Cloud Computing (ITCS-6190) Project Deliverable 4

Group No: 10
Team members:
Roshna Sanjana Kommareddy
Tharun Abhinav Suraneni
Shashank Bijarapu
Swathi Geerlakunta
Archana Avancha

Optimization, Scalability, and Deployment for Heart Attack Prediction Project

Introduction
This document provides an overview of Phase 4 of the Allen Brain Observatory project, AWS Cost Analysis and Optimization, Pipeline scalability, and Deployment Strategies. The goal is to make the project cost-efficient, scalable, and deployable in the AWS cloud.

AWS Cost Analysis and Optimization
Objective: Optimize your project-related AWS spend without sacrificing performance and dependability.
1. Analyzing AWS Costs
- **Methodology**: Use the AWS Cost Explorer tool to see how your project spends money on AWS. Keep track of your monthly and daily expenses, and look at usage reports for every service.
- **Action Steps**:
    - clearly look into the AWS Cost Explorer dashboard.
    - note down the high spending services and examine weather they are in line with your project goal.
    - examine the costs that occur unwantedly..
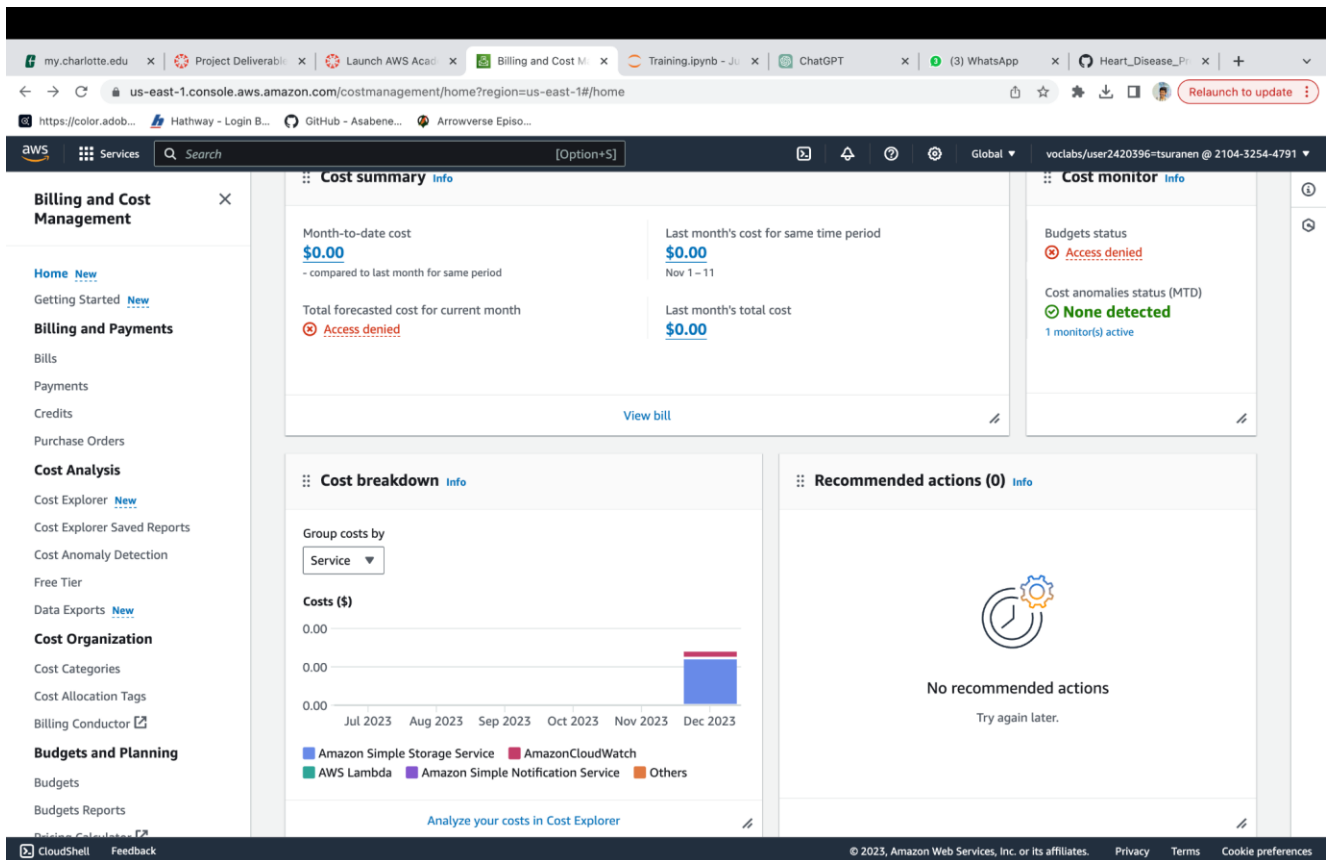
2. Cost Optimization Strategies

