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

**Install Hadoop**

Pig runs on top of Hadoop, so you need Hadoop set up first.

* **Download Hadoop**: <https://hadoop.apache.org/releases.html>
* **Install Hadoop in local or pseudo-distributed mode** (good for testing)

**Install Apache Pig**

* **Download Pig**: <https://pig.apache.org/releases.html>
* Extract it and set environment variables:

export PIG\_HOME=/path/to/pig

export PATH=$PATH:$PIG\_HOME/bin

Verify installation:

pig –version

### ****Prepare Your Datasets****

Create two sample CSV files: students.csv

1,John,18,A

2,Alice,19,B

3,Bob,17,A

4,David,20,B

scores.csv

1,Math,85

2,Math,78

3,Science,90

4,Math,88

### ****Upload Files to HDFS (if using HDFS)****

hdfs dfs -mkdir /pigdata

hdfs dfs -put students.csv /pigdata/

hdfs dfs -put scores.csv /pigdata/

## Pig Latin Script

Save the following as script.pig in the same folder:

-- Load student data

students = LOAD 'students.csv' USING PigStorage(',')

AS (student\_id:int, name:chararray, age:int, grade:chararray);

-- Load score data

scores = LOAD 'scores.csv' USING PigStorage(',')

AS (student\_id:int, subject:chararray, score:int);

-- FILTER: Students older than 18

filtered\_students = FILTER students BY age > 18;

-- PROJECT: Only student name and grade

projected\_students = FOREACH filtered\_students GENERATE name, grade;

-- SORT: Students by name

sorted\_students = ORDER projected\_students BY name ASC;

-- GROUP: Scores grouped by subject

grouped\_scores = GROUP scores BY subject;

-- JOIN: Students and scores on student\_id

joined\_data = JOIN students BY student\_id, scores BY student\_id;

-- Show results

DUMP sorted\_students;

DUMP grouped\_scores;

DUMP joined\_data;

## Run the Script in Command Prompt

Open your Command Prompt:

cd C:\PigProjects

pig -x local script.pig

This will print:

* Sorted students (name & grade, where age > 18)
* Grouped scores (by subject)
* Joined data from students and scores