AWS CDK Python Technical Test - Data Engineering

Overview

This technical test evaluates your knowledge of AWS services for data engineering using AWS CDK with Python. You will build a complete data pipeline that extracts data from a public API, stores it in S3, catalogs it with AWS Glue, and makes it queryable through Amazon Athena, all while managing permissions through AWS Lake Formation.

Duration

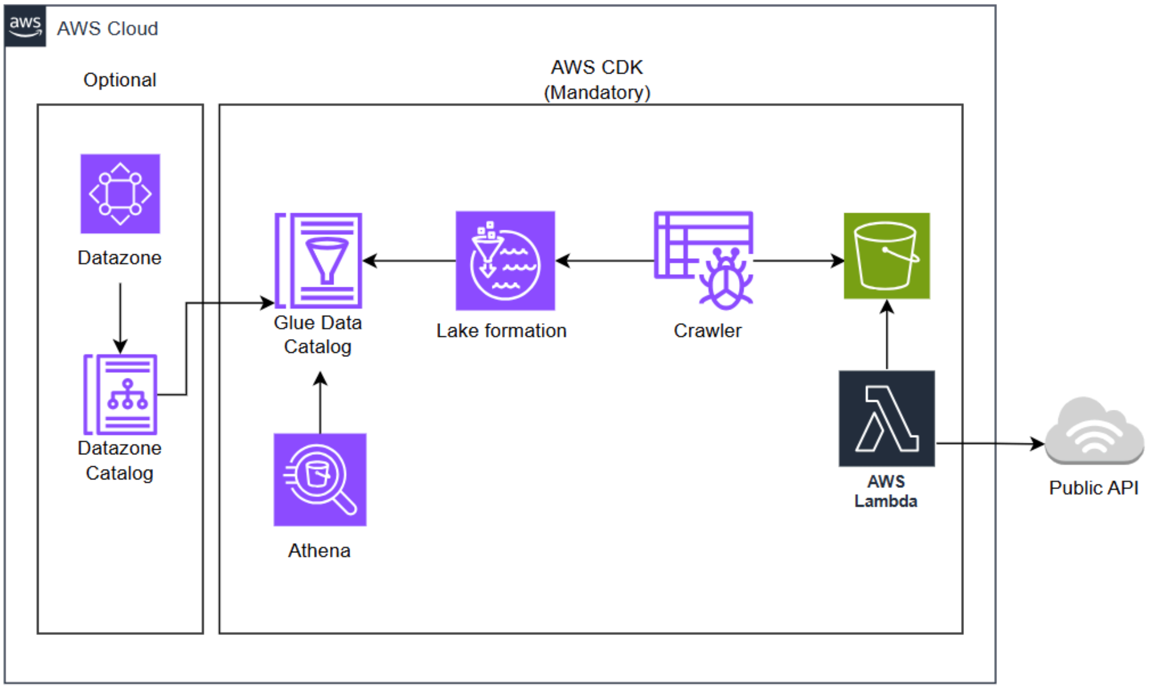
Time Limit: 5 Days

Prerequisites

- AWS CLI configured with appropriate permissions
- AWS CDK v2 installed
- Python 3.10+ with pip
- Node.js (for CDK)
- Basic understanding of data engineering concepts

Architecture Overview

You will build a serverless data pipeline with the following components:



Part 1: Core Requirements (Mandatory – AWS CDK Implementation)

1.1 Project Setup

- Create a new AWS CDK project using Python
- Structure your project with appropriate separation of concerns

1.2 Lambda Function for Data Extraction

Create a Lambda function that:

- Calls a public API (It can be any of those suggested below or any of your choice)
- Processes the response data
- Saves the data to S3

Suggested Public APIs:

- JSONPlaceholder API: https://jsonplaceholder.typicode.com/users
- Random User API: https://randomuser.me/api/?results=100

1.3 S3 Storage

- Create an S3 bucket for data storage
- Choose an appropriate file format (CSV, Parquet, Avro, etc)

1.4 AWS Glue Integration

- Create a Glue Database
- Create a Glue Crawler that:
  - Discovers the schema of your S3 data
  - Creates/updates table definitions automatically
  - Runs on a schedule or can be triggered manually

1.5 Lake Formation Setup

- Create appropriate permissions for:
  - The Lambda execution role to write to S3
  - The Glue Crawler role to catalog the data
  - Athena users to query the data
- Implement table-level and column-level permissions

1.6 Amazon Athena Configuration

- Set up Athena to query the Glue catalog
- Configure a query results location in S3
- Ensure queries work properly against your cataloged data

Part 2: Bonus Challenge (Optional)

2.1 Amazon DataZone Integration (Manually through the AWS Console)

If you want to showcase additional skills, you can:

- Create a DataZone Domain (version 1)

- Set up a Project within the domain
- Create an Environment
- Add a Data Source pointing to your Glue database
- Create a Data Asset from your table

Note: This can be done manually through the AWS Console - CDK implementation is not required for this bonus section.

Deliverables

Required Files

1. **CDK Code**: Complete CDK application with all stacks
2. **Lambda Source Code**: The function code for data extraction

Sample Test Scenarios

After deployment, you should be able to:

1. Trigger the Lambda function and see data appear in S3
2. Run the Glue Crawler and see a table created in the Glue Catalog
3. Query the data using Athena
4. Demonstrate that Lake Formation permissions are working correctly