DATA+AI SUMMIT
BY databricks

What's new in Databricks Workflows

R.R. Fäustlin, Product Management, Databricks

June 12, 2024

Product safe harbor statement

This information is provided to outline Databricks’ general product direction and is for **informational purposes only**. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all

Agenda

- The Databricks Workflows Story
- Recent innovations
- Looking ahead
- Demo

- **2015 Cron-based jobs**
- **2016 Notebook workflows**
- **2020 Jobs with multiple tasks**
  - Reliability
  - Monitoring
- **2022 The best lakehouse orchestrator**
  - Integration with the lakehouse
  - Streaming
  - Cluster reuse
- **2023 Serverless, performance and ease of use**
- **2024 AI powered ETL**

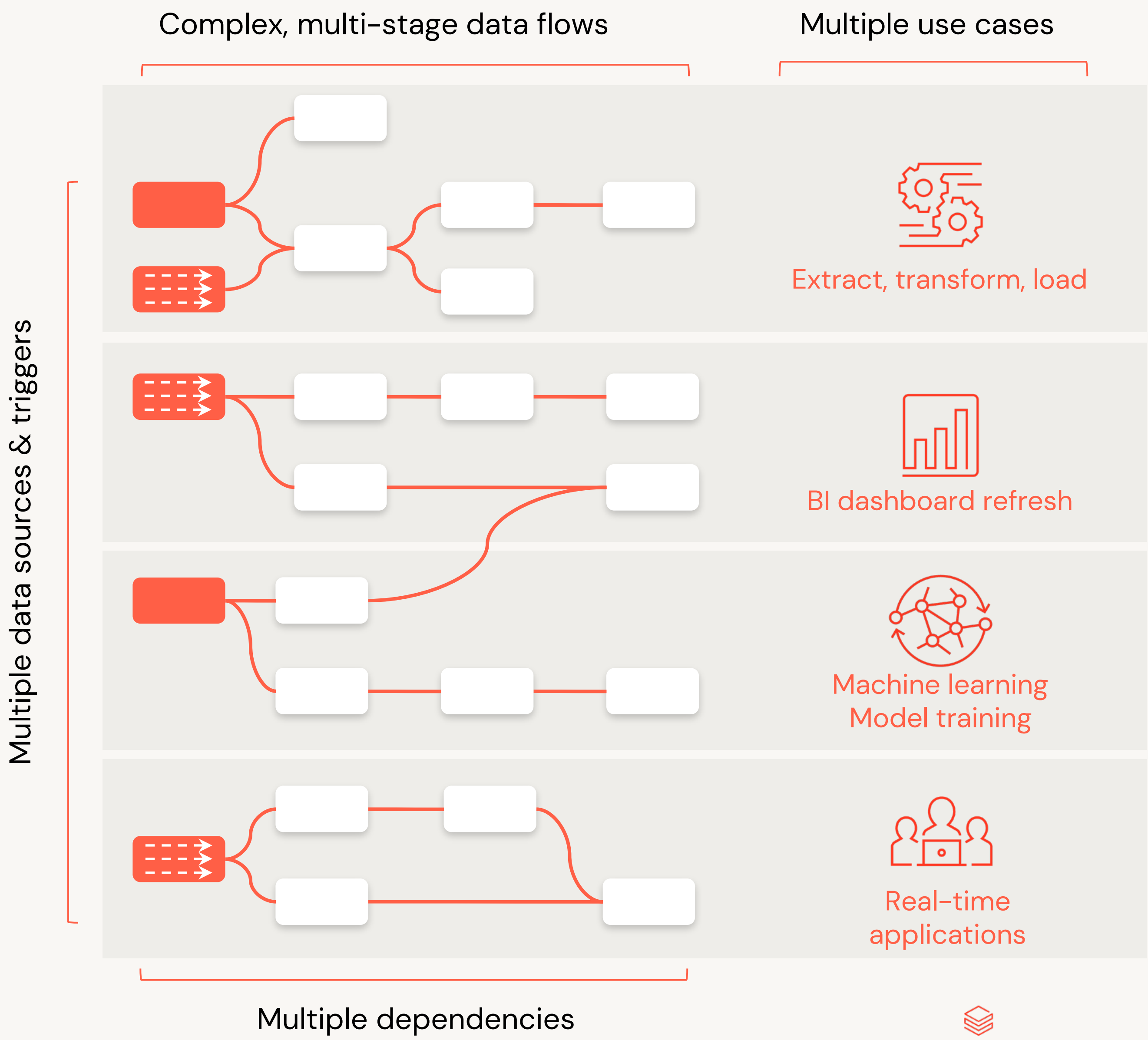
Modern data engineering requires modern data orchestration

Modern data engineering requires modern data orchestration

Orchestrating processes across all data, analytics and AI use cases is business critical

> “Data pipelines are growing in size, volume, and complexity, with multistage processing and dependencies between various data assets.”*

**Gartner Data Engineering Essentials, Patterns and Best Practices, September 2022*



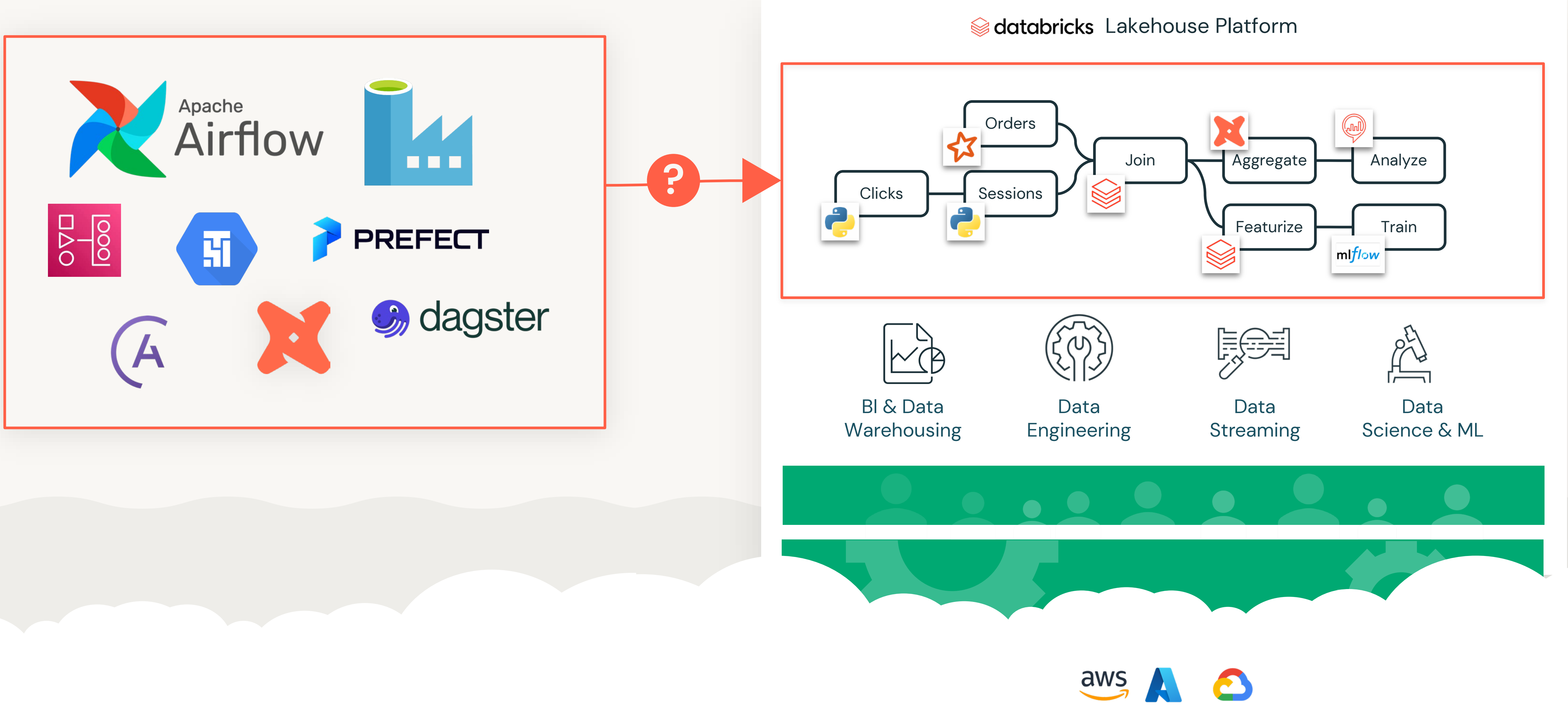
But organizations struggle with so many tools

>65%

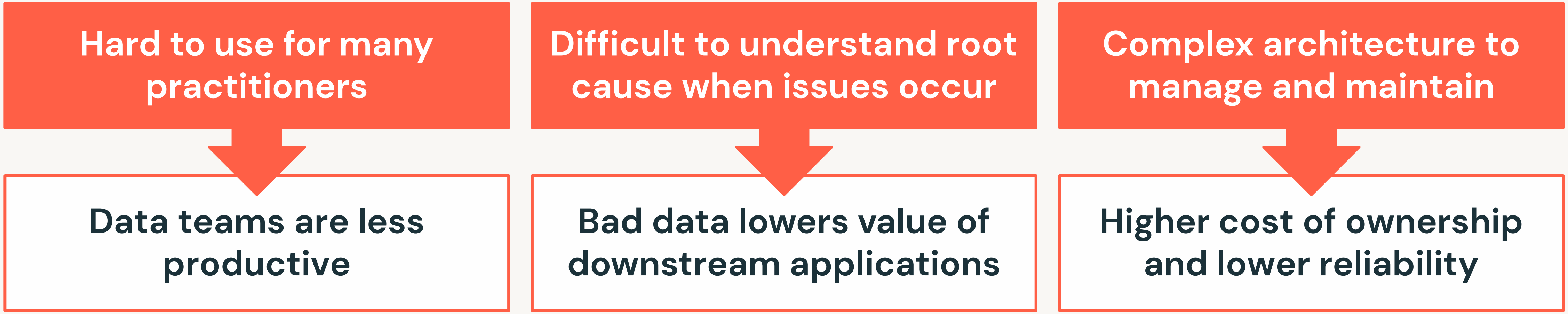
of organizations are using 10+ data engineering and intelligence tools

Source: IDC DataOps Survey, 2020

Many ways to orchestrate your Lakehouse



External orchestrators create challenges



Hard to use for many practitioners

Data teams are less productive

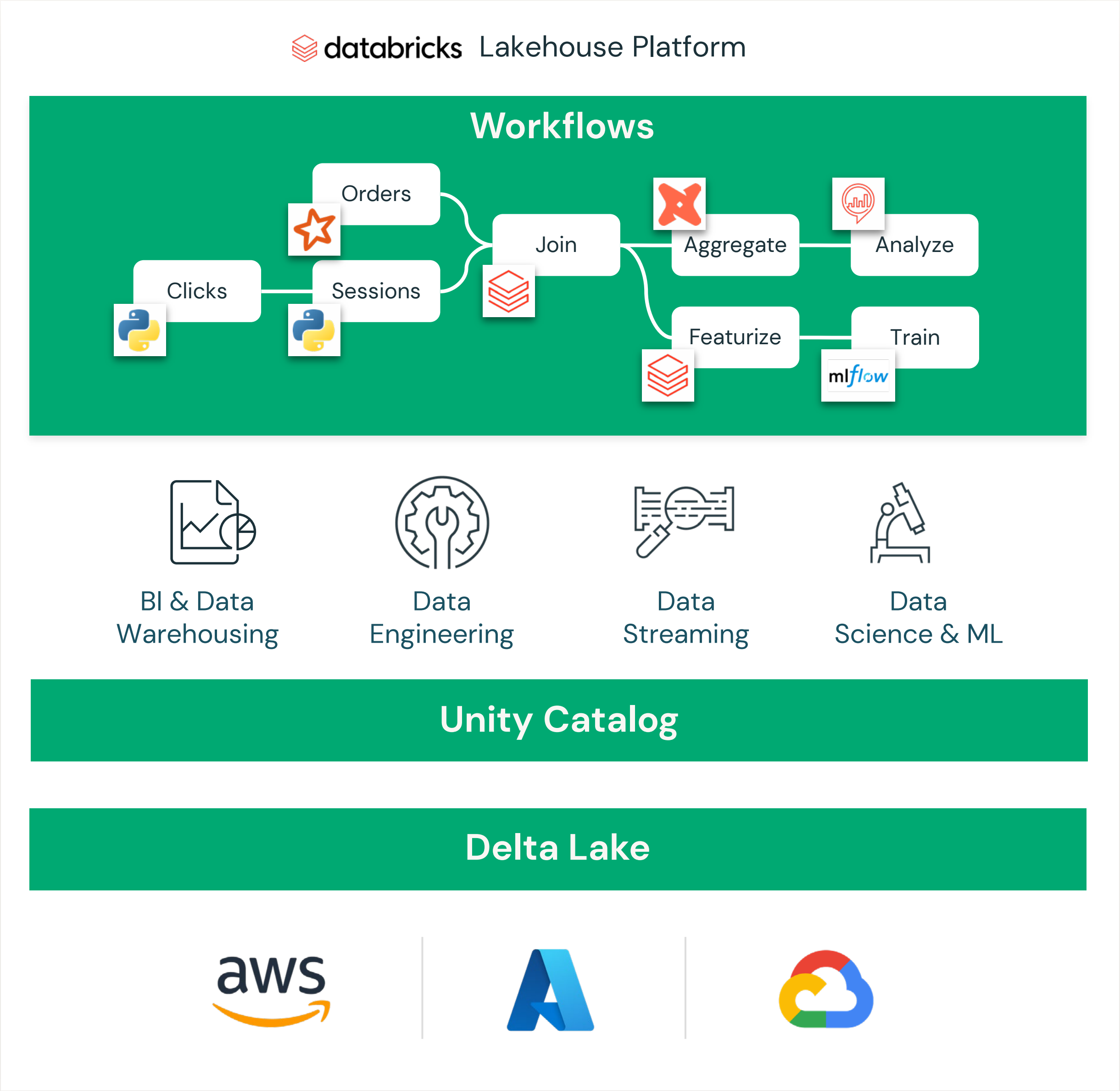
Difficult to understand root cause when issues occur

