Spark on Jupyter

Task Descrpition

In this coursework you will write several code snippets in Spark for an example dataset. You can test your queries on the Jupyter Notebook (with Spark on Azure). You will need to turn in your Jupyter Notebooks as *.ipynb files via the Hub.

Data Description

The dataset used is "Airline On-Time Statistics and Delay Causes" collected by U.S. Department of Transportation. The dataset can be downloaded form this link (only accessible inside Imperial's network; use VPN if connecting from outside). The file is compressed. Uncompress before use (may take a while) and move it to Azure so you can analyse it with Spark.

http://146.169.47.39/air_transits.csv.gz

Each row of the dataset contains 29 values, separated by comma.

TABLE 1 VARIABLE DESCRIPTIONS

#	Name	Description
1	Year	1987-2008
2	Month	12-Jan
3	DayofMonth	31-Jan
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?

23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Tasks

You should write Spark programs using Python (or other languages). Below is sample code for parsing the CSV file and count the total number of flights with an actual elapsed time between 60 and 90 minutes:

```
sc.textFile("PATH TO CSV FILE") \
    .map(lambda line: line.split(",")) \
    .filter(lambda line: (60 <= line[11] <= 90)) \
    .count()
```

Write spark codes to fulfill the following tasks and print the result on the console.

1) Compute the Total number of records
2) Find total number of flights per year
3) Find the plane with the most number of flights. Each plane has a unique TailNum.
4) Compute the total flight time of each plane.
5) Find the busiest airport (in terms of number of departures) for each year.
6) Find the airline with the highest average delay in 2006.
7) Find all possible one-stop flights between any two destinations (not necessarily with the same carrier, but excluding the returning flights) with a stopover between 1 hour and 2 hours.
8) Assume that a passenger wants to travel from Los Angeles International Airport (airport code: LAX) to Orlando International Airport (airport code: MCO), and then go back to Los Angeles (LAX). They depart LAX no earlier than 5:59 (scheduled time), stay at least 3:01 hours in Orlando and then arrive at LAX no later than 11pm. Based on the "scheduled" times, find which carrier has the most flights satisfying these constraints.

Submission

You need to turn in the Jupyter notebook with the code as well as the printed answers to each question. Submission is through the Hub.