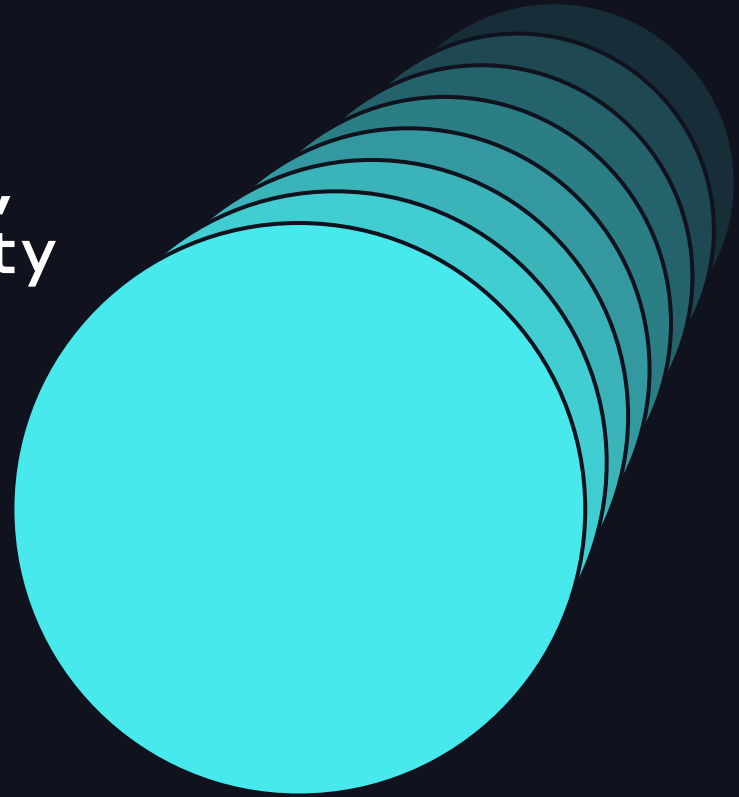


# Scaling DBSQL Performance, Observability, and Security



---

Jeremy Lewallen, PM - SQL  
Alex Esibov, PM - Security  
13 June 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**

# Who is this talk for?

If you have DBSQL implemented and you're considering adding new workloads, new users, or you want a better understanding of how Databricks largest customers scale

# Databricks largest customers focus on 3 areas



## Performance

Effectively manage your performance and cost objectives across several business units.



## Observability

Enable tens of thousands of users to adopt Databricks SQL and unlock value from your data.



## Security

Simply secure the SQL platform and make serverless your private network

# First topic, scaling workloads



## Performance

Effectively manage your large scale workloads with world class performance and TCO



## Observability

Enable tens of thousands of users to adopt Databricks SQL and unlock value from your data.



## Security

Simply secure the SQL platform and make serverless your private network

# Designed to be a warehouse that can handle all OLAP workloads

5 features (aka the AI Engine) that ensure optimal performance and best TCO for any workload you throw at it



# AI Engine—Learning from your data to improve performance

Collecting query stats and data stats throughout your data's lifetime

PREDICTIVE OPTIMIZATION

QUERY OPTIMIZER

DATA RETRIEVAL

PHOTON

INTELLIGENT WORKLOAD MGMT

- Detailed stats on your tables, partitions, and data layouts
- Optimizes file sizes
- Vacuums / deletes data that is no longer referenced

```
ALTER {SCHEMA | DATABASE} schema_name {ENABLE |  
DISABLE} PREDICTIVE OPTIMIZATION;
```

# AI Engine—Crafting efficient query plans

Query Optimizer creates highly efficient query plans with inputs from Predictive Optimization

PREDICTIVE OPTIMIZATION

QUERY OPTIMIZER

DATA RETRIEVAL

PHOTON

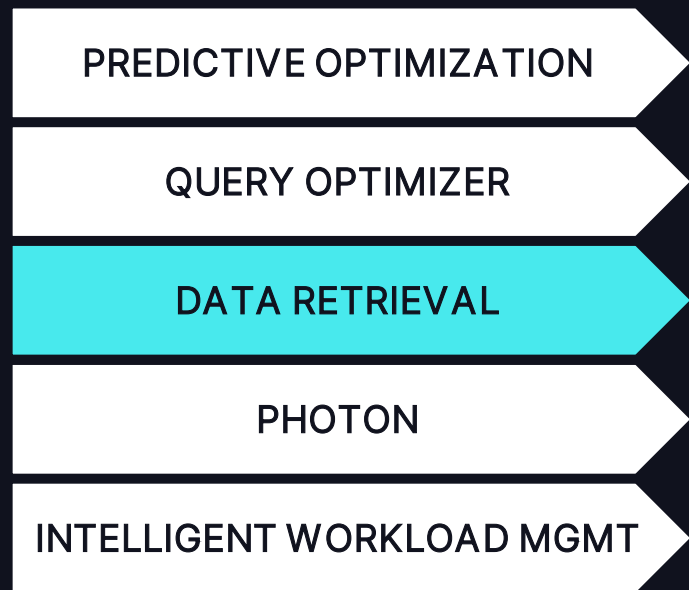
INTELLIGENT WORKLOAD MGMT

- SQL is declarative
- Inputs from PO are used to inform the Optimizer (*~18% better latency on benchmarks with PO data*)
- Generative, in nature, 1000s of plans created and the best one is chosen
- **No Action Needed!**



# AI Engine—State-of-the-art data retrieval with Liquid & Predictive IO

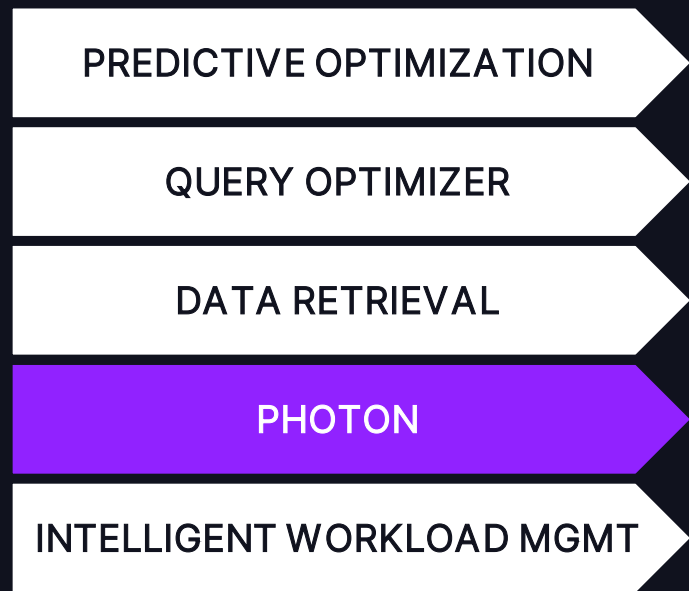
## High performance data retrieval



- Rapid pruning, filtering, and data retrieval from cloud storage
- Enhanced by Liquid Clustering
- Point lookups via ML predictions
- Optimized native DBSQL drivers
- **Tip:** Enable Liquid Clustering & Use Pro or Serverless

