Data Platform Architectures & Machine Learning Operations (MLOps)

Dr. Christoph Gröger
Arnold Lutsch
Robert Bosch GmbH
IoT & Digitalization - Data Strategy

Prof. Dr. Jan Kirenz
HdM Stuttgart

Data platform architectures & MLOps

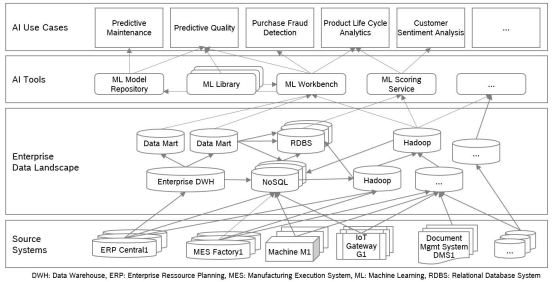
Modul-Nr: 338025-338027

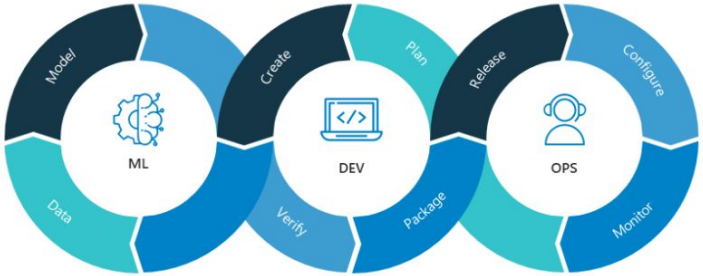
SWS/ECTS: 5/10

Prüfungsform: PP

Projektpartner:
Dr. Christoph Gröger
Arnold Lutsch
Robert Bosch GmbH

Themenschwerpunkte
- Datenplattform-architektur
- Open Source Technologien (bspw. Delta Lake, Kubeflow, TFX)
- Machine Learning Operations
- Entwicklung in Python





Im Rahmen des Projekts soll mit Hilfe von Open-Source-Software eine prototypische state-of-the-art **Data Lake bzw. Datenplattformarchitektur** (bspw. Lakehouse) für die Realisierung unterschiedlicher Machine Learning Anwendungsfälle konzipiert und implementiert werden.

In der Architektur sollen für eine möglichst umfassende Automatisierung des **Machine Learning Lifecycles** Komponenten der Disziplin **Machine Learning Operations (MLOps)** berücksichtigt werden (bspw. feature store, model registry, data pipelines).

“… [we need to] help companies progress on their AI journey, from one-off AI experimentation to gaining a robust organization-wide capability that acts as a source of competitive agility and growth.”

Accenture (2019)

The Proof of Concept Factory

Most companies...

- … conduct AI experiments and pilots but achieve a low scaling success rate
- … have significant AI under investments, yielding low returns

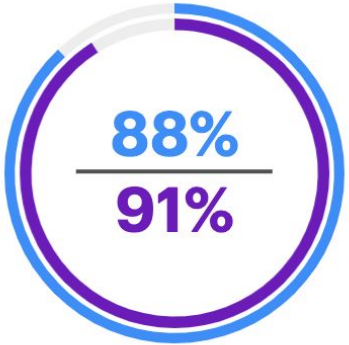
accenture

80-85% PoC Factory

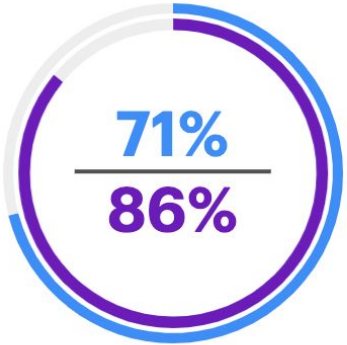
Source: Accenture (2019) https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments

How crucial is scaling AI to your business?

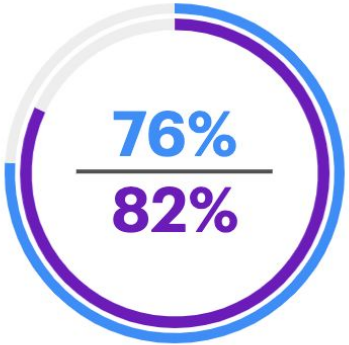
UNITED STATES vs. GERMANY



of executives say they won't achieve their growth objectives without scaling AI.



of executives believe they risk going out of business in 5 years if they don't scale AI.



