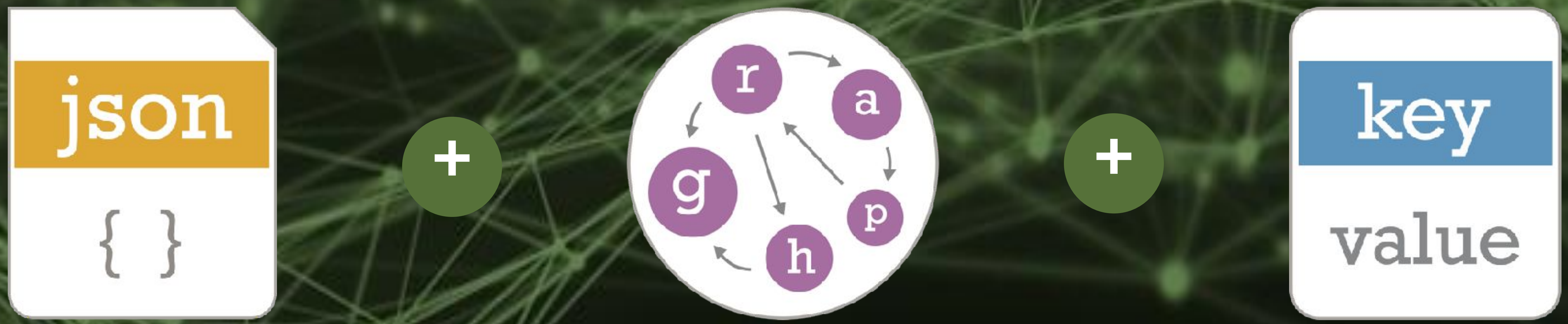
The case for a common metadata layer for machine learning platforms

