Managing Data with Azure Data Lake Store Gen2

Xavier Morera

PASSIONATE ABOUT ENTERPRISE SEARCH AND BIG DATA

@xmorera www.xaviermorera.com


BIG DATA

Security

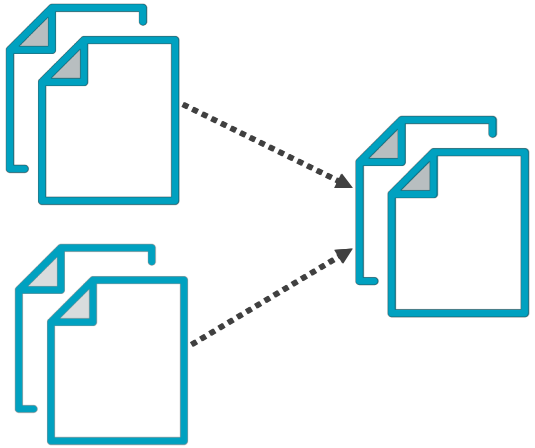
Supports ACL and POSIX permissions

- Extra granularity
- Through admin tools or frameworks

Role-based Access Control

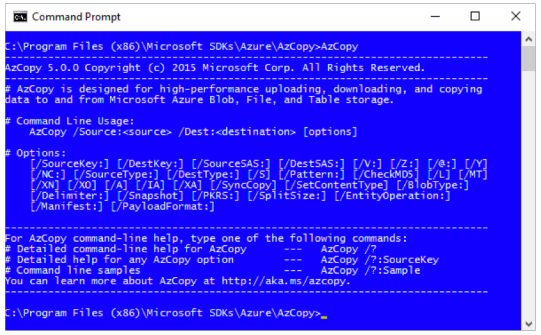
User Managed Service Identity

Ingesting Data

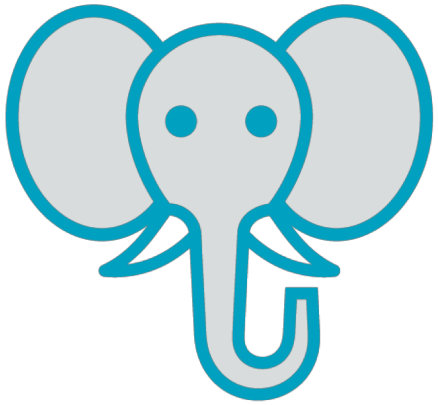


distcp

Azure Data Factory



AzCopy

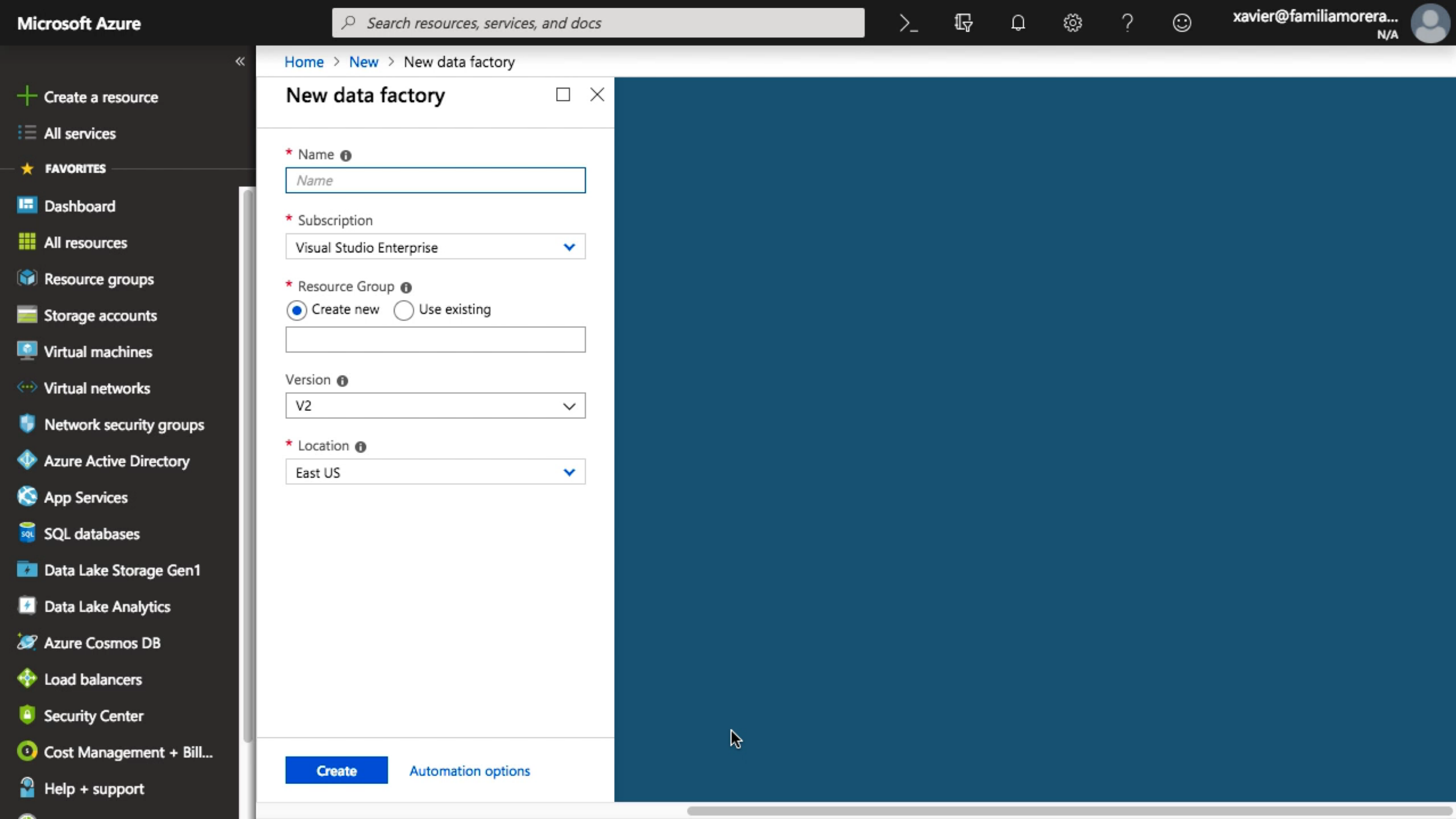


Hadoop Commands

REST API

Demo

Ingesting Data into ADLS Gen2 from AWS S3 using Azure Data Factory


Microsoft Azure
Search resources, services, and docs
xavier@familiamorera...
N/A
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
Storage accounts
Virtual machines
Virtual networks
Network security groups
Azure Active Directory
App Services
SQL databases
Data Lake Storage Gen1
Data Lake Analytics
Azure Cosmos DB
Load balancers
Security Center
Cost Management + Bill...
Help + support
Home
New
New data factory
New data factory
Name
Name
Subscription
Visual Studio Enterprise
Resource Group
Create new
Use existing
Version
V2
Location
East US
Create
Automation options

Microsoft Azure

Copy Data

1 Properties
One time copy

2 Source
Connection
Dataset

3 Destination
Connection
Dataset

4 Settings

5 Summary

6 Deployment

Choose the input file or folder

Select a source folder or file to be copied to the destination data store.

File or folder *

Browse

blackbook
max-vms
pscs
stackexchangex
usadoscrawler

Cancel

Choose

Previous

Next

Microsoft Azure

Copy Data

1 Properties
One time copy

2 Source
Amazon S3
Connection
Dataset

3 Destination
Connection
Dataset

4 Settings

5 Summary

6 Deployment

Destination data store

Specify the destination data store for the co

All Azure Database File Generi

All

Try c

Previous Next

New Linked Service

Search

All Azure File


Azure Blob Storage


Azure Data Lake Storage Gen1


Azure Data Lake Storage Gen2 (Preview)


Azure File Storage


File System

Cancel

Continue

Demo

Copying data from ADLS Gen1 to Gen2 with Azure Data Factory

Microsoft Azure

adfmovetogen2

Help us improve. Click here to tell us how we are doing.

Azure Data Factory

Let's get started

Create pipeline

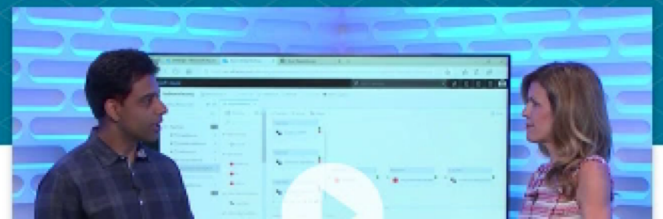
Copy Data

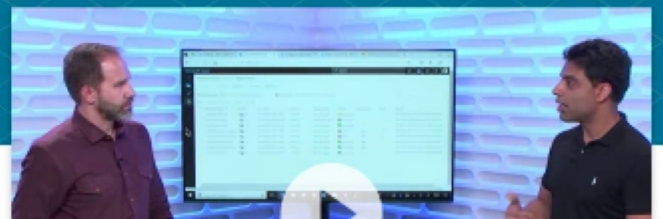
Configure SSIS Integration Runtime

Set up Code Repository

Videos

View All Videos





Microsoft Azure

Copy Data

1 Properties
One time copy

2 Source
Connection
Dataset

3 Destination
Connection
Dataset

4 Settings

5 Summary

6 Deployment

Source data store

Specify the source data store for the copy ta

All Azure Database File Generic

All

AmazonS3data

Previous Next

New Linked Service

Search

All Azure Database File Generic Protocol NoSQL Services and apps

Azure Blob Storage

Azure Cosmos DB

Azure Data Lake Storage Gen1

Azure Data Lake Storage Gen2 (Preview)

Azure Database for MySQL

Azure Database for PostgreSQL

Cancel Continue

Microsoft Azure

Copy Data

1 Properties
One time copy

2 Source
Azure Data Lake Storage Gen1
- Connection
- Dataset

3 Destination
- Connection
- Dataset

4 Settings

5 Summary

6 Deployment

Choose the output file or folder

Specify a folder that will contain output files or a specific output file in the destination data store.

Folder path

Browse

> datalakegen2hdifs >

- HdiNotebooks
- HdiSamples
- ams
- amshbase
- app-logs
- apps
- atshistory
- custom-scriptaction-logs

Cancel

Choose

Previous

Next

Demo

Using the Azure Data Lake Store Gen2 REST API

Azure Data Lake Store REST API

11/15/2016 • 2 minutes to read • Contributors

Use the Azure Data Lake Storage Gen2 REST APIs to interact with Azure Blob storage through a file system interface. This interface allows you to create and manage file systems, as well as to create and manage directories and files.

Authorization/Authentication

Azure Data Lake Storage Gen2 APIs support authorization with the Azure Storage Shared Key mechanism described in the Authorize with Shared Key article.

Operations

- Filesystem
  - Create
  - Delete
  - Get Properties
  - Set Properties
  - List
- Path

Authorization/Authentication

Azure Data Lake Storage Gen2 APIs support authorization with the Azure Storage Shared Key mechanism described in the Authorize with Shared Key article.

Operations

- Filesystem
  - Create
  - Delete
  - Get Properties
  - Set Properties
  - List
- Path
  - Create
  - Delete
  - Get Properties
  - Lease
  - List
  - Read
  - Update

See Also

- Azure Data Lake Storage Gen2 Documentation
- Azure Blob Storage REST API

Access Token

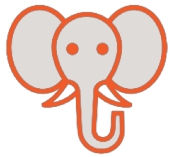
```
ACCESS_TOKEN=$(curl -X POST
-H "Content-Type: application/x-www-form-urlencoded"
-d "client_id=$CLIENT_ID1&client_secret=$CLIENT_SECRET1
&scope=https%3A%2F%2Fstorage.azure.com%2F.default
&grant_type=client_credentials"
"https://login.microsoftonline.com/$TENTANT_NAME/oauth2/v2.0/token"
| jq -r '.access_token')
```

Demo

Moving Data from Blobs to Azure Data Lake Store Gen2 Using distcp

Working with Data

file:///



hdfs:///

http:///

Or many more

Storage Drivers

WASB

Windows Azure Storage Blob driver

ABFS

Azure Blob Filesystem driver

Hadoop Azure Support: Azure Blob Storage

Introduction

The hadoop-azure module provides support for integration with Azure Blob Storage. The built jar file, named hadoop-azure.jar, also declares transitive dependencies on the additional artifacts it requires, notably the Azure Storage SDK for Java.

Features

- Read and write data stored in an Azure Blob Storage account.
- Present a hierarchical file system view by implementing the standard Hadoop `FileSystem` interface.
- Supports configuration of multiple Azure Blob Storage accounts.

