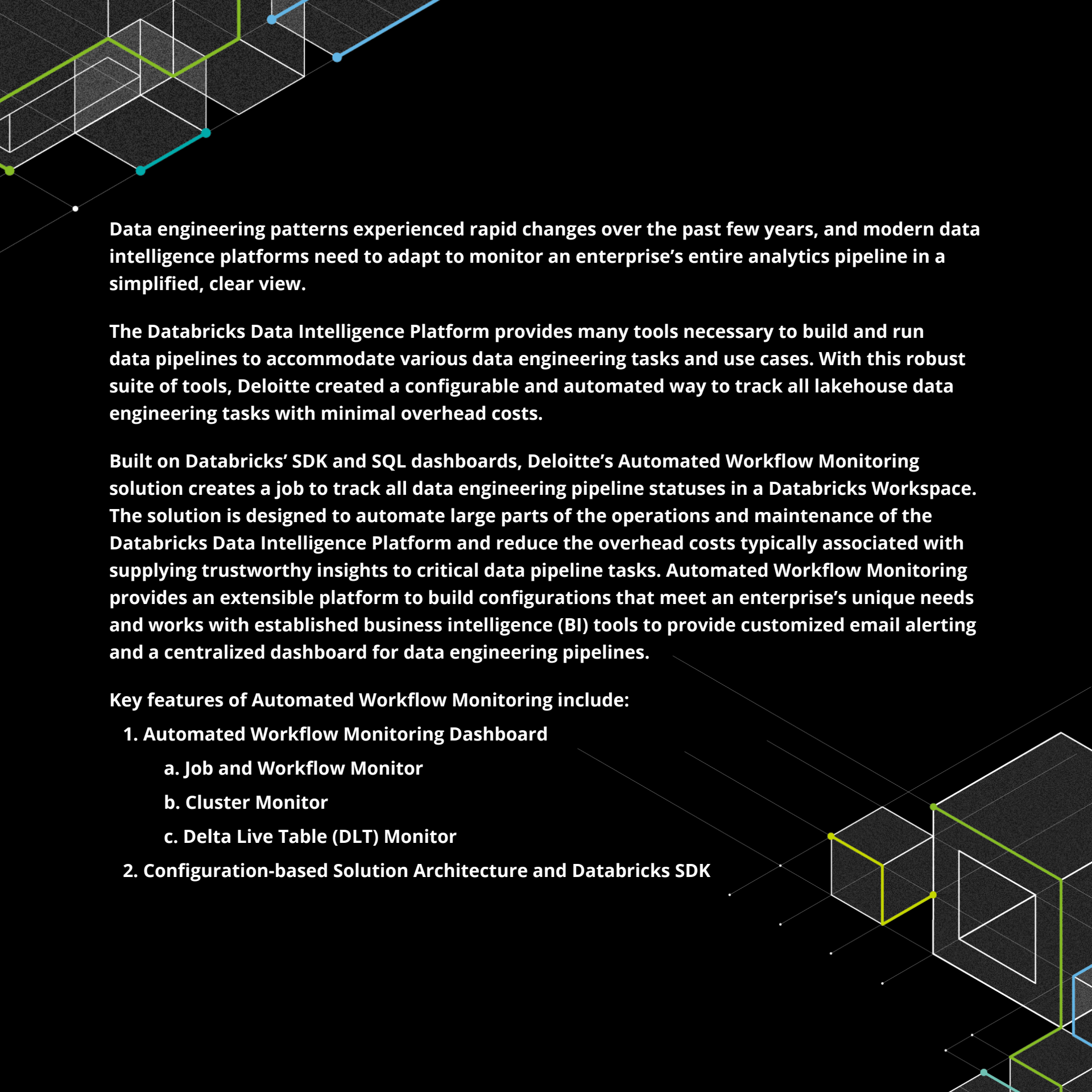databricks | **Deloitte.**

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.

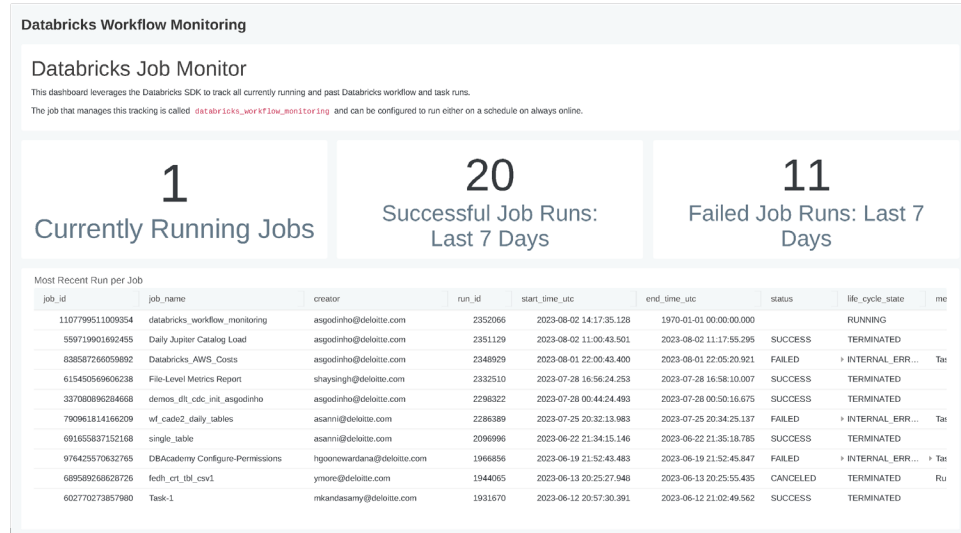Key features of Automated Workflow Monitoring include:
1. Automated Workflow Monitoring Dashboard
   a. Job and Workflow Monitor
   b. Cluster Monitor
   c. Delta Live Table (DLT) Monitor
2. Configuration-based Solution Architecture and Databricks SDK

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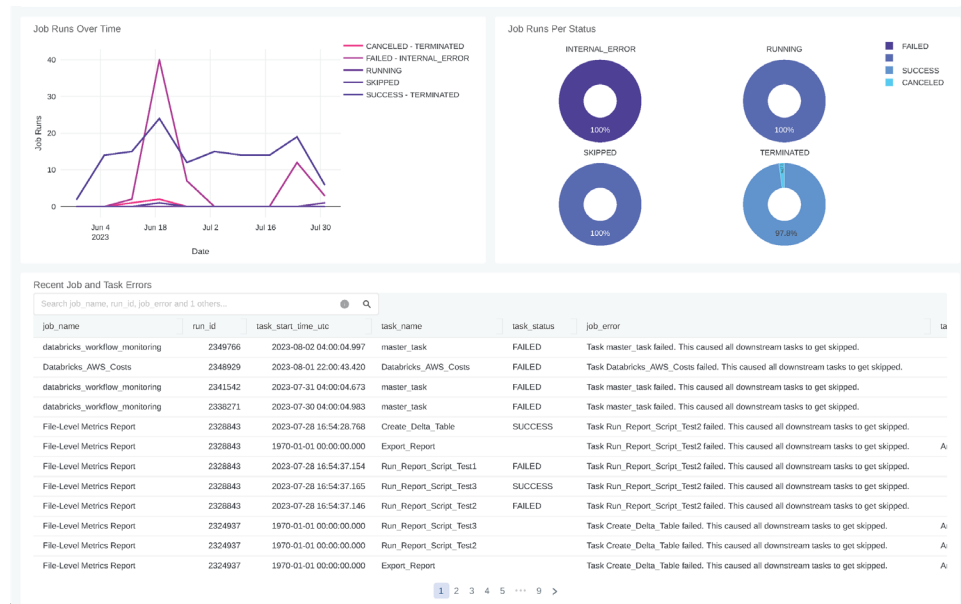
Figure 1. The Monitoring Dashboard provides KPI metrics for a summarized view of the health of the Databricks lakehouse.

Databricks Workflow Monitoring

Databricks Job Monitor

This dashboard leverages the Databricks SDK to track all currently running and past Databricks workflow and task runs.

The job that manages this tracking is called `databricks_workflow_monitoring` and can be configured to run either on a schedule on always online.

1	20	11
Currently Running Jobs	Successful Job Runs: Last 7 Days	Failed Job Runs: Last 7 Days

Most Recent Run per Job

job_id	job_name	creator	run_id	start_time_utc	end_time_utc	status	life_cycle_state	me
1107799511009354	databricks_workflow_monitoring	asgodinho@deloitte.com	2352066	2023-08-02 14:17:35.128	1970-01-01 00:00:00.000		RUNNING	
559719901692455	Daily Jupiter Catalog Load	asgodinho@deloitte.com	2351129	2023-08-02 11:00:43.501	2023-08-02 11:17:55.295	SUCCESS	TERMINATED	
838587266059892	Databricks_AWS_Costs	asgodinho@deloitte.com	2348929	2023-08-01 22:00:43.400	2023-08-01 22:05:20.921	FAILED	▸ INTERNAL_ERR...	Tas
615450569606238	File-Level Metrics Report	shaysingh@deloitte.com	2332510	2023-07-28 16:56:24.253	2023-07-28 16:58:10.007	SUCCESS	TERMINATED	
337080896284668	demos_dlt_cdc_init_asgodinho	asgodinho@deloitte.com	2298322	2023-07-28 00:44:24.493	2023-07-28 00:50:16.675	SUCCESS	TERMINATED	
79096181416209	wf_cade2_daily_tables	asanni@deloitte.com	2286389	2023-07-25 20:32:13.983	2023-07-25 20:34:25.137	FAILED	▸ INTERNAL_ERR...	Tas
691655837152168	single_table	asanni@deloitte.com	2096996	2023-06-22 21:34:15.146	2023-06-22 21:35:18.785	SUCCESS	TERMINATED	
97642557063276	DBAcademy Configure-Permissions	hgoonewardana@deloitte.com	1966856	2023-06-19 21:52:43.483	2023-06-19 21:52:45.847	FAILED	▸ INTERNAL_ERR...	▸ Tas
689589268628726	fedh_crt_tbl_csv1	ymore@deloitte.com	1944065	2023-06-13 20:25:27.948	2023-06-13 20:25:55.435	CANCELED	TERMINATED	Ru
602770273857980	Task-1	mkandasamy@deloitte.com	1931670	2023-06-12 20:57:30.391	2023-06-12 21:02:49.562	SUCCESS	TERMINATED	

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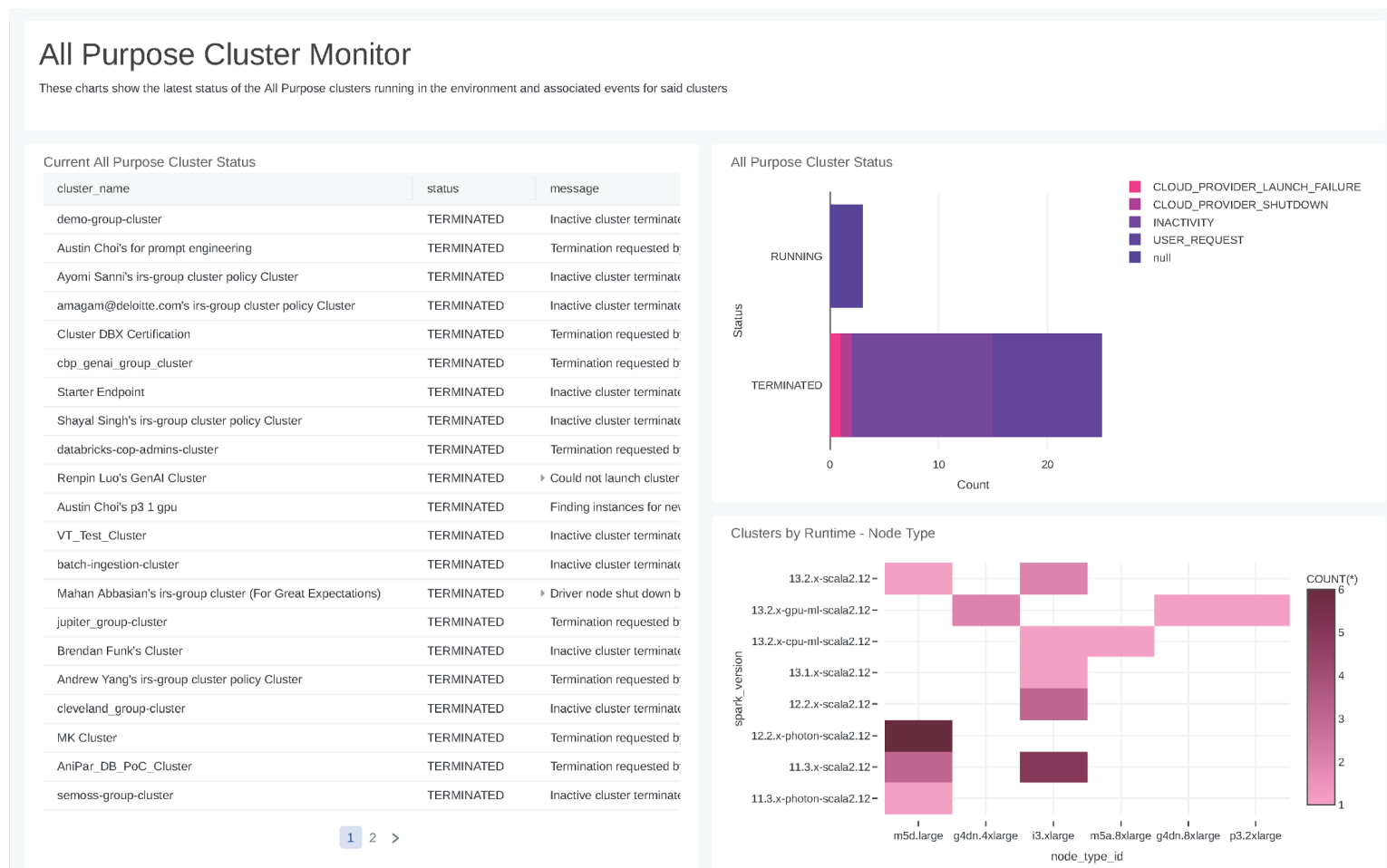
Figure 2. Workflow Monitoring tracks how many jobs are currently running, failed, succeeded, or skipped. This accelerates ticket resolution time, which lowers operating and maintenance costs and monitoring overhead for an enterprise.

Job Runs Over Time

Legend: CANCELED - TERMINATED, FAILED - INTERNAL_ERROR, RUNNING, SKIPPED, SUCCESS - TERMINATED

Job Runs Per Status

INTERNAL_ERROR 100% — RUNNING 100% — SKIPPED 100% — TERMINATED 97.4%

Legend: FAILED, SUCCESS, CANCELED

Recent Job and Task Errors

job_name	run_id	task_start_time_utc	task_name	task_status	job_error	ta
databricks_workflow_monitoring	2349766	2023-08-02 04:00:04.997	master_task	FAILED	Task master_task failed. This caused all downstream tasks to get skipped.	
Databricks_AWS_Costs	2348929	2023-08-01 22:00:43.420	Databricks_AWS_Costs	FAILED	Task Databricks_AWS_Costs failed. This caused all downstream tasks to get skipped.	
databricks_workflow_monitoring	2341542	2023-07-31 04:00:04.673	master_task	FAILED	Task master_task failed. This caused all downstream tasks to get skipped.	
databricks_workflow_monitoring	2338271	2023-07-30 04:00:04.983	master_task	FAILED	Task master_task failed. This caused all downstream tasks to get skipped.	
File-Level Metrics Report	2328843	2023-07-28 16:54:28.768	Create_Delta_Table	SUCCESS	Task Run_Report_Script_Test2 failed. This caused all downstream tasks to get skipped.	
File-Level Metrics Report	2328843	1970-01-01 00:00:00.000	Export_Report		Task Run_Report_Script_Test2 failed. This caused all downstream tasks to get skipped.	A
File-Level Metrics Report	2328843	2023-07-28 16:54:37.154	Run_Report_Script_Test1	FAILED	Task Run_Report_Script_Test2 failed. This caused all downstream tasks to get skipped.	
File-Level Metrics Report	2328843	2023-07-28 16:54:37.165	Run_Report_Script_Test3	SUCCESS	Task Run_Report_Script_Test2 failed. This caused all downstream tasks to get skipped.	
File-Level Metrics Report	2328843	2023-07-28 16:54:37.146	Run_Report_Script_Test2	FAILED	Task Run_Report_Script_Test2 failed. This caused all downstream tasks to get skipped.	
File-Level Metrics Report	2324937	1970-01-01 00:00:00.000	Run_Report_Script_Test3		Task Create_Delta_Table failed. This caused all downstream tasks to get skipped.	A
File-Level Metrics Report	2324937	1970-01-01 00:00:00.000	Run_Report_Script_Test2		Task Create_Delta_Table failed. This caused all downstream tasks to get skipped.	A
File-Level Metrics Report	2324937	1970-01-01 00:00:00.000	Export_Report		Task Create_Delta_Table failed. This caused all downstream tasks to get skipped.	A

1 2 3 4 5 ⋯ 9 ›

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.

Delta Live Table Monitor

This table shows the latest status of the Delta Live Table pipelines that have run in the Databricks environment

Latest Delta Live Table Pipeline Run Status

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.
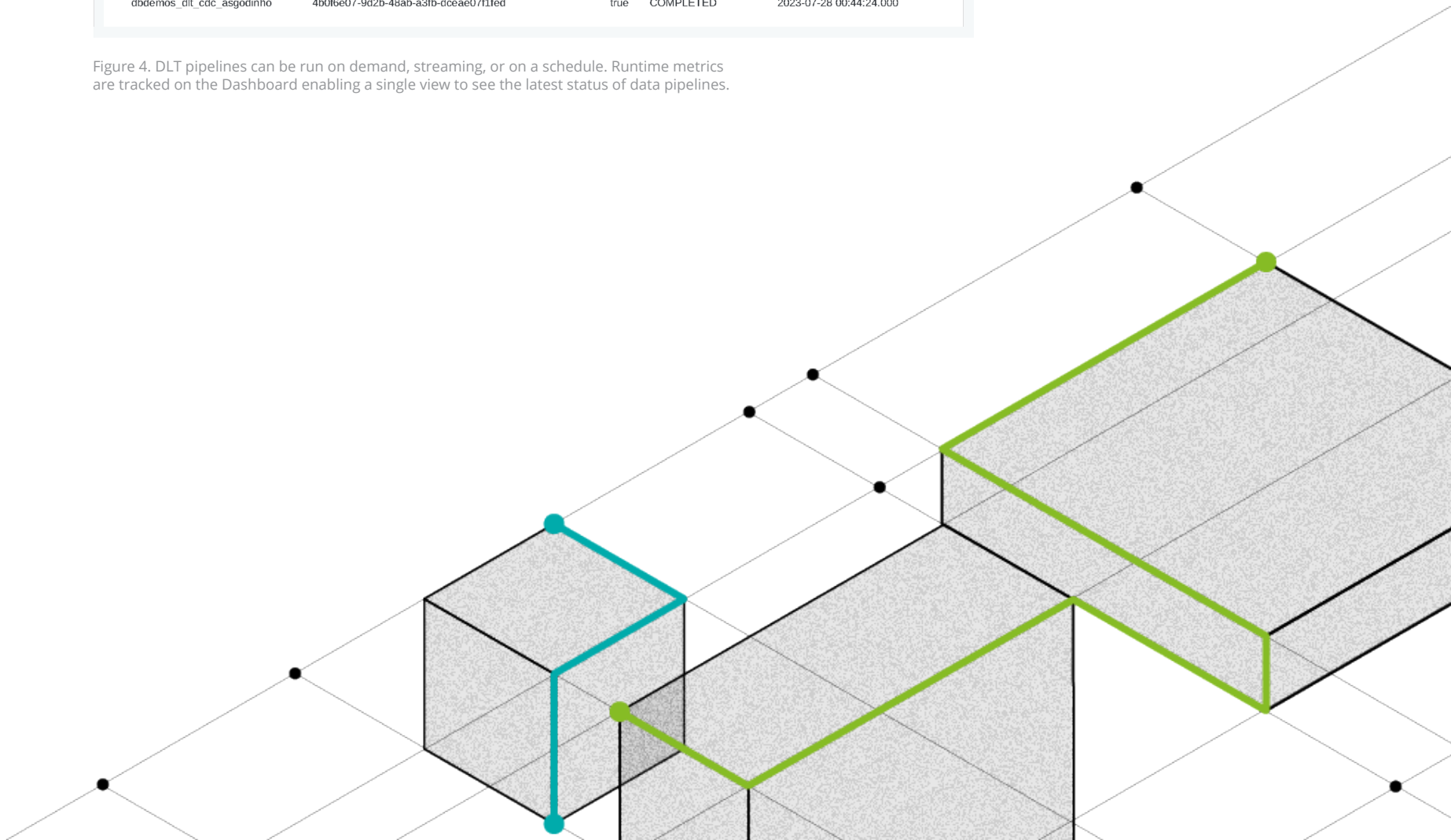
Configuration-based Solution Architecture and Databricks SDK

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.
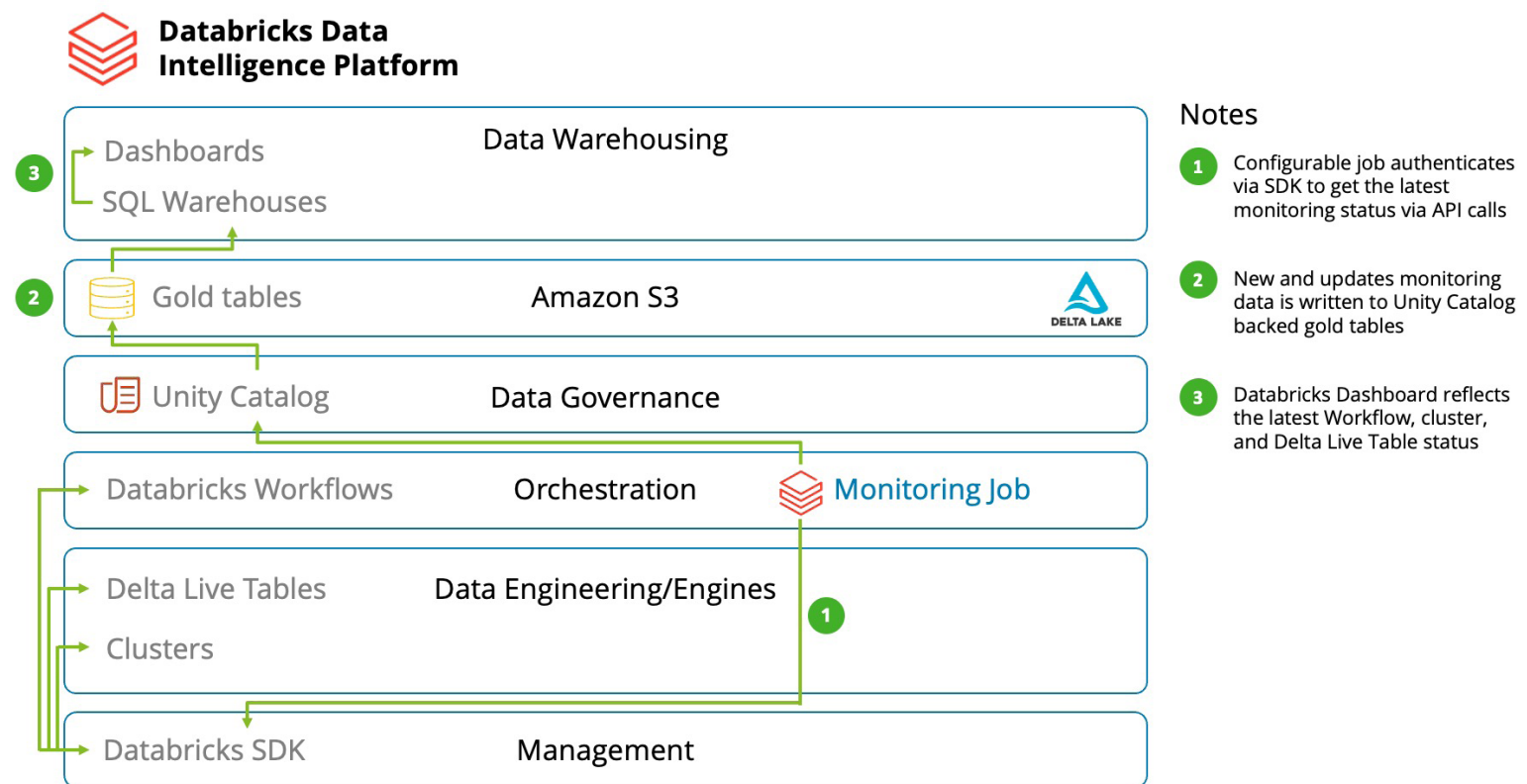
Figure 5. Architecture Reference Diagram of the Workflow Monitoring Framework

