BIG DATA
ACADEMY

LABORATORIO 47

SOLUCIONES DE BIG DATA CON DATABRICKS SOBRE SYNAPSE

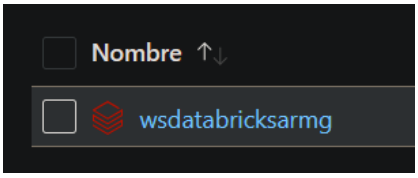
FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

LABORATORIO 47 – SOLUCIONES DE BIG DATA CON DATABRICKS SOBRE SYNAPSE

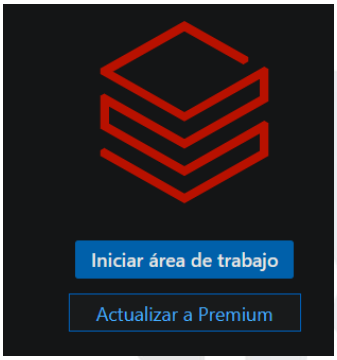
1. Desde el buscador de servicios, buscamos:

Azure Databricks

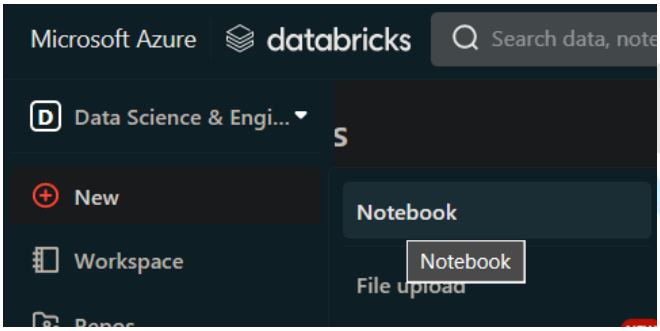
2. Damos clic en la instancia “wsdatabricks**XXX**”



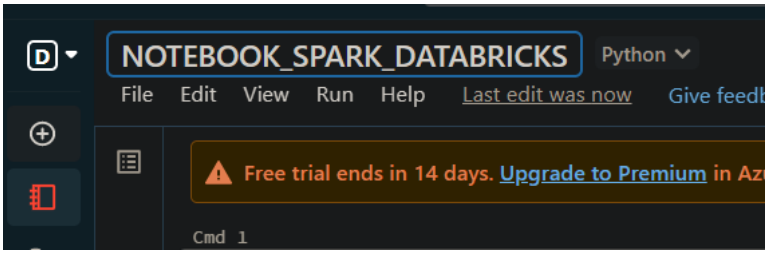
3. Damos clic en “Iniciar área de trabajo”



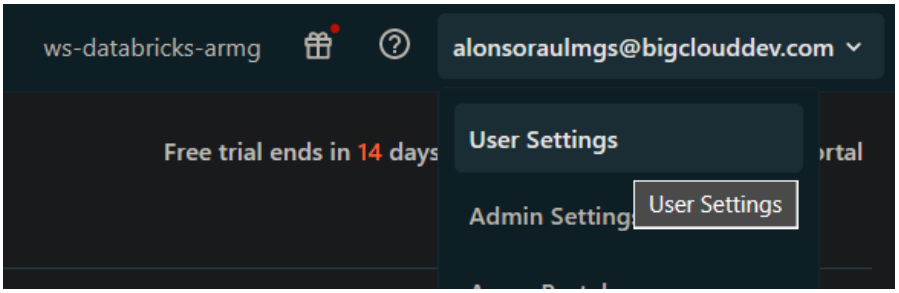
4. Crearemos un notebook de prueba, seleccionamos la opción “New / Notebook”



