

LAB - 7

Aim: To Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter the data.

Procedure:

1. Navigate to Apache Pig downloads: <https://downloads.apache.org/pig/latest/>
2. Download '*pig-0.17.0.tar.gz*'
3. Extract the .tar.gz file using 7-Zip
4. Open System Environment Variables
5. Add PIG_HOME User Variable
 - Variable Name: *PIG_HOME*
 - Variable Value: *C:\pig-0.17.0*
6. Update System PATH Variable
 - Add: *C:\pig-0.17.0\bin* to PATH
7. Locate 'pig.cmd' in *C:\pig-0.17.0\bin*
8. Modify HADOOP_BIN_PATH configuration
 - Original: ``set HADOOP_BIN_PATH=%HADOOP_HOME%\bin``
 - Updated: ``set HADOOP_BIN_PATH=%HADOOP_HOME%\libexec``
9. Open Command Prompt
10. Run Pig in Local Mode:
pig -x local
11. **Pig Latin Scripts:**
 - a. **Data Loading**

students.csv [Read-Only] - Excel

File Home Insert Page Layout Formulas Data Review View Help

Paste Cut Copy Format Painter Clipboard

Calibri 11 A⁺ A⁻ B I U Font

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-d

I4

	A	B	C	D	E	F	G	H
1	student_id	name	age	subject	score	grade_level		
2	1	John Smith	18	Math	85	12		
3	2	Emily John	17	Science	92	11		
4	3	Michael Br	16	English	78	10		
5	4	Sarah Davi	18	Math	88	12		
6	5	David Wils	17	Science	95	11		
7	6	Jessica Lee	16	English	82	10		
8	7	Kevin Cher	17	Math	90	11		
9	8	Amanda R	18	Science	87	12		
10	9	Ryan Thon	16	English	75	10		
11	10	Lisa Kim	17	Math	93	11		

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

b. Filtering Data

```
high_performers =
    FILTER students BY
    score > 85;
```

```
(2,Emily Johnson,17,Science,92,11)
(4,Sarah Davis,18,Math,88,12)
(5,David Wilson,17,Science,95,11)
(7,Kevin Chen,17,Math,90,11)
(8,Amanda Rodriguez,18,Science,87,12)
(10,Lisa Kim,17,Math,93,11)
```

c. Grouping Data

grouped_by_subject = GROUP students BY subject;

```
(Math, {(10,Lisa Kim,17,Math,93,11), (7,Kevin Chen,17,Math,90,11), (4,Sarah Davis,18,Math,88,12), (1,John Smith,18,Math,85,12)})
(English, {(9,Ryan Thompson,16,English,75,10), (6,Jessica Lee,16,English,82,10), (3,Michael Brown,16,English,78,10)})
(Science, {(8,Amanda Rodriguez,18,Science,87,12), (5,David Wilson,17,Science,95,11), (2,Emily Johnson,17,Science,92,11)})
(subject, {(,name,,subject,,)})
```

d. Projecting

Columns

name_score =
FOREACH students
GENERATE name,
score;

```
(John Smith,85)
(Emily Johnson,92)
(Michael Brown,78)
(Sarah Davis,88)
(David Wilson,95)
(Jessica Lee,82)
(Kevin Chen,90)
(Amanda Rodriguez,87)
(Ryan Thompson,75)
(Lisa Kim,93)
```

e. Sorting Data

sorted_by_score = ORDER students BY score DESC;

```
(5,David Wilson,17,Science,95,11)
(10,Lisa Kim,17,Math,93,11)
(2,Emily Johnson,17,Science,92,11)
(7,Kevin Chen,17,Math,90,11)
(4,Sarah Davis,18,Math,88,12)
(8,Amanda Rodriguez,18,Science,87,12)
(1,John Smith,18,Math,85,12)
(6,Jessica Lee,16,English,82,10)
(3,Michael Brown,16,English,78,10)
(9,Ryan Thompson,16,English,75,10)
(,name,,subject,,)
```

Output Screenshots

Pig Installed

```
C:\>pig -x local
2024-11-21 18:14:34,942 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-11-21 18:14:34,943 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-11-21 18:14:35,194 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-11-21 18:14:35,194 [main] INFO org.apache.pig.Main - Logging error messages to: C:\hadoop\logs\pig_1732193075190.log
2024-11-21 18:14:35,212 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file C:\Users\Anil kumar singh/.pigbootstrap not found
2024-11-21 18:14:35,326 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-11-21 18:14:35,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-11-21 18:14:35,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-11-21 18:14:35,435 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-3831ff88-163e-4d4a-aac3-f3821896654e
2024-11-21 18:14:35,435 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
```

Results and Conclusion

- Pig was successfully installed and configured.
- Pig Latin scripts for sorting, grouping, joining, projecting, and filtering data were executed successfully in local mode.