- **Right-Sizing Resources**:Evaluate the EC2 instances, RDS databases and other AWS services used in the project based on their usage patterns. Resize them to meet actual workload requirements.
- **Reserved Instances and Savings Plans**: For consistent and predictable workloads such as database instances, consider purchasing reserved instances or economy plans.
- **Spot Instances**:Deploy EC2 Spot instances for all flexible, non-critical processing tasks such as batch jobs, significantly reducing computing costs.
- **Auto Scaling**: combine auto-scaling to dynamically adjust capacity and ensure cost-effective handling of workload fluctuations.
- **S3 Lifecycle Management**: Establish lifecycle policies to transfer aged data to more economical storage classes such as S3 Standard-IA or Glacier.
- **AWS Trusted Advisor**: Frequently utilize AWS Trusted Advisor to obtain suggestions on cost reduction through the removal of unused or idle resources.

Scaling the Pipeline

Objective: Automate and expand the data processing pipeline for the Heart Attack Prediction project in a streamlined and scalable manner.

- **AWS Lambda for ETL Jobs**: Utilize AWS Lambda to manage ETL tasks, triggered by events like new data uploads to S3 buckets. This serverless method enables efficient handling of data processing workloads without the need for server management.
  - **Automating ETL Pipeline**: Implement Lambda functions to automate the ETL process, activating when new data becomes available, guaranteeing prompt processing and storage of the data.

- **Integration with AWS Services**: Effortlessly incorporate Lambda with additional AWS services like S3 for data storage and either RDS or DynamoDB for database administration.
- **AWS Step Functions**: Make use of AWS Step Functions to coordinate intricate workflows, especially in cases where the ETL process encompasses numerous steps or involves conditional logic.

Deployment

AWS SageMaker for Model Deployment
- **Objective**: To deploy the Allen Brain Observatory model in a scalable, managed environment.
- **SageMaker Features**: Make use of SageMaker's functionalities for deploying models, which include straightforward scalability, establishment of real-time inference endpoints, and the implementation of A/B testing for various model versions.
- **Deployment Process**:
  - Format the machine learning model to be compatible with SageMaker.
  - Set up and launch the model on SageMaker endpoints.
  - Integrate monitoring using SageMaker's tools to guarantee the performance and reliability of the model.

Alternative Deployment Options
- **Amazon ECS/EKS**: For finer control over the environment, contemplate deploying the model in containerized environments using either Amazon ECS or EKS.
- **CI/CD Pipelines**:Set up a continuous deployment and integration pipeline utilizing AWS CodePipeline and AWS CodeBuild for the model and its related applications.

Monitoring and Logging
- Integrate Amazon CloudWatch for thorough monitoring and logging. This involves monitoring application and model performance, establishing alerts for unusual activities or errors, and logging for auditing and debugging needs.

Conclusion

During Phase 4, our goal is to enhance cost-effectiveness, guarantee scalability, and deploy the Allen Brain Observatory model efficiently within the AWS cloud. Continuous attention is essential to adapt to evolving requirements and usage patterns, ensuring that the project remains both cost-effective and scalable while maintaining efficiency.