Explorar Azure Synapse Analytics

Azure Synapse Analytics proporciona una plataforma de análisis de datos única y consolidada para análisis de datos de un extremo a otro. En este ejercicio, explorará varias formas de ingerir y explorar datos. Este ejercicio está diseñado como una descripción general de alto nivel de las diversas capacidades principales de Azure Synapse Analytics. Hay otros ejercicios disponibles para explorar capacidades específicas con más detalle.

Este ejercicio debería tardar aproximadamente **60** minutos en completarse.

Antes de que empieces

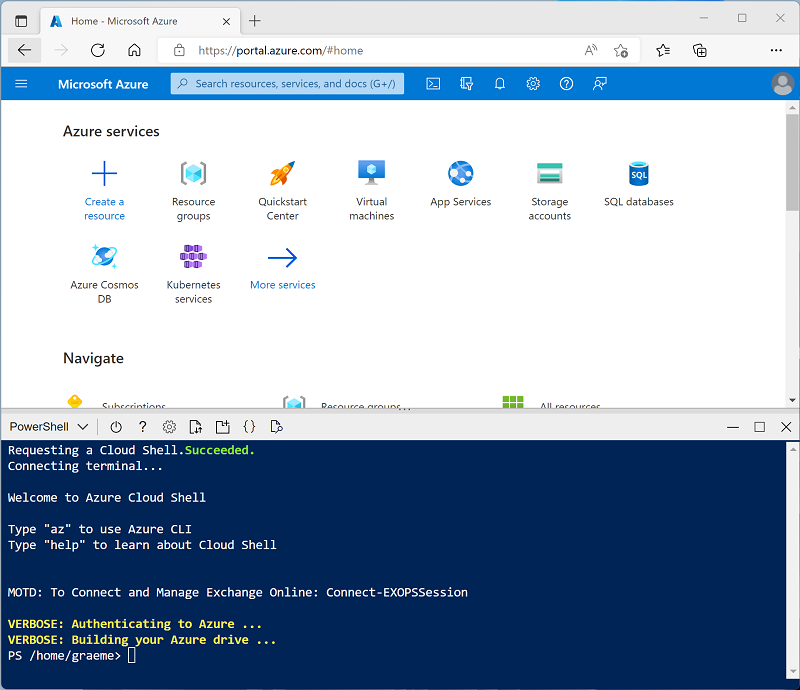
Necesitará una [suscripción de Azure](https://azure.microsoft.com/free) a la que tenga acceso de nivel administrativo.

Aprovisionar un área de trabajo de Azure Synapse Analytics

*Un área de trabajo* de Azure Synapse Analytics proporciona un punto central para administrar datos y tiempos de ejecución de procesamiento de datos. Puede aprovisionar un área de trabajo mediante la interfaz interactiva en Azure Portal o puede implementar un área de trabajo y recursos dentro de ella mediante un script o una plantilla. En la mayoría de los escenarios de producción, es mejor automatizar el aprovisionamiento con scripts y plantillas para poder incorporar la implementación de recursos en un proceso de desarrollo y operaciones ( *DevOps ) repetible.*

En este ejercicio, utilizará una combinación de un script de PowerShell y una plantilla de ARM para aprovisionar Azure Synapse Analytics.

1. En un explorador web, inicie sesión en [Azure Portal](https://portal.azure.com/) en https://portal.azure.com.
2. Utilice el botón **[>\_]** a la derecha de la barra de búsqueda en la parte superior de la página para crear un nuevo Cloud Shell en Azure Portal, seleccionando un entorno ***de PowerShell*** y creando almacenamiento si se le solicita. Cloud Shell proporciona una interfaz de línea de comandos en un panel en la parte inferior de Azure Portal, como se muestra aquí:

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/cloud-shell.png)

**Nota** : si anteriormente creó un shell de nube que usa un entorno *Bash* , use el menú desplegable en la parte superior izquierda del panel de shell de nube para cambiarlo a ***PowerShell*** .

