**Data Engineering – Final Assessment**

**Business Scenario:**

The COVID-19 pandemic, caused by the novel coronavirus (SARS-CoV-2), has had a significant impact on the world since its emergence in late 2019. The virus spreads primarily through respiratory droplets when an infected person coughs, sneezes, talks, or breathes. It has led to millions of infections and fatalities globally, prompting governments and health organizations to implement various measures to control its spread

**Challenges:**

In this assessment, An individual has to collect and analyse various COVID-related data sets, such as infection rates, vaccination rates, hospitalization data, and demographic information. They could gather this data from AWS Data Exchange platform. Using advanced data analytics techniques, individual could provide valuable insights and trends with respect to COVID

**What is Expected?**

After completing this final project, you will apply various data engineering skills and techniques. You've recently joined the organization, facing a business challenge that requires data engineering skills on real data sets. Based on your findings, the next level of analysis will be charted out.

Here are some indicative types of data engineering techniques you can perform. Please note that this is not an exhaustive list, you may add more

You will perform the various tasks that professional data analysts do as part of their jobs, including:

* Data collection from multiple sources
* Data wrangling and data preparation
* Exploratory data analysis
* Data Storage and retrieval

**Data Discovery ,Data Preparation, Data Manipulation, Data Storage**

* Identify data source types
* Identify duplicate values in the dataset
* Remove duplicate values from the dataset.
* Identify missing values in the dataset.
* Determine the missing values in the dataset.
* Normalize data in the dataset.
* Identify silver , bronze , gold layer attributes

The project will come to a close with a presentation of your data engineering report to relevant organisation stakeholders. An executive summary, your analysis, and a conclusion will all be included in the report. Both the final output and your work for the various Data engineering process phases will be taken into consideration as part of evaluation.

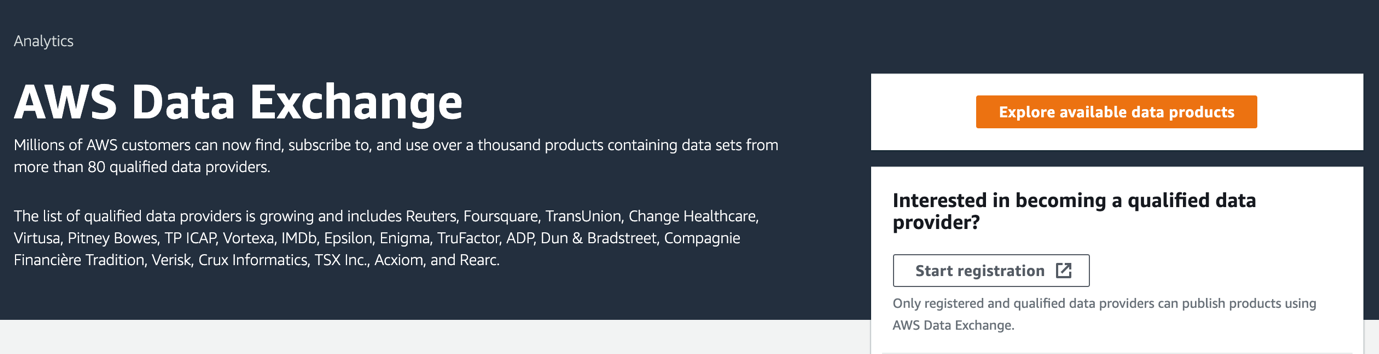
Your ability to use Jupyter Notebooks, SQL, Relational Database Management Systems (RDBMS), Big Data Tools , Kafka , Airflow , NoSQL ,cloud computing environments like Azure and GCP, and Python libraries like Pandas, Numpy, and others will be put to the test as part of this project.

