IMAAD HAJWANE

202101132 / 23

BIG DATA ANALYTICS

ASSIGNMENT: PERFORM LOAD AND STORE OF DATA USING PIG

Loading and Storing Data in Apache Pig

In Apache Pig, **LOAD** and **STORE** are fundamental operations for data input and output. They allow you to read data from various sources (e.g., HDFS, local file system) and store results back into these systems after processing. The operations are part of the broader Pig Latin scripting language, which enables high-level data transformations.

Loading Data Using Pig

The **LOAD** function in Pig is used to read data from a file or a directory and bring it into the Pig environment for processing. The data can be located on the local file system or in HDFS.

Storing Data Using Pig

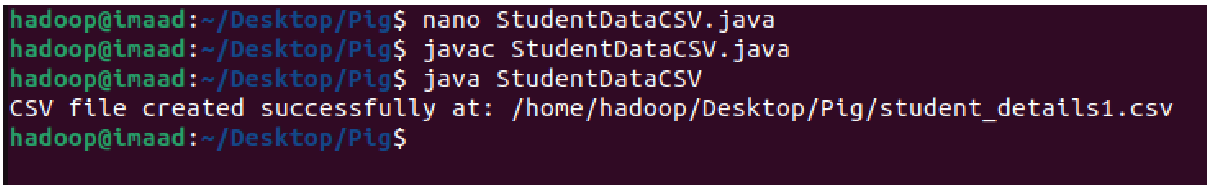
After processing the data using Pig, you can save the results back to a file or directory using the **STORE** command. Like the LOAD command, STORE allows you to specify the storage format and the output path.

Common Pig Operations Beyond LOAD and STORE:

- **FILTER**:
  - Filters records based on a condition.
- **FOREACH**:
  - Applies a transformation to each record.
- **GROUP**:
  - Groups records by a specified field.
- **JOIN**:
  - Joins two or more datasets based on a common key.
- **ORDER BY**:
  - Sorts data based on specified fields.

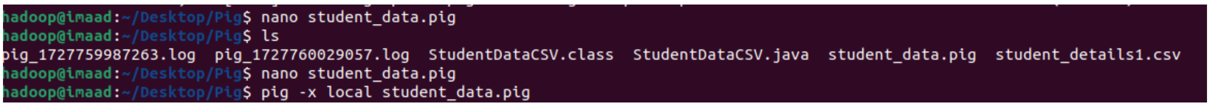
EXECUTION:

- Creation of CSV file & storing



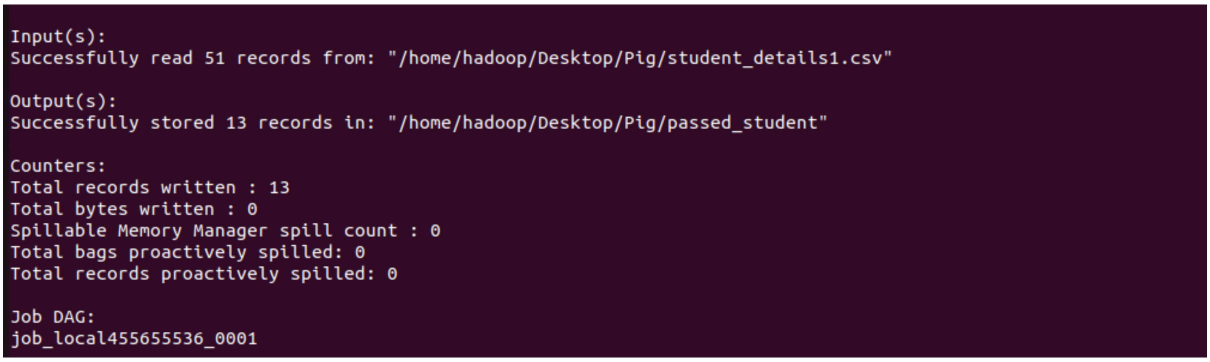
```
hadoop@imaad:~/Desktop/Pig$ nano StudentDataCSV.java
hadoop@imaad:~/Desktop/Pig$ javac StudentDataCSV.java
hadoop@imaad:~/Desktop/Pig$ java StudentDataCSV
CSV file created successfully at: /home/hadoop/Desktop/Pig/student_details1.csv
hadoop@imaad:~/Desktop/Pig$
```

