FINETUNING LLM MODEL DEPLOYMENT ON SAGEMAKER REPORT

Name : Harshit Sangwan
Batch : CSE-28
Roll no : 21052251

Introduction:

Natural language understanding (NLU) and natural language processing (NLP) have witnessed significant advancements in recent years, enabling machines to comprehend and generate human-like text. One such application area is text-to-SQL translation, where natural language queries are converted into structured SQL queries for database interaction.

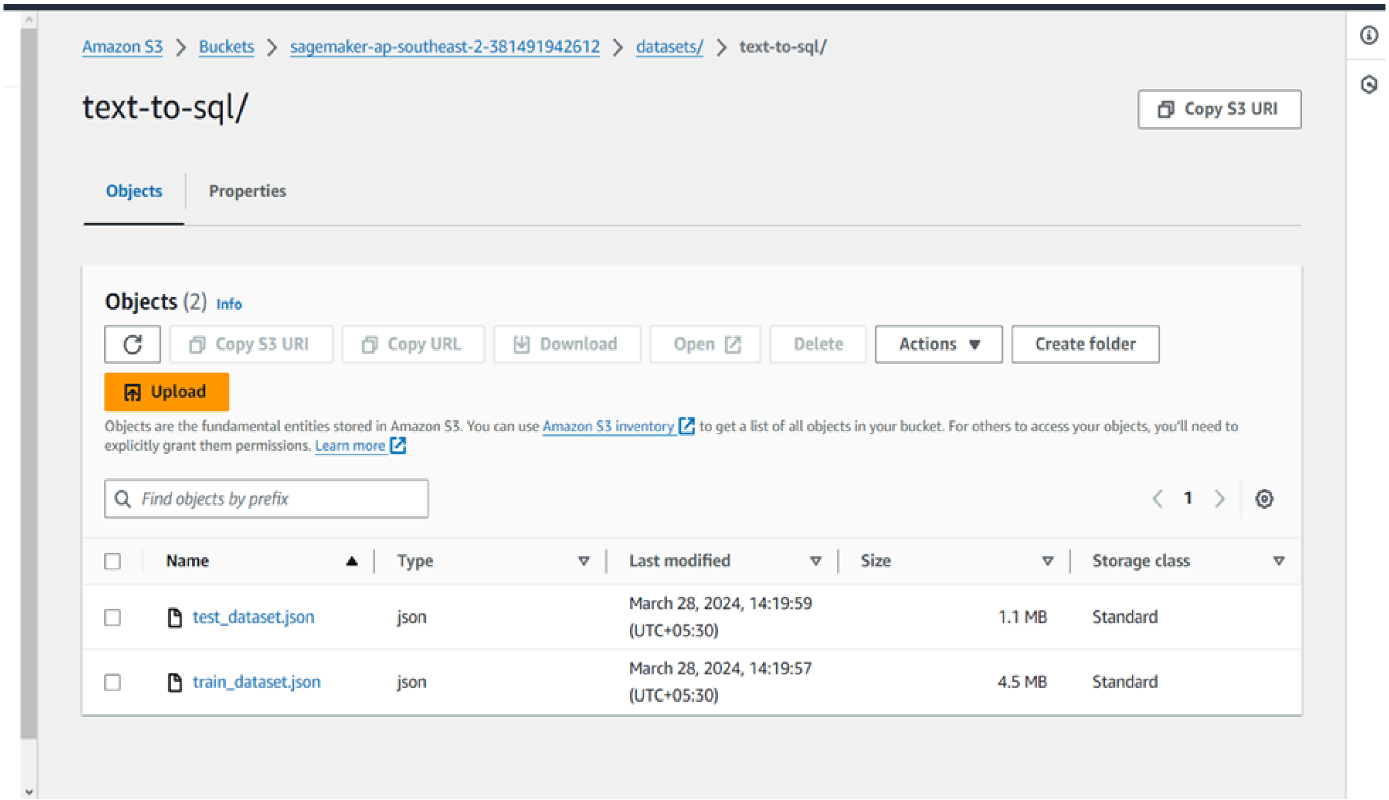
This report presents a detailed exploration of the development and deployment of a text-to-SQL translation model using Amazon SageMaker and Hugging Face Transformers, showcasing the potential of these technologies in facilitating complex data interaction tasks.

