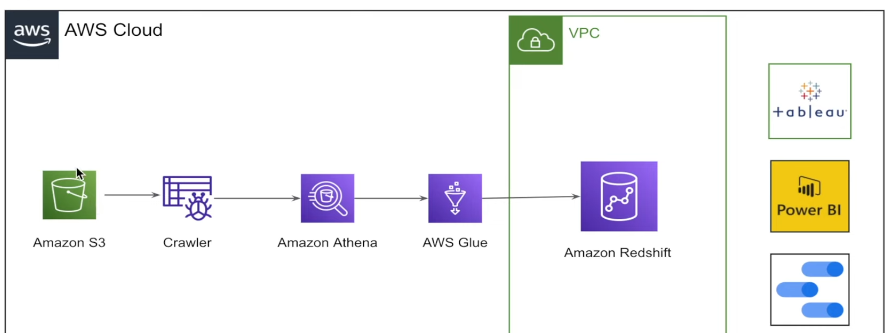
**COVID\_19 PROJECT**



**Step 1:** Initially, you directly download the data from the AWS S3 explorer. Then upload to S3 bucket that you have create(**praful-de-covid-project**)

Now we have data in S3.

===========================================================

Step 1: Then we should understand the data. So that we use crawler to know about columns, data types and all. That how by understand data we can build DATA MODEL.

GLUE - ETL service of AWS, You crete workflows using Glue, Create endpoint(Jupiter note book)

Now using glue for crawling the data. Here you create a crawler and IAM role and IAM role should be given little permission like S3 full access and all.

**Now I have created a Crawler called praful-de-covid-project and Create IAM role with few permissions and also selected the database for it to store all the tables.**

**It also created the table(enigma\_jhud) inside the database(covid\_dataset).**

**Similarly it done to all the csv files and created 10 tables.**

**===================================================================**

Then we use athena to query the date. Also make sure in athena settings you have give output s3 bucket. Which will store metadata of all the things you query using athena.

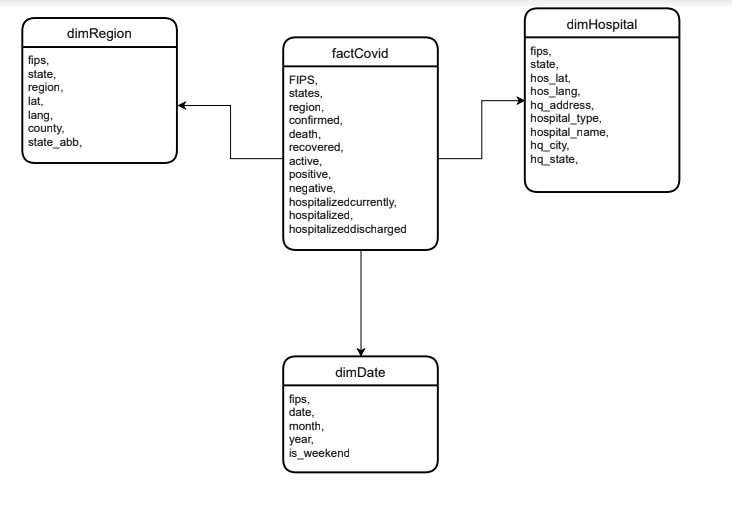
Using athena you can see all the 10 tables and also you can query about tables.

===================================================================

Step 2: Create a DATA MODEL in data.io website(free). Just in athena generate DDL and then copy all the columns and paste in table in data.io. Like that you have to do for all the 10 tables.

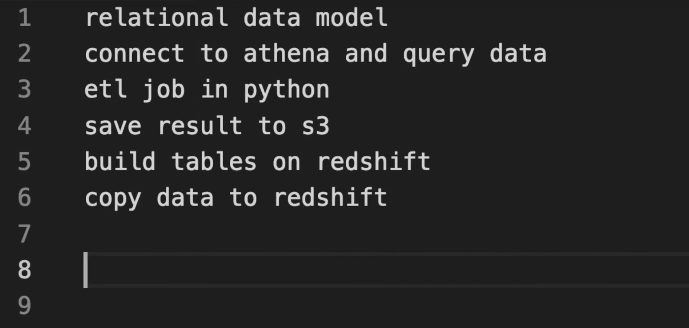
====================================================================

Step 3: Create Dimension Schema

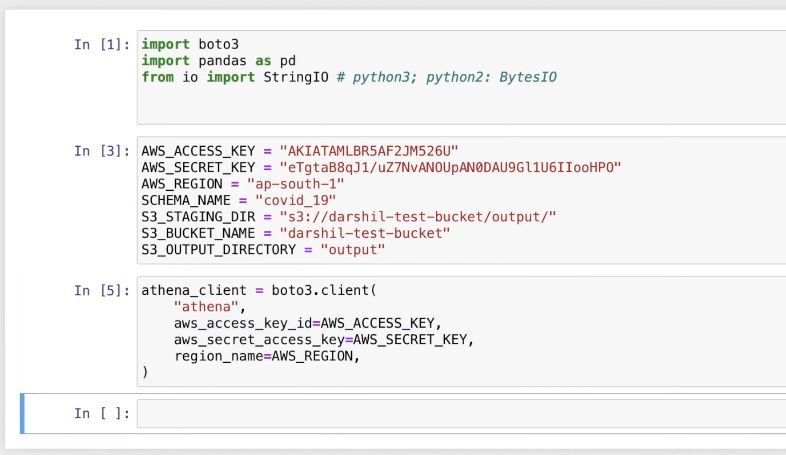


=================================================================

Step 1: Relation data model(nothing but the tables)

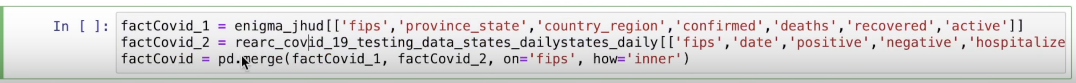


Step 2: Connect Athena using boto3 to query the data using pandas in Jupyter notebook.

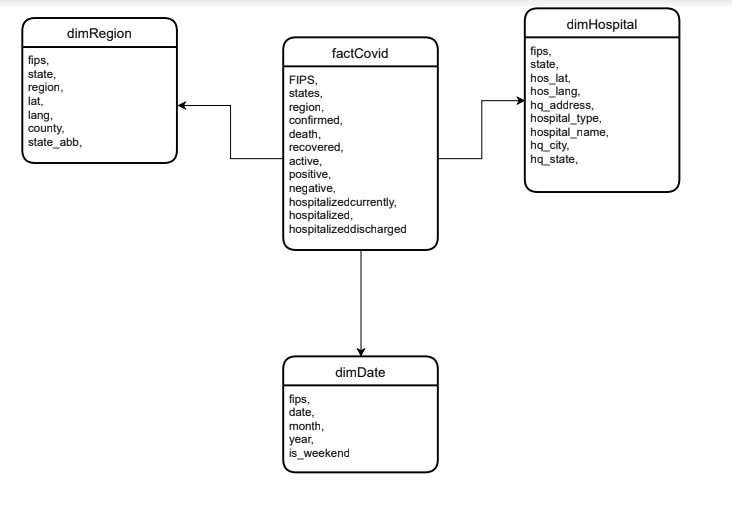


=================================================================

3. ETL job using python.



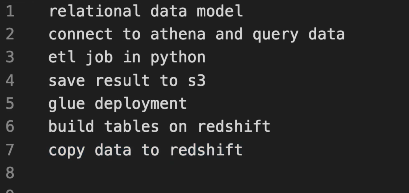
Now created fact table called **factcovid** using above script**.** Similarly you you have do all the dim tables.



After extracting required columns are store into tables. The dimension schema is ready.

==========================================================

Then store the output to S3 bucket.



================================================================

RedShift -

Now we use library called redshift\_connector

Create a table on redshift.

This is one way to do it but using glue job you do the same.

===============================================================

AWS Glue JOb is the one generate script automatically.//(remember this can be done in jupiter using python)

