Proposal: Smart Autoscaling Governance for AWS Glue Jobs

# Objective

Implement a serverless automation framework to enforce autoscaling on all newly created AWS Glue jobs and monitor configuration changes that may impact performance, cost efficiency, or compliance. Stakeholders will be notified via email when such changes occur to ensure proactive governance.

# Current Challenge

Currently, there is no policy in place enforcing the use of autoscaling for AWS Glue jobs. This means job creators can manually configure fixed worker counts without accountability or cost awareness. As a result:

• Jobs may be overprovisioned, leading to inflated Glue costs.

• Teams lack visibility into when critical job settings (e.g., worker type, autoscaling) are modified.

• There is no centralized enforcement or alerting mechanism tied to resource optimization or owner responsibility.

# Proposed Solution

We propose developing a lightweight, automated governance layer with the following features:

1. 1. Autoscaling Enforcement

- Detect creation of new Glue jobs using AWS EventBridge + CloudTrail.  
 - A Lambda function automatically updates jobs to use:  
 - WorkerType: G.1X or G.2X  
 - EnableAutoScaling: true with NumberOfWorkers mode  
 - Ensures cost-efficient configuration from the start.

1. 2. Configuration Change Monitoring

- Detect updates to existing Glue jobs (via the UpdateJob API).  
 - A second Lambda function compares key settings (e.g., autoscaling status, worker type).  
 - If significant changes are detected (e.g., autoscaling disabled), an email is sent via Amazon SES to the job owner based on the NotifyEmail tag.

1. 3. Tag-Based Ownership & Alerting

- Job owners are identified through a standard NotifyEmail tag.  
 - Real-time alerts ensure accountability for any changes that could lead to cost increases or performance degradation.

# Cost Benefits of Enforcing Autoscaling

Autoscaling helps optimize Glue job execution by dynamically adjusting the number of workers during runtime. This leads to:

1. 1. Optimized Compute Usage

- Avoid overprovisioning with fixed worker counts.  
 - Pay only for the compute resources actually used.

1. 2. Reduced Idle Time

- Autoscaling releases excess workers during low-demand phases, lowering overall DPU-hour consumption.

1. 3. Better Handling of Variable Loads

- For jobs with fluctuating data volumes or processing times, autoscaling ensures resources match real-time requirements — reducing cost and runtime.

1. 4. Performance Without Overspend

- Jobs can scale up temporarily to meet demand and complete faster.  
 - While peak usage may be higher, the total DPU-minutes consumed can be significantly lower.  
  
 Example:  
 - A job with fixed 10 workers running for 30 minutes = 5 DPU-hours  
 - With autoscaling (averaging 7 workers) finishing in 15 minutes = ~1.75 DPU-hours  
 - Result: Lower cost and faster execution.

1. 5. Governance Against Cost Regressions

- Automated alerts ensure that disabling autoscaling or increasing resource limits triggers owner review.  
 - Prevents silent configuration changes that could increase monthly costs.

# Outcome

By implementing this governance framework, we will:  
• Ensure all Glue jobs start with cost-efficient, scalable configurations.  
• Provide transparency and accountability for configuration changes.  
• Improve resource efficiency and reduce manual oversight.  
• Deliver measurable cost savings across ETL pipelines.