1. Tenga en cuenta que puede cambiar el tamaño del shell de la nube arrastrando la barra separadora en la parte superior del panel o usando los íconos **—** , **◻** y **X** en la parte superior derecha del panel para minimizar, maximizar y cerrar el panel. Para obtener más información sobre el uso de Azure Cloud Shell, consulte la [documentación de Azure Cloud Shell](https://docs.microsoft.com/azure/cloud-shell/overview) .
2. En el panel de PowerShell, ingrese los siguientes comandos para clonar este repositorio:

Código

rm -r demaster -f

git clone https://github.com/ReinyerFerra/mbit-azure-data-engineer-labs demaster

1. Una vez clonado el repositorio, ingrese los siguientes comandos para cambiar a la carpeta para este ejercicio y ejecutar el script **setup.ps1** que contiene:

Código

cd demaster/Laboratorios/01

./setup.ps1

1. Si se le solicita, elija qué suscripción desea usar (esto solo sucederá si tiene acceso a varias suscripciones de Azure).
2. Cuando se le solicite, ingrese una contraseña adecuada que se establecerá para su Azure Synapse SQL pool.

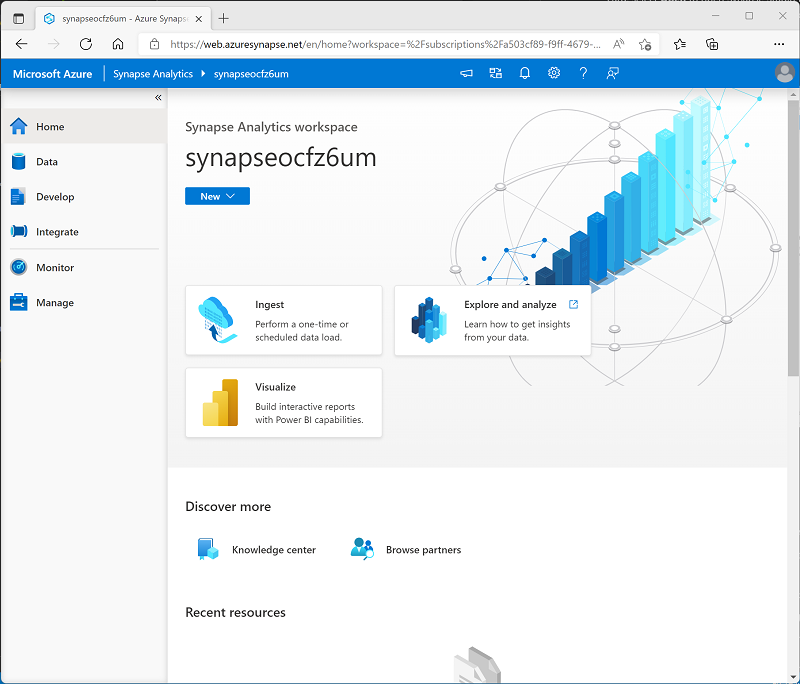
**Nota** : ¡Asegúrese de recordar esta contraseña!

1. Espere a que se complete el script; esto suele tardar unos 20 minutos.

Explorar Synapse Studio

*Synapse Studio* es un portal basado en web en el que puede administrar y trabajar con los recursos de su área de trabajo de Azure Synapse Analytics.

1. Cuando el script de instalación haya terminado de ejecutarse, en Azure Portal, vaya al grupo de recursos **demaster- *xxxxxxx*** que creó y observe que este grupo de recursos contiene su espacio de trabajo de Synapse, una cuenta de almacenamiento para su Data Lake, un Apache Spark pool, un Data Explorer pool y un SQL pool dedicado.
2. Seleccione su espacio de trabajo de Synapse y, en su página **Información 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.
3. En el lado izquierdo de Synapse Studio, use el ícono **››** para expandir el menú; esto muestra las diferentes páginas dentro de Synapse Studio que usará para administrar recursos y realizar tareas de análisis de datos, como se muestra aquí:

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/synapse-studio.png)

