**Project Name:** ETL pipeline in AWS to analyze Covid-19 Datasets

**Project Short Description**: The Covid-19 Analysis project centralizes COVID-19 datasets on AWS. The data engineer's tasks involve cleaning data with a Spark pipeline, creating Glue tables, previewing and validating with Athena for analysis.

**Project Long Description**: The Covid-19 Analysis project involves curating and centralizing up-to-date datasets related to the novel coronavirus (SARS-CoV-2) and COVID-19. Hosted on AWS, the curated data lake includes case tracking data, hospital bed availability, and research articles. As a data engineer, the task is to implement a Spark pipeline to clean the source data, upload it to an S3 staging directory, create Glue tables and use Athena for preview and validation for analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Milestone Sr. No or Order | Milestone Name | Milestone Description | Task Name | Task Description |
| 1 | Data Cleaning and Staging | Clean the source data and upload the cleaned data into a staging directory on S3. | Spark Pipeline for Data Cleaning | Implement a Spark pipeline to clean the source data and upload the cleaned data to the staging directory on S3. |
| 2 | Glue Table Creation | Create Glue tables using a Crawler in the Glue Catalog for the cleaned data | Glue Crawler Configuration | Configure and execute a Glue Crawler on the cleaned data and Generate Glue tables in the Glue Catalog. |
| 3 | Athena Preview and Validation | Utilize Athena for previewing and validating the created tables | Athena Query Execution | Write and execute Athena queries to preview data from the created Glue tables and validate the correctness of the data. |