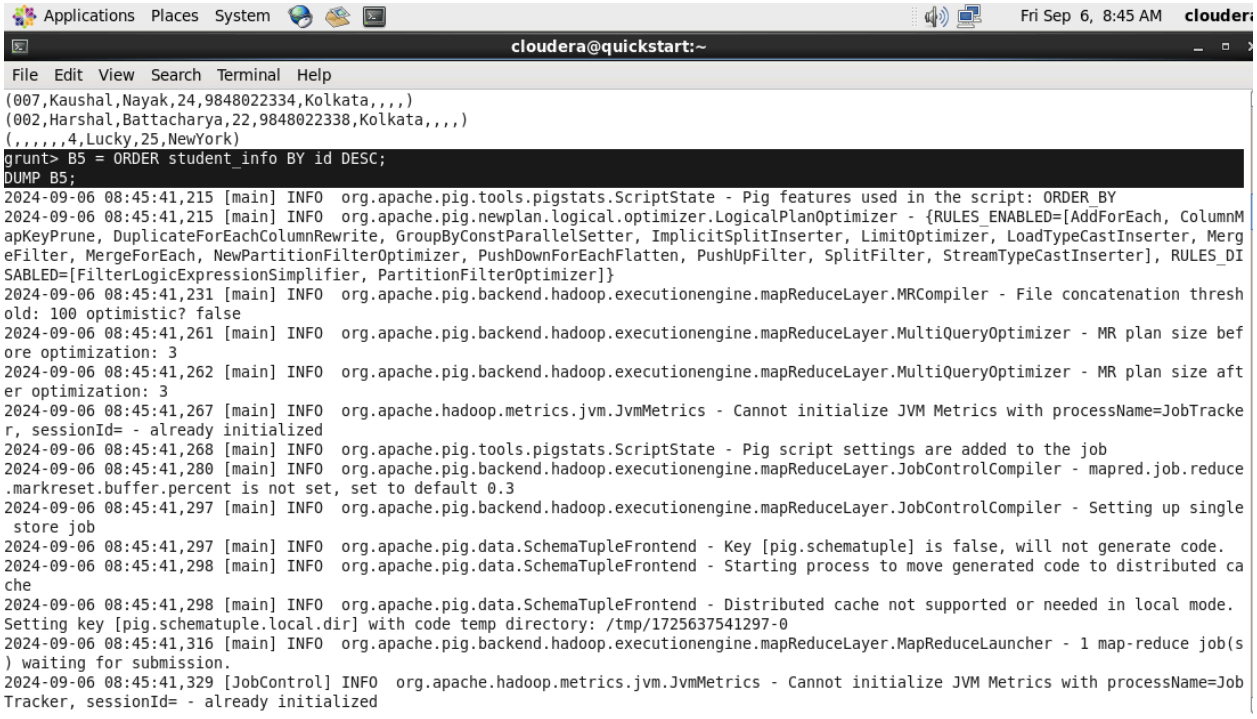

Successfully read records from: "/home/cloudera/Desktop/interns_data.txt"
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-751038793"
Job DAG:
job_local1974257114_0013
2024-09-06 08:41:37,874 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:41:37,874 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:41:37,874 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:41:37,874 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:41:37,875 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:41:37,898 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:41:37,898 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(004,Atharva,Agarwal,21,9848022330,Pune,,,,)
(005,Raj,Jain,23,9848022336,Mumbai,5,Pradeep,23,Mumbai)
(005,Raj,Jain,23,9848022336,Mumbai,1,Shreyas,22,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,5,Pradeep,23,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,1,Shreyas,22,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,5,Pradeep,23,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,1,Shreyas,22,Mumbai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,6,Baldev,22,Chennai)
(006,Shubham,Mishra,23,9848022335,Chennai,6,Baldev,22,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata,,,,)
(002,Harshal,Battacharya,22,9848022338,Kolkata,,,,)

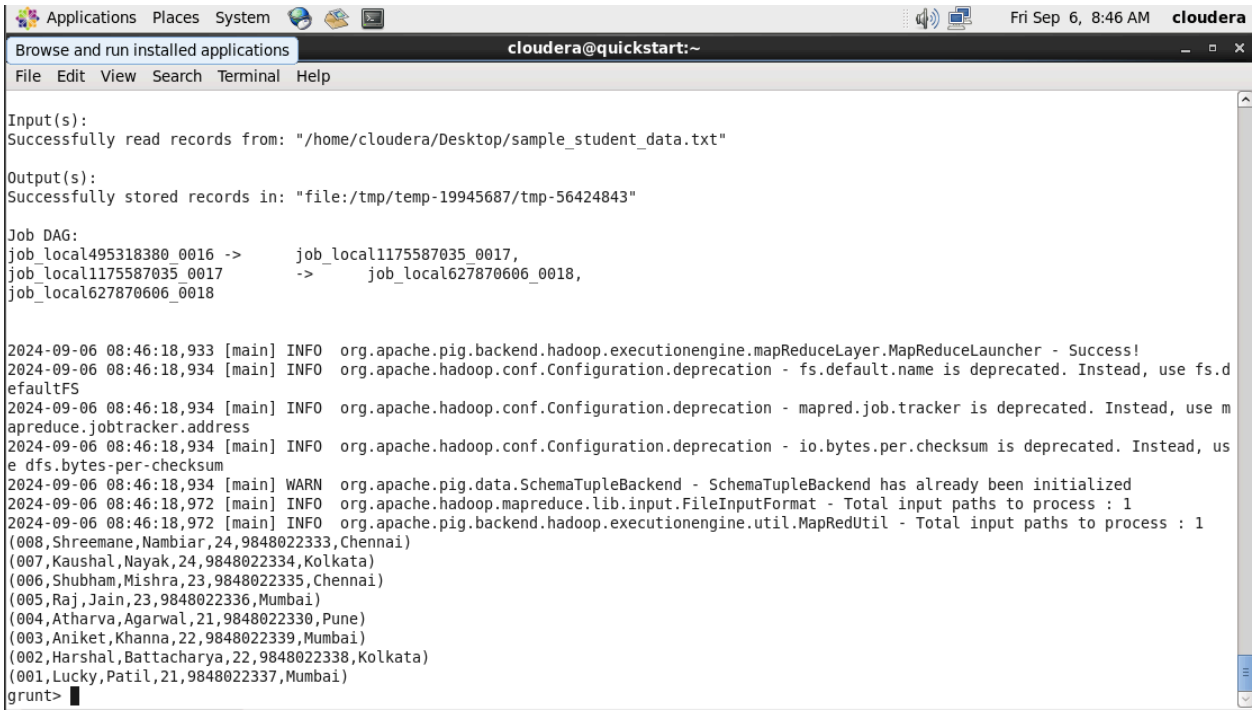

Applications Places System
Fri Sep 6, 8:43 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
(002,Harshal,Battacharya,22,9848022338,Kolkata,,,,)
grunt> RightJoin = JOIN student_info BY city RIGHT OUTER, I BY city;
DUMP RightJoin;
2024-09-06 08:43:08,191 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2024-09-06 08:43:08,192 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:43:08,203 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:43:08,204 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:43:08,204 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:43:08,205 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:43:08,205 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:43:08,215 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:43:08,216 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-06 08:43:08,216 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:43:08,225 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=439
2024-09-06 08:43:08,225 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-06 08:43:08,247 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:43:08,248 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:43:08,248 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed ca

