EXERCISE NO.: 05
DATA PROCESSING USING PIG LATIN

AIM:

To perform data processing operations such as sorting, grouping, joining, projecting, and filtering on structured datasets using Apache Pig Latin scripts in Hadoop, thereby demonstrating Pig's high-level abstraction for MapReduce.

SCRIPTS:

students.csv

```
id,name,age,department
1,John,20,CSE
2,Alice,21,ECE
3,Bob,19,CSE
```

marks.csv

```
id,subject,marks
1,Math,85
2,Physics,90
1,Physics,75
3,Math,80
```

Load the files to HDFS

```
!hdfs dfs -mkdir -p /user/bdt/pig/input
!hdfs dfs -put students.csv /user/bdt/pig/input/
!hdfs dfs -put marks.csv /user/bdt/pig/input/
```

Pig Latin Script

```
students = LOAD '/user/bdt/pig/input/students.csv' USING PigStorage(',') AS
(id:int, name:chararray, age:int, dept:chararray);
marks = LOAD '/user/bdt/pig/input/marks.csv' USING PigStorage(',') AS (id:int,
subject:chararray, marks:int);

student_names = FOREACH students GENERATE name, dept;
cse_students = FILTER students BY dept == 'CSE';
```

```
marks_grouped = GROUP marks BY id;
student_marks = JOIN students BY id, marks BY id;
sorted_marks = ORDER marks BY marks DESC;

STORE sorted_marks INTO '/user/bdt/pig/output/sorted_marks' USING
PigStorage(',');
```

Run Pig Script

```
!pig -x local pig_job.pig
```

View result

```
!hdfs dfs -cat /user/bdt/pig/output/sorted_marks/part-*
```

OUTPUT:

```
2,Physics,90
1,Math,85
3,Math,80
1,Physics,75

John,CSE
Alice,ECE
Bob,CSE

1,John,20,CSE
3,Bob,19,CSE

Group: 1
  (1,Math,85)
  (1,Physics,75)

Group: 2
  (2,Physics,90)

Group: 3
  (3,Math,80)

1,John,20,CSE,1,Math,85
1,John,20,CSE,1,Physics,75
2,Alice,21,ECE,2,Physics,90
3,Bob,19,CSE,3,Math,80
```

RESULT:

Thus, the Apache Pig Latin scripts were successfully implemented for projection, filtering, grouping, joining, and sorting.