Gule Workflows

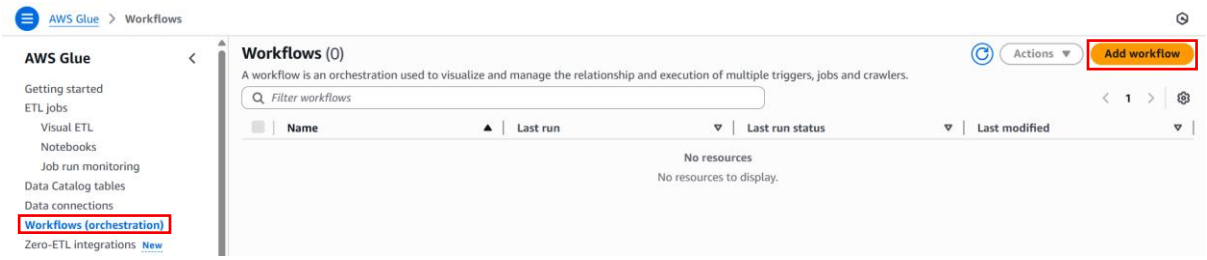
To Begin with the Lab

Summary of the Lab

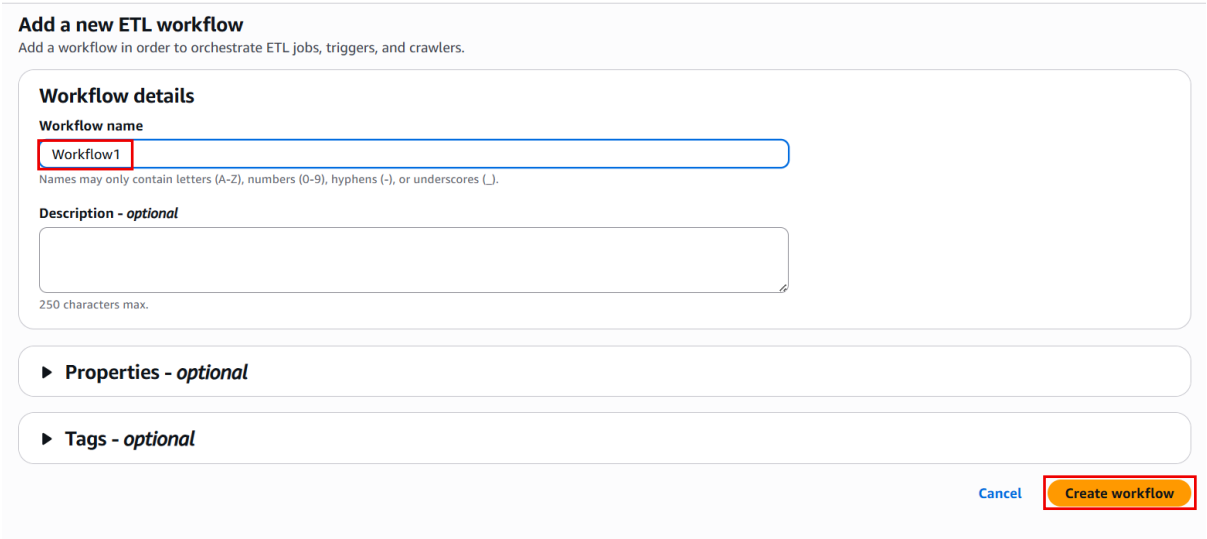
In this lab, you learn how to create and manage **AWS Glue Workflows** to automate and orchestrate multiple Glue components. You begin by creating a new workflow and adding an **On-demand trigger** as the starting point. Then, you add nodes such as **ETL jobs** and **crawlers**, linking them through event-based triggers to control execution order. Each trigger can be configured to activate after specific job outcomes, like "Job succeeded." Finally, you can manually run the workflow, monitor execution statuses (Succeeded, Failed, Running), and verify the outputs in S3 or the Glue Data Catalog. This demonstrates seamless job orchestration.

Note: This ensures no scheduled runs or resources continue to execute and incur charges. So we won't run this. This is just demonstration purpose.

- Go to the **AWS Console**.
- Navigate to **AWS Glue → Workflows**.
- Click **Add workflow**.

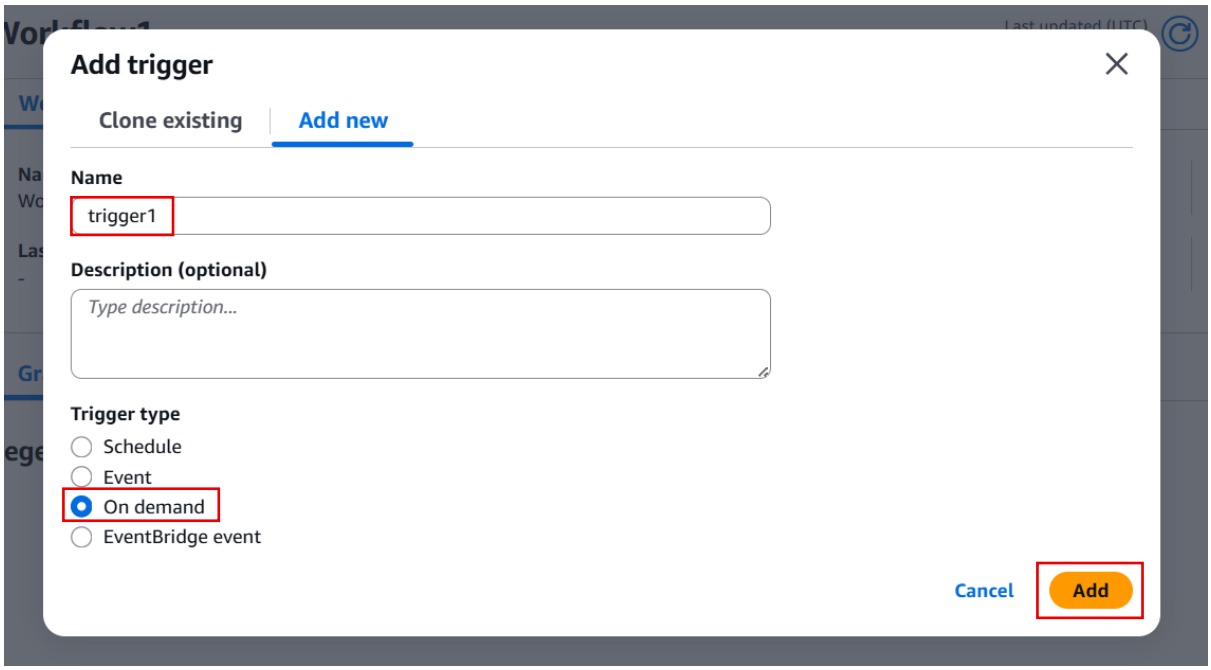


- Enter a **Workflow Name** (e.g., workflow1).

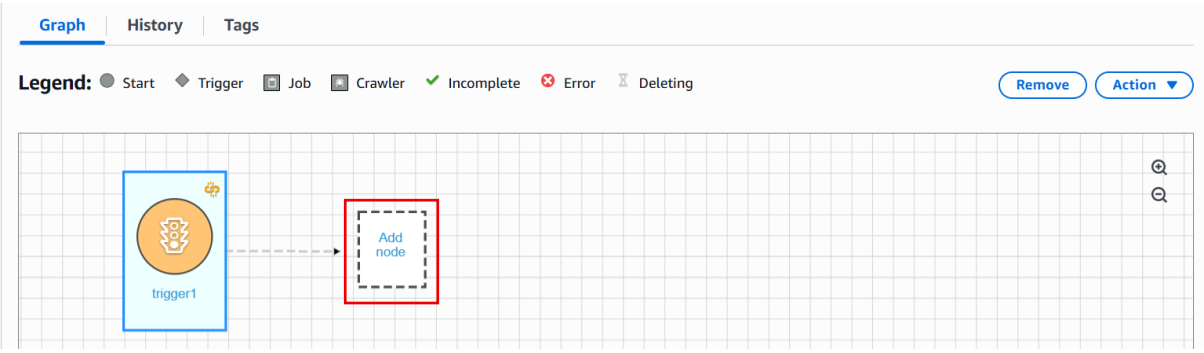


- Click **Create workflow**.
- Every workflow must start with a **trigger**.

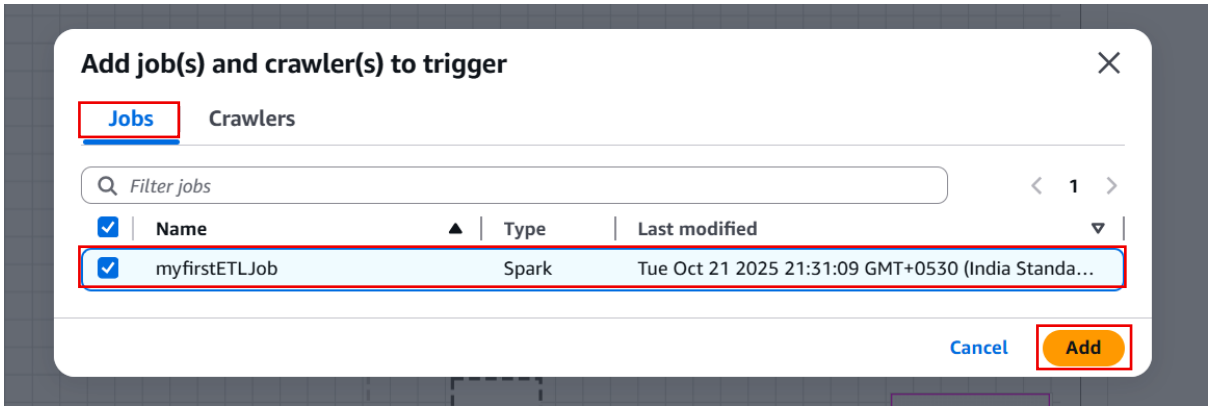
- Click **Add trigger** on the canvas.
- In the setup window:
    - Give the trigger a **name** (e.g., trigger1).
    - Choose the **trigger type**:
        - **On demand** – manually start the workflow.
        - **Scheduled** – run automatically at a specific time (daily, weekly, or cron).
        - **Event-based** – triggered after another job/crawler event.
    - For this lab, select **On demand**.
- Click **Add** to create the trigger.



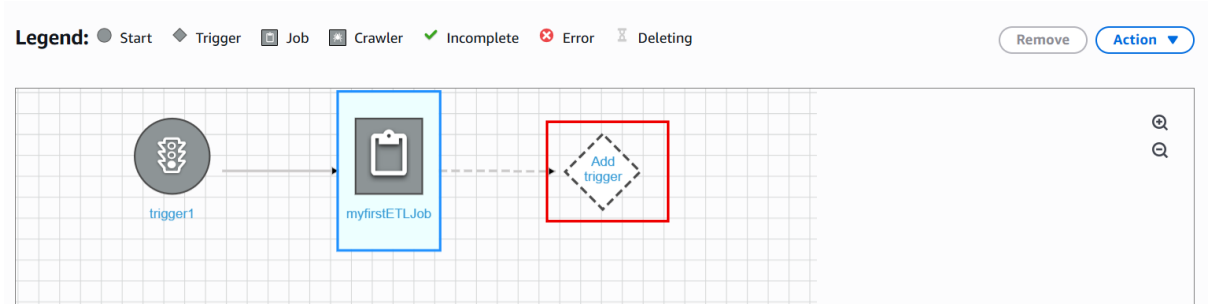
- Now we will select the our ETL Job
- Click on Add node.



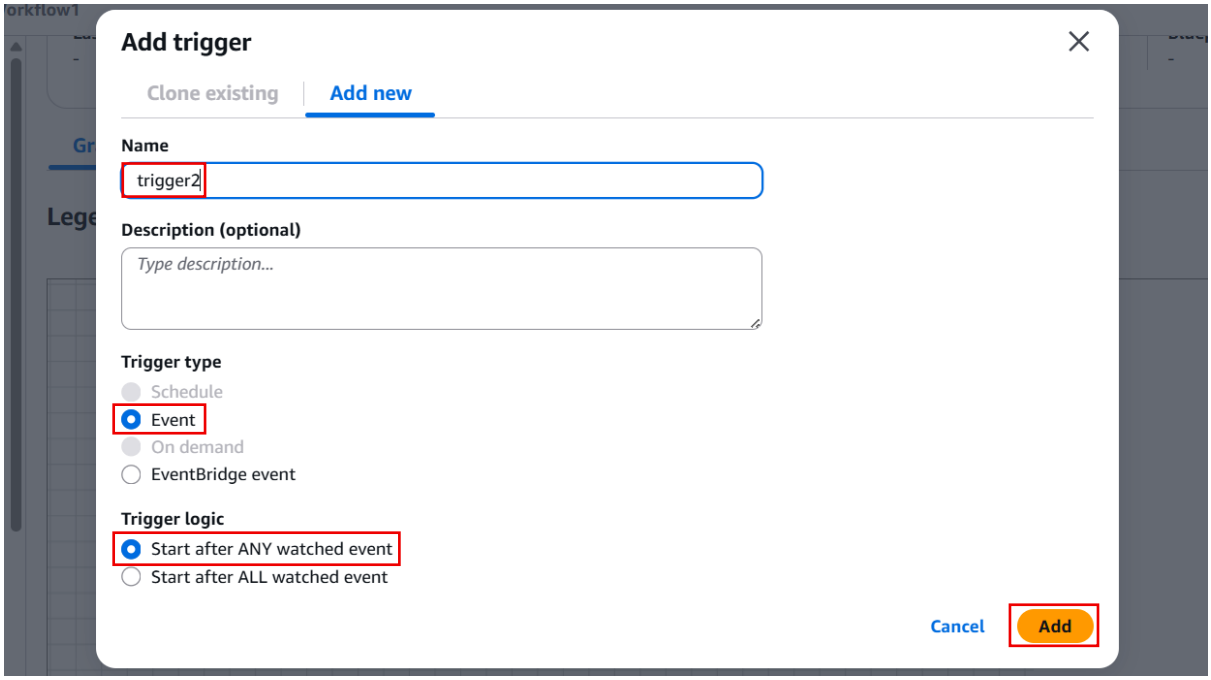
- We can select the Job and Crawler both



- To chain another component, select the **job node** you added.

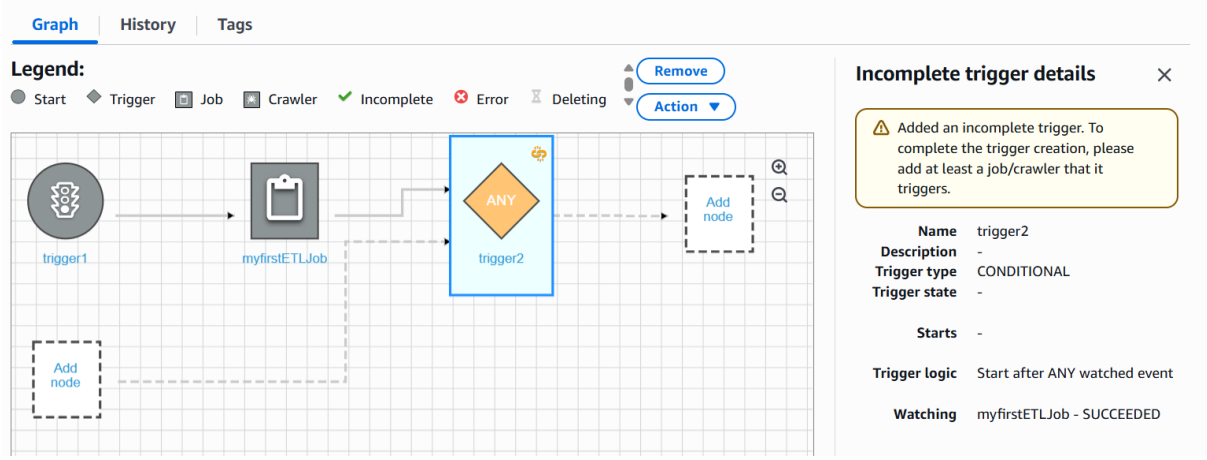


- Click **Add trigger → Event trigger**.
- Name the trigger (e.g., trigger2).