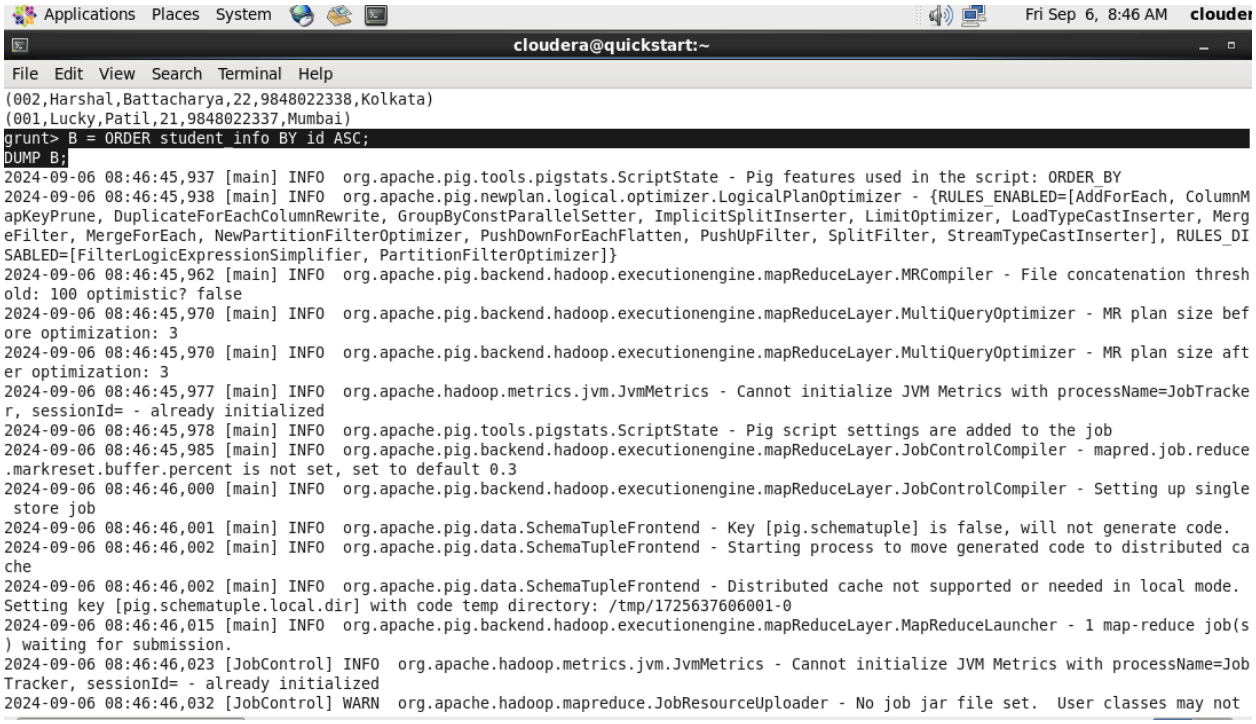

Applications Places System
Fri Sep 6, 8:43 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Successfully read records from: "/home/cloudera/Desktop/interns_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-1489754504"
Job DAG:
job_local697689685_0014
2024-09-06 08:43:27,116 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:43:27,116 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:43:27,117 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:43:27,117 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:43:27,117 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:43:27,143 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:43:27,144 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(,,,,,,2,Shreedhar,23,Delhi)
(,,,,,,3,Harshit,23,Tokyo)
(005,Raj,Jain,23,9848022336,Mumbai,5,Pradeep,23,Mumbai)
(005,Raj,Jain,23,9848022336,Mumbai,1,Shreyas,22,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,5,Pradeep,23,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,1,Shreyas,22,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,5,Pradeep,23,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,1,Shreyas,22,Mumbai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,6,Baldev,22,Chennai)
(006,Shubham,Mishra,23,9848022335,Chennai,6,Baldev,22,Chennai)
(,,,,4,Lucky,25,NewYork)

