

วิธีใช้งาน boto3 ใน Python เพื่อเชื่อมต่อกับ Amazon S3
<https://youtu.be/8MrUrP8uiWo>

วิธีใช้งาน dbt กับ Amazon Redshift
<https://youtu.be/P-4PBG69tF0>

วิธี Copy ข้อมูลจาก S3 เข้า Table ใน Amazon Redshift
<https://youtu.be/zgvROm5H2Vk>

GitHub โค้ดตัวอย่าง
<https://github.com/zkan/swu-ds525/tree/main/08-capstone-project>

ลองดูอันนี้ครับ
<https://anant.us/blog/modern-business/airflow-and-spark-running-spark-jobs-on-airflow-docker-based-solution/>

อันนี้ repo ในบทความ <https://github.com/yTek01/docker-spark-airflow>

อาจารย์คะ ขอโทษนะคะ เราเริ่มงงstep ที่ต้องทำอะไรคะ

คือประมาณนี้รีเพล่าคะ

1. เราต้องใช้ boto3 เพื่อเชื่อมต่อเครื่องเรากับ S3 <https://youtu.be/8MrUrP8uiWo>
2. นำข้อมูลเข้า s3
3. สร้างตารางแบบที่หัดทำวันนี้
4. ใช้ dbt เหมือนในคลิปวิธีใช้งาน dbt กับ Amazon Redshift <https://youtu.be/P-4PBG69tF0>
5. ทำ data modelling
6. ทำ dashboard โดย tableau อาจเป็นการ โหลดcsv file มาใช้ก็ได้

```
ddd_v1_w_foQ_1383611@runweb69697:~$ cat ~/.aws/credentials
[default]
aws_access_key_id =
aws_secret_access_key =
aws_session_token =
```

Boto3 client s3

root

```
|-- Employee_Name: string (nullable = true)
|-- EmpID: string (nullable = true)
|-- MarriedID: string (nullable = true)
|-- MaritalStatusID: string (nullable = true)
|-- GenderID: string (nullable = true)
|-- EmpStatusID: string (nullable = true)
|-- DeptID: string (nullable = true)
|-- PerfScoreID: string (nullable = true)
|-- FromDiversityJobFairID: string (nullable = true)
|-- Salary: string (nullable = true)
|-- Termd: string (nullable = true)
|-- PositionID: string (nullable = true)
|-- Position: string (nullable = true)
|-- State: string (nullable = true)
|-- Zip: string (nullable = true)
|-- DOB: string (nullable = true)
|-- Sex: string (nullable = true)
|-- MaritalDesc: string (nullable = true)
|-- CitizenDesc: string (nullable = true)
|-- HispanicLatino: string (nullable = true)
|-- RaceDesc: string (nullable = true)
|-- DateofHire: string (nullable = true)
|-- DateofTermination: string (nullable = true)
|-- TermReason: string (nullable = true)
|-- EmploymentStatus: string (nullable = true)
|-- Department: string (nullable = true)
|-- ManagerName: string (nullable = true)
|-- ManagerID: string (nullable = true)
|-- RecruitmentSource: string (nullable = true)
|-- PerformanceScore: string (nullable = true)
|-- EngagementSurvey: string (nullable = true)
|-- EmpSatisfaction: string (nullable = true)
|-- SpecialProjectsCount: string (nullable = true)
|-- LastPerformanceReview_Date: string (nullable = true)
|-- DaysLateLast30: string (nullable = true)
|-- Absences: string (nullable = true)
```

https://public.tableau.com/app/profile/peeyapak.somvitoon/viz/finalProject_16713669641350/Story1

ปัญหา แผนกHR ต้องการdashboard รายละเอียดเงินเดือนพนักงาน, จำนวนคน, แหล่งหางานของพนักงาน, ประสิทธิภาพการทำงานของพนักงาน โดยทางแผนกปกติจะมีการทำรายละเอียดใส่ในcsv ไฟล์อยู่แล้ว

Columnในไฟล์ HRDataset_v14.csv

Employee_Name
EmpID
MarriedID
MaritalStatusID
GenderID
EmpStatusID
DeptID
PerfScoreID
FromDiversityJobFairID
Salary
Termd
PositionID
Position
State
Zip
DOB
Sex
MaritalDesc
CitizenDesc
HispanicLatino
RaceDesc
DateofHire
DateofTermination
TermReason
EmploymentStatus
Department
ManagerName
ManagerID
RecruitmentSource
PerformanceScore
EngagementSurvey
EmpSatisfaction
SpecialProjectsCount
LastPerformanceReview_Date
DaysLateLast30
Absences

จึงออกแบบ data model ใหม่เป็น

employee_details	salary	sex	recruitment
Employee_Name	EmpID	EmpID	EmpID
EmpID	Employee_Name	Employee_Name	Employee_Name
Salary	Salary	Sex	RecruitmentSource
Position	Department	Department	
Sex			
MaritalDesc			
EmploymentStatus			
Department			
RecruitmentSource			
PerformanceScore			
Absences			

ขั้นตอนการทำงาน

1. สร้าง S3 เพื่อเป็นที่เก็บข้อมูล
2. สร้างredshift
3. Upload HRDataset_v14.csv มาที่ gitpod
4. ติดตั้งboto3 และเตรียมaws เพื่อconnect ไป s3
5. Run main.py เพื่อส่งไฟล์ HRDataset_v14.csv ไปยัง s3
6. สร้างdata model ที่ spark local
ใช้docker-compose
ใช้งานspark local
Run file “etl_local_final.py” เพื่อสร้างdata model ชุดใหม่
ติดปัญหาsave ใน folder work ไม่ได้ ต้อง save ไว้ด้านนอก folder work
7. Download “etl_local_final.py” และ ไฟล์datamodel ใหม่มาที่ git pot
8. Run senddatamodeltos3.py เพื่อส่ง ไฟล์datamodel ใหม่ไปยัง S3
9. Download ไฟล์ เพื่อทำDashboard จาก Tableau
10. Publish
https://public.tableau.com/app/profile/peeyapak.somvitoon/viz/finalProject_16713669641350/Story1

รายละเอียด

1. สร้าง S3 เพื่อเป็นที่เก็บข้อมูล
2. Upload HRDataset_v14.csv มาที่ gitpod
3. ติดตั้งboto3 และเตรียมaws เพื่อconnect ไป s3
4. Run main.py เพื่อส่งไฟล์ HRDataset_v14.csv ไปยัง s3

The screenshot shows a Gitpod workspace interface. On the left is the Explorer sidebar with a file tree for 'SWU-DS525' and '08-capstone-project'. The main editor area shows the 'requirements.txt' file with the following content:

```
1 boto3==1.26.32
2 botocore==1.29.32
3 jmespath==1.0.1
4 s3transfer==0.6.0
5
```

Below the editor is a terminal window with the following output:

```
gitpod /workspace/swu-ds525/08-capstone-pr
gitpod /workspace/swu-ds525/08-cap
gitpod /workspace/swu-ds525/08-capstone
gitpod /workspace/swu-ds525/08-capstone-
gitpod /workspace/swu-ds525/08-capstone-
gitpod /workspace/swu-ds525/08-capstone-project (main)
gitpod /workspace/swu-ds525/08-capstone-project (main)
gitpod /workspace/swu-ds525/08-capstone-project (main)
$
```

The screenshot shows the AWS Learner Lab interface. The top navigation bar includes 'ALLv1-24...' and 'Modules > Learner Lab'. The main content area shows the 'Learner Lab' section with a 'Start Lab' button and a 'Used \$8 of \$100' indicator. The bottom section shows a terminal window with the command 'ddd_v1_w_foQ_1383611@runweb69697:~\$'.

```

ddd_v1_w_foQ_1383611@runweb69697:~$ cat ~/.aws/credentials
[default]
aws_access_key_id = ASIARC3CKZR2UOP0SJMU
aws_secret_access_key = jYq71XRw4OMvJ+f1TXsdnQMT+eBN4TfTRx3irP/5
aws_session_token = FwoGZXIvYXdzEAcADdgO+f4wWHiPh48Q9iLPAfkBEt2v
1LmgfiFRNhJ97UwhpZUwtya6Rwg35mk1zHd6iGEJwYdPXBYLBBEi7bBGs08rYNcv
JUDEuLM2AHd01acInDjvS1/iB/u+sh1RIx4FiysvGUPJ2WjmY5Vd09zoDnFau6/F
G8trUIueChui0ZGXXKDYgvBT1CA4rKp3jcn1Hu3ZgWky6xEQBG7+ldfThepIdMMU
ewB9eZFq/mQcMRBu0a2pTAbbyq9rVrHz4CwJaekKCuH/ECdUuzm7501bhbE/ujMwo
jc0GeM+D2Zp47S131/ecBjIthM0HoZ/y3cVg1jKG07pNEIqeb4wub0vMjgx+07+s
nSnX+Uj0n1ViqrUK1p5Q
ddd_v1_w_foQ_1383611@runweb69697:~$ ^C
ddd_v1_w_foQ_1383611@runweb69697:~$ 

```

08-capstone-project > main.py

```

1  import boto3
2
3  aws_access_key_id = "ASIARC3CKZR2UOP0SJMU"
4  aws_secret_access_key = "jYq71XRw4OMvJ+f1TXsdnQMT+eBN4TfTRx3irP/5"
5  aws_session_token = "FwoGZXIvYXdzEAcADdgO+f4wWHiPh48Q9iLPAfkBEt2v"
6
7  client = boto3.client(
8      's3',
9      aws_access_key_id=aws_access_key_id,
10     aws_secret_access_key=aws_secret_access_key,
11     aws_session_token=aws_session_token
12 )
13 print(client)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

gitpod /workspace/swu-ds525/08-capstone
gitpod /workspace/swu-ds525/08-capstone-
gitpod /workspace/swu-ds525/08-capstone-project (main
gitpod /workspace/swu-ds525/08-capstone-project (main)
gitpod /workspace/swu-ds525/08-capstone-project (main) $ p
ip freeze > requirements.txt
gitpod /workspace/swu-ds525/08-capston
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<boto3.client.S3 object at 0x7f4b75685640>
gitpod /workspace/swu-ds525/08-capstone-project (main) ^

```

```

main.py U x
08-capstone-project > main.py
1  import boto3
2
3  aws_access_key_id = "ASIARC3CKZR2UOPOSJMU"
4  aws_secret_access_key = "jYq7lXRw40MvJ+f1TXsdnQMT+eBN4TfTRx3irP/5"
5  aws_session_token = "FwoGZXIvYXZlEAcaDDg0+f4wWlHiph48Q9iLPAfkBEt2v1LmgfifRNhJ97UwhpZUwtya6Rwg35mk1zHd6iGEJwYdPXBYLBBEi7b8Gs08rYNcvJl"
6
7  client = boto3.client(
8      's3',
9      aws_access_key_id=aws_access_key_id,
10     aws_secret_access_key=aws_secret_access_key,
11     aws_session_token=aws_session_token
12 )
13 print(client)
14
15 response = client.list_objects(
16     Bucket = "kikbucket",
17     MaxKeys=2,
18 )
19
20 print(response)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

    print(response["Contents"])
KeyError: 'Contents'
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7f2a77c73640>
{'ResponseMetadata': {'RequestId': 'EBSPD3HG1YN2Y23T', 'HostId': 'qMmv7M+rGqk0xhlj7F6PrL5+P7/xNFBuckg3hBt4qfyoGP+szQMZcYgzWob1t4ALBNfOdIcvXe0=', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-amz-id-2': 'qMmv7M+rGqk0xhlj7F6PrL5+P7/xNFBuckg3hBt4qfyoGP+szQMZcYgzWob1t4ALBNfOdIcvXe0=', 'x-amz-request-id': 'EBSPD3HG1YN2Y23T', 'date': 'Sat, 17 Dec 2022 14:58:27 GMT', 'x-amz-bucket-region': 'us-east-1', 'content-type': 'application/xml', 'transfer-encoding': 'chunked', 'server': 'AmazonS3'}, 'RetryAttempts': 0}, 'IsTruncated': False, 'Marker': '', 'Name': 'kikbucket', 'Prefix': '', 'MaxKeys': 2, 'EncodingType': 'url'}

```

main.py U x {} github_events_01.json 1

08-capstone-project > main.py

```

5  aws_session_token = "FwoGZXIvYXZlEAcADDgO+f4wWHipH48Q9iLPAfkBE
6
7  client = boto3.client(
8      's3',
9      aws_access_key_id=aws_access_key_id,
10     aws_secret_access_key=aws_secret_access_key,
11     aws_session_token=aws_session_token
12 )
13 print(client)
14
15 response = client.list_objects(
16     Bucket = "kikbucket",
17     MaxKeys=2,
18 )
19
20 #print(response["Contents"])
21
22 contents = response["Contents"]
23
24 for content in contents:
25     print(content["Key"], content["Size"])

```

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

KeyError: 'contents'
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7fcae5159640>
Traceback (most recent call last):
  File "main.py", line 25, in <module>
    print(contents["Key"], contents["Size"])
TypeError: list indices must be integers or slices, not str
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7f730b6ca640>
Traceback (most recent call last):
  File "main.py", line 25, in <module>
    print(contents["Key"])
TypeError: list indices must be integers or slices, not str
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7f8cfd0c5640>
events_json_path.json 125
github_events_01.json 70104
gitpod /workspace/swu-ds525/08-capstone-project (main) $

```


main.py U X

HRDataset_v14.csv U

github_events_01.json 1

08-capstone-project > main.py

```
1  import boto3
2
3  aws_access_key_id = "ASIARC3CKZR2UOP0SJMU"
4  aws_secret_access_key = "jYq7lXRw40MvJ+f1TXsdnQMT+eBN4TfTRx3irP/5"
5  aws_session_token = "FwoGZXIvYXdzEAcADDgO+f4wWHiph48Q9iLPAfkBEt2v1Lmgfi"
6
7  s3 = boto3.resource(
8      "s3",
9      aws_access_key_id=aws_access_key_id,
10     aws_secret_access_key=aws_secret_access_key,
11     aws_session_token=aws_session_token
12 )
13 s3.meta.client.upload_file(
14     "HRDataset_v14.csv",
15     "kik-bucket-public",
16     "HRDataset_v14.csv",
17 )
18
19
20
21 # client = boto3.client(
22 #     "s3",
```

PROBLEMS

1

OUTPUT

DEBUG CONSOLE

TERMINAL

PORTS

```
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7fcae5159640>
```

```
Traceback (most recent call last):
```

```
  File "main.py", line 25, in <module>
    print(contents["Key"], contents["Size"])
```

```
TypeError: list indices must be integers or slices, not str
```

```
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7f730b6ca640>
```

```
Traceback (most recent call last):
```

```
  File "main.py", line 25, in <module>
    print(contents["Key"])
```

```
TypeError: list indices must be integers or slices, not str
```

```
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
<botocore.client.S3 object at 0x7f8cfd0c5640>
```

```
events_json_path.json 125
```

```
github_events_01.json 70104
```

```
gitpod /workspace/swu-ds525/08-capstone-project (main) $ python main.py
```

```
gitpod /workspace/swu-ds525/08-capstone-project (main) $
```

kik-bucket-public [Info](#)[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (1)

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](#)< 1 > [Settings](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	HRDataset_v14.csv	csv	December 17, 2022, 22:33:40 (UTC+07:00)	75.1 KB	Standard

aws

Services

Search

[Alt+S]

N. Virginia

voclabs/user1591029=peeyapak.somlitoon@g.sww.ac.th @ 0748-312

Amazon Redshift

Redshift serverless [New](#)

Provisioned clusters dashboard

Clusters

Reserved nodes

Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Configurations

AWS Partner Integration

Informatica Data Loader

Advisor

AWS Marketplace

Alarms

Events

What's new

Amazon Redshift is creating redshift-cluster-1

Sample data load
After the cluster is created, Amazon Redshift starts to load the sample data.

Try new Amazon Redshift features in preview.

Create a cluster with preview features. Production use of the cluster is not supported. Use this cluster for testing only.

Create preview cluster

Amazon Redshift > Clusters

In my account | From other accounts

Connect to Redshift clusters

Query data using Redshift query editor

Use the query editor v2 to run queries in your Redshift cluster.

Query data

Work with your client tools

You can connect to Amazon Redshift from your client tools, such as SQL clients, business intelligence (BI) tools, and extract, transform, load (ETL) tools, using JDBC or ODBC drivers.

Cluster

Cluster identifier

Copy JDBC URL

Copy ODBC URL

Choose your JDBC or ODBC driver

Use JDBC or ODBC drivers to connect to Amazon Redshift from your client tools, such as SQL clients, BI tools, and ETL tools. We recommend using the new Amazon Redshift-specific drivers for better performance and scalability.

Driver

JDBC 4.2 without AWS SDK (jar)

Download driver

Clusters (1) Info

Refresh

Query data

Actions

Create cluster

Filter clusters by property or value

Cluster

Status

Cluster namespace

Availability Zone

Multi-AZ

Storage capacity us...

CPU utilization

redshift-cluster-1

ra3.4xlarge | 2 nodes | 256 ...

Creating

560c1d29-fe1a-4fb7-...

-

No

Amazon Redshift

Redshift serverless

Provisioned clusters dashboard

Clusters

- Reserved nodes
- Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Configurations

Amazon Redshift query editor v2 is now available

Query editor v2 provides new features such as multistatement query execution, query parameterization, query versioning, visualizations, and query sharing. [Learn more](#)

Go to query editor v2

Amazon Redshift > Clusters > redshift-cluster-1

redshift-cluster-1

ActionsEditAdd partner integrationQuery data

General information

Cluster identifier redshift-cluster-1	Status Available	Node type ra3.4xlarge	Endpoint redshift-cluster-1.cbdkb6riw...
Cluster namespace 560c1d29-fe1a-4fb7-9d22-deaeddab65ca	Date created December 18, 2022, 10:31 (UTC+07:00)	Number of nodes 2	JDBC URL jdbc:redshift://redshift-clust...
Cluster configuration Production	Storage used 0.00% (0.00 of 256 TB used)		ODBC URL Driver={Amazon Redshift (x...
	Multi-AZ No		

Edit cluster redshift-cluster-1

Cluster configuration

Cluster identifier

This is the unique key that identifies a cluster.

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

▼ Backup

Automated snapshots are periodic backups of the cluster.

Automated snapshot retention period

Specify how many days to retain automated snapshots.

The retention period must be 1-35 days.

Manual snapshot retention period

Specify how long do you want to retain your snapshot.

The retention period must be 1-3653 days.

Cluster relocation

Enable the ability to relocate your cluster in another Availability Zone. After you enable relocation, you use the VPC endpoint of the cluster to determine the cluster IP address, instead of the leader node IP address. You can find the VPC endpoint in the Network and security section of the cluster details page. [Learn more](#) 

☒ Disabled

☐ Enabled

► Network and security [Info](#)

Edit network and security

▼ Network and security [Info](#)

Virtual private cloud (VPC)

This VPC defines the virtual networking environment for this cluster.

vpc-0644f8397501301da

VPC security groups

This VPC security group defines which subnets and IP ranges the cluster can use in the VPC.

Choose one or more security groups ▼

default



sg-0b7184600e1e31554

Cluster subnet group

Choose the Amazon Redshift subnet group to launch the cluster in.

default

Availability Zone

Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

Enhanced VPC routing

Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. [Learn more](#)

☒ Turn off

☐ Turn on

Publicly accessible

☒ Turn on Publicly accessible

Allow public connections to Amazon Redshift.



When you turn on this feature, clients can connect to the database from outside the VPC.

redshift-cluster-1

Actions ▼

Edit

Add partner integration

Query data ▼

General information

Cluster identifier
redshift-cluster-1Status
✔ AvailableNode type
ra3.4xlarge

✔ Endpoint copied

redshift-cluster-1.cbdkb6riw...

Cluster namespace
560c1d29-fe1a-4fb7-9d22-
deaeddab65caDate created
December 18, 2022, 10:31
(UTC+07:00)Number of nodes
2

JDBC URL

jdbc:redshift://redshift-clust...

Cluster configuration
ProductionStorage used
0.00% (0.00 of 256 TB used)

ODBC URL

Driver={Amazon Redshift (x...

Multi-AZ
No

Connect to database



Connection

Select a recent database connection or create a new database connection.

☒ Use a recent connection☐ Create a new connection

Recent connection

redshift-cluster-1 (Available)

Database: dev User: awsuser Authentication: Temporary credentials ▼

Cancel

Connect

Create notebook



Error

User: arn:aws:sts::074831285365:assumed-role/voclabs/user1591029=peeyapak.somvitoon@g.swu.ac.th is not authorized to perform: iam:PassRole on resource: arn:aws:iam::074831285365:role/LabRole because no identity-based policy allows the iam:PassRole action

- สร้างdata model ที่ spark local
ใช้docker-compose
ใช้งานspark local
Run file “etl_local_final.py” เพื่อสร้างdata model ชุดใหม่
ติดปัญหาsave ใน folder work ไม่ได้ ต้อง save ไว้ด้านนอก folder work
- Download “etl_local_final.py” และ ไฟล์datamodel ใหม่มาที่ git pot
- Run senddatamodeltos3.py เพื่อส่ง ไฟล์datamodel ใหม่ไปยัง S3

```

[28]: from pyspark.sql import SparkSession
      # from pyspark.sql.types import StructType, StructField, DoubleType, StringType, IntegerType, DateType, TimestampType
      # import pyspark.sql.functions as F

[29]: import pandas as pd
      import glob

[30]: p = glob.glob("data/*.csv")

[31]: p

[32]: ['data/HRDataset_v14.csv']

[33]: data = "HRDataset_v14.csv"

[34]: spark = SparkSession.builder \
      .appName("ETL") \
      .getOrCreate()

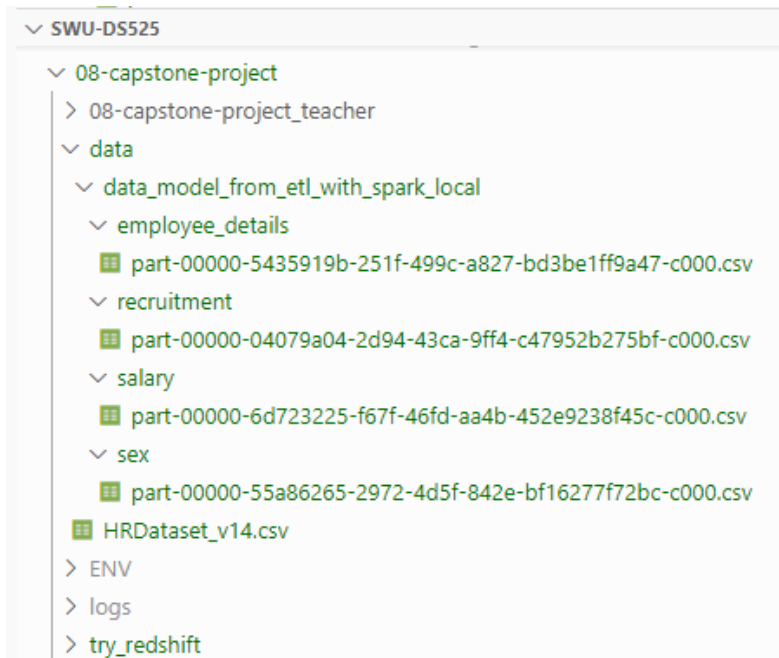
[35]: data_folder = "data"

[36]: data = spark.read.option("header", "true").option("multiline", "true").csv(data_folder)

[37]: data.show()

+-----+
| Employee_Name|EmpID|MarriedID|MaritalStatusID|GenderID|EmpStatusID|DeptID|PerfScoreID|FromDiversity|JobFairID|Salary|TermID|PositionID|
| Position(State| Zip| DOB|Sex|MaritalDesc| CitizenDesc|HispanicLatino| RaceDesc|DateofHire|DateofTermination|
| TermReason| EmploymentStatus| Department| ManagerName|ManagerID| RecruitmentSource|PerformanceScore|EngagementSurvey|EmpSatisfaction|SpecialProjectsCount|LastPerformanceReview_Date|DaysLateLast30|Absences|
+-----+

```



Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3

Buckets

kik-bucket-public

Info

Objects

Properties

Permissions

Metrics

Management

Access Points

Objects (5)

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](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

< 1 >

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	employee_details.csv	csv	December 18, 2022, 17:05:10 (UTC+07:00)	33.2 KB	Standard
<input type="checkbox"/>	HRDataset_v14.csv	csv	December 17, 2022, 22:33:40 (UTC+07:00)	75.1 KB	Standard
<input type="checkbox"/>	recruitment.csv	csv	December 18, 2022, 17:06:49 (UTC+07:00)	10.6 KB	Standard
<input type="checkbox"/>	salary.csv	csv	December 18, 2022, 17:06:49 (UTC+07:00)	12.1 KB	Standard
<input type="checkbox"/>	sex.csv	csv	December 18, 2022, 17:06:49 (UTC+07:00)	10.8 KB	Standard

SWU-DS525

- > 01-data-modeling-i
- > 02-data-modeling-ii
- > 03-building-a-data-warehouse
- > 04-building-a-data-lake
- > 05-creating-and-scheduling-data-pipelines
- > 06-analytics-engineering
- > 07-Data Visualization and DashboardPage

08-capstone-project

- > 08-capstone-project_teacher

data

- > data_model_from_etl_with_spark_local
- > ENV
- employee_details.csv
- HRDataset_v14.csv
- recruitment.csv
- salary.csv
- sex.csv

ENV

logs

try_redshift

aws_access_key_id

docker-compose.yml

etl_local_final.ipynb

etl_local.ipynb

main.py

8. Download ไฟล์ เพื่อทำDashboard จาก Tableau
9. Publish

https://public.tableau.com/app/profile/peeyapak.somvitoon/viz/finalProject_16713669641350/Story1