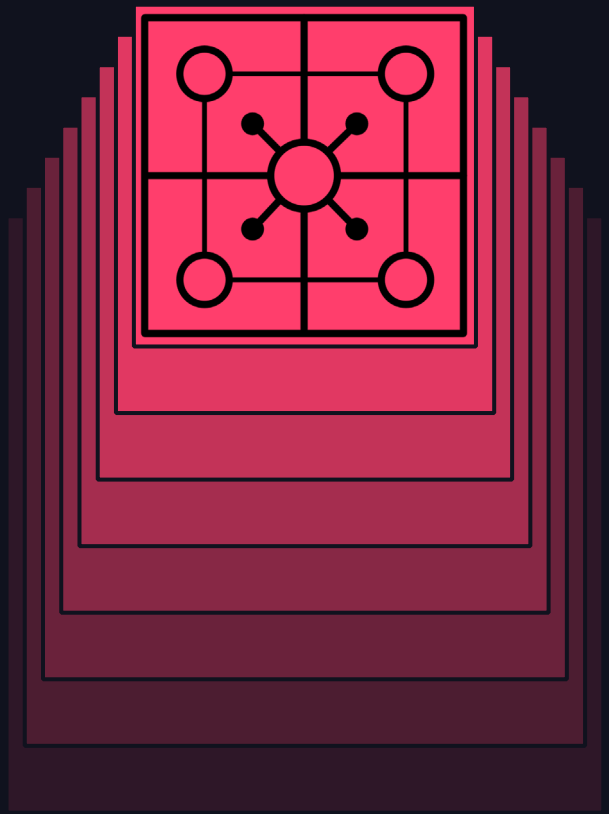
DATA+AI SUMMIT BY databricks

Unilever

BLUEPRINT



Roberto Alejandro Flores Meregote
2024-06-08

©2024 Databricks Inc. — All rights reserved

NICE TO MEET YOU

©2024 Databricks Inc. — All rights reserved

BLUEPRINT

THE DE TEAM!

©2024 Databricks Inc. — All rights reserved

BLUEPRINT

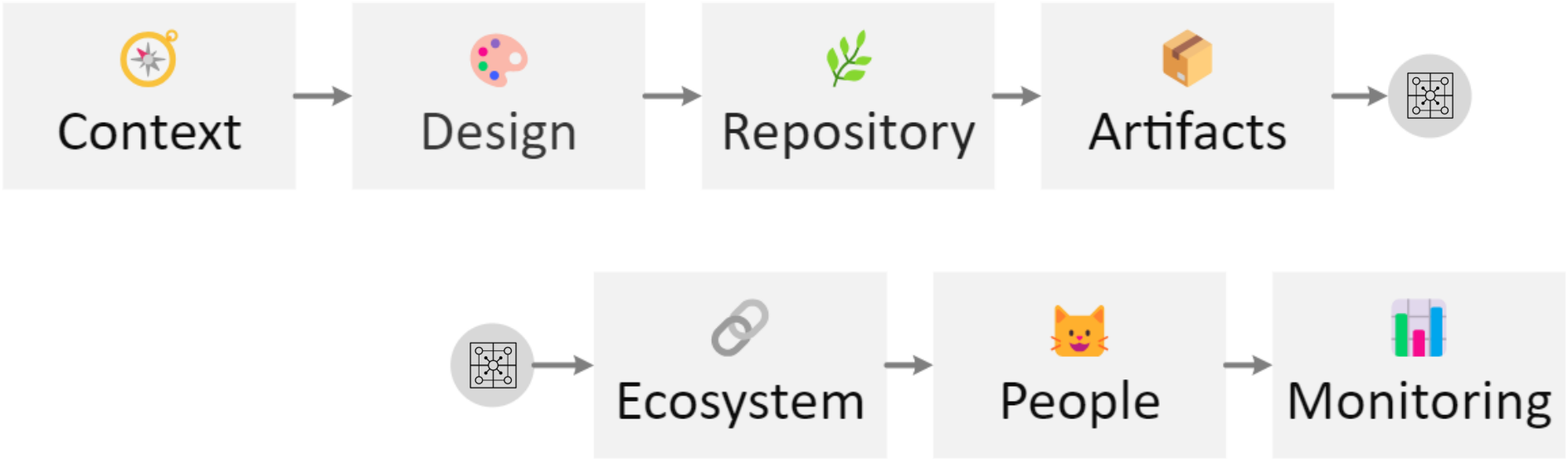
Roberto Flores Meregote

Europe Head of Data Engineering @Unilever

©2024 Databricks Inc. — All rights reserved

BLUEPRINT

AGENDA!

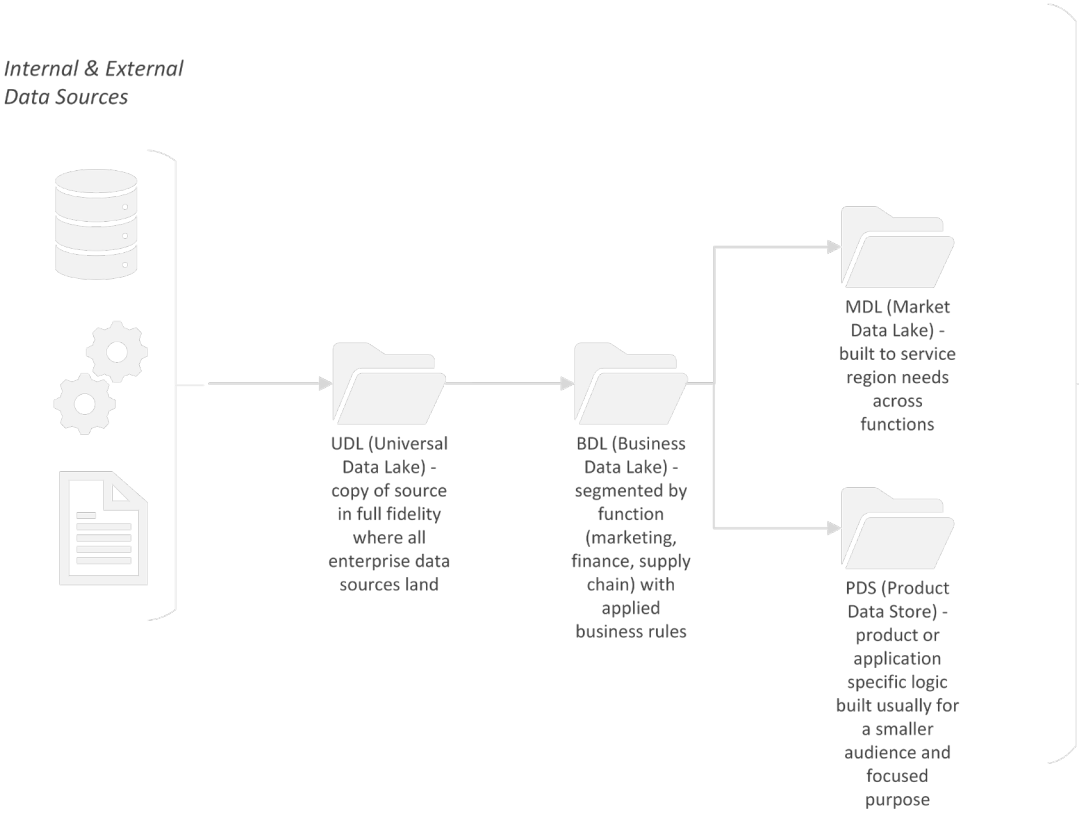


©2024 Databricks Inc. — All rights reserved

CONTEXT

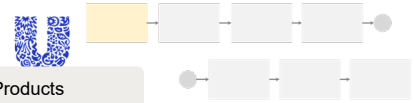
©2024 Databricks Inc. — All rights reserved

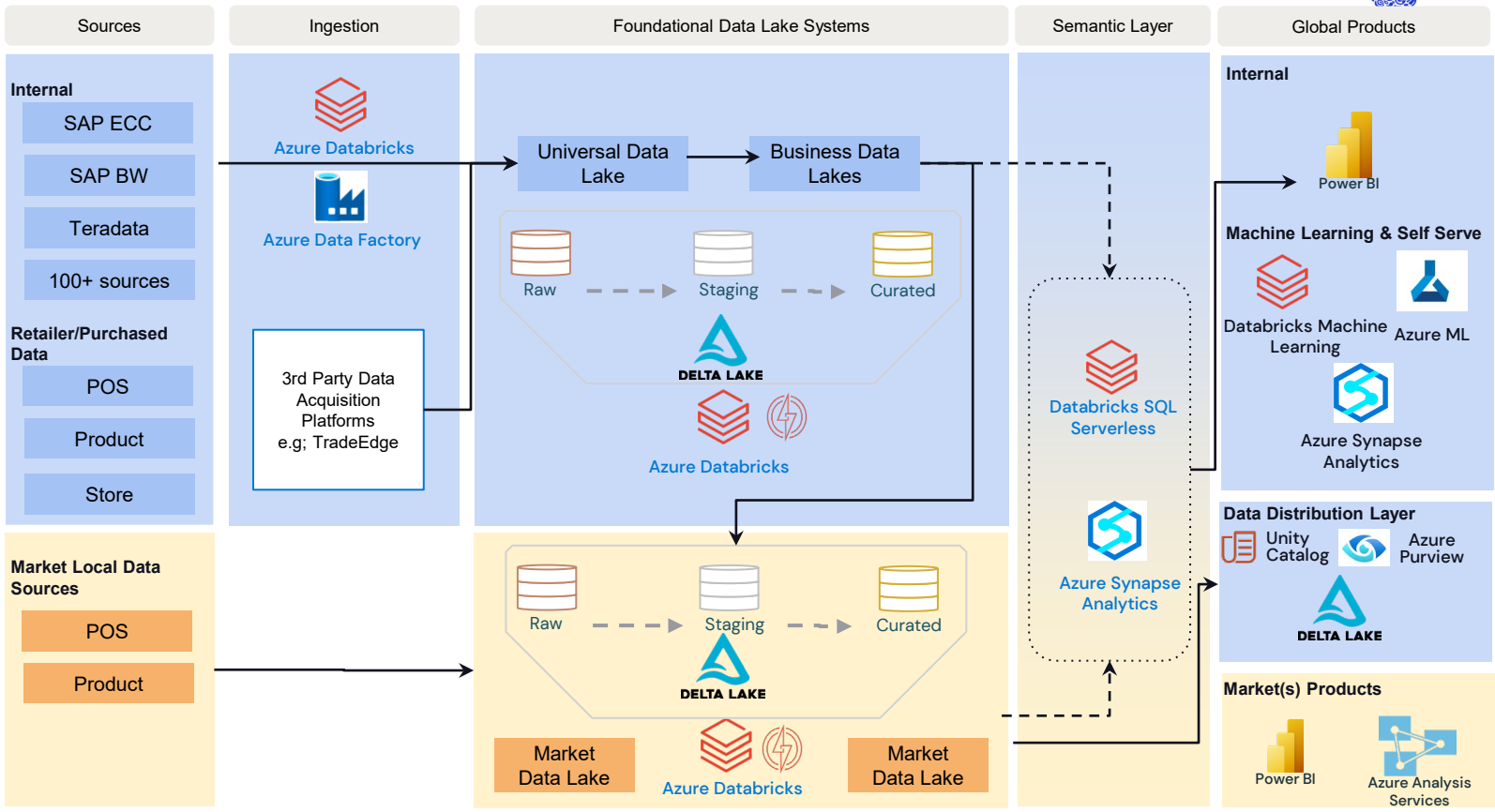
UNILEVER'S DATA ESTATE



©2024 Databricks Inc. — All rights reserved

UNILEVER'S DATA ESTATE

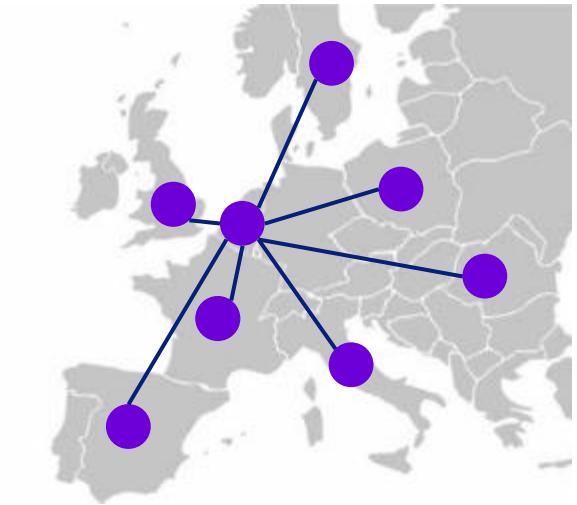




©2024 Databricks Inc. — All rights reserved

EUROPE IS A COMPLEX MARKET

PRIOR TO MDL, COUNTRIES HAD DIFFERENT DATA MATURITY LEVELS



©2024 Databricks Inc. — All rights reserved

EUROPE IS A COMPLEX MARKET

SIMILAR IN SIZE TO NORTH AMERICA, BUT WITH HIGH DATA COMPLEXITY

DATA COMPLEXITY

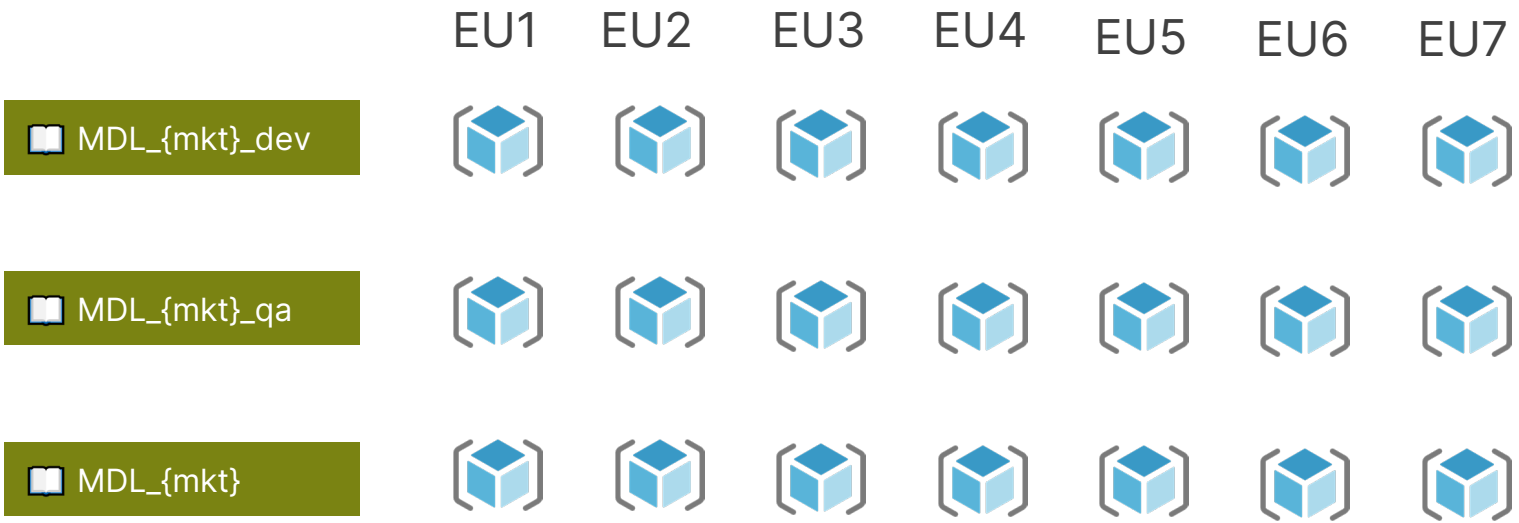
UL Markets	2	38
UL BG Cells	10	129
Official Languages	1	24
Databases (external)	20	316

Source: Europe MDL | Country / Category Cells with a Nielsen subscription

©2024 Databricks Inc. — All rights reserved

EUROPE IS A COMPLEX MARKET

PREVIOUS SETUP MEANT THAT WE STARTED WITH MANY ENVIRONMENTS



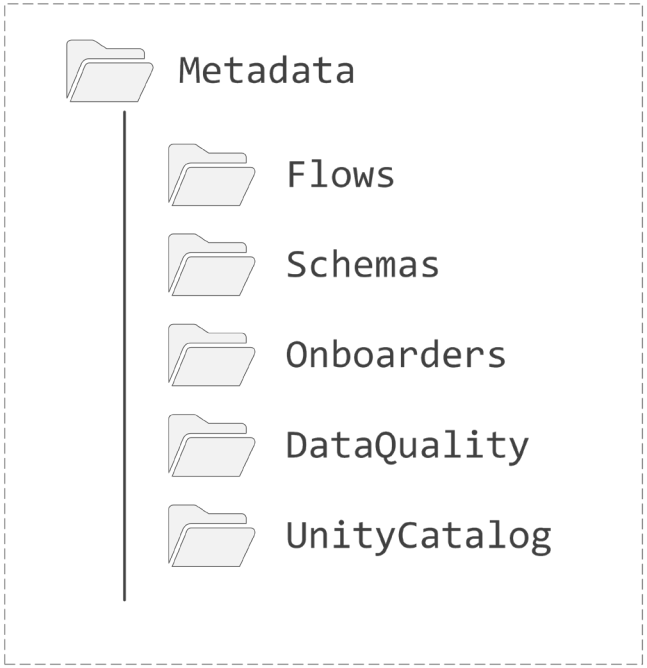
©2024 Databricks Inc. — All rights reserved

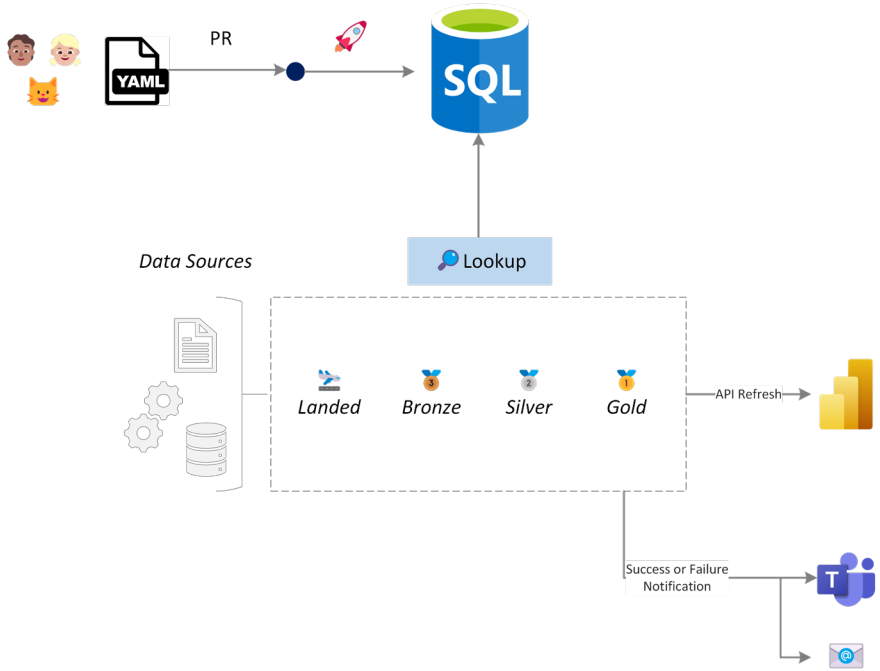
FRAMEWORK

©2024 Databricks Inc. — All rights reserved

FAMEWORK OVERVIEW

WHAT WE WANTED TO MANAGE IN A TECH AGNOSTIC WAY





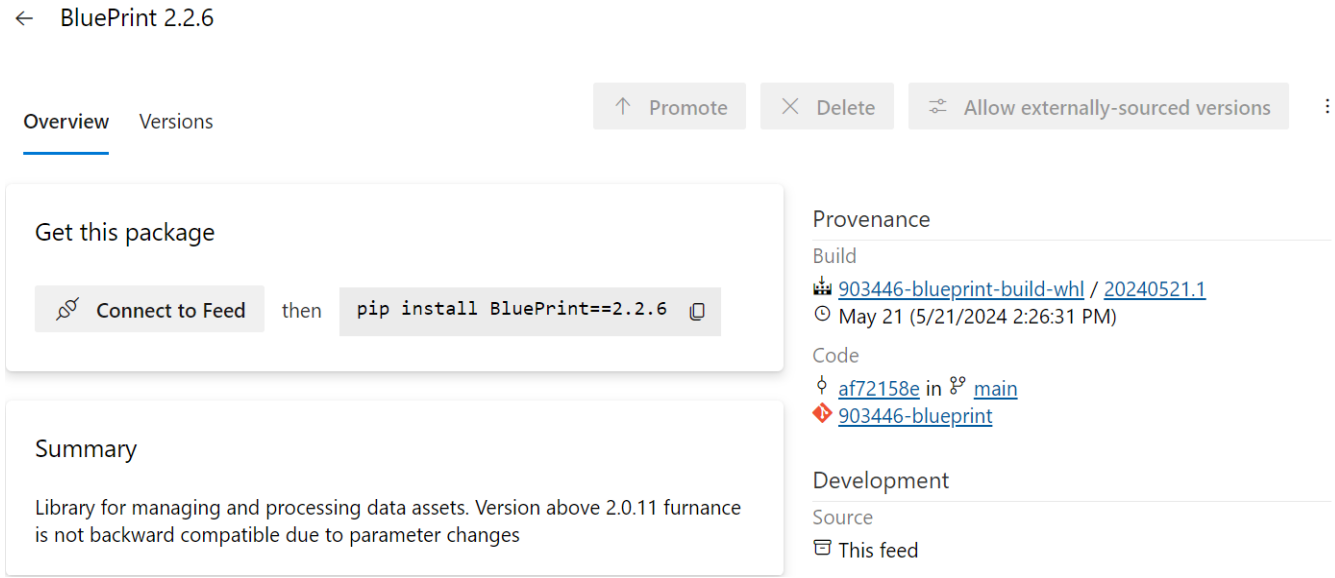
©2024 Databricks Inc. — All rights reserved

FAMEWORK OVERVIEW

WHAT WE WANTED TO MANAGE IN A TECH AGNOSTIC WAY

Common Artifacts

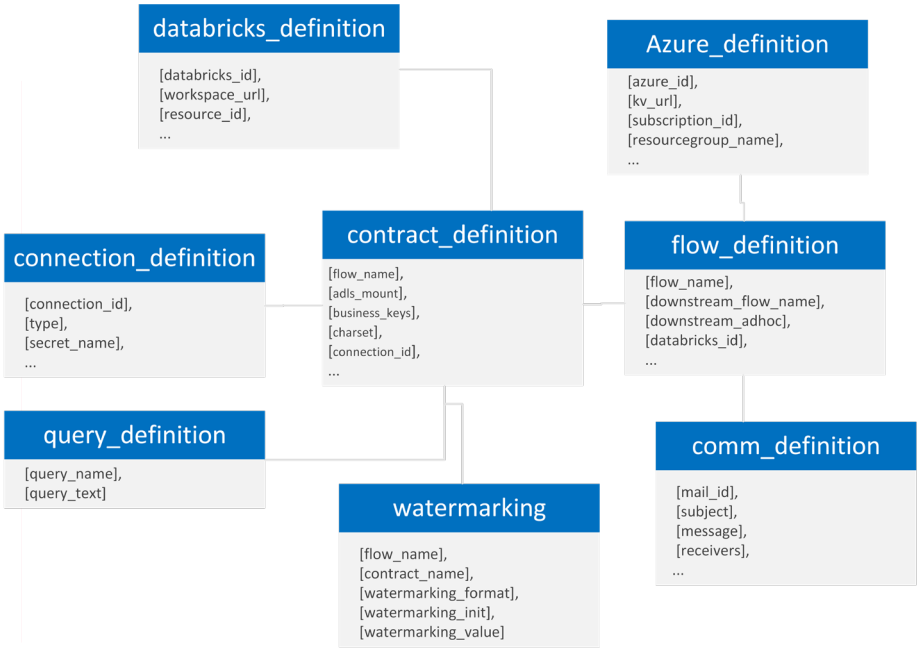
1. Onboarding
2. Unity Catalog Management
3. Power Platform Integrations
4. Data Factory Pipelines



©2024 Databricks Inc. — All rights reserved

FRAMEWORK OVERVIEW

DATABASE



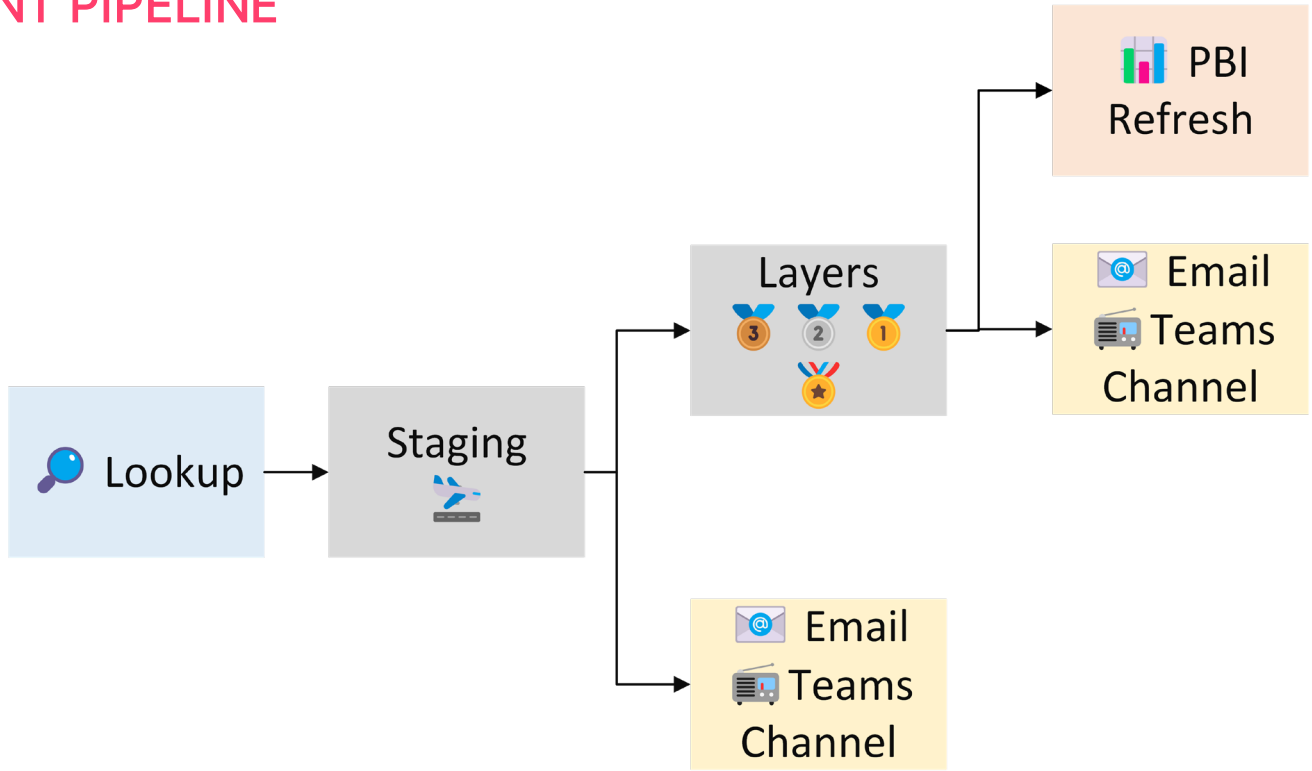
Stored Procedure Sample

```
CREATE OR ALTER PROCEDURE [utils].[sp_staging_info]
    @flow_group_name NVARCHAR(255),
    @staging_type NVARCHAR(255),
    @meta_schema NVARCHAR(255),
    @flow_filter NVARCHAR(255) = NULL

    AS
    SELECT DISTINCT
   fw.flow_name,
       fw.blueprint_version,
       az.kv_url,
       az.landing_adls_url,
       az.delta_lake_path,
       az.subscription_id,
       az.tenant_id,
       az.resourcegroup_name,
       az.secret_scope,
       az.azure_devops_organization,
       ct.user_id,
       ct.user_name,
    (...)
```


©2024 Databricks Inc. — All rights reserved

WHAT DOES IT LOOK LIKE

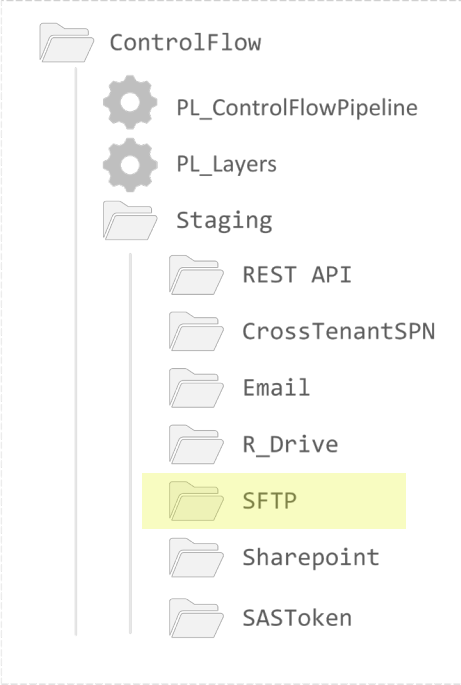
PARENT PIPELINE

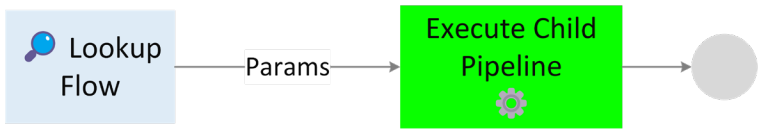


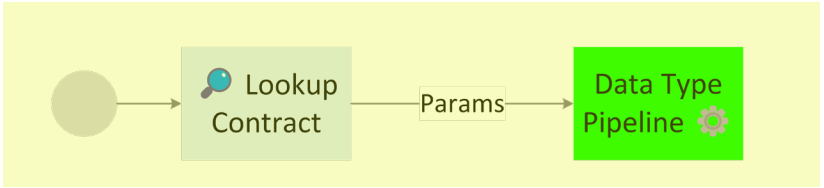
©2024 Databricks Inc. — All rights reserved

WHAT DOES IT LOOK LIKE

STAGING PIPELINES



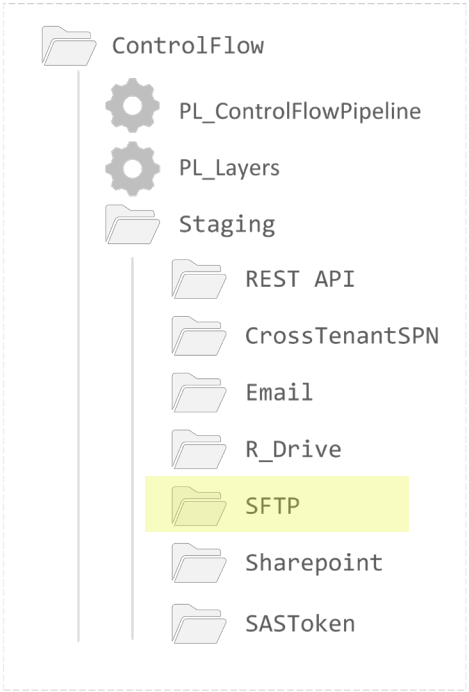


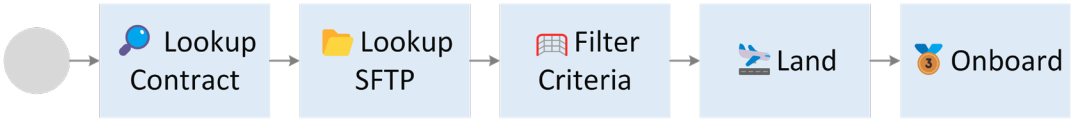


©2024 Databricks Inc. — All rights reserved

WHAT DOES IT LOOK LIKE

STAGING PIPELINE SAMPLE - SFTP

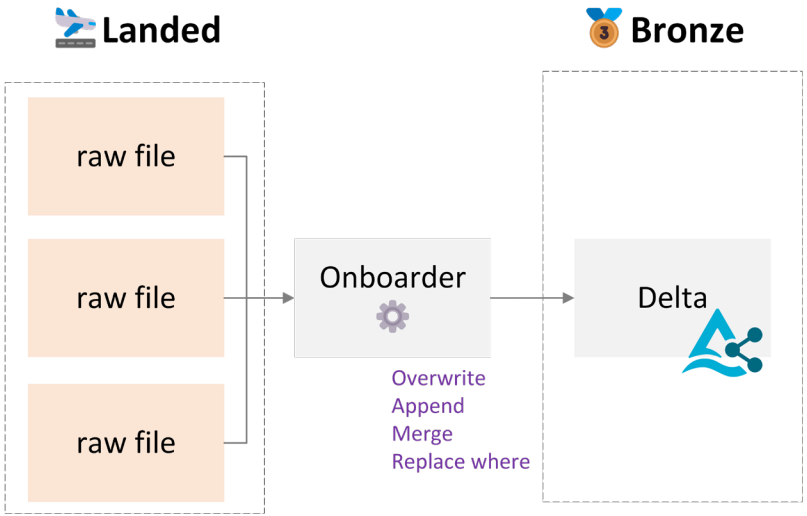




©2024 Databricks Inc. — All rights reserved

WHAT DOES IT LOOK LIKE

ONBOARDING

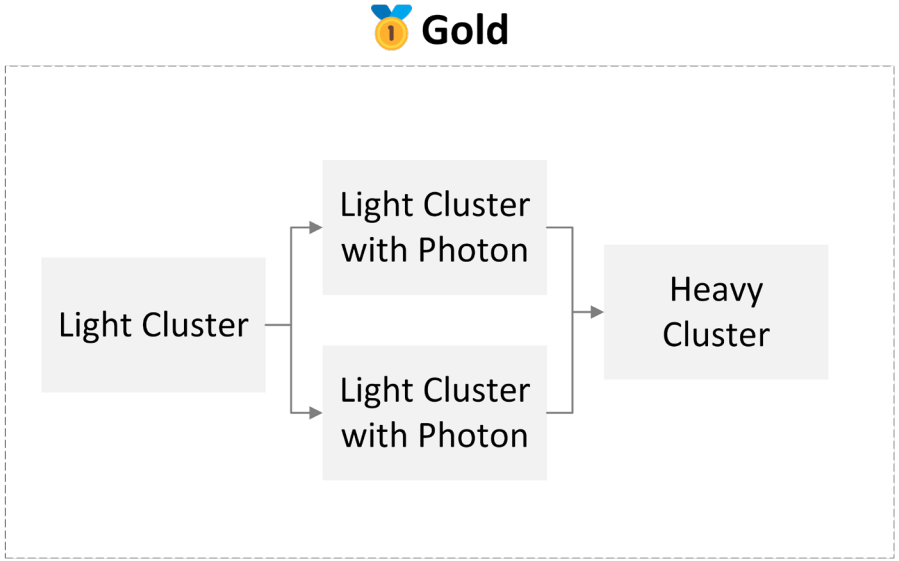



```yaml
- contract_name: BNL_PDD_LOOKUP_LIST_INGRED
  adls_mount: /dbfs/mnt/landingzone/
  business_keys: null
  connection_id: SAP_DQ_SP
  delimiter: null
  file_type: .xlsx
  header: 1
  layer: bronze
  load_type: F
  mapping_id: null
  onFail: Fail
  range: A1
  region: Benelux
  schema_name: bronze_mdl_bnlsharepoint
  sheet_name: ingredientGPC
  skip_line: null
  source_file_name: lookup_lists.xlsx
  source_path: sites/PDDsourcedata/Shared%20Documents/PDD%20lookup%20lists
  source_system_type: AdHoc
  source_system: Sharepoint
  stagingType: PL_Sharepoint_MetadataBased
  table_name: bdl_pdd_lov_ingred
```


©2024 Databricks Inc. — All rights reserved

WHAT DOES IT LOOK LIKE

LAYER EXECUTION, DEPENDENCIES AND RUN ORDER




```yaml
gold_notebooks:
  #Finance Dim
- path: /MarketDataLake/PRO/05-Gold/Finance/gold_finance_dim
  flow_filter: Finance
  parameters:
    DefaultParm: NA
  run_order: 1
  databricks_id: Normal Cluster
  #Nielsen Dim
- path: /MarketDataLake/PRO/05-Gold/Nielsen/gold_nielsen_dim
  flow_filter: Poland
  parameters:
    param_country_2: Poland
    param_country_cd: PL
  databricks_id: Heavy Cluster
  run_order: 2
- path: /MarketDataLake/PRO/05-Gold/Nielsen/gold_nielsen_dim
  flow_filter: Hungary
  parameters:
    param_country_2: Hungary
    param_country_cd: HU
  run_order: 2
  databricks_id: Normal Cluster
```

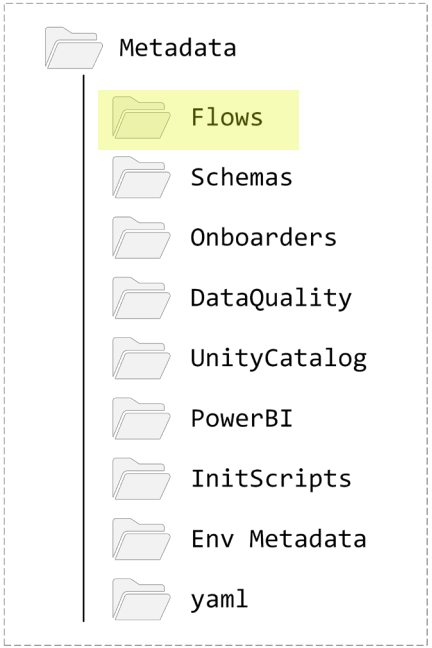

©2024 Databricks Inc. — All rights reserved

REPO STRUCTURE

©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

FLOW FILES



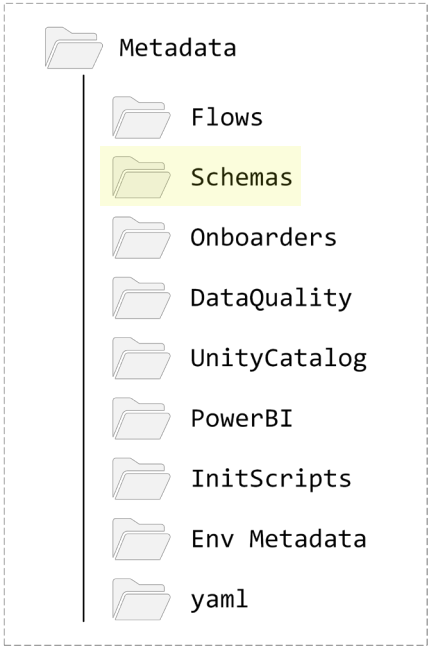
sample: Flow.yaml

```yaml
Flows:
- flow_name: DACH_Retailer
  azure_id: Azure Info
  bi_id: 'EPOS_DE_PBI'
  databricks_id: UC_Enabled_Cluster
  downstream_adhoc: ''
  downstream_flow_name: ''
  mail_id: 'DACH ePos Mail'
  run_downstream: True
  master_pipeline: PL_Master_MetadataDriven
  blueprint_version: 2.0.12
Contracts:
  - contract_name: DACH_REWE_SALES_AS
    adls_mount: /dbfs/mnt/landingzone/
    business_keys: null
    connection_id: DACH SFTP
    delimiter: ;
    file_type: .csv
    header: 0
    layer: bronze
    (...)
```


©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

SCHEMAS



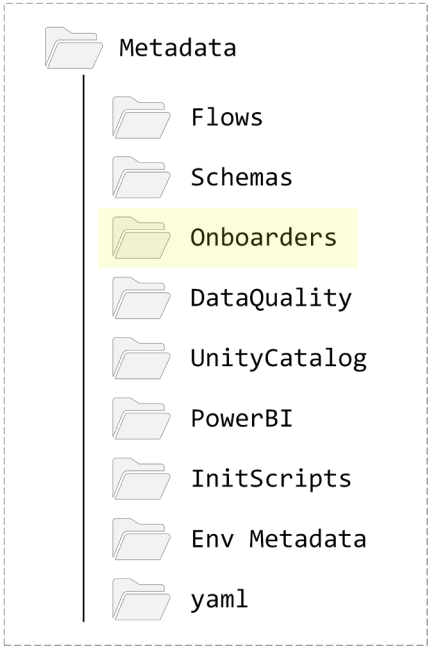
Sample: Schemas.json

```
{
    "$schema": "http://json-schema.org/draft-07/schema#",
    "type": "object",
    "properties": {
        "Flows": {
            "type": "array",
            "items": {
                "type": "object",
                "properties": {
                    "flow_name": {
                        "type": ["integer", "string"]
                    },
                    "downstream_flow_name": {
                        "type": ["string", "null"]
                    },
                    "downstream_adhoc": {
                        "type": ["string", "null"]
    (...)
```


©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

GENERALIZED ONBOARDER



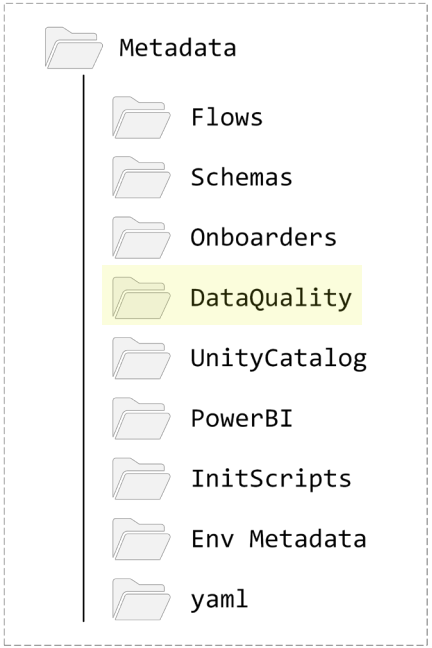
Sample: Onboarder.py

```
# Databricks notebook source
from blueprint.engineering.kiln import Kiln
import os
# COMMAND ----------
dbutils.widgets.removeAll()

    flow_name = dbutils.widgets.get("flow_name")
contract_name = dbutils.widgets.get("contract_name")
meta_schema = dbutils.widgets.get("meta_schema")

    print(f"Flow Name: {flow_name}")
print(f"Onboarding Contract: {contract_name}")
print(f"Metadata Schema: {meta_schema}")

    # COMMAND ----------

    print(f'Running with version of Blueprint:
    {os.environ["PKG_VERSION"]}')


    (...)
```


©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

DATA QUALITY (GREAT EXPECTATIONS)



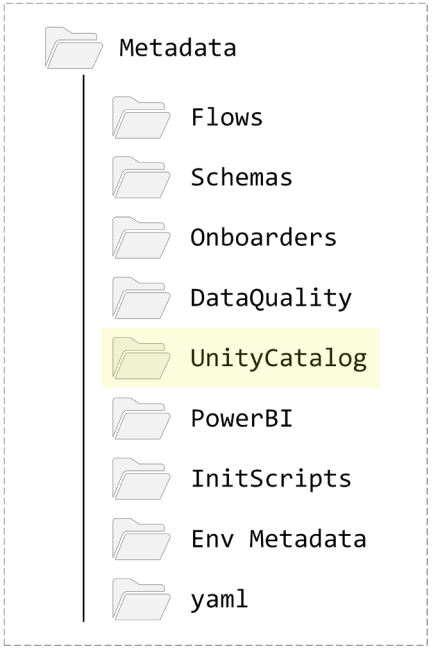
Sample: DataQuality.py

```
import pyspark.sql.functions as F
import json
from pyspark.sql.types import IntegerType, StringType, DoubleType,
    LongType
from delta.tables import *
from great_expectations.core.batch import RuntimeBatchRequest
from great_expectations.util import get_context
from great_expectations.data_context.types.base import (
    DataContextConfig,
    FilesystemStoreBackendDefaults,
)

    import great_expectations as ge
from great_expectations.dataset import (
    SparkDFDataset,
    MetaSparkDFDataset,
)


    (...)
```


©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

UNITY CATALOG METADATA



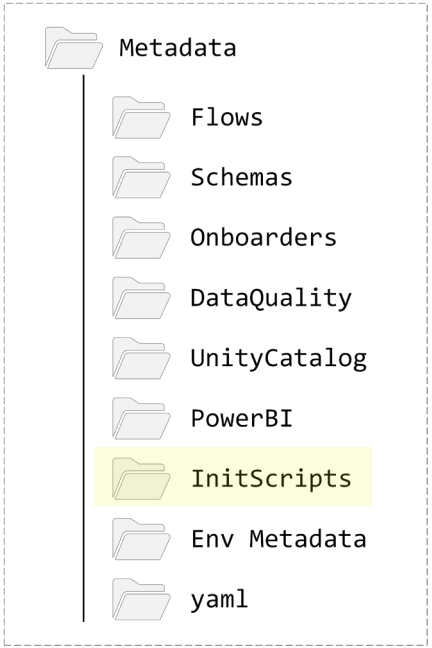
Sample: UnityCatalogDefinition.yaml

```yaml
UC_Def:
- flow_name: Init_schema
  schemas:
  - name: init_schema
    comment: "This is just a dummy file to init the correct schema"
    tags:
    - Test
    access:
      - acls_group_name: "SEC-ES-DA-p-903444-europe-analyst"
        acls_access_type: "SELECT, MODIFY"
        acls_access_catalogue: "mdl_europe_anz_dev"
    tables:
    - name: init_she
      comment: "### Supports Markdown 1. First item 2. Second item"
      tags:
      - Test      columns:
      - name: "Country_Code"
        comment: "Country code key as of 07/03/2024"
        tags:
          - PK
```


©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

INIT SCRIPTS



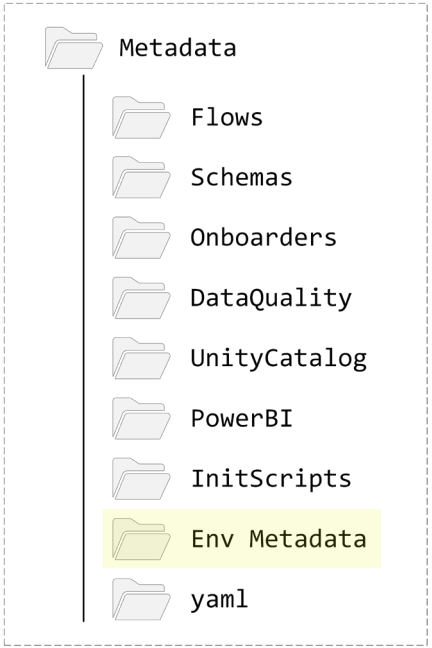
Sample: BluePrintInit.sh

```
curl -sL https://aka.ms/InstallAzureCLIDeb | bash

python -m pip install --upgrade pip

az login --service-principal -u ${AZ_DEVOPS_SP_APP_ID} -p
    ${AZ_DEVOPS_SP_SECRET} --tenant ${AZ_TENANT_ID}

AZ_SP_TOKEN=$(az account get-access-token --query accessToken -o tsv)

pip install BluePrint==${PKG_VERSION} --index-url
    "https://token:${AZ_SP_TOKEN}@pkgs.dev.azure.com/${AZ_DEVOPS_ORG_NA
    ME}/_packaging/${AZ_DEVOPS_FEED_NAME}/pypi/simple/"


    (...)
```


©2024 Databricks Inc. — All rights reserved

METADATA REPOSITORY STRUCTURE

UNITY CATALOG METADATA



Sample: EnvMetadata_dev.yaml

```
DatabricksDefinition:
- databricks_id: Strong Cluster
  default_catalog: mdl_europe_anz_dev
  workspace_url: https://adb-xxxxxxxx.yy.azuredatabricks.net
  resource_id: /subscriptions/blablabla
  policy_id: 0000EB555555
  cluster_version: 12.2.x-scala2.12
  cluster_type: Standard_E8d_v4
  scaling: '5:20'
  init_script: /Metadata/InitScripts/BluePrintInit.sh
AzureDefinition:
- azure_id: Azure Info
  kv_url: https://xxxxx.vault.azure.net/
  landing_adls_url: https:/xxxxxxx.dfs.core.windows.net/
  delta_lake_path: 'abfss://unilever@
    xxxxxxx.dfs.core.windows.net/datalake'
  subscription_id: bbbbb-aaaa-zzzz-yyyy-xxxxxx
  resourcegroup_name: abceasyasonetwothree-rg
  tenant_id: aaaaaaa-bbbbb-cccccccc-ddddd
  azure_devops_organization: bnxxxxxxxxxxxsts
    (...)
```

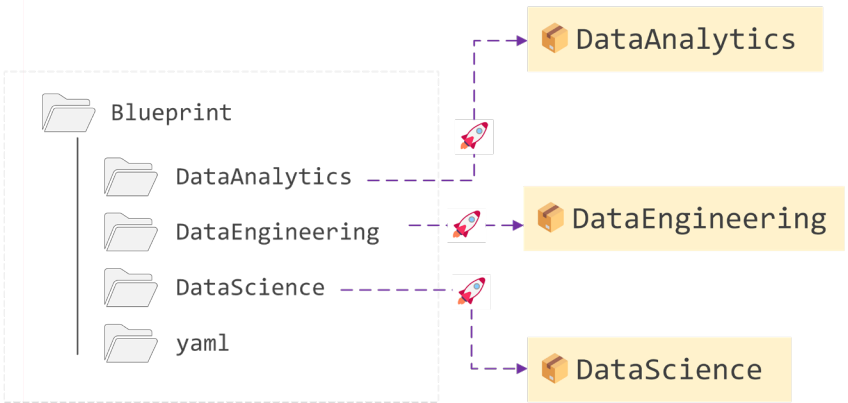

©2024 Databricks Inc. — All rights reserved

ARTIFACTS

©2024 Databricks Inc. — All rights reserved

HOW DO WE DEPLOY IT

PYTHON ARTIFACTS



YAML sample: Blueprint Python Artifact Creation

```yaml
stages:
- stage: Building_new_version
      jobs:
  - job: Build
    pool:
        name: Azure Pipelines
        vmImage: windows-2019

    steps:
    - checkout: self
      persistCredentials: true
      clean: true

        - script: |
      pip install twine
      pip install build
    displayName: 'Install twine and build'

        - script: |
      python -m build -w blueprint
    displayName: 'Build wheel'

        - task: TwineAuthenticate@1
    inputs:
      artifactFeed: Europe_MDL

        - script: |
      python -m twine upload -r Europe_MDL --config-file $(PYPIRC_PATH) blueprint/dist/*.whl
    displayName: 'Upload build whl to feed'
```


©2024 Databricks Inc. — All rights reserved

PYTHON PACKAGE ARTIFACT FEED

PACKAGE

BluePrint

Version 2.2.6

Version	Views	Publish date	Source
2.2.6 (currently selected)	Local	Tuesday	This feed
2.2.4	Local	Apr 15	This feed
2.2.3	Local	Apr 10	This feed
2.2.1	Local	Apr 2	This feed

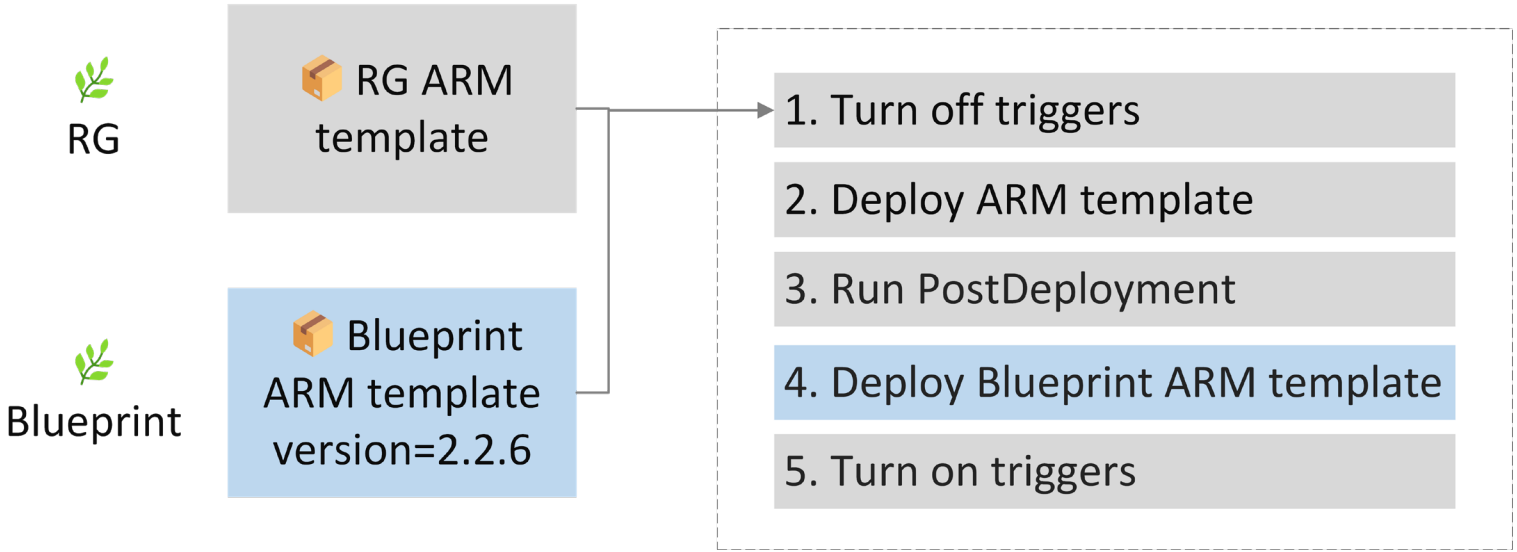
Get this package

Connect to Feed then `pip install BluePrint==2.0.12`

©2024 Databricks Inc. — All rights reserved

PYTHON PACKAGE ARTIFACT FEED

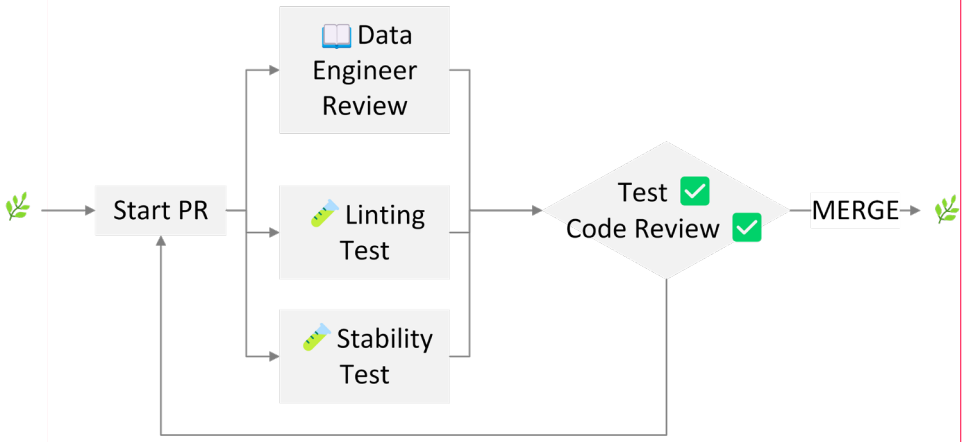
DATA FACTORY ARM TEMPLATES



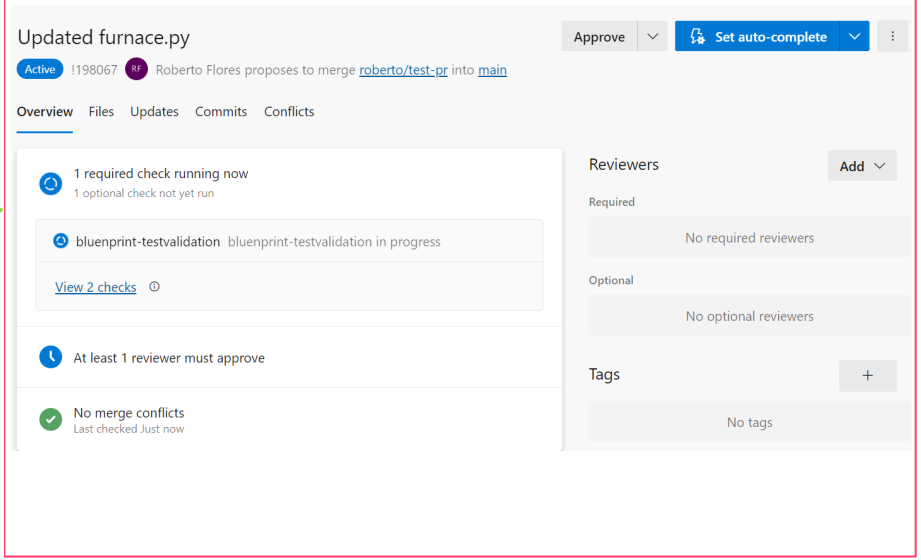
©2024 Databricks Inc. — All rights reserved

REUSABILITY REQUIRES RESILIENCE

TESTING AIMS TO ENSURE STABILITY ACROSS VERSIONS



Testing Pipeline Screenshot



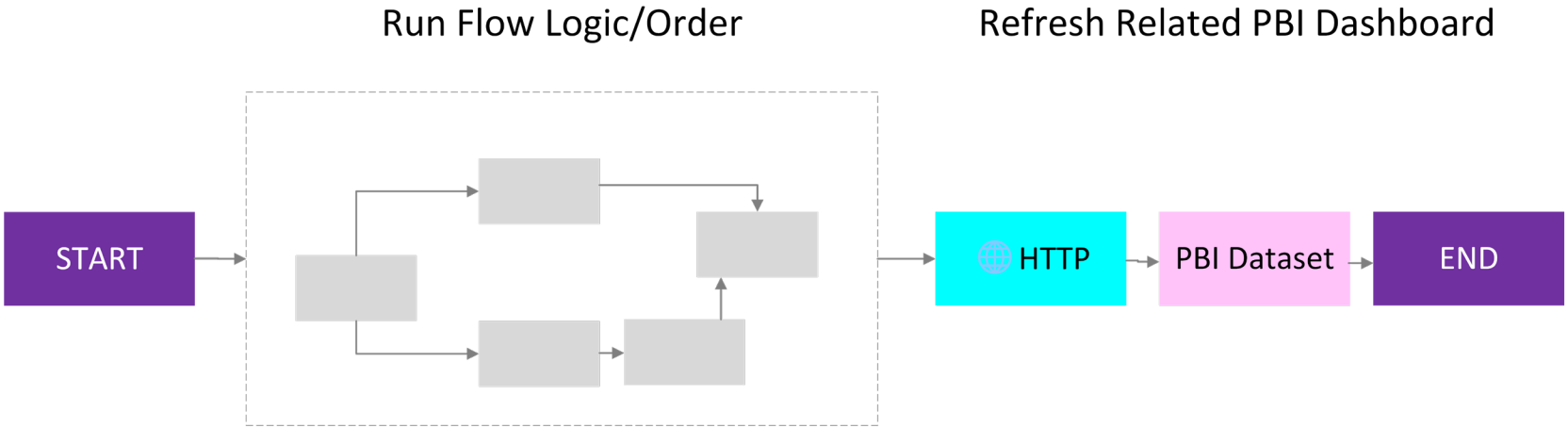
©2024 Databricks Inc. — All rights reserved

INTEGRATION WITH DATA ESTATE

©2024 Databricks Inc. — All rights reserved

PBI PREMIUM CAPACITY

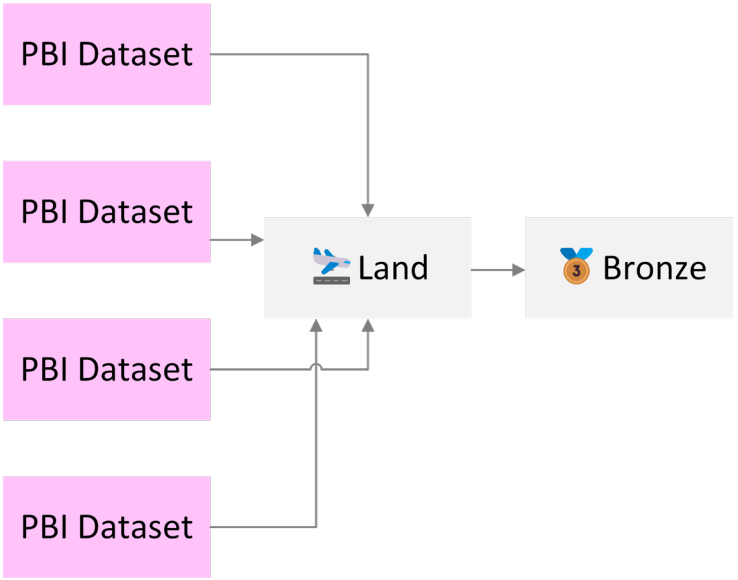
REFRESHING VIA ENHANCED API



©2024 Databricks Inc. — All rights reserved

PBI PREMIUM CAPACITY

USAGE STATISTICS

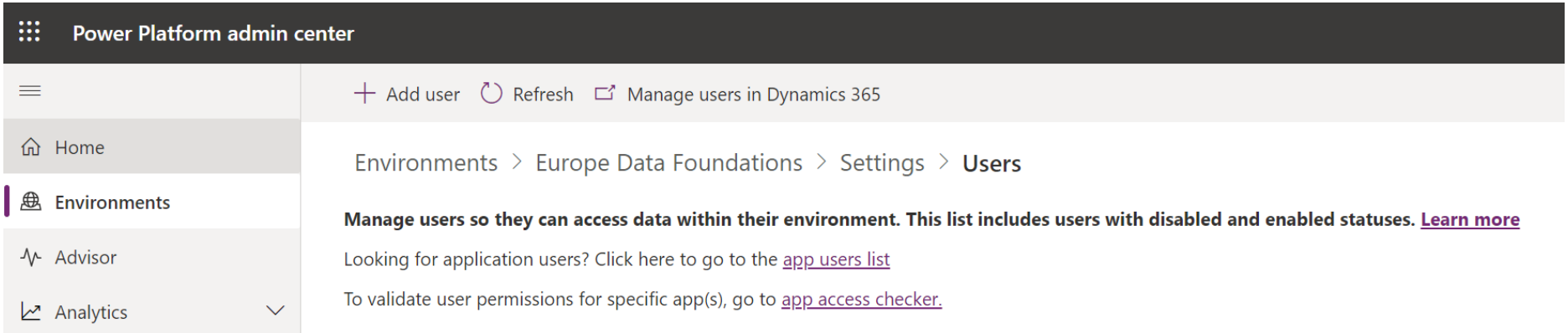


©2024 Databricks Inc. — All rights reserved

D365 INTEGRATION

D365 API – SERVICE PRINCIPAL

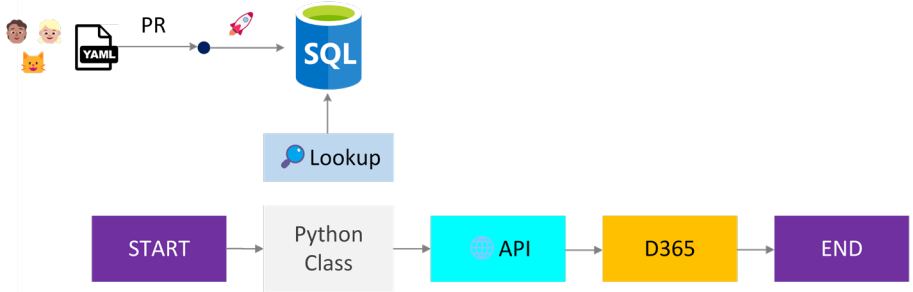
API / Permissions na...	Type	Description
Dynamics CRM (1)		
user_impersonation	Delegated	Access Common Data Service as orga...



©2024 Databricks Inc. — All rights reserved

D365 INTEGRATION

TABLE CREATION BASED ON YAML DEFINITION



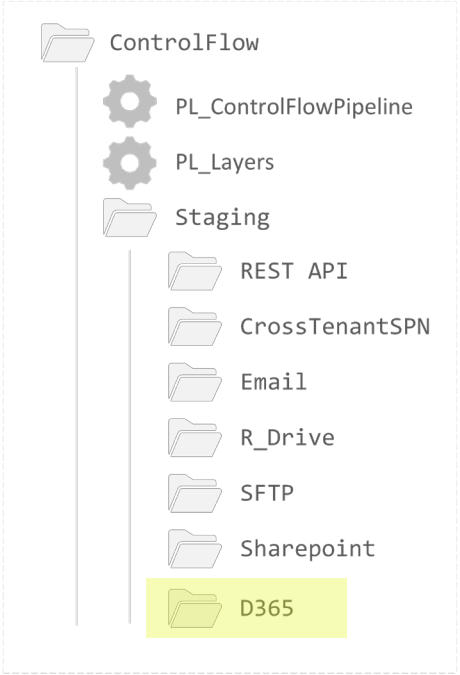
API Permissions

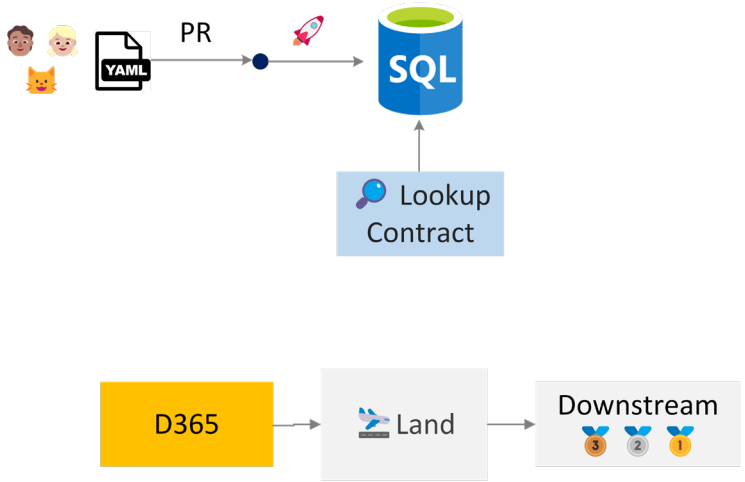
```
dv_target_yaml = """
table:
    edf_SnOP_Target_Test:
      unique_column: ID
      column_definitions:
          - column: Timeframe
            data_type: string_100
          - column: TargetMetric
            data_type: string_100
          - column: Portfolio
            data_type: choice_portfolio
          - column: Region
            data_type: string_100
          - column: ProductCategory
            data_type: string_50 """
```


©2024 Databricks Inc. — All rights reserved

D365 INTEGRATION

DATA FLOW – INBOUND

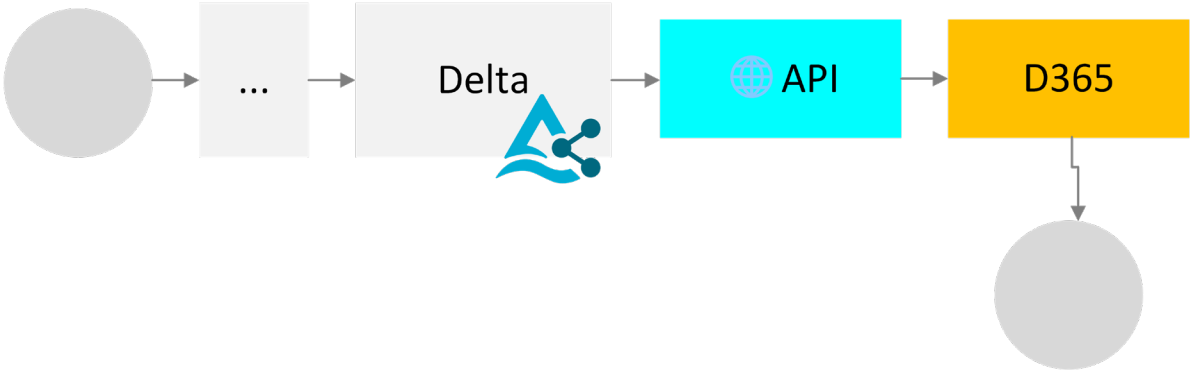




©2024 Databricks Inc. — All rights reserved

D365 INTEGRATION

DATA FLOW – OUTBOUND



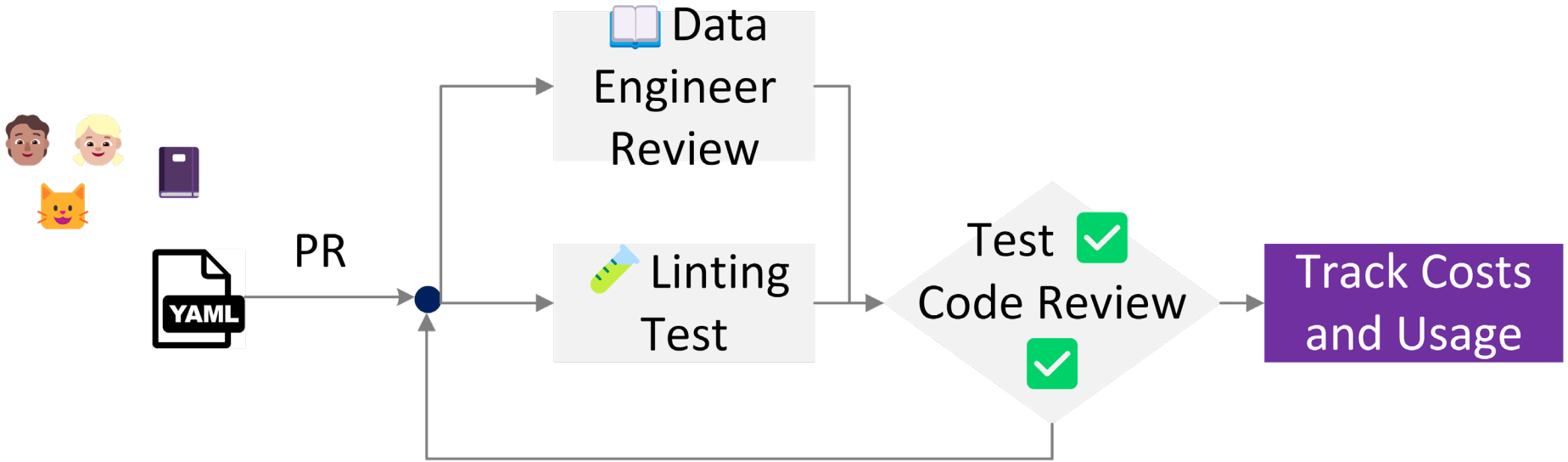
©2024 Databricks Inc. — All rights reserved

SELF SERVICE ENGINEERING

©2024 Databricks Inc. — All rights reserved

SELF SERVICE

REDUCES BOTTLENECKS



©2024 Databricks Inc. — All rights reserved

MONITORING

©2024 Databricks Inc. — All rights reserved

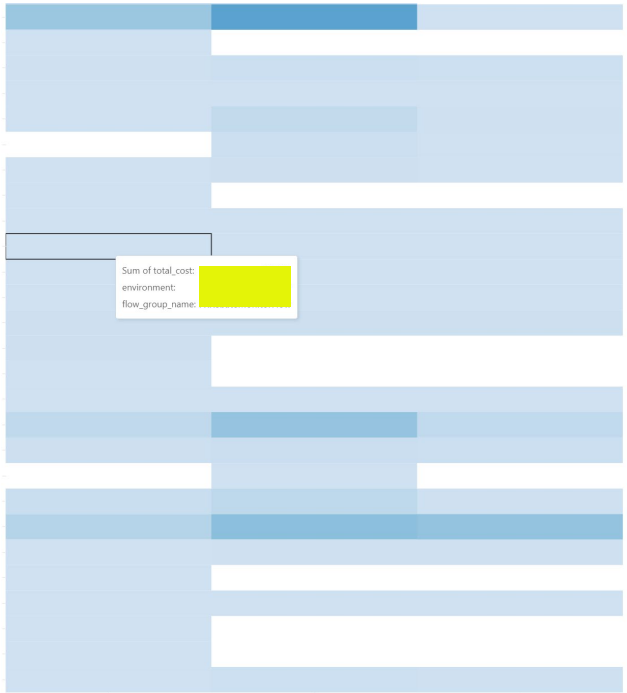
COST MONITORING

ACTIVITY BASED COSTING

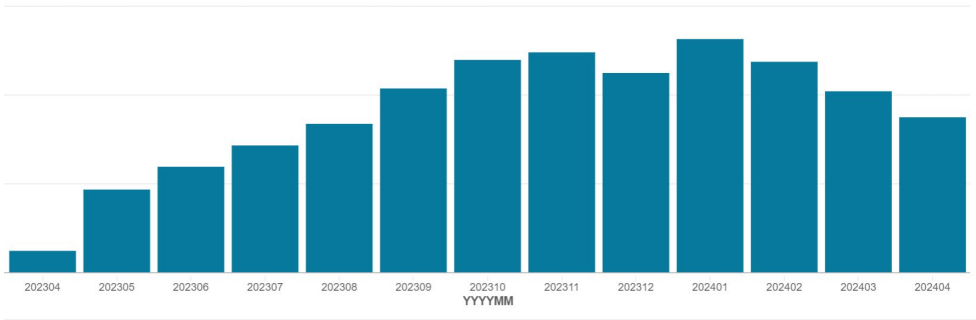
Flow A

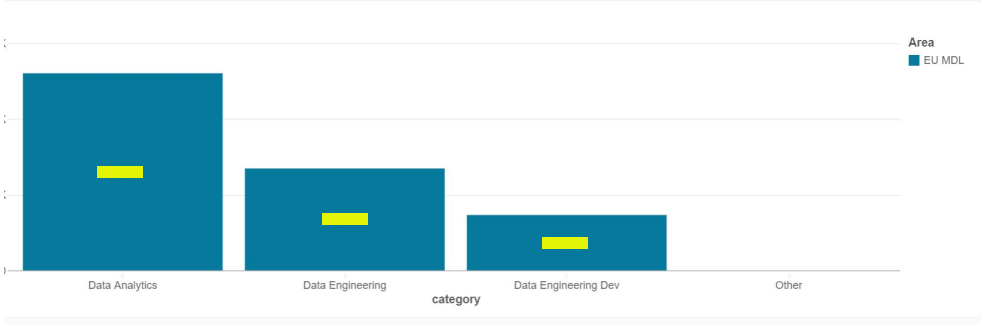
Flow B

...



Flow X

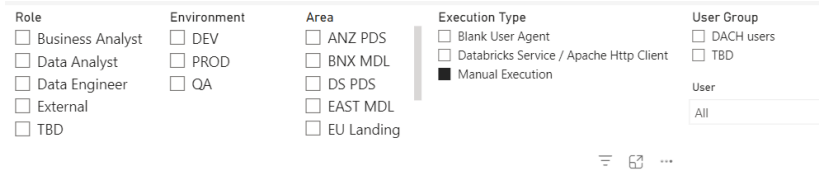


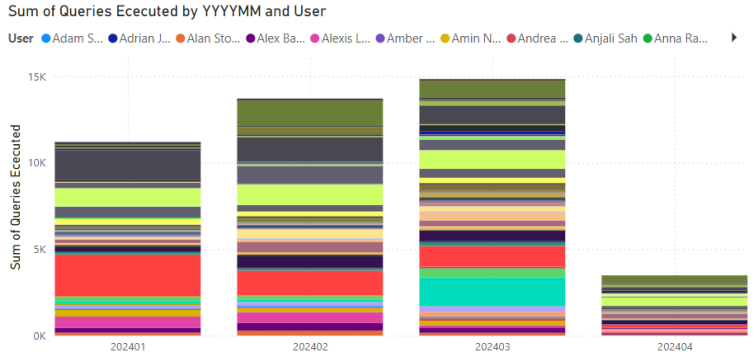


©2024 Databricks Inc. — All rights reserved

USAGE MONITORING

USAGE SEGMENTED BY PERSONA AND TOOL





©2024 Databricks Inc. — All rights reserved

THANK YOU!

Q&A

©2024 Databricks Inc. — All rights reserved