Data Scientist Assignment: Building a GPT-Based Tool for Boiler Data

Objective: You are tasked with building a GPT-based tool that will allow users to query a database and retrieve insights about coal-fired boiler operations. The tool should be able to answer questions based on data provided in an Excel file, which you will save in a SQL database.

Steps to Complete the Assignment:

1. **Data Preprocessing:**
    - You will be provided with an Excel file containing coal-fired boiler data.
    - Save the data in an SQL database (e.g., MySQL, PostgreSQL, or SQLite).
    - Ensure proper data types and normalization of the data where needed (e.g., date fields, numerical fields, etc.).
2. **Data Cleaning and Transformation:**
    - Clean the data to handle missing values, outliers, and other inconsistencies.
    - Transform the data into a format that can be used to answer specific queries, such as aggregating daily, weekly, or monthly statistics.
    - Ensure the data is ready for use in building a question-answering system.
3. **Dataset Creation for GPT Model:**
    - Create a dataset that will be used to train or fine-tune a GPT-based model. This dataset should consist of question-answer pairs based on the provided data.
    - Example questions:

        - “What is today’s generation?”
        - “Which loss is impacting efficiency?”
        - “What was my steam temperature average for the last week?”
    - Ensure that the dataset includes a variety of questions, and the correct answers derived from the SQL database.

4. **Model Implementation:**
    - Choose an appropriate GPT or language model (you can use Hugging Face Transformers, OpenAI API, or any other open-source models).
    - Fine-tune the model using the dataset you have created. If fine-tuning is not feasible, use a pre-trained model and provide a way to retrieve answers based on the SQL data.

5. **Tool Development:**
    - Develop a simple interface (e.g., using Streamlit or Flask) that allows users to input questions.
    - The system should query the database, process the data, and provide accurate answers to user queries.
    - Example queries include:
        - “What is my steam temperature average for last week?”
        - “Which loss is impacting the efficiency the most?”
        - “What was the generation today?”

6. **Evaluation and Reporting:**
    - Evaluate the performance of the GPT model in answering the questions correctly.

- Provide a report and PPT detailing the methodology, challenges, and insights from the project. Also, include any assumptions made during the development.

Deliverables:

1. Python code for data preprocessing, model training (or fine-tuning), and building the question-answering interface.
2. The SQL database containing the data.
3. A report explaining the approach, challenges, and the results.

Bonus Points:

- Integrating real-time data updates into the model.
- Using advanced techniques like few-shot learning or zero-shot learning.