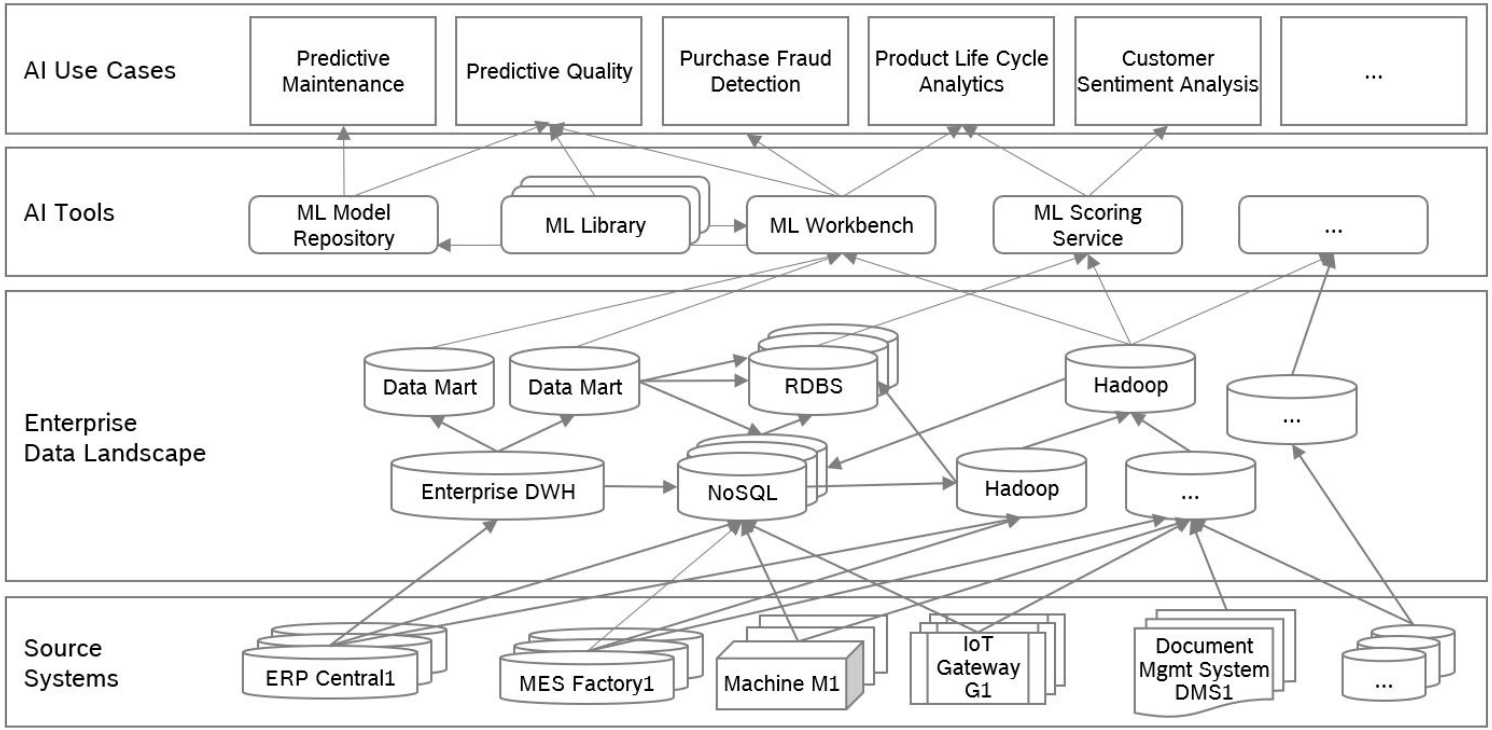
of executives acknowledge they know how to pilot, but struggle to scale AI across the business.

Source: Accenture (2019); n = 1,500 C-suite executives from organizations across 16 industries https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments

“[…] AI is currently done in an insular fashion leading to a polyglot and heterogeneous enterprise data landscape. This makes systematic data management, comprehensive data democratization and an overall data governance considerably challenging and prevents the wide-spread use of AI in industrial enterprises.”

Gröger (2021)

Insular AI and Enterprise Data Landscape



DWH: Data Warehouse, ERP: Enterprise Ressource Planning, MES: Manufacturing Execution System, ML: Machine Learning, RDBS: Relational Database System