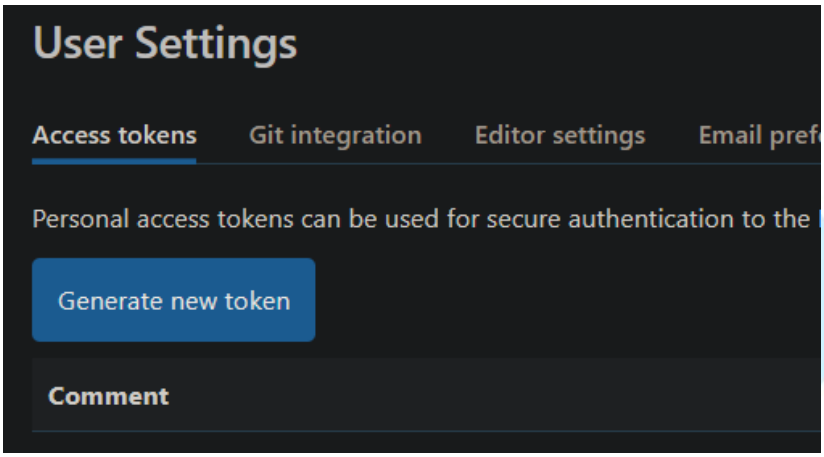
5. De nombre colocamos “NOTEBOOK_SPARK_DATABRICKS”.



6. Necesitaremos crear un token de acceso para conectar la cuenta de Databricks con Synapse. Desde la esquina superior derecha seleccionamos “Nombre de nuestro usuario / User Settings”



7. Seleccionamos la pestaña “Access tokens” y damos clic en “Generate new token”

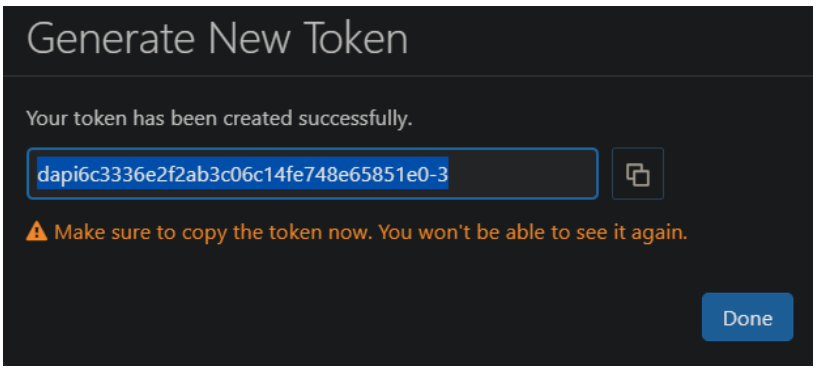


8. Configuramos:

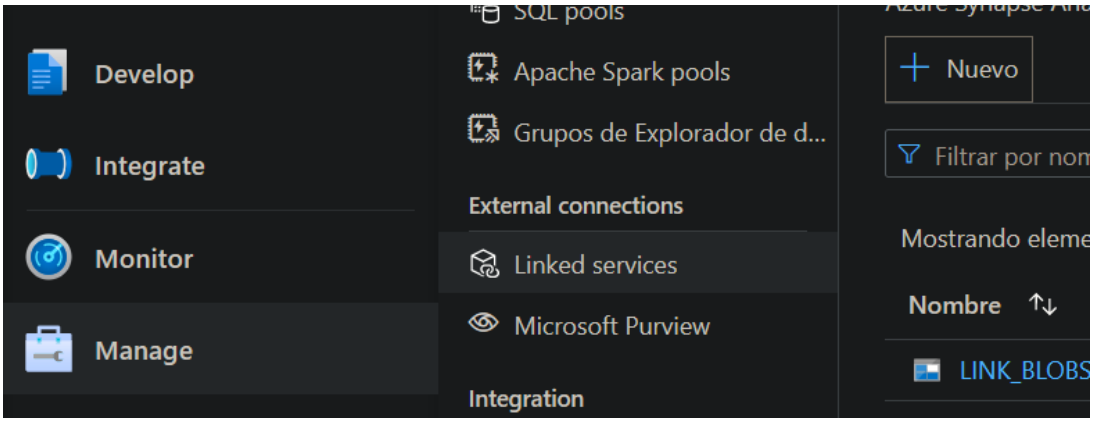
Comment	TOKEN_SYNAPSE
Lifetime (days)	90

Damos clic en “Generate”

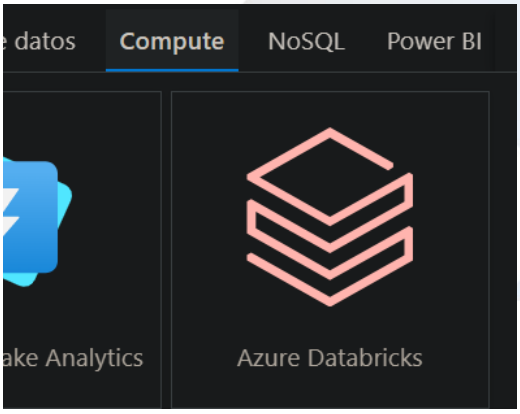
9. Copiamos el token mostrado y lo guardamos por un momento en cualquier lugar



10. Vincularemos el servicio de DATABRICKS en SYNAPSE. Desde Synapse Studio, seleccionamos “Manage / Linked services / + Nuevo”



11. Desde la pestaña “Compute” seleccionamos “Azure Databricks” y damos clic en “Continuar”

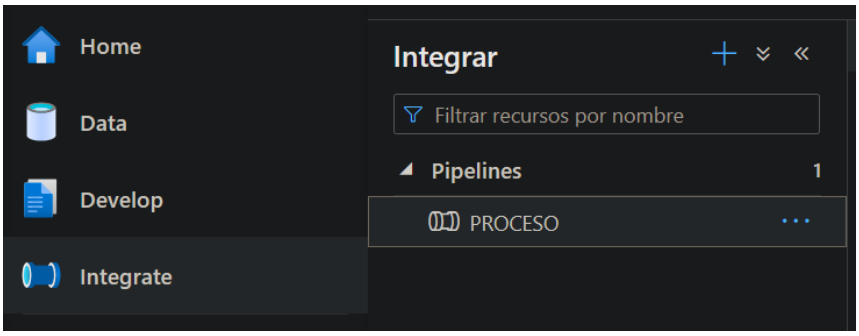


12. Configuramos:

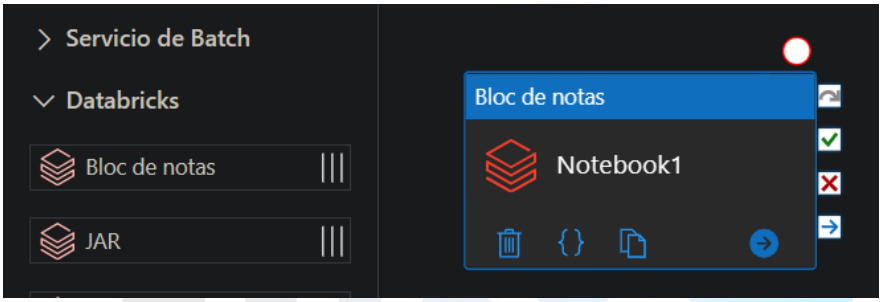
Nombre	LINK_DATABRICKS
Área de trabajo de Databricks	wsdatabricks**XXX**
Seleccionar clúster	Nuevo clúster de trabajo
Token de acceso	**PEGAR_EL_TOKEN_COPIADO**
Versión de clúster	10.4 LTS (includes Apache Spark 3.2.1, Scala 2.12)
Tipo de nodo del clúster	Standard_DS3_v2
Versión de Python	3
Opciones de trabajo	Fixed
Trabajos (número de servidores del clúster)	1

Damos clic en “Crear” (**IMPORTANTE: AL DARLE CLIC A “CREAR” EL CLÚSTER NO SE CREA, SÓLO HEMOS CREADO LA CONFIGURACIÓN, EL CLÚSTER SÓLO SE CREARÁ CUANDO UN PIPELINE EJECUTE CÓDIGO DENTRO DE ÉL Y SE AUTODESTRUYE CUANDO EL PIPELINE FINALICE**)

