Databricks

Matthew Edwards

Content

1. Believing in Unicorns
2. Data Science Platform
3. Machine Learning

Databricks

1. Believing in Unicorns

Data Science Unicorn

- Scope Projects
- Import Data from Data Sources
- Clean and Wrangle Data
- Build and Evaluate Data Products
- Deploy Data Products
- Monitor and Maintain Data Products

Scope Projects

1. Objectives
2. Deliverables
3. Resources

- Data Science Team Lead (Data Science Process Alliance)
- Project Management (Google Coursera)

Import Data from Data Sources

- Import Data from Databases
- Import Data from Directories
- Change Data Capture

Clean and Wrangle Data

- Clean and Wrangle Data
- Build Data Pipelines
- Manage Storage and Data

Build and Evaluate Data Products

1. Build Data Products
   - Statistical Models
   - Machine Learning Models
   - Deep Learning Models
   - Dashboards
2. Evaluate Data Products
   - Model Performance Testing
   - Dashboard Usability Testing

Deploy Data Products

1. Deploy Models
   - Model Registries
   - Feature Stores
2. Deploy Dashboards
   - Design

- Designing Machine Learning Systems (Chip Huyen)

Monitor and Maintain Data Products

- Monitor
  - Model Performance
  - System Performance
  - Data Shift
- Maintain
  - Repair and Retrain Models
  - Production Testing

- Designing Machine Learning Systems (Chip Huyen)

End-to-End Data Science Specialist

- A **data science specialist** is someone that specialises in either statistics, machine learning, deep learning or dashboarding
- An **end-to-end data scientist** is someone that can develop, deploy, monitor and maintain data pipelines from data source to data product
- An end-to-end data data science specialist can exist with the right **data science platform**

Databricks

2. Data Science Platform

Background

- Data Warehouse
  - Data: Structured (e.g., tables)
  - Storage: High-cost / Managed
  - Processing: Schema on Write (ETL)
- Data Lake
  - Data: Unstructured (e.g., images)
  - Storage: Low-cost / Unmanaged
  - Processing: Schema on Read (ELT)

Data Lakehouse

- Data: Structure and Unstructured
- Storage: Low-cost / Managed
- Processing: Schema on Write*
- **Best of Both Worlds**

