

PROGRAM NO:6

Write Pig Latin scripts to sort, group, join, project, and filter the data.

PIG INSTALLATION

Extract to: C:\pig

Set Environment Variables:

PIG_HOME = C:\pig

HADOOP_HOME = (path to your Hadoop, e.g., C:\hadoop)

JAVA_HOME = (path to JDK, e.g., C:\Program Files\Java\jdk1.8.0_311)

Add to Path:

%PIG_HOME%\bin

Run Pig

Open Command Prompt:

For Local Mode:

pig -x local

Step1: First create a Folder with name pigdata in D Drive then create students.txt with this content as in below

101,John,CS,80

102,Alice,EC,90

103,Bob,CS,75

104,David,EC,85

105,Eve,ME,70

Then add departments.txt and add below content

CS,Computer Science

EC,Electronics

ME,Mechanical

Step2

Now open cmd prompt as administrator and run command

```
pig -x local
```

Step3

LOADING OF DATA

```
students = LOAD 'D:/pigdata/students.txt'  
    USING PigStorage(',')  
    AS (id:int, name:chararray, dept:chararray, marks:int);  
  
departments = LOAD 'D:/pigdata/departments.txt'  
    USING PigStorage(',')  
    AS (code:chararray, dept_name:chararray);  
  
Project Specific Columns  
projected_data = FOREACH students GENERATE name, marks;  
DUMP projected_data;  
  
Filter Rows  
high_scorers = FILTER students BY marks > 80;  
  
DUMP high_scorers;
```

Group by Department
grouped_by_dept = GROUP students BY dept;

```
DUMP grouped_by_dept;
```

A) To get average marks per department:
avg_marks = FOREACH grouped_by_dept GENERATE
group AS dept, AVG(students.marks) AS avg_score;

```
DUMP avg_marks;
```

B) Sort by Marks Descending
sorted_students = ORDER students BY marks DESC;

```
DUMP sorted_students;
```

C) Join Students with Departments
joined_data = JOIN students BY dept, departments BY code;

```
DUMP joined_data;
```

D) To project joined output:
result = FOREACH joined_data GENERATE

```
students::id,  
students::name,  
  
departments::dept_name,  
      students::marks;  
DUMP result;
```