DATA+AI SUMMIT
BY databricks

MLOps at WGU:

Solutions to Production ML with Databricks

Zach Clement, Jonathan Bown

Date: TBD

OVERVIEW

- Challenges
- Project Goals and Features
- Architecture CI/CD
- User and Developer Components
- Demo
- Discussion

KEY CHALLENGES

Lack Of

Governance

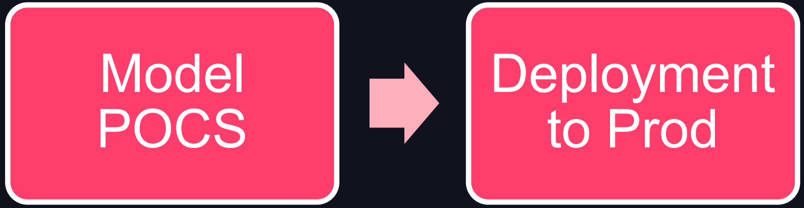
Automation

Traceability

Repeatability

Monitoring

A Gap Between



Model POCS

Deployment to Prod

GOALS AND FEATURES

Goals	Features
Self-governed data science environment	• Automated Project level resources and permissions access
Version Control	• Everything as code (ETL, workflows, compute, permissions)
Auditable	• Lineage tracking of data, workflows, experiments, models, code, permissions
Simplify productionalization using repeatable and standardized processes	• Automation via CI/CD - Dev/stage/prod environments • Orchestration of pipelines
Maintain model performance.	• Monitor batch inference models • Compare candidate models to current models • Data validation and profiling tools
Balance MLOps needs with Data Scientists skills	• Accommodate notebooks, widgets in workflows • Make it as usable as possible

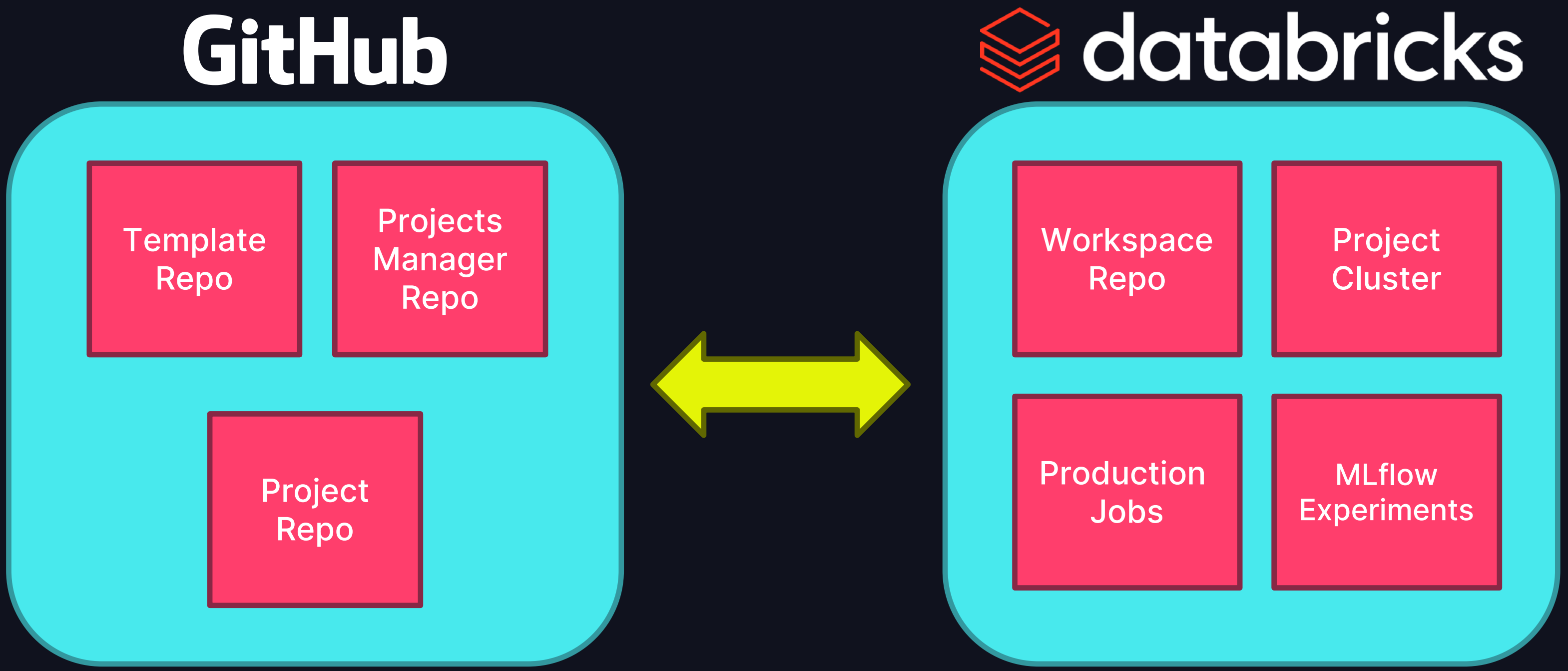
MARVIN – ML AND DATA OPS PLATFORM

Monitored, Auditable, Automated, Repeatable, Versioned, Intelligent

- The MARVIN platform is built to put models into production
  - Creating and maintaining project infrastructure
  - Providing tools for data scientists to streamline their development
  - Monitoring for workflow failures and communicating to stakeholders
  - Integrating Databricks features
  - Compatibility with wide variety of model types and frameworks
  - Modularity to allow rapid integration of new or changing feature requirements

MARVIN - MAJOR COMPONENTS

Monitored, Auditable, Automated, Repeatable, Versioned, Intelligent



PLATFORMS

Users Perspective

databricks

- Compute clusters
- Managed DEV/STAGE/PROD environments
- Notebook interface
- Python package to use in notebooks
- Integration of MLflow, data, workflows, model registry, experiments

GitHub

- CICD actions
- Templated project repository
- Collaborators, branch protections
- YML configuration files
- Automatic versioning with template & package
- Documentation

PLATFORMS

Developers perspective

databricks

- DBX deploy
- Databricks SDK
- DBFS for MLflow artifacts
- Unity Catalog for Data storage
- MLOps cluster for package development
- Jobs clusters

GitHub

- Template repo to instantiate projects
- Manager repo to maintain projects
- AWS Lambda for updating
- Copier templating
- GitHub API
- Test project for integration checks
- Version management
- Mkdocs documentation

MARVIN - Components

Projects Manager Repository

- Creates Databricks Assets
- Creates GitHub Assets
- Handles asset permissions & updates
- Python Package
- Token management
- AWS Lambda

Template Repository

- Copier Template
- Python Package
- Project Example Code
- Mkdocs Documentation
- Github Workflows

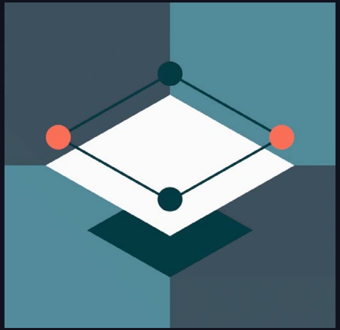
Open Source Tooling

- DBX
- MLflow
- Delta Tables
- Evidently
- Great Expectations
- Databricks SDK/CLI

