Keerthi Ravikanti - Assignment 1

// Assignment 1: what is the average delay for each day of the week.

1. Assumptions considered for this assignment:

1) **Arrival Delay:** Given that business is highly focussed to meet the passenger expectations by ensuring the passenger is arrived at destination ON TIME. For understanding of "Flight delay", I have considered only the factor of Arrival Delay as it do matter for passengers.

2) **Cancelled & Diverted flights** are not considered for arriving the flight delay numbers because we don't have any data for analytical purpose with respect to Flight Arrival Delays

2. Scala Code:

//Step 1: Start all the hadoop processes by following commands

$start-dfs.sh

$start-yarn.sh

Enter into Spark environment

$spark-shell

//Loading the CSV file into spark with the help of a scala varibale SC

val aircraft = sc.textFile("file:///home/hduser/Datasets/flights2007.csv")

```
scala> val aircraft = sc.textFile("file:///home/hduser/Datasets/flights2007.csv")
17/05/22 03:05:31 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 86.5 KB, free 86.5 KB)
17/05/22 03:05:31 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 19.6 KB, free 106.1 KB)
17/05/22 03:05:31 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:51157 (size: 19.6 KB, free: 517.4 MB)
17/05/22 03:05:31 INFO spark.SparkContext: Created broadcast 0 from textFile at <console>:27
aircraft: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27
```

//Step 2: Removing the header column and splitting the lines to fields delimited by comma

val records = aircraft.filter(line=> !line.contains("Year")).map(line => line.split(",").map(elem => elem.trim))

println(records.take(10).deep)

```
scala> val records = aircraft.filter(line=> !line.contains("Year")).map(line => line.split(",").map(elem => elem.trim))
records: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[3] at map at <console>:29

scala> println(records.take(10).deep)
17/05/22 03:08:32 INFO mapred.FileInputFormat: Total input paths to process : 1
17/05/22 03:08:32 INFO spark.SparkContext: Starting job: take at <console>:32
17/05/22 03:08:33 INFO scheduler.DAGScheduler: Got job 0 (take at <console>:32) with 1 output partitions
```

//Step 3: Filtering out the cancelled flights containing values as 1

val filterArrDelay1 = records.filter(rec => (rec(21) != 1))

```
scala> val filterArrDelay1 = records.filter(rec => (rec(21) != 1))
filterArrDelay1: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[4] at filter at <console>:31
```

//Step 4: Filtering out the diverted flights containing values as 1

val filterArrDelay2 = filterArrDelay1.filter(rec => (rec(23) != 1))

```
scala> val filterArrDelay2 = filterArrDelay1.filter(rec => (rec(23) != 1))
filterArrDelay2: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[5] at filter at <console>:33
```

//Step 5: Removed the flights containing "NA" values under ArrDelay column

val filterArrDelay3 = filterArrDelay2.filter(rec => (rec(14) !="NA"))

//Step 6: As per the requirement grouping the arrival delay values for each DayOfWeek

val AircraftArrDelayMap = filterArrDelay3.map(rec=> (rec(3),rec(14).toInt))

```
scala> val filterArrDelay3 = filterArrDelay2.filter(rec => (rec(14) !="NA" ))
filterArrDelay3: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[13] at filter at <console>:35

scala> val AircraftArrDelayMap = filterArrDelay3.map(rec=> (rec(3),rec(14).toInt))
AircraftArrDelayMap: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[14] at map at <console>:37
```

//Step 7: Calculating the average of arrival delay at each DayOfWeek

val AvgArrDelay = AircraftArrDelayMap.mapValues(x => (x, 1)).reduceByKey((x, y) => (x._1 + y._1, x._2 + y._2)).mapValues(y => 1.0 * y._1 / y._2).collect

```
AvgArrDelay: Array[(String, Double)] = Array((1,10.513502556550229), (2,8.263684434009868), (3,9.962943847767281), (4,12.685980155261941), (5,13
.067675000697863), (6,5.846600031017031), (7,10.32957740663109))
```

3. Analyzing the output

Flights on Day of the week (5) which is Friday has the highest delay with 13.06 minutes where as Day of the week (6) which is Saturday has lowest Delay with 5.84 minutes.