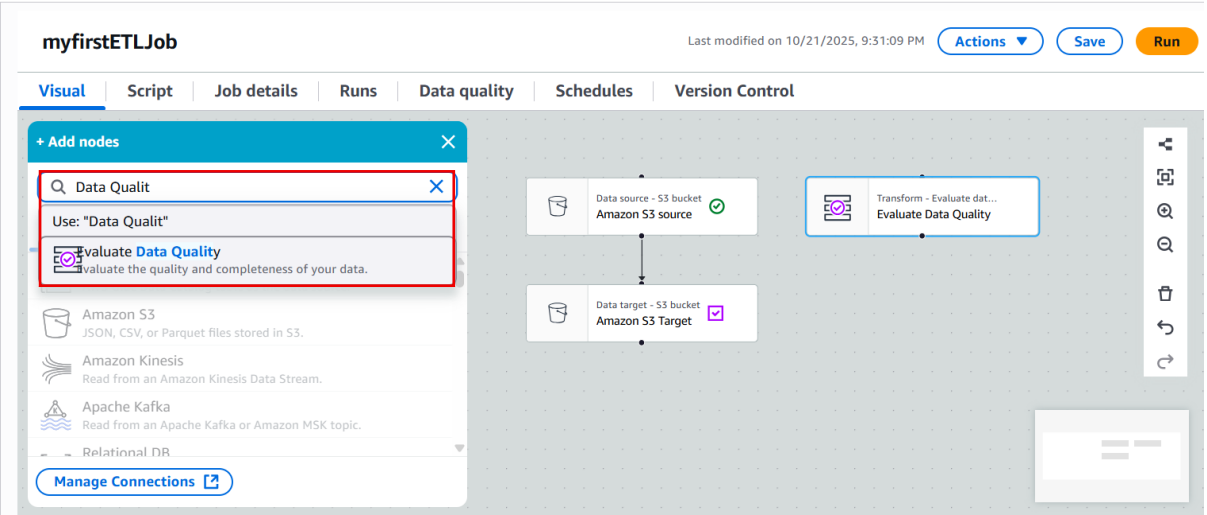
Glue Data Quality

To Begin with the Lab

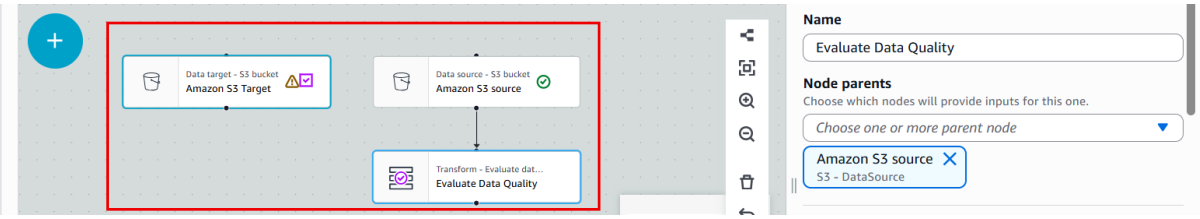
Summary of the Lab

In this lab, you enhance an existing AWS Glue ETL job by adding a **Data Quality evaluation** step to ensure data integrity before loading it to the target. After opening the ETL job in the **Visual ETL editor**, you add an **Evaluate Data Quality** node, create a new **Rule Set**, and define validation rules such as column count, completeness, and data types. You preview and validate these rules, configure output options to store quality reports, and set failure handling to stop the job on rule violations. Finally, you run the job and verify the quality results in S3.