**NOTE:** Results must be backed with appropriate inferences and insights.

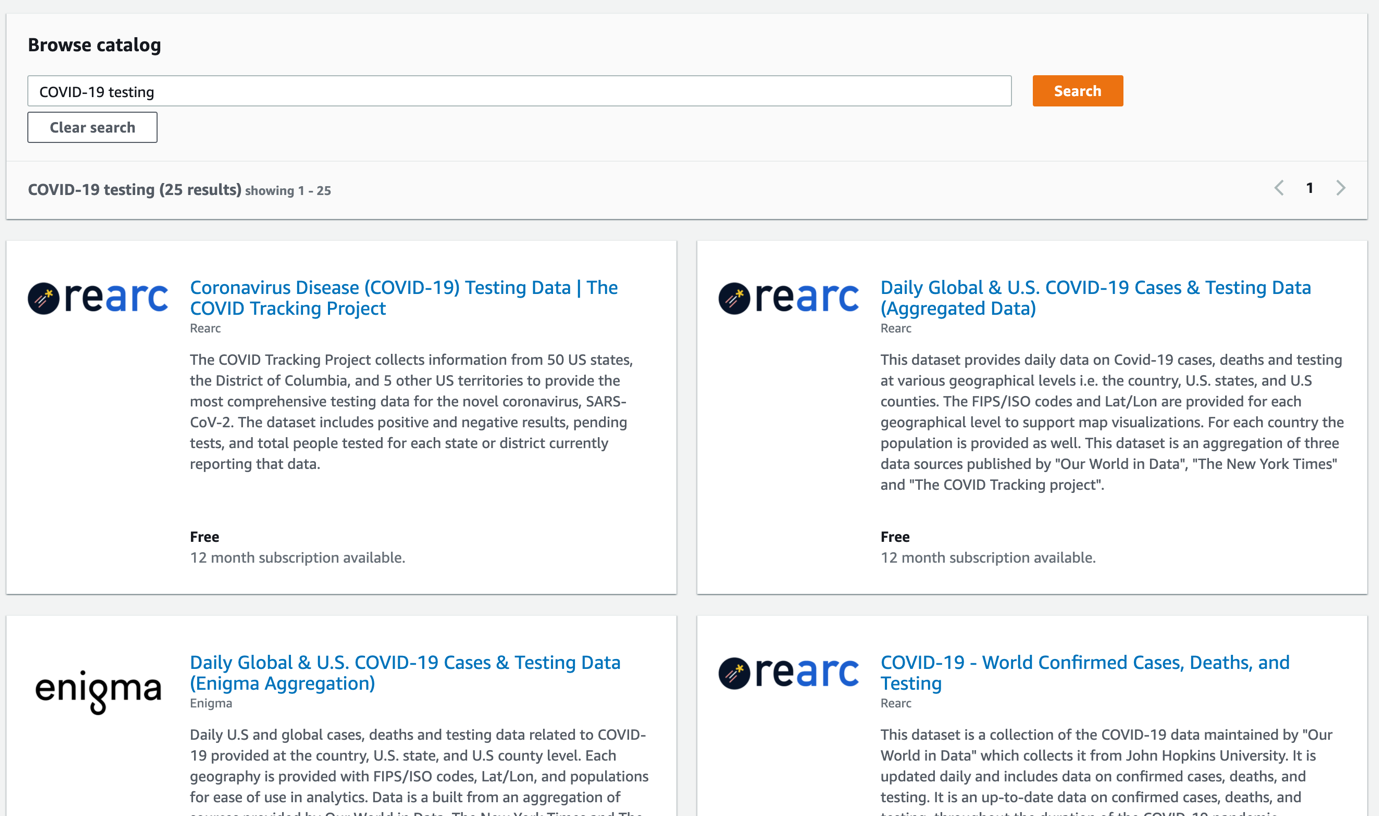
**Data Dictionary:**

Register on AWS Data Exchange for COVD data set as explained below and the data dictionary is part of below registration