From Data to Metadata



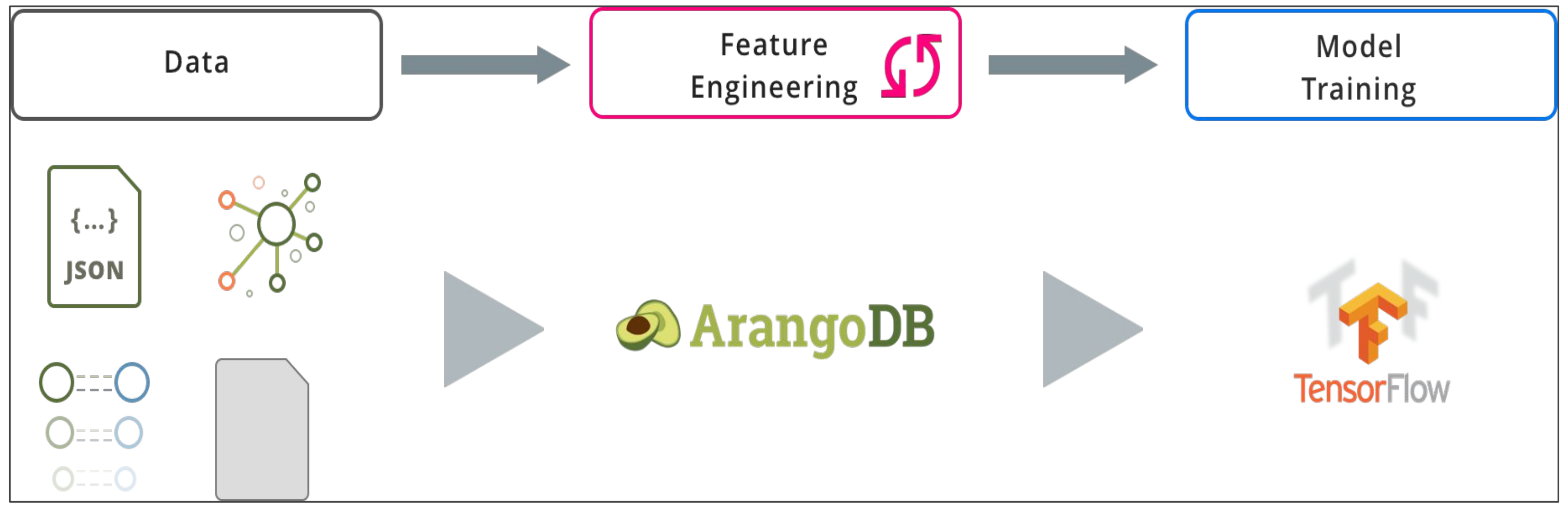
Databases and Machine Learning

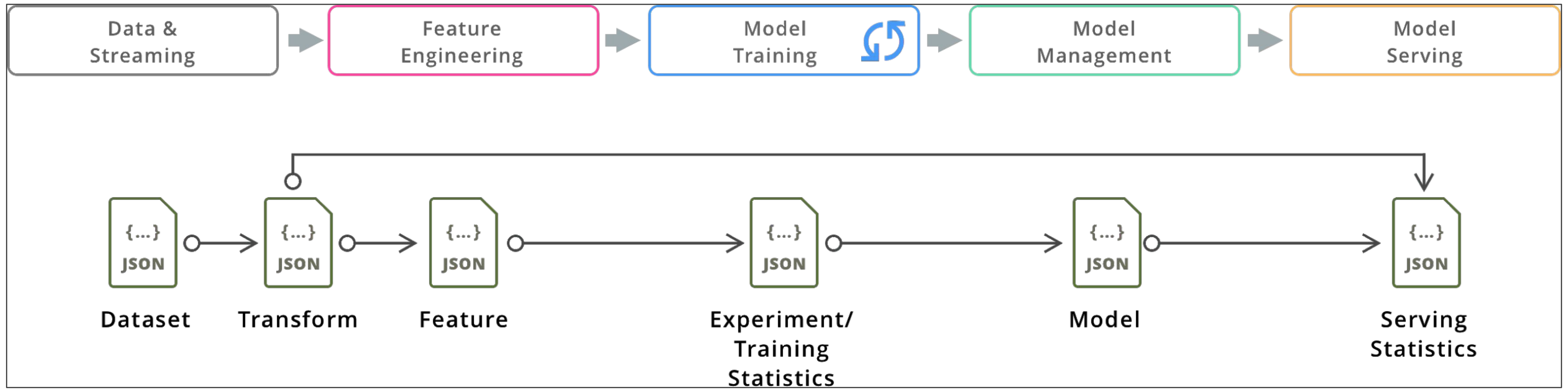
Multi-Model-Powered Machine Learning

- Feature and Model Engineering

Databases for Machine Learning Infrastructure

- Utilize Multi-Model for managing heterogeneous metadata across Machine Learning Pipelines

