

Data Engineering with Azure

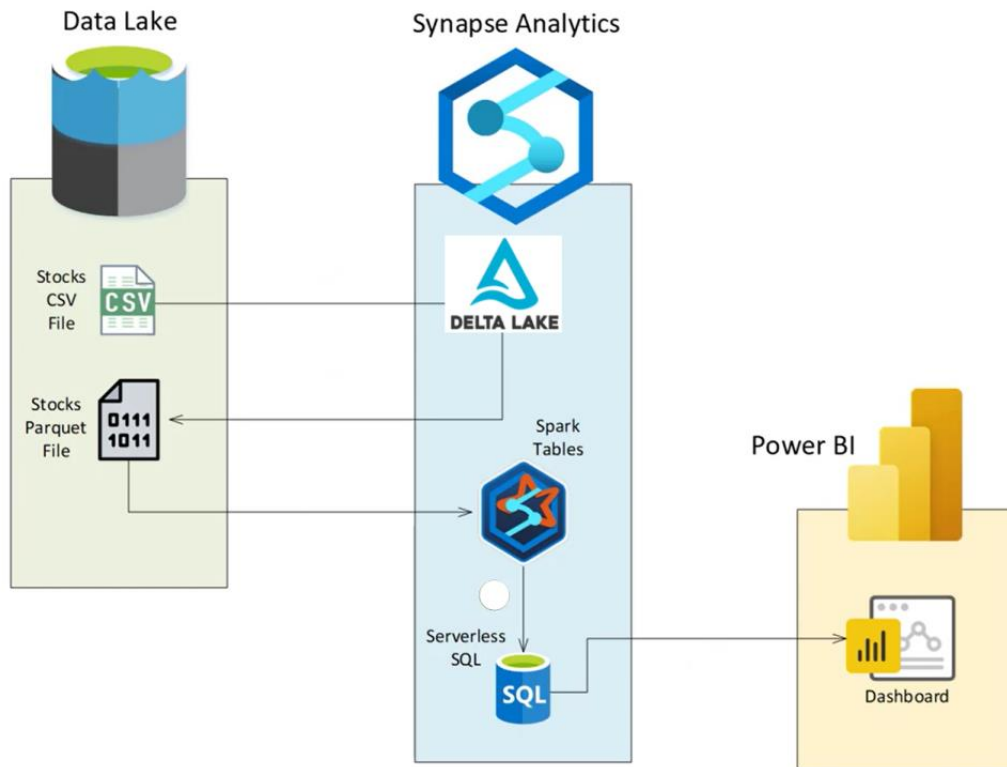
Proyecto Final, Caso: Retail.SA

Estudiante: Gael Velasquez

Docente: Kremlin Huaman

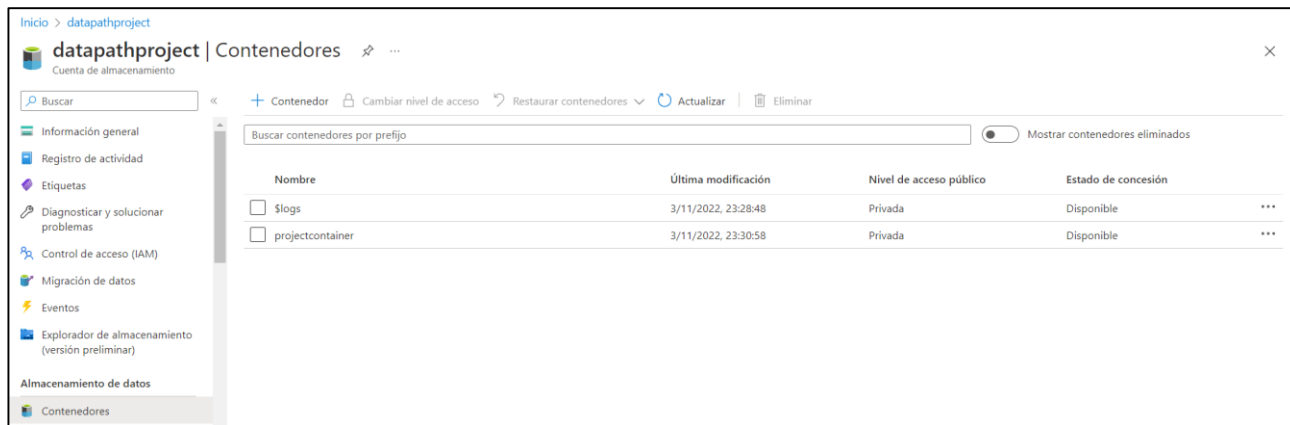
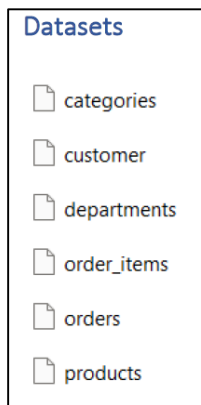
Noviembre, 2022

