RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Go, change the world

Databricks

A Unified Analytics Platform

Faculty Mentor:
Dr.S.Anupama Kumar
Associate Professor
Dept of AIML

Presented by:
Aditya tekriwal
P Shreyas
Jaswanth
Gagan gowda V S

databricks

RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

What is DataBricks ?

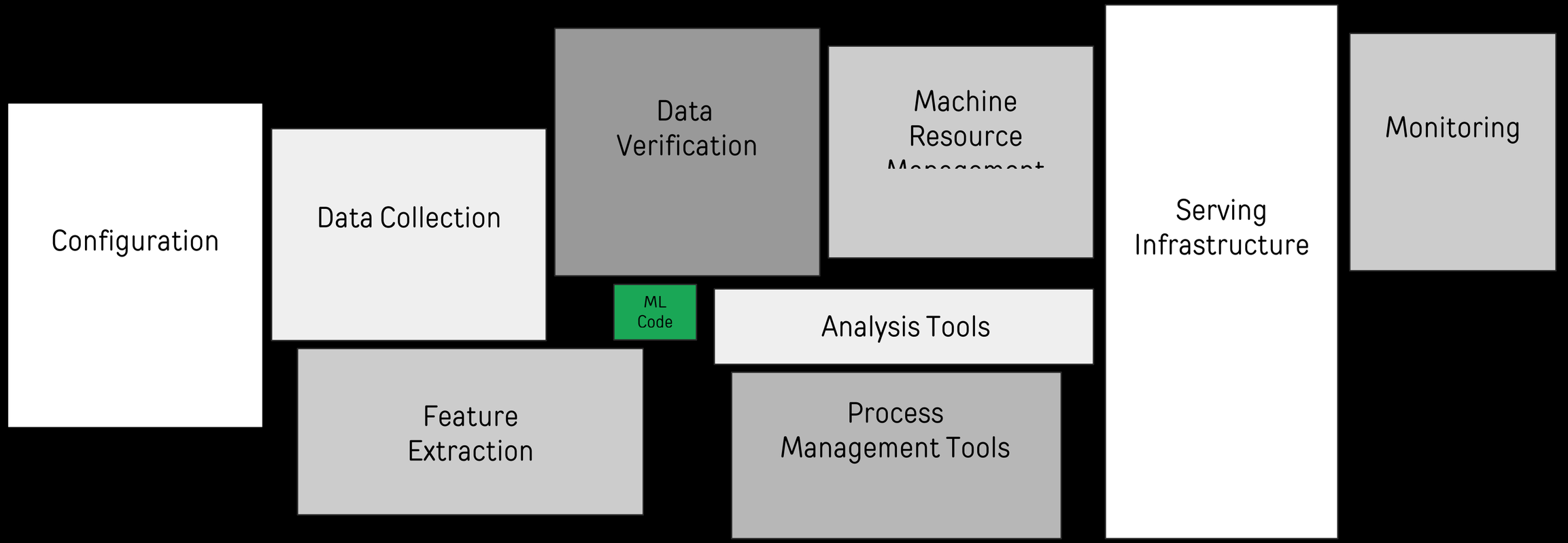
Databricks is a unified, cloud-based platform designed for data analytics and artificial intelligence (AI). Designed to have all the features necessary for an MLOPs life cycle when combined with a cloud service provider.

Databricks offers a platform for other workloads, including machine learning, data storage and processing, streaming analytics, and business intelligence