# AI Engine—Lightning-fast query processing

## Unmatched query speeds with Photon



- Purpose built Vectorized Execution Engine that can handle Spectrum of workloads
- Innovative Optimization Techniques complete with Adaptivity - GPU
- Reduces computational overhead and thus Total Cost of Ownership.
- No code change necessary

# AI Engine—Scheduling that optimizes for latency and cost

**DBSQL's IWM ensures high concurrency and efficient compute allocation**

PREDICTIVE OPTIMIZATION

- Ensures workloads are distributed across available compute to protect latency

QUERY OPTIMIZER

- Rapid allocation of compute when warehouse is full

DATA RETRIEVAL

- Rapid downscale when demand decreases, optimizing cost efficiency

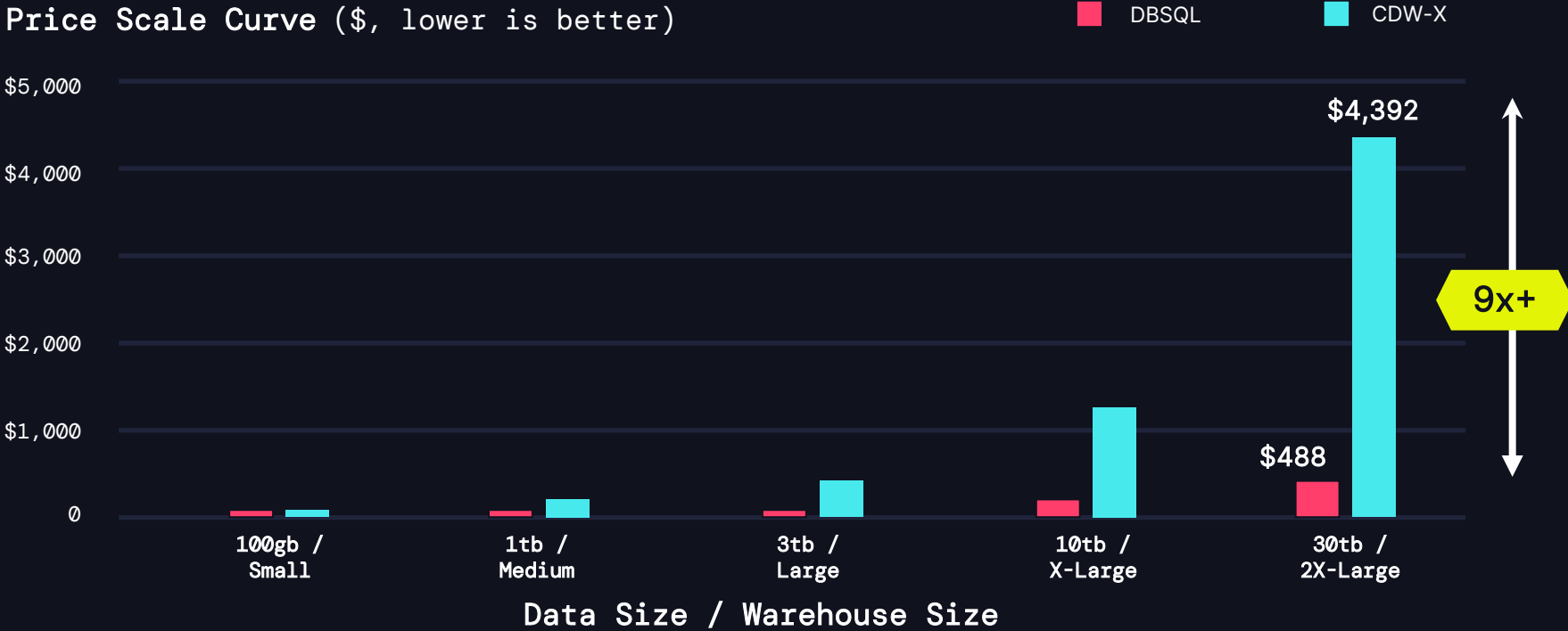
PHOTON

- Tip: Use Serverless with Autoscaling

INTELLIGENT WORKLOAD MGMT

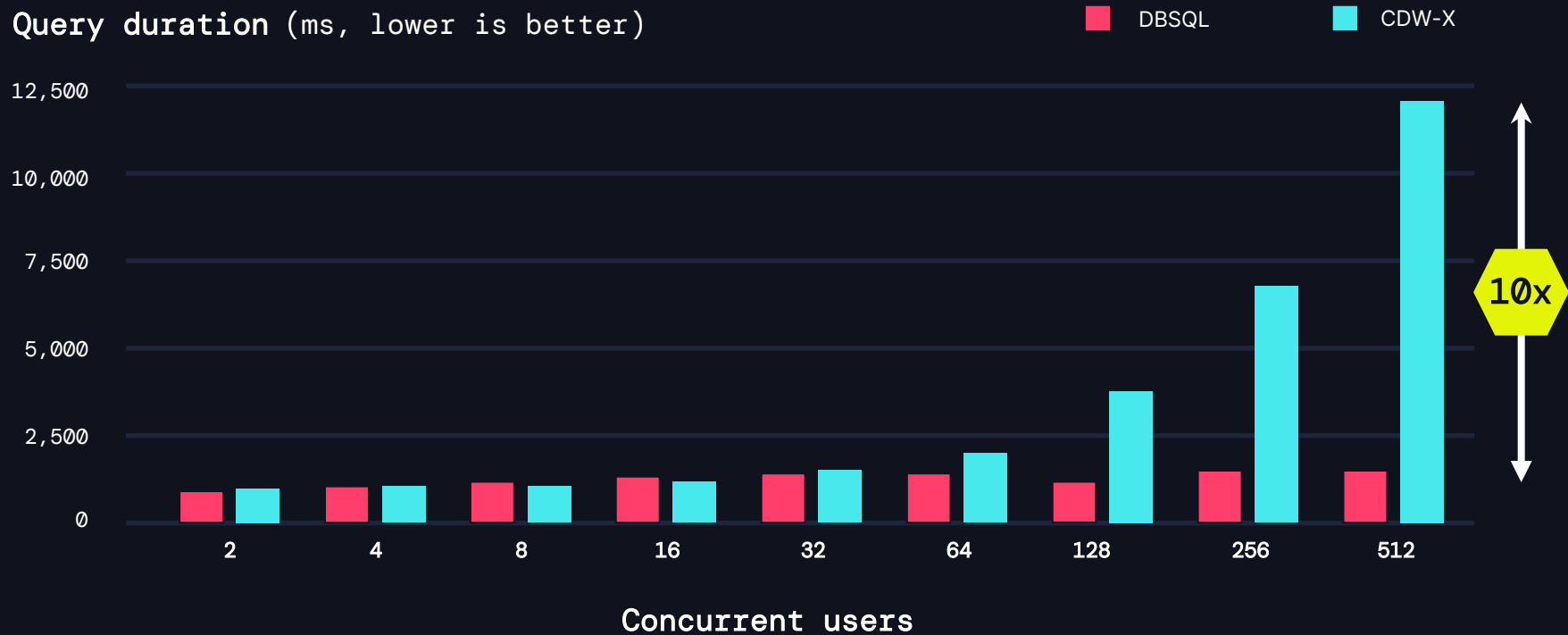
# Scalable ETL

Price Scale Curve (\$, lower is better)



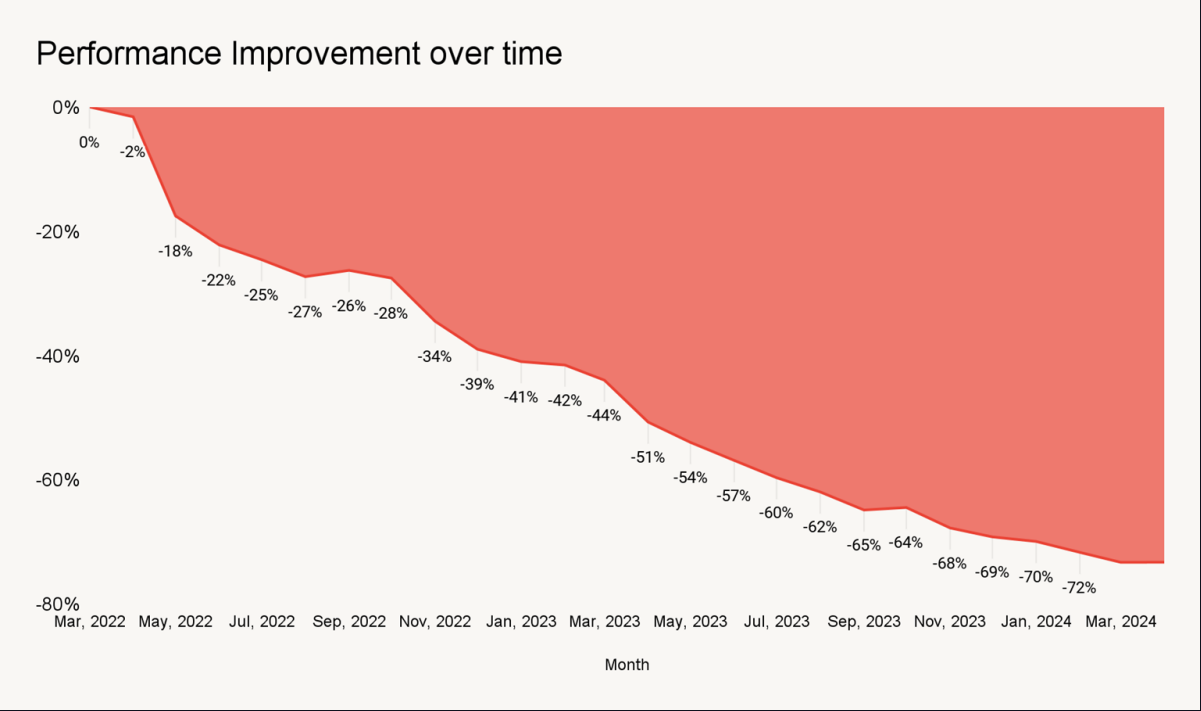
# Highly Concurrent BI Queries

Query duration (ms, lower is better)



# ACTUAL PRODUCTION WORKLOADS

# Actual customer queries improved 73% over last 2 years



**73%  
Faster!**



# What you need to scale DBSQL performance?

- 1 - Use Serverless with autoscaling enabled
- 2 - Enable Predictive Optimization and Liquid Clustering
- 3 - Nothing Else!!!



# OBSERVABILITY



## Performance

Effectively manage your performance and cost objectives across several business units.



## Observability

Enable tens of thousands of users to adopt Databricks SQL and unlock value from your data.



