Data Used: https://github.com/deepanshuMeteor/Databricks-AWS/blob/main/covid\_data.csv

1. Create a new notebook and rename it to Covid Pipeline Notebook.
2. Next to the notebook’s name, select **Python** and change the default language to **SQL**.
3. Enter the following code in the first cell *without running it*. All cells will be executed after the pipeline is created. This code defines a materialized view that will be populated by the raw data previously downloaded:

import dlt

from pyspark.sql.functions import col

@dlt.table(

    comment="COVID sample dataset. This data was ingested from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University."

)

def raw\_covid\_data():

    return (

        spark.read.format("csv")

        .option("header", True)

        .load("dbfs:/FileStore/tables/covid\_data.csv")

        .select(

            col("Last\_Update"),

            col("Country\_Region"),

            col("Confirmed"),

            col("Deaths"),

            col("Recovered")

        )

    )

1. Under the first cell, use the **+ Code** icon to add a new cell, and enter the following code to query, filter, and format the data from the previous table before analysis.

sql

CREATE OR REFRESH MATERIALIZED VIEW processed\_covid\_data(

CONSTRAINT valid\_country\_region EXPECT (Country\_Region IS NOT NULL) ON VIOLATION FAIL UPDATE

)

COMMENT "Formatted and filtered data for analysis."

AS

SELECT

TO\_DATE(Last\_Update, 'MM/dd/yyyy') as Report\_Date,

Country\_Region,

Confirmed,

Deaths,

Recovered

FROM live.raw\_covid\_data;

1. In a third new code cell, enter the following code that will create an enriched data view for further analysis once the pipeline is successfully executed.

sql

CREATE OR REFRESH MATERIALIZED VIEW aggregated\_covid\_data

COMMENT "Aggregated daily data for the US with total counts."

AS

SELECT

Report\_Date,

sum(Confirmed) as Total\_Confirmed,

sum(Deaths) as Total\_Deaths,

sum(Recovered) as Total\_Recovered

FROM live.processed\_covid\_data

GROUP BY Report\_Date;

1. select **Create pipeline** from output.
2. In the **Create pipeline** page, create a new pipeline with the following settings:
   * **Pipeline name**: Covid Pipeline
   * Enable Serverless checkbox
   * **Pipeline mode**: Triggered
   * **Source code**: *Browse to your* Covid Pipeline Notebook *notebook in the* Users/user@name *folder*.
   * **Default catalog -**  rps\_demo
   * **Target schema**: *Enter* default
3. Select **Create** and then **Start**. Then wait for the pipeline to run (which may take some time).

View results as a visualization

Go to the Data Ingestion and Exploration notebook, add a new code cell and run the following code to load the raw\_covid\_data into a dataframe:

%sql

SELECT \* FROM raw\_covid\_data

Above the table of results, select + and then select Visualization to view the visualization editor, and then apply the following options:

Visualization type: Line

X Column: Report\_Date

Y Column: Add a new column and select Total\_Confirmed. Apply the Sum aggregation.

Save the visualization and view the resulting chart in the notebook.