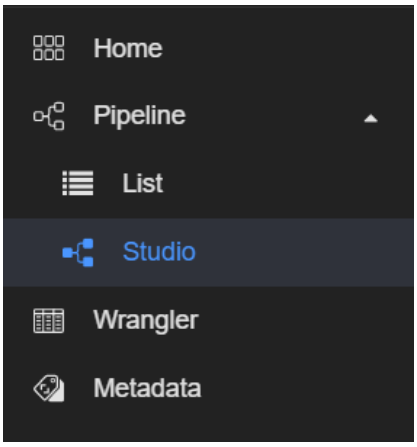
BIG DATA
ACADEMY

LABORATORIO 58
SOLUCIONES DE BIG DATA CON SPARK

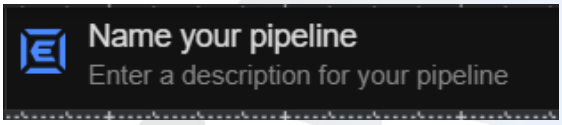
FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

LABORATORIO 58 – SOLUCIONES DE BIG DATA CON SPARK

1. Seleccionamos la opción “Pipeline / Studio”



2. En “Name your pipeline” colocamos el nombre “TO_GOLD” y damos clic en “Save”

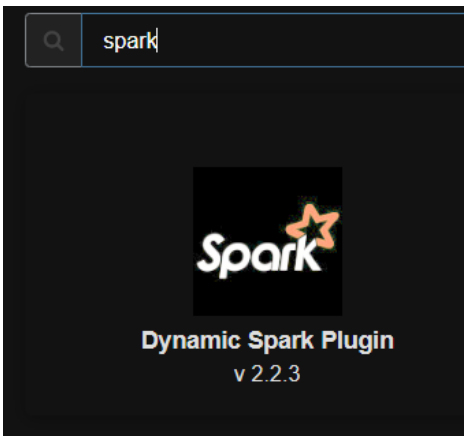


3. El script de la solución está escrito en SPARK, por defecto el componente de SPARK está desactivado, deberemos activarlo desde el HUB. Damos clic en “HUB”

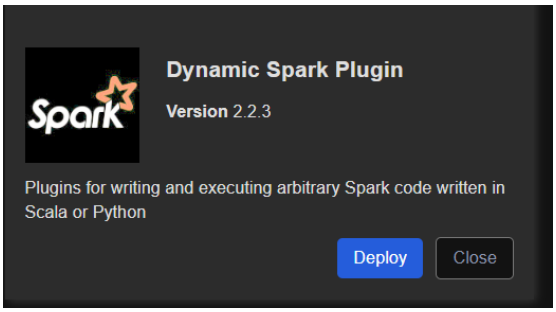
4. Buscamos el componente:

Spark

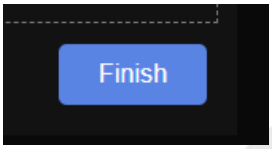
5. Damos clic sobre el componente



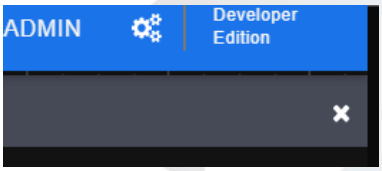
6. Seleccionamos la opción “Deploy”



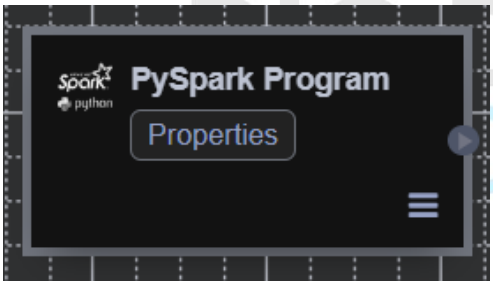
7. Y confirmamos dando clic en “Finish”



8. Damos clic en “x” para volver al graficador



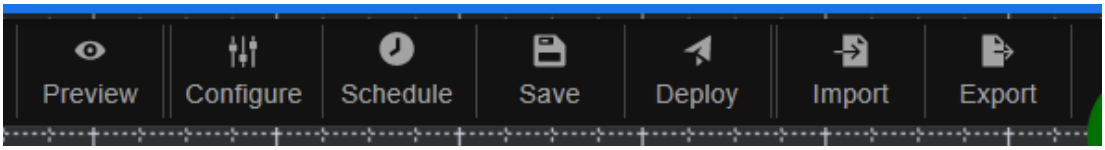
9. Desde la sección “Conditions and Actions” hacemos clic sobre “PySpark Program” para agregar un componente de ejecución de código SPARK. Damos clic en “Properties”



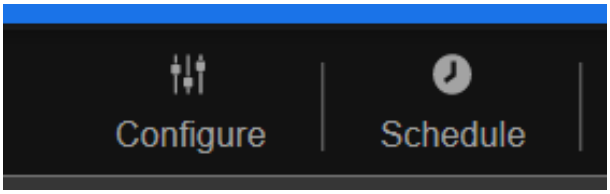
10. En la sección “Script” colocamos el contenido del **“SCRIPT_1.py”**

11. Damos clic en “Validate” para verificar que las configuraciones estén correctas y damos clic en “x” para salir de la configuración.

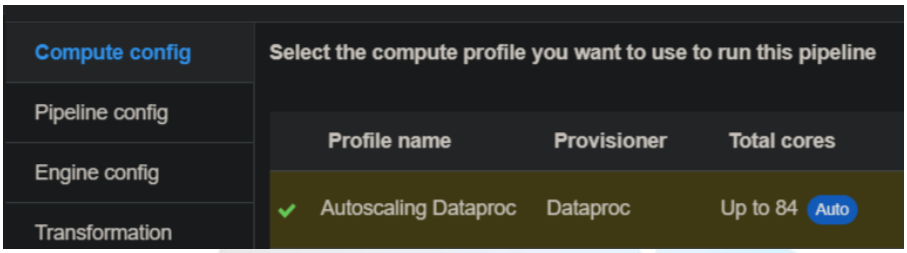
12. Guardamos el job dando clic en “Save” y luego en “Deploy”



13. Configuramos el clúster de Big Data “Dataproc” asociado al proceso. Damos clic en “Configure”



14. Desde la pestaña “Compute config” seleccionamos “Autoscaling Dataproc” para obtener el clúster



Damos clic en “Save”