



BIG DATA
ACADEMY

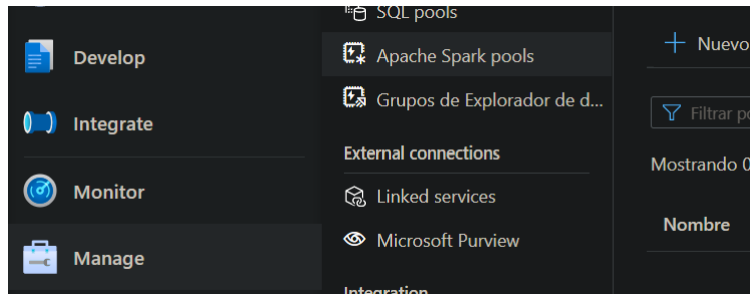
LABORATORIO 46

SOLUCIONES DE BIG DATA CON SYNAPSE SPARK POOL

FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

LABORATORIO 46 – SOLUCIONES DE BIG DATA CON SYNAPSE SPARK POOL

1. Dentro de Synapse podemos usar los clústers de “Spark Pool” para ejecutar código Spark y no depender de servicios externos como Databricks. Damos clic a “Manage / Apache Spark pools / Nuevo”



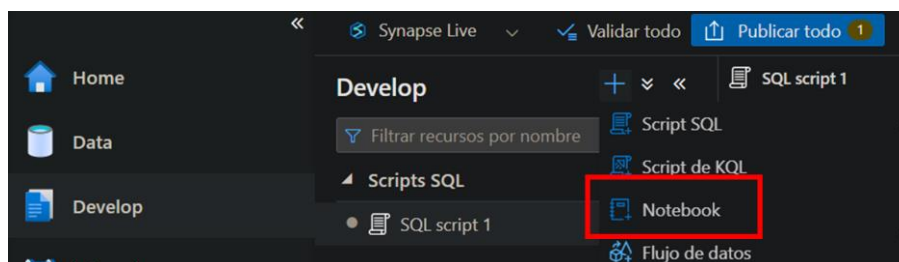
2. Configuramos:

En la pestaña “Aspectos básicos”

Nombre del grupo de Apache Spark	sparkpoolXXX
Familia de tamaño de nodo	Con optimización para memoria
Tamaño del nodo	Small (4 núcleos virtuales / 32 GB RAM)
Escalabilidad automática	Deshabilitada
Número de nodos	3

Damos clic en “Revisar y crear” y luego en “Crear”

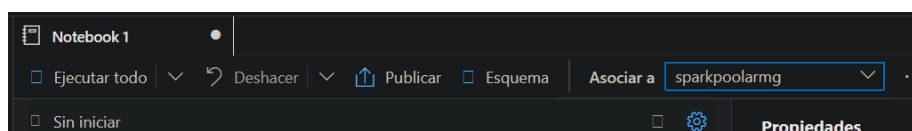
3. Desde la opción “Develop” seleccionamos “+ / Notebook”



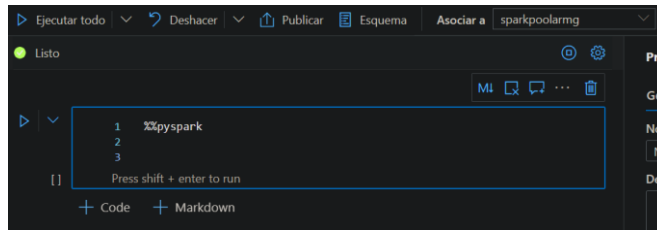
4. Colocamos como nombre de Notebooks “NOTEBOOK_SPARK_POOL”



5. Seleccionamos el grupo de SPARK “sparkpoolXXX”



6. Dentro de la sección de código ejecutaremos código de Spark con Python (PySpark), para eso como primera línea dentro de la sección de código escribimos “%%pyspark”



```
1 %%pyspark
2
3
```

7. En “Spark Pool” para referenciar a un archivo de datos, lo hacemos de la siguiente manera:

abfss://BLOB_STORAGE@ACCOUNT_STORAGE.dfs.core.windows.net/ruta/al/archivo.data

Donde:

abfss://	Es el sistema de archivos de Azure (Azure Blob File System)
BLOB_STORAGE	Nombre del blob storage en donde se encuentra nuestro archivo de datos
ACCOUNT_STORAGE	Nombre de la cuenta de almacenamiento donde se encuentra el blob storage
dfs.core.windows.net	Dirección de dominio que usa Azure
/ruta/al/archivo.data	Ruta del archivo de datos dentro del blob storage

Leeremos el archivo “persona.data” de la capa “bronze” del Delta Lake:

abfss://deltalake@storagebdaXXX.dfs.core.windows.net/bronze/persona/persona.data

8. Copiamos y pegamos el contenido del script “**SCRIPT 1.py**”, el cual lee un archivo de datos y lo carga dentro de un dataframe.



```
1 %%pyspark
2
3 #IMPORTANTE, EN EL BLOB STORAGE CAMBIAR "XXX" POR LAS INICIALES DE TU NOMBRE
4
5 #Leemos el archivo
6 dfData = spark.read.format("csv").option("header", "true").option("delimiter", "|").option("inferSchema", "true").load(abfss://deltalake@storagebdaXXX.dfs.core.windows.net/bronze/persona/persona.data)
```

[5] ✓ 4 sec - alonsoraulmgs ejecutó el comando en 4 s 31 ms el 11:37:52 AM, 7/04/23

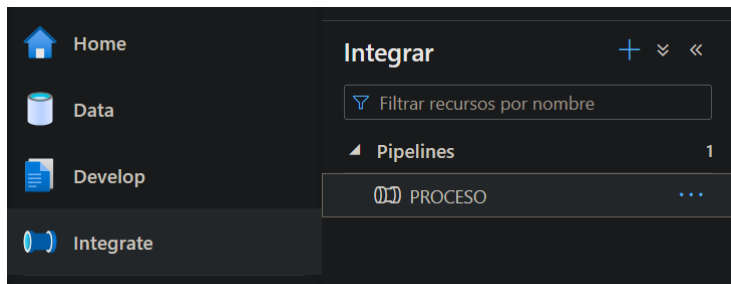
Ejecución del trabajo Correcto Spark 2 ejecutores y 8 núcleos Ver en supervisión Abrir interfaz

(IMPORTANTE: NO EJECUTAR EL CÓDIGO, SINO SE CREARÁ UN CLÚSTER DE PAGA, SI LO EJECUTAS LA CREACIÓN DEL CLÚSTER DE PAGA TOMA 10 MINUTOS)

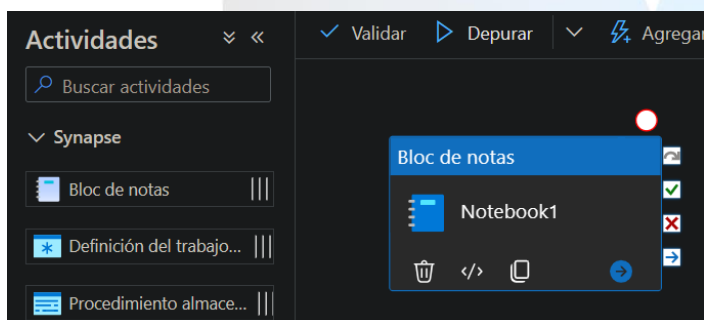
9. El resto del código es el mismo que vimos en DATABRICKS, guardamos el notebook dando clic en “Publicar todo”



10. Ahora agregaremos el notebook al pipeline de ejecución, vamos a “Integrate” y en “Pipelines” damos clic sobre “PROCESO” para abrir el pipeline.



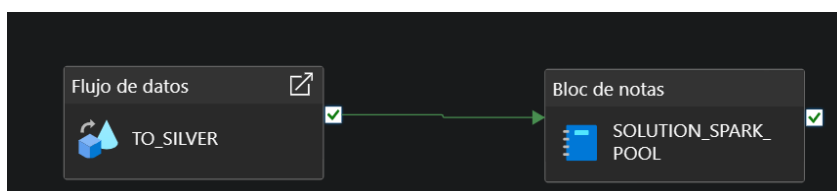
11. Desde la sección “Synapse” agregamos la actividad “Bloc de notas”



Configuramos:

Pestaña	Opción	Valor
General	Nombre	SOLUTION_SPARK_POOL
Configuración	Bloc de notas	NOTEBOOK_SPARK_POOL
Configuración	Grupo de Spark	sparkpoolXXX

12. Conectamos las actividades



13. Guardamos los cambios dando clic en “Publicar todo”

