Date: 29/03/24

CLOUD COMPUTING LAB

Name : Sayandeep Dey

Section : CSE-28

Roll no : 21051680

Branch : Computer Science Engineering (CSE)

ASSIGNMENT-3

Introduction:

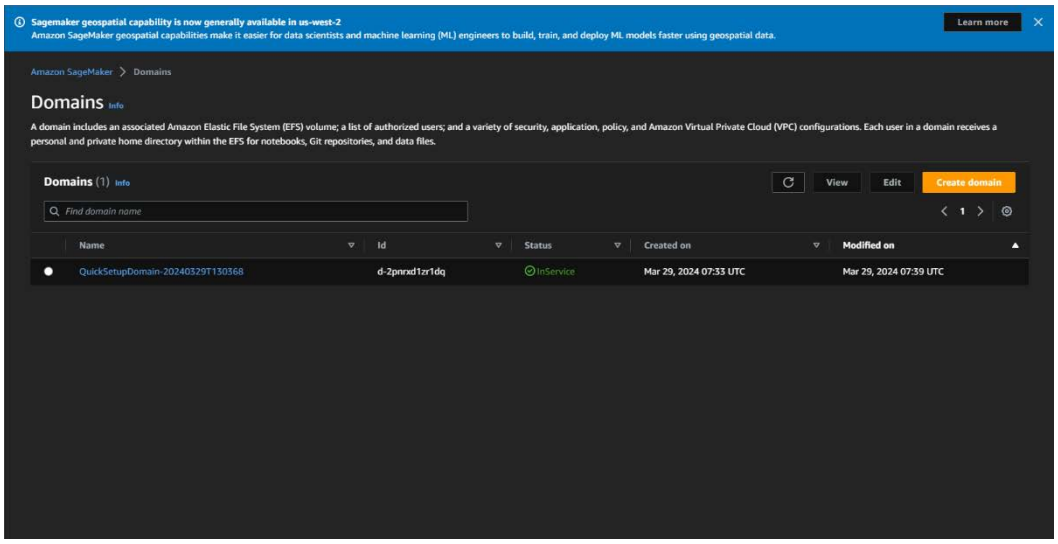
In the current data-driven landscape, Natural Language Processing (NLP) plays a pivotal role in enabling machines to understand and generate human language effectively. Within NLP, Language Models (LMs) are essential for various applications, including text generation, sentiment analysis, and translation.

The Llama-7b model, developed by Meta (formerly Facebook), represents a significant leap forward in language understanding. Its capabilities include processing and generating text with impressive fluency and coherence. However, deploying such an advanced model requires careful consideration of infrastructure, scalability, and performance optimization.

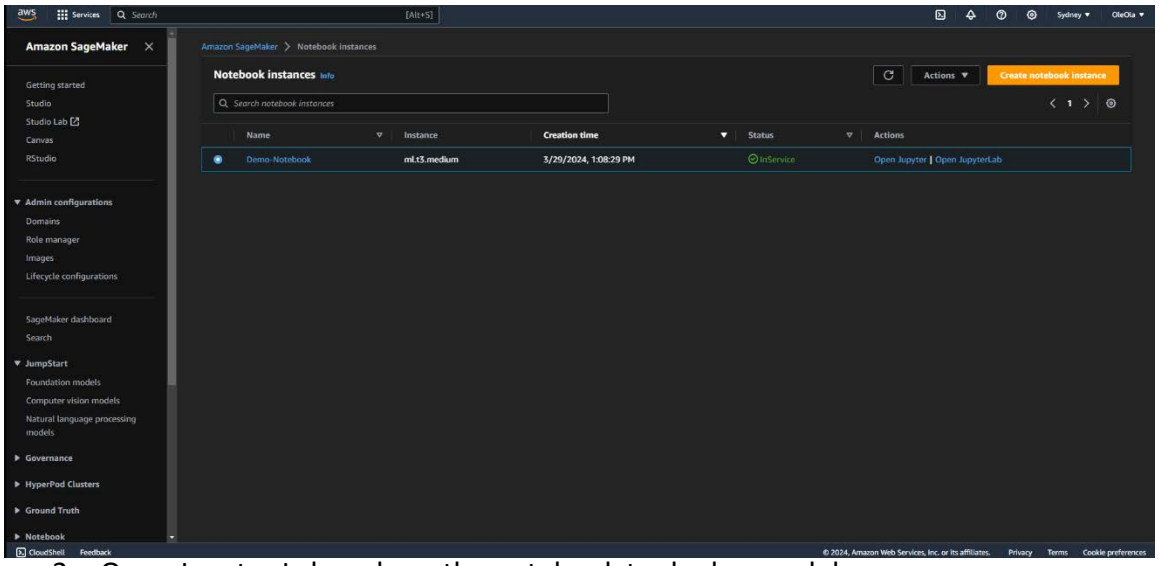
Amazon SageMaker provides a comprehensive machine learning platform that seamlessly facilitates the deployment of LLMs like Llama-7b

