Lab 4 Streaming Data Analytics

Handout: 31 March, Hand-in: 21 April 23:59 (Thursday)

In this lab, you will work **in a group** to learn how to **set up a streaming data analytics pipeline** using Amazon SageMaker, Amazon Kinesis Data stream, Amazon Kinesis Data Analytics, Amazon API Gateway and a Lambda function.

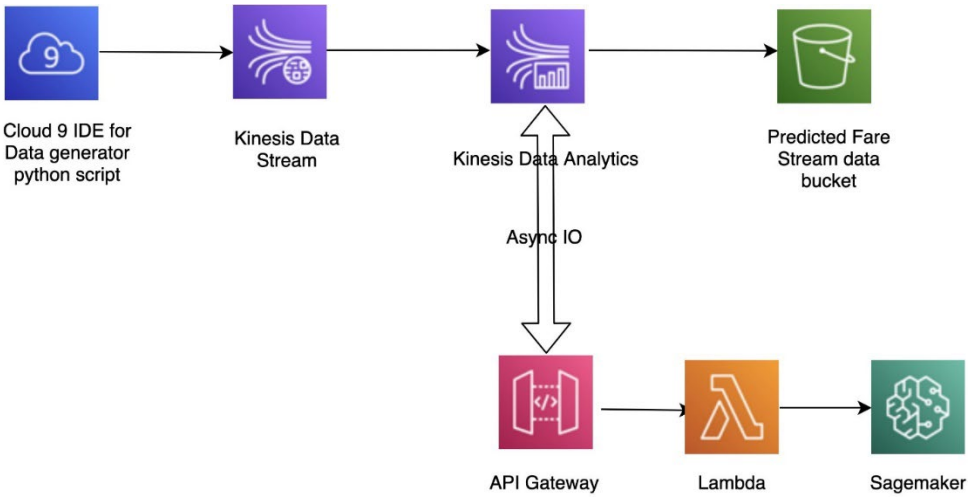


Fig. 1 Streaming data analytics pipeline

Overview:

In this lab, we will use a subset of sample data from nyc taxi ride dataset to train a toy model. The model will predict the taxi fare based on taxi rides. We will deploy this model behind SageMaker and will use API gateway and lambda function to invoke the SageMaker endpoint in real time. To generate the real time streaming data of taxi rides, we will use Python script running in Cloud 9. And for streaming processing, we will use Kinesis Data stream and analytics. The predicted output can be stored in a S3 bucket. Main components in this pipeline are shown in Fig. 1, including

- **Streaming Data Generator**: a streaming data generator (data generator python script running on AWS Cloud9 IDE) that generate data stream. AWS Cloud9 is a cloud-based integrated development environment (IDE) that lets you write, run, and debug your code with just a browser. The data stream will be sent to AWS Kinesis Data Streams which collects data for analytics.
- **AWS Kinesis Data stream**: a massively scalable and durable real-time data streaming service. It can continuously capture gigabytes of data per second from hundreds of thousands of sources.
- **AWS Kinesis Data Analytics:** a tool to transform and analyze streaming data in real time using Apache Flink.
- **AWS SageMaker**: a cloud machine-learning platform that enables developers to create, train, and deploy machine-learning (ML) models in the cloud.

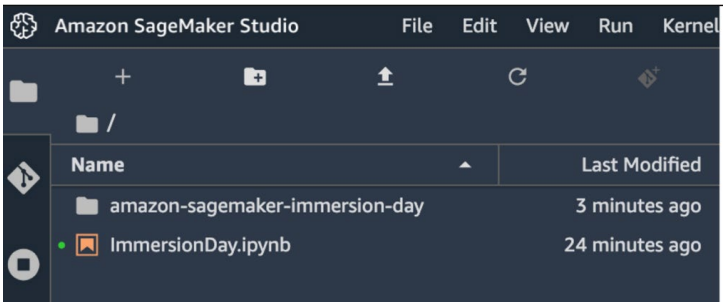
Tasks for the Lab:

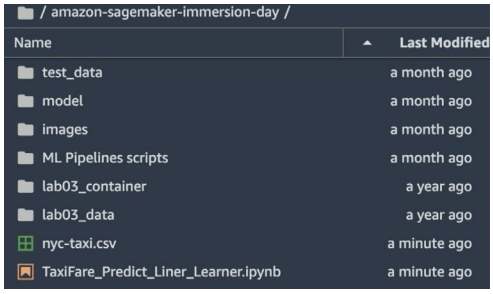
Prerequisite: Setup SageMaker Studio

Before you start the lab, you need to set up the SageMaker Studio to create IAM role, load the dataset used in the lab, and create a Notebook development environment. Please follow the '**self-paced lab**' guide below to finish the setup.

https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US/prerequisites/option2#downloading-the-content-of-the-github-repository-needed-for-the-labs

At the end of the setup, you will obtain a Jupyter Notebook environment with start code and a folder with necessary dataset and scripts.





After you have set up the environment, please follow the instructions to work on each task in the lab.
https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US/lab7

Task 1. Streaming Data Generation

In this task, you need to generate the real time streaming of taxi rides. It is achieved by using python data generator script which will run in Cloud 9 for streaming data injection.

Task 2. Streaming Data Processing

After the data ingestion, you will use Amazon Kinesis Data Stream service to process the data stream and use the Amazon Kinesis Data Analytics service to call analytics model developed in the SageMaker. Within the Amazon Kinesis Data Analytics module, an Apache Flink java application will be used to asynchronously invoke SageMaker endpoint for all incoming streaming data.

Task 3. ML Model Training

In this lab, you can train your model with SageMaker linear learner built-in algorithm. You can deploy this model behind SageMaker endpoint and will use API gateway and lambda function to invoke the SageMaker endpoint in real time.

Task 4. Storing the Prediction

Once fare prediction data is generated for each incoming streaming data hen resultant dataset with predicted fare will be stored in S3 bucket.

Reflection Questions

After finishing the lab tasks, please think about and answer the following questions:

1. What's the job of AWS Kinesis Data Stream in this pipeline? Why can't we link the python data generator script directly to the AWS Kinesis Data Analytics (KDA) engine?

2. What's the relation between AWS Kinesis Data Analytics (KDA) and Apache Flink? What are the roles played by the AWS KDA and SageMaker respectively in this pipeline?

3. Why does an *Asynchronous* I/O operator is used between the AWS Kinesis Data Analytics and API Gateway in this pipeline?

4. This demo is using simulated data created by a python script. What changes in the pipeline does it need to ingest real-world data in a production environment? (You don't need to implement that)

Reminder

Note that, keeping a shard will cause a charge of $0.015 per hour, or $0.36 per day ($0.015 x 24hrs). **Once you finished the lab, do delete those services to avoid unnecessary charging!**

Submission & Rubrics:

Please **check off your lab** with Prof. Wenchao or TA and **submit your report in groups** no later than **21 April.** In your report, please briefly **describe how you solved each task** and **answer the reflection questions**.

Total marks for this lab: 30pt

- Completion of lab tasks (20pt, 5pt for each task).
- Report (10pt including reflection questions)

Reference:

https://catalog.us-east-1.prod.workshops.aws/workshops/63069e26-921c-4ce1-9cc7-dd882ff62575/en-US/lab7