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
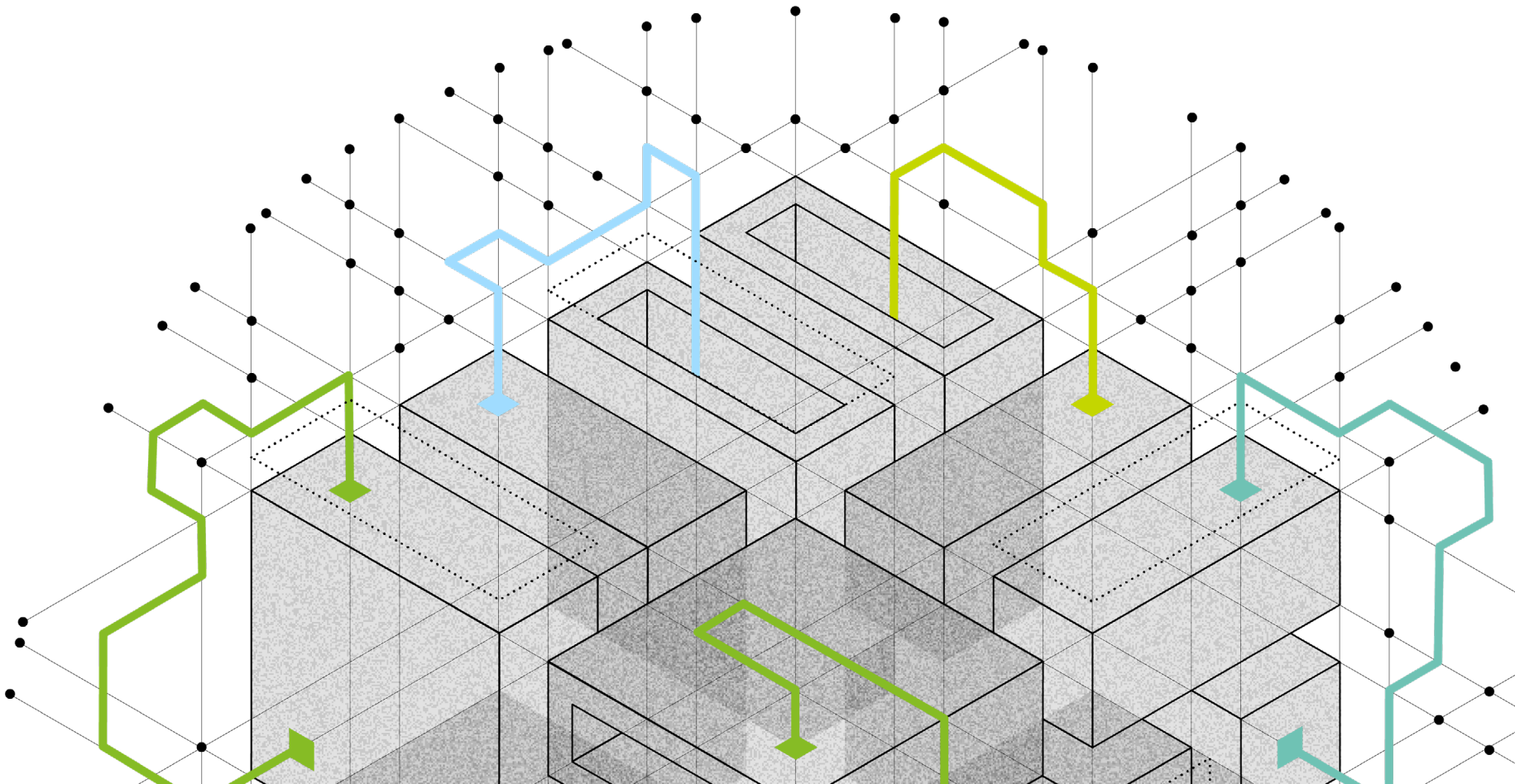
Conclusion

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 Eningeering
Technology Fellow

Deloitte Consulting LLP
mkandasamy@deloitte.com