REVEFI

Technical Brief

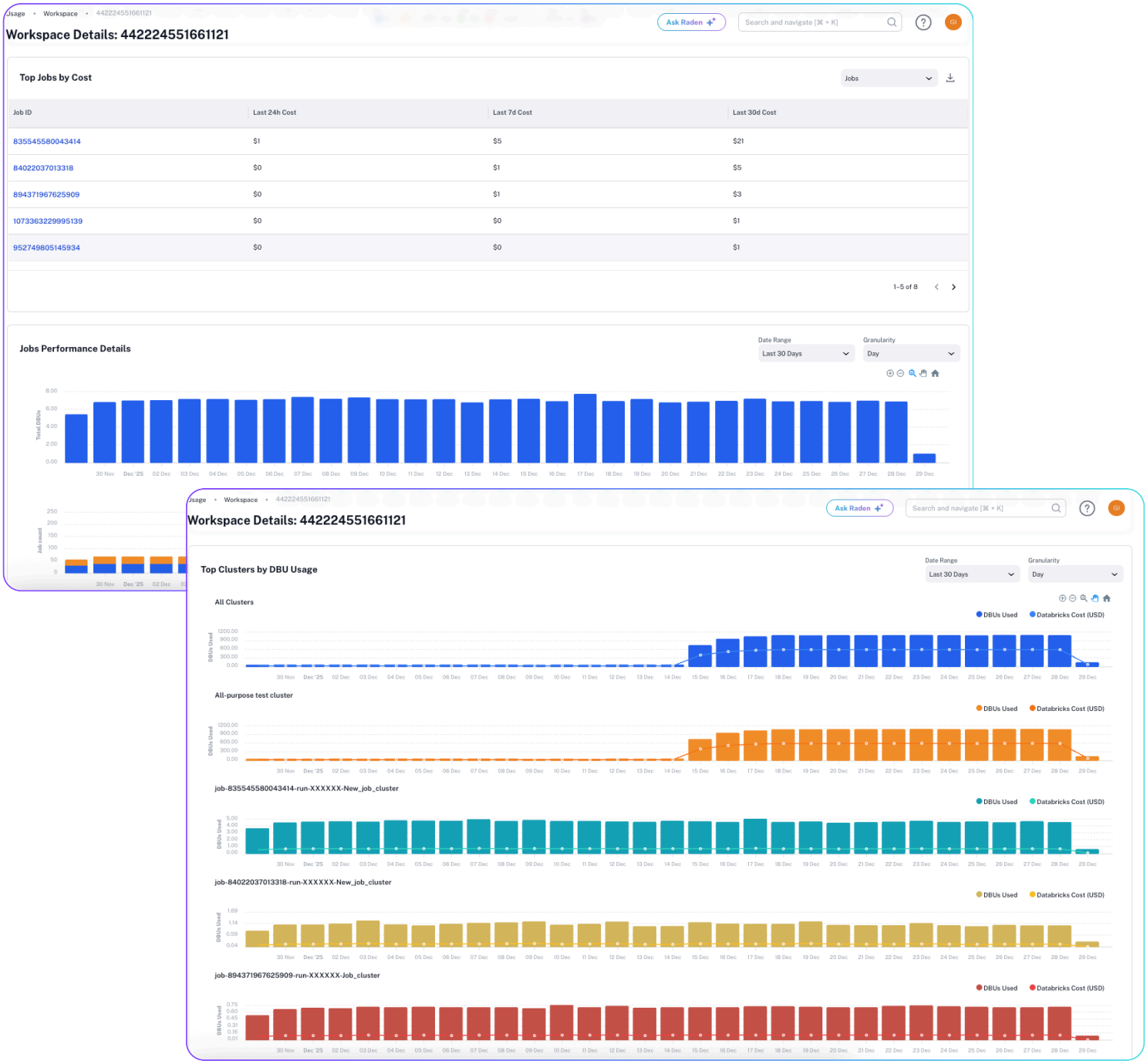
Databricks Job Optimization

Gartner COOL VENDOR 2025

This document outlines how Revefi helps optimize Databricks Jobs with detailed job-level observability and automated optimization, bridging the gap between cost visibility, performance diagnostics, optimization and actionable remediation.

Revefi integrates with Databricks workspaces in a few minutes and data teams can quickly analyze job execution behavior, DBU consumption, cost attribution, failure patterns, and cluster efficiency to optimize job configurations through controlled experiments.

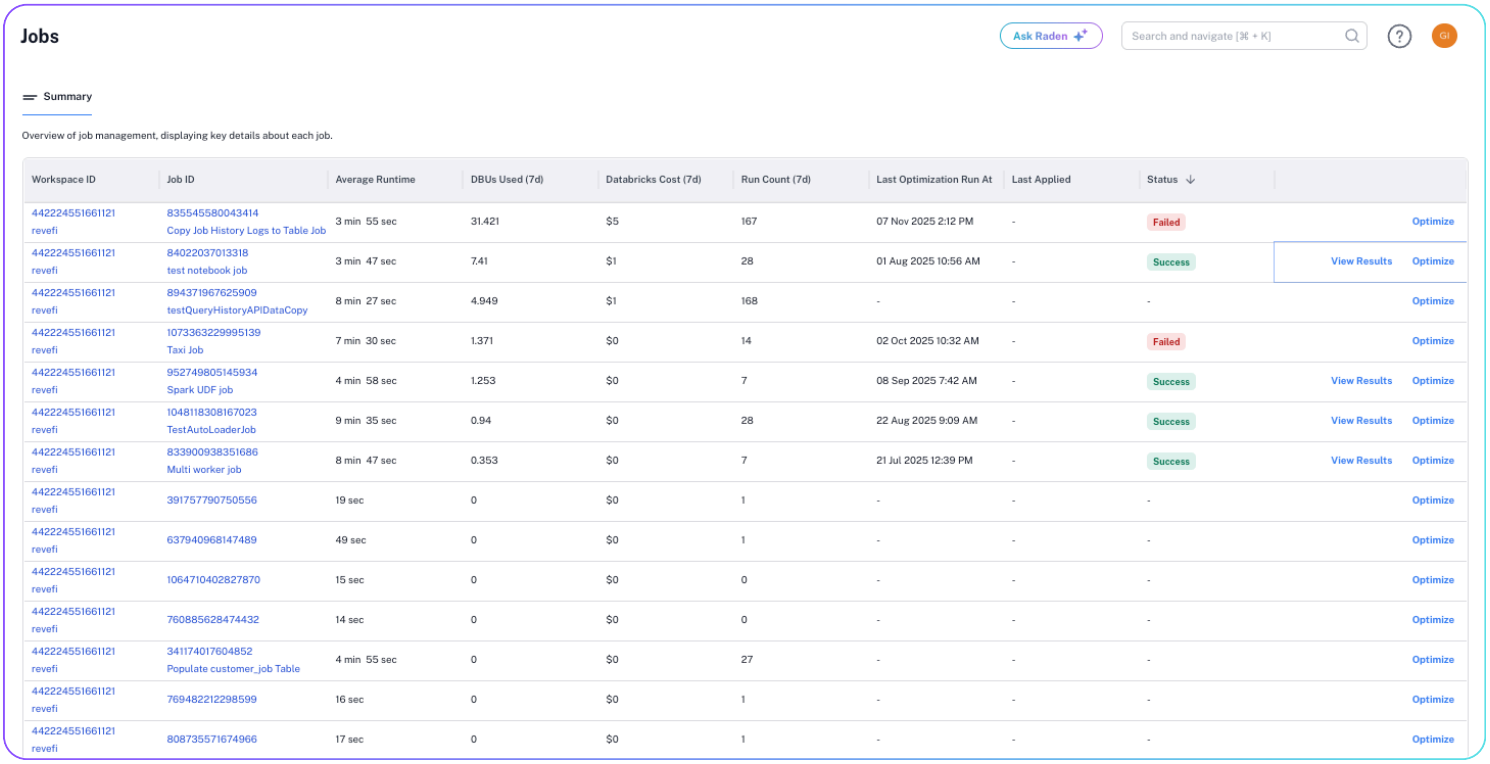
The resultant closed-loop system from observe to diagnose to experiment to optimize and validate, with quantified savings and SLA-aware decisioning extends the core Databricks platform capabilities.



Job Inventory and Operational Visibility

Revefi provides a centralized Jobs view across Databricks workspaces that aggregates job-level operational, performance, and cost telemetry. Each job is mapped to its Workspace ID and Job ID, with clear identification of the job and its execution context. The view surfaces average runtime, run frequency, and DBUs consumed over a rolling window.

Databricks cost is attributed at the job level for the same time period, enabling direct cost-to-workload analysis. Optimization lifecycle state is tracked per job, including success, failure, or not yet optimized. The timestamp of the most recent optimization run is retained for auditability. Inline actions allow users to inspect optimization results or trigger new optimization experiments directly from the job list.

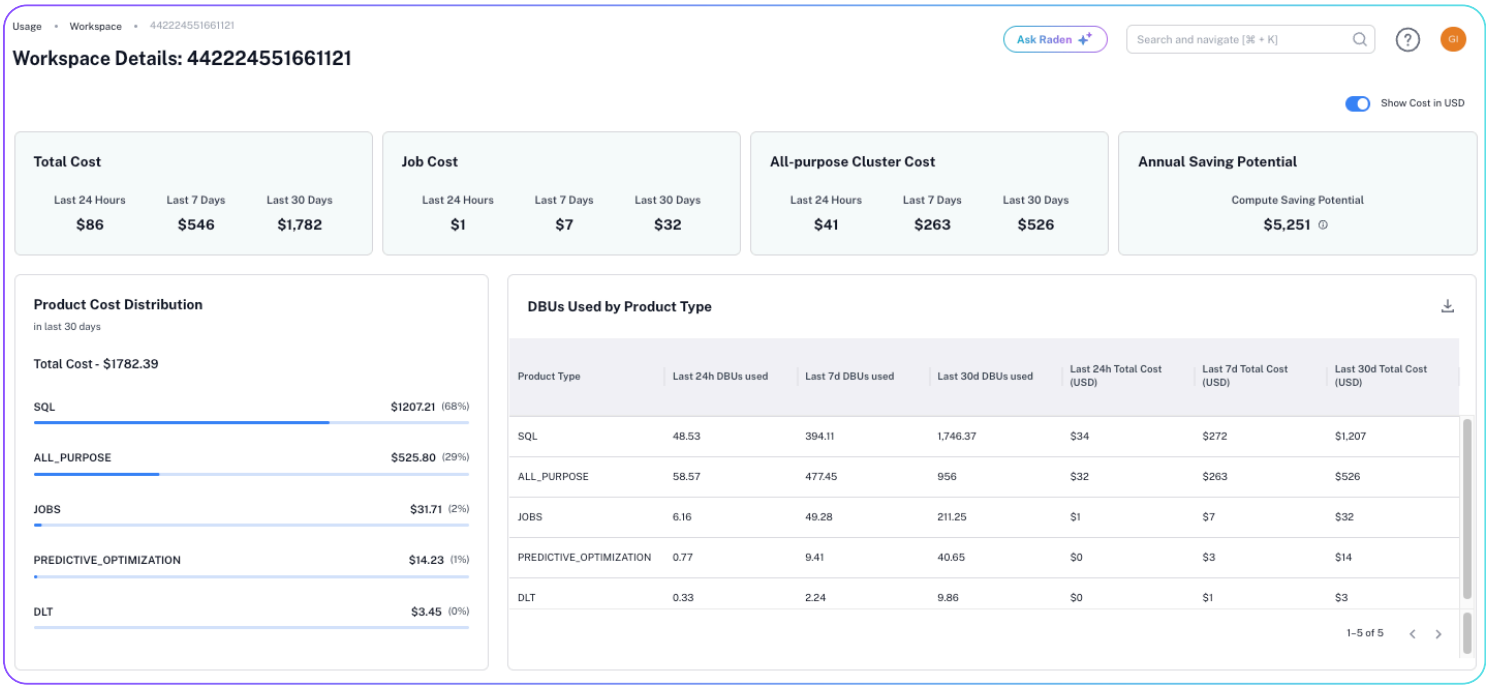


Workspace ID	Job ID	Average Runtime	DBUs Used (7d)	Databricks Cost (7d)	Run Count (7d)	Last Optimization Run At	Last Applied	Status	
442224551661121 revefi	835545580043414 Copy Job History Logs to Table Job	3 min 55 sec	31.421	$5	167	07 Nov 2025 2:12 PM	-	Failed	Optimize
442224551661121 revefi	84022037013318 test notebook job	3 min 47 sec	7.41	$1	28	01 Aug 2025 10:56 AM	-	Success	View Results Optimize
442224551661121 revefi	894371967625909 testQueryHistoryAPIDataCopy	8 min 27 sec	4.949	$1	168	-	-	-	Optimize
442224551661121 revefi	1073363229995139 Taxi Job	7 min 30 sec	1.371	$0	14	02 Oct 2025 10:32 AM	-	Failed	Optimize
442224551661121 revefi	952749805145934 Spark UDF job	4 min 58 sec	1.253	$0	7	08 Sep 2025 7:42 AM	-	Success	View Results Optimize
442224551661121 revefi	1048118308167023 TestAutoLoaderJob	9 min 35 sec	0.94	$0	28	22 Aug 2025 9:09 AM	-	Success	View Results Optimize
442224551661121 revefi	833900938351686 Multi worker job	8 min 47 sec	0.353	$0	7	21 Jul 2025 12:39 PM	-	Success	View Results Optimize
442224551661121 revefi	391757790750556	19 sec	0	$0	1	-	-	-	Optimize
442224551661121 revefi	637940968147489	49 sec	0	$0	1	-	-	-	Optimize
442224551661121 revefi	1064710402827870	15 sec	0	$0	0	-	-	-	Optimize
442224551661121 revefi	760885628474432	14 sec	0	$0	0	-	-	-	Optimize
442224551661121 revefi	341174017604852 Populate customer_job Table	4 min 55 sec	0	$0	27	-	-	-	Optimize
442224551661121 revefi	769482212298599	16 sec	0	$0	1	-	-	-	Optimize
442224551661121 revefi	808735571674966	17 sec	0	$0	1	-	-	-	Optimize

This view functions as a control plane for Databricks Jobs, enabling users to quickly identify high-frequency jobs with disproportionate DBU consumption, recurring job failures, and workloads that are strong candidates for optimization based on runtime or cost characteristics. By explicitly tracking both successful and failed optimization attempts, Revefi treats optimization as key execution artifacts rather than static or advisory recommendations.

Workspace Cost Attribution

Revefi consolidates job level telemetry into Workspace details that supports FinOps oriented cost analysis. It provides cost rollups across multiple time horizons, including total workspace spend over the last 24 hours, 7 days, and 30 days. Costs are further segmented into job-specific spend and all-purpose cluster spend for attribution. An estimated annual savings potential is computed to contextualize optimization impact.



This separation helps teams distinguish inefficiencies in batch jobs from interactive or exploratory cluster usage. In addition, costs are distributed by Databricks product category, including SQL, all-purpose clusters, jobs, predictive optimization, and Delta Live Tables. This breakdown enables platform owners to identify dominant cost drivers and prioritize optimization efforts accordingly. i.e., Instead of optimizing every job, Revefi helps identify which jobs deliver the highest ROI when tuned.

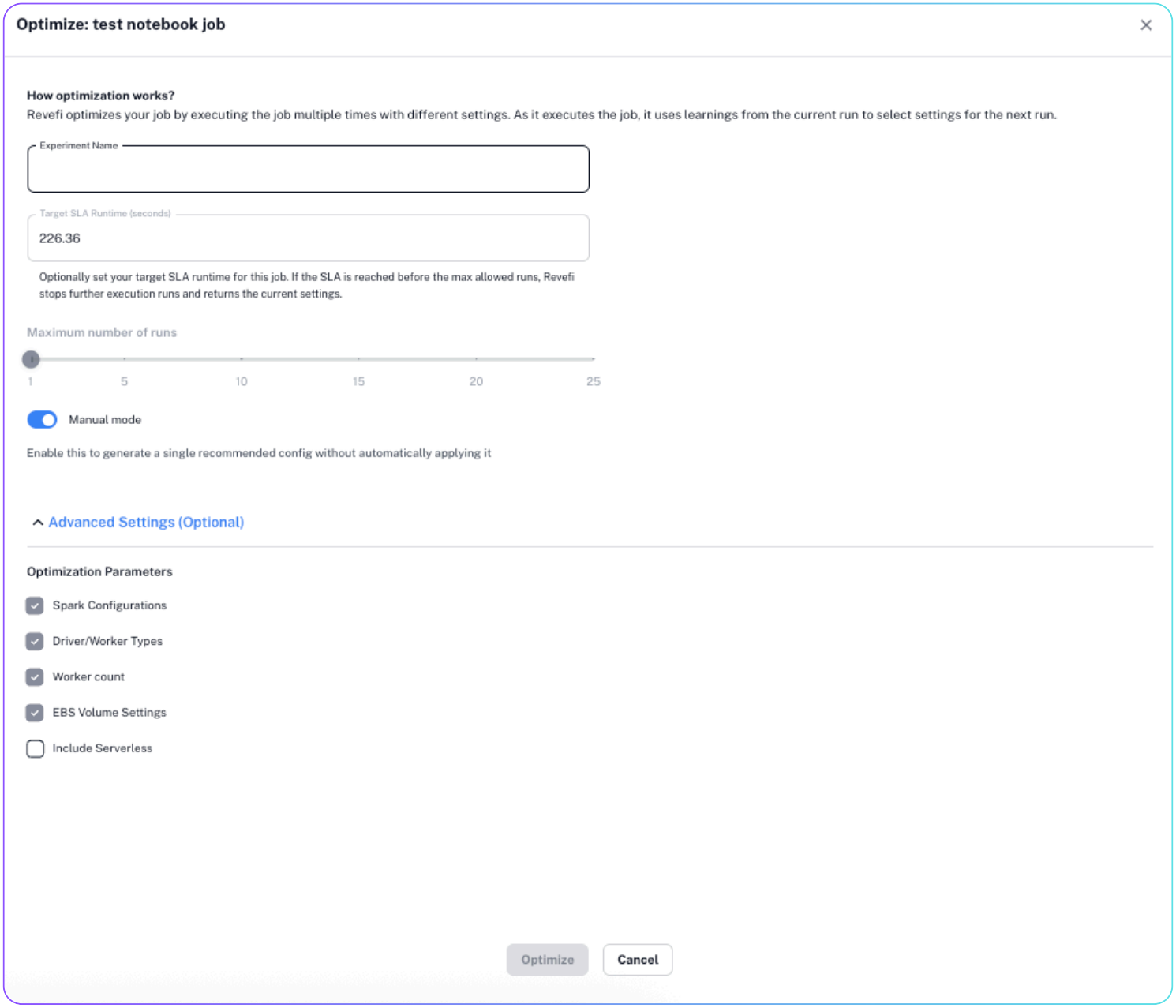
Job Performance & Reliability Analysis

Revefi provides detailed insights into job performance with time-series visualizations across configurable time ranges and granularities. It tracks daily DBU consumption alongside successful and failed job runs, directly correlating reliability with cost impact. By exposing failures as wasted DBUs, the view highlights the economic cost of instability. Spikes in failures relative to DBU usage help teams identify regressions, data anomalies, or cluster under-provisioning.

Revefi analysis extends to cluster-level attribution, ranking clusters by DBU usage across all clusters, all-purpose clusters, and job-scoped ephemeral clusters. For each cluster, DBUs and associated Databricks costs are tracked over time. This makes transient inefficiencies from short-lived job clusters observable and actionable.

Automated Job Optimization Engine

Revefi uses an experimentation-driven optimization workflow where selected Databricks jobs are executed multiple times under varying configurations. Results from each run are used to iteratively refine subsequent experiments.

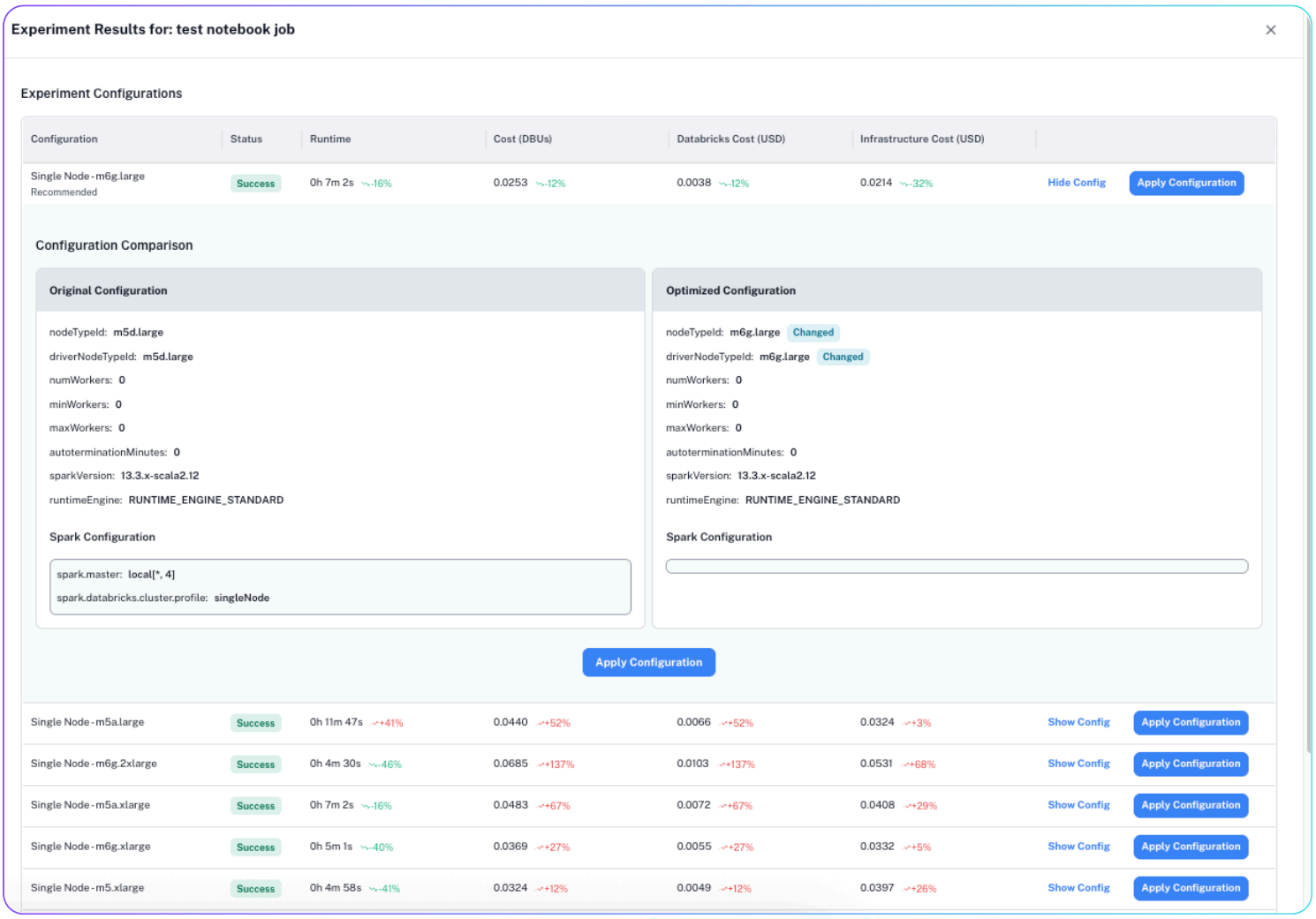


The optimization process terminates when a defined SLA runtime is met or a maximum number of runs is reached. Users can control target runtime thresholds, experiment limits, and whether results are applied manually or automatically. Advanced tuning parameters include Spark configuration settings, driver and worker node types, worker counts, storage configuration, and optional serverless execution.

This Optimization approach makes it measurable, controlled, and reversible, supporting safe application and workload deployment in production environments.

Experiment Results and Configuration

Users gain side-by-side comparisons of baseline and optimized job configurations. Key metrics include runtime, DBUs consumed, Databricks cost, and underlying infrastructure cost. Configuration differences are explicitly highlighted, such as changes in node types, driver sizing, worker topology, and Spark settings.



Each optimization experiment is tracked as a discrete run with a clear success status. Performance and cost improvements are quantified using percentage deltas for runtime and cost.

This comparison model enables engineers to validate optimization outcomes and understand the impact of each configuration change.With this transparency, data teams gain trust by understanding the recommendations, not just what to change.

Summary

Revefi makes managing Databricks Jobs measurable, optimizable, and economically transparent. By combining observability, experimentation, and automated configuration tuning and remediation, Revefi turns Databricks and Spark optimization from a black-box approach into a repeatable and manageable engineering discipline.

Learn More

Navigate to https://www.revefi.com/videos#category=Databricks to learn more about Databricks cost, performance, data quality and data operations use cases solved by Revefi.

Integrations

Organizations That Trust Revefi