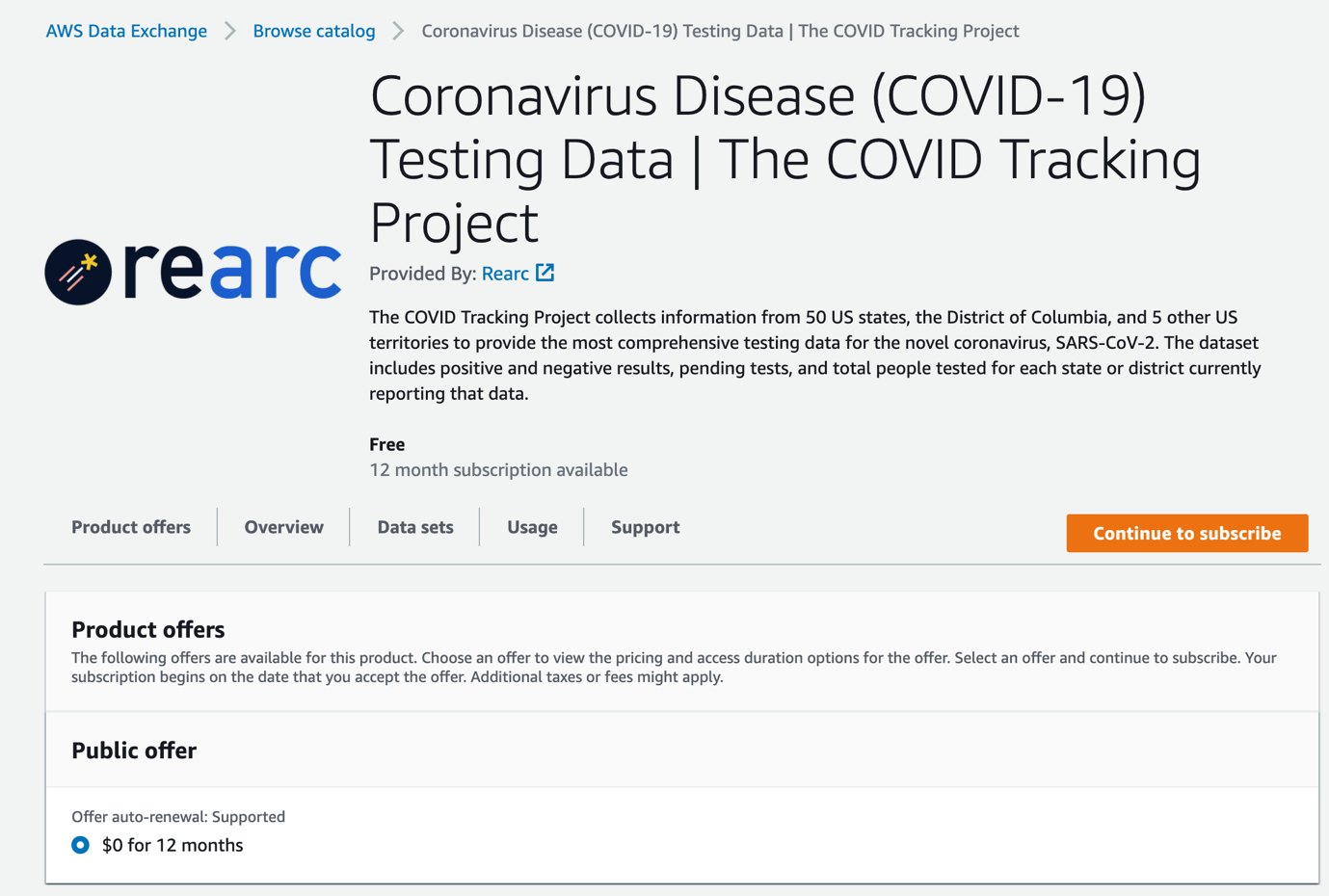
1. Go to aws console from the LAB and click Explore available data products.



1. In Browse catalog search bar, put in **COVID-19 testing** as the search string. You should see the following search results being returned



1. Click the first returned result, which should be Coronavirus Disease (COVID-19) Testing Data | The COVID Tracking Project. This dataset is free for subscription as the cost is $0 for 12 months. In the Overview section of the dataset, you could find details about the source, metadata, data provider and support information



1. Click Continue to subscribe button to the subscription page. You can choose the terms for your subscription here and click Subscribe on the bottom to subscribe to the dataset

Follow the onscreen prompts to move the data to S3 bucket.

**Tasks :**

**Data Discovery , Processing , Loading Task :**

**Task 1**

* + 1. Move the dataset onto S3.
    2. Create Glue database and 2 crawlers to crawl CSV and JSON folders using AWS CLI only , provide screen shots of using AWS CLI (Please note Console cannot be used to create Database and Crawlers)
    3. Run the Crawlers from CLI
    4. Confirm the data is been able to query from Athena

**Task 2**

1. Deploy GLUE locally and not use AWS GLUE. I have shared a walkthrough of this on Last class
2. Develop a GLUE JOB to convert the CSV files to PARQUET from Local Development only (Do not use AWS GLUE Console). If there are memory limitations then limit the dataset to 100 records only
3. Create crawler using AWS CLI on the parquet file
4. Run the crawler from CLI and confirm data from Athena

**Task 3**

* 1. Run basic Analytical queries on the dataset using Athena from AWS Console

**Data Versioning Task :**

**Task 1 Create a repository in github and upload the dataset**

* 1. Create a repository in github.com

**Task 2 Generate the Reports and upload them to github**

1. Clone the repository that you created in task 1
2. Create a branch for this activity ,Create a file with all the queries from the above tasks.
3. List the version history ( commit history) , status of the local repository
4. Commit the file to the branch
5. Push the changes to the github