SECURITY CONSIDERATIONS

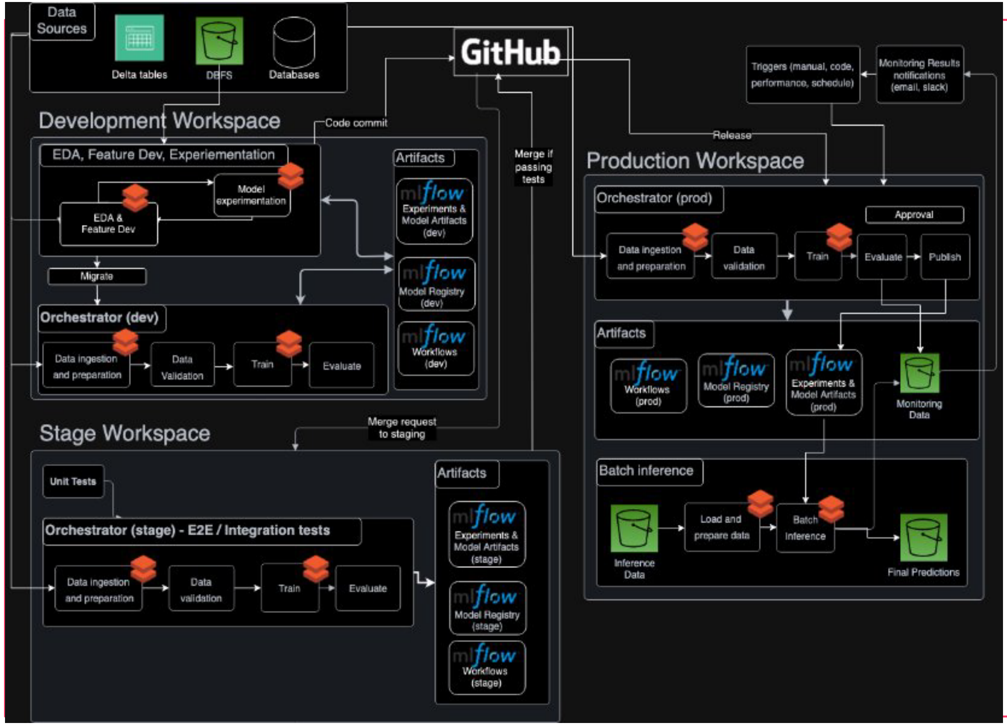
Constraints From WGU

- GitHub Organization Rules
  - Fine Grained Tokens
  - Approvals
  - Runner Limits
- Databricks Service Principals for Automation
- Databricks Permissions
  - Workspace
  - Experiment
  - Model
  - Data (Unity Catalog)



MARVIN ARCHITECTURE

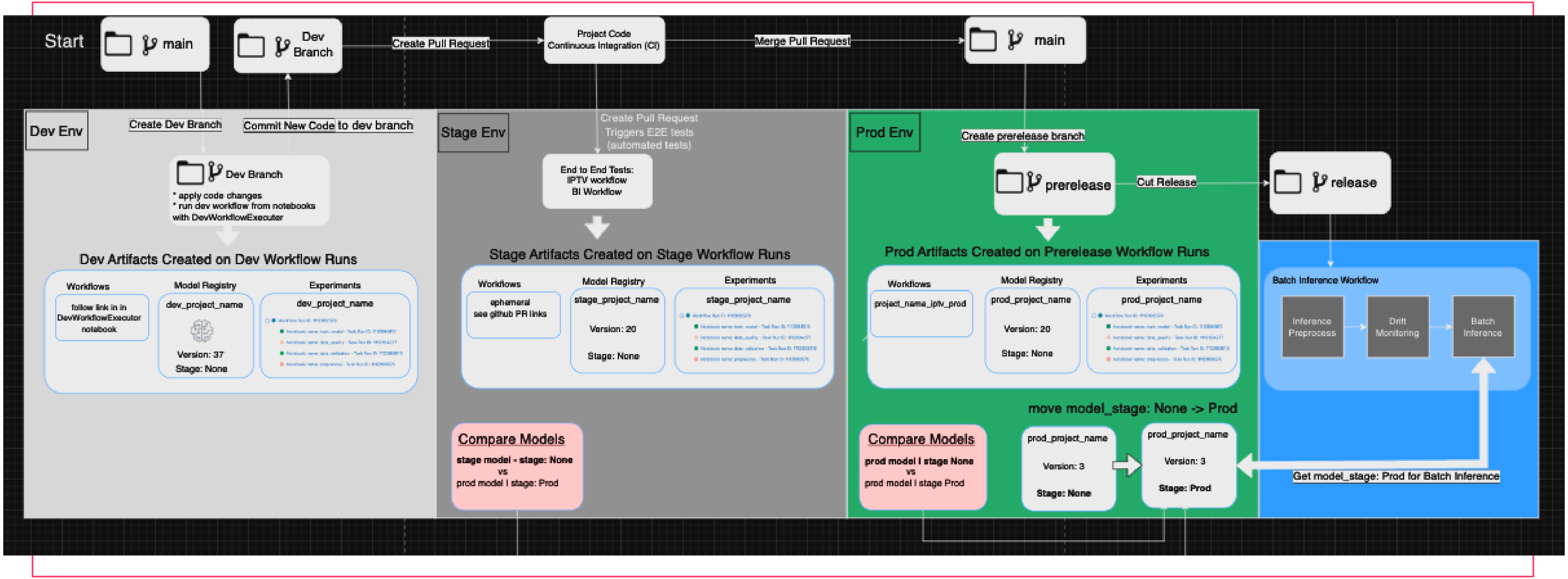
Single Project



- Input Data Sources
- CI/CD & Environments
- EDA & Experimentation (Dev)
- Orchestration
- Integration Tests (Stage)
- Inference
- Monitoring

CI/CD

Overview



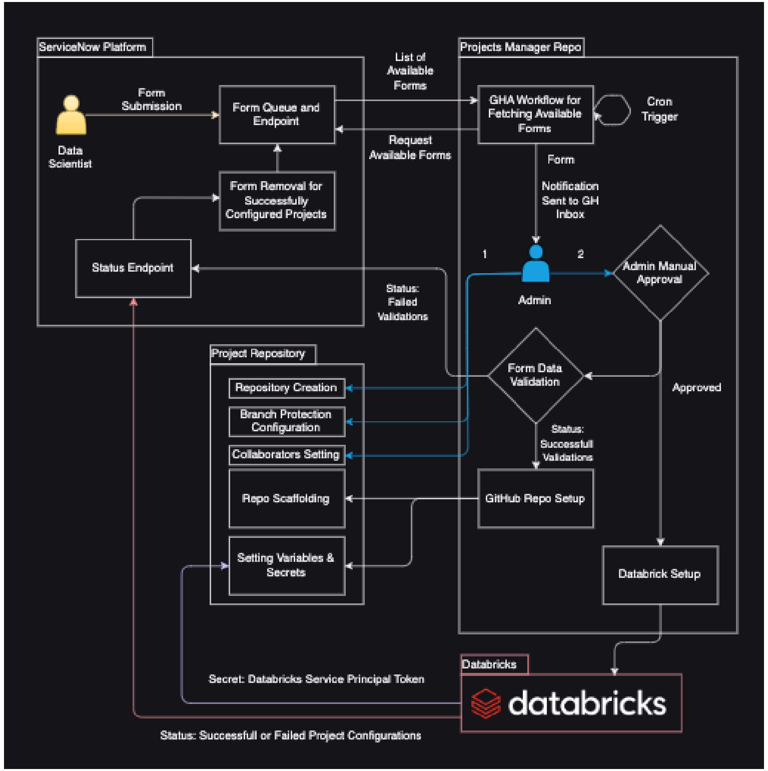
USER WORKFLOW

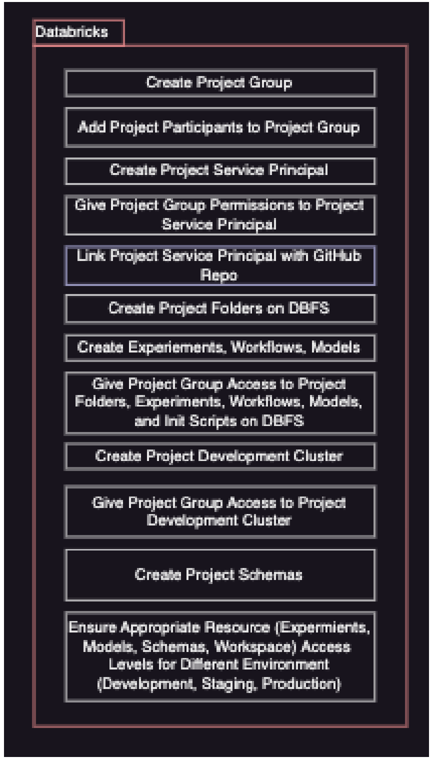
Steps to move from DEV to PROD

- User pulls repository into Databricks workspace
- Notebooks imported into 'notebooks' folder
- Update config/yml files
- Test in DEV using dev_workflow_executor notebook
- Move into STAGE by creating PR
    - Trigger e2e tests
    - Run as service principal
- Cut release using GitHub UI -> Deployed to Databricks in PROD environment

MARVIN ARCHITECTURE

Project Manager – New Project





Create Project Group

Add Project Participants to Project Group

Create Project Service Principal

Give Project Group Permissions to Project Service Principal

Link Project Service Principal with GitHub Repo

Create Project Folders on DBFS

Create Experiements, Workflows, Models

Give Project Group Access to Project Folders, Experiments, Workflows, Models, and Init Scripts on DBFS

Create Project Development Cluster

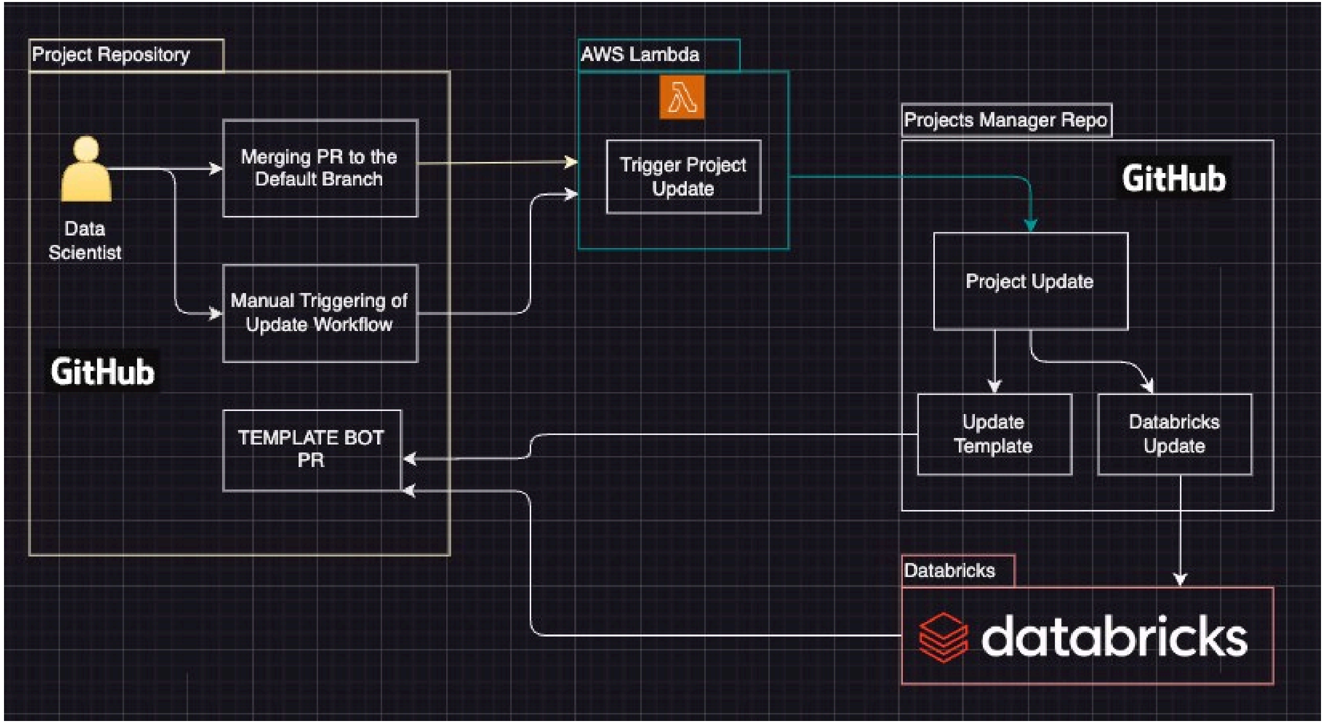
Give Project Group Access to Project Development Cluster

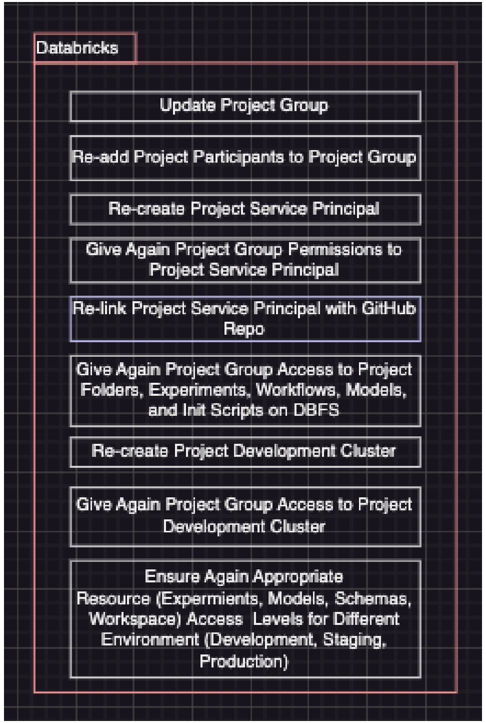
Create Project Schemas

Ensure Appropriate Resource (Expermients, Models, Schemas, Workspace) Access Levels for Different Environment (Development, Staging, Production)

MARVIN ARCHITECTURE

Project Manager – Existing Project Update





Update Project Group

Re-add Project Participants to Project Group

Re-create Project Service Principal

Give Again Project Group Permissions to Project Service Principal

Re-link Project Service Principal with GitHub Repo

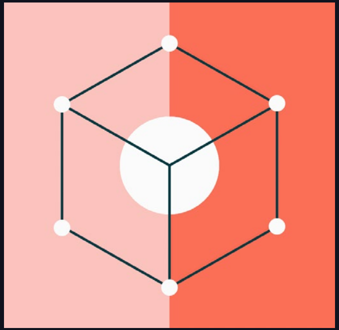
Give Again Project Group Access to Project Folders, Experiments, Workflows, Models, and Init Scripts on DBFS

Re-create Project Development Cluster

Give Again Project Group Access to Project Development Cluster

Ensure Again Appropriate Resource (Expermients, Models, Schemas, Workspace) Access Levels for Different Environment (Development, Staging, Production)

DEMO



SUCCESSES

How MARVIN is changing WGU

- Took existing project migration timeline from 2 months down to 1 week in 2023
- In 6 months, MARVIN already maintains 11 production models (goal was 5 for year 1)
- Errors in production are caught quickly with Opsgenie integrations
- Pending expansion to Dataops (after unity catalog integration)
- More collaboration between data science and engineering departments
  - Upskilling Analyst and Data Scientist roles in GitHub and Databricks
  - Less separation between data science experimentation and model production
- WGU is now moving toward more inhouse ML solutions instead of 3rd party tools

QUESTIONS?

RESOURCES

DATA+AI SUMMIT

THANK YOU

- Zach Clement – zach.clement@wgu.edu
- Jonathan Bown – jonathan.bown@wgu.edu

WGU
WESTERN GOVERNORS UNIVERSITY