URI Syntax

abfs[s]://<file_system>@<account_name>
.dfs.core.windows.net/<path>/<file_name>

Full URI Syntax

abfs://datalakegen2hdifs@datalakesaps
.dfs.core.windows.net/data/posts.csv

Default Filesystem URI Syntax

/data/posts.csv

Demo

Moving Data from Blobs to Azure Data Lake Store Gen2 Using distcp

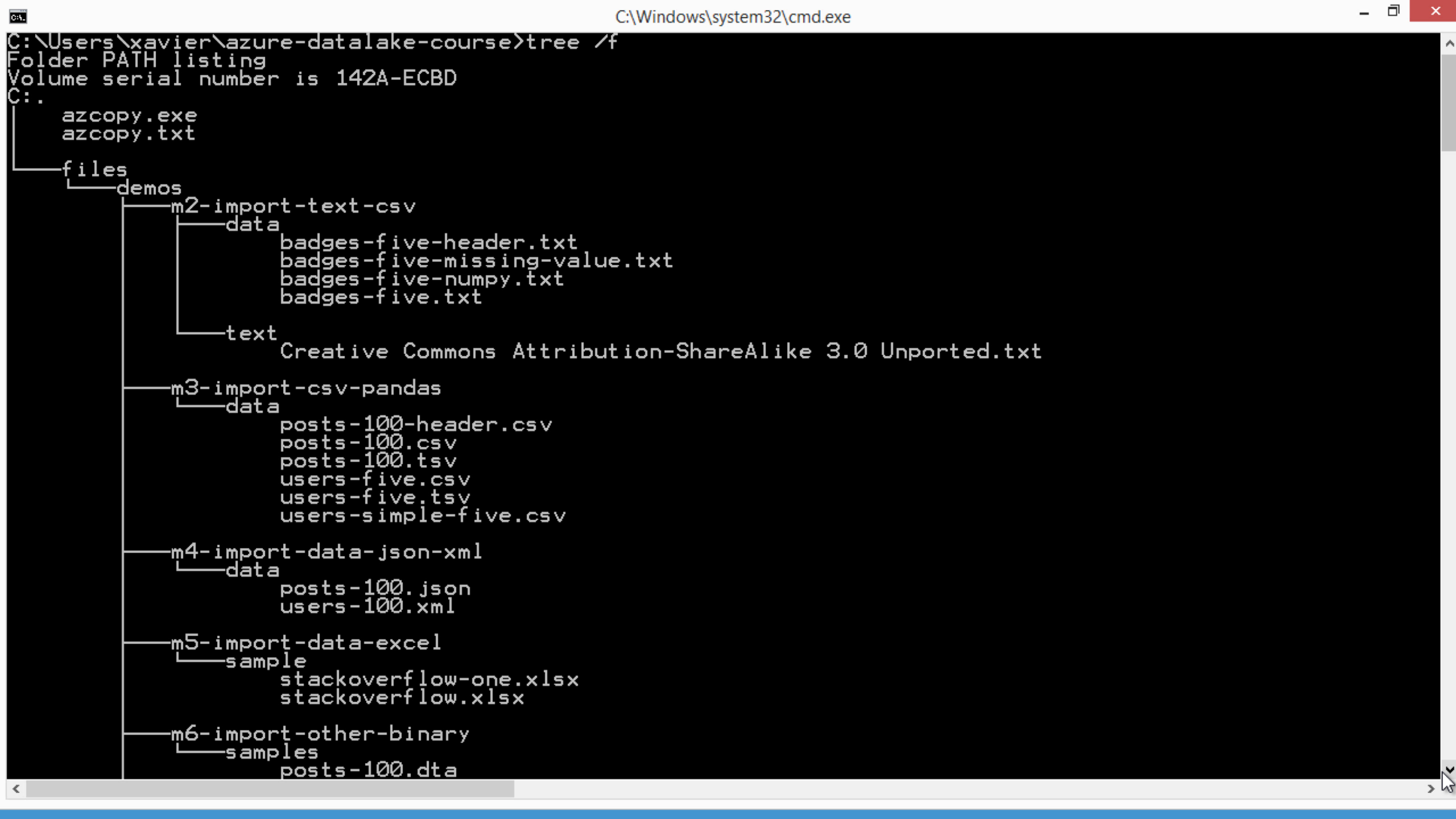
```
sshuser@hn0-datala:~$ hdfs dfs -ls wasb://datablob@datalakesaps.blob.core.windows.net/
Found 7 items
drwxrwxrwx   -          0 1970-01-01 00:00 wasb://datablob@datalakesaps.blob.core.windows
.net/data
-rwxrwxrwx   1       7566 2018-11-07 21:44 wasb://datablob@datalakesaps.blob.core.windows
.net/posts-100.dta
-rwxrwxrwx   1     413500 2018-11-07 21:44 wasb://datablob@datalakesaps.blob.core.windows
.net/posts-100.h5
-rwxrwxrwx   1       8608 2018-11-07 21:44 wasb://datablob@datalakesaps.blob.core.windows
.net/posts-100.mat
-rwxrwxrwx   1       7139 2018-11-07 21:44 wasb://datablob@datalakesaps.blob.core.windows
.net/posts-100.pkl
-rwxrwxrwx   1      42185 2018-11-07 21:44 wasb://datablob@datalakesaps.blob.core.windows
.net/posts-100.pkl.gz
-rwxrwxrwx   1     131072 2018-11-07 21:44 wasb://datablob@datalakesaps.blob.core.windows
.net/posts-100.sas7bdat
sshuser@hn0-datala:~$
```

```
sshuser@hn0-datala:~$ hdfs dfs -mkdir -p abfs://datalakegen2hdifs@datalakesaps.dfs.core.windows.net/datafromblob
sshuser@hn0-datala:~$ hdfs dfs -ls /
Found 20 items
drwxr-xr-x   - sshuser sshuser          0 2018-11-07 18:36 /HdiNotebooks
drwxr-xr-x   - sshuser sshuser          0 2018-11-07 18:44 /HdiSamples
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /ams
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /amshbase
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /app-logs
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /apps
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /atshistory
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:44 /custom-scriptaction-logs
drwxr-xr-x   - sshuser sshuser          0 2018-11-07 21:58 /datafromblob
drwxr-x---   - sshuser sshuser          0 2018-11-07 20:29 /datafromgen1
drwxr-x---   - sshuser sshuser          0 2018-11-07 19:55 /datafroms3
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:44 /example
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /hbase
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /hdp
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /hive
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /mapred
drwxr-x---   - sshuser sshuser          0 2018-11-07 21:39 /mapreducestaging
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /mr-history
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /tmp
drwxr-x---   - sshuser sshuser          0 2018-11-07 18:28 /user
sshuser@hn0-datala:~$
```

```
sshuser@hn0-datala:~$ hadoop distcp wasb://datablob@datalakesaps.blob.core.windows.net/po
sts* abfs://datalakegen2hdifs@datalakesaps.dfs.core.windows.net/datafromblob
```

Demo

Copying or Moving Data to Azure Data Lake Store Gen2 with AzCopy


C:\Windows\system32\cmd.exe
C:\Users\xavier\azure-datalake-course>tree /f
Folder PATH listing
Volume serial number is 142A-ECBD
C:.
    azcopy.exe
    azcopy.txt
    files
        demos
            m2-import-text-csv
                data
                    badges-five-header.txt
                    badges-five-missing-value.txt
                    badges-five-numpy.txt
                    badges-five.txt
                text
                    Creative Commons Attribution-ShareAlike 3.0 Unported.txt
            m3-import-csv-pandas
                data
                    posts-100-header.csv
                    posts-100.csv
                    posts-100.tsv
                    users-five.csv
                    users-five.tsv
                    users-simple-five.csv
            m4-import-data-json-xml
                data
                    posts-100.json
                    users-100.xml
            m5-import-data-excel
                sample
                    stackoverflow-one.xlsx
                    stackoverflow.xlsx
            m6-import-other-binary
                samples
                    posts-100.dta

C:\Windows\system32\cmd.exe

```
C:\Users\xavier\azure-datalake-course>set ACCOUNT_NAME=datalakesaps

C:\Users\xavier\azure-datalake-course>set ACCOUNT_KEY=5ZmcjecBsHSdjdKTxyBwte00tys/5+qJXqKkH/ow/M+1/6WtopbA

C:\Users\xavier\azure-datalake-course>azcopy cp "files/*" "https://datalakesaps.dfs.core.windows.net/data"

Job f85a7aaf-3679-c24e-56cb-3e86f4160a43 has started

 0 Active Connections, 26 Done, 0 Failed, 0 Pending, 26 Total,  2-sec Throughput (MB/s): 0.4895
Job f85a7aaf-3679-c24e-56cb-3e86f4160a43 summary
Elapsed Time (Minutes) 0.05
Total Number Of Transfers  26
Number of Transfers Completed  26
Number of Transfers Failed  0
Final Job Status  Completed

C:\Users\xavier\azure-datalake-course>_
```

```
sshuser@hn0-datala:~$ hdfs dfs -ls /datafromazcopy
Found 1 items
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos
sshuser@hn0-datala:~$ hdfs dfs -ls /datafromazcopy/demos
Found 6 items
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos/m2-impor
t-text-csv
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos/m3-impor
t-csv-pandas
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos/m4-impor
t-data-json-xml
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos/m5-impor
t-data-excel
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos/m6-impor
t-other-binary
drwxr-x---   - sshuser sshuser          0 2018-11-08 00:03 /datafromazcopy/demos/m7-impor
t-database
sshuser@hn0-datala:~$
```

Takeaway

- Many ways of ingesting data
  - distcp
  - Azure Data Factory
  - AzCopy
- REST API
- Don't take security for granted!
- Hadoop commands
  - ABFS driver