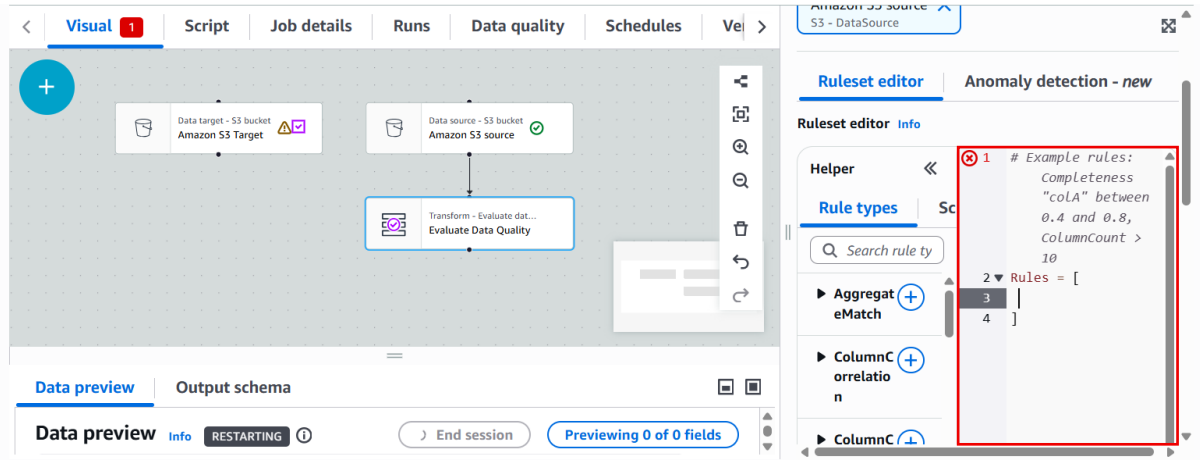
- **Prerequisite**
    - An **S3 bucket** with a documents folder (source data) and a **target** folder (destination).
    - A working **AWS Glue ETL job** that reads data from the source and writes to the target.
    - Proper **IAM permissions** (AWSGlueServiceRole with S3FullAccess).
- Go to the **AWS Console → AWS Glue → ETL Jobs**.
- Locate the ETL job you previously created for moving data from the **documents** folder to the **target** folder.
- Click on the job name to **open** it in the **Visual ETL editor**.
- In the **Visual ETL editor**, click on the **'+' (Add node)** icon.
- Search for **"Evaluate Data Quality"** and select it.
    - This node checks data integrity and applies validation rules.



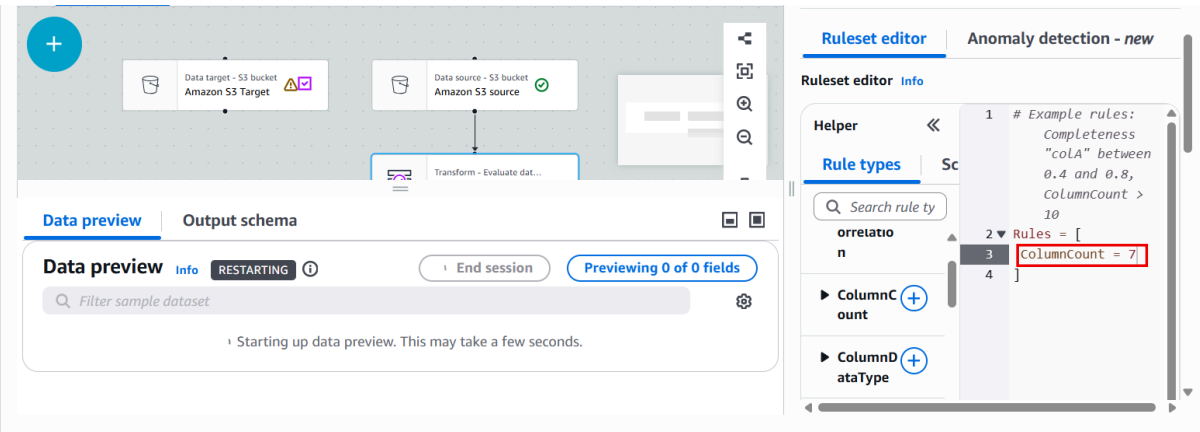
- Connect your **S3 Source** node to the **Evaluate Data Quality** node.
- (Optional) Disconnect your S3 target temporarily to configure the quality node.



