IMAAD IMRAN HAJWANE

202101132 / 21

BIG DATA ASSIGNMENT

TOPIC: PIG

---

1. Grouping

Concept: Grouping is a fundamental operation that collects data with the same key into a single group. This allows for aggregate functions to be applied to each group.

- **Theory**: In many data analysis scenarios, you often want to analyze data in groups. For example, if you have sales data, you might want to group sales by region or by product category. In Pig, the `GROUP` operator collects all the records that have the same value for the specified field(s).
- **Use Case**: Count the number of occurrences of each category in a dataset or calculate the sum of values for each group.

2. Joining

Concept: Joining combines two or more datasets based on a common key. This is akin to SQL joins (inner join, outer join).

- **Theory**: When working with relational databases or datasets, it is common to have related information stored in separate tables. Joining allows you to bring this related data together for analysis.
- **Types of Joins**:
  - **Inner Join**: Returns records that have matching values in both datasets.
  - **Outer Join**: Can be left, right, or full, returning all records from one dataset and the matched records from the other.
- **Use Case**: Combine user information with their respective transactions for analysis.

3. Combining

Concept: Combining refers to merging multiple datasets into one. This can be done using the `UNION` operator in Pig.

- **Theory**: When you have datasets that share the same schema (same structure), you might want to merge them into a single dataset for ease of analysis. The `UNION` operation stacks datasets on top of each other.
- **Use Case**: Merging sales data from different regions or time periods into a single dataset for overall analysis.

4. Splitting

Concept: Splitting divides a dataset into multiple datasets based on specified conditions.

- **Theory**: This operation is useful when you want to create separate subsets of data for different analysis paths. The `SPLIT` operator allows you to define multiple output relations based on a condition.
- **Use Case**: Divide a dataset of students into those who passed and those who failed based on their scores.

5. Filtering

Concept: Filtering reduces the dataset by removing records that do not meet specified criteria.

- **Theory**: In data analysis, you often need to focus on a specific subset of data. The `FILTER` operator allows you to specify conditions that records must satisfy to be included in the resulting dataset.
- **Use Case**: Selecting records of customers whose purchases exceed a certain amount.

6. Sorting

Concept: Sorting arranges records in a specified order, either ascending or descending.

- **Theory**: Sorting is a common operation in data processing that helps to organize data for better readability and analysis. In Pig, the `ORDER` operator is used to sort data based on one or more fields.
- **Use Case**: Sort a list of products by price or sort user registrations by date.

File Generating Commands:

- Creation of all files

```
hadoop@imaad:~/Desktop/Pig_Commands$ nano GroupingCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ nano group.pig
hadoop@imaad:~/Desktop/Pig_Commands$ nano joinCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ nano join.pig
hadoop@imaad:~/Desktop/Pig_Commands$ nano CombineCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ nano combine.pig
hadoop@imaad:~/Desktop/Pig_Commands$ nano SplitCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ nano split.pig
hadoop@imaad:~/Desktop/Pig_Commands$ nano FilterCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ nano filter.pig
hadoop@imaad:~/Desktop/Pig_Commands$ nano SortCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ nano sort.pig
hadoop@imaad:~/Desktop/Pig_Commands$
```

CSV Files & Code Files:

1. Grouping

- Csv file creation & Execution & Output

```
hadoop@imaad:~/Desktop/Pig_Commands$ javac GroupingCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ java GroupingCSV
CSV file created: group_data.csv
```

```
Input(s):
Successfully read 9 records from: "file:///home/hadoop/Desktop/Pig_Commands/group_data.csv"

Output(s):
Successfully stored 4 records in: "file:/tmp/temp561830826/tmp-766887973"

Counters:
Total records written : 4
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local611667652_0001
```

```
(A,{(A,30),(A,20),(A,10)})
(B,{(B,50),(B,40)})
(C,{(C,80),(C,70),(C,60)})
(category,{(category,)})
```

2. Joining

- CSV File generation, Execution & Output

```
hadoop@imaad:~/Desktop/Pig_Commands$ javac joinCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ java joinCSV
CSV files created: data1.csv and data2.csv
hadoop@imaad:~/Desktop/Pig_Commands$ pig -x local
2024-10-15 11:01:28,719 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-15 11:01:28,721 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-10-15 11:01:28,836 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-10-15 11:01:28,836 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/Desktop/Pig_Commands/pig_1728970288829.log
2024-10-15 11:01:28,860 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2024-10-15 11:01:29,032 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-10-15 11:01:29,034 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-10-15 11:01:29,187 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-10-15 11:01:29,220 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-6329d6b7-7ea4-4c5e-98ff-3faa858fb6fa
2024-10-15 11:01:29,244 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> grunt> exec join.pig
```

```
(1,John,1,5000)
(2,Alice,2,6000)
(3,Bob,3,5500)
grunt>
```