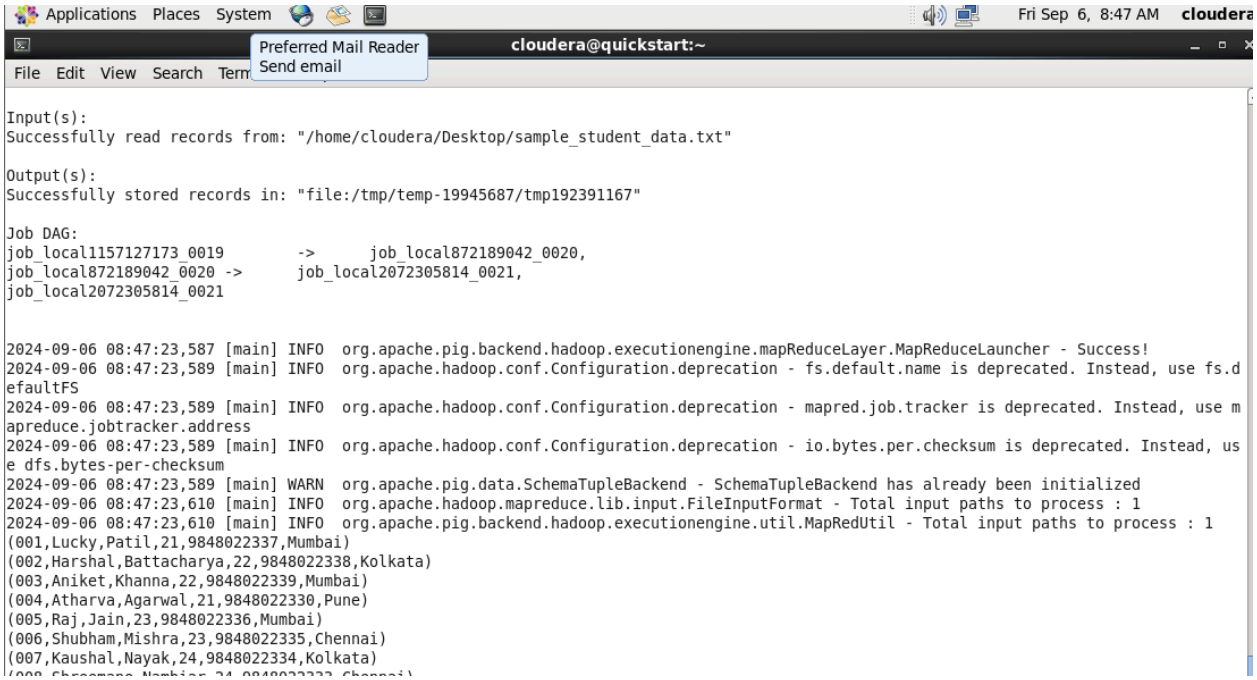

Applications Places System
Fri Sep 6, 8:44 AM cloudera
cloudera@quickstart:~
Change account settings and status
File Edit View Search Terminal Help
(,,,,,,4,Lucky,25,NewYork)
grunt> FullJoin = JOIN student_info BY city FULL OUTER, I BY city;
DUMP FullJoin;
2024-09-06 08:44:22,290 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: HASH_JOIN
2024-09-06 08:44:22,291 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:44:22,307 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:44:22,308 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:44:22,308 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:44:22,310 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:44:22,310 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:44:22,317 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:44:22,322 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-06 08:44:22,323 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:44:22,327 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=439
2024-09-06 08:44:22,327 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-06 08:44:22,338 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:44:22,343 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:44:22,343 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache

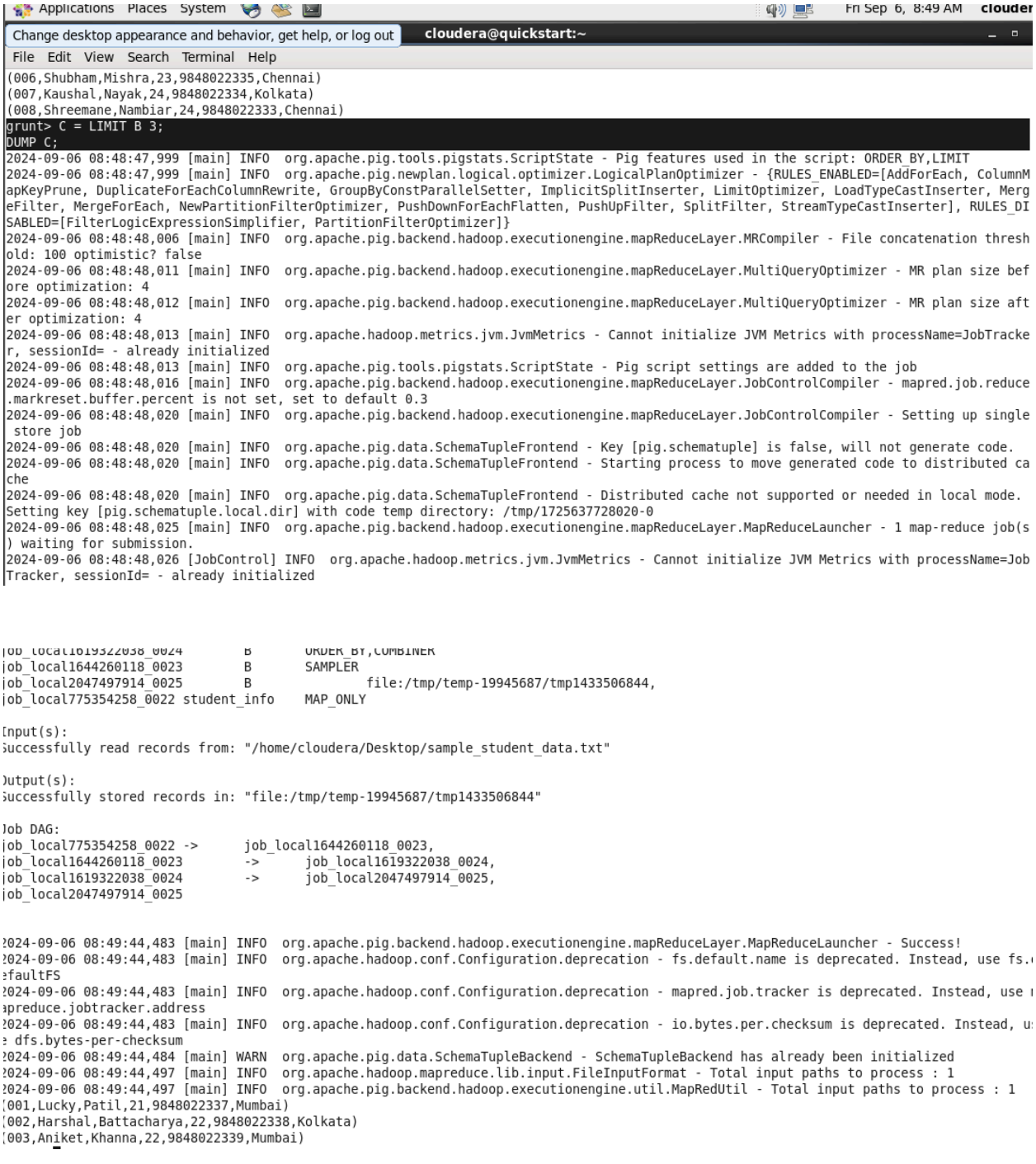

Applications Places System
Fri Sep 6, 8:45 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-491666410"
Job DAG:
job_local213834154_0015
2024-09-06 08:44:41,098 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:44:41,099 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:44:41,099 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:44:41,099 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:44:41,099 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:44:41,111 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:44:41,111 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(004,Atharva,Agarwal,21,9848022330,Pune,,,,)
(,,,,,,2,Shreedhar,23,Delhi)
(,,,,,,3,Harshit,23,Tokyo)
(005,Raj,Jain,23,9848022336,Mumbai,5,Pradeep,23,Mumbai)
(005,Raj,Jain,23,9848022336,Mumbai,1,Shreyas,22,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,5,Pradeep,23,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai,1,Shreyas,22,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,5,Pradeep,23,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai,1,Shreyas,22,Mumbai)
(008,Shreemane,Nambiar,24,9848022333,Chennai,6,Baldev,22,Chennai)
(006,Shubham,Mishra,23,9848022335,Chennai,6,Baldev,22,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata,,,,)
(002,Harshal,Battacharya,22,9848022338,Kolkata,,,,)
(,,,,,,4,Lucky,25,NewYork)
grunt>

