

Deloitte.

 databricks



DELOITTE'S AUTOMATED WORKFLOW MONITORING:

Enhanced Lakehouse Monitoring with single-window Databricks dashboards

Data engineering patterns experienced rapid changes over the past few years, and modern data intelligence platforms need to adapt to monitor an enterprise's entire analytics pipeline in a simplified, clear view.

The Databricks Data Intelligence Platform provides many tools necessary to build and run data pipelines to accommodate various data engineering tasks and use cases. With this robust suite of tools, Deloitte created a configurable and automated way to track all lakehouse data engineering tasks with minimal overhead costs.

Built on Databricks' SDK and SQL dashboards, Deloitte's Automated Workflow Monitoring solution creates a job to track all data engineering pipeline statuses in a Databricks Workspace. The solution is designed to automate large parts of the operations and maintenance of the Databricks Data Intelligence Platform and reduce the overhead costs typically associated with supplying trustworthy insights to critical data pipeline tasks. Automated Workflow Monitoring provides an extensible platform to build configurations that meet an enterprise's unique needs and works with established business intelligence (BI) tools to provide customized email alerting and a centralized dashboard for data engineering pipelines.

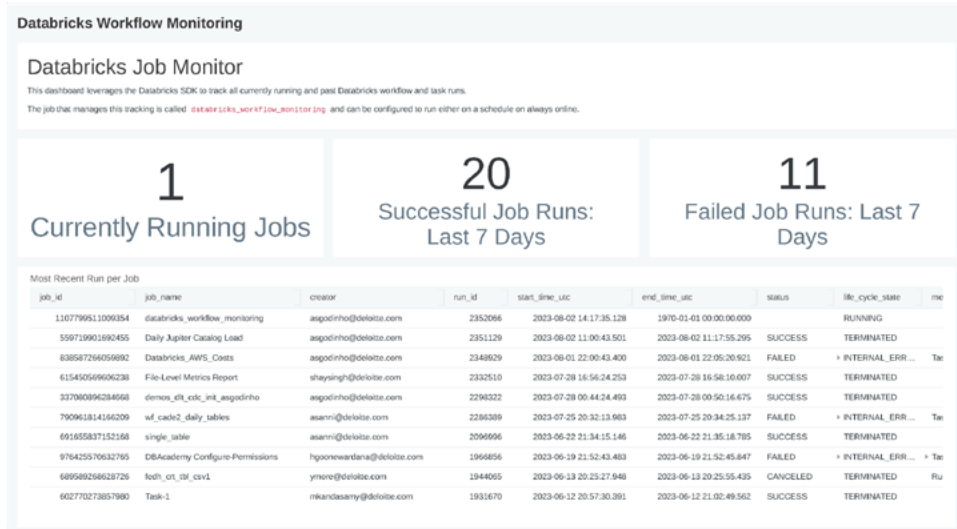
Key features of Automated Workflow Monitoring include:

1. Automated Workflow Monitoring Dashboard
 - a. Job and Workflow Monitor
 - b. Cluster Monitor
 - c. Delta Live Table (DLT) Monitor
2. Configuration-based Solution Architecture and Databricks SDK

AUTOMATED WORKFLOW MONITORING DASHBOARD

The Automated Workflow Monitoring Dashboard provides platform administrators with a single view into the health of all the data engineering-related tasks running in their Databricks Workspace. The Dashboard can offer near real-time insights in the latest status of Databricks Workflows, Clusters, and Delta Live Tables.

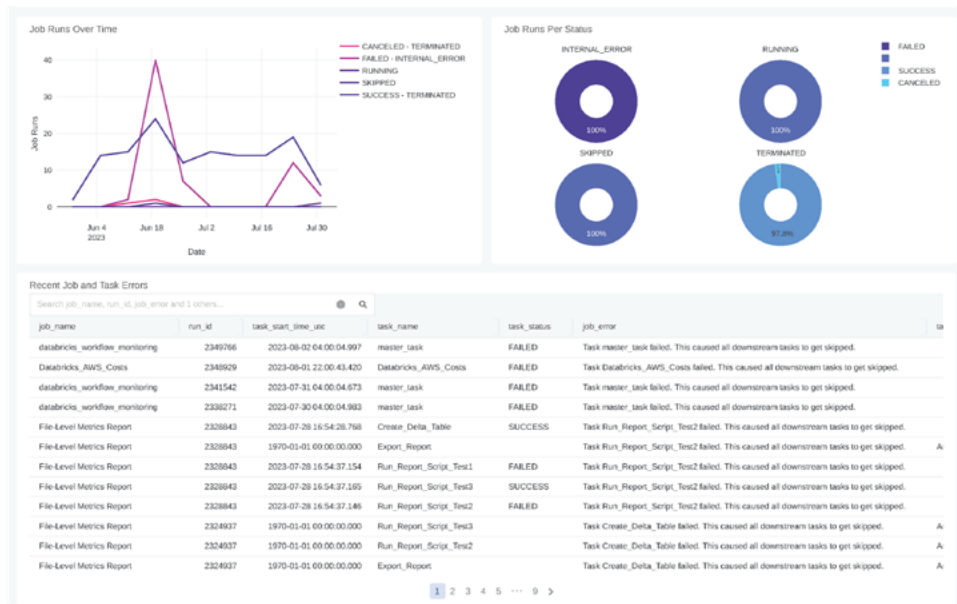
Figure 1. The Monitoring Dashboard provides KPI metrics for a summarized view of the health of the Databricks lakehouse.



Job and Workflow Monitor

Databricks Workflows consist of jobs that contain multiple, individual tasks that each run a data pipeline operation. During runtime, any component in a Workflow may fail, necessitating quick observation of the error and job status. The Job section of the Dashboard outputs metrics on current and historical job runs to quickly pinpoint the overall health of the environment. Job information such as the name, run ID, more detailed task name, and error messages are all stored on lakehouse Unity Catalog tables.

Figure 2. Workflow Monitoring tracks how many jobs are currently running, failed, succeeded, or skipped. This accelerates ticket resolution time, which lowers operating and maintenance costs and monitoring overhead for an enterprise.



Cluster Monitor

All Purpose clusters are ad hoc computing resources that analyze data collaboratively in a Databricks notebook and are used often in the development process when making Databricks Jobs. Since these clusters can fail or emit error messages, organizations can gain insights about the development process by tracking and observing these changes. Because Databricks Runtimes are constantly updated and clusters scale resources on demand, it's useful to track how clusters perform over time. The Cluster Monitoring dashboard collects metrics on All Purpose cluster events to allow administrators to see all other cluster-related events, including when clusters turn on, turn off, and scale up. Additionally, it tracks the run times and node sizes of All Purpose clusters to easily see if any cluster needs updating.

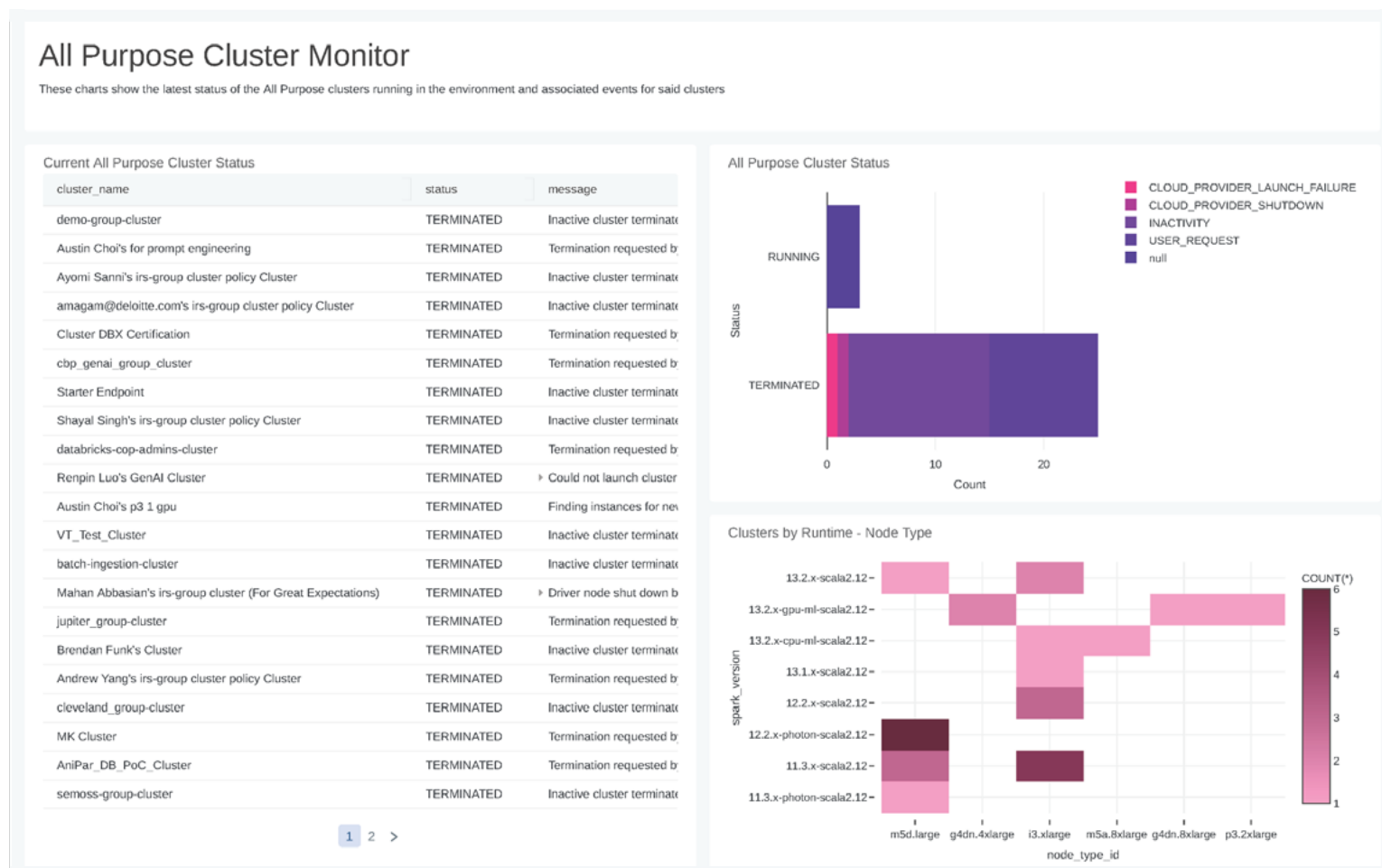
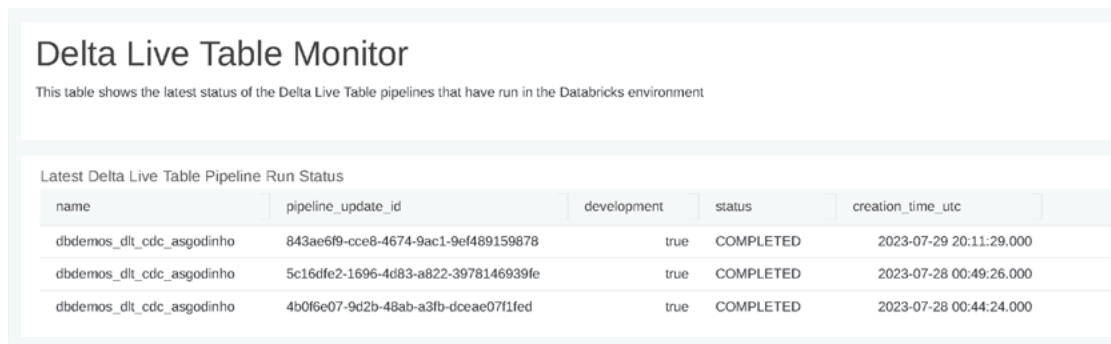


Figure 3. Workflow Monitoring tracks how many jobs are currently running, failed, succeeded, or skipped. This accelerates ticket resolution time, which lowers operating and maintenance costs and monitoring overhead for an enterprise.

Delta Live Table (DLT) Monitor

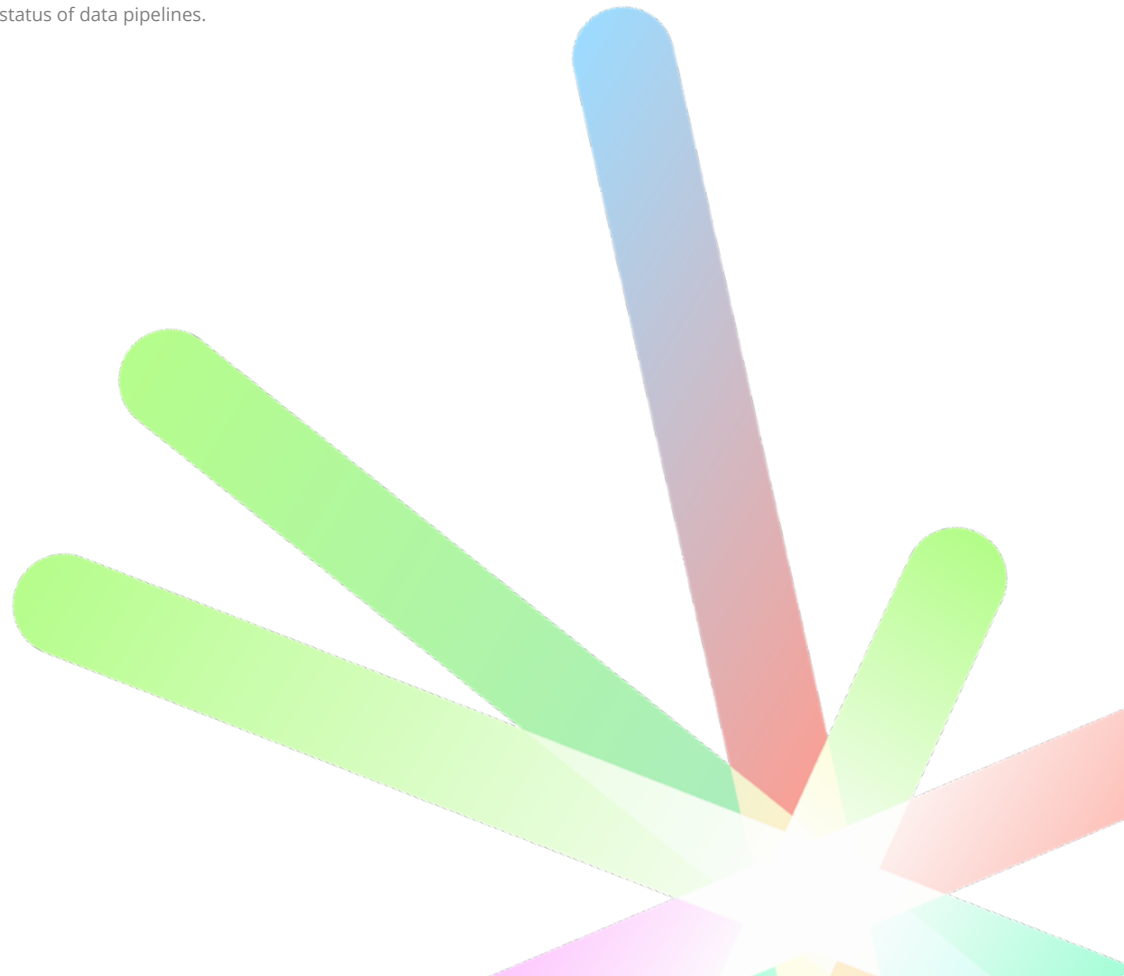
DLT makes it easier for data engineers to build data pipelines and provide data quality metrics for both batch and streaming processes in the lakehouse. The dashboard enables administrators to see the latest pipeline runs, monitor which pipelines are in production, and view the status of a pipeline.



The screenshot shows a dashboard titled "Delta Live Table Monitor" with a subtitle "This table shows the latest status of the Delta Live Table pipelines that have run in the Databricks environment". Below the subtitle is a table titled "Latest Delta Live Table Pipeline Run Status". The table has five columns: "name", "pipeline_update_id", "development", "status", and "creation_time_utc". There are three rows of data, all showing a status of "COMPLETED".

name	pipeline_update_id	development	status	creation_time_utc
dbdemos_dlt_cdc_asgodinho	843ae6f9-cce8-4674-9ac1-9ef489159878	true	COMPLETED	2023-07-29 20:11:29.000
dbdemos_dlt_cdc_asgodinho	5c16dfe2-1696-4d83-a822-3978146939fe	true	COMPLETED	2023-07-28 00:49:26.000
dbdemos_dlt_cdc_asgodinho	4b0f6e07-9d2b-48ab-a3fb-dceae07f1fed	true	COMPLETED	2023-07-28 00:44:24.000

Figure 4. DLT pipelines can be run on demand, streaming, or on a schedule. Runtime metrics are tracked on the Dashboard enabling a single view to see the latest status of data pipelines.



CONFIGURATION-BASED SOLUTION ARCHITECTURE AND DATABRICKS SDK

The core of Deloitte's Automated Workflow Monitoring solution leverages a customizable job, which runs via a script that uses YAML to configure the job settings. Mature lakehouse environments that need up-to-date information can set the schedule of the job to "streaming" via YAML, which causes the job to run constantly—always delivering the latest status update to Unity Catalog. When immediate data is not necessary, organizations can set the schedule to run "daily," "weekly," or at a given number of minutes. For testing purposes, a configuration value called "cluster_id" can be set for an All Purpose cluster to run the job immediately.

Data is generated by the Databricks' SDK, a tool used to automate operations, including accounts, workspaces, and related resources such as jobs. The Monitoring job updates the latest workflow status data by following this procedure:

1. Querying the data: Databricks SDK calls are used to get the relevant data to populate a Python dictionary.
2. Generating the latest data frame: The Python dictionary is parsed through to populate a Apache Spark™ data frame with the essential columns from the API call.
3. Merge into: The Spark data frame is merged into the Unity Catalog table to update records with the latest status or insert new records for brand new data.

Using this method to gather the status metrics pushes most of the processing to the API call and Python dictionaries, which allows the tool to still process a large volume of data in seconds. The longest operation—writing to the lakehouse—is done only once so Spark can appropriately distribute the operation as needed.

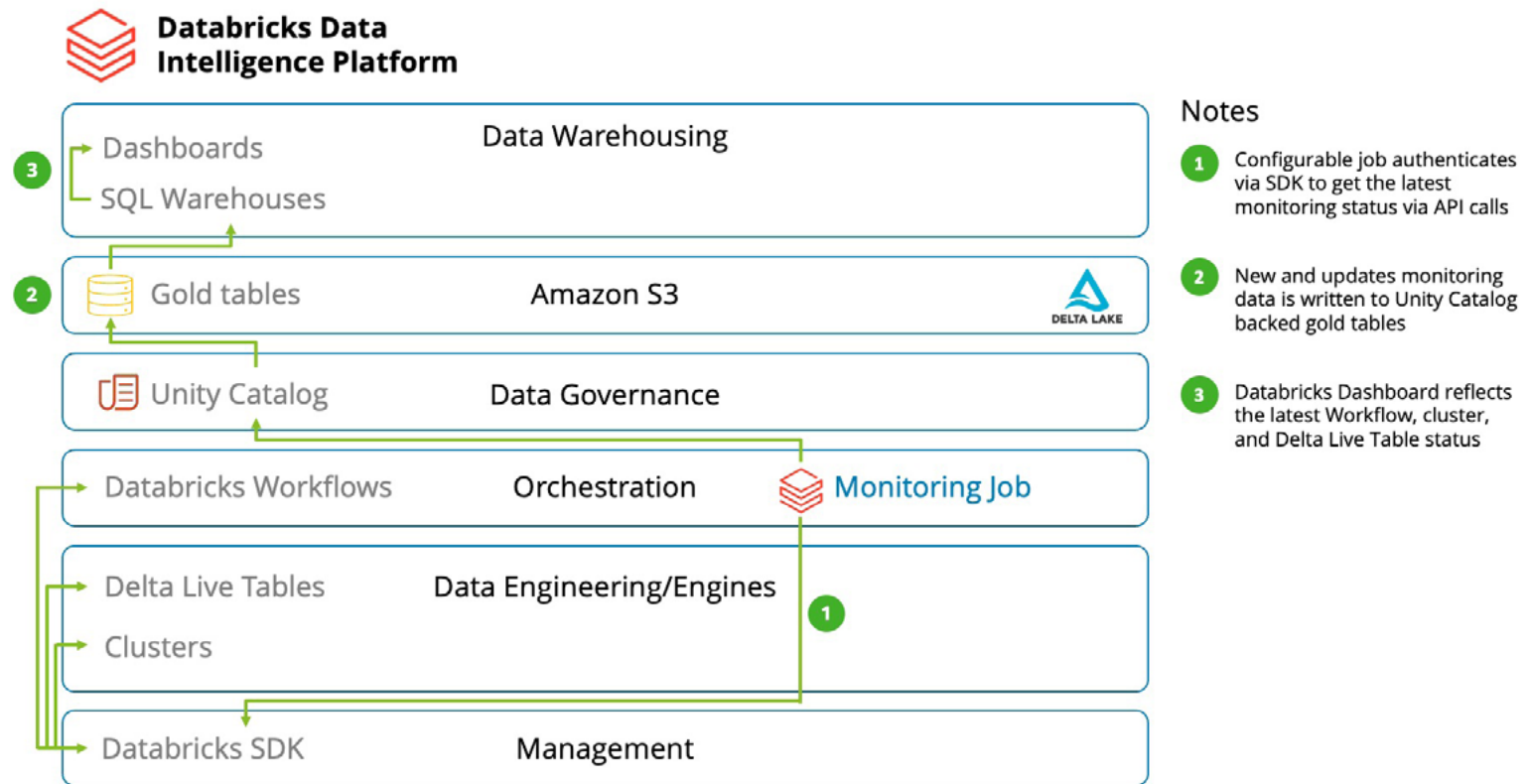
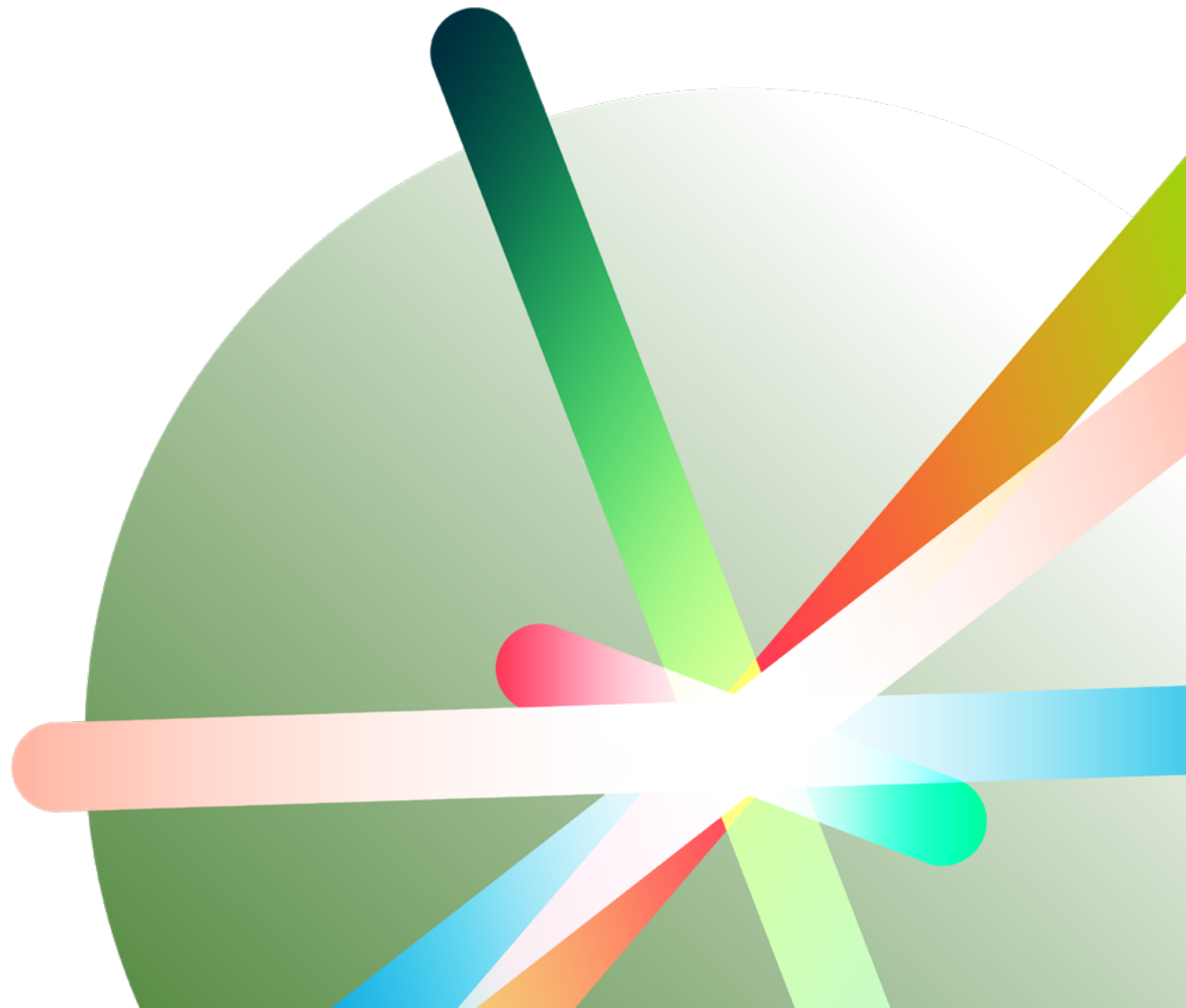


Figure 5. Architecture Reference Diagram of the Workflow Monitoring Framework

GETTING STARTED WITH AUTOMATED WORKFLOW MONITORING

To add Automated Workflow Monitoring to your Databricks Data Intelligence Platform, follow these four steps:

- ① Reach out to the [Deloitte Databricks alliance](#) to get access to Deloitte's [Enterprise GitHub repository](#)
- ② After gaining access to the code, pull the Automated Workflow Monitoring code from GitHub and upload the Python files to your Databricks Workspace
- ③ Fill out the `configuration_script.py` file with the appropriate config values
- ④ Run the `job_creation_script.py` to Automated Workflow Monitoring job



CONCLUSION

Automated Workload Monitoring can help organizations improve efficiency and cost savings. By automating task monitoring and customizing alerts, organizations can accelerate time to resolution for data quality issues and quickly disseminate critical information to support teams. This also reduces maintenance overhead, leads to fewer operational tasks, and provides long-term cost savings to upkeep Databricks. Automated Workflow Monitoring is also designed to complement existing workflows, as developers can integrate the extensible solution seamlessly in their design patterns. With Automated Workload Monitoring, Deloitte can provide more effective support to clients' Databricks implementations.

[Learn more](#) about the Deloitte and Databricks alliance, or contact us to discuss adopting Deloitte's Automated Workflow Monitoring.

Ashvic Godinho

AI & Data Engineering
Specialist Master

Deloitte Consulting LLP
asgodinho@deloitte.com

Mani Kandasamy

AI & Data Engineering
Technology Fellow

Deloitte Consulting LLP
mkandasamy@deloitte.com

About Deloitte

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms or their related entities (collectively, the "Deloitte Network"), is, by means of this communication, rendering professional advice or services. Before making any decisions or taking any action that may affect your finances, or your business, you should consult a qualified professional adviser. No entity in the Deloitte Network shall be responsible for any loss whatsoever sustained by any person who relies on this communication. As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting. Copyright 2024 Deloitte Development LLC. All rights reserved.