DATA+AI SUMMIT BY databricks

Healthcare Data with Unity Catalog: Providence's Journey

Anna Erickson, Satish Marripelli, Janet Vickers, Lawrence Yapp

©2024 Databricks Inc. — All rights reserved

Providence St Joseph Health

122K CAREGIVERS

38K NURSES

34K PHYSICIANS

$2.1B COMMUNITY BENEFIT

51 HOSPITALS

1000 CLINICS

29M TOTAL PATIENT VISITS

2.6M COVERED LIVES

1700+ PUBLIC RESEARCH STUDIES

1 HEALTH PLAN

18 SUPPORTIVE HOUSING FACILITIES

HIGH SCHOOL NURSING SCHOOLS & UNIVERSITIES

178 PARTNER SITES

87 COMMUNITY CONNECT PARTNERS

Providence

Our Problem - The Wild West

- 131 unique workspaces across 10 subscriptions
- No visibility into cost and usage
- Experimentation with no governance, no cluster policies
- Duplicate data as each workspace had its own source data
- Opportunities for optimization
- Opportunities for tighter collaboration across teams

©2024 Databricks Inc. — All rights reserved

The Team

We are a cross functional group spanning multiple IS teams.

- Healthcare Intelligence
- Cloud Infrastructure
- Cyber Security
- Networking & Firewall
- And Databricks!

©2024 Databricks Inc. — All rights reserved

Onsite Sessions

- We prepped and planned for multiple in-person sessions that lasted a week each time
- Focus was on collaboration and achieving a specific set of deliverables
- Because we were in one room, we could quickly clear barriers
- “Let the architects, architect”

©2024 Databricks Inc. — All rights reserved

Team Building, Even After Hours!

While we were onsite, we continued to collaborate and discussed project plans over dinner.

- It may seem insignificant but social gatherings were needed for team forming and building trust and rapport.

©2024 Databricks Inc. — All rights reserved

Setting up Infrastructure

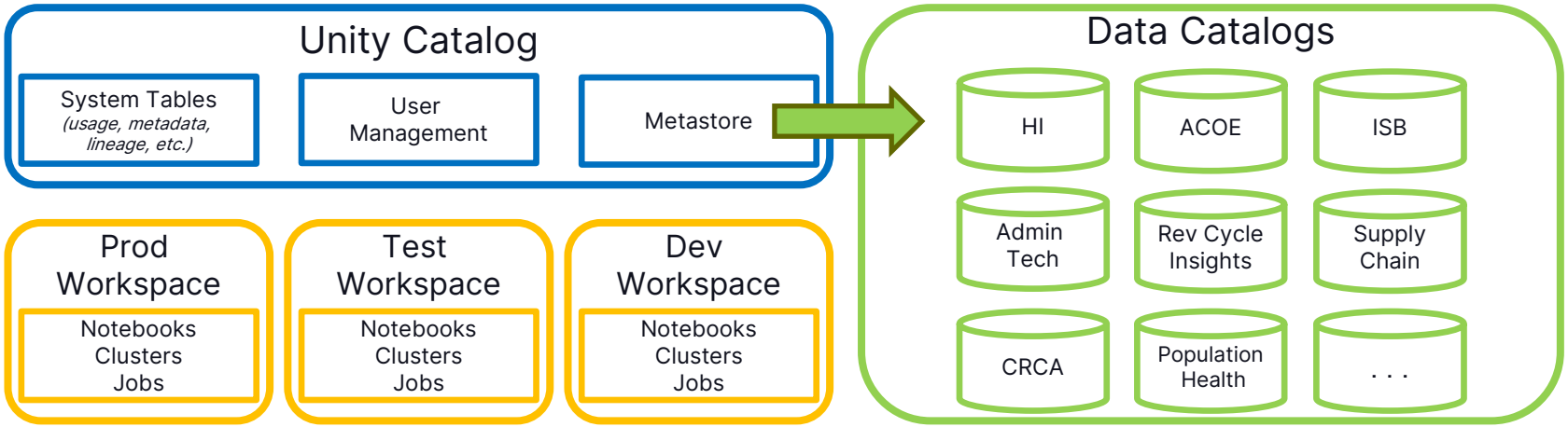
- Networking and firewall configuration
- Involved Cyber Security throughout entire process
- Infrastructure clean up
  - Source data stored in ADLS instead of DBFS, Code Repo and Monitoring
  - Workspace by workspace information gathering
- Deployed single workspace for dev, test, and production in the central subscription

