Big Data Developer Homework

Introduction

- To complete the task, we recommend using Spark engine. However, you are free to choose any other tool or programming language.
- Completing all the parts of the homework is not mandatory. We strongly encourage you to submit your results even if you did not manage to complete everything.
- To submit your homework, please send back a zip file with an application source code or a link to a GitHub repository.
- Your application should be deployable and executable from the command line on a Linux environment.
- We will take into consideration not only the correctness of algorithms but also compliance with good coding practices and overall code quality.
- Covering code with unit tests is highly preferred

Data

- Raw data is in the raw_data folder.
- Assume that new data files will be uploaded constantly to the raw_data directory.
- Files are filled with information on impressions (impressions_processed_dk_) and clicks (clicks_processed_dk_*).
- One record in the files represents one impression/click.
- The file format is parquet https://parquet.apache.org

Task

Create an application that calculates the count of impressions and clicks by date and each hour of the day, for a specific user-agent value „some user agent“. The user agent field is in the device_settings structure field user_agent.

- Application can take user agent as a parameter and produce results to Apache Kafka that should be running in Docker container.
- If application is run multiple times with the same user-agent parameter – it should not process already processed files, but only process newly added files since last run.
- Output Kafka topic should contain user-agent as a key.
- Each Kafka message value should contain Protobuf encoded aggregated data by date/hour of impression and click counts. Additionally calculate average time between subsequent events (for both impressions and clicks separately) within aggregation period using creation_time field.
- If data is missing for some of the hours of the day, the application should fill in these rows with zero impressions and clicks.
- datetimes of data inside the files can be determined from a file name: impressions_processed_dk_20220526113212045_172845633-172845636_1.parquet , date and time format - „2022-05-26 11:32“.

Also please provide suggestion and example how to access and check the output data that is in Kafka topic.