Big Data Analytics and Visualization Lab

Assignment 5:

Pig Latin Basics

1. Write a Pig command to display all employee data in ascending order.Command:

[cloudera@quickstart ~]$ cat /home/cloudera/employee_data.txt

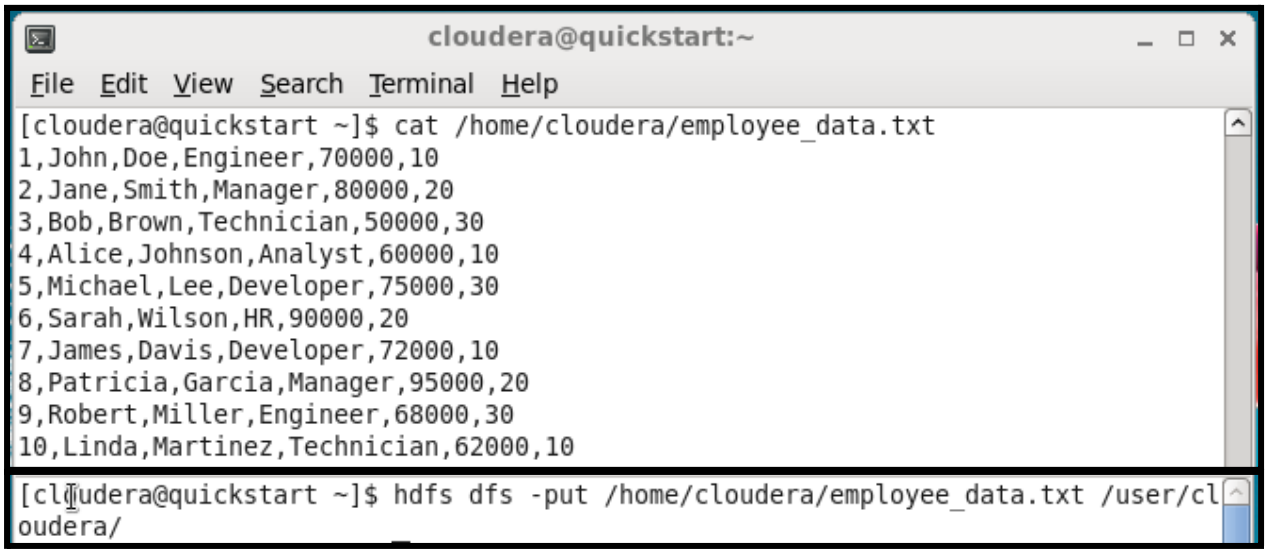
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/employee_data.txt /user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

[cloudera@quickstart ~]$ pig

grunt> employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int,emp_fname:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);

grunt> sorted_employees = ORDER employees BY emp_id ASC;grunt> DUMP sorted_employees;

Output:



```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 4 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-13 09:58 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
```

```
2024-10-20 04:53:42,898 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 04:53:43,097 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 04:53:43,180 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

```
grunt> employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int
, emp_fname:chararray, emp_lname:chararray, job:chararray, salary:float, deptcod
e:int);
grunt> sorted_employees = ORDER employees BY emp_id ASC;
grunt> DUMP sorted_employees;
2024-10-20 04:57:31,339 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: ORDER_BY
```

```
2024-10-20 05:04:27,529 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,John,Doe,Engineer,70000.0,10)
(2,Jane,Smith,Manager,80000.0,20)
(3,Bob,Brown,Technician,50000.0,30)
(4,Alice,Johnson,Analyst,60000.0,10)
(5,Michael,Lee,Developer,75000.0,30)
(6,Sarah,Wilson,HR,90000.0,20)
(7,James,Davis,Developer,72000.0,10)
(8,Patricia,Garcia,Manager,95000.0,20)
(9,Robert,Miller,Engineer,68000.0,30)
(10,Linda,Martinez,Technician,62000.0,10)
```

2. Write a Pig Latin script to display the full name of each employee and thefirst three characters of their job title.

Pig Latin Script:

> **full_names.pig**