1. Vea la página **Data** y tenga en cuenta que hay dos pestañas que contienen fuentes de datos:
   * Una pestaña **Workspace** que contiene bases de datos definidas en el espacio de trabajo (incluidas bases de datos SQL dedicadas y bases de datos del Explorador de datos)
   * Una pestaña **Linked** que contiene orígenes de datos vinculados al área de trabajo, incluido el almacenamiento de Azure Data Lake.
2. Vea la página **Develop**, que actualmente está vacía. Aquí es donde puede definir scripts y otros activos utilizados para desarrollar soluciones de procesamiento de datos.
3. Vea la página **Integrate**, que también está vacía. Utilice esta página para gestionar la ingesta de datos y los activos de integración; como canalizaciones para transferir y transformar datos entre fuentes de datos.
4. Vea la página **Monitor**. Aquí es donde puede observar los trabajos de procesamiento de datos mientras se ejecutan y ver su historial.
5. Vea la página **Manage** . Aquí es donde administra los grupos, los tiempos de ejecución y otros recursos utilizados en su área de trabajo de Azure Synapse. Vea cada una de las pestañas en la sección **Analytics pools** y tenga en cuenta que su espacio de trabajo incluye los siguientes grupos:
   * **SQL** **pools**:
     + **Integrado** : un SQL pool *sin servidor* que puede usar bajo demanda para explorar o procesar datos en un lago de datos mediante comandos SQL.
     + **sql *xxxxxxx*** : un SQL pool *dedicado* que aloja una base de datos de almacén de datos relacional.
   * **Apache Spark** **pools**:
     + **spark *xxxxxxx*** : que puede usar bajo demanda para explorar o procesar datos en un lago de datos mediante lenguajes de programación como Scala o Python. <!— - **Grupos de Explorador de datos** :
     + **adx *xxxxxxx*** : un grupo de Explorador de datos que puede utilizar para analizar datos mediante Kusto Query Language (KQL). —>

Ingerir datos con una canalización

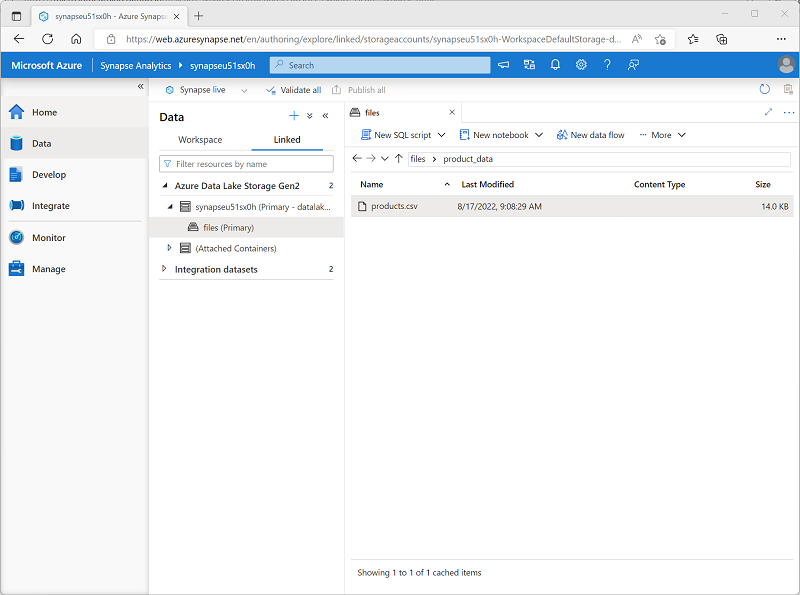
Una de las tareas clave que puede realizar con Azure Synapse Analytics es definir *canalizaciones* que transfieran (y, si es necesario, transformen) datos de una amplia gama de orígenes a su espacio de trabajo para su análisis.

Utilice la tarea Copiar datos para crear una canalización

1. En Synapse Studio, en la página **de inicio (Home)**, seleccione **Ingerir** para abrir la **herramienta** **de** **Copiar datos.**
2. En la herramienta de 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 **Origen** , en el subpaso **Conjunto de datos** , seleccione la siguiente configuración:
   * **Tipo de origen** : Todo
   * **Conexión**: *cree una nueva conexión y, en el panel****Servicio vinculado****que aparece, en la pestaña* ***Generic protocol****, seleccione****HTTP****. Luego continúe y cree una conexión a un archivo de datos usando la siguiente configuración:*
     + **Nombre** : Productos
     + **Descripción** : Lista de productos vía HTTP
     + **Conectar mediante Integration Runtime** : AutoResolveIntegrationRuntime
     + **URL base**: https://raw.githubusercontent.com/ReinyerFerra/mbit-azure-data-engineer-labs/main/Laboratorios/01/adventureworks/products.csv
     + **Validación del certificado del servidor**: habilitar
     + **Tipo de autenticación**: Anónimo
4. Después de crear la conexión, en la página **del almacén de datos de origen**, asegúrese de que estén seleccionadas las siguientes configuraciones y luego seleccione **Siguiente >**:
   * **Dirección URL relativa**: *dejar en blanco*
   * **Método de solicitud**: GET
   * **Encabezados adicionales**: *dejar en blanco*
   * **Copia binaria**: no seleccionada
   * **Tiempo de espera de solicitud**: *dejar en blanco*
   * **Número máximo de conexiones simultáneas**: *dejar en blanco*
5. En el paso **Origen**, en el subpaso **Configuración** **del formato de archivo**, 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 la página **de configuración de formato de archivo**, asegúrese de que estén seleccionadas las siguientes configuraciones y luego seleccione **Siguiente >**:
   * **Formato de archivo**: DelimitedText
   * **Delimitador de columnas**: Comma (,)
   * **Delimitador de fila**: Line feed (\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**: files/product\_data
   * **Nombre del archivo**: productos.csv
   * **Comportamiento de copia**: Ninguno
   * **Máximo de conexiones simultáneas**: *dejar en blanco*
   * **Tamaño del bloque (MB)**: *dejar en blanco*
9. En el paso **Destino**, en el subpaso **Configuración** , en la página **Configuración de formato de archivo** , asegúrese de que las siguientes propiedades estén seleccionadas. Luego seleccione **Siguiente >**:
   * **Formato de archivo**: DelimitedText
   * **Delimitador de columna**: Comma (,)
   * **Delimitador de fila**: Line Feed (\n)
   * **Agregar encabezado al archivo**: Seleccionado
   * **Tipo de compresión**: Ninguno
   * **Número máximo de filas por archivo**: *dejar en blanco*
   * **Nombre de archivo prefijo**: *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 errores**: *dejar en blanco*
    * **Habilitar registro**: no seleccionado
    * **Habilitar el almacenamiento provisional**: 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 paso **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. Vea la página **Integrate** y verifique que ahora contenga una canalización denominada **Copiar productos**.

Ver los datos ingeridos

1. En la página **Data** , seleccione la pestaña **Linked** y expanda la jerarquía del contenedor del lago de datos ***xxxxxxx* (principal) de Synapse** hasta que vea el almacenamiento **de archivos** para su espacio de trabajo de Synapse. Luego seleccione el almacenamiento de archivos para verificar que se haya copiado en esta ubicación una carpeta llamada **product\_data** que contiene un archivo llamado **products.csv, como se muestra aquí:**

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/product_files.png)

1. Haga clic con el botón derecho en el archivo de datos **products.csv** y seleccione **Versión preliminar** para ver los datos ingeridos. Luego cierre la vista previa.

Utilice un SQL pool sin servidor 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 SQL pool sin servidor para ejecutar código SQL en datos en un Data Lake.

1. En Synapse Studio, haga clic con el botón derecho en el archivo **products.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:

Código

-- This is auto-generated code

SELECT

    TOP 100 \*

FROM

    OPENROWSET(

        BULK 'https://datalakexxxxxxx.dfs.core.windows.net/files/product\_data/products.csv',

        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.

1. En la lista **Conectarse a ,** asegúrese de que **Integrado** esté seleccionado; esto representa el SQL pool integrado que se creó con su espacio de trabajo.
2. 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 | C4 |
| --- | --- | --- | --- |
| 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 |
| … | … | … | … |

1. 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 *datalakexxxxxxx* con el nombre de su cuenta de almacenamiento del lago de datos) y luego vuelva a ejecutar la consulta:

Código

SELECT

    TOP 100 \*

FROM

    OPENROWSET(

        BULK 'https://datalakexxxxxxx.dfs.core.windows.net/files/product\_data/products.csv',

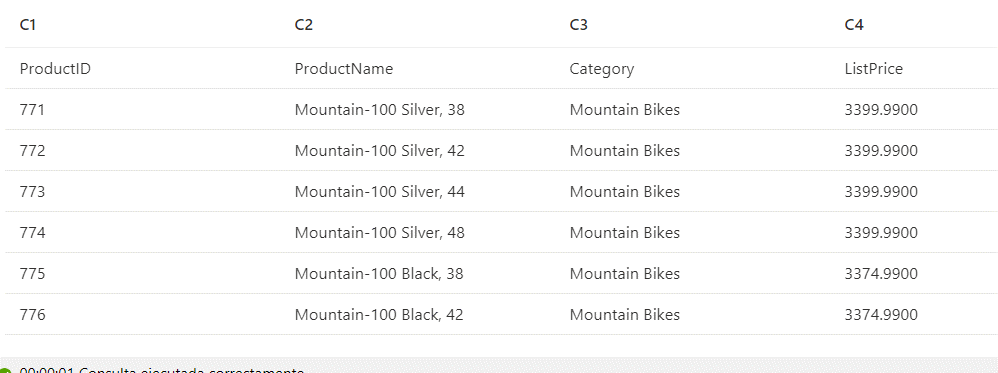
        FORMAT = 'CSV',

        PARSER\_VERSION='2.0',

HEADER\_ROW = TRUE

) AS [result]

Ahora los resultados se ven así:



1. Modifique la consulta de la siguiente manera (reemplazando *datalakexxxxxxx* con el nombre de su cuenta de almacenamiento del lago de datos):

Código

SELECT

    Category, COUNT(\*) AS ProductCount

FROM

    OPENROWSET(

        BULK 'https://datalakexxxxxxx.dfs.core.windows.net/files/product\_data/products.csv',

        FORMAT = 'CSV',

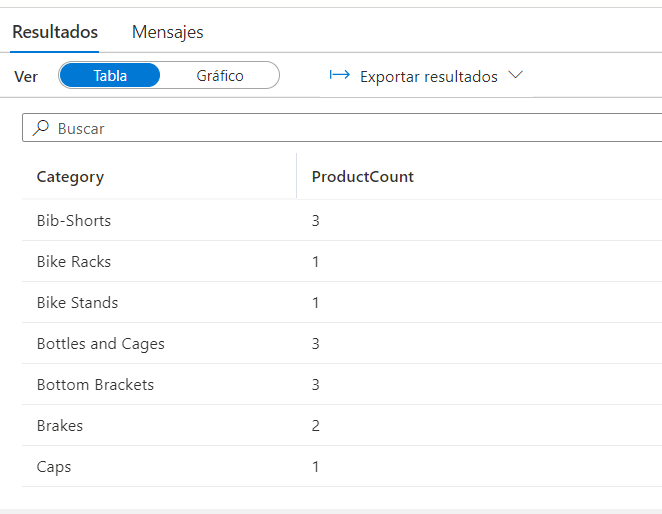
        PARSER\_VERSION='2.0',

HEADER\_ROW = TRUE

) AS [result]

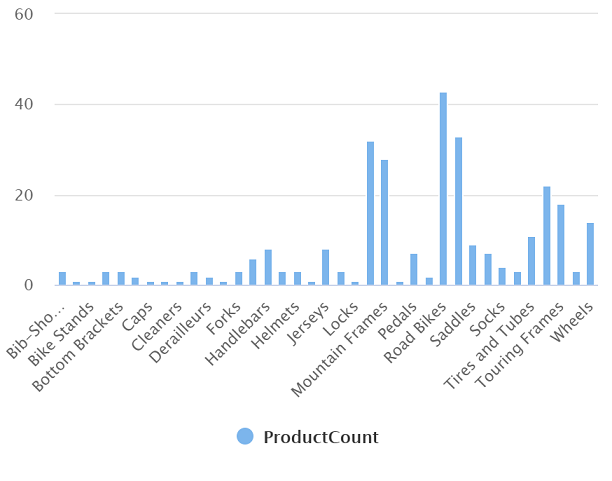
GROUP BY Category;

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



1. En el panel **Propiedades de SQL Script 1**, cambie el **Nombre** para **Contar productos por categoria**. Luego, en la barra de herramientas, seleccione **Publicar** para guardar el script.
2. Cierre el panel de secuencia de comandos **Contar productos por categoría**.
3. En Synapse Studio, seleccione la página **Develop** y observe que su script SQL **de recuento de productos por categoria** publicado se ha guardado allí.
4. Seleccione el script SQL **Contar productos por categoria** para volver a abrirlo. Luego asegúrese de que el script esté conectado al SQL pool **integrado** y ejecútelo para recuperar los recuentos de productos.
5. En el panel **Resultados** , seleccione la vista **Gráfico** y luego seleccione las siguientes configuraciones para el gráfico:
   * **Tipo de gráfico** : columna
   * **Columna de categoría** : Category
   * **Columnas de leyenda (serie)** : ProductCount
   * **Posición de la leyenda** : parte inferior del centro
   * **Etiqueta de leyenda (serie)** : *dejar en blanco*
   * **Valor mínimo de leyenda (serie)** : *dejar en blanco*
   * **Valor máximo de leyenda (serie)** : *dejar en blanco*
   * **Etiqueta de categoría** : *dejar en blanco*

El gráfico resultante debería parecerse a este:

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/column-chart.png)

Utilice un Spark pool 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 *Spark* pool; que utiliza un motor de procesamiento de datos distribuido basado en Apache Spark.

1. En Synapse Studio, si la pestaña de **files** que abrió anteriormente y que contiene el archivo **products.csv** ya no está abierta, en la página **Data**, busque la carpeta **product\_data.**Luego, haga clic con el botón derecho en **products.csv**, seleccione **Nuevo cuaderno** y seleccione **Cargar en Dataframe**.
2. En el panel **Notebook 1** que se abre, en la lista **Asociar a**, seleccione el grupo **Sparkxxxxxxx** Spark y asegúrese de que el **lenguaje** esté configurado en **PySpark (Python)** .
3. Revise el código en la primera (y única) celda del cuaderno, que debería verse así:

Código

%%pyspark

df = spark.read.load('abfss://files@datalakexxxxxxx.dfs.core.windows.net/product\_data/products.csv', format='csv'

## If header exists uncomment line below

##, header=True

)

display(df.limit(10))

1. Utilice el icono **▷** 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.
2. Finalmente, los resultados deberían aparecer debajo de la celda y deberían ser similares a este:
3. Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

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

Código

%%pyspark

df = spark.read.load('abfss://files@datalakexxxxxxx.dfs.core.windows.net/product\_data/products.csv', format='csv'

## If header exists uncomment line below

, header=True

)

display(df.limit(10))

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

Interfaz de usuario gráfica, Texto, Tabla

Descripción generada automáticamente

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

Código

df\_counts = df.groupby(df.Category).count()

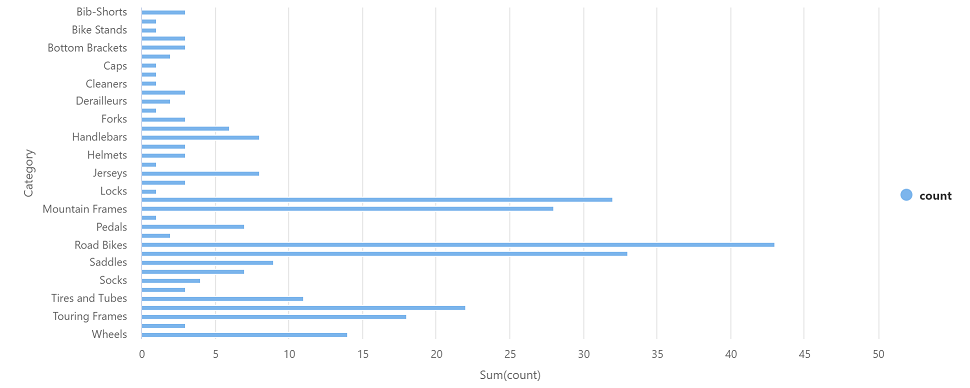
display(df\_counts)

1. Ejecute la nueva celda de código haciendo clic en su ícono **▷** y revise los resultados, que deberían verse similares a este:

Tabla

Descripción generada automáticamente

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

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/bar-chart.png)

1. Si aún no está visible, muestre la página **Propiedades** seleccionando el botón **Propiedades** (que se parece a **🗏 \*** ) en el extremo derecho de la barra de herramientas. Luego, en el panel **Propiedades** , cambie el nombre del cuaderno a **Explorar productos** y use el botón **Publicar** en la barra de herramientas para guardarlo.
2. Cierre el panel del cuaderno y detenga la sesión de Spark cuando se le solicite. Luego vea la página **Develop** para verificar que el cuaderno se haya guardado.

