Data Engineering in the Cloud

Serverless SQL Pools

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Files Used in Azure
- External Table
- Lab: Loading Data into External Table

Files Used in Azure

There are three types of files can be accessible in Azure

- CSV files
- Parquet files
- JSON files

Parquet files

- Parquet is a columnar storage file format
- Parquet files typically use efficient compression algorithms (like Snappy, Gzip, or Brotli) on a per-column basis, leading to significant storage savings.
- Parquet files are used for
  - Data Warehousing for storing large volumes of data
  - Big Data Analytics for use with big data frameworks like Apache Spark
  - Useful in ETL (Extract, Transform, Load) workflows for intermediate storage of data

JSON files

- JavaScript Object Notation (JSON) file is Semistructured
  - It is widely used in APIs and web services
- Data is represented in key/value pairs
  - Key must be a string
  - Value can be String, number, Object, Array, Boolean, Null

```
{
  "name": "John",
  "age": 30
}
```

JSON files

- An example

```
{
  "name": "Alice",
  "age": 28,
  "isStudent": false,
  "courses": ["Math", "Science"],
  "address": {
    "street": "123 Main St",
    "city": "Wonderland"
  }
}
```

JSON files

- JSON log file examples:
  - https://jsonplaceholder.typicode.com/posts
  - https://jsonlint.com/?url=https://jsonlint.com/datasets/us-states-with-detail.json
- Dealing bad records:
  - https://jsonlint.com/

External Table

- In Azure Storage, external tables are used to write data to files and read data from files.
- External table can be used to access different files formats like CSV and Parquet.
- The CREATE EXTERNAL TABLE command creates an external table that allows Synapse SQL to access data from Azure Data Lake Storage.

Serverless SQL Pool Authentication

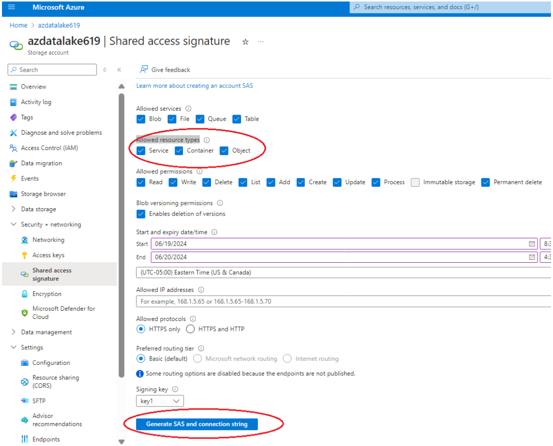
- Authentication and authorization are crucial aspects of managing access to serverless SQL pools, ensuring that users can only perform actions they are permitted to within the environment.
- SQL Authentication: Users can authenticate using SQL Server authentication. This involves creating a SQL user in the serverless SQL pool. Once authenticated, the SQL user is mapped to the appropriate permissions within the pool.
- Database Role Authorization Determines what actions users can perform within the serverless SQL pool. Roles can be assigned to users, granting them permissions to execute certain operations like Read, Write and Execute.

Lab: Loading Data into External Table

- Prerequisites (Lab in Lecture04):
    - Create a datalake (storage account with hierarchical namespace enabled)
    - Create two containers: `container1` and `rawdata`
    - Upload the data set `population.csv` to the container `rawdata`
    - Create a data base
    - Create an Azure Synapse Workspace
        - Add a dedicated SQL pool

Lab: Loading Data into External Table

- Go to the datalake storage account → security+networking → Shared access signature → select all Allowed resource types



Allowed services ⓘ
☑ Blob ☑ File ☑ Queue ☑ Table

Allowed resource types ⓘ
☑ Service ☑ Container ☑ Object

Allowed permissions ⓘ
☑ Read ☑ Write ☑ Delete ☑ List ☑ Add ☑ Create ☑ Update ☑ Process □ Immutable storage ☑ Permanent delete

Blob versioning permissions ⓘ
☑ Enables deletion of versions

Start and expiry date/time ⓘ
Start 06/19/2024 8:
End 06/20/2024 4:
(UTC-05:00) Eastern Time (US & Canada)

Allowed IP addresses ⓘ
For example, 168.1.5.65 or 168.1.5.65-168.1.5.70

Allowed protocols ⓘ
● HTTPS only ○ HTTPS and HTTP

Preferred routing tier ⓘ
● Basic (default) ○ Microsoft network routing ○ Internet routing
ⓘ Some routing options are disabled because the endpoints are not published.

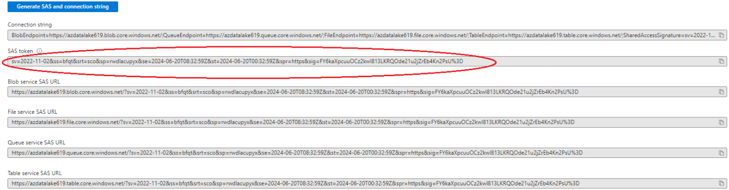
Signing key ⓘ
key1

Generate SAS and connection string

- Then click Generate SAS and connection string

Lab: Loading Data into External Table

- Copy the SAS token which is a very long string
  - My token was `sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T09:26:32Z`



Connection string

```
BlobEndpoint=https://azdatalake619.blob.core.windows.net/;QueueEndpoint=https://azdatalake619.queue.core.windows.net/;FileEndpoint=https://azdatalake619.file.core.windows.net/;TableEndpoint=https://azdatalake619.table.core.windows.net/;SharedAccessSignature=sv=2022-1...
```

SAS token

```
sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T08:32:59Z&st=2024-06-20T00:32:59Z&spr=https&sig=FY6kaXpcuuOCz2kwl813LKRQOde21u2jZrEb4Kn2PsU%3D
```

Blob service SAS URL

```
https://azdatalake619.blob.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T08:32:59Z&st=2024-06-20T00:32:59Z&spr=https&sig=FY6kaXpcuuOCz2kwl813LKRQOde21u2jZrEb4Kn2PsU%3D
```

File service SAS URL

```
https://azdatalake619.file.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T08:32:59Z&st=2024-06-20T00:32:59Z&spr=https&sig=FY6kaXpcuuOCz2kwl813LKRQOde21u2jZrEb4Kn2PsU%3D
```

Queue service SAS URL

```
https://azdatalake619.queue.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T08:32:59Z&st=2024-06-20T00:32:59Z&spr=https&sig=FY6kaXpcuuOCz2kwl813LKRQOde21u2jZrEb4Kn2PsU%3D
```

Table service SAS URL

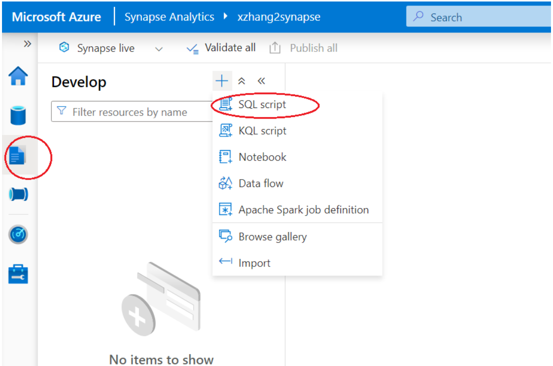
```
https://azdatalake619.table.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T08:32:59Z&st=2024-06-20T00:32:59Z&spr=https&sig=FY6kaXpcuuOCz2kwl813LKRQOde21u2jZrEb4Kn2PsU%3D
```

Lab: Loading Data into External Table

- Let's open the Synapse Studio
- In the left Navigation Pane, choose Develop and add an SQL script
- Create a user database:

```
CREATE DATABASE population_db;
```

- you can connect to the built-in serverless SQL pool or the Dedicated pool?



Lab: Loading Data into External Table

- In the SQL script file, create a master key due to the error message `Please create a master key in the database or open the master key in the session before performing this operation.`

```
CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'EsuMath23';
```

- Create the database scoped credential with the SAS token

```
CREATE DATABASE SCOPED CREDENTIAL SasToken
WITH IDENTITY='SHARED ACCESS SIGNATURE',
SECRET = 'sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-06-20T
```

Lab: Loading Data into External Table

- Create the external data source

```
CREATE EXTERNAL DATA SOURCE population_DS_DB
WITH
(
  LOCATION = 'https://azdatalake619.blob.core.windows.net/rawdata',
  CREDENTIAL = SasToken
);
```

Lab: Loading Data into External Table

- Create fileformat name

```
CREATE EXTERNAL FILE FORMAT TextFileFormat WITH (
    FORMAT_TYPE = DELIMITEDTEXT,
FORMAT_OPTIONS (
  FIELD_TERMINATOR = ',',
   FIRST_ROW = 2));
```

Lab: Loading Data into External Table

- Create the external table with the correct file format

```
CREATE EXTERNAL TABLE populationtable
 (
    [Country] nvarchar(4000),
    [Population] nvarchar(4000),
    [YearlyChange] nvarchar(4000),
    [NetChange] nvarchar(4000),
    [Density] nvarchar(4000),
    [LandArea] nvarchar(4000),
    [Migrants] nvarchar(4000),
    [Fert.Rate] nvarchar(4000),
    [Med.Age] nvarchar(4000),
    [UrbanPop%] nvarchar(4000),
    [Worldshare] nvarchar(4000)
    )
    WITH (
    LOCATION = 'population.csv',
    DATA_SOURCE = population_DS_DB,
    FILE_FORMAT = TextFileFormat
    );
```

Lab: Loading Data into External Table

```
select top 10 * from populationtable;
```

Lab: Loading Data into External Table

- Delete all what we have created.

License