Arquitectura en Azure



1. ADLS

La 1er etapa es la carga de los archivos a un Container (Data Lake)



2. Estructura Delta Lake

Posteriormente se propone la estructura Delta Lake en el Contenedor

Inicio > datapathproject | Contenedores >

projectcontainer

Contenedor

⌕ Buscar

«

⬆ Cargar

⊕ Agregar directorio

🔄 Actualizar

|

🔄 Cambiar nombre

🗑 Eliminar

↔ Cambiar nivel

🔑 Adquirir concesión

🔑 Interrumpir concesión

📄 Información general

🔧 Diagnosticar y solucionar problemas

🔑 Control de acceso (IAM)

Configuración

🔑 Tokens de acceso compartido

🔑 Administrar dirección ACL

🔑 Directiva de acceso

📄 Propiedades

📄 Metadatos

Método de autenticación: Clave de acceso [\(Cambiar a la cuenta de usuario de Azure AD\)](#)

Ubicación: projectcontainer

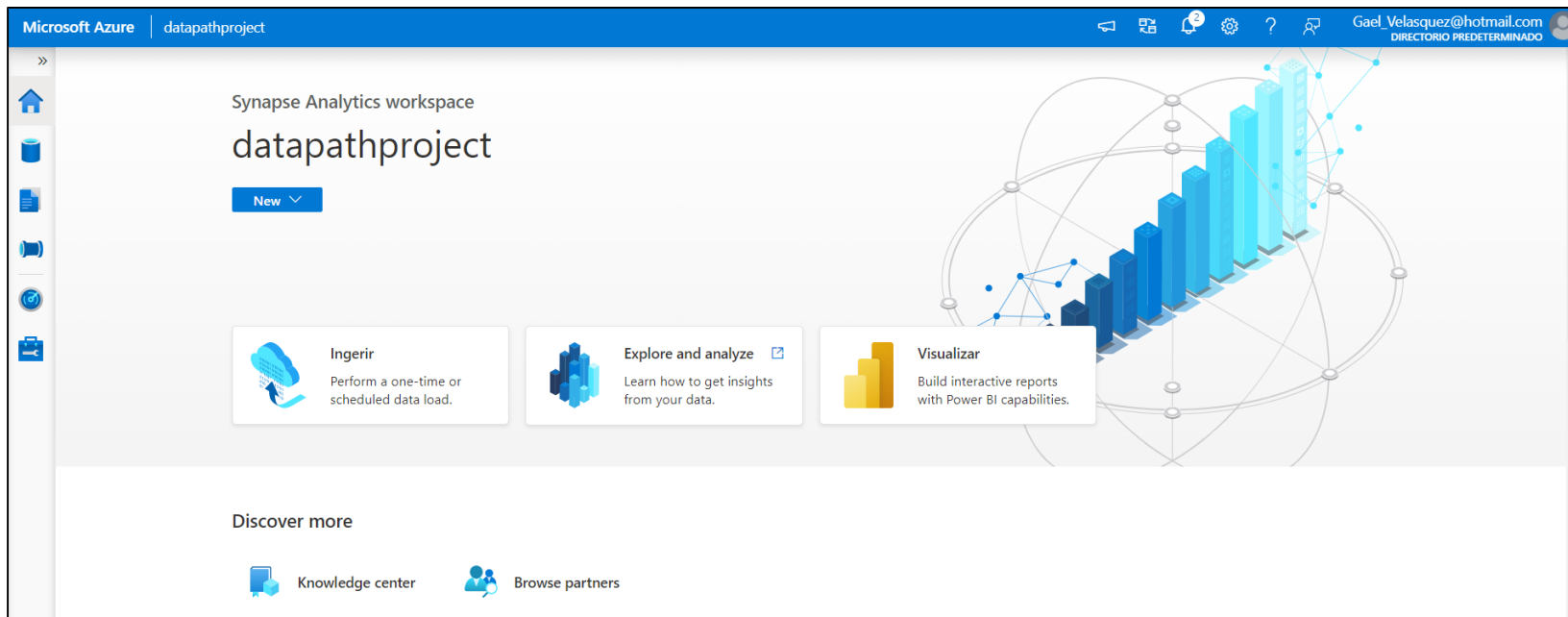
🔍 Buscar blobs por prefijo (distingue mayúsculas de minúsculas)