```
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS
(emp_id:int,emp_fname:chararray, emp_lname:chararray, job:chararray,
salary:float, deptcode:int);

-- Create full name and extract first three characters of job
full_names = FOREACH employees GENERATE CONCAT(emp_fname, ' ',
emp_lname) ASfull_name, SUBSTRING(job, 0, 3) AS job_prefix;

-- Dump results to
displayDUMP
full_names;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/full_names.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/full_names.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec full_names.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/full_names.pig
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int, emp_f
name:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);

-- Create full name and extract first three characters of job
full_names = FOREACH employees GENERATE CONCAT(emp_fname, ' ', emp_lname) AS ful
l_name, SUBSTRING(job, 0, 3) AS job_prefix;

-- Dump results to display
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/full_names.pig /user/cloud
era/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 5 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 05:04 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
```

```
grunt> exec full_names.pig
2024-10-20 05:23:43,833 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 05:23:44,694 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2024-10-20 05:23:44,697 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
```

```
nputFormat - Total input paths to process : 1
2024-10-20 05:25:28,995 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(John Doe,Eng)
(Jane Smith,Man)
(Bob Brown,Tec)
(Alice Johnson,Ana)
(Michael Lee,Dev)
(Sarah Wilson,HR)
(James Davis,Dev)
(Patricia Garcia,Man)
(Robert Miller,Eng)
(Linda Martinez,Tec)
```

3. Write a Pig Latin script to split the students file into two parts: batchA andbatchB, based on rollno (use the students.txt file).

Pig Latin Script:

> **batch.pig**

```
-- Load student data
students =   LOAD   'students.txt' USING   PigStorage(',') AS   (roll:int,
            fname:chararray,lname:chararray, gender:chararray, program:chararray,
specialization:chararray);

-- Split into batchA (roll < 100) and batchB (roll >=
100)batchA = FILTER students BY roll < 100;
batchB = FILTER students BY roll >= 100;

-- Dump results to
displayDUMP batchA;
DUMP batchB;

-- Store results
STORE batchA INTO 'batchA' USING
PigStorage(',');STORE batchB INTO 'batchB'
USING PigStorage(',');
```

Command:

[cloudera@quickstart ~]$ cat

/home/cloudera/students.txt[cloudera@quickstart

~]$ cat /home/cloudera/batch.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/students.txt

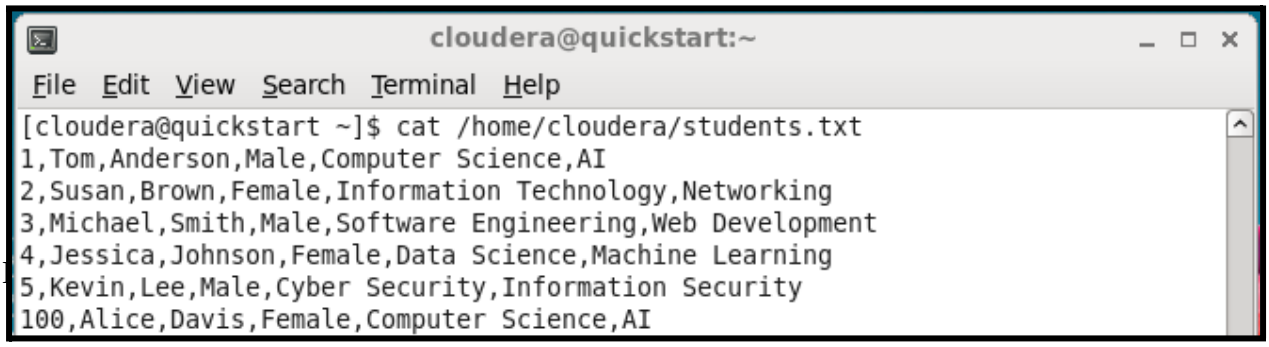
/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -put

/home/cloudera/batch.pig /user/cloudera/ [cloudera@quickstart ~]$ hdfs

dfs -ls /user/cloudera/

grunt> exec batch.pig

Output:



```
100,Alice,Davis,Female,Computer Science,AI
101,Bob,White,Male,Information Technology,Networking
6,Emily,Clark,Female,Data Science,Data Analysis
7,Matthew,Robinson,Male,Computer Science,Cloud Computing
8,Ava,Young,Female,Information Technology,Web Development
9,Oliver,King,Male,Software Engineering,Mobile Development
10,Sophia,Wright,Female,Cyber Security,Forensics
102,Liam,Hall,Male,Information Technology,DevOps
103,Isabella,Lopez,Female,Computer Science,AI
104,Noah,Scott,Male,Data Science,Big Data
105,Emma,Adams,Female,Software Engineering,Quality Assurance
106,Lucas,Baker,Male,Cyber Security,Penetration Testing
```

```
[cloudera@quickstart ~]$ cat /home/cloudera/batch.pig
-- Load student data
students = LOAD 'students.txt' USING PigStorage(',') AS (roll:int, fname:chararr
ay, lname:chararray, gender:chararray, program:chararray, specialization:chararr
ay);

-- Split into batchA (roll < 100) and batchB (roll >= 100)
batchA = FILTER students BY roll < 100;
batchB = FILTER students BY roll >= 100;

-- Dump results to display
DUMP batchA;
DUMP batchB;

-- Store results
STORE batchA INTO 'batchA' USING PigStorage(',');
STORE batchB INTO 'batchB' USING PigStorage(',');
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/students.txt /user/clouder
a/
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/batch.pig /user/cloudera/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 7 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 05:36 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        508 2024-10-20 05:56 /user/cloudera/batc
h.pig
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
```

```
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
-rw-r--r--   1 cloudera cloudera        899 2024-10-20 05:28 /user/cloudera/stud
ents.txt
```

```
grunt> exec batch.pig
2024-10-20 06:00:07,761 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 06:00:08,584 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: FILTER
2024-10-20 06:00:08,588 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
```

```
2024-10-20 06:01:51,434 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,Tom,Anderson,Male,Computer Science,AI)
(2,Susan,Brown,Female,Information Technology,Networking)
(3,Michael,Smith,Male,Software Engineering,Web Development)
(4,Jessica,Johnson,Female,Data Science,Machine Learning)
(5,Kevin,Lee,Male,Cyber Security,Information Security)
(6,Emily,Clark,Female,Data Science,Data Analysis)
(7,Matthew,Robinson,Male,Computer Science,Cloud Computing)
(8,Ava,Young,Female,Information Technology,Web Development)
(9,Oliver,King,Male,Software Engineering,Mobile Development)
(10,Sophia,Wright,Female,Cyber Security,Forensics)
2024-10-20 06:01:51,662 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: FILTER
```

```
2024-10-20 06:03:39,122 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(100,Alice,Davis,Female,Computer Science,AI)
(101,Bob,White,Male,Information Technology,Networking)
(102,Liam,Hall,Male,Information Technology,DevOps)
(103,Isabella,Lopez,Female,Computer Science,AI)
(104,Noah,Scott,Male,Data Science,Big Data)
(105,Emma,Adams,Female,Software Engineering,Quality Assurance)
(106,Lucas,Baker,Male,Cyber Security,Penetration Testing)
2024-10-20 06:03:39,702 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: FILTER
```

```
Input(s):
Successfully read 17 records (1276 bytes) from: "hdfs://quickstart.cloudera:8020
/user/cloudera/students.txt"

Output(s):
Successfully stored 10 records (532 bytes) in: "hdfs://quickstart.cloudera:8020/
user/cloudera/batchA"
Successfully stored 7 records (350 bytes) in: "hdfs://quickstart.cloudera:8020/u
ser/cloudera/batchB"