## Security

Simply secure the SQL platform and make serverless your private network.

# DBSQL now has 4 out-of-the-box system tables

Currently in Public Preview



**Billing**

Hourly cost data



**Warehouse events**

Warehouse start, stop, upscale, and downscale events



**Query history**

All queries executed on the warehouse



**Warehouses**

Warehouse change history, size information

# Cost controls #1: Proactive resource management

Set alerts for warehouses running longer than expected, and warehouses that are upscaled longer than usual

COST CONTROLS: COMPUTE


COST CONTROLS:  
INEFFICIENT QUERIES

USER EXPERIENCE:  
QUERY PATTERNS

USER EXPERIENCE: LATENCY

Best practices. Alerts on:

- Long running warehouses (20 hr)
- Warehouses running at an unusual time (weekends)
- Warehouses that have been upscaled longer than expected (max clusters > 2 hours)

	<sup>A</sup> <sub>C</sub> warehouse_id	 event_time	.00 running_hours	<sup>1</sup> <sub>3</sub> cluster_count
1	bc0c0000d1c725125	2024-06-12T01:05:40.025	19.77	1
2	9e2d1c0a0c0400000	2024-06-12T18:47:07.525	2.08	1
3	00000000000000000	2024-06-12T20:01:27.287	0.83	1
4	00000000000000000	2024-06-12T20:01:51.850	0.83	1
5	00000000000000000	2024-06-12T20:32:28.757	0.32	1
6	c21d0000e70000000	2024-06-12T20:12:52.322	0.65	1
7	00000000000000000	2024-06-12T20:27:45.699	0.4	1
8	00000000000000000	2024-06-12T18:33:43.718	2.3	1
9	00000000000000000	2024-06-12T18:31:34.746	2.33	1
10	00000000000000000	2024-06-12T20:24:47.348	0.45	1
11	00000000000000000	2024-06-12T19:52:00.599	1	1

