Big Data Analytics Techniques and Applications

Homework 2

Due Date: 2020/11/10 23:59:59

Goal

Use PySpark to analyze the given dataset to answer the following questions.

Questions

1. Find the maximal delays (you should consider both ArrDelay and DepDelay) for each month of 2008.
2. How many flights were delayed caused by weather between 2000 ~ 2005? Please show the counting for each year.
3. List Top 5 airports which occur delays most in 2007. (Please show the IATA airport code)

Dataset

**Airline on-time performance dataset
The datasets are on New E3.

Requirements

1. A report (Hw2_StudentID.pdf), which should include:
    - The execution results by using PySpark
    - Answers to the questions given above
    - Anything else worth mentioning (e.g. other valuable observations) or difficulties encountered in this work.

Data Description

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
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