Explore el análisis de datos en Azure con Azure Synapse Analytics

En este ejercicio, aprovisionará un área de trabajo de Azure Synapse Analytics en su suscripción de Azure y la usará para ingerir y consultar datos.

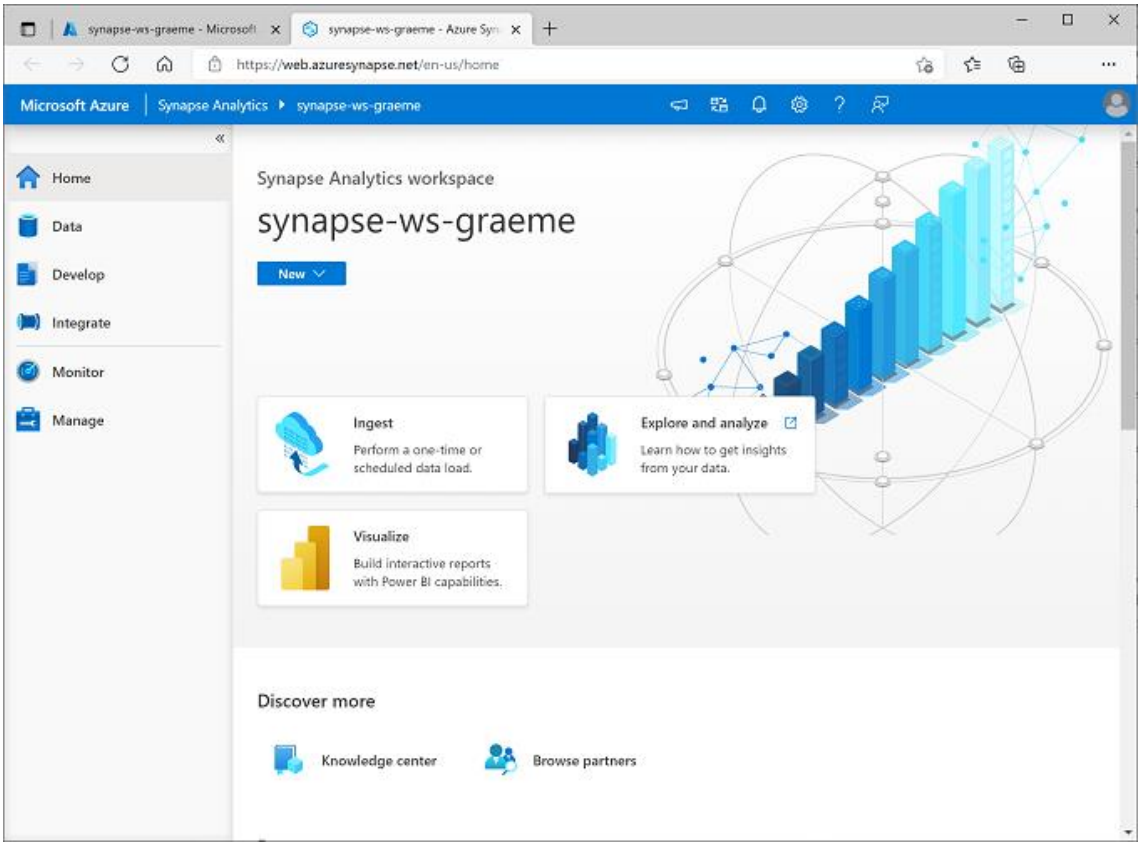
Aprovisionar un área de trabajo de Azure Synapse Analytics

Para usar Azure Synapse Analytics, debe aprovisionar un recurso de área de trabajo de Azure Synapse Analytics en su suscripción de Azure.