- Locating & initializing of Pig for execution



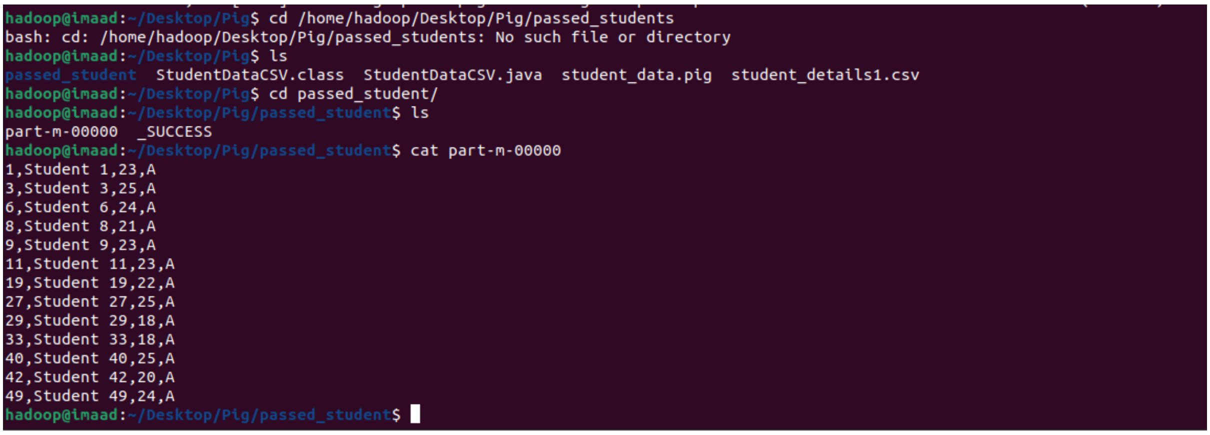
```
hadoop@imaad:~/Desktop/Pig$ nano student_data.pig
hadoop@imaad:~/Desktop/Pig$ ls
pig_1727759987263.log  pig_1727760029057.log  StudentDataCSV.class  StudentDataCSV.java  student_data.pig  student_details1.csv
hadoop@imaad:~/Desktop/Pig$ nano student_data.pig
hadoop@imaad:~/Desktop/Pig$ pig -x local student_data.pig
```

- Successful Execution of Pig Script



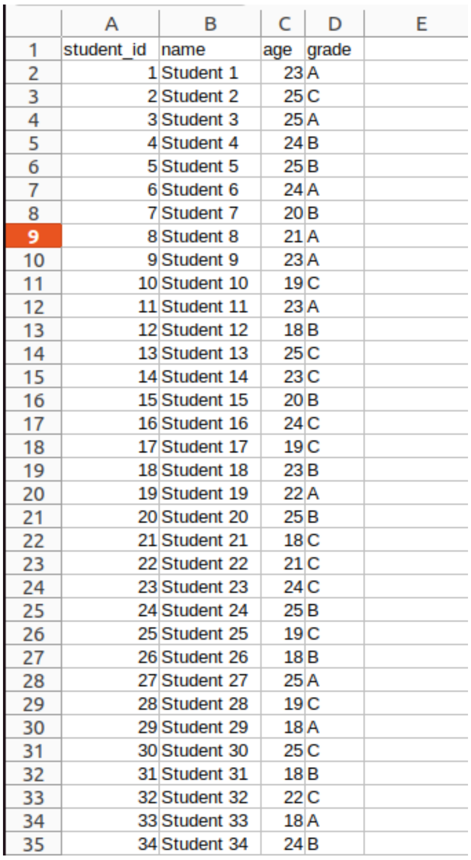
```
Input(s):
Successfully read 51 records from: "/home/hadoop/Desktop/Pig/student_details1.csv"

Output(s):
Successfully stored 13 records in: "/home/hadoop/Desktop/Pig/passed_student"

Counters:
Total records written : 13
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local455655536_0001
```

- Display of data, after execution of pig script



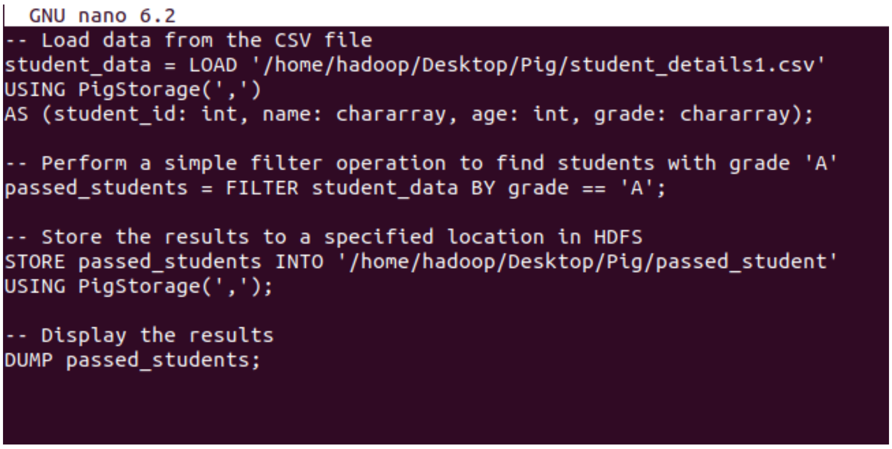
```
hadoop@imaad:~/Desktop/Pig$ cd /home/hadoop/Desktop/Pig/passed_students
bash: cd: /home/hadoop/Desktop/Pig/passed_students: No such file or directory
hadoop@imaad:~/Desktop/Pig$ ls
passed_student  StudentDataCSV.class  StudentDataCSV.java  student_data.pig  student_details1.csv
hadoop@imaad:~/Desktop/Pig$ cd passed_student/
hadoop@imaad:~/Desktop/Pig/passed_student$ ls
part-m-00000  _SUCCESS
hadoop@imaad:~/Desktop/Pig/passed_student$ cat part-m-00000
1,Student 1,23,A
3,Student 3,25,A
6,Student 6,24,A
8,Student 8,21,A
9,Student 9,23,A
11,Student 11,23,A
19,Student 19,22,A
27,Student 27,25,A
29,Student 29,18,A
33,Student 33,18,A
40,Student 40,25,A
42,Student 42,20,A
49,Student 49,24,A
hadoop@imaad:~/Desktop/Pig/passed_student$ 
```

DATASET SAMPLE:



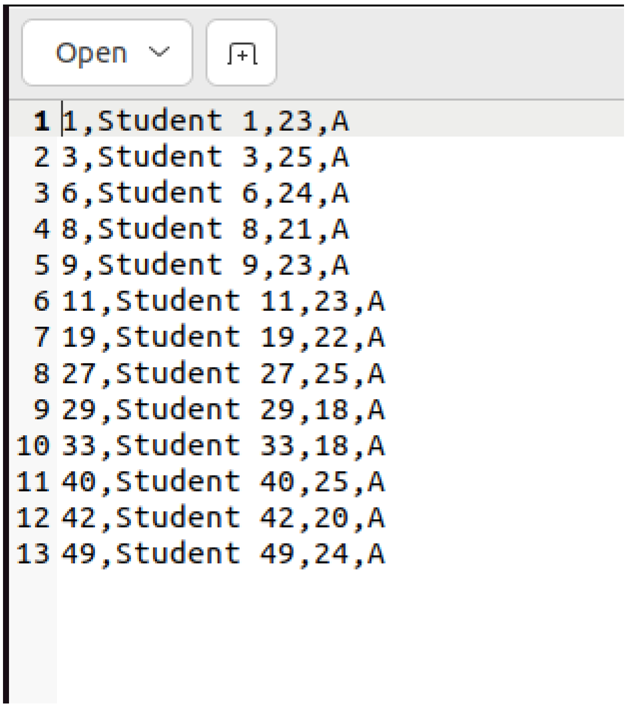
	A	B	C	D	E
1	student_id	name	age	grade	
2	1	Student 1	23	A	
3	2	Student 2	25	C	
4	3	Student 3	25	A	
5	4	Student 4	24	B	
6	5	Student 5	25	B	
7	6	Student 6	24	A	
8	7	Student 7	20	B	
9	8	Student 8	21	A	
10	9	Student 9	23	A	
11	10	Student 10	19	C	
12	11	Student 11	23	A	
13	12	Student 12	18	B	
14	13	Student 13	25	C	
15	14	Student 14	23	C	
16	15	Student 15	20	B	
17	16	Student 16	24	C	
18	17	Student 17	19	C	
19	18	Student 18	23	B	
20	19	Student 19	22	A	
21	20	Student 20	25	B	
22	21	Student 21	18	C	
23	22	Student 22	21	C	
24	23	Student 23	24	C	
25	24	Student 24	25	B	
26	25	Student 25	19	C	
27	26	Student 26	18	B	
28	27	Student 27	25	A	
29	28	Student 28	19	C	
30	29	Student 29	18	A	
31	30	Student 30	25	C	
32	31	Student 31	18	B	
33	32	Student 32	22	C	
34	33	Student 33	18	A	
35	34	Student 34	24	B	

PIG SCRIPT:



```
  GNU nano 6.2
-- Load data from the CSV file
student_data = LOAD '/home/hadoop/Desktop/Pig/student_details1.csv'
USING PigStorage(',')
AS (student_id: int, name: chararray, age: int, grade: chararray);

-- Perform a simple filter operation to find students with grade 'A'
passed_students = FILTER student_data BY grade == 'A';

-- Store the results to a specified location in HDFS
STORE passed_students INTO '/home/hadoop/Desktop/Pig/passed_student'
USING PigStorage(',');

-- Display the results
DUMP passed_students;
```

PIG OUTPUT:



```
1,Student 1,23,A
3,Student 3,25,A
6,Student 6,24,A
8,Student 8,21,A
9,Student 9,23,A
11,Student 11,23,A
19,Student 19,22,A
27,Student 27,25,A
29,Student 29,18,A
33,Student 33,18,A
40,Student 40,25,A
42,Student 42,20,A
49,Student 49,24,A
```