3. Combining

- CSV File generation, Execution & Output

```
hadoop@imaad:~/Desktop/Pig_Commands$ javac CombineCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ java CombineCSV
CSV files created: data1.csv and data2.csv
hadoop@imaad:~/Desktop/Pig_Commands$ pig -x local
```

```
Input(s):
Successfully read 3 records from: "file:///home/hadoop/Desktop/Pig_Commands/data1.csv"
Successfully read 3 records from: "file:///home/hadoop/Desktop/Pig_Commands/data2.csv"

Output(s):
Successfully stored 6 records in: "file:/tmp/temp735777084/tmp495779239"

Counters:
Total records written : 6
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2143731647_0001


2024-10-15 11:03:02,102 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:03:02,103 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:03:02,104 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:03:02,110 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 4 time(s).
2024-10-15 11:03:02,118 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-10-15 11:03:02,138 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-10-15 11:03:02,138 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-10-15 11:03:02,152 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 2
2024-10-15 11:03:02,156 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 2
(,)
(1,10)
(2,20)
(,)
(3,30)
(4,40)
grunt>
```

4. Splitting

- CSV File generation, Execution & Output

```
hadoop@imaad:~/Desktop/Pig_Commands$ javac SplitCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ java SplitCSV
CSV file created: split_data.csv
hadoop@imaad:~/Desktop/Pig_Commands$ pig -x local
```

```
Input(s):
Successfully read 6 records from: "file:///home/hadoop/Desktop/Pig_Commands/split_data.csv"

Output(s):
Successfully stored 3 records in: "file:/tmp/temp1209956120/tmp-808715237"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2097982809_0002


2024-10-15 11:04:23,904 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:04:23,906 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:04:23,914 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:04:23,919 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 2 time(s).
2024-10-15 11:04:23,924 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-10-15 11:04:23,925 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-10-15 11:04:23,926 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-10-15 11:04:23,942 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-10-15 11:04:23,947 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,50)
(2,30)
(4,20)
```

5. Filtering
    - CSV File generation, Execution & Output

```
hadoop@imaad:~/Desktop/Pig_Commands$ javac FilterCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ java FilterCSV
CSV file created: filter_data.csv
hadoop@imaad:~/Desktop/Pig_Commands$ pig -x local
2024-10-15 11:06:00,530 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-15 11:06:00,530 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
```

```
Input(s):
Successfully read 6 records from: "file:///home/hadoop/Desktop/Pig_Commands/filter_data.csv"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp2106802612/tmp-1478338435"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local354848849_0001


2024-10-15 11:06:07,253 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:06:07,258 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:06:07,261 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:06:07,274 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 2 time(s).
2024-10-15 11:06:07,277 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-10-15 11:06:07,279 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-10-15 11:06:07,283 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-10-15 11:06:07,300 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-10-15 11:06:07,303 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(3,80)
(5,60)
grunt>
```

6. Sorting
    - CSV File generation, Execution & Output

```
hadoop@imaad:~/Desktop/Pig_Commands$ javac SortCSV.java
hadoop@imaad:~/Desktop/Pig_Commands$ java SortCSV
CSV file created: sort_data.csv
hadoop@imaad:~/Desktop/Pig_Commands$ pig -x local
2024-10-15 11:06:49,916 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-15 11:06:49,917 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
```

```
Input(s):
Successfully read 6 records from: "file:///home/hadoop/Desktop/Pig_Commands/sort_data.csv"

Output(s):
Successfully stored 6 records in: "file:/tmp/temp-467285359/tmp-1421296812"

Counters:
Total records written : 6
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1280134976_0001        ->      job_local1801100351_0002,
job_local1801100351_0002        ->      job_local121291750_0003,
job_local121291750_0003


2024-10-15 11:07:01,535 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,537 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,538 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,561 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,594 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,596 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,611 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,613 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,614 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2024-10-15 11:07:01,629 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 2 time(s).
2024-10-15 11:07:01,648 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-10-15 11:07:01,650 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-10-15 11:07:01,651 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-10-15 11:07:01,654 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-10-15 11:07:01,654 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(3,80)
(5,60)
(1,50)
(2,30)
(4,20)
(,)
grunt>
```