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

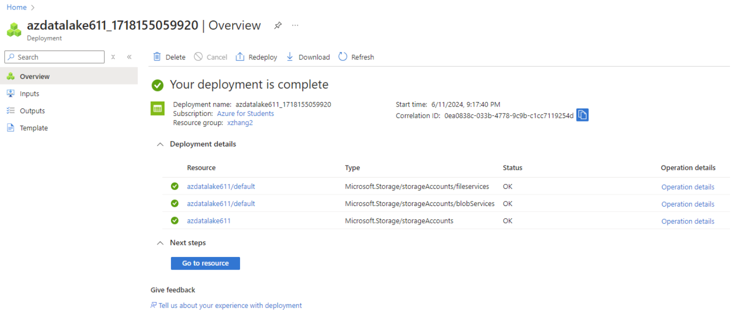
- Data Lake Storage Gen2 includes the following capabilities.
  - Hadoop-compatible access
  - Hierarchical directory structure
  - Optimized cost and performance
  - Finer grain security model
  - Massive scalability
- The hierarchical namespace is a key feature that enables Azure Data Lake Storage 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 creat 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

Lab

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



azdatalake611_1718155059920 | Overview

Your deployment is complete

Deployment name: azdatalake611_1718155059920
Subscription: Azure for Students
Resource group: xzhang2

Start time: 6/11/2024, 9:17:40 PM
Correlation ID: 0ea0838c-033b-4778-9c9b-c1cc7119254d

Deployment details

Resource	Type	Status	Operation details
✓ azdatalake611/default	Microsoft.Storage/storageAccounts/fileservices	OK	Operation details
✓ azdatalake611/default	Microsoft.Storage/storageAccounts/blobServices	OK	Operation details
✓ azdatalake611	Microsoft.Storage/storageAccounts	OK	Operation details

Next steps

Give feedback

Tell us about your experience with deployment

Lab

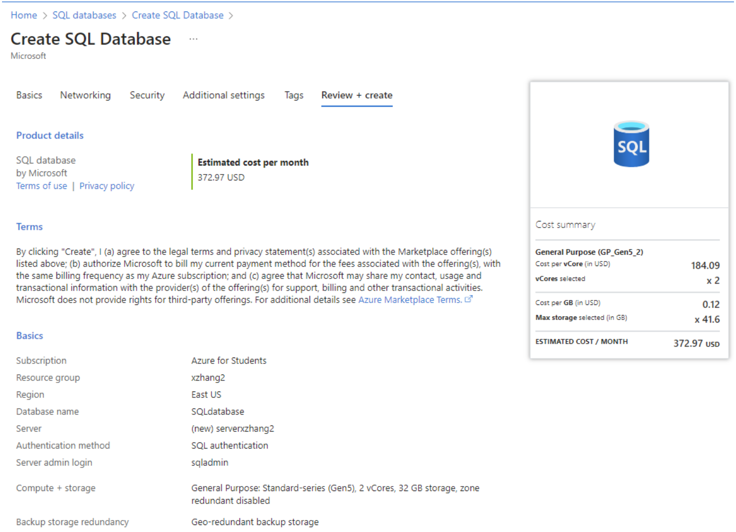
- Home → SQL databases and click `Create SQL database`
- We need to create a server
  - Authentication method: **Use SQL authentication**
  - Set Username: sqladmin and pwd: EsuMath23
- SQL elastic pool: No
- Workload environment: Production or Development(more Cost-Effective)
- `Compute + storage`: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

Lab

- Review + create



Create SQL Database ···

Microsoft

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.

Basics

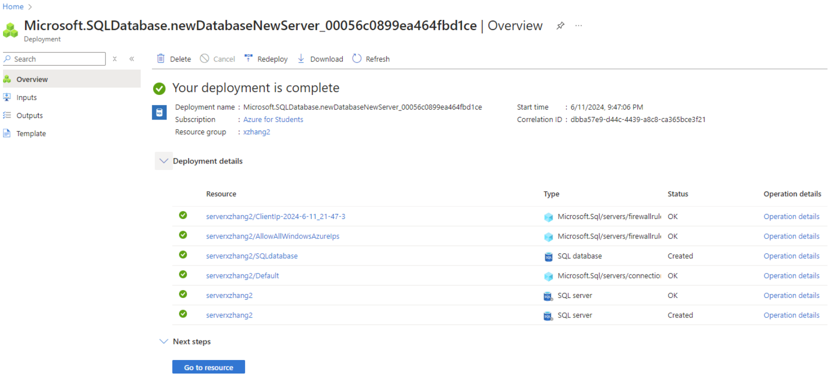
Subscription	Azure for Students
Resource group	xzhang2
Region	East US
Database name	SQLdatabase
Server	(new) serverxzhang2
Authentication method	SQL authentication
Server admin login	sqladmin
Compute + storage	General Purpose: Standard-series (Gen5), 2 vCores, 32 GB storage, zone redundant disabled
Backup storage redundancy	Geo-redundant backup storage

Cost summary

General Purpose (GP_Gen5_2)

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
ESTIMATED COST / MONTH	372.97 USD

Lab



Microsoft.SQLDatabase.newDatabaseNewServer_00056c0899ea464fbd1ce | Overview

Deployment

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 : dbba57e9-d44c-4439-a8c8-ca365bce3f21

Deployment details

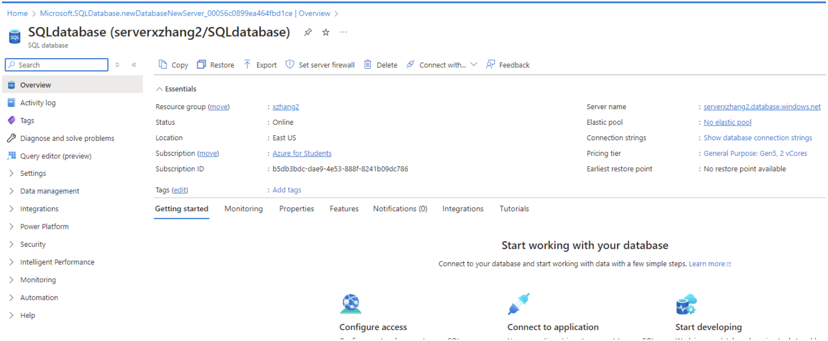
	Resource	Type	Status	Operation details
✓	serverxzhang2/ClientIp-2024-6-11_21-47-3	Microsoft.Sql/servers/firewallrul	OK	Operation details
✓	serverxzhang2/AllowAllWindowsAzureIps	Microsoft.Sql/servers/firewallrul	OK	Operation details
✓	serverxzhang2/SQLdatabase	SQL database	Created	Operation details
✓	serverxzhang2/Default	Microsoft.Sql/servers/connectio	OK	Operation details
✓	serverxzhang2	SQL server	OK	Operation details
✓	serverxzhang2	SQL server	Created	Operation details

Next steps

Go to resource

Lab

- Click Go to resource



Home > Microsoft.SQLDatabase.newDatabaseNewServer_00056c0899ea464fbd1ce | Overview >

SQLdatabase (serverxzhang2/SQLdatabase)

SQL database

Resource group (move) : xzhang2

Status : Online

Location : East US

Subscription (move) : Azure for Students

Subscription ID : b5db3bdc-dae9-4e53-888f-8241b09dc786

Tags (edit) : Add tags

Server name : serverxzhang2.database.windows.net

Elastic pool : No elastic pool

Connection strings : Show database connection strings

Pricing tier : General Purpose: Gen5, 2 vCores

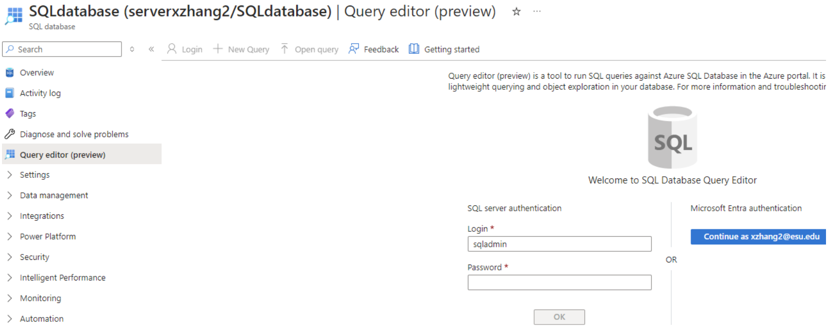
Earliest restore point : No restore point available

Start working with your database

Connect to your database and start working with data with a few simple steps. Learn more

Lab

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



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

SQL database

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 troubleshootir

Welcome to SQL Database Query Editor

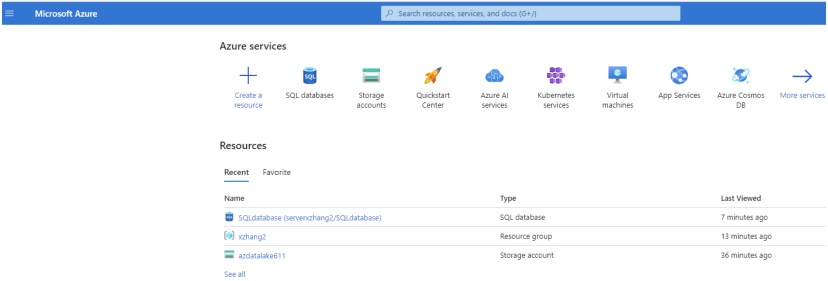
SQL server authentication

Login *: sqladmin

Password *:

Lab

- Now go back to Home



Azure services

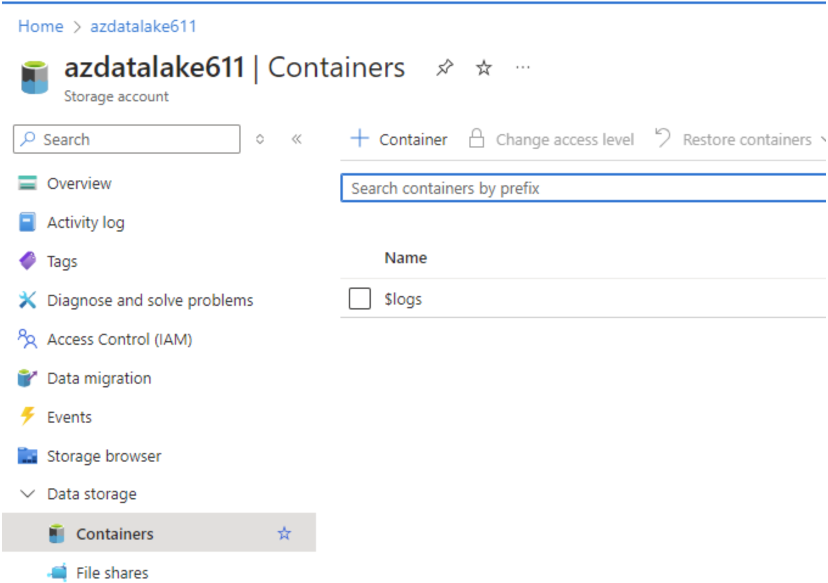
Resources

Name	Type	Last Viewed
SQLdatabase (serverxzhang2/SQLdatabase)	SQL database	7 minutes ago
xzhang2	Resource group	13 minutes ago
azdatalake611	Storage account	36 minutes ago

See all

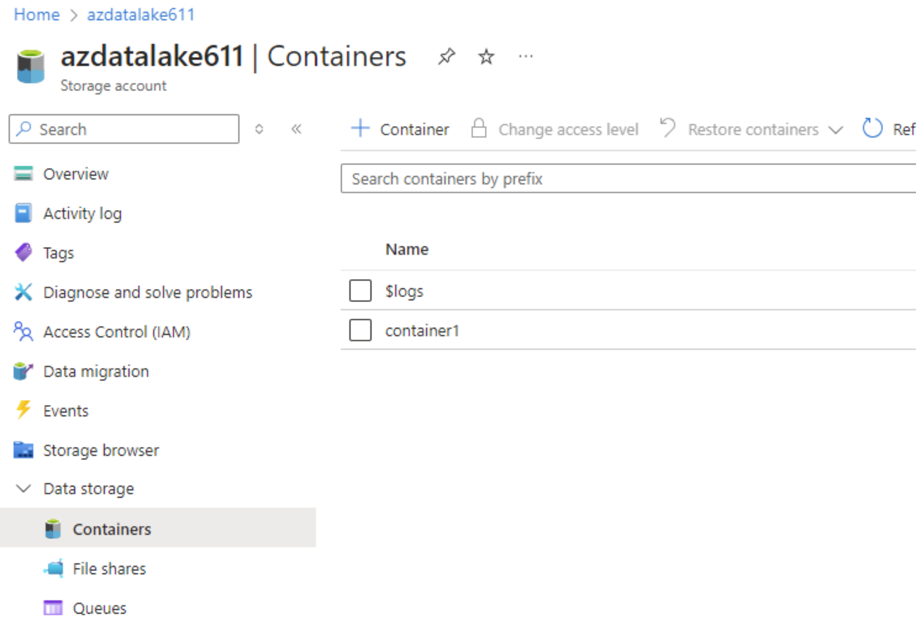
Lab

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



	Name
□	$logs

Lab



Home > azdatalake611

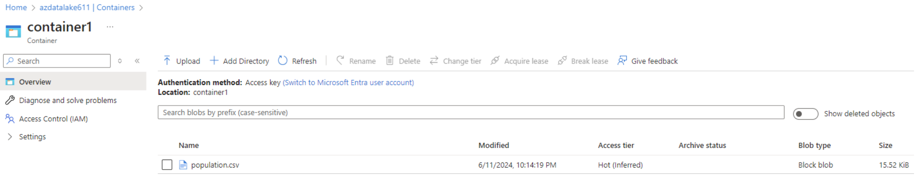
azdatalake611 | Containers

Storage account

	Name
□	$logs
□	container1

Lab

- 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



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: container1

	Name	Modified	Access tier	Archive status	Blob type	Size
□	population.csv	6/11/2024, 10:14:19 PM	Hot (Inferred)		Block blob	15.52 KiB

License