Stream Processing – 2018/19

Project 2

Goal

The goal of this project is to exercise the use of Spark Streaming and/or Apache Flink.

Description

Alternative 1: solve the same 6 + 2 problems as in project 1 by using either Spark Streaming or Spark Flink.

Alternative 2: select 2 + 1 problems to solve and solve them both in Spark Streaming and Spark Flink.

Setting

File project2.zip (https://drive.google.com/file/d/1d7Aw4po-ojqpZQpeTuchU_wxs-t87IUC/view?usp=sharing) includes two batch files:

- start-kafka.sh – starts Kafka in a docker container in network "ps-net" in sserver "kafka:9092"
- start-publisher.sh – starts a process in docker to publish events in Kafka. This script can receive two paramenter. The first controls the file to be ingested in kafka (default: /debs/sample.csv.gz); the second controls the rate of ingestion (default: 60 times faster than realtime). The docker mount the local directory logs into /debs in the container.

Report

The report should present your work and have the following structure:
Introduction : presents the objectives of the project;
System design : presents the architecture of the systems and the algorithms used;
Evaluation : presents the evaluation
Conclusions.

Dates

31/May – delivery of the project.