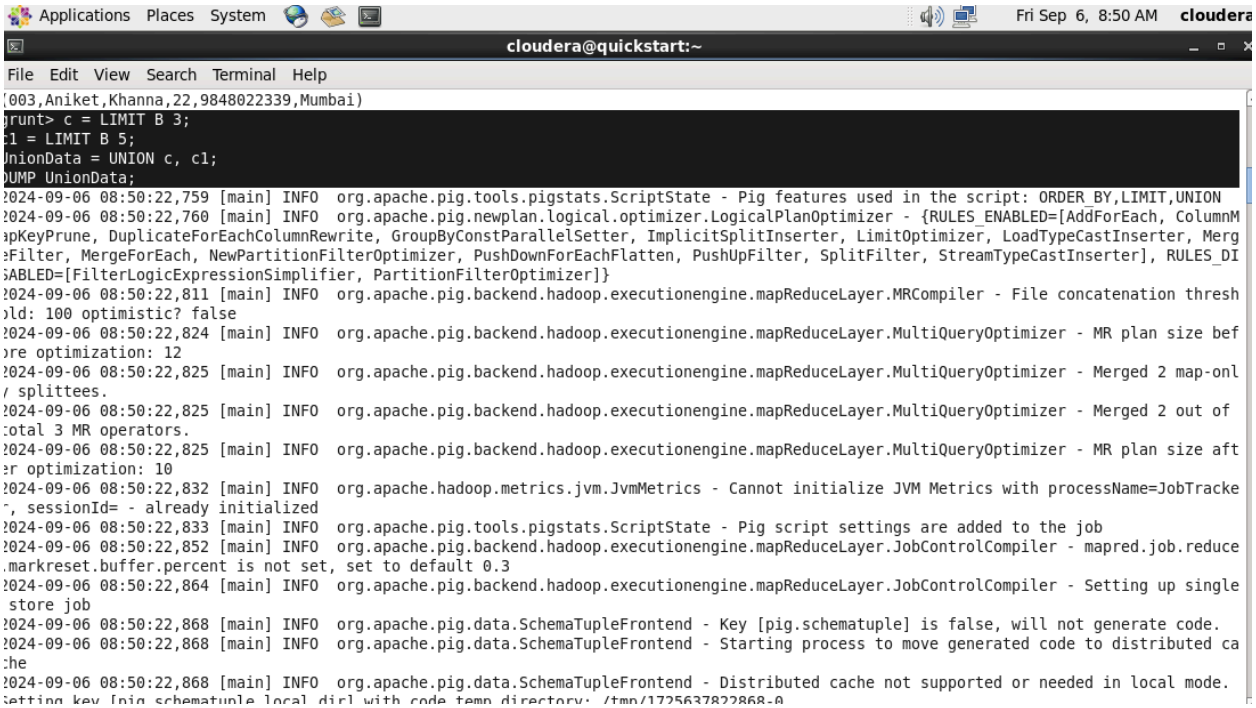

Applications Places System
Fri Sep 6, 8:45 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
(007,Kaushal,Nayak,24,9848022334,Kolkata,,,,)
(002,Harshal,Battacharya,22,9848022338,Kolkata,,,,)
(,,,,,,4,Lucky,25,NewYork)
grunt> B5 = ORDER student_info BY id DESC;
DUMP B5;
2024-09-06 08:45:41,215 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY
2024-09-06 08:45:41,215 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:45:41,231 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:45:41,261 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2024-09-06 08:45:41,262 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2024-09-06 08:45:41,267 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:45:41,268 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:45:41,280 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:45:41,297 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:45:41,297 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:45:41,298 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-09-06 08:45:41,298 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725637541297-0
2024-09-06 08:45:41,316 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2024-09-06 08:45:41,329 [JobControl] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

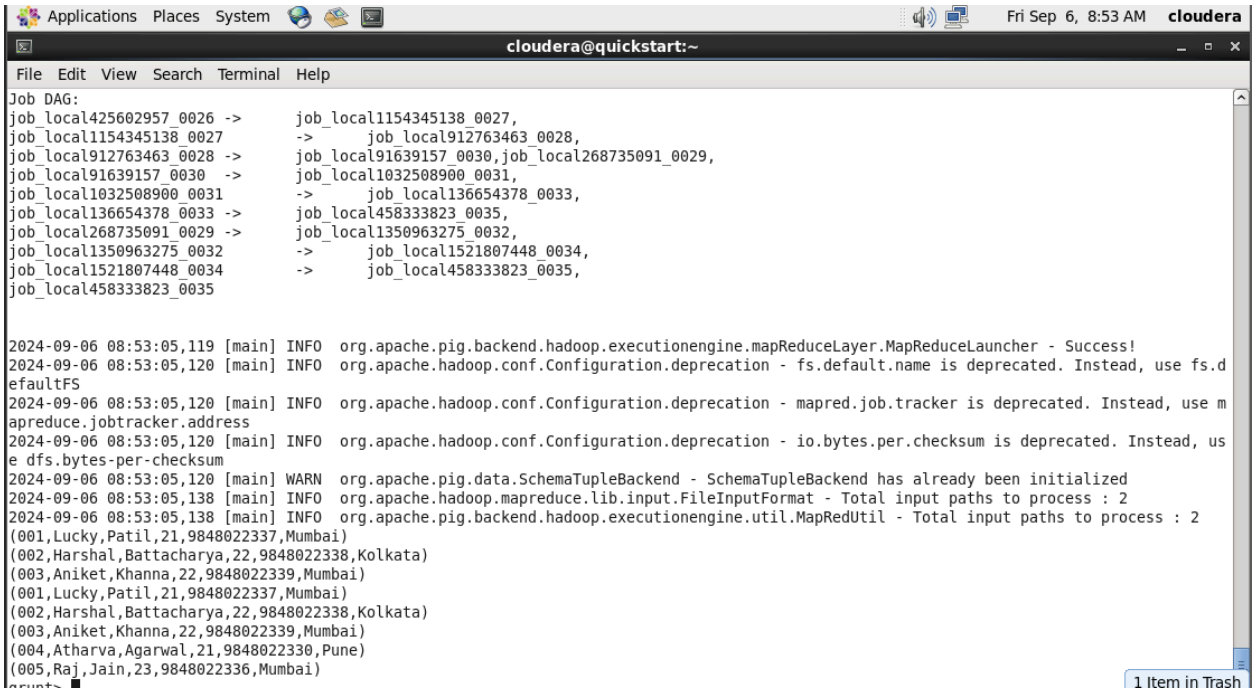

Applications Places System
Fri Sep 6, 8:46 AM cloudera
Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-56424843"
Job DAG:
job_local495318380_0016 -> job_local1175587035_0017,
job_local1175587035_0017 -> job_local627870606_0018,
job_local627870606_0018
2024-09-06 08:46:18,933 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:46:18,934 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:46:18,934 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:46:18,934 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:46:18,934 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:46:18,972 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:46:18,972 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(008,Shreemane,Nambiar,24,9848022333,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata)
(006,Shubham,Mishra,23,9848022335,Chennai)
(005,Raj,Jain,23,9848022336,Mumbai)
(004,Atharva,Agarwal,21,9848022330,Pune)
(003,Aniket,Khanna,22,9848022339,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(001,Lucky,Patil,21,9848022337,Mumbai)
grunt>