☐ Mostrar objetos eliminados

Nombre	Modificado	Nivel de acceso	Estado del archivo	Tipo de blob	Tamaño	Estado de concesión
<input type="checkbox"/> 📁 bronze						- ...
<input type="checkbox"/> 📁 gold						- ...
<input type="checkbox"/> 📁 silver						- ...
<input type="checkbox"/> 📁 synapse						- ...

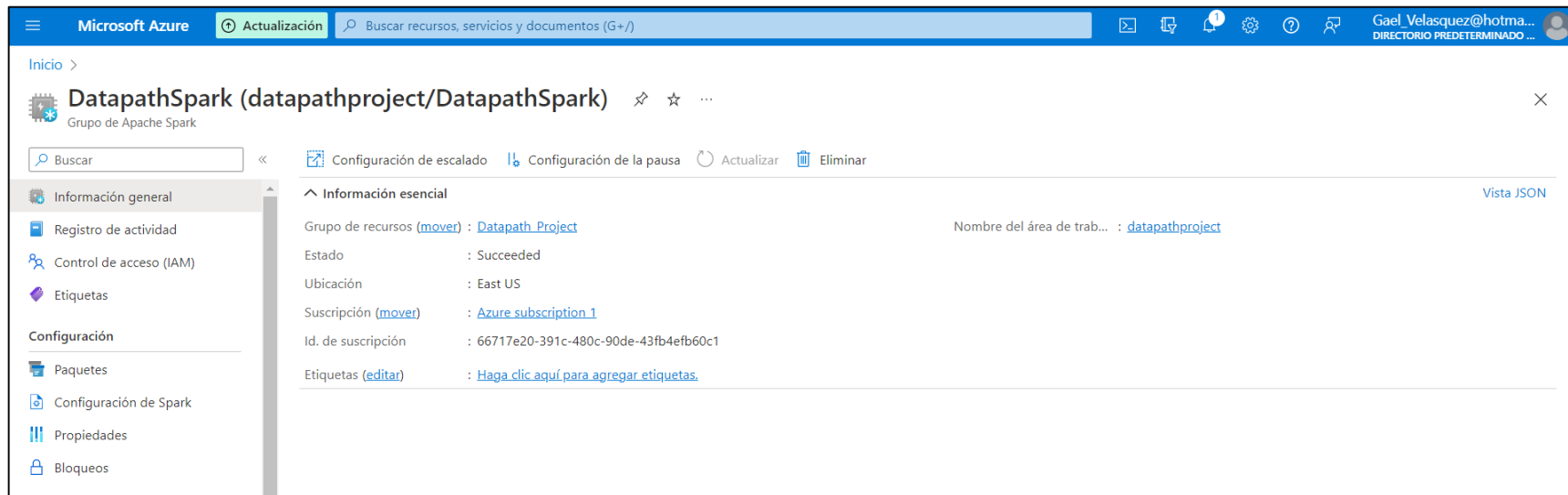
3. Workspace Synapse

Habilitar el espacio de trabajo en Azure Synapse, con los permisos respectivos



4. Spark Pool

Habilitar el el Spark Pool para el Procesamiento



The screenshot displays the Microsoft Azure portal interface. At the top, the header includes the 'Microsoft Azure' logo, a green 'Actualización' (Update) button, and a search bar. The user's profile 'Gael Velasquez@hotmail...' is visible in the top right corner. The main content area shows the 'DatapathSpark (datapathproject/DatapathSpark)' resource page. The left sidebar contains navigation links for 'Información general', 'Registro de actividad', 'Control de acceso (IAM)', 'Etiquetas', 'Configuración', 'Paquetes', 'Configuración de Spark', 'Propiedades', and 'Bloqueos'. The main panel displays the 'Información esencial' (Essential information) section, which includes a search bar, action buttons for 'Configuración de escalado', 'Configuración de la pausa', 'Actualizar', and 'Eliminar', and a 'Vista JSON' link. The resource details are as follows:

Propiedad	Valor
Grupo de recursos (mover)	Datapath Project
Nombre del área de trab...	datapathproject
Estado	Succeeded
Ubicación	East US
Suscripción (mover)	Azure subscription 1
Id. de suscripción	66717e20-391c-480c-90de-43fb4efb60c1
Etiquetas (editar)	Haga clic aquí para agregar etiquetas.

5. Definir el Proceso ETL

Generar los Notebooks necesarios para el Procesamiento y armar el Pipeline con el Flujo Logico del Proyecto

Data Engineer with Azure

Datapath Project

Author: Gael Velasquez

+ Code

+ Markdown

1 spark

✓ - Command executed in 188 ms on 7:55:42 PM, 11/04/22

SparkSession - hive

SparkContext

[Spark UI](#)

Versionv3.2.2.5.0-73283859

Master yarn

AppNameProj_Extract_DatapathSpark_1667689560

EXTRACT DATA

Loading...

1 from pyspark.sql.types import StructField, StructType, StringType, LongType, IntegerType, FloatType

2 from pyspark.sql.functions import substring, col, when

✓ - Command executed in 180 ms on 7:55:42 PM, 11/04/22

Extract

Transform

1 bronze_path = 'abfss://projectcontainer@datapathproject.dfs.core.windows.net/bronze'

2 silver_path = 'abfss://projectcontainer@datapathproject.dfs.core.windows.net/silver'

3 gold_path = 'abfss://projectcontainer@datapathproject.dfs.core.windows.net/gold'

✓ - Command executed in 192 ms on 7:58:42 PM, 11/04/22

Reading Silver Layer

1 ddepartments = spark.read.format("delta").load(f"{silver_path}/departments")

2 dcategories = spark.read.format("delta").load(f"{silver_path}/categories")

3 dproducts = spark.read.format("delta").load(f"{silver_path}/products")

4 dorder_items = spark.read.format("delta").load(f"{silver_path}/order_items")

5 dorders = spark.read.format("delta").load(f"{silver_path}/orders")

6 dcustomers = spark.read.format("delta").load(f"{silver_path}/customers")

✓ - Command executed in 42 sec 12 ms on 7:59:25 PM, 11/04/22

1 ddepartments.createOrReplaceTempView("departments")

2 dcategories.createOrReplaceTempView("categories")

3 dproducts.createOrReplaceTempView("products")

4 dorder_items.createOrReplaceTempView("order_items")

5 dorders.createOrReplaceTempView("orders")

6 dcustomers.createOrReplaceTempView("customers")

✓ - Command executed in 17 sec 60 ms on 7:59:42 PM, 11/04/22

Making Transformations

Reading Gold Layer

1 R1 = spark.read.format("delta").load(f"{gold_path}/R1")

2 R2 = spark.read.format("delta").load(f"{gold_path}/R2")

3 R31 = spark.read.format("delta").load(f"{gold_path}/R31")

4 R32 = spark.read.format("delta").load(f"{gold_path}/R32")

5 R33 = spark.read.format("delta").load(f"{gold_path}/R33")

6 R4 = spark.read.format("delta").load(f"{gold_path}/R4")

7 R5 = spark.read.format("delta").load(f"{gold_path}/R5")

✓ - Command executed in 43 sec 810 ms on 8:02:54 PM, 11/04/22

Load Parquet Format

1 R1.write.mode("overwrite").parquet(f"{gold_path}/parquet/R1")

2 R2.write.mode("overwrite").parquet(f"{gold_path}/parquet/R2")

3 R31.write.mode("overwrite").parquet(f"{gold_path}/parquet/R31")

4 R32.write.mode("overwrite").parquet(f"{gold_path}/parquet/R32")

5 R33.write.mode("overwrite").parquet(f"{gold_path}/parquet/R33")

6 R4.write.mode("overwrite").parquet(f"{gold_path}/parquet/R4")

7 R5.write.mode("overwrite").parquet(f"{gold_path}/parquet/R5")

✓ - Command executed in 14 sec 956 ms on 8:06:32 PM, 11/04/22

Load

5. Definir el Proceso ETL

Microsoft Azure | datapathproject

Buscar

Synapse Live Validar todo Publicar todo

Integrar

Filtrar recursos por nombre

Pipelines 1

Datapath_Project

Datapath_Project

Validar Depurar Agregar desencadenador

Bloc de notas

Extract Transform Load

Expandir el panel del cuadro de herramientas

Parámetros Variables Configuración Salida

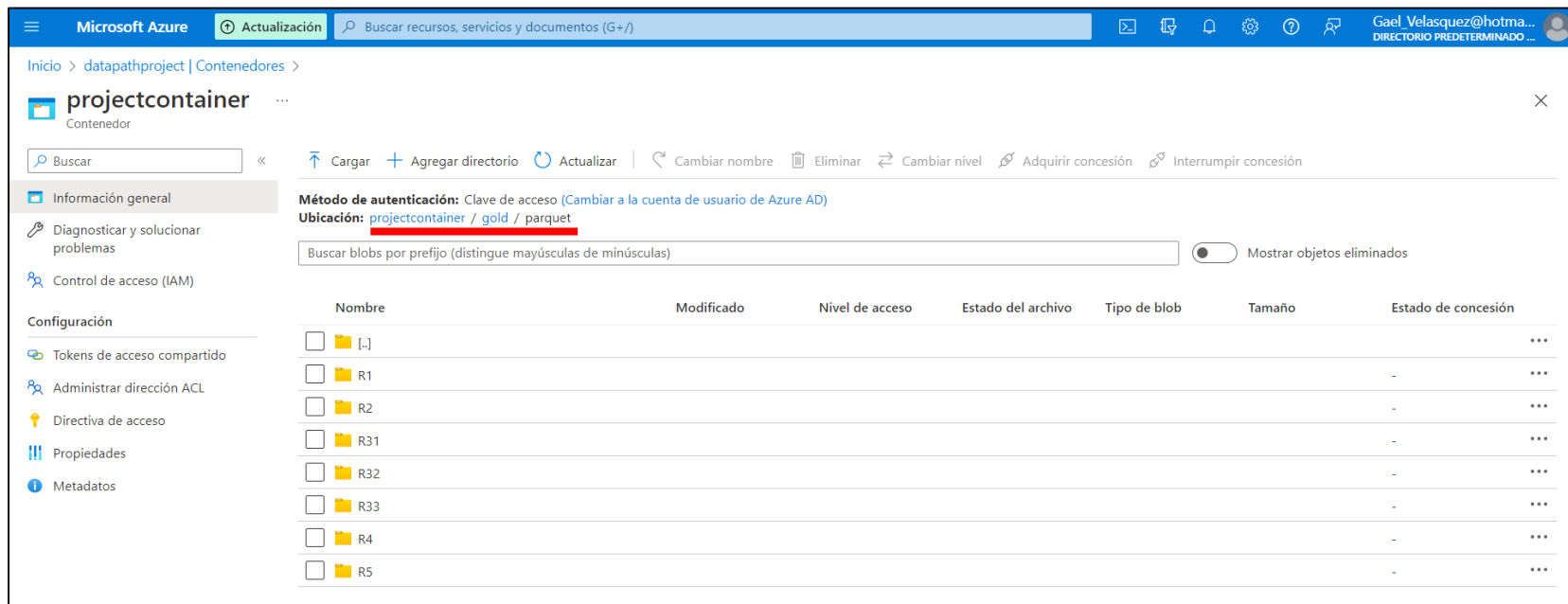
Id. de ejecución de canalización: **b800e833-e322-45e3-b749-c4ba8e32a6d9**

Ver consumo de la ejecución de depuración

Tipo	Inicio de la ejecución	Duración	Estado	Entorno de ejecución de in	Id. de ejecución
Bloc de notas	2022-11-05T01:13:24.1924	00:02:07	✓ Correcto	AutoResolveIntegrationRu	e2b59f5f-3d22-4412-bf48-:
Bloc de notas	2022-11-05T01:15:32.2010	00:02:53	✓ Correcto	AutoResolveIntegrationRu	ca348558-a797-445f-85a4-.
Bloc de notas	2022-11-05T01:18:27.4951	00:02:06	✓ Correcto	AutoResolveIntegrationRu	2d7c164c-5264-48c4-987e-

6. Generación de la capa Gold del Proceso ETL

Con el Pipeline ejecutado, se obtiene los archivos delta y parquet esperados, en la capa Gold



The screenshot shows the Microsoft Azure portal interface for a storage account named 'projectcontainer'. The 'Ubicación' (Location) is highlighted as 'projectcontainer / gold / parquet'. The table below lists the contents of the storage account.

Nombre	Modificado	Nivel de acceso	Estado del archivo	Tipo de blob	Tamaño	Estado de concesión
<input type="checkbox"/> [.]						...
<input type="checkbox"/> R1					-	...
<input type="checkbox"/> R2					-	...
<input type="checkbox"/> R31					-	...
<input type="checkbox"/> R32					-	...
<input type="checkbox"/> R33					-	...
<input type="checkbox"/> R4					-	...
<input type="checkbox"/> R5					-	...

6. Generación de la capa Gold del Proceso ETL

Inicio > datapathproject | Contenedores >

projectcontainer ...

Contenedor

Buscar

Cargar + Agregar directorio Actualizar Cambiar nombre Eliminar Cambiar nivel Adquirir concesión Interrumpir concesión

Información general

Diagnosticar y solucionar problemas

Control de acceso (IAM)

Configuración

Tokens de acceso compartido

Administrar dirección ACL

Directiva de acceso

Propiedades

Metadatos

Método de autenticación: Clave de acceso (Cambiar a la cuenta de usuario de Azure AD)

Ubicación: projectcontainer / gold / parquet / R1

Buscar blobs por prefijo (distingue mayúsculas de minúsculas) ☐ Mostrar objetos eliminados

Nombre	Modificado	Nivel de acceso	Estado del archivo	Tipo de blob	Tamaño	Estado de concesión
<input type="checkbox"/> [.]						...
<input type="checkbox"/> _SUCCESS	4/11/2022, 21:20:08	Frecuente (inferido)		Blob en bloques	0 B	Disponible ...
<input type="checkbox"/> part-00000-949fe972-d4b8-436b-a574-2e8a57339...	4/11/2022, 21:20:05	Frecuente (inferido)		Blob en bloques	49.73 KiB	Disponible ...

Microsoft Azure Actualización Buscar recursos, servicios y documentos (G+J)

Inicio > datapathproject | Contenedores >

projectcontainer ...

Contenedor

Buscar

Cargar + Agregar directorio Actualizar Cambiar nombre Eliminar Cambiar nivel Adquirir concesión Interrumpir concesión

Información general

Diagnosticar y solucionar problemas

Control de acceso (IAM)

Configuración

Tokens de acceso compartido

Administrar dirección ACL

Directiva de acceso

Propiedades

Metadatos

Método de autenticación: Clave de acceso (Cambiar a la cuenta de usuario de Azure AD)

Ubicación: projectcontainer / gold / R1

Buscar blobs por prefijo (distingue mayúsculas de minúsculas) ☒ Mostrar objetos eliminados

Nombre	Modificado	Nivel de acceso	Estado del archivo	Tipo de blob	Tamaño	Estado de concesión
<input type="checkbox"/> [.]						...
<input type="checkbox"/> _delta_log						...
<input type="checkbox"/> part-00000-c134740c-e98d-44fa-9eec-09cc49a3ee...	4/11/2022, 21:17:28	Frecuente (inferido)		Blob en bloques	49.73 KiB	Disponible ...

7. Generacion del SQL Pool

Posteriormente se define una base de datos con tablas Spark correspondientes a cada Resultado

The screenshot shows the Microsoft Azure Data Studio interface for a project named 'datapathproject'. The left sidebar displays the 'Data' section with a 'Workspace' and a 'Lake database'. The 'Project_Database' is expanded, showing a list of tables: Db_R1, Db_R2, Db_R3, Db_R31, Db_R32, Db_R33, Db_R4, and Db_R5. Each table is associated with a specific Spark result. The 'Properties' panel on the right shows the configuration for the 'Project_Database', including the linked service 'datapathproject-WorkspaceDefault...', the input folder 'projectcontainer/Project_Datab...', and the data format 'Delimited Text'.

Table Name	Columns
Db_R1	customer_full_name, Total_Orders
Db_R2	year_order, month_order, category_name, Total_Quantity
Db_R3	day_order, Income_Mean
Db_R31	year_order, Income_Mean
Db_R32	month_order, Income_Mean
Db_R33	day_order, Income_Mean
Db_R4	order_id, order_status, Total_Income
Db_R5	year_order, quarter_order, category_name, Total_Income

7. Generacion del SQL Pool

Posteriormente se define una base de datos con tablas Spark correspondientes a cada Resultado

The screenshot shows the Microsoft Azure Data Studio interface for a project named 'datapathproject'. The left sidebar displays the 'Data' section with a 'Workspace' and a 'Lake database'. The 'Project_Database' is expanded, showing a list of tables: Db_R1, Db_R2, Db_R3, Db_R31, Db_R32, Db_R33, Db_R4, and Db_R5. Each table is associated with a specific Spark result. The 'Properties' panel on the right shows the configuration for the 'Project_Database', including the linked service 'datapathproject-WorkspaceDefault...', the input folder 'projectcontainer/Project_Datab...', and the data format 'Delimited Text'.

Table Name	Columns
Db_R1	customer_full_name, Total_Orders
Db_R2	year_order, month_order, category_name, Total_Quantity
Db_R3	day_order, Income_Mean
Db_R31	year_order, Income_Mean
Db_R32	month_order, Income_Mean
Db_R33	day_order, Income_Mean
Db_R4	order_id, order_status, Total_Income
Db_R5	year_order, quarter_order, category_name, Total_Income

8. Generacion de las Consultas

Es posible obtener las consultas desde cada una de las tablas generadas en la BD

The screenshot displays the Microsoft Azure Data Studio interface. The top bar shows the 'Microsoft Azure' logo and the project name 'datapathproject'. A search bar is present. The left sidebar contains a 'Develop' section with a search filter and a list of 'Scripts SQL' (SQL script 1 through 7) and 'Cuadernos' (Proj_Extract, Proj_Load, Proj_Transform). The main editor area shows 'SQL script 1' with the following SQL code:

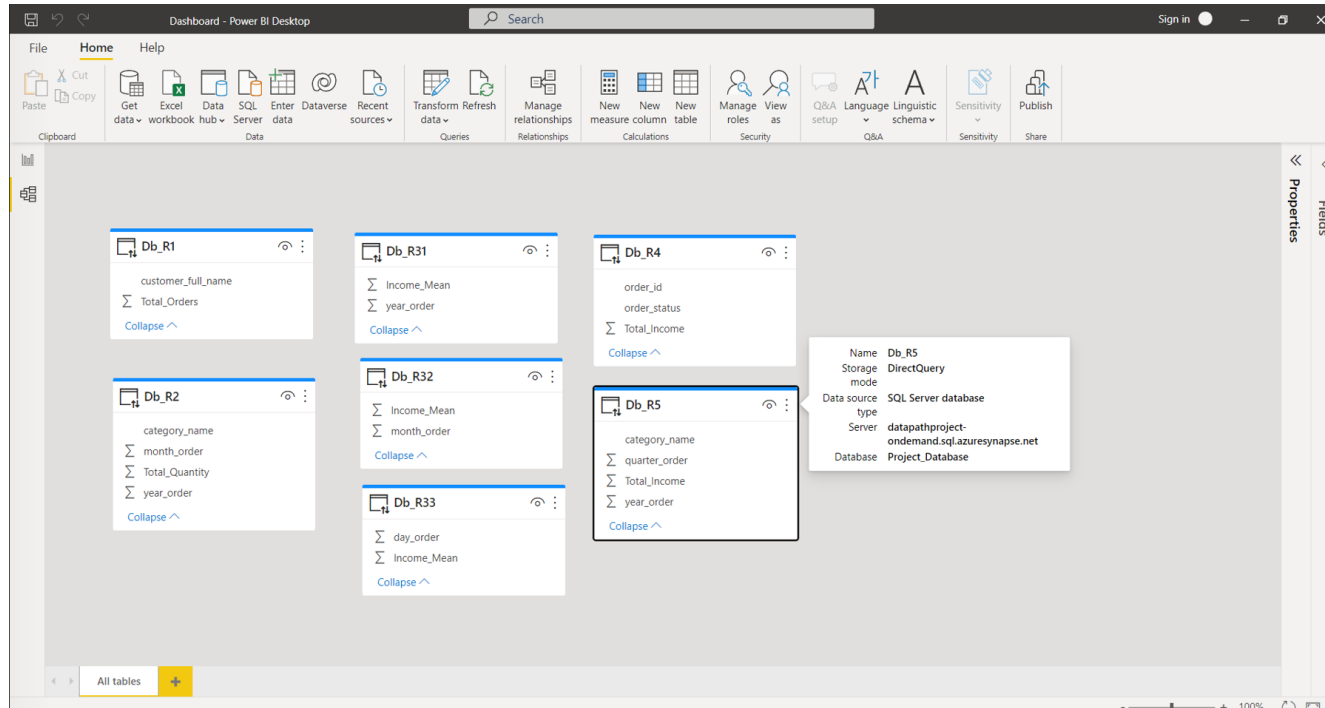
```
1 SELECT TOP (10) [customer_full_name]
2 , [Total_Orders]
3 FROM [Project_Database].[dbo].[Db_R1]
```

Below the script, the 'Results' tab is active, showing a table with two columns: 'customer_full_name' and 'Total_Orders'. The table contains 10 rows of data.

customer_full_name	Total_Orders
Mary Smith	9511
Robert Smith	291
James Smith	289
David Smith	278
John Smith	253
William Smith	214
Mary Jones	189
Michael Smith	172
Elizabeth Smith	170

9. Conexión con Power BI

Como ultimo paso, se procede a conectar las tablas generadas al servicio de Power BI para Visualización



10. Visualización / Dashboard

Ordenes realizadas agrupadas por customer

Mary Smith

9511

Sum of Total_Orders

Robert Smith

291

Sum of Total_Orders

James Smith

289

Sum of Total_Orders

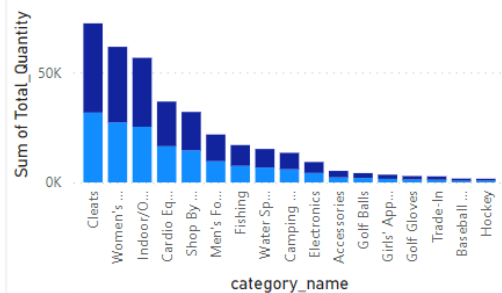
David Smith

278

Productos Vendidos

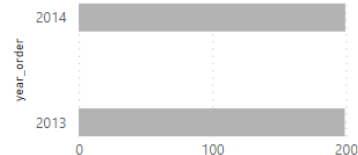
Sum of Total_Quantity by category_name and year_order

year_order ● 2013 ● 2014



Ingresos Promedio

Sum of Income_Mean by year_order



Sum of Income_Mean by month_order

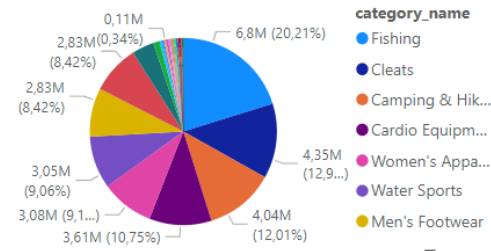


Pedidos Cancelados

10065 order_id	CANCELED order_status	1.009,95 Sum of Total_In...
10290 order_id	CANCELED order_status	1.119,93 Sum of Total_In...
10936 order_id	CANCELED order_status	1.099,93 Sum of Total_In...
11943 order_id	CANCELED order_status	1.149,90 Sum of Total_In...

Ingresos por Categoria

Sum of Total_Income by category_name



Gracias