**Project Book — Databricks Asset Bundle (DAB) for GenAI & ML**

Production-ready patterns for ML and LLM/GenAI on Azure Databricks

Version: 1.0 | Date: 2025-09-10

# 1. Executive Summary

This Project Book defines a production-grade template and runbook for delivering machine learning (ML) and generative AI (GenAI) applications using Databricks Asset Bundles (DAB). It separates concerns between ML (data science, modeling, evaluation) and MLOps (platform, deployment, governance, observability), and codifies the repository structure, bundle configuration (databricks.yml), CI/CD, security, and operations.

# 2. Scope & Audience

* Audience: ML Engineers, Data Scientists, Platform/MLOps Engineers, Security & Governance, Product Owners.
* Scope: Applies to both traditional ML (forecasting, classification, etc.) and GenAI (RAG, agents, copilots).
* Cloud/Platform: Azure Databricks with Unity Catalog, MLflow, Mosaic AI, Entra ID, Serverless Compute.

# 3. Architecture Overview

3.1 Four-layer Reference (from Project Charter)

* Data Foundation (GraphRAG): Delta Lake + Knowledge Graph (Spark GraphX/GraphFrames) for long-term memory.
* Tooling & Integration (Model Context Protocol, MCP): standardized, secure tool connectors.
* Intelligence & Reasoning (Multi-Agent System): OpenAI Assistants/Agents orchestration with LangChain.
* Foundational LLM (Hosted in North America): choice of Llama, GPT, Claude; governed access via Mosaic AI.

3.2 Confirmed Stack (from Project Charter)

* Cloud: Microsoft Azure; Platform: Databricks (Serverless where possible).
* Data & AI Platform: Mosaic AI; Governance: Unity Catalog; Tracking: MLflow.
* AuthN/Z: Microsoft Entra ID (SSO, OAuth/M2M).
* Agent Framework: OpenAI Assistants, LangChain; UI: Gradio (or app frameworks).

# 4. ML vs MLOps — Separation of Concerns

4.1 ML (Product/Use-case scope)

* Define business problem & KPIs; curate datasets; feature engineering; model training & evaluation.
* Prompt/agent design, RAG pipelines, tool selection (MCP), offline & online evaluation.
* Register models & app artifacts in MLflow / Unity Catalog; author notebooks & src modules; tests.

4.2 MLOps (Platform/Operations scope)

* Bundle configuration (databricks.yml), environments (dev/stage/prod) and deployment modes.
* Provision & manage compute (serverless where possible), permissions, and secret management.
* Observability (MLflow Tracing), model/app serving, cost/usage tracking via Mosaic AI Gateway.
* CI/CD automation, policy checks, data/AI governance via Unity Catalog.

# 5. Databricks Asset Bundles (DAB) Overview

5.1 Why Bundles

* Declarative, versioned deployments for jobs, pipelines, models, serving endpoints, and related resources.
* Environment-specific targets (dev/stage/prod) with validation and deployment modes.
* Repeatable automation in CI/CD with bundle validate/deploy/run commands.

5.2 Key Files & Concepts

* databricks.yml at repo root defines bundle name, targets, artifacts, variables, workspace settings.
* resources section defines jobs, pipelines, models, connections, volumes, dashboards, apps.
* targets allow environment-specific overrides (compute, permissions, schedules, variables).

# 6. Repository & Project Folder Structure

Use the following structure for any ML or GenAI project. The /bundle folder contains DAB configuration; ML code is modularized under /src; notebooks are for exploration; pipelines and tests are first-class citizens.

/ (repo root)  
├─ app/  
│ ├─ app.py # Gradio/Streamlit or FastAPI UI entry point  
│ ├─ settings.py # UI config (reads from env/Unity Catalog connections)  
│ └─ prompts/ # Prompt templates (governed via UC Volumes)  
│ ├─ system/agent\_system.md  
│ └─ rag/retrieval\_prompt.md  
│  
src/  
├─ config/ # Centralized configuration & registries  
│ ├─ settings.py # App/env settings (env vars, UC connections)  
│ ├─ secrets\_provider.py # Secure secret access helpers (UC connections/Scopes)  
│ ├─ registry.py # Paths, model names, feature tables, UC objects  
│ └─ \_\_init\_\_.py  
│  
├─ logging/ # Structured logging, tracing, and instrumentation  
│ ├─ logger.py # Python logging config (JSON, levels, handlers)  
│ ├─ mlflow\_logging.py # MLflow run helpers, params/metrics/artifacts  
│ ├─ tracing.py # MLflow Tracing spans for LLM/agent tool calls  
│ └─ \_\_init\_\_.py  
│  
├─ data/ # Data IO, preprocessing, schemas  
│ ├─ loaders/  
│ │ ├─ delta\_reader.py # Read/write Delta tables (Unity Catalog)  
│ │ ├─ sql\_loader.py # SQL/Databricks SQL queries  
│ │ ├─ file\_loader.py # PDFs/Docs/JSON/CSV ingestion  
│ │ └─ \_\_init\_\_.py  
│ ├─ processing/  
│ │ ├─ cleaners.py # Text/column cleaning, null handling  
│ │ ├─ transformers.py # Feature transforms, tokenization  
│ │ └─ \_\_init\_\_.py  
│ ├─ schemas/  
│ │ ├─ pydantic\_models.py # Pydantic data contracts for IO & API  
│ │ ├─ table\_schemas.py # Spark/Delta schemas, expectations  
│ │ └─ \_\_init\_\_.py  
│ ├─ graph/  
│ │ ├─ graph\_builder.py # GraphRAG: GraphFrames/GraphX builders  
│ │ └─ \_\_init\_\_.py  
│ └─ \_\_init\_\_.py  
│  
├─ evaluation/ # Metrics, evaluators, and reports  
│ ├─ metrics/  
│ │ ├─ classification.py # Precision/Recall/F1/ROC  
│ │ ├─ regression.py # RMSE/MAE/R2  
│ │ ├─ rag.py # Retrieval precision/recall, MRR, nDCG  
│ │ ├─ generative.py # LLM quality: BLEU/ROUGE/BERTScore/hallucination  
│ │ └─ \_\_init\_\_.py  
│ ├─ evaluators/  
│ │ ├─ offline\_eval.py # Batch eval against gold data  
│ │ ├─ online\_eval.py # Shadow/AB eval hooks, SLO checks  
│ │ └─ \_\_init\_\_.py  
│ ├─ reports/  
│ │ ├─ report\_writer.py # HTML/Markdown/Delta‑logged reports  
│ │ └─ \_\_init\_\_.py  
│ └─ \_\_init\_\_.py  
│  
├─ ai/ # Unified area for traditional ML + GenAI  
│ ├─ training/  
│ │ ├─ sklearn\_trainer.py # Classical ML trainers  
│ │ ├─ pytorch\_trainer.py # Deep learning trainers  
│ │ ├─ finetune\_trainer.py # LLM fine‑tuning / PEFT trainers  
│ │ └─ \_\_init\_\_.py  
│ ├─ inference/  
│ │ ├─ batch\_scoring.py # Batch inference  
│ │ ├─ realtime\_scoring.py # Real‑time/serving adapters  
│ │ ├─ router.py # Route to model vs. LLM/agent  
│ │ └─ \_\_init\_\_.py  
│ ├─ embeddings/  
│ │ ├─ embedding\_client.py # Mosaic AI/OpenAI/HF embeddings  
│ │ └─ \_\_init\_\_.py  
│ ├─ retrieval/  
│ │ ├─ vector\_index.py # Vector Search index CRUD  
│ │ ├─ retriever.py # Hybrid retrieval strategies  
│ │ └─ \_\_init\_\_.py  
│ ├─ prompts/  
│ │ ├─ templates.py # Programmatic prompt assembly  
│ │ └─ \_\_init\_\_.py  
│ ├─ agents/  
│ │ ├─ assistants.py # Multi‑agent orchestration  
│ │ ├─ tools\_mcp.py # MCP tool registrations & guards  
│ │ └─ \_\_init\_\_.py  
│ ├─ generation/  
│ │ ├─ llm\_client.py # Gateway client (Mosaic AI Gateway/provider SDK)  
│ │ └─ \_\_init\_\_.py  
│ ├─ registry/  
│ │ ├─ mlflow\_registry.py # Register/promote models, alias mgmt  
│ │ └─ \_\_init\_\_.py  
│ └─ \_\_init\_\_.py  
│  
├─ pipelines/ # Reusable orchestration units (invoked by jobs)  
│ ├─ training\_pipeline.py  
│ ├─ validation\_pipeline.py  
│ ├─ testing\_pipeline.py  
│ ├─ inference\_pipeline.py  
│ └─ \_\_init\_\_.py  
│  
├─ serving/ # Serving/adapters and request validation  
│ ├─ endpoints.py  
│ ├─ validators.py # Pydantic validators for API payloads  
│ └─ \_\_init\_\_.py  
│  
├─ utils/ # Small cross‑cutting helpers  
│ ├─ io.py  
│ ├─ decorators.py  
│ ├─ constants.py  
│ └─ \_\_init\_\_.py  
│  
└─ \_\_init\_\_.py  
│  
├─ bundle/  
│ ├─ databricks.yml # Bundle root configuration  
│ └─ resources/ # Split resource files (optional)   
│ ├─ jobs.yml  
│ ├─ pipelines.yml  
│ ├─ connections.yml  
│ ├─ serving\_endpoints.yml  
│ └─ permissions.yml  
│  
├─ docs/  
│ ├─ index.md  
│ └─ architecture.md  
│  
├─ .github/  
│ └─ workflows/  
│ └─ databricks\_deploy.yml # CI/CD for validate/deploy/run  
│  
├─ requirements.txt  
├─ README.md  
└─ LICENSE

# 7. Sample Files — Purpose & Starters

7.1 bundle/databricks.yml (skeleton)

bundle:  
 name: genai-ml-bundle  
 # optional: pin CLI version  
 # databricks\_cli\_version: '0.218.0'  
  
workspace:  
 host: https://<your-workspace>.azuredatabricks.net  
 artifact\_path: /Volumes/<catalog>/<schema>/<volume>/artifacts  
  
variables:  
 project\_name:  
 default: genai-app  
 env:  
 default: dev  
  
artifacts:  
 app\_wheel:  
 path: .  
 build: 'python -m pip wheel -w dist .'  
  
resources:  
 jobs:  
 training\_job:  
 name: "${var.project\_name}-train"  
 job\_clusters:  
 - job\_cluster\_key: serverless  
 serverless: true  
 tasks:  
 - task\_key: train  
 job\_cluster\_key: serverless  
 python\_wheel\_task:  
 package\_name: genai\_app  
 entry\_point: pipelines.training\_pipeline  
 libraries:  
 - whl: ./dist/\*.whl  
 permissions:  
 - level: CAN\_MANAGE\_RUN  
 group\_name: ml-engineers  
  
 serving\_endpoints:  
 rag\_service:  
 name: "${var.project\_name}-endpoint"  
 config:  
 served\_models:  
 - model\_name: uc://<catalog>.<schema>.models.rag\_app  
 model\_version: '1'  
 workload\_size: Small  
 scale\_to\_zero\_enabled: true  
  
 connections:  
 openai\_gateway:  
 name: ai-gateway  
 connection\_type: ai\_gateway  
  
targets:  
 dev:  
 mode: development  
 variables:  
 env: dev  
 stage:  
 mode: development  
 variables:  
 env: stage  
 prod:  
 mode: production  
 variables:  
 env: prod

7.2 .github/workflows/databricks\_deploy.yml (CI/CD skeleton)

name: DAB Validate & Deploy  
on:  
 push:  
 branches: [ main ]  
 pull\_request:  
 branches: [ main ]  
  
jobs:  
 validate-deploy:  
 runs-on: ubuntu-latest  
 steps:  
 - uses: actions/checkout@v4  
 - name: Set up Python  
 uses: actions/setup-python@v5  
 with:  
 python-version: '3.11'  
 - name: Install Databricks CLI  
 run: |  
 pip install databricks-cli  
 - name: Auth (OAuth M2M)  
 run: |  
 databricks auth login --host ${{ secrets.DATABRICKS\_HOST }} --oauth-client-id ${{ secrets.AZURE\_SP\_CLIENT\_ID }} --oauth-client-secret ${{ secrets.AZURE\_SP\_CLIENT\_SECRET }} --interactive false  
 - name: Validate bundle  
 run: databricks bundle validate -t prod  
 - name: Deploy bundle  
 if: github.ref == 'refs/heads/main'  
 run: databricks bundle deploy -t prod

# 8. Governance & Security

* Unity Catalog: single source of truth for data, models, volumes; assign privileges to groups; avoid user-level grants.
* Use service principals for automation; manage identities in Entra ID; prefer OAuth tokens over PATs.
* Store secrets in Unity Catalog connections or secret scopes; never commit secrets.
* Adopt serverless compute for least-privilege data plane and simpler ops; lock down production endpoints.

# 9. Observability & Monitoring

* MLflow: experiment tracking, model registry (aliases like @champion/@challenger), lineage to runs.
* Mosaic AI Gateway: usage tracking, payload logging (Delta tables), guardrails, rate limits, fallbacks.
* MLflow Tracing: capture prompts, tool calls, latencies; wire into dashboards and alerts.
* Quality monitors: schedule evaluation jobs; track costs/latency & business KPIs.

# 10. Testing & Quality Gates

* Unit tests for src modules; integration tests for end-to-end pipelines; security tests for auth and inputs.
* Bundle validation in CI; policy checks (lint YAML, prohibited settings).
* Offline eval datasets; automated acceptance thresholds (retrieval precision/recall, hallucination rate).
* Blue/green model rollout using model registry aliases; automatic rollback on SLO breaches.

# 11. Runbooks & Operational Procedures

Common Databricks Asset Bundle commands:

databricks bundle validate -t <target>  
databricks bundle deploy -t <target>  
databricks bundle run <resource-name> -t <target>  
databricks bundle destroy -t <target>

# 12. Appendices

A. Sample RAG application structure (adapted)

See notebooks and src layout above; align with your current RAG structure to ensure parity with model tracking and retrieval components.

B. MCP Tooling Layer

* Expose enterprise tools via MCP servers; register tools in agents/tools\_mcp.py; enforce auth via Unity Catalog connections.
* Log tool I/O via MLflow Tracing for auditability and evaluation.