

# Administration & Identity Best Practices

Future-proofing Your Databricks Account



**Siddharth Bhai**  
Product Management,  
Databricks



**Gaurav Bhatnagar**  
Product Management,  
Databricks



**Vicky Avison**  
Staff Data Engineer,  
Plexure

# 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.

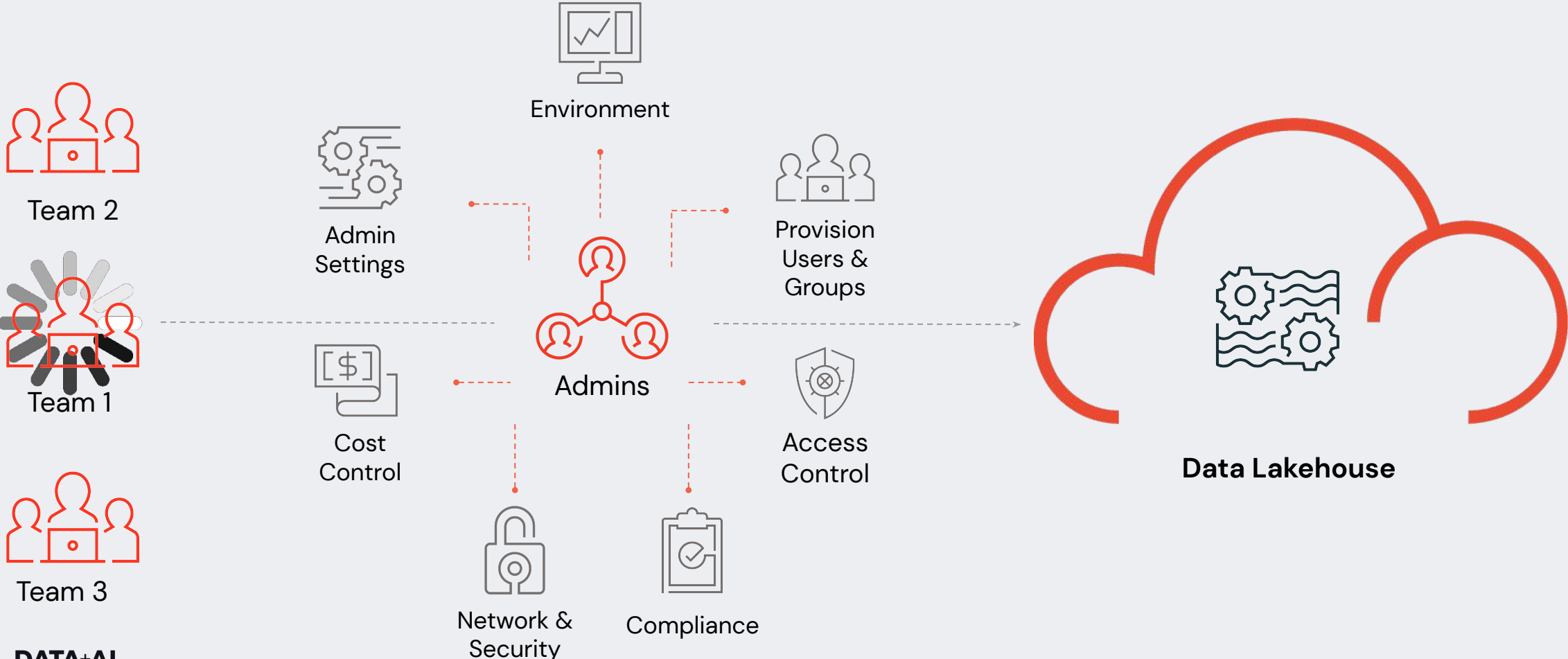
# Session objectives

Managing Databricks at scale as your DB usage deepens

1. Reference **architectures** to organize Databricks
2. **Case Study**—Databricks adventures in the real world
3. New product **updates** and evolved **best practices** with a demo

# Customer journey

Practitioners in the organization have to wait for admins to configure the environment just right, before they can get access.



# Setting up & Managing Databricks

# Setting up & Managing Databricks

1. Roles and responsibilities
2. Patterns for organizing workspaces
3. Scope of management

# Roles and Responsibilities

## Core personas



Databricks admin



Team admin



Data admin



Analyst



Data scientists



Data engineers

## Stakeholder personas



Cloud Ops  
Admin



Identity  
Admins

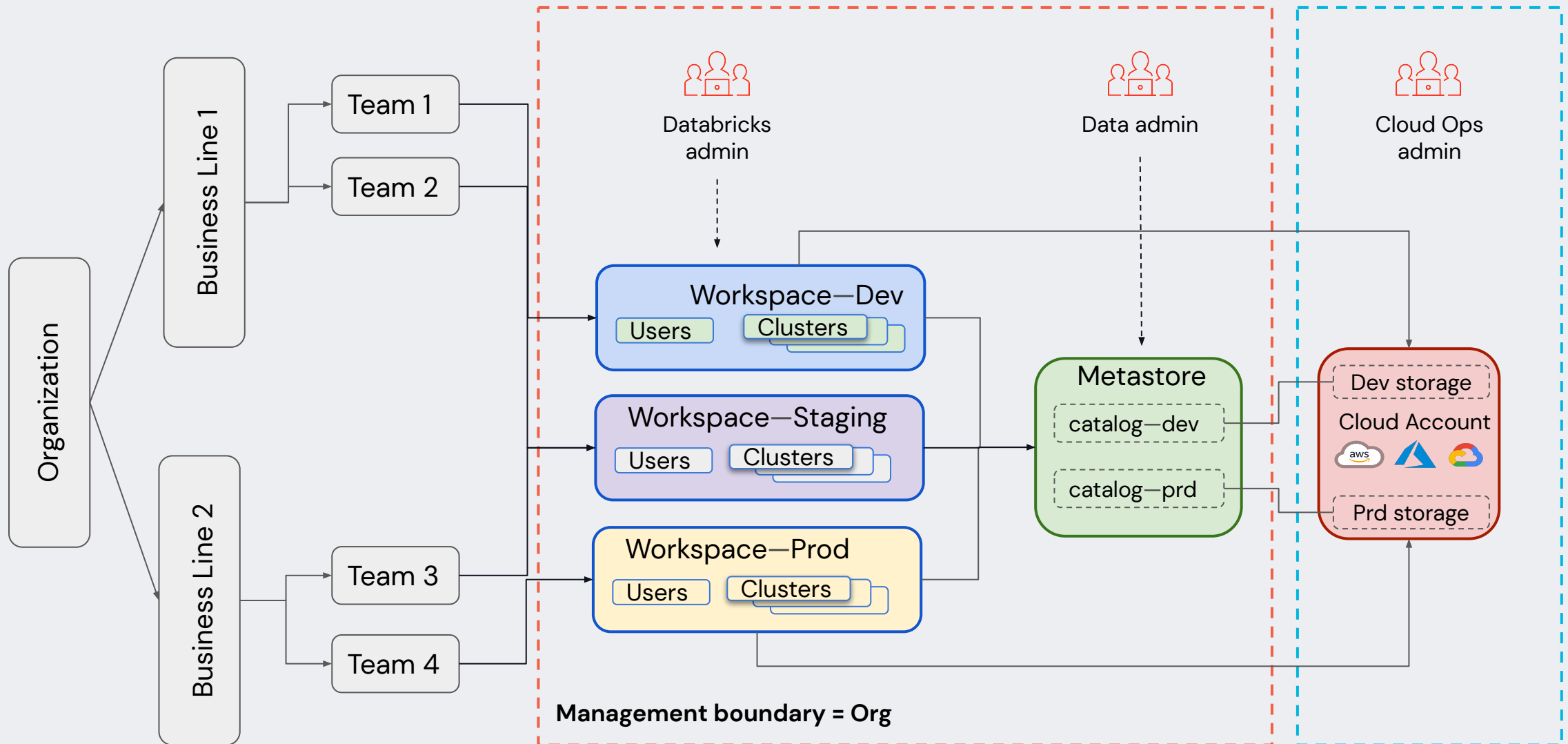


Billing &  
procurement



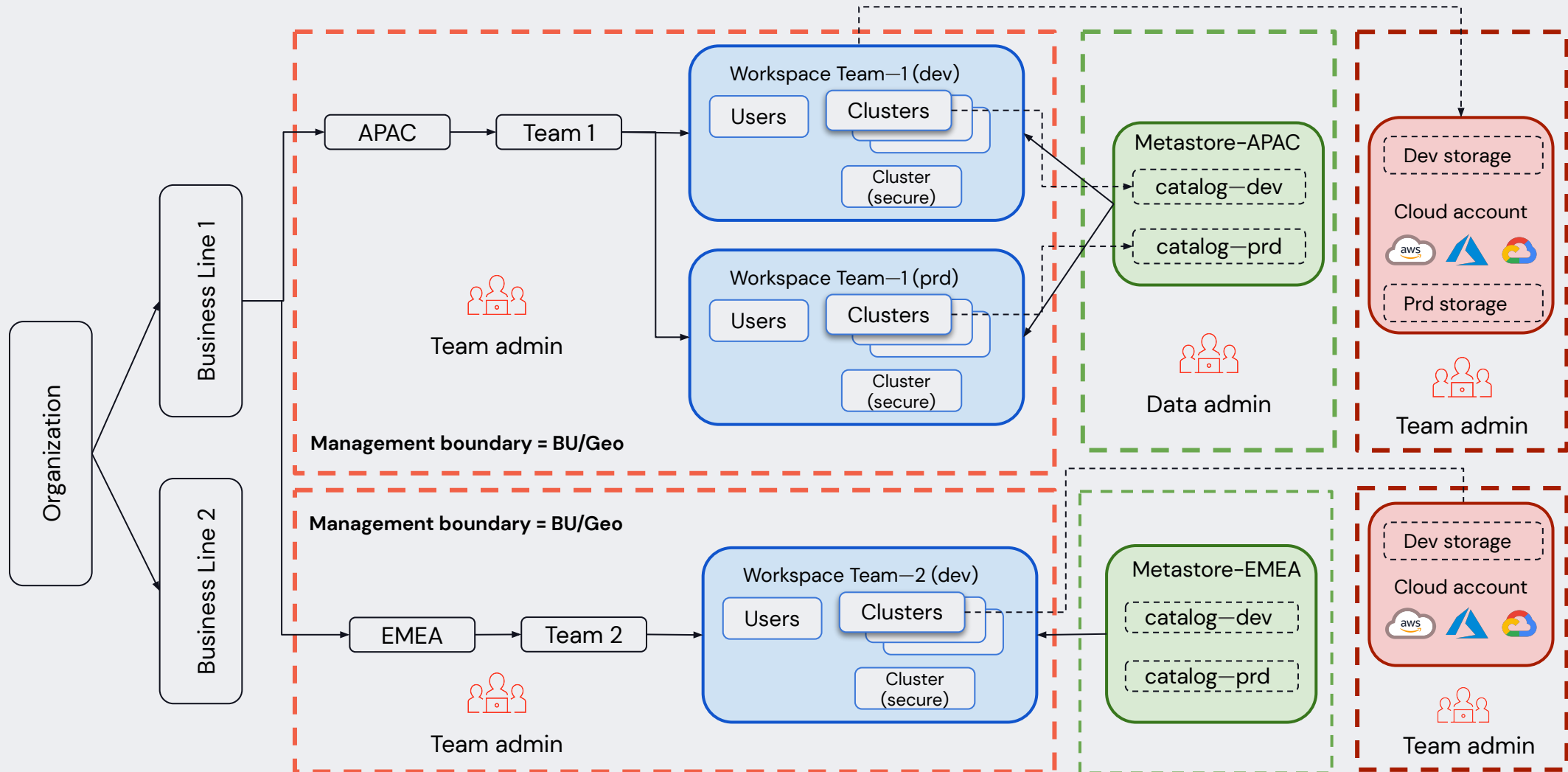
Security &  
compliance

# Centralized set up





# Decentralized set up (multi-geo/multi-team)



# Best practices: scaling administration



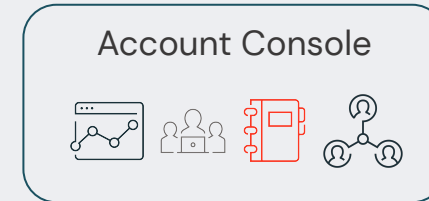
## Set up Central Data platform team

Onboards teams  
Ongoing governance



## Invest in automation

Workspaces  
Pipelines  
SPN or Service Accounts



## Manage centrally

SCIM provisioning  
SSO  
Audit and billing  
Enable features



## Cost management

Resource tags  
Cluster policies



## Networking & security

Ent or Prem SKU  
BYOVPC and private link



## Limit # of Workspaces

Strict network/data isolation  
Data products in separate regions  
Account/workspace limits

# Case Study: Databricks deployment

## Customer nearby

Plexure identifies a customer's location and formulates right offer at right time to influence a purchase. A push notification is sent via QSR's mobile app

## Customer at home



Customer redeems offer at POS via mobile app

Plexure starts gathering behavioural and activity data against the customer profile

Occasional customer downloads the mobile app

Plexure updates customer profile with purchase data to optimize future product/ offer recommendations

## Customer arrives

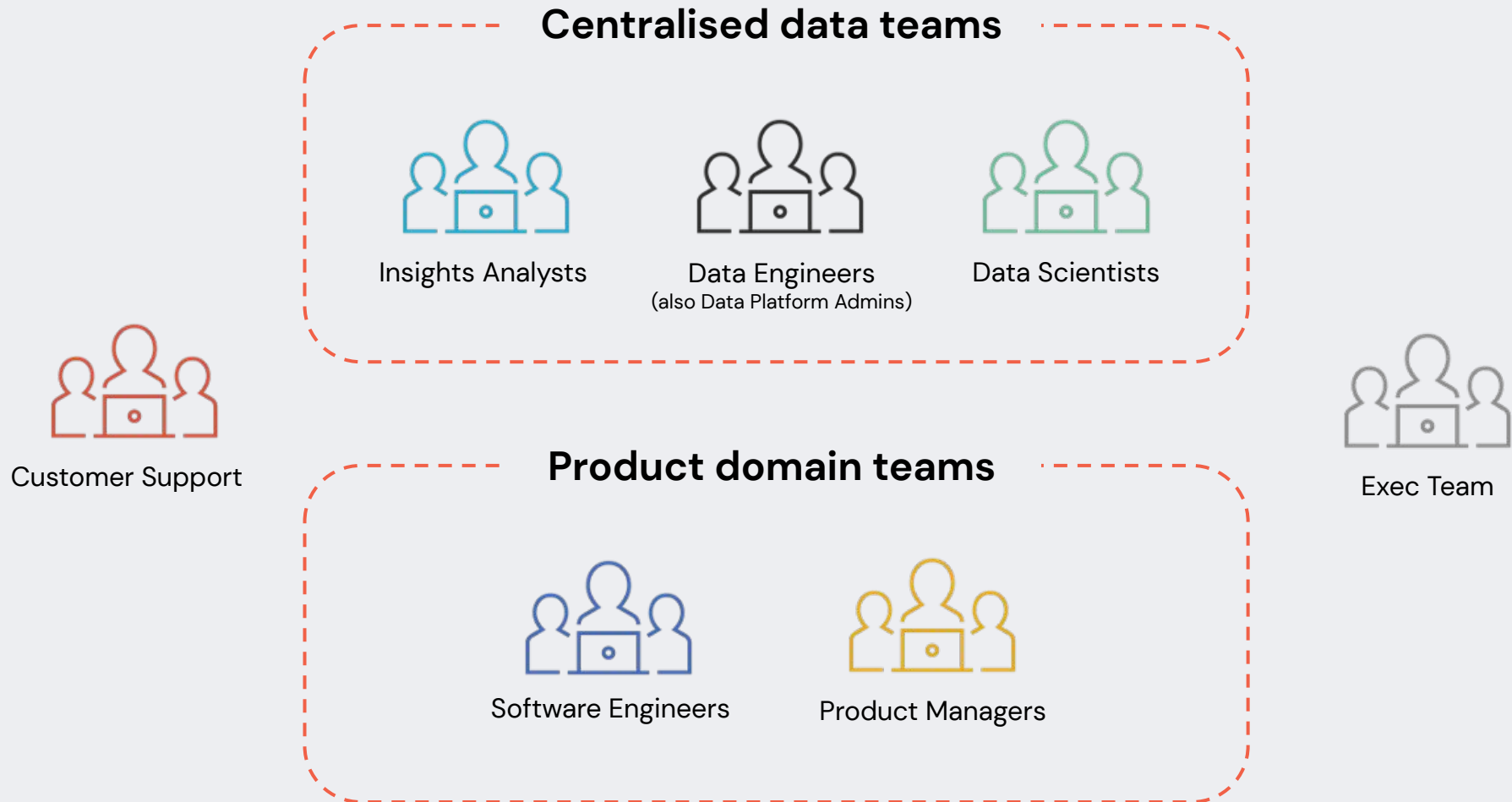
Plexure utilizes location data to identify their arrival

Customer loyalty points can be redeemed at any time by exchanging them for a special reward (e.g. an extra cheeseburger). As a result, it encourages consumers to return and use the points to earn rewards, etc

Consumers (who sign up to the app) earn loyalty points at POS or MOP

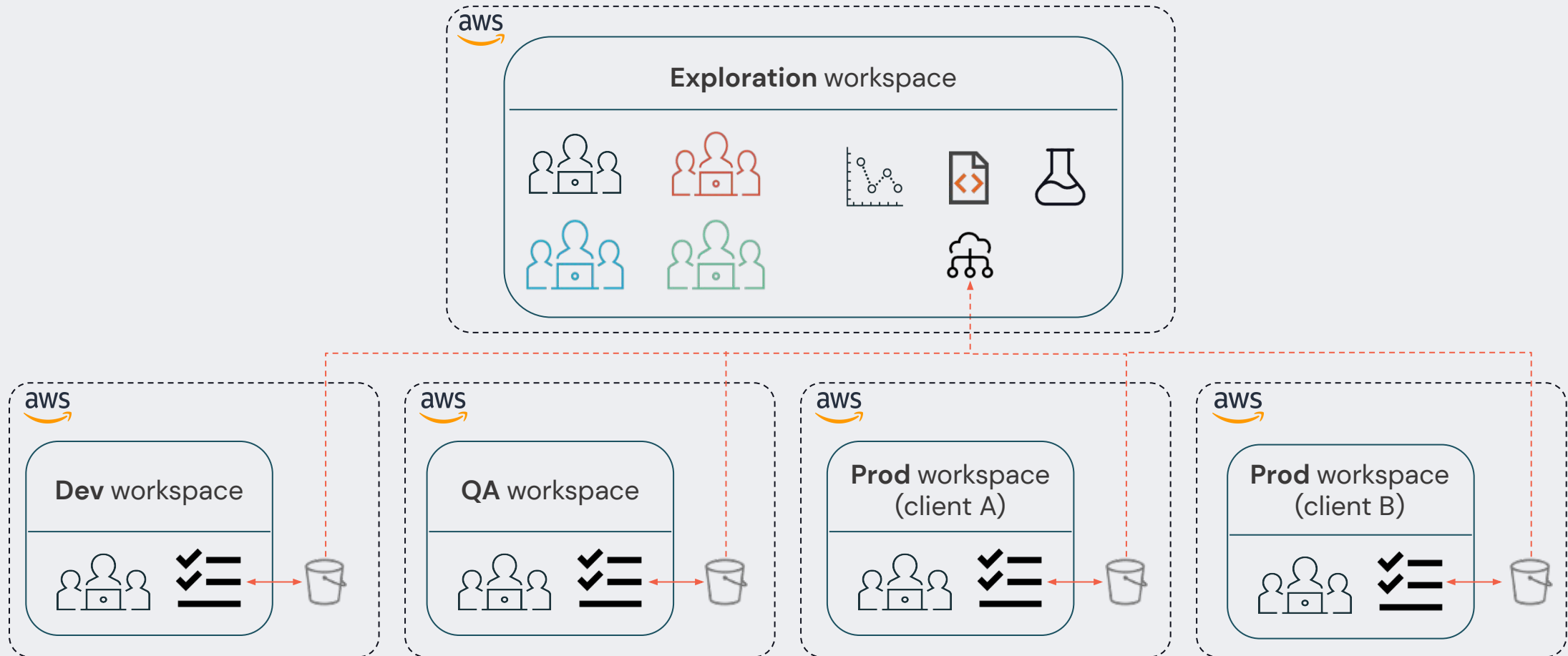
Customer orders and pays via mobile app and specifies desire fulfilment method (ie. curbside/ drivethrough/ table service)

Plexure starts gathering behavioural and activity data against customer profile



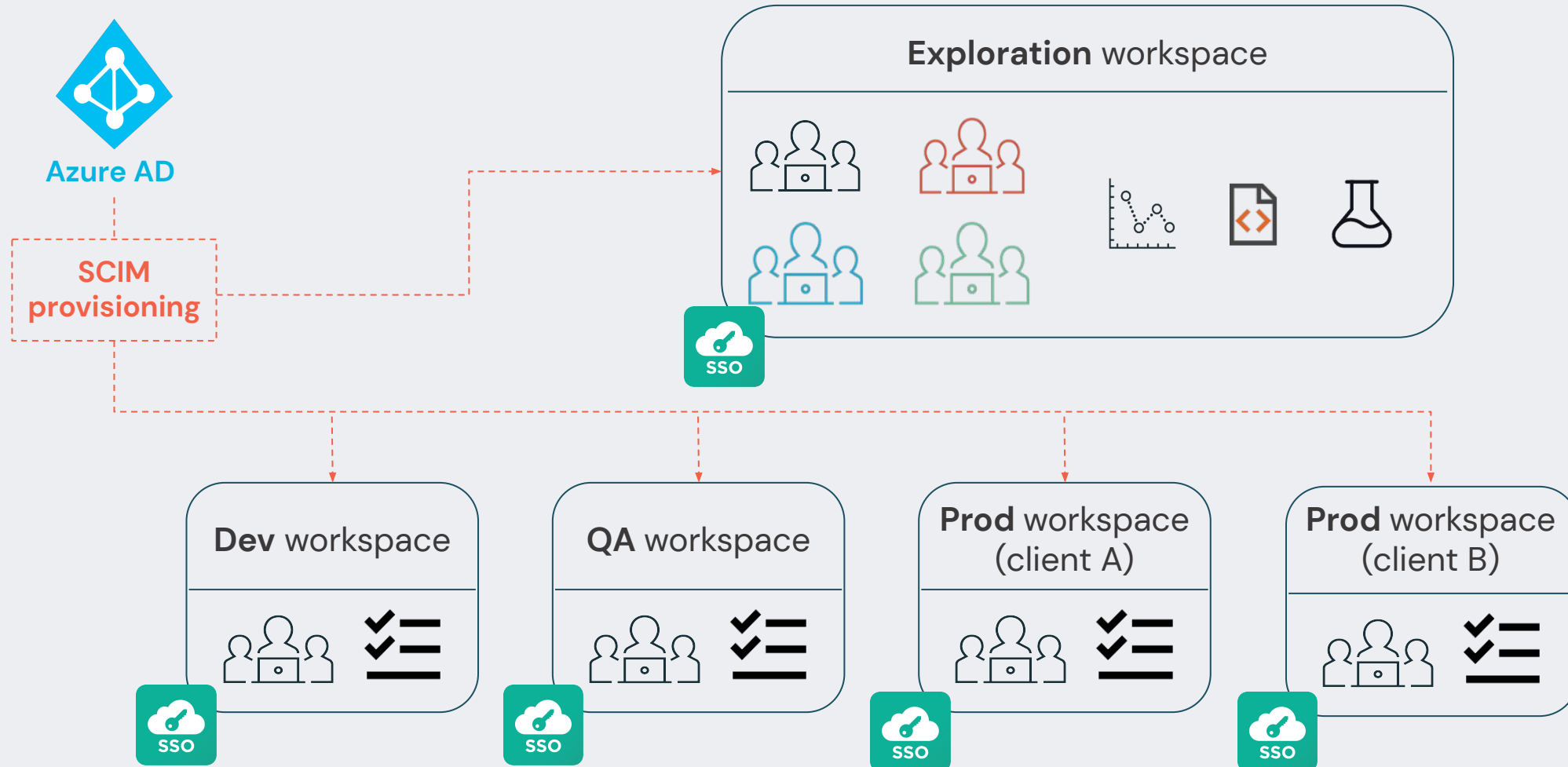


## Workspace architecture





# Identity management



# Using a single workspace

## An alternative approach

### Pros:

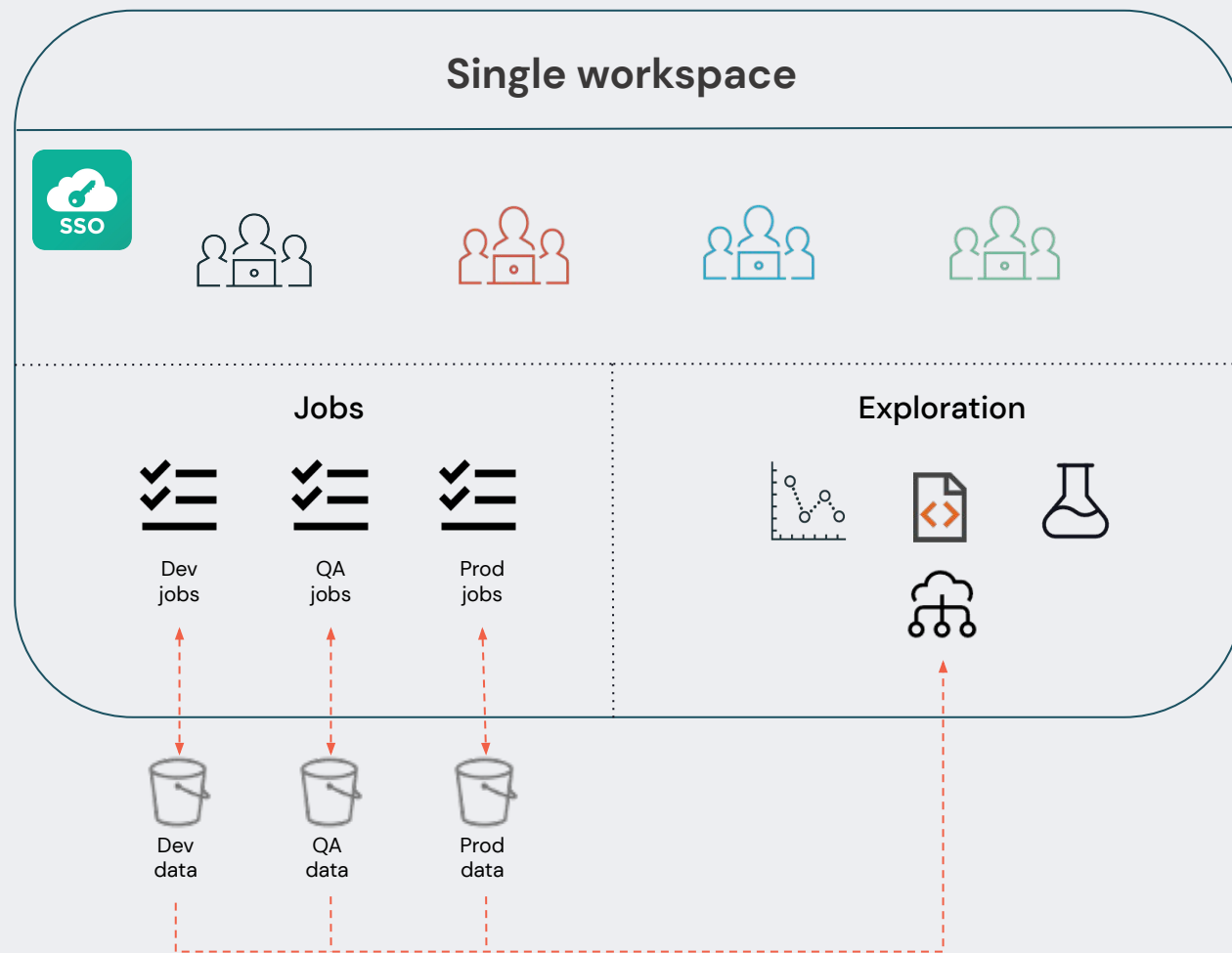
Everything in one place helps with monitoring/visibility

Simpler admin e.g. only have to configure SSO/SCIM provisioning once

### Cons:

Don't have the same level of data isolation

Can only configure a single region per workspace





# Automated administration

## Terraform

Almost everything fully automated via **terraform**, including:

- Workspace creation
- Cluster & SQL endpoint creation
- Workspace permissions
- Unity catalog (including data grants)

Manually administered:

- SSO and SCIM provisioning
- Workspace settings

```
resource "databricks_cluster" "general" {
  cluster_name           = "General"
  spark_version          = data.databricks_spark_version.latest_lts.id
  driver_node_type_id   = data.databricks_node_type.driver.id
  node_type_id          = data.databricks_node_type.worker.id
  autotermination_minutes = 30
  enable_elastic_disk    = true
  enable_local_disk_encryption = true
  spark_conf             = local.general_spark_conf
  is_pinned              = true
  autoscale {
    min_workers = 1
    max_workers = 10
  }

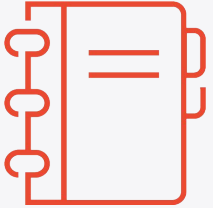
  data_security_mode = "USER_ISOLATION"
}

resource "databricks_permissions" "general_cluster_usage" {
  cluster_id = databricks_cluster.general.id

  access_control {
    group_name       = data.databricks_group.data_engineers.display_name
    permission_level = "CAN_RESTART"
  }

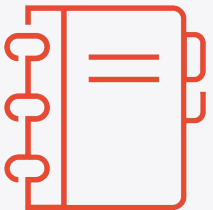
  access_control {
    group_name       = data.databricks_group.ds_and_insights.display_name
    permission_level = "CAN_RESTART"
  }
}
```

# Lessons learned

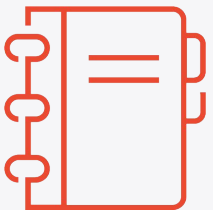


Terraform provider is **worth its weight in gold**

...But not everything is supported and we have to revert to **manual administration** (e.g. workspace settings)



SCIM provisioning and SSO make it really **easy to onboard** new users/groups



... but it's a **pain** to set up with multiple workspaces

# New Product Updates

# New product updates

Manage Databricks easily at scale

## 1. Easy administration

Revamped Account Console

Identity Federation

## 2. Easy manageability

Single Cross-workspace metastore

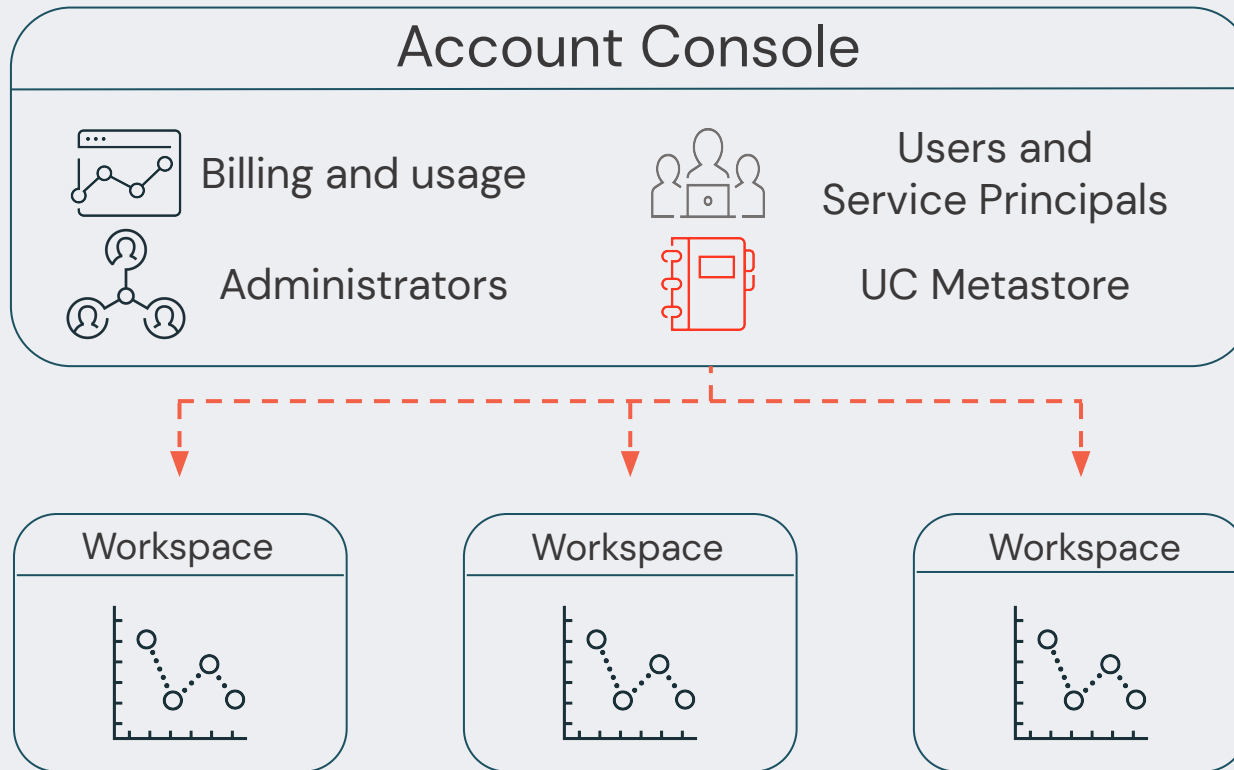
Faster onboarding of new teams

**Demo** to view these in action & Updated Best Practices

Coming Soon : Account Console  
(Some settings - Still In Development)

# Account Console

Single pane of glass to administer workspaces



Account-Level Identities

Centralized Billing  
Information

Cross-workspace Unity  
Catalog metastore

IP Allow Lists

# Identity Federation

Account level users and groups usable in workspaces

Workspaces > DAIS22-dev

## DAIS22-dev

Configuration **Role assignments**

Assign users, groups, and service principals to this workspace. To add users, groups, and service principals to your account, go to the [Users & Groups](#) tab.

Q Search Search Add role assignments

Name	Type	Roles
Data Analysts	Group	User
Data Engineers	Group	User
Siddharth Bhai (siddharth.bhai@databricks.com)	User	Admin

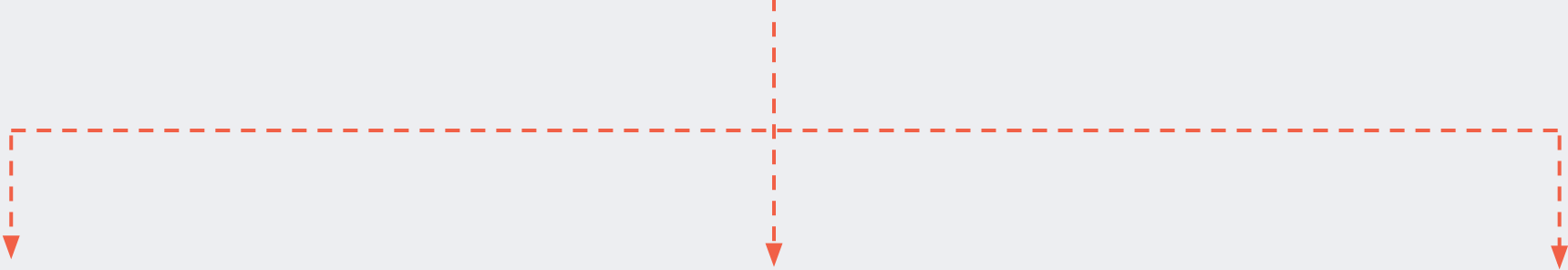
Identity Federation refers to assigning account-level users, service principals, and groups from the account to workspaces

Existing model: WS identities  
(Generally available)



Identity sync is reconfigured **repeatedly**  
from an Identity Provider for workspaces



Workspace  
SCIM APIs





**Workspace 1**

Workspace Users + Service Principals 	<b>Workspace local</b> Groups 
---	---

**Workspace 2**

Workspace Users + Service Principals 	<b>Workspace local</b> Groups 
---	---

**Workspace 3**

Workspace Users + Service Principals 	<b>Workspace local</b> Groups 
---	---

Coming Soon : Identity Federation  
(In Preview now)

Identity sync is configured **once** from an  
Identity Provider



Account  
SCIM APIs

Account console

Account users +  
service principles

Account  
Groups

Non-ID Fed workspace

Account users +  
service principles

Workspace local  
Groups



ID Fed workspace with existing groups

Account users +  
service principles

Workspace  
local Groups

Account  
Groups



ID Fed workspace

Account users +  
service principles

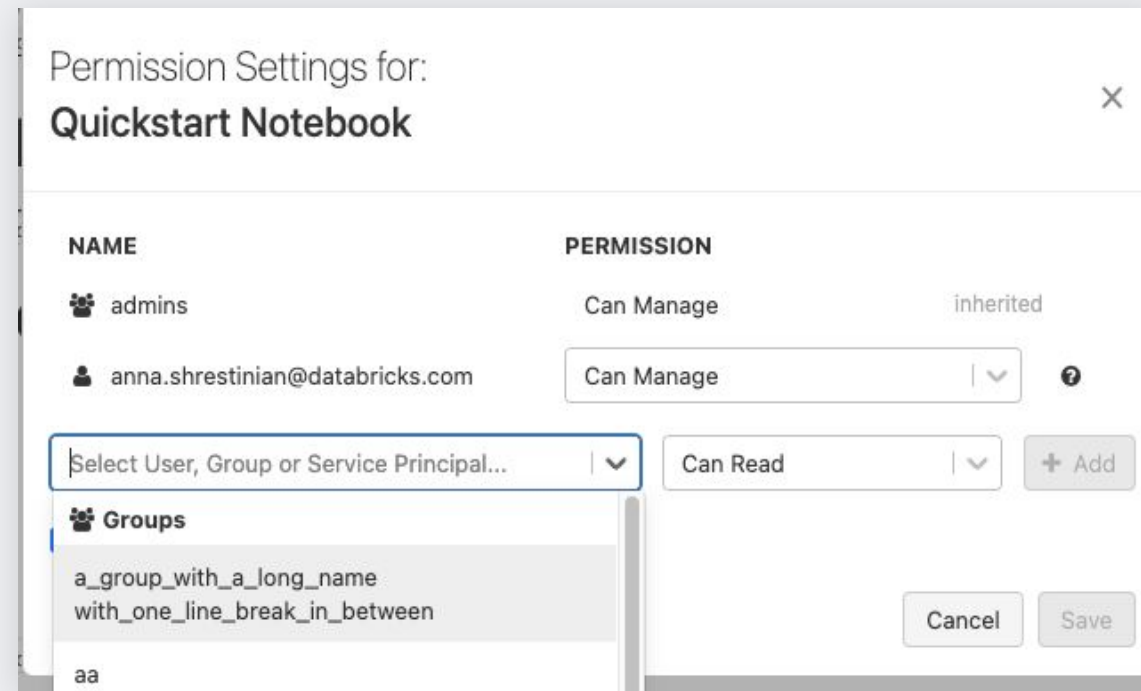
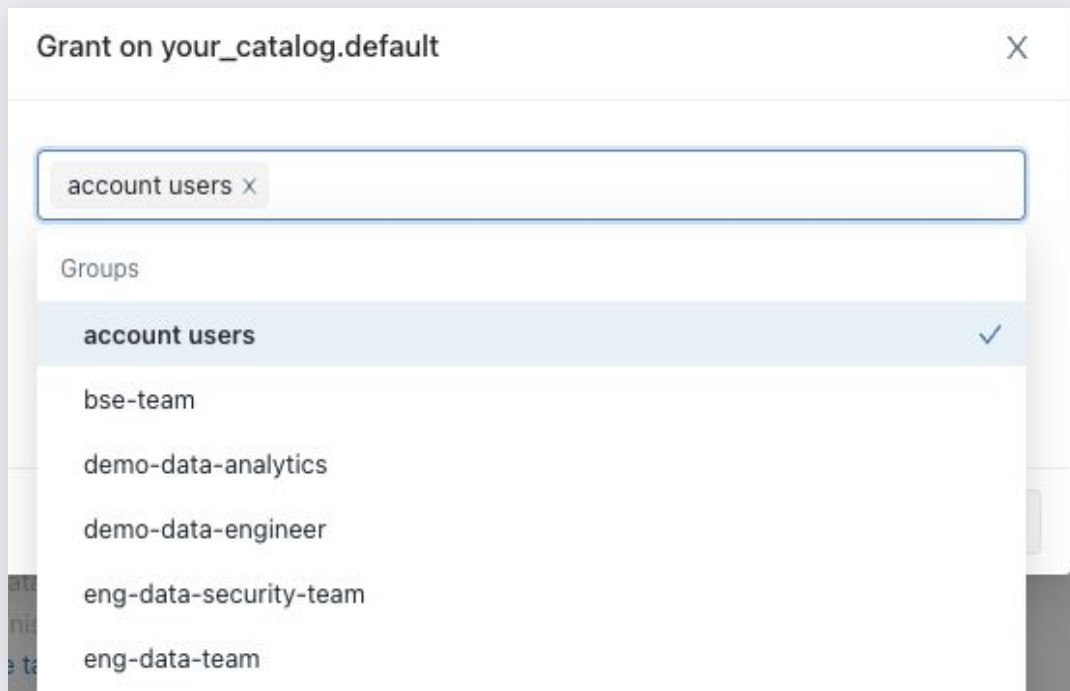
Account  
Groups





# Identity Federation

Account level users and groups usable in workspaces



## Sharing account-level objects

Users can search/resolve all account users, service principals, and groups (even when not assigned to the workspace)

## Sharing workspace-level objects

Users can search/resolve only the account users, service principals, and groups that have been explicitly assigned to the workspace.

# Best Practices with account level identities

- 1 —• **Sync all users** & relevant groups to account, subject to scale limits.
- 2 —• **Use a scoped SCIM token** to set up sync from IdP to account level.
- 3 —• **Watch audit logs** for SCIM or SSO configuration issues.
- 4 —• **Configure an IdP** at the account console, similar to workspaces.
- 5 —• **MFA for end-users**—Configure at your Identity Provider.
- 6 —• **Trust your admins**—Account and Workspace admins.

# Simplifying config management

Manage and govern settings easily from a single interface

Central management of all available settings at the Account layer

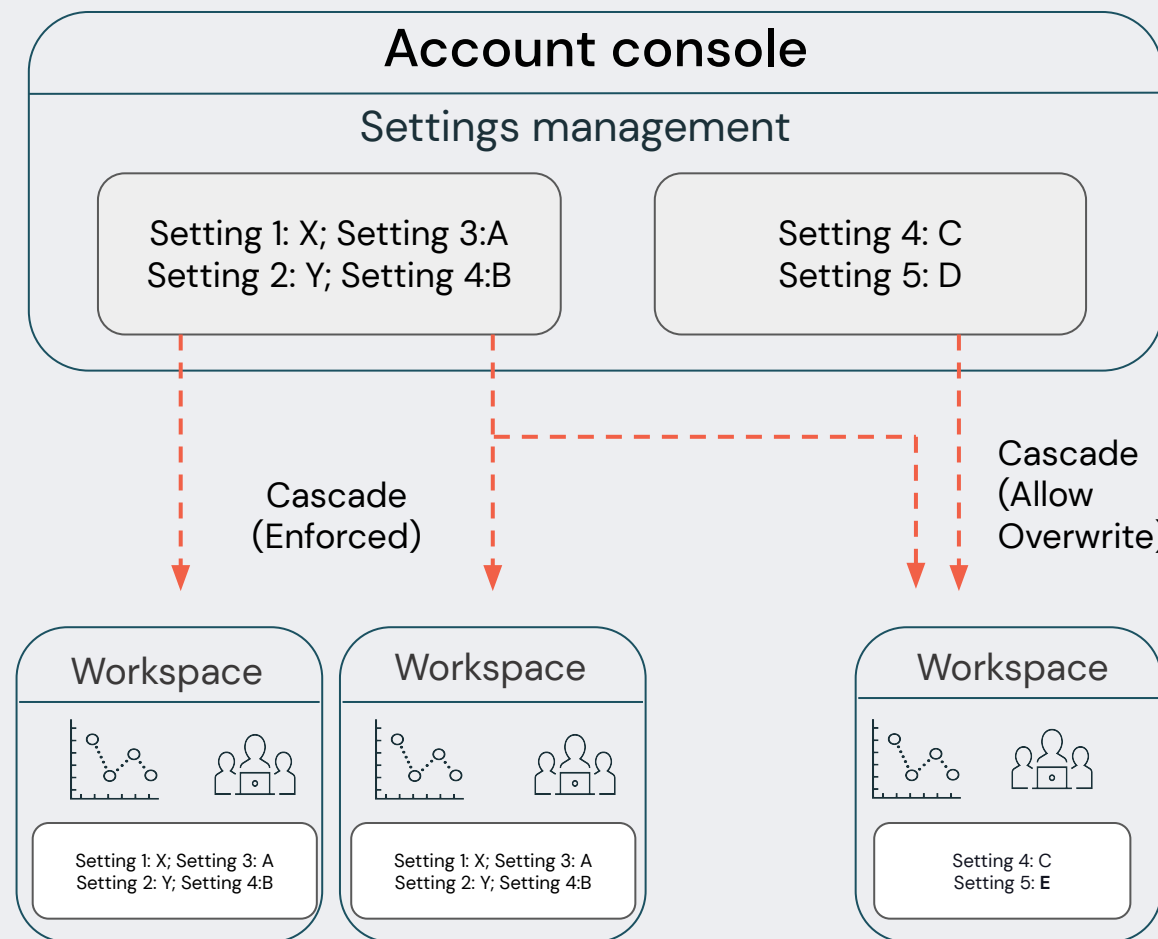
Set defaults once, apply everywhere

Enforce defaults or delegate control

At-scale management

Automatable & auditable

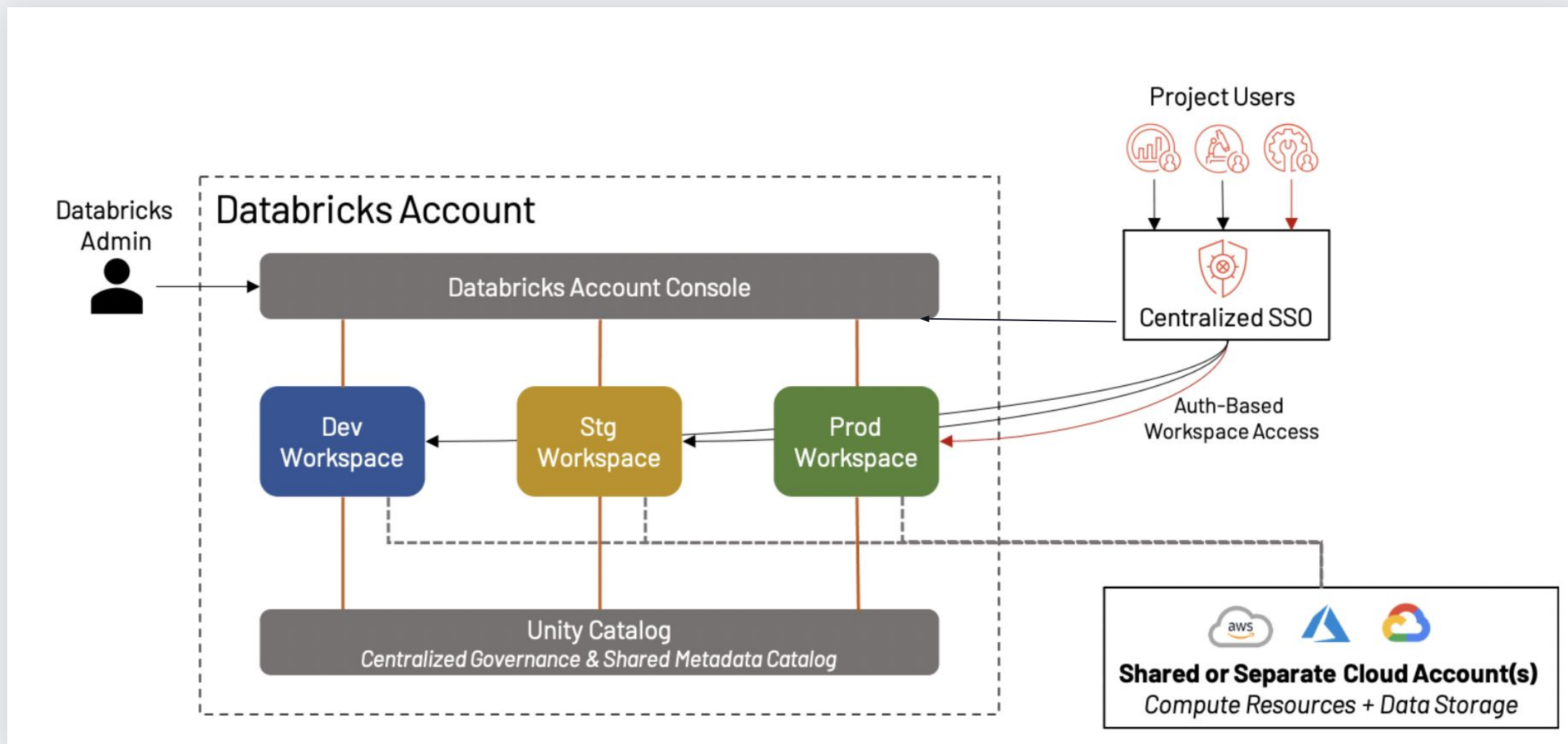
Constraints are correctly and consistently applied



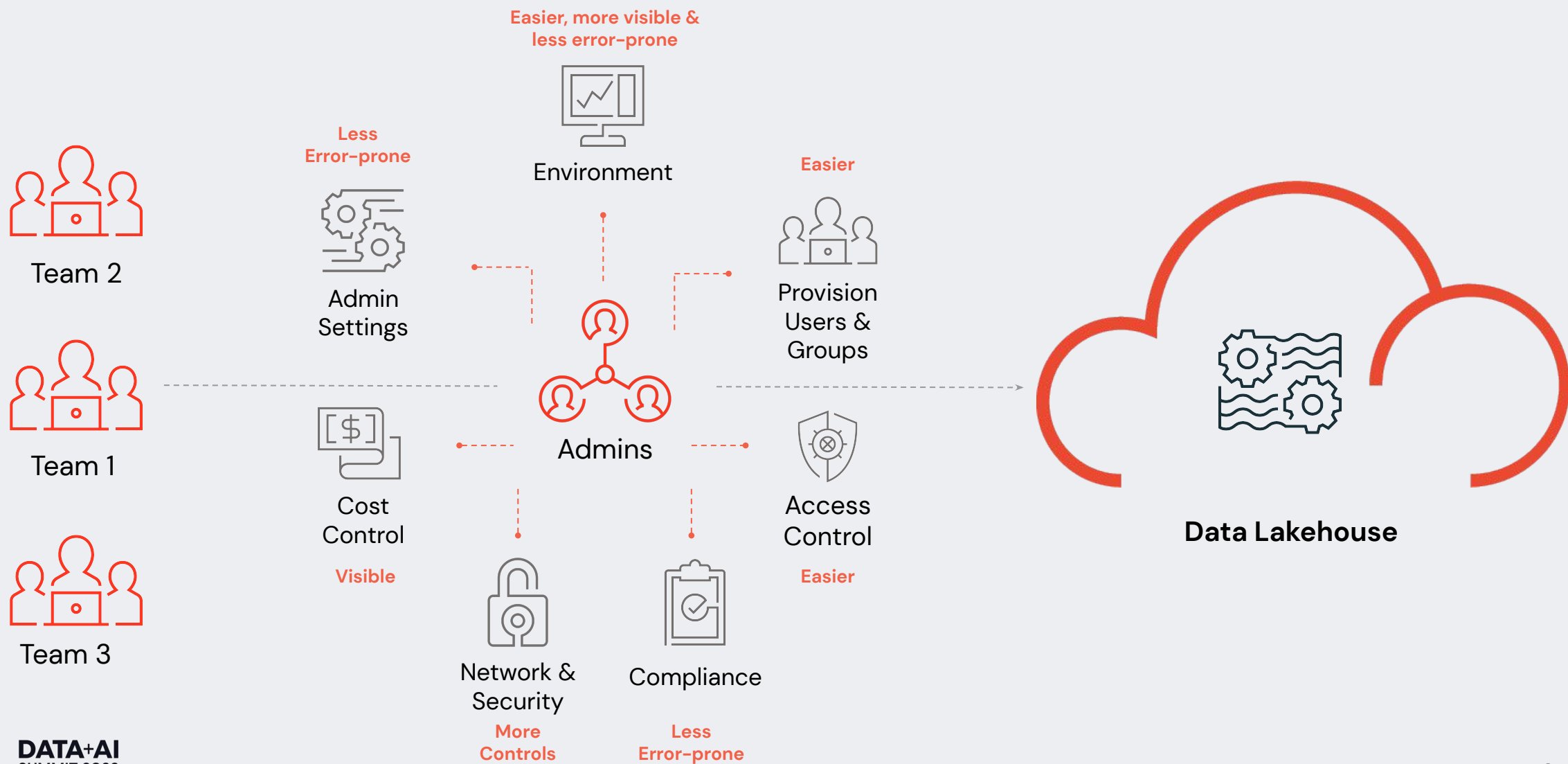
# Demo



# (Revamped!) Account console



# (New!) Faster onboarding



# Session takeaways

## **Establish identity at the account level, and sync all users and groups**

- This is the foundation of the lakehouse; users and data in one place
- Sharing data assets within the organization becomes easy

## **Establish UC at the account level; this allows you to leverage the identities**

- UC policies will cascade across workspaces, as well as across clouds
- Centralized identity powers UC

## **Centralized settings are key for simplicity, security and scale**

- Cascading settings from the account console enables you to scale
- Workspaces will inherit identities & settings for easy expansion



**DATA+AI**  
**SUMMIT 2022**

**Thank you**



**Siddharth Bhai**  
Product Management,  
Databricks



**Gaurav Bhatnagar**  
Product Management,  
Databricks



**Vicky Avison**  
Staff Data Engineer,  
Plexure