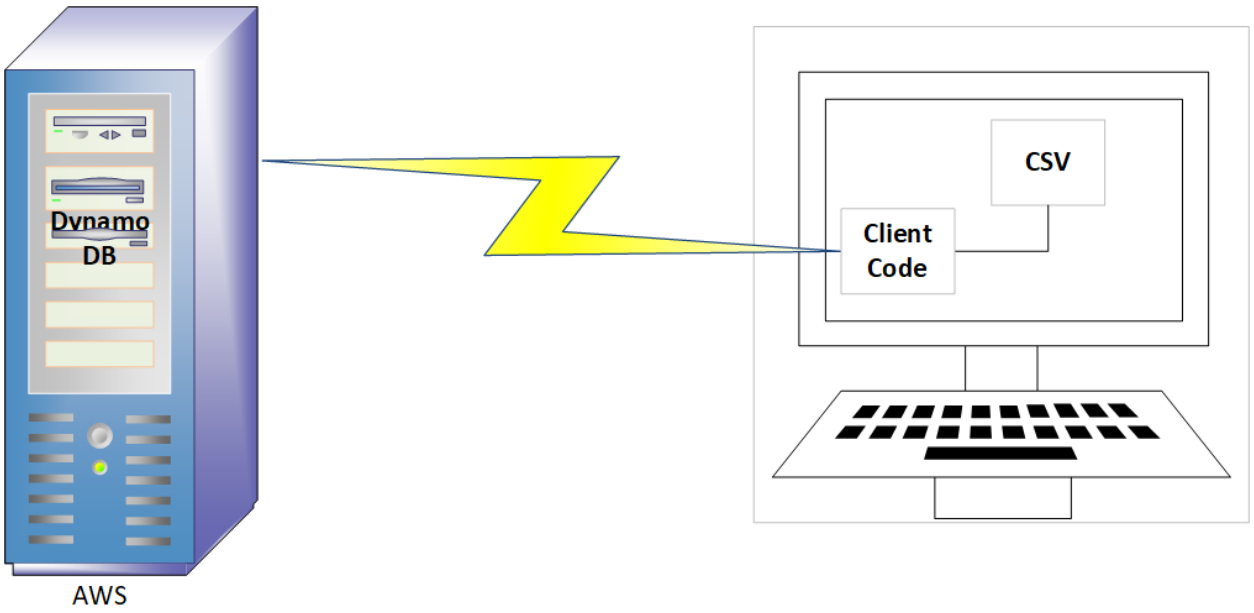
[Fall-2021]

HW-3

Deadline: October 21st 11:59PM EST (8:59PM PST)

General Description:

In this homework, you will create NoSQL database and populate it with data.



Problem Description

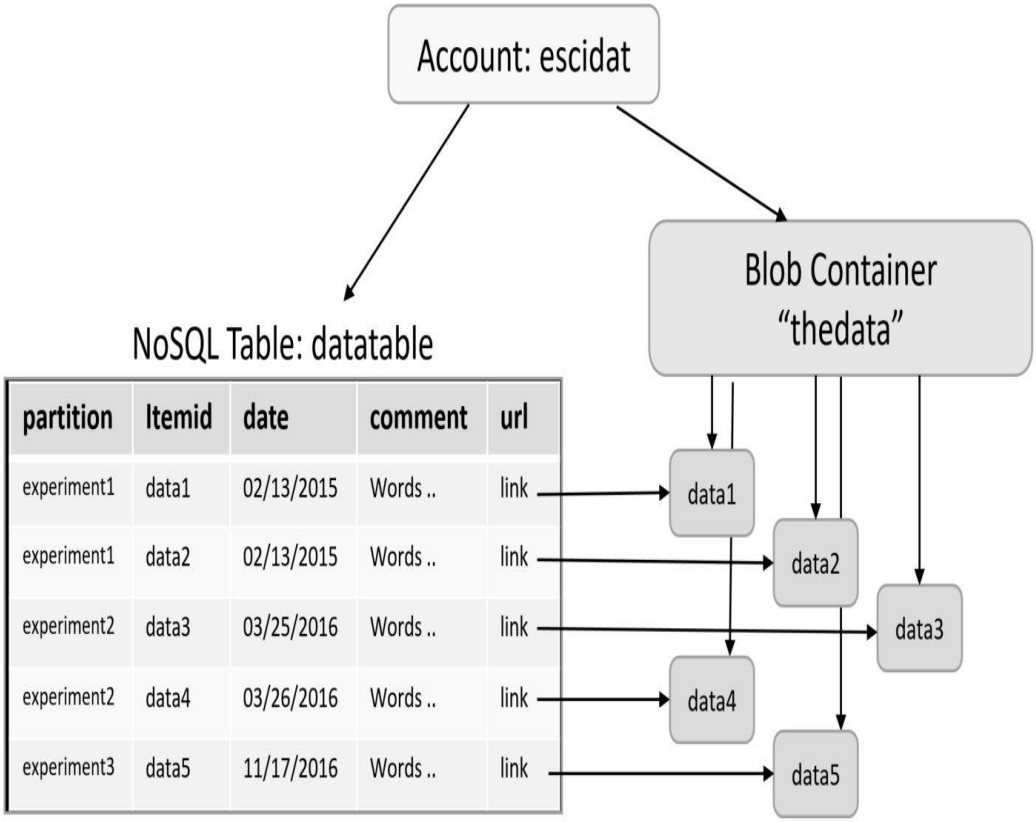
You have conducted few experiments and you would like to capture the results of your experiments. These experiments share common attributes (data points) so you decided to build regular table to hold these common attributes. However, these experiments may have different attributes as well. Therefore, they should be captured separately.

You have identified that NoSQL database is the best method to store your data given their lack of structure. NoSQL DB allows you to upload your experiment data – with their distinctions- to a one NoSQL DB.

General information about all experiments are stored in master CSV file. Each record in the master CSV provides general information about all experiments. This CSV is useful to capture common data points among all experiments.

However, each experiment has its own set of data points that differ from other experiments. So, each record in the master CSV will include a pointer to a file on the local hard disk to provide all the special data points that are unique to each experiment.

Below, a screenshot is provided for clarification on how the data shall be stored on your NoSQL database.



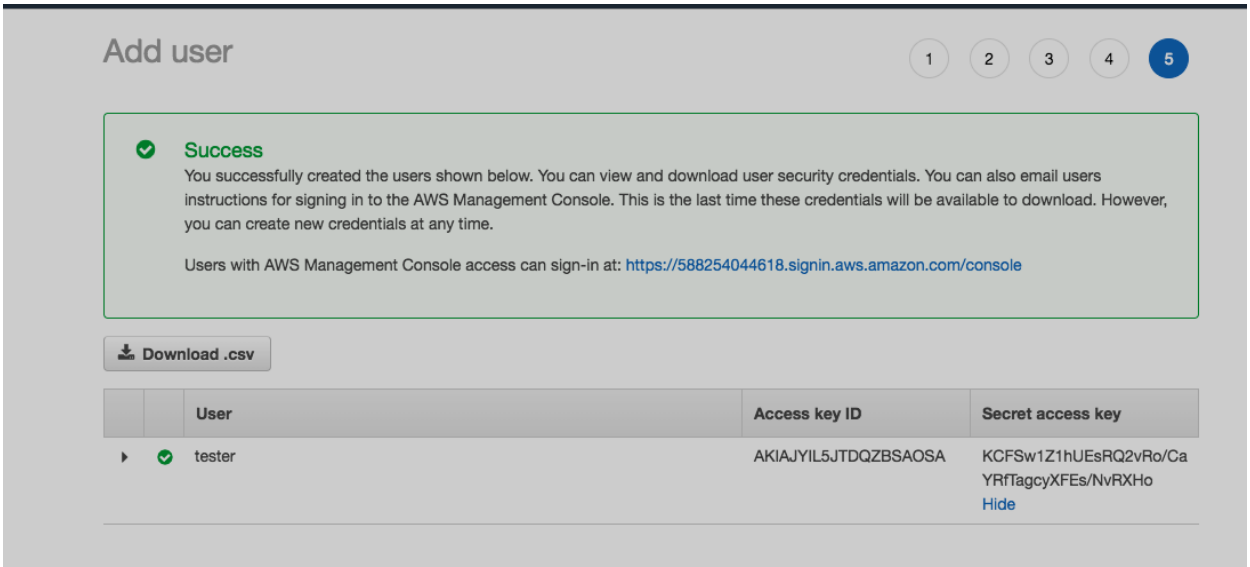
You are requested to develop an application to upload your CSV data (Master CSV and other CSVs) to one NoSQL DynamoDB.

In order to do so, please execute the following steps:

- Create DynamoDB table.
- Read the master CSV for Experiments
- Read the data for each experiment based on the location stored in the master CSV
- Upload the BLOB data to your NoSQL DB and Fill your NoSQL DB with experiment data

To complete this assignment, you need to:

- Setup AWS account and download security credential from the Amazon IAM (**I**dentity and **A**ccess **M**anagement)



- You may choose the programming language to use. However, if you decided to use Python, you can use the Amazon Python Boto3 SDK (https://pypi.org/project/boto3/)
- Boto3 Python library will allow you to:
    - access AWS resources using a pair of access and secret keys
    - Create AWS S3 bucket to host your DynamoDB
        - *Hint#1: make your S3 bucket publicly accessible.*
        - *Hint#2: You may face some problems with creating your S3 buckets due to selecting a reserved bucket name, so try different bucket names if you face such issue. Use the following naming conventions guidelines*

- Upload your data as blobs to the S3 bucket

```
import boto3
s3 = boto3.resource('s3',
    aws_access_key_id='YOUR ACCESS KEY',
    aws_secret_access_key='your secret key' )
```

```
s3.create_bucket(Bucket='datacont', CreateBucketConfiguration={
    'LocationConstraint': 'us-west-2'})
```

```
# Upload a file, 'test.jpg' into the newly created bucket
s3.Object('datacont', 'test.jpg').put(
    Body=open('/home/mydata/test.jpg', 'rb'))
```

- Create a DynamoDB table to store metadata and references to S3 objects

```
dyndb = boto3.resource('dynamodb', region_name='us-west-2' )

# The first time that we define a table, we use
table = dyndb.create_table(
    TableName='DataTable',
    KeySchema=[
        { 'AttributeName': 'PartitionKey', 'KeyType': 'HASH'},
        { 'AttributeName': 'RowKey', 'KeyType': 'RANGE' }
    ],
    AttributeDefinitions=[
        { 'AttributeName': 'PartitionKey', 'AttributeType': 'S' },
        { 'AttributeName': 'RowKey',       'AttributeType': 'S' }
    ]
)

# Wait for the table to be created
table.meta.client.get_waiter('table_exists')
                            .wait(TableName='DataTable')

# If the table has been previously defined, use:
# table = dyndb.Table("DataTable")
```

- Read the metadata from a CSV file, Move the data objects into the blob store, and Enter the metadata row into the table

```
import csv
urlbase = "https://s3-us-west-2.amazonaws.com/datacont/"
with open('\path-to-your-data\experiments.csv', 'rb') as csvfile:
    csvf = csv.reader(csvfile, delimiter=',', quotechar='|')
    for item in csvf:
        body = open('path-to-your-data\datafiles\\'+item[3], 'rb')
        s3.Object('datacont', item[3]).put(Body=body)
        md = s3.Object('datacont', item[3]).Acl()
            .put(ACL='public-read')
        url=urlbase +item[3]
        metadata_item={'PartitionKey': item[0], 'RowKey': item[1],
            'description' : item[4], 'date' : item[2], 'url':url}
        table.put_item(Item=metadata_item)
```

P.S. Full code (not fully runnable) is provided the support guide with this homework.

P.S. you need to try AWS Free trial account instead of the AWS student account to see all policies. You might be asked to put your credit card information, but you shouldn't be charged. The amount of AWS credits that are available for free trials surpasses by far the credits you need to complete this experiment.

P.S. I have uploaded sample dummy data for clarity with this assignment but feel free to choose your column names and dataset.

Submission Guidelines:

- Post URL for your GitHub repository to Canvas. Make sure to keep your GitHub repository public.
- Create a folder in your GitHub repository and name it "NoSQL". Keep all the Homework related materials under this folder.
- For this homework, submit the following:
    1. Your Code file(s) (50%)
    2. Screenshot of the query you use on your local machine to pull the data from your Dynamo DB. (20%)
    3. Screenshot of the results of the above query from your local machine's terminal. (30%)

Common Penalties:

- Your GitHub repository is not public: 100% reduction (won't be graded)
- Late submissions on Canvas or GitHub: 100% reduction (won't be graded)