Retrieve User Activity Data on an Online Forum Using SQL

95%

Project brief

Courses

Resources

Assessment

40 hoursLast updated on Thursday, October 13, 2022

You’ve just been hired as a junior data analyst at **ChatData**, a popular network of question-and-answer (Q&A) websites in the fields of data analytics, data science and artificial intelligence. They help budding data analysts stay up-to-date with current innovations, find answers to burning questions, and stay active in the data community!

ChatData

Your company logo

Only a few days into your new job at ChatData, you receive an email from Melanie, the lead data analyst, who briefs you on your first task:

From: Melanie
To: Me
Subject: User Activity Dataset

Hey!

Sorry to throw this on you during your first week. We received a pretty urgent request from **Oliver**, our social media manager. He is putting together the roadmap for next year and has some specific questions about **user activity (interaction)** on our sites.

He wants to learn how ChatData sites are being used in the real world to understand which features are useful to the users and what additional features might be worth introducing.

I already found the data set on interactions. I managed to clean and format the data into three separate CSV files attached: posts, comments and users.

To ensure that you understand how this simple database works, could you prepare an entity relationship diagram (ERD) for it? Here is a quick video that explains how to do that.

Then, in order to respond to Oliver, I need you to execute some SQL queries that will help us find the answers to his questions. Jeremy, one of your team members, started working on this before he left on vacation. He created the attached Excel spreadsheet that will guide you through the task. Here is some information he sent:

- Column B is the action column, and it lists the steps you are expected to complete. Each action requires you to write a query.
- Column C provides some hints to help you.

Use this spreadsheet as your requirements document. To complete the assignment, you will be using the **chata_data_student.ipynb** Jupyter Notebook Jeremy also started working in, which you can also find attached to this email. You will create a database using a lightweight database management system called SQLite. Then write and test your queries against the SQLite database directly in the Jupyter Notebook.

The notebook will also prompt you to insert your SQL queries into a table in SQLite. By keeping track of them in SQLite, you will help other data analysts in the company extract similar information from the database in the future. The work you are doing here is highly requested, and we would not want every analyst to redo all the work each time, right? Teamwork is key!

I've also attached ChatData's Data Standards document, which will provide you with some best-practices for this project.

Here is a recap of everything you'll have to do:

- Create an ERD.
- Create an SQLite database and load the data.
- Create single table queries to analyse engagement.
- Create cross table queries to analyse engagement further.
- Check the queries table to ensure you have recorded all the queries.

I will present the results of your work in the chat_Data_Student.ipynb notebook to Oliver in our next meeting.

With all this information, I think you should be good to go!

Thanks again.
Cheers!
Melanie

Attachments:

- Requirements spreadsheet
- Dataset
- chat_data_ipynb notebook
- ChatData_Data_Standards

Task 1: Create the ERD and Database and Load the Data

After reading the email, you grab a coffee and get to work.

First, you create the ERD as requested by Melanie. Next, you open up the spreadsheet and Jupyter Notebook and start going through it, answering the "to-dos" as they are posed. You realize that the sections in the spreadsheet are nicely aligned with the steps in the notebook. Jeremy has done a lot of the initial legwork for you!

At the end of task 1, you should have:

- Created an ERD.
- Created a SQLite database.
- Loaded data into the tables.
- Spent some time thinking about and taking some personal notes on the data structure, how well it works in a relational model, and the security and ethical considerations.

For background on designing a database please refer to the following videos:

- **First Normal Form (1NF) | Database Normalization | DBMS**
- **Second Normal Form (2NF) | Database Normalization | DBMS**
- **Third Normal Form (3NF) | Database Normalization | DBMS**

Task 2: Create Single Table Queries to Analyze Engagement

Next, you decide to work on the Single Table Queries. These will not require any join logic. You realize that you'll be able to use the Single Table Queries section of the spreadsheet as a guide and add the queries to the corresponding section of the Jupyter Notebook.

At the end of task 2 you should have:

- Completed all of the Single Table Queries.

Task 3: Create Cross table queries to Further Analyze Engagement

Your third task is to write and test all of the queries that require joins. You make a mental note not to forget: **a join creates another table in memory**. You use the cross table queries section of the spreadsheet as a guide and add the queries to the corresponding section of the Jupyter Notebook.

This is a more involved section than task 2 as it requires joining the posts, comments and users tables. The joins you will need are always simple inner joins. There is no need to specify left, right or inner joins.

The best resources to help you are **Stack Overflow** and the SQLite documentation.

- **SQLite Documentation**
- **SQLite Tutorial**
- **SQLite Date And Time Functions**

Both SQLite and PostgreSQL (and virtually every other RDB on the planet) use SQL as their DML. However, there are slight differences between them. One big one is the lack of dates datatype in SQLite. You need to use strftime in order to manipulate dates and times.
At the end of task 3, you should have:

- Completed all of the queries that require joins.

Task 4: Check the Queries table

Finally, you are ready to finish up. You look at the queries that you have been adding to the table and feel pretty good that your hard work will be available to your work colleagues over time.

In the Jupyter Notebook, check that all the queries are present in the table and remove any duplicates.

However, you also wonder if you can improve them by making them more readable via formatting, so you have a go at this! Ensure all the queries in the queries table are in upper case.

Finally, you do some queries against the queries table! By having them managed in this way, you discover it's easy to find particular types of queries, such as those that use a GROUP BY clause.

At the end of task 4, you should have:

- All queries working and formatted in the queries table.
- Searched the queries table for particular types of queries.

Task 5: Report back on your process

Melanie is pleased with everything you've done and has told you several times that she is looking forward to seeing the finished notebook!

She sends you an email with a list of questions about how you worked. She wants several decisions documented because Oliver often has a lot of questions, and she also wants to make sure that, as a new hire, you are ready to work more independently in the future. So, she asks you to create a series of slides that answer all of her questions. You read them closely and get to work!

For information on privacy by design, please refer to **UK ICO - data protection by design and default**.

At the end of task 5, you should have:

- Created your presentation, which includes at least one slide for each of the included topics.

Deliverables

1. The ERD.
2. The completed chat_data_student notebook.
3. The presentation that covers all of the items requested by Melanie:
    - Data analysis lifecycle
    - Requirements
    - Tools
    - Quality
    - Security, ethics and legislation

To make it easier for your work to be reviewed by the jury, upload all the project deliverables to the platform in a zip folder called ‘Project_title_LastName_FirstName’. Use the following naming convention for each of your deliverable: LastName_FirstName_number of deliverable_name of deliverable__your start date of project. This is how it should be named:

- LastName_FirstName_1_ERD_mmyyyy
- LastName_FirstName_1_notebook_mmyyyy
- etc.

For example, the deliverable here would be named: Smith_Mary_1_ERD_042022

Project Presentation

Your oral presentation will last 30 minutes, during which your assessor will play the role of Melanie. The assessor will challenge your decisions, so be prepared to defend your work. The presentation will be structured as follows:

- **Presentation of deliverables (10 minutes).**

- **Review of notebook with all to dos completed (10 minutes).**
- **Discussion (5 minutes):**
    - Playing the role of Melanie, the assessor will ask you questions about your methodology and your deliverables, such as:
        - What did you learn about the data analysis lifecycle?
        - Which stages are predominant in this project?
        - What issues arose around data privacy and which types of data were in this project?
        - Can you explain the business drivers for relational database technology and some of the design issues?
- **Debrief (5 minutes):**
    - At the end of the session, the assessor will stop playing the role of Melanie so you can debrief together.

Your presentation should last 20 minutes (+/- 5 minutes). Respecting presentation time requirements is important in professional environments. In consequence, if your presentation is under 15 minutes or over 25 minutes, you may be asked to redo the assessment.

Skills

- Describe the data life cycle and the steps involved in carrying out routine tasks
- Apply organizational architecture requirements to data analysis activities
- Describe fundamentals of data structures, database design, implementation and maintenance
- Describe the fundamentals of organisational data architecture
- Identify and escalate quality risks in data analysis with suggested mitigation/resolutions
- Use data systems securely to meet requirements and in line with procedures and legislation
- Select and apply the most appropriate data tools to achieve the best outcome
- Undertake customer requirements analysis and implement findings in planning and outputs