Fall 2024 CS 224P **Instructor: Prof. Chen Li** **TAs: Xiaozhen Liu, Jiadong Bai**

Homework Assignment #6

(Flink)

After completing previous assignments successfully, your reputation as a ZotMusic Data Engineer has reached new heights! This time, the startup has asked you to dive into real-time stream processing with Flink. Your goal is to apply your Flink skills to analyze music streaming data in real-time, extract insights, and answer key questions about the ZotMusic data.

You will be working with several JSON files containing the following collections: Records, ReviewLikes, Reviews, Sessions, and Users. Each file contains JSON objects, one per line (i.e. in the “JSON Lines” format).

Get Ready…

Start by reading through the HW6 setup instructions carefully and ensure that you have PyFlink installed and configured. You will need a Jupyter Notebook to execute the Python code using PyFlink.

Get (Data) Set…

Download the hw6_template.ipynb and zot-music-dataset-hw6.zip.

Go...!

You will start by creating a Flink environment, reading the JSON data as a DataStream, and implementing stream processing tasks.

1. Start by running some initial setup checks and exploring the schema of the provided data files using PyFlink. This will help you understand the structure of the data before diving into more complex queries.

   A. Use the given function read_json_as_datastream to create the datastream for users, and use print() from PyFlink datastream to output the user data. For each user data that comes out of the stream, what is its type? (Hint: use type() to check.)

   B. In the provided codes, the function:

   ```
   def read_json_as_datastream(file_path: str, env:
   StreamExecutionEnvironment)
   ```

   is used to read a given file into a data stream. The class:

```
class JsonObjectMapFunction(MapFunction)
```

is used by the `read_json_as_datastream` as one of the processing steps. What is it for? Write your answer as a comment in the notebook.

C. In 1.A's output, you should see something like:

```
3 > some data here
```

A number with a '>' is annotated before each user data. Explain the meaning of this annotation by referring to the Flink's document. Write your answer as a comment in the notebook.

2. Now that you have explored the basic schema, it's time to start analyzing the data! Answer the following questions using PyFlink.

   A. ZotMusic's extensive library of music includes a diverse collection of albums and singles. The team wants to focus on a specific subset of the records for analysis. Process the input stream of records to keep only **R&B albums with 12 or more songs.** For each qualified album, the output should include its record_id, title, genre, and number_of_songs it contains. You should answer this question **TWICE, using <u>PyFlink DataStream API and Table API</u> respectively.**

   B. ZotMusic is interested in analyzing user engagement by tracking how many likes a specific user has given to reviews on the platform. This information helps the team understand the activity levels of individual users and identify the most active participants in the community. Your task is to count the number of likes given by a user with id "**user_G91ZrXr4QOuT**" to other reviews. For this task, you need to filter the dataset for likes associated with the given user_id, aggregate the count of those likes, and output the total count. You should answer this question **TWICE, using <u>PyFlink DataStream API and Table API</u> respectively.**

   C. ZotMusic aims to enhance user engagement by better understanding how listeners interact with the platform. Your task is to compute the total listening duration (in minutes) for the user with id "**user_wQD7tpOzS564**". The output should include the user_id as well as the total_listening_time in minutes calculated from all the sessions created by this user. You should answer this question **TWICE, using <u>PyFlink DataStream API and Table API</u> respectively.**

   D. Replay counts on ZotMusic serve as a pulse for identifying trending records that capture the community's attention. Your colleagues are particularly interested in tracking the cumulative replay counts of songs to highlight records as they become popular. The task is to identify records whose cumulative replay counts exceed 1,000 and print out their record IDs. Importantly, the output **must follow the time order, i.e., the time of surpassing the 1,000-replay threshold**. You should use the **PyFlink DataStream API** to write this job.

E. ZotMusic's goal is to gain deeper insights into user behavior by analyzing daily engagement. Each day there are some sessions initiated, and the team wants to **identify the dates with 70 or more sessions in real time**. Flink's window operator is perfect for getting such insight in real time. The sessions in the Sessions.jsonl file are already sorted by the initiated_at timestamp, which can be used as the EventTime for Flink to process. Output the date and the number of sessions whenever a qualified date is identified. You should use the **PyFlink window-related DataStream API** to write this job.

What To Turn In

When you have finished your assignment, submit a PDF file with all the answers and code snippets, including the results from running each query. Follow these steps for submission:

1. Keep your answers in a Jupyter Notebook using PyFlink.
2. For each question, provide the full PyFlink code snippet and ensure it runs correctly.
3. Make sure all your code is visible without horizontal scrolling. Break long lines if necessary.
4. Save your .ipynb file, with all of the cells and their results, as a PDF. To do that nicely in Jupyter, if you have LaTex available, you should use File -> Download as -> PDF via LaTex (.pdf). If not, you can use File -> Print Preview and then PDF-print the result. Once you have done this, be sure to double-check to see that all of your code is visible in the resulting PDF file.
5. Double-check your PDF to ensure all code and outputs are visible.
6. Submit the PDF to Gradescope.