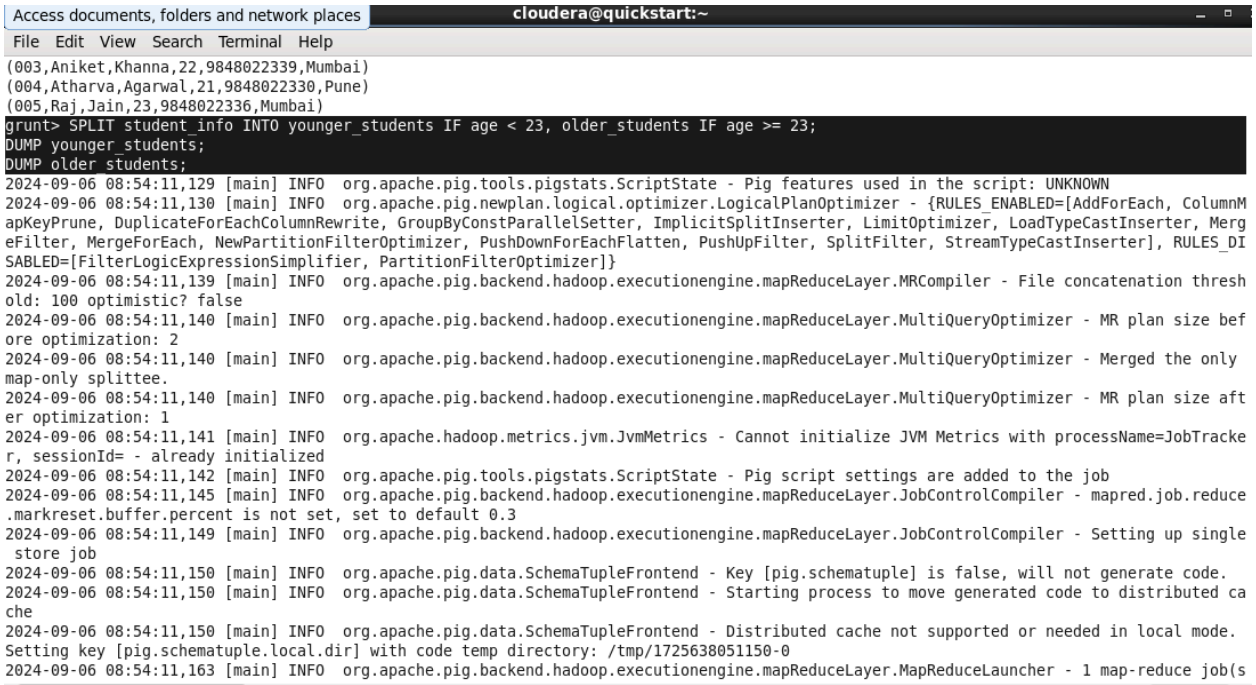

Applications Places System
Fri Sep 6, 8:46 AM
cloudera@quickstart:~
File Edit View Search Terminal Help
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(001,Lucky,Patil,21,9848022337,Mumbai)
grunt> B = ORDER student_info BY id ASC;
DUMP B;
2024-09-06 08:46:45,937 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY
2024-09-06 08:46:45,938 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:46:45,962 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:46:45,970 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2024-09-06 08:46:45,970 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2024-09-06 08:46:45,977 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:46:45,978 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:46:45,985 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:46:46,000 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:46:46,001 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:46:46,002 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-09-06 08:46:46,002 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725637606001-0
2024-09-06 08:46:46,015 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2024-09-06 08:46:46,023 [JobControl] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:46:46,032 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not

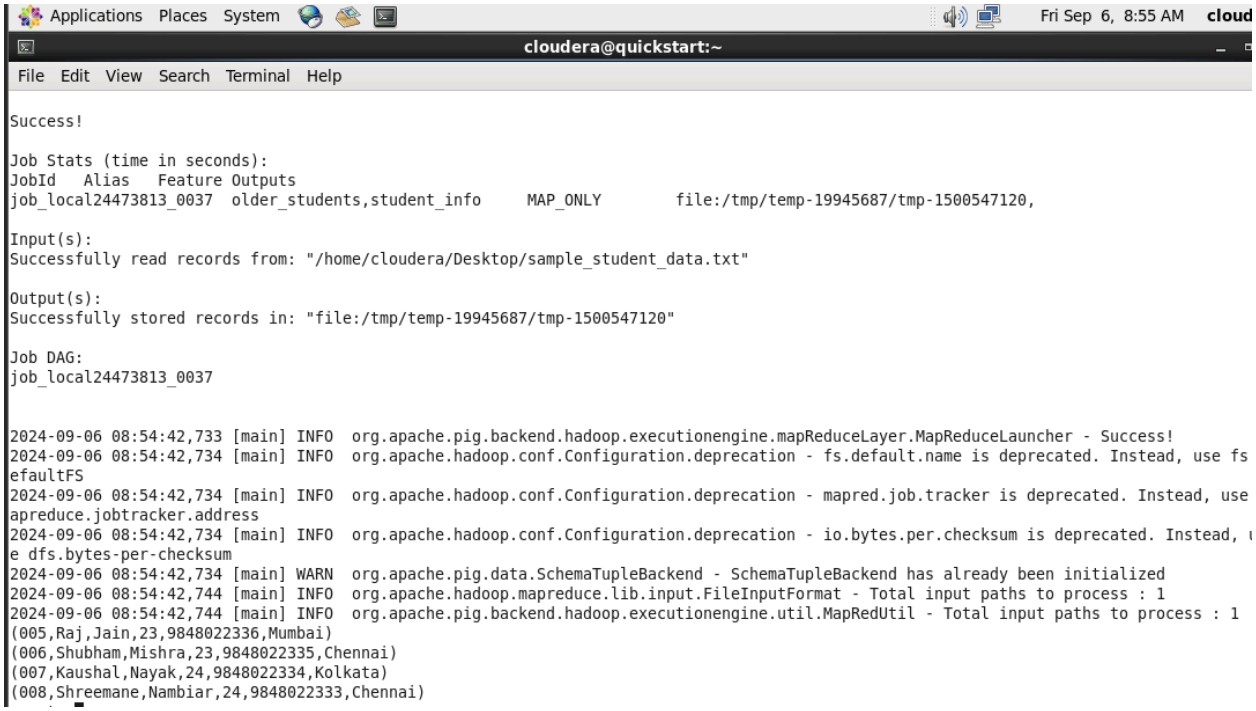

Applications Places System
Fri Sep 6, 8:47 AM
cloudera@quickstart:~
Preferred Mail Reader
Send email
File Edit View Search Term
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp192391167"
Job DAG:
job_local1157127173_0019 -> job_local872189042_0020,
job_local872189042_0020 -> job_local2072305814_0021,
job_local2072305814_0021
2024-09-06 08:47:23,587 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:47:23,589 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:47:23,589 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:47:23,589 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:47:23,589 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:47:23,610 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:47:23,610 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(001,Lucky,Patil,21,9848022337,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(003,Aniket,Khanna,22,9848022339,Mumbai)
(004,Atharva,Agarwal,21,9848022330,Pune)
(005,Raj,Jain,23,9848022336,Mumbai)
(006,Shubham,Mishra,23,9848022335,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata)