©2024 Databricks Inc. — All rights reserved

Databricks Unity Catalog Overview

Unity Catalog enabled in a centrally managed subscriptions for the enterprise. This makes it easier to govern, manage users, and permissions.

- Data Catalogs structured similarly to Snowflake Databases
- Metadata layer that interacts with data – schemas, tables, views, etc.
- Data resides in ADLSGen2



Consult with Data Architect to avoid data swamp

©2024 Databricks Inc. — All rights reserved

Unity Catalog - Benefits and Impact

Why are we consolidating, cleaning up, and migrating to Unity Catalog?

- Reduces the number of workspaces from 131 to 3 (one per environment)
- Centralizing makes it easier to manage and remove duplicate data
- Remove low usage & low value workspaces and clusters
- Easier to secure, govern, and apply policies and manage users/permissions
- Easier to enable CICD and Infrastructure as Code (IaC)
- Easier to monitor and to enable lifecycle management (remove unused clusters, etc.)
- Easier to enable tagging for cost and usage reporting
- Easier to enable new AI features

©2024 Databricks Inc. — All rights reserved

Migration is not a lift and shift

HI Migration Counts	Workspace
Starting Count	76
Completed/Removed	48
Percentage Complete	63%

Additional Stats	Workspace
Current Count	33
Consoldated	43
Removed	5
Not Started	1
In Process	32
Remaining	28

NON-HI Migration Counts	Workspace
Starting Count	55
Completed/Removed	33
Percentage Complete	60%

Additional Stats	Workspace
Current Count	39
Consoldated	12
Removed	21
Not Started	15
In Process	3
Remaining	17

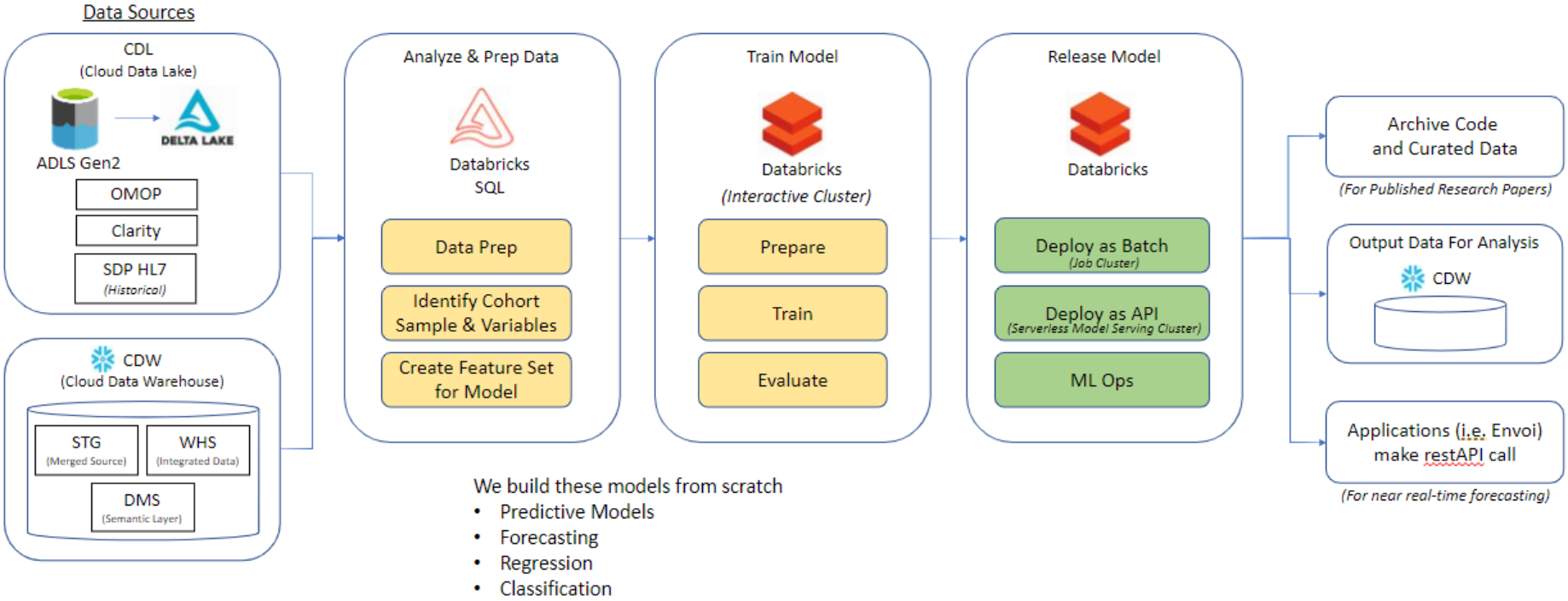
We began the migration in Jan 2024 and it's a work in progress