# Cost controls #2: Find and alert on poor performing queries

## Deep dive into query performance

COST CONTROLS: COMPUTE

COST CONTROLS:  
INEFFICIENT QUERIES

USER EXPERIENCE:  
QUERY PATTERNS

USER EXPERIENCE: LATENCY

- Alerts for queries causing significant disk spill
- Monitor queries with excessive shuffle operations
- Top n queries in execution time

```
select
  warehouse_id,
  statement_id,
  sum(spilled_local_bytes) as spilled_local_bytes
from system.query.history
where start_time >= date_sub(CURRENT_DATE, 30)
group by all
having spilled_local_bytes > 0
```

Raw results ▾ +

	$A^B_C$ warehouse_id	$A^B_C$ statement_id	$1^2_3$ spilled_local_bytes
1	██	01ef1cd3-a188-146e-930e-61626e2f59b9	10650694110237
2	██	01ef1d0d-7f42-1b6b-a441-712868fb9dd2	171553649695
3	██	01ef1d10-afc6-1925-a0dc-ed27912ab0c8	167574471553
4	██	01ef21b3-b44e-1c52-a108-ef0f636e188b	128213400767
5	██	01ef22c7-1021-166e-ab71-ff87893bf569	127393026612
6	██	01ef2348-8bf0-1ac0-9c6f-94aea23c241e	125153787980
7	██	01ef1f4d-df43-14dd-88e9-c0ee661c9dda	120445305569
8	██	01ef1f27-d5ce-16f6-8f9b-2dd06921ad33	119141006499
9	██	01ef19c5-2199-14cb-8f34-53bb0e1df6e1	115325434498
10	██	01ef16f0-6e98-158b-a354-7e6a214ead7f	115199659378
11	██	01ef16b2-3628-19f3-820b-03cc8a9699b9	115118675168
12	██	01ef1771-f3e3-183b-a61b-689a3c5d5b5b	113359311979

# User experience #1: Identify changing query patterns

Identify key usage patterns to address performance issues

COST CONTROLS: COMPUTE



COST CONTROLS:  
INEFFICIENT QUERIES

USER EXPERIENCE:  
QUERY PATTERNS

USER EXPERIENCE: LATENCY

- Increases in warehouse usage
- Changes in total throughput
- Changes in queueing

```
SELECT
  date(start_time) as date,
  count(*) as daily_queries
FROM
  `system`.`query`.`history`
where
  start_time >= date_sub(now(), 60) and warehouse_id='x'
GROUP BY
  date(start_time)
```

	 start_time 	$1^2_3$ daily_queries
1	2024-06-12	41008
2	2024-06-11	51361
3	2024-06-10	49845
4	2024-06-09	43837
5	2024-06-08	45145
6	2024-06-07	51083
7	2024-06-06	51102
8	2024-06-05	51432
9	2024-06-04	53860
10	2024-06-03	49708
11	2024-06-02	40010



# User experience #2: Easily understand end user experience

Identify and alert on changes that impact users

COST CONTROLS: COMPUTE

- Monitor warehouse latency, including median and extreme values like p99, to catch performance bottlenecks

COST CONTROLS:  
INEFFICIENT QUERIES

- Track execution time of critical queries to identify slow operations

USER EXPERIENCE:  
QUERY PATTERNS

- Monitor overall system latency to ensure responsive performance for end users

USER EXPERIENCE: LATENCY

	📅 date ⚙️	A <sup>B</sup> C warehouse_id	1.2 p50_latency	1.2 p75_latency	1.2 p95_latency	1.2 p99_latency
1	2024-06-12	2024061210011718116	0.2	0.24	0.34	5.68
2	2024-06-12	1190181117511611	0.18	0.38	0.48	1.05
3	2024-06-12	511604510111541111	0.39	0.56	1.48	15.42
4	2024-06-12	1111111111111111111	233.17	253.53	365.67	369.49
5	2024-06-12	1111111111111111111	0.33	0.44	2.82	209.06
6	2024-06-12	1111111111111111111	4.17	6.48	11.53	19.27
7	2024-06-12	1111111111111111111	262.23	262.23	262.23	262.23
8	2024-06-12	1111111111111111111	13.46	14.17	15.23	15.44
9	2024-06-12	1111111111111111111	11.76	14.31	608.67	727.54
10	2024-06-12	1111111111111111111	19.61	19.61	19.61	19.61
11	2024-06-12	1111111111111111111	5.62	7.12	10.42	11.14

# RESOURCES



HELP DOCUMENT



PREWRITTEN  
QUERIES

# SECURITY



## Performance

Effectively manage your performance and cost objectives across several business units.



## Observability

Enable tens of thousands of users to adopt Databricks SQL and unlock value from your data.



## Security

Simply secure the SQL platform and make serverless your private network

# Best Practices for Scaling Security



**Identity**



**Governance**



**Networking**



**Compliance**

**Ease of Use**

# Identity & Governance

# Best practices for identity & governance

## Enable SSO

Authenticate via single sign-on (SSO) at the account level.

Use SSO via OAuth federation to your favorite BI tools.

## Sync Identities

Sync users and groups from your identity provider, once.

## Manage permissions

Use groups to manage permissions to catalogs

## Secure data access

Filter sensitive table data with row-level security and column-level security, including masking



# Provision all your users and groups

SSO for your users, with multi-factor authentication (MFA)

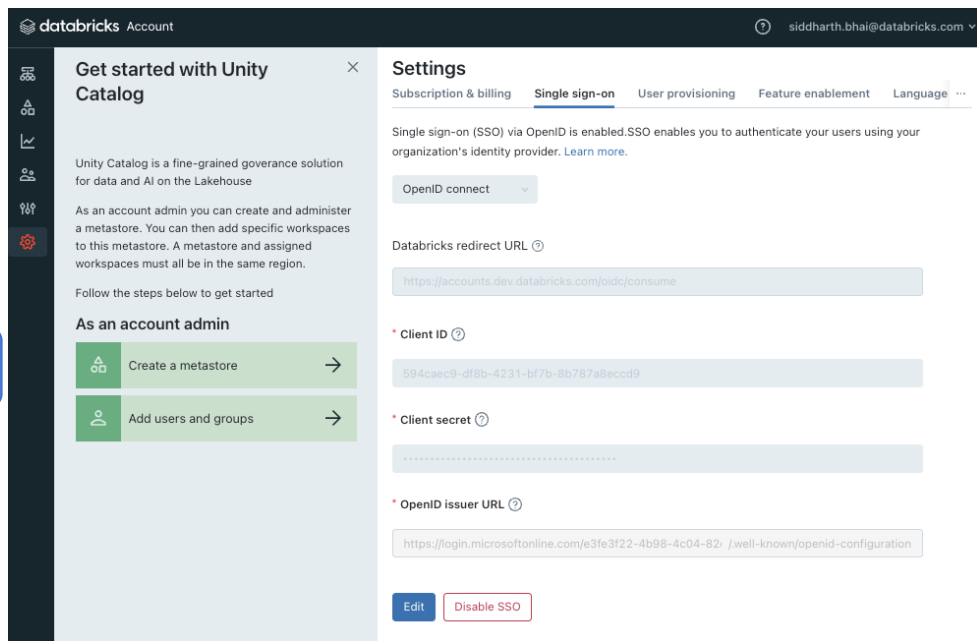
Single Sign on is provisioned **once** with your Identity Provider



automatically applies to all **new workspaces created** in these accounts



## Account Console



The screenshot shows the Databricks Account Console interface. The top navigation bar includes the Databricks logo, the word "Account", and the user's email address "siddharth.bhai@databricks.com". The main content area is split into two panels. The left panel, titled "Get started with Unity Catalog", provides introductory text and two primary actions: "Create a metastore" and "Add users and groups". The right panel, titled "Settings", has tabs for "Subscription & billing", "Single sign-on", "User provisioning", "Feature enablement", and "Language". The "Single sign-on" tab is active, showing that SSO via OpenID is enabled. It includes a dropdown menu for "OpenID connect", a text field for "Databricks redirect URL" (containing "https://accounts.dev.databricks.com/oidc/consume"), and fields for "Client ID" (594caec9-df8b-4231-bf7b-8b787a8eccd9) and "Client secret". There is also a field for "OpenID issuer URL" (https://login.microsoftonline.com/e3fe3f22-4b98-4c04-821.../well-known/openid-configuration). At the bottom of the settings panel are "Edit" and "Disable SSO" buttons.





# Users can use their favorite BI tools

## Enable SSO with OAuth

Query the freshest data in SQL, and build **apps** and **dashboards** with **any tools** powered by the lakehouse



 **databricks**



Sign in to continue to Databricks

noe.derakhshani@databricks.com

.....

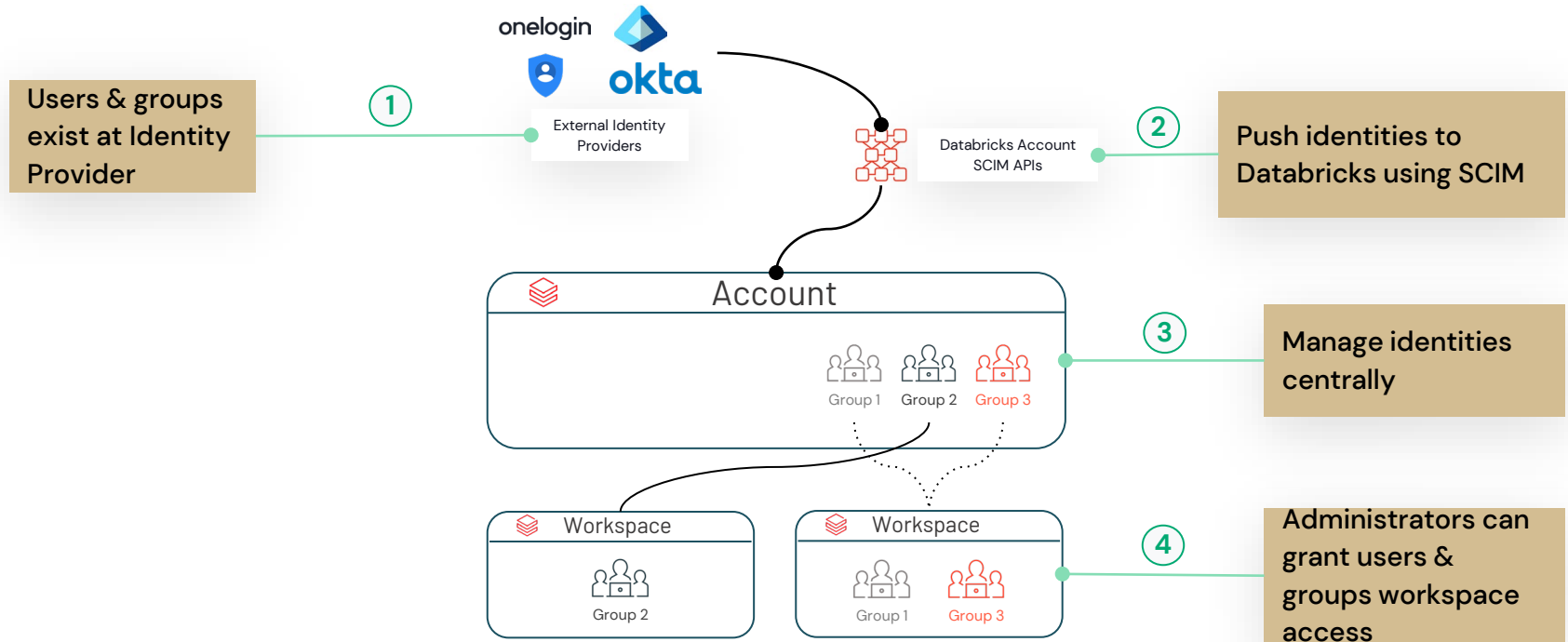
Sign In

The screenshot shows the Tableau 'Connect' dialog box with 'Databricks' selected. The 'General' tab is active, showing the connection URL 'd.databricks.com' and a dropdown menu. A red box highlights the dropdown menu. Below the dropdown, the text 's://[Server Hostname]/oidc' and 'ng.cloud.databricks.com/oidc' is visible. A 'Sign In' button is at the bottom right. The background shows the Tableau 'Open' dialog and a list of 'Accelerators'.



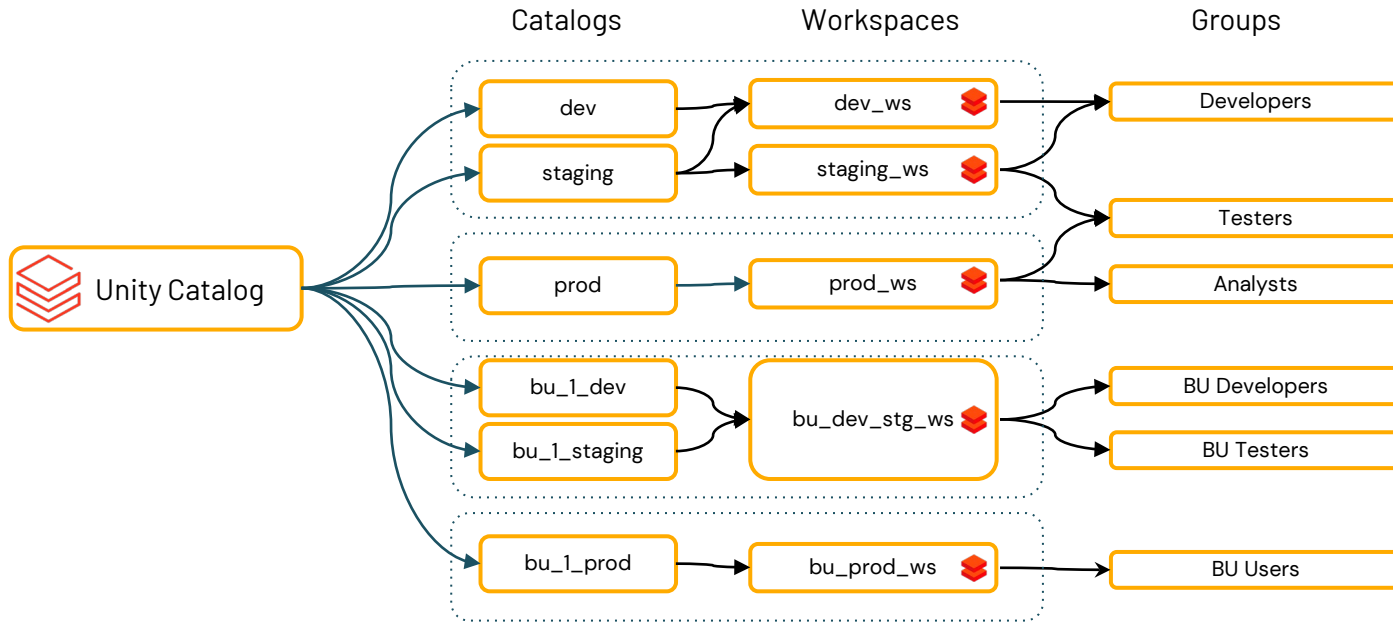
# Synchronize identities to the account

Simplify sync pipelines to one per account



# Restrict data access to authorized environments

Access to Catalogs can be isolated to specific Workspaces & Groups



# Row Level Security and Column Level Masking

Provide differential fine grained access to datasets

Coming soon: Use attribute-based access control (ABAC) to scale application of masks

## Only show specific rows

```
CREATE FUNCTION <name> ( <parameter_name >  
<parameter_type> .. )  
RETURN {filter clause whose output must be a boolean}
```

```
CREATE FUNCTION us_filter(region STRING)  
RETURN IF(IS_MEMBER('admin'), true, region="US");
```

```
ALTER TABLE sales SET ROW FILTER us_filter ON region;
```

Test for group membership

Assign reusable filter to table

Specify filter predicates

## Mask or redact sensitive columns

```
CREATE FUNCTION <name> (<parameter_name>,  
<parameter_type>, [, <column>...])  
RETURN {expression with the same type as the first  
parameter}
```

```
CREATE FUNCTION ssn_mask(ssn STRING)  
RETURN IF(IS_MEMBER('admin'), ssn, "*****");
```

```
ALTER TABLE users ALTER COLUMN table_ssn SET MASK  
ssn_mask;
```

Test for group membership

Assign reusable mask to column

Specify mask or function to mask



# Networking

# Classic Architecture

Customer Cloud providers

Managed by Databricks

Clients

Your assets



Cloud



Internet



SaaS



On-prem

Classic (non-serverless) compute

Control Plane

Databricks UI

Unity Catalog

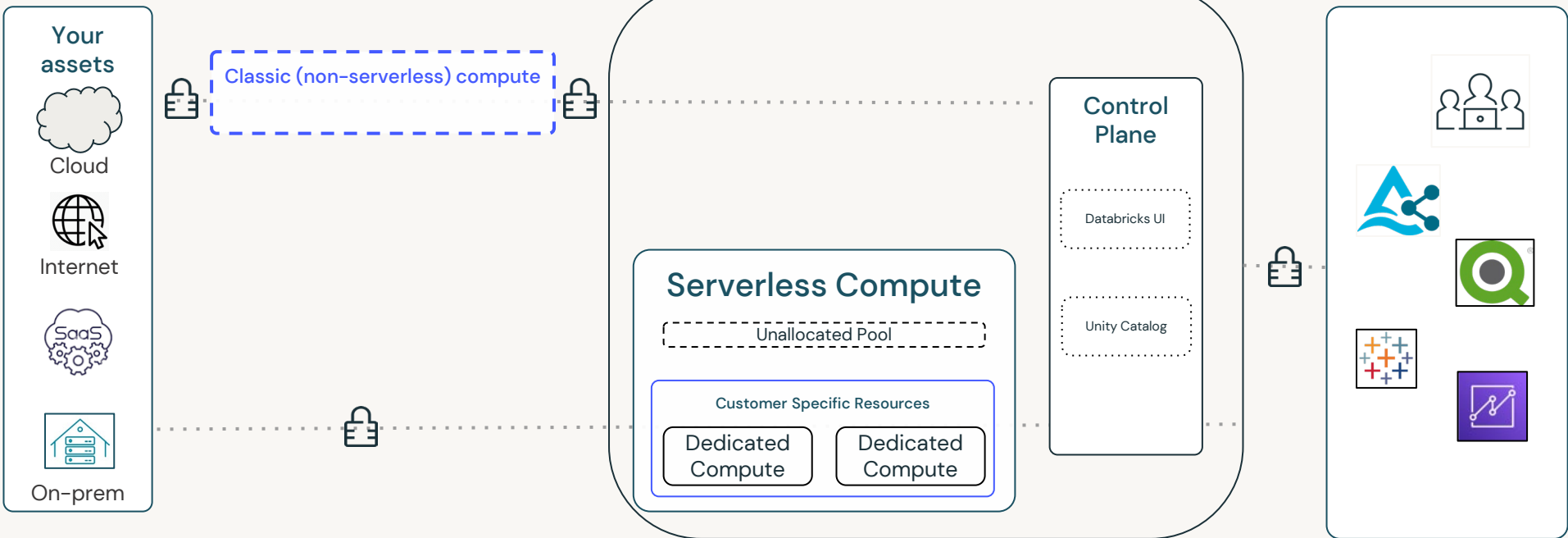


# Serverless Architecture

Customer Cloud providers

Managed by Databricks

Clients

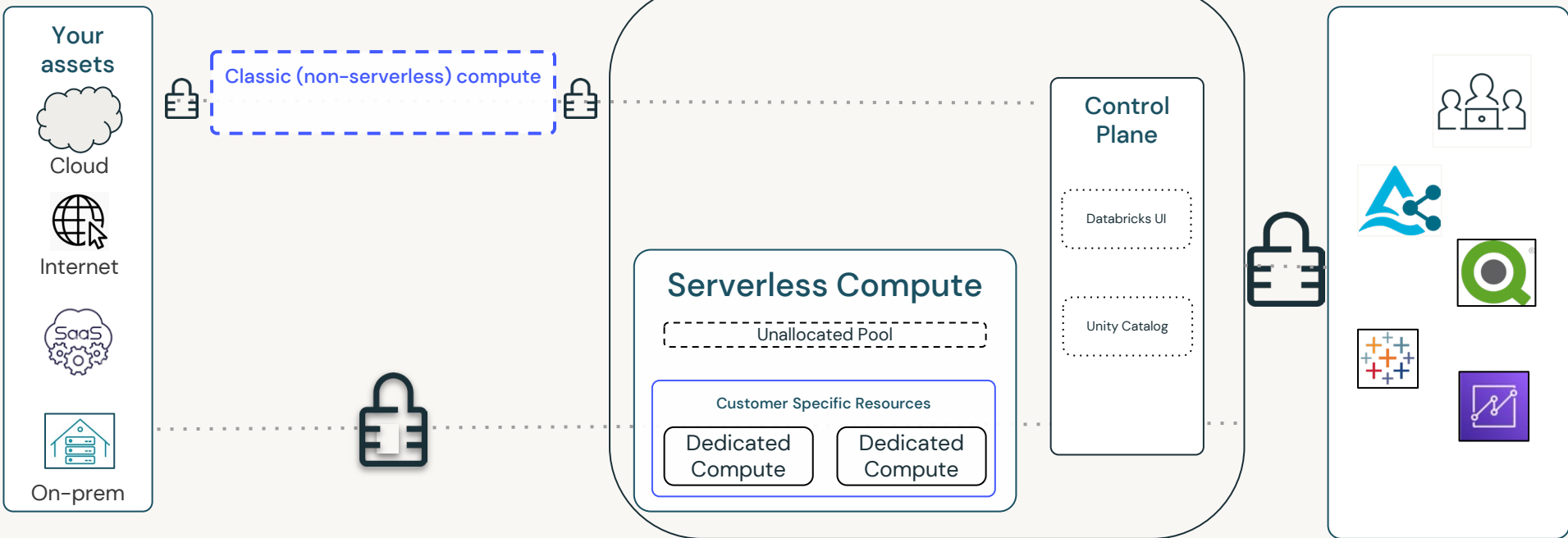


# Serverless Architecture

Customer Cloud providers

Managed by Databricks

Clients

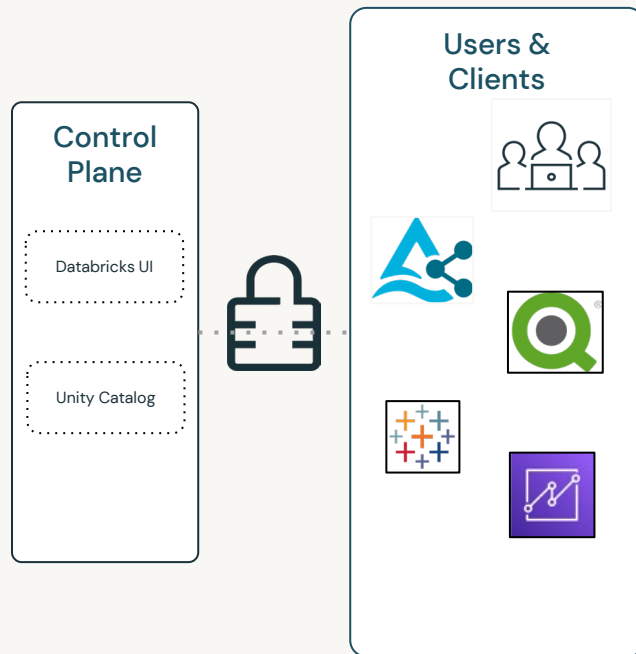




# Users & Clients to Databricks

## Defense-in-depth security to prevent unauthorized access to a Databricks Workspace

1. *Mitigate user credential compromise risks*  
Already addressed **SSO** with **multi-factor authentication** (MFA) with unified login
1. *Mitigate token replay risks*  
Configure **private link** and / or **IP ACLs** to prevent access from unauthorized networks



