Part 1 - Design for streaming analytics (weighted factor for grades = 4)

1. Select a dataset (https://data.cityofnewyork.us/resource/t29m-gskq.json) suitable for streaming analytics for a customer as a running example (thus the basic unit of the data should be a discrete record/event data). Explain the dataset and at least two different analytics for the customer: (i) streaming analytics which analyzes streaming data from the customer (customerstreamapp) and (ii) a batch analytics which analyzes historical results outputted by the streaming analytics. The explanation should be at a high level to allow us to understand the data and possible analytics so that, later on, you can implement and use them in answering other questions. (1 point).

The selected dataset for this project is the NYC Taxi Dataset, with 17 columns for each Row. These variables are:

VondorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID , DOLocationID, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount.

All these variables are numbers, except for tpep_pickup_datetime, tpep_dropoff_datetime which are Date & time format; and the store_and_fwd_flag which is plain text. To know more details, please visit https://data.cityofnewyork.us/api/views/t29m-gskq/files/89042b9b-8280-4339-bda2-d68f428a7499?download=true&filename=data_dictionary_trip_records_yellow.pdf

Each row of the dataset correspond to one trip, with all the info related to it. From the pick up date and time, to the method used to pay, the rate applied or the total amount paid at the end.

From the client point of view, one useful analysis is the average time between trips uploaded to the service and the average number of trips upload per driver each day. These two are for the streaming analysis. The analysis for the historical are the total amount at the end of each day and the average price per trip.

2. Customers will send data through message brokers/messaging systems which become data stream sources. Discuss and explain the following aspects for the streaming analytics: (i) should the analytics handle keyed or non-keyed data streams for the customer data, and (ii) which types of delivery guarantees should be suitable. (1 point).

The stream is a non-keyed stream, so the original stream will not be split into multiple logical streams and all the windowing logic will be performed by a single task.

The deliveries including the variables: VondorID, tpep_pickup_datetime, tpep_dropoff_datetime, trip_distance and total_amount. This deliveries should be the minum suitable to perform the streaming analytics.

3. Given streaming data from the customer (selected before). Explain the following issues: (i) which types of time should be associated with stream sources for the analytics and be considered in stream processing (if the data sources have no timestamps associated with events, then what would be your solution), and (ii) which types of windows should be developed for the analytics (if no window, then why). Explain these aspects and give examples. (1 point).

In this dataset, we have two different dates, in order to produce analytics. But, in case we don't have any timestamp on the data I will use the ingestion time, when the event entered the platform. In case we can have the event time, it will perform better analytics and close to reality than the ingestion time. For those cases where the communication is weak or problems appear.

The windows should be fixed in order to simplify the platform and the amount of messages from the producer will always be more or less the same amount, because the regulation of taxis at the same time.

4. Explain which performance metrics would be important for the streaming analytics for your customer cases. (1 point).

Two key streaming analytics for the customer could be the amount of different ID in real time as the average time travel for every hour. This two metrics will help the customer to incorporate more taxis in case the average time per hour increases. Or it could save money to the drivers by knowing the demand.

5. Provide a design of your architecture for the streaming analytics service in which you clarify: customer data sources, mysimbdp message brokers, mysimbdp streaming computing service, customer streaming analytics app, mysimbdp-coredms, and other components, if needed. Explain your choices of technologies for implementing your design and reusability of existing assignment works. Note that the result from customerstreamapp will be sent back to the customer in near real-time. (1 point).

Note that questions 1-4 are very much based on the selected datasets and customers. Thus you must give concrete examples based on the data and customer.

Part 3 - Connection (weighted factor for grades = 2)

Notes: no software implementation is required for this part

6. If you would like the analytics results to be stored also into mysimbdp-coredms as the final sink, how would you modify the design and implement this (better to use a figure to explain your design). (1 point)

7. Given the output of streaming analytics stored in mysimbdp-coredms for a long time. Explain a batch analytics (see also Part 1, question 1) that could be used to analyze such historical data. How would you implement it? (1 point)

8. Assume that the streaming analytics detects a critical condition (e.g., a very high rate of alerts) that should trigger the execution of a batch analytics to analyze historical data. How would you extend your architecture in Part 1 to support this (use a figure to explain your work)? (1 point)

9. If you want to scale your streaming analytics service for many customers and data, which components would you focus and which techniques you want to use? (1 point)

I will focus on the windows types, studying and testing where the sliding type or the session-based suits the problem. If I would have more clients Ill focus on the API delivered to the customer and the ingestion system. If the number of clients is the same but the amount of data increases I will focus on the storage and Database systems and the Processing systems.

10. Is it possible to achieve end-to-end exactly once delivery in your current implementation? If yes, explain why. If not, what could be conditions and changes to make it happen? If it is impossible to have end-to-end exactly once delivery in your view, explain why. (1 point)