Big Data Analytics Techniques and Applications

Homework 2

Due Date: 2023/04/09 23:59:59

- **Dataset:**

  Airline on-time performance datasets (in **2000~2005**, **2007~2008**). You can find these datasets on E3.

  You need to use PySpark to analyze the given dataset to answer the following questions.
  (PySpark Documentation: https://spark.apache.org/docs/latest/api/python/)
  (PySpark Tutorial: https://sparkbyexamples.com/pyspark-tutorial/)

  * Please download all datasets on E3 and make sure that you use the correct dataset.

- **Questions:**

  - Q1: Find the maximal delays (you should consider both ArrDelay and DepDelay) for each month of 2008.
  - Q2: How many flights were delayed caused by security between 2000 ~ 2005? Please show the counting for each year.
  - Q3: List Top 5 airports which occur delays most and least in 2007. (Please show the IATA airport code)

- **Requirements:**

  - Submit a report named "HW2_{StudentID}.pdf" (e.g., HW2_310456099.pdf) to E3 and describe clearly the following items:

    - You can use PySpark via Google Colab or other platforms.
      - Refer to "PySpark_Colab.pptx" for the usage of PySpark in Colab.
    - The execution results by using PySpark (Attach source code).
    - Descriptions of how you solve each question in detail.
    - Some figures or tables to illustrate your analyzed answers to each question.
    - Anything else worth mentioning (e.g. other valuable observations or difficulties encountered in this work and how you resolve them).

  - You need to submit source code files to E3.

    - You should **zip** all source code files in a file named "HW2_{StudentID}_Code.**zip**" (e.g., HW2_310456099_Code.zip).

- **Penalty for late submission:**
  - If your work is submitted within one day after the deadline, you will get only **80%** of original score.
  - If your work is submitted within two days after the deadline, you will get only **50%** of original score.
  - If your work is submitted over two days after the deadline, you will get **zero score** on this homework.
- **Penalty for format error:**
  - The report file name has any format error. (-5%)
  - The report is not in pdf. (-5%)
  - The source code file is not in zip. (-5%)
- **Data Description:**

Name	Description
Year	1987~2008
Month	1~12
DayofMonth	1~31
DayOfWeek	1 (Monday) - 7 (Sunday)
DepTime	actual departure time (local, hhmm)
CRSDepTime	scheduled departure time (local, hhmm)
ArrTime	actual arrival time (local, hhmm)
CRSArrTime	scheduled arrival time (local, hhmm)
UniqueCarrier	unique carrier code
FlightNum	flight number
TailNum	plane tail number
ActualElapsedTime	in minutes
CRSElapsedTime	in minutes
AirTime	in minutes
ArrDelay	arrival delay, in minutes
DepDelay	departure delay, in minutes

Origin	origin IATA airport code
Dest	destination IATA airport code
Distance	in miles
TaxiIn	taxi in time, in minutes
TaxiOut	taxi out time in minutes
Cancelled	was the flight cancelled?
CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
Diverted	1 = yes, 0 = no
CarrierDelay	in minutes
WeatherDelay	in minutes
NASDelay	in minutes
SecurityDelay	in minutes
LateAircraftDelay	in minutes