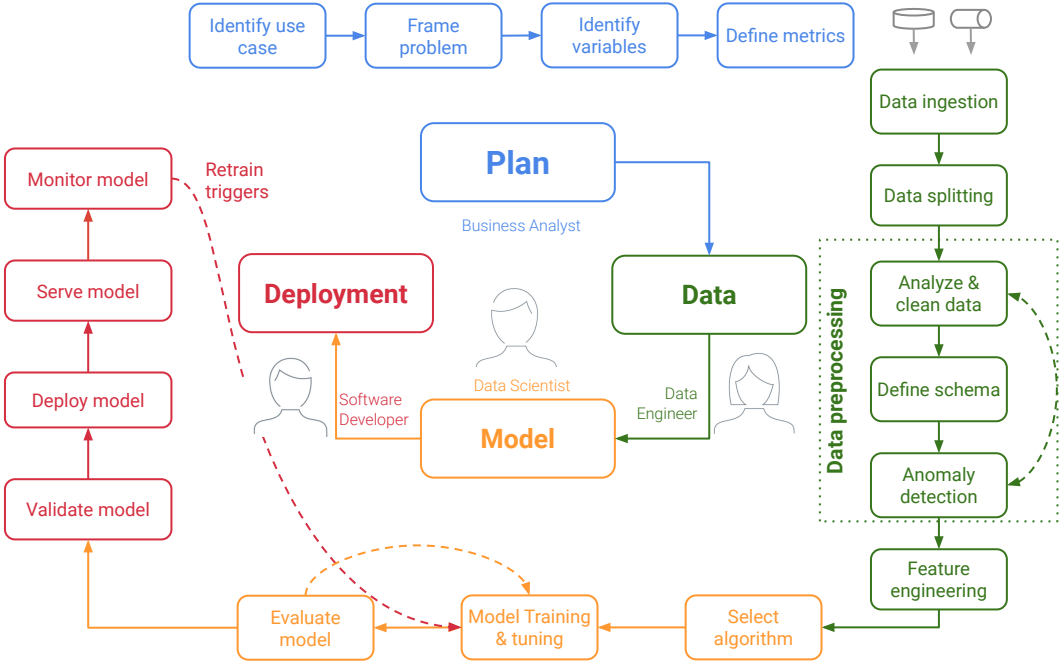
Source: Gröger (2021) http://christophgroeger.de/download/Groeger_There_Is_No_AI_Without_Data.pdf

Machine Learning Lifecycle, Data Platform Architectures & Machine Learning Operations (MLOps)

Prof. Dr. Jan Kirenz

Lifecycle
of an ML System

Plan | Data | Model | Deployment



Source: Kirenz (2021)

Data platform architectures

Evolution

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

Michael Armbrust[1], Ali Ghodsi[1,2], Reynold Xin[1], Matei Zaharia[1,3]
[1]Databricks, [2]UC Berkeley, [3]Stanford University

Abstract

This paper argues that the data warehouse architecture as we know it today will wither in the coming years and be replaced by a new architectural pattern, the Lakehouse, which will (i) be based on open direct-access data formats, such as Apache Parquet, (ii) have first-class support for machine learning and data science, and (iii) offer state-of-the-art performance. Lakehouses can help address several major challenges with data warehouses, including data staleness, reliability, total cost of ownership, data lock-in, and limited use-case support. We discuss how the industry is already moving toward Lakehouses and how this shift may affect work in data management. We also report results from a Lakehouse system using Parquet that is competitive with popular cloud data warehouses on TPC-DS.

1 Introduction

This paper argues that the data warehouse architecture as we know it today will wane in the coming years and be replaced by a new architectural pattern, which we refer to as the Lakehouse, characterized by (i) open direct-access data formats, such as Apache Parquet and ORC, (ii) first-class support for machine learning and data science workloads, and (iii) state-of-the-art performance.

The history of data warehousing started with helping business leaders get analytical insights by collecting data from operational databases into centralized warehouses, which then could be used for decision support and business intelligence (BI). Data in these warehouses would be written with schema-on-write, which ensured that the data model was optimized for downstream BI consumption. We refer to this as the first generation data analytics platforms.

A decade ago, the first generation systems started to face several challenges. First, they typically coupled compute and storage into an on-premises appliance. This forced enterprises to provision and pay for the peak of user load and data under management, which became very costly as datasets grew. Second, not only were datasets growing rapidly, but more and more datasets were completely unstructured, e.g., video, audio, and text documents, which data warehouses could not store and query at all.

To solve these problems, the second generation data analytics platforms started offloading all the raw data into *data lakes*: low-cost storage systems with a file API that hold data in generic and usually open file formats, such as Apache Parquet and ORC [8, 9]. This approach started with the Apache Hadoop movement [5], using the Hadoop File System (HDFS) for cheap storage. The data lake was a schema-on-read architecture that enabled the agility of storing any data at low cost, but on the other hand, punted the problem of data quality and governance downstream. In this architecture, a small subset of data in the lake would later be ETLed to a downstream data warehouse (such as Teradata) for the most important decision support and BI applications. The use of open formats also made data lake data directly accessible to a wide range of other analytics engines, such as machine learning systems [30, 37, 42].

From 2015 onwards, cloud data lakes, such as S3, ADLS and GCS, started replacing HDFS. They have superior durability (often >10 nines), geo-replication, and most importantly, extremely low cost with the possibility of automatic, even cheaper, archival storage, e.g., AWS Glacier. The rest of the architecture is largely the same in the cloud as in the second generation systems, with a downstream data warehouse such as Redshift or Snowflake. This two-tier data lake + warehouse architecture is now dominant in the industry in our experience (used at virtually all Fortune 500 enterprises).

This brings us to the challenges with current data architectures. While the cloud data lake and warehouse architecture is ostensibly cheap due to separate storage (e.g., S3) and compute (e.g., Redshift), a two-tier architecture is highly complex for users. In the first generation platforms, all data was ETLed from operational data systems directly into a warehouse. In today's architectures, data is first ETLed into lakes, and then again ELTed into warehouses, creating complexity, delays, and new failure modes. Moreover, enterprise use cases now include advanced analytics such as machine learning, for which *neither* data lakes nor warehouses are ideal. Specifically, today's data architectures commonly suffer from four problems:

Reliability. Keeping the data lake and warehouse consistent is difficult and costly. Continuous engineering is required to ETL data between the two systems and make it available to high-performance decision support and BI. Each ETL step also risks incurring failures or introducing bugs that reduce data quality, e.g., due to subtle differences between the data lake and warehouse engines.

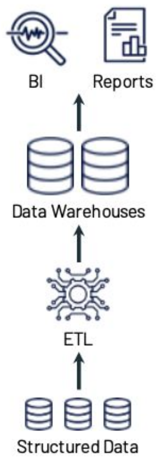
Data staleness. The data in the warehouse is stale compared to that of the data lake, with new data frequently taking days to load. This is a step back compared to the first generation of analytics systems, where new operational data was immediately available for queries. According to a survey by Dimensional Research and Fivetran, 86% of analysts use out-of-date data and 62% report waiting on engineering resources numerous times per month [47].

Limited support for advanced analytics. Businesses want to ask predictive questions using their warehousing data, e.g., "which customers should I offer discounts to?" Despite much research on the confluence of ML and data management, none of the leading machine learning systems, such as TensorFlow, PyTorch and XGBoost, work well on top of warehouses. Unlike BI queries, which extract a small amount of data, these systems need to process large datasets using complex non-SQL code. Reading this data via ODBC/JDBC is inefficient, and there is no way to directly access the internal

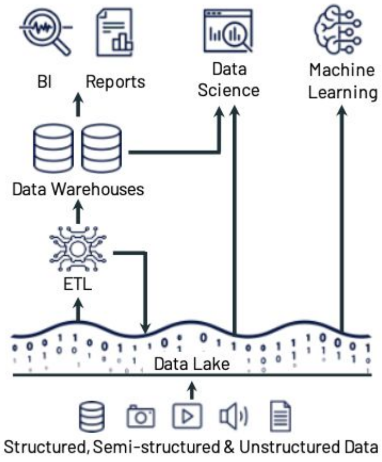
 11th Annual Conference on Innovative Data Systems Research (CIDR '21), January 11–15, 2021, Online.

https://databricks.com/de/wp-content/uploads/2020/12/cidr_lakehouse.pdf

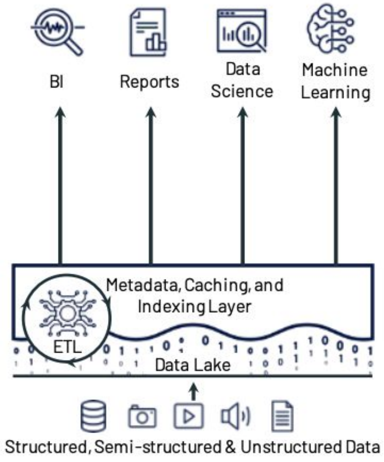
First generation platforms



Current two-tiers architectures



Lakehouse platform

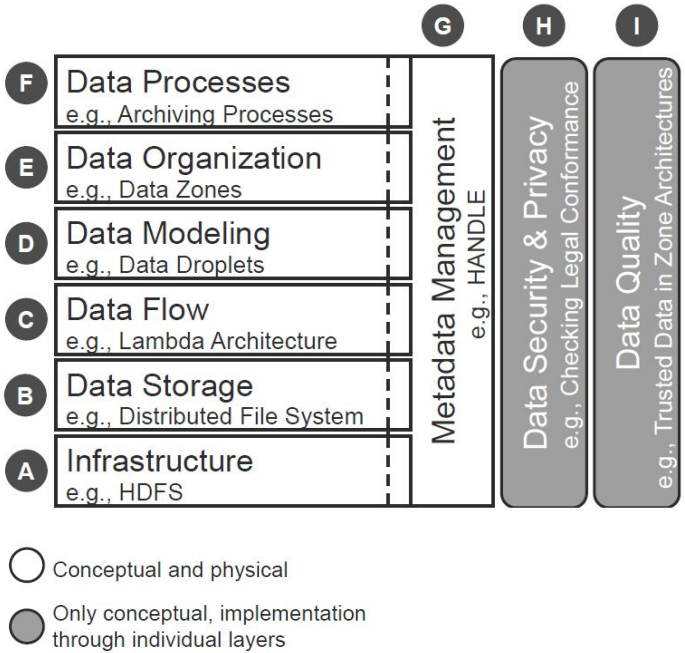


Source: Armbrust et al (2021)

Data platform architectures

Architecture aspects

Aspects

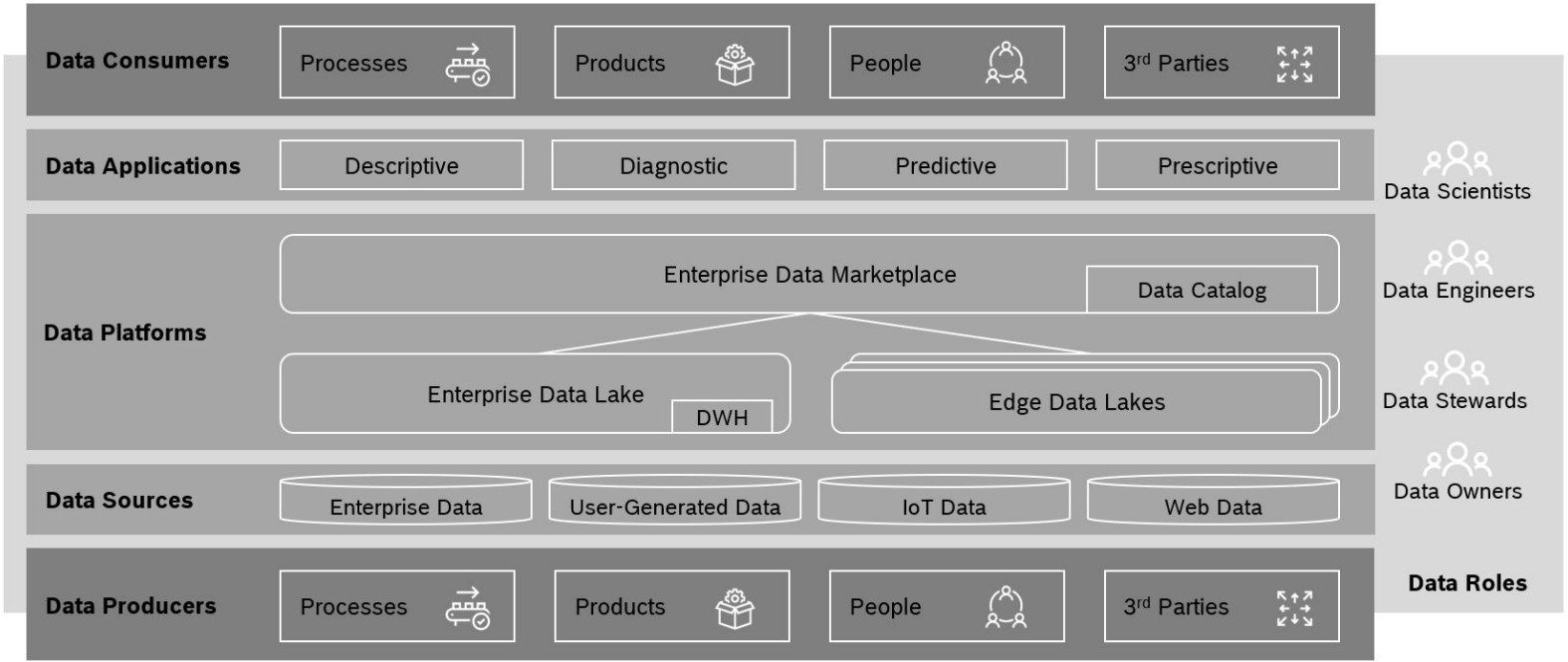


Example

DLAF Layer	AIRPORTS DL [Ma17a]
A. Infrastructure	Hadoop (HDFS, MapReduce), Apache Flume, Apache Spark, Apache Oozie, Apache Pig, Apache Atlas, R Studio, Shiny, Apache Sqoop
B. Data Storage	Single File System
C. Data Flow	Data are ingested as streams, but processed as batches
D. Data Modeling	Raw Messages, AIRPORTS Data Model
E. Data Organization	Four Zone Architecture
F. Data Processes	Processing Pipeline for Messages (ETL Processes), Processes for Ingestion and Use
G. Metadata Management	Managed by Apache Atlas
H. Data Security & Privacy	Tracking manipulation of data
I. Data Quality	Tracking manipulation of data, Quality through Zones

Source: Giebler et al. (2021) https://dx.doi.org/10.18420/btw2021-19

Data platforms
for industrial enterprises



Core elements of a data ecosystem for industrial enterprises

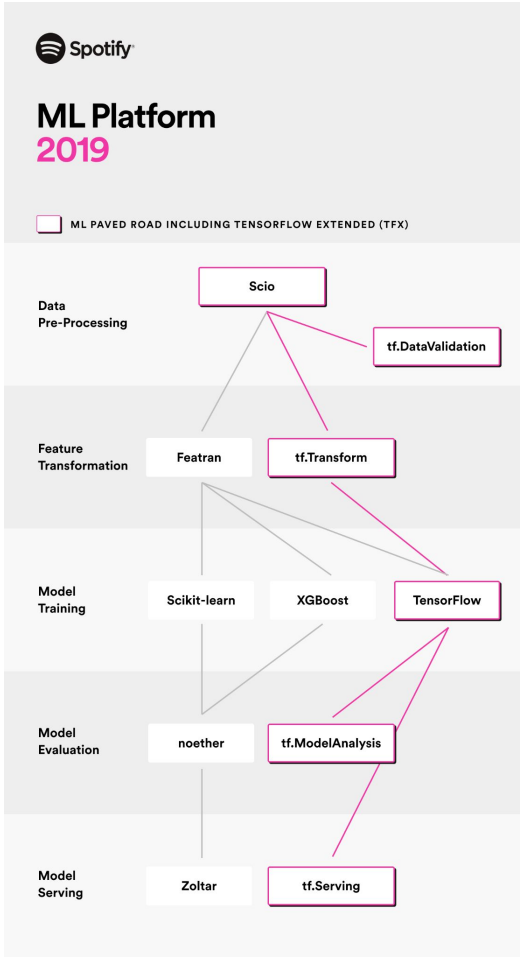
Source: Gröger (2021)

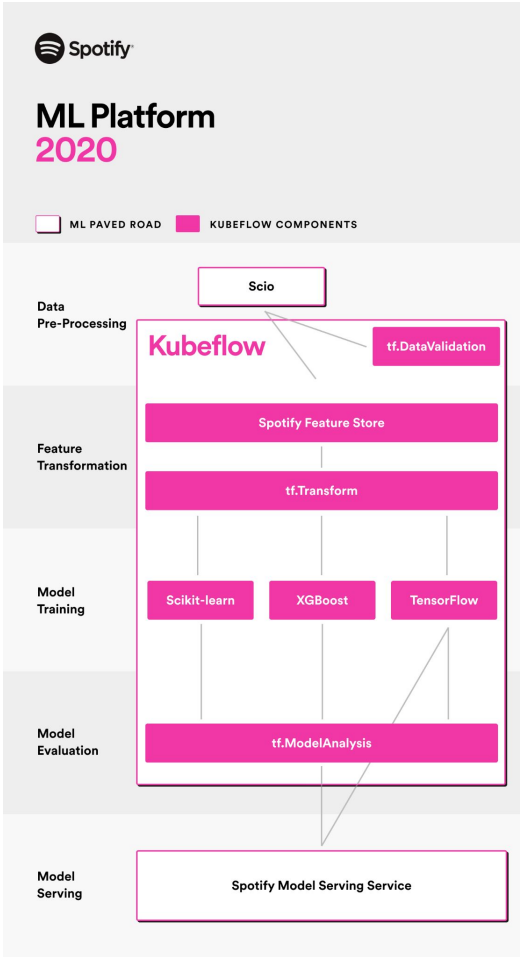
ML platform @ Spotify
Reference architecture

The Winding Road to Better Machine Learning Infrastructure Through Tensorflow Extended and Kubeflow

December 13, 2019
Published by Josh Baer, Samuel Ngahane

https://engineering.atspotify.com/2019/12/13/the-winding-road-to-better-machine-learning-infrastructure-through-tensorflow-extended-and-kubeflow/

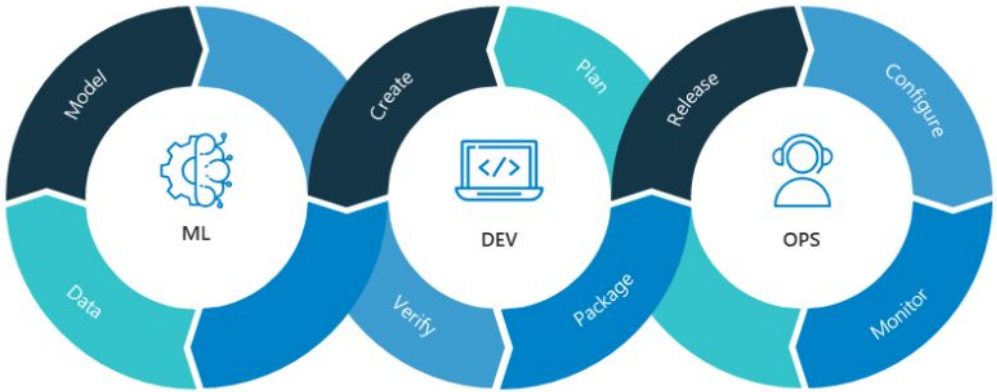




Source: Baer & Ngahane (2019)

Machine learning operations (MLOps)

- ML Engineering culture and practice that aims at **unifying** ML System **development** (Dev) and ML system **operations** (Ops)
- Tools and principles to support workflow **standardization** and **automation** through the ML system lifecycle**.**

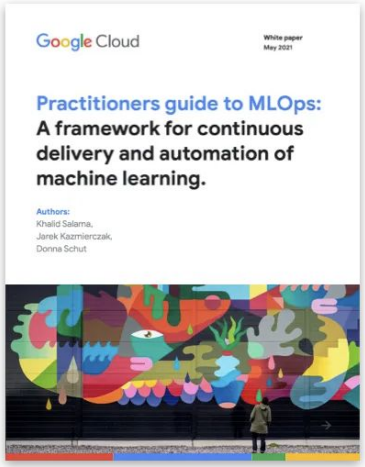


Source: Nvidia (2021) https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/

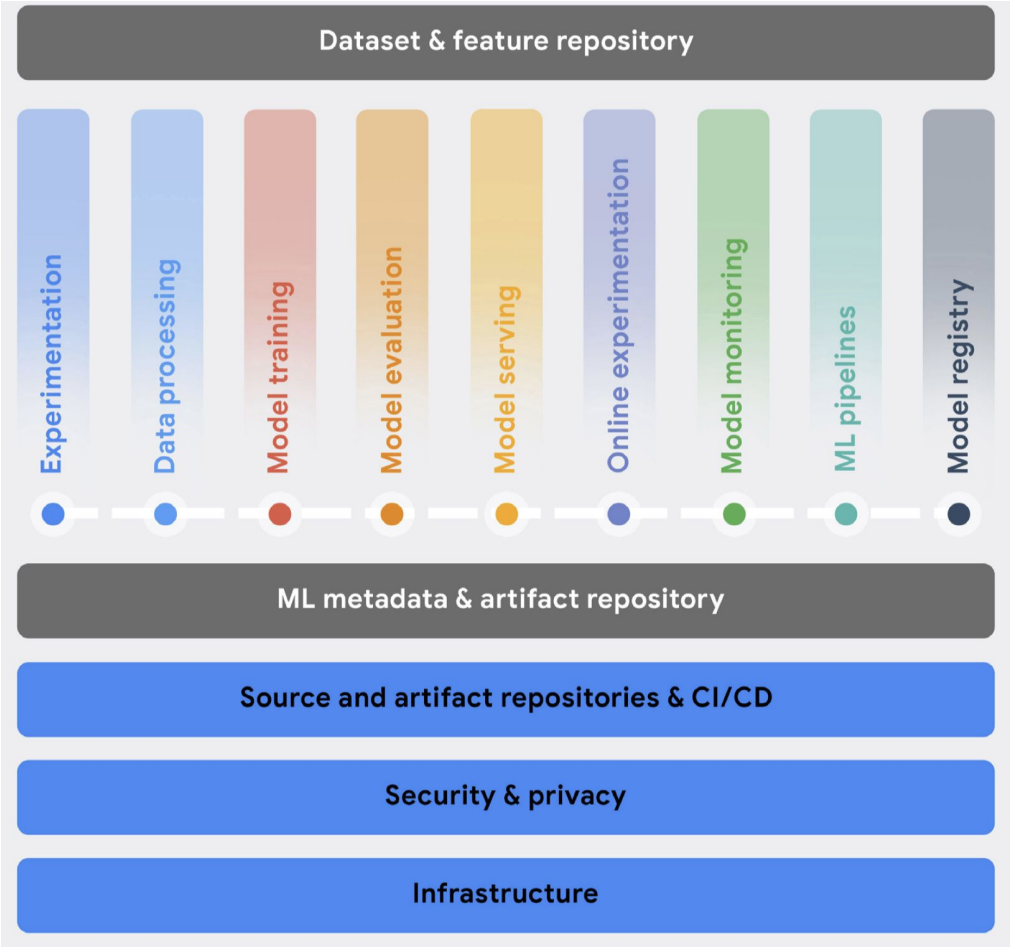
Source: Salama, Kazmierczak & Schut (2021)

MLOps
Introduction

Learn the basics of MLOps



https://cloud.google.com/resources/mlops-whitepaper



MLOps Core technical capabilities

Source: Salama, Kazmierczak & Schut (2021)

Introduction to Machine Learning Operations

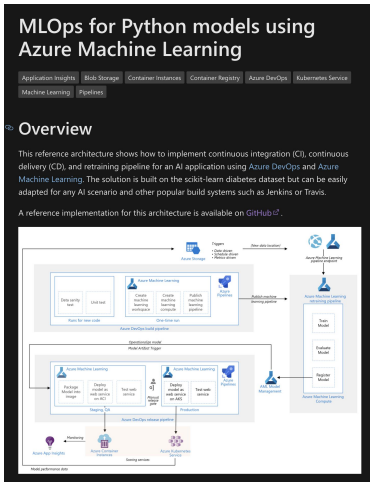
Source: Keating (2020)

MLOps demo (including Kubeflow)

MLOps in Azure

Reference architectures

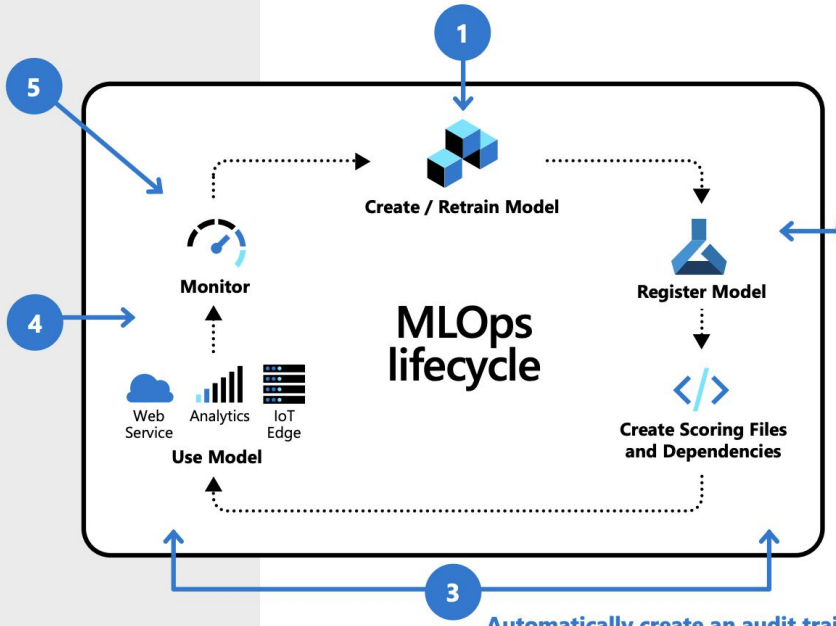
https://azure.microsoft.com/de-de/services/machine-learning/mlops/



https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/mlops-python

5 Best Practices to optimize your MLOps lifecycle on Azure:

Create models with reusable ML pipelines using the Azure Machine Learning extension for Azure DevOps. Store your code in GitHub so it automatically integrates into your MLOps pipeline.



Observe data drift and feed back model information to improve future training.

Deploy and monitor performance so you can release models with confidence and know when to retrain.

Automate your MLOps rollout using Azure DevOps + Azure Machine Learning for version models with rich metadata and event management.

Automatically create an audit trail for all artifacts in your MLOPs pipeline ensure asset integrity and meet regulatory requirements.

Source: Microsoft (2021)

Resources

Armbrust, M., Ghodsi, A., Xin, R. & Zaharia, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. 11th Annual Conference on Innovative Data Systems Research (CIDR '21), January 11–15, 2021, Online https://databricks.com/de/wp-content/uploads/2020/12/cidr_lakehouse.pdf

Baer, J. & Samuel Ngahane, S. (2019). The Winding Road to Better Machine Learning Infrastructure Through Tensorflow Extended and Kubeflow. https://engineering.atspotify.com/2019/12/13/the-winding-road-to-better-machine-learning-infrastructure-through-tensorflow-extended-and-kubeflow/

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., Mitschang, B. The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture. In: Proceedings of the 19th Conference for Database Systems for Business, Technology and Web (BTW), pp. 351-370. https://dx.doi.org/10.18420/btw2021-19

Gröger, C. (2021). There Is No AI Without Data. Industry Experiences on the Data Challenges of AI and Call for a Data Ecosystem for Industrial Enterprises. In: Communications of the ACM, 2021, to appear. http://christophgroeger.de/download/Groeger_There_Is_No_AI_Without_Data.pdf

Kirenz, J. (2021). MLOps with Tensor Flow Extended & Kubeflow. Presentation https://de.slideshare.net/JulianKirenz/mlops-build-pipelines-with-tensorflow-extended-kubeflow

Salama, K., Kazmierczak, J. & Schut, D. (2021). Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning. Google Whitepaper. https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf