Hogwarts Datalakes

Getting started

What is a Datalake?

A datalake (a.k.a Storage Account) is Microsoft's object storage solution for the cloud. Blob storage is optimized for storing massive amounts of unstructured data. Unstructured data is data that doesn't adhere to a particular data model or definition, such as text or binary data.

Azure Data Lake includes all the capabilities required to make it easy for developers, data scientists, and analysts to store data of any size, shape, and speed, and do all types of processing and analytics across platforms and languages.

What are the available Datalakes and Containers?

Environment	Datalake	Containers
Unclass	hogwartsdatalakeu	data, warehouse
	hogwartsdatalakeusersu	users, analytics
PB	hogwartsdatalakepb	data, warehouse, users, analytics

Here is a description of the different containers (and the permissions applied):

- **data**: no access for alpr-members-sg. Used for transitory un-processed data (for example: data downloaded from Internet that will be loaded into a production Hive table)
- **warehouse**: read-only to alpr-members-sg. Holds the production Hive table and underlying parquet files. This is default hive metastore location when none is specified.
- **users**: full access to alpr-members-sg. This is your playground space, we highly recommend that you create a folder with your username underneath this container to store any data or files that you use to develop your analytics. Please note that all members of alpr-members-sg can read/modify files and folders stored in that container.
- **analytics**: full access to alpr-members-sg. This is to store the input and results of your production analytics. Please create a meaningful folder structure within that container. Please note that all members of alpr-members-sg can read/modify files and folders

stored in that container.

If you need to create another container (for example to further restrict access permissions) please contact us for instructions.

Accessing the Datalake

How to access the Hogwarts Datalake through the Azure portal?

Visit the Unclassified datalake portal or the PB datalake portal

How to access the Hogwarts Datalake using Spark?

Please refer to the Spark User Guide for an example.

How to access the Hogwarts Datalake in Python?

Tutorial: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-directory-file-acl-python

Reference: https://docs.microsoft.com/en-us/python/api/azure-storage-file-datalake/azure.storage.filedatalake?view=azure-python

Example Notebook: https://github.com/CybercentreCanada/hogwarts-notebooks/blob/master/3-ExploreAzureDatalake.ipynb - Connect your Github account

How to access the Datalake in Java?

Tutorial: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-directory-file-acl-java

Reference: https://docs.microsoft.com/en-us/java/api/?view=azure-java-stable

How to access the Datalake using Azure CLI?

Here is an example to list all the containers that you have access to in the hogwartsdatalakeu:

```
az login
az storage container list --account-name hogwartsdatalakeu --subscription Chimera-U --auth-mode login
```

For a list of other storage commands, try out to `az storage --help` or visit https://docs.microsoft.com/en-us/cli/azure/storage?view=azure-cli-latest

How to access the Datalake using the Azure REST API?

Please refer to: https://docs.microsoft.com/en-us/rest/api/storageservices/data-lake-storage-gen2

How to copy files from/to the Datalake?

Azure has a tool called azcopy that can be used for that purpose: https://docs.microsoft.com/en-us/azure/storage/common/storage-ref-azcopy?toc=/azure/storage/blobs/toc.json

Advanced topics

RBAC and ACL permissions on container/folder/file

Azure provides 2 ways to secure your files and folders: Role-Based Access Control (RBAC) and Access Contol List (ACL).

The smallest granularity for RBAC is at the container level and this will be evaluated at a higher priority than ACLs. Therefore, if you assign a role to a security principal in the scope of a container, that security principal has the authorization level associated with that role for ALL directories and files in that container, regardless of ACL assignments.

It is therefore important to note that all ALPR users have read+write access to the `users` and `analytics` containers which means **that they can see/modify all files stored in those containers**.

If you need to restrict permissions on a folder that contains sensitive information, please contact us for guidance.

How to create a user SAS token to access a datalake container?

Please contact us for instructions.