1. Abra el portal de Azure en https://portal.azure.com e inicie sesión con las credenciales asociadas con su suscripción de Azure.
2. En Azure Portal, en la página **de inicio** , use el ícono **+ Crear un recurso** para crear un nuevo recurso.
3. Busque *Azure Synapse Analytics* y cree un nuevo recurso **de Azure Synapse Analytics** con la siguiente configuración:
    - **Suscripción** : *Su suscripción a Azure*
        - **Grupo de recursos** : *cree un nuevo grupo de recursos con un nombre adecuado, como "synapse-rg"*.
        - **Grupo de recursos administrado** : *ingrese un nombre apropiado, por ejemplo "synapse-managed-rg"* .
    - **Nombre del espacio de trabajo** : *Ingrese un nombre de espacio de trabajo único, por ejemplo "synapse-ws-"*.
    - **Región** : *seleccione cualquiera de las siguientes regiones* :
        - Australia Este
        - Centro de EE. UU.
        - Este de EE. UU. 2
        - norte de Europa
        - Centro-sur de EE. UU.
        - El sudeste de Asia
        - Sur del Reino Unido
        - Europa occidental
        - Oeste de EE. UU.
        - OesteUS 2
    - **Seleccione Data Lake Storage Gen 2** : desde suscripción
        - **Nombre de cuenta** : *Cree una cuenta nueva con un nombre único, por ejemplo "datalake"*.

- **Nombre del sistema de archivos** : *Crea un nuevo sistema de archivos con un nombre único, por ejemplo "fs"*.

4. Cuando haya ingresado estos detalles, seleccione **Revisar + crear** y luego seleccione **Crear** para crear el espacio de trabajo.
5. Espere a que se cree el espacio de trabajo; esto puede tardar unos cinco minutos.
6. Cuando se complete la implementación, vaya al grupo de recursos que se creó y observe que contiene su área de trabajo de Synapse Analytics y una cuenta de almacenamiento de Data Lake.
7. Seleccione su espacio de trabajo de Synapse y, en su página **Descripción general** , en la tarjeta **Abrir Synapse Studio** , seleccione **Abrir** para abrir Synapse Studio en una nueva pestaña del navegador. Synapse Studio es una interfaz basada en web que puede utilizar para trabajar con su espacio de trabajo de Synapse Analytics.
8. En el lado izquierdo de Synapse Studio, use el ícono ›› para expandir el menú; esto revela las diferentes páginas dentro de Synapse Studio que usará para administrar recursos y realizar tareas de análisis de datos, como se muestra aquí:



Ingerir datos

Una de las tareas clave que puede realizar con Azure Synapse Analytics es definir *canalizaciones* que transfieran (y, si es necesario, transformen) datos de una amplia gama de orígenes a su espacio de trabajo para su análisis.

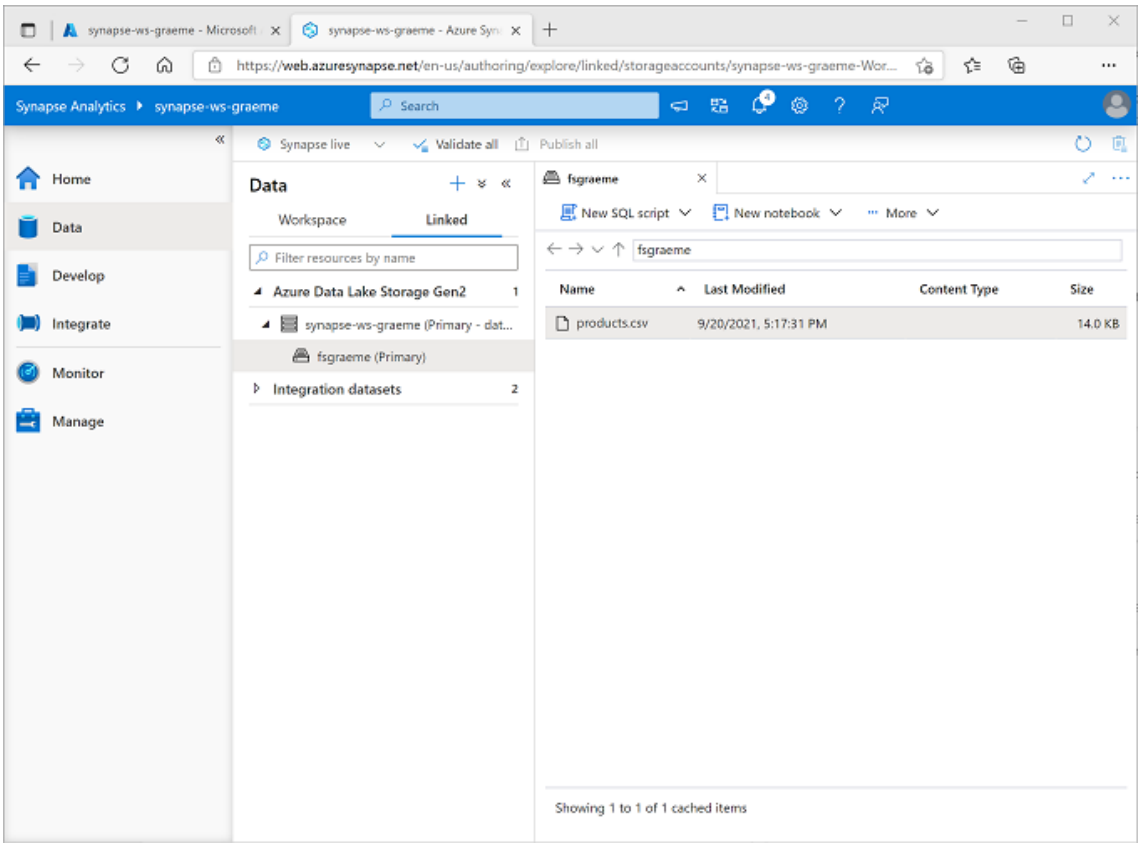
1. En Synapse Studio, en la página **de inicio** , seleccione **Ingerir** para abrir la **herramienta Copiar datos** .
2. En la herramienta Copiar datos, en el paso **Propiedades , asegúrese de que Tarea de copia integrada** y **Ejecutar una vez ahora** estén seleccionados y haga clic en **Siguiente >** .
3. En el paso **Fuente** , en el subpaso **Conjunto de datos** , seleccione la siguiente configuración:
    - **Tipo de fuente** : Todos
    - **Conexión** : *cree una nueva conexión y, en el panel* ***Nueva conexión*** *que aparece, en la pestaña* ***Protocolo genérico*** *, seleccione* ***HTTP*** *. Luego continúe y cree una conexión a un archivo de datos usando la siguiente configuración:*
        - **Nombre** : Productos AdventureWorks
        - **Descripción** : Lista de productos vía HTTP
        - **Conéctese a través del tiempo de ejecución de integración** : AutoResolveIntegrationRuntime
        - **URL básica** :`https://raw.githubusercontent.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/master/Azure-Synapse/products.csv`
        - **Validación del certificado del servidor** : habilitar
        - **Tipo de autenticación** : Anónimo
4. Después de crear la conexión, en el subpaso **Fuente/Conjunto de datos** , asegúrese de que estén seleccionadas las siguientes configuraciones y luego seleccione **Siguiente >** :
    - **URL relativa** : *dejar en blanco*
    - **Método de solicitud** : OBTENER
    - **Encabezados adicionales** : *dejar en blanco*
    - **Copia binaria** : no seleccionada
    - **Solicitar tiempo de espera** : *dejar en blanco*
    - **Máximo de conexiones simultáneas** : *dejar en blanco*
5. En el paso **Fuente** , en el subpaso **Configuración , seleccione Vista previa de datos** para ver una vista previa de los datos del producto que su canalización consumirá y luego cierre la vista previa.
6. Después de obtener una vista previa de los datos, en el paso **Fuente/Configuración** , asegúrese de que estén seleccionadas las siguientes configuraciones y luego seleccione **Siguiente >** :

- **Formato de archivo** : texto delimitado
- **Delimitador de columna** : coma (,)
- **Delimitador de fila** : avance de línea (\n)
- **Primera fila como encabezado** : Seleccionado
- **Tipo de compresión** : Ninguno

7. En el paso **Destino** , en el subpaso **Conjunto de datos** , seleccione la siguiente configuración:
   - **Tipo de destino** : Azure Data Lake Storage Gen 2
   - **Conexión** : *seleccione la conexión existente a su almacén de lago de datos (esto se creó para usted cuando creó el espacio de trabajo).*
8. Después de seleccionar la conexión, en el paso **Destino/Conjunto de datos** , asegúrese de que estén seleccionadas las siguientes configuraciones y luego seleccione **Siguiente >** :
   - **Ruta de la carpeta** : *busque la carpeta de su sistema de archivos*
   - **Nombre del archivo** : productos.csv
   - **Comportamiento de copia** : Ninguno
   - **Máximo de conexiones simultáneas** : *dejar en blanco*
   - **Tamaño de bloque (MB)** : *dejar en blanco*
9. En el paso **Destino** , en el subpaso **Configuración** , asegúrese de que estén seleccionadas las siguientes propiedades. Luego seleccione **Siguiente >** :
   - **Formato de archivo** : texto delimitado
   - **Delimitador de columna** : coma (,)
   - **Delimitador de fila** : avance de línea (\n)
   - **Agregar encabezado al archivo** : Seleccionado
   - **Tipo de compresión** : Ninguno
   - **Máximo de filas por archivo** : *dejar en blanco*
   - **Prefijo del nombre del archivo** : *dejar en blanco*
10. En el paso **Configuración** , ingrese la siguiente configuración y luego haga clic en **Siguiente >** :
    - **Nombre de la tarea** : Copiar productos
    - **Descripción de la tarea** Copiar datos de productos
    - **Tolerancia a fallos** : *dejar en blanco*
    - **Habilitar registro** : no seleccionado
    - **Habilitar preparación** : no seleccionado
11. En el paso **Revisar y finalizar** , en el subpaso **Revisar** , lea el resumen y luego haga clic en **Siguiente >** .
12. En el subpaso **Implementación** , espere a que se implemente la canalización y luego haga clic en **Finalizar** .
13. En Synapse Studio, seleccione la página **Monitor** y, en la pestaña **Ejecuciones de canalización** , espere a que se complete la canalización **de Copiar productos** con un estado de **Correcto** (puede

usar el botón ↻ **Actualizar** en la página Ejecuciones de canalización para actualizar el estado).

14. En la página **Datos** , seleccione la pestaña **Vinculado** y expanda la jerarquía **de Azure Data Lake Storage Gen 2** hasta que vea el almacenamiento de archivos para su área de trabajo de Synapse. Luego seleccione el almacenamiento de archivos para verificar que se haya copiado un archivo llamado **productos.csv en esta ubicación, como se muestra aquí:**



Utilice un grupo de SQL para analizar datos

Ahora que ha ingerido algunos datos en su espacio de trabajo, puede utilizar Synapse Analytics para consultarlos y analizarlos. Una de las formas más comunes de consultar datos es usar SQL y en Synapse Analytics puede usar un *grupo de SQL* para ejecutar código SQL.