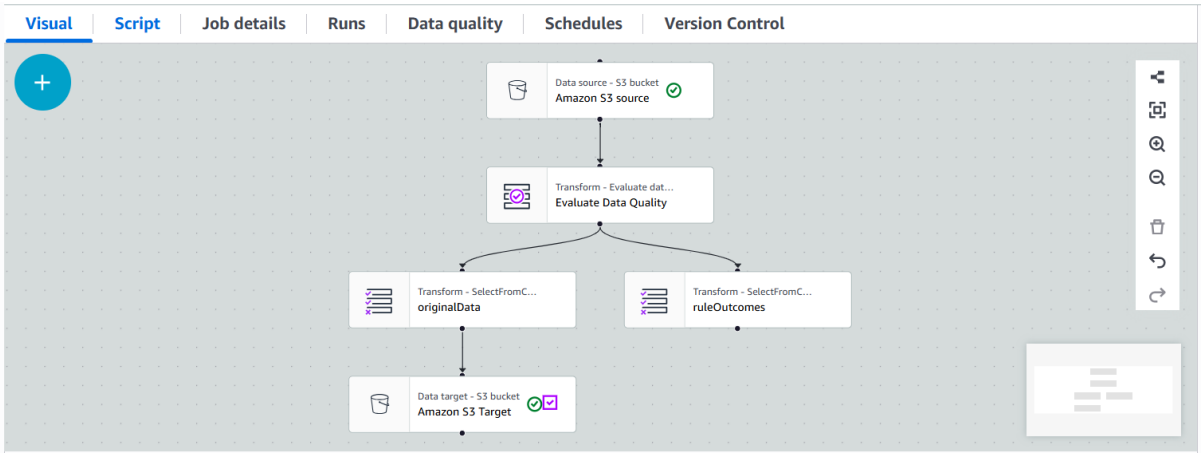
- Click on the **Evaluate Data Quality** node to open its configuration panel.
- Under **Rule Set Editor**, choose **Create new rule set**.
- You'll see two options:
  - **Anomaly Detection (ML-based)** – auto-detects unusual patterns.
  - **Rule Set Editor (manual rules)** – define specific quality checks.
    → **Select Rule Set Editor** for this lab.



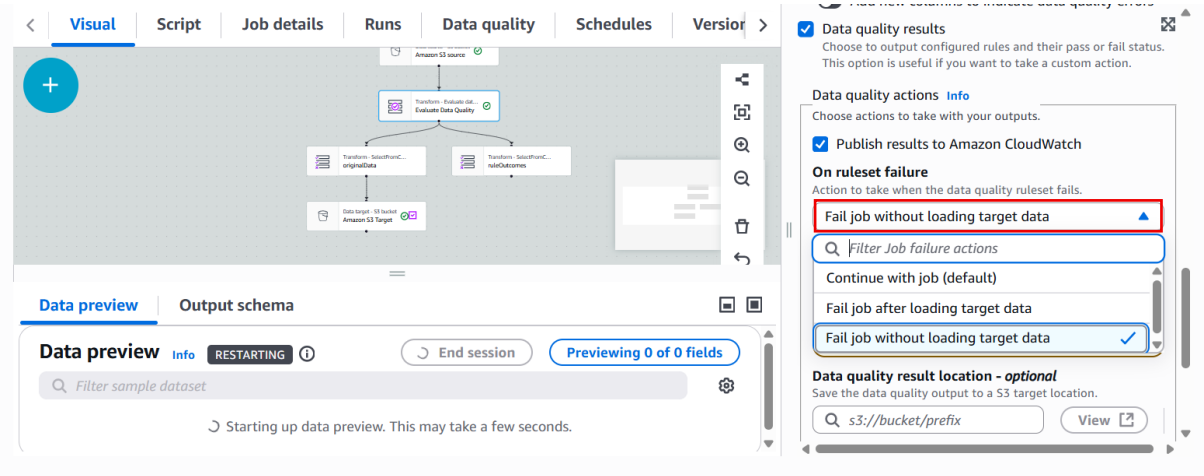
- In the **Rule Set Editor**, click **Add rule**.
- Use the **Rule Helper** to define rules, for example:
  - ColumnCount == 7 → ensures the dataset has 7 columns.
  - Completeness between 0.4 and 1.0 → checks data completeness.
  - ColumnDataType("name", "string") → ensures the "name" column is of type string.
- Click **Preview** to validate rules.
  - If the preview shows **failed**, adjust values (e.g., correct column count).
  - When rules pass, you'll see a green checkmark.



- Scroll down in the configuration panel.
- Under **Output Options**, enable:
    - **Output original data** – keep the source dataset output.
    - **Output data quality results** – output a report of quality checks.
- Choose an **S3 bucket** location for the results (optional).



- Under **Failure Handling**, configure what should happen if rules fail:
    - Select **"Fail job on rule set failure"**.
    - This ensures the ETL job **stops** if data quality checks fail (to prevent bad data loading).



- Click **Run** to execute the ETL job.
- Wait for the job to complete.
- If the job **fails**, review the rule set to identify data quality issues.
- If it **succeeds**, go to your **target S3 folder** to verify: