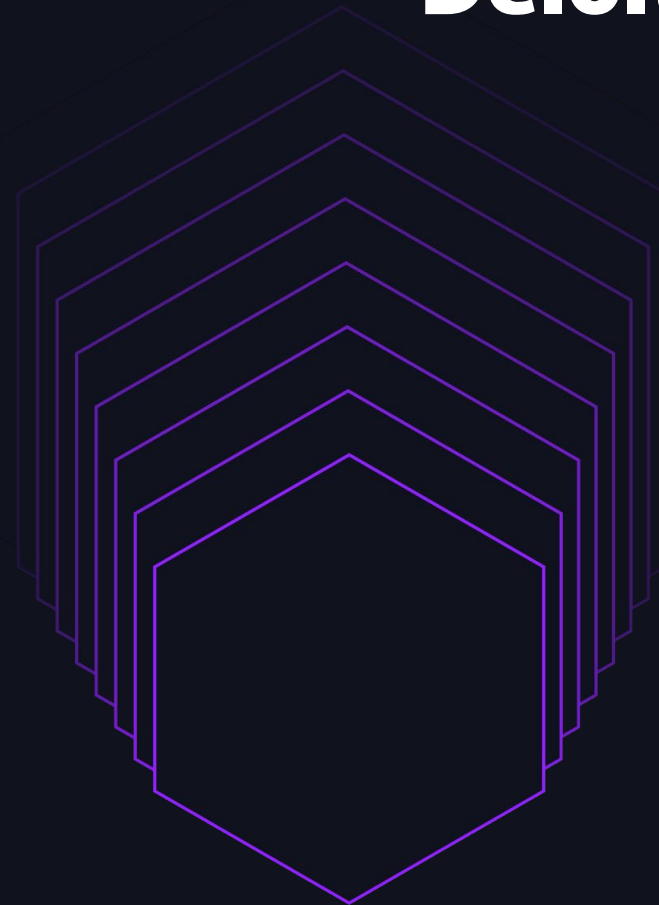


STREAMLINED MONITORING OF DATABRICKS WORKFLOWS WITH DELOITTE



Ashvic Godinho
June 12th, 2024

CENTRALIZED ANALYTICS DATASTORE FOR MONITORING

Current State

- Data Pipelines have become more complex
- More data workflows created over time
- Databricks adds new features often

Compute

All-purpose compute Job compute SQL warehouses Vector Search Pools Policies

Workflows

Jobs Job runs Delta Live Tables

** Databricks ETL features have grown*

Challenges

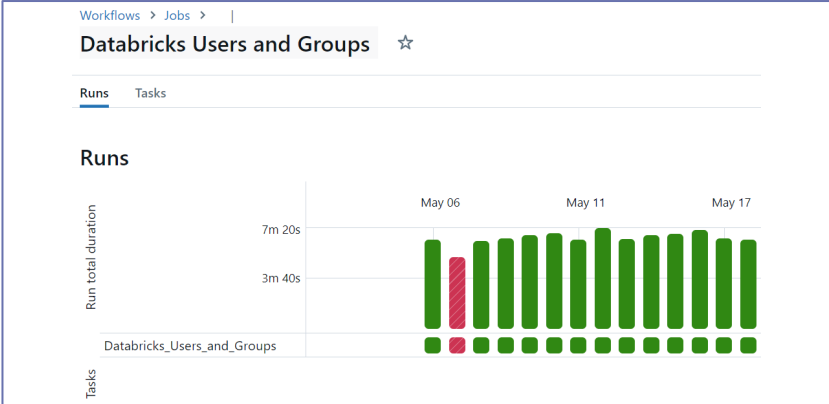
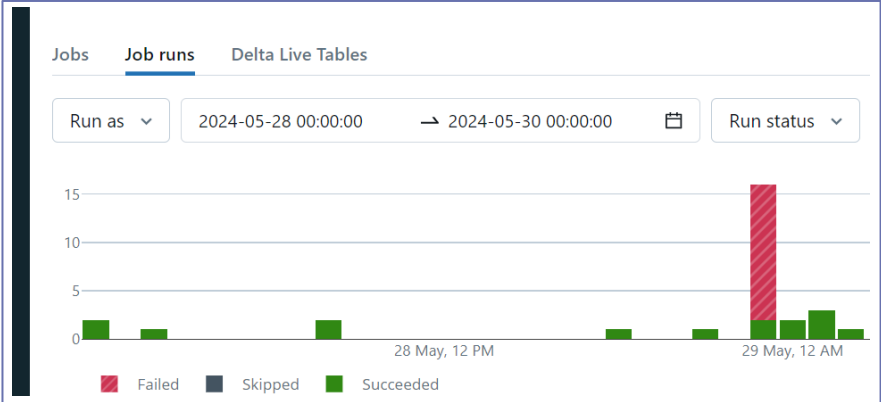
- Time spent investigating issues
- API and Meta data can be transient
- Monitoring for anomalies is hard
- Customer's ecosystem don't always integrate with Databricks easily

MONITORING IN DATABRICKS

What tools are out there currently?



Jobs



Good overview, but limited or manual

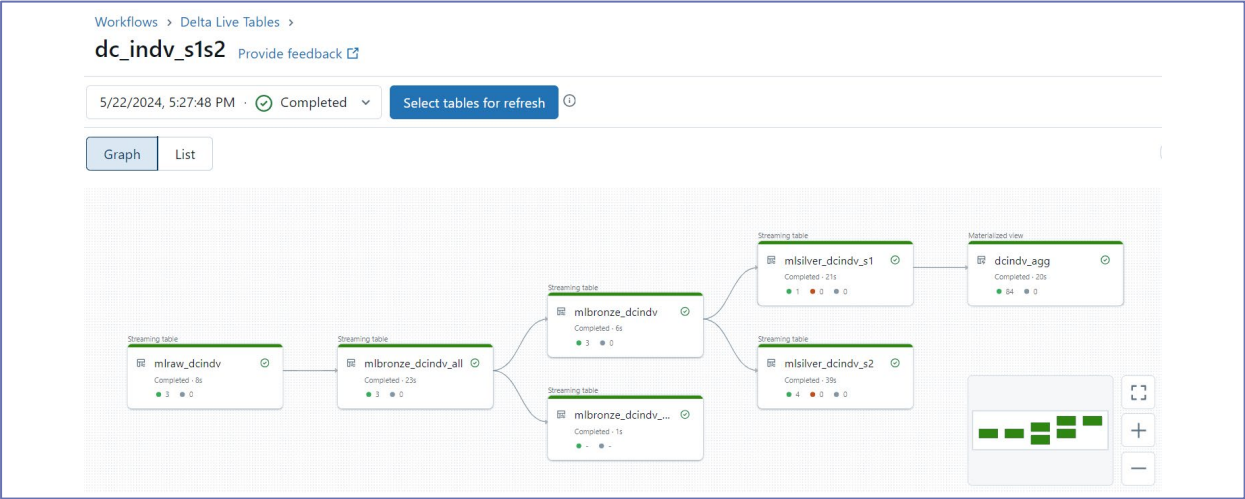


MONITORING IN DATABRICKS

What tools are out there currently?



Delta Live Tables



Detailed, but not persistent

MONITORING IN DATABRICKS

What tools are out there currently?



All-Purpose Clusters

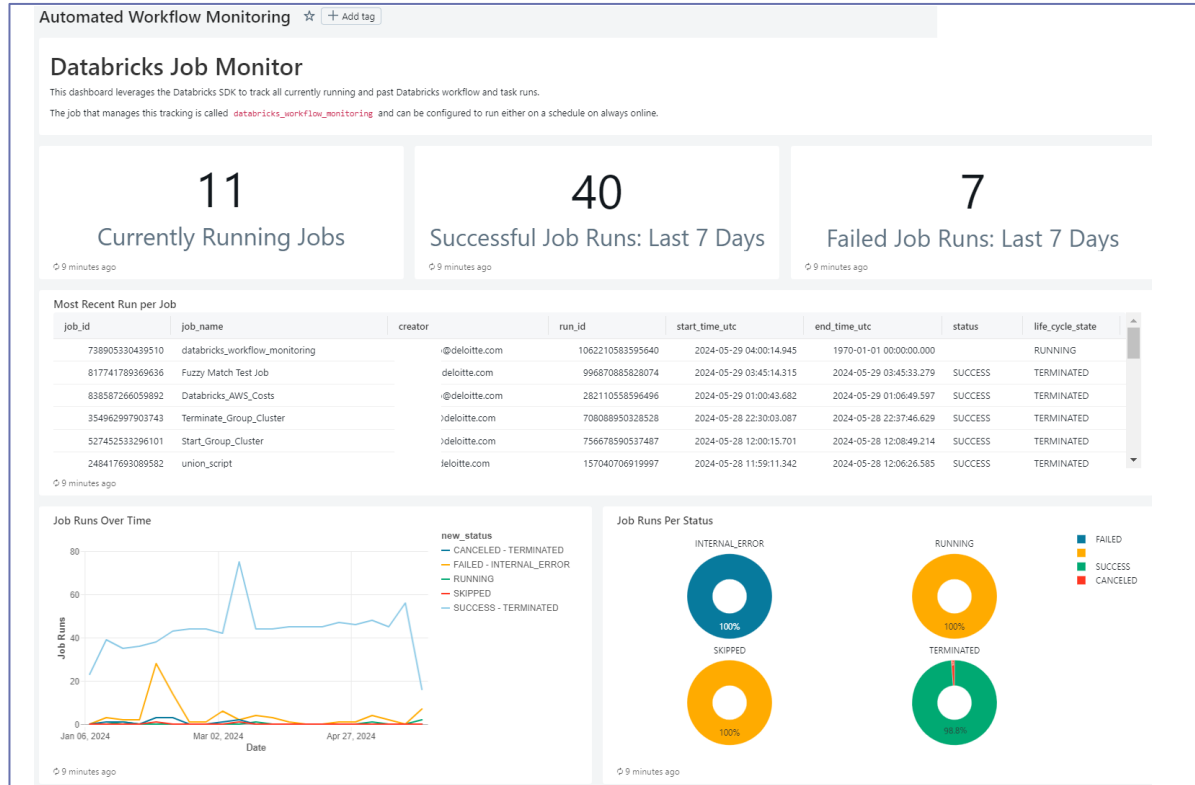
State	Name	Policy	Runtime
	Ash GPU Cluster	databricks-cop-admins clu:	14.3 ML
	Ashvic Godinho's Cluster	databricks-cop-admins clu:	13.3
	genai_group_cluster	genai_group cluster	13.2 ML
	databricks-cop-admins-cluster	databricks-cop-admins clu:	13.2
	demo_group_cluster	demo_group cluster policy	14.3

All manual



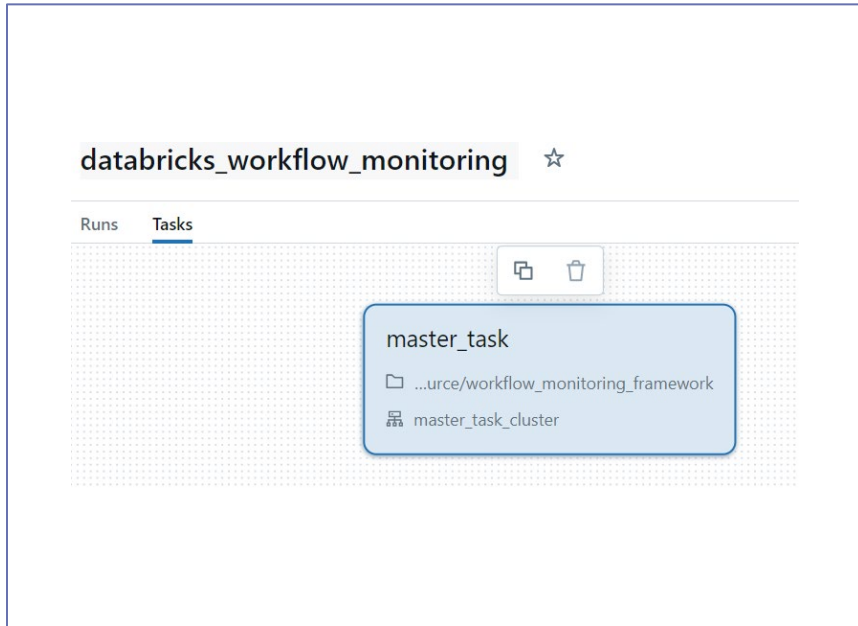
AUTOMATED MONITORING SOLUTION

Single Dashboard to view all Databricks Workflows



AUTOMATED MONITORING SOLUTION

Single Job to collect all Databricks Workflow data



The screenshot displays the Databricks interface for a workflow named "databricks_workflow_monitoring". The "Tasks" tab is active, showing a task named "master_task". The task's source code path is indicated as "...urce/workflow_monitoring_framework" and the cluster used is "master_task_cluster".

PYTHON

```
def processAll():  
    """  
    Runs all the necessary processing  
    """  
    if config['workflow_tracking'] == 'Y':  
        processWorkflowMonitoring()  
    if config['all_purpose_cluster_tracking'] == 'Y':  
        processClusterMonitoring()  
    if config['dlt_tracking'] == 'Y':  
        processDLTMonitoring()
```



DEMONSTRATION



CUSTOMIZABLE JOB

Configuration values to customize what you want to track

Unity Catalog Locations

- workflow_table:
admin_catalog.....workflows
- task_table:
admin_catalog.....tasks
- cluster_events_table:
admin_catalog.....cluster_events
- etc.

Tracking Indicators

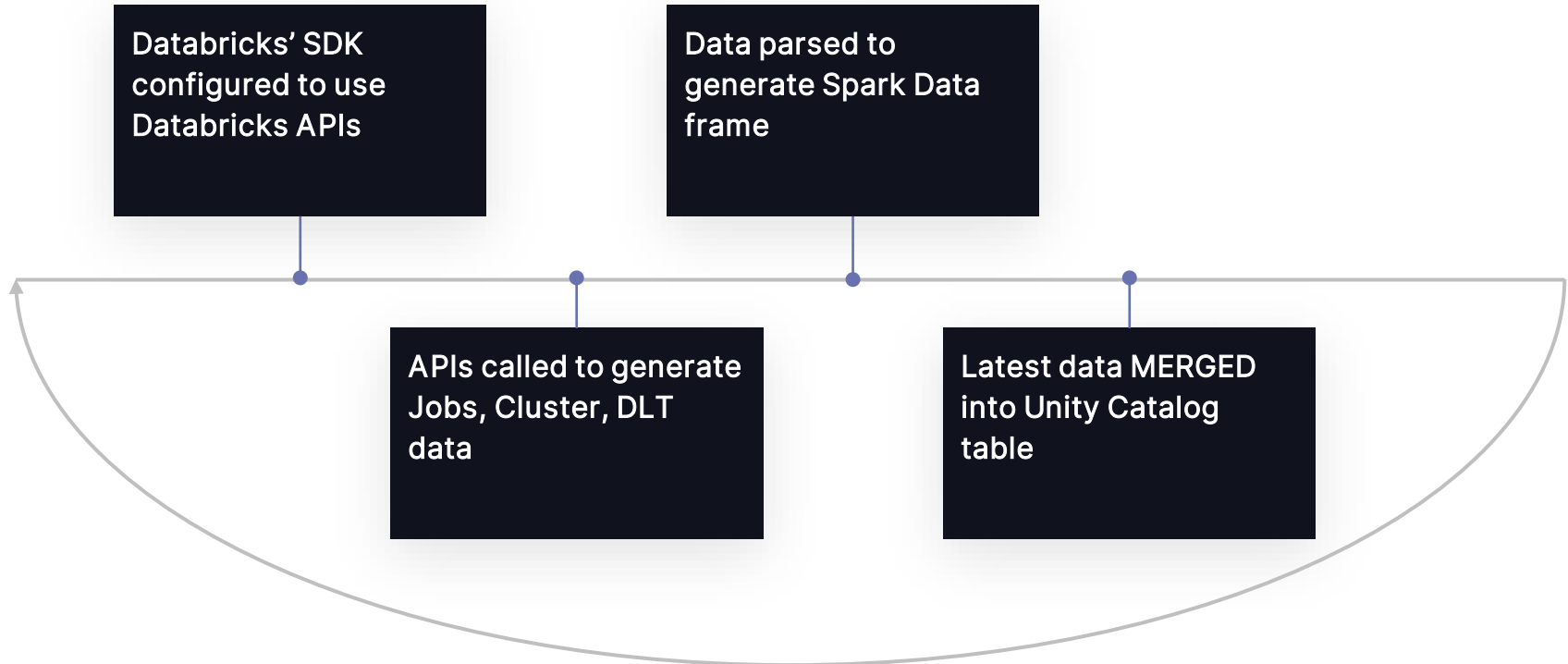
- workflow_tracking: Y
- all_purpose_cluster_tracking: Y
- dlt_tracking: Y

Job Configs

- type: scheduled
- frequency: daily
- cluster: 1030-192001-w4tfjs0i

JOB PROCESS

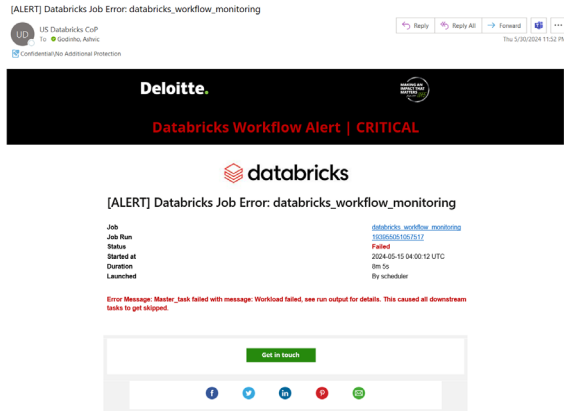
Databricks' new SDK is the linchpin to query essential data



EXTENSIBILITY FOR YOUR CUSTOMER

Build upon framework to customize tool for your specific use case

Email Alerting



Business Intelligence



Connectors: Salesforce and ServiceNow

Status	ID	Subject	Priority	Type	Updated At
P	#46	Another sample	low	question	Jul 18, 2019 11:45 pm
O	#45	Sample ticket #3	normal	task	Jul 18, 2019 11:40 pm
O	#41	Sample ticket	normal	problem	Jul 18, 2019 11:42 pm
P	#39	Sample ticket #2	urgent	task	Jul 18, 2019 11:40 pm

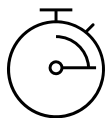
Source: https://zen-marketing-documentation.s3.amazonaws.com/docs/en/Salesforce_ticket_view2.png



Source: <https://sysdig.com/wp-content/uploads/service-now-img-1170x761.jpg>

WHAT VALUE DOES THIS BRING?