Bad data lowers value of downstream applications

Complex architecture to manage and maintain

Higher cost of ownership and lower reliability

These tools are not unified with your Lakehouse



Databricks Workflows

Unified orchestration for data, analytics, and AI on the Lakehouse Platform

- Simple authoring
- Actionable insights
- Proven reliability

Top 3 reasons why customers love Databricks Workflows

Simple authoring

for all data practitioners

Any data practitioner can accelerate their development by easily orchestrating Workflows from inside their Databricks workspace in just a few clicks. Advanced users can use their favorite IDEs with full support for CI/CD.

Actionable insights

from real-time monitoring

Full visibility into every task in every workflow. See the **health** of all your production workloads in **real-time** with detailed metrics and analytics to identify, troubleshoot, and fix issues fast.

Proven reliability

for production workloads

A **fully managed** service with serverless data processing and years of **99.95% uptime**. Workflows is trusted by thousands of Databricks customers running millions of production workloads.

>10k customers | >25 million VMs/day | >99.95% uptime

Disney streaming services

Adobe

AT&T

ABN·AMRO

John Deere

COMCAST

FedEx®

zalando

Walmart

Nielsen

HSBC

Wood Mackenzie
A Verisk Business

Improved collaboration
80-90% faster processing

Ahold Delhaize

ADF → Workflows
4.5x faster deployment
50% cost reduction

yipit DATA

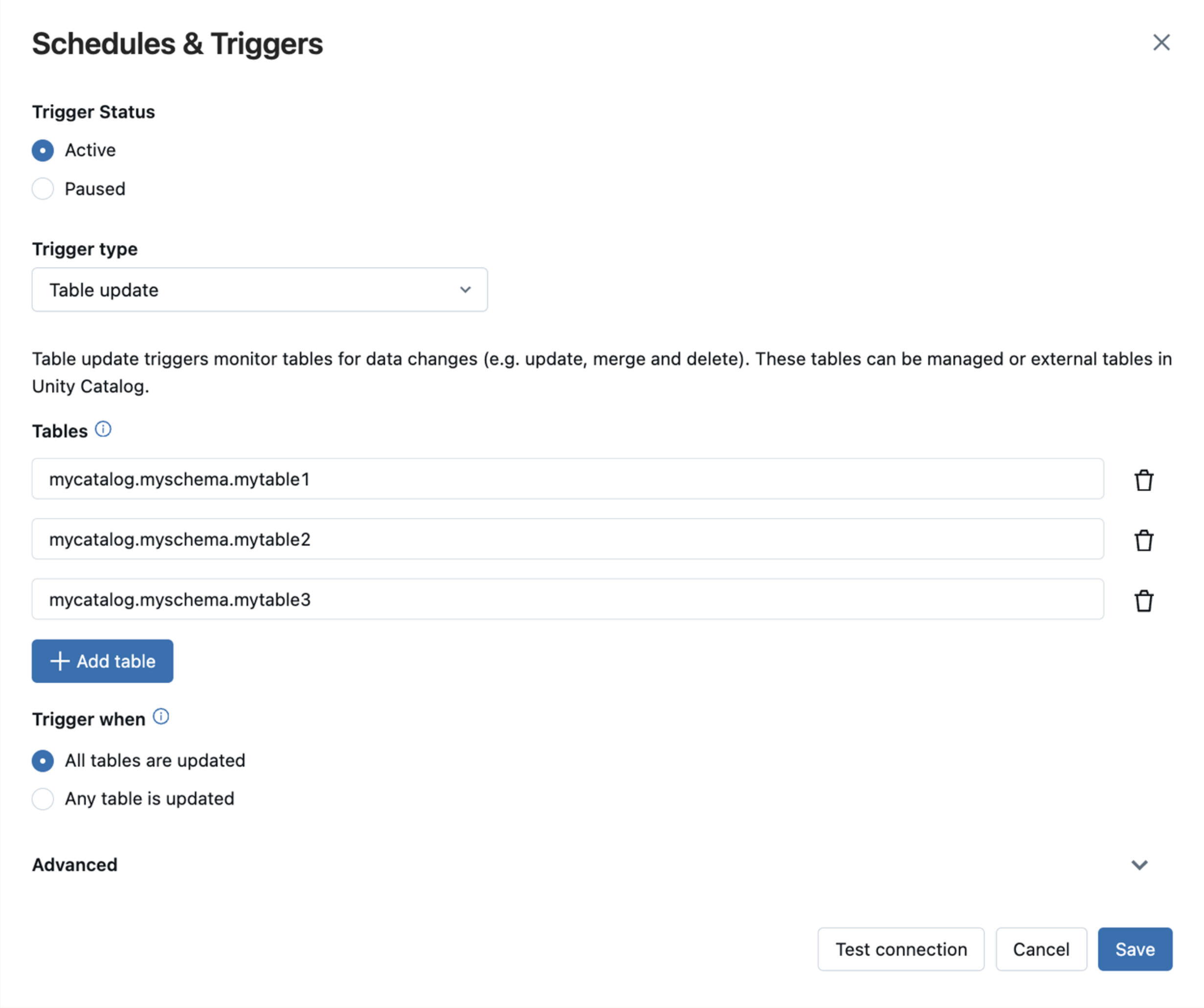
Airflow → Workflows
60% cost reduction
90% faster processing

>70 new features 🚀 shipped 🚀

This is the highlight reel.

Data triggers: Run only when you need to

- ✅ Trigger based on table change
- SOON Trigger when new files arrive
  Now unlimited files count and increasing number of triggers
- ✅ Trigger another job
  “job as a task”



Schedules & Triggers

Trigger Status

● Active

○ Paused

Trigger type

Table update

Table update triggers monitor tables for data changes (e.g. update, merge and delete). These tables can be managed or external tables in Unity Catalog.

Tables

mycatalog.myschema.mytable1

mycatalog.myschema.mytable2

mycatalog.myschema.mytable3

Trigger when

● All tables are updated

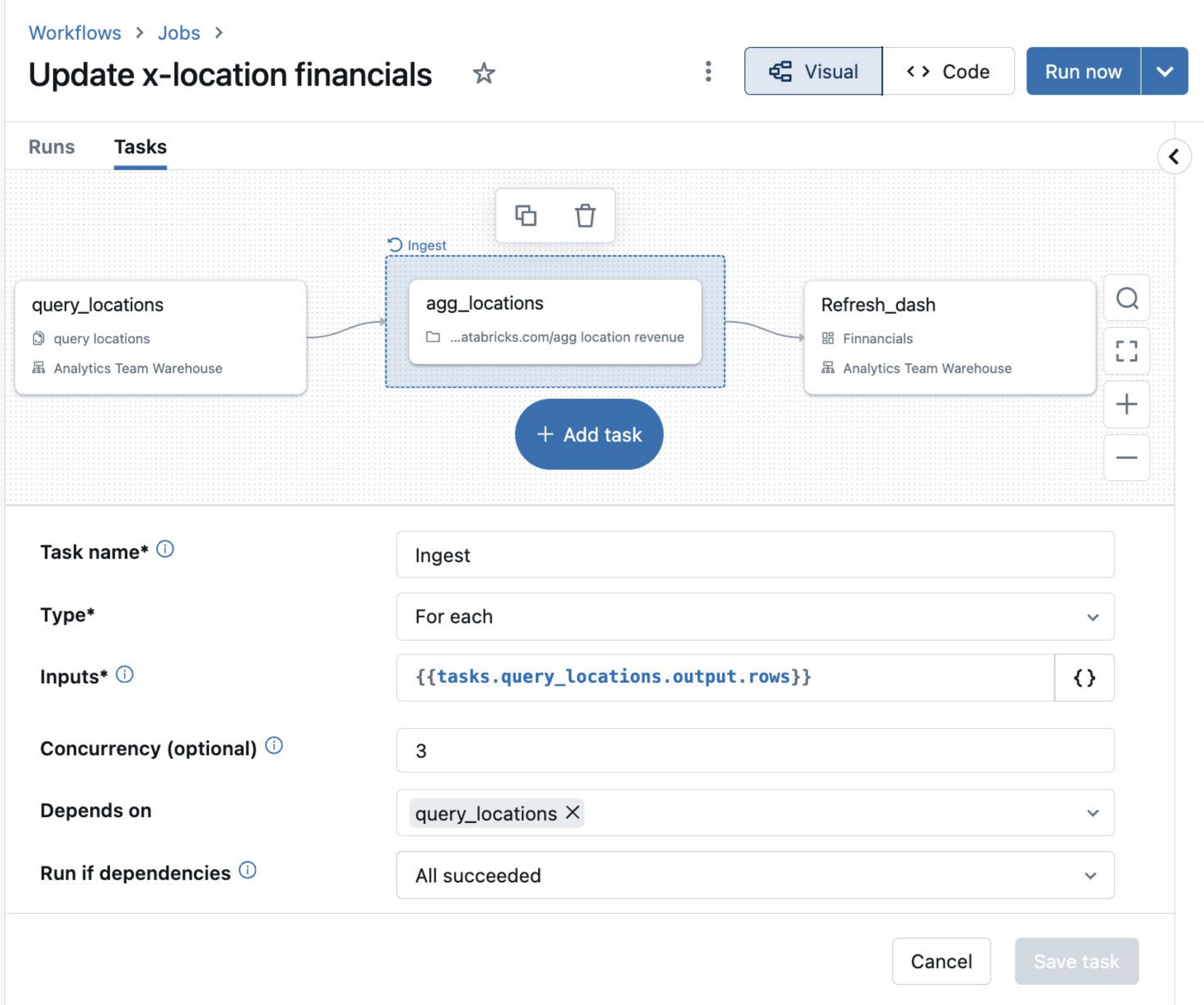
○ Any table is updated

Advanced

Advanced SQL orchestration

- ✅ Referencing of SQL query results in other tasks, e.g. for conditional execution
- ✅ Multi-SQL statement support

