Practica 9. Solucion ETL para batch processing con Hadoop en ambiente Azure con los servicios **HDInsight + Azure Data Lake Storage Gen 2 + HIVE+ SQL Database + Sqoop**

Nombre de la empresa

Descripción generada automáticamente con confianza media

Diagrama

Descripción generada automáticamente

Crear un grupo de recursos:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza media

Click en Create.

Crear Source: Azure Data Lake Storage Gen2

Empezamos por la Storage Account

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

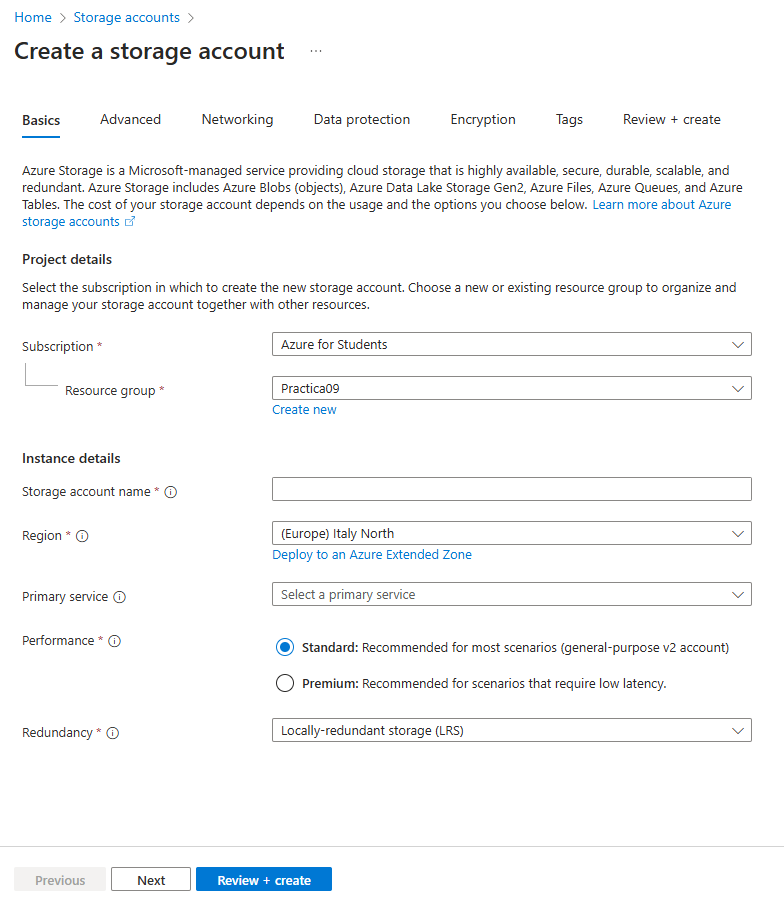
Descripción generada automáticamente

Pagina principal

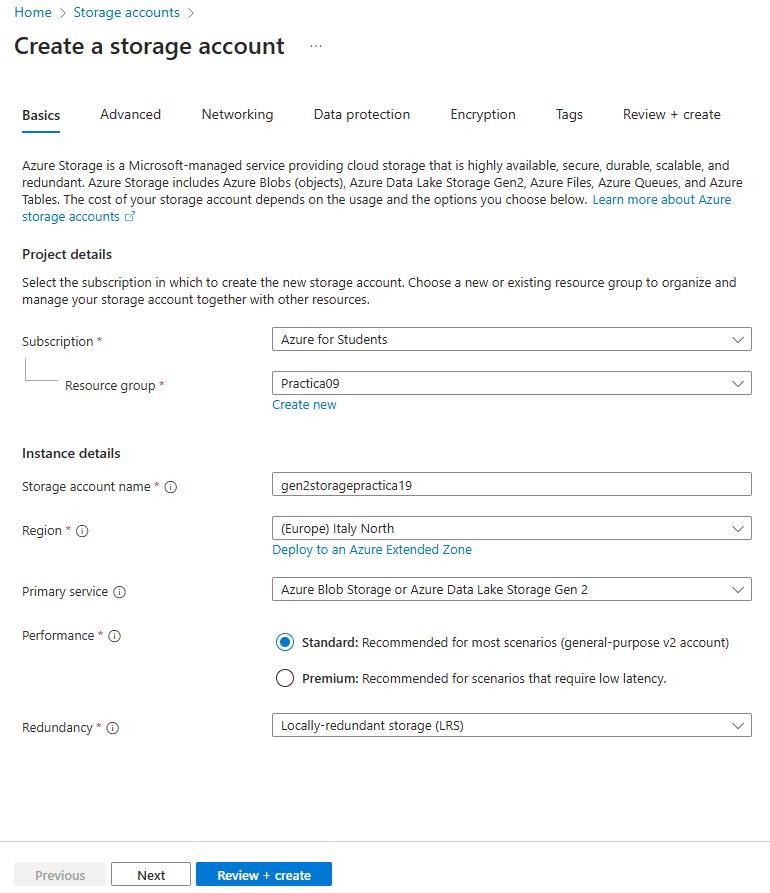
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Create.

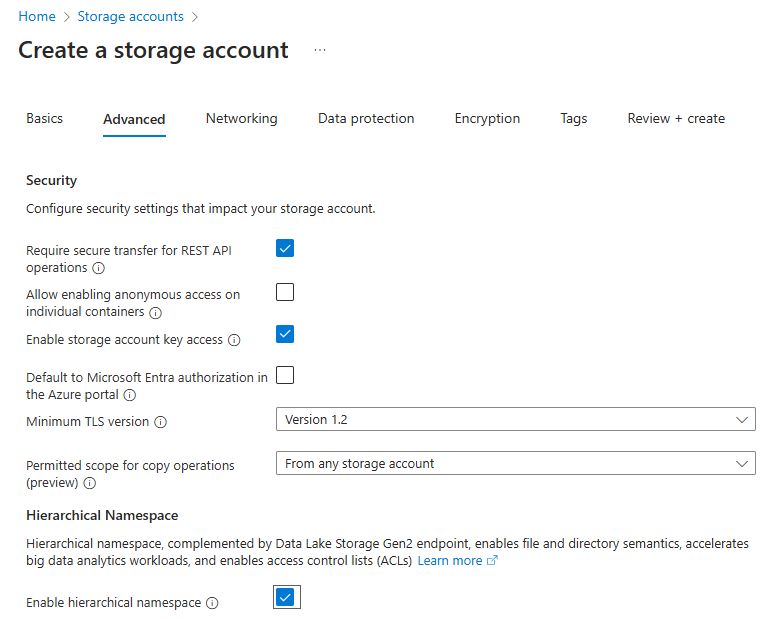


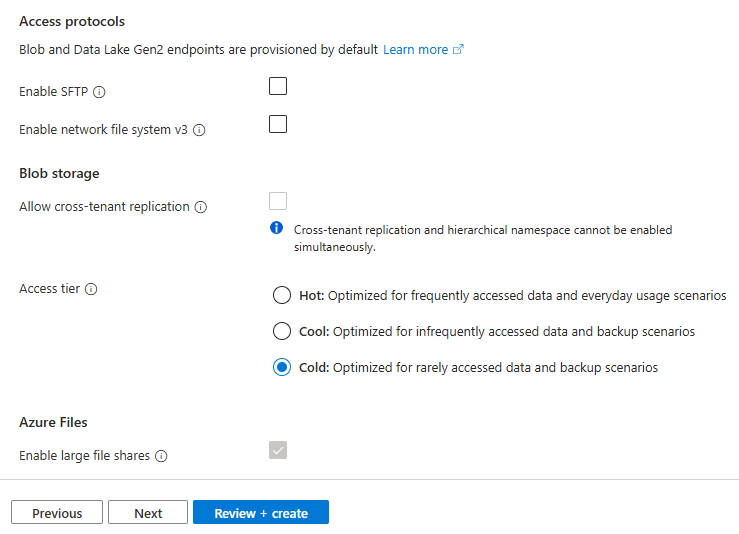
En Basics, debe estar lleno con lo siguiente:



Click en Next

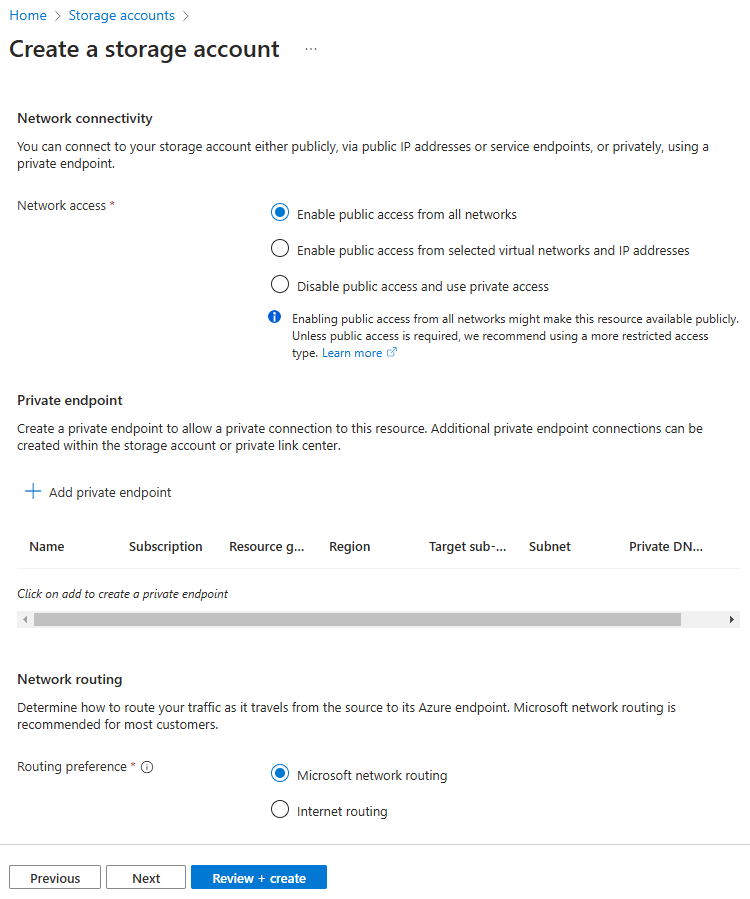
En Advanced lo llenamos con lo siguiente:





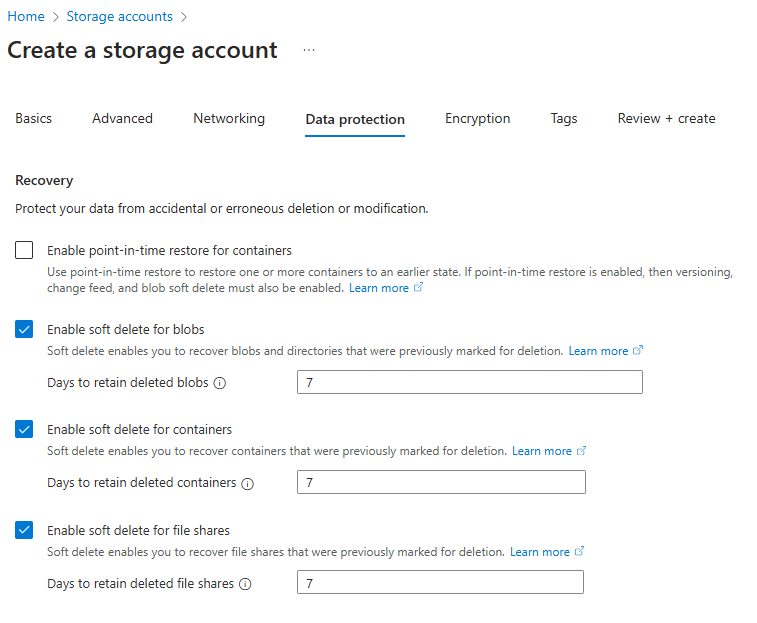
Click en Next

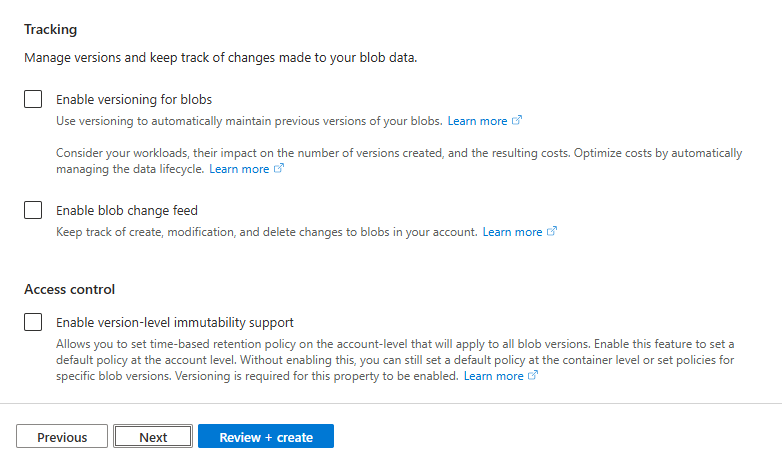
En Networking, se llena así:



Click en Next

Data Protection





Click en Next

Encryption

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Click en NExt

Tags, lo podemos dejar en blanco

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

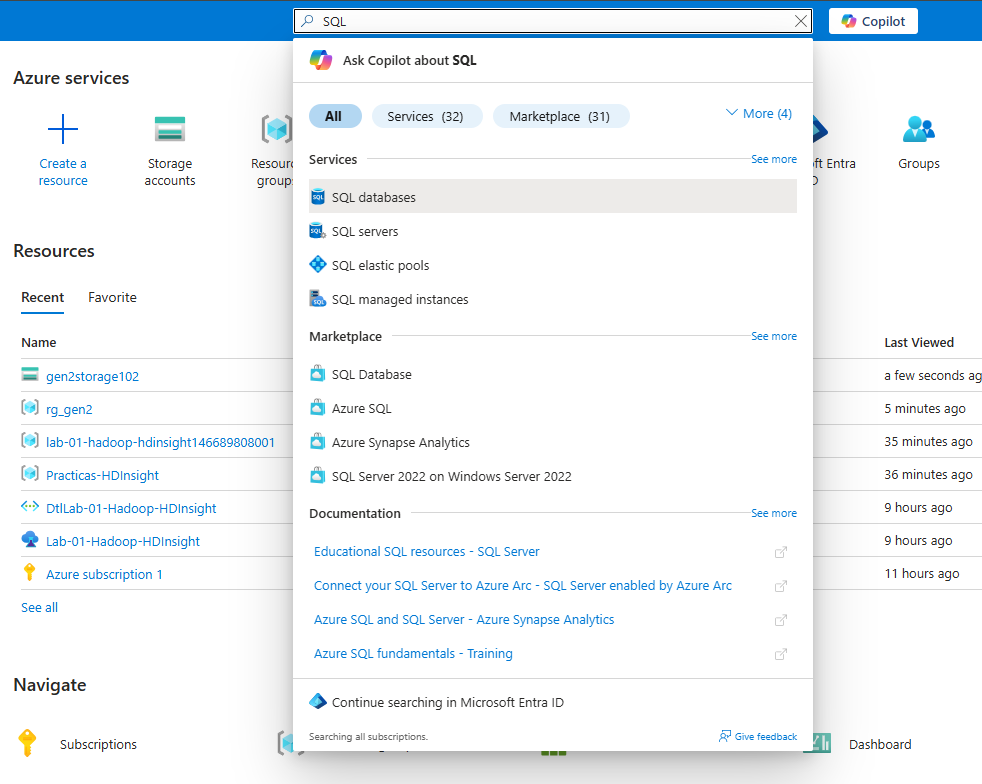
Click en Next

Review + Create: click en Create

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Vamos a crear el servicio de SQL Server (Data Destination):

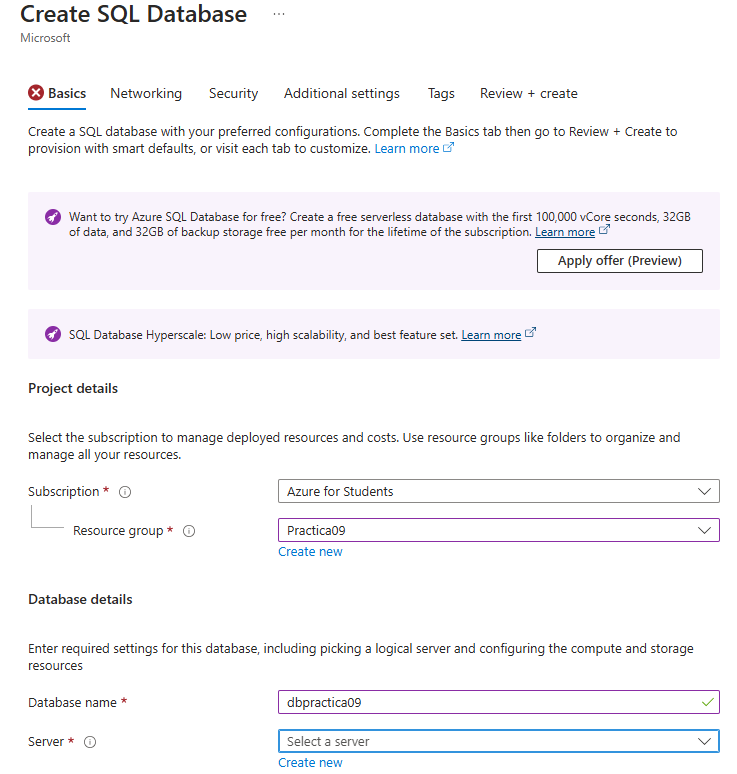


Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

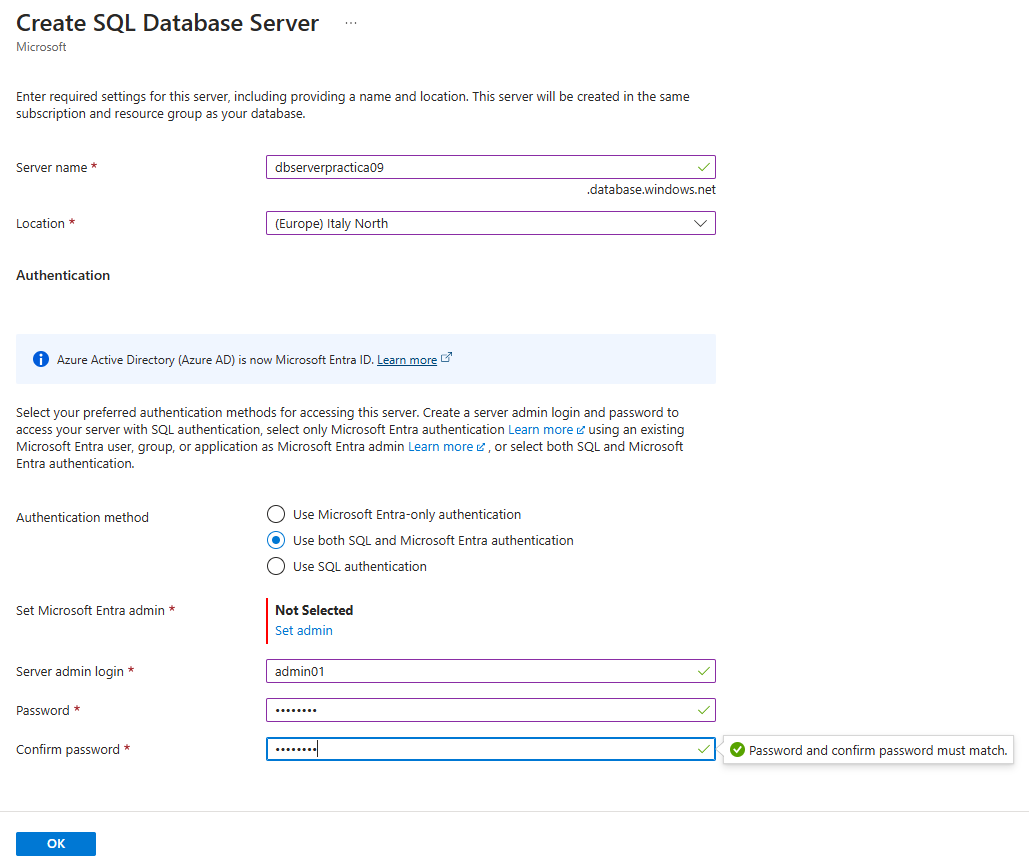
Descripción generada automáticamente

Click en Create.

Basic:



En Server, dar click en Create New:



Click en Set admin, usa tu email de azure for students:

Interfaz de usuario gráfica, Texto, Aplicación

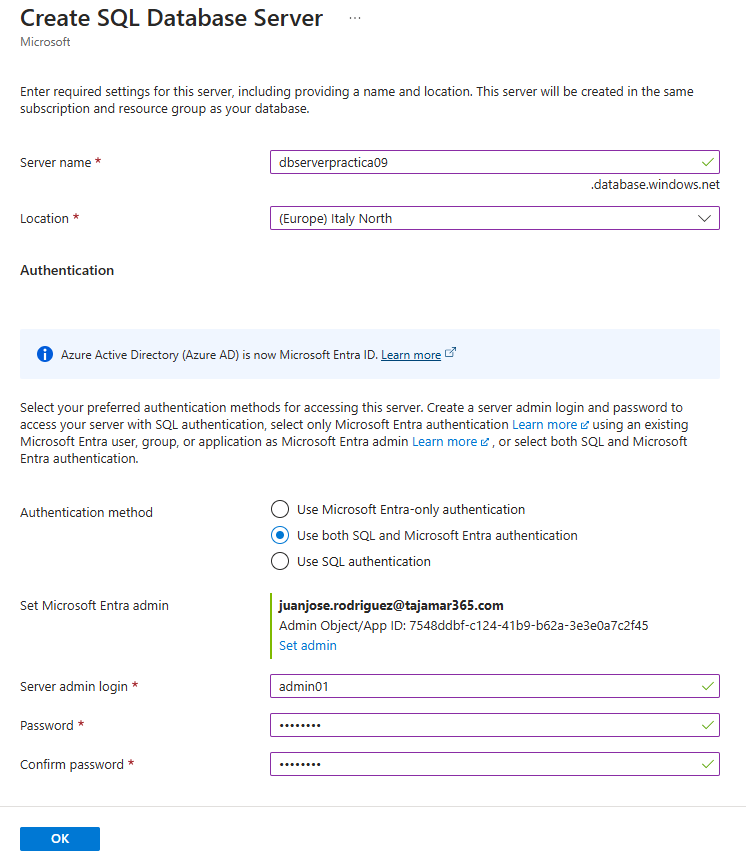
Descripción generada automáticamente

Imagen que contiene Rectángulo

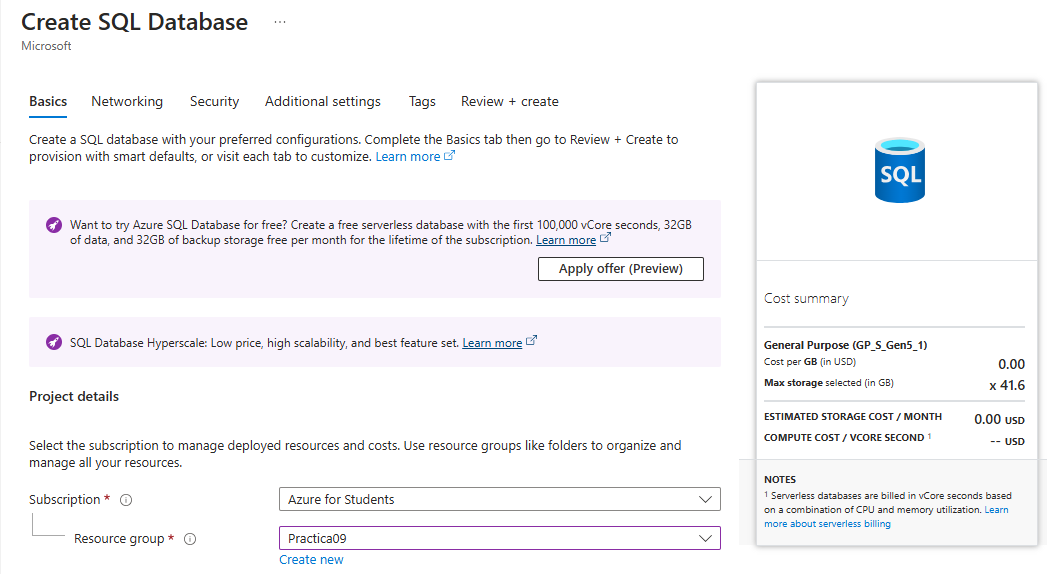
Descripción generada automáticamente

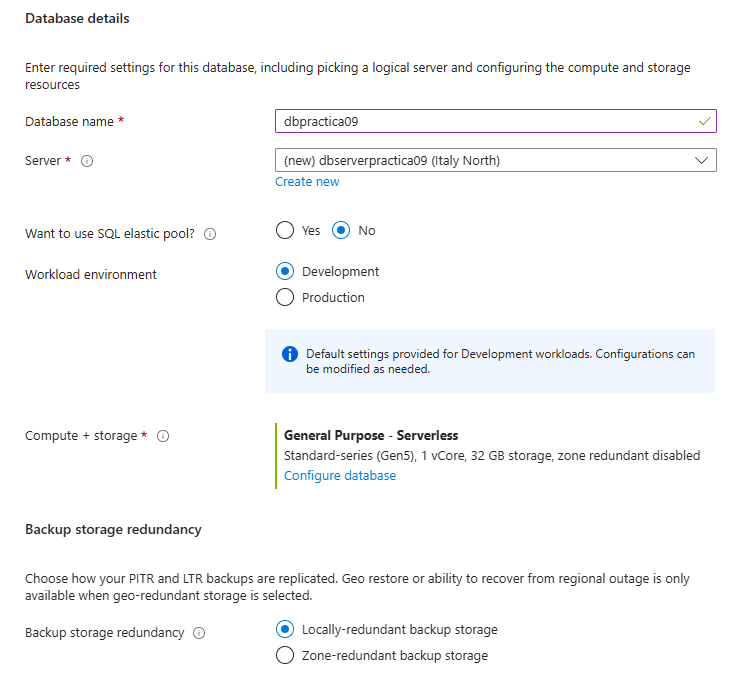
Click en Select.

Debería verse algo similar a esto:

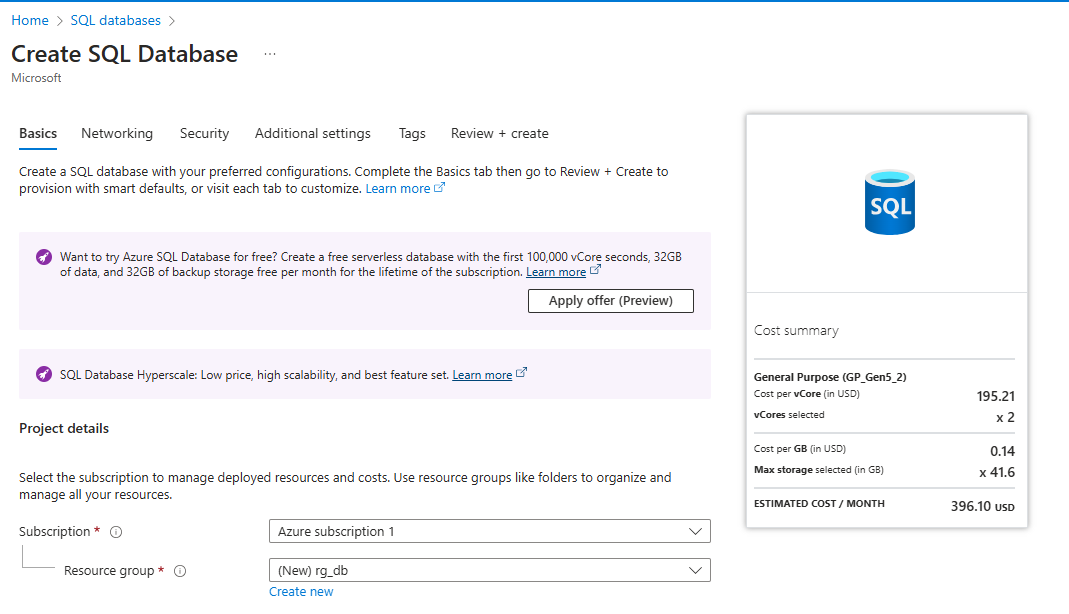


Click en Ok.





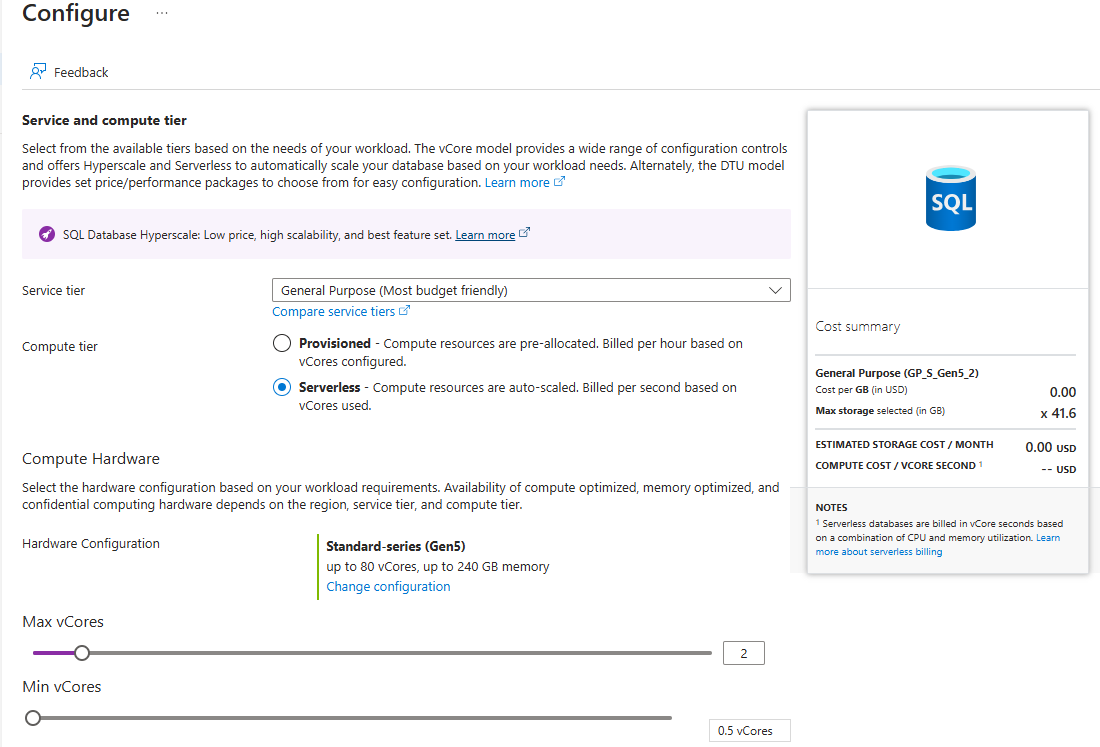
Click en Configure database:



Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Configure database:

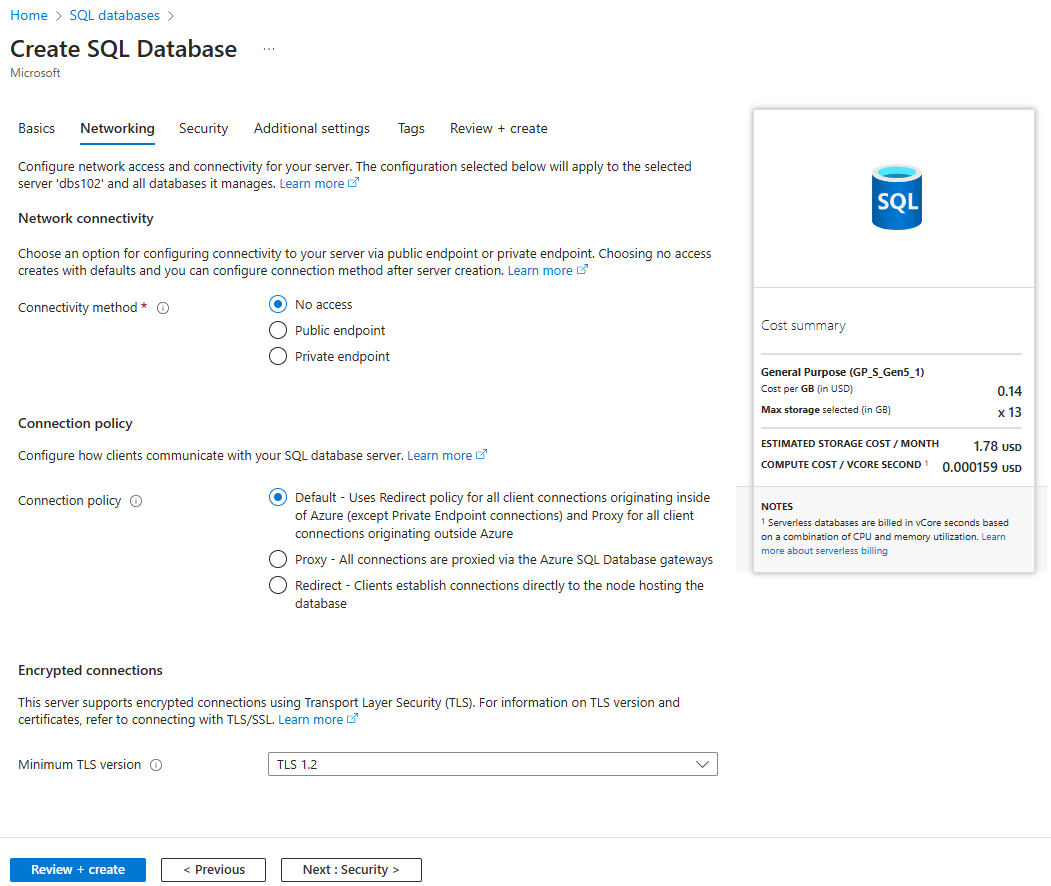


Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

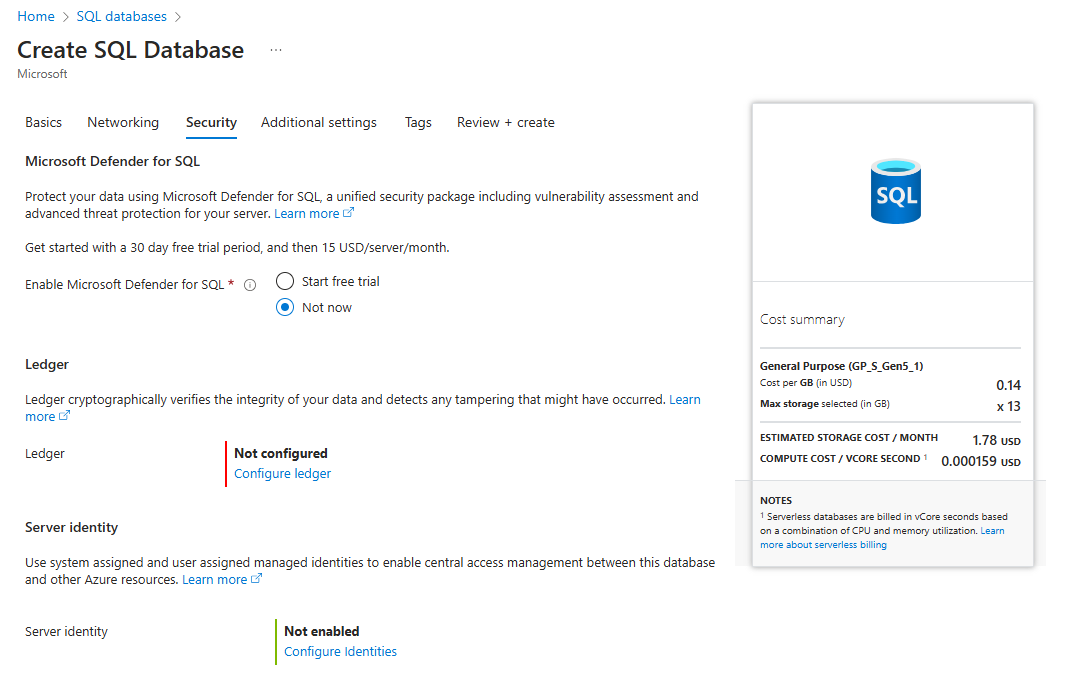
Descripción generada automáticamente

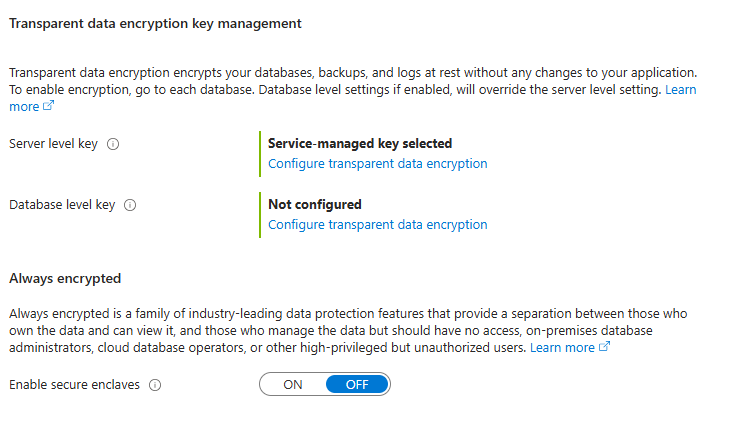
Click en Apply

Next: Networking

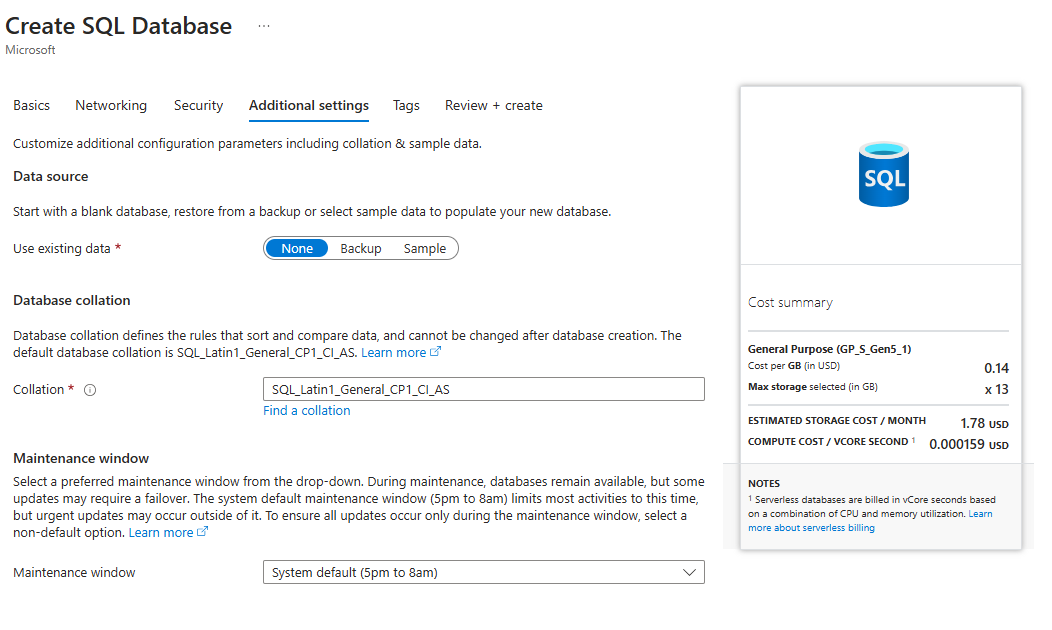


Next: Security





Click en: Next: Additional Settings

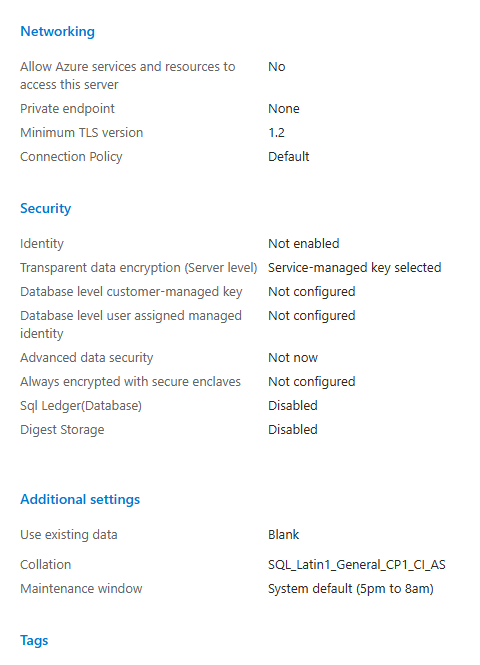


Next: Tags, dejar en blanco

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Next: Review + Create Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente  
  


Interfaz de usuario gráfica, Aplicación

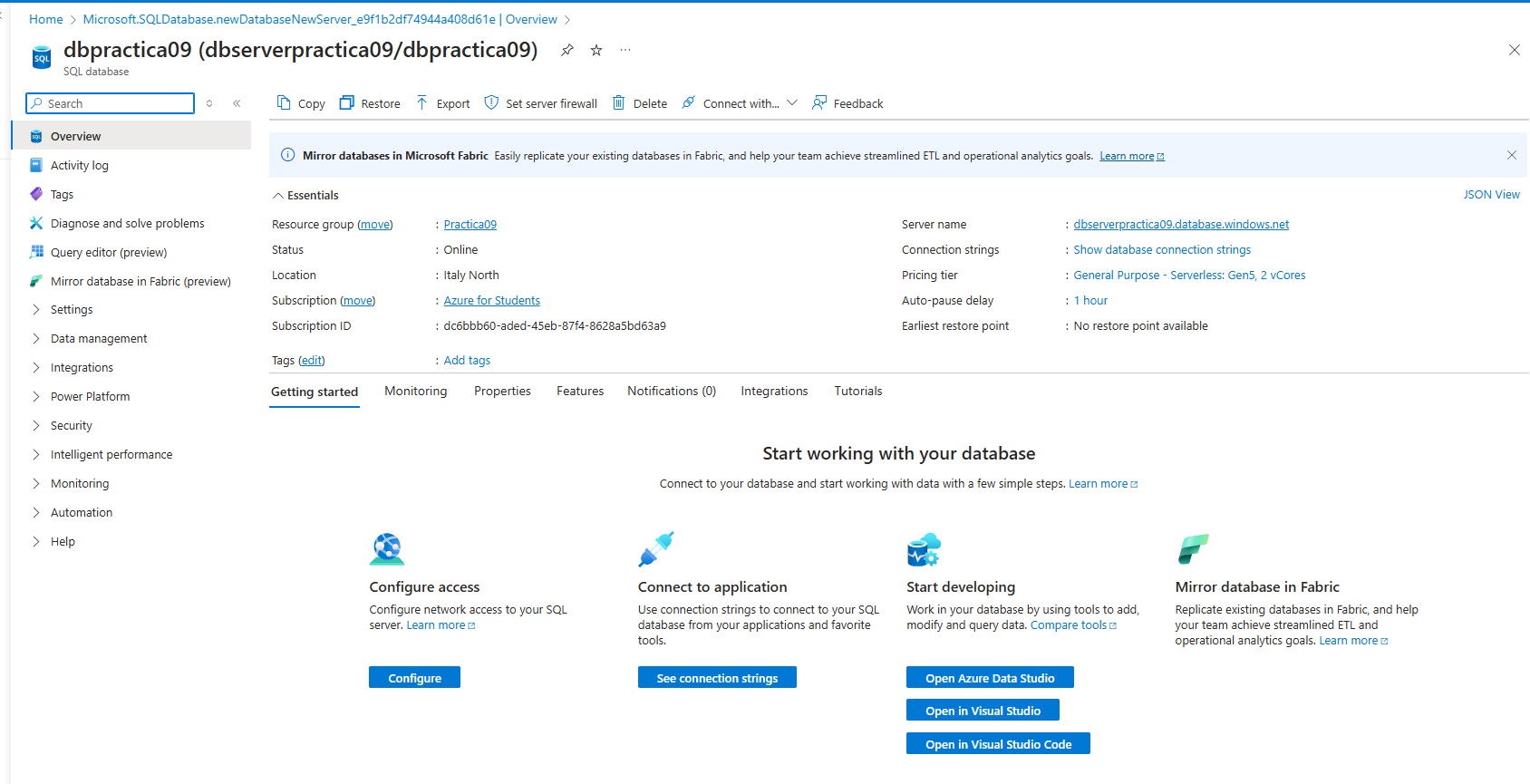
Descripción generada automáticamente con confianza media

Click en Create

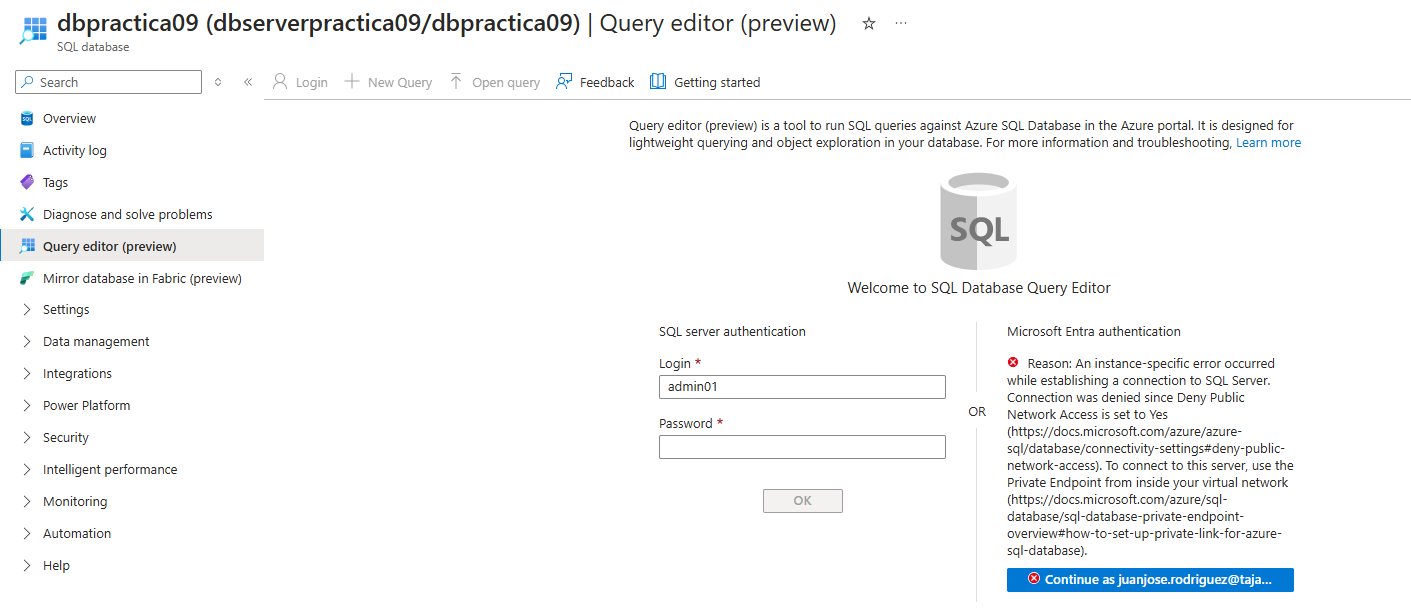
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

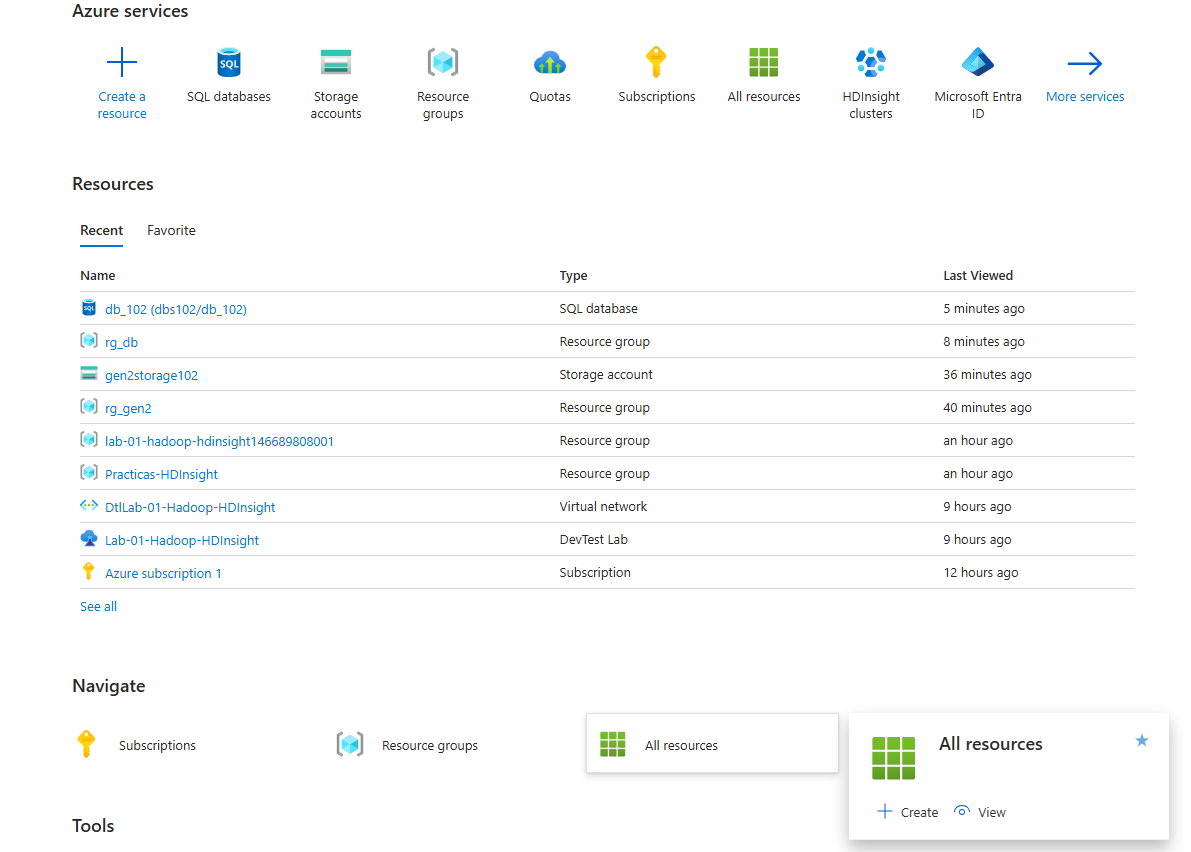
Click en Go to resource:

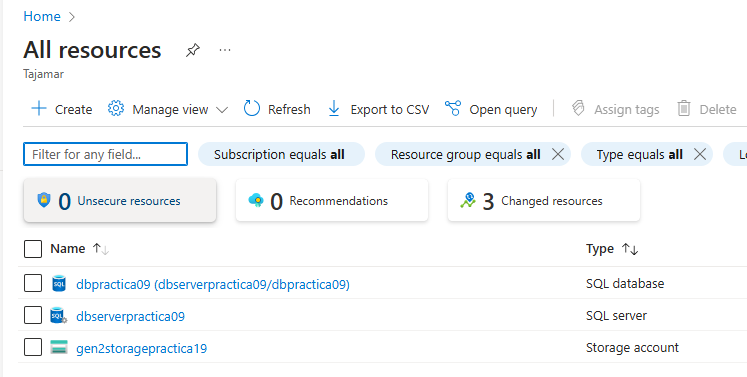


Click en Query editor (preview):



Para resolver el error nos vamos a Home y luego click en All resources:





Click en dbserverpractica09 (SQL Server).

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Networking:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Activar Selected networks

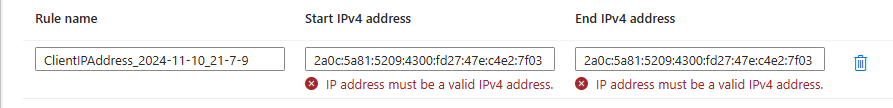
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

En Firewall rules, dar click en Add your client IPv4 address…

Imagen que contiene Logotipo

Descripción generada automáticamente



Si falla lo anterior , dar click en:



Hay que Introducir manualmente la IP, escribir en el navegador whats myip y anotar la dirección ip y rellenar. Luego Ok.

Interfaz de usuario gráfica, Aplicación, Word

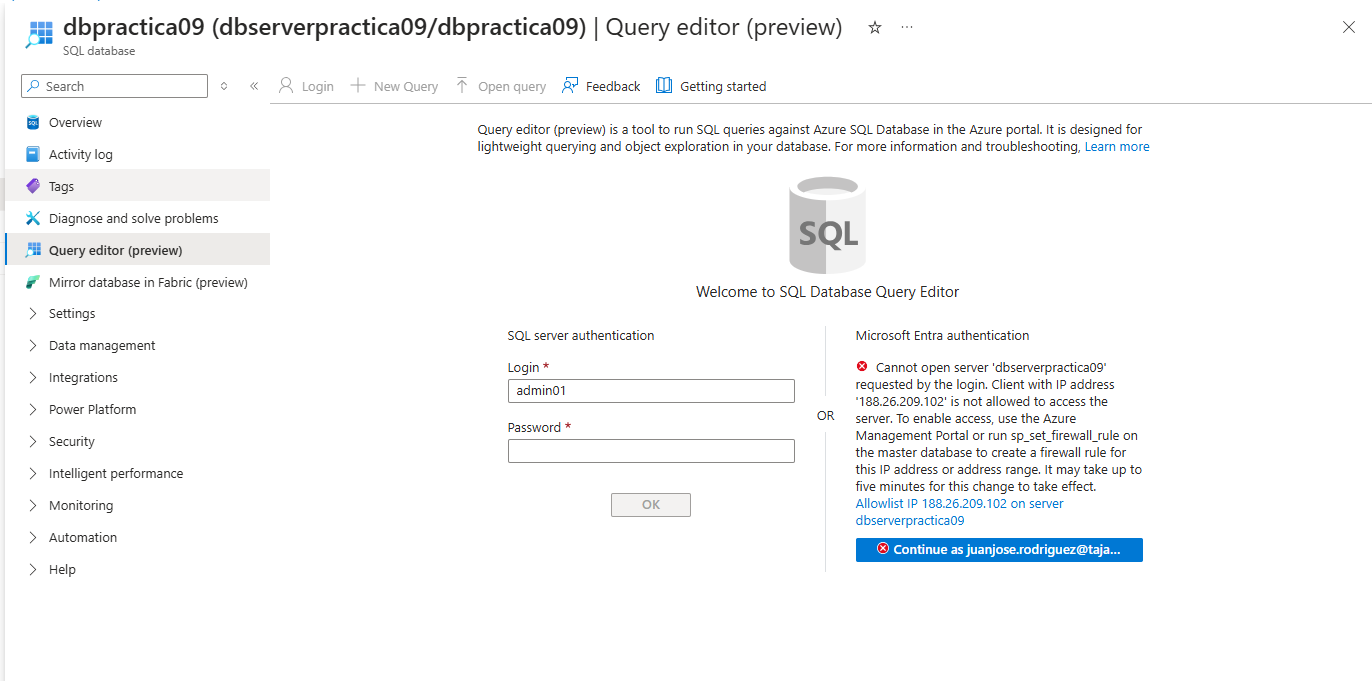
Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Save.

Volver a sql database (dbpractica09) y click en Query editor (preview):



Click en allowlist IP … on server dbserverpractica09

Continue as [tuemail@tajamar365.com](mailto:tuemail@tajamar365.com)

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Añadir Managed Identity a Gen2 y a DataBase. ¿ Cuál es el propósito de usar Managed Identity?

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Create

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Basics

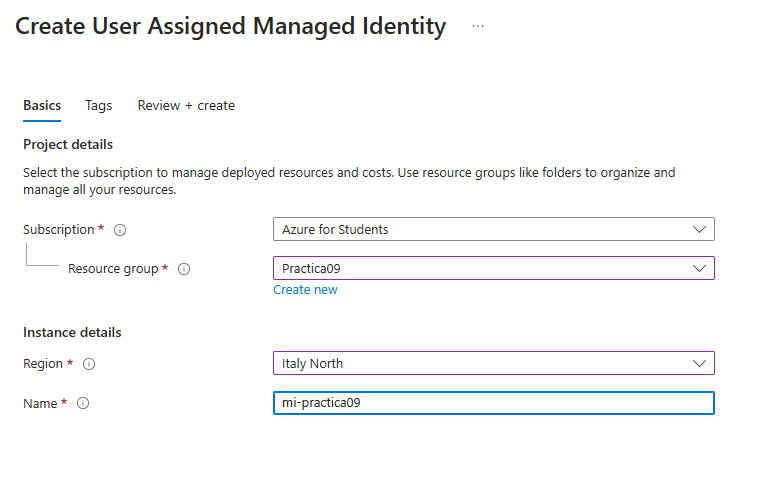


Imagen que contiene Diagrama

Descripción generada automáticamente

Click en Next

Tags, dejar en blanco. Click en Next.

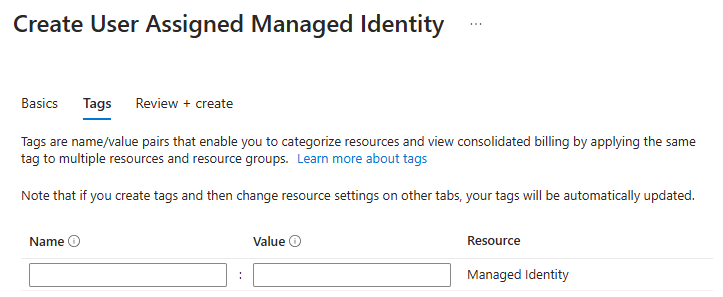


Imagen que contiene Texto

Descripción generada automáticamente

Click en Next

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente

Click en create.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Una vez creado nos vamos a Home y buscamos gen2storagepractica09

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Access Control (IAM)

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Add, luego add role assignment

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Creamos uno que se llama Storage Blob Data Owner.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Click en Next

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Seleccionar Managed Identity, luego Select members:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media



Click en Select

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen que contiene Rectángulo

Descripción generada automáticamente

Click en Next.

Conditions: Dejar como aparece por defecto

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen que contiene Forma

Descripción generada automáticamente

Click en Next

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



Click en Review + assign

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Nos vamos a Home y buscamos dbserverpractica09 (SQL Server) , luego Access Control (IAM):

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Add, luego add role assignment

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Seleccionamos SQL DB Contributor:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente  
Click en Next

Add role assignment



Managed Identity y luego Select members:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



Click en Select

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente



Click en Next.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



Click en review + assign

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ahora, vamos a crear con esas identidades asignadas un Cluster SQL Interactivo con HDInsight. Nos vamos a Home y escribimos: HDInsight clusters

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en create:

Basics

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Select cluster type:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Escoger el que dice Interactive Query (Select)

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Asignar login y password para el Cluster

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

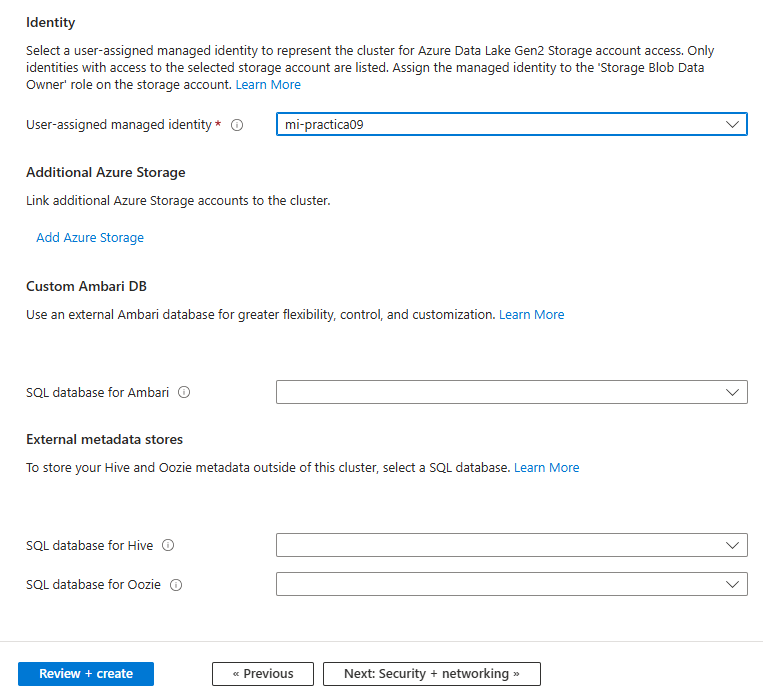
Descripción generada automáticamente

Click en Next: Storage >>

Storage

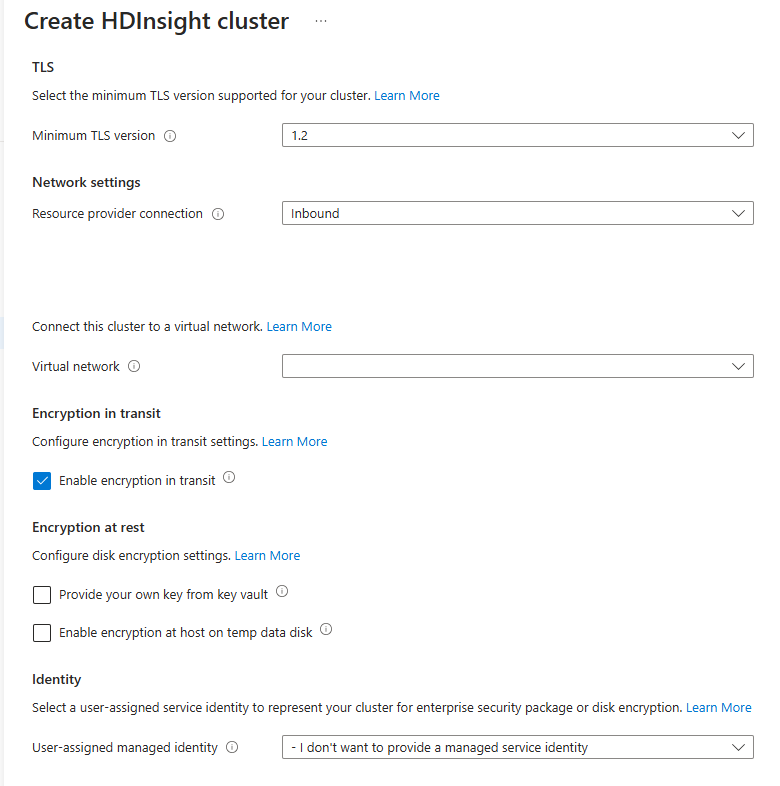
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente



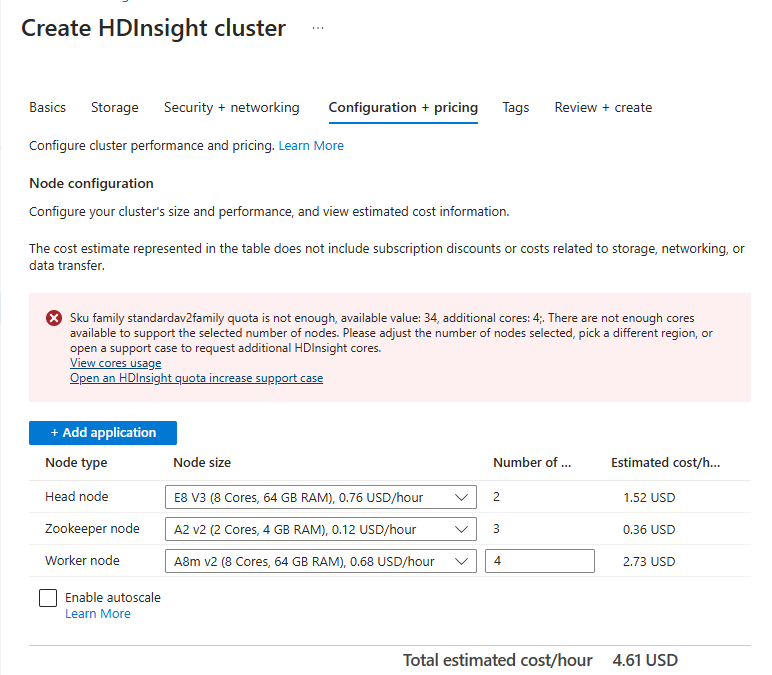
Click en Next: Security + Networking

Dejar lo que aparece por defecto:





Click en Next: Configuration + pricing >>



En el worker node cambiar el 4 por 1



Click en Next: Tags >>

Dejar lo que aparece por defecto

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen que contiene Texto

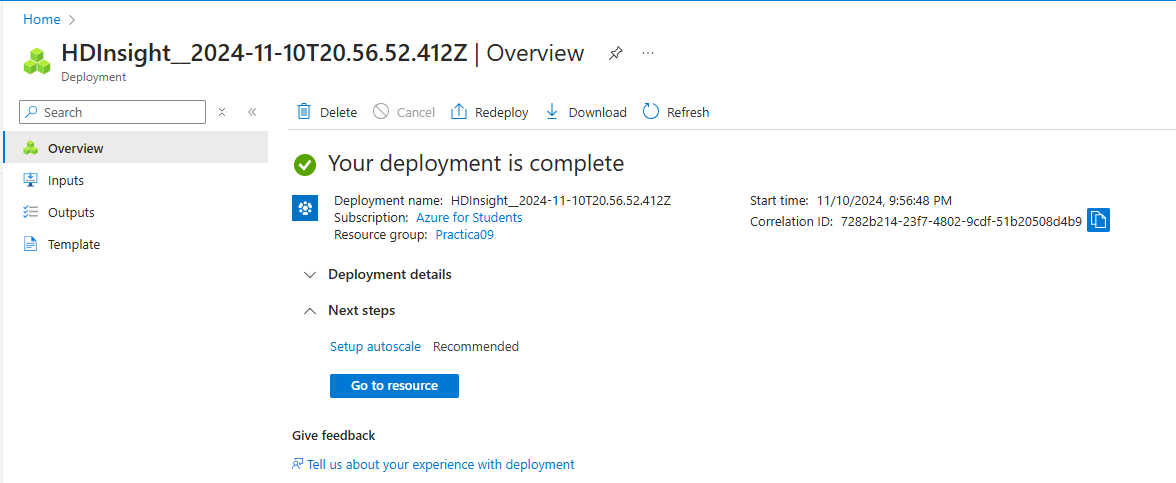
Descripción generada automáticamente

Click en Next: Review + create >>

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Create. Dependiendo de la conexión, tardará unos minutos.



“Data Ingestion”.

Ingestar nuestros datos al Data Lake Storage

Buscamos el archivo FlightDelayData.csv.

Luego en Azure nos vamos a gen2storagepractica09

Una captura de pantalla de una computadora

Descripción generada automáticamente

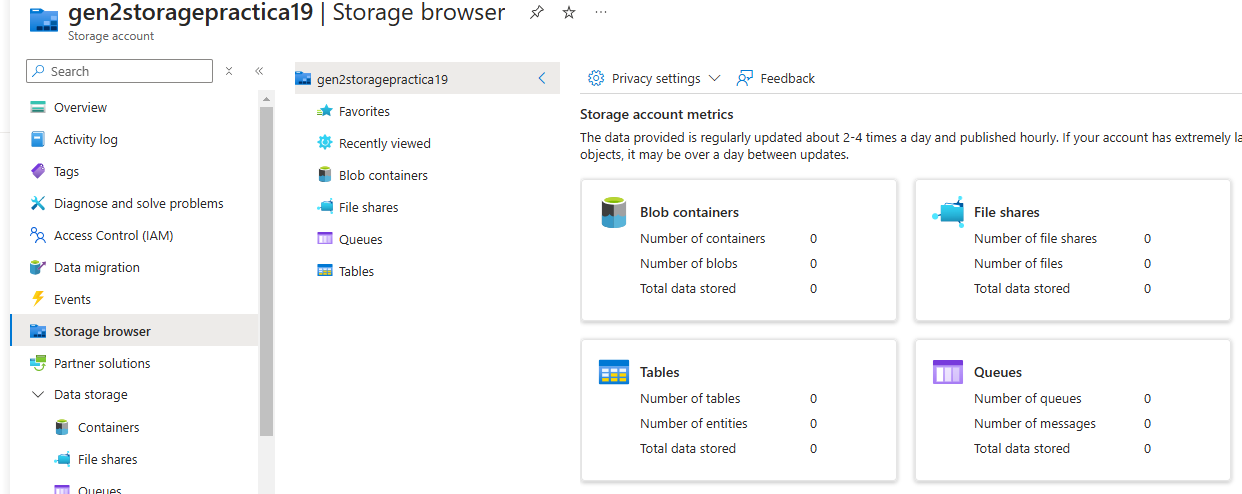
Click en Data Lake Storage:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

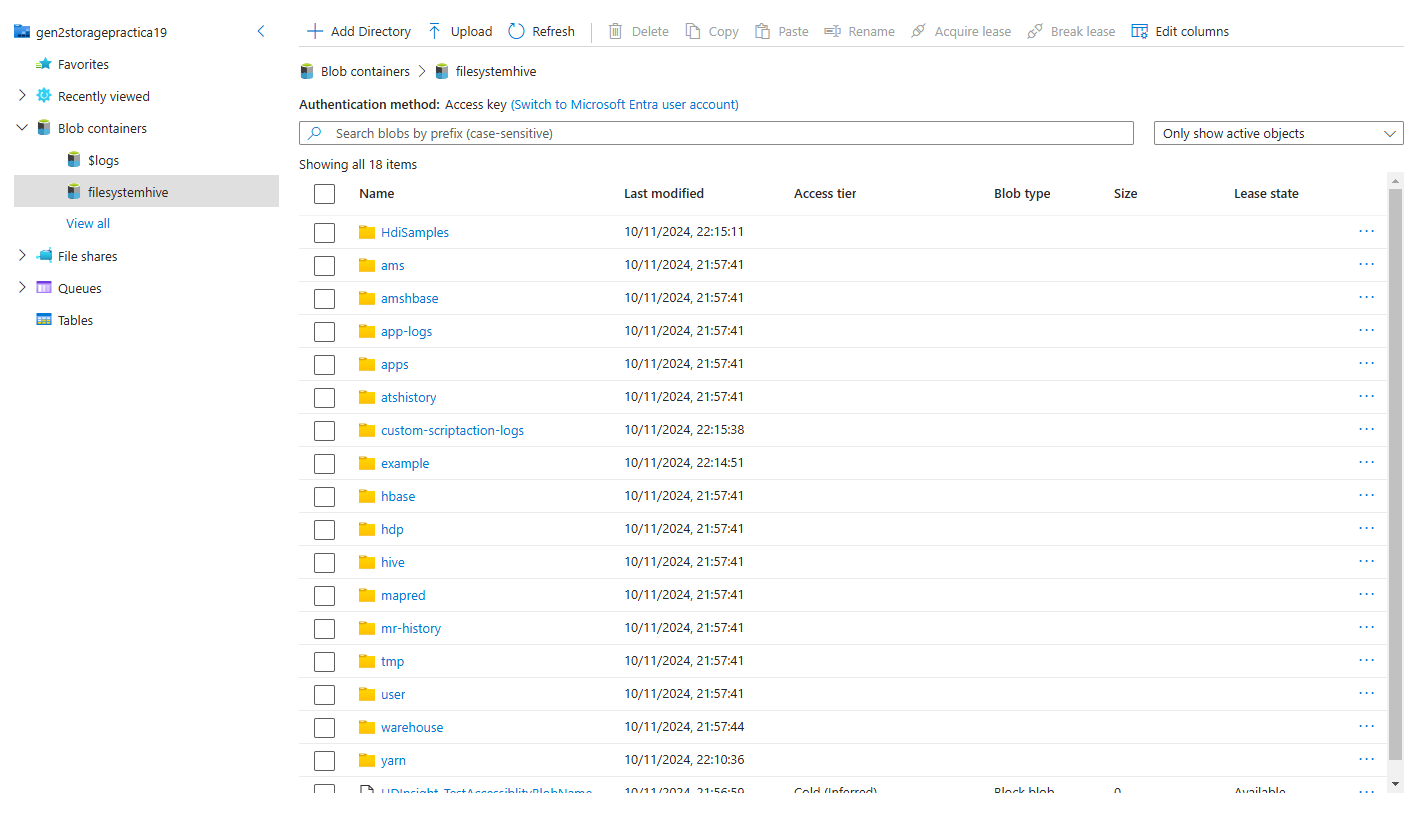
Descripción generada automáticamente

Click en Storage Browser





Click en Blob Containers y luego en filesystemhive



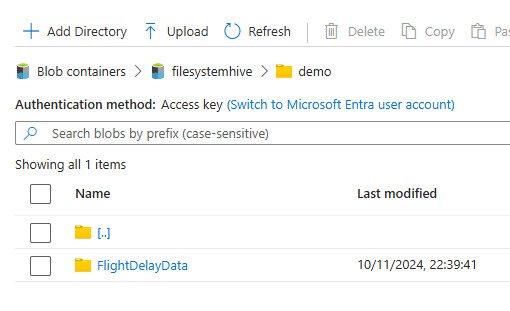
Dando click en Add Directory, creamos una carpeta con nombre demo y luego FlightDelayData.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



Click en Save



Entramos a FlightDelayData y creamos una carpeta con nombre InputData

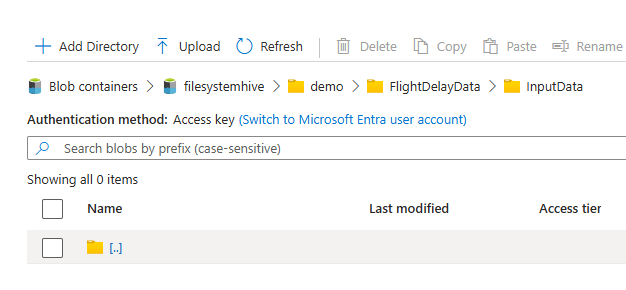
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



Click en Save.

Entramos a la carpeta InputData y subimos el fichero FlightDelayData.



Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Upload

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

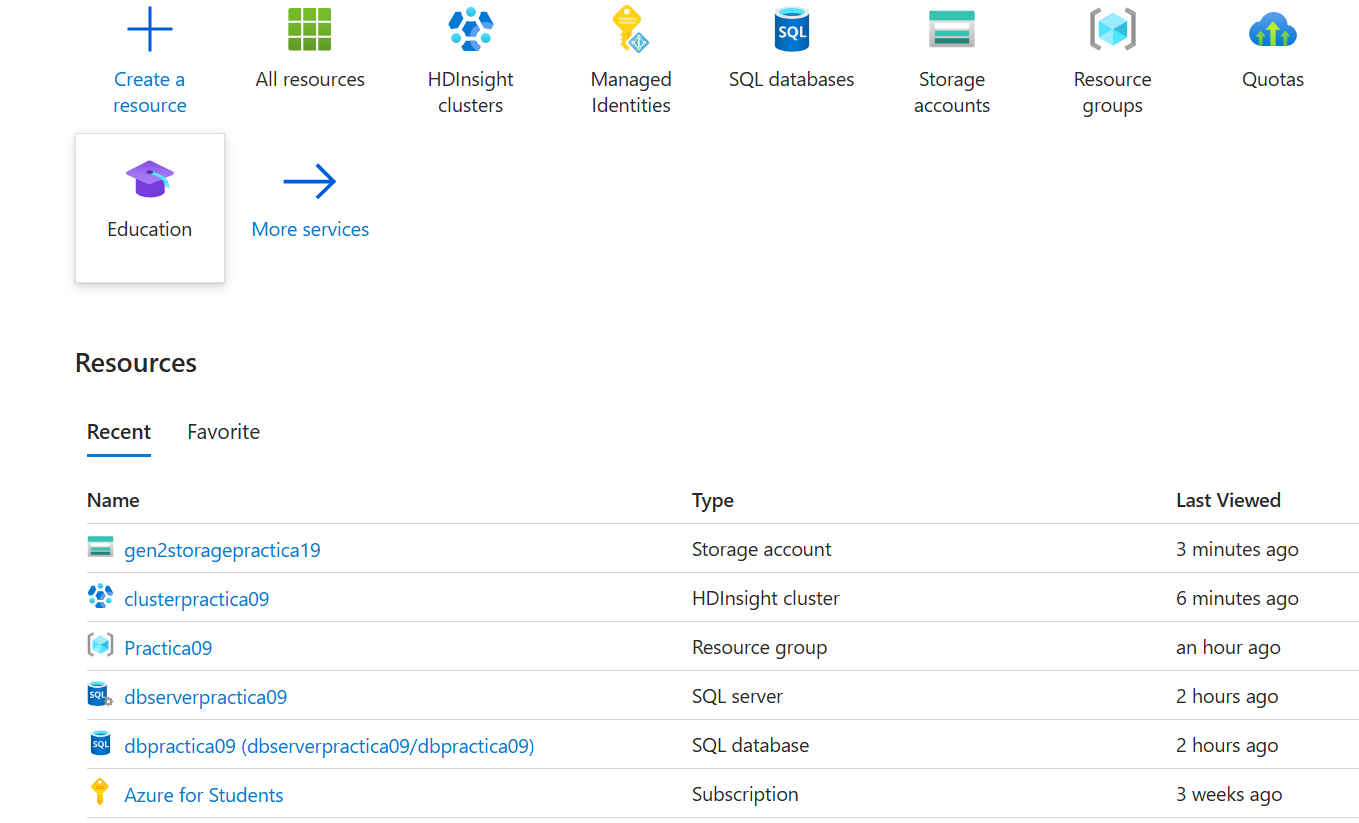
Esperar hasta completarse.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

“Data Extraction” con Hive

Desde Home, damos click en nuestro cluster (clusterpractica09):

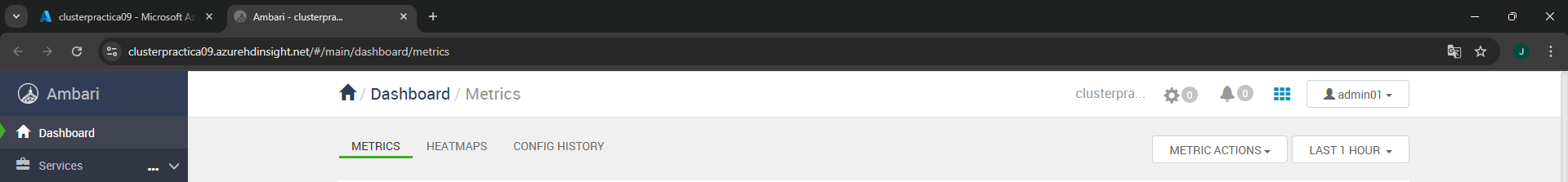


Click en Ambari home:

Captura de pantalla de computadora

Descripción generada automáticamente

Te pedirá el usuario y clave que creaste al momento de levantar el Clúster.



Una vez dentro haz click en:

Icono

Descripción generada automáticamente

Escoger Hive View 2.0

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Después de validar el hdfs se mostrará el editor de ‘queries’ de hive:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Copiar y pegar el contenido del fichero query-1.txt.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

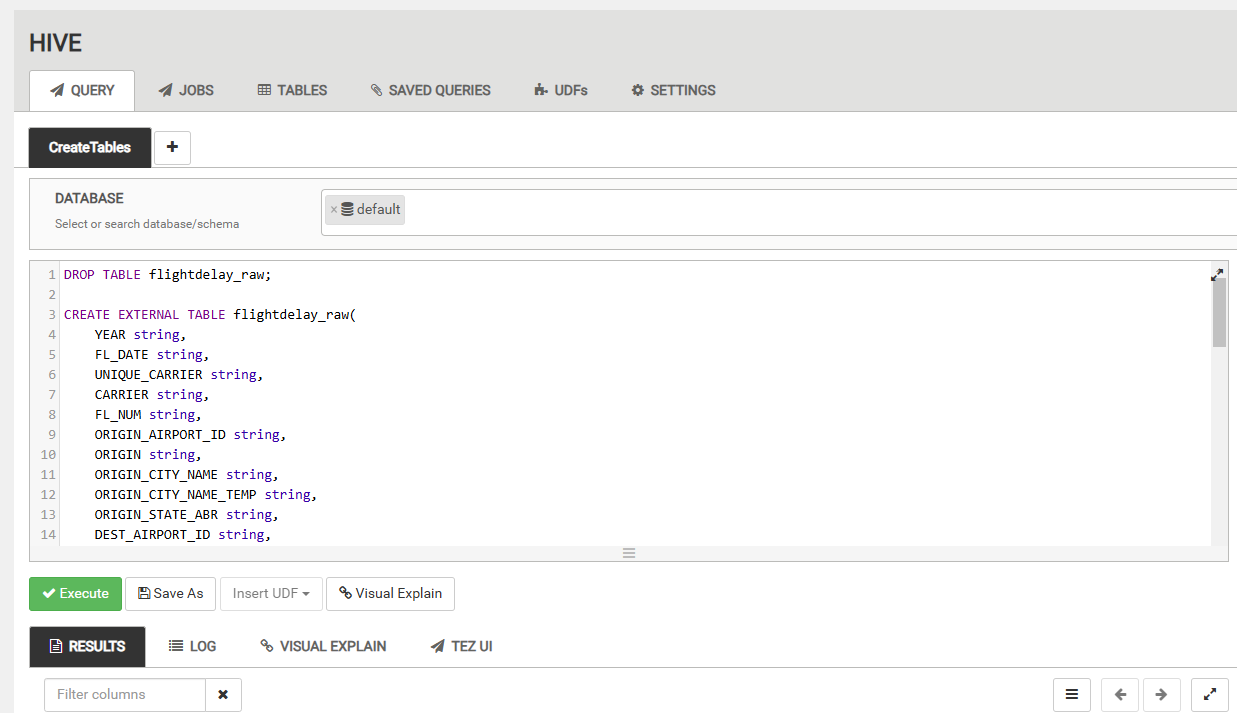
Luego click en execute

Ir a Tables

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Volvemos a Query



Abrimos un nuevo worksheet en el símbolo +:

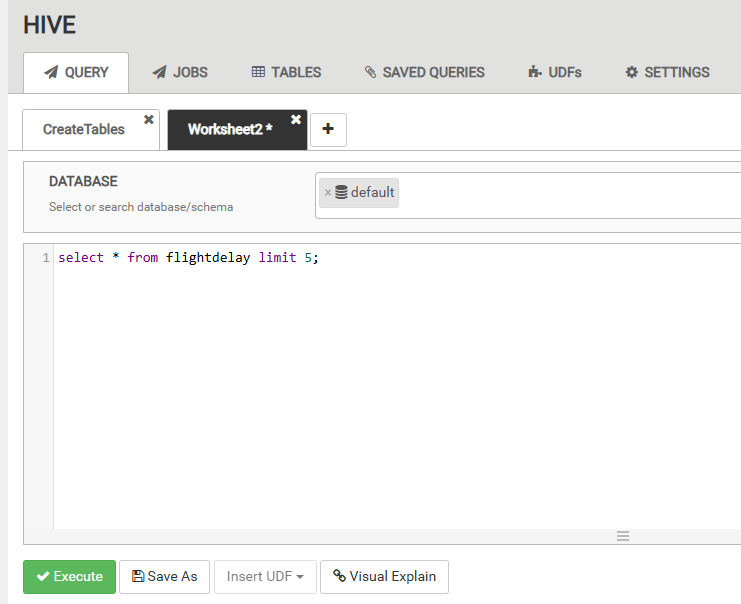
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Hagamos una consulta simple:



Click en Execute



La salida de la consulta se puede ver en Results:

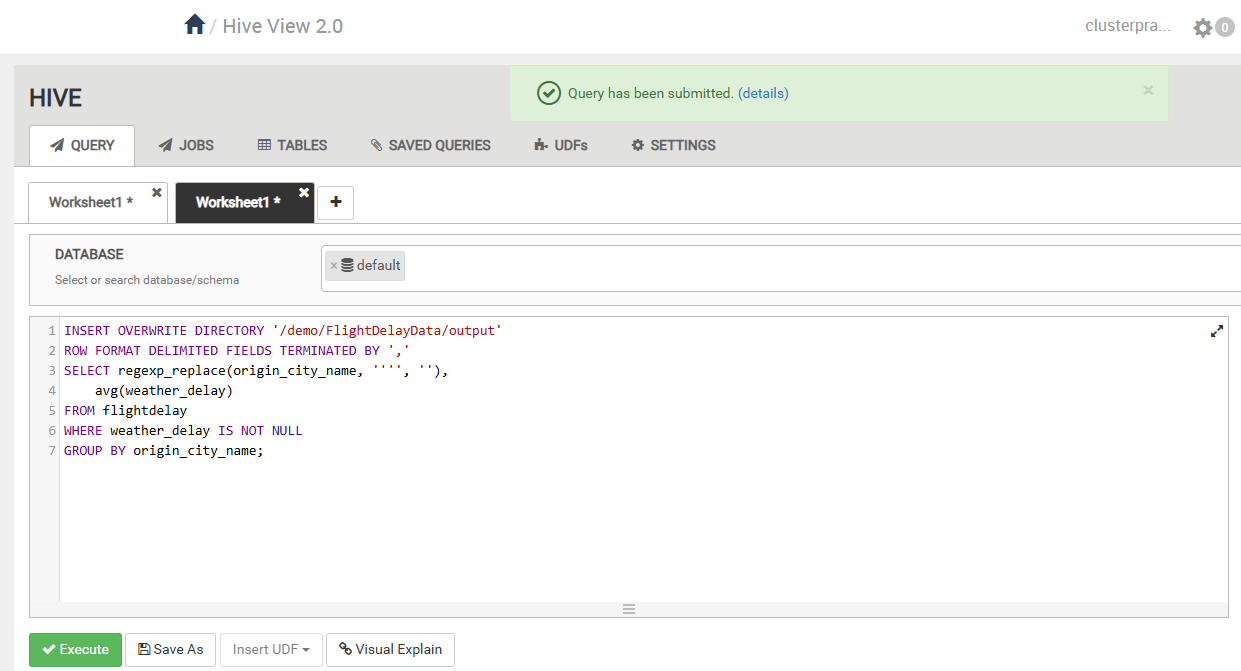
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

En esta sección hemos realizado una ingestión desde Data Lake Storage Gen 2 hasta un lugar dentro del cluster utilizando T-SQL mediante Hive.

“Data Transformation” con Hive

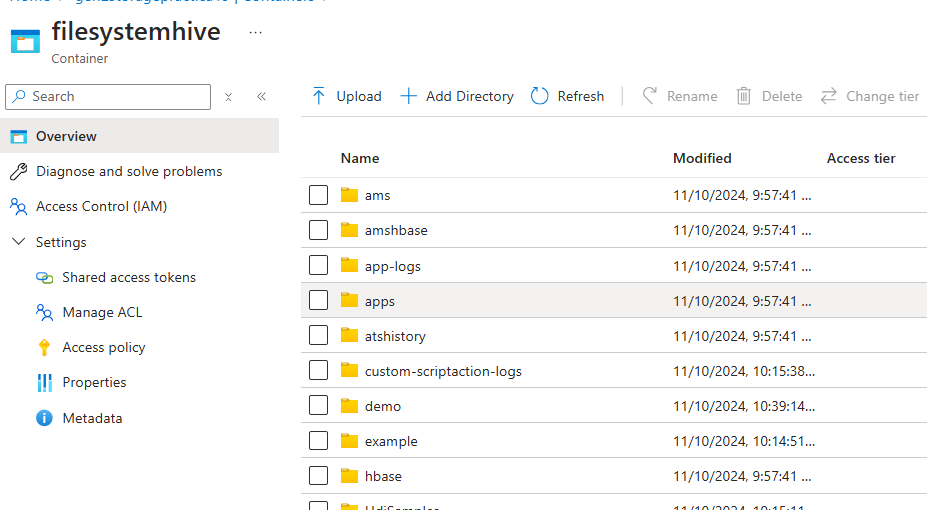
Abrimos un nuevo worksheet y ejecutamos la siguiente query (query-3.txt)



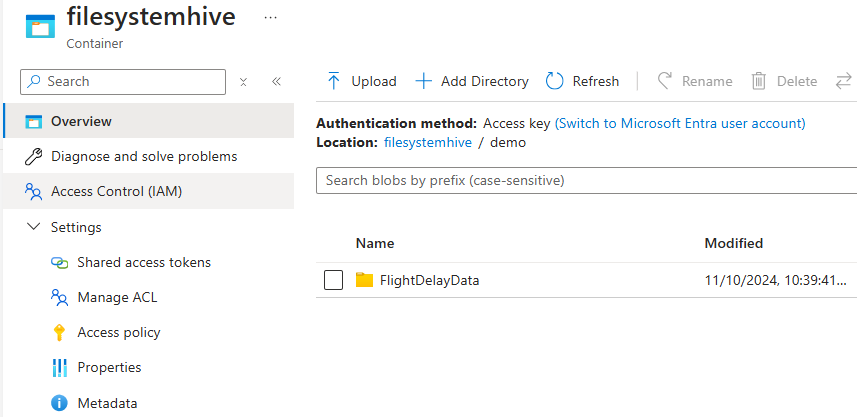
Explicar paso a paso que hace la query-3 de arriba.

Nos vamos a Azure, hasta gen2storagepractica09

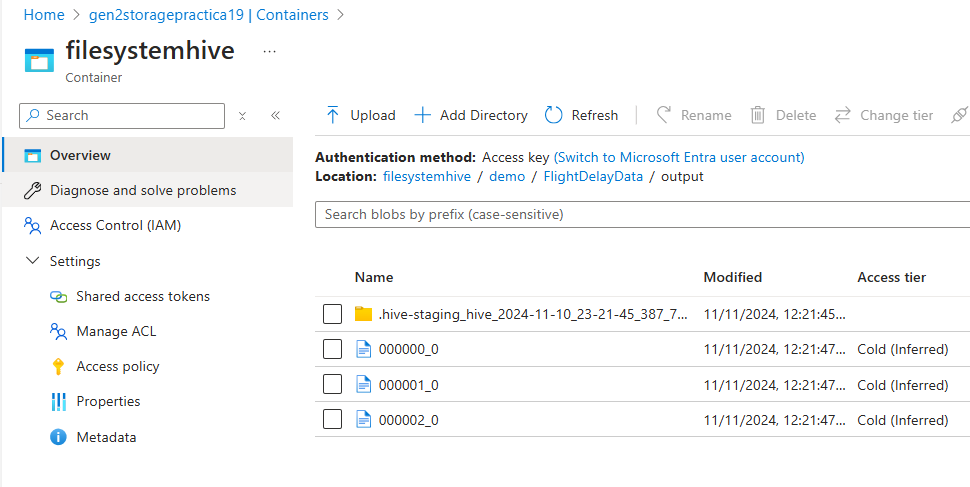
Click en containers y luego click a filesystemhive:



Click en demo:



Click en FlightDelayData y luego output:



La información que tenemos en output no tiene una estructura SQL, hive guardó la salida en chunks. Para tener nuestros datos en un formato SQL usaremos Sqoop.

“Data Export” mediante Sqoop.

Antes de usar sqoop vamos a nuestra base de datos desde azure yendo a dbpractica09.

Home

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en dbpractica09 (SQL Database)

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Query editor (preview)

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

En el editor creamos nuestra tabla mediante la query:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en run:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ahora, vamos de nuevo a nuestro clúster:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Click en SSH + Cluster login

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

En Hostname seleccionamos el ssh:

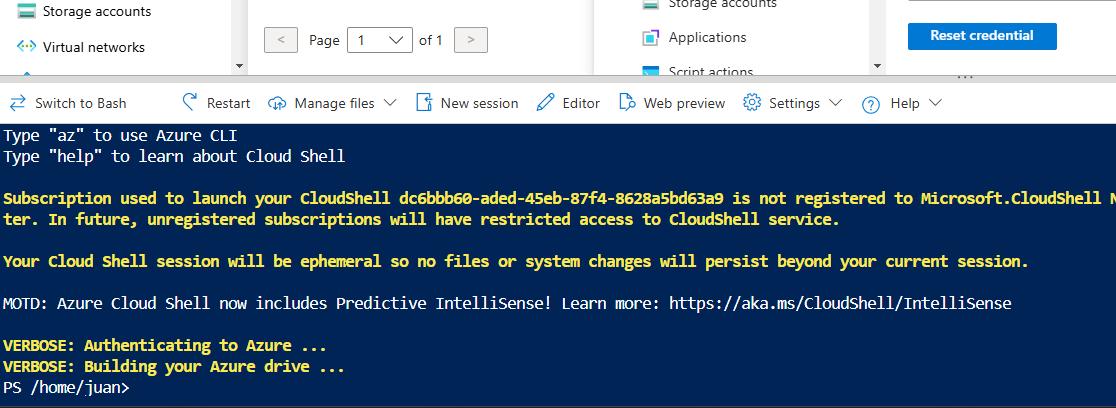
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

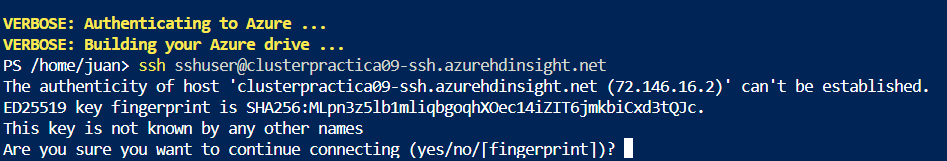
Copiamos el texto.

Nos vamos al powershell de Azure.





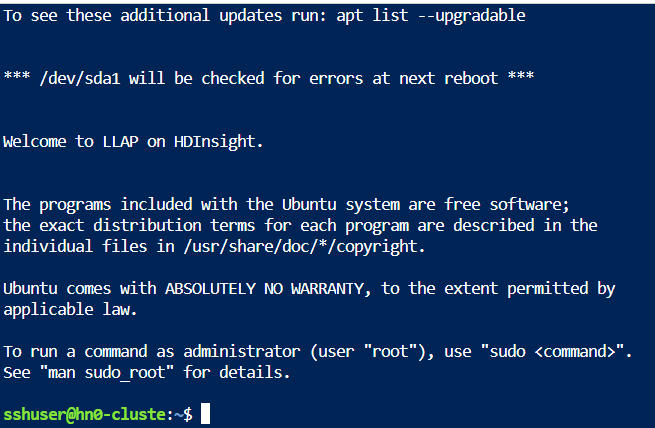
Pegar el string (ssh) que tenemos copiado:



Yes



Introducir cluster password:



Escribimos:

sqoop list-databases --connect jdbc:sqlserver://dbserverpractica09.database.windows.net:1433 --username admin01 -P

Donde: dbserverpractica09.database.windows.net es el nombre de tu servidor

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Si no esta instalado Sqoop en tu HDInsight hay buscar en la documentación o en ChatGPT como hacerlo.



Te pide la contraseña de tu base de datos. Si da error es seguramente por el firewall. Copiamos la IP que se muestra en powershell y luego vamos a nuestro sql server (dbserverpractica09):

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Networking:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Add a firewall rule:

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Añadimos la ip copiada que muestra el error en powershell:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Click en OK

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Click en Save

Nos vamos de nuevo a PowerShell

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Volvemos a escribir el comando de sqoop list -databases…

sqoop list-databases --connect jdbc:sqlserver://dbserverpractica09.database.windows.net:1433 --username admin01 -P

Texto

Descripción generada automáticamente

Una vez dentro, escribimos:

sqoop export --connect 'jdbc:sqlserver://dbserverpractica09.database.windows.net:1433;database=dbpractica09' --username admin01 -P --table 'delays' --export-dir '/demo/FlightDelayData/output' --fields-terminated-by ',' -m 1

El proceso tardará unos 5 o 10 minutos.

Una vez que se termina la ejecución, nos vamos a nuestra base de datos y verificamos que la tabla se ha llenado:

