DATA+AI SUMMIT BY databricks

REPSOL

Next-Gen Energy AI: Asset Bundles/MLOps on Repsol's Data & AI Platform

Agenda

- Introduction
- Architecture in Repsol
- Why Databricks?
- DAB (Databricks Asset Bundles) for MLOps
- Demo CI/CD
- Issues and Solutions
- Next Steps

Introduction

About Us

Alfonso Fernandez Barandiaran
Repsol
Head of Bigdata & AI Cloud Platform

Carlos Rosado Moral
Repsol
Machine Learning & MLOps Specialist

About REPSOL

A global multi-energy company based in Madrid, Spain. We put the customer at the heart of everything we do with the aim to meet all energy needs by offering new solutions.

- **Comprehensive company:** Present throughout the entire value chain, and market a wide range of products in over **90 countries** worldwide.
- **Technology and innovation:** Cutting-edge technology to obtain best solutions in the energy industry; launched **> 670 digital initiatives** to improve efficiency & safety and optimize resources.
- **Talented team:** Diverse team of **> 25,000 employees** representing **77 nationalities,** across **27 countries**.
- **Net zero emissions by 2050:** The **first energy company** to set this ambitious objective in-line with the Paris Agreement, and we are aligning our entire value chain to achieve it.

Architecture in Repsol

ARiA Platform

What is ARiA?

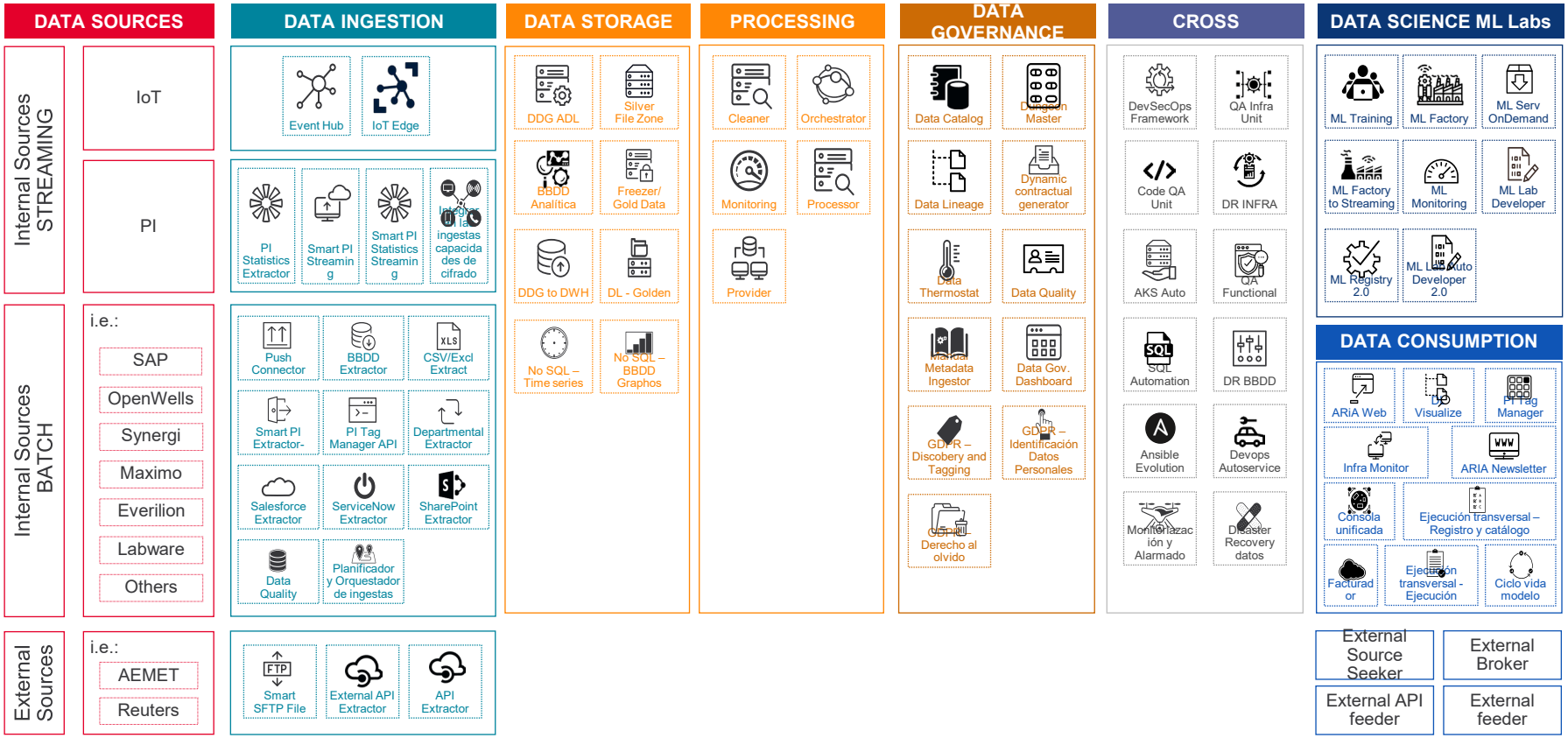
ARiA is an integrated technological solution that allows data collected from different external sources and hosted in its persistence systems (databases, file systems, etc.) to be governed, accessed and made available users, data applications, analytics and/or any other technology to meet strategic business objectives.

In this sense, ARiA is **Repsol's Centralized Data, Analytics & AI Platform.**

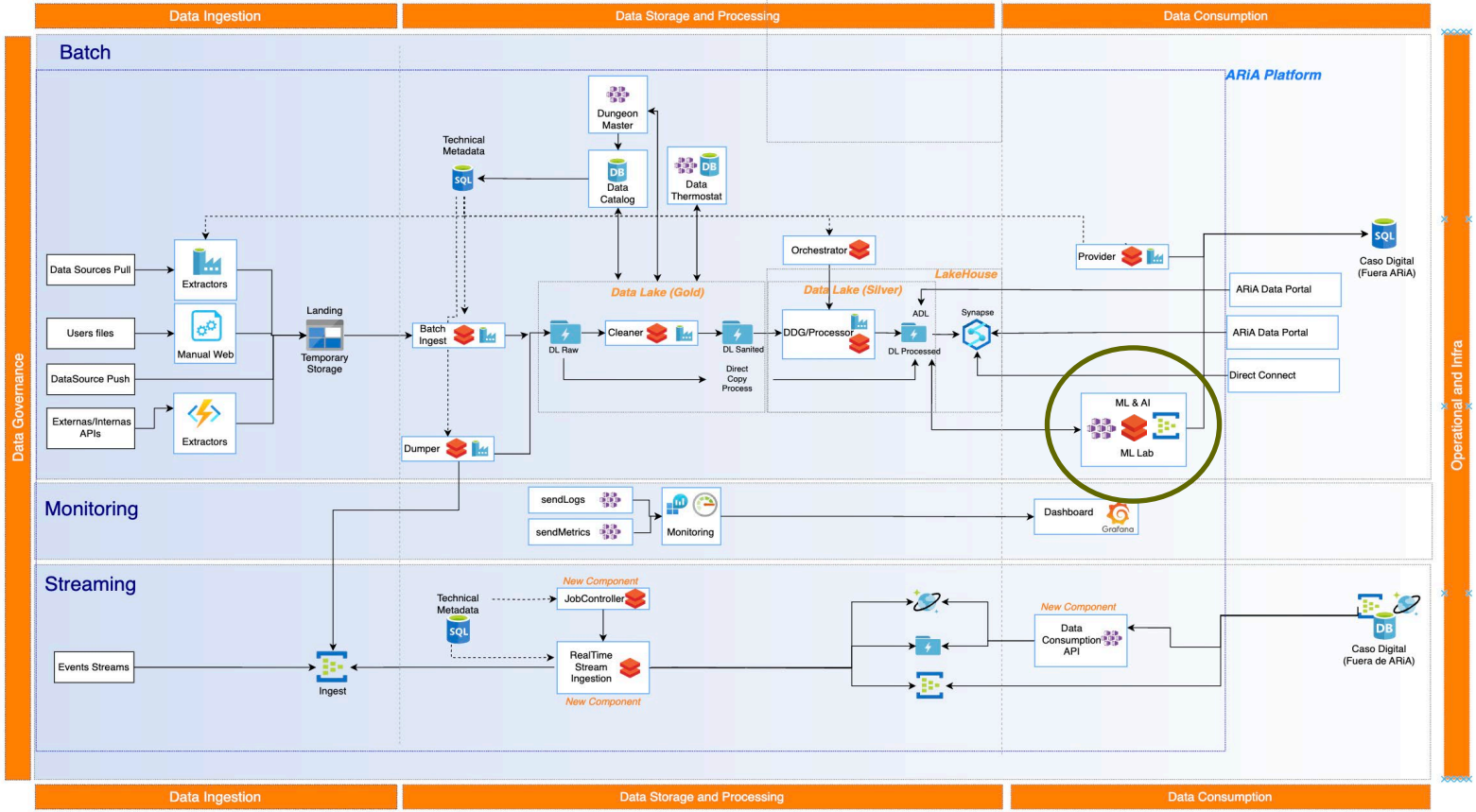
ARiA Design Principles

	ARiA principles	Vs.	Traditional approach
1	Vision Platform		Case by case
2	A Multi-business		Many vertical platforms
3	Cloud Platform (Microsoft Azure)		On premise
4	Own development (PaaS)		Commercial platforms
5	Case by case		Big bang
6	Components Oriented Architecture		Monolithics Architecture
7	DevSecOps by design		Data point
8	Lakehouse Vision		Data Lake

ARiA Architecture



ARiA Architecture



Why Databricks?

MLOps Evolution at Repsol

Kubern

- Endpoints in
- Execution in
- ML Console

L + Databricks

ubernetes

ML and
ks

Veb application)

Data Citizens

New MLOps at Repsol

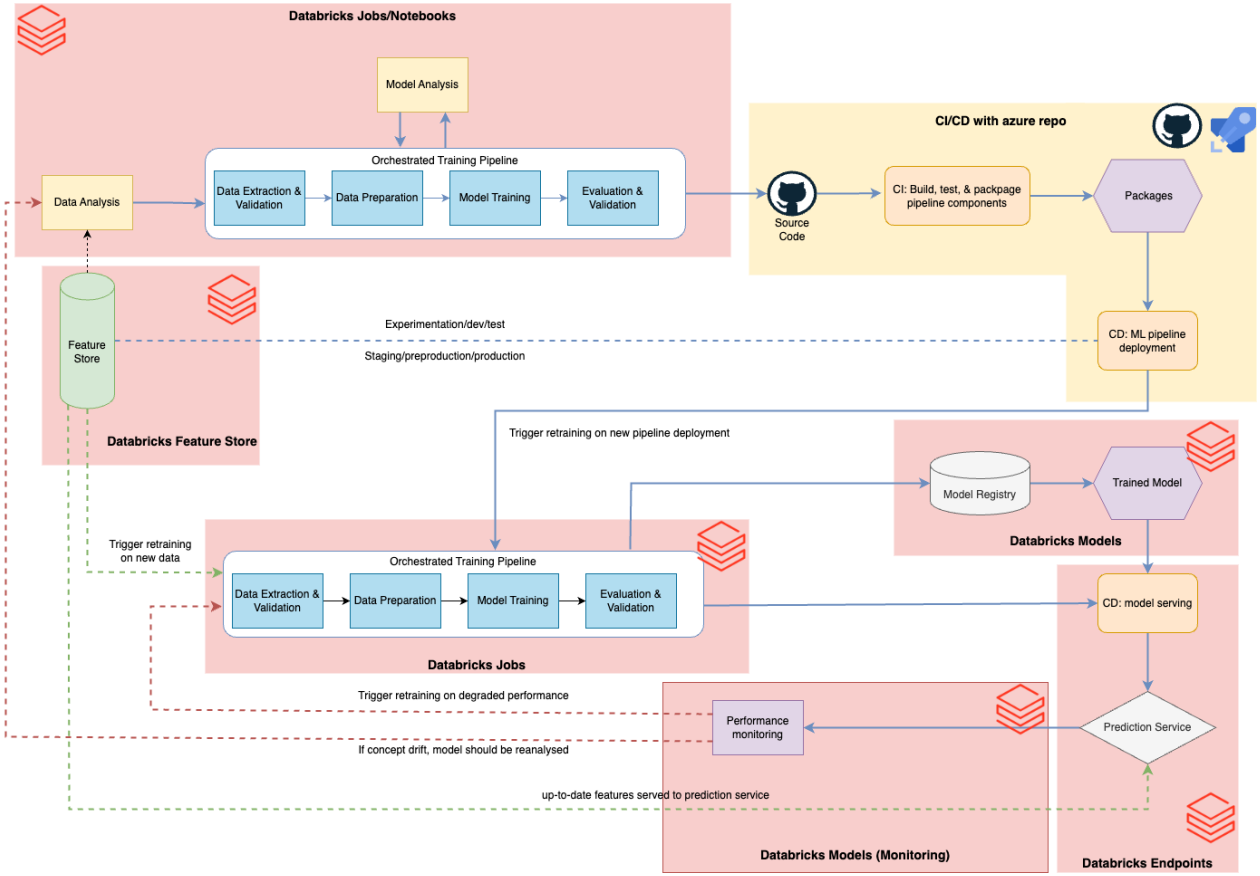
Databricks

- MLOps + LLMOps
- Gen AI process
- Execution of all kind of models
- New tools for Data Citizens

Kubernetes

- MLOps APIs Microservices
- Gen AI APIs
- ML Console (Web application)
- Optimization Models in Kubernetes

Our Model Life Cycle



(*) Book of MLOps Google

Kubeflow vs Databricks Costs

REPSOL

Kubeflow

$48k
Infra cost

$245k
Develop Cost

$7,5k
Training cost

$20k
Migration models

Databricks

$58k
Infra cost

$110k
Develop Cost

$10k
Training cost

$7k
Migration models

$280k
Current MLOps
Infra cost

VS

$58k
New MLOps
Infra cost

DAB (Databricks Asset Bundles) for MLOps

DAB (Databricks Asset Bundles)

Databricks
Asset
Bundles

What are DAB?

YAML files that specify the artifacts, resources, and configurations of a Databricks Project.

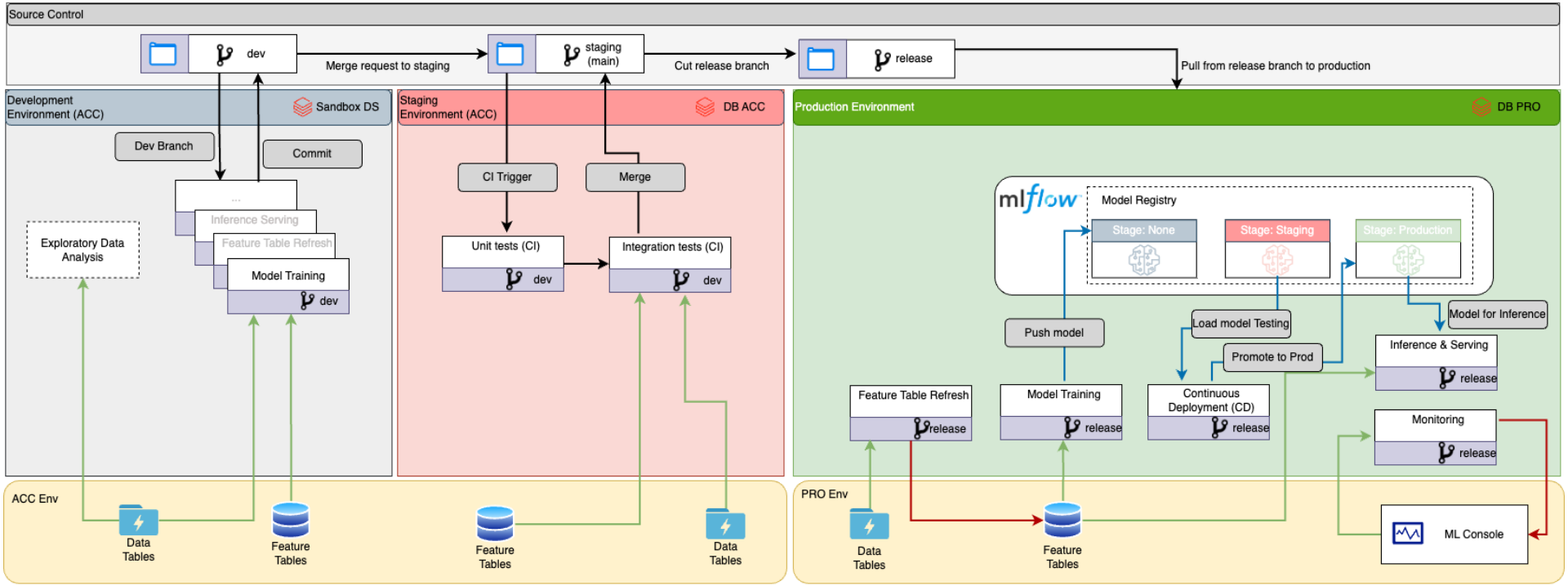
How do bundles work?

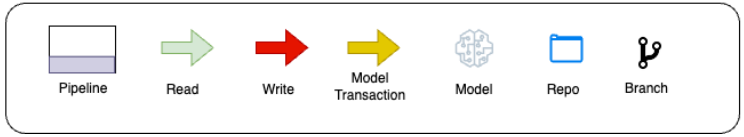
The **Databricks cli** has functions to **validate, deploy and run** Databricks Asset Bundles using bundle.yml files

Where are bundles used?

Bundles are useful during **development and CI/CD** processes

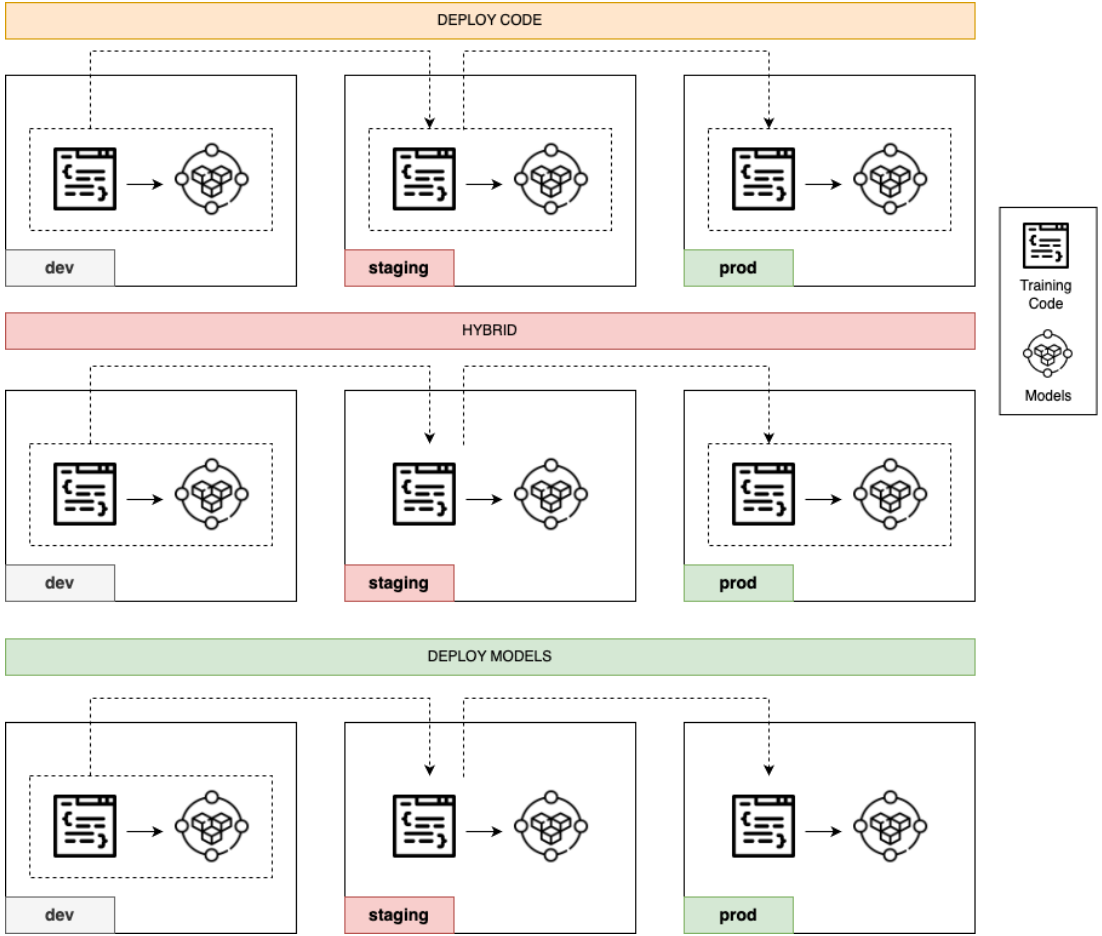
MLOps Steps





(*) The Big Book of MLOps

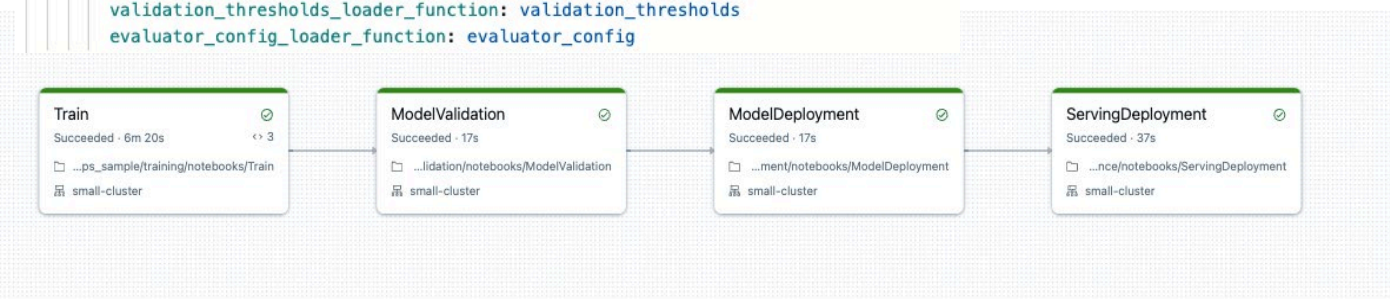
ML Deployment Patterns



Bundles Usage Example

```yaml
tasks:
- task_key: Train
  job_cluster_key: small-cluster
  notebook_task:
    notebook_path: ./../dap_mllab_mlops_sample/training/notebooks/Train.py
    base_parameters:
      env: ${bundle.environment}
      training_data_path: abfss://processed@datahub01dextdapdls.dfs.core.windows.net/PS
      experiment_name: ${var.experiment}
      model_name: ${var.model}
- task_key: ModelValidation
  job_cluster_key: small-cluster
  depends_on:
  - task_key: Train
  notebook_task:
    notebook_path: ./../dap_mllab_mlops_sample/validation/notebooks/ModelValidation.py
    base_parameters:
      env: ${bundle.environment}
      experiment_name: ${var.experiment}
      enable_baseline_comparison: "false"
      validation_input: abfss://processed@datahub01dextdapdls.dfs.core.windows.net/PS_S
      model_type: regressor
      targets: mean_squared_error
      custom_metrics_loader_function: custom_metrics
      validation_thresholds_loader_function: validation_thresholds
      evaluator_config_loader_function: evaluator_config
```

Taks generated by Use Cases to orchestrate in Databricks



DAG Tasks in Databricks

Bundles Usage Example

```yaml
# Validate bundle to be deployed to the staging workspace
- script: |
    databricks bundle validate -t test
  workingDirectory: $(repo_name)
  displayName: Validate bundle for test environment
  env:
    DATABRICKS_HOST: $(mllab-databricks-mlops-url)
    ARM_TENANT_ID: $(tenant-id)
    ARM_CLIENT_ID: $(case-sp-id)
    ARM_CLIENT_SECRET: $(case-sp-secret)

# Deploy bundle to staging workspace
- script: |
    databricks bundle deploy -t test
  workingDirectory: $(repo_name)
  displayName: Deploy bundle to test environment in staging workspace
  env:
    DATABRICKS_HOST: $(mllab-databricks-mlops-url)
    ARM_TENANT_ID: $(tenant-id)
    ARM_CLIENT_ID: $(case-sp-id)
    ARM_CLIENT_SECRET: $(case-sp-secret)

# Run Integration tests
- script: |
    databricks bundle run integration-tests-job -t test
  workingDirectory: $(repo_name)
  displayName: Run Integration Tests
  env:
    DATABRICKS_HOST: $(mllab-databricks-mlops-url)
    ARM_TENANT_ID: $(tenant-id)
    ARM_CLIENT_ID: $(case-sp-id)
    ARM_CLIENT_SECRET: $(case-sp-secret)
```

1. Validate bundle before deploy all code

2. Deploy code in correct environment

3. Run code (integration test, training, inference, etc)

DEMO CI/CD

DEMO

EXPLORER

DAP-MLLAB-MLOPS-CICD-DAIS

- dap-mllab-mlops-dais
  - azuredevops_cicd.yaml
  - azuredevops_test_ci.yml
- .flake8
- .gitignore
- clusters_definitions.yml
- cookiecutter.json
- copy_model.py
- deploy_job.py
- mypy.ini
- project_validation.py
- README.md
- requirements.txt
- test-requirements.txt

mypy.ini | deploy_job.py | copy_model.py | cookiecutter.json | clusters_definitions.yml | .gitignore | .flake8 | azuredevops_cicd.yaml | azuredevops_test_ci.yml

dap-mllab-mlops-dais > azuredevops_cicd.yaml

cicd Aa ab .* ? of 4

```yaml
        echo "##vso[task.setvariable variable=case-sp-id;]$CASE_SP_ID"
        echo "##vso[task.setvariable variable=case-sp-secret;]$CASE_SP_SECRET"
      workingDirectory: $(repo_name)
      displayName: 'Get Secrets & Set Secrets'
      env:
        DATABRICKS_TOKEN: $(dbmlops-databricks-token)
        DATABRICKS_HOST: $(mllab-databricks-mlops-url)

    # Validate bundle to be deployed to the prod workspace
    - script: |
        databricks bundle validate -t prod
      workingDirectory: $(repo_name)
      displayName: Validate bundle for prod environment
      env:
        DATABRICKS_HOST: $(mllab-databricks-mlops-url)
        ARM_TENANT_ID: $(tenant-id)
        ARM_CLIENT_ID: $(case-sp-id)
        ARM_CLIENT_SECRET: $(case-sp-secret)

    # Deploy bundle to prod workspace
    - script: |
        databricks bundle deploy -t prod
      workingDirectory: $(repo_name)
      displayName: Deploy bundle to prod environment
      env:
        DATABRICKS_HOST: $(mllab-databricks-mlops-url)
        ARM_TENANT_ID: $(tenant-id)
        ARM_CLIENT_ID: $(case-sp-id)
        ARM_CLIENT_SECRET: $(case-sp-secret)
```

PROBLEMS | OUTPUT | DEBUG CONSOLE | TERMINAL

```
 create mode 100644 test-requirements.txt
carlosrosado@carloss-MacBook-Pro dap-mllab-mlops-cicd-dais % git push origin feat/poc_cicd_dais
Enumerating objects: 18, done.
Counting objects: 100% (18/18), done.
Delta compression using up to 8 threads
Compressing objects: 100% (15/15), done.
Writing objects: 100% (16/16), 9.27 KiB | 9.27 MiB/s, done.
Total 16 (delta 0), reused 0 (delta 0), pack-reused 0
remote: Analyzing objects... (16/16) (21 ms)
remote: Validating commits... (1/1) done (14 ms)
remote: Storing packfile... done (58 ms)
remote: Storing index... done (89 ms)
To https://dev.azure.com/repsol-digital-team/DAPlatform01/_git/dap-mllab-mlops-cicd-dais
 * [new branch]      feat/poc_cicd_dais -> feat/poc_cicd_dais
carlosrosado@carloss-MacBook-Pro dap-mllab-mlops-cicd-dais %
```

OUTLINE

TIMELINE

feat/poc_cicd_dais | Team | 0 0 | Ln 335, Col 41 (24 selected) | Spaces: 2 | UTF-8 | LF | YAML

Issues and Solutions

Issues

MLFlow Recipes: Not for our DS

Models and endpoints deployed without control in names and capabilities

Cost reduction of the Deployed Databricks

Deploy Users groups for models and endpoints

Solutions

Model, Endpoints and Experiments auditory

Cost reduction of deployed Databricks: Hierarchical Endpoints

Deploy Users group with Terraform

Next Steps

Next Steps Roadmap

Feature Store

Data & Model Drift

Gen AI Visualization

Migrate All Models

Responsible AI

Cost Monitoring

Thank You

REPSOL