Microsoft

DataOps for the Modern Data Warehouse on Microsoft Azure

Lace Lofranco
Microsoft

Microsoft

Lace Lofranco

Senior Software Engineer
Commercial Software Engineering
Microsoft
@lacelofranco
github.com/devlace

Agenda

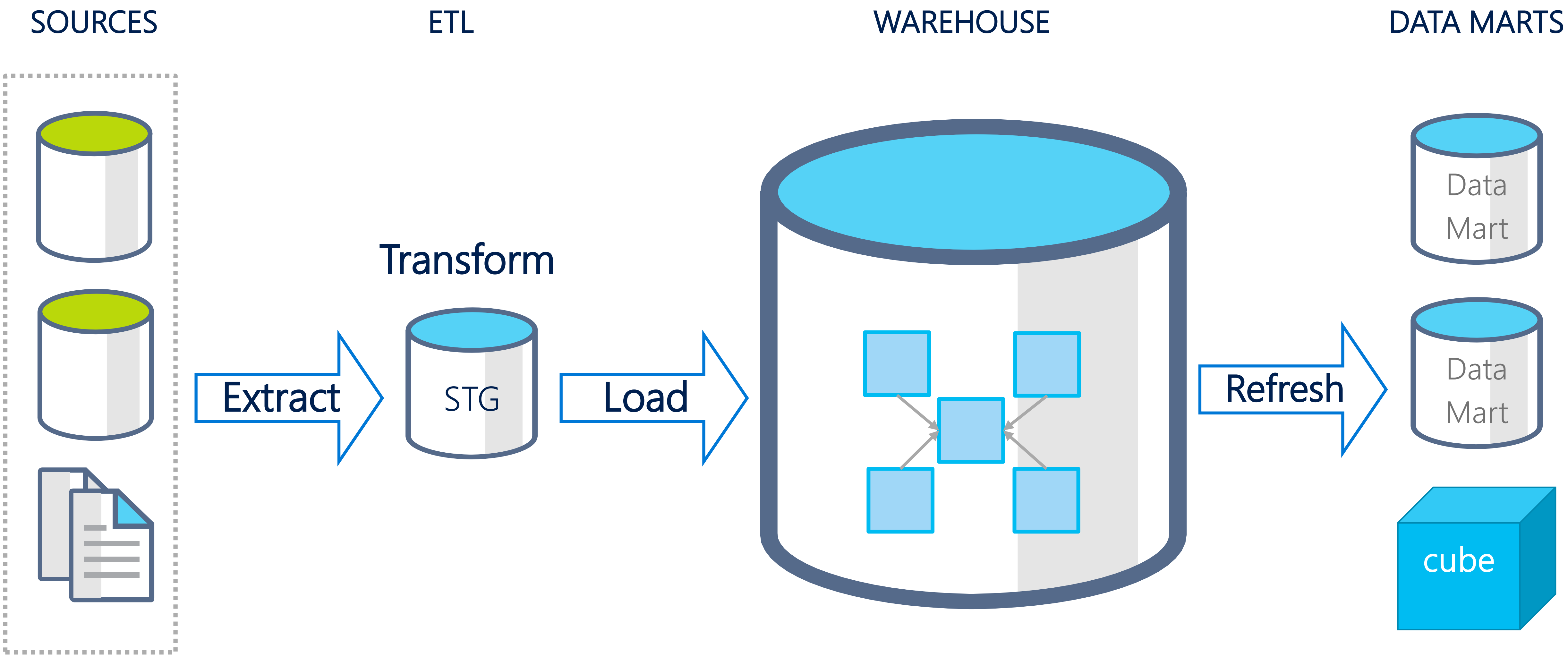
Modern Data Warehousing on Azure

How to operationalize? DevOps!

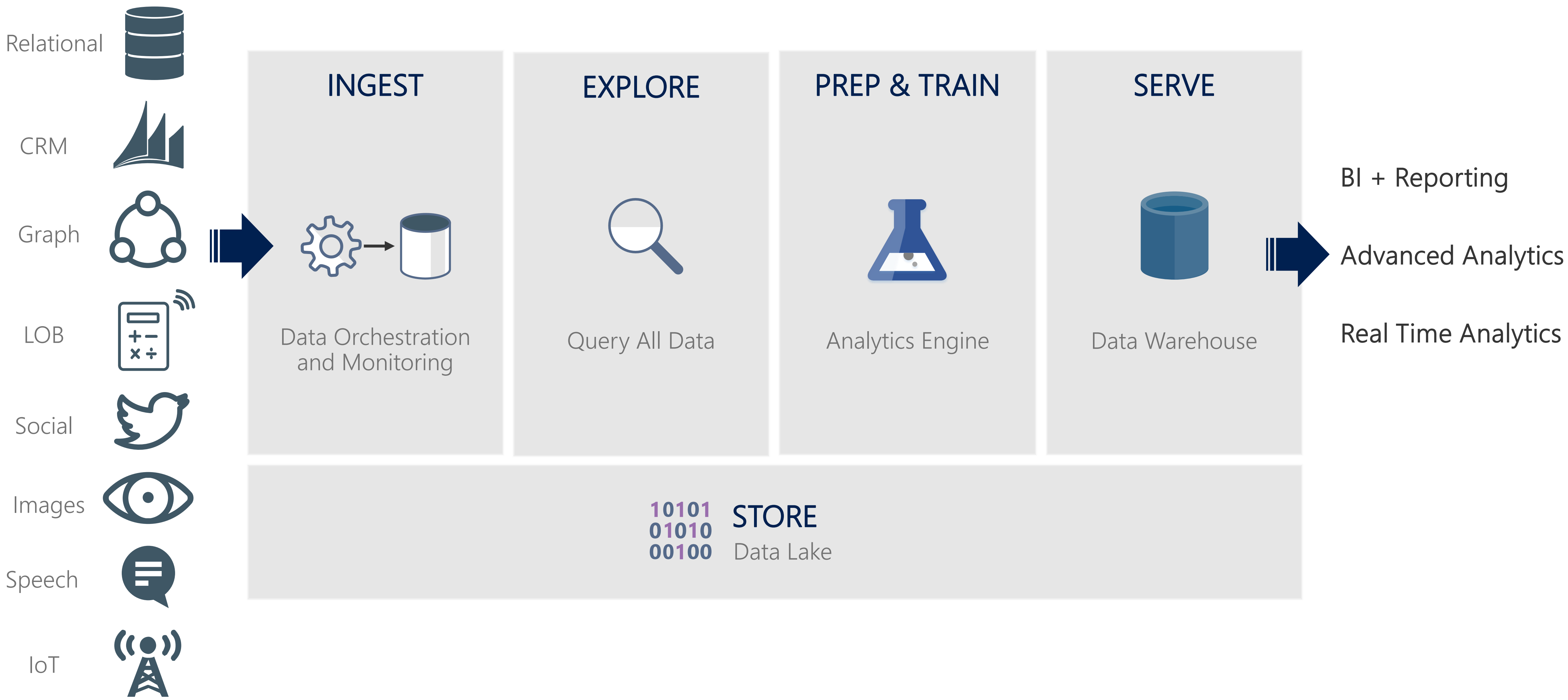
Modern Data Warehouse

Overview

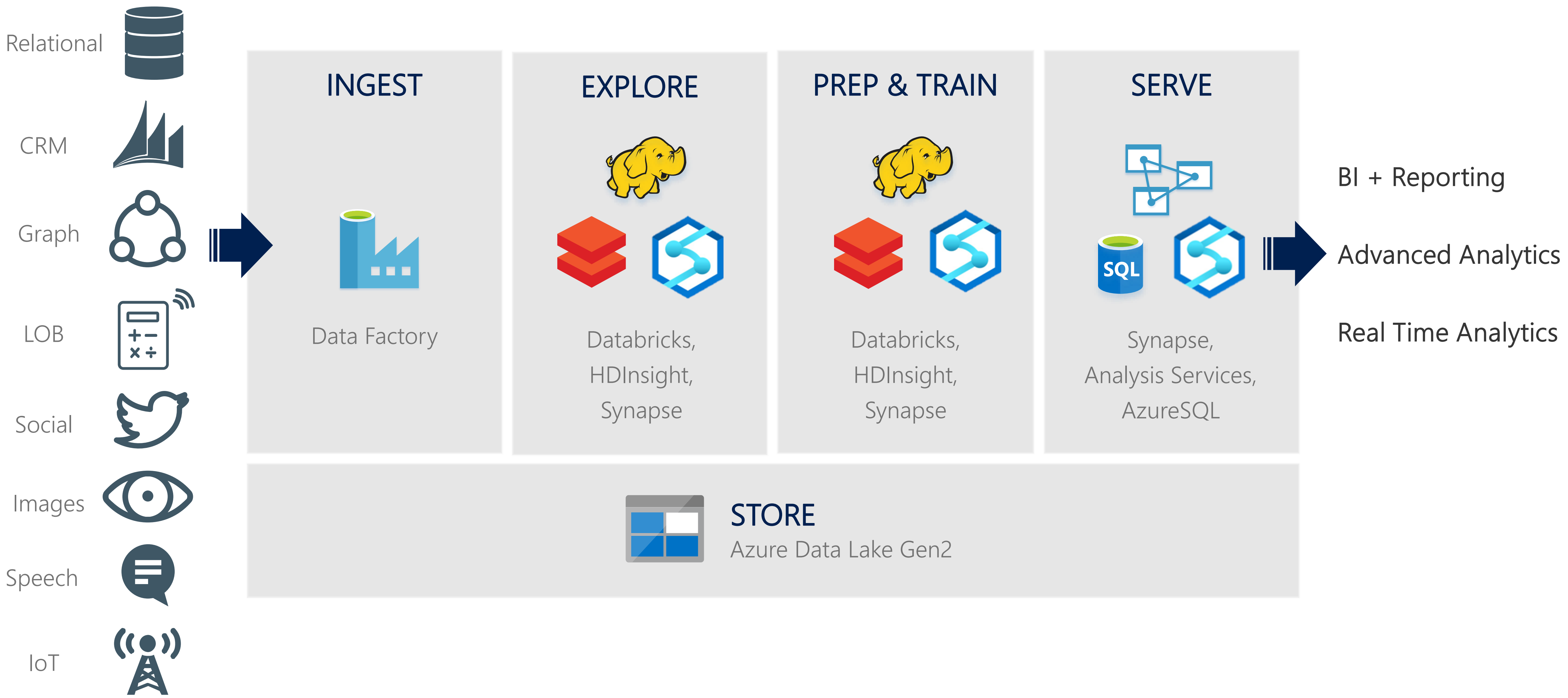
Traditional Data Warehousing



Modern Data Warehouse



Modern Data Warehouse on Azure



Relational

CRM

Graph

LOB

Social

Images

Speech

IoT

INGEST

Data Factory

EXPLORE

Databricks, HDInsight, Synapse

PREP & TRAIN

Databricks, HDInsight, Synapse

SERVE

Synapse, Analysis Services, AzureSQL

STORE

Azure Data Lake Gen2

BI + Reporting

Advanced Analytics

Real Time Analytics

Data Tiers

Bronze

Raw, unprocessed

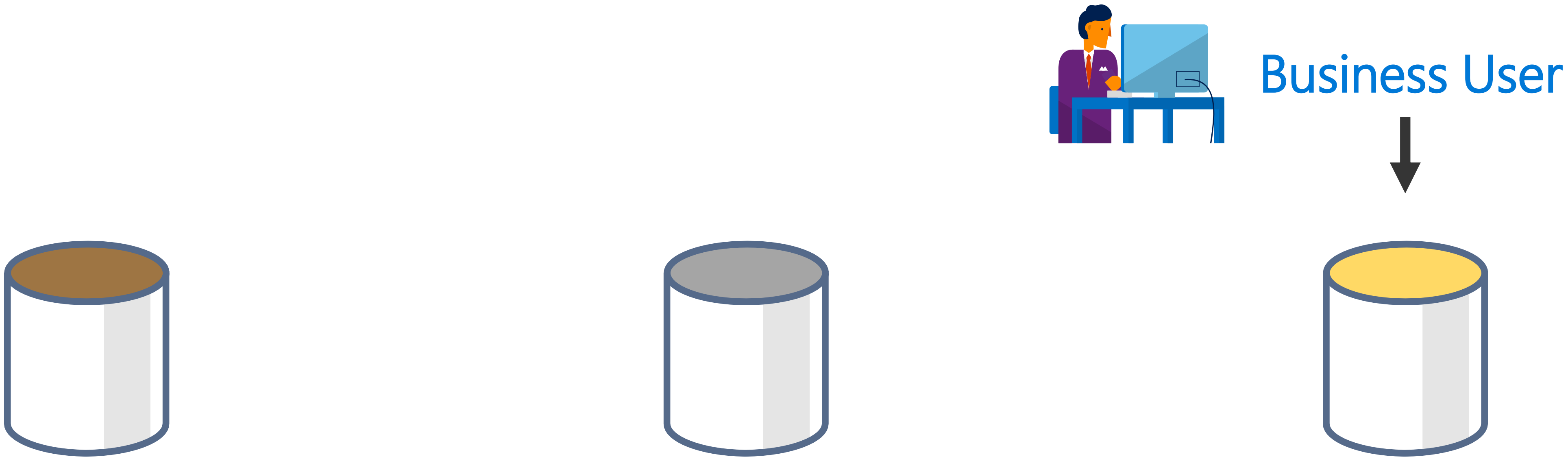
Silver

Cleansed,
augmented

Gold

Optimized
for consumption

Data Tiers - Users



Business User

Bronze

Raw, unprocessed

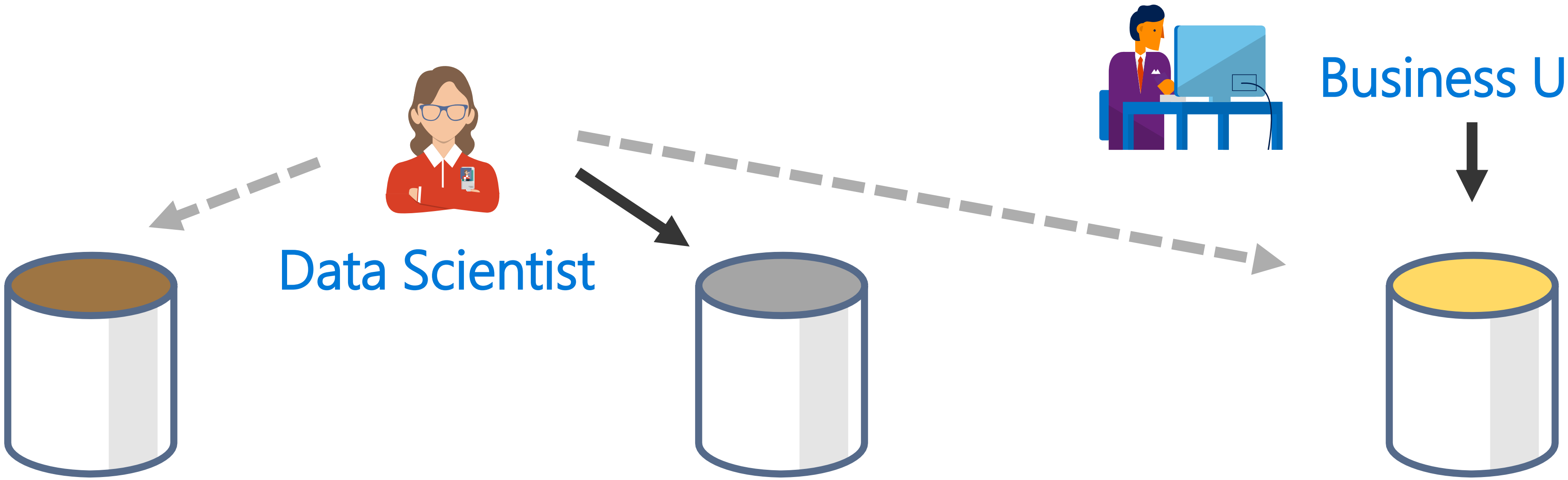
Silver

Cleansed,
augmented

Gold

Optimized
for consumption

Data Tiers - Users



Business User

Data Scientist

Bronze

Raw, unprocessed

Silver

Cleansed,
augmented

Gold

Optimized
for consumption

Learnings

Validate early in your pipeline.

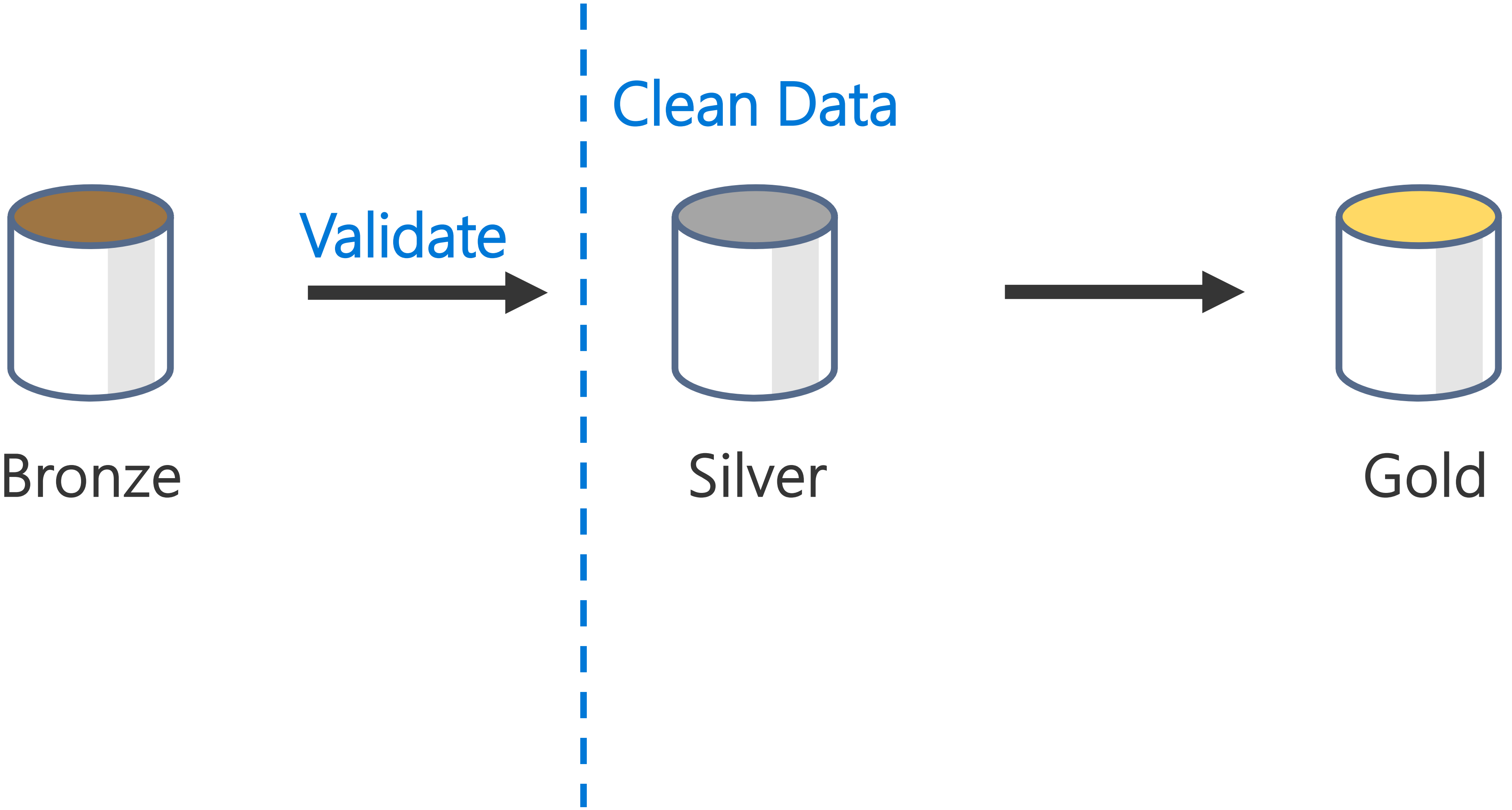
Validate data early in the Pipeline

Bronze

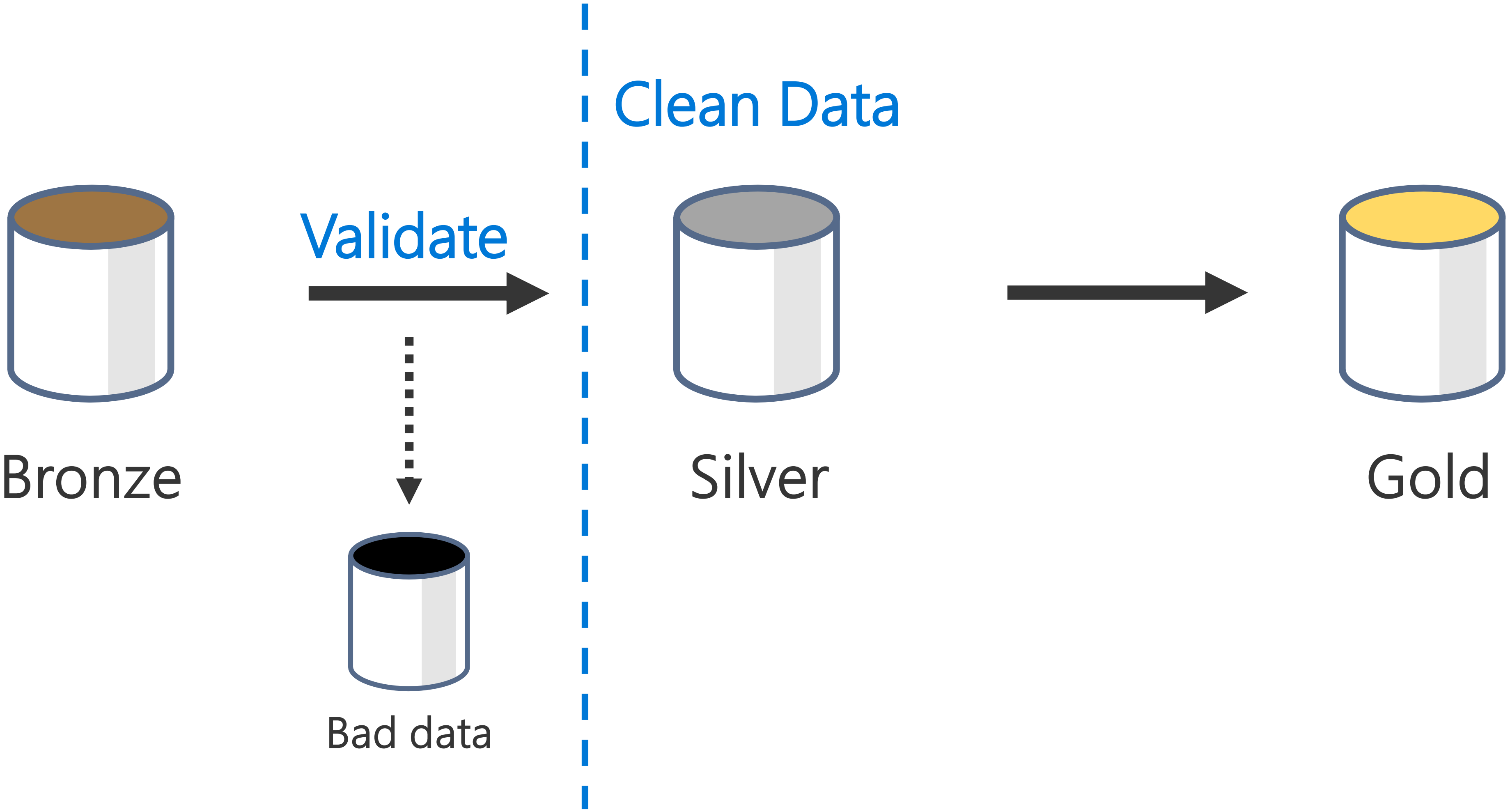
Silver

Gold

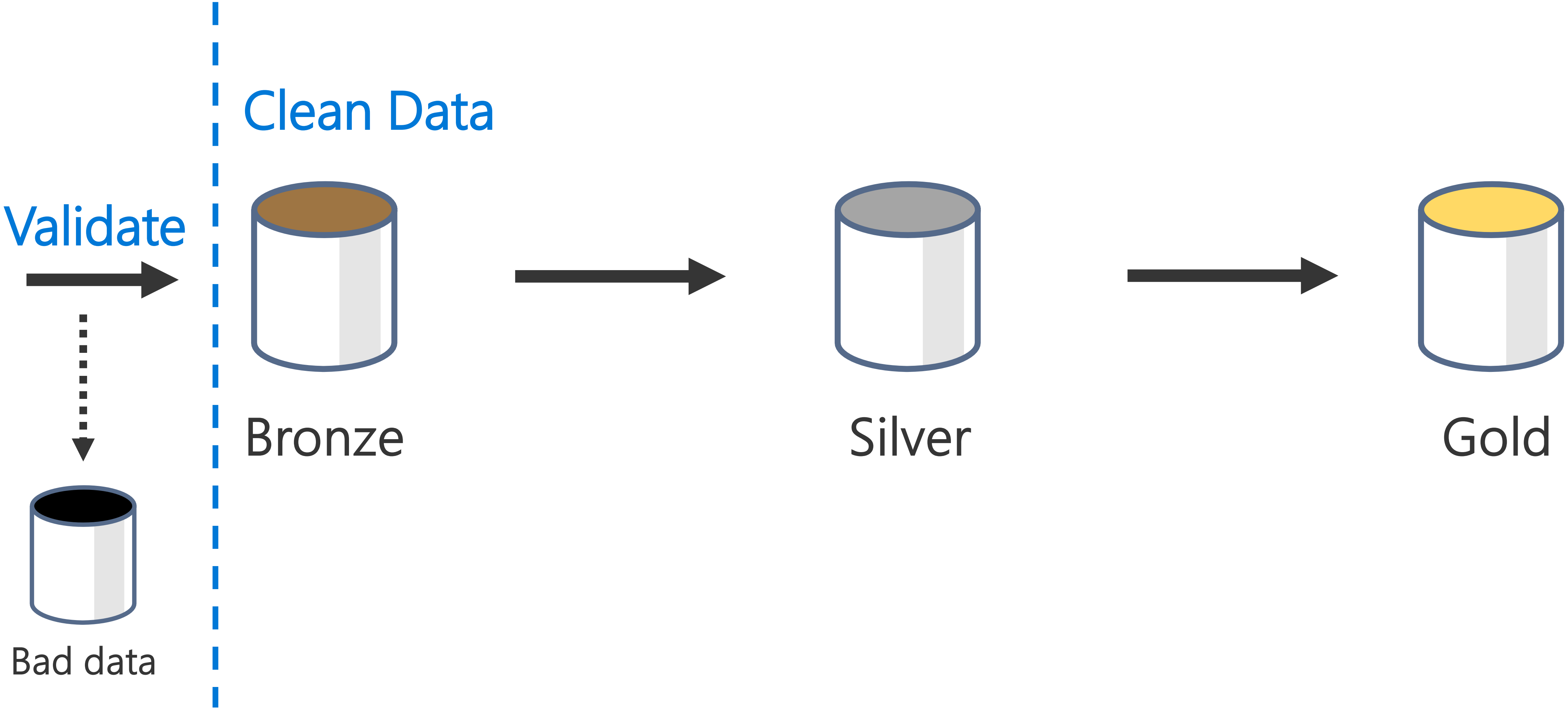
Validate data early in the Pipeline

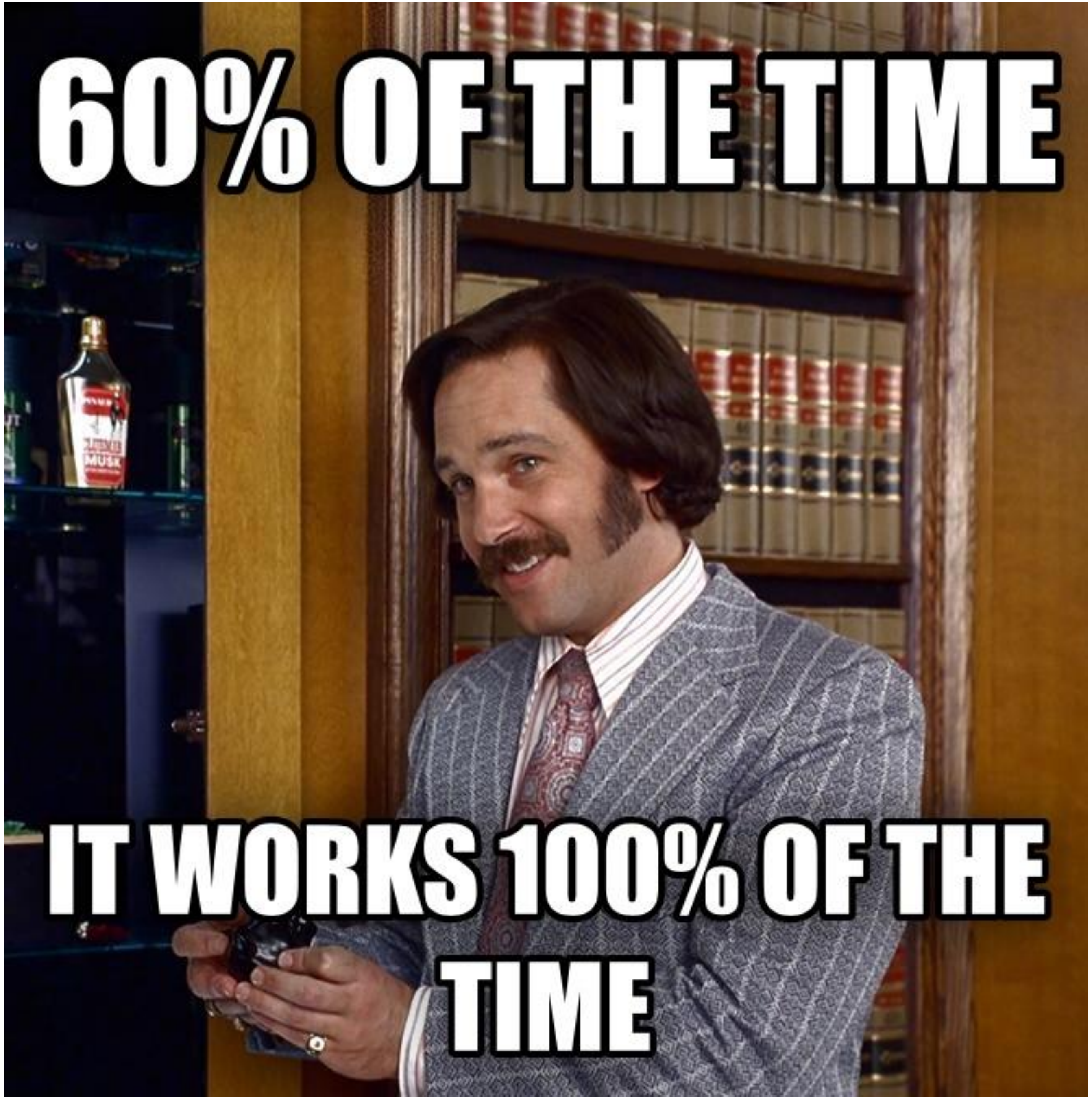


Validate data early in the Pipeline

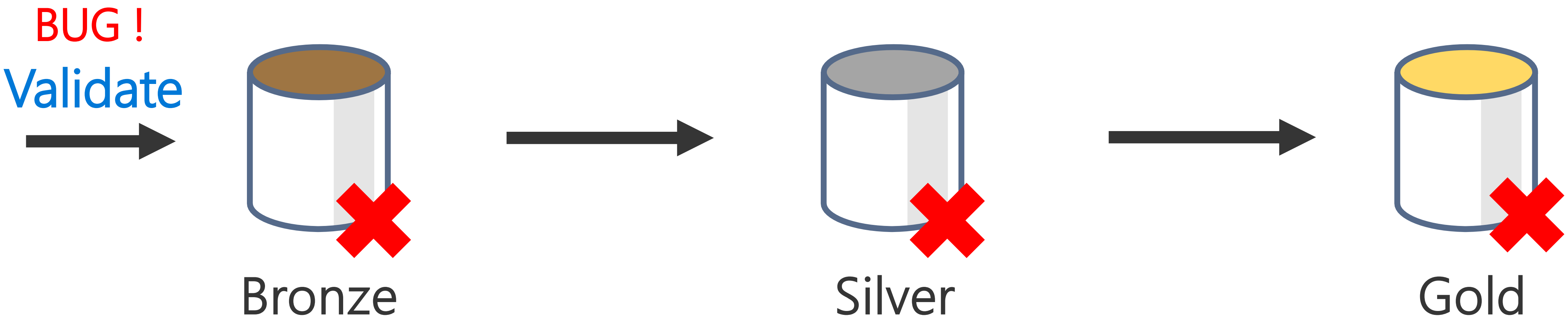


Why not here?




60% OF THE TIME
IT WORKS 100% OF THE TIME

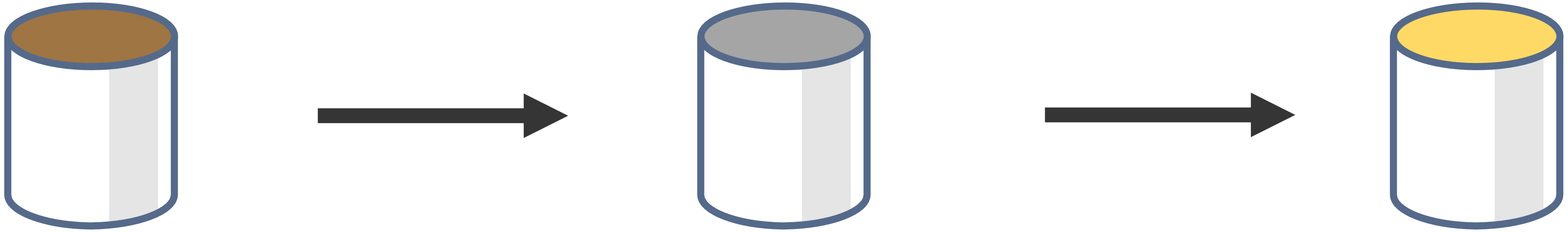
Why not here?.. Because code is not perfect



Learnings

Ensure data pipeline is replayable.

Data Tiers - Replayability



Bronze

Append only,
immutable

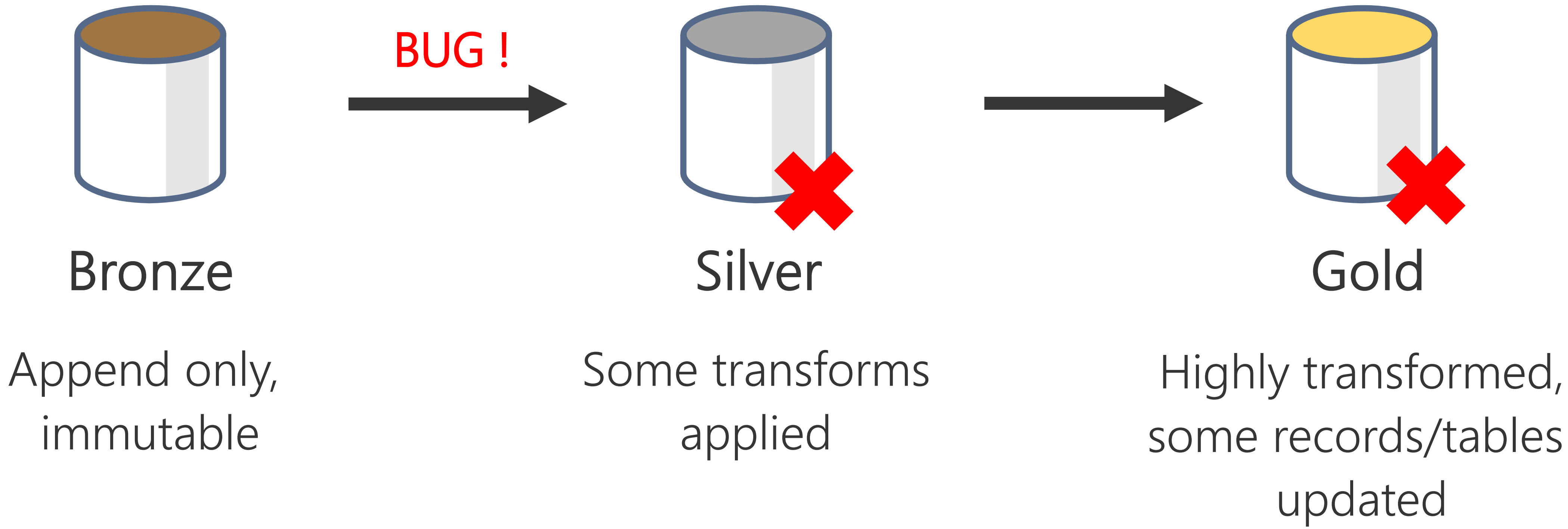
Silver

Some transforms
applied

Gold

Highly transformed,
some records/tables
updated

Data Tiers - Replayability



BUG !

Bronze

Append only, immutable

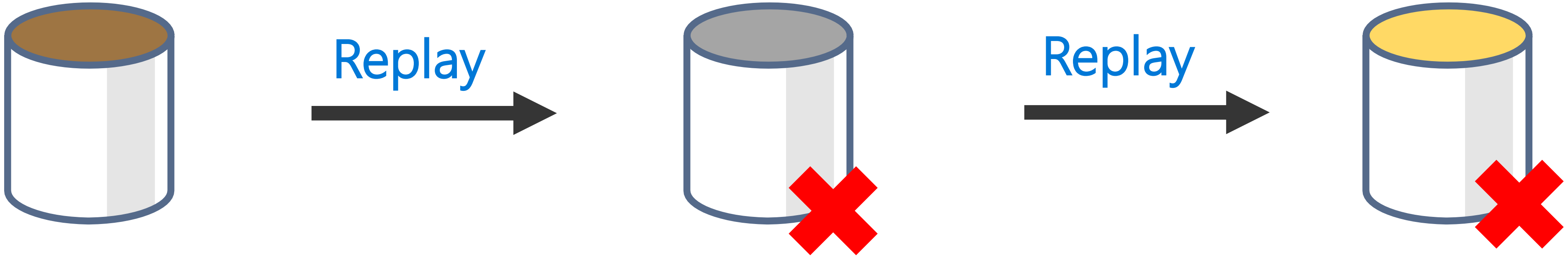
Silver

Some transforms applied

Gold

Highly transformed, some records/tables updated

Data Tiers - Replayability



Bronze

Append only,
immutable

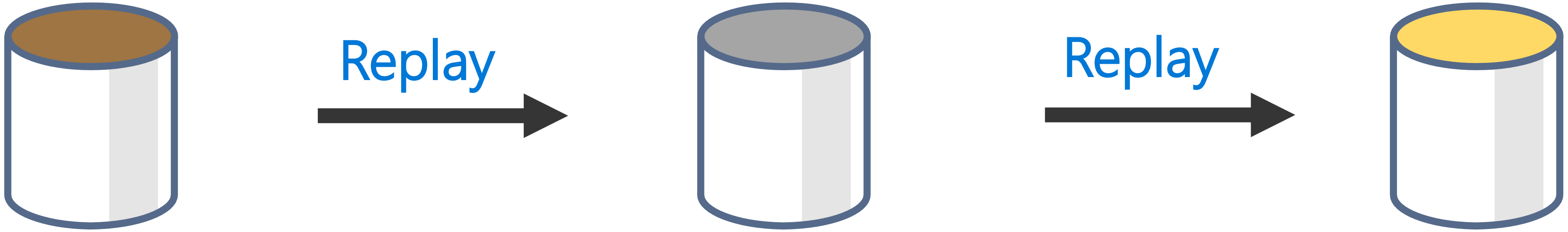
Silver

Some transforms
applied

Gold

Highly transformed,
some records/tables
updated

Data Tiers - Replayability



Replay

Replay

Bronze

Append only, immutable

Silver

Some transforms applied

Gold

Highly transformed, some records/tables updated

Melbourne Parking Data

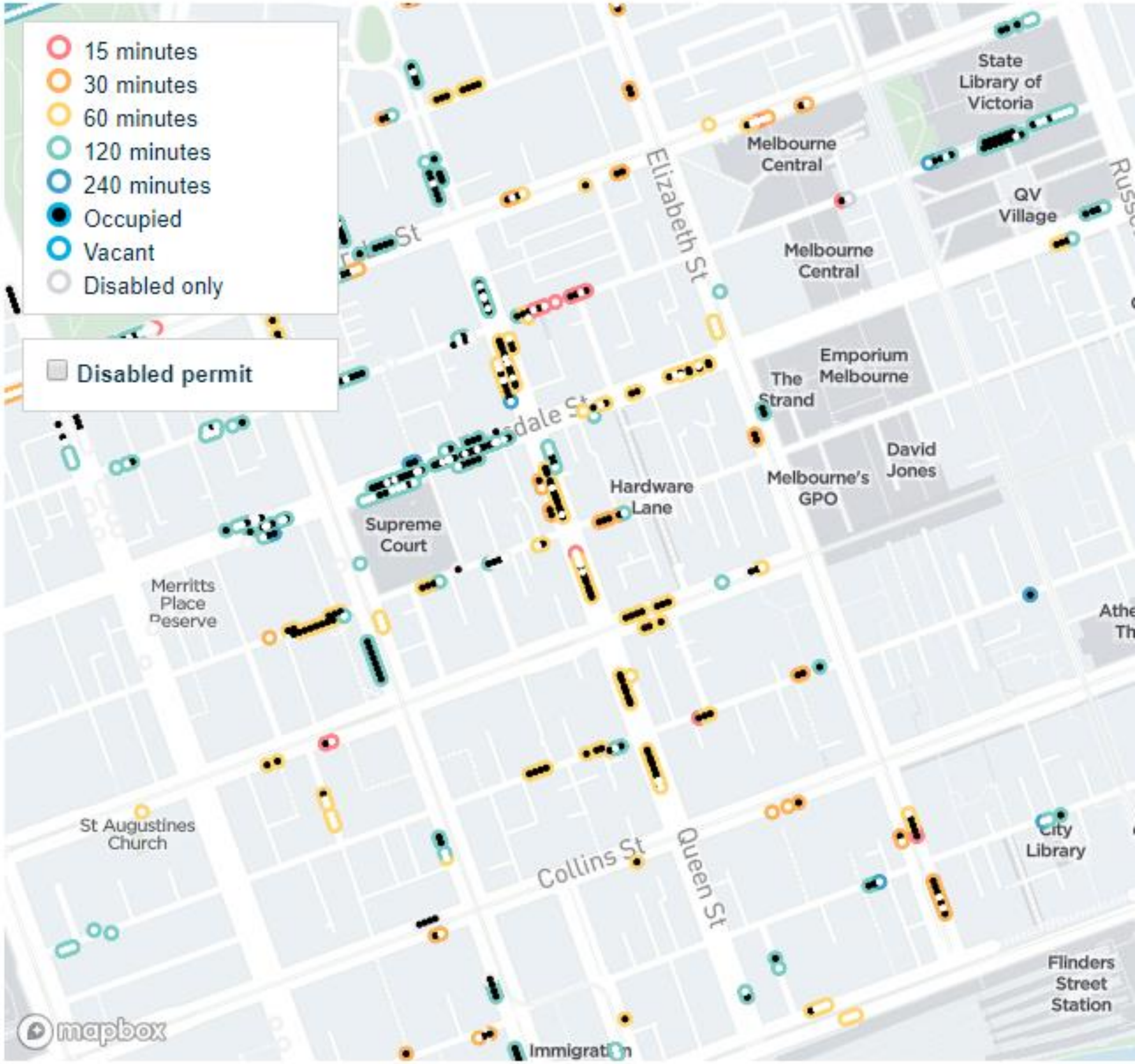
4300 in-ground sensors in our on-street parking bays available through Melbourne Open Data Platform.

Public API available.

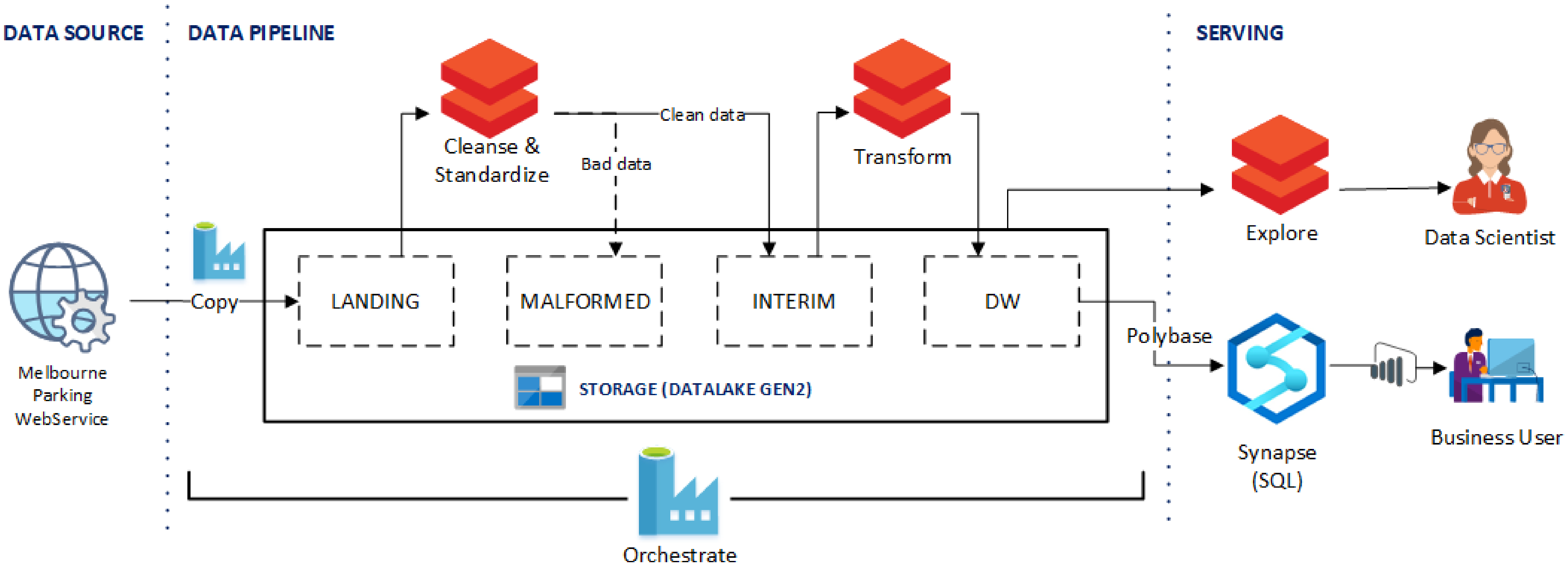
Source:

https://www.melbourne.vic.gov.au/about-council/governance-transparency/open-data/Pages/on-street-parking-data.aspx

Map of on-street parking data



Demo Pipeline



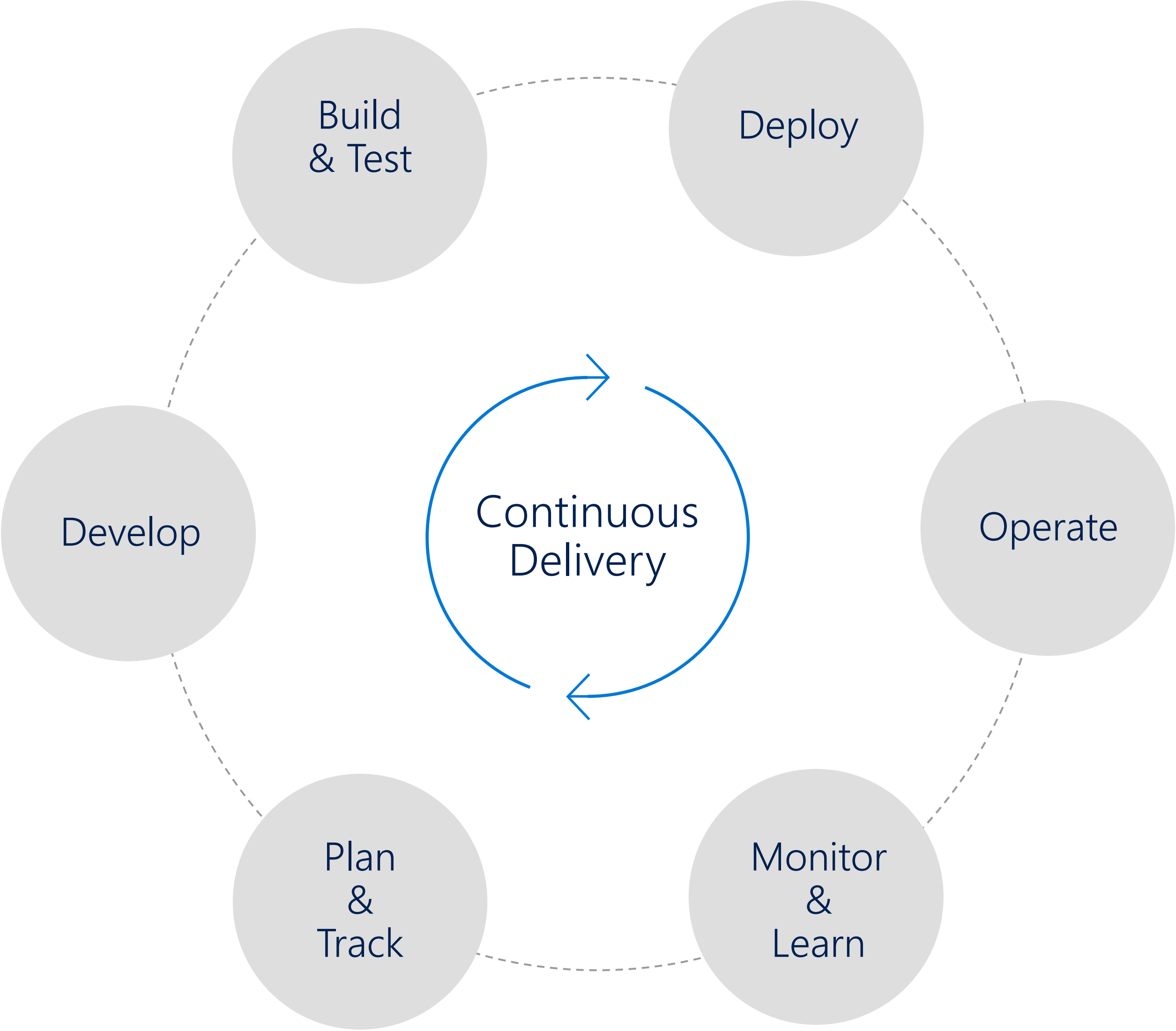
How do you operationalize?

#RubDevOpsOnIt

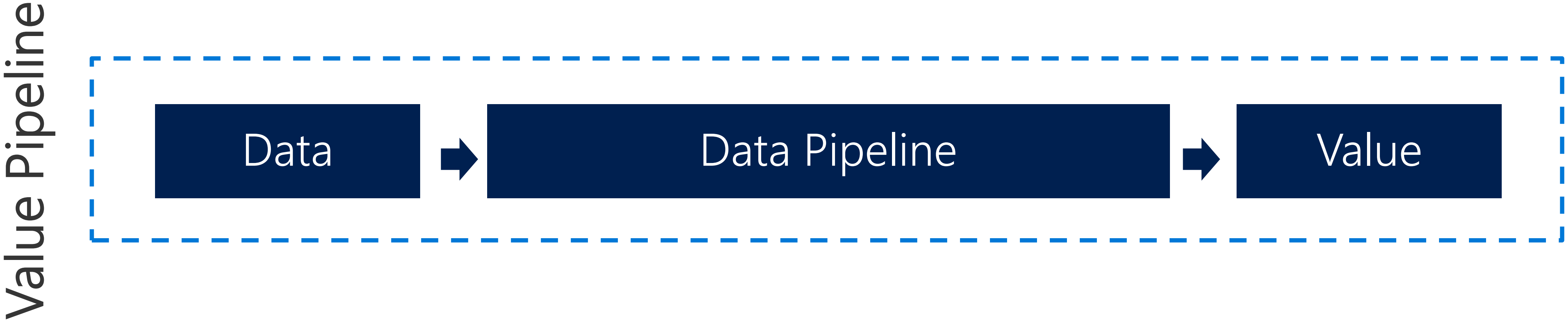
What is DevOps?

- People. Process. Products.

> “DevOps is the union of **people**, **process**, and **products** to enable continuous delivery of value to your end users.”

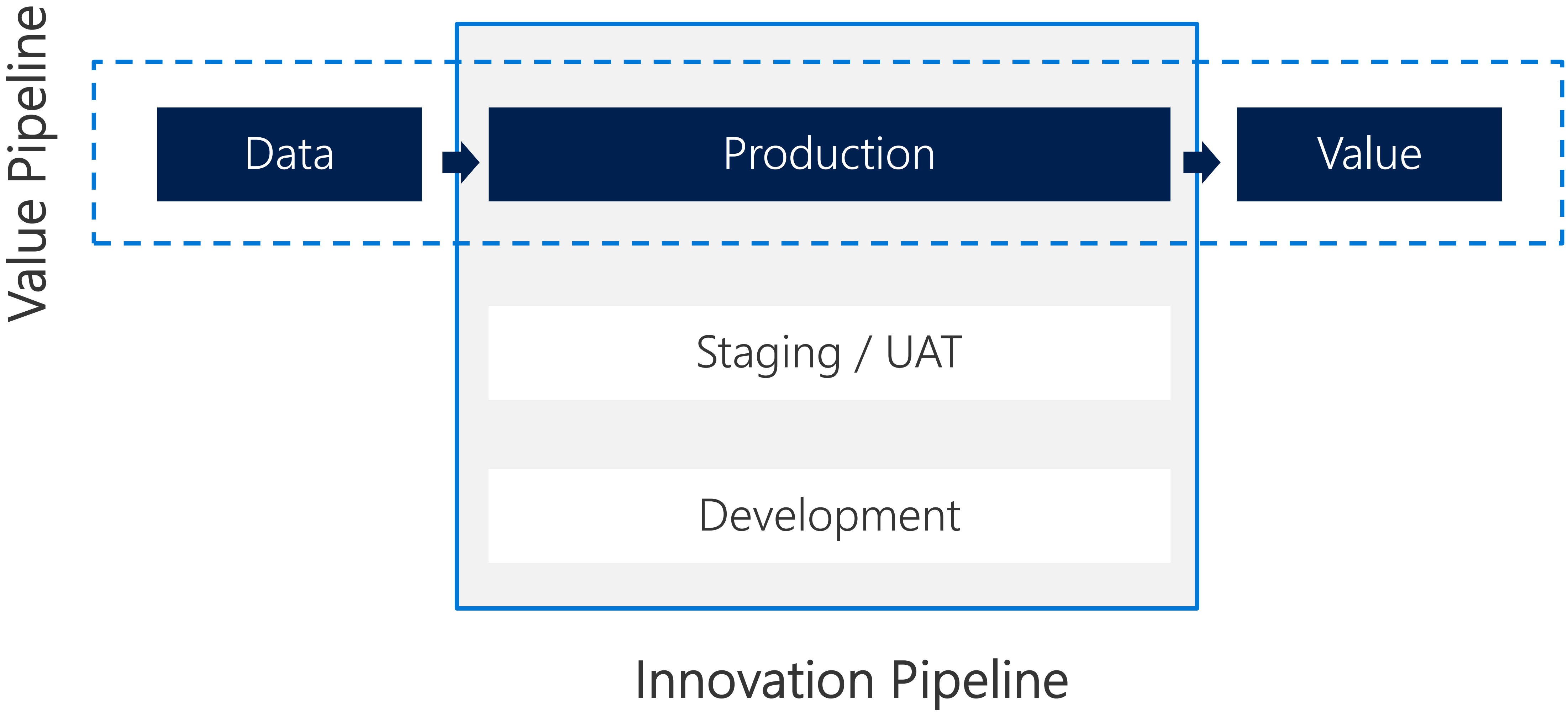


DataOps



DataOps is NOT Just DevOps for Data, Data Kitchen

DataOps

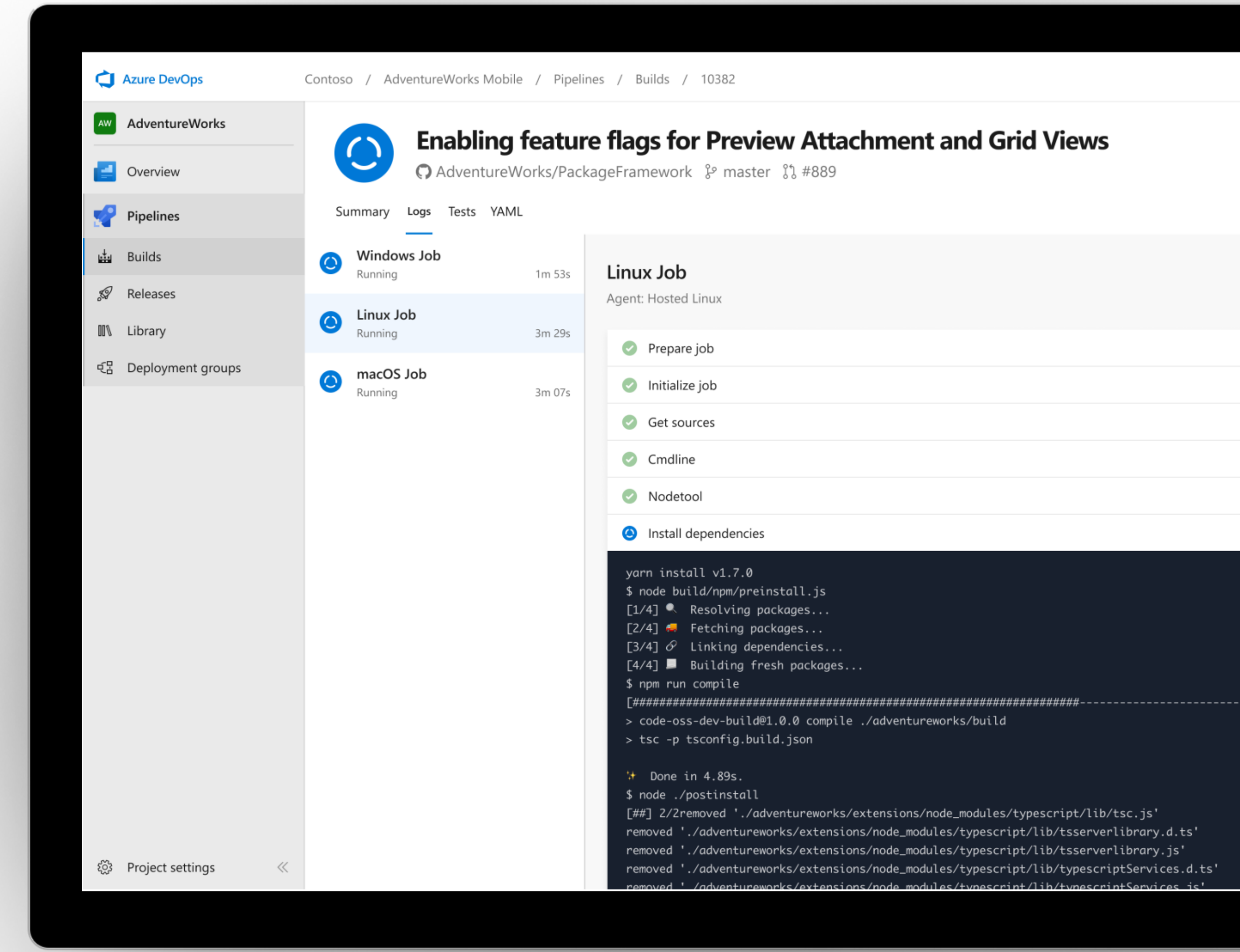


DataOps is NOT Just DevOps for Data, Data Kitchen

Azure Pipelines

Cloud-hosted pipelines for Linux, Windows and macOS, with unlimited minutes for open source

- Multi language, platform, and cloud support
- Extensible
- Containers and Kubernetes
- Best-in-class for open source



Enabling feature flags for Preview Attachment and Grid Views

AdventureWorks/PackageFramework master #889

Windows Job	Running	1m 53s
Linux Job	Running	3m 29s
macOS Job	Running	3m 07s

Linux Job

Agent: Hosted Linux

- ✓ Prepare job
- ✓ Initialize job
- ✓ Get sources
- ✓ Cmdline
- ✓ Nodetool
- Install dependencies

```
yarn install v1.7.0
$ node build/npm/preinstall.js
[1/4] 🔍  Resolving packages...
[2/4] 🚚  Fetching packages...
[3/4] 🔗  Linking dependencies...
[4/4] 📃  Building fresh packages...
$ npm run compile
[#######################################################################-------------------------
> code-oss-dev-build@1.0.0 compile ./adventureworks/build
> tsc -p tsconfig.build.json

✨  Done in 4.89s.
$ node ./postinstall
[##] 2/2removed './adventureworks/extensions/node_modules/typescript/lib/tsc.js'
removed './adventureworks/extensions/node_modules/typescript/lib/tsserverlibrary.d.ts'
removed './adventureworks/extensions/node_modules/typescript/lib/tsserverlibrary.js'
removed './adventureworks/extensions/node_modules/typescript/lib/typescriptServices.d.ts'
removed './adventureworks/extensions/node_modules/typescript/lib/typescriptServices.js'
```

https://azure.com/pipelines

Azure DevOps

Azure Boards

Deliver value to your users faster using proven agile tools to plan, track, and discuss work across your teams.

Azure Pipelines

Build, test, and deploy with CI/CD that works with any language, platform, and cloud. Connect to GitHub or any other Git provider and deploy continuously.

Azure Repos

Get unlimited, cloud-hosted private Git repos and collaborate to build better code with pull requests and advanced file management.

Azure Test Plans

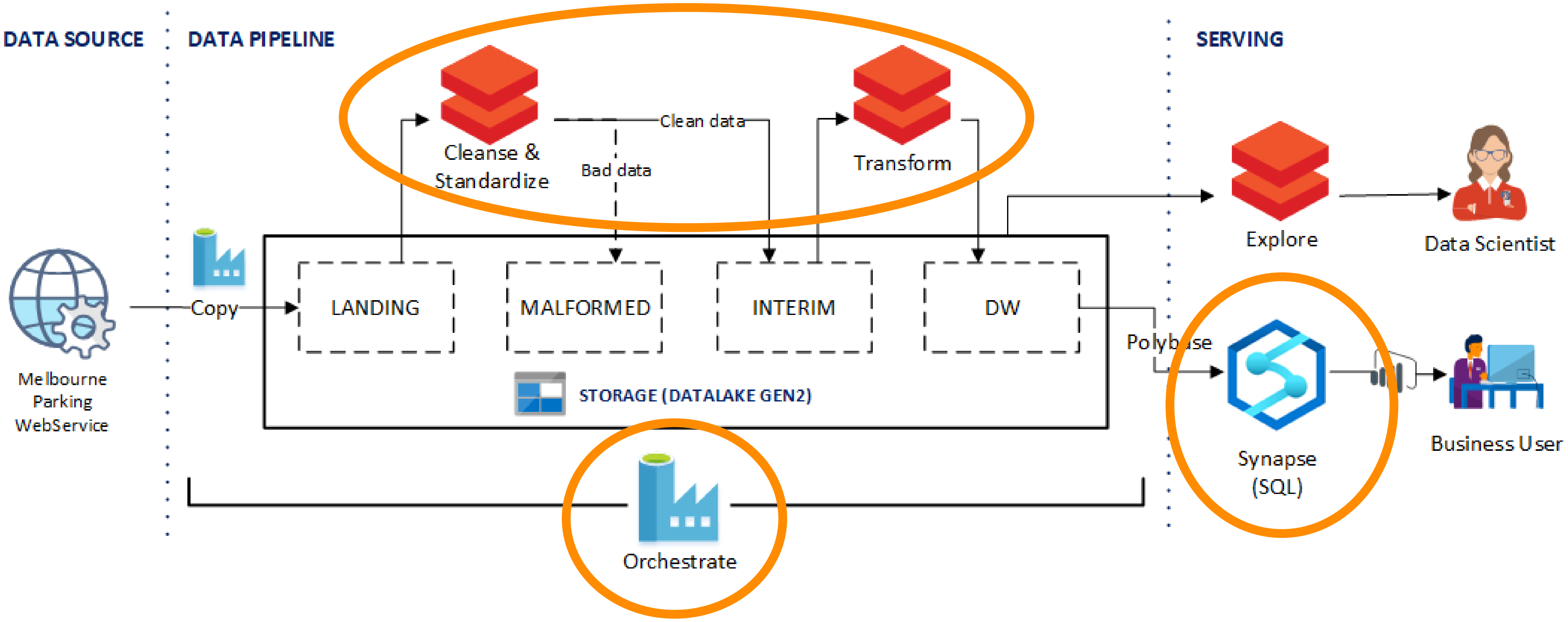
Test and ship with confidence using manual and exploratory testing tools.

Azure Artifacts

Create, host, and share packages with your team, and add artifacts to your CI/CD pipelines with a single click.

https://azure.com/devops

Demo Value Pipeline: Modern Data Warehouse on Azure



Azure Databricks

Data transformation code belong in packages, not notebooks

Notebooks are a ‘light-wrapper’ around packages (whl, jar, etc)

Higher-quality, reusable, testable code

Run unit tests ‘out-of-workspace’

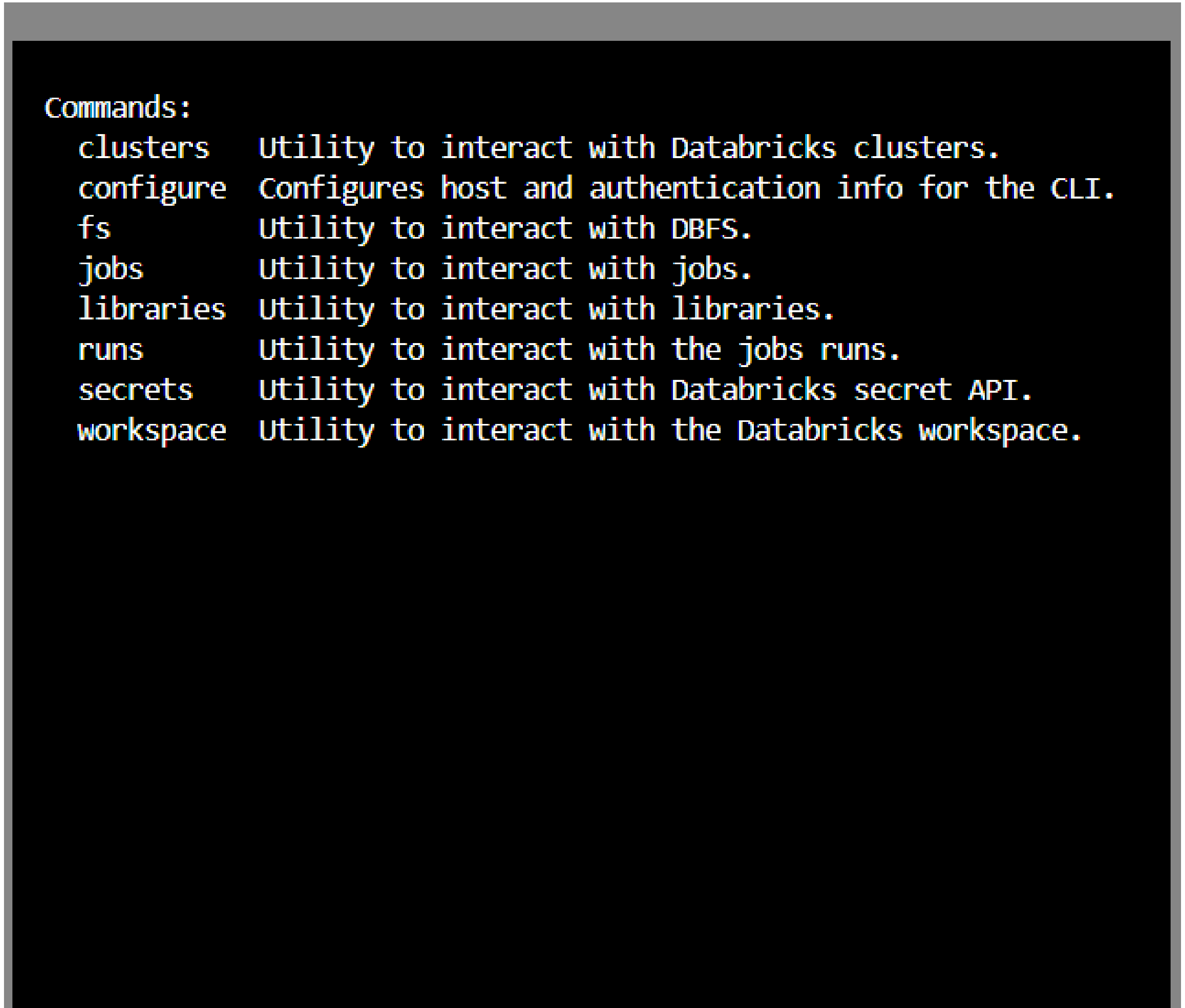
Faster development feedback cycle

Azure Databricks

Version control notebooks

- Integration with Github, Azure DevOps, BitBucket

Use Databricks CLI / REST APIs to automate work.



```
Commands:
  clusters   Utility to interact with Databricks clusters.
  configure  Configures host and authentication info for the CLI.
  fs         Utility to interact with DBFS.
  jobs       Utility to interact with jobs.
  libraries  Utility to interact with libraries.
  runs       Utility to interact with the jobs runs.
  secrets    Utility to interact with Databricks secret API.
  workspace  Utility to interact with the Databricks workspace.
```

SQL Server Data Tools (SSDT)

Keep database objects/schema definitions in source control.

Build and validate deployable **DACPAC**.

Schema-compare to auto-detect changes.

- Generate T-SQL script w/ incremental changes and publish changes to server

Azure SQL Data Warehouse support (Preview)

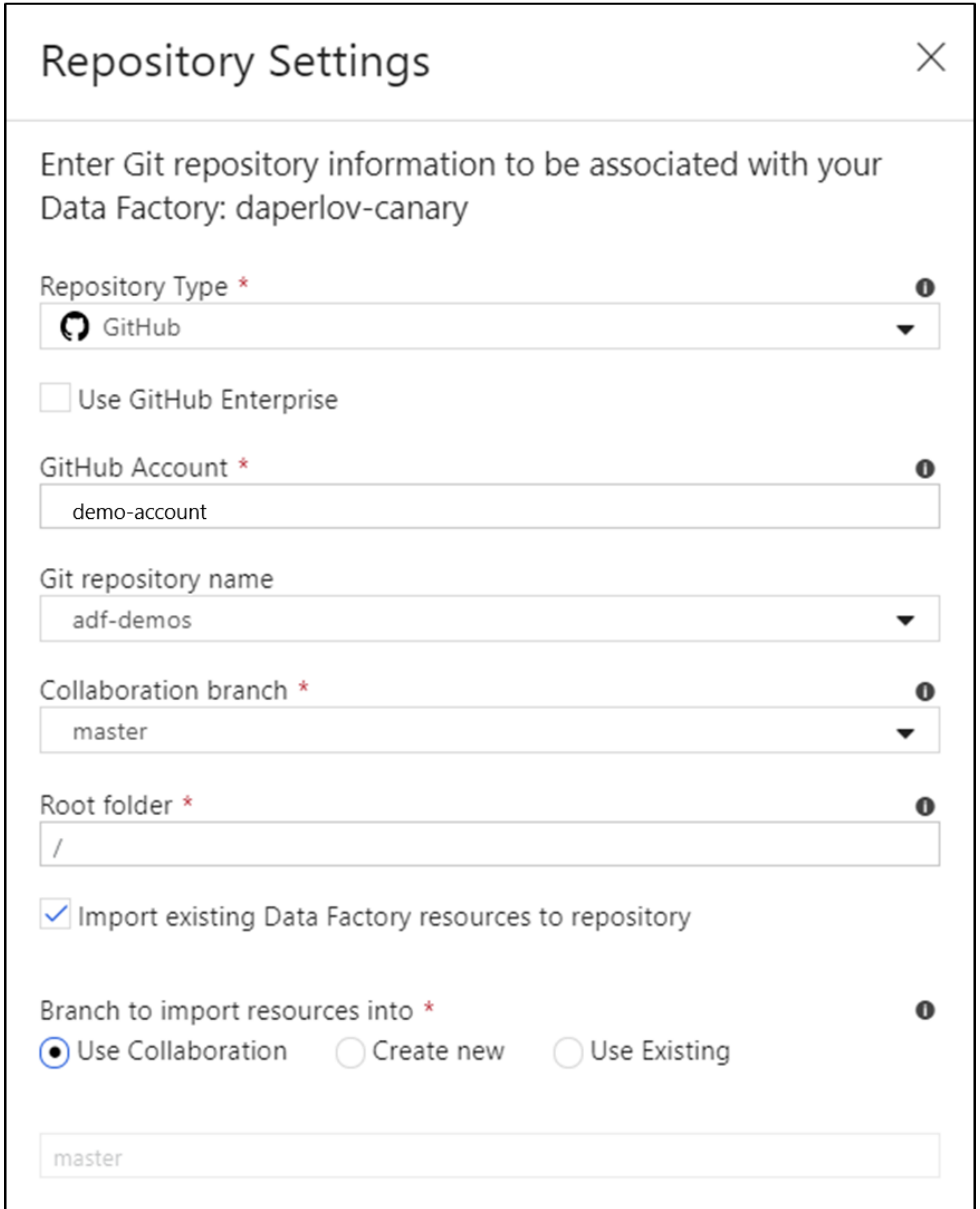
CLI - sqlpackage.exe

```
SqlPackage.exe
/Action:Publish
/SourceFile:ProjectName.dacpac
/TargetServerName:Server\Instance
/TargetDatabaseName:DBName
/Variables:Foo=Bar
```

Azure Data Factory

Ability to export pipeline definitions as ARM templates.

Integration with Azure DevOps and Github.



Repository Settings

Enter Git repository information to be associated with your Data Factory: daperlov-canary

Repository Type *
GitHub

□ Use GitHub Enterprise

GitHub Account *
demo-account

Git repository name
adf-demos

Collaboration branch *
master

Root folder *
/

☑ Import existing Data Factory resources to repository

Branch to import resources into *
● Use Collaboration ○ Create new ○ Use Existing

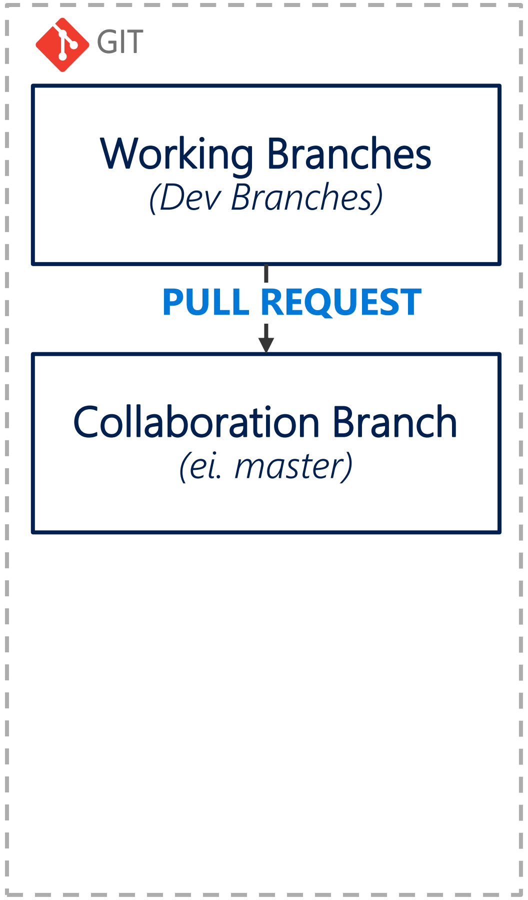
master

Azure Data Factory – CI / CD

GIT

Working Branches
(Dev Branches)

Azure Data Factory – CI / CD



GIT

Working Branches

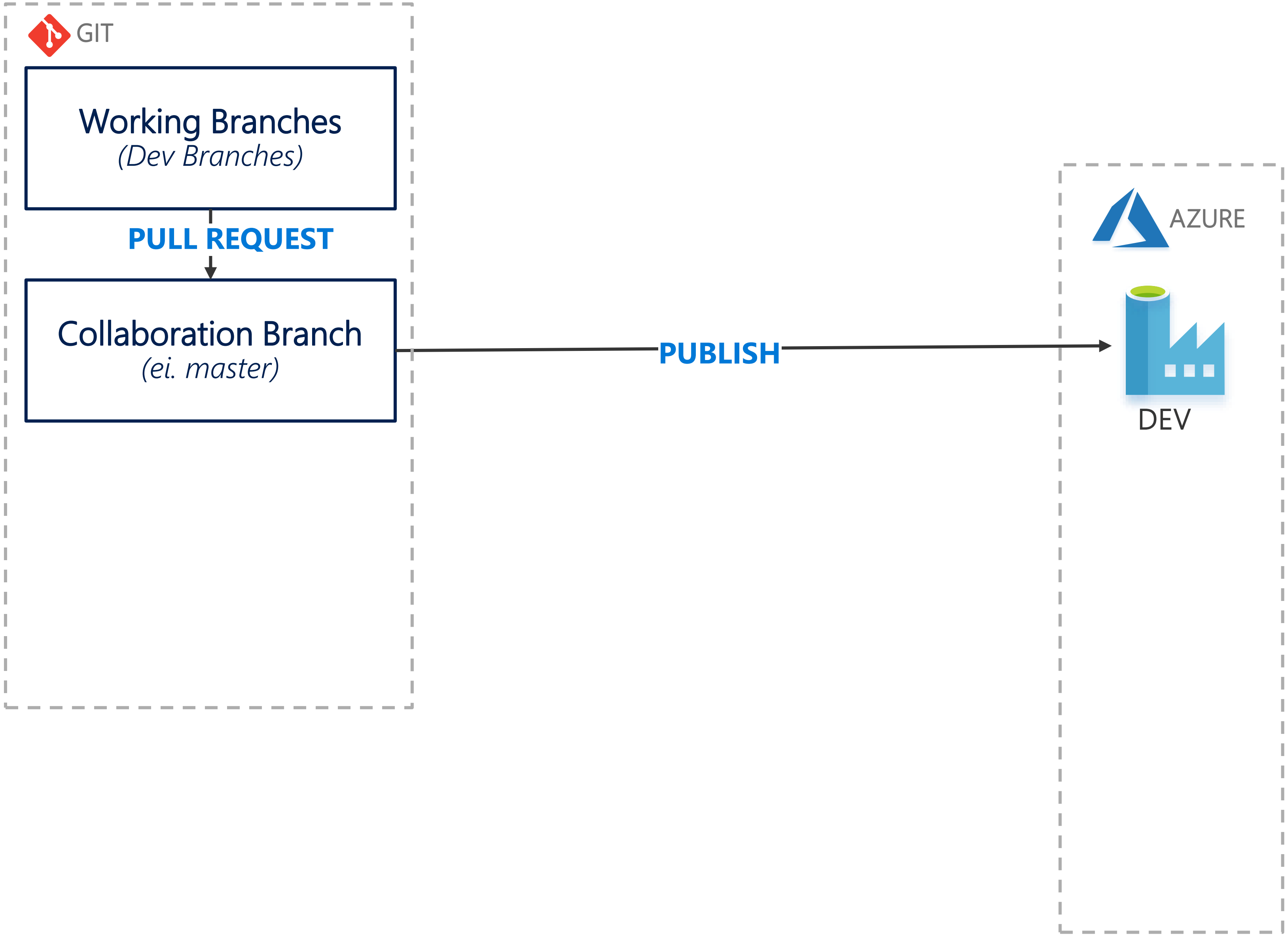
(Dev Branches)

PULL REQUEST

Collaboration Branch

(ei. master)

Azure Data Factory – CI / CD



GIT

Working Branches
(Dev Branches)

PULL REQUEST

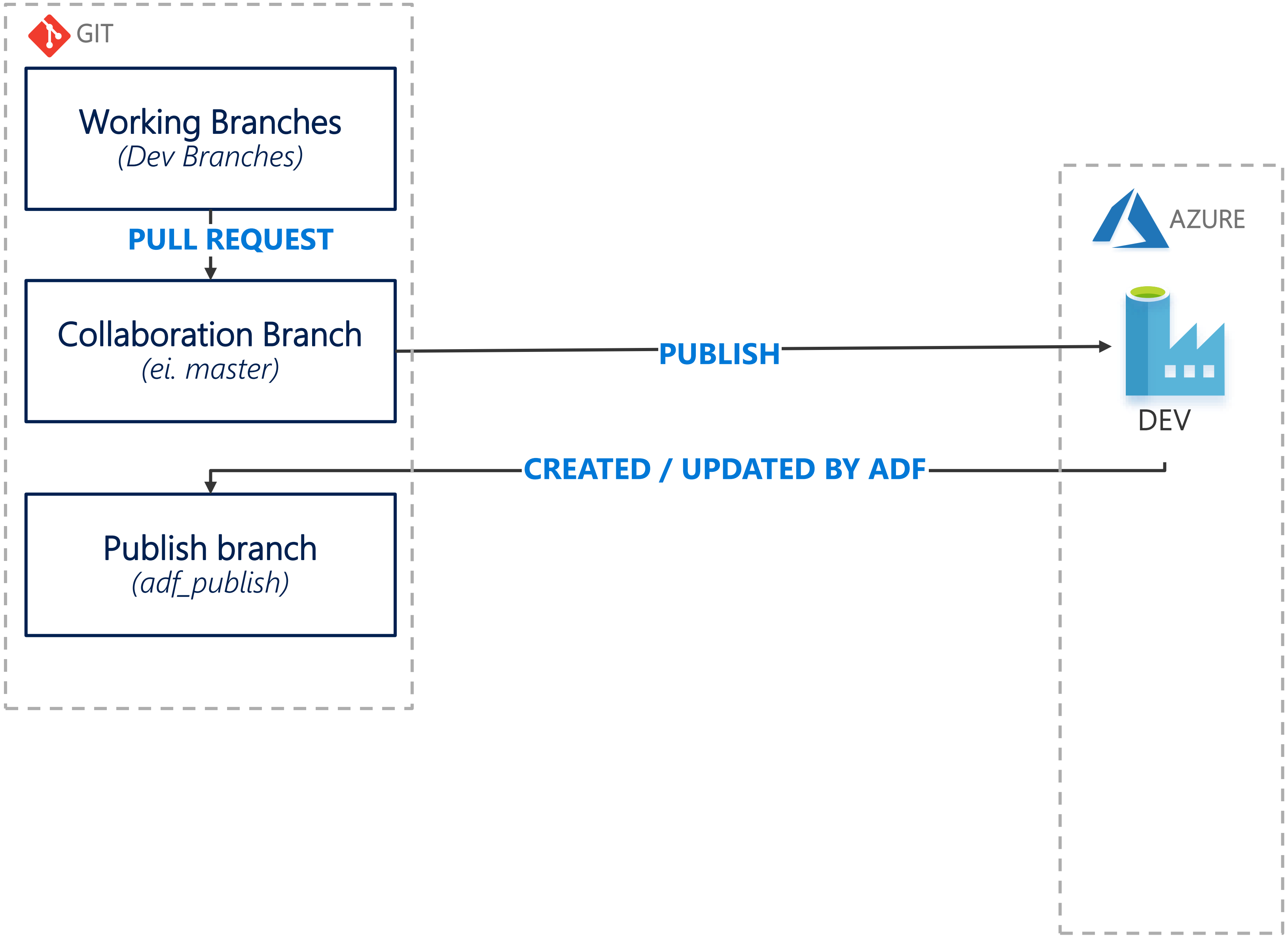
Collaboration Branch
(ei. master)

PUBLISH

AZURE

DEV

Azure Data Factory – CI / CD



GIT

Working Branches
(Dev Branches)

PULL REQUEST

Collaboration Branch
(ei. master)

PUBLISH

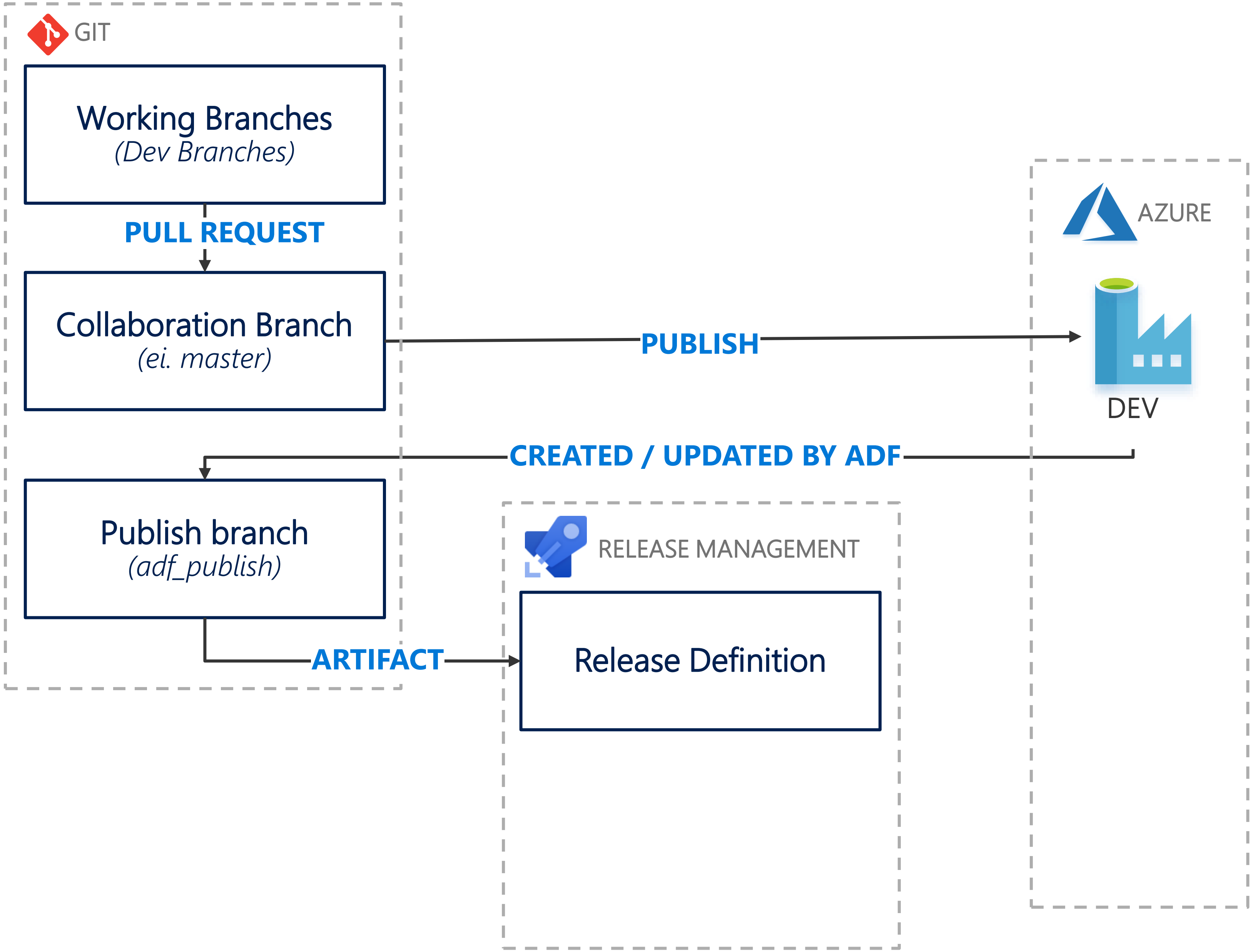
AZURE

DEV

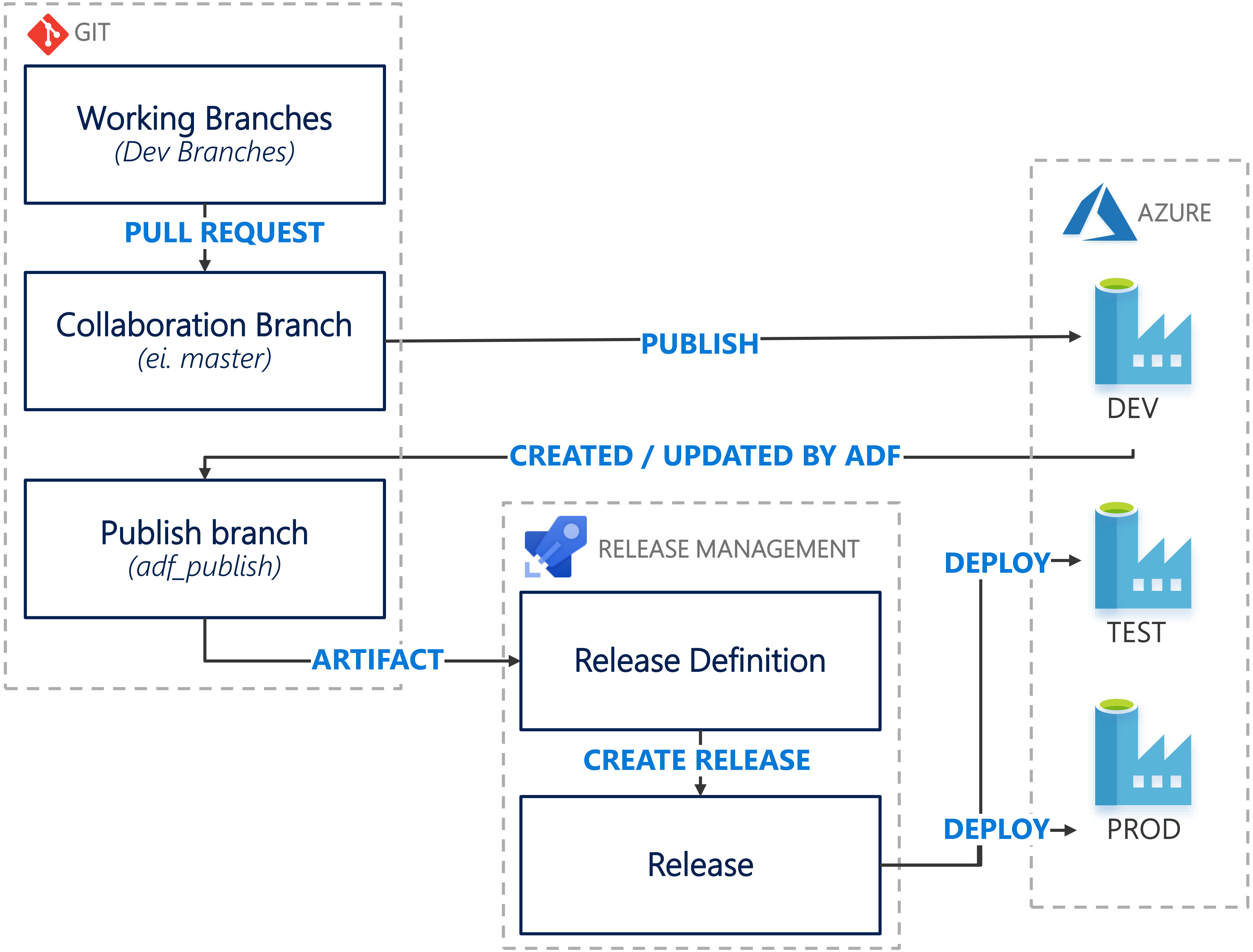
CREATED / UPDATED BY ADF

Publish branch
(adf_publish)

Azure Data Factory – CI / CD



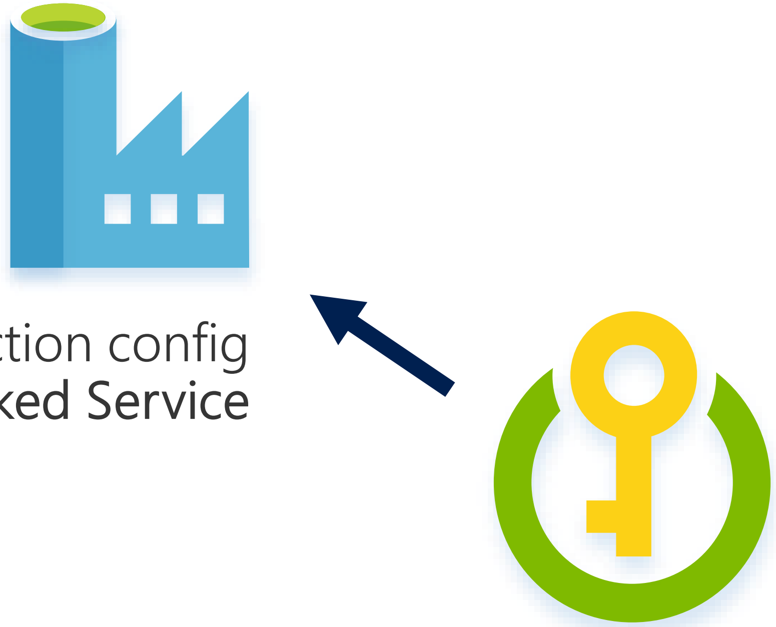
Azure Data Factory – CI / CD



Learnings

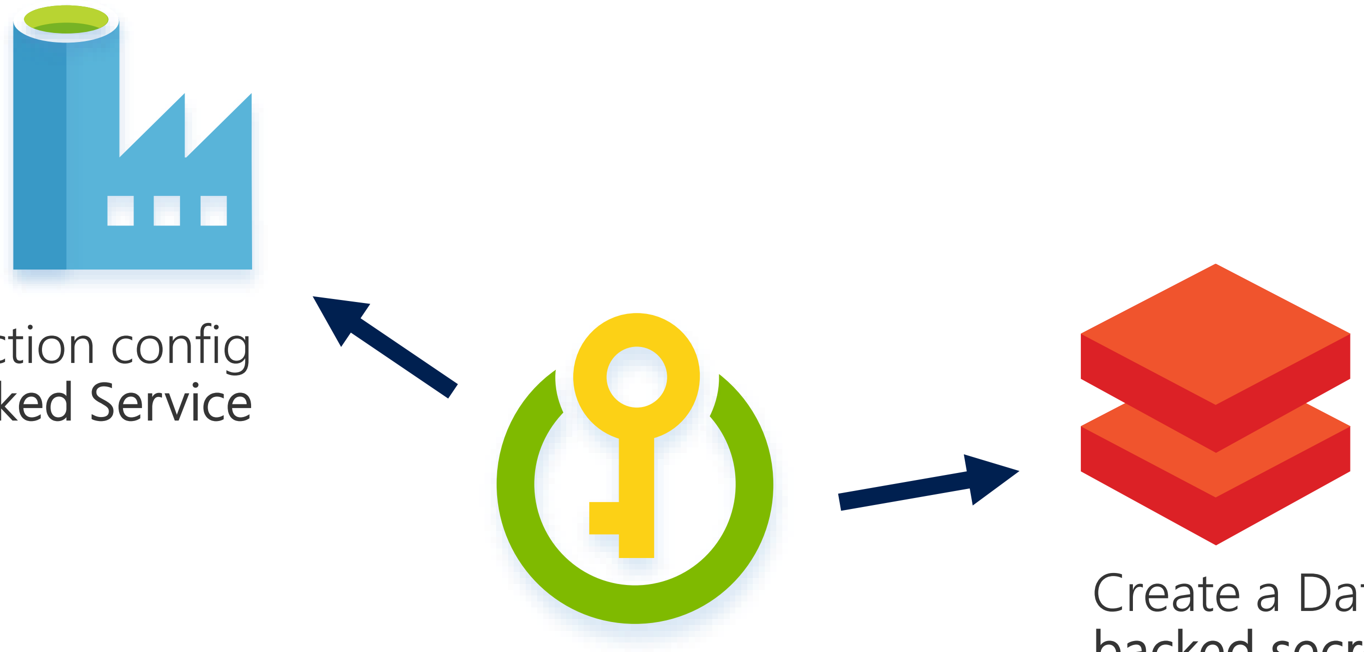
Secure and Centralize Configuration.

Secure and Centralize Configuration in Azure KeyVault



Pull sensitive connection config
w/ **KeyVault Linked Service**

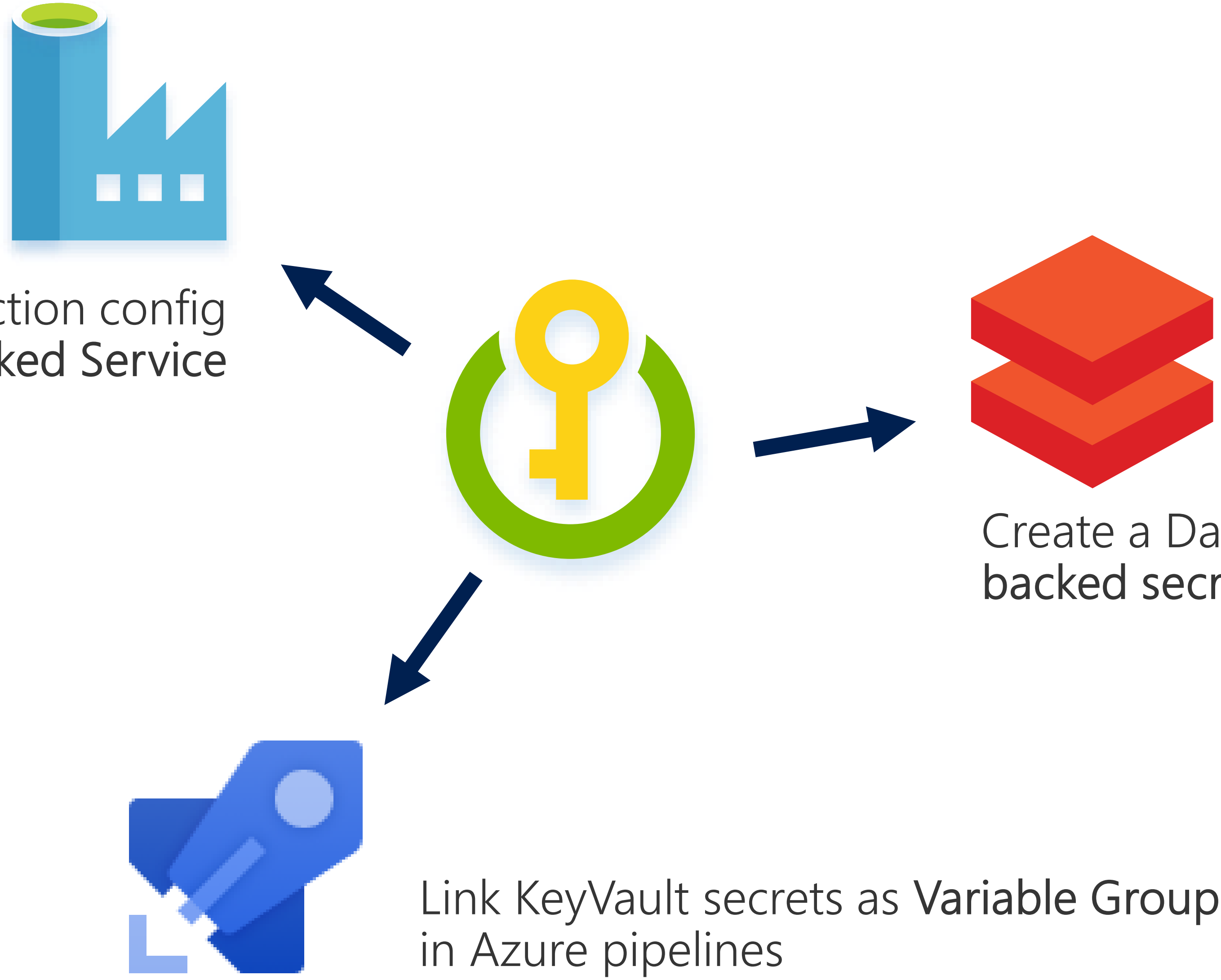
Secure and Centralize Configuration in Azure KeyVault



Pull sensitive connection config
w/ **KeyVault Linked Service**

Create a Databricks **KeyVault-backed secret scope** (PREVIEW)

Secure and Centralize Configuration in Azure KeyVault

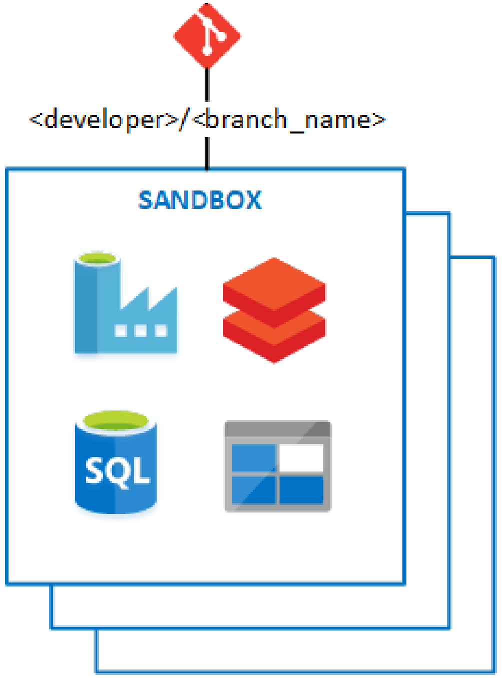


Pull sensitive connection config w/ **KeyVault Linked Service**

Create a Databricks **KeyVault-backed secret scope** (PREVIEW)

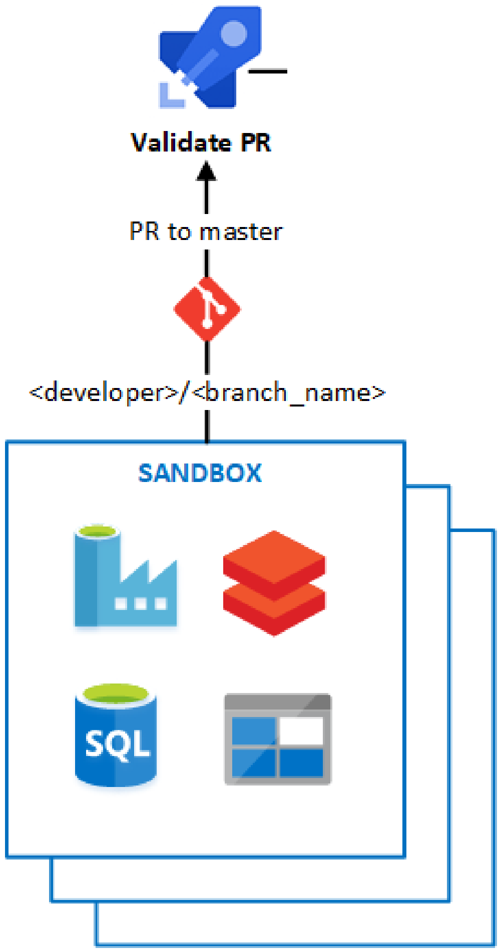
Link KeyVault secrets as **Variable Groups** in Azure pipelines

Innovation Pipeline (CI/CD)

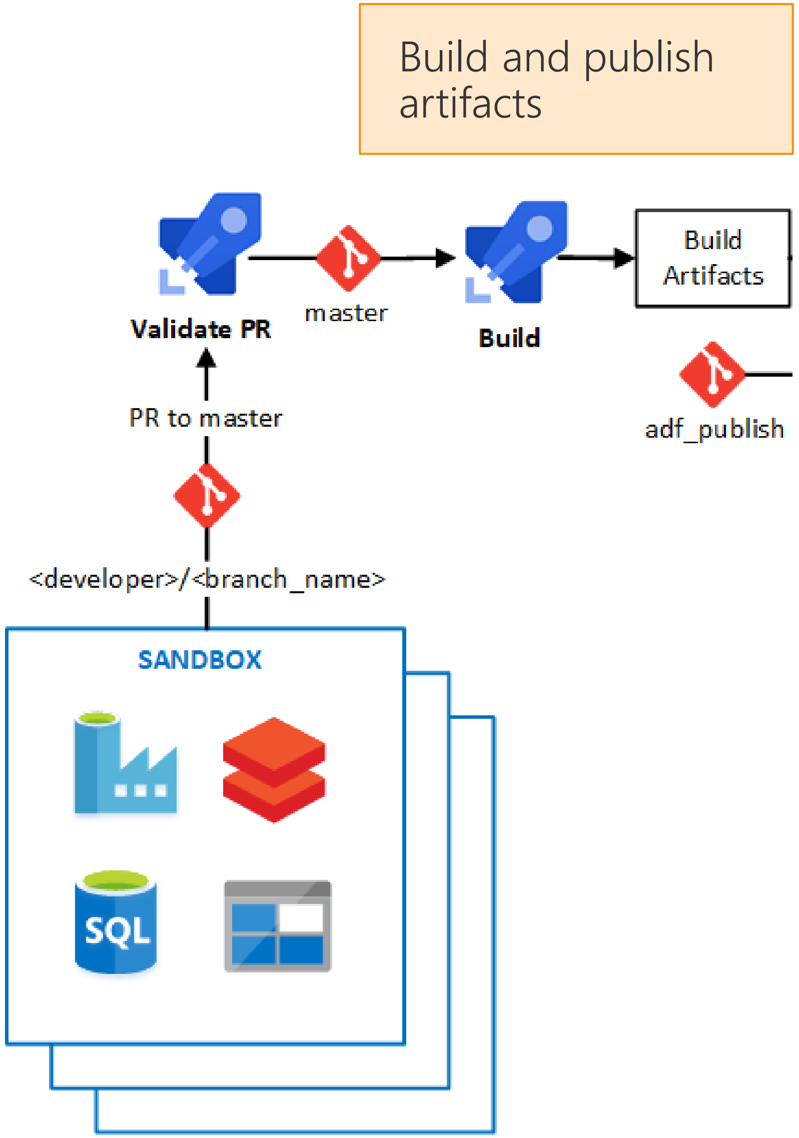


Innovation Pipeline (CI/CD)

Run unit tests,
linting,
DACPAC build

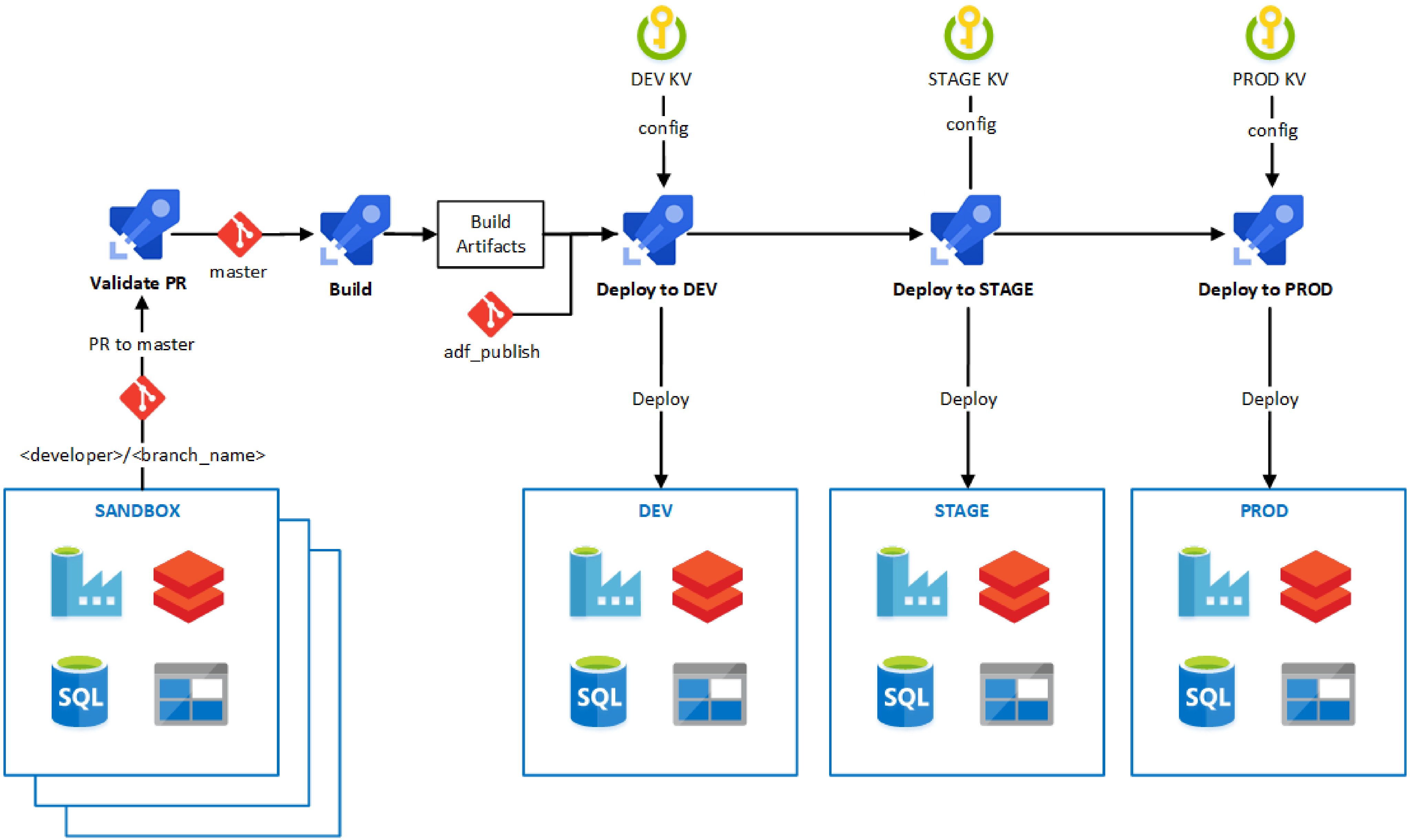


Innovation Pipeline (CI/CD)

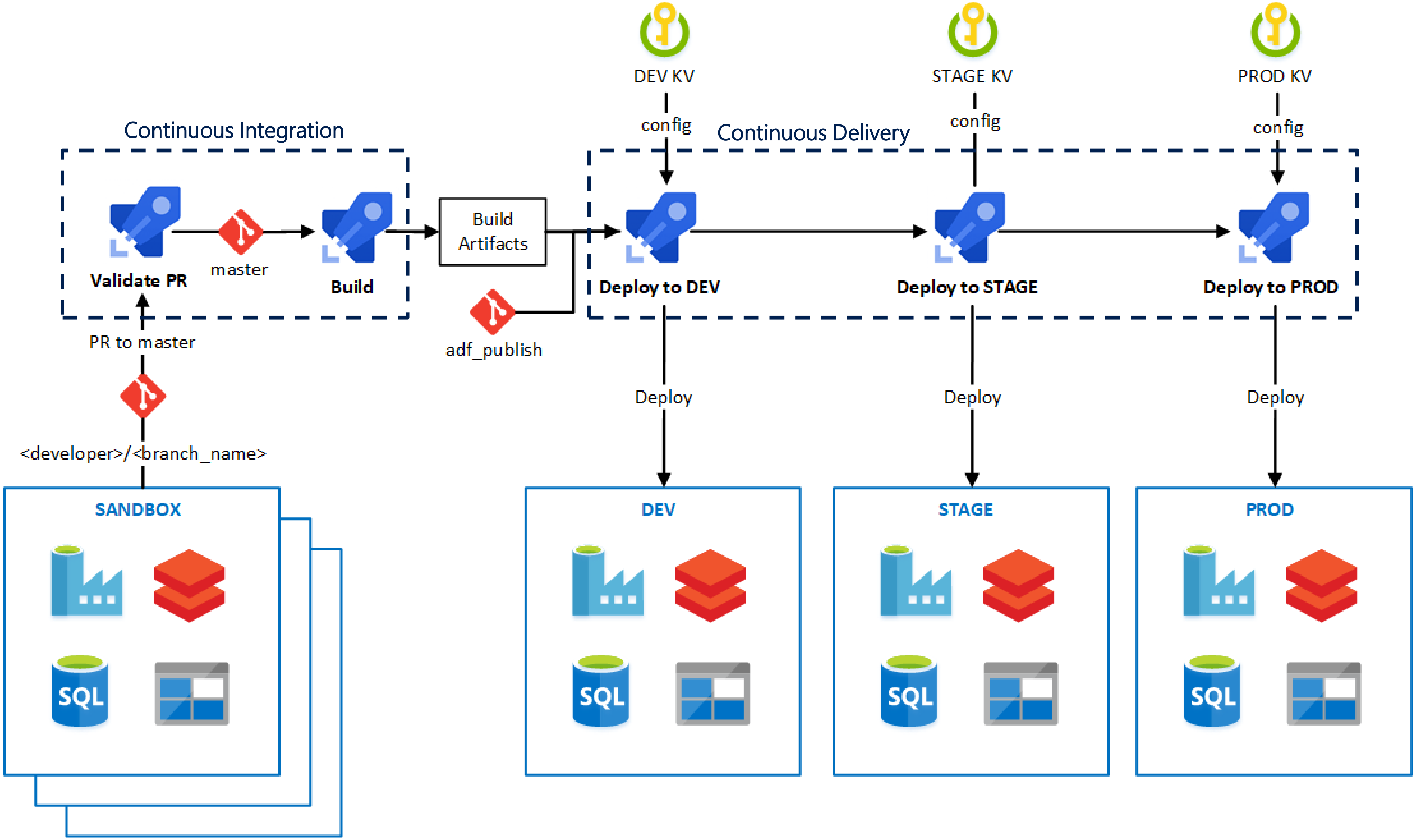


Innovation Pipeline (CI/CD)

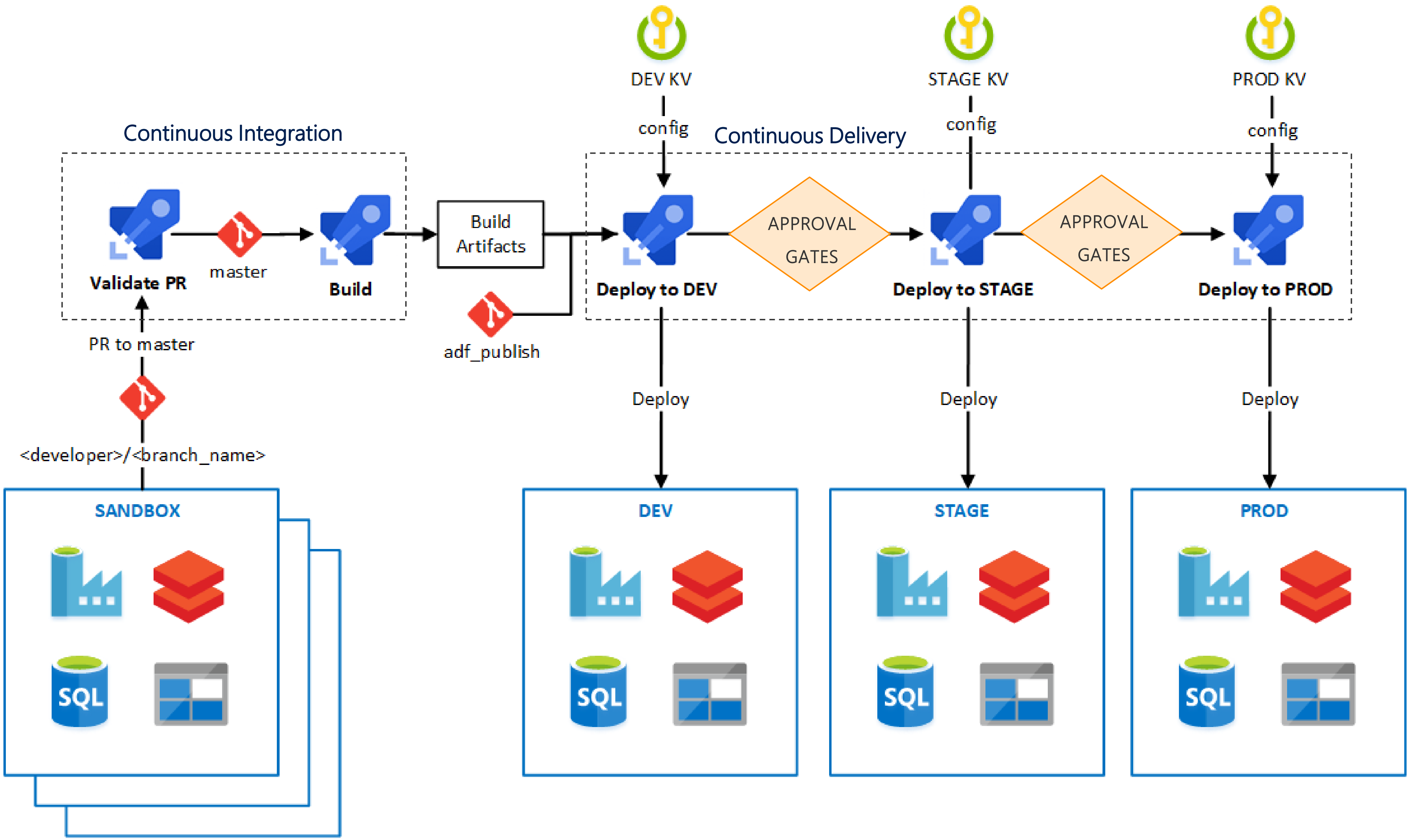
Deploy across environments



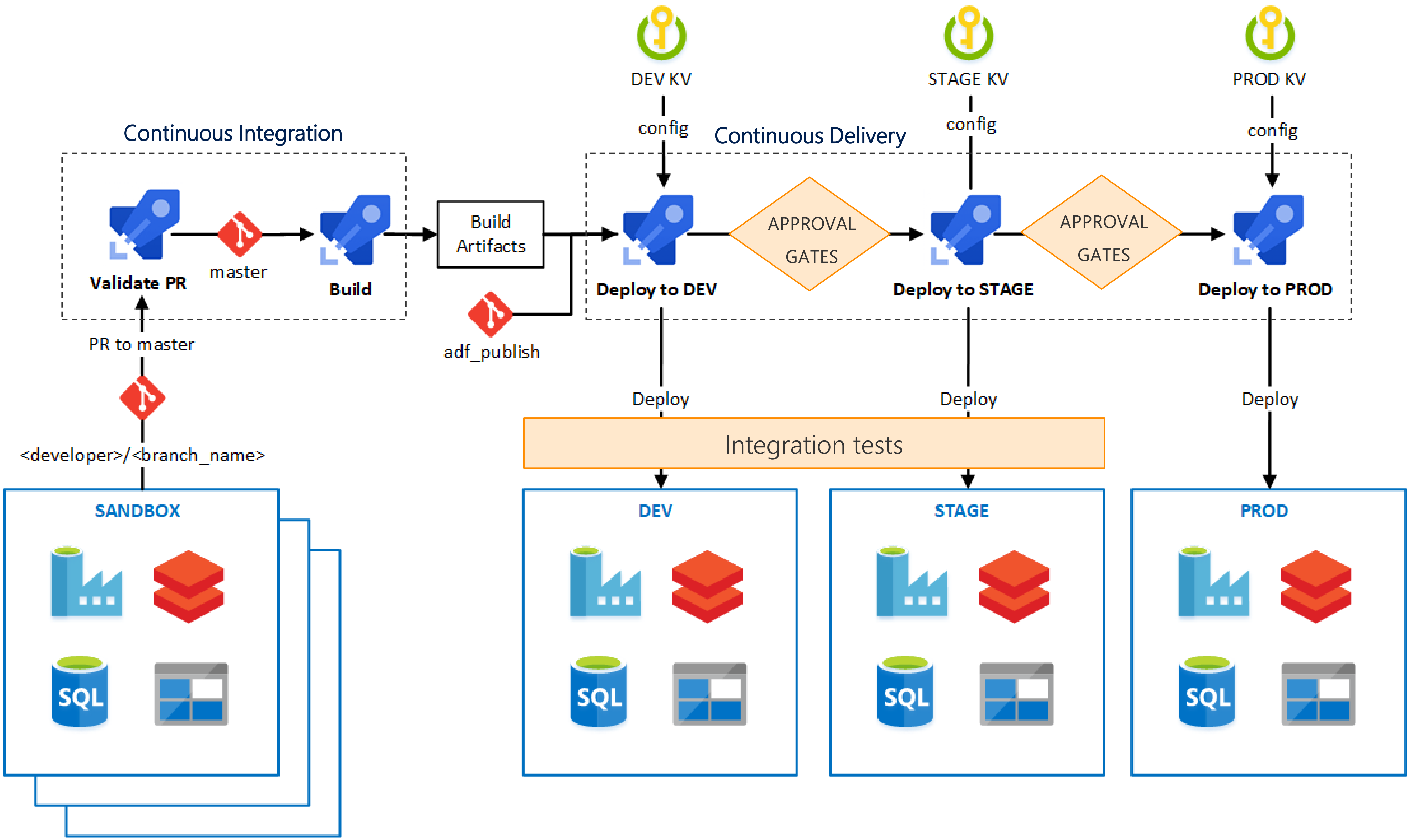
Innovation Pipeline (CI/CD)



Innovation Pipeline (CI/CD)

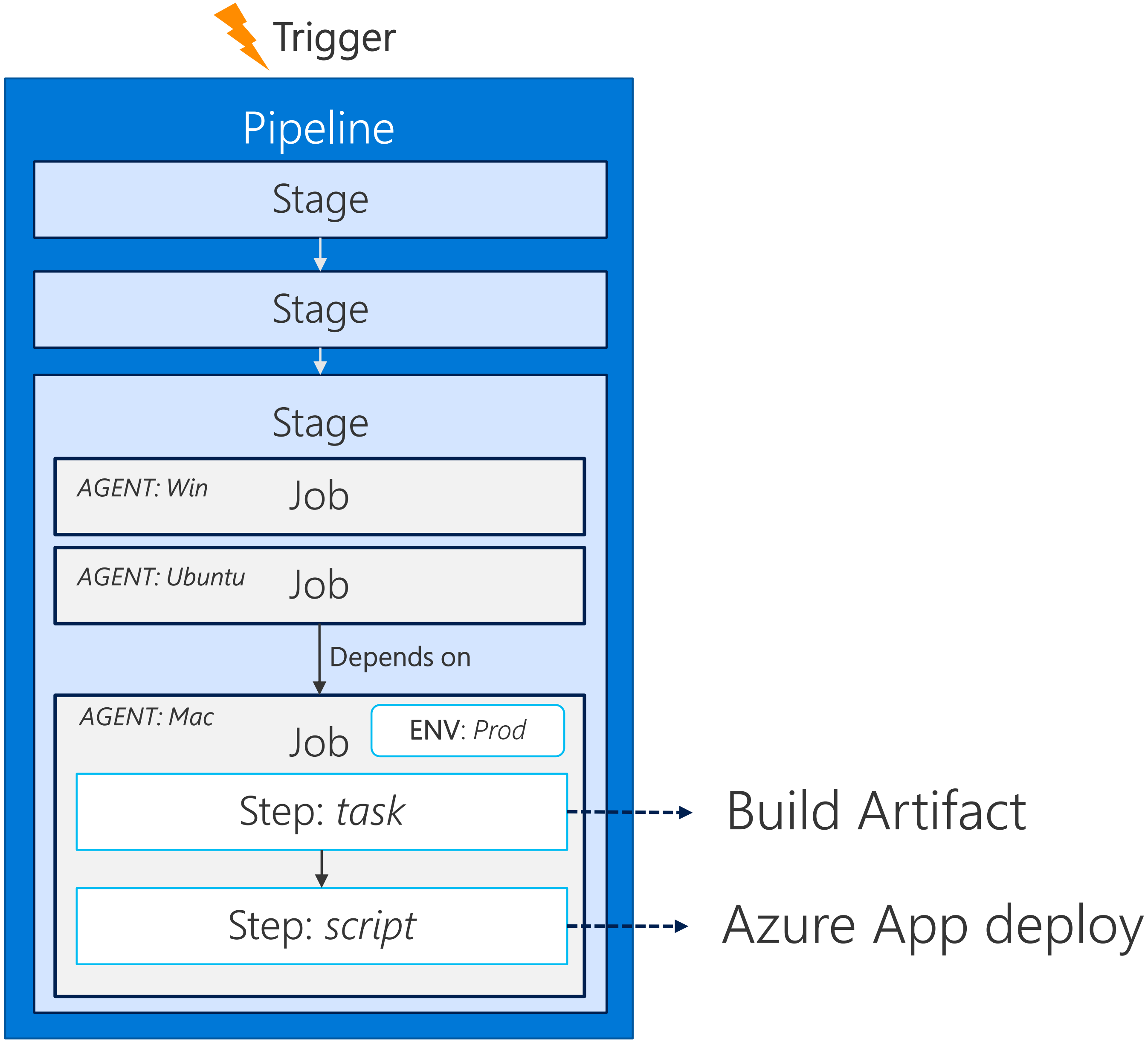


Innovation Pipeline (CI/CD)



Demo: CI/CD

Azure Pipelines Concepts

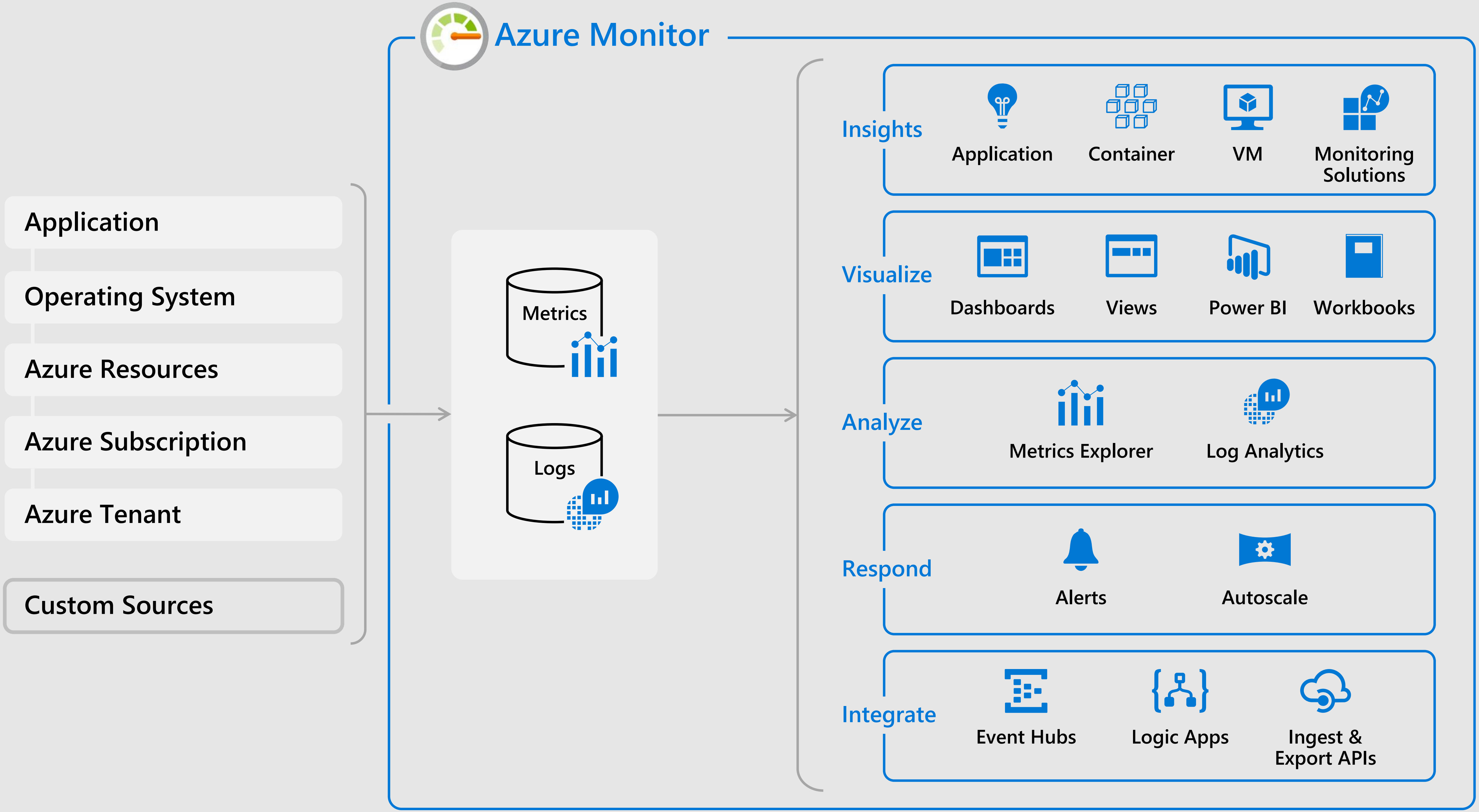


```
trigger:
- master

variables:
- foo: bar

pool:
  vmImage: ubuntu-18.04

stages:
- stage:
  jobs:
  - job:
    steps:
    - script: echo $foo
```

Demo: CI/CD – continued

Learnings

Monitor your infrastructure, pipeline *and* data.


Azure Monitor
Application
Operating System
Azure Resources
Azure Subscription
Azure Tenant
Custom Sources
Metrics
Logs
Insights
Application
Container
VM
Monitoring Solutions
Visualize
Dashboards
Views
Power BI
Workbooks
Analyze
Metrics Explorer
Log Analytics
Respond
Alerts
Autoscale
Integrate
Event Hubs
Logic Apps
Ingest & Export APIs

Key Learnings

- Leverage data-tiering in the data lake.
- Validate early in the data pipeline.
- Ensure the data pipeline is replayable (idempotent).
- Automate* deployments (CI/CD).
- Ensure data transformation code is testable.
- Secure and centralize configuration.
- Monitor infrastructure, pipelines *and* data.

https://aka.ms/mdw-dataops

Microsoft

Thank you!

Lace Lofranco

Senior Software Engineer, Microsoft

🐦 @lacelofranco

github.com/devlace