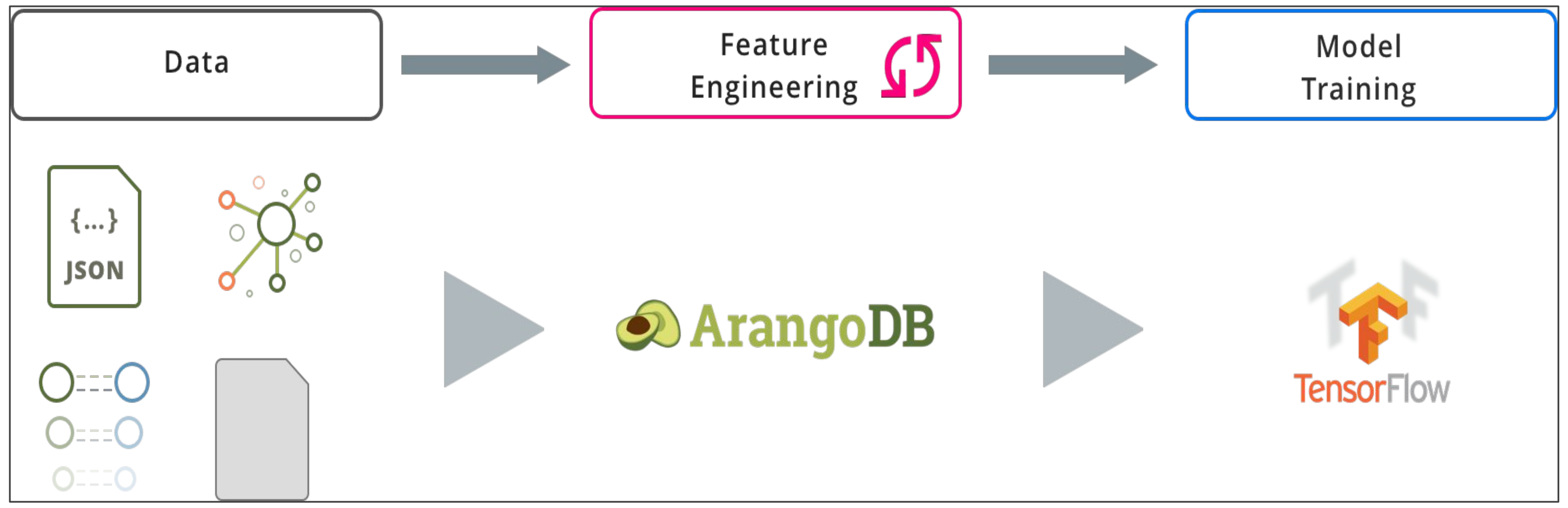




ArangoML

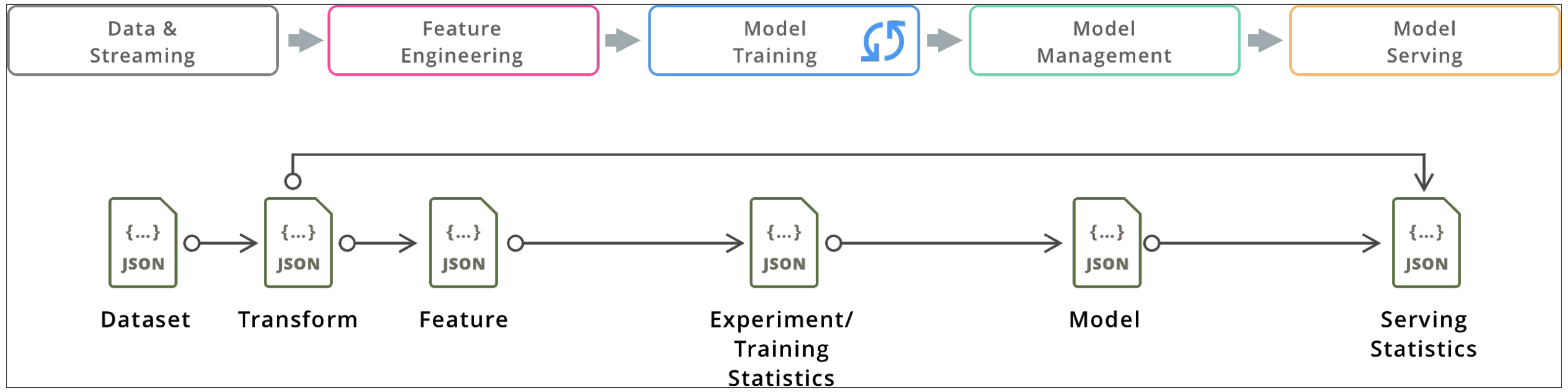
Multi-Model-Powered Machine Learning

- Feature and Model Engineering



Databases for Machine Learning Infrastructure

- Utilize Multi-Model for managing heterogeneous metadata across Machine Learning Pipelines

