



Challenge Data engineer
Azure
Documentación técnica

Índice

| | |
|---------------------------|----------|
| 1. Objetivos | 3 |
| 2. Infraestructura | 3 |
| 3. Diseño | 8 |
| 4. Versión | 8 |
| 5. Fuentes | 9 |

1. Objetivos

El objetivo de este documento es detallar el desarrollo que se realizó para este challenge de la universidad Siglo XXI, en la cual consistía obtener información de una API pública, procesarla y disponibilizar la información en un contenedor de Azure Data Lake. También, requería la obtención de información desde una base de datos de Azure Database y persistir en Azure Data Lake en formato parquet. A continuación se destacan los pasos seguidos para deployar la infraestructura para este proceso y las decisiones de diseño que se tomaron.

2. Infraestructura

En este punto se destacan la infraestructura desplegada.

Grupo de Recursos

Dentro de nuestra suscripción creamos un grupo de recursos:

[Inicio](#) > [Grupos de recursos](#) >

Crear un grupo de recursos

[Datos básicos](#) [Etiquetas](#) [Revisar y crear](#)

Grupo de recursos - Contenedor que incluye los recursos relacionados para una solución de Azure. El grupo de recursos puede contener todos los recursos de la solución o solamente los recursos que quiere administrar en grupo. Debe decidir cómo quiere asignar los recursos a los grupos de recursos según lo que resulte más pertinente para su organización. [Más información](#)

Detalles del proyecto

Suscripción * ⓘ Azure subscription 1 ▼

Grupo de recursos * ⓘ sigloXXI_RG ✓

Detalles del recurso

Región * ⓘ (US) East US ▼

Observamos que se creó

Grupos de recursos

Directorio predeterminado

[+ Crear](#) [Administrar vista](#) [Actualizar](#) [Exportar a CSV](#) [Abrir consulta](#) [Asignar etiquetas](#)

Filtrar por cualquier ca... [Suscripción es igual a todo](#) [Ubicación es igual a todo](#) [Agregar filtro](#)

Mostrando de 1 a 2 de 2 registros.

| Nombre ⓘ | Suscripción ⓘ | Ubicación ⓘ |
|---|----------------------|-------------|
| <input type="checkbox"/> 03_management_rg | Azure subscription 1 | East US |
| <input type="checkbox"/> sigloXXI_RG | Azure subscription 1 | East US |

Cuenta de almacenamiento

Creamos una cuenta de almacenamiento

[Inicio](#) > [Cuentas de almacenamiento](#) >

Crear una cuenta de almacenamiento ...

Datos básicos

[Opciones avanzadas](#)

[Redes](#)

[Protección de datos](#)

[Cifrado](#)

[Etiquetas](#)

[Revisar](#)

Azure Storage es un servicio administrado por Microsoft que proporciona almacenamiento en la nube altamente disponible, seguro, duradero, escalable y redundante. Azure Storage incluye Azure Blob (objetos), Azure Data Lake Storage Gen2, Azure Files, Azure Queues y Azure Tables. El costo de una cuenta de Storage depende del uso y de las opciones que elija a continuación. [Más información sobre las cuentas de almacenamiento de Azure](#)

Detalles del proyecto

Seleccione la suscripción en la que se creará la nueva cuenta de almacenamiento. Elija un grupo de recursos nuevo o uno ya existente para organizar y administrar la cuenta de almacenamiento junto con otros recursos.

| | |
|---------------------|---|
| Suscripción * | <input type="text" value="Azure subscription 1"/> |
| Grupo de recursos * | <input type="text" value="sigloXXI_RG"/> |

[Crear nuevo](#)

