EXPERIMENT 9: *To use Apache Pig and implement group by, order by, filter and join operations on databases inside HDFS.*

Name: Adwait Purao
UID : 2021300101
Division: COMPS B

1. **Start Pig in Local Mode:**
    a. **Command:**

```
pig -x local
```

```
grunt>
```

2. **Load data into Apache Pig:**
    a. **Command:**

```
cricketers_data = LOAD
'file:///home/hadoop/Desktop/Pig/cricketers_data.txt' USING
PigStorage(',') AS (player_id:int, name:chararray, team:chararray,
score:int);

dump cricketers_data;
```

```
2024-04-03 01:29:59,926 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-04-03 01:29:59,927 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,John,DHAKA,85)
(2,Alex,CHITTAGONG,92)
(3,Michael,KHULNA,78)
(4,Chris,DHAKA,102)
(5,Steve,KHULNA,115)
(6,David,CHITTAGONG,88)
(,,,)
grunt>
```

3. **Load data into Apache Pig:**
    a. **Command:**

```
team_data = LOAD
'file:///home/hadoop/Desktop/Pig/team_data.txt' USING
PigStorage(',') AS (team:chararray, coach:chararray);

dump team_data;
```

```
2024-04-03 09:59:44,428 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(DHAKA,Coach1)
(CHITTAGONG,Coach2)
(KHULNA,Coach3)
(,)
grunt>
```

4. **Group By: To group the data by team:**
    a. **Command:**

```
grouped_data = GROUP cricketers_data BY team;

dump grouped_data;
```

```
2024-04-03 01:31:05,737 [main] INFO  org.apache.pig.backend.hadoop.exe
(DHAKA,{(4,Chris,DHAKA,102),(1,John,DHAKA,85)})
(KHULNA,{(5,Steve,KHULNA,115),(3,Michael,KHULNA,78)})
(CHITTAGONG,{(6,David,CHITTAGONG,88),(2,Alex,CHITTAGONG,92)})
(,{(,,,)})
grunt>
```

5. **Order By: To order the data by score:**
    a. **Command:**

```
ordered_data = ORDER cricketers_data BY score DESC;

dump ordered_data;
```

```
2024-04-03 01:31:46,652 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-04-03 01:31:46,652 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(5,Steve,KHULNA,115)
(4,Chris,DHAKA,102)
(2,Alex,CHITTAGONG,92)
(6,David,CHITTAGONG,88)
(1,John,DHAKA,85)
(3,Michael,KHULNA,78)
(,,,)
grunt>
```

6. **Filter: To filter the data for a specific team, say KHULNA**
    a. **Command:**

```
filtered_data = FILTER cricketers_data BY team == 'KHULNA';

dump filtered_data;
```

```
2024-04-03 01:33:06,074 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-04-03 01:33:06,075 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(3,Michael,KHULNA,78)
(5,Steve,KHULNA,115)
grunt>
```

7. **Join: To join two datasets,on the basis of a common attribute.**
    a. **Command:**

```
joined_data = JOIN cricketers_data BY team, team_data BY
team;

dump joined_data;
```

```
2024-04-03 09:53:44,766 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(4,Chris,DHAKA,102,DHAKA,Coach1)
(1,John,DHAKA,85,DHAKA,Coach1)
(5,Steve,KHULNA,115,KHULNA,Coach3)
(3,Michael,KHULNA,78,KHULNA,Coach3)
(6,David,CHITTAGONG,88,CHITTAGONG,Coach2)
(2,Alex,CHITTAGONG,92,CHITTAGONG,Coach2)
grunt>
```