**When it comes to connecting to your  
resources from serverless...**

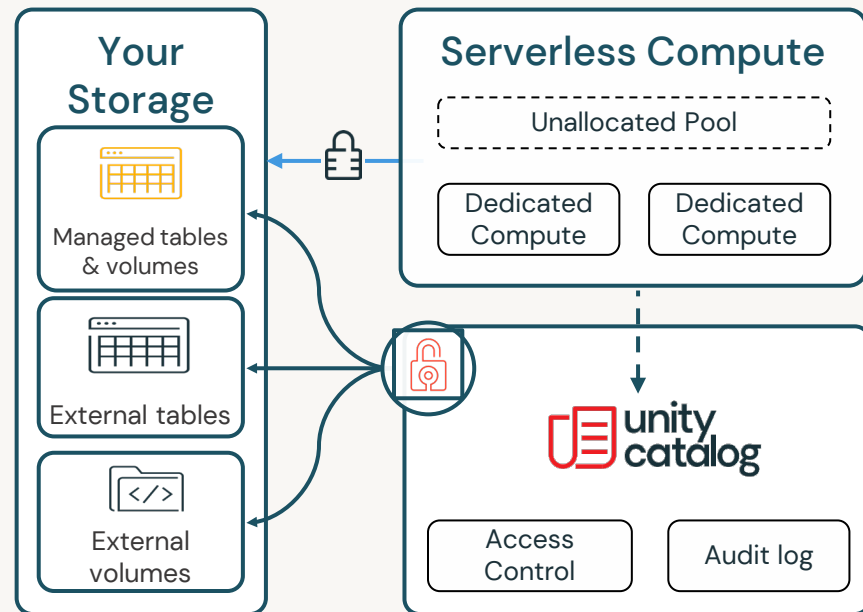


# Securing access to Cloud Storage

## Unity Catalog network security

### Network security

1. Traffic over cloud provider's backbone, TLS 1.2+ encrypted
1. [Azure] Dedicated **Private Link** for network layer defense in depth, with free data processing!



# Securing egress to the internet

- Serverless DBSQL does not have access to the internet by default
- **[New!] Serverless Egress Controls** enable you to control access to the internet from user code (e.g., network UDFs, model serving, DLT, notebooks, etc.)
- Access via Unity Catalog governed paths is always allowed

The screenshot shows the Databricks interface for configuring network policies. The breadcrumb trail is 'Cloud resources > Network > Network policy > AMEA\_data\_team\_network\_policy'. A notification box states: 'No restrictions on the destinations accessible from your serverless environment. [details](#)'. Below this, there are two radio buttons: 'Full access' (unselected) and 'Restricted access' (selected). The 'Restricted Access Rules' section contains two tables. The first table, 'Allowed Internet Destinations', lists rules for \*.salesforce.com (FQDN), 12.46.78.0 (IP Address), and 203.30.18.128/25 (IP Range). The second table, 'Allowed S3 Storage Destinations', lists rules for s3://YourBucket (us-west-2) and s3://YourBucket2 (us-east-1). At the bottom, a 'Workspace Selector' table lists several workspaces, with 'Dev-ws-westus2' and another workspace selected.

Allowed Internet Destinations	Type	
*.salesforce.com	FQDN	🗑️
12.46.78.0	IP Address	🗑️
203.30.18.128/25	IP Range	🗑️
+ < Previous Next >		

Allowed S3 Storage Destinations	Region	
s3://YourBucket	us-west-2	🗑️
s3://YourBucket2	us-east-1	🗑️
+ < Previous Next >		

Workspace	Created
<input checked="" type="checkbox"/> Dev-ws-westus2	06/01/2022
<input type="checkbox"/> staging-ws-data-eastus	
<input type="checkbox"/> prod-ws-data-retaildept	
<input checked="" type="checkbox"/> [Redacted]	
<input type="checkbox"/> [Redacted]	

# Compliance

# Data security and compliance is essential...

## Health Care

### HIPAA/HITRUST

All companies that accept, process, store or transmit health care data (PHI)

- Insurance companies
- Hospitals
- Health and Life Science
- Self-insured employers

## Government

### FedRAMP / DoD IL

#### US Federal Agencies

Companies & contractors working with US Federal Government

## Financial

### PCI-DSS

All companies that accept, process, store or transmit credit card information

- Payment Gateways
- Retail Investment Platforms
- Card Issuing Institutions



# ...but compliance can be quite overwhelming

- What controls do I need when **processing regulated data**?
  - Data in transit/at rest encryption, FIPS 140-2
  - AV, file integrity monitoring, vulnerability scanning
  - OS hardening
  - Patching & Update
- How do I correctly **implement those controls**?
- How do I ensure my environment **remains correctly configured**?



# How Databricks helps with compliance

DBSQL is available for all major certifications and on AWS GovCloud!

## AWS

- **Commercial:** HIPAA, IRAP, PCI-DSS, FedRAMP Moderate
- **GovCloud:** ITAR, FedRAMP High, DoD IL5 (coming soon)

## Azure

- **Commercial:** HIPAA, PCI-DSS (coming soon)

## GCP

- **Commercial:** HIPAA

All applicable controls for all available standards in 4 clicks

The screenshot shows the 'Security and compliance' configuration page in Databricks. It features three tabs: 'Configuration', 'Permissions', and 'Security and compliance'. The 'Security and compliance' tab is active. The page is divided into three sections, each with a 'Configure' button and a status indicator.

- Compliance security profile:** Status is 'Enabled'. Description: 'Applies a highly secure baseline to the workspace, making it easier to meet and manage applicable compliance control requirements. [Learn more](#)'. Compliance standards selected: PCI-DSS.
- Enhanced security monitoring:** Status is 'Enabled' (indicated by a toggle switch). Description: 'Provides security monitoring capabilities to the workspace. If the compliance security profile is enabled, this setting is permanently enforced. [Learn more](#)'.
- Automatic cluster update:** Status is 'Enabled' (indicated by a toggle switch). Description: 'Monthly on the 1st [Sunday at 1:00 AM UTC](#)'.





# Security – Call to Action

**Interested in joining our previews?  
Interested in joining a customer  
focus group on security and  
compliance?**

Fill out this 10 second survey - QR code or:  
<https://tinyurl.com/DAIS2024Sec>



# Conclusions

# How can you scale with Databricks?

You learned how DBSQL warehouses ensure:



## Performance

Effectively manage your performance and cost objectives across several business units.

Enable predictive optimization

Enable liquid clustering

Use auto-scaling serverless warehouses



## Observability

Enable tens of thousands of users to adopt Databricks SQL and unlock value from your data.

Use DBSQL system tables for monitoring and alerting



## Security

Simply secure the SQL platform and make serverless your private network

Sync identities, provision users and BI tools

Use Unity Catalog to secure access to your data

Enable private connections, and disallow unauthorized resources



Thank you!  
Questions?

---

# Learn more at the summit!



Databricks  
Events App



## Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.



## What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!



## Get trained and certified

- Visit the Learning Hub Experience at **Moscone West, 2nd Floor!**
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! [databricks.com/learn](https://databricks.com/learn)



# APPENDIX