Detalles de la instancia

| | |
|---|---|
| Nombre de la cuenta de almacenamiento ⓘ * | <input type="text" value="sigloxxietlsa"/> |
| Región ⓘ * | <input type="text" value="(US) East US"/> |
| | Implementar en una zona perimetral |
| Rendimiento ⓘ * | <input checked="" type="radio"/> Estándar: Opción recomendada para la mayoría de los escenarios (cuenta de uso general v2) <input type="radio"/> Prémium: Se recomienda para escenarios que requieren una latencia baja. |
| Redundancia ⓘ * | <input type="text" value="Almacenamiento con redundancia local (LRS)"/> |

Habilitamos la opción de nombres jerárquicos

Espacio de nombres jerárquico

El espacio de nombres jerárquico, complementado con el punto de conexión de Data Lake Storage Gen2, habilita la semántica de archivos y directorios, acelera las cargas de trabajo de análisis de macrodatos y habilita las listas de control de acceso (ACL) [Más información](#)

| | |
|--|-------------------------------------|
| Habilitar el espacio de nombres jerárquico | <input checked="" type="checkbox"/> |
|--|-------------------------------------|

Observamos que se creó

The screenshot shows the Azure portal interface for an implementation. At the top, the title is 'sigloxxietlsa_1698852766186 | Información general'. Below the title, there's a search bar and a row of action buttons: 'Eliminar', 'Cancelar', 'Volver a implementar', 'Descargar', and 'Actualizar'. On the left, a sidebar lists 'Información general' (selected), 'Entradas', 'Salidas', and 'Plantilla'. The main content area shows a green checkmark and the text 'Se completó la implementación'. Below this, it lists implementation details: 'Nombre de implementación: sigloxxietlsa_1698852766186', 'Suscripción: Azure subscription 1', and 'Grupo de recursos: sigloXXI_RG'. To the right, it shows 'Hora de inicio: 1/11/2023, 12:32:52' and 'Id. de correlación: 421c2e3b-f565-400f-8577-61daac8b15b6'. There are expandable sections for 'Detalles de implementación' and 'Pasos siguientes', with a 'Ir al recurso' button under the latter. At the bottom, there's a link to 'Enviar comentarios' and a feedback prompt 'Cuéntenos su experiencia con la implementación'.

Creamos un contenedor que se llame “input”:

Nuevo contenedor

Nombre *

Nivel de acceso anónimo ⓘ

Privada (sin acceso anónimo) ▼



El nivel de acceso está definido como privado porque el acceso anónimo está deshabilitado en esta cuenta de almacenamiento.

▼ Avanzado

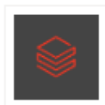
Azure DataBricks

Creamos un nuevo recurso que es Azure Databricks

[Inicio](#) > [Crear un recurso](#) > [Marketplace](#) >

Azure Databricks

Microsoft



Azure Databricks [Agregar a Favoritos](#)

Microsoft | Azure Service

★ 4.5 (354 clasificaciones)

Plan

Azure Databricks

Crear

[Información general](#)

[Planes](#)

[Información de uso y soporte técnico](#)

[Clasificaciones + reseñas](#)

[Inicio](#) > [Crear un recurso](#) > [Marketplace](#) > [Azure Databricks](#) >

Creación de un área de trabajo de Azure Databricks

[Datos básicos](#)

[Redes](#)

[Cifrado](#)

[Etiquetas](#)

[Revisar y crear](#)

Detalles del proyecto

Seleccione la suscripción para administrar recursos implementados y los costes. Use los grupos de recursos como carpetas para organizar y administrar todos los recursos.

Suscripción * ⓘ

Azure subscription 1

Grupo de recursos * ⓘ

sigloXXI_RG

[Crear nuevo](#)

Detalles de instancia

Nombre del área de trabajo *

siglo_XXI_AD

Región *

East US

Plan de tarifa * ⓘ

Standard (Apache Spark, Secure with Microsoft Entra ID)

Nombre del grupo de recursos administrados

Enter name for managed resource group

Observamos que se creó correctamente

Inicio >

siglo_XXI_AD
Servicio de Azure Databricks

Buscar Eliminar

Información general

- Registro de actividad
- Control de acceso (IAM)
- Etiquetas
- Diagnosticar y solucionar problemas

Configuración

- Emparejamiento de Virtual Network
- Cifrado
- Redes
- Propiedades
- Bloqueos
- Automation
 - CU / PS

Información esencial

Estado : Active

Grupo de recursos : [sigloXXI_RG](#)

Ubicación : East US

Suscripción : [Azure subscription 1](#)


Id. de suscripción : 301619bd-ccb9-45b8-9fa9-9e33dacc67c5

Etiquetas ([editar](#)) : [Agregar etiquetas](#)

Grupo de recursos admin... : [databricks-rg-siglo_XXI_AD-rgjw2vb2hkv6](#)

URL : <https://adb-415444347716914.14.azuredatabricks.net>

Plan de tarifa : [Standard \(Apache Spark, Secure with Microsoft Entra ID\) \(Click to change\)](#)



[Iniciar área de trabajo](#)

[Actualizar a Premium](#)

Iniciamos Azure Databricks:

Microsoft Azure databricks Buscar datos, cuadros, recientes y más... CTRL + P

Nuevo

- Workspace
- Recientes
- Catálogo
- Flujos de trabajo
- Cómputo
- Ingeniería de datos
- Ejecuciones
- Tablas Delta Live
- Machine Learning
- Experimentos
- Características
- Modelos
- Servicio

Workspace

- Inicio
- Workspace
 - Shared
 - Users
- Repos
- Favoritos
- Papelera

Workspace

| Nombre | Tipo | Propietario |
|--------|---------|-------------|
| Shared | Carpeta | Desconocido |
| Users | Carpeta | Desconocido |

Dentro creamos un cluster para poder realizar las pruebas:

Cómputo > Nuevo cómputo > Vista previa de la IU [Enviar comentarios](#)

Siglo XXI Cluster [✎](#)

☒ Multi-nodo ☐ Nodo único

Modo de acceso [?](#) Acceso de usuario único [?](#)

Usuario único [▼](#)

Nicolas Herrera [▼](#)

Rendimiento

Versión de Databricks Runtime [?](#)

Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1) [▼](#)

☒ Utilizar aceleración Photon [?](#)

Tipo de worker [?](#)

Standard_DS3_v2

14 GB de memoria y 4 núcleos [▼](#)

Workers mín.

1

Workers máx.

1

☐ Instancias de spot [?](#)

Tipo de driver

El mismo que el worker

14 GB de memoria y 4 núcleos [▼](#)

☒ Habilitar la auto expansión [?](#)

☒ Terminar después de minutos de inactividad [?](#)

Etiquetas [?](#)

Añadir etiquetas

Key

Value

Añadir

> Etiquetas añadidas automáticamente

► Opciones avanzadas

Conexión entre Databricks y Azure Storage

Para realizar la conexión entre Databricks y Azure Storage se utilizaron los siguientes servicios de Azure:

- Microsoft Entra ID
- Key Vault
- IAM
- Databrick Scope

Para la configuración se siguió la documentación oficial de Microsoft Azure:

<https://learn.microsoft.com/es-es/azure/databricks/getting-started/connect-to-azure-storage>

Azure SQL Database

Creamos el servidor

[Inicio](#) > [SQL Database](#) > [Crear base de datos SQL](#) >

Crear un servidor de SQL Database ...

Microsoft

Detalles del servidor

Especifique la configuración necesaria para este servidor, incluida la inclusión de un nombre y una ubicación. Este servidor se creará en la misma suscripción y grupo de recursos que la base de datos.

Nombre del servidor *

siglo-xxi-server




.database.windows.net

Ubicación *

(US) East US



Autenticación

 Azure Active Directory (Azure AD) is now Microsoft Entra ID. [Más información](#)

Seleccione los métodos de autenticación preferidos para acceder a este servidor. Cree un servidor inicio de sesión y una contraseña de administrador para acceder a su servidor con autenticación de SQL, seleccione solo Azure AD autenticación. [Más información](#) use un usuario, grupo o aplicación de Azure AD existente como administrador de Azure AD [Más información](#), o seleccione la autenticación de SQL y Azure AD.

Método de autenticación



Usar solo la autenticación de Azure Active Directory (Azure AD)



Uso de la autenticación de SQL y Azure AD



Uso de la autenticación de SQL

Establecer administrador de Azure AD *

No seleccionado

[Establecer administrador](#)

Eliminar Cancelar Volver a implementar Descargar Actualizar

La implementación está en curso

Nombre de implementación : Microsoft.SQLDatabase.newDatabaseNewServer_e282287027d04ddcbf4a4
 Suscripción : Azure subscription 1
 Grupo de recursos : sigloXXI_RG

Detalles de implementación

| Recurso |
|------------------|
| siglo-xxi-server |

Enviar comentarios

Cuéntenos su experiencia con la implementación

Una vez creada necesitamos crear una regla de firewall con nuestra ip para que permita conectarnos.

Nos conectamos

DB_SigloXXI (siglo-xxi-server/DB_SigloXXI) | Editor de consultas (version preliminar) ☆ ...

Base de datos SQL

Buscar Inicio de sesión Nueva consulta Abrir consulta Comentarios Introducción

Información general
 Registro de actividad
 Etiquetas
 Diagnosticar y solucionar problemas
 Editor de consultas (versión preliminar)
 Configuración
 Proceso y almacenamiento
 Cadenas de conexión
 Propiedades
 Bloqueos
 Administración de datos
 Réplicas

Query editor (preview) is a tool to run SQL queries against Azure SQL Database in the Azure portal. It is designed for lightweight querying and object exploration in your database. For more information and troubleshooting, [Más información](#)

SQL

Le damos la bienvenida al editor de consultas de SQL Database

Autenticación de SQL Server

Inicio de sesión *

adminsqli

Contraseña *

.....

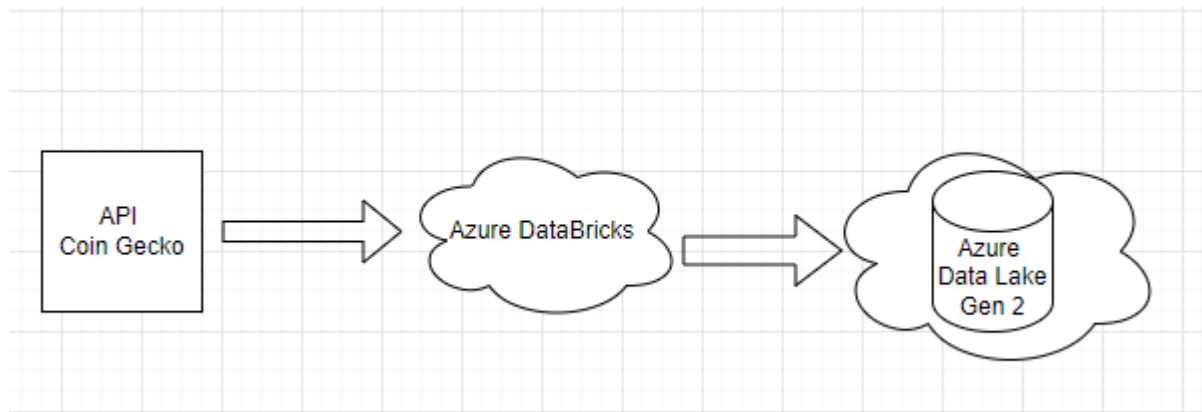
Aceptar

Autenticación de Active Directory

Continuar como nahdevelopment@hotmail...

3. Diseño

Para realizar este challenge se decidió utilizar la API de CoinGecko, la cual proporciona información sobre precios de criptomonedas.