1. En Synapse Studio, haga clic con el botón derecho en el archivo **productos.csv** en el almacenamiento de archivos de su espacio de trabajo de Synapse, seleccione **Nuevo script SQL** y seleccione **Seleccionar las 100 filas SUPERIORES** .
2. En el panel **SQL Script 1** que se abre, revise el código SQL que se ha generado, que debería ser similar a este:

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(

       BULK 'https://datalakexx.dfs.core.windows.net/fsxx/products.cs
v',
        FORMAT = 'CSV',
        PARSER_VERSION='2.0'
    ) AS [result]
```

Este código abre un conjunto de filas del archivo de texto que importó y recupera las primeras 100 filas de datos.

3. En la lista **Conectar a , asegúrese de que Integrado** esté seleccionado; esto representa el grupo de SQL integrado que se creó con su espacio de trabajo.
4. En la barra de herramientas, use el botón ▷ **Ejecutar** para ejecutar el código SQL y revise los resultados, que deberían verse similares a este:

C1	c2	c3
ID del Producto	Nombre del producto	Categoría
771	Montaña-100 Plata, 38	Bicicletas de montaña
772	Montaña-100 Plata, 42	Bicicletas de montaña
…	…	…

5. Tenga en cuenta que los resultados constan de cuatro columnas denominadas C1, C2, C3 y C4; y que la primera fila de los resultados contenga los nombres de los campos de datos. Para solucionar este problema, agregue un parámetro HEADER_ROW = TRUE a la función OPENROWSET como se muestra aquí (reemplazando *datalakexx* y *fsxx* con los nombres de su cuenta de almacenamiento del lago de datos y sistema de archivos) y luego vuelva a ejecutar la consulta:

```
SELECT
    TOP 100 *
FROM
    OPENROWSET(

       BULK 'https://datalakexx.dfs.core.windows.net/fsxx/products.cs
v',
```

```
        FORMAT = 'CSV',
        PARSER_VERSION='2.0',
        HEADER_ROW = TRUE
    ) AS [result]
```

Ahora los resultados se ven así:

ID del Producto	Nombre del producto	Categoría
771	Montaña-100 Plata, 38	Bicicletas de montaña
772	Montaña-100 Plata, 42	Bicicletas de montaña
…	…	…

6. Modifique la consulta de la siguiente manera (reemplazando *datalakexx* y *fsxx* con los nombres de su cuenta de almacenamiento del lago de datos y sistema de archivos):

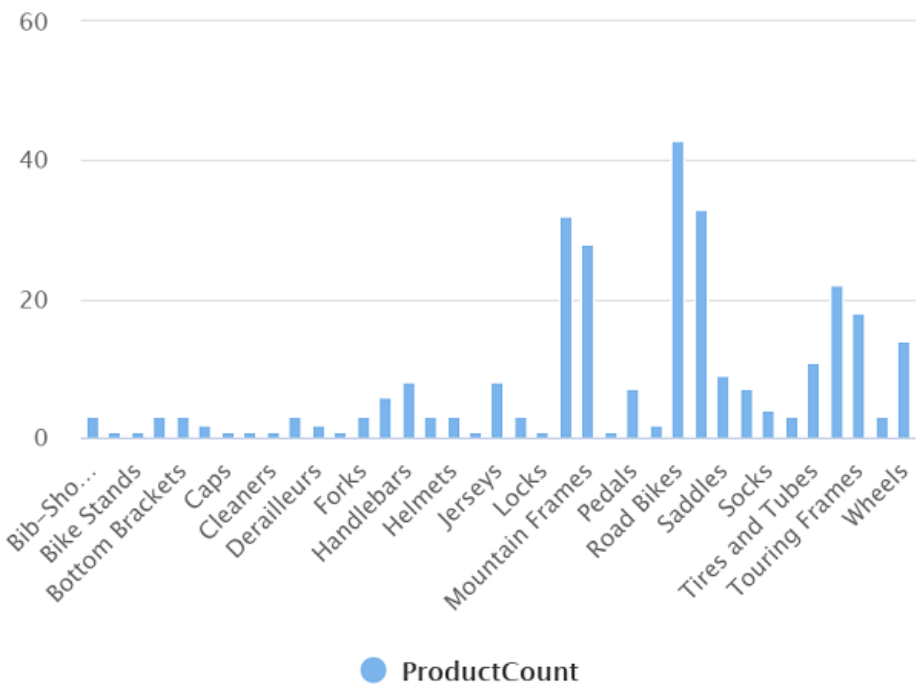
```
SELECT
    Category, COUNT(*) AS ProductCount
FROM
    OPENROWSET(

