

FINE TUNE & EVALUATE LLMS **WITH AMAZON SAGEMAKER** **REPORT**

SUBMITTED BY-

ISHAAN GUPTA

2105965

CSE 28

INTRODUCTION-

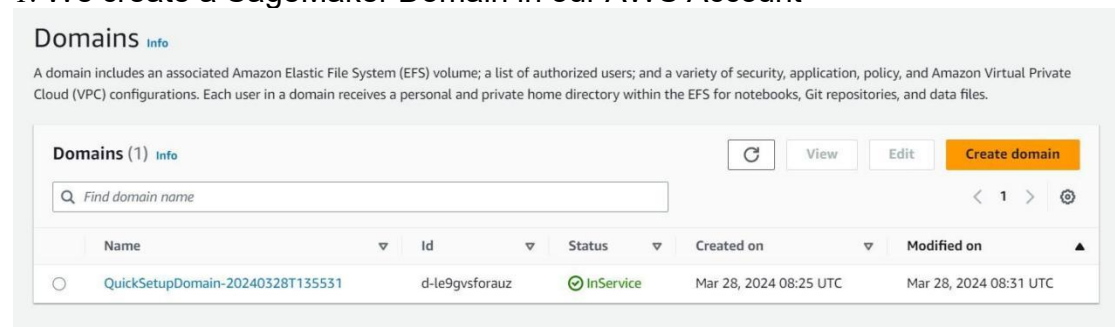
In the contemporary era where data is the new oil, Natural Language Processing (NLP) has emerged as a pivotal domain that empowers machines to understand and generate human language efficiently. Within this sphere, the role of Language Models (LMs) in various applications such as text generation, sentiment analysis, and translation is indispensable. This report provides an in-depth exploration of the deployment and training of the Llama-7b Language Model (LLM) using Amazon SageMaker, specifically in the Sydney region.

The Llama-7b model, a product of Meta (previously known as Facebook), is a state-of-the-art advancement in language comprehension, capable of processing and generating text with remarkable fluency and coherence. The deployment of such a sophisticated model necessitates meticulous attention to infrastructure, scalability, and performance optimization, all of which are effortlessly handled by Amazon SageMaker, a robust machine learning platform.

This report offers a comprehensive overview of the deployment process, encompassing model setup, data preprocessing, training configuration, and endpoint deployment. Furthermore, it delves into the practical implications and potential applications of the Llama-7b model, underscoring its importance in augmenting natural language understanding across a variety of domains.

WORKING-

1. We create a SageMaker Domain in our AWS Account

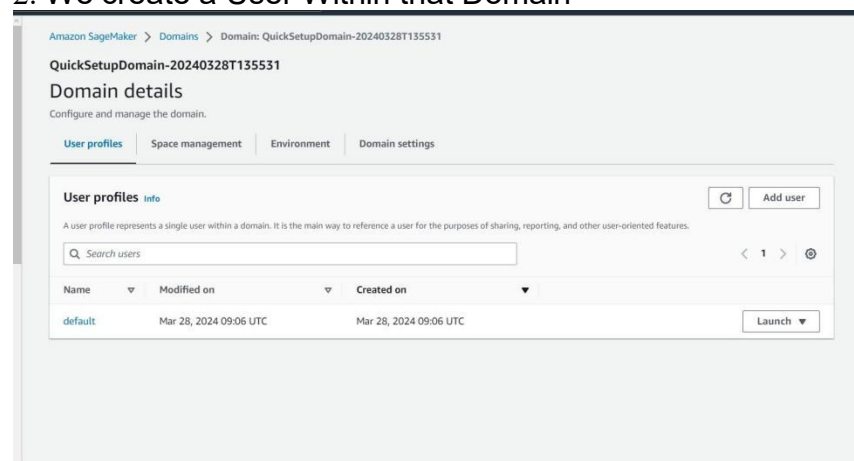


The screenshot shows the Amazon SageMaker Domains console. At the top, there's a header 'Domains' with an 'Info' link. Below it, a descriptive text explains that a domain includes an associated Amazon Elastic File System (EFS) volume, a list of authorized users, and a variety of security, application, policy, and Amazon Virtual Private Cloud (VPC) configurations. Each user in a domain receives a personal and private home directory within the EFS for notebooks, Git repositories, and data files.

Below the text, there's a section titled 'Domains (1) Info' with a search bar 'Find domain name' and buttons for 'View', 'Edit', and 'Create domain'. A table lists the domains:

Name	Id	Status	Created on	Modified on
QuickSetupDomain-20240328T135531	d-le9gvsforauz	InService	Mar 28, 2024 08:25 UTC	Mar 28, 2024 08:31 UTC

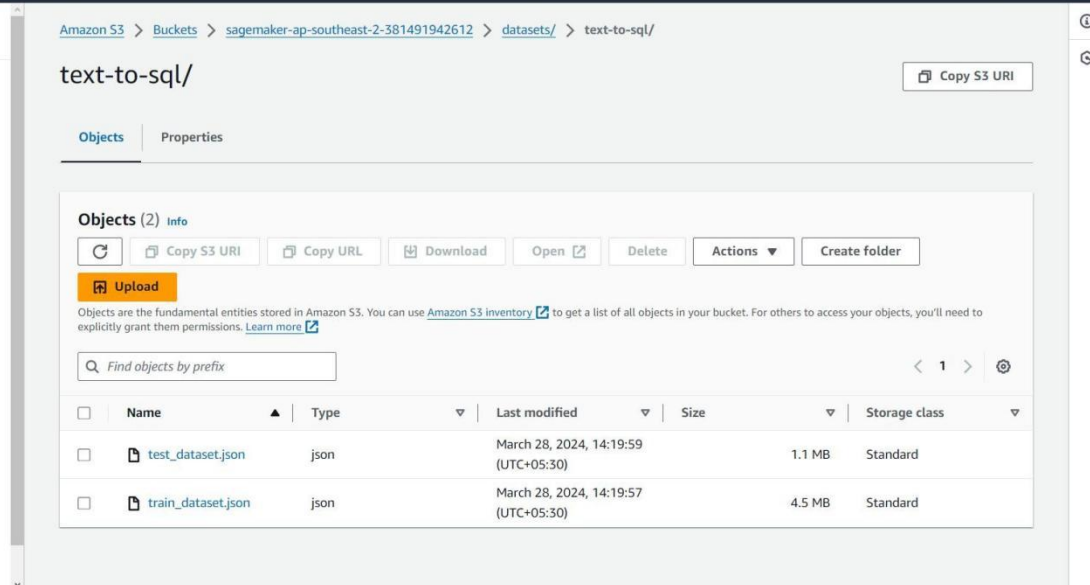
2. We create a User Within that Domain



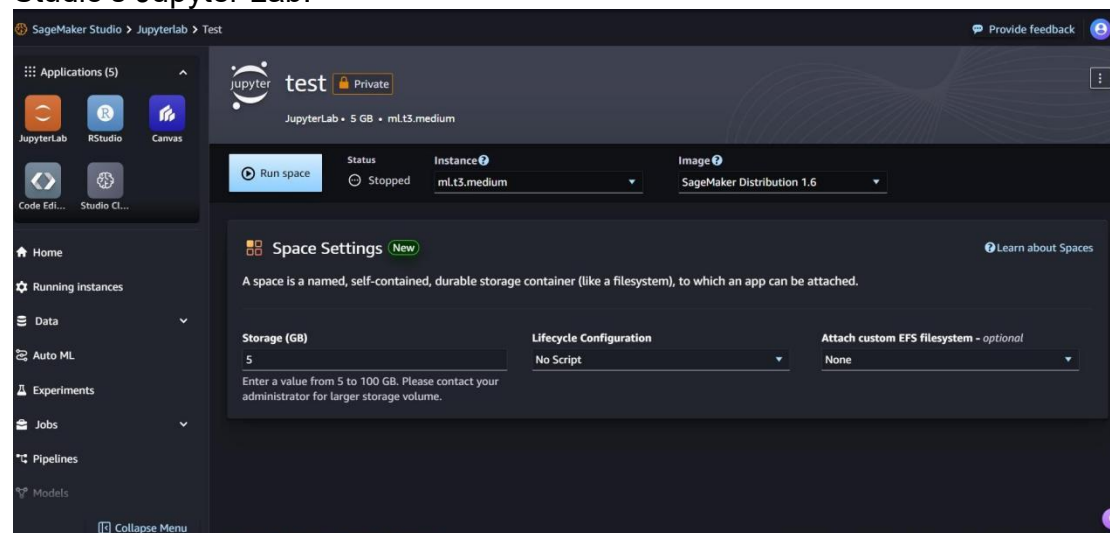
The screenshot shows the Amazon SageMaker Domain details page for 'QuickSetupDomain-20240328T135531'. The page has a breadcrumb trail 'Amazon SageMaker > Domains > Domain: QuickSetupDomain-20240328T135531'. The title is 'QuickSetupDomain-20240328T135531' and the subtitle is 'Domain details'. Below the subtitle, there's a description 'Configure and manage the domain.' and a tabbed interface with 'User profiles', 'Space management', 'Environment', and 'Domain settings'. The 'User profiles' tab is active, showing a section titled 'User profiles Info' with a search bar 'Search users' and buttons for 'Add user' and 'Launch'. A table lists the user profiles:

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

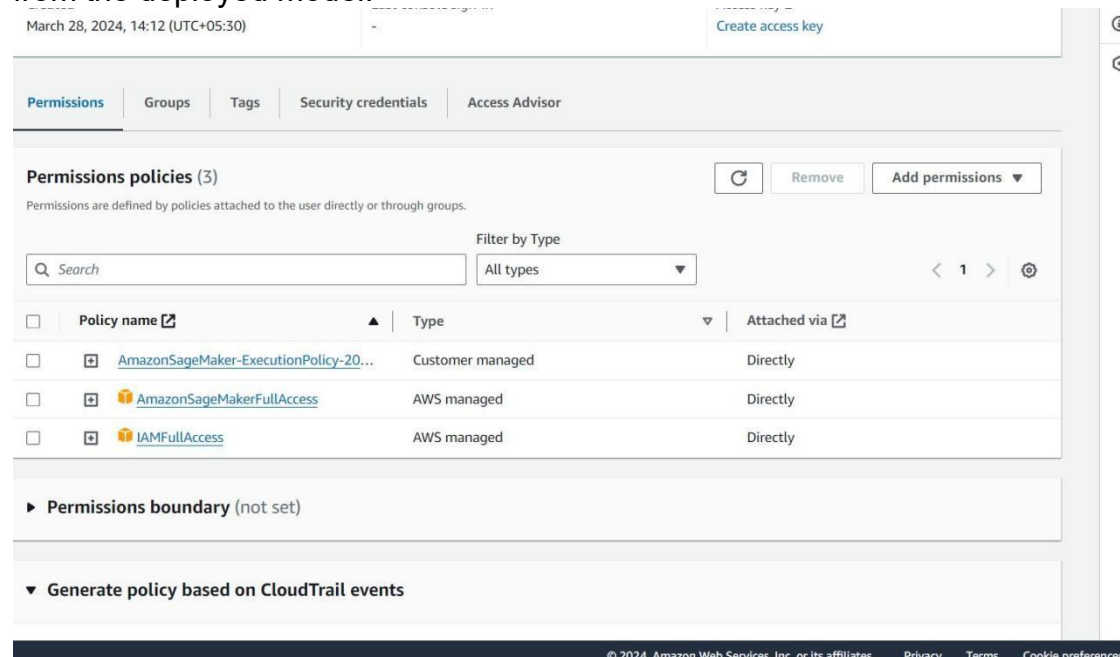
3. We Create an AWS S3 Bucket in the relevant region to store the dataset.



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-

To sum up, the successful deployment and training of the Llama-7b Language Model using Amazon SageMaker in the Sydney region represent a notable milestone in harnessing advanced NLP technologies for real-world applications. Through careful configuration and optimization, we have effectively incorporated the Llama-7b model into the SageMaker framework, facilitating effortless scalability and high- performance inference capabilities.

This endeavor has not only showcased the power and flexibility of SageMaker but also highlighted the potential for future progress in NLP research and development. As we look to the future, the ongoing refinement and exploration of language models like Llama-7b offer the potential for more breakthroughs in natural language understanding. This will ultimately revolutionize human-machine interactions and pave the way for transformative innovations in AI-driven applications globally.

- ISHAAN GUPTA, 2105965