



BIG DATA
ACADEMY

LABORATORIO 29

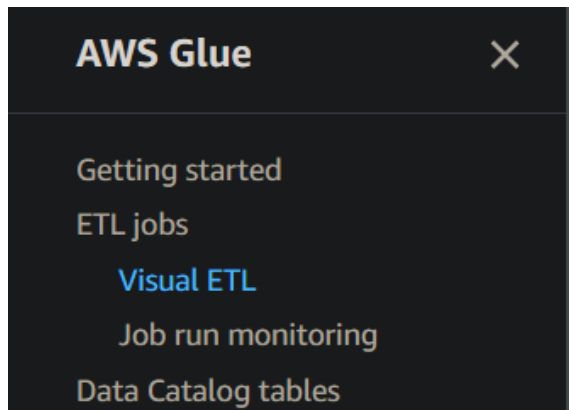
PROCESO TO_SILVER PARA
REGLAS DE CALIDAD CON
GLUE

FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

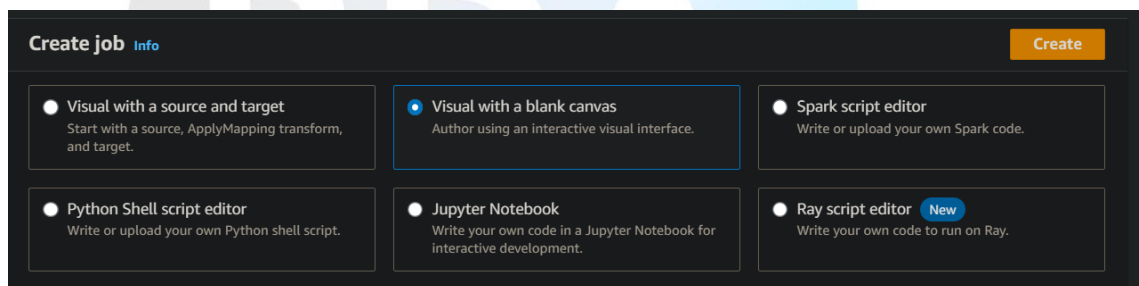
LABORATORIO 29 – PROCESO TO_SILVER PARA REGLAS DE CALIDAD CON GLUE

1. Desde el buscador de servicios, buscamos:

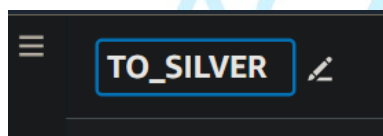
2. Desde la sección “ETL jobs”, seleccionamos “Visual ETL”



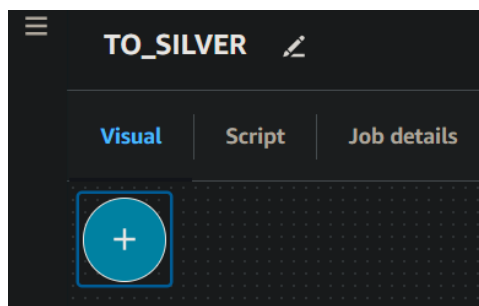
3. Crearemos un job visual, seleccionamos “Visual with a blank canvas” y damos clic en “Create”.



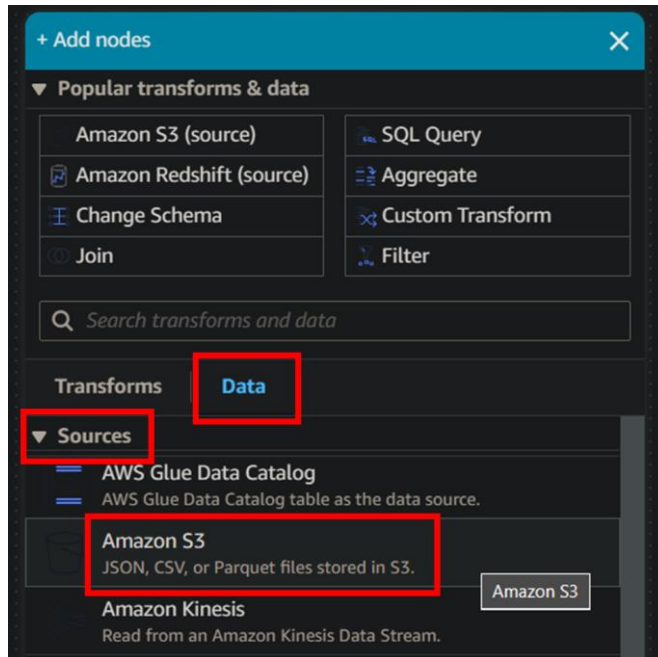
4. De nombre de job colocamos “TO_SILVER”