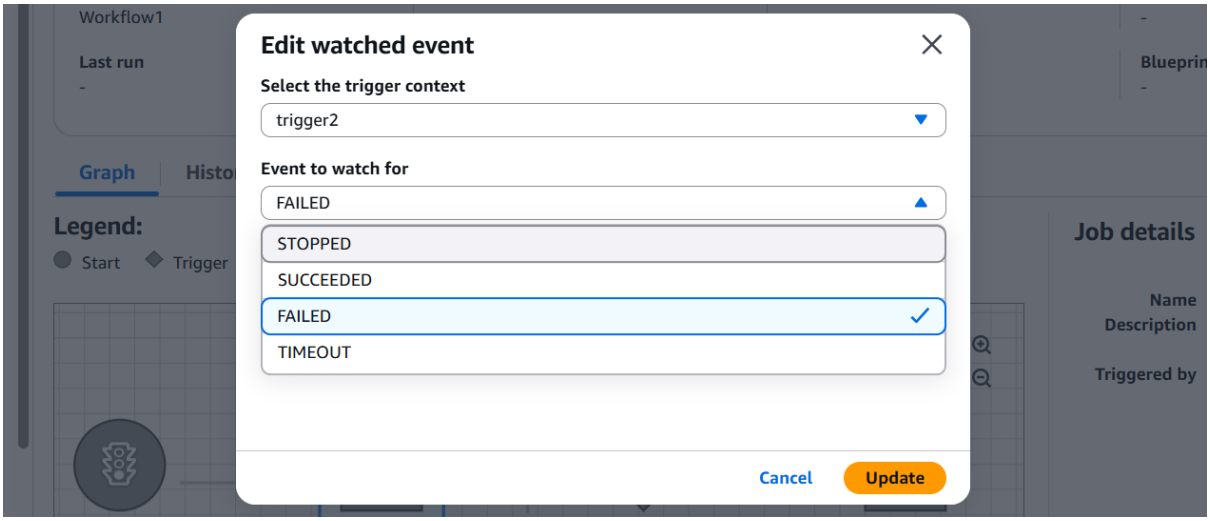


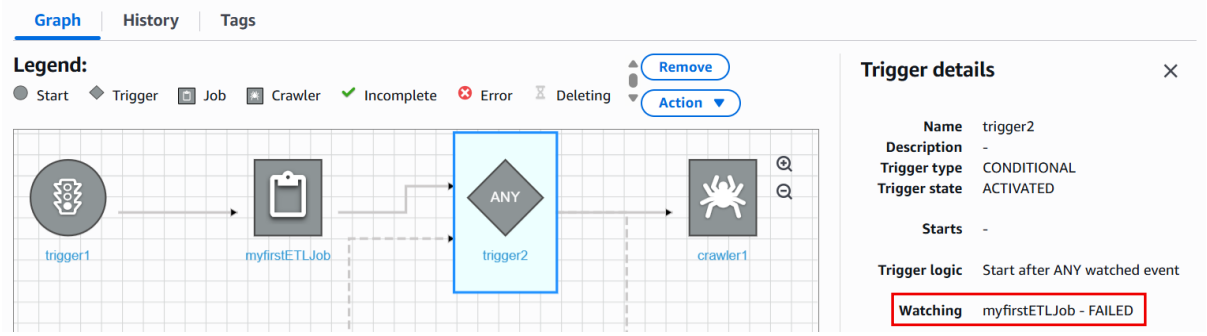
- Under **Trigger Logic**, choose when it should activate:
  - **Start after ANY watched event** (if you have multiple dependencies).
  - **Start after ALL watched events**.
- Under **Watched event**, select **Job succeeded** (so it only runs on success).
- Click **Add**.

- Now we will add a crawler.
- Click on Add node.



- Select any **trigger node** to view its configuration.
- You can change:
  - **Watched events** (e.g., trigger on job failed instead of succeeded).
  - **Trigger conditions** (any vs all watched events).
- Save your changes after modifying.





- Click **Run workflow** (if On-demand trigger is selected).
- Monitor the **workflow runs** in the **Runs tab**.

- You’ll see each job and crawler’s execution status (Succeeded, Failed, or Running).
- Once complete, check your **S3 output** or **Glue Data Catalog** for updated results.