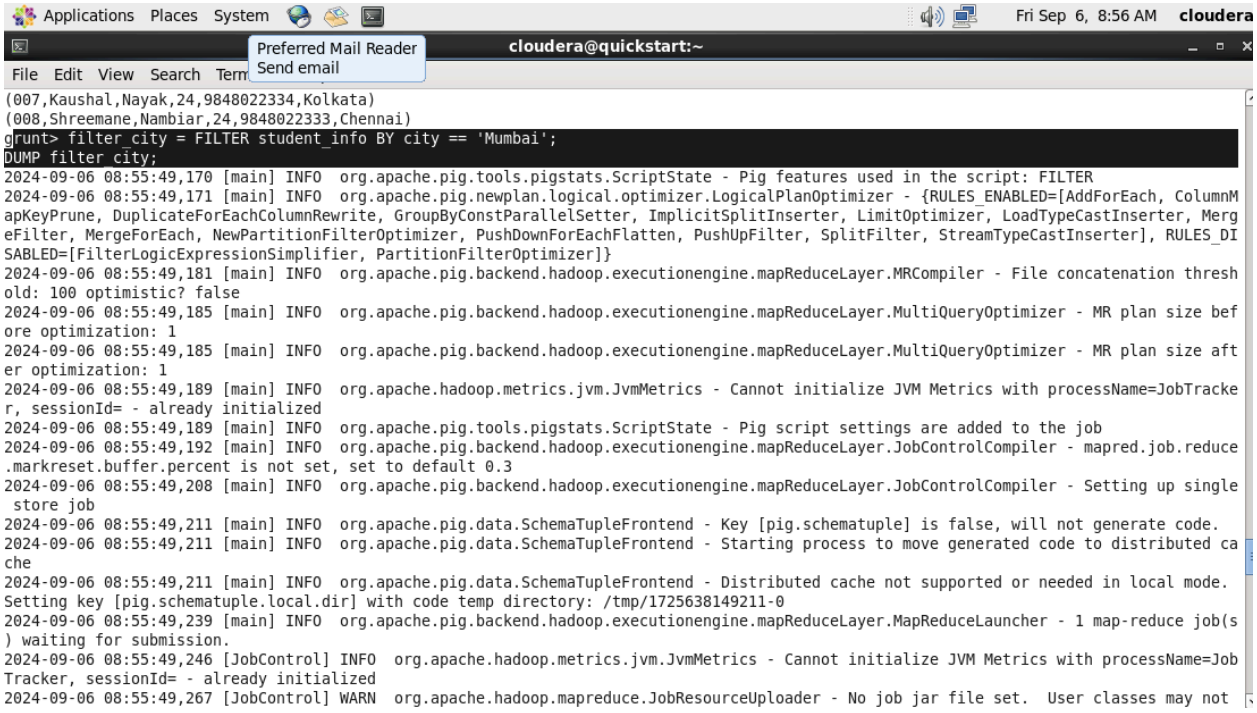

Applications Places System
Fri Sep 6, 8:49 AM cloudera
Change desktop appearance and behavior, get help, or log out
cloudera@quickstart:~
File Edit View Search Terminal Help
(006,Shubham,Mishra,23,9848022335,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata)
(008,Shreemane,Nambiar,24,9848022333,Chennai)
grunt> C = LIMIT B 3;
DUMP C;
2024-09-06 08:48:47,999 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,LIMIT
2024-09-06 08:48:47,999 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnM
apKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, Merg
eFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DI
SABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:48:48,006 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
old: 100 optimistic? false
2024-09-06 08:48:48,011 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
ore optimization: 4
2024-09-06 08:48:48,012 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
er optimization: 4
2024-09-06 08:48:48,013 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2024-09-06 08:48:48,013 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:48:48,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:48:48,020 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single
store job
2024-09-06 08:48:48,020 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:48:48,020 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed ca
che
2024-09-06 08:48:48,020 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode.
Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725637728020-0
2024-09-06 08:48:48,025 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s
) waiting for submission.
2024-09-06 08:48:48,026 [JobControl] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=Job
Tracker, sessionId= - already initialized
job_local1644260118_0023 B SAMPLER
job_local2047497914_0025 B file:/tmp/temp-19945687/tmp1433506844,
job_local775354258_0022 student_info MAP_ONLY
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp1433506844"
Job DAG:
job_local775354258_0022 -> job_local1644260118_0023,
job_local1644260118_0023 -> job_local1619322038_0024,
job_local1619322038_0024 -> job_local2047497914_0025,
job_local2047497914_0025
2024-09-06 08:49:44,483 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:49:44,483 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
efaultFS
2024-09-06 08:49:44,483 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
apreduce.jobtracker.address
2024-09-06 08:49:44,483 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
e dfs.bytes-per-checksum
2024-09-06 08:49:44,484 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:49:44,497 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:49:44,497 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(001,Lucky,Patil,21,9848022337,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(003,Aniket,Khanna,22,9848022339,Mumbai)

