Assignment - 'Live' data analysis

Introduction

We would like you to create a simple “real-time” stream processing pipeline. We will use a fixed flight route dataset as a source. Our pipeline should provide us the most popular source airports per time window.

Here is a link with a dataset which you use for this assignment: https://raw.githubusercontent.com/jpatokal/openflights/master/data/routes.dat . The documentation for this data can be found here: https://openflights.org/data.html .

Assignment

Using the flight routes dataset you will stream the given data into our pipeline, process it and display results. Keep in mind the resulting implementation needs to be somehow demoable. Use tooling you think is best fitted for the task (e.g. Docker, public cloud, etc.) and give us an explanation why you decided for it.

For the sake of this assignment, we limit the choice of processing framework to Apache Spark. We strongly suggest the use of Spark Structured Streaming over Spark Streaming. Use either Python or Scala.

Tasks:

1. Create a batch Spark job that read in the routes dataset. It should create an overview of the top 10 airports used as source airport. Write the output to a filesystem.
2. Use Spark structured streaming to change your job into a streaming job, and use the dataset file as a source.
3. Next change your streaming job so the aggregations are done using sliding windows. Pick any window and sliding interval. The end result should be the top 10 airports used as source airport within each window. When choosing the window interval, keep the size of the dataset in mind.
4. Productionize your code by adding unit tests.

We would like to be able to check your assignment. Code produced plus a description of your steps and thinking process should be the output of your assignment.

The goal of the exercise is not to get perfect solutions but rather to test candidate’s ability to implement the assignment, flexibility, willingness to learn and explain technical solutions they implemented.

Good luck!