Realized Customer Benefits by using Automated Monitoring Solution



Improving efficiency

- Accelerate time to resolution of data pipeline issues
- Disseminate critical information to support teams

Priority setting for issues resulting in streamlined alerts and time savings



Cost Savings

- Reduce maintenance overhead
- Fewer operational support tasks
- Long term cost savings

Cluster event tracking enabled optimized cluster sizing and reduce compute costs

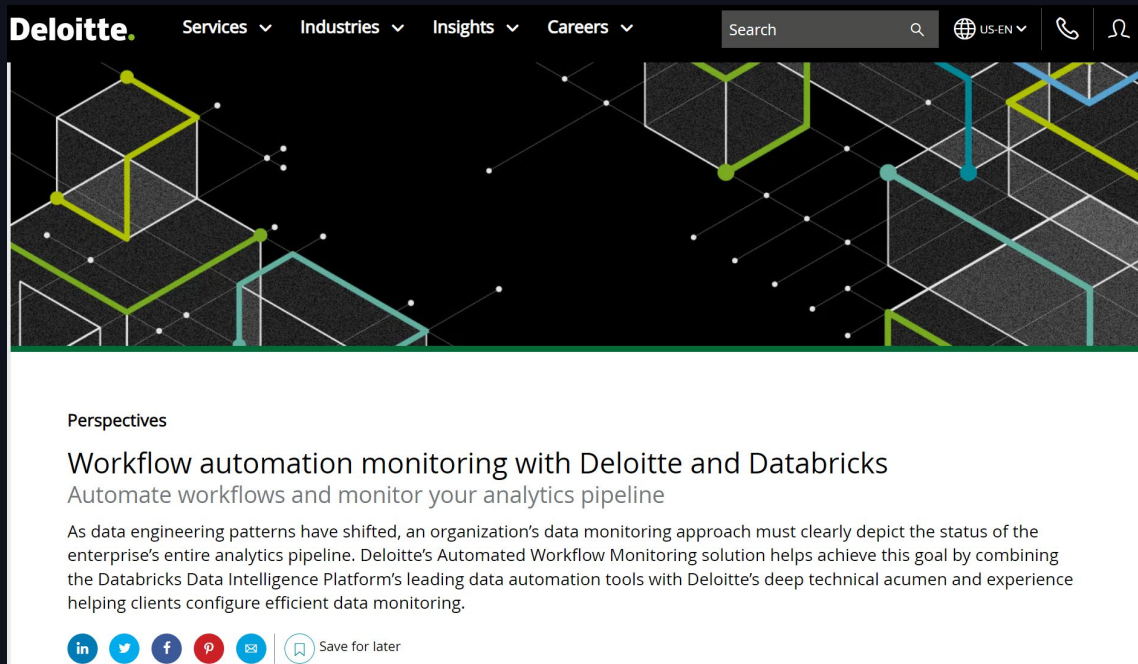


Extensibility

- Complements existing workflows
- “Hook in” to data to customize monitoring for your enterprise solution

Integrated Databricks Alerting into customer’s proprietary tool for 24/7 monitoring

RESOURCES & Q&A



The screenshot shows the top navigation bar of the Deloitte website with links for Services, Industries, Insights, and Careers. A search bar and a language selector (US-EN) are also visible. The main content area features a header image with a network diagram and a section titled 'Perspectives' containing the article title and a brief description.

Deloitte. Services ▾ Industries ▾ Insights ▾ Careers ▾ Search US-EN

Perspectives

Workflow automation monitoring with Deloitte and Databricks

Automate workflows and monitor your analytics pipeline

As data engineering patterns have shifted, an organization's data monitoring approach must clearly depict the status of the enterprise's entire analytics pipeline. Deloitte's Automated Workflow Monitoring solution helps achieve this goal by combining the Databricks Data Intelligence Platform's leading data automation tools with Deloitte's deep technical acumen and experience helping clients configure efficient data monitoring.

in twitter facebook pinterest email Save for later

<https://www2.deloitte.com/us/en/pages/consulting/articles/databricks-alliance-workflow-automation.html>