Counters:
Total records written : 17
```

```
Counters:
Total records written : 17
Total bytes written : 882
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1729424705485_0009


2024-10-20 06:05:22,198 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
```

4. Write a Pig Latin script to display the union of batchA and batchB.Pig Latin Script:

> **union_batches.pig**

```
-- Load batch data
batchA = LOAD 'batchA' USING PigStorage(',') AS (roll:int, fname:chararray,
lname:chararray,gender:chararray, program:chararray, specialization:chararray);
batchB = LOAD 'batchB' USING PigStorage(',') AS (roll:int, fname:chararray,
lname:chararray,gender:chararray, program:chararray, specialization:chararray);

-- Union of batchA and batchB
union_batches = UNION batchA,
batchB;

-- Dump results to
displayDUMP
union_batches;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/union_batches.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/union_batches.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec union_batches.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/union_batches.pig
-- Load batch data
batchA = LOAD 'batchA' USING PigStorage(',') AS (roll:int, fname:chararray, lnam
e:chararray, gender:chararray, program:chararray, specialization:chararray);
batchB = LOAD 'batchB' USING PigStorage(',') AS (roll:int, fname:chararray, lnam
e:chararray, gender:chararray, program:chararray, specialization:chararray);

-- Union of batchA and batchB
union_batches = UNION batchA, batchB;

-- Dump results to display
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/union_batches.pig /user/cl
oudera/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 10 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
```

```
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        508 2024-10-20 05:56 /user/cloudera/batc
h.pig
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hA
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hB
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
-rw-r--r--   1 cloudera cloudera        899 2024-10-20 05:28 /user/cloudera/stud
ents.txt
-rw-r--r--   1 cloudera cloudera        458 2024-10-20 06:02 /user/cloudera/unio
n_batches.pig
```

```
grunt> exec union_batches.pig
2024-10-20 06:17:47,861 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 06:17:47,971 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 06:17:49,561 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNION
```

```
2024-10-20 06:19:43,270 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 2
(1,Tom,Anderson,Male,Computer Science,AI)
(2,Susan,Brown,Female,Information Technology,Networking)
(3,Michael,Smith,Male,Software Engineering,Web Development)
(4,Jessica,Johnson,Female,Data Science,Machine Learning)
(5,Kevin,Lee,Male,Cyber Security,Information Security)
(6,Emily,Clark,Female,Data Science,Data Analysis)
(7,Matthew,Robinson,Male,Computer Science,Cloud Computing)
(8,Ava,Young,Female,Information Technology,Web Development)
(9,Oliver,King,Male,Software Engineering,Mobile Development)
(10,Sophia,Wright,Female,Cyber Security,Forensics)
(100,Alice,Davis,Female,Computer Science,AI)
(101,Bob,White,Male,Information Technology,Networking)
(102,Liam,Hall,Male,Information Technology,DevOps)
(103,Isabella,Lopez,Female,Computer Science,AI)
(104,Noah,Scott,Male,Data Science,Big Data)
(105,Emma,Adams,Female,Software Engineering,Quality Assurance)
(106,Lucas,Baker,Male,Cyber Security,Penetration Testing)
```

5. Write a Pig Latin script to check the total number of rows in the employeedata.

Pig Latin Script:

> **total_count.pig**

```
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS
(emp_id:int,emp_fname:chararray, emp_lname:chararray, job:chararray,
salary:float, deptcode:int);

-- Group all the data into one group (to count the total number of
rows)grouped_data = GROUP employees ALL;

-- Count the number of rows
total_count = FOREACH grouped_data GENERATE COUNT(employees);

-- Dump the result to display the total number of
rowsDUMP total_count;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/total_count.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/total_count.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec total_count.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/total_count.pig
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int, emp_f
name:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);

-- Group all the data into one group (to count the total number of rows)
grouped_data = GROUP employees ALL;

-- Count the number of rows
total_count = FOREACH grouped_data GENERATE COUNT(employees);

-- Dump the result to display the total number of rows
DUMP total_count;
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/total_count.pig /user/clou
dera/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 11 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 06:19 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        508 2024-10-20 05:56 /user/cloudera/batc
h.pig
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hA
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hB
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
-rw-r--r--   1 cloudera cloudera        899 2024-10-20 05:28 /user/cloudera/stud
ents.txt
-rw-r--r--   1 cloudera cloudera        458 2024-10-20 06:31 /user/cloudera/tota
l_count.pig
-rw-r--r--   1 cloudera cloudera        458 2024-10-20 06:02 /user/cloudera/unio
n_batches.pig
```

```
grunt> exec total_count.pig
2024-10-20 06:31:56,808 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 06:31:57,729 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY
```

```
2024-10-20 06:35:06,706 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(10)
```

6. Write a Pig Latin script to find the department-wise maximum salary (usethe employee_data.txt structure).

Pig Latin Script:

> **max_salary.pig**

```
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS
(emp_id:int,emp_fname:chararray, emp_lname:chararray, job:chararray,
salary:float, deptcode:int);

-- Group by deptcode and find maximum
salarygrouped = GROUP employees BY
deptcode;
max_salary = FOREACH grouped GENERATE group AS deptcode,
MAX(employees.salary)AS max_salary;

-- Dump results to
displayDUMP
max_salary;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/max_salary.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/max_salary.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec max_salary.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/max_salary.pig
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int, emp_f
name:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);

-- Group by deptcode and find maximum salary
grouped = GROUP employees BY deptcode;
max_salary = FOREACH grouped GENERATE group AS deptcode, MAX(employees.salary) A
S max_salary;
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/max_salary.pig /user/cloud
era/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 12 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 06:35 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        508 2024-10-20 05:56 /user/cloudera/batc
h.pig
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hA
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hB
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
-rw-r--r--   1 cloudera cloudera        415 2024-10-20 06:43 /user/cloudera/max_
salary.pig
-rw-r--r--   1 cloudera cloudera        899 2024-10-20 05:28 /user/cloudera/stud
ents.txt
```

```
grunt> exec max_salary.pig
2024-10-20 06:47:40,703 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 06:47:41,595 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY
```

```
2024-10-20 06:50:00,118 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(10,72000.0)
(20,95000.0)
(30,75000.0)
```

7. Write a Pig Latin script to find the employee details for those in department30.

Pig Latin Script:

> **dept30_employees.pig**

```
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS
(emp_id:int,emp_fname:chararray, emp_lname:chararray, job:chararray,
salary:float, deptcode:int);

-- Filter employees in department 30
dept30_employees = FILTER employees BY deptcode == 30;

-- Dump results to
display DUMP
dept30_employees;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/dept30_employees.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/dept30_employees.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec dept30_employees.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/dept30_employees.pig
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int, emp_f
name:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);

-- Filter employees in department 30
dept30_employees = FILTER employees BY deptcode == 30;

-- Dump results to display
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/dept30_employees.pig /user
/cloudera/
```

```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 13 items
```

```
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 06:49 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
-rw-r--r--   1 cloudera cloudera        508 2024-10-20 05:56 /user/cloudera/batc
h.pig
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hA
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hB
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 07:01 /user/cloudera/dept
30_employees.pig
```

```
grunt> exec dept30_employees.pig
2024-10-20 07:05:07,766 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 07:05:13,803 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: FILTER
2024-10-20 07:05:14,301 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
```

```
2024-10-20 07:07:32,409 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(3,Bob,Brown,Technician,50000.0,30)
(5,Michael,Lee,Developer,75000.0,30)
(9,Robert,Miller,Engineer,68000.0,30)
```

8. Write a Pig Latin script to display the first 5 rows of the employee data.Pig Latin Script:

> **first_five.pig**

```
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS
(emp_id:int,emp_fname:chararray, emp_lname:chararray, job:chararray,
salary:float, deptcode:int);

-- Limit to first 5 rows
first_five = LIMIT employees 5;

-- Dump results to
displayDUMP first_five;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/first_five.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/first_five.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec first_five.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/first_five.pig
-- Load employee data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int, emp_f
name:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);

-- Limit to first 5 rows
first_five = LIMIT employees 5;

-- Dump results to display
DUMP first_five;
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/first_five.pig /user/cloud
era/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 14 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
drwx------   - cloudera cloudera          0 2024-10-20 07:07 /user/cloudera/.sta
ging
-rw-r--r--   1 cloudera cloudera         62 2024-07-31 08:28 /user/cloudera/14Ak
ashChoudhary.txt
```

```
-rw-r--r--   1 cloudera cloudera        508 2024-10-20 05:56 /user/cloudera/batc
h.pig
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hA
drwxr-xr-x   - cloudera cloudera          0 2024-10-20 06:05 /user/cloudera/batc
hB
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 07:01 /user/cloudera/dept
30_employees.pig
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        293 2024-10-20 07:13 /user/cloudera/firs
t_five.pig
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
```

```
grunt> exec first_five.pig
2024-10-20 07:14:44,511 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 07:14:45,015 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: LIMIT
2024-10-20 07:14:45,016 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
```

```
2024-10-20 07:19:02,900 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,John,Doe,Engineer,70000.0,10)
(2,Jane,Smith,Manager,80000.0,20)
(3,Bob,Brown,Technician,50000.0,30)
(4,Alice,Johnson,Analyst,60000.0,10)
(5,Michael,Lee,Developer,75000.0,30)
```

9. Write a Pig Latin script in the file WordCount.pig and execute the script toget the required output.

Pig Latin Script:

> **word_count.pig**

```
-- Load the input data
data = LOAD 'input.txt' USING PigStorage('\n') AS (line:chararray);

-- Split each line into words
words = FOREACH data GENERATE FLATTEN(TOKENIZE(line)) AS word;

-- Remove any null or empty words
filtered_words = FILTER words BY word IS NOT NULL AND word != '';

-- Group the words
grouped_words = GROUP filtered_words BY word;

-- Count the occurrences of each word
word_count = FOREACH grouped_words GENERATE group AS
word,COUNT(filtered_words) AS count;

-- Display the results directly in the
terminalDUMP word_count;
```

Command:

[cloudera@quickstart ~]$ cat /home/cloudera/input.txt

[cloudera@quickstart ~]$ cat

/home/cloudera/word_count.pig

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/input.txt /user/cloudera/

[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/word_count.pig

/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/

grunt> exec word_count.pig

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/input.txt
hello world
hello big data
hello pig
pig is great for big data
big data is the future
```

```
[cloudera@quickstart ~]$ cat /home/cloudera/word_count.pig
-- Load the input data
data = LOAD 'input.txt' USING PigStorage('\n') AS (line:chararray);

-- Split each line into words
words = FOREACH data GENERATE FLATTEN(TOKENIZE(line)) AS word;

-- Remove any null or empty words
filtered_words = FILTER words BY word IS NOT NULL AND word != '';

-- Group the words
grouped_words = GROUP filtered_words BY word;

-- Count the occurrences of each word
word_count = FOREACH grouped_words GENERATE group AS word, COUNT(filtered_words)
 AS count;

-- Display the results directly in the terminal
DUMP word_count;
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/input.txt /user/cloudera/
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/word_count.pig /user/cloud
era/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 16 items
```

```
-rw-r--r--   1 cloudera cloudera         86 2024-10-20 08:23 /user/cloudera/inpu
t.txt
-rw-r--r--   1 cloudera cloudera        415 2024-10-20 06:43 /user/cloudera/max_
salary.pig
-rw-r--r--   1 cloudera cloudera        899 2024-10-20 05:28 /user/cloudera/stud
ents.txt
-rw-r--r--   1 cloudera cloudera        458 2024-10-20 06:31 /user/cloudera/tota
l_count.pig
-rw-r--r--   1 cloudera cloudera        458 2024-10-20 06:02 /user/cloudera/unio
n_batches.pig
-rw-r--r--   1 cloudera cloudera        564 2024-10-20 08:24 /user/cloudera/word
 count.pig
```

```
grunt> exec word_count.pig
2024-10-20 08:24:40,933 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 08:24:49,385 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY,FILTER
```

```
2024-10-20 08:27:36,504 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(is,2)
(big,3)
(for,1)
(pig,2)
(the,1)
(data,3)
(great,1)
(hello,3)
(world,1)
(future,1)
grunt>
```

10. Write a Pig Latin script to demonstrate different join operations, such as inner join, left join, right join, and full join, using employee_data and dept_data files based on deptcode.

Pig Latin Script:

> **join.pig**

```
-- Load employee and department data
employees = LOAD 'employee_data.txt' USING PigStorage(',')
AS (emp_id:int,emp_fname:chararray, emp_lname:chararray,
job:chararray, salary:float, deptcode:int); departments = LOAD
'dept_data.txt' USING PigStorage(',') AS
(deptcode:int,deptname:chararray);

-- Inner Join
inner_join = JOIN employees BY deptcode, departments BY
deptcode;DUMP inner_join;

-- Left Join
left_join = JOIN employees BY deptcode LEFT, departments BY
deptcode;DUMP left_join;

-- Right Join
right_join = JOIN employees BY deptcode RIGHT, departments BY
deptcode;DUMP right_join;

-- Full Join
full_join = JOIN employees BY deptcode FULL, departments BY
deptcode;DUMP full_join;
```

Command:

```
[cloudera@quickstart ~]$ cat
/home/cloudera/dept_data.txt[cloudera@quickstart
~]$ cat /home/cloudera/join.pig
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/dept_data.txt
/user/cloudera/[cloudera@quickstart ~]$ hdfs dfs -put
/home/cloudera/join.pig /user/cloudera/ [cloudera@quickstart ~]$ hdfs dfs -ls
/user/cloudera/


grunt> exec join.pig
```

Output:

```
[cloudera@quickstart ~]$ cat /home/cloudera/dept_data.txt
10,Engineering
20,Management
30,Technical Support
```

```
[cloudera@quickstart ~]$ cat /home/cloudera/join.pig
-- Load employee and department data
employees = LOAD 'employee_data.txt' USING PigStorage(',') AS (emp_id:int, emp_f
name:chararray, emp_lname:chararray, job:chararray, salary:float, deptcode:int);
departments = LOAD 'dept_data.txt' USING PigStorage(',') AS (deptcode:int, deptn
ame:chararray);

-- Inner Join
inner_join = JOIN employees BY deptcode, departments BY deptcode;
DUMP inner_join;

-- Left Join
left_join = JOIN employees BY deptcode LEFT, departments BY deptcode;
DUMP left_join;

-- Right Join
right_join = JOIN employees BY deptcode RIGHT, departments BY deptcode;
DUMP right_join;

-- Full Join
full_join = JOIN employees BY deptcode FULL, departments BY deptcode;
DUMP full_join;
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/dept_data.txt /user/cloude
ra/
```

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/join.pig /user/cloudera/
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/
Found 18 items
drwx------   - cloudera cloudera          0 2024-07-31 08:45 /user/cloudera/.Tra
sh
```

```
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 07:01 /user/cloudera/dept
30_employees.pig
-rw-r--r--   1 cloudera cloudera         53 2024-10-20 08:34 /user/cloudera/dept
_data.txt
-rw-r--r--   1 cloudera cloudera        334 2024-10-20 04:49 /user/cloudera/empl
oyee_data.txt
-rw-r--r--   1 cloudera cloudera        293 2024-10-20 07:13 /user/cloudera/firs
t_five.pig
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
```

