

DATABRICKS DATA PLATFORM & APPS

HEALTH CHECK REPORT

A comprehensive view of productivity, cost and performance optimizations.

PREPARED FOR ACME CORPORATION

DATA FROM JULY 3 TO JULY 16, 2024

TABLE OF CONTENTS

These are just a small subset of the insights you get with Unravel to help you get familiarized with the platform, let's get started!

This is a health check report for the types of cost and performance insights Unravel provides you with. In this report, you will learn about the automation capabilities of Unravel's FinOps and Data Observability platform.

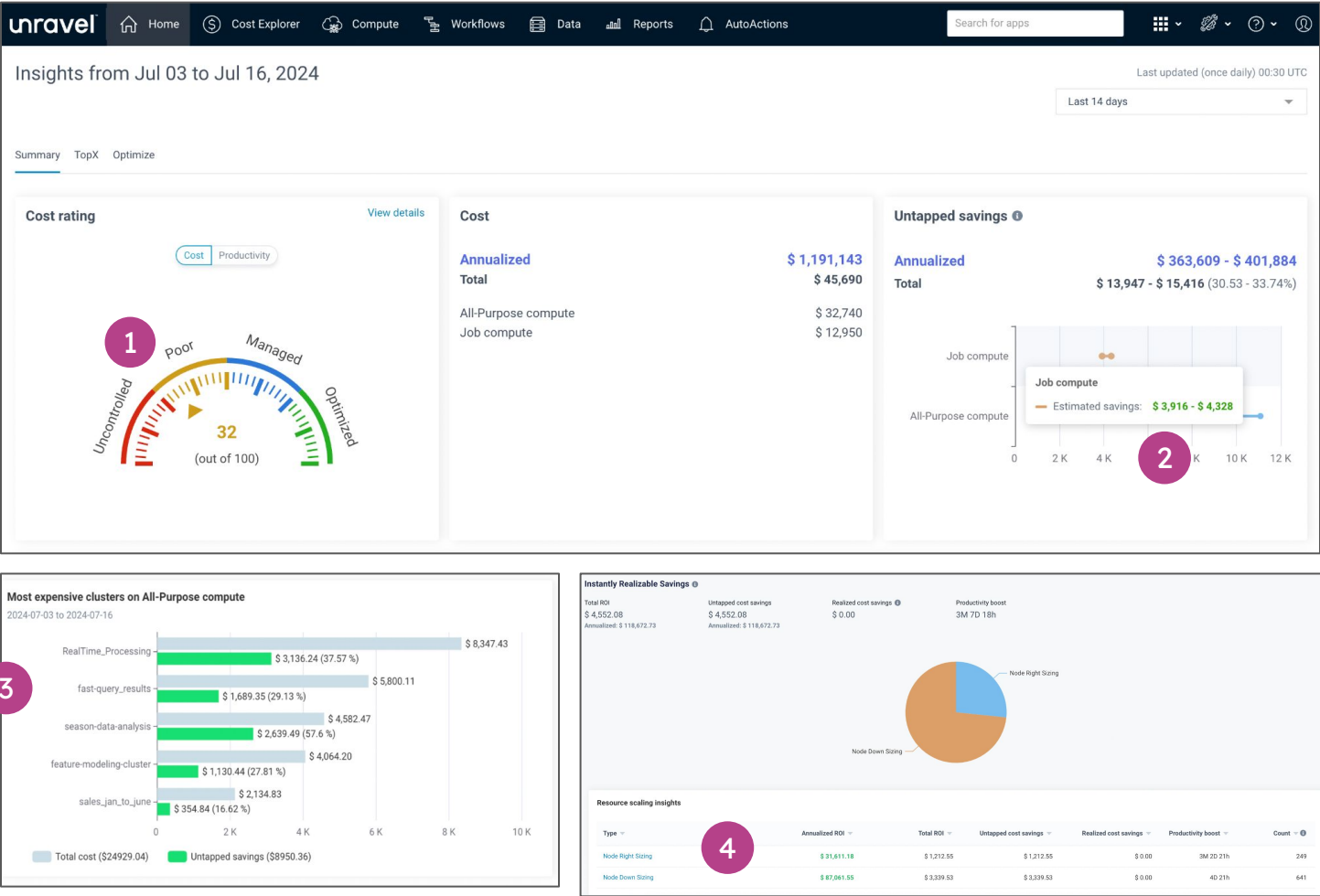
We will review a few examples of issues and insights automatically caught by Unravel and how simple it is to detect, root cause and resolve issues across your data platform, pipelines, and queries.

- 03** Cost Optimization
- 04** Performance Optimization
- 05** Data Optimization
- 06** Proactive Governance

Cost Optimization

COMPREHENSIVE OPTIMIZATION ACROSS JOBS, WORKSPACES, USERS

Unravel calculates your current ‘cost rating’, then provides clarity on where to focus (jobs, data, users, infra, etc) and what to do in order to improve your Databricks cost. Unravel even tracks the improvements made so you can keep on top of things.



Key findings

- 32 ‘cost rating’ indicates a poor cost health indicating that there are many recommendations from Unravel that can be applied to improve the ‘cost rating’
- 34% cost improvement (\$401K annually) can be realized by applying recommendations from Unravel.
- Get clarity on where to focus optimization effort by understanding top users, clusters, workspaces. Example: The ‘RealTime_Processing’ cluster can be improved by 37% by tuning a few things.
- Node rightsizing insights can help instantly realize over \$118k in annualized savings.

Performance Optimization

DEBUG AND TUNE JOBS OR WORKFLOWS FAST.
SAVE TIME TO FIND AND RESOLVE ERRORS, CODE INEFFICIENCIES, MISCONFIGURATIONS & MORE.

Unravel automatically pinpoints bottlenecks, errors, inefficiencies in complex jobs. Saving you time and effort, while guaranteeing performance SLAs.

unravel

Home

Cost Explorer

Compute

Workflows

Data

Reports

AutoActions

Search for apps

Jobs

Job Runs

Insights

PREVIEW

Bottlenecks (310)

\$ 2057.37 untapped savings

3Y 9M 3D 16h productivity boost

Inefficiencies (37)

\$ 47.13 untapped savings

1M 22D 23h productivity boost

Over-provisioned (249)

\$ 1212.55 untapped savings

312 Recs

Resource contention for CPU (8)

Resource contention for Disk (240)

Excessive IO activity detected (170)

Slow SQL Query execution detected (170)

Slow tasks detected in shuffle (204)

Job	Runs	Insight type	Total cost (USD)	Untapped cost savings (USD)	Realized cost savings (USD)
IFW-2 21885193892900	364	Bottlenecks (242) Over-provisioned (71) Inefficiencies (51)	\$ 881.01	\$ 386.17	\$ 4.60
PROD_TO_QA_SIT_DATA_COPY_TASK 107659794024666	416	Bottlenecks (537) Over-provisioned (273) Inefficiencies (15)	\$ 585.45	\$ 344.86	\$ 2.11
Optimize_goods_movement_vw 554884428569439	7	Bottlenecks (6)	\$ 310.89	\$ 211.50	\$ 1.49
AGW Rolling File Log 1107465903625637	310	Bottlenecks (593) Over-provisioned (284) Inefficiencies (1)	\$ 201.01	\$ 116.48	\$ 0.00
IFW_SCM_O9_GLOBAL_HANAPGE 663207972063795	161	Bottlenecks (223) Inefficiencies (7) Over-provisioned (3)	\$ 154.14	\$ 95.12	\$ 0.00
AGW Remelt Yeongju PI 69184050068041	155	Bottlenecks (281) Over-provisioned (190)	\$ 164.17	\$ 92.66	\$ 0.00

Analysis

Resources

Errors

Configuration

Logs

Update Insights

APP FAILURE

2 insights for failure analysis

DRIVER FAILED WITH OUTFOMEMORY ERROR

EXECUTOR FAILED WITH OUTFOMEMORY ERROR

4 executors in the application have failed with Java OutOfMemory error

Check the [usedHeap, maxHeap, committedHeap, totalPhysicalMemory] on the Resource Usage tab
Consider changing worker type with high memory resource

App Summary > Attempt-1

STATUS

COST

DURATION

DATA I/O

START TIME

END TIME

JOB COUNT

STAGES COUNT

SUCCESS

\$ 9.38

33m 34s

13.01 GB

04/22/2024

04/22/2024

1400

1400

Analysis

Resources

Errors

Configuration

Logs

Tags

Program

SQL

Timings

BOTTLENECKS

1 insight to make this app faster

BAD JOIN

2 distinct queries affected

Query 655 ***

Query 655 ***

Query 655 ***

Query 655 ***

Query 655 ***

Inefficient join condition is detected in operator SortMergeJoin [icast(d31f12907... of stage 375)

Inefficient join detected with join condition icast(d31f12907 as bigint) = icast(m73442221 as bigint) and icast(d2f12878 as bigint) = icast(m197442345 as bigint)

Parameter	Value
Tables involved	ntms_audit_log_1_ss_hist,ntms_audit_log_1_ss_curr
Number of rows of left relation	651400
Number of rows of right relation	597488
Join output rows	2409099

The total number of joined rows are (3.70) times more than maximum of left & right child.

This indicates join keys (icast(d31f12907 as bigint) = icast(m73442221 as bigint) and icast(d2f12878 as bigint) = icast(m197442345 as bigint)) are of low cardinality. Check if extra conditions can be applied to r more rows

Key findings

- Unravel has automatically detected and diagnosed several performance issues including bottlenecks and inefficiencies.
- This complex workflow with 1400 jobs took 34% more time to execute than usual. To understand the cause, would take 3 - 8 hours to diagnose manually. However, Unravel has automatically surfaced “Bad Joins” in the code and provides you a way to improve performance dramatically.
- In the second example, Unravel automatically diagnosed an “Out-of-memory error”, which caused the application to fail. This means users don’t have to waste time digging through logs to resolve errors.

DATA PLATFORM & APPS HEALTH CHECK REPORT

4

Data Optimization

DATA QUALITY, USAGE, ACCESS INTELLIGENCE & OPTIMIZATION.

Your guide to quickly understand which tables/files are being used, by whom, and when. How to improve data quality, and storage efficiency.

Metastore TypehiveWorkspaceData-ScienceMetastorejdbc:mariadb://consolidated-eastusc3...

unitycatalog

1

SearchDownload CSV

Total tables : 169

Table state

Hot (107)

Warm (25)

Cold (37)

Unknown (0)

Table Events

Missing ... (169)

Ineffici...(156)Only

Quality ... (Inefficient Storage Format)

Database	Table	Owner	Path	Table Type	File System	Storage Format	Created	Latest Access	Size	Apps	Partitions	Users
<input checked="" type="checkbox"/>	default1m_delta_table_repar...	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	15:25:37 07/31/2024	19:23:54 07/31/2024	---	5	0	2
<input type="checkbox"/>	default1m_delta_table_repar...	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	15:25:45 07/31/2024	19:23:54 07/31/2024	---	5	0	2
<input type="checkbox"/>	usecase data_quality_transac...	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	09:56:12 05/29/2024	09:09:56 07/25/2024	---	4	0	1
<input type="checkbox"/>	default5m_delta_shufflehash...	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	16:56:43 08/01/2024	12:10:31 08/02/2024	---	3	0	1
<input type="checkbox"/>	default5m_delta_shufflehash...	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	17:07:21 08/01/2024	12:10:31 08/02/2024	---	3	0	1
<input type="checkbox"/>	default1m_delta_table_repar...	root	dbfs:/user/hive/ware...	EXTERNAL_TABLE	DBFS	SEQUENCE_FILE	12:24:33 08/01/2024	13:11:32 08/01/2024	---	2	0	1
<input type="checkbox"/>	default1m_delta_table_repar...	root	dbfs:/user/hive/ware...	EXTERNAL_TABLE	DBFS	SEQUENCE_FILE	12:26:06 08/01/2024	13:11:32 08/01/2024	---	2	0	1
<input type="checkbox"/>	default500k_delta_3wayjoin	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	22:38:31 08/01/2024	05:52:07 08/05/2024	---	2	0	1
<input type="checkbox"/>	default500k_delta_3wayjoin....	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	22:45:06 08/01/2024	05:52:07 08/05/2024	---	2	0	1
<input type="checkbox"/>	default10k_delta_3wayjoin	root	dbfs:/user/hive/ware...	MANAGED_TABLE	DBFS	SEQUENCE_FILE	22:46:36 08/01/2024	05:52:07 08/05/2024	---	2	0	1

Key findings

1. Connect your Unity catalog or Hive Metastore for usage and access analytics.. This provides you clarity on which user used a dataset, when and in which job.
2. 21% of tables are“cold” and haven’t been used in a while. These could be deleted or moved to cold storage to save cost and improve efficiency.
3. Automatically surfaced inefficiencies such as 156 tables have “Inefficient storage format” which if fixed improves performance.

Proactive FinOps

PREVENT COST OVERRUNS.
AUTOMATE DETECTION, DIAGNOSES AND FIXES TO STAY UNDER BUDGET WHILE SCALING YOUR DATA ANALYTICS.

Unravel’s FinOps agent acts as your team member and works for you or with you to resolve cost issues.

1

Unravel Agent digest

13 new feeds

Today 12:02 PM

Unravel Agent applied 77 cost saving recommendations for **fraud-detection-ml** to reduce cost by 21%.

Cost vs untapped savings trend

Cost (\$USD)

4/16 4/17 4/18

Time

Legend: All-Purpose Clusters, Jobs Compute, Jobs Compute with Photon, SQL Compute, All-Pur, Job Compute untapped savings, Job Compute with Photon untapped savings, SQL Compute untapped savings, Delta Line Tables untapped savings, Productivity boost (hrs)

Review

Today 1:14 PM

Budget for **Marketing** business unit is at risk of overrun in the next 7 days.

unravel-prod

View

Business unit: Marketing

Cost center: unravel-prod

Cost center

\$ 22,047 / \$ 24,000

At risk

Auto-fix

Create ticket

Review

Today 3:25 PM

1,059 cold tables identified and deleted. Saved 3.03 TB of space.

Table state

Total 2,193

Hot (320)

Warm (814)

Cold (1,059)

1,059

ΔR 7%

2

Unmanaged cost centers

Cost center	Prev. month cost	Cost trend	Action
production	\$ 27,397	3.1 % ↑	Budget
staging	\$ 18,308	13.0 % ↑	Budget
dev	\$ 11,773	6.8 % ↑	Budget
red team	\$ 4,208	4.2 % ↑	Budget

3

Cost savings

Cost Untapped Realized

\$ 940k-1.2M

Annualized

\$ 80k-85k

This period

Type	Count	Savings
Warehouses	69	\$ 20.2k
Jobs	1.3k	\$ 18.5k
Clusters	810	\$ 15.1k
Tables	79	\$ 11.8k
Queries	989	\$ 10.5k
Photon jobs	381	\$ 7.7k

Key findings

1. Unravel’s FinOps agent taking cost-saving actions on your behalf. Or surfacing important areas needing your attention.
2. Proactive budget trackers to keep your teams and projects efficient.
3. Automatically surface every type of cost improvement possible across infrastructure, jobs, tables, users.

DATA PLATFORM & APPS HEALTH CHECK REPORT

6

Proactive Operations

PREVENT PERFORMANCE ISSUES AND DOWNTIME.

AUTOMATE DETECTION, DIAGNOSES AND FIXES TO ENSURE PERFORMANCE AND RELIABILITY STANDARDS.

Unravel's DataOps agent acts as your team member and works for you or with you to resolve issues.

1

Unravel Agent digest 7 new feeds

Today 12:41 PM
55 anomalous errors detected in **fraud_prevention** job created by [redacted]

Daily trend by business unit

Auto-fix Create ticket Review

Today 3:11 PM
13 of 55 bottlenecks on **JobsStagesSimulator** put the running job runs at risk of missing SLA in 7 hrs.

Impacted job runs

Start time	Run ID	Insights	Duration	User
2024-03-15 12:39 UTC	1047604054020035	Resource contention in driver (join condition is inefficient)	21m 14s	knad@unravel.net
2024-03-15 11:58 UTC	1047604054021182	Join condition is inefficient	47m 33s	dlad@unravel.net
2024-03-15 11:35 UTC	1047604054048750	Slow SQL Spark operator detected	18m 25s	nyad@unravel.net
2024-03-15 07:05 UTC	1047603980172076	Resource contention in driver	1h 13m 4s	nyad@unravel.net
2024-03-15 06:14 UTC	1047604054054705	Slow SQL Spark operator detected (join condition is inefficient)	38m 18s	prasad@unravel.net

Auto-fix Create ticket Review

Today 5:09 PM
24% additional issues were created this week for workspace **data-ml-forecast**.

Integrations

Tickets created	Incidents created	Notifications sent
133	23	258

CH/CO PRs with recommendations

Approved	In review	Merged
29	24	20

Unravel Agent recommendations

19	25	13
----	----	----

2

Over-provisioned (18)
\$ 3.5k untapped savings
\$ 1.7 realized savings

Failures (5)
77 hrs productivity savings

Bottlenecks (30)
102 hrs productivity savings
26% total speedup

Inefficiencies (13)
\$ 8.2k untapped savings
131 hrs productivity savings

☒ View all

☐ Custom

Top 5 Speedup

13% total speedup

☐ Selection

- ☒ Resource contention in driver (20) 18% total speedup
- ☒ Join condition is inefficient (1) 43% total speedup
- ☒ Slow SQL Spark operator (9) 6 SLAs at risk

Cancel Apply

Key findings

1. Unravel's DataOps surfacing errors, bottlenecks, and other problems proactively. From here Unravel can take action for you, or provide a guided remedy.
2. Proactive issues tracker with automatic diagnosis and remediation to help you keep your environment running optimally.

GET YOUR OWN HEALTH CHECK REPORT IN THREE EASY STEPS



Sign Up

Start the process using the
form below



Share

Securely connect to your
Databricks telemetry



Review

Receive your Health Check
report in your inbox

REQUEST YOUR FREE REPORT

About Unravel Data

Unravel Data is the first AI driven data observability and FinOps platform designed for the data stack. Unravel enables data driven organizations to identify and resolve performance, cost and reliability problems across platforms and pipelines. Its customers include Citibank, Maersk, Johnson & Johnson, UBS and Intel among others. Unravel Data is backed by Microsoft, GGV Capital, Menlo Ventures, Point72, Third Point Ventures and Harmony Ventures.

Visit www.unraveldata.com for more information.