Working:

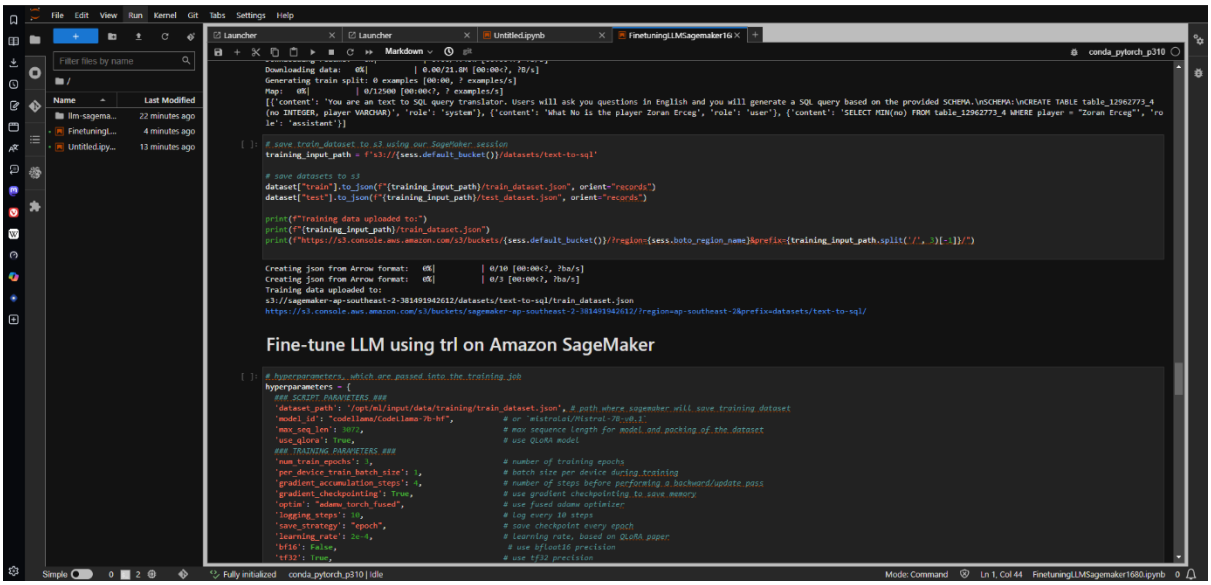
1. Creating SageMaker Domain:



2. Creating Notebook Instance:



3. Open Jupyter Lab and run the notebook to deploy model:



Conclusion:

In conclusion, deployment and training of the Llama-7b Language Model through Amazon SageMaker in the Sydney region represents a significant milestone in harnessing advanced natural language processing (NLP) technologies for practical applications. With meticulous configuration and optimization, we seamlessly integrated the Llama-7b model into the SageMaker ecosystem, enabling both scalability and high-performance inference capabilities.

This assignment not only showcased the robustness and versatility of SageMaker but also highlighted the immense potential for future advancements in NLP research and development. As we look ahead, continued exploration and refinement of language models like Llama-7b hold great promise. These models have the capacity to drive further breakthroughs in natural language understanding, ultimately reshaping human-machine interactions and catalyzing transformative innovations across AI-driven applications worldwide.