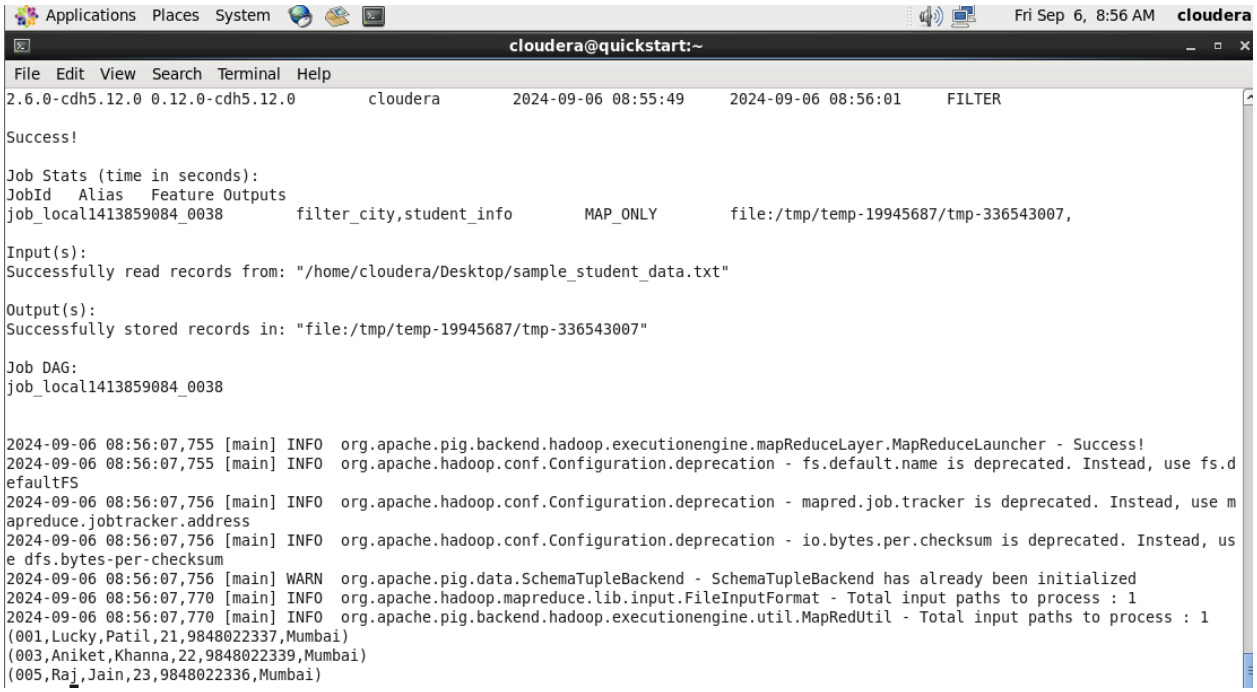

Applications Places System
Fri Sep 6, 8:50 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
(003,Aniket,Khanna,22,9848022339,Mumbai)
grunt> c = LIMIT B 3;
c1 = LIMIT B 5;
UnionData = UNION c, c1;
DUMP UnionData;
2024-09-06 08:50:22,759 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,LIMIT,UNION
2024-09-06 08:50:22,760 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnM
apKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, Merg
eFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DI
SABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:50:22,811 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
old: 100 optimistic? false
2024-09-06 08:50:22,824 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
ore optimization: 12
2024-09-06 08:50:22,825 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 2 map-onl
y splittees.
2024-09-06 08:50:22,825 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 2 out of
total 3 MR operators.
2024-09-06 08:50:22,825 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
er optimization: 10
2024-09-06 08:50:22,832 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2024-09-06 08:50:22,833 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:50:22,852 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:50:22,864 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single
store job
2024-09-06 08:50:22,868 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:50:22,868 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed ca
che
2024-09-06 08:50:22,868 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode.
Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725637822868-0