Utilice un SQL pool dedicado para consultar un almacén de datos

Hasta ahora ha visto algunas técnicas para explorar y procesar datos basados ​​en archivos en un lago de datos. En muchos casos, una solución de análisis empresarial utiliza un lago de datos para almacenar y preparar datos no estructurados que luego se pueden cargar en un almacén de datos relacional para respaldar cargas de trabajo de inteligencia empresarial (BI). En Azure Synapse Analytics, estos almacenes de datos se pueden implementar en un de SQL Pool dedicado.

1. En Synapse Studio, en la página **Manage,** en la sección**SQL pools**, seleccione la fila del SQL pool dedicado **sql *xxxxxxx*** y luego use su ícono **▷** para reanudarlo.
2. Espere a que se inicie el SQL pool. Esto puede tardar unos minutos. Utilice el botón **↻ Actualizar** para comprobar su estado periódicamente. El estado se mostrará como **En línea** cuando esté listo.
3. Cuando se haya iniciado el SQL pool, seleccione la página **Data ;**y en la pestaña **Workspace** , expanda **las bases de datos SQL** y verifique que **sql *xxxxxxx*** esté en la lista (use el ícono **↻** en la parte superior izquierda de la página para actualizar la vista si es necesario).
4. Expanda la base de datos **sql *xxxxxxx*** y su carpeta **Tablas** y luego, en el menú **...** de la tabla **FactInternetSales** , seleccione **Nuevo script SQL** y seleccione **Seleccionar las primeras 100 filas**.
5. Revise los resultados de la consulta, que muestran las primeras 100 transacciones de ventas en la tabla. Estos datos se cargaron en la base de datos mediante el script de configuración y se almacenan permanentemente en la base de datos asociada con el SQL pool dedicado.
6. Reemplace la consulta SQL con el siguiente código:

SQLCopiar

SELECT d.CalendarYear, d.MonthNumberOfYear, d.EnglishMonthName,

p.EnglishProductName AS Product, SUM(o.OrderQuantity) AS UnitsSold

FROM dbo.FactInternetSales AS o

JOIN dbo.DimDate AS d ON o.OrderDateKey = d.DateKey

JOIN dbo.DimProduct AS p ON o.ProductKey = p.ProductKey

GROUP BY d.CalendarYear, d.MonthNumberOfYear, d.EnglishMonthName, p.EnglishProductName

ORDER BY d.MonthNumberOfYear

1. Utilice el botón **▷ Ejecutar** para ejecutar la consulta modificada, que devuelve la cantidad de cada producto vendido por año y mes.
2. Si aún no está visible, muestre la página **Propiedades** seleccionando el botón **Propiedades** (que se parece a **🗏 \*** ) en el extremo derecho de la barra de herramientas. Luego, en el panel **Propiedades** , cambie el nombre de la consulta a **Ventas agregadas de productos** y use el botón **Publicar** en la barra de herramientas para guardarla.
3. Cierre el panel de consulta y luego vea la página **Develop** para verificar que se haya guardado el script SQL.
4. En la página **Mange**, seleccione la fila del SQL pool dedicado **sql *xxxxxxx*** y use su ícono ❚❚ para pausarlo.

Eliminar recursos de Azure

Ahora que ha terminado de explorar Azure Synapse Analytics, debe eliminar los recursos que creó para evitar costos innecesarios de Azure.

1. Cierre la pestaña del navegador Synapse Studio y regrese a Azure Portal.
2. En Azure Portal, en la página de **inicio**, seleccione **Grupos de recursos**.
3. Seleccione el grupo de recursos **demaster- *xxxxxxx*** para su área de trabajo de Synapse Analytics (no el grupo de recursos administrado) y verifique que contenga el área de trabajo de Synapse, la cuenta de almacenamiento, el SQL pool, el grupo de Data Explorer y el grupo de Spark para su área de trabajo.
4. En la parte superior de la página **Descripción general** de su grupo de recursos, seleccione **Eliminar grupo de recursos**.
5. Ingrese el nombre del grupo de recursos **demaster- *xxxxxxx*** para confirmar que desea eliminarlo y seleccione **Eliminar**.

Después de unos minutos, se eliminarán el grupo de recursos del área de trabajo de Azure Synapse y el grupo de recursos del área de trabajo administrada asociado a él.