       BULK 'https://datalakexx.dfs.core.windows.net/fsxx/products.csv',
        FORMAT = 'CSV',
        PARSER_VERSION='2.0',
        HEADER_ROW = TRUE
    ) AS [result]
GROUP BY Category;
```

7. Ejecute la consulta modificada, que debería devolver un conjunto de resultados que contenga la cantidad de productos en cada categoría, como este:

Categoría	Recuento de productos
Culottes con tirantes	3
Portabicicletas	1
…	…

8. En el panel **Propiedades de SQL Script 1** , cambie el **Nombre** para **Contar productos por categoría** . Luego, en la barra de herramientas, seleccione **Publicar** para guardar el script.
9. Cierre el panel de secuencia de comandos **Contar productos por categoría** .
10. En Synapse Studio, seleccione la página **Desarrollar** y observe que su script SQL **de recuento de productos por categoría** publicado se ha guardado allí.
11. Seleccione el script SQL **Contar productos por categoría** para volver a abrirlo. Luego asegúrese de que el script esté conectado al grupo de SQL **integrado** y ejecútelo para recuperar los recuentos de productos.
12. En el panel **Resultados** , seleccione la vista **Gráfico** y luego seleccione las siguientes configuraciones para el gráfico:
    - **Tipo de gráfico** : columna
    - **Columna de categoría** : Categoría
    - **Columnas de leyenda (serie)** : ProductCount
    - **Posición de la leyenda** : abajo - centro
    - **Etiqueta de leyenda (serie)** : *dejar en blanco*
    - **Valor mínimo de leyenda (serie)** : *dejar en blanco*
    - **Máximo de leyenda (serie)** : *dejar en blanco*
    - **Etiqueta de categoría** : *dejar en blanco*

    El gráfico resultante debería parecerse a este:



Utilice un grupo de Spark para analizar datos

Si bien SQL es un lenguaje común para consultar conjuntos de datos estructurados, muchos analistas de datos encuentran útiles lenguajes como Python para explorar y preparar datos para el análisis. En Azure Synapse Analytics, puede ejecutar código Python (y otros) en un *grupo de Spark* ; que utiliza un motor de procesamiento de datos distribuido basado en Apache Spark.

1. En Synapse Studio, seleccione la página **Administrar** .
2. Seleccione la pestaña **Grupos de Apache Spark** y luego use el ícono + **Nuevo** para crear un nuevo grupo de Spark con la siguiente configuración:
    - **Nombre del grupo de Apache Spark** : chispa
    - **Familia de tamaños de nodos** : Memoria optimizada
    - **Tamaño de nodo** : Pequeño (4 vCores / 32 GB)
    - **Escala automática** : habilitada
    - **Número de nodos** 3—-3
3. Revise y cree el grupo de Spark y luego espere a que se implemente (lo que puede tardar unos minutos).
4. Cuando se haya implementado el grupo de Spark, en Synapse Studio, en la página **Datos** , busque el sistema de archivos de su espacio de trabajo de Synapse. Luego, haga clic con el botón derecho en **Products.csv** , seleccione **Nuevo cuaderno** y seleccione **Cargar en marco de datos** .
5. En el panel **Notebook 1** que se abre, en la lista **Adjuntar a , seleccione el grupo de Spark** que se creó anteriormente y asegúrese de que el **Idioma** esté configurado en **PySpark (Python)** .
6. Revise el código en la primera (y única) celda del cuaderno, que debería verse así:

```
%%pyspark
df =
spark.read.load('abfss://fsxx@datalakexx.dfs.core.windows.net/products
.csv', format='csv'
## If header exists uncomment line below
##, header=True
)
display(df.limit(10))
```

7. Seleccione ▷ **Ejecutar** a la izquierda de la celda del código para ejecutarlo y espere los resultados. La primera vez que ejecuta una celda en un cuaderno, se inicia el grupo de Spark, por lo que puede tardar aproximadamente un minuto en obtener resultados.
8. Finalmente, los resultados deberían aparecer debajo de la celda y deberían ser similares a este:

c0	*c1*	*c2*	*c3*
ID del Producto	Nombre del producto	Categoría	Precio de lista
771	Montaña-100 Plata, 38	Bicicletas de montaña	3399.9900
772	Montaña-100 Plata, 42	Bicicletas de montaña	3399.9900
…	…	…	…

9. Descomente la línea *,header=True* (porque el archivo productos.csv tiene los encabezados de columna en la primera línea), para que su código se vea así:

```
%%pyspark
df =
spark.read.load('abfss://fsxx@datalakexx.dfs.core.windows.net/products
.csv', format='csv'
## If header exists uncomment line below
, header=True
)
display(df.limit(10))
```

10. Vuelva a ejecutar la celda y verifique que los resultados se vean así:

ID del Producto	**Nombre del producto**	**Categoría**	**Precio de lista**
771	Montaña-100 Plata, 38	Bicicletas de montaña	3399.9900
772	Montaña-100 Plata, 42	Bicicletas de montaña	3399.9900
…	…	…	…

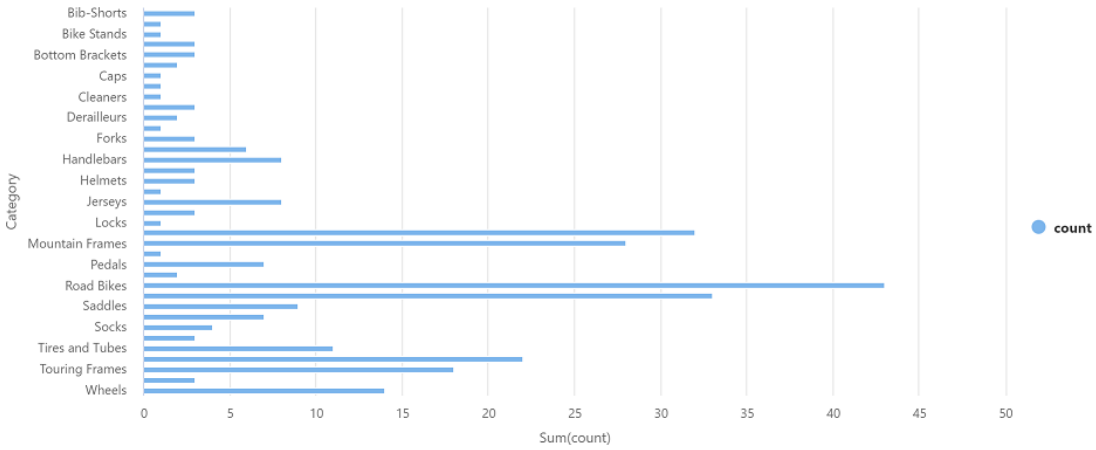
11. Tenga en cuenta que ejecutar la celda nuevamente lleva menos tiempo porque el grupo de Spark ya está iniciado.
12. Debajo de los resultados, use el ícono **+ Código** para agregar una nueva celda de código al cuaderno.
13. En la nueva celda de código vacía, agregue el siguiente código:

```
df_counts = df.groupBy(df.Category).count()
display(df_counts)
```

14. Seleccione ▷ **Ejecutar** a la izquierda para ejecutar la nueva celda de código y revise los resultados, que deberían verse similares a este:

Categoría	contar
Auriculares	3
Ruedas	14
…	…

15. En la salida de resultados de la celda, seleccione la vista **Gráfico** . El gráfico resultante debería parecerse a este:



16. Cierre el panel **Notebook 1** y descarte los cambios.