{

"cells": [

{

"cell\_type": "markdown",

"id": "12b93a92",

"metadata": {},

"source": [

"# COVID-19 Data Engineering Project"

]

},

{

"cell\_type": "markdown",

"id": "5c3ddcf7",

"metadata": {},

"source": [

"Refreshing the work I did in previous company.\n",

"\n",

"Performing data modeling, data wrangling and extract-load-transform on the COVID-19 Data Lake available on registry of open data AWS using various AWS tools such as boto3, Glue, S3, Athena and Redshift."

]

},

{

"cell\_type": "markdown",

"id": "b053e4f7",

"metadata": {},

"source": [

"## Code"

]

},

{

"cell\_type": "code",

"execution\_count": 61,

"id": "411b9af1",

"metadata": {},

"outputs": [],

"source": [

"import boto3\n",

"import pandas as pd\n",

"from io import StringIO\n",

"import time"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "d76e5894",

"metadata": {},

"outputs": [],

"source": [

"AWS\_ACCESS\_KEY = ''\n",

"AWS\_SECRET\_KEY = ''\n",

"AWS\_REGION = 'eu-central-1'\n",

"SCHEMA\_NAME = 'covid\_19'\n",

"S3\_STAGING\_DIR = 's3://oovk-covid-project-output-buck/output/'\n",

"S3\_BUCKET\_NAME = 'oovk-test-bucket'\n",

"S3\_OUTPUT\_DIRECTORY = 'output'\n"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "cf0a8e01",

"metadata": {},

"outputs": [],

"source": [

"athena\_client = boto3.client(\"athena\",\n",

" aws\_access\_key\_id = AWS\_ACCESS\_KEY,\n",

" aws\_secret\_access\_key\_id = AWS\_SECRET\_KEY,\n",

" region\_name = AWS\_REGION,)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "9b6ea807",

"metadata": {},

"outputs": [],

"source": [

"#Gets the data from athena and converts it into pandas dataframe\n",

"Dict = {}\n",

"def download\_and\_load\_query\_results(\n",

" client: boto3.client, query\_response: Dict\n",

") -> pd.DataFrame:\n",

" while True:\n",

" try:\n",

" client.get\_query\_results(QueryExecutionId=query\_response[\"QueryExecutionId\"])\n",

" break\n",

" except Exception as err:\n",

" if \"not yet finished\" in str(err):\n",

" time.sleep(0.01)\n",

" else:\n",

" raise err\n",

" temp\_file\_location: str = \"athena\_query\_results.csv\"\n",

" s3\_client = boto3.client(\"s3\",\n",

" aws\_access\_key\_id = AWS\_ACCESS\_KEY,\n",

" aws\_secret\_access\_key\_id = AWS\_SECRET\_KEY,\n",

" region\_name = AWS\_REGION,)\n",

" s3\_client.download\_file(S3\_BUCKET\_NAME,\n",

" f\"{S3\_OUTPUT\_DIRECTORY}/{query\_response['QueryExecutionId']}.csv\",\n",

" temp\_file\_location,)\n",

" return pd.read\_csv(temp\_file\_location)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "ae70f540",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM enigma\_jhud\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"enigma\_jhud = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "faa4315c",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM nytimes\_data\_in\_usa\_us\_county\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"nytimes\_data\_in\_usa\_us\_county = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "52d2b2ba",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM nytimes\_data\_in\_usaus\_states\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"nytimes\_data\_in\_usaus\_states = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "8a5b3e30",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM countypopulation\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"countypopulation = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "69e22a65",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM rearc\_usa\_hospital\_beds\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"rearc\_usa\_hospital\_beds = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "f8161734",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM state\_abv\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"state\_abv = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "d42e5c69",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM states\_daily\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"states\_daily = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "db1630c1",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM us\_daily\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"us\_daily = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "7d3978f3",

"metadata": {},

"outputs": [],

"source": [

"response = athena\_client.start\_query\_execution(\n",

" QueryString=\"SELECT \* FROM us\_total\_latest\",\n",

" QueryExecutionContext={\"Database\":SCHEMA\_NAME},\n",

" ResultConfiguration={\"OutputLocation\":S3\_STAGING\_DIR\n",

" \"EncryptionConfiguration\":{\"EncryptionOption\":\"SSE\_S3\"},\n",

" },\n",

")\n",

"\n",

"us\_total\_latest = download\_and\_load\_query\_results(athena\_client, response)"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "4e63607a",

