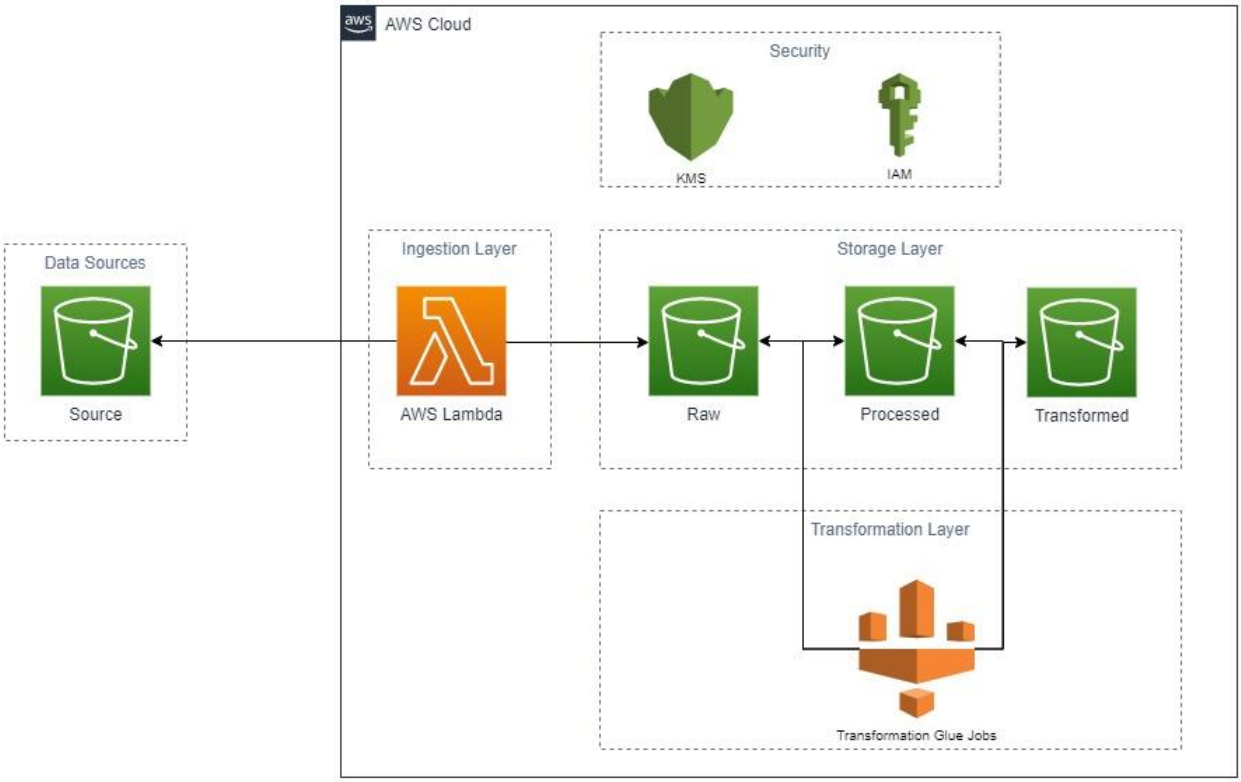
Data Lake Implementation

A large farming company has tasked you with implementing their data lake design using CloudFormation.



Currently, they store their data in several CSV files. These files are located in the Source S3 bucket. For now, the Source S3 bucket will be in the same AWS account as the data lake.

I expect the storage layer to be encrypted.
Appropriate IAM roles, policies, etc. are created with the minimal permissions, i.e., required permissions only.

The ingestion Layer will consist of lambda function that is run on demand.
Once run, it will copy the files in the source as is and upload them to the Raw bucket of the Storage Layer.

Upon upload completion, a python shell Glue job will automatically start. This job will combine both files and upload the combined file to the Processed S3 bucket.
Upon upload completion, another python shell Glue job will be automatically kicked off. This job will create a new column, namely yield_per_hectare. This column is the result of dividing the yield column by the area. This job will also convert the file to parquet format and store the resulting file in the Transformed S3 bucket.

Analytics and Machine Learning

You are also task with looking through this data and find patterns and make predictions that will help the customer make more money.

Dataset

Information about yield from 1000 farms across a country.

- Id - Identifier
- Water - the average amount of water received by hectare
- UV - the average amount of light received by hectare
- Area - the size of the farm in hectares
- Fertilizer_usage - the level of fertilization
- Yield - total crop yield by farm
- Pesticides - the amount of pesticide used per hectare
- Region - region code
- Pesticides Used - comma-separated list of pesticides used
- Yield_per_hectare

Requirements

- analyze this dataset and try to find some interesting patterns and statistics.
- find the target variable, build a prediction model for it, and evaluate its performance.
- provide insights into the data.
- use python 3 with pandas in a Jupyter notebook
- the notebook must be written well and clearly highlights your insights, findings, etc. and describes your thought process. I expect your Jupyter notebook to serve the purpose of a report.
- Please inform the farming company with what your next steps would be if you were to proceed further with this project.