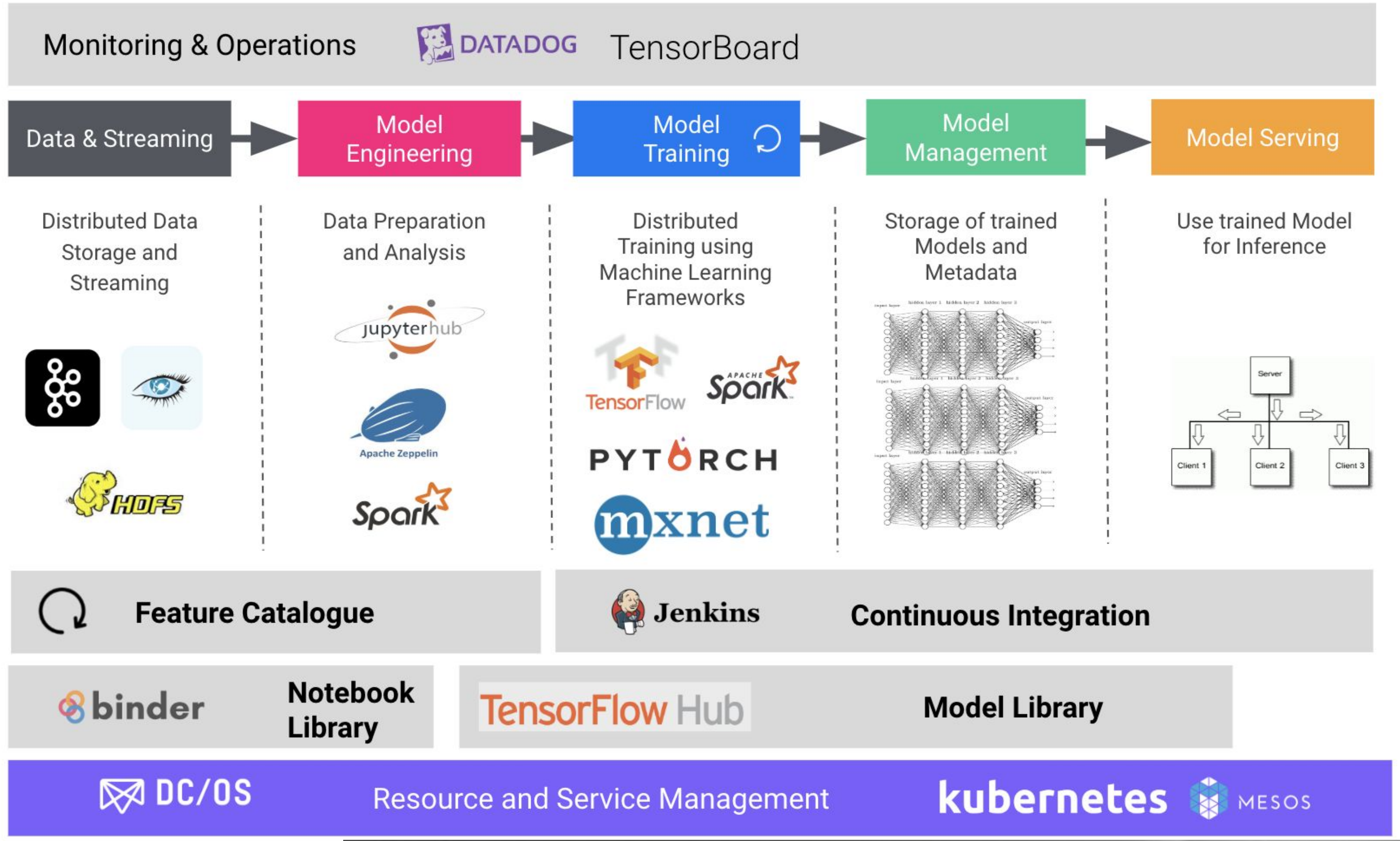


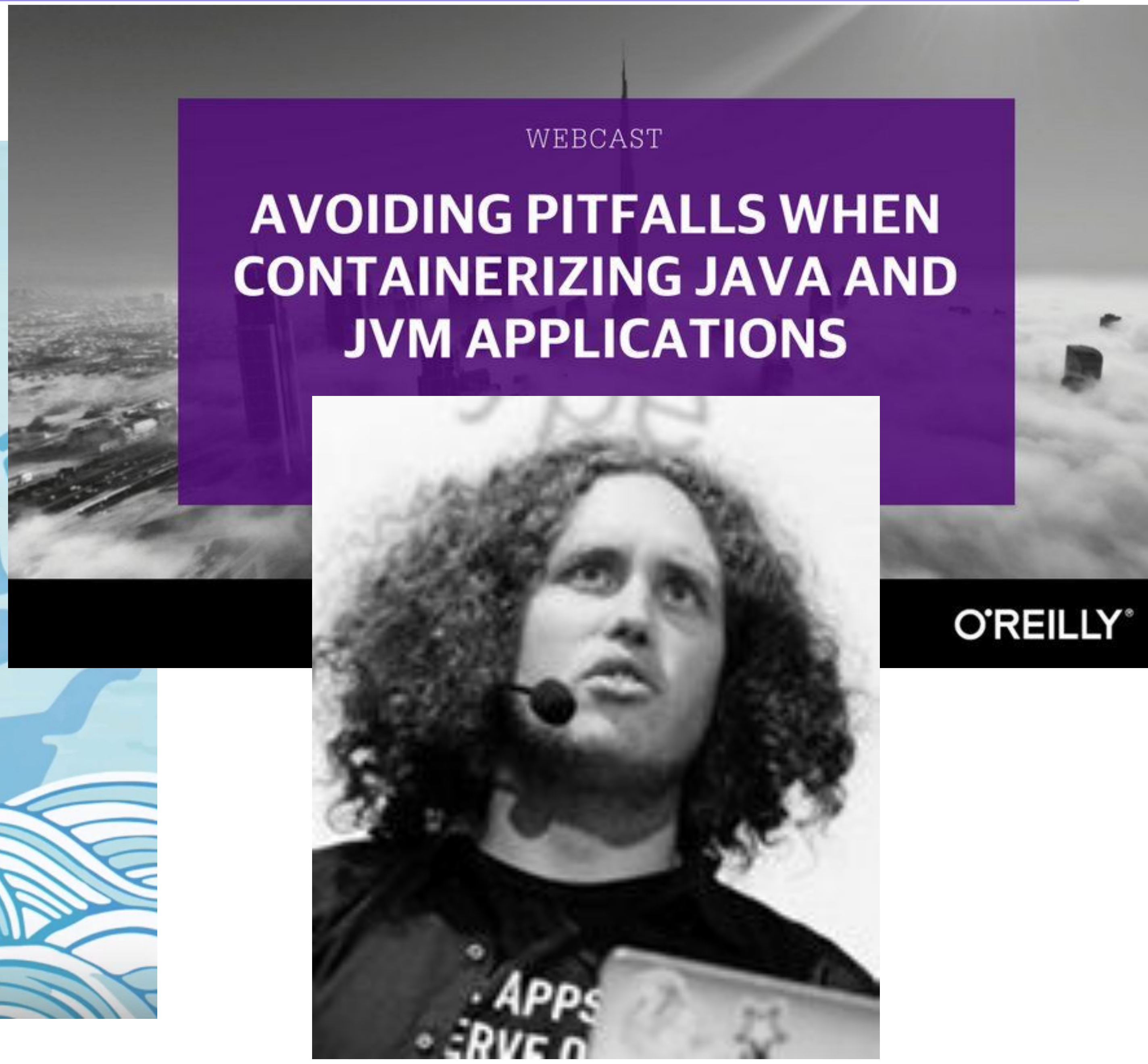
Jörg Schad, PhD

Head of Engineering and ML @ArangoDB

- **Suki.ai**
- **Mesosphere**
- **Architect @SAP Hana**
- **PhD Distributed DB Systems**

