

# Working with Azure Data Lake Gen2 account

## Exercise 1 – Create Azure Data Lake Gen2 Account

1. Go to Azure portal (portal.azure.com)
2. In the search bar, search for Storage Accounts. And select it.
3. Click on **Create**.
4. Fill up the properties to create account:
  - a. [Basics Tab]
    - i. Select subscription
    - ii. On Resource Group, click Create new and provide a name
    - iii. Provide a unique storage account name
    - iv. Select region of your choice (example – East US 2)
    - v. You have options for selecting performance & redundancy options – keep it as is.
    - vi. Click Next

The screenshot shows the 'Basics' tab of the Azure portal's storage account creation wizard. The 'Project details' section includes a 'Subscription' dropdown set to 'MSDN Platforms' and a 'Resource group' dropdown set to '(New) OReilly', with a 'Create new' link below it. The 'Instance details' section includes a 'Storage account name' text box with 'oreillydatalake', a 'Region' dropdown set to '(US) East US 2', and 'Performance' radio buttons where 'Standard' is selected. The 'Redundancy' dropdown is set to 'Geo-redundant storage (GRS)', and the checkbox 'Make read access to data available in the event of regional unavailability' is checked.

Basics Advanced Networking Data protection Encryption Tags Review

**Project details**

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription \* MSDN Platforms

Resource group \* (New) OReilly  
[Create new](#)

**Instance details**

If you need to create a legacy storage account type, please click [here](#).

Storage account name ⓘ \* oreillydatalake

Region ⓘ \* (US) East US 2

Performance ⓘ \*

☒ Standard: Recommended for most scenarios (general-purpose v2 account)

☐ Premium: Recommended for scenarios that require low latency.

Redundancy ⓘ \* Geo-redundant storage (GRS)

☒ Make read access to data available in the event of regional unavailability.

- b. [Advanced Tab]
  - i. In **Hierarchical Namespace** section -> Select **Enable Hierarchical Namespace** checkbox.
  - ii. Click **Review + create**.

Basics **Advanced** Networking Data protection Encryption Tags Review + create

### Security

Configure security settings that impact your storage account.

Require secure transfer for REST API operations ☒

Allow enabling anonymous access on individual containers ☐

Enable storage account key access ☒

Default to Microsoft Entra authorization in the Azure portal ☐

Minimum TLS version

Permitted scope for copy operations (preview)

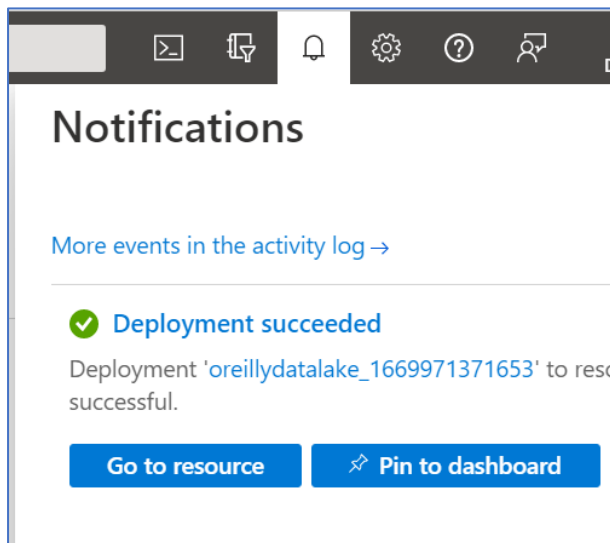
### Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) [Learn more](#)

Enable hierarchical namespace ☒

c. Click **Create**.

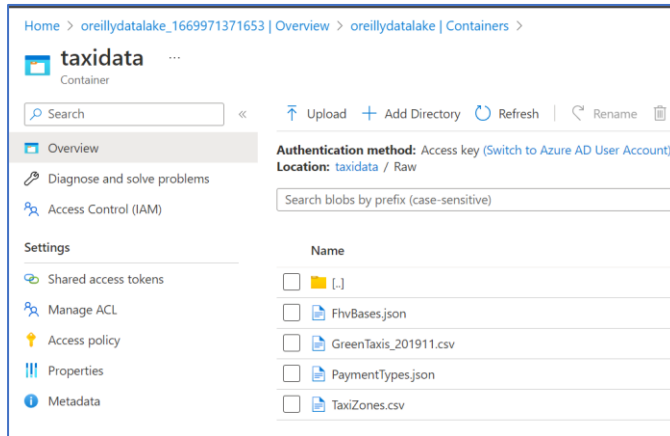
5. This will create a new Azure Data Lake Gen2 account.



## Exercise 2 – Upload Files to Azure Data Lake Gen2 Account

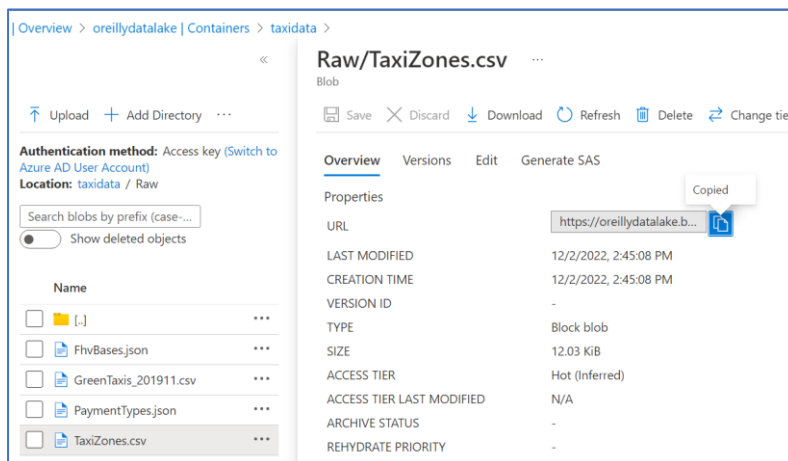
0. In a separate browser window, go to following URL and download sample files:  
<https://tinyurl.com/data-lake-bootcamp-2022>
1. Open Data Lake account created in the previous step.
2. From left pane, under **Data storage**, go to **Containers**.
3. Click on **Add container**, fill up the properties, and click **Create**:
  - a. Name: taxidata
  - b. Public access level: Private

4. Once created, open **taxidata** container.
5. To store raw data, click on **Add Directory**, fill up the properties, and click **Save**:
  - a. Name: Raw
6. Click on **Raw** directory to open it.
7. In Raw directory, upload the files downloaded in Step 0.



## Exercise 3 – Generate and use SAS Token

1. In taxidata container, click on TaxiZones.csv file.
2. Copy file URL for TaxiZones.csv.



3. Paste URL in browser and see if its accessible.
  - This should not work, since no access is available.
4. Click on Generate SAS, and define the permissions:
  - a. Permissions: Read
  - b. Start: Keep one day before to avoid Timezone issues
  - c. End: One month ahead in future
  - d. Click on Generate SAS token and URL

**Raw/TaxiZones.csv** ...

Blob

Save Discard Download Refresh Delete

Overview Versions Edit **Generate SAS**

A shared access signature (SAS) is a URL that grants restricted access to an Azure Storage blob. Use it when you want to share access to a blob without sharing your storage account key. [Learn more about creating an account SAS](#)

Signing method

☒ Account key ☐ User delegation key

Signing key ⓘ

Key 1 ▼

Stored access policy

None ▼

Permissions \* ⓘ

Read ▼

Start and expiry date/time ⓘ

Start

12/01/2022 2:52:50 PM

(UTC+05:30) Chennai, Kolkata, Mumbai, New Delhi ▼

Expiry

01/07/2023 10:52:50 PM

(UTC+05:30) Chennai, Kolkata, Mumbai, New Delhi ▼

Allowed IP addresses ⓘ

for example, 168.1.5.65 or 168.1.5.65-168.1....

Allowed protocols ⓘ

☒ HTTPS only ☐ HTTPS and HTTP

**Generate SAS token and URL**

5. Copy Blob SAS URL.
  - This will include file URL and SAS token.
6. Paste URL in browser and see if its accessible.
  - This should work, since access is provided via SAS token.

## Exercise 4 – Configure Data Lifecycle

1. From left pane of storage account, under **Data Management**, navigate to **Lifecycle management** tab.

Home > oreillydatalake\_1669971371653 | Overview > oreillydatalake

**oreillydatalake | Lifecycle management** ☆ ...

Storage account

Search << + Add a rule ✓ Enable □ Disable Refresh Delete

**Data management**

- Redundancy
- Data protection
- Blob inventory
- Static website
- Lifecycle management**

**Settings**

- Configuration

Lifecycle management offers a rich, rule-based policy for general purpose storage. Lifecycle management policy may take up to 48 hours to complete. [Learn more](#)

**List View** Code View

Enable access tracking ⓘ ☐

☐ Name

No rules

2. Click on **Add a rule**.
3. Provide the details:
  - a. Rule name: Rule1
  - b. Rule scope: Apply rule to all blobs in your storage account
  - c. Blob type: Block blobs
  - d. Blob subtype: Base blobs
  - e. Click Next

4. Define rule conditions, and click on **Save**:
- If Base blobs haven't been modified in 60 days, then move to Cool storage
  - If Base blobs haven't been modified in 180 days, then move to Archive storage
  - If Base blobs haven't been modified in 365 days, then Delete the blob

Details

Base blobs

Lifecycle management uses your rules to automatically move blobs to cooler tiers or to delete them. If you create multiple rules, the associated actions must be implemented in tier order (from hot to cool storage, then archive, then deletion).

If

Base blobs haven't been modified in 60 days

↓

Then

Move to cool storage

If

Base blobs haven't been modified in 180 days

↓

Then

Move to archive storage

If

Base blobs haven't been modified in 365 days

↓

Then

Delete the blob

↓

+ Add conditions