→ Full support for control flow, e.g. conditionals, for-each



Task name* ⓘ Ingest

Type* For each

Inputs* ⓘ `{{tasks.query_locations.output.rows}}`

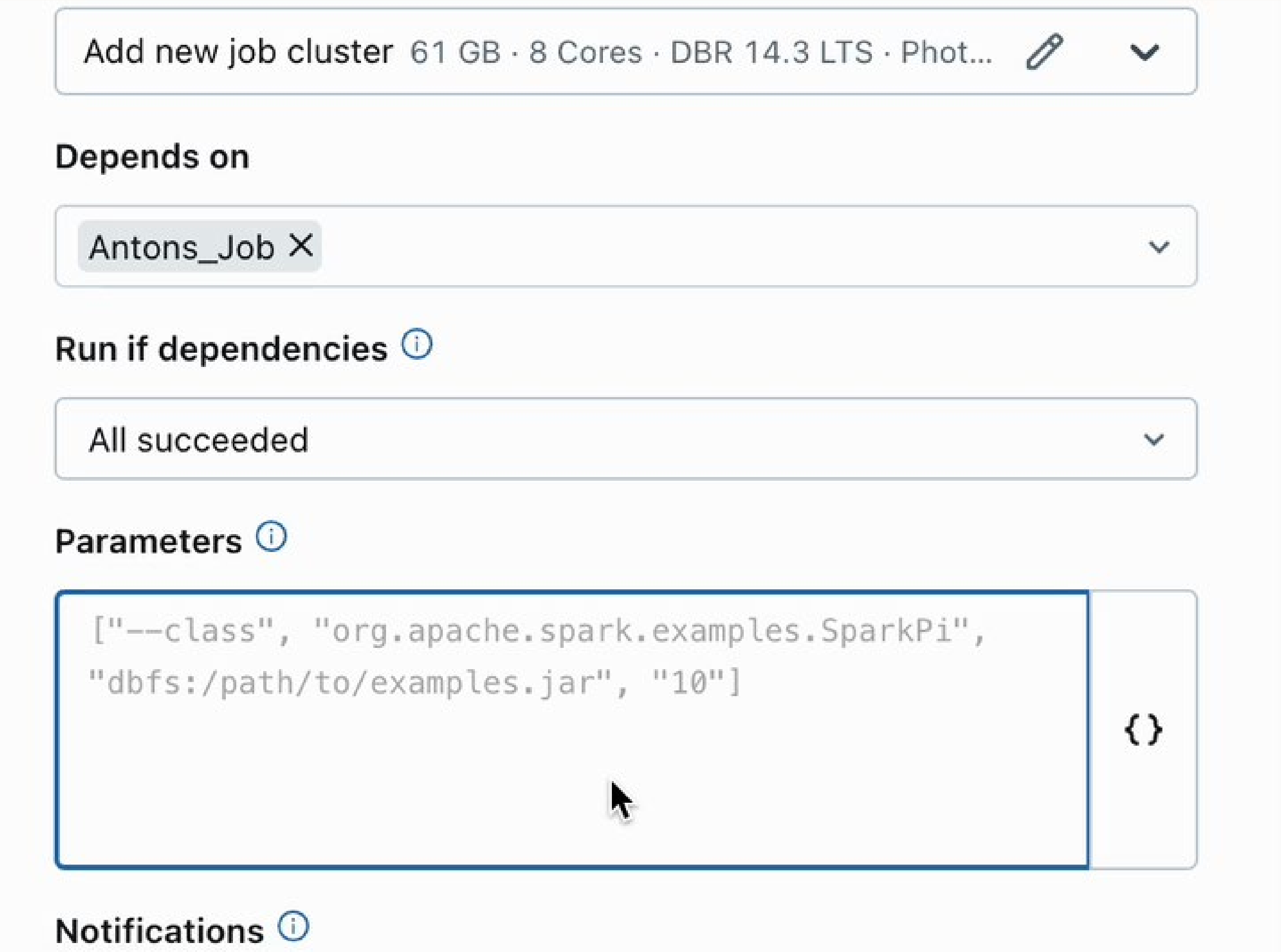
Concurrency (optional) ⓘ 3

Depends on query_locations ✕

Run if dependencies ⓘ All succeeded

Faster and easier

- ✅ Now up to 1k tasks per job
- ✅ Improved cluster defaults to set you up for success
- ✅ Easily exchange messages across tasks, now with a simplified UI and auto-complete



Add new job cluster 61 GB · 8 Cores · DBR 14.3 LTS · Phot...

Depends on

Antons_Job ✕

Run if dependencies ⓘ

All succeeded

Parameters ⓘ

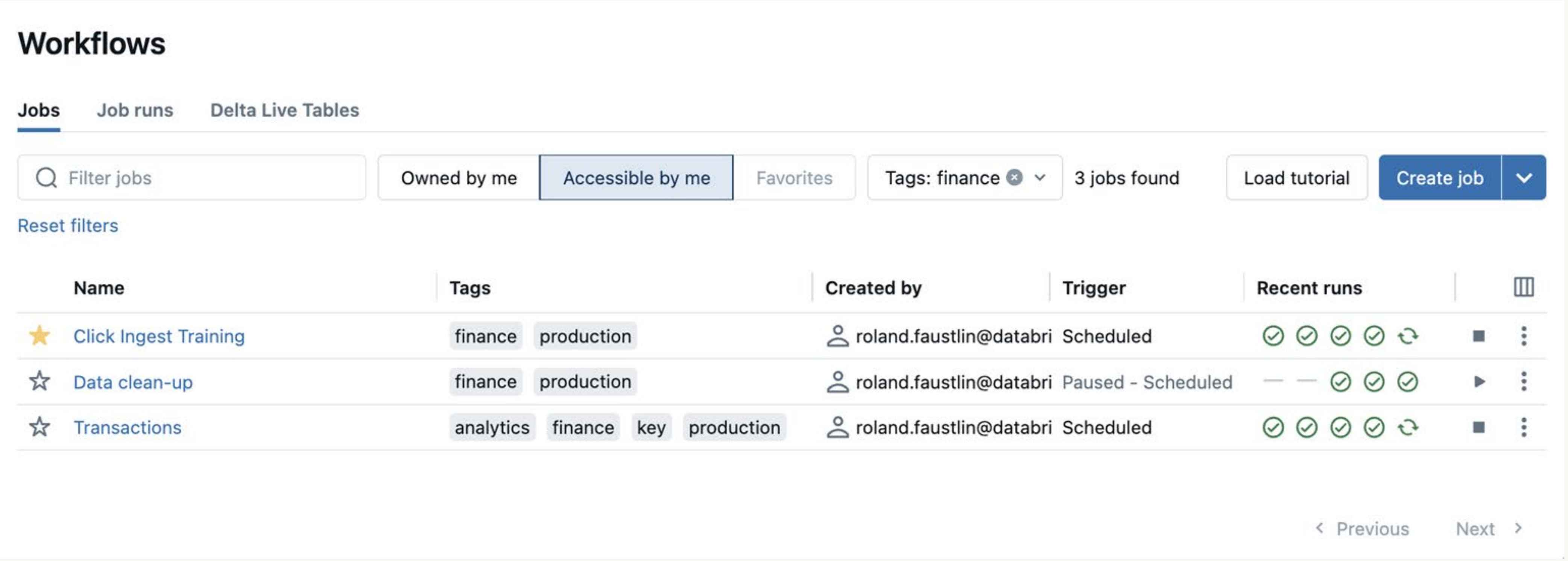
```
["--class", "org.apache.spark.examples.SparkPi",
"dbfs:/path/to/examples.jar", "10"]
```

Notifications ⓘ

Find your assets quickly

✅ Easily filter to the job or run that you care about, e.g. with job favourites ⭐

✅ Descriptions for your jobs and tasks



Workflows

Jobs | Job runs | Delta Live Tables

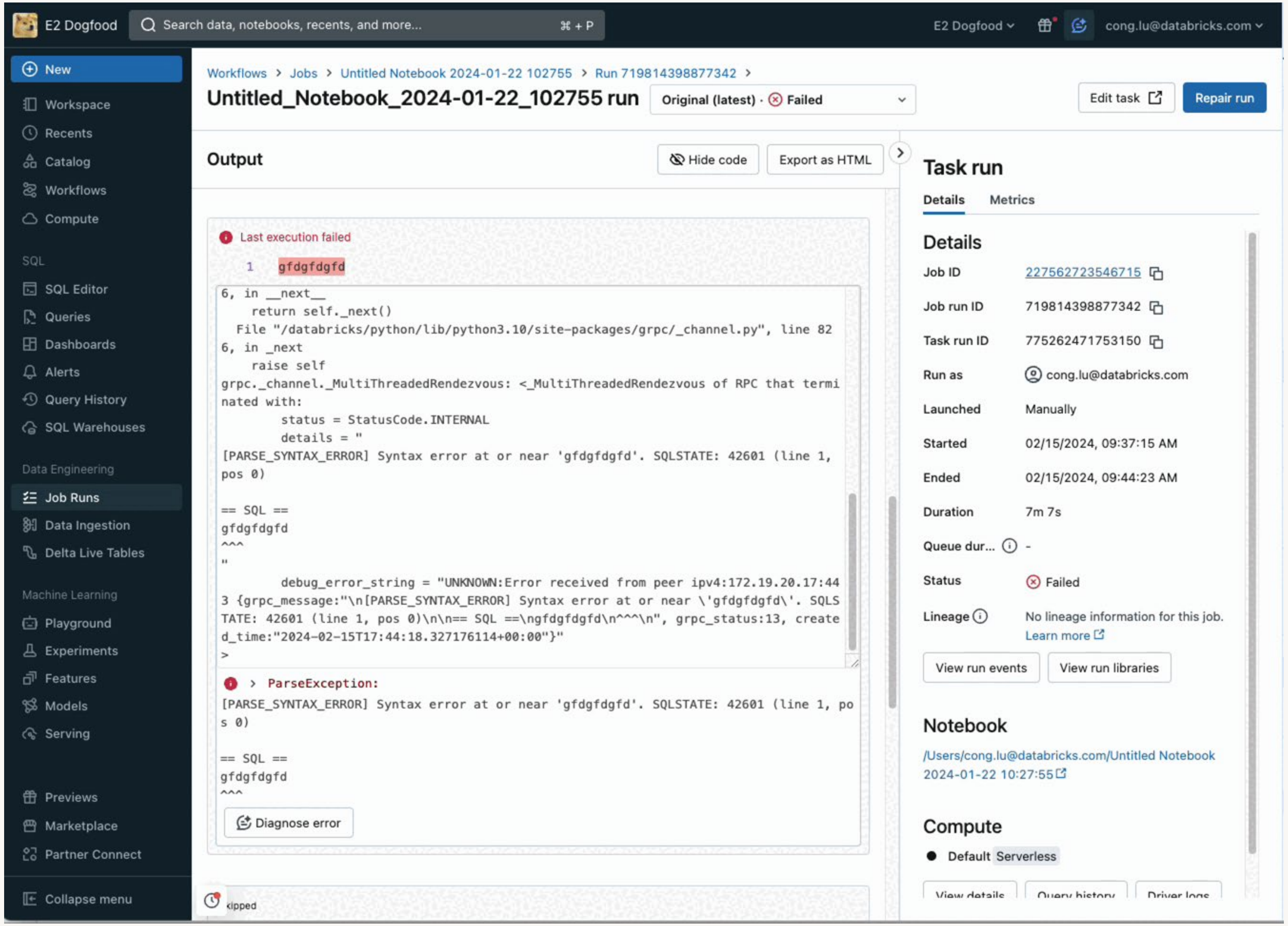
Filter jobs | Owned by me | Accessible by me | Favorites | Tags: finance | 3 jobs found | Load tutorial | Create job

Reset filters

	Name	Tags	Created by	Trigger	Recent runs		
★	Click Ingest Training	finance production	roland.faustlin@databri	Scheduled	✓ ✓ ✓ ✓ ↻	■	⋮
☆	Data clean-up	finance production	roland.faustlin@databri	Paused - Scheduled	— — ✓ ✓ ✓	▶	⋮
☆	Transactions	analytics finance key production	roland.faustlin@databri	Scheduled	✓ ✓ ✓ ✓ ↻	■	⋮

‹ Previous | Next ›

✅ AI assisted debugging integrated



Workflows > Jobs > Untitled Notebook 2024-01-22 102755 > Run 719814398877342 >

Untitled_Notebook_2024-01-22_102755 run   Original (latest) · Failed

Output

Last execution failed

```
gfdgfdgfd
```

```
6, in __next__
    return self._next()
  File "/databricks/python/lib/python3.10/site-packages/grpc/_channel.py", line 82
6, in _next
    raise self
grpc._channel._MultiThreadedRendezvous: <_MultiThreadedRendezvous of RPC that termi
nated with:
        status = StatusCode.INTERNAL
        details = "
[PARSE_SYNTAX_ERROR] Syntax error at or near 'gfdgfdgfd'. SQLSTATE: 42601 (line 1,
pos 0)

== SQL ==
gfdgfdgfd
^^^
"
        debug_error_string = "UNKNOWN:Error received from peer ipv4:172.19.20.17:44
3 {grpc_message:"\n[PARSE_SYNTAX_ERROR] Syntax error at or near \'gfdgfdgfd\'. SQLS
TATE: 42601 (line 1, pos 0)\n\n== SQL ==\ngfdgfdgfd\n^^^\n", grpc_status:13, create
d_time:"2024-02-15T17:44:18.327176114+00:00"}"
>
```

```
> ParseException:
[PARSE_SYNTAX_ERROR] Syntax error at or near 'gfdgfdgfd'. SQLSTATE: 42601 (line 1, po
s 0)

== SQL ==
gfdgfdgfd
^^^
```

Task run

Details

Job ID	227562723546715
Job run ID	719814398877342
Task run ID	775262471753150
Run as	cong.lu@databricks.com
Launched	Manually
Started	02/15/2024, 09:37:15 AM
Ended	02/15/2024, 09:44:23 AM
Duration	7m 7s
Queue dur...	-
Status	Failed
Lineage	No lineage information for this job. Learn more

