EXP-6 : Develop Pig Latin scripts to sort, group, join, project and filter the data

A. Install Hadoop and Pig

- Install Java
- Install Hadoop
- Install Apache Pig (which uses Pig Latin scripts).

1. Download Apache Pig

Open a terminal and run:

```
cd /opt
```

```
sudo wget https://downloads.apache.org/pig/latest/pig-0.17.0.tar.gz
```

2. **Extract the Pig archive**

```
sudo tar -xvzf pig-0.17.0.tar.gz
```

Rename for simplicity (optional)

```
sudo mv pig-0.17.0 pig
```

3. Set Environment Variables

Edit your .bashrc

```
nano ~/.bashrc
```

Add these lines at the bottom:

```
# Pig Environment

export PIG_HOME=/opt/pig

export PATH=$PATH:$PIG_HOME/bin

export PIG_CLASSPATH=$HADOOP_HOME/etc/Hadoop
```

Note:

- HADOOP_HOME must already be set (you said Hadoop is installed).
- If not, you might have to set it too:

```
export HADOOP_HOME=/opt/hadoop

export PATH=$PATH:$HADOOP_HOME/bin
```

Now, reload the bash profile:

```
source ~/.bashrc
```

4. **Test Pig Installation**

```
pig -version
```

You should see something like :

Apache Pig version 0.17.0 (rUnknown)

compiled May 2016

5. Running Pig

You can run Pig in two modes:

- **Local mode** (no need for Hadoop):

```
pig -x local
```

- **MapReduce mode** (with Hadoop cluster)

```
pig
```

You’ll get into the grunt> shell to start writing Pig Latin scripts.

B. Prepare Your Data Files

Create a directory to work:

```
mkdir pig_project
cd pig_project
```

Create a data file : `nano students.txt`

```
1,Jayan,Math,85
2,Akshara,English,78
3,Bararth,Math,92
4,John,English,88
5,Charan,Math,90
```

Save and exit from editor.

C. Write the Pig Script

Create a script file:

```
nano student_operations.pig
```

type the following script :

```
students = LOAD 'students.txt' USING PigStorage(',')
        AS (student_id:int, name:chararray, subject:chararray, score:int);
```

```
-- Filter students with score > 80
high_scores = FILTER students BY score > 80;

-- Project only name and score
projected = FOREACH high_scores GENERATE name, score;

-- Group by name
grouped = GROUP projected BY name;

-- Sort by score descending (flatten needed after group)
flattened = FOREACH grouped {
   sorted = ORDER projected BY score DESC;
   GENERATE group AS name, sorted;
};

DUMP flattened;
```

Save and exit;

D. Run the Pig Script

Since we are not using a Hadoop cluster, run in **local mode**

```
pig -x local student_operations.pig
```

Output will be printed on the terminal.

E. (Optional) Save output to file

If you want to **store** output instead of DUMP:
Add at end of script:

```
STORE flattened INTO 'output_folder' USING PigStorage(',');
```

Then after running, check:

```
ls output_folder/
cat output_folder/part-m-00000
```