Unit 9 Homework: Employee Database

Background

It’s a beautiful spring day, and it’s been two weeks since you were hired as a new data engineer at Pewlett Hackard. Your first major task is a research project on employees of the corporation from the 1980s and 1990s. All that remains of the database of employees from that period are six CSV files.

In this assignment, you will design the tables to hold data in the CSVs, import the CSVs into a SQL database, and answer questions about the data. In other words, you will perform **data modeling**, **data engineering**, and **data analysis**.

Before You Begin

1. Create a new repository for this project called `sql-challenge`. **Do not add this homework assignment to an existing repository**.

2. Clone the new repository to your computer.

3. Inside your local Git repository, create a directory for the SQL challenge. Use a folder name that corresponds to this assignment, like `EmployeeSQL`.

4. Add your files to this folder.

5. Push these changes to GitHub.

Instructions

This assignment is divided into three parts: data modeling, data engineering, and data analysis.

Data Modeling

Inspect the CSVs and sketch out an ERD of the tables. Feel free to use a tool like http://www.quickdatabasediagrams.com.

Data Engineering

- Use the provided information to create a table schema for each of the six CSV files. Remember to specify data types, primary keys, foreign keys, and other constraints.

  - For the primary keys, verify that the column is unique. Otherwise, create a composite key, which takes two primary keys to uniquely identify a row.

  - Be sure to create tables in the correct order to handle foreign keys.

- Import each CSV file into the corresponding SQL table.

  > **Hint:** To avoid errors, be sure to import the data in the same order that the tables were created. Also remember to account for the headers when importing.

Data Analysis

Once you have a complete database, perform these steps:

1. List the following details of each employee: employee number, last name, first name, sex, and salary.

2. List first name, last name, and hire date for employees who were hired in 1986.

3. List the manager of each department with the following information: department number, department name, the manager's employee number, last name, first name.

4. List the department of each employee with the following information: employee number, last name, first name, and department name.

5. List first name, last name, and sex for employees whose first name is "Hercules" and last names begin with "B."

6. List all employees in the Sales department, including their employee number, last name, first name, and department name.

7. List all employees in the Sales and Development departments, including their employee number, last name, first name, and department name.

8. List the frequency count of employee last names (i.e., how many employees share each last name) in descending order.

Bonus (Optional)

As you examine the data, you begin to suspect that the dataset is fake. Maybe your boss gave you spurious data in order to test the data engineering skills of a new employee? To confirm your hunch, you decide to create a visualization of the data to present to your boss. Follow these steps:

1. Import the SQL database into Pandas. (Yes, you could read the CSVs directly in Pandas, but you are, after all, trying to prove your technical mettle.) This step may require some research. Feel free to use the following code to get started. Be sure to make any necessary modifications for your username, password, host, port, and database name:

   ```
   from sqlalchemy import create_engine
   engine = create_engine('postgresql://localhost:5432/<your_db_name>')
   connection = engine.connect()
   ```

   - Consult the SQLAlchemy documentation for more information.

   - If you’re using a password, do not upload your password to your GitHub repository. Review this video and the GitHub website for more information.

2. Create a histogram to visualize the most common salary ranges for employees.

3. Create a bar chart of average salary by title.

Submission

- Create an image file of your ERD.

- Create a `.sql` file of your table schemata.

- Create a `.sql` file of your queries.

- (Optional) Create a Jupyter notebook of the bonus analysis.

- Create and upload a repository with the above files to GitHub and post a link on BootCamp Spot.

- Ensure your repository has regular commits and a thorough README.md file

Rubric

Unit 9 Homework Rubric

References

Mockaroo, LLC. (2021). Realistic Data Generator. https://www.mockaroo.com/