Notebook

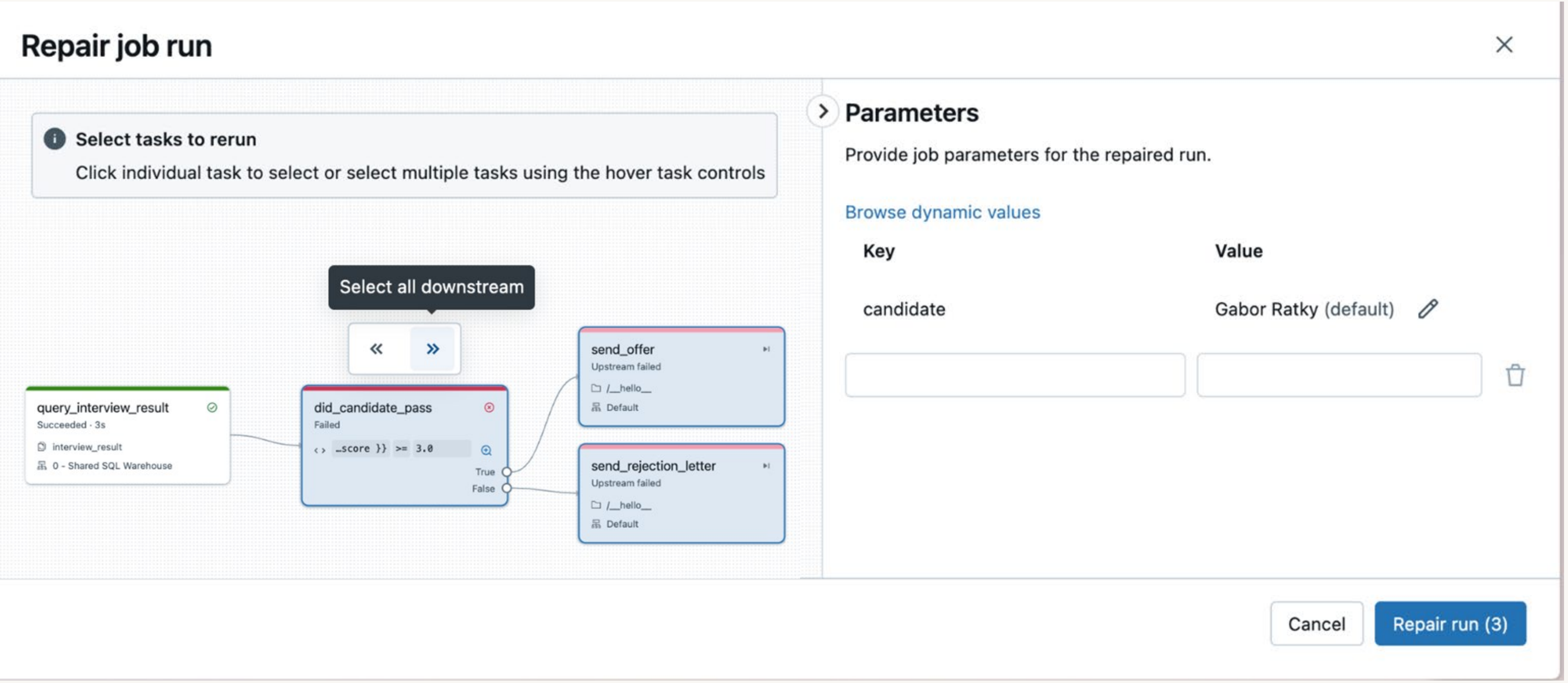
/Users/cong.lu@databricks.com/Untitled Notebook 2024-01-22 10:27:55

Compute

● Default Serverless

Only run what you need

- ✅ Run only the tasks you need to get back on track, now also single and successful tasks



Repair job run

Select tasks to rerun

Click individual task to select or select multiple tasks using the hover task controls

Parameters

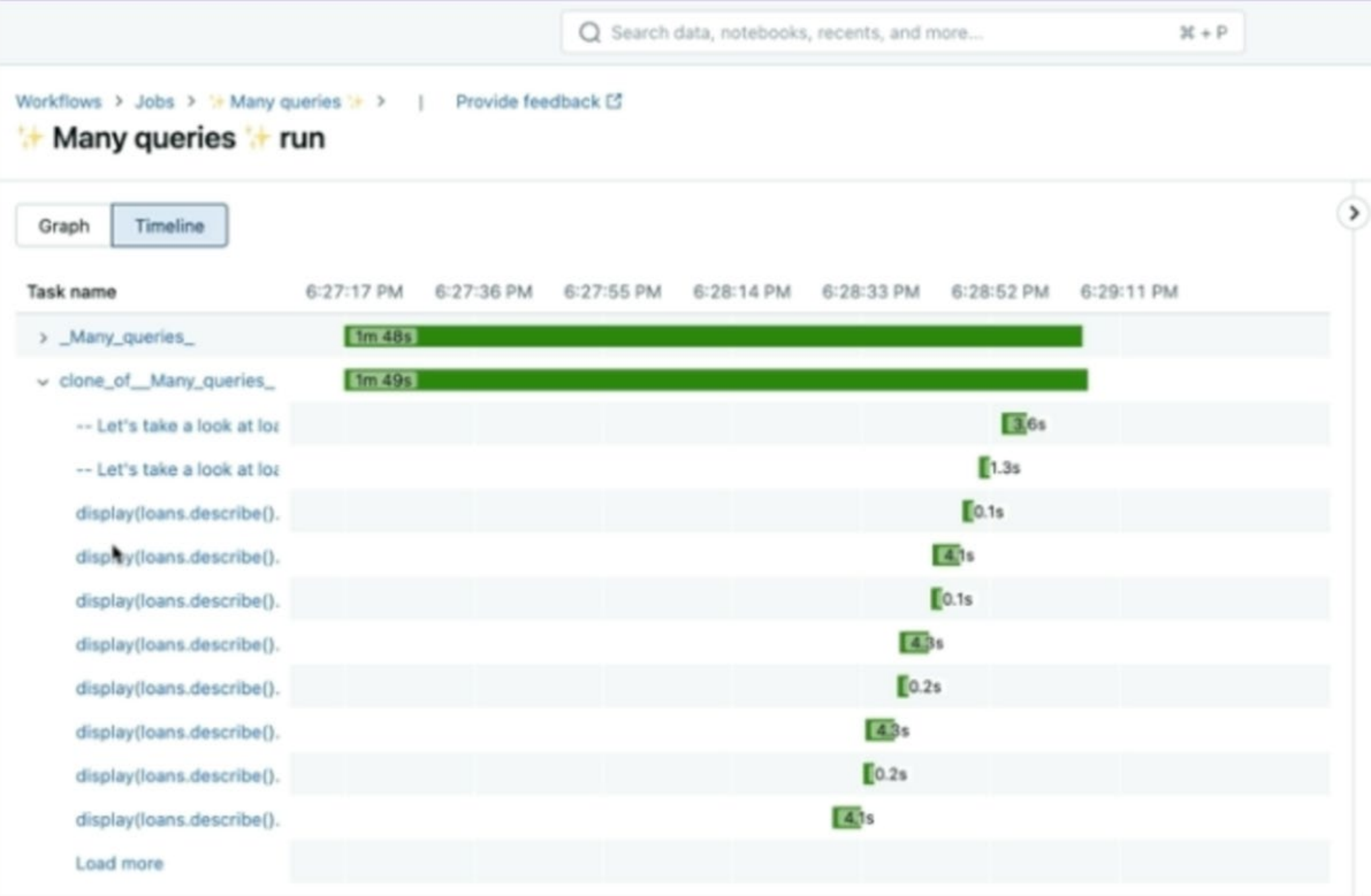
Provide job parameters for the repaired run.

Browse dynamic values

Key	Value
candidate	Gabor Ratky (default)

Easily optimize price/performance

- ☑ Timeline view across tasks and queries
- ☑ Query profile integration
- ☑ Track streaming lag and alert on deviations
- ☑ Alert when jobs are running late



Search data, notebooks, recents, and more... ⌘ + P

Workflows > Jobs > Many queries > | Provide feedback

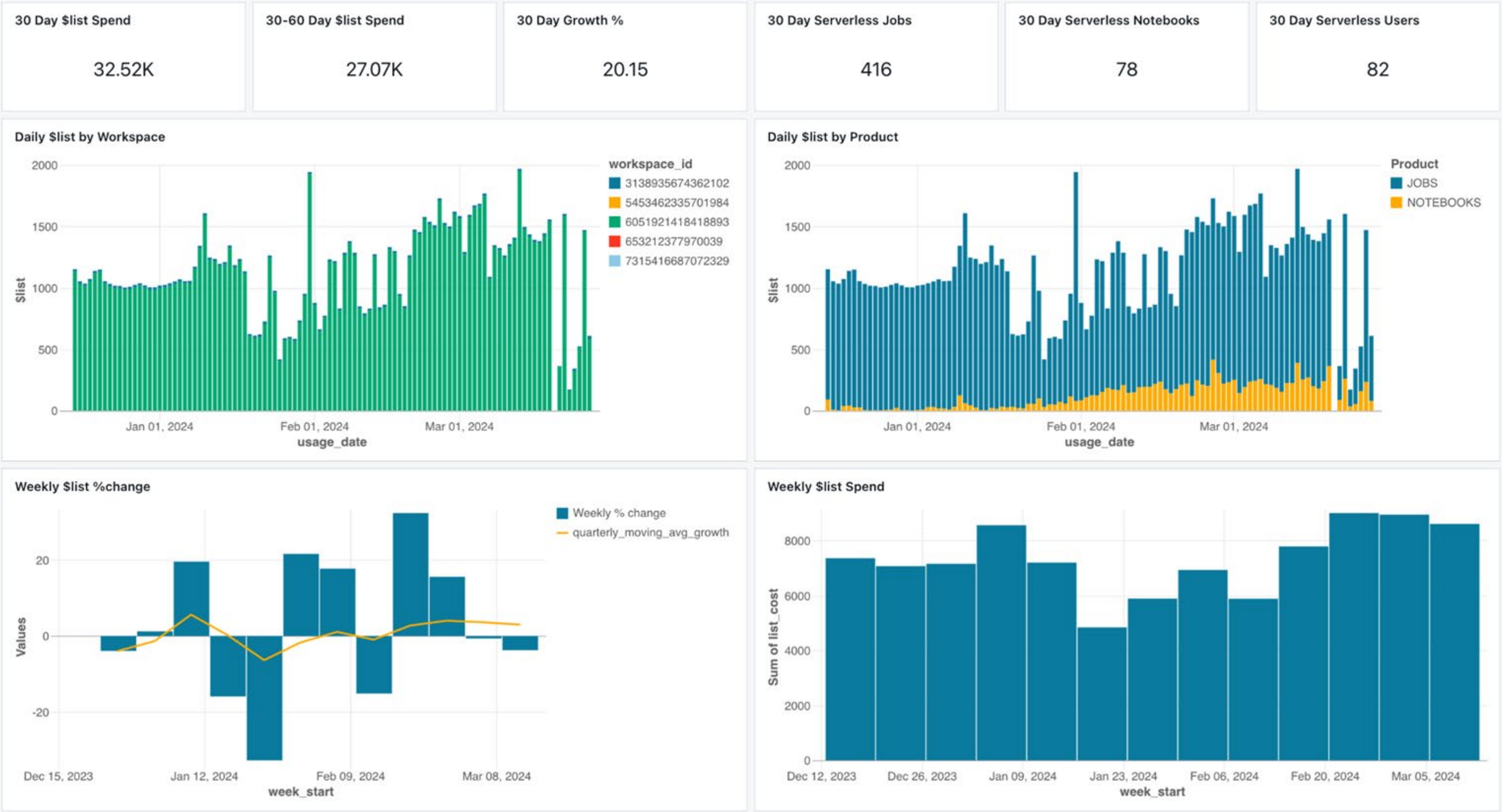
Many queries run

Task name | 6:27:17 PM | 6:27:36 PM | 6:27:55 PM | 6:28:14 PM | 6:28:33 PM | 6:28:52 PM | 6:29:11 PM

- > _Many_queries_ — 1m 48s
- ⌄ clone_of__Many_queries_ — 1m 49s
  - -- Let's take a look at loa — 3.6s
  - -- Let's take a look at loa — 1.3s
  - display(loans.describe(). — 0.1s
  - display(loans.describe(). — 4.1s
  - display(loans.describe(). — 0.1s
  - display(loans.describe(). — 4.3s
  - display(loans.describe(). — 0.2s
  - display(loans.describe(). — 4.3s
  - display(loans.describe(). — 0.2s
  - display(loans.describe(). — 4.1s
  - Load more

Track cost and long-term trends

- ✅ System tables and templated/customizable dashboards
- SOON In-UI budget monitoring and alerting



30 Day $list Spend
32.52K

30-60 Day $list Spend
27.07K

30 Day Growth %
20.15

30 Day Serverless Jobs
416

30 Day Serverless Notebooks
78

30 Day Serverless Users
82

PyDABs: Anything in Databricks as code

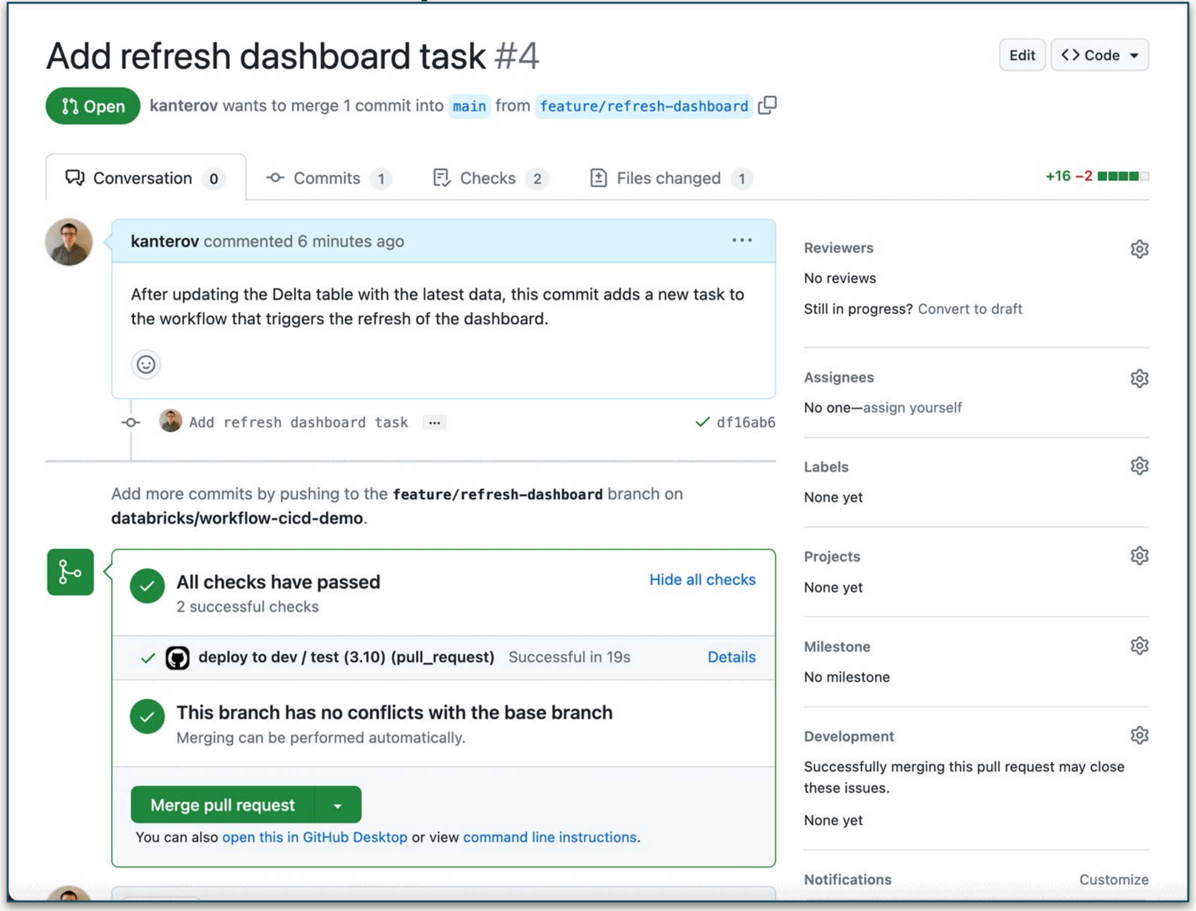
- ☑ Python SDK
- ☑ Terraform support
- ☑ Run jobs as service principal

→ Easily develop Workflows in your IDE as Python code

→ Compare changes

→ Collaborate with UI-only users

Code review your Workflows



Add refresh dashboard task #4

kanterov wants to merge 1 commit into `main` from `feature/refresh-dashboard`

kanterov commented 6 minutes ago

After updating the Delta table with the latest data, this commit adds a new task to the workflow that triggers the refresh of the dashboard.

`Add refresh dashboard task` ✓ `df16ab6`

Add more commits by pushing to the **`feature/refresh-dashboard`** branch on **databricks/workflow-cicd-demo**.

All checks have passed
2 successful checks

✓ **deploy to dev / test (3.10) (pull_request)** Successful in 19s

This branch has no conflicts with the base branch
Merging can be performed automatically.

You can also open this in GitHub Desktop or view command line instructions.

Reviewers
No reviews
Still in progress? Convert to draft

Assignees
No one—assign yourself

Labels
None yet

Projects
None yet

Milestone
No milestone

Development
Successfully merging this pull request may close these issues.
None yet

Notifications
Customize

INGESTION CONNECTORS:

Breakout session on Thursday at 11:20 AM

Efficient data ingestion for everyone

Simple and low-maintenance ➡ Fewer headaches, quicker time to value, democratized data

Unified with the lakehouse ➡ Secure and healthy pipelines that live where you do your work

Efficient end-to-end ➡ Lower costs, better performance, better scalability

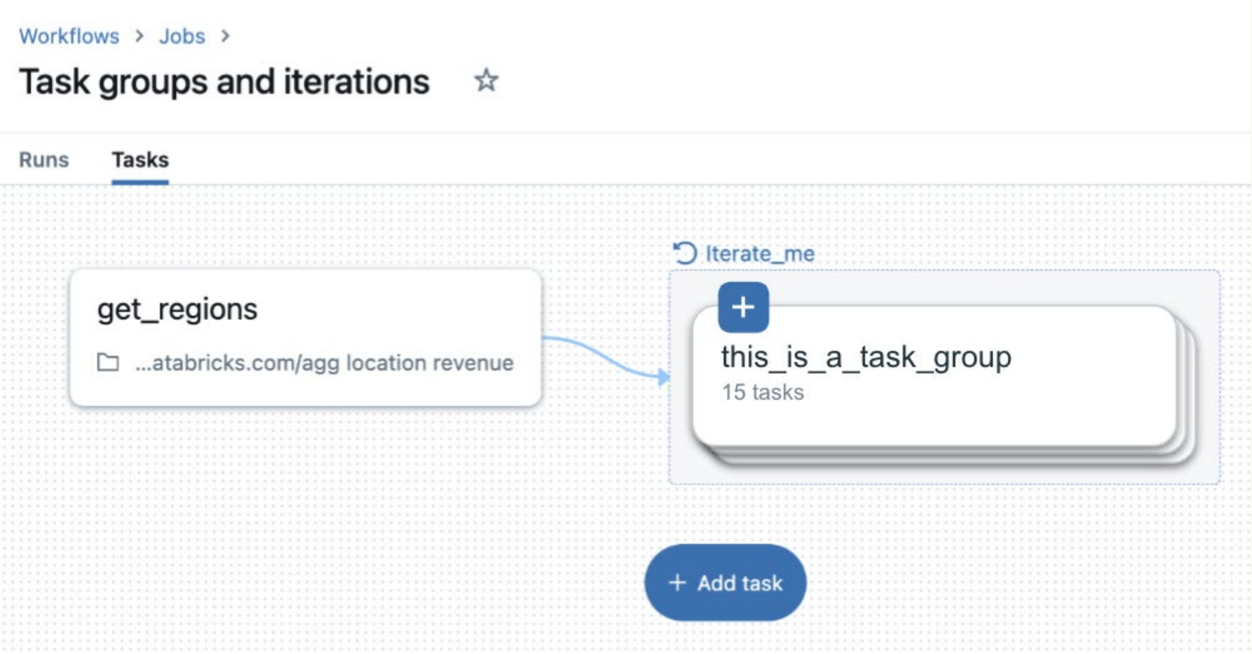
salesforce	Private preview
workday.	Private preview
SQL	Private preview
ORACLE NETSUITE	Coming soon
servicenow	Coming soon
Google Analytics	Coming soon
SharePoint	Coming soon
PostgreSQL	Coming soon

And we are not done. 💪

preliminary

More trigger and control-flow options

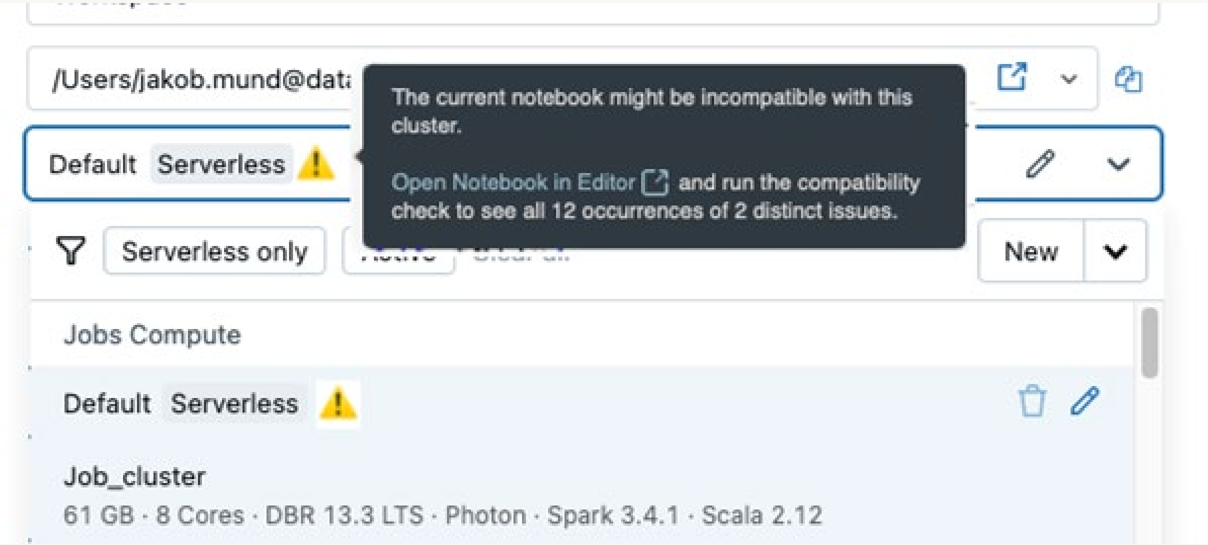
- ⏭ For-each loops across multiple tasks
- ⏭ Task groups: Visual segmentation of large DAGs
- ⏭ Periodic triggers: Run every week, day, hour
- ⏭ Multiple triggers per job
- ⏭ Queuing on infrastructure resources, e.g. instance pool availability



preliminary

Unity Catalog and Serverless compatibility

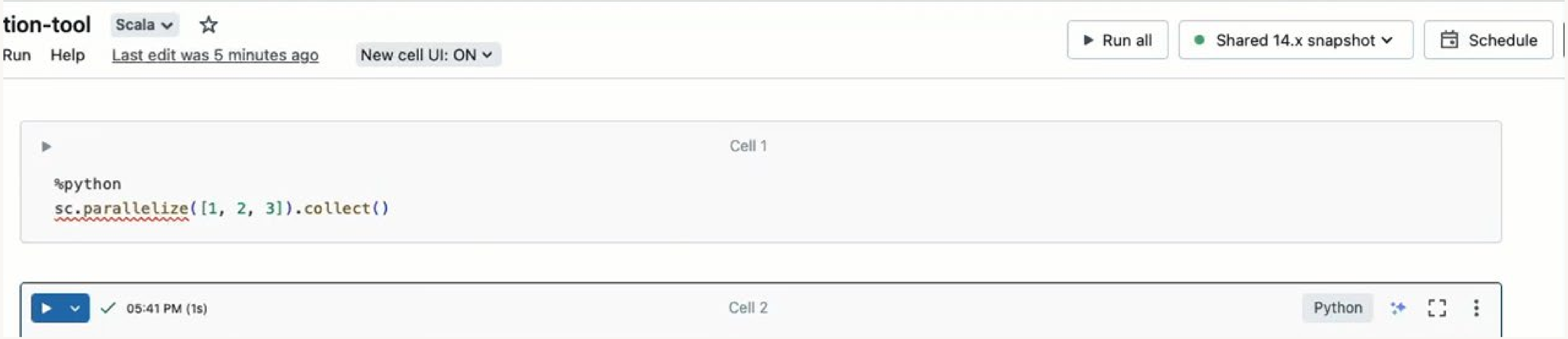
- ⏭ Compatibility checks for Unity Catalog and serverless
- ⏭ AI assisted code updates



The current notebook might be incompatible with this cluster.

Open Notebook in Editor and run the compatibility check to see all 12 occurrences of 2 distinct issues.

61 GB · 8 Cores · DBR 13.3 LTS · Photon · Spark 3.4.1 · Scala 2.12



```
%python
sc.parallelize([1, 2, 3]).collect()
```

preliminary

Operational health across the Lakehouse

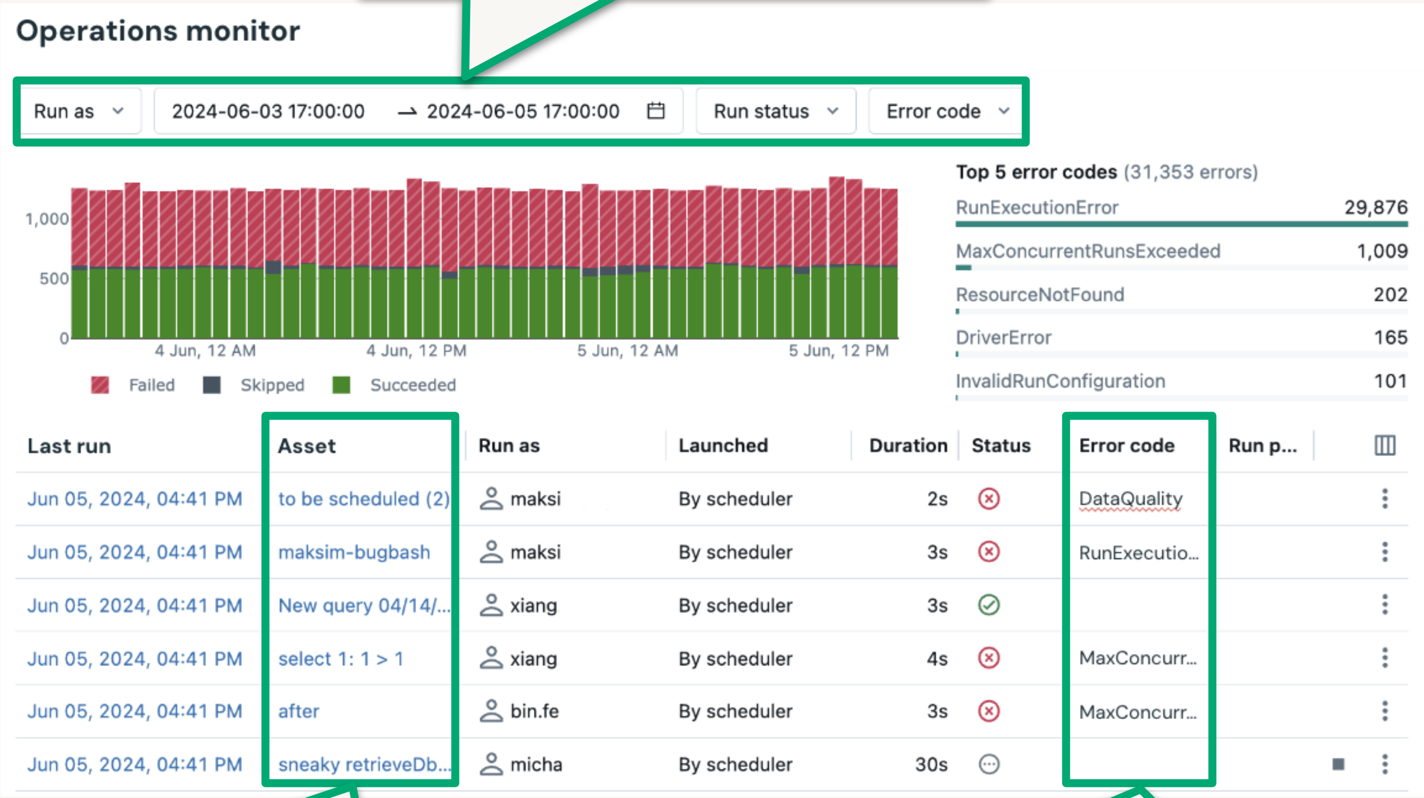
Single operational view

→ Health metrics across all assets

→ Data health monitoring

→ Anomaly detection

Filter, e.g. by team or alert type



Top 5 error codes (31,353 errors)

RunExecutionError	29,876
MaxConcurrentRunsExceeded	1,009
ResourceNotFound	202
DriverError	165
InvalidRunConfiguration	101

Last run	Asset	Run as	Launched	Duration	Status	Error code	Run p...
Jun 05, 2024, 04:41 PM	to be scheduled (2)	maksi	By scheduler	2s	⊗	DataQuality	
Jun 05, 2024, 04:41 PM	maksim-bugbash	maksi	By scheduler	3s	⊗	RunExecutio...	
Jun 05, 2024, 04:41 PM	New query 04/14/...	xiang	By scheduler	3s	✓		
Jun 05, 2024, 04:41 PM	select 1: 1 > 1	xiang	By scheduler	4s	⊗	MaxConcurr...	
Jun 05, 2024, 04:41 PM	after	bin.fe	By scheduler	3s	⊗	MaxConcurr...	
Jun 05, 2024, 04:41 PM	sneaky retrieveDb...	micha	By scheduler	30s	⋯		■

Any asset, e.g. job, pipeline, table

Any issue, e.g. data quality

Serverless Compute

SIMPLE and FAST

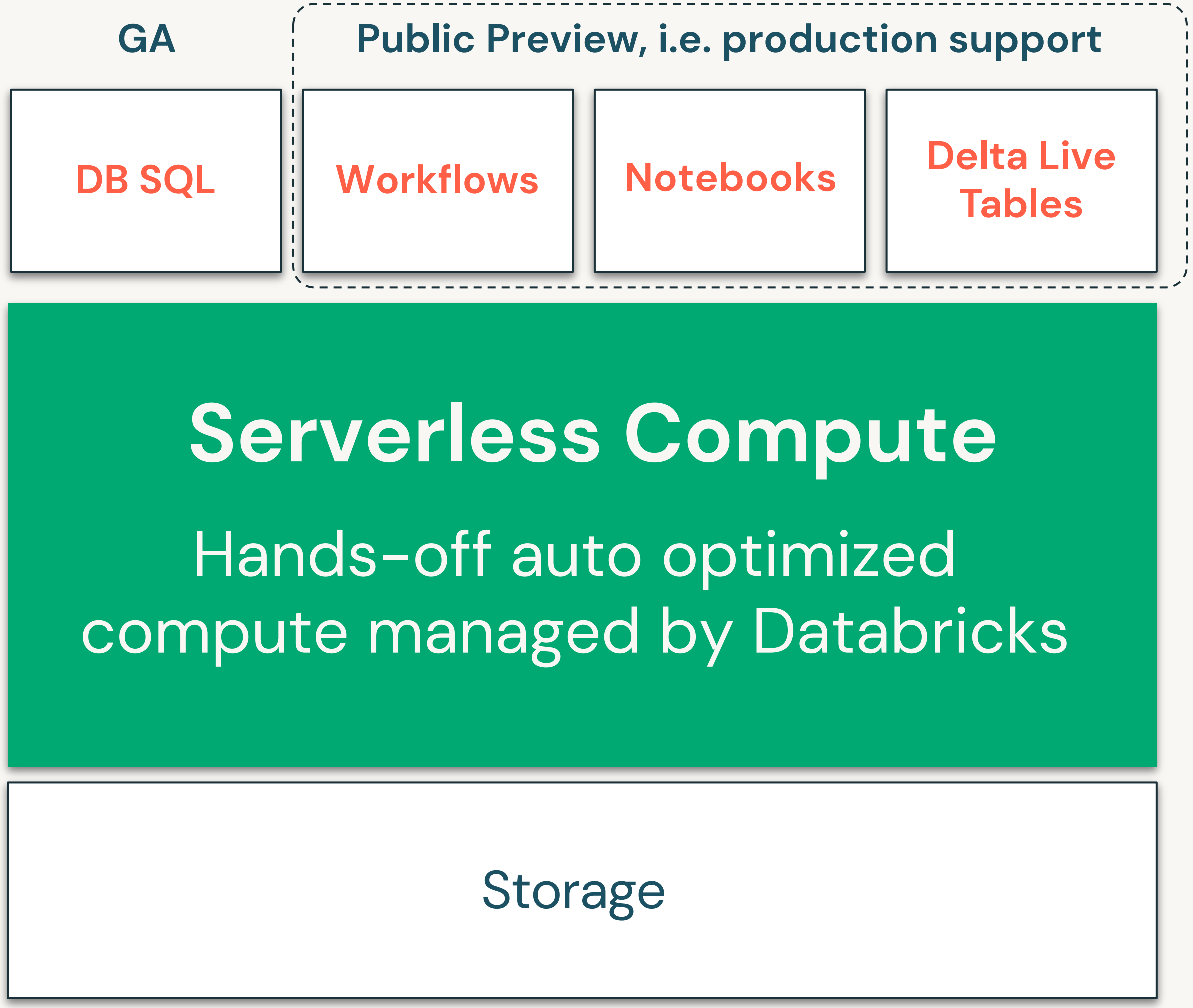
No knobs
Fast startup
For any practitioner

EFFICIENT

Fully managed and versionless
Paying only what you use
Strong cost governance

RELIABLE

Secure by default
Stable with smart fail-overs



GA

Public Preview, i.e. production support

DB SQL

Workflows

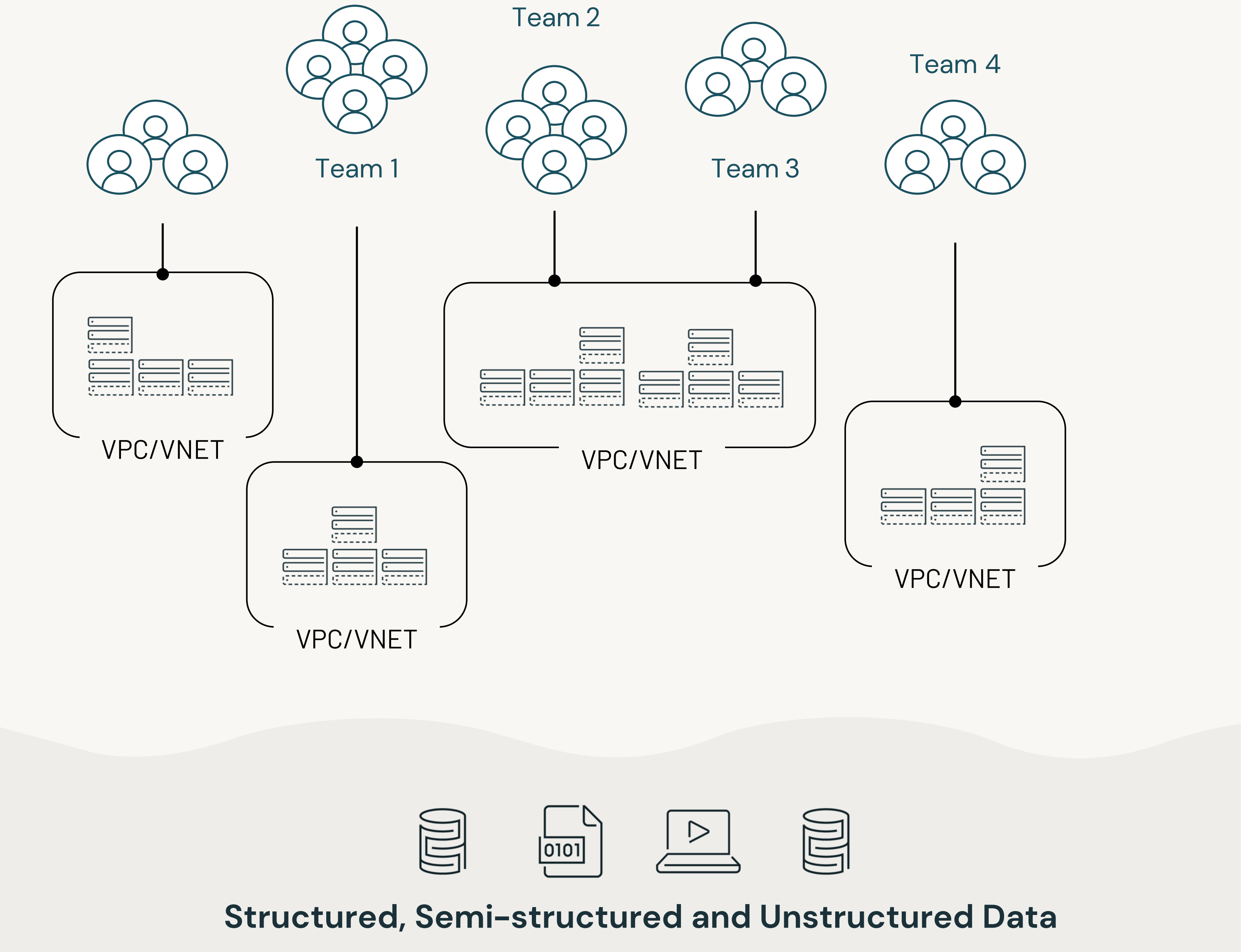
Notebooks

Delta Live Tables

Serverless Compute

Hands-off auto optimized compute managed by Databricks

Storage



STOP spending time on…

Setting up networks

Create and configure VNets
Set up gateways and firewall rules
Setup and manage private endpoints
X-tenant identities
IP address / subnet management

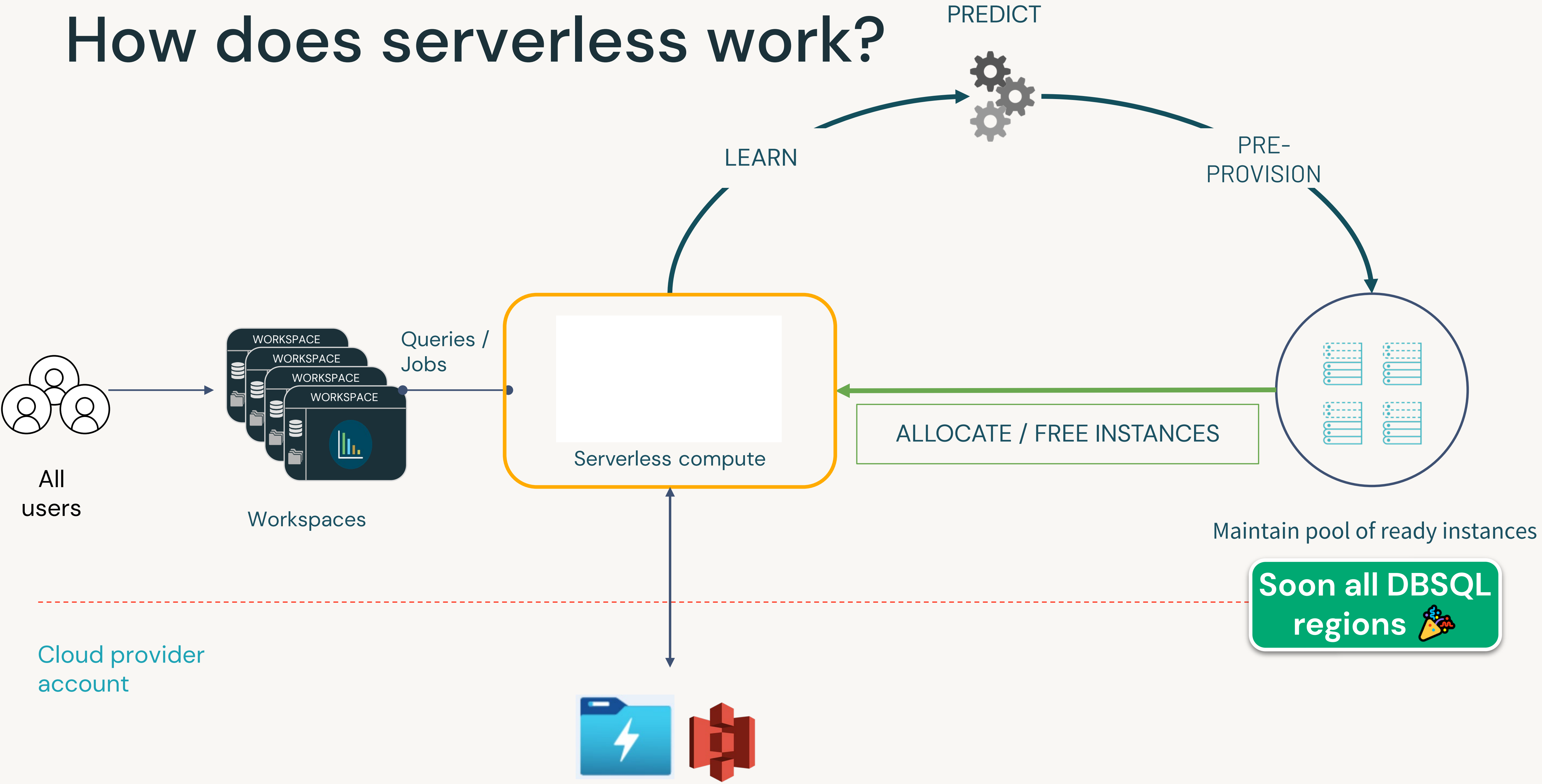
Security and Compliance

Vulnerability management
Encryption and key management
Intrusion detection and monitoring Data exfiltration protection

Managing efficiency

Capacity projections and reservations
Right sizing instances for workloads
Maintaining high utilization
Managing instance pools
Vacuum / compaction of Delta tables

How does serverless work?



Soon all DBSQL regions 🎉

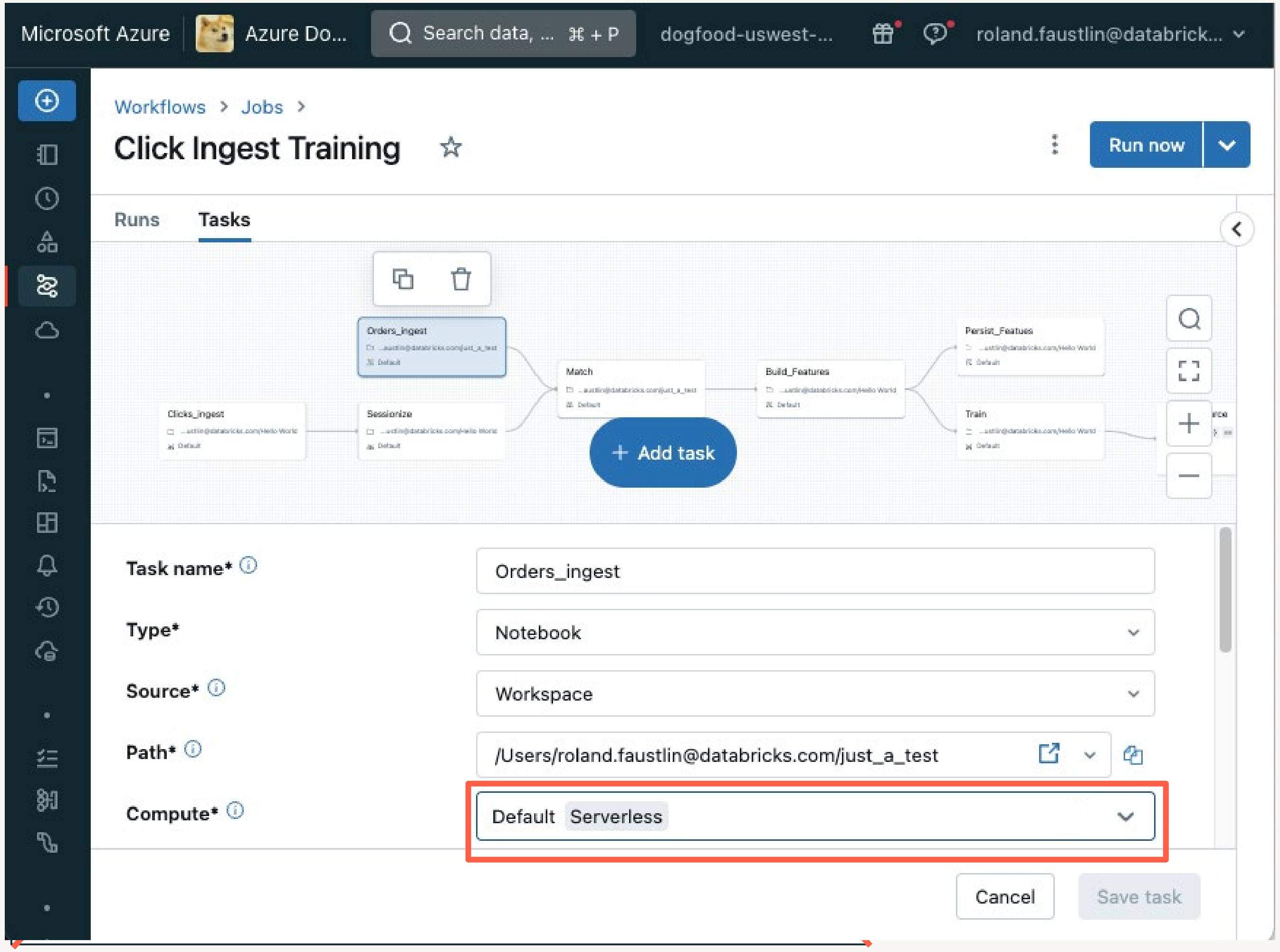
Key technologies in Serverless

- AI managed warm-pool of VMs enabling faster up- and down-scaling
- Enhanced horizontal autoscaler
- Versionless: automatically latest features (DBR, photon, etc.)
- First and only secure multi-user Spark w/ fully isolated user code
- Environment caching
- Automatic vertical scaling (soon)

Fully managed

Automatically improving

Serverless Compute in Workflows


Microsoft Azure
Azure Do...
Search data, ... ⌘ + P
dogfood-uswest-...
roland.faustlin@databrick...
Workflows > Jobs >
Run now
Runs
Tasks
Orders_ingest
...austin@databricks.com/just_a_test
Default
Clicks_ingest
...ustin@databricks.com/Hello World
Default
Sessionize
...ustin@databricks.com/Hello World
Default
Match
...austin@databricks.com/just_a_test
Default
Build_Features
...ustin@databricks.com/Hello World
Default
Persist_Featues
...ustin@databricks.com/Hello World
Default
Train
...ustin@databricks.com/Hello World
Default
+ Add task
Cancel
Save task

Click Ingest Training ☆

Task name*	Orders_ingest
Type*	Notebook
Source*	Workspace
Path*	/Users/roland.faustlin@databricks.com/just_a_test
Compute*	Default Serverless

Fully managed and reliable

- <60s startup
- Automatic failover
- Cost optimized + development mode (soon)

When to use Serverless

	Use cases
▶️ **Use now** 🥳	● Interactive Pythons or SQL (no Scala yet) ● New jobs ● Existing jobs compatible with Unity Catalog shared access mode ● Performance/startup time, streaming is important ● Instead of instance pools or all-purpose compute
⏭️ Later this year	● Cost-optimized mode ● Team level cost attribution ● Internet access controls ● GPUs

- Soon all serverless SQL regions!
- Built on UC Lakeguard

General availability of serverless compute for Notebooks, Workflows, DLT

SIMPLE and FAST

No knobs

Fast startup

For any practitioner

EFFICIENT

Fully managed and versionless

Paying only what you use

Strong cost governance

RELIABLE

Secure by default

Stable with smart fail-overs

Breakout session today at 4:00 PM

...rolling out next few weeks

...in all regions with serverless SQL

Demo

More to explore – Databricks Workflows

At DAIS

- **Practitioners Guide to Serverless Compute**
  Today, 4:00 PM PDT, South, Level 3, Rm 305
- **Ingestion Connectors**
  Thursday, 11:20 AM PDT, West, Level 2, Rm 2009
- **Workflows: Practical How-Tos and Demos**
  Thursday, 12:30 PM PDT, South, Level 3, Rm 307
- **Keynote on Data Engineering**
  Thursday, Jun 13, 8:30 AM PDT, South, Expo Level, Hall C

…and beyond

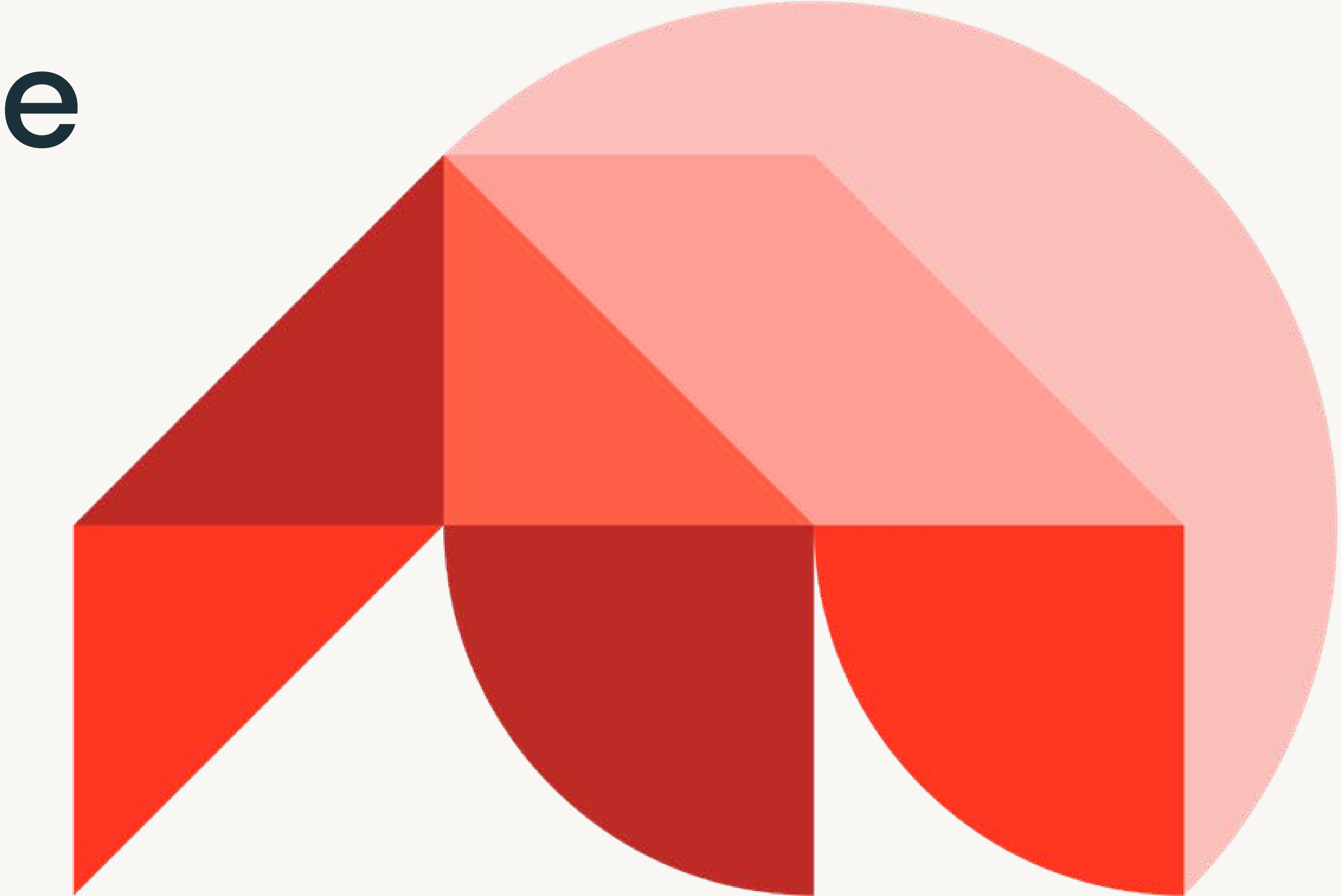
- databricks.com/product/workflows
- databricks.com/demos

🙏 Thank you

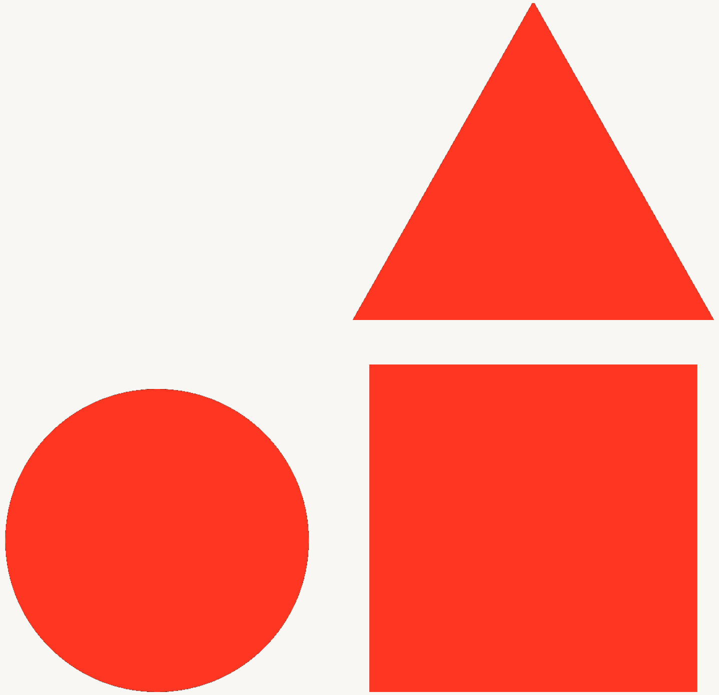
OUT

Your short but interesting slide title goes here

Author Name
Date



Your short but interesting slide title goes here



Author Name
Date

Resources

Table samples

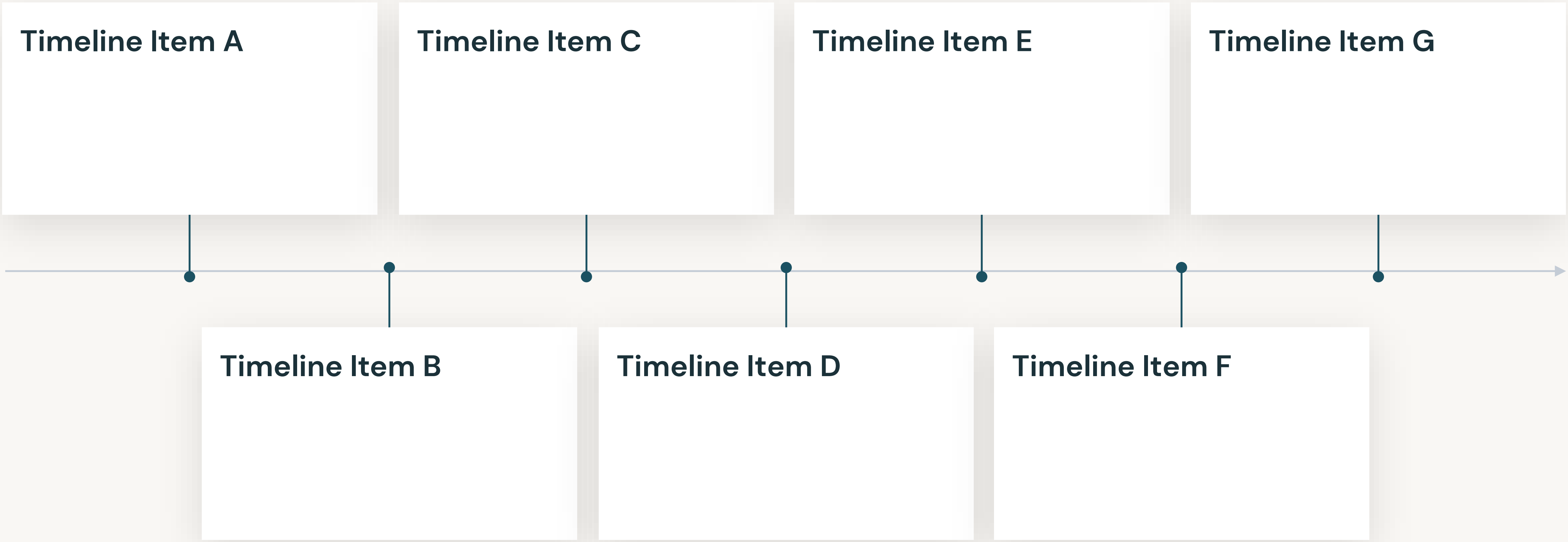
Your subtitle here

Table Header 1	Table Header 2	Table Header 3
Table Content	Table Content	Table Content
Table Content	Table Content	Table Content
Table Content	Table Content	Table Content

Table Header 1	Table Header 2	Table Header 3	Table Header 4	Table Header 5	Table Header 6
Table Content	Table Content	Table Content	Table Content	Table Content	Table Content
Table Content	Table Content	Table Content	Table Content	Table Content	Table Content
Table Content	Table Content	Table Content	Table Content	Table Content	Table Content

Timeline style

Your subtitle here



Timeline Item A

Timeline Item B

Timeline Item C

Timeline Item D

Timeline Item E

Timeline Item F

Timeline Item G

Primary icons

Examples

Included are a few various icons and illustrations. To access the full library of icons, please follow this link:

Click for primary icons

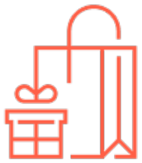
Life Sciences

Cloud Security

Analytics

Data Sharing

Collaboration



Retail

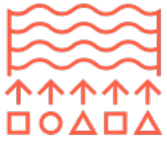
Multi-cloud

Gaming

Public Sector

Prediction

Data Science



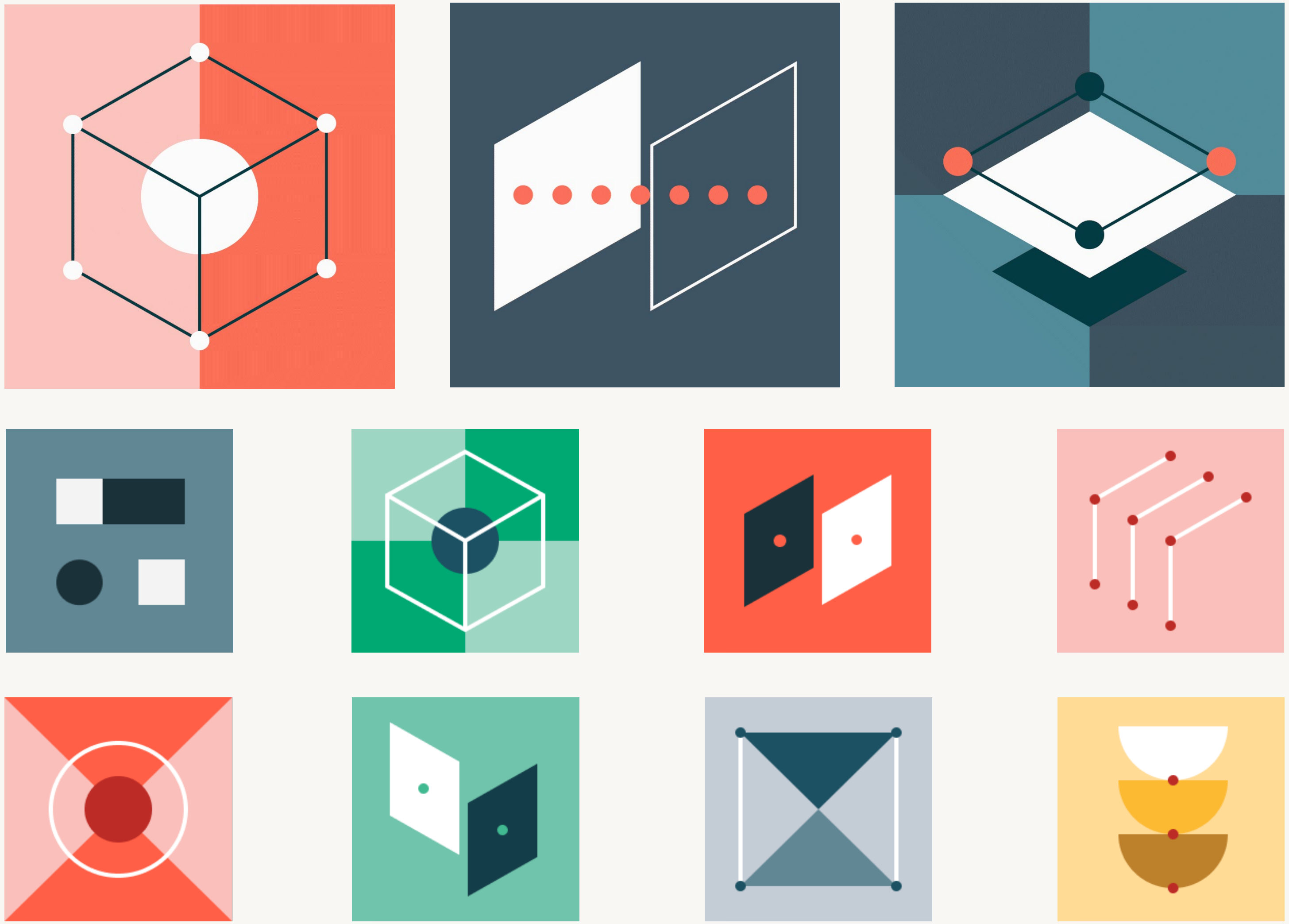
Data Lake

Secondary icons

Examples

Included are a few various icons and illustrations. To access the full library of icons, please follow this link:

Click for secondary icons



Illustrations

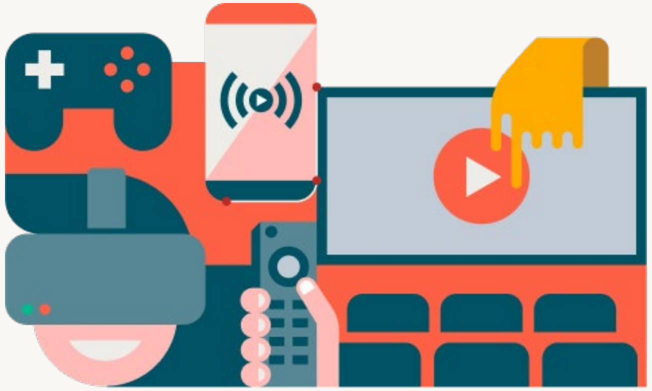
Examples

Included are a few various icons and illustrations. To access the full library of icons, please follow this link:

Click for illustrations

Manufacturing

Retail


Media & Entertainment

Healthcare and Life Sciences

Public Sector

Financial Services