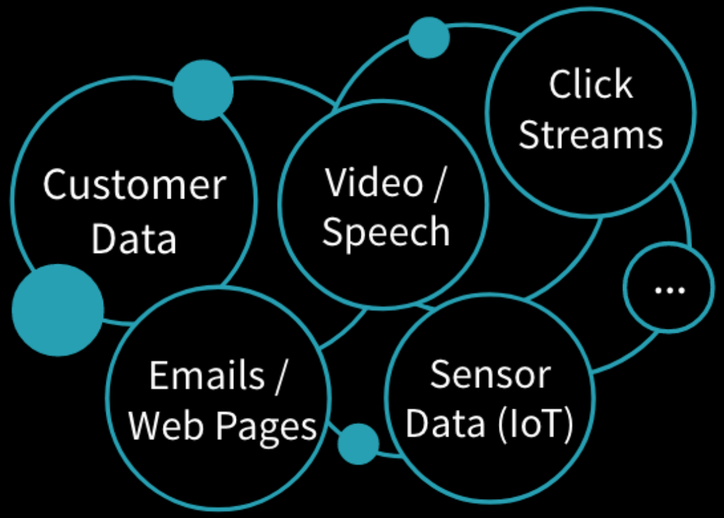
databricks

Hardest Part of AI isn't AI, it's Data

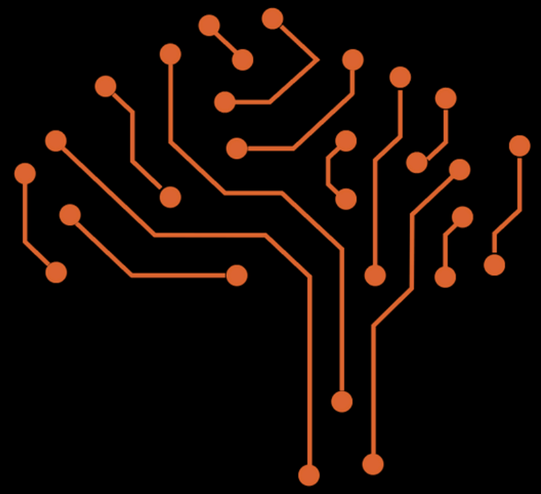
"Hidden Technical Debt in Machine Learning Systems," Google NIPS 2015



Data & AI Technologies are in Silos



X



SQL hadoop mongoDB Parquet

kafka ORACLE TERADATA

Great for Data, but not AI

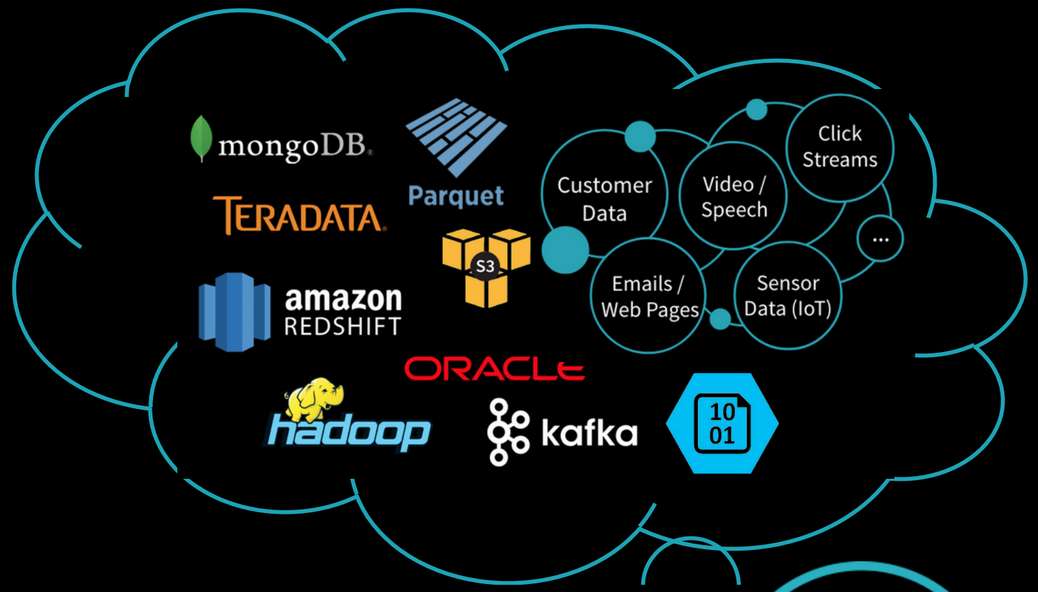
K Keras scikit learn PYTORCH

TensorFlow™ mxnet

Great for AI, but not for data

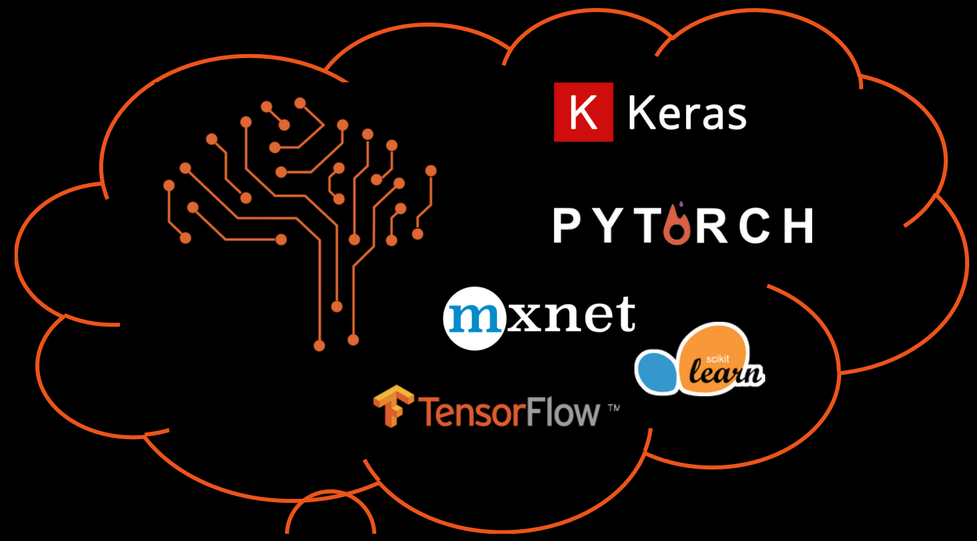
databricks

Data & AI People are in Silos



DATA ENGINEERS

X



DATA SCIENTISTS

RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Go, change the world

DataBricks and Cloud

AZURE

Blob Storage

Data Lake Store

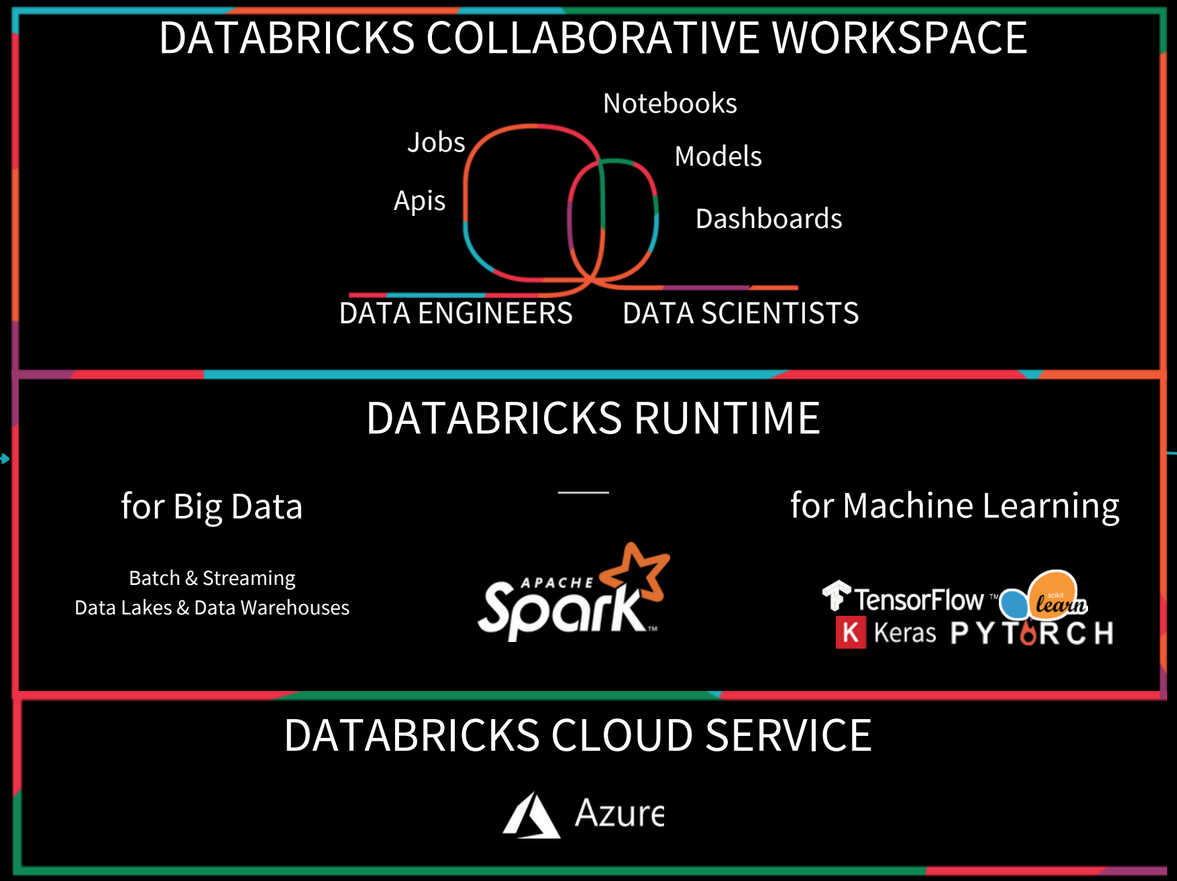
SQL Data Warehouse

Cosmos DB

Event Hub

IoT Hub

Azure Data Factory



Power BI

BI Reporting
Dashboards

Microsoft Active Directory

Security Integration

databricks

RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Free vs Paid

Features	Free	Paid
Experiment Tracking	Full Support	Full support
Pipeline Orchestration	None	Custom Options
Dataset Managment	Basic	Custom Options
Model Deployment	None	Custom Support
Colaboration	Can share only Notebooks	Full Support
Security	Basic	Can Integrate third party options
Scalability	Just for experiments	Scales As the compute
Support	Community(stackOverflow,Reddit)	Custom Call Support
Custom Storage	Free 15 gb	Fully customizable
Cost	Free	Depeds on Workload
ChatBot Support for Notebooks	Available	Avalaible

Compatibility:

Entirely Cloud based

Cloud Partners For:

- Compute
- Storage
- Deployment

Languages:

Default Version: 3.10

Default Version: 4.3.1

Default Version: 2.12

For Experimentation

Compatible with ANSI SQL:2003

For Data Manipulation

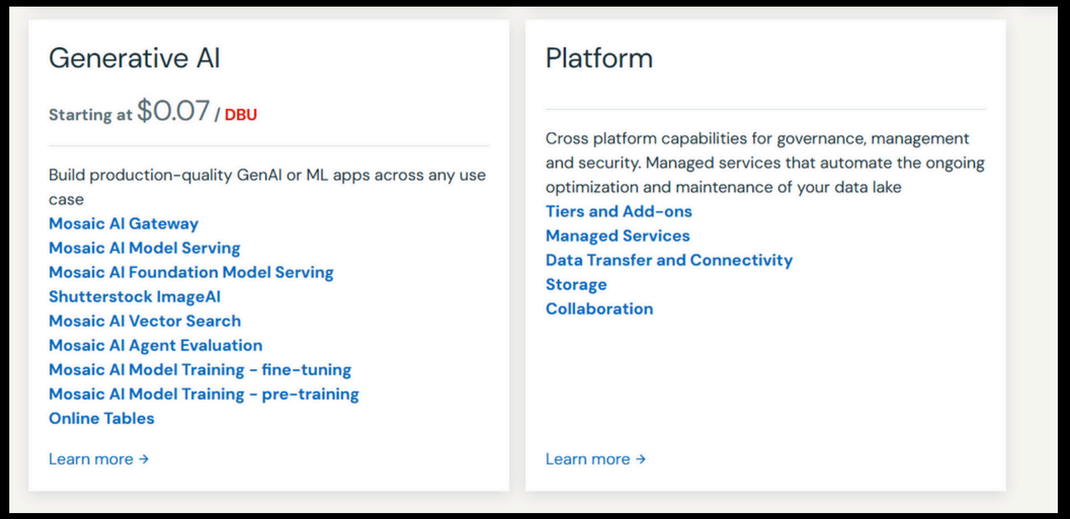
Enterprise Cost Options:

What is a DBU?

A Databricks Unit (DBU) is a normalized unit of processing power on the Databricks Lakehouse Platform used for measurement and pricing purposes. The number of DBUs a workload consumes is driven by processing metrics, which may include the compute resources used and the amount of data processed.

Practical Examples:

- Standard_F4 (4 cores, 8GB RAM) = 0.75 DBUs/hour
- Standard_DS3_v2 (4 cores, 14GB RAM) = 1 DBU/hour
- Memory Optimized (8 cores, 64GB RAM) = 2.4 DBUs/hour



How to create an account and get started on Databricks?

DEMO

Key Features of DataBricks

Data Processing & Analytics

- Apache Spark IntegrationNative Spark execution environment
- Optimized Spark runtime
- Interactive data processing
- Support for batch and streaming data

Delta Lake

- SupportACID transactions
- Schema enforcement
- Time travel (data versioning)
- Optimized performance with Delta engine
- Merge, update, and delete operations

Development Environment

- Collaborative NotebooksMultiple language support (Python, R, SQL, Scala)
- Real-time collaboration
- Version control integration
- Code snippets and templates
- Markdown documentation support

Integrated Development Tools

- Interactive data visualization
- Built-in SQL query editor
- Git integration
- Job scheduling and monitoring
- Command palette for quick actions

Machine Learning Features

- MLflow IntegrationExperiment tracking
- Model versioning
- Model registry
- Deployment management
- Parameter logging

AutoML

- CapabilitiesAutomated feature engineering
- Model selection
- Hyperparameter tuning
- Model evaluation
- Feature importance analysis

Key Features of DataBricks

Data Science Tools

- Built-in LibrariesPopular ML frameworks (scikit-learn, TensorFlow, PyTorch)
- Data manipulation libraries (Pandas, NumPy)
- Visualization tools (Matplotlib, Seaborn)
- Statistical analysis packages

Workspace Management

- Project organization
- Access control
- Resource management
- Cluster configuration
- Environment management

Enterprise Features

- Security & GovernanceRole-based access control
- Audit logging
- Data encryption
- Compliance controls
- Token-based authentication

Integration Capabilities

- Cloud service integration
- CI/CD pipeline support
- API access
- External tool connectivity
- Data source connections

Performance Features

- Optimization ToolsQuery optimization
- Caching mechanisms
- Resource allocation
- Performance monitoring
- Cost management

Data Engineering

- ETL/ELT CapabilitiesData pipeline creation
- Workflow orchestration
- Job scheduling
- Error handling
- Data quality checks

Key Features of DataBricks

Cost Management

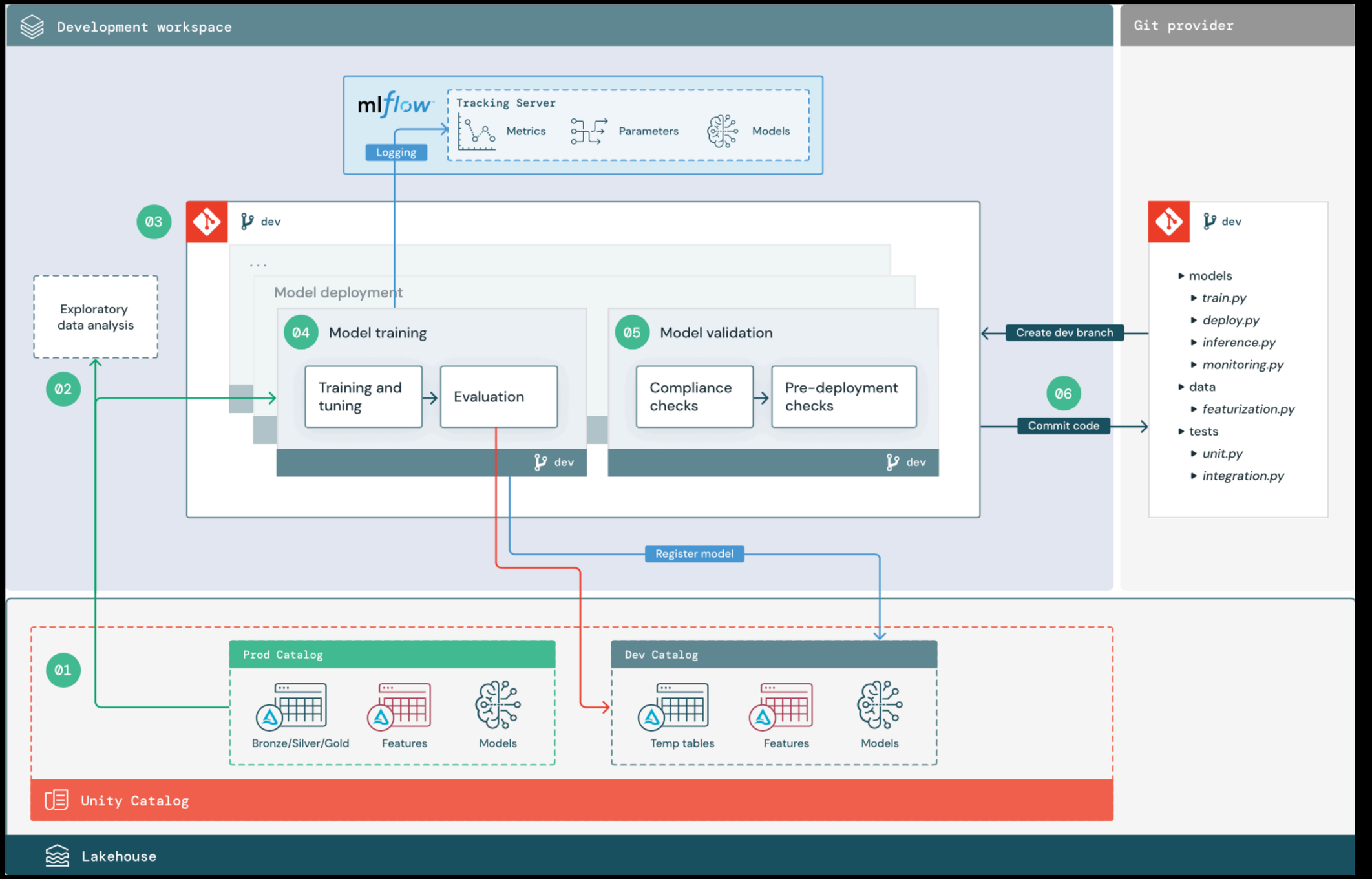
- Resource Monitoring Cluster utilization tracking
- Cost analysis
- Usage reporting
- Budget controls
- Optimization recommendations

Community Support

- Learning Resources Documentation
- Sample notebooks
- Training materials
- Community forums
- Knowledge base

databricks

In an MLOPS workflow



RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Limitation and Drawbacks

1. Resource Limitations

- Challenge: Community Edition is restricted to single-node clusters with limited computing power and 15GB storage.
- Impact: Teams working with large datasets or complex computations may experience performance bottlenecks.
- Mitigation: Optimize code for efficiency and use data sampling techniques for development.

2. Enterprise Features Restriction

- Challenge: Many advanced features (Docker, Kubernetes, MLflow serving, job scheduling) are only available in paid versions.
- Impact: Teams may find themselves limited in deployment options and production-grade capabilities.
- Mitigation: Consider upgrading to paid version for production deployments or explore alternative open-source solutions.

3. Collaboration and Version Control

- Challenge: Limited collaboration features in Community Edition, with restrictions on sharing and workspace management.
- Impact: Teams may struggle with code sharing, version control, and maintaining collaborative workflows.
- Mitigation: Utilize external version control systems (Git) and maintain clear documentation for team coordination.

RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Limitation and Drawbacks

5. Security and Access Control

- Challenge: Basic security features only; advanced security, SSO, and fine-grained access control require paid versions.
- Impact: Organizations with strict security requirements may find Community Edition insufficient.
- Mitigation: Implement additional security measures at the application level or consider paid versions for sensitive data.

6. Learning Curve and Support

- Challenge: Complex ecosystem with multiple components (Spark, Delta Lake, MLflow) requires significant learning.
- Impact: New users may take time to become productive, especially with distributed computing concepts.
- Mitigation: Utilize available documentation, community resources, and start with simpler workflows before advancing.

RV College of Engineering®
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Our Model

Demo

End