"metadata": {},

"outputs": [],

"source": [

"us\_total\_latest.head() #Checking the dataframe"

]

},

{

"cell\_type": "markdown",

"id": "1d95de12",

"metadata": {},

"source": [

"## Building Dimensional Model"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "ac7b01f6",

"metadata": {},

"outputs": [],

"source": [

"factCovid\_1 = enigma\_jhud[['fips','province\_state','country\_region','confirmed','deaths','recovered','active']]\n",

"factCovid\_2 = us\_daily[['fips','date','positive','negative','hospitalizedcurrently','hospitalized','hospitalizeddischarged']]\n",

"factCovid = pd.merge(factCovid\_1,factCovid\_2,on='fips',how='inner')"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "c71063e4",

"metadata": {},

"outputs": [],

"source": [

"factCovid.shape"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "4ba90142",

"metadata": {},

"outputs": [],

"source": [

"dimRegion\_1 = enigma\_jhud[['fips','province\_state','country\_region','latitude','longitude']]\n",

"dimRegion\_2 = nytimes\_data\_in\_usa\_us\_county[['fips','county','state']]\n",

"dimRegion = pd.merge(dimRegion\_1,dimRegion\_2,on='fips',how='inner')"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "1e77bdf2",

"metadata": {},

"outputs": [],

"source": [

"dimHospital = rearc\_usa\_hospital\_beds[['fips','state\_name','latitude','longitude','hq\_address','hospital\_name','hospital\_type','hq\_city','hq\_state']]"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "239af5d2",

"metadata": {},

"outputs": [],

"source": [

"dimDate = states\_daily[['fips','date']]"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "03aa5482",

"metadata": {},

"outputs": [],

"source": [

"dimDate.head()"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "fe793a1e",

"metadata": {},

"outputs": [],

"source": [

"dimDate['date'] = pd.to\_datetime(dimDate['date'], format='%Y%m%d')"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "996271c1",

"metadata": {},

"outputs": [],

"source": [

"dimDate['year'] = dimDate['date'].dt.year\n",

"dimDate['date'] = dimDate['date'].dt.month\n",

"dimDate['day\_of\_week'] = dimDate['date'].dt.dayofweek"

]

},

{

"cell\_type": "markdown",

"id": "810d9e73",

"metadata": {},

"source": [

"## Saving to S3"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "4401b70e",

"metadata": {},

"outputs": [],

"source": [

"bucket = 'oovk-covid-project-output-buck'"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "6fcfef87",

"metadata": {},

"outputs": [],

"source": [

"csv\_buffer = StringIO()\n",

"factCovid.to\_csv(csv\_buffer)\n",

"s3\_resource = boto3.resource('s3')\n",

"s3\_resource.Object(bucket,'output/factCovid.csv').put(Body=csv\_buffer.getvalue())\n"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "2f6592cc",

"metadata": {},

"outputs": [],

"source": [

"csv\_buffer.getvalue()"

]

},

{

"cell\_type": "markdown",

"id": "7ade61b9",

"metadata": {},

"source": [

"## Extracting schema from dataset"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "966f9a79",

"metadata": {},

"outputs": [],

"source": [

"dimDatesql = pd.io.sql.get\_schema(dimDate.reset\_index(),'dimDate')\n",

"print(''.join(dimDatesql))"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "aef53cdf",

"metadata": {},

"outputs": [],

"source": [

"factCovidsql = pd.io.sql.get\_schema(factCovid.reset\_index(),'factCovid')\n",

"print(''.join(factCovidsql))"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "fd270d0e",

"metadata": {},

"outputs": [],

"source": [

"dimRegionsql = pd.io.sql.get\_schema(dimRegion.reset\_index(),'dimRegion')\n",

"print(''.join(dimRegionsql))"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "d35eda3d",

"metadata": {},

"outputs": [],

"source": [

"dimHospitalsql = pd.io.sql.get\_schema(dimHospital.reset\_index(),'dimHospital')\n",

"print(''.join(dimHospitalsql))"

]

},

{

"cell\_type": "markdown",

"id": "b5f8be7f",

"metadata": {},

"source": [

"## Redshift Connector"

]

},

{

"cell\_type": "code",

"execution\_count": 64,

"id": "e676f3a8",

"metadata": {},

"outputs": [],

"source": [

"import redshift\_connector"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "0dff6e9c",