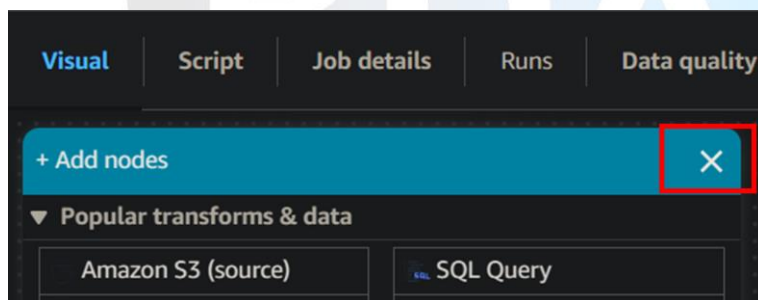
5. Desde la pestaña “Visual” damos clic en “+” para agregar un paso de procesamiento



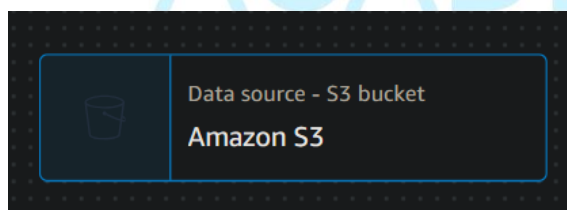
- Agregaremos un paso para leer los datos a procesar. Desde la pestaña “Data”, en la sección “Sources”, seleccionamos “Amazon S3”.



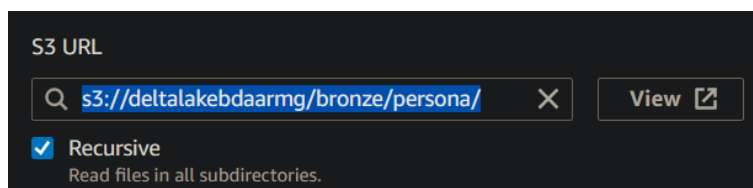
- Damos clic en “X” para cerrar el cuadro de diálogo



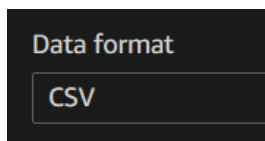
- Damos clic en el paso agregado para configurarlo



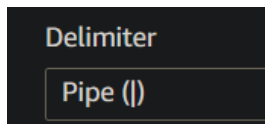
- En “S3 URL” escribimos “s3://deltalakebdaXXX/bronze/persona/”



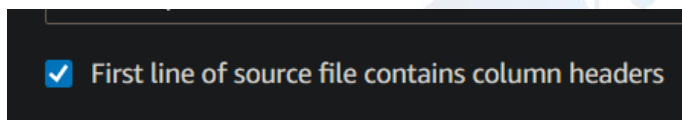
10. En “Data format” seleccionamos “CSV”



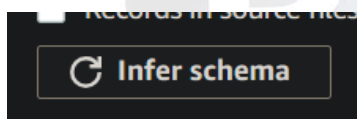
11. En “Delimiter”, seleccionamos “|”



12. Activamos la casilla “First line of source file contains column headers” para indicar que la primera fila del archivo es la cabecera.



13. Damos clic en “Infer schema” para que se obtengan los nombres de las columnas del archivo desde la cabecera



14. Verificamos el esquema de metadatos desde la pestaña “Output schema”, por defecto los campos son colocados como “string”, luego lo solucionaremos.

