Overview

Create an AI agent solution on AWS that hosts a FastAPI endpoint via Agentcore runtime that users can query real-time or historical stock prices and receive stream responses.

Technical Requirements

AWS Service (Minimum):

- AWS Agentcore
- AWS Cognito

Backend:

- Written in Python
- Agentcore runtime hosted via FastAPI
- Setup Cognito user pool for inbound user authorization
- Setup Langfuse cloud free tier for observability
- Uses langgraph for agent orchestration (ReAct type agent)
- A knowledge base with document retrieval of the following documents:
    - https://s2.q4cdn.com/299287126/files/doc_financials/2025/ar/Amazon-2024-Annual-Report.pdf
    - https://s2.q4cdn.com/299287126/files/doc_financials/2025/q3/AMZN-Q3-2025-Earnings-Release.pdf
    - https://s2.q4cdn.com/299287126/files/doc_financials/2025/q2/AMZN-Q2-2025-Earnings-Release.pdf
- Have minimum 2 finance tools for stock retrieval(use yfinance api for price retrieval):
    - retrieve_realtime_stock_price
    - retrieve_historical_stock_price
- Streams events via .astream()
    - Reference:
        - https://langchain-ai.github.io/langgraph/how-tos/streaming/#filter-by-llm-invocation
- Infrastructure written in Terraform
- Event responses must be streamed

User Acceptance Criteria

- Source code in a repository with clear Readme on how to deploy the infrastructure
- A notebook demonstrating invocation of your deployed endpoint (that can be executed by our team) of the following query:
    - What is the stock price for Amazon right now?
    - What were the stock prices for Amazon in Q4 last year?

    - Compare Amazon's recent stock performance to what analysts predicted in their reports
    - I’m researching AMZN give me the current price and any relevant information about their AI business
    - What is the total amount of office space Amazon owned in North America in 2024?
- Notebook should contain screenshots or api response showing Langfuse traces
- Notebook should show user authentication from Cognito user pool