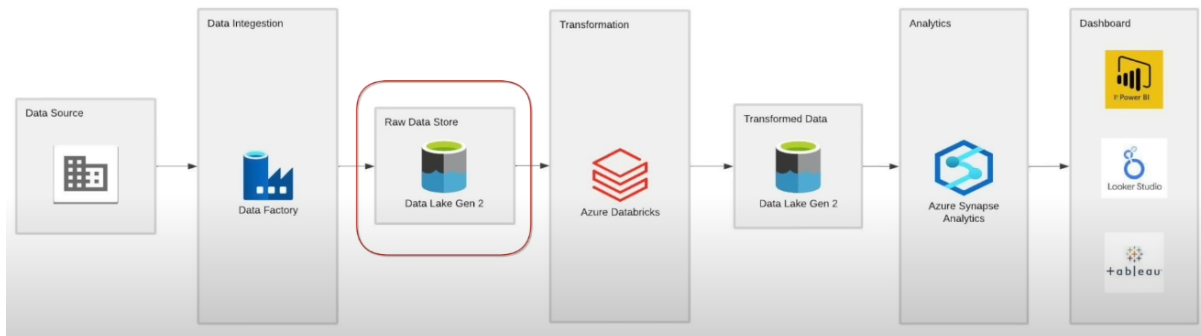
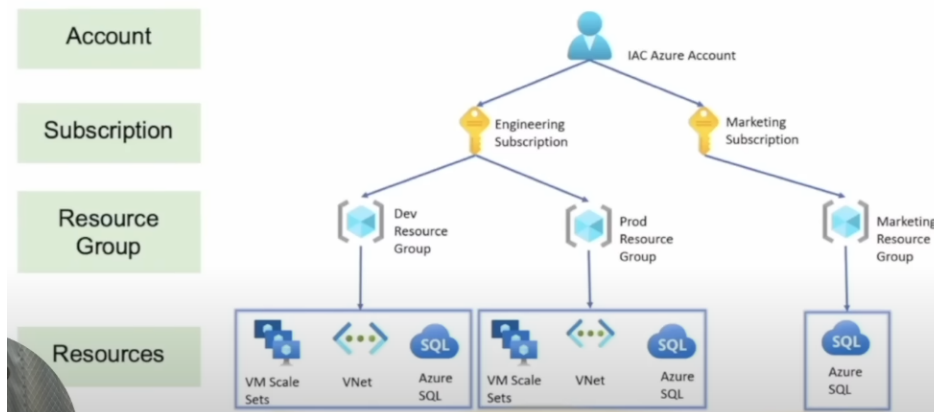


Tokyo Olympic Data Analytics | End-To-End Azure Data Engineering Project



Azure Account Structure



Data Factory

Data integration service that enables you to create, schedule, and manage data pipelines for efficient data movement and transformation between various sources and destinations in Azure and beyond. It simplifies ETL (Extract, Transform, Load) and data integration tasks.



Data Lake Gen 2

Data lake solution that combines the capabilities of a data lake with the power of Azure Blob Storage, allowing you to store and analyze large volumes of structured and unstructured data with enhanced performance, security, and analytics capabilities.



Azure Databricks

Databricks is a unified analytics platform built on top of Apache Spark, designed to help data engineers and data scientists collaborate on big data processing and machine learning tasks. It provides tools for data exploration, data processing, and building machine learning models in a collaborative and scalable environment.



Synapse Analytics

SQL Data Warehouse, is a cloud-based analytics service provided by Microsoft Azure. It combines big data and data warehousing into a single integrated platform, allowing organizations to analyze and process large volumes of data for business intelligence and data analytics purposes.

Ingresamos a nuestra cuenta Azure y creamos una storage account.

Microsoft Azure

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

Home > Storage accounts >

Create a storage account

BasicsAdvancedNetworkingData protectionEncryptionTagsReview

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *

Azure subscription 1

Resource group *

(New) tokyo-olympic

Create new

Instance details

Storage account name ⓘ *

tokyoolympicdatastorage1

Region ⓘ *

(US) East US

Deploy to an edge zone

Performance ⓘ *

☒ Standard: Recommended for most scenarios (general-purpose v2 account)

☐ Premium: Recommended for scenarios that require low latency.

Redundancy ⓘ *

Geo-redundant storage (GRS)

☒ Make read access to data available in the event of regional unavailability.

Review

< Previous

Next: Advanced >

En la sección Advanced habilitamos el Hierarchical Namespace, con el fin de organizar nuestros archivos como se hace normalmente en nuestro directorio local.

Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) [Learn more](#)

Enable hierarchical namespace ☒

Los demás valores no los modificamos. Finalmente creamos la cuenta de almacenamiento.

Microsoft Azure

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

Home > tokyoolympicdatastorage1_1695394530402 | Overview >

tokyoolympicdatastorage1

Storage account

Upload

Open in Explorer

Delete

Move

Refresh

Open in mobile

CU / PS

Feedback

Overview

Activity log

Tags

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Data storage

Containers

File shares

Queues

Tables

Security + networking

Networking

Access keys

Shared access signature

Encryption

Microsoft Defender for Cloud

Data management

Redundancy

Data protection

Blob inventory

Static website

Essentials

Resource group (move) : tokyo-olympic

Location : East US

Primary/Secondary Location : Primary: East US, Secondary: West US

Subscription (move) : Azure subscription 1

Subscription ID : 68d2a34a-fc41-49b2-89b2-dc68dbed9512

Disk state : Primary: Available, Secondary: Available

Tags (edit) : Add tags

PropertiesMonitoringCapabilities (5)Recommendations (0)TutorialsTools + SDKs

Data Lake Storage

Hierarchical namespace : Enabled

Default access tier : Hot

Blob anonymous access : Disabled

Blob soft delete : Enabled (7 days)

Container soft delete : Enabled (7 days)

Versioning : Disabled

Change feed : Disabled

NFS v3 : Disabled

SFTP : Disabled

File service

Large file share : Disabled

Active Directory : Not configured

Default share-level permissions : Disabled

Soft delete : Enabled (7 days)

Share capacity : 5 TiB

Performance : Standard

Replication : Read-access geo-redundant storage (RA-GRS)

Account kind : StorageV2 (general purpose v2)

Account type : Standard

Provisioning state : Succeeded

Created : 9/22/2023, 9:55:33 AM

JSON View

Security

Require secure transfer for REST API operations : Enabled

Storage account key access : Enabled

Minimum TLS version : Version 1.2

Infrastructure encryption : Disabled

Networking

Allow access from : All networks

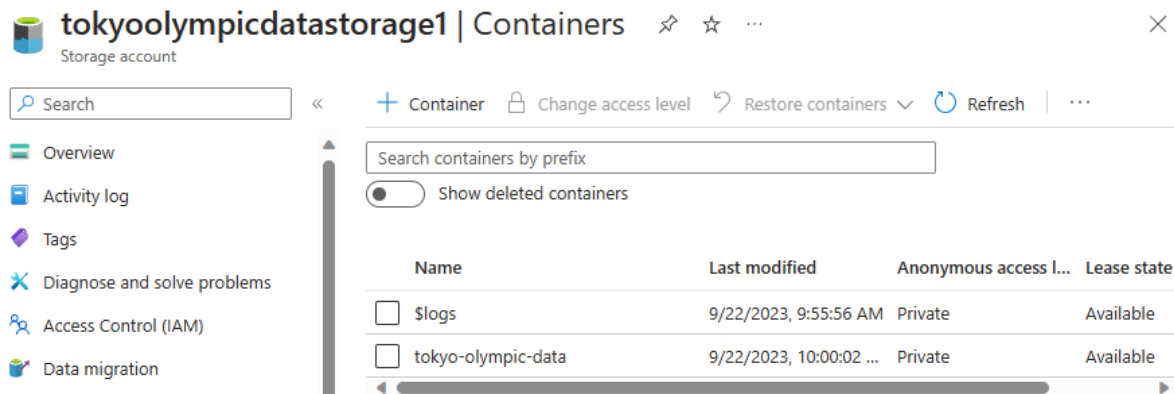
Number of private endpoint connections : 0

Network routing : Microsoft network routing

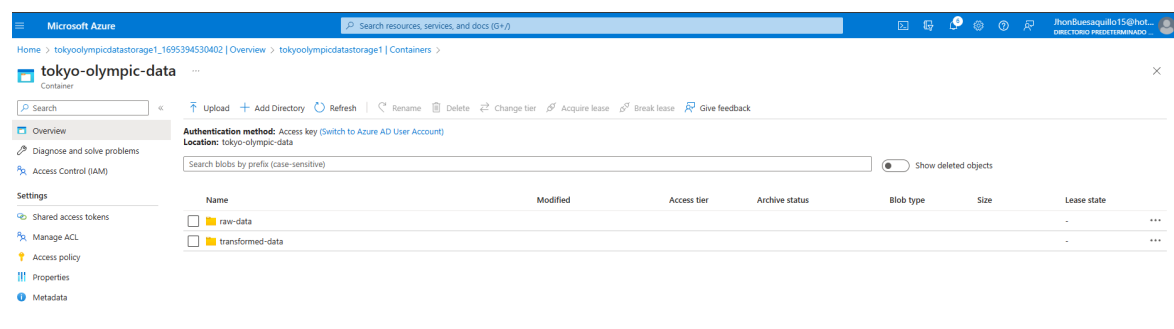
Access for trusted Microsoft services : Yes

Endpoint type : Standard

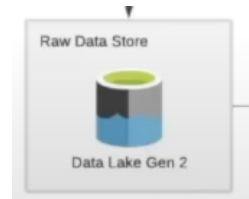
En este laboratorio nos concentraremos en la sección de Containers. Creamos uno nuevo llamado tokyo-olympic-data.



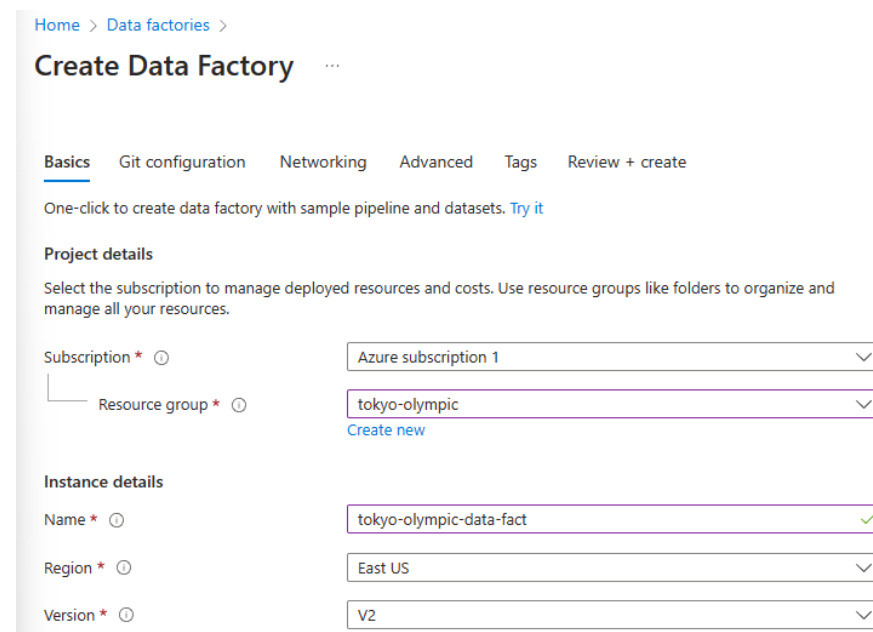
Dentro del container creamos dos directorios para almacenar los datos.



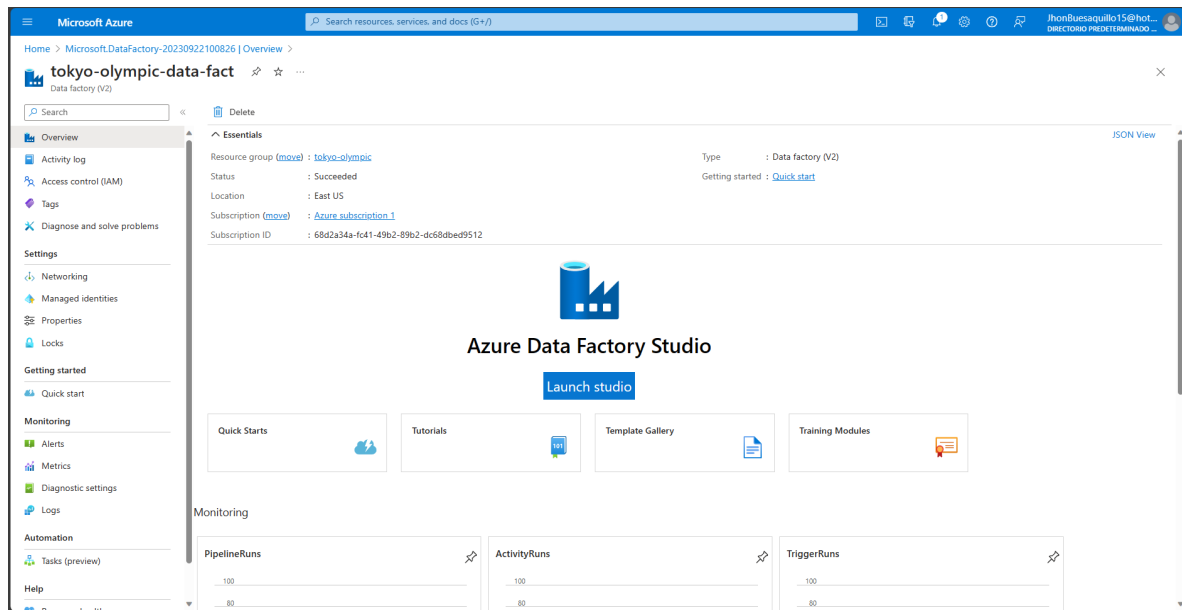
Lo que hemos hecho hasta ahora se encuentra abarcado en esta parte del esquema.