Databricks Data Science Platform

- Simple
- Open
- Multi-cloud

Simple

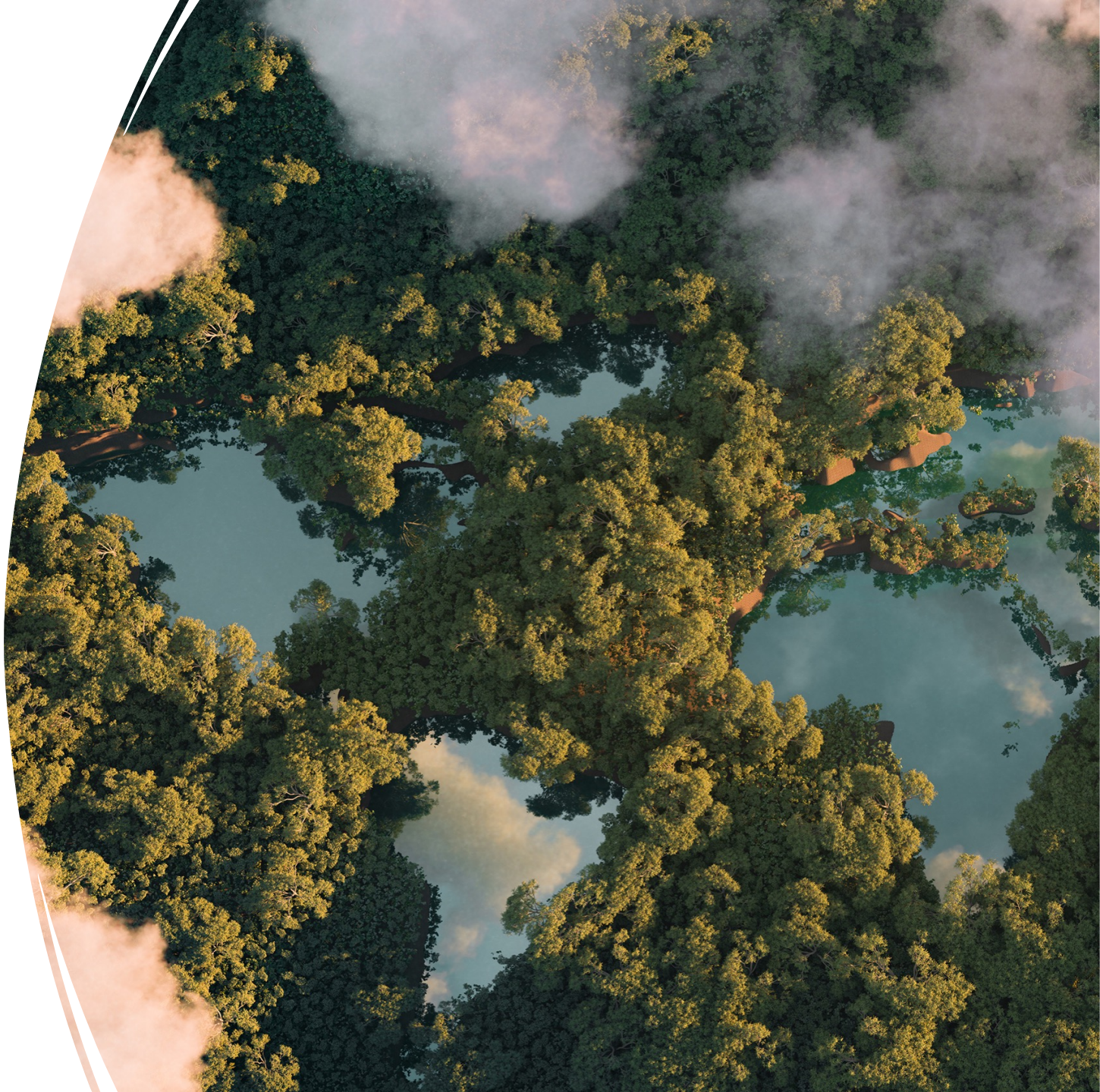
- Data Lakehouse
- Combines Lake and Warehouse
- Unity Catalog (Governance)

Open

- Apache Spark for Computation
- Delta Lake for Storage
- MLflow for Machine Learning

