PROJECT, AGGREGATE, REMOVE , COUNT, LIMIT, SKIP AND SORT JSON FILE USING HADOOP PIG

COMMAND:

```
grunt> employees = LOAD '/user/hadoop/emp.json' USING
JsonLoader('name:chararray,age:int,department:chararray,salary:float');
```

```
2024-09-13 17:31:43,759 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
```

```
grunt> projected = FOREACH employees GENERATE name, salary;
```

```
grunt> dump projected;
```

```
2024-09-13 17:32:12,309 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: UNKNOWN
2024-09-13 17:32:12,356 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2024-09-13 17:32:12,371 [main] INFO  org.apache.pig.data.SchemaTupleBackend
- Key [pig.schematuple] was not set... will not generate code.
2024-09-13 17:32:12,406 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter,
MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2024-09-13 17:32:12,435 [main] INFO
org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for
employees: $1, $2
2024-09-13 17:32:12,505 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler -
File concatenation threshold: 100 optimistic? false
2024-09-13 17:32:12,538 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOpti
mizer - MR plan size before optimization: 1
2024-09-13 17:32:12,539 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOpti
mizer - MR plan size after optimization: 1
2024-09-13 17:32:18,125 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.MapTask - Processing split: Number of splits :1
```

```
Total Length = 269
Input split[0]:
  Length = 269
  ClassName: org.apache.hadoop.mapreduce.lib.input.FileSplit
  Locations:

-----------------------

2024-09-13 17:32:18,174 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigRecordRead
er - Current split being processed
hdfs://localhost:9000/user/hadoop/emp.json:0+269
2024-09-13 17:32:18,186 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapreduce.lib.output.PathOutputCommitterFactory - No output
committer factory defined, defaulting to FileOutputCommitterFactory
2024-09-13 17:32:18,186 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output
Committer Algorithm version is 2
2024-09-13 17:32:18,186 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter -
FileOutputCommitter skip cleanup _temporary folders under output directory:false,
ignore cleanup failures: false
2024-09-13 17:32:18,263 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen)
of size 699400192 to monitor. collectionUsageThreshold = 489580128,
usageThreshold = 489580128
2024-09-13 17:32:18,267 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set...
will not generate code.
2024-09-13 17:32:18,279 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$
Map - Aliases being processed per job phase (AliasName[line,offset]): M:
employees[1,12],projected[2,12] C:  R:
2024-09-13 17:32:18,369 [LocalJobRunner Map Task Executor #0] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigHadoopLog
ger - org.apache.pig.builtin.JsonLoader(UDF_WARNING_1): Bad record, could
not find start of record [
2024-09-13 17:32:18,376 [LocalJobRunner Map Task Executor #0] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigHadoopLog
ger - org.apache.pig.builtin.JsonLoader(UDF_WARNING_1): Encountered
exception org.codehaus.jackson.JsonParseException: Unexpected close marker ']':
expected '}' (for ROOT starting at [Source:
java.io.ByteArrayInputStream@57193cb3; line: 1, column: 0])
```

```
 at [Source: java.io.ByteArrayInputStream@57193cb3; line: 1, column: 2]. Bad
record, returning null for ]
2024-09-13 17:32:18,381 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner -
2024-09-13 17:32:18,890 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task -
Task:attempt_local1742288717_0001_m_000000_0 is done. And is in the process
of committing
2024-09-13 17:32:18,898 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner -
2024-09-13 17:32:18,898 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task
attempt_local1742288717_0001_m_000000_0 is allowed to commit now
2024-09-13 17:32:18,982 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of
task 'attempt_local1742288717_0001_m_000000_0' to
hdfs://localhost:9000/tmp/temp727912759/tmp-208615080
2024-09-13 17:32:18,985 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - map
2024-09-13 17:32:18,985 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task
'attempt_local1742288717_0001_m_000000_0' done.
2024-09-13 17:32:18,993 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Final Counters for
attempt_local1742288717_0001_m_000000_0: Counters: 23
        File System Counters
        FILE: Number of bytes read=429
        FILE: Number of bytes written=6409860
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=5610724
        HDFS: Number of bytes written=5610545
        HDFS: Number of read operations=46
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=15
        HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
        Map input records=6
        Map output records=6
        Input split bytes=369
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=0
        Total committed heap usage (bytes)=399507456
        File Input Format Counters
        Bytes Read=0
        File Output Format Counters
        Bytes Written=0
        org.apache.pig.PigWarning
        UDF_WARNING_1=1
        org.apache.pig.builtin.JsonLoader
        UDF_WARNING_1=1
2024-09-13 17:32:18,993 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - Finishing task:
attempt_local1742288717_0001_m_000000_0
2024-09-13 17:32:18,994 [Thread-30] INFO
org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2024-09-13 17:32:19,195 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLau
ncher - 50% complete
2024-09-13 17:32:19,195 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLau
ncher - Running jobs are [job_local1742288717_0001]
2024-09-13 17:32:22,717 [main] WARN
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
already initialized!
2024-09-13 17:32:22,723 [main] WARN
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
already initialized!
2024-09-13 17:32:22,724 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is
deprecated. Instead, use mapreduce.job.maps
2024-09-13 17:32:22,724 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is
deprecated. Instead, use mapreduce.job.reduces
2024-09-13 17:32:22,725 [main] WARN
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
already initialized!
2024-09-13 17:32:22,869 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLau
ncher - 100% complete
2024-09-13 17:32:22,872 [main] INFO
org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion    PigVersion   UserId       StartedAt    FinishedAt   Features
```

```
3.3.6  0.16.0hadoop        2024-09-13 17:32:12       2024-09-13 17:32:22
UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces      MaxMapTime       MinMapTime        AvgMapTime
MedianMapTime   MaxReduceTime   MinReduceTime    AvgReduceTime
MedianReducetime        Alias  Feature       Outputs
job_local1742288717_0001      1       0       n/a     n/a     n/a     n/a     00      0
0       employees,projected       MAP_ONLY
hdfs://localhost:9000/tmp/temp727912759/tmp-208615080,

Input(s):
Successfully read 6 records (5610724 bytes) from: "/user/hadoop/emp.json"

Output(s):
Successfully stored 6 records (5610545 bytes) in:
"hdfs://localhost:9000/tmp/temp727912759/tmp-208615080"

Counters:
Total records written : 6
Total bytes written : 5610545
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1742288717_0001


2024-09-13 17:32:22,874 [main] WARN
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
already initialized!
2024-09-13 17:32:22,876 [main] WARN
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
already initialized!
2024-09-13 17:32:22,877 [main] WARN
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
already initialized!
2024-09-13 17:32:22,879 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLau
ncher - Encountered Warning UDF_WARNING_1 1 time(s).
```

```
2024-09-13 17:32:22,879 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLau
ncher - Success!
2024-09-13 17:32:22,882 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2024-09-13 17:32:22,882 [main] WARN
org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already
been initialized
2024-09-13 17:32:22,927 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to
process : 1
2024-09-13 17:32:22,928 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths to process : 1
```

OUTPUT:

```
(,)
(Jhon Deer,50000.0)
(Jhon Deer,60000.0)
(Celena,70000.0)
(Vinoth,70000.0)
(,)
grunt>
```

COMMAND:

```
grunt> employees = LOAD '/user/hadoop/emp.json' USING
JsonLoader('name:chararray,age:int,department:chararray,salary:float');
grunt> total_salary = FOREACH (GROUP employees ALL) GENERATE
SUM(employees.salary) AS total_salary;
grunt> dump total_salary;
```

OUTPUT:

```
(250000.0)
```

COMMAND:

```
grunt>skipped_employees = LIMIT employees 1000000;
grunt>dump skipped_employees;
```

OUTPUT:

```
(,,,)
(Jhon Deer,30,HR,50000.0)
(Jhon Deer,40,Marketing,60000.0)
(Celena,19,Finance,70000.0)
(Vinoth,20,IT,70000.0)
(,,,)
```

COMMAND:

```
top_3_employees = LIMIT employees 3;

DUMP top_3_employees;
```

OUTPUT:

```
2024-09-13 17:40:28,634 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths to process : 1
(,,,)
(Jhon Deer,30,HR,50000.0)
(Jhon Deer,40,Marketing,60000.0)
(Celena,19,Finance,70000.0)
```

COMMAND:

```
-- Count the number of employees
employee_count = FOREACH (GROUP employees ALL) GENERATE
COUNT(employees) AS
total_count;

DUMP employee_count;
```

OUTPUT:

```
2024-09-13 17:42:54,524 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to
process : 1
```

```
2024-09-13 17:42:54,526 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths to process : 1
(4)
```

COMMAND:

```
-- Remove employees from the 'IT' department
filtered_employees = FILTER employees BY department != 'IT';

DUMP filtered_employees;
```

OUTPUT:

```
2024-09-13 17:44:05,392 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to
process : 1
2024-09-13 17:44:05,393 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths to process : 1
(Jhon Deer,30,HR,50000.0)
(Jhon Deer,40,Marketing,60000.0)
(Celena,19,Finance,70000.0)
```

PYTHON FILE HDFS ACCESS

```
hadoop@vinoth-ubuntu:~$ python3 -m venv .venv
hadoop@vinoth-ubuntu:~$ source .venv/bin/activate
(.venv) hadoop@vinoth-ubuntu:~$ python3 -m pip install hdfs
(.venv) hadoop@vinoth-ubuntu:~/Documents$ python3 process_data.py
```

OUTPUT:

```
Raw JSON Data: [
{"name":"Jhon Deer","age":30,"department":"HR", "salary":50000},
{"name":"Jhon Deer","age":40,"department":"Marketing", "salary":60000},
{"name":"Celena","age":19,"department":"Finance", "salary":70000},
{"name":"Vinoth","age":20,"department":"IT", "salary":70000}
]

Filtered JSON file saved successfully.
Projection: Select only name and salary columns
        name  salary
```

```
0  Jhon Deer   50000
1  Jhon Deer   60000
2       Celena   70000
3       Vinoth   70000
Aggregation: Calculate total salary
Total Salary: 250000


# Count: Number of employees earning more than 50000
Number of High Earners (>50000): 3


limit Top 5 highest salary
Top 5 Earners:
        name  age department  salary
2       Celena   19  Finance   70000
3       Vinoth   20          IT   70000
1  Jhon Deer   40  Marketing   60000
0  Jhon Deer   30           HR   50000


Skipped DataFrame (First 2 rows skipped):
        name  age department  salary
2  Celena   19       Finance   70000
3  Vinoth   20       IT   70000


Filtered DataFrame (Sales department removed):
        name  age department  salary
0  Jhon Deer   30           HR   50000
1  Jhon Deer   40  Marketing   60000
2       Celena   19  Finance   70000
(.venv) hadoop@vinoth-ubuntu:~/Documents$ ^C
```