Data source properties - S3

Output schema

Data preview

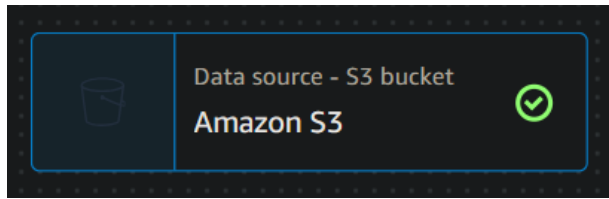
Schema

Info

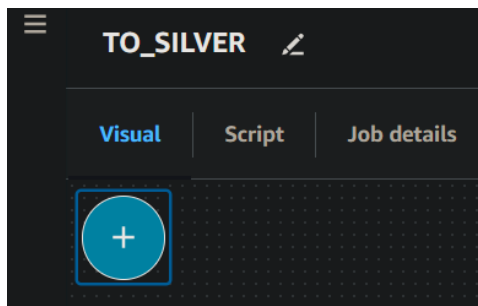
Edit

Key	Data type	Partition
ID	string	-
NOMBRE	string	-
TELEFONO	string	-
CORREO	string	-
FECHA_INGRESO	string	-
EDAD	string	-
SALARIO	string	-
ID_EMPRESA	string	-

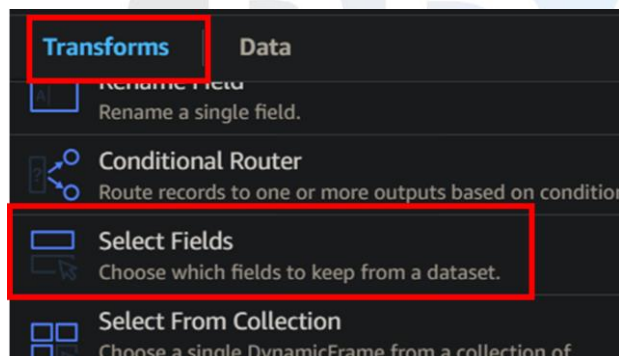
15. Agregaremos un segundo paso de procesamiento, damos clic en el paso de lectura de “S3” para seleccionarlo



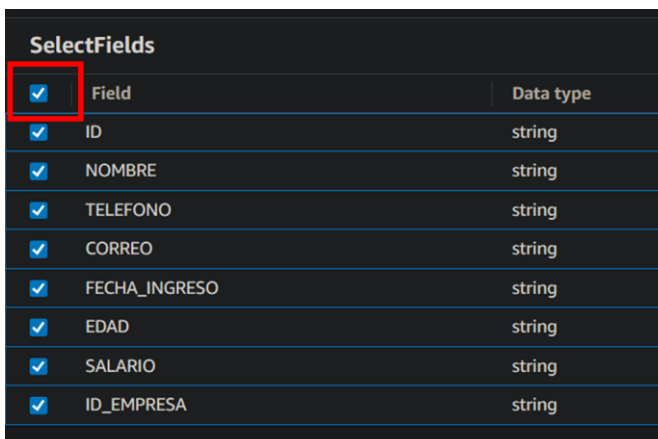
16. Volvemos a seleccionar “+” para agregar un siguiente paso



17. Desde la sección “Transforms”, agregamos el paso “Select Fields”



18. En un escenario real el modelador puede indicarnos procesar sólo algunos campos, para este ejemplo seleccionaremos todos los campos, activamos la casilla para seleccionar todos los campos.