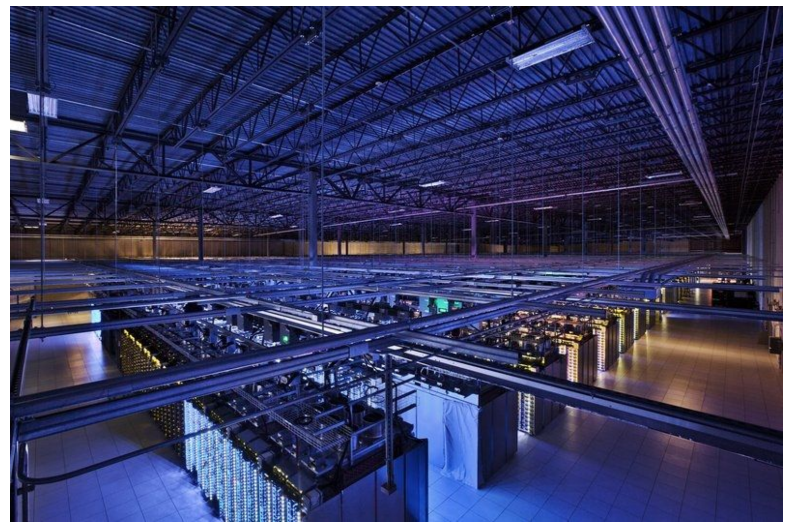
- **Twitter: @joerg_schad**

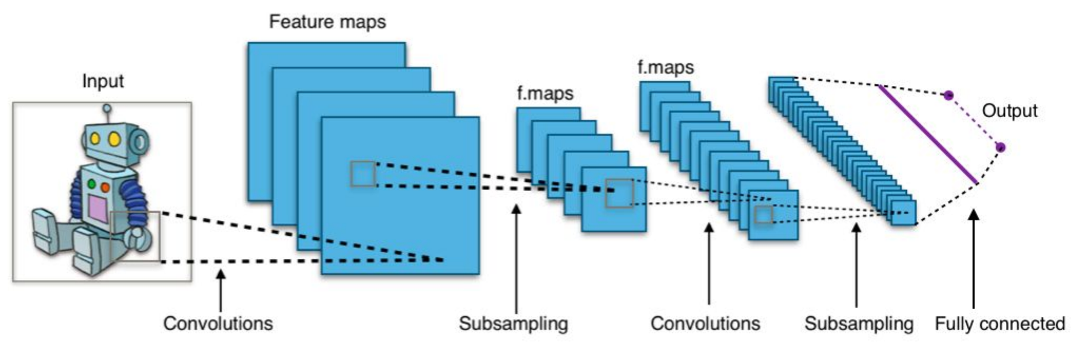




Why is machine learning taking off?


Science
Projects
Data
Free
Dataset
Websites
Gather
Here
Python
Sas
Big
Bank
Use
Datascience
Top
Lists
Best
Excel
Part
World
Curated
Software

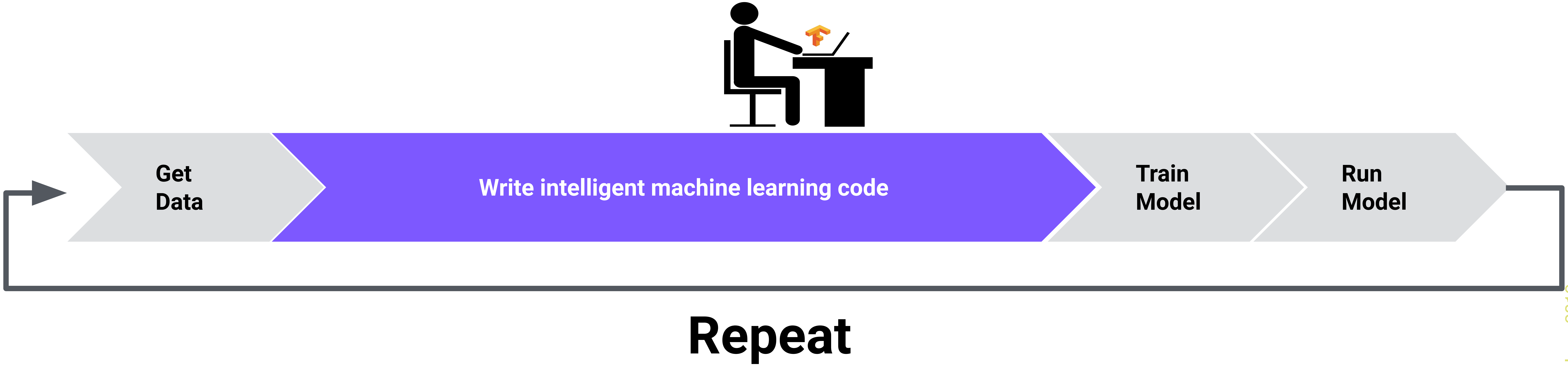



Input
Feature maps
f.maps
f.maps
Output
Convolutions
Subsampling
Convolutions
Subsampling
Fully connected