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



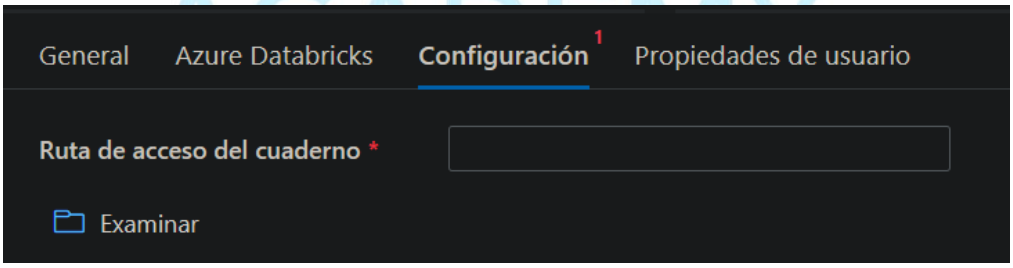
14. Desde el grupo de actividades “Databricks” agregamos “Bloc de notas”



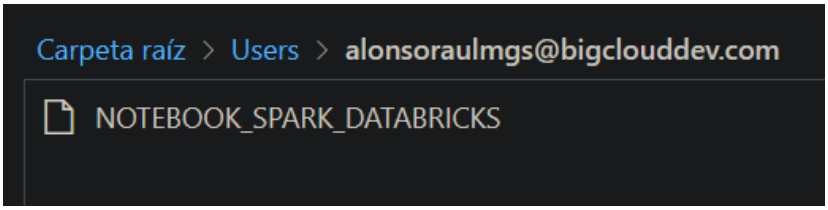
15. Configuramos:

Pestaña	Opción	Valor
General	Nombre	NOTEBOOK_SPARK_DATABRICKS
Azure Databricks	Servicio vinculado de Databricks	LINK_DATABRICKS

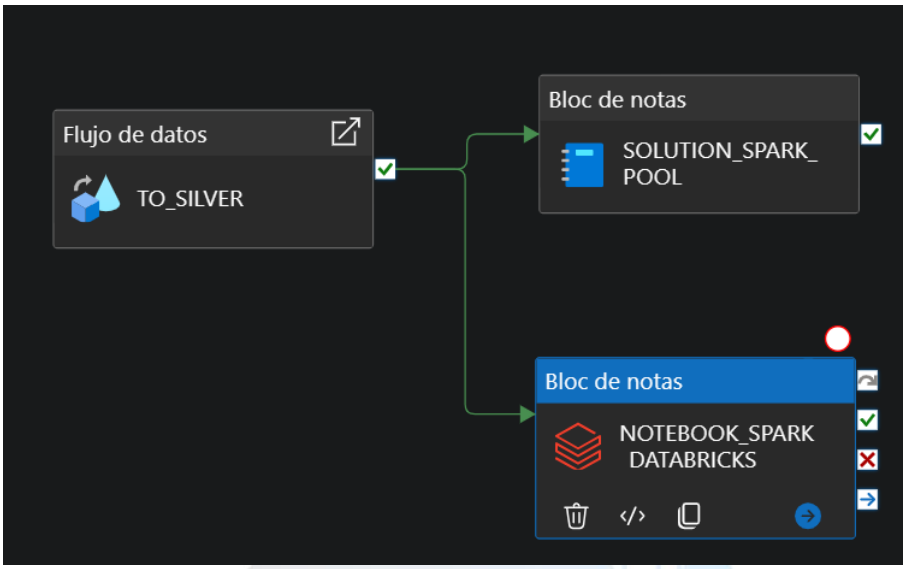
16. En la pestaña “Configuración”, seleccionamos “Examinar” para buscar el notebook



17. Entramos al directorio “/Users/TU_CORREO” y seleccionamos el notebook “NOTEBOOK_SPARK_DATABRICKS”



18. Conectamos las actividades



19. Damos clic en “Validar todo” y luego en “Publicar todo”