D ONE

WE MAKE SENSE.

d-one.ai

GTC23

GROWING TOGETHER CONFERENCE

GROWING
TOGETHER
CONFERENCE

Track 3:
Analytic Innovations

Databricks brick-by-brick:

Data, Analytics and ML in one platform

d-one.ai

GROWING
TOGETHER
CONFERENCE

After this session, you will have a holistic overview of the Databricks capabilities in the Data & AI space.

You will work with hands-on examples that you can reuse for your own data projects.

Your Hosts:

Spyros Cavadias
spyros.cavadias@done.ai

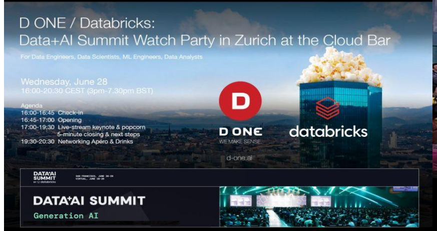
Robert Yousif
robert.yousif@done.ai

Agenda

- Introducing the Databricks Lakehouse
- Data capabilities:
  - Databricks Workspace
  - Delta + Unity Catalog
  - Medallion Architecture & Workflow Orchestration
- ML capabilities:
  - ML Development
  - ML Operations

D ONE and Databricks

- Swiss Partner of the Year 2023
- First Champion in Switzerland
- Databricks Expert Group (20+ professionals)


D ONE / Databricks:
Data+AI Summit Watch Party in Zurich at the Cloud Bar

For Data Engineers, Data Scientists, ML Engineers, Data Analysts

Wednesday, June 28
16:00-20:30 CEST (3pm-7.30pm BST)

Agenda
16:00-16:45 Check-in
16:45-17:00 Opening
17:00-19:30 Live-stream keynote & popcorn
5-minute closing & next steps
19:30-20:30 Networking Apéro & Drinks

D ONE
WE MAKE SENSE.
databricks
d-one.ai

DATA+AI SUMMIT
Generation AI

Full House @ Databricks Lakehouse Day

D|ONE
WE MAKE SENSE.

The 2023 Award goes to:

D|ONE
WE MAKE SENSE.

d-one.ai

GROWING
TOGETHER
CONFERENCE

Introduction

What is a data platform?

A data platform is an **integrated set of technologies** that collectively meets an organization’s end-to-end data needs.

A data analytics platform is an **ecosystem of services and technologies** that needs to perform analysis on voluminous, complex and dynamic data

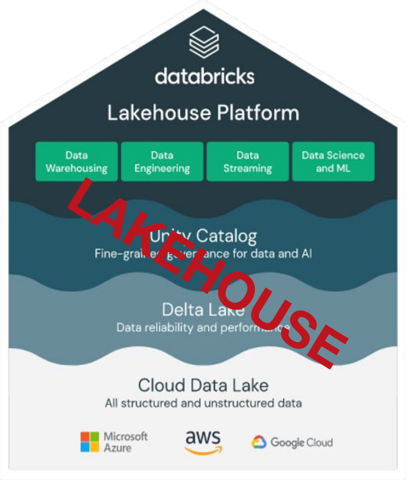
-

that allows you to retrieve, combine, interact with, explore, and visualize data from the various sources a company might have.

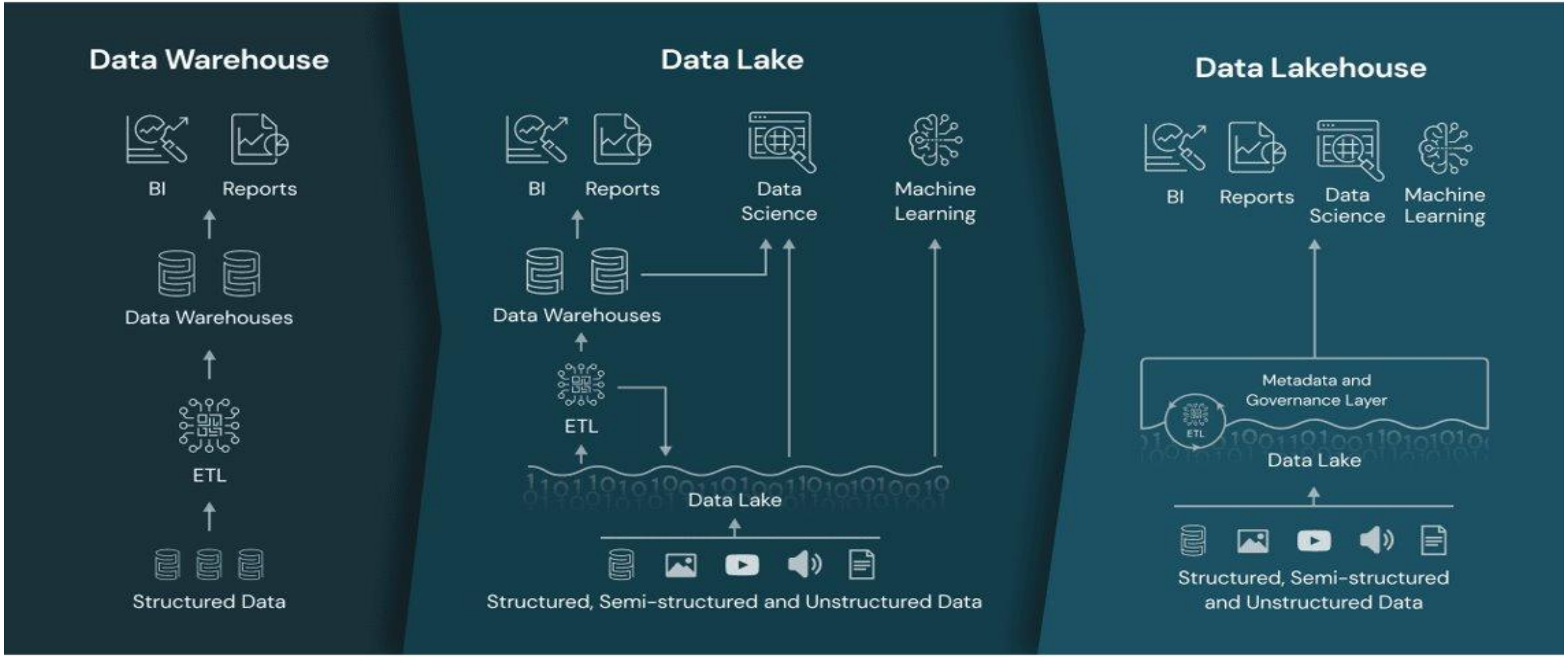
Data platforms encompass a **range of elements** required to support the data management cycle.

What do you need from a data platform?

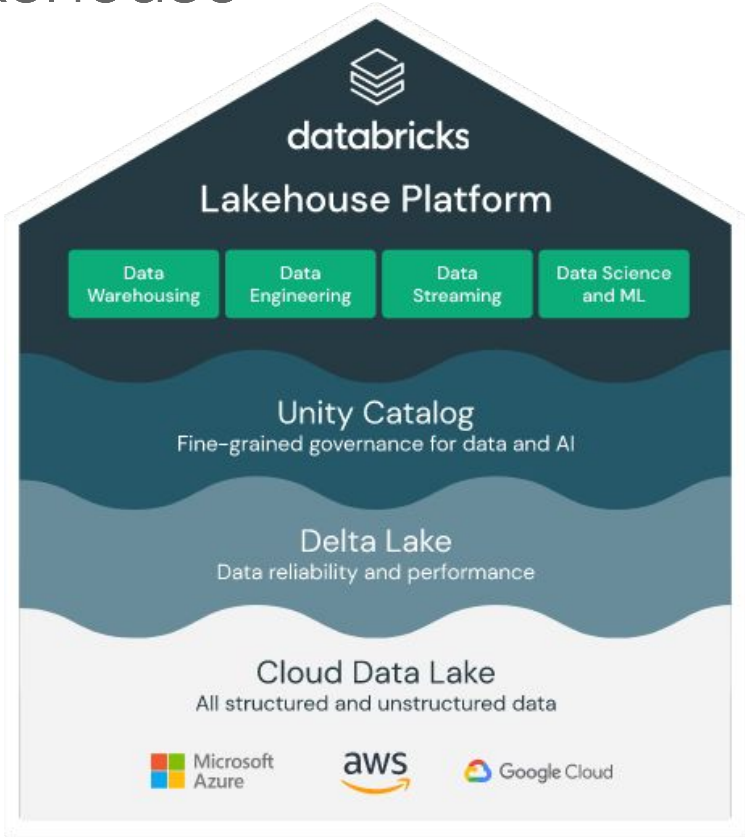
- Warehouse
  - BI & Reporting
  - Structured Data
- Data Lake
  - Data Science & Machine Learning
  - Unstructured Data
- Others
  - Infrastructure
  - Governance
  - Operations



Why the Lakehouse?



Databricks Lakehouse



databricks

Lakehouse Platform

Data Warehousing | Data Engineering | Data Streaming | Data Science and ML

Unity Catalog

Fine-grained governance for data and AI

Delta Lake

Data reliability and performance

Cloud Data Lake

All structured and unstructured data

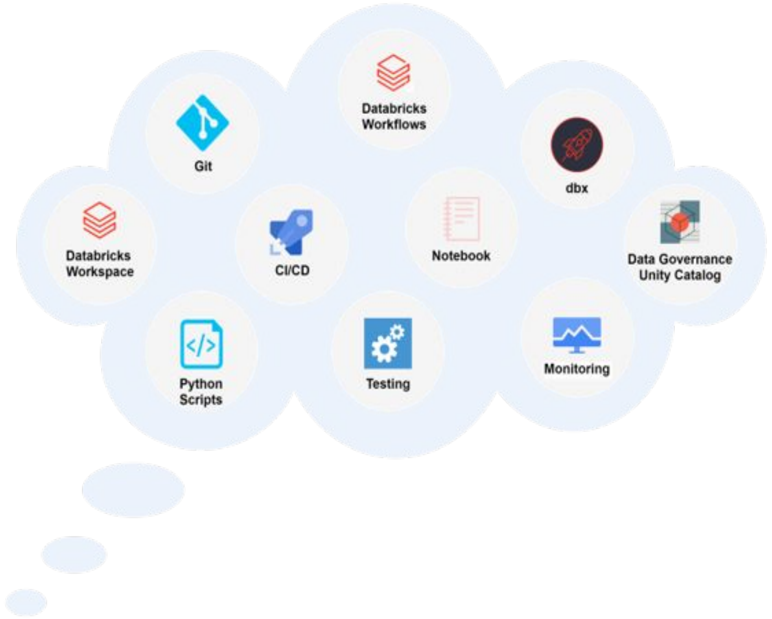
Microsoft Azure | aws | Google Cloud

What is Databricks

A unified set of tools for building, deploying, sharing and maintaining enterprise-grade data solutions at scale.

Combining

- Data Engineering
- Machine Learning, AI & Data Science
- Data Warehousing, analytics & BI
- Data Governance and Secure Data Sharing



Platform Integrations, services/technologies

- Delta Lake 🟢
- Apache Spark 🟢
- Workflows 🟢
- Databricks SQL 🟢
- Unity Catalog 🟢
- Delta Sharing 🔴
- Delta Live Tables 🔴

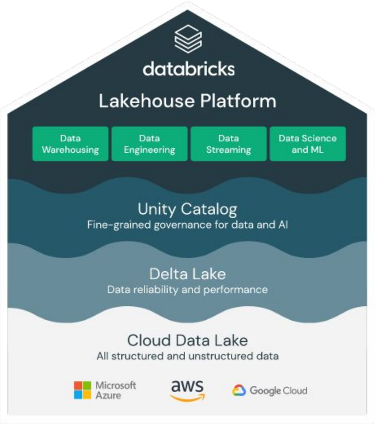
covered in workshop

not covered in workshop

GROWING
TOGETHER
CONFERENCE

Delta + Unity Catalog

Delta Lake



ACID Transactions

Protect your data with serializability, the strongest level of isolation

Scalable Metadata

Handle petabyte-scale tables with billions of partitions and files with ease

Time Travel

Access/revert to earlier versions of data for audits, rollbacks, or reproduce

Open Source

Community driven, open standards, open protocol, open discussions

Unified Batch/Streaming

Exactly once semantics ingestion to backfill to interactive queries

Schema Evolution / Enforcement

Prevent bad data from causing data

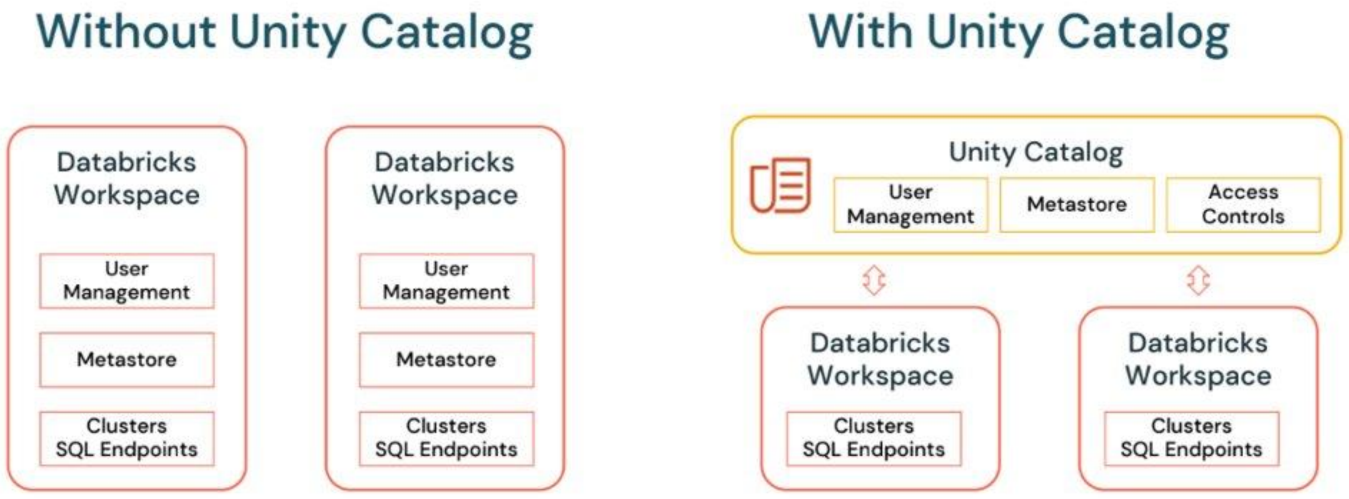
Audit History

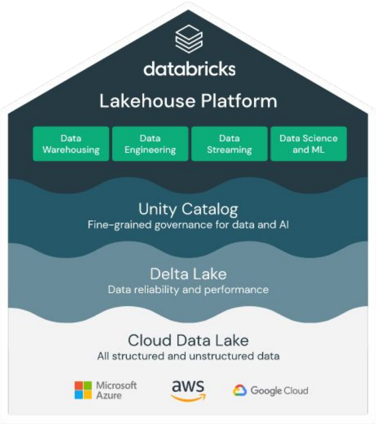
Delta Lake log all change details providing a fill audit trail

DML Operations

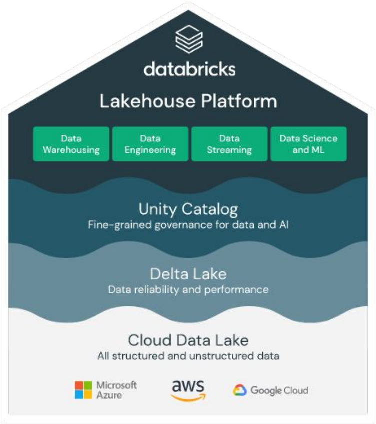
SQL, Scala/Java and Python APIs to merge, update and delete datasets

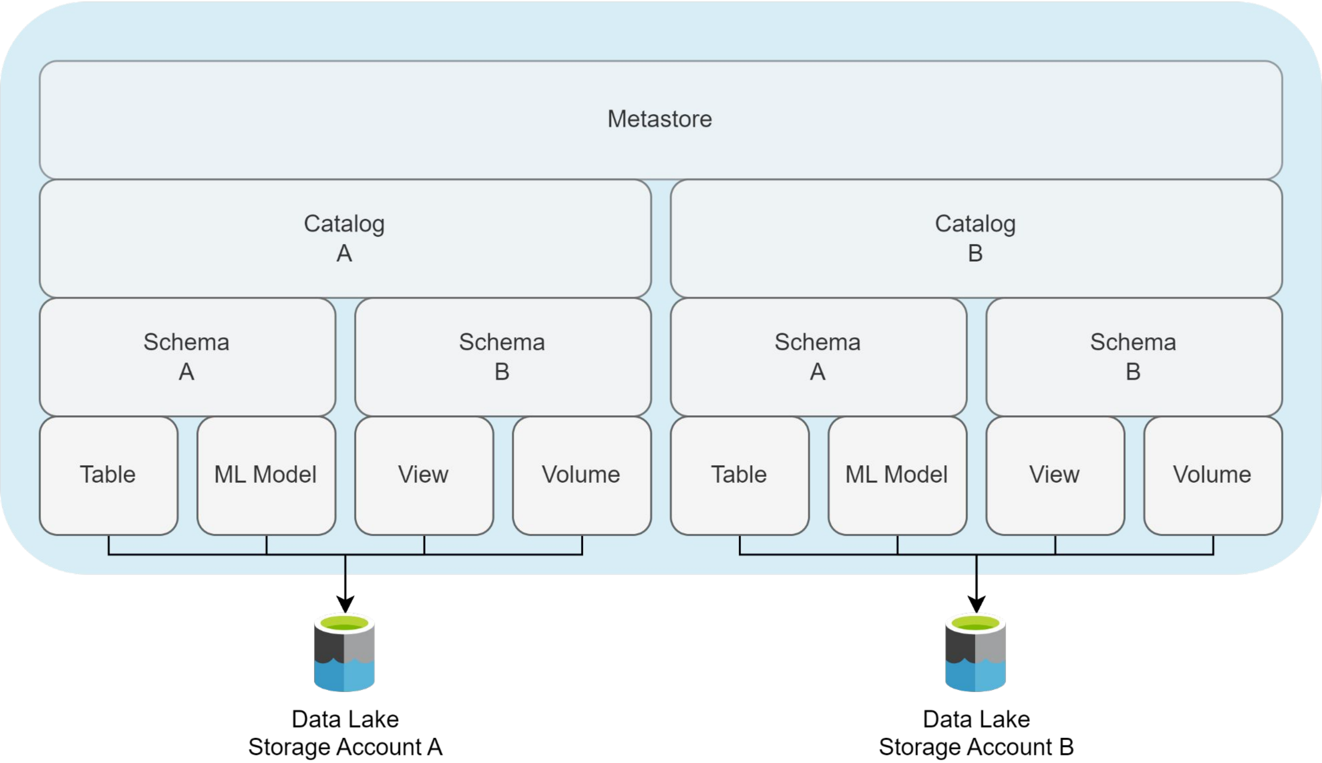
Unity Catalog





Unity Catalog - Structure





Demo - Workspace, Unity & Delta

Exercise

- Get to know Databricks:
  - `1. Navigate your workspace`
  - `2. Clone from the repository: https://github.com/d-one/brick-by-brick`
  - `3. Create own cluster`
  - `4. Read the data and display it using the Delta + Unity Catalog Notebook`

LINKS:

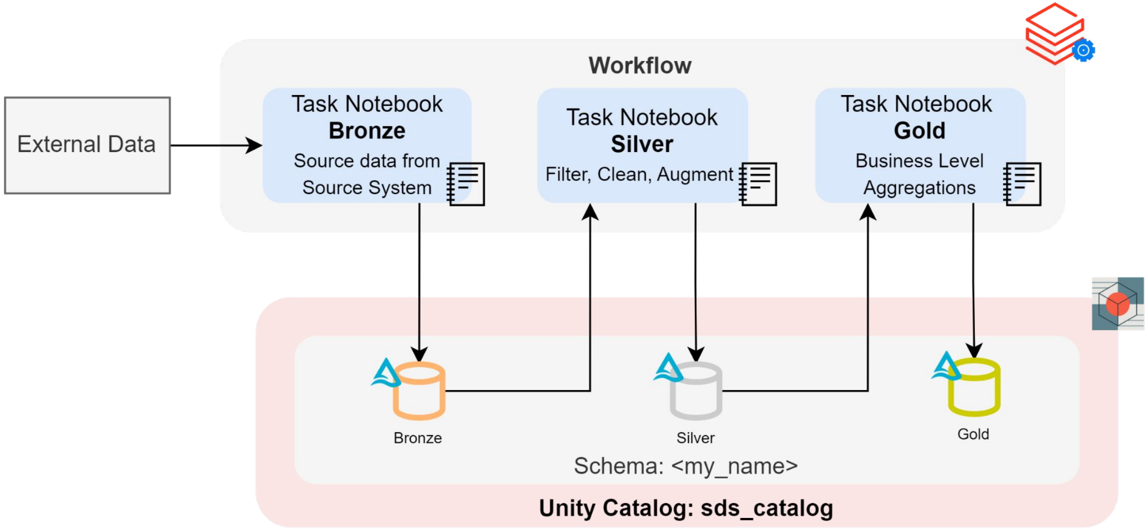
- Github Repository
- Databricks Workspace

GROWING
TOGETHER
CONFERENCE

Medallion Architecture & Workflow Orchestration

Databricks Workflows

- Orchestrate & Automate end-to-end data pipelines
- GUI & API for defining and managing complex workflows
- Supports multiple task types
  - Python Script/Wheel file
  - Notebooks
  - dbt & dlt
  - Databricks SQL Queries



Demo - Workflow orchestration and Medallion Architecture

GROWING
TOGETHER
CONFERENCE

ML & MLOps in Databricks

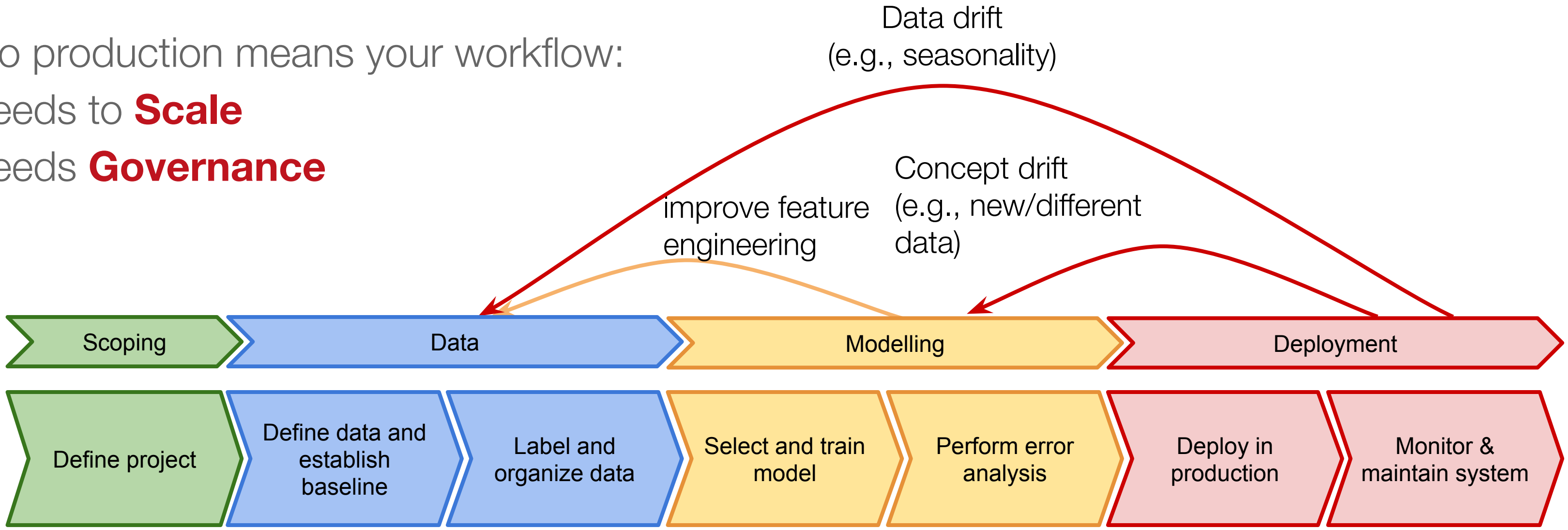
Databricks Lakehouse AI capabilities

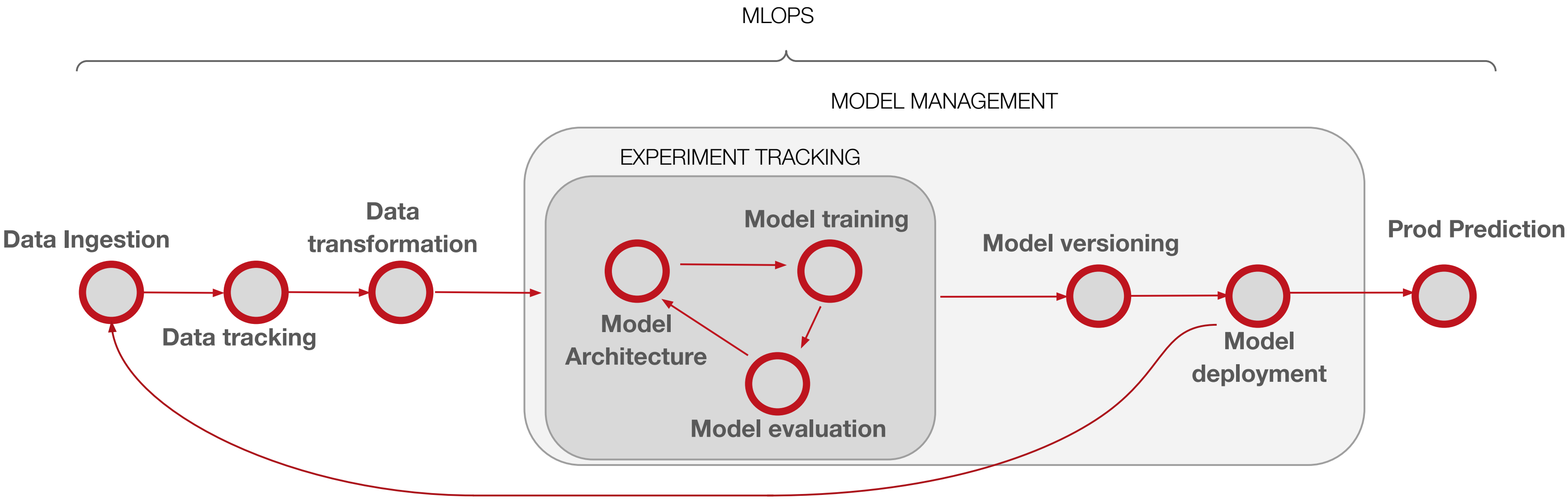
- Tracking Experiments	🟢	covered in workshop
- Model Registry	🟢	
- Model Serving	🟢	
- AutoML	🟢	
- Feature Store	🔴	not covered in workshop
- LLMs and GenAI	🔴	

Machine Learning Pipeline

Going to production means your workflow:

- Needs to **Scale**
- Needs **Governance**



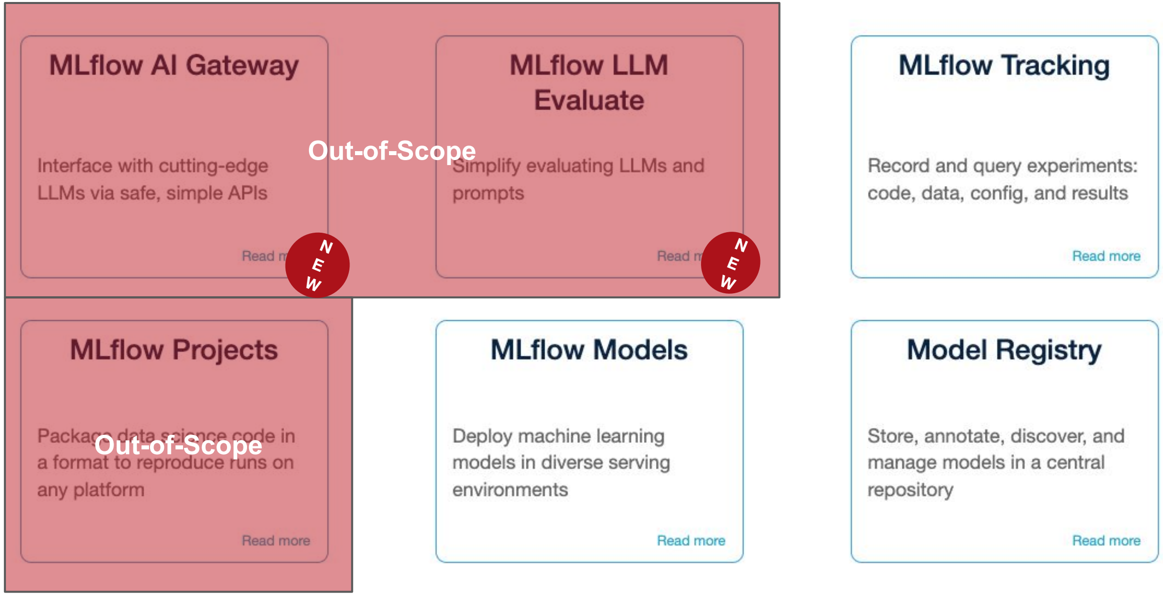

MLOPS
MODEL MANAGEMENT
EXPERIMENT TRACKING
Data Ingestion
Data tracking
Data transformation
Model training
Model Architecture
Model evaluation
Model versioning
Model deployment
Prod Prediction

mlflow and databricks

- Open source
- Runs the same way everywhere (locally or in the cloud)
- Useful from 1 developer to 100+ developers

- Design philosophy:
  1. API-First
  2. Integration with popular libraries
  3. Modular design (can use DISTINCT components separately)

MLflow Components



Out-of-Scope

MLflow AI Gateway

Interface with cutting-edge LLMs via safe, simple APIs

Read m

NEW

MLflow LLM Evaluate

Simplify evaluating LLMs and prompts

Read m

NEW

MLflow Tracking

Record and query experiments: code, data, config, and results

Read more

Out-of-Scope

MLflow Projects

Package data science code in a format to reproduce runs on any platform

Read more

MLflow Models

Deploy machine learning models in diverse serving environments

Read more

Model Registry

Store, annotate, discover, and manage models in a central repository

Read more

Useful links:
- www.mlflow.org
- www.github.com/mlflow
- www.databricks.com/mlflow

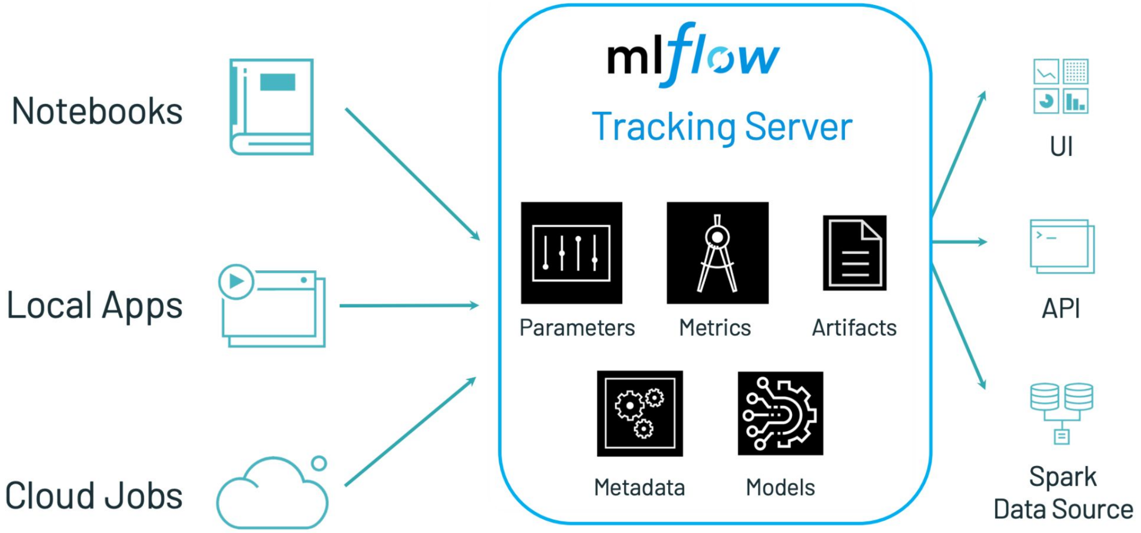
MLflow Tracking

What do we track?

- Parameters : inputs to our code `mlflow.log_param()..`
- Metrics : numeric values to access our models `mlflow.log_metric()..`
- Tags/Notes: info about the run `mlflow.set_tag()..`
- Artifacts: files,data and models produced `mlflow.log_artifact(), mlflow.log_artifacts()..`
- Source: what code run
- Version: what version of the code run (github)
- Run: the particular code instance (id) captured by MLflow `mlflow.start_run()..`
- Experiment: the set of runs `mlflow.create_experiment(), mlflow.set_experiment()..`

More on : https://www.mlflow.org/docs/latest/tracking.html

MLflow Tracking



Demo - Experiments

MLflow Model

```
# Directory written by
mlflow.sklearn.save_model(model,
"my_model")

my_model/
├── MLmodel
├── model.pkl
├── conda.yaml
└── requirements.txt
```

```
# in MLmodel file
time_created:
2021-10-25T17:28:53.35

flavors:
  sklearn:
    sklearn_version: 0.24.1
    pickled_model: model.pkl
  python_function:
    loader_module: mlflow.sklearn
```

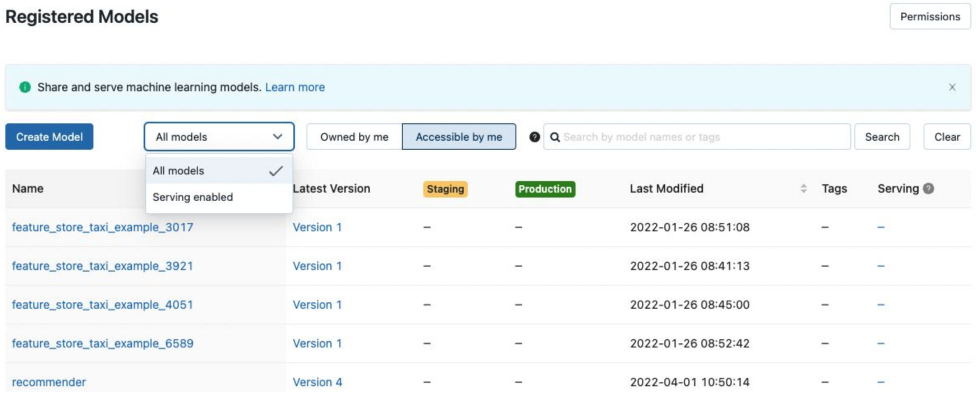
MLflow Model Registry

It enables:

- model lineage
- model versioning
- stage transitions
- annotations

You register a model through:

1. API Workflow
2. UI Workflow



Registered Models

Share and serve machine learning models. Learn more

Name	Latest Version	Staging	Production	Last Modified	Tags	Serving
feature_store_taxi_example_3017	Version 1	–	–	2022-01-26 08:51:08	–	–
feature_store_taxi_example_3921	Version 1	–	–	2022-01-26 08:41:13	–	–
feature_store_taxi_example_4051	Version 1	–	–	2022-01-26 08:45:00	–	–
feature_store_taxi_example_6589	Version 1	–	–	2022-01-26 08:52:42	–	–
recommender	Version 4	–	–	2022-04-01 10:50:14	–	–

```
# register model
res = mlflow.register_model(my_model_uri, "my_model")
```

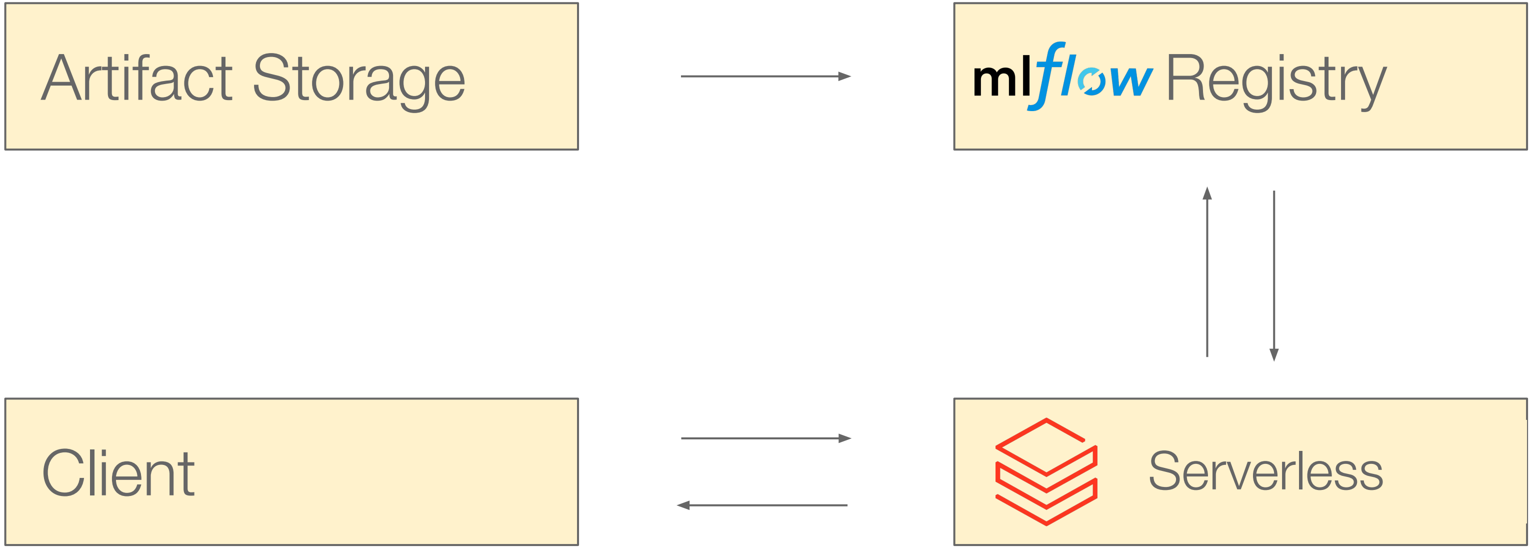
Demo - Model Registry

Deployment

Batch Prediction

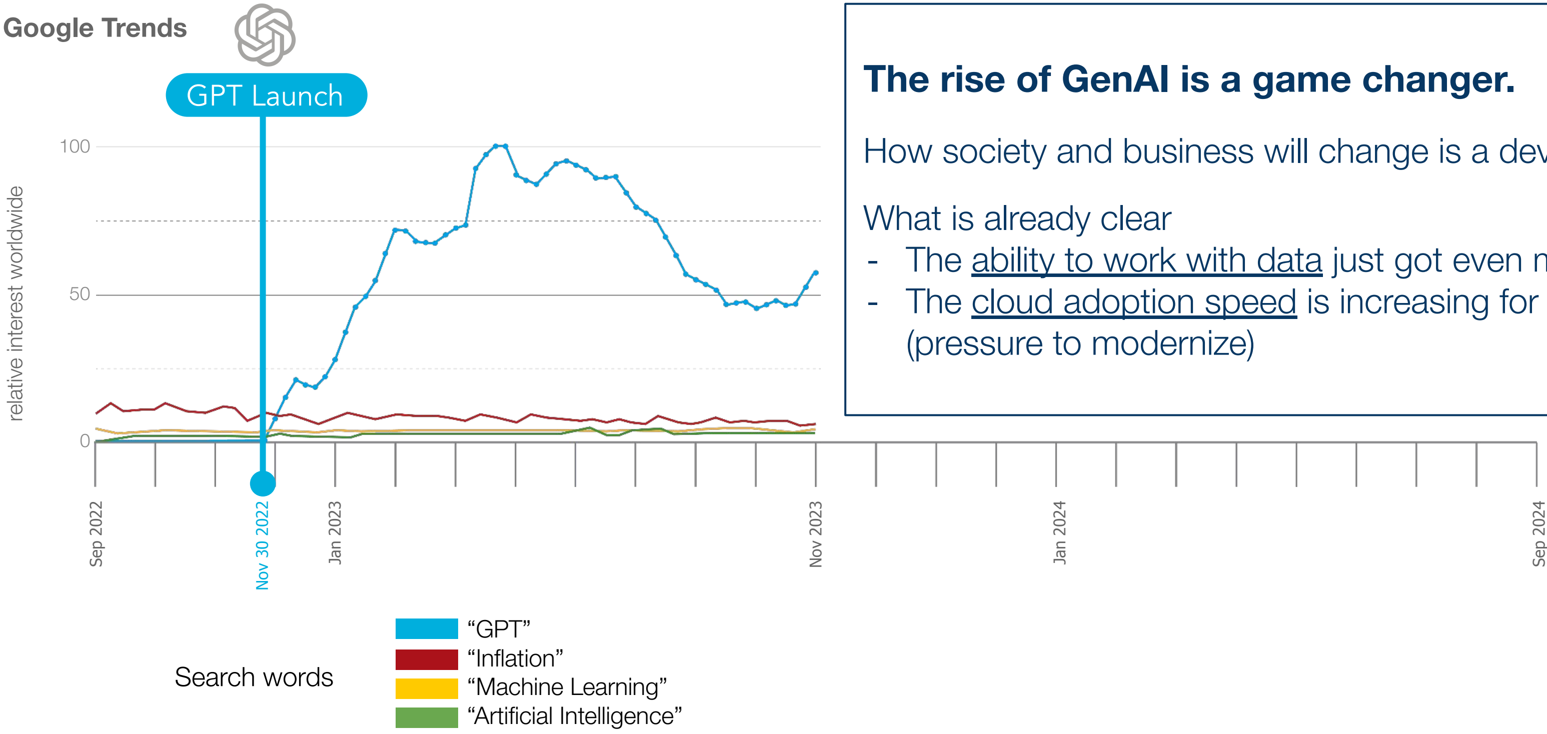
Online Prediction

Model Serving



Demo - Model Serving

Enter GenAI



The rise of GenAI is a game changer.

How society and business will change is a developing story.

What is already clear

- The ability to work with data just got even more important
- The cloud adoption speed is increasing for all industries (pressure to modernize)

What are the game changing elements of GenAI

Interaction via natural language

Generating content

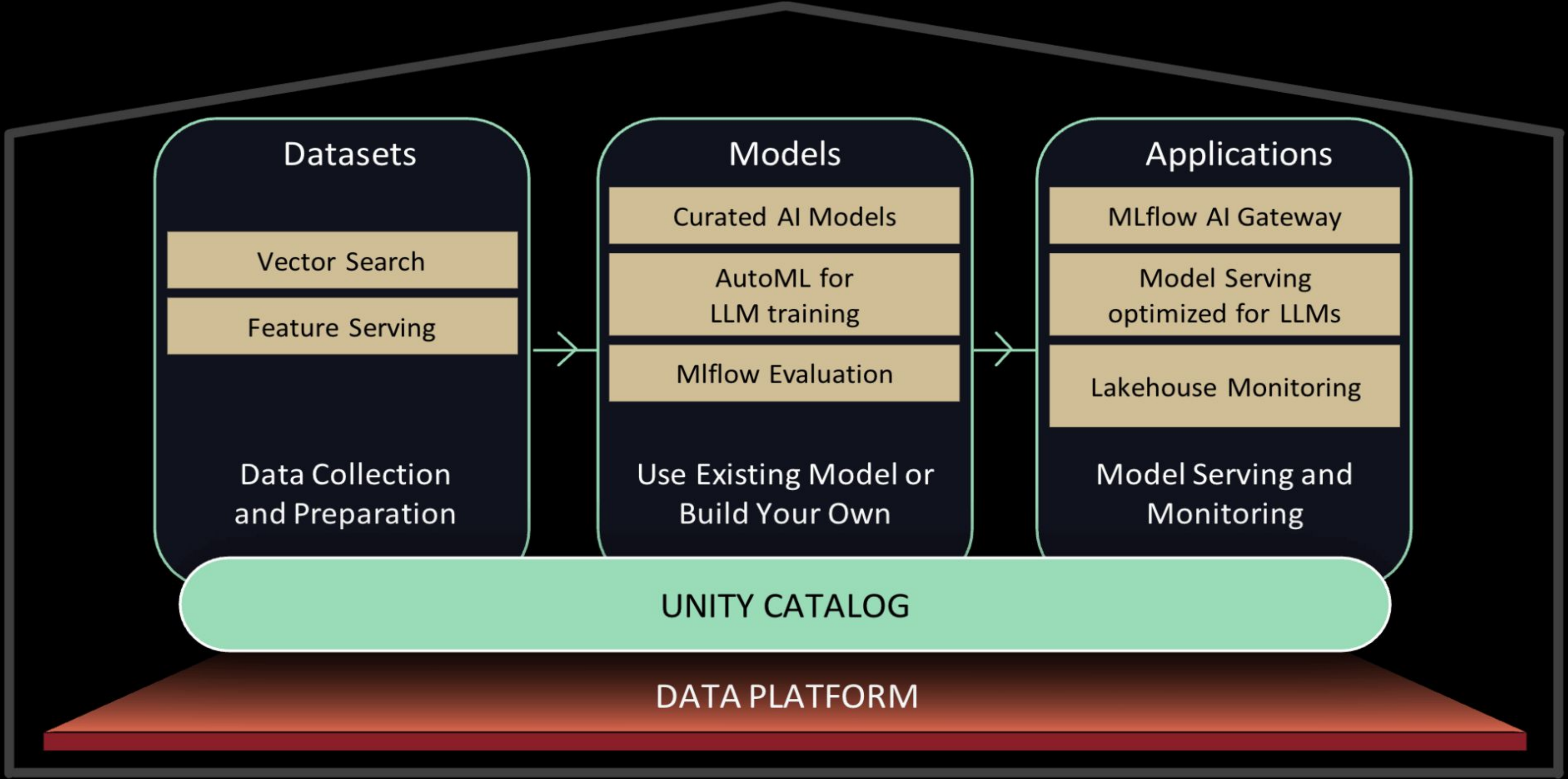
Based on existing content: natural language, visuals, code

Zero shot learning

Foundational model fits all kind of different questions without dedicated training

… and everyone has already used it in consequence

Lakehouse AI — optimized for Generative AI



Datasets

Vector Search

Feature Serving

Data Collection and Preparation

Models

Curated AI Models

AutoML for LLM training

Mlflow Evaluation

Use Existing Model or Build Your Own

Applications

MLflow AI Gateway

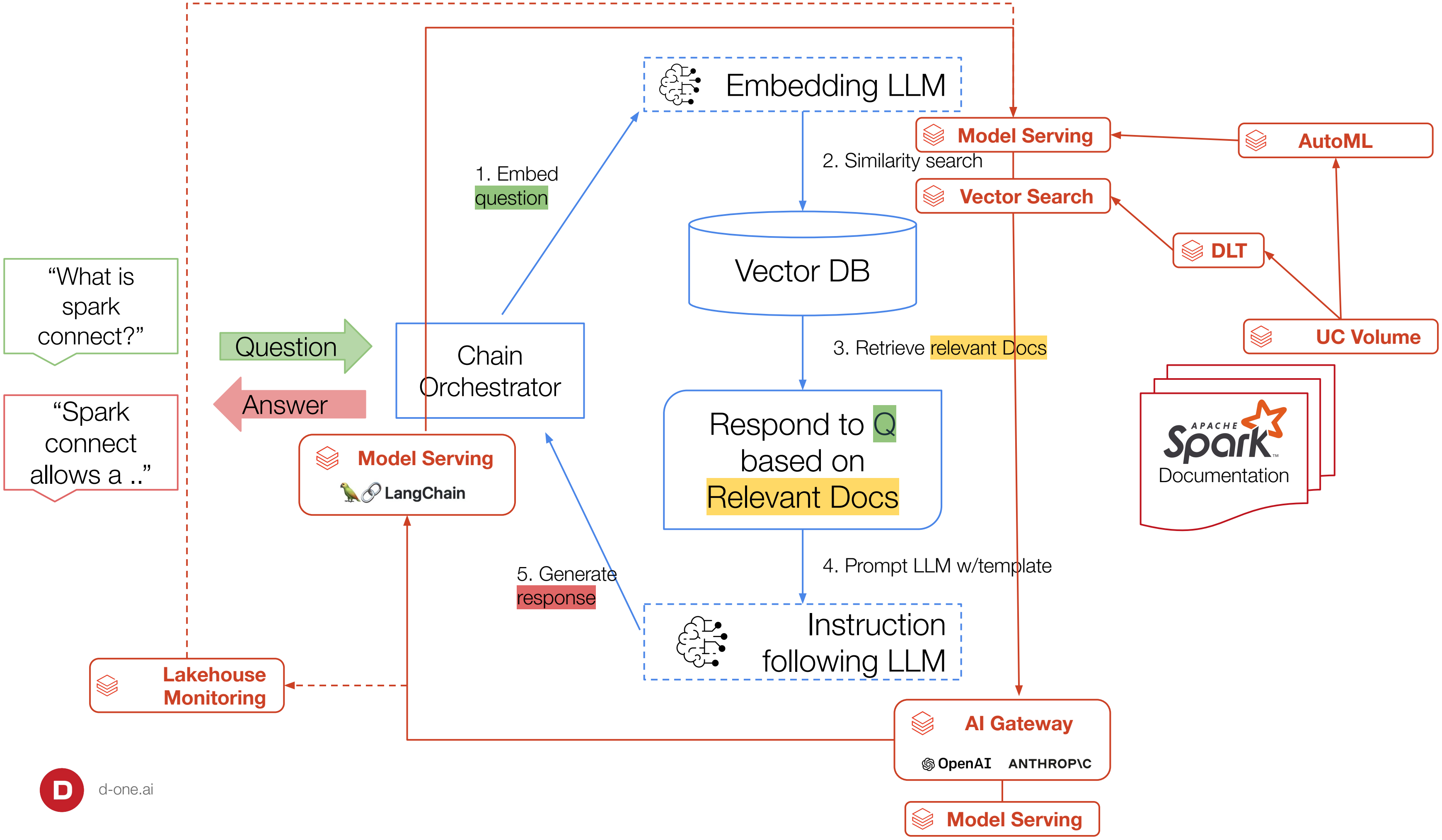
Model Serving optimized for LLMs

Lakehouse Monitoring

Model Serving and Monitoring

UNITY CATALOG

DATA PLATFORM



Embedding LLM
Model Serving
AutoML
1. Embed question
2. Similarity search
Vector Search
DLT
"What is spark connect?"
Vector DB
UC Volume
Question
Chain Orchestrator
3. Retrieve relevant Docs
"Spark connect allows a .."
Answer
Respond to Q based on Relevant Docs
APACHE Spark Documentation
Model Serving
LangChain
5. Generate response
4. Prompt LLM w/template
Instruction following LLM
Lakehouse Monitoring
AI Gateway
OpenAI
ANTHROP\C
Model Serving

d-one.ai

GROWING
TOGETHER
CONFERENCE

Recap

Recap

- Introducing the Databricks Lakehouse
- Data capabilities:
    - Databricks Workspace
    - Delta + Unity Catalog
    - Medallion Architecture & Workflow Orchestration
- ML capabilities:
    - ML Development
    - ML Operations

D ONE Databricks Expert Group

At D ONE we have profound knowledge & experience on solving real problems leveraging data platform best practices. The Databricks expert group consists of Solution Architect Champions and professionals with extensive ecosystem experience.

Shared Content:

- Brick-by-brick (Swiss Data Science Conference 2023)
- Streamline Data Pipelines with Databricks (Medium)
- Metadata driven Lakehouse Data Pipelines (Lakehouse Days 2023)
- Databricks Workspace Migration (Medium)

Upcoming:

- Data & AI World Tour Zurich (November 23rd 2023)
- Applied Machine Learning Days (March 2024)

D ONE

WE MAKE SENSE.

databricks

Get in touch : databricks@d-one.ai

GROWING
TOGETHER
CONFERENCE

Q & A

GROWING
TOGETHER
CONFERENCE

Track 2:

Analytics Innovations

Databricks brick-by-brick: Data, Analytics and ML in one platform

Your thoughts are important to us! Would you give us feedback?

Exercise (optional)

- Data Engineering and Unity Catalog:
  - 1. Run bronze, silver and gold notebooks, setting the parameters so you write to your personal catalogs
  - 2. Link the 3 notebooks together with a workflow as described in https://github.com/d-one/brick-by-brick#3-creating-a-workflow-job

- **LINKS**: Github Repository , Databricks Workspace

Exercise (optional)

- Machine Learning:
    - 1. Run the ML Preprocessing notebook in your catalog to create the features table
    - 2. Move on to the ML MLflow Tracking notebook and walk through the steps to understand how to interact with MLflow experiments inside the Databricks workspace
    - 3. Move on to the ML Model Registry notebook and walk through the steps to understand how to interact with the model registry via python APIs or via the directly using the UI
    - 4. Tie steps 1-3 together by creating a new ML workflow! See the results of the workflow run in the UI.
    - 5. Finally move on to the AutoML notebook and see for yourself how easy it is to use databricks AutoML as a quick way to create baseline models.

- **LINKS**: Github Repository , Databricks Workspace

D ONE

WE MAKE SENSE.

d-one.ai

DATA DRIVEN VALUE CREATION

DATA SCIENCE & ANALYTICS | DATA MANAGEMENT | VISUALIZATION & DATA EXPERIENCE

D ONE, Sihlfeldstrasse 58, 8003 Zürich

GROWING
TOGETHER
CONFERENCE

Appendix

Titel: Helvetica Neue Light 28 (min. 24)

Font size:

- 18 for the slide content
- 16 for the slide content
- 14 for the slide content
- 12 minimum size content

- 8 can be used for diagrams, notes and references

Font type: Helvetica Neue Light

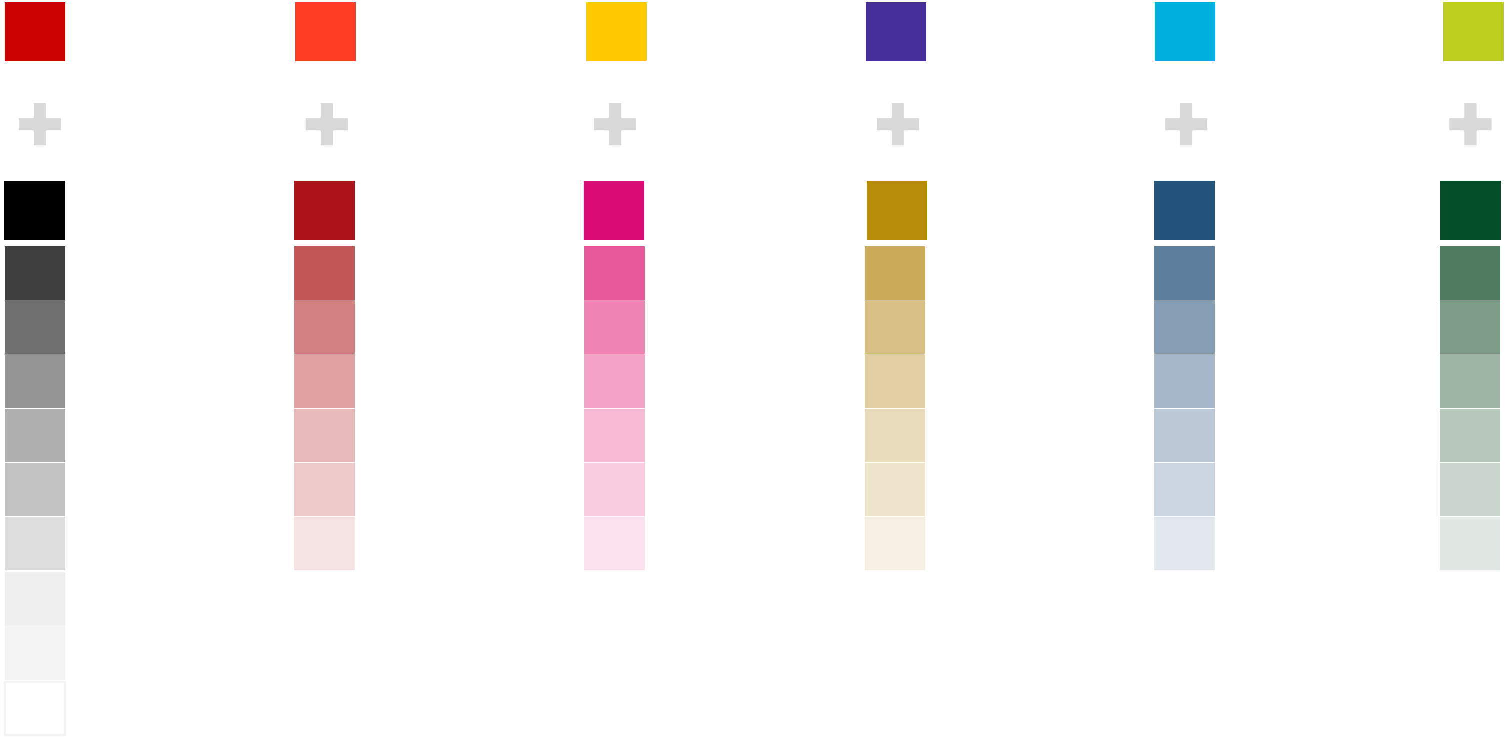
- **Helvetica Neue Bold** can be used to **highlight** an element in a text or as a text box title

Font color:

- Dark grey 3 from D ONE - 12-2022 color pallette
- another color can be used to highlight an element in a text, red “Accent 1”

ONE SINGLE COLOR PER SLIDE

If you need a range:



d-one.ai

Tabelle

Method	Description	Example for “Peter”	Example for 42	Limitation
Create synthetic data	*Generates artificial(fake) data that resembles the original dataset.*	Peter	17	Difficult to mimic the noise of “normal” data.
Pseudonymization	*Replaces the identifying fields by pseudonyms(fictional identifiers).*	Dalerf	21	as above
Data Masking/Data Obfuscation	*Replaces some attributes with similar values; keep relationships and statistical distribution.*			
Generalization	*Substitutes an original value with a more abstract one.*	A*	40-50	Loss of granularity
Shuffling	*Randomizes the existing values vertically across a data set.*	Thomas	57	
Removing/Nulling	*Replaces the sensitive values with a generic value (e.g. ‘*’, ’X’).*	NULL	XXX	Loss of information; possibly makes data unusable for testing.
Hashing / Tokenization	*Replaces the sensitive values with a hash value.*	2b348a84	90e2a5170	Not suitable for testing purposes.

Kacheln

Job 1

4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt

Job 2

4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt

Job 3

4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt

Job 4

4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt
4 Zeilen: 11pt

Job 1

3 Zeilen: 12pt
3 Zeilen: 12pt
3 Zeilen: 12pt

Job 2

3 Zeilen: 12pt
3 Zeilen: 12pt
3 Zeilen: 12pt

Job 3

3 Zeilen: 12pt
3 Zeilen: 12pt
3 Zeilen: 12pt

Job 4

3 Zeilen: 12pt
3 Zeilen: 12pt
3 Zeilen: 12pt

Job 1

4 Zeilen: 11pt
- Einzug1
- Einzug2
- Einzug 3

Job 2

4 Zeilen: 11pt
- Einzug1
- Einzug2
- Einzug 3

Job 3

4 Zeilen: 11pt
- Einzug1
- Einzug2
- Einzug 3

Job 4

4 Zeilen: 11pt
- Einzug1
- Einzug2
- Einzug 3

3 Box Grid

Zeile 1

- Einzug1
- Einzug2
- Einzug 3

Zeile 2

- Einzug1
- Einzug2
- Einzug 3

Zeile 3

- Einzug1
- Einzug2
- Einzug 3

Text/Image ½ Seite

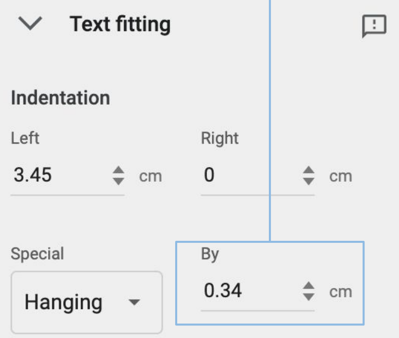
Zeile 1 - Einzug1 - Einzug2 - Einzug 3	Zeile 1 - Einzug1 - Einzug2 - Einzug 3
Zeile 1 - Einzug1 - Einzug2 - Einzug 3	Zeile 1 - Einzug1 - Einzug2 - Einzug 3

Bullet points formatting

Formatting details if case needed: modify the indentation parameters in the Text Fitting section and the text size as follows

- Hyphen color set to red “Accent 1”, always (dark red in D ONE color palette)
- “By” set to **0.34** cm
  always (this is the distance between the hyphen and the text)

- First Level, 18pt - “Left” set to **0.24** cm (this is to define the hyphen location)
    - First Second Level, 16pt, “Left” set to **1.41** cm
        - First Third Level, 14pt, “Left” set to **2.47** cm
            - First Fourth Level, 14pt, “Left” set to **3.45** cm



Text fitting

Indentation

Left: 3.45 cm

Right: 0 cm

Special: Hanging

By: 0.34 cm

D ONE

WE MAKE SENSE.

d-one.ai

DATA DRIVEN VALUE CREATION

DATA SCIENCE & ANALYTICS | DATA MANAGEMENT | VISUALIZATION & DATA EXPERIENCE

D ONE, Sihlfeldstrasse 58, 8003 Zürich