BIG DATA
ACADEMY

LABORATORIO 57
PROCESO TO_SILVER PARA REGLAS DE CALIDAD

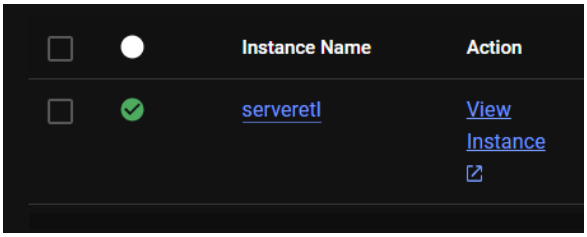
FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

LABORATORIO 57 – PROCESO TO_SILVER PARA REGLAS DE CALIDAD

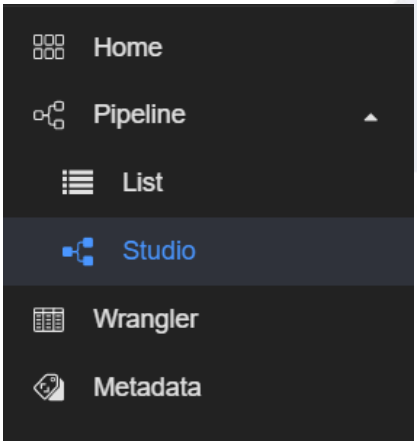
1. Desde el buscador de servicios, buscamos:

Data Fusion

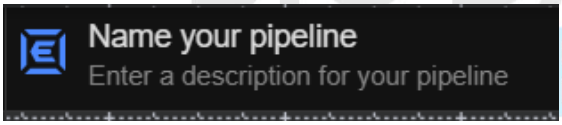
2. Damos clic en “View Instance”



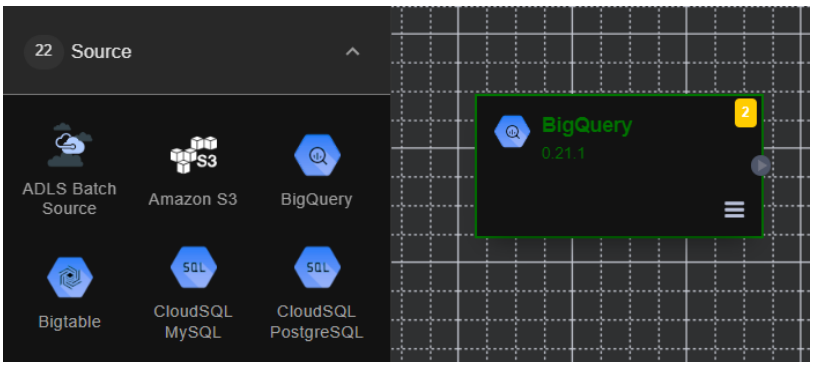
3. Seleccionamos la opción “Pipeline / Studio”



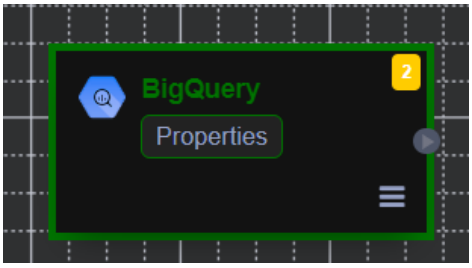
4. En “Name your pipeline” colocamos el nombre “TO_SILVER” y damos clic en “Save”



5. En la sección “Source” hacemos clic sobre “BigQuery” para agregar un componente de lectura de tabla.



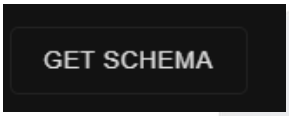
6. Damos clic sobre “Properties”



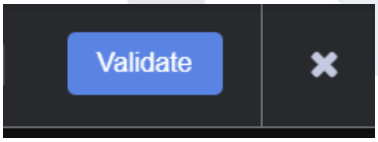
7. Configuramos:

Dataset	bronze
Table	persona

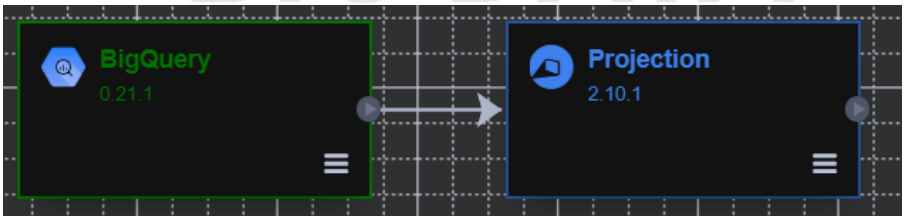
Damos clic sobre “Get Schema” para obtener el esquema de metadatos



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

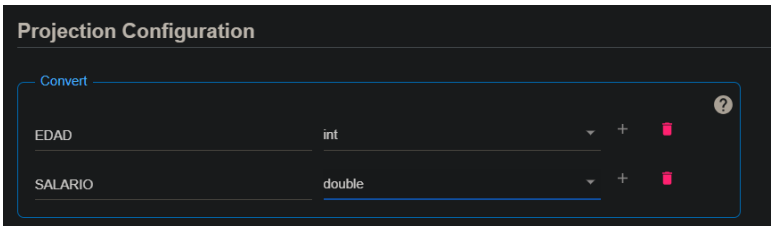


9. Desde la sección “Transform” hacemos clic sobre “Projection” para agregar un componente de casteo de campos y lo conectamos con el componente anterior.

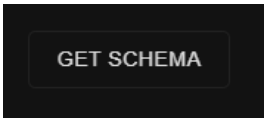


10. Damos clic en “Properties”

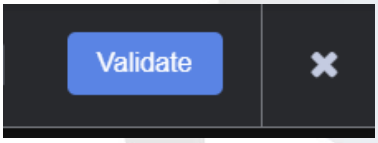
11. En la sección “Convert” casteamos los campos “EDAD” y “SALARIO” a “int” y “double” respectivamente



12. Damos clic en “Get Schema” para definir el esquema de salida



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



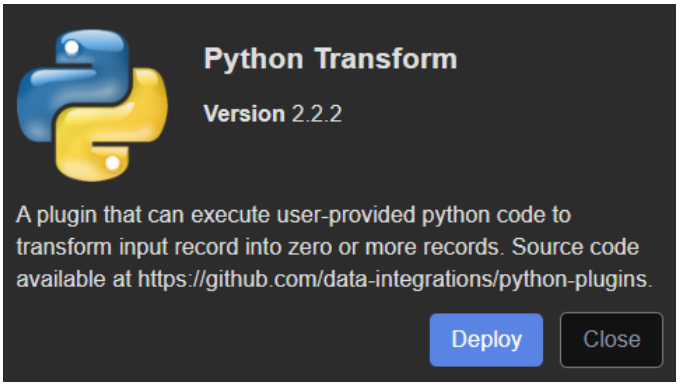
14. Las operaciones de limpieza las realizaremos con código PYTHON, por defecto el componente de PYTHON está desactivado, deberemos activarlo desde el HUB. Damos clic en “HUB”

15. Buscamos el componente:

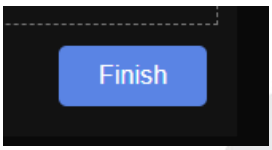
Python

16. Damos clic sobre la última versión disponible del componente

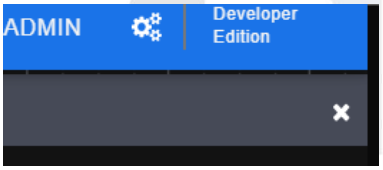
17. Seleccionamos la opción “Deploy”



18. Y confirmamos dando clic en “Finish”

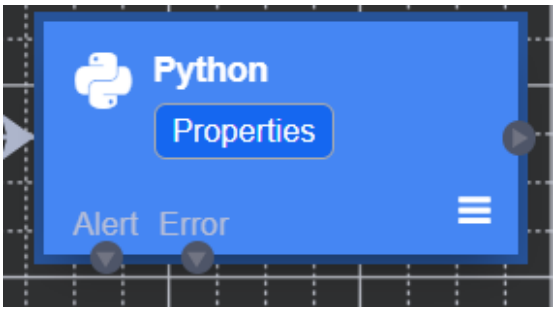


19. Damos clic en “x” para volver al graficador



20. Desde la sección “Transform” hacemos clic sobre “Python” para agregar un componente de limpieza y lo conectamos con el componente anterior.

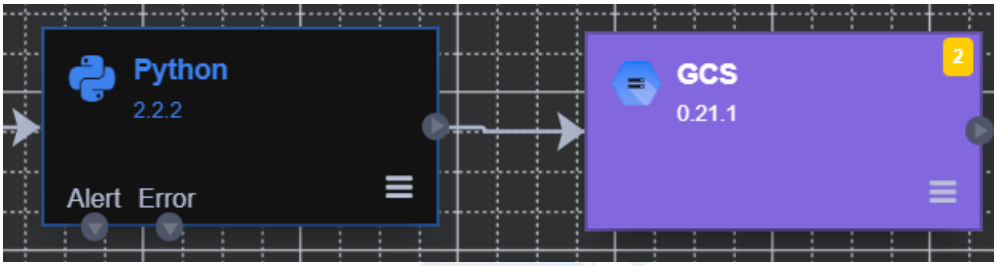
21. Damos clic en “Properties”



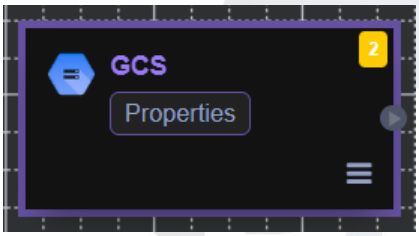
22. En la sección “Script” colocamos el contenido del **“SCRIPT_1.py”**

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

24. En la sección “Sink” hacemos clic sobre “GCS” para agregar un componente de binarización y escritura sobre cloud storage y lo conectamos con el componente anterior.



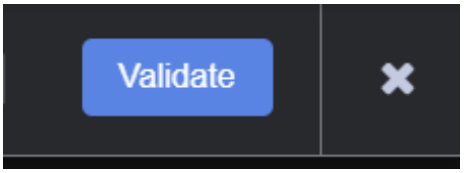
25. Damos clic en “Properties”



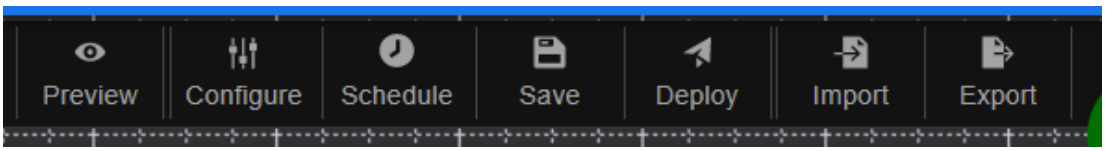
26. Configuramos:

Path	gs://storagebda**XXX**/silver/persona
Path Suffix	yyyy-MM-dd-HH-mm
Format	Parquet

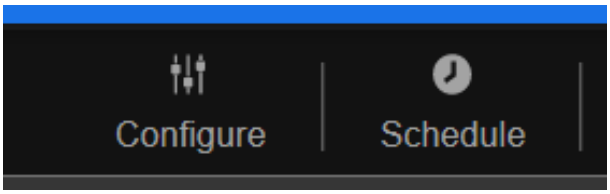
27. Damos clic en “Validate” para verificar que las configuraciones estén correctas y damos clic en “x” para salir de la configuración.



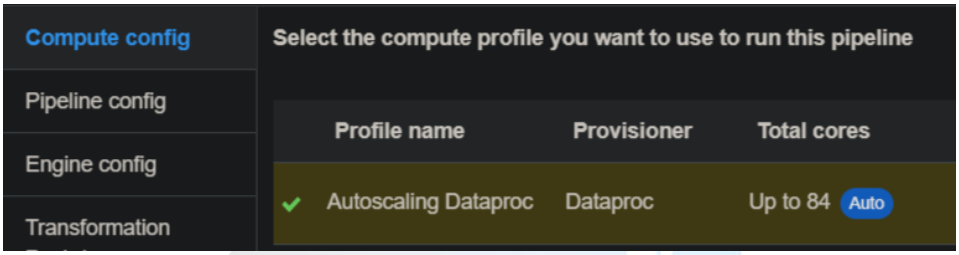
28. Guardamos el job dando clic en “Save” y luego en “Deploy”



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



30. Desde la pestaña “Compute config” seleccionamos “Autoscaling Dataproc” para asignar un clúster en función de la volumetría de los archivos



Damos clic en “Save”