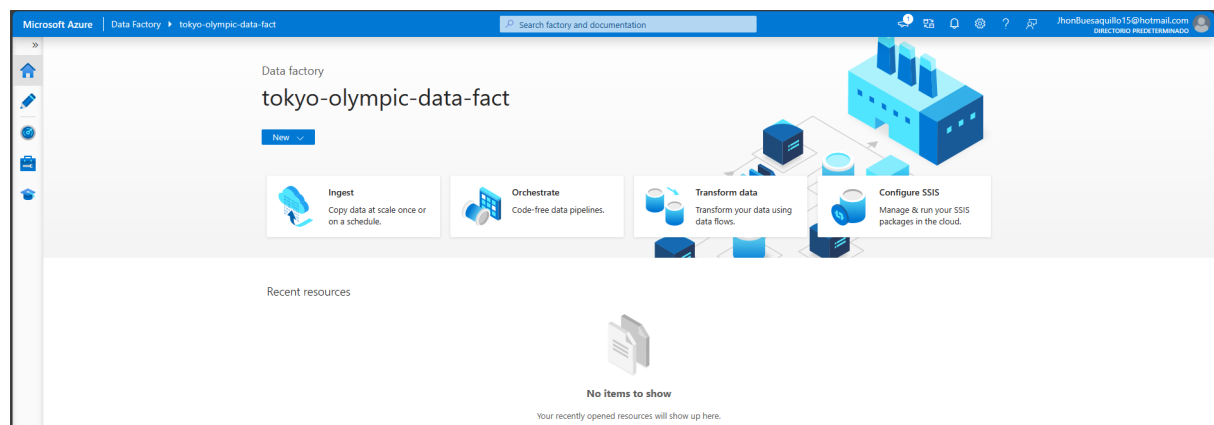
Ahora, ingresamos al servicio de Data factories y creamos una nueva Data Factory.



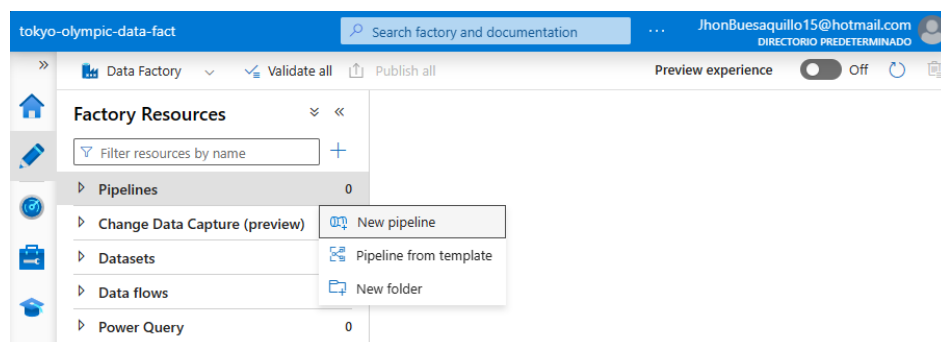
El servicio creado se puede ver a continuación.



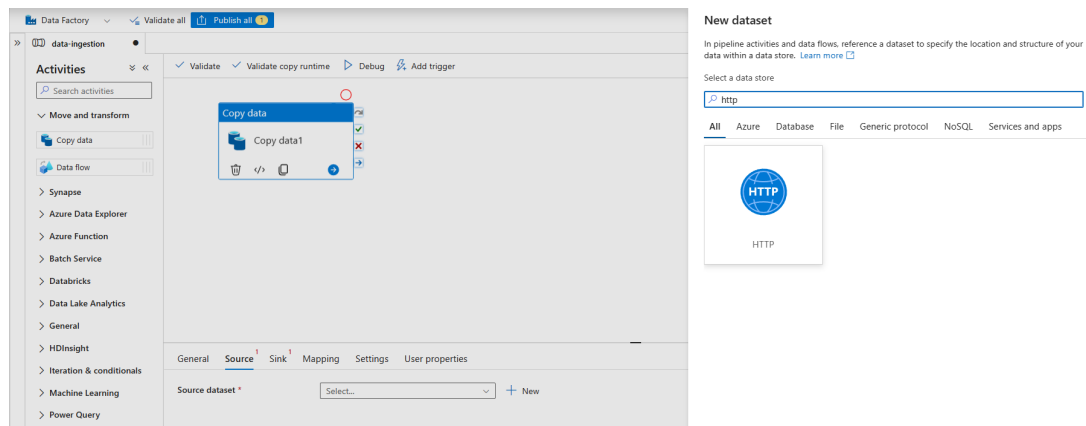
Ingresamos a Data Factory Studio.



Vamos a la sección Author y creamos un nuevo pipeline.



Agregamos un bloque de Copy data. En la sección *source* seleccionamos HTTP como fuente. Luego, en formato, elegimos DelimitedText, ya que nuestros archivos son .csv.



En la configuración, ingresamos la url del dataset y una autenticación anónima, ya que es información abierta al público. Finalmente queda creado este vínculo. Podemos ver la data con la opción *Preview data*.

New linked service

HTTP [Learn more](#)

Name *

AthletesHTTP

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Base URL *

https://raw.githubusercontent.com/darshilparmar/tokyo-olympic-azure-data-engineering-prc

Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server Certificate Validation

☒ Enable ☐ Disable

Authentication type *

Anonymous

Ahora, para cargar nuestros datos en la storage account ya creada, vamos a la sección sink y seleccionamos Azure Data Lake Storage Gen2. Luego, en formato, elegimos DelimitedText.

New linked service

Azure Data Lake Storage Gen2 [Learn more](#)

Name *

AzureDataLakeStorage1

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication type

Account key

Account selection method

☒ From Azure subscription ☐ Enter manually

Azure subscription

Azure subscription 1 (68d2a34a-fc41-49b2-89b2-dc68dbed9512)

Storage account name *

tokyoolympicdatastorage1

Set properties

Name

ADLS

Linked service *

AzureDataLakeStorage1

File path

tokyo-olympic-data

/ raw-data

/ athletes.csv

First row as header



Import schema



From connection/store



From sample file



None

Validamos nuestro pipeline.

Hacemos el debug y verificamos que todo se ejecute correctamente.

Name	Modified	Access tier	Archive status
athletes.csv	9/22/2023, 10:52:42 AM	Hot (Inferred)	

Realizamos el mismo proceso para los demás datasets, reutilizando el sink de Azure Data Lake Storage Gen2 para todos.

Microsoft Azure | Data Factory | tokyo-olympic-data-fact

Activities: Copy data (Athletes, Coaches, EntriesGender, Medals, Teams)

Pipeline run ID: a5d1b4a9-825d-4753-bac0-d2df773011f8

Pipeline status: Succeeded

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID	Log
Teams	Succeeded	Copy data	9/22/2023, 11:23:06 AM	11s	AutoResolveIntegration		a981cb3-5f0d-4b9d-88cb-49811f430ae4	
Medals	Succeeded	Copy data	9/22/2023, 11:22:56 AM	9s	AutoResolveIntegration		de2e33ce-b32f-4f87-8f99-6043732ad406	
EntriesGender	Succeeded	Copy data	9/22/2023, 11:22:43 AM	12s	AutoResolveIntegration		4309967-839e-4435-9fed-d194bf387375	
Coaches	Succeeded	Copy data	9/22/2023, 11:22:31 AM	11s	AutoResolveIntegration		22015d1d-8761-4803-b031-6415e2f86268	
Athletes	Succeeded	Copy data	9/22/2023, 11:22:18 AM	12s	AutoResolveIntegration		655ae8a7-6eda-4367-bb7a-a0b4ee5f698	

- Name**
- ☐ athletes.csv
 - ☐ coaches.csv
 - ☐ entries_gender.csv
 - ☐ medals.csv
 - ☐ teams.csv

Continuando con la arquitectura, ahora creamos un servicio Azure Databricks.

Microsoft Azure | tokyo-olympic-tokyo-olympic-db | Overview

tokyo-olympic-db
Azure Databricks Service

Overview

Status: Active

Resource group: tokyo-olympic

Location: East US

Subscription: Azure subscription 1

Subscription ID: 68d2a34a-fc41-49b2-89b2-dc68dbed9512

Managed Resource Group: databricks-rg-tokyo-olympic-db-ga4yuzvyaef4g

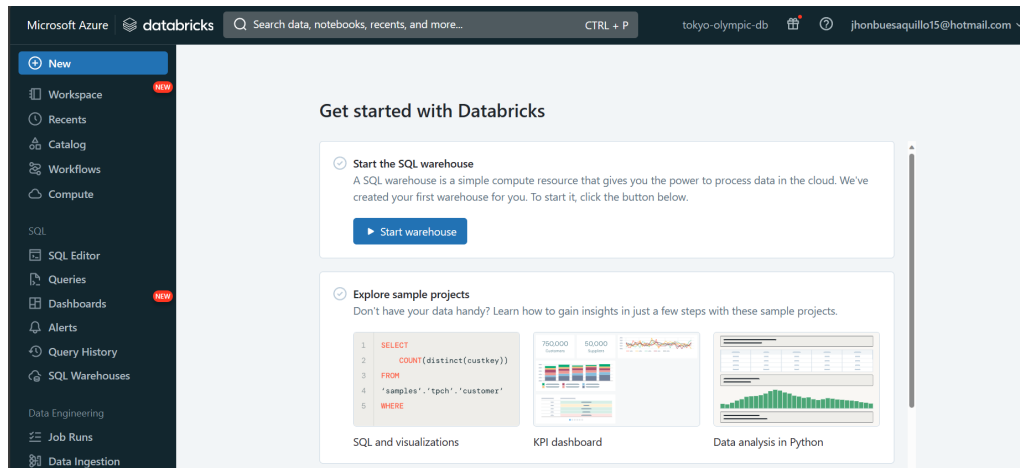
URL: https://adb-4161556213235154.14.azure-databricks.net

Pricing Tier: Premium (+ Role-based access control) (Click to change)

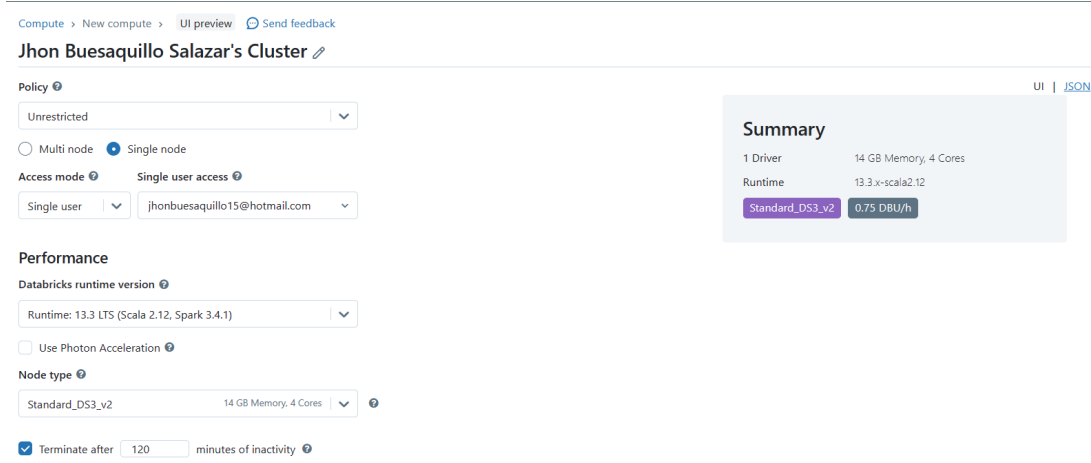
Launch Workspace

Documentation, Getting Started, Import Data from File, Import Data from Azure Storage, Notebook, Admin Guide, Link Azure ML workspace

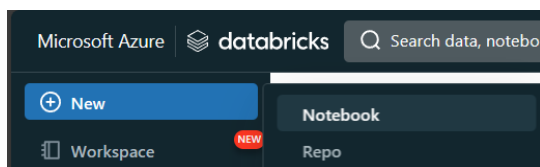
Lanzamos el workspace.



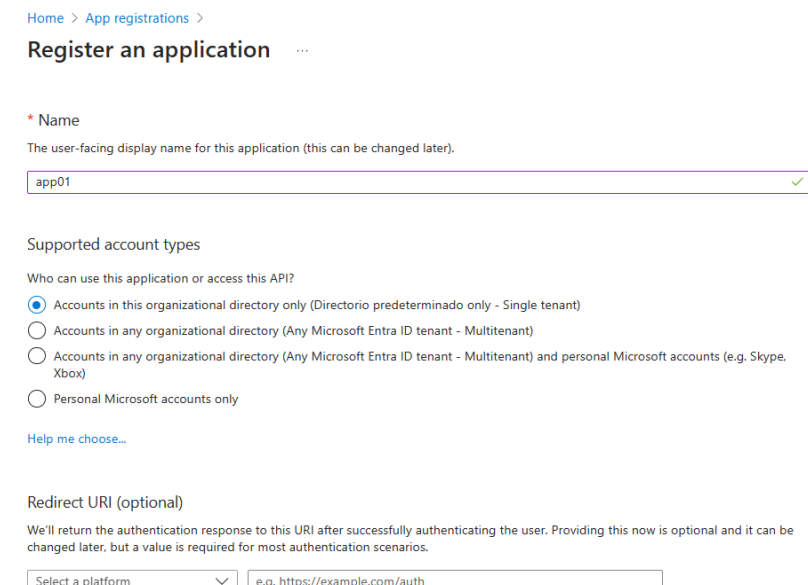
Creamos una instancia Compute para poder correr Spark.



Luego, creamos un nuevo notebook para desarrollar nuestro código Spark.



Creamos un registro de app para poder hacer una conexión entre Azure Databricks y ADLS.



Creamos un nuevo secret.

Home > App registrations > app01

app01 | Certificates & secrets

Search

Got feedback?

Overview

Quickstart

Integration assistant

Manage

Branding & properties

Authentication

Certificates & secrets

Token configuration

API permissions

Credentials enable confidential applications to identify themselves to the authentication service when receiving tokens at a web addressable location (using an HTTPS scheme). For a higher level of assurance, we recommend using a certificate (instead of a client secret) as a credential.

Certificates (0)

Client secrets (1)

Federated credentials (0)

A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.

+ New client secret

Description	Expires	Value	Secret ID
secretkey	3/20/2024	nuV8Q~Nr76Kl06nM_yXu9zmYMY-DJj...	f974c3f9-ad8e-4410-9a1d-918e0043ff79

Tomamos el Client ID y Tenant ID del registro, como también el Value del secret para crear la conexión.

```
1  configs = {"fs.azure.account.auth.type": "OAuth",
2  "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.
  ClientCredsTokenProvider",
3  "fs.azure.account.oauth2.client.id": "6ea29bf7-ef62-4955-a342-1cdbae07f74",
4  "fs.azure.account.oauth2.client.secret":
  'nuV8Q~Nr76Kl06nM_yXu9zmYMY-DJjmbHI6ebbh',
5  "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/
  22b59d6a-4ebf-4603-a1c8-65dfc64b4f04/oauth2/token"}
6
```

Estos valores pueden ser protegidos usando un Key Vault de Azure, para tener en cuenta en un ambiente empresarial.

La conexión ahora nos exige crear un rol en el container.

Home > Storage accounts > tokyoolympicdatastorage1 | Containers > tokyo-olympic-data

tokyo-olympic-data | Access Control (IAM)

Search

+ Add

Download role assignments

Edit columns

Refresh

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

View my access

Add role assignment

Add co-administrator

My access

View my level of access to this resource.

Asignamos la configuración correspondiente.

Add role assignment

Role

Members

Review + assign

Role

Storage Account Contributor

Scope

/subscriptions/68d2a34a-fc41-49b2-89b2-dc68dbed9512/resourceGroups/tokyo-olympic/providers/Microsoft.Storage/storageAccounts/tokyoolympicdatastorage1/blobServices/default/containers/tokyo-olympic-data

Members

Name	Object ID	Type
app01	41347d2f-43a6-476a-b3ba-b22ee0099307	App

Description

No description






▼ Storage Blob Data Contributor

<input type="checkbox"/>	 app01	App	Storage Blob Data Contributor ⓘ
--------------------------	---	-----	---------------------------------

Realizamos la transformación de los datos con Pyspark. Una vez hecho esto, enviamos estos archivos al directorio transformed-data de nuestro container.

Location: [tokyo-olympic-data](#) / [transformed-data](#) / athletes






Search blobs by prefix (case-sensitive)

	Name	Modified	Access tier
<input type="checkbox"/>	 [..]		
<input type="checkbox"/>	 _committed_877905915125892948	9/23/2023, 10:39:22 ...	Hot (Inferred)
<input type="checkbox"/>	 _started_877905915125892948	9/23/2023, 10:39:22 ...	Hot (Inferred)
<input type="checkbox"/>	 _SUCCESS	9/23/2023, 10:39:22 ...	Hot (Inferred)
<input type="checkbox"/>	 part-00000-tid-877905915125892948-a881b9a0-c39b-4a22-92dc-412257dea06...	9/23/2023, 10:39:22 ...	Hot (Inferred)

Location: [tokyo-olympic-data](#) / transformed-data

Search blobs by prefix (case-sensitive)

☐ Show deleted objects

	Name	Modified
<input type="checkbox"/>	 athletes	
<input type="checkbox"/>	 coaches	
<input type="checkbox"/>	 entries_gender	
<input type="checkbox"/>	 medals	
<input type="checkbox"/>	 teams	

Como siguiente paso, usaremos los datos transformados en **Azure Synapse Analytics**.

Create Synapse workspace ...

Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.

Subscription * ⓘ ⓘ The Synapse and SQL resource providers are now registered with this subscription.

Resource group * ⓘ [Create new](#)

Managed resource group ⓘ

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * ✓

Region *

Select Data Lake Storage Gen2 * ⓘ ☒ From subscription ☐ Manually via URL

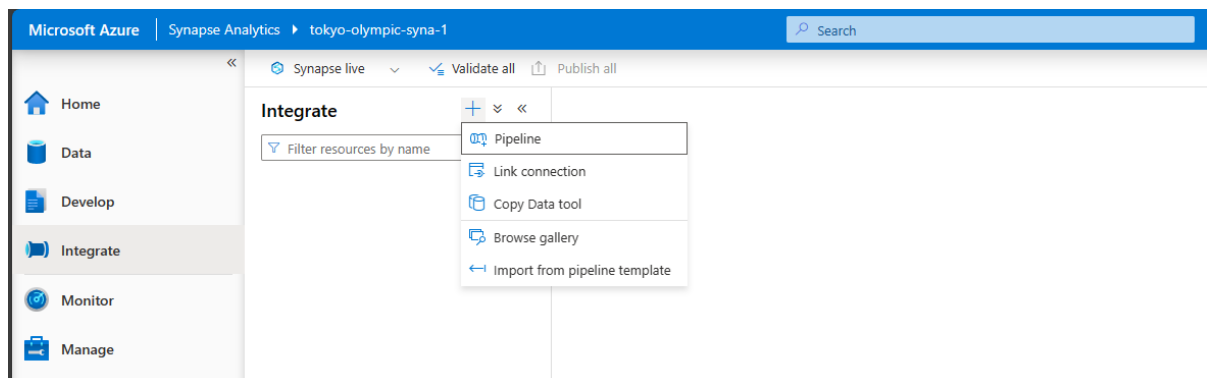
Account name * ⓘ [Create new](#)

File system name * [Create new](#)

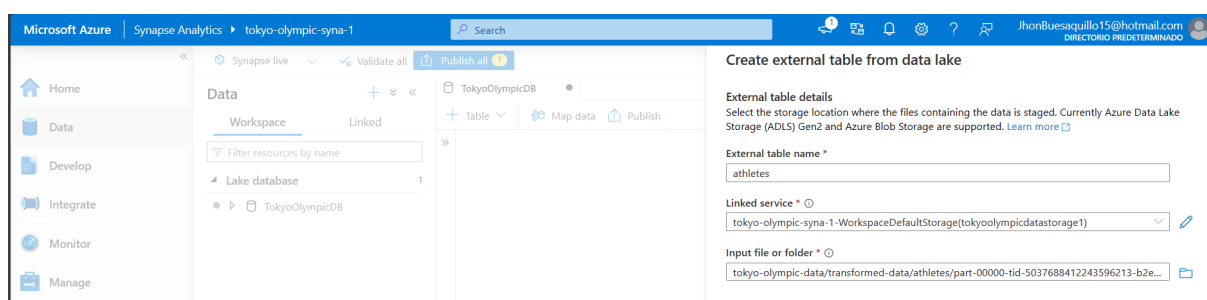
☒ Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.

[Review + create](#) [< Previous](#) [Next: Security >](#)

Nos damos cuenta que el servicio Synapse nos ofrece todo lo que hicimos anteriormente. En un caso real se podría usar solo este servicio.



Creamos un nuevo workspace de Data (lake database). Vamos agregando las tablas provenientes de transformed-data



Microsoft Azure | Synapse Analytics | tokyo-olympic-syna-1

Synapse live Validate all Publish all

Home Data Develop Integrate Monitor Manage

Data Workspace Linked

Filter resources by name

Lake database 1

TokyoOlympicsDB

Tables

Filter by keyword

Others 1

athletes

abc: PersonName

abc: Country

abc: Discipline

See less ^

General Columns Relationships

Filter by keyword + Column Clone Convert type Delete

<input type="checkbox"/>	Name	Keys	Description	Nullability	Data type	Format / Length
Standard column (3)						
<input type="checkbox"/>	PersonName	<input type="checkbox"/> PK		<input checked="" type="checkbox"/> Null	abc: string	8000
<input type="checkbox"/>	Country	<input type="checkbox"/> PK		<input checked="" type="checkbox"/> Null	abc: string	8000
<input type="checkbox"/>	Discipline	<input type="checkbox"/> PK		<input checked="" type="checkbox"/> Null	abc: string	8000
Partition column (0)						

Synapse live Validate all Publish all

Data Workspace Linked

Filter resources by name

Lake database 1

TokyoOlympicsDB

Tables

athletes

SQL script 1

Run Undo Publish Query plan Connect to Built-in Use database TokyoOlympicsDB

```
1 SELECT TOP (100) [PersonName]
2 ,[Country]
3 ,[Discipline]
4 FROM [TokyoOlympicsDB].[dbo].[athletes]
```

New SQL script Select TOP 100 rows

New notebook Chart Export results

Machine Learning

PersonName	Country	Discipline
AALERUD Katrine	Norway	Cycling Road
ABAD Nestor	Spain	Artistic Gymnastics
ABAGNALE Giovanni	Italy	Rowing
ABALDE Alberto	Spain	Basketball
ABALDE Tamara	Spain	Basketball

Properties

General Related (0)

Name * SQL script 1

Description

Type .sql script

Size 95 bytes

Results settings per query

☒ First 5000 rows (default)

☐ All rows

Synapse live Validate all Publish all

Data Workspace Linked

Filter resources by name

Lake database 1

TokyoOlympicsDB

Tables

athletes

abc: PersonName

abc: Country

abc: Discipline

See less ^

coaches

abc: Name

abc: Country

abc: Discipline

abc: Event

See less ^

entries_gender

abc: Discipline

123: Female

123: Male

123: Total

See less ^

medals

123: Rank

abc: Team_Country

123: Gold

123: Silver

123: Bronze

123: Total

123: Rank_by_Total

See less ^

teams

abc: TeamName

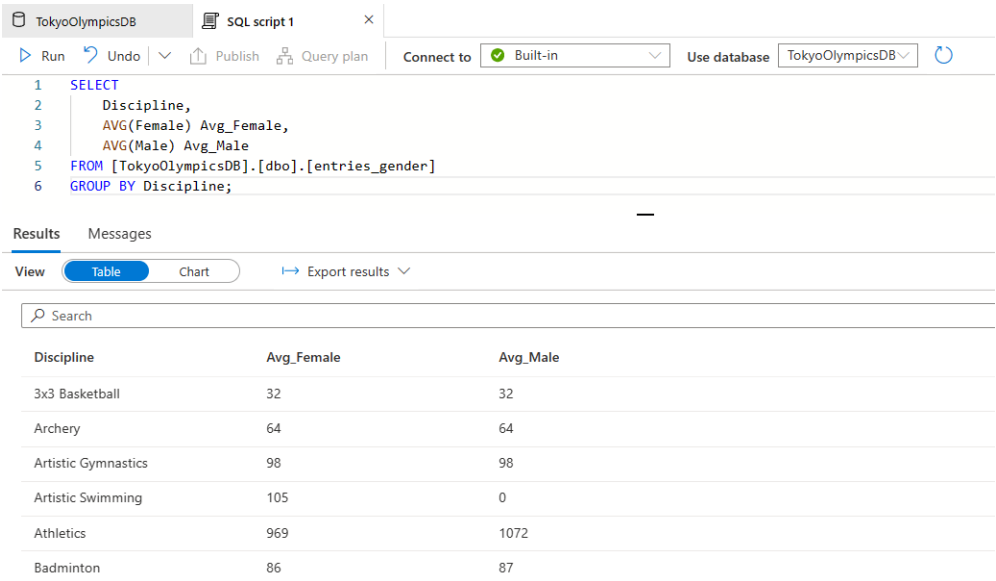
abc: Discipline

abc: Country

abc: Event

See less ^

Usamos el script para probar algunos queries.



También nos ofrece gráficas para visualizar mejor los datos.