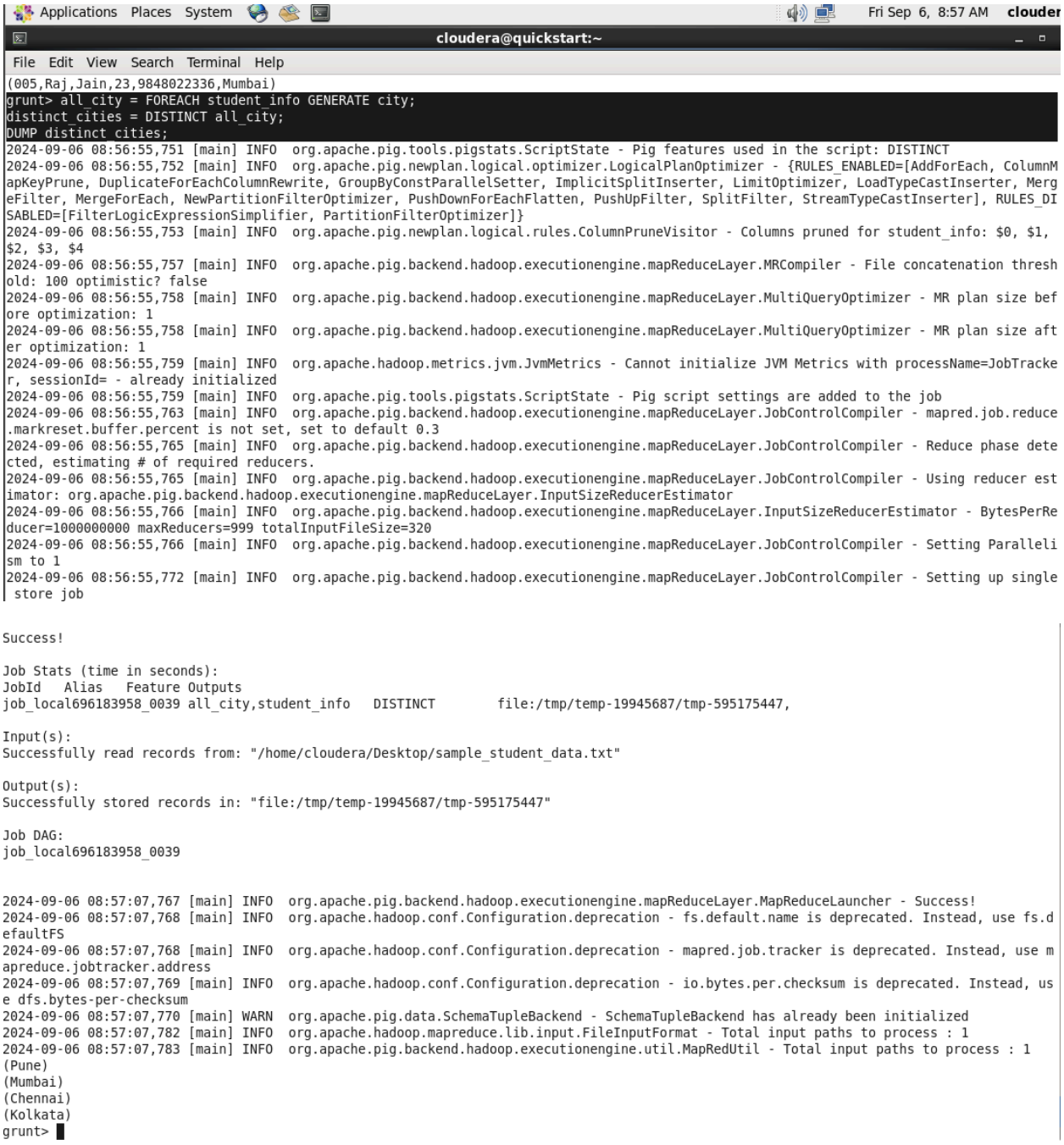

Applications Places System
Fri Sep 6, 8:53 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Job DAG:
job_local425602957_0026 -> job_local1154345138_0027,
job_local1154345138_0027 -> job_local912763463_0028,
job_local912763463_0028 -> job_local91639157_0030,job_local268735091_0029,
job_local91639157_0030 -> job_local1032508900_0031,
job_local1032508900_0031 -> job_local136654378_0033,
job_local136654378_0033 -> job_local458333823_0035,
job_local268735091_0029 -> job_local1350963275_0032,
job_local1350963275_0032 -> job_local1521807448_0034,
job_local1521807448_0034 -> job_local458333823_0035,
job_local458333823_0035
2024-09-06 08:53:05,119 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:53:05,120 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2024-09-06 08:53:05,120 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use m
apreduce.jobtracker.address
2024-09-06 08:53:05,120 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, us
e dfs.bytes-per-checksum
2024-09-06 08:53:05,120 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:53:05,138 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 2
2024-09-06 08:53:05,138 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 2
(001,Lucky,Patil,21,9848022337,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(003,Aniket,Khanna,22,9848022339,Mumbai)
(001,Lucky,Patil,21,9848022337,Mumbai)
(002,Harshal,Battacharya,22,9848022338,Kolkata)
(003,Aniket,Khanna,22,9848022339,Mumbai)
(004,Atharva,Agarwal,21,9848022330,Pune)
(005,Raj,Jain,23,9848022336,Mumbai)
grunt>
1 Item in Trash


Access documents, folders and network places
cloudera@quickstart:~
File Edit View Search Terminal Help
(003,Aniket,Khanna,22,9848022339,Mumbai)
(004,Atharva,Agarwal,21,9848022330,Pune)
(005,Raj,Jain,23,9848022336,Mumbai)
grunt> SPLIT student_info INTO younger_students IF age < 23, older_students IF age >= 23;
DUMP younger_students;
DUMP older_students;
2024-09-06 08:54:11,129 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-06 08:54:11,130 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:54:11,139 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:54:11,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2024-09-06 08:54:11,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged the only map-only splittee.
2024-09-06 08:54:11,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:54:11,141 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:54:11,142 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:54:11,145 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:54:11,149 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:54:11,150 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:54:11,150 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-09-06 08:54:11,150 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725638051150-0
2024-09-06 08:54:11,163 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s


Applications Places System
Fri Sep 6, 8:55 AM cloud
cloudera@quickstart:~
File Edit View Search Terminal Help
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local24473813_0037 older_students,student_info MAP_ONLY file:/tmp/temp-19945687/tmp-1500547120,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-1500547120"
Job DAG:
job_local24473813_0037
2024-09-06 08:54:42,733 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:54:42,734 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:54:42,734 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:54:42,734 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:54:42,734 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:54:42,744 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:54:42,744 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(005,Raj,Jain,23,9848022336,Mumbai)
(006,Shubham,Mishra,23,9848022335,Chennai)
(007,Kaushal,Nayak,24,9848022334,Kolkata)
(008,Shreemane,Nambiar,24,9848022333,Chennai)


Applications Places System
Fri Sep 6, 8:56 AM cloudera
Preferred Mail Reader
Send email
cloudera@quickstart:~
File Edit View Search Term
(007,Kaushal,Nayak,24,9848022334,Kolkata)
(008,Shreemane,Nambiar,24,9848022333,Chennai)
grunt> filter_city = FILTER student_info BY city == 'Mumbai';
DUMP filter_city;
2024-09-06 08:55:49,170 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2024-09-06 08:55:49,171 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:55:49,181 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:55:49,185 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:55:49,185 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:55:49,189 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:55:49,189 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:55:49,192 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:55:49,208 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-09-06 08:55:49,211 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-09-06 08:55:49,211 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-09-06 08:55:49,211 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1725638149211-0
2024-09-06 08:55:49,239 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2024-09-06 08:55:49,246 [JobControl] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:55:49,267 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not


Applications Places System
Fri Sep 6, 8:56 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0 cloudera 2024-09-06 08:55:49 2024-09-06 08:56:01 FILTER
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1413859084_0038 filter_city,student_info MAP_ONLY file:/tmp/temp-19945687/tmp-336543007,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-336543007"
Job DAG:
job_local1413859084_0038
2024-09-06 08:56:07,755 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:56:07,755 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:56:07,756 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:56:07,756 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:56:07,756 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:56:07,770 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:56:07,770 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(001,Lucky,Patil,21,9848022337,Mumbai)
(003,Aniket,Khanna,22,9848022339,Mumbai)
(005,Raj,Jain,23,9848022336,Mumbai)


Applications Places System
Fri Sep 6, 8:57 AM clouder
cloudera@quickstart:~
File Edit View Search Terminal Help
(005,Raj,Jain,23,9848022336,Mumbai)
grunt> all_city = FOREACH student_info GENERATE city;
distinct_cities = DISTINCT all_city;
DUMP distinct_cities;
2024-09-06 08:56:55,751 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: DISTINCT
2024-09-06 08:56:55,752 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-09-06 08:56:55,753 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for student_info: $0, $1, $2, $3, $4
2024-09-06 08:56:55,757 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-06 08:56:55,758 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-06 08:56:55,758 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-09-06 08:56:55,759 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2024-09-06 08:56:55,759 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-09-06 08:56:55,763 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-09-06 08:56:55,765 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2024-09-06 08:56:55,765 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2024-09-06 08:56:55,766 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=320
2024-09-06 08:56:55,766 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2024-09-06 08:56:55,772 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local696183958_0039 all_city,student_info DISTINCT file:/tmp/temp-19945687/tmp-595175447,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/sample_student_data.txt"
Output(s):
Successfully stored records in: "file:/tmp/temp-19945687/tmp-595175447"
Job DAG:
job_local696183958_0039
2024-09-06 08:57:07,767 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-06 08:57:07,768 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-06 08:57:07,768 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-06 08:57:07,769 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-09-06 08:57:07,770 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-06 08:57:07,782 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-09-06 08:57:07,783 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Pune)
(Mumbai)
(Chennai)
(Kolkata)
grunt>

Conclusion-

In this assignment, we successfully created a Pig Data Model to efficiently process large datasets using Pig Latin. We explored Pig's diverse data types, including simple types like int and chararray as well as complex types like tuples and bags. After downloading and preparing a sample dataset, we designed a Pig script to load, transform, and analyze the data. By executing the script, we performed various Pig operations, such as filtering, grouping, joining, and sorting the data. Additionally, we utilized diagnostic operators to better understand the data structure and execution plan, gaining a comprehensive understanding of Pig's capabilities for big data processing. This assignment solidified our practical knowledge of Pig as a powerful tool for large-scale data analytic