Problem Statement

As already mentioned, you will have to perform the POC on three use cases i.e. single record lookup, filter and GroupBy accompanied with ordering in ascending order. So their respective problem statements are mentioned below. The following problems you will have to solve using Pig and Spark RDDs:

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

Guidelines

By now you must have gone through the problem statement at least once. As this assignment is a real-life industry problem, will instruct you to solve it exactly the way how it is done in industry. In industry, you have to be very careful while executing any job in the Hadoop or Spark cluster because the clusters are shared resources.

Apart from our jobs, there are other jobs which are dependent on the same cluster. So, before running any job in the cluster, in your case EC2 instance, you have to be precise about the correctness of the job because it would be a highly unpleasant situation if the entire cluster crashes due to some trivial bug in your code. To avoid such bugs, the codes are first developed and tested in local on a sample dataset. Once the codes produce valid output then only they are run on the real data set in the cluster. For solving this assignment, you are also expected to follow the same industrial approach, i.e. in the first part of the assignment, you will have to

develop the codes(pig and spark) and test your code on a sample data file of size ~2MB. In the second part of the assignment, you will be expected to run the same codes on the 7GB data in EC2 instance.

Steps for solving the first part of the assignment is given below:

- Install pig in your local
- Download eclipse with Java 1.8
- In a text editor develop pig codes for each given problem. Three different pig codes i.e. one for each problem statement
- Using eclipse develop spark codes for each given problem. Three different spark codes or java projects i.e. one for each problem statement. Please refer to the spark coding lectures for setting up the Java project for running spark codes
- Run the pig and spark codes on the sample data
- Capture the following stats in a tabular format:

Job Run Time

	Single Record Lookup	Filter	Group by Accompanied with Order by
Pig			
Spark RDD			

You may not see much difference in processing time for Pig and Spark as the size of the input data is very small(~2MB).

Total Count of Records in Output(The cells in each column should have the same value)

	Single Record Lookup	Filter	Group by
Pig			
Spark RDD			

Please note that the number of records for each column has to be the same. If the values are different in each column then either your Pig or Spark code is incorrect and you will have to revisit them again

The link for downloading the sample dataset is mentioned at bottom of the page.

As of now, please complete this exercise in local. Once you are done with this exercise and confident that your codes are giving the expected results you will have to run the same codes in the EC2 instance as well. The 7GB dataset will be available soon on EC2.