- Requires prep work with legacy workspace owners
- Understand existing workflows, data sources and destinations, and how to organize their data in new UC
- Test new jobs and pipelines before shutting down legacy workspaces
- Onboarded 20+ different teams to UC!

©2024 Databricks Inc. — All rights reserved

Design Patterns - Machine Learning

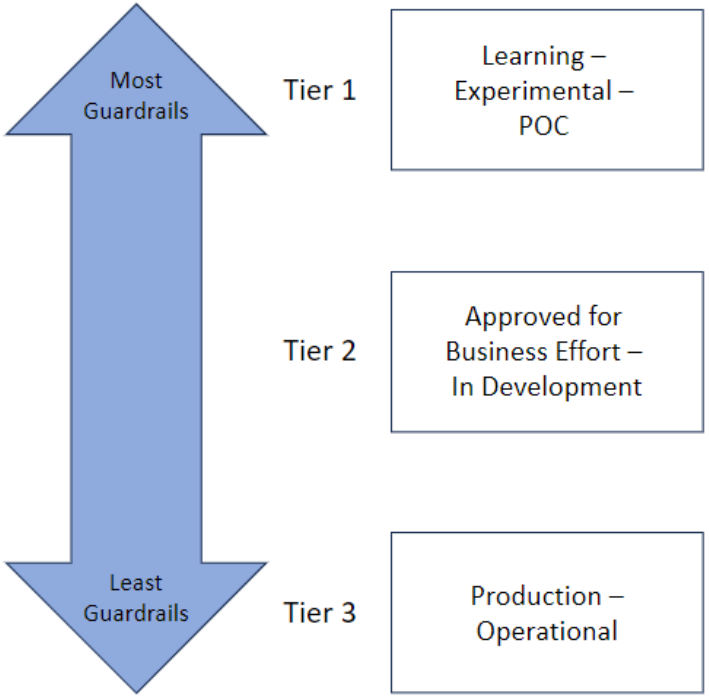
- Reviewed existing HI Workspaces and created 7 design patterns
- Helped inform cluster policy recommendations and cross collaboration team on use cases



©2024 Databricks Inc. — All rights reserved

Enable Guardrails via Policies

- 40 Policies put in place thus far
- Cluster Policies are tied to a users' AAD groups



Cluster Guardrails
Restrict to types of clusters (job, interactive, dlt, etc.)?
Restrict from scheduling interactive clusters
Restrict Interactive Clusters in Prod (only for development)
Set default to small cluster/VM size
Set default to auto-terminate in 30 minutes
Are there certain types of VMs with reservations we want to promote the use of (make default)?
How many machines in the cluster? (min/max range)
Default # nodes
Do we want to restrict dbus used in hour?
Assign tags with default values based on type of processing/pattern

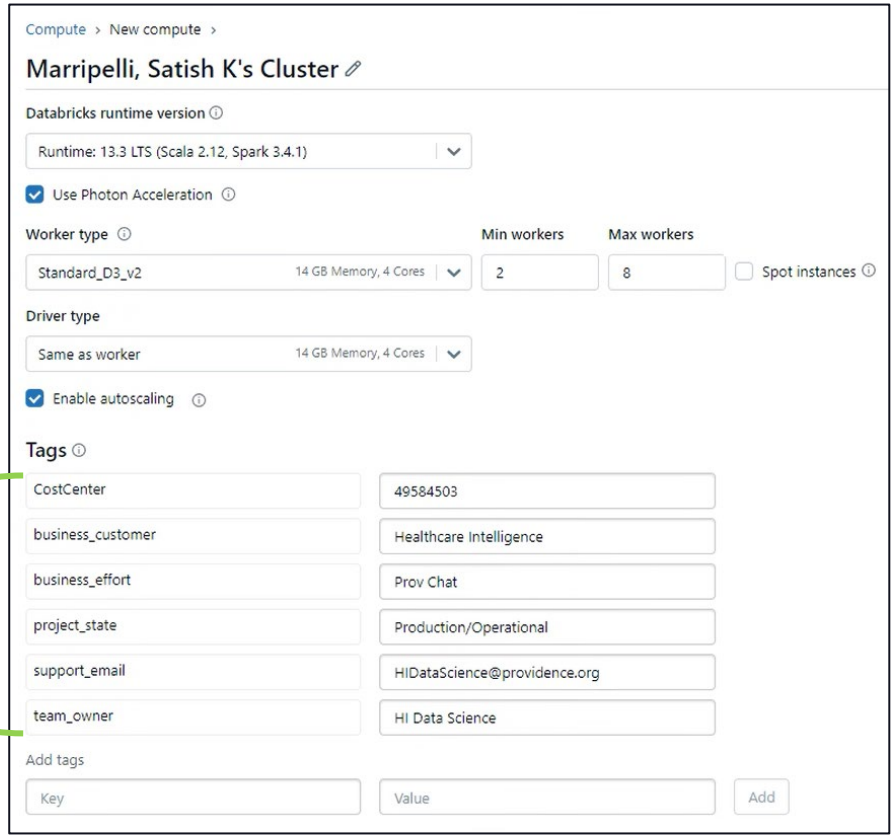
Other Guardrails
Restrict new workspace creation outside of Unity Catalog
Ability to opt-in to new features (deny preview items)

©2024 Databricks Inc. — All rights reserved

Reinforced Tagging Through Policy

- We used to have one workspace per project
- Now that we're consolidating workspaces, it was important for us to use tags for visibility into telemetry, cost/usage, and lifecycle management.
  - Policies by users' AAD Group, Pattern, and cluster type
  - Automatically assign tags or have defined list of values (no more free text)

Business Customer	Who is this for? Which downstream customer benefits from this processing?
Business Effort	Business friendly name for the project/effort
Project State	Used to indicate where the cluster is in the lifecycle and to manage guardrails.
Support Email	Contact email for operational support
Team Owner	Functional team that manages the cluster



©2024 Databricks Inc. — All rights reserved

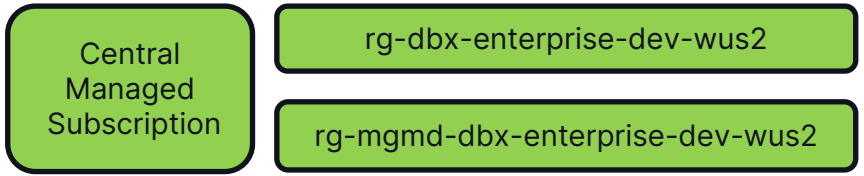
Cloud Value Optimization Goal

- For Azure Databricks, we saved ~ $420K in 2023
  - Partnered with Databricks and Microsoft to understand Total Cost of Ownership – DBU, Support, Influenced Costs
  - Reviewed highest cost workspaces
    - Understand value the data processing provides
    - Collaborate with Databricks to identify key data patterns to inform guardrails and provide visibility
    - "Show backs" leading to changes in behavior
    - Identified low value workspaces that should be sunset
  - Focus on optimization
    - Changing timeouts and right-sizing high cost VMs
    - Optimizing Foundational Data (Clarity Ingest, De-Identification, Truveta Trucking Service)
    - Remove low cost/unused workspaces
  - Use data to better inform P3 Reservations and future commit agreements
- Continued focus on optimization in 2024
  - Unity Catalog Migration: focus on rightsizing VMs, optimization, and turning off low value clusters.
  - Interactive vs Job Clusters: policy put in place to restrict scheduling clusters (up to 5X the cost of job clusters). Work on changing behavior to use interactive clusters for development only.

©2024 Databricks Inc. — All rights reserved

Databricks Total Cost of Ownership

When a Databricks cluster is running, it incurs DBU, Support, and Influenced Costs. Use Azure Portal for big picture view of total cost of ownership.



Non-Managed Resource Group Includes Databricks DBU + Support Costs

Managed resource group includes Databricks Influenced Costs

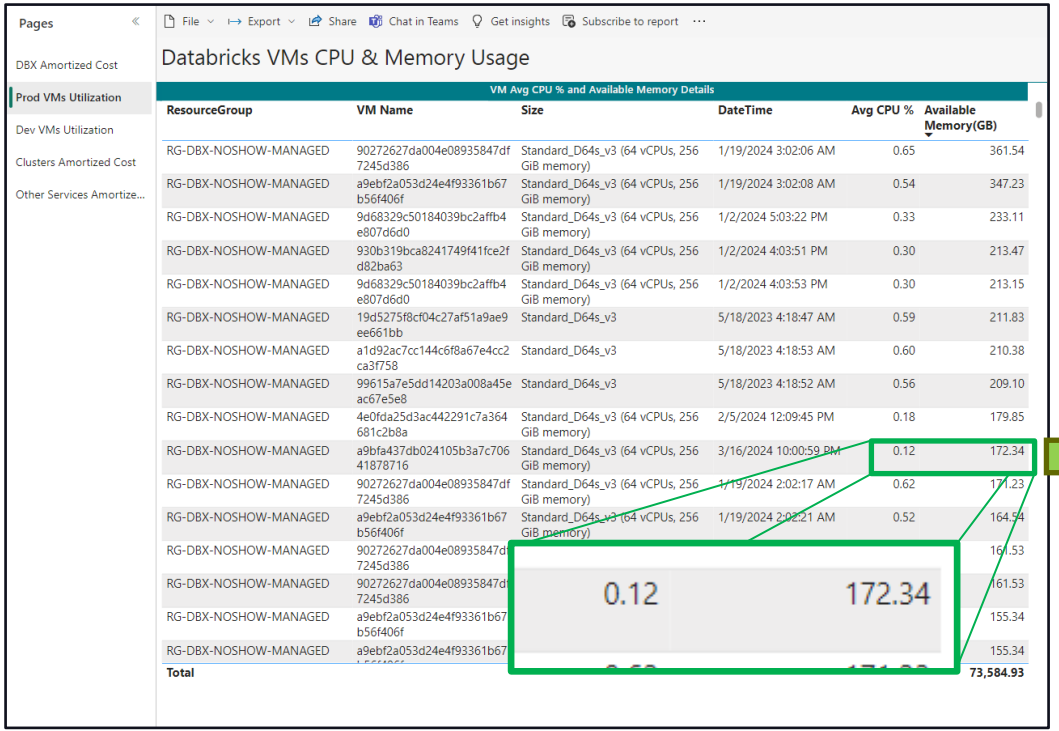
Resource Group	Databricks Cost Type	Service Name	Amortized Cost Amt*	% TCO
rg-dbx-enterprise-dev-wus2	DBU	Azure Databricks	126,057	63.03%
	Support	Virtual Network	223	0.11%
		Azure DNS	10	0.00%
rg-mgmd-dbx-enterprise-dev-wus2	Influenced	Virtual Machines	52,375	26.19%
		Virtual Network	13,205	6.60%
		Storage	7,901	3.95%
		Bandwidth	178	0.09%
		Microsoft Defender for Cloud	50	0.02%
		Grand Total	**200,000**	**100%**

* Synthetic data

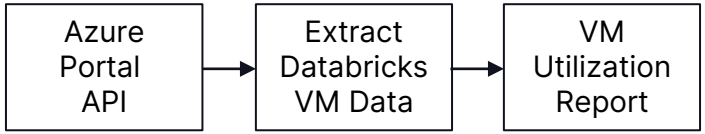
Note: Serverless clusters will not show Influenced Costs in Azure Portal. Expect DBU spend to increase to reflect these costs.

©2024 Databricks Inc. — All rights reserved

Cost Optimization - Rightsizing VMs



Extract data to find oversized and underutilized Databricks Virtual Machines



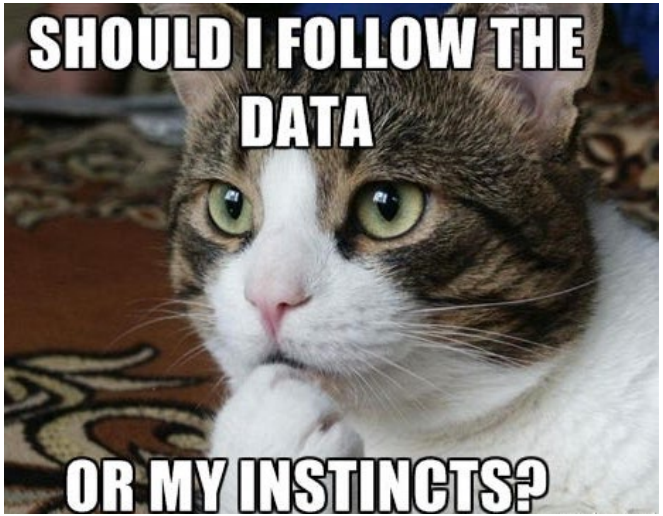
Low Average CPU % Utilized and High Available Memory (GB) indicated we need to follow up with the project teams to right size the clusters/VMs.

©2024 Databricks Inc. — All rights reserved

Unity Catalog - System Tables

Allows for reporting of both Unity Catalog and non-Unity Catalog workspaces

- Identified clusters with time out of 120+ minutes
- Jobs usage (the exact DBUs that each jobs cluster consumes)
- DBUs by cluster type/SKU
- Audit logs – tracks user activities



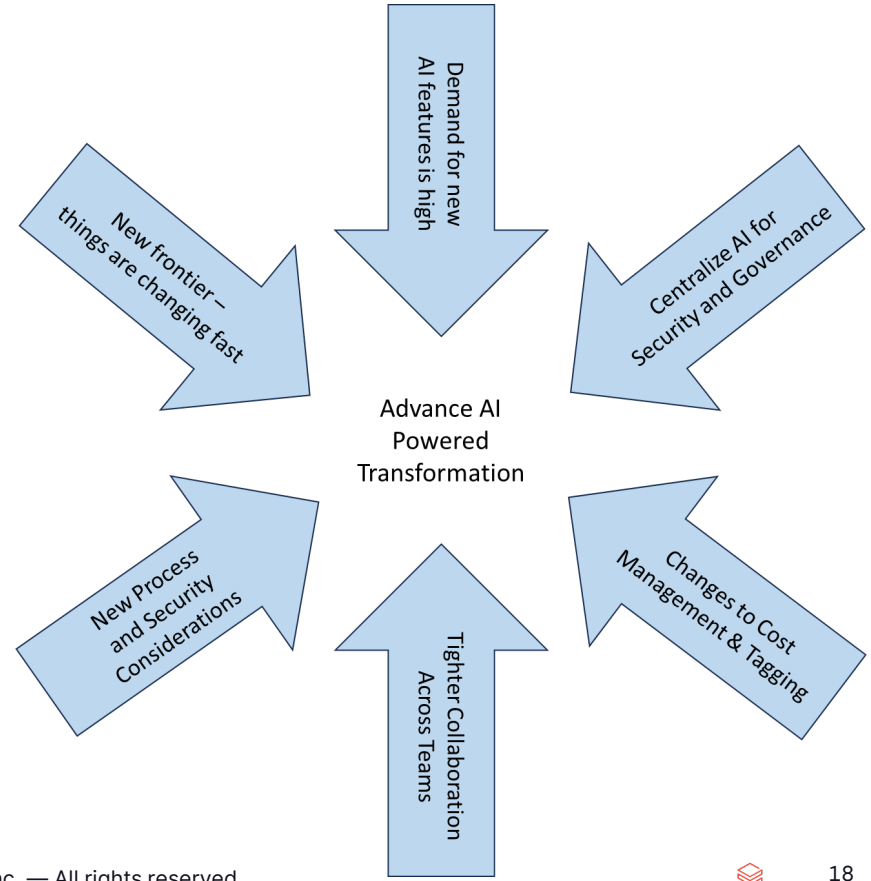
©2024 Databricks Inc. — All rights reserved

What about serverless? What's Next?

We are focused on serverless and other Databricks features that advance our AI work.

Enabled Serverless Model Serving

- Initial use case is to deploy one of our machine learning models as a restAPI so it can be called by downstream applications



Emerging AI Forces

Healthcare is a highly regulated environment.
Tension between:

Being Innovative ⟺ Being Safe/Secure

©2024 Databricks Inc. — All rights reserved

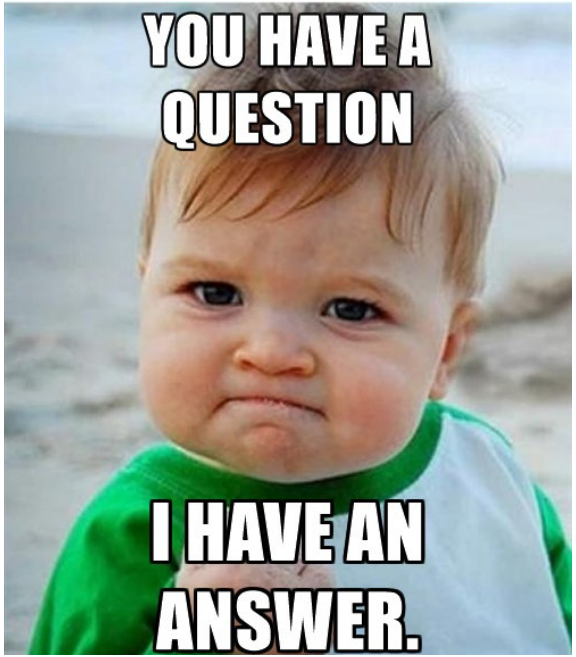
Enabling AI Features

- Partner with Databricks to share our AI use cases and learn about upcoming features
- We are focused on infrastructure to enable Azure Open AI Models
- Keep Databricks Feature Road Map up to date and plan quarterly for next set of features to enable

©2024 Databricks Inc. — All rights reserved

Questions?

Thank you for attending our presentation!



©2024 Databricks Inc. — All rights reserved