MSiA 423: Cloud Engineering

Team Project

Project

Developing and Implementing an end-to-end Machine Learning Solution on AWS Cloud

In this project, students will develop and implement an end-to-end machine learning solution on AWS Cloud using best practices in cloud architecture and coding. The goal of the project is to build a machine learning model and deploy it in a production environment on the cloud. The project will require students to use AWS services like EC2, S3, among others.

Project Requirements

1. Planning and Budgeting: High level estimates of project timelines and cloud infrastructure costs with accompanying Architecture Diagram (Diagrams.net) and Cost Estimate (AWS Cost Calculator).
2. Data Collection and Preparation: Identify and collect relevant data for their project. They will then clean and preprocess the data to make it ready for use in training the machine learning model. Raw and processed data should be stored in the cloud using services such as S3 or RDS.
3. Model Building and Training: Students will use popular machine learning libraries like SkLearn, TensorFlow, etc. to build and train their machine learning model. They will experiment with different model architectures and hyperparameters to find the best-performing model. Model training should leverage cloud compute like EC2 or ECS and cloud storage like S3 or RDS for artifacts and data.
4. Model Deployment: Once the machine learning model is trained, students will deploy it on AWS using best practices in cloud architecture. They will use services like EC2, ECS, or Lambda to host their model in a scalable and secure manner.
5. Configuration Files, Logging, and Monitoring: Students will use configuration management to **securely** store parameters used in their project, like hyperparameters, API keys, and AWS credentials. They will implement logging and monitoring to track the performance of their model and ensure it is working correctly. **NOTE:** Significant penalty will be applied for hard coding/exposing any API or AWS keys.
6. Good Coding Practices: Students will follow best coding practices, like using version control, writing readable code (PEP 8), and commenting their code. They will also use tools like linters and formatters to ensure their code meets quality standards.

Overall, this project will give students hands-on experience in developing and deploying a machine learning model on the cloud using best practices in cloud architecture and coding. They will learn how to use popular AWS services and build a scalable and secure solution.

Because the focus of this course is *Cloud Engineering*, students are encouraged to focus their efforts on the architecture and orchestration of a cloud-based solution. Do not choose a difficult dataset or overly complex machine learning problem if it will detract from your ability to complete the cloud engineering requirements. Your grade will be primarily based on infrastructure, not on model performance or novelty.

Data

Use any publicly available or open source (structured and/or semi-structured) dataset. You are highly encouraged to fuse data from multiple sources which could provide insights that may not be obvious from a single source.

Infrastructure

The implementation **must involve use of AWS** for primary data processing and model building, and possibly other cloud services.

Team Size

The team size is **3-4 students per team**. Due to the multitude and complexity of the technologies involved, great teamwork is critical to the success of big data projects.

Project Timelines

- **Week 2** – Form Teams
- **Week 3** – Finalize Business problem and datasets
- **Week 4** – Complete loading of dataset in to cloud data storage
- **Week 5** – Complete data clean up and processing
- **Week 6** – Complete exploratory data analysis and visualization
- **Week 7** – Complete baseline machine learning models and model metrics
- **Week 8** – Complete challenger/advanced machine learning models
- **Week 9** – Stretch goal: Automation of data pipelines

Project Submission

- Single submission per team within 1 day of the final presentation
- Upload all source code (ETL, DDL/DML, Notebooks, Pipelines, etc.) as a single zip file
- Upload project presentation as a separate PPT/PDF file
- Demonstrate deployment of model into a cloud environment. At least one of the following:
    - Event-driven data processing (S3 –> Lambda)
    - Model Training and Artifact Storage (ECS –> S3)
    - Inference via exposed web application (ALB/APIGW + ECS/EC2/Lambda)