```
-rw-r--r--   1 cloudera cloudera        420 2024-10-20 05:22 /user/cloudera/full
_names.pig
-rw-r--r--   1 cloudera cloudera         86 2024-10-20 08:23 /user/cloudera/inpu
t.txt
-rw-r--r--   1 cloudera cloudera        714 2024-10-20 08:35 /user/cloudera/join
.pig
-rw-r--r--   1 cloudera cloudera        415 2024-10-20 06:43 /user/cloudera/max_
salary.pig
```

```
grunt> exec join.pig
2024-10-20 08:35:30,703 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-10-20 08:35:31,889 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: HASH_JOIN
```

```
2024-10-20 08:38:14,214 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(10,Linda,Martinez,Technician,62000.0,10,10,Engineering)
(7,James,Davis,Developer,72000.0,10,10,Engineering)
(4,Alice,Johnson,Analyst,60000.0,10,10,Engineering)
(1,John,Doe,Engineer,70000.0,10,10,Engineering)
(8,Patricia,Garcia,Manager,95000.0,20,20,Management)
(6,Sarah,Wilson,HR,90000.0,20,20,Management)
(2,Jane,Smith,Manager,80000.0,20,20,Management)
(9,Robert,Miller,Engineer,68000.0,30,30,Technical Support)
(5,Michael,Lee,Developer,75000.0,30,30,Technical Support)
(3,Bob,Brown,Technician,50000.0,30,30,Technical Support)
2024-10-20 08:38:14,781 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: HASH_JOIN
```

```
2024-10-20 08:40:54,939 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(10,Linda,Martinez,Technician,62000.0,10,10,Engineering)
(7,James,Davis,Developer,72000.0,10,10,Engineering)
(4,Alice,Johnson,Analyst,60000.0,10,10,Engineering)
(1,John,Doe,Engineer,70000.0,10,10,Engineering)
(8,Patricia,Garcia,Manager,95000.0,20,20,Management)
(6,Sarah,Wilson,HR,90000.0,20,20,Management)
(2,Jane,Smith,Manager,80000.0,20,20,Management)
(9,Robert,Miller,Engineer,68000.0,30,30,Technical Support)
(5,Michael,Lee,Developer,75000.0,30,30,Technical Support)
(3,Bob,Brown,Technician,50000.0,30,30,Technical Support)
2024-10-20 08:40:55,648 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: HASH_JOIN
```

*

```
2024-10-20 08:43:24,933 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(10,Linda,Martinez,Technician,62000.0,10,10,Engineering)
(7,James,Davis,Developer,72000.0,10,10,Engineering)
(4,Alice,Johnson,Analyst,60000.0,10,10,Engineering)
(1,John,Doe,Engineer,70000.0,10,10,Engineering)
(8,Patricia,Garcia,Manager,95000.0,20,20,Management)
(6,Sarah,Wilson,HR,90000.0,20,20,Management)
(2,Jane,Smith,Manager,80000.0,20,20,Management)
(9,Robert,Miller,Engineer,68000.0,30,30,Technical Support)
(5,Michael,Lee,Developer,75000.0,30,30,Technical Support)
(3,Bob,Brown,Technician,50000.0,30,30,Technical Support)
2024-10-20 08:43:25,499 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: HASH_JOIN
```

```
2024-10-20 08:45:56,391 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(10,Linda,Martinez,Technician,62000.0,10,10,Engineering)
(7,James,Davis,Developer,72000.0,10,10,Engineering)
(4,Alice,Johnson,Analyst,60000.0,10,10,Engineering)
(1,John,Doe,Engineer,70000.0,10,10,Engineering)
(8,Patricia,Garcia,Manager,95000.0,20,20,Management)
(6,Sarah,Wilson,HR,90000.0,20,20,Management)
(2,Jane,Smith,Manager,80000.0,20,20,Management)
(9,Robert,Miller,Engineer,68000.0,30,30,Technical Support)
(5,Michael,Lee,Developer,75000.0,30,30,Technical Support)
(3,Bob,Brown,Technician,50000.0,30,30,Technical Support)
```