Try it out objective

Use this hands-on to get started with the analytical service Athena.

The goal

The following are the goals of this hands-on:

1. Create a bucket and upload CSV data
2. Query data stored in S3 (data lake) via Athena
3. Develop understanding of schema definition and data analysis

Important - this exercise **will not work in AWS Academy**, a personal AWS account is mandatory. This exercise is for technical learners only.

Please note if a field (short for text field/text area/checkbox/radio/dropdown/list or any other UI element) is not specified in the following steps, it means the default value of the field set by AWS needs to be used. No change is needed for those fields as part of this hands-on.

This exercise will work with multiple services, please use a dedicated browser instance with only the tabs that are needed for this exercise, otherwise it may/will lead to confusion.

A. Hands-on: Setup S3 bucket (Athena cache location)

1. Go to the S3 management console at https://console.aws.amazon.com/s3/ (you will be required to sign in)
2. Observe the page will show the list of buckets, alternatively will prompt to create a bucket
3. Click on the **Create bucket** button
4. Under the **General configuration** card make the following changes
   a) For **Bucket name** field use the folowing -

   ```
   pgpccmmddhhmm
   ```

   mmddhhmm represents Month, day, hour and minute, a valid value could be pgpcc12251538 (25th December 15hrs 38mins), do not use this example value from here, use the actual date and time at the time of doing this exercise.

   b) For the **AWS Region** dropdown ensure the value is **N Virginia**
5. Scroll to the bottom of the page and click on the **Create bucket** button
6. Click on the **Link** of the **bucket** which will now show in the list of buckets
7. Click on the **Create folder** button
8. Under the **Folder** card, for the **Folder name** field, paste the following value

   ```
   cache
   ```

9. Click on the **Create folder** button
10. Repeat the above folder creation process for the following folder as well

   ```
   ccdata
   ```

11. **Click** on the link of the folder **ccdata**
12. **Download** the **ccspend.csv** from **Olympus** on your laptop/desktop (remember this location)
13. Go back to the **S3 management console** and **Upload** the CSV in the **ccdata** folder using the instructions below
    a) Click on the **Upload** button
    b) Click on **Add files** button in the **Files and folders** card
    c) Scroll down to the bottom and click on **Upload** button
    d) Once the upload completes (message on the top of the screen turns green), click on the **Close** button (right top of the screen)
14. Click on **Amazon S3** (blue link on the left top of the page) breadcrumb trail, this will show the bucket listing again

B. Hands-on: Athena cache setup

1. Go to the Athena management console at https://console.aws.amazon.com/athena/ (you will be required to sign in if not already)
2. Ensure the region is **N Virginia**
3. In the left navigation click on the ≡ (hamburger) menu on the left top of the screen to expand it (if needed)
4. Click on **Query Editor** menu option
5. The screen defaults to the Editor tab (will be highlighted), **click** on the **Settings** tab
6. Under the **Settings** card **click** on the **Manage** button on the right side of the card
7. In the **Manage settings** card click on **Browse S3** button
8. Click on the **Link** of the **bucket created in the earlier step**

9. Click on the **Radio** to the left of the **cache** folder
10. Click on the **Choose** button and the screen will change
11. Click on the **Save** button and the screen will change back to display the tabs (**Settings** tab will be **highlighted** this time)
12. Click on the **Editor** tab

C. Hands-on: Athena schema definition (powered by Glue)

1. Observe the layout of the editor which has a **left navigation card Data** and the right side is the editor with one **New query 1** tab open
2. **Paste** the following **statement** in the **New query 1** tab

```
create database appdata;
```

3. **Click** on the **Run** button (this was disabled earlier, it was enabled after pasting the query)
4. Under the **Data** card (left side) **click** on the **Database** dropdown
5. Select **appdata** from the dropdown
6. **Delete** the **previous statement** (create database ...) from the **New query 1** tab (after this step the tab should be completely blank)
7. **Paste** the following statement in the **New query 1** tab after making the modification described below

Modification details - the text **YOURBUCKET** in the statement below needs to be **replaced** with the **bucket** name that was **created earlier** using the pgpccmmddhhmm format. Ensure the shash around the bucket name or any other info is not removed accidently.

```
CREATE EXTERNAL TABLE IF NOT EXISTS appdata.ccspend (
year_month bigint,
agency_number bigint,
last_name string,
first_name string,
agency_name string,
amount double,
vendor string,
txn_date string,
posted_date string,
description string
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES ('serialization.format' = ','field.delim' = ',')
LOCATION 's3://YOURBUCKET/ccdata/'
TBLPROPERTIES ('has_encrypted_data'='false');
```

8. **Click** on the **Run** button
9. Notice on the left navigation under the **Data card** and **Tables and views** section under the **Tables** title **ccspend** table has been **created**
10. **Delete** the **previous statement** (create external table ...) from the **New query 1** tab (after this step the tab should be completely blank)
11. **Paste** the following statement in the **New query 1** tab

```
select * from ccspend limit 10;
```

12. **Click** on the **Run** button
13. **Delete** the **previous statement** (create external table ...) from the **New query 1** tab (after this step the tab should be completely blank)
14. **Paste** the following statement in the **New query 1** tab

```
select count(1) NumTxn, agency_name
from ccspend
group by agency_name
having count(1) > 5;
```

15. **Click** on the **Run** button

D. Hands-On: Cleaning up!

1. In the Athena management console cleanup the database and the table
2. **Delete** the **previous statement** (create external table ...) from the **New query 1** tab (after this step the tab should be completely blank)
3. **Paste** the following statement in the **New query 1** tab

```
drop table ccspend;
```

4. **Click** on the **Run** button
5. **Delete** the **previous statement** (create external table ...) from the **New query 1** tab (after this step the tab should be completely blank)
6. **Paste** the following statement in the **New query 1** tab

```
drop database appdata;
```

7. **Click** on the **Run** button
8. Go to the S3 management console listing the buckets
9. Empty the bucket by clicking on the **Empty** button (in the confirmation window type **Permanently delete**), once the delete is done click on **Exit** button
10. Delete the bucket by clicking on the **Delete** button (in the confirmation window type the bucket name) and **click** on **Delete bucket** button