Multi-Cloud

- Microsoft Azure
- Amazon Web Service
- Google Cloud Platform



Personas

- Data Science & Engineering
- Machine Learning
- SQL

Data Science & Engineering

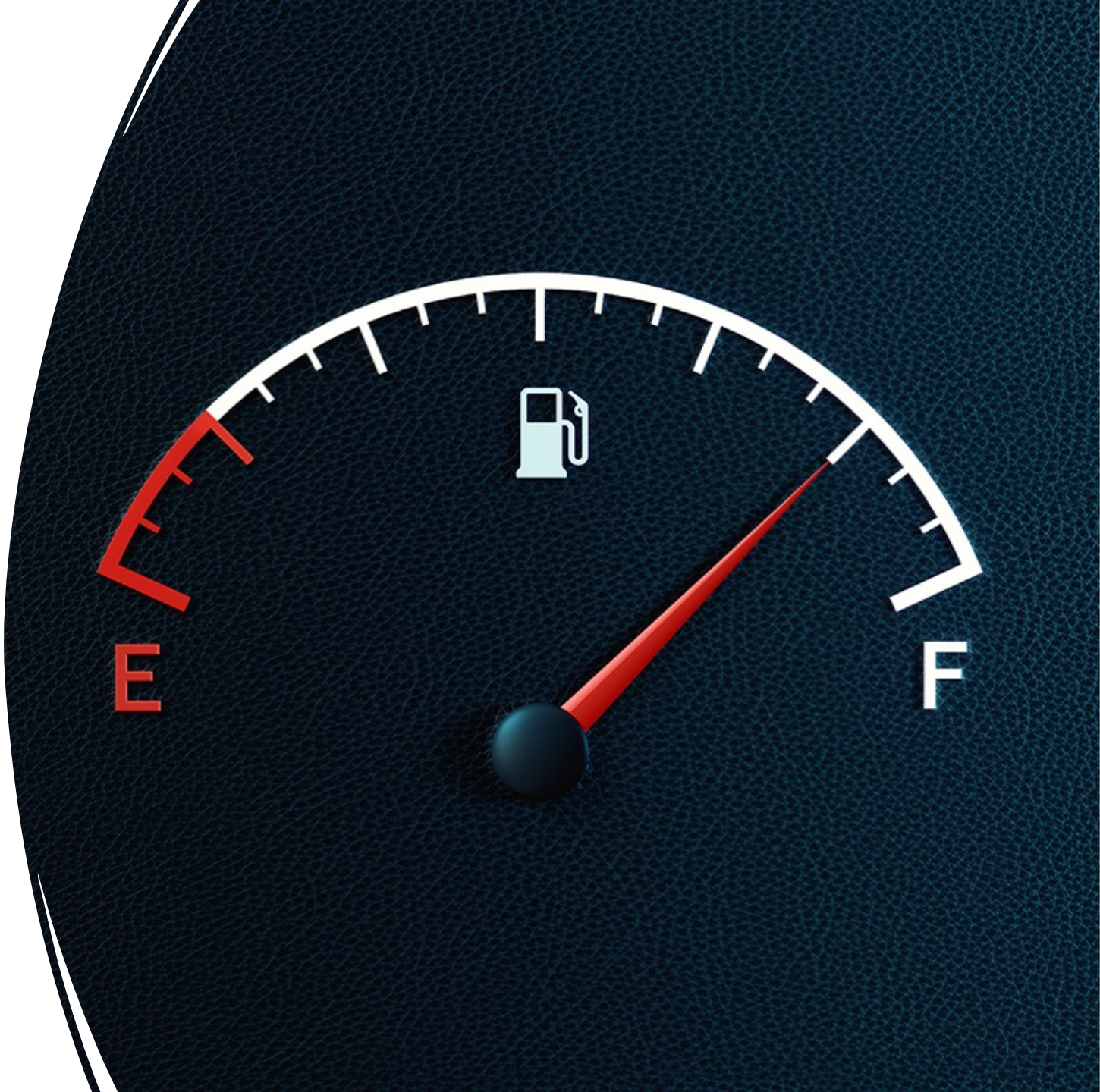
- Runtimes
- Clusters
- Notebooks
- Workflows
- Repos
- DBFS

Machine Learning

- Data Preparation
- Model Training
- Deployment

SQL

- Query Management
- Visualisation
- Dashboards
- Alerts



Databricks

3. Machine Learning

Machine Learning Data Pipeline

1. Experiment Tracking
2. Delta Live Tables
3. Feature Stores
4. Model Registry
5. Model Deployment

Experiment Tracking

- Automatic
- Parameters
- Metrics
- Artifacts
- Models

Delta Live Tables

- Data Pipeline Framework
- Change Data Capture
- Stream Data Processing
- Data Quality
- Publish Data

Feature Stores

- Data Skew
- Point-in-time Correct
- Feature Discovery
- Server-side Computation

Model Registry

- Model Versioning
- Staged and Production Models
- Archived Models
- Access Across Databricks

Model Deployment

- REST API
- Real-time Inference
- Serverless Compute
- Containerised
- Feature Store

Skills Required

- Learn Python (e.g., PySpark + Scikit-learn)
- Learn ML (e.g., XGBoost + Random Forest)
- Learn Databricks (e.g., MLflow + Delta Lake)

Pricing

- Pay As You Go
- No Up-front Costs
- Pay for Compute Resources
- Spot Instances

Getting Started

- Community Edition
- 14 Day Free Trail
- Free Storage and Compute
- Limited Functionality

Conclusion

An **End-to-end Data Science Machine Learning Specialist** can exist with with the **Databricks Platform** given the right skills (e.g., ML and PySpark).

Questions?