Fall 2025: CS 224P, UCI   Instructor: Prof Chen Li   TA's: Sarah Asad, Ali Risheh

Homework Assignment #6

Deadline: 11/30/2025 (Sunday) 11:45PM

After completing previous assignments successfully, your reputation as a ZotPets 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 pets data in real-time, extract insights, and answer key questions about the ZotPets data.

You will be working with several JSON files containing the following collections: PetOwners, Pets, Products, Groomers, Users, Appointments, Services. Each file contains JSONL objects, one per line (i.e. in the “JSON Lines” format).

Get Ready…

We will be using a service called Google Colab**.** You can use your local Notebook environment as well.

Get (Data) Set…

Download the hw6_template.ipynb and zotpets_data.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 data stream for users, and use print() from PyFlink datastream to output the

user data. For each user record that comes out of the stream, what is its type? (Hint: use type() to check.)

b. In the provided code, the following function

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

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

```Python
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. When using Flink’s print() function, earlier versions displayed a prefix such as 1>, indicating the subtask ID that produced the output. Explain what a **subtask** is in Flink’s DataStream API and how it relates to operator parallelism.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. ZotPets keeps data about pets. Your task is to process the pets stream and keep only the pets whose species is “Rabbit” and whose breed is “Golden Retriever”. For each qualifying appointment, output the pet_id, name, and dob. **Complete this task twice: once using the PyFlink DataStream API and once using the PyFlink Table API.**
    b. ZotPets is interested in analyzing how many grooming appointments a specific pet has had. This information helps the team understand engagement levels and identify the most active pet owners on the platform. Your task is to count the total number of grooming appointments taken by a pet with pet_id "pet_011". For this task, you need to filter the appointments dataset for entries associated with the given pet_id, aggregate the number of appointments, and output the final count. **You should answer this question twice, using the PyFlink DataStream API and the PyFlink Table API respectively.**

c. ZotPets aims to better understand customer engagement by analyzing how much pet owners spend on grooming services. Your task is to compute the total amount spent by the pet owner with id "user_001" across all of their pets' appointments.To complete this task, join the owner's pets with their appointments and then join with the services dataset to retrieve each service's price. Sum all prices to produce the final total. The output should include owner_id and total_amount_spent.

d. ZotPets aims to better understand groomer performance by analyzing how much revenue each groomer brings to the platform. Your task is to compute the total revenue (in dollars) generated by the groomer with id "user_002" across all grooming appointments they completed. Your output should include groomer_id and total_revenue.

e. ZotPets aims to understand user reach for each groomer on the platform. Your last task is to compute the number of unique pets who have booked at least one appointment with each groomer. Your output should include groomer_id and unique_pet_owners_served.

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.