

# Data Engineering in the Cloud

## Azure Data Lake Storage Gen2

Xuemao Zhang  
East Stroudsburg University

January 18, 2025

# Outline

- Introduction to Azure Data Lake Storage Gen2
- Lab: Create an Azure Database

# Introduction to Azure Data Lake

- A **data lake** is a single, centralized repository where you can store all your data, both structured and unstructured.
- Key Benefits:
  - ▶ Centralized Storage: Quickly and more easily store, access, and analyze a wide variety of data in a single location.
  - ▶ Flexible Data Storage: Store your data in its raw or native format, usually as files or binary large objects (blobs).
  - ▶ No Predefined Structure: No need to conform your data to fit an existing structure.
- Advantages:
  - ▶ Unified Data Management: Simplifies data storage and management across the organization.
  - ▶ Enhanced Data Analysis: Facilitates advanced analytics and insights from diverse data types.
  - ▶ Scalability: Accommodates growing data volumes efficiently.

# Introduction to Azure Data Lake

- Azure Data Lake Storage Gen2 is a set of capabilities dedicated to **big data analytics**, built on Azure Blob Storage.
  - ▶ The data that you ingest persist as blobs in the storage account.
- It combines the scalable and cost-effective storage capabilities of Blob Storage with advanced data management, security, and performance features required for large-scale analytics.
- *Azure Data Lake Storage Gen2* refers to the current implementation of Azure's Data Lake Storage solution. The previous implementation, Azure Data Lake Storage Gen1 was retired on February 29, 2024.

# Introduction to Azure Data Lake

- Data Lake Storage Gen2 includes the following capabilities.
  - ▶ Hadoop-compatible access
  - ▶ Hierarchical directory structure
  - ▶ Optimized cost and performance
  - ▶ Finer grain security model
  - ▶ Massive scalability
- The hierarchical namespace is a key feature that enables Azure Data Lake Storage Gen2 to provide high-performance data access at object storage scale and price.

# Introduction to Azure Data Lake

- You can use the `hierarchical namespace` feature to organize all the objects and files within your storage account into a hierarchy of directories and nested subdirectories.
  - ▶ Data is organized in much the same way that files are organized on your computer.
- To unlock Data Lake Storage Gen2 capabilities, we need to enable the hierarchical namespace setting when we create a storage account.
- When you use Azure Data Lake Storage to store data, the data is stored using a hierarchical file system that is compatible with Hadoop Distributed File System (HDFS).
  - ▶ We'll discuss HDFS later.
- For more information, read [Introduction to Azure Data Lake Storage Gen2](#)

- Create a storage account with **Enable hierarchical namespace** checked.

Home > **azdatalake611\_1718155059920** | Overview ⚙️ ...

Deployment

Search x < [Delete] [Cancel] [Redeploy] [Download] [Refresh]

**Overview**

Inputs

Outputs

Template

✓ **Your deployment is complete**

Deployment name: azdatalake611\_1718155059920      Start time: 6/11/2024, 9:17:40 PM  
Subscription: Azure for Students      Correlation ID: 0ea0838c-033b-4778-9c9b-c1cc7119254d [Info]

Resource group: xzhang2

Deployment details

Resource	Type	Status	Operation details
✓ azdatalake611/default	Microsoft.Storage/storageAccounts/fileservices	OK	<a href="#">Operation details</a>
✓ azdatalake611/default	Microsoft.Storage/storageAccounts/blobServices	OK	<a href="#">Operation details</a>
✓ azdatalake611	Microsoft.Storage/storageAccounts	OK	<a href="#">Operation details</a>

Next steps

[Go to resource](#)

Give feedback

[Tell us about your experience with deployment](#)

# Lab

- Home → SQL databases and click Create SQL database
- We need to create a server
  - ▶ Authentication method: **Use SQL authentication**
  - ▶ Set Username: sqladmin and pwd: EsuMath23
- SQL elastic pool: No
- Workload environment: Production or Development(more Cost-Effective)
- Compute + storage:
  - ▶ Service tier: General purpose



