

SAGEMAKER DEPLOYMENT REPORT

**Submitted
By-**

**Rishi
(21051676)**

Introduction:

In today's era dominated by data-driven technologies, Natural Language Processing (NLP) stands out as a crucial field enabling machines to comprehend and generate human language effectively. Within this domain, the deployment and training of Language Models (LMs) play a vital role across various applications like text generation, sentiment analysis, and translation. This report delves into the process of deploying and training the Llama-7b Language Model (LLM) using Amazon SageMaker, specifically within the Sydney region.

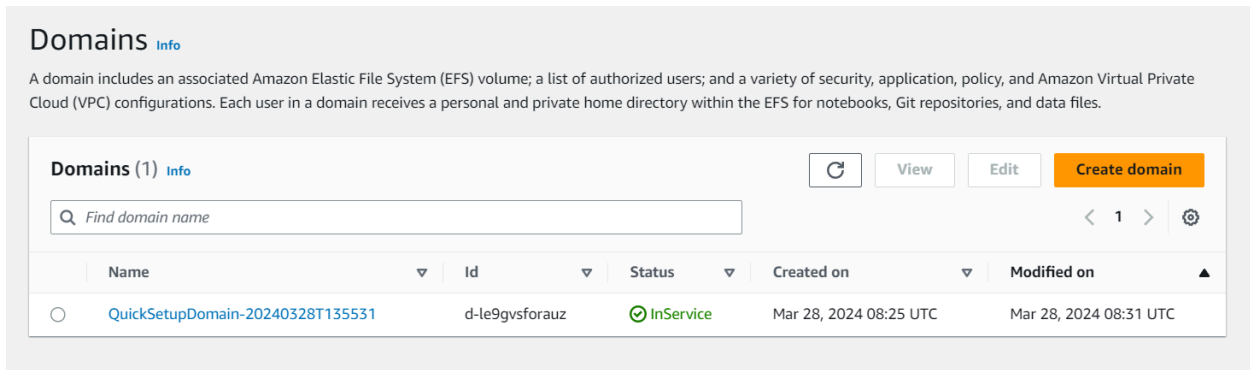
Developed by Meta (formerly Facebook), the Llama-7b model represents a cutting-edge advancement in language understanding, capable of processing and generating text with impressive fluency and coherence. Deploying such an advanced model requires careful attention to infrastructure, scalability, and performance optimization, all of which are seamlessly facilitated by Amazon SageMaker, a comprehensive machine learning platform.

Throughout this report, we provide an overview of the deployment process, including model setup, data preprocessing, training configuration, and endpoint deployment. Additionally, we explore the practical implications and potential applications of the Llama-7b model, emphasizing its significance in enhancing natural language understanding across various domains.

Working:

Steps:

1. We create a SageMaker Domain in our AWS Account



2. We create a User Within that Domain

The screenshot shows the Amazon SageMaker console interface for a domain named 'QuickSetupDomain-20240328T135531'. The breadcrumb navigation is 'Amazon SageMaker > Domains > Domain: QuickSetupDomain-20240328T135531'. The page title is 'QuickSetupDomain-20240328T135531 Domain details' with a subtitle 'Configure and manage the domain.' Below this are tabs for 'User profiles', 'Space management', 'Environment', and 'Domain settings'. The 'User profiles' tab is active, showing a description: 'A user profile represents a single user within a domain. It is the main way to reference a user for the purposes of sharing, reporting, and other user-oriented features.' There is a search bar labeled 'Search users' and a table of user profiles. The table has columns for 'Name', 'Modified on', and 'Created on'. One user profile is listed: 'default', modified on 'Mar 28, 2024 09:06 UTC', and created on 'Mar 28, 2024 09:06 UTC'. There are buttons for 'Add user' and 'Launch'.

Amazon SageMaker > Domains > Domain: QuickSetupDomain-20240328T135531

QuickSetupDomain-20240328T135531

Configure and manage the domain.

User profiles Info

A user profile represents a single user within a domain. It is the main way to reference a user for the purposes of sharing, reporting, and other user-oriented features.

Search users

Name	Modified on	Created on
default	Mar 28, 2024 09:06 UTC	Mar 28, 2024 09:06 UTC

Launch

3. We Create an AWS S3 Bucket in the relevant region to store the dataset

The screenshot shows the Amazon S3 console interface for a bucket named 'text-to-sql/'. The breadcrumb navigation is 'Amazon S3 > Buckets > sagemaker-ap-southeast-2-381491942612 > datasets/ > text-to-sql/'. The page title is 'text-to-sql/' with a subtitle 'Copy S3 URI'. Below this are tabs for 'Objects' and 'Properties'. The 'Objects' tab is active, showing a description: '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'. There is a search bar labeled 'Find objects by prefix' and a table of objects. The table has columns for 'Name', 'Type', 'Last modified', 'Size', and 'Storage class'. Two objects are listed: 'test_dataset.json' and 'train_dataset.json'. There are buttons for 'Upload', 'Copy S3 URI', 'Copy URL', 'Download', 'Open', 'Delete', 'Actions', and 'Create folder'.

