PROJECT OVERVIEW: SUMMARIZING GITHUB SOURCE CODE WITH VERTEX AI LLM AND BIGQUERY

objective

This project explores how to perform automatic summarization of source code from GitHub repositories and identify the programming language in each repository using Vertex AI's Large Language Model (LLM). The implementation is done on Google Cloud, where Vertex AI's PaLM API is connected to BigQuery for efficient data processing.

Key Components and Workflow

1. **BigQuery Dataset Creation**:
   - Set up a BigQuery dataset to store the model that will interface with Vertex AI for text generation tasks.
   - This dataset is used to host functions, manage data, and perform SQL queries as part of the summarization process.
2. **BigQuery Model with Vertex AI as Remote Function**:
   - Create a BigQuery model that leverages Vertex AI's PaLM API through a remote function.
   - This integration allows BigQuery to handle complex language tasks, such as summarizing code and detecting programming languages, using Vertex AI's LLM.
3. **Establishing Connection between BigQuery and Vertex AI**:
   - Set up an external connection to link BigQuery with Vertex AI, enabling direct access to the LLM for executing text-based tasks.
   - This connection facilitates real-time language processing and summarization within the BigQuery environment.

Tasks Completed

- **Data Summarization**: The LLM generates concise summaries of source code in various repositories, making it easier to understand the purpose and function of each repo without examining each file individually.
- **Language Identification**: The model can detect and label the programming language used in the GitHub repositories, which is useful for categorization and metadata creation.

Skills Demonstrated

- Working knowledge of **BigQuery and SQL** for managing data storage and performing queries.
- Experience using **Vertex AI and the paLM API** to integrate advanced machine learning functions with SQL operations.
- Understanding of **external connections** in Google Cloud to facilitate cross-platform workflows.

Outcome

The project successfully automates the summarization and language identification process for GitHub source code, achieving a 100% grade in this lab. It showcases the potential of combining machine learning capabilities with cloud-based data management systems to streamline code analysis and improve project documentation.