# Lab

- Networking tab:
  - ▶ Connectivity method: Public endpoint
  - ▶ Firewall rules: Yes and Yes
- Security tab:
  - ▶ Enable Microsoft Defender for SQL: Not now
- Additional settings: No changes
- Tags: No changes

## ● Review + create

Home > SQL databases > Create SQL Database >

### Create SQL Database ...

Microsoft

Basics Networking Security Additional settings Tags Review + create

#### Product details

SQL database  
by Microsoft  
[Terms of use](#) | [Privacy policy](#)

**Estimated cost per month**  
372.97 USD

#### Terms

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#). <sup>13</sup>

#### Basics

Subscription	Azure for Students
Resource group	xzhang2
Region	East US
Database name	SQLdatabase
Server	(new) servexzhang2
Authentication method	SQL authentication
Server admin login	sqladmin
Compute + storage	General Purpose: Standard-series (Gen5), 2 vCores, 32 GB storage, zone redundant disabled
Backup storage redundancy	Geo-redundant backup storage



#### Cost summary

<b>General Purpose (GP_Gen5_2)</b>	
Cost per vCore (in USD)	184.09
vCores selected	x 2
Cost per GB (in USD)	0.12
Max storage selected (in GB)	x 41.6
<b>ESTIMATED COST / MONTH</b>	<b>372.97 USD</b>

[Home](#) >

## Microsoft.SQLDatabase.newDatabaseNewServer\_00056c0899ea464fbd1ce | Overview

Deployment

[Delete](#) [Cancel](#) [Redeploy](#) [Download](#) [Refresh](#)

Overview

Inputs

Outputs

Template

## ✓ Your deployment is complete



Deployment name : Microsoft.SQLDatabase.newDatabaseNewServer\_00056c0899ea464fbd1ce

Subscription : Azure for Students

Resource group : xzhang2

Start time : 6/11/2024, 9:47:06 PM

Correlation ID : ddba57e9-d44c-4439-a8c8-ca363bce3f21

## Deployment details

Resource	Type	Status	Operation details
✓ serverxzhang2/Clientip-2024-6-11_21-47-3	Microsoft.Sql/servers/firewallrul	OK	<a href="#">Operation details</a>
✓ serverxzhang2/AllowAllWindowsAzureIps	Microsoft.Sql/servers/firewallrul	OK	<a href="#">Operation details</a>
✓ serverxzhang2/SQLdatabase	SQL database	Created	<a href="#">Operation details</a>
✓ serverxzhang2/Default	Microsoft.Sql/servers/connectio	OK	<a href="#">Operation details</a>
✓ serverxzhang2	SQL server	OK	<a href="#">Operation details</a>
✓ serverxzhang2	SQL server	Created	<a href="#">Operation details</a>

## Next steps

[Go to resource](#)

- Click Go to resource

Home > Microsoft.SQLDatabase.newDatabaseNewServer\_00056c0899ea464fbd1ce | Overview >

## SQLdatabase (serverxzhang2/SQLdatabase)

SQL database

Search

Copy Restore Export Set server firewall Delete Connect with... Feedback

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Query editor (preview)
- Settings
- Data management
- Integrations
- Power Platform
- Security
- Intelligent Performance
- Monitoring
- Automation
- Help


### Essentials

Resource group (move)	: <a href="#">xzhang2</a>	Server name	: <a href="#">serverxzhang2.database.windows.net</a>
Status	: Online	Elastic pool	: <a href="#">No elastic pool</a>
Location	: East US	Connection strings	: <a href="#">Show database connection strings</a>
Subscription (move)	: <a href="#">Azure for Students</a>	Pricing tier	: General Purpose: Gen5, 2 vCores
Subscription ID	: b5db3bdc-dae9-4e53-888f-8241b09dc796	Earliest restore point	: No restore point available
Tags (edit)	: <a href="#">Add tags</a>		


**Getting started** Monitoring Properties Features Notifications (0) Integrations Tutorials

### Start working with your database


Connect to your database and start working with data with a few simple steps. [Learn more](#)



Configure access



Connect to application



Start developing

- To access Query editor, you need the username and pwd you created before.


SQLdatabase (serverxzhang2/SQLdatabase) | Query editor (preview) ☆ ...

SQL database

Search ◯ ◀ ◻ Login + New Query ↗ Open query ↗ Feedback ⓘ Getting started

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Query editor (preview)**
- > Settings
- > Data management
- > Integrations
- > Power Platform
- > Security
- > Intelligent Performance
- > Monitoring
- > Automation

Query editor (preview) is a tool to run SQL queries against Azure SQL Database in the Azure portal. It is lightweight querying and object exploration in your database. For more information and troubleshoot



Welcome to SQL Database Query Editor

SQL server authentication

Login \*

Password \*

OK

Microsoft Entra authentication

[Continue as xzhang2@esu.edu](#)

OR

- Now go back to Home

The screenshot shows the Microsoft Azure portal interface. At the top is a blue header bar with the Microsoft Azure logo on the left and a search bar on the right containing the text "Search resources, services, and docs (G+I)". Below the header, the "Azure services" section displays a row of icons for various services: "Create a resource", "SQL databases", "Storage accounts", "Quickstart Center", "Azure AI services", "Kubernetes services", "Virtual machines", "App Services", "Azure Cosmos DB", and "More services". Below this is the "Resources" section, which has two tabs: "Recent" (selected) and "Favorite". It contains a table with three columns: "Name", "Type", and "Last Viewed". The table lists three resources: "SQLdatabase (servenzhang2/SQLdatabase)" of type "SQL database" viewed 7 minutes ago, "xzhang2" of type "Resource group" viewed 13 minutes ago, and "azdatalake611" of type "Storage account" viewed 36 minutes ago. A "See all" link is located at the bottom of the resources list.

Microsoft Azure

Search resources, services, and docs (G+I)

Azure services

- Create a resource
- SQL databases
- Storage accounts
- Quickstart Center
- Azure AI services
- Kubernetes services
- Virtual machines
- App Services
- Azure Cosmos DB
- More services

Resources

Recent Favorite




Name	Type	Last Viewed
SQLdatabase (servenzhang2/SQLdatabase)	SQL database	7 minutes ago
xzhang2	Resource group	13 minutes ago
azdatalake611	Storage account	36 minutes ago

See all



# Lab


- Then click on the storage account and add a container container1


Home > azdatalake611


 **azdatalake611** | Containers   ...


Storage account


  [+ Container](#) [🔒 Change access level](#) [↶ Restore containers ↷](#)


 Overview


 Activity log


 Tags

 Diagnose and solve problems



 Access Control (IAM)


 Data migration

 Events

 Storage browser

▼ Data storage

 **Containers** 

 File shares

Name
<input type="checkbox"/> \$logs

Home > azdatalake611



## azdatalake611 | Containers

Storage account



Container



Change access level



Restore containers



Ref

Overview

Activity log

Tags

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

▼ Data storage



Containers



File shares



Queues

Name



\$logs



container1



- Open container1, and upload the data set `population.csv` to the container
  - We did not upload any data to a storage account in the last lab

Home > azdatalake611 | Containers >

## container1

Container

Search

[Upload](#)
[+ Add Directory](#)
[Refresh](#)
[Rename](#)
[Delete](#)
[Change tier](#)
[Acquire lease](#)
[Break lease](#)
[Give feedback](#)

### Overview

[Diagnose and solve problems](#)

[Access Control \(IAM\)](#)

[Settings](#)

**Authentication method:** Access key ([Switch to Microsoft Entra user account](#))

**Location:** container1

Search blobs by prefix (case-sensitive)

☐ Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> <a href="#">population.csv</a>	6/11/2024, 10:14:19 PM	Hot (inferred)		Block blob	15.52 KiB

# License



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).