PIG

- **Fetch the record having VendorID as '2' AND tpep_pickup_datetime as '2017-10-01 00:15:30' AND tpep_dropoff_datetime as '2017-10-01 00:25:11' AND passenger_count as '1' AND trip_distance as '2.17'**

```
MovieLensData = LOAD '/user/root/spark/input_data/yellow_tripdata*'
using PigStorage(',') as
(VendorID,
tpep_pickup_datetime,
tpep_dropoff_datetime,
passenger_count,
trip_distance,
RatecodeID,
store_and_fwd_flag,
PULocationID,
DOLocationID,
payment_type,
fare_amount,
extra,
mta_tax,
tip_amount,
tolls_amount,
improvement_surcharge,
total_amount
);

--VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count
,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationI
D,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,impr
ovement_surcharge,total_amount


MovieLensData_Clean = FILTER MovieLensData BY NOT (VendorID ==
'VendorID');

ReqData = FILTER MovieLensData_Clean BY VendorID == '2' AND
tpep_pickup_datetime == '2017-10-01 00:15:30' AND
                tpep_dropoff_datetime == '2017-10-01 00:25:11' AND
passenger_count == '1' AND trip_distance == '2.17';
```

```
STORE ReqData INTO
'/user/root/spark/pig_output_data/single_row_lookup/' USING
PigStorage(',');
```

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.14.0 0.12.0-cdh5.14.0 root 2018-04-12 07:14:37 2018-04-12 07:17:11 FILTER

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTIme AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1523459984514_0005 36 0 12 7 9 9 n/a n/a n/a n/a MovieLensData,MovieLensData_Clean MAP_ONLY /user/root/spark/pig_output_data/single_row_lookup,

Input(s):
Successfully read 54518004 records (4810073225 bytes) from: "/user/root/spark/input_data/yellow_tripdata*"

Output(s):
Successfully stored 1 records (90 bytes) in: "/user/root/spark/pig_output_data/single_row_lookup"

```
[root@ip-10-0-0-59 spark_assignment]# hadoop dfs -cat
/user/root/spark/pig_output_data/single_row_lookup/part*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

2,2017-10-01 00:15:30,2017-10-01 00:25:11,1,2.17,1,N,141,142,1,9,0.5,0.5,2.06,0,0.3,12.36
```

- **Filter all the records having RatecodeID as 4.**

```
MovieLensData = LOAD '/user/root/spark/input_data/yellow_tripdata*'
using PigStorage(',') as
(VendorID,
tpep_pickup_datetime,
tpep_dropoff_datetime,
passenger_count,
trip_distance,
RatecodeID,
store_and_fwd_flag,
PULocationID,
```

```
DOLocationID,
payment_type,
fare_amount,
extra,
mta_tax,
tip_amount,
tolls_amount,
improvement_surcharge,
total_amount
);

--VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count
,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationI
D,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,impr
ovement_surcharge,total_amount


MovieLensData_Clean = FILTER MovieLensData BY NOT (VendorID ==
'VendorID');

ReqData = FILTER MovieLensData_Clean BY RatecodeID == '4';

STORE ReqData INTO '/user/root/spark/pig_output_data/filter/' USING
PigStorage(',');
```

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.14.0 0.12.0-cdh5.14.0 root 2018-04-12 07:06:01 2018-04-12 07:08:39 FILTER

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTIme	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReducetime	Alias	Feature	Outputs
job_1523459984514_0003	36	0	12	7	10	10	n/a	n/a	n/a	n/a	MovieLensData,MovieLensData_Clean	MAP_ONLY	/user/root/spark/pig_output_data/filter,

Input(s):
Successfully read 54518004 records (4810073225 bytes) from: "/user/root/spark/input_data/yellow_tripdata*"

Output(s):
Successfully stored 31029 records (2811007 bytes) in: "/user/root/spark/pig_output_data/filter"

- **Group By all the records based on payment type and find the count for each group. Sort the payment types in ascending order of their count.**

```
MovieLensData = LOAD '/user/root/spark/input_data/yellow_tripdata*'
using PigStorage(',') as
(VendorID,
tpep_pickup_datetime,
tpep_dropoff_datetime,
passenger_count,
trip_distance,
RatecodeID,
store_and_fwd_flag,
PULocationID,
DOLocationID,
payment_type,
fare_amount,
extra,
mta_tax,
tip_amount,
tolls_amount,
improvement_surcharge,
total_amount
);

--VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count
,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationI
D,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,impr
ovement_surcharge,total_amount


MovieLensData_Clean1 = FILTER MovieLensData BY NOT (VendorID ==
'VendorID');

MovieLensData_Clean = FILTER MovieLensData_Clean1 BY NOT (VendorID ==
‘’);
ReqData = FOREACH  MovieLensData_Clean GENERATE payment_type AS
payment_type, 1L AS CNT;
ReqDataGrp = GROUP ReqData BY payment_type;
```

```
FinalData = FOREACH ReqDataGrp GENERATE FLATTEN(group),
SUM(ReqData.CNT) AS CNT;
FinalDataSort = ORDER FinalData BY CNT;

STORE FinalDataSort INTO
'/user/root/spark/pig_output_data/groupbysort/' USING
PigStorage(',');
```

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.14.0 0.12.0-cdh5.14.0 root 2018-04-12 07:22:50 2018-04-12 07:28:04 GROUP_BY,ORDER_BY,FILTER

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTIme	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReducetime	Alias	Feature	Outputs
job_1523459984514_0006	36	5	18	11	16	17	42	4	12	5	FinalData,MovieLensData,MovieLensData_Clean,ReqData,ReqDataGrp	GROUP_BY,COMBINER	
job_1523459984514_0007	1	1	2	2	2	2	3	3	3	3	FinalDataSort	SAMPLER	
job_1523459984514_0008	1	1	2	2	2	2	3	3	3	3	FinalDataSort	ORDER_BY	/user/root/spark/pig_output_data/groupbysort,

Input(s):
Successfully read 54518004 records (4810073225 bytes) from: "/user/root/spark/input_data/yellow_tripdata*"

Output(s):
Successfully stored 4 records (39 bytes) in: "/user/root/spark/pig_output_data/groupbysort"

```
[root@ip-10-0-0-59 spark_assignment]# hadoop dfs -cat
/user/root/spark/pig_output_data/groupbysort/part*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

4,85287
3,292090
2,17648209
1,36492406
```

Java

- **Fetch the record having VendorID as '2' AND tpep_pickup_datetime as '2017-10-01 00:15:30' AND tpep_dropoff_datetime as '2017-10-01 00:25:11' AND passenger_count as '1' AND trip_distance as '2.17'**

```
package Spark.Spark_Single_Row_Lookup;


import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;

public class SparkSingleRowLookup {

	public static void main(String[] args) {
		SparkConf conf = new SparkConf();
		JavaSparkContext sc = new JavaSparkContext(conf);

		JavaRDD<String> tripData = sc.textFile(args[0]);


		JavaRDD<String> tripDataClean = tripData.filter(
				x -> {
					String[] vals = x.split(",");
					if (vals[0].equals("VendorID")){
						return false;
					} else {
						return true;
					}
				}

				);
		JavaRDD<String> reqdData = tripDataClean.filter(
				x -> {
					String[] vals = x.split(",");
					if (vals[0].equals("2") &&
vals[1].equals("2017-10-01 00:15:30") && vals[2].equals("2017-10-01
00:25:11")
							&& vals[3].equals("1") &&
vals[4].equals("2.17")){
						return true;
					} else {
						return false;
					}
				}

				);
```

```
            reqdData.saveAsTextFile(args[1]);
            sc.close();
      }


}
```

```
spark-submit --class
Spark.Spark_Single_Row_Lookup.SparkSingleRowLookup \
--master yarn --deploy-mode client \
--name spark_singlerow_lookup --conf
"spark.app.id=spark_singlerow_lookup" \
Spark_Single_Row_Lookup-0.0.1-SNAPSHOT-jar-with-dependencies.jar
/user/root/spark/input_data/yellow_tripdata*
/user/root/spark/output_data/single_row_lookup/
```

```
18/04/12 07:40:43 INFO spark.SparkContext: Running Spark version 1.6.0
18/04/12 07:42:09 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-4d0373af-9e1e-48a4-b68e-a1ad14105360
```

```
[root@ip-10-0-0-59 spark_assignment]# hadoop fs -cat
/user/root/spark/output_data/single_row_lookup/part*
2,2017-10-01 00:15:30,2017-10-01 00:25:11,1,2.17,1,N,141,142,1,9,0.5,0.5,2.06,0,0.3,12.36
```

- **Filter all the records having RatecodeID as 4.**

```
spark-submit --class spark.filter.SparkFilter \
> --master yarn --deploy-mode client \
> --name spark_filter --conf "spark.app.id=spark_filter" \
> filter-0.0.1-SNAPSHOT-jar-with-dependencies.jar
/user/root/spark/input_data/yellow_tripdata* /user/root/spark/output_data/filter/
```

```
package spark.filter;
```

```
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;

public class SparkFilter {

	public static void main(String[] args) {
		SparkConf conf = new SparkConf();
		JavaSparkContext sc = new JavaSparkContext(conf);

		JavaRDD<String> tripData = sc.textFile(args[0]);


		JavaRDD<String> tripDataClean = tripData.filter(
				x -> {
					String[] vals = x.split(",");
					if (vals[0].equals("VendorID") ||
vals[0].equals("") || vals[0]==null){
						return false;
					} else {
						return true;
					}
				}

		);
		JavaRDD<String> reqdData = tripDataClean.filter(

				x -> {
					String[] vals = x.split(",");
					if (vals[5].equals("4")){
						return true;
					} else {
						return false;
					}
				}

			);

		reqdData.saveAsTextFile(args[1]);
		sc.close();
	}
```

```
}
```

```
18/04/11 15:24:27 INFO spark.SparkContext: Running Spark version 1.6.0

18/04/11 15:26:05 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-2c7ec5ba-fb16-409c-bea0-59c08ca96782
```

```
[root@ip-10-0-0-59 spark_assignment]# hadoop fs -cat
/user/root/spark/output_data/filter/part* | wc -l
31029
```

- **Group By all the records based on payment type and find the count for each group. Sort the payment types in ascending order of their count.**

```
[root@ip-10-0-0-59 spark_assignment]# spark-submit --class
spark.groupbywithsort.SparkGroupBySort \
> --master yarn --deploy-mode client \
> --name spark_groupby --conf "spark.app.id=spark_groupby" \
> groupbywithsort-0.0.1-SNAPSHOT-jar-with-dependencies.jar
/user/root/spark/input_data/yellow_tripdata* /user/root/spark/output_data/groupbysort/
```

```
18/04/12 06:17:46 INFO spark.SparkContext: Running Spark version 1.6.0

18/04/12 06:19:18 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-c12f5cba-72d1-4308-acd4-eb04b81b915c
```

```
package spark.groupbywithsort;

import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaPairRDD;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;

import scala.Tuple2;

public class SparkGroupBySort {
```

```
	public static void main(String[] args) {
		SparkConf conf = new SparkConf();
		JavaSparkContext sc = new JavaSparkContext(conf);

		JavaRDD<String> tripData = sc.textFile(args[0]);


		JavaRDD<String> tripDataClean = tripData.filter(
				x -> {
					String[] vals = x.split(",");
					if (vals[0].equals("VendorID") || vals[0].equals("") ||
vals[0]==null){
						return false;
					} else {
						return true;
					}
				}

				);
		JavaPairRDD<String, Long> paymentType = tripDataClean.mapToPair(

				x -> {
					String[] vals = x.split(",");
					return new Tuple2<String, Long>(vals[9], 1L);
				}

				);
		JavaPairRDD<String, Long> finalStat =  paymentType.reduceByKey((x,y) -> x+y);
		JavaPairRDD<Long, String> xchangeKey = finalStat.mapToPair(

				tuple -> {
					Long key = tuple._2;
					String value = tuple._1;
					return new Tuple2<Long, String>(key,value);
				}
				);

		JavaPairRDD<Long, String> orderData = xchangeKey.sortByKey(false);
		JavaPairRDD<String, Long> xchangeKeyAgain =        orderData.mapToPair(

				tuple -> {
					Long value = tuple._1;
					String key = tuple._2;
					return new Tuple2<String, Long>(key,value);
				}
				);


		xchangeKeyAgain.saveAsTextFile(args[1]);
		sc.close();
	}

}
```

```
[root@ip-10-0-0-59 spark_assignment]# hadoop fs -cat
/user/root/spark/output_data/groupbysort/part*
(4,85287)
(3,292090)
(2,17648209)
(1,36492406)
```