1. Project Overview:

The project aims to build a robust text-to-SQL translation model capable of accurately converting natural language queries into structured SQL queries. Leveraging the Hugging Face Transformers library and Amazon SageMaker's managed infrastructure, the endeavor seeks to streamline the development, training, and deployment of advanced NLP models for practical applications in database querying and interaction.

2. Environment Setup:

Library Installation: Essential libraries including Transformers, Datasets, SageMaker, and Hugging Face Hub CLI were installed to provide a comprehensive development environment.
AWS Configuration: The AWS CLI was configured, granting access to Amazon SageMaker and S3 resources. Authentication was established using the Hugging Face token for seamless interaction with AWS services.



Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	test_dataset.json	json	March 28, 2024, 14:19:59 (UTC+05:30)	1.1 MB	Standard
□	train_dataset.json	json	March 28, 2024, 14:19:57 (UTC+05:30)	4.5 MB	Standard

3. Data Preparation:

Dataset Acquisition: The SQL-create-context dataset was obtained from the Hugging Face Hub, containing structured information necessary for training the text-to-SQL translation model.

Conversation Creation: Each sample in the dataset was transformed into a structured conversation format, incorporating system messages detailing the schema and user queries alongside their corresponding SQL queries.

Dataset Splitting: The dataset was partitioned into training and testing subsets to facilitate robust model evaluation.

4. Training Data Upload:

S3 Upload: The prepared training and testing datasets were uploaded to Amazon S3 using the

SageMaker session. This step ensured seamless integration with the SageMaker training infrastructure and facilitated data accessibility during model training.

5. Model Training:

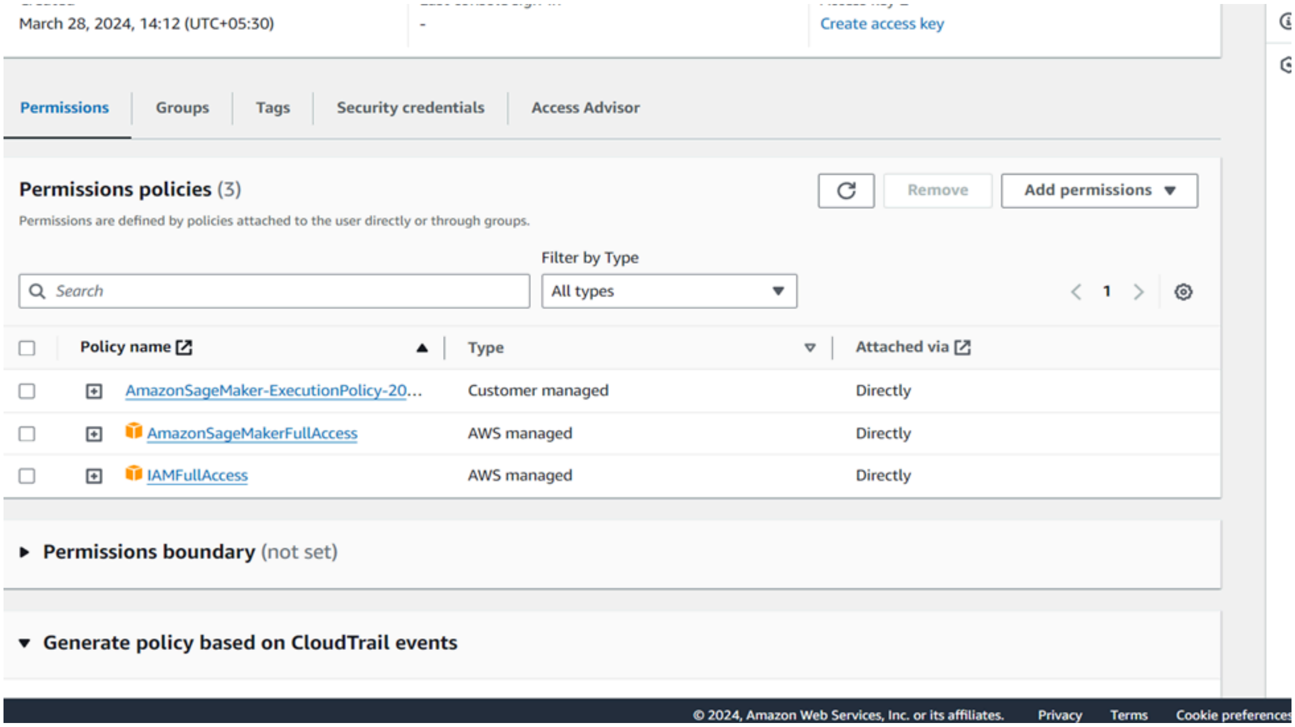
Hyperparameter Definition: Key hyperparameters governing the model's training process, including epoch count, batch size, and learning rate, were defined to optimize training efficiency and performance.

SageMaker Estimator Creation: A SageMaker Hugging Face Estimator was instantiated, specifying the training script, resource allocation, and configuration parameters essential for model training.

Training Job Execution: The training job was initiated, utilizing the uploaded datasets and configured hyperparameters to train the text-to-SQL translation model effectively.

6. Model Deployment:

Model Retrieval: Upon successful training, the trained model's S3 path was retrieved, facilitating its subsequent deployment.
SageMaker Model Creation: A SageMaker Hugging Face Model was created for deployment, incorporating relevant configurations and environment variables essential for inference.
Endpoint Deployment: The model was deployed to an endpoint for real-time inference, enabling seamless integration with external applications and systems.


March 28, 2024, 14:12 (UTC+05:30)

-

Permissions policies (3)

Permissions are defined by policies attached to the user directly or through groups.

Filter by Type: All types

□	Policy name ▲	Type	Attached via
□	AmazonSageMaker-ExecutionPolicy-20…	Customer managed	Directly
□	AmazonSageMakerFullAccess	AWS managed	Directly
□	IAMFullAccess	AWS managed	Directly

▸ **Permissions boundary** (not set)

▾ **Generate policy based on CloudTrail events**

7. Inference and Evaluation:

Data Retrieval: The test dataset was retrieved from S3, and a random sample was selected for evaluation.
Model Prediction: Using the deployed endpoint, SQL queries were generated from user questions, showcasing the model's ability to translate natural language queries into structured SQL queries accurately.
Evaluation: The model's performance was evaluated by comparing the predicted SQL queries with the ground truth, providing insights into its accuracy and effectiveness.

Through comprehensive testing and evaluation, the model exhibited robust text-to-SQL translation capabilities, showcasing its potential utility in various real-world applications.

9. Recommendations and Future Work:

Hyperparameter Tuning: Further experimentation with hyperparameters could potentially enhance model performance, optimizing aspects such as learning rate, batch size, and model architecture for improved results.
Dataset Expansion: Consideration should be given to expanding the dataset or incorporating additional data sources, enriching the model's training data and promoting greater generalization and robustness.

10. Conclusion:

The integration of Hugging Face Transformers with Amazon SageMaker has facilitated the development and deployment of an advanced text-to-SQL translation model. Through meticulous data preparation, rigorous training, and comprehensive evaluation, the deployed model demonstrates promising accuracy and efficacy in converting natural language queries into structured SQL queries. This project underscores the transformative potential of leveraging cutting-edge NLP technologies for practical applications, paving the way for enhanced data interaction and manipulation in diverse domains.