Se extrajo la información para la siguientes monedas:

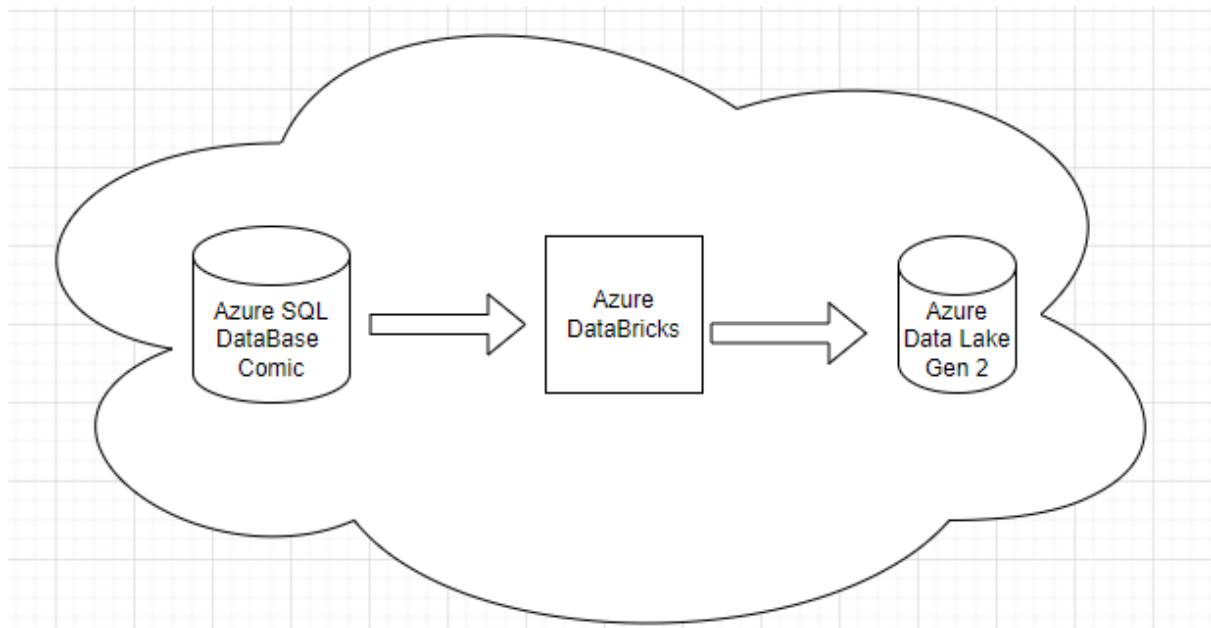
- ethereum
- ripple
- litecoin
- cardano

Los campos extraídos fueron:

- fecha actual
- nombre de la moneda
- valor en dólares
- valor en dólares de las últimas 24 hs

Para el ítem 2, se creó una base de datos Comic, la cual tiene una tabla llamada superheroes que contiene los campos:

- Personaje
- Nombre_Alter_Ego
- Edad
- Altura_en_CM
- Peso_en_kg
- Superpoder
- Ciudad



Mejoras por hacer:

Si bien se extrajeron los datos requeridos para cumplir con el MVP se puede extraer mucha más información para ser explotada.

Con respecto a las conexiones, por una cuestión de tiempo se optó por incluir las credenciales en el mismo notebook. Esto se debería mejorar con un notebook que contenga las variables de entorno y utilizar un servicio como Azure KeyVault para guardar esta información.

4. Versión

| Versión | Autor | Fecha Creación | Fecha Modificación |
|---------|-----------------|----------------|--------------------|
| v1 | Nicolás Herrera | 31/10/2023 | 01/11/2023 |

5. Fuentes

- CoinGecko: <https://www.coingecko.com/api/documentation>
- KeyVault:
<https://learn.microsoft.com/es-es/azure/databricks/getting-started/connect-to-azure-storage>