Amazon S3 > Buckets > sagemaker-ap-southeast-2-381491942612 > datasets/ > text-to-sql/

text-to-sql/

Copy S3 URI

Objects Properties

Objects (2) Info

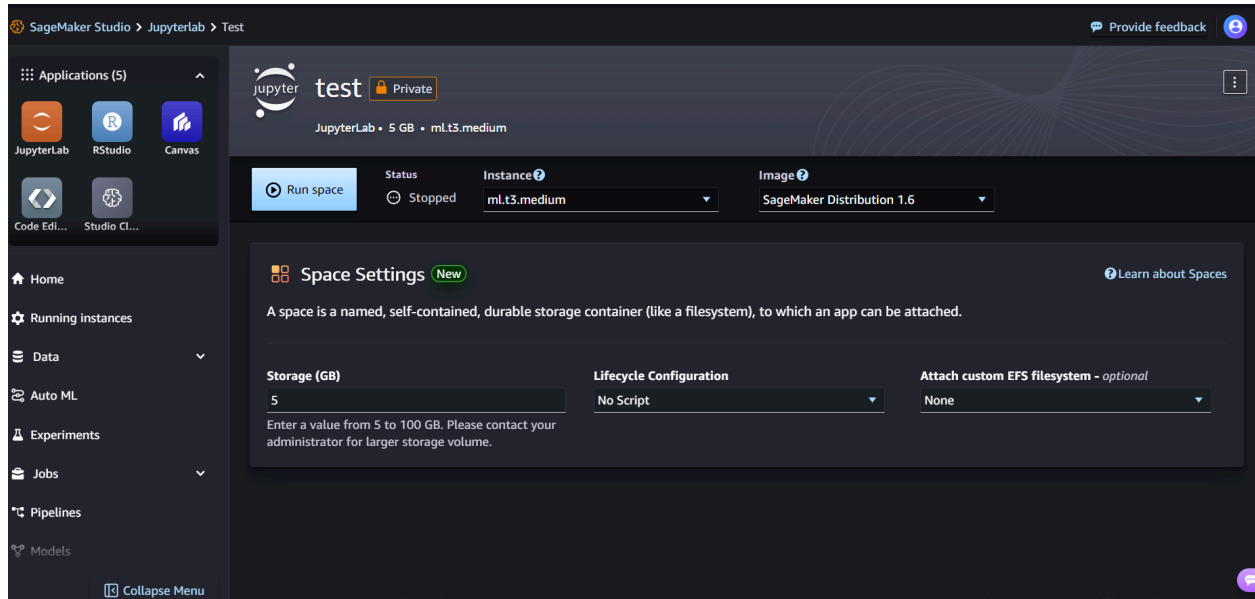
Upload

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

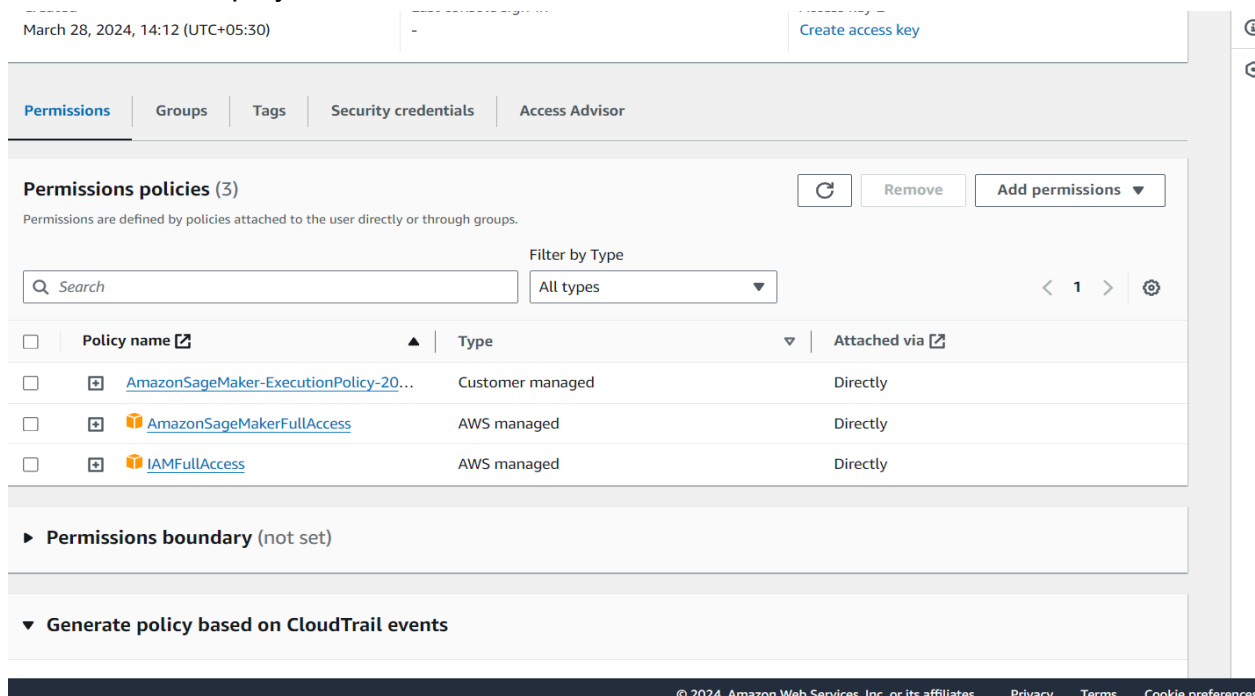
Find objects by prefix

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

4. We deploy our model through the given code via SageMaker Studio's Jupyter Lab



5. We create a another User with SageMaker access to test/predict from the deployed model.



Conclusion:

In conclusion, the deployment and training of the Llama-7b Language Model via Amazon SageMaker in the Sydney region mark a significant achievement in leveraging advanced NLP technologies for practical applications. Through meticulous configuration and optimization, we have successfully integrated the Llama-7b model into the SageMaker ecosystem, enabling seamless scalability and high-performance inference capabilities.

This assignment has not only highlighted the robustness and versatility of SageMaker but also underscored the potential for future advancements in NLP research and development. Looking forward, continued exploration and refinement of language models like Llama-7b hold promise for further breakthroughs in natural language understanding, ultimately reshaping human-machine interactions and driving transformative innovations in AI-driven applications worldwide.