"metadata": {},

"outputs": [],

"source": [

"conn = redshift\_connector.connect(\n",

" host='',\n",

" databse='dev',\n",

" user='awsuser',\n",

" password='Passw0rd123'\n",

")"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "8ee0fc5f",

"metadata": {},

"outputs": [],

"source": [

"conn.autocommit = True"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "fd42e592",

"metadata": {},

"outputs": [],

"source": [

"cursor=redshift\_connector.Cursor = conn.cursor()"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "44a62cc5",

"metadata": {},

"outputs": [],

"source": [

"cursor.execute(\"\"\"\n",

"CREATE TABLE \"dimDate\" (\n",

"\"index\" INTEGER,\n",

"\"fips\" INTEGER,\n",

"\"date\" TIMESTAMP,\n",

"\"year\" INTEGER,\n",

"\"month\" INTEGER,\n",

"\"day\_of\_week\" INTEGER\n",

")\n",

"\"\"\")"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "160eb9fb",

"metadata": {},

"outputs": [],

"source": [

"cursor.execute(\"\"\"\n",

"CREATE TABLE \"dimHospital\" (\n",

"\"index\" INTEGER,\n",

"\"fips\" REAL,\n",

"\"state\_name\" TEXT,\n",

"\"longitutde\" REAL,\n",

"\"latitude\" REAL,\n",

"\"hq\_address\" TEXT,\n",

"\"hospital\_name\" TEXT,\n",

"\"hospital\_type\" TEXT,\n",

"\"hq\_city\" TEXT,\n",

"\"hq\_state\" TEXT,\n",

")\n",

"\"\"\")\n",

"\n",

"cursor.execute(\"\"\"\n",

"CREATE TABLE \"factCovid\" (\n",

"\"index\" INTEGER,\n",

"\"fips\" REAL,\n",

"\"province\_state\" TEXT,\n",

"\"country\_region\" TEXT,\n",

"\"confirmed\" REAL,\n",

"\"deaths\" REAL,\n",

"\"recovered\" REAL,\n",

"\"active\" REAL,\n",

"\"date\" INTEGER,\n",

"\"positive\" REAL,\n",

"\"negative\" REAL,\n",

"\"hospitalizedcurrently\" REAL,\n",

"\"hospitalized\" REAL,\n",

"\"hospitalizeddischarged\" REAL\n",

")\n",

"\"\"\")"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "3a5f86f8",

"metadata": {},

"outputs": [],

"source": [

"cursor.execute(\"\"\"\n",

"CREATE TABLE \"dimRegion\" (\n",

"\"index\" INTEGER,\n",

"\"fips\" REAL,\n",

"\"province\_state\" TEXT,\n",

"\"country\_region\" TEXT,\n",

"\"latitude\" REAL,\n",

"\"longititude\" REAL,\n",

"\"county\" TEXT,\n",

"\"state\" TEXT\n",

")\n",

"\"\"\")"

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "13d221c1",

"metadata": {},

"outputs": [],

"source": [

"cursor.execute(\"\"\"\n",

"copy dimDate from 's3\_uri'\n",

"credentials 'aws\_iam\_role=arn:aws:iam:iamrole'\n",

"delimiter ','\n",

"region 'ap-south-1'\n",

"IGNOREHEADER 1\n",

"\"\"\")"

]

},

{

"cell\_type": "markdown",

"id": "5c31f831",

"metadata": {},

"source": [

"## We can also upload this script into glue and run glue jobs."

]

},

{

"cell\_type": "code",

"execution\_count": null,

"id": "de1604b1",

"metadata": {},

"outputs": [],

"source": []

}

],

"metadata": {

"kernelspec": {

"display\_name": "Python 3.8.6 64-bit ('myenv': conda)",

"language": "python",

"name": "python386jvsc74a57bd0857970f990130bbcaee778cf1846f7875676d945310dca1379fe4b5ef3d258a5"

},

"language\_info": {

"codemirror\_mode": {

"name": "ipython",

"version": 3

},

"file\_extension": ".py",

"mimetype": "text/x-python",

"name": "python",

"nbconvert\_exporter": "python",

"pygments\_lexer": "ipython3",

"version": "3.8.6"

}

},

"nbformat": 4,

"nbformat\_minor": 5

}