

Task 2 - Loading and Saving Data on AWS S3

Q1: Updates based on the instructor's comments and more brainstorming.

a. Comments:

Integrate the data with something else - ex. insurance data, hospitalization cases, the number of doctors for each specialty in each state, opioid addiction cases, etc.

b. Updates:

We are investigating more datasets to augment the open payments information. Currently, we add the data of prescription drugs prescribed by individual physicians and other health care providers. We are thinking that there might be a relationship between two datasets. We select the year 2019 as our analyzing subject. So the datasets uploaded are as follows,

- BigSets/MUP_DPR_RY21_P04_V10_DY19_NPIBN_1.csv: Medicare Provider Utilization and Payment Data in 2019 [\[link\]](#)
- BigSets/OP_DTL_GNRL_PGYR2019_P06302021.csv: Open Payment data in 2019 [\[link\]](#)
- Datasets/OP_PH_PRFL_SPLMTL_P06302021.csv: Open Payment Physician Information data in 2019
- Datasets/VSRR_Provisional_Drug_Overdose_Death_Counts.csv: Provisional Drug Overdose Death Counts [\[link\]](#)

Q2: Load data to S3

Amazon S3 > usf-msds694-openpayments

usf-msds694-openpayments [Info](#)

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (4)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	BigSets/	Folder	-	-	-
<input type="checkbox"/>	Datasets/	Folder	-	-	-
<input type="checkbox"/>	Notebooks/	Folder	-	-	-
<input type="checkbox"/>	README.md	md	November 11, 2021, 23:28:20 (UTC-08:00)	14.0 B	Standard

Q3: Check the size of the bucket with the loaded data

```
$ aws s3api list-objects --bucket "usf-msds694-openpayments" --output json --query "[sum(Contents[].Size), length(Contents[])]"
```

```
1 ! aws s3api list-objects --bucket "usf-msds694-openpayments" \  
2 --output json --query "[sum(Contents[].Size), length(Contents[])]"  
  
[  
  9732296517,  
  16  
]
```

Q4: Connection between EMR and S3 (by EMR)

The screenshot shows a PySpark Jupyter Notebook interface. The left sidebar displays a file explorer with a file named 'Task2_Load_Test.ipynb' modified 'a minute ago'. The main area shows the notebook content with three code cells:

```
[1]: import pyspark  
import os  
os.environ['PYSPARK_SUBMIT_ARGS'] = '--packages "org.apache.hadoop:hadoop-aws:3.3.1" pyspark-shell'
```

Below the code, the output shows the Spark application starting successfully:

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
13	application_1636769299535_0015	pyspark	idle			✓

SparkSession available as 'spark'.

```
[2]: sc = pyspark.SparkContext.getOrCreate()  
rdds = sc.wholeTextFiles("s3n://usf-msds694-openpayments/Datasets/")  
rdds.count()
```

Below this code, the output shows the Spark job progress bar.

```
[3]: sc.stop()
```

Below this code, the output shows the Spark job progress bar.

```
[ ]:
```