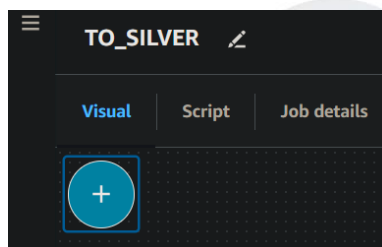


SelectFields		
<input checked="" type="checkbox"/>	Field	Data type
<input checked="" type="checkbox"/>	ID	string
<input checked="" type="checkbox"/>	NOMBRE	string
<input checked="" type="checkbox"/>	TELEFONO	string
<input checked="" type="checkbox"/>	CORREO	string
<input checked="" type="checkbox"/>	FECHA_INGRESO	string
<input checked="" type="checkbox"/>	EDAD	string
<input checked="" type="checkbox"/>	SALARIO	string
<input checked="" type="checkbox"/>	ID_EMPRESA	string

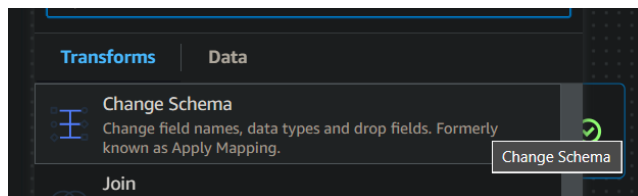
19. Seleccionamos el paso “Select Fields” para agregar un siguiente paso



20. Seleccionamos “+”



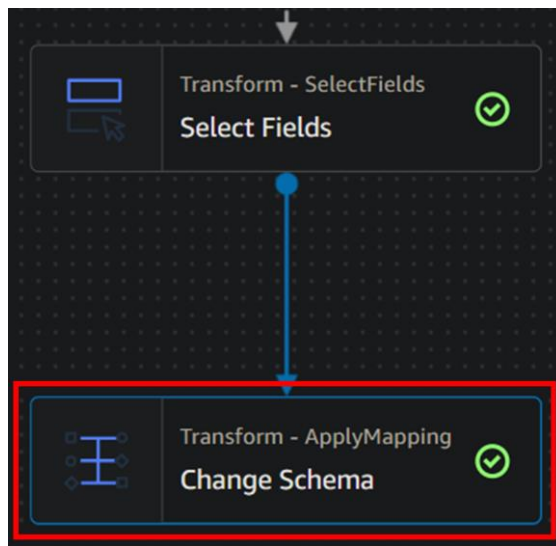
21. Colocaremos los tipos de datos correctos, desde la pestaña “Transforms” seleccionamos “Change Schema”.



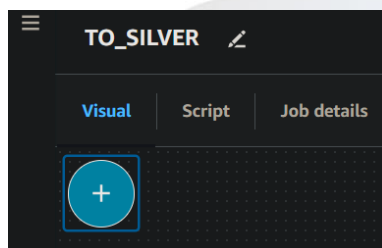
22. Modificamos los campos “EDAD” y “SALARIO” a los tipos “int” y “double” respectivamente.

Change Schema (Apply mapping)			
Source key	Target key	Data type	Drop
ID	ID	string	<input type="checkbox"/>
NOMBRE	NOMBRE	string	<input type="checkbox"/>
TELEFONO	TELEFONO	string	<input type="checkbox"/>
CORREO	CORREO	string	<input type="checkbox"/>
FECHA_INGRESO	FECHA_INGRESO	string	<input type="checkbox"/>
EDAD	EDAD	int	<input type="checkbox"/>
SALARIO	SALARIO	double	<input type="checkbox"/>
ID_EMPRESA	ID_EMPRESA	string	<input type="checkbox"/>

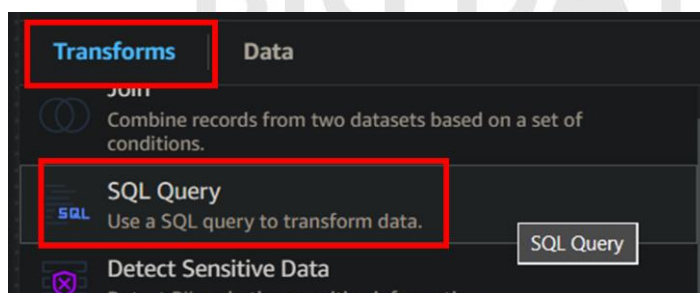
23. Seleccionamos el paso “Change Schema” para agregar un siguiente paso



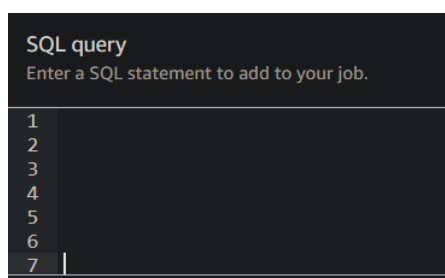
24. Seleccionamos “+”



25. Aplicaremos un paso de limpieza de datos, desde la pestaña “Transforms” seleccionamos “SQL Query”



26. En la sección “SQL query”

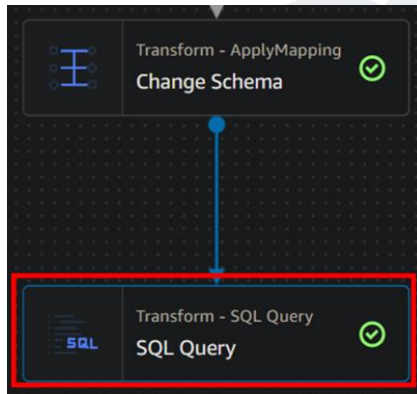


Ingresamos un query para quedarnos con los registros que no tengan identificadores nulos y los campos numéricos mayores a cero:

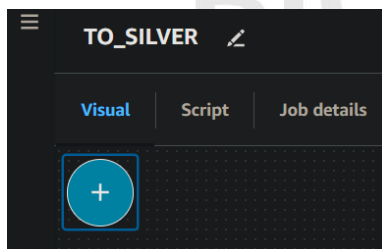
```
SELECT
  *
FROM
  myDataSource
WHERE
  ID IS NOT NULL AND
  ID_EMPRESA IS NOT NULL AND
  EDAD > 0 AND
  SALARIO > 0
```

Donde “myDataSource” hace referencia a los registros entregados por el paso anterior

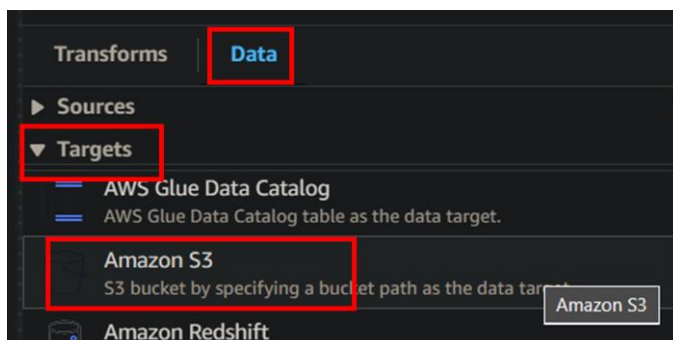
27. Por último, escribiremos el resultado en la zona “SILVER” del “DELTA LAKE”.
Seleccionamos el paso “SQL Query” para agregar un siguiente paso.



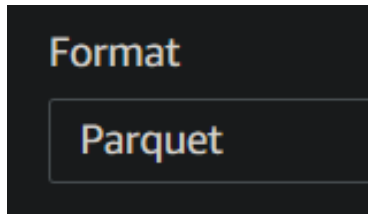
28. Seleccionamos “+”



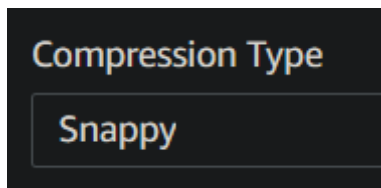
29. En la pestaña “Data”, en la sección “Targets”, seleccionamos “Amazon S3”



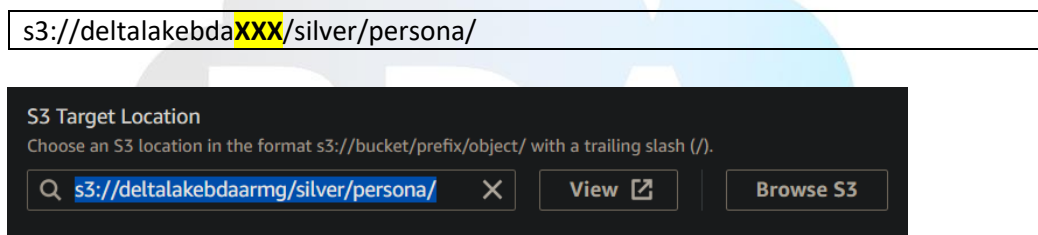
30. En “Format” seleccionamos “Parquet” como formato de escritura



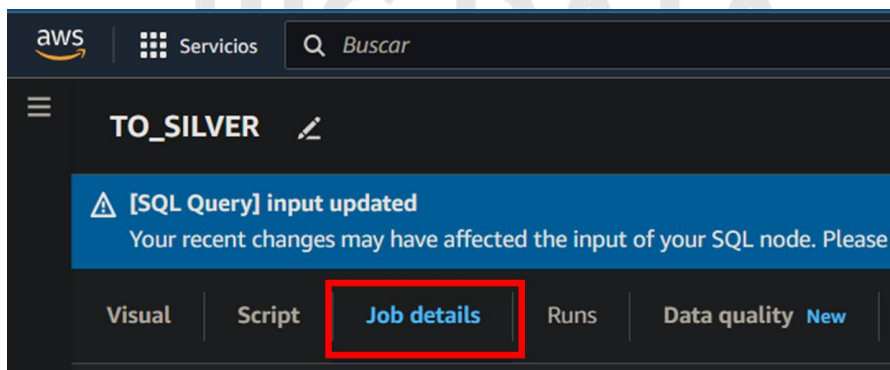
31. En “Compression Type” seleccionamos “Snappy”



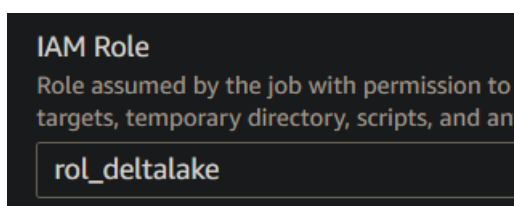
32. En “S3 Target Location” ingresamos el directorio de escritura:



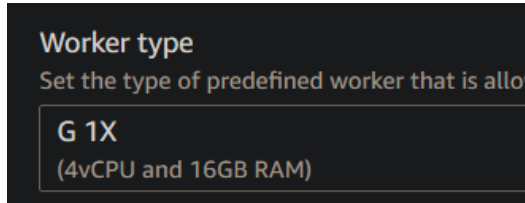
33. Asignaremos el clúster de Big Data para que ejecute el proceso dibujado. Seleccionamos “Job details”



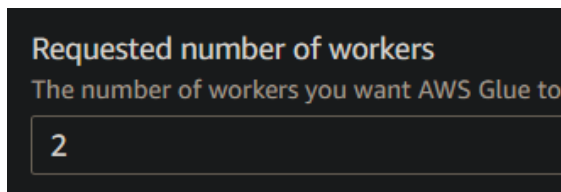
34. En “IAM Role” seleccionamos el rol “rol_deltalake” que tiene los permisos de acceso a “S3”



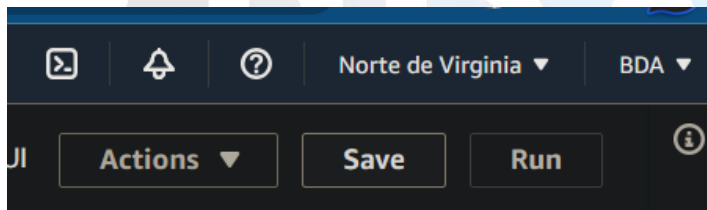
35. En “Worker type” seleccionamos el tipo de servidores que tendrá el clúster, en un escenario real deberíamos seleccionar “G 8X (128 GB RAM | 32 vCPU)”, para no salir de la capa gratuita seleccionamos “G 1X (16 GB RAM | 4 vCPU)”



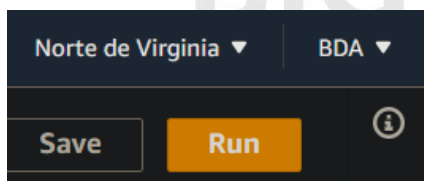
36. En “Requested number of workers” indicamos la cantidad de servidores del clúster, para no salir de la capa gratuita seleccionamos “2”



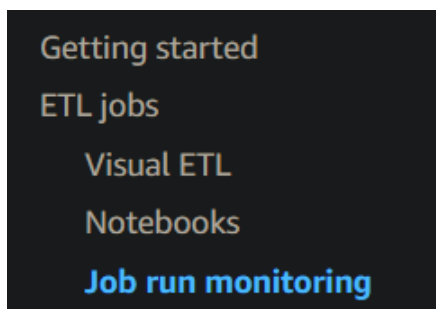
37. Guardamos los cambios del proceso dando clic en “Save”



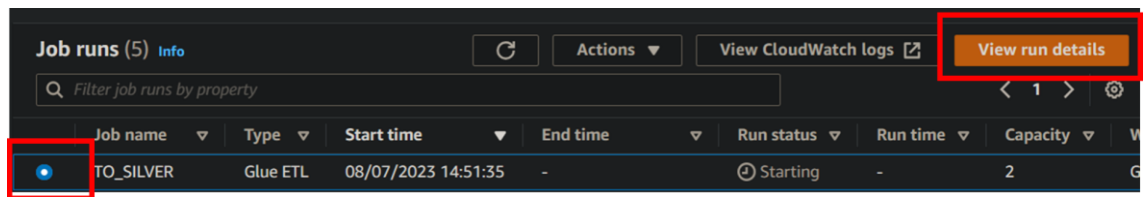
38. Ejecutaremos manualmente el proceso dando clic en “Run”



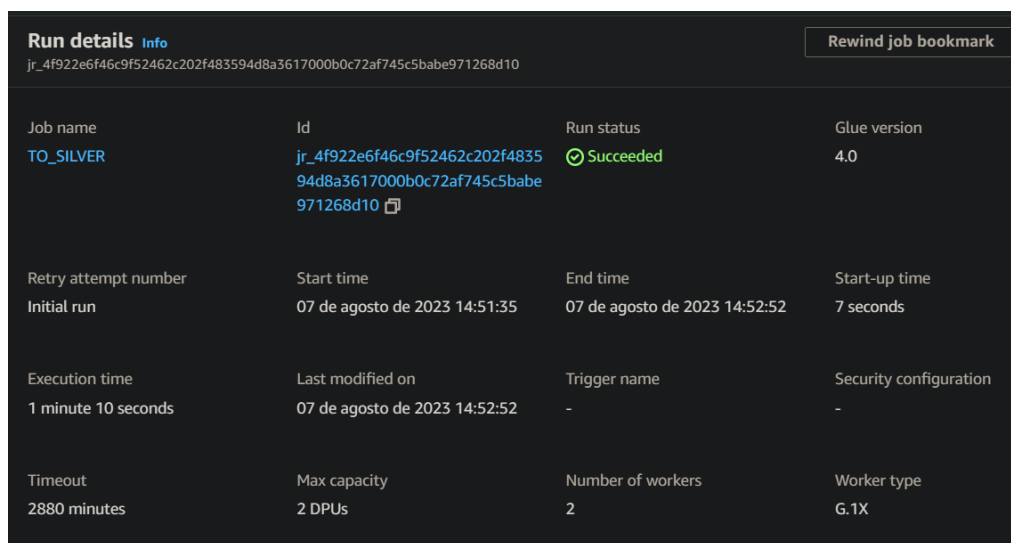
39. Para monitorear el proceso en la sección “ETL jobs” seleccionamos “Job run monitoring”



40. Seleccionamos el proceso “TO_SILVER” y damos clic en “View run details” para ver los detalles de ejecución



41. Desde aquí podemos la hora de inicio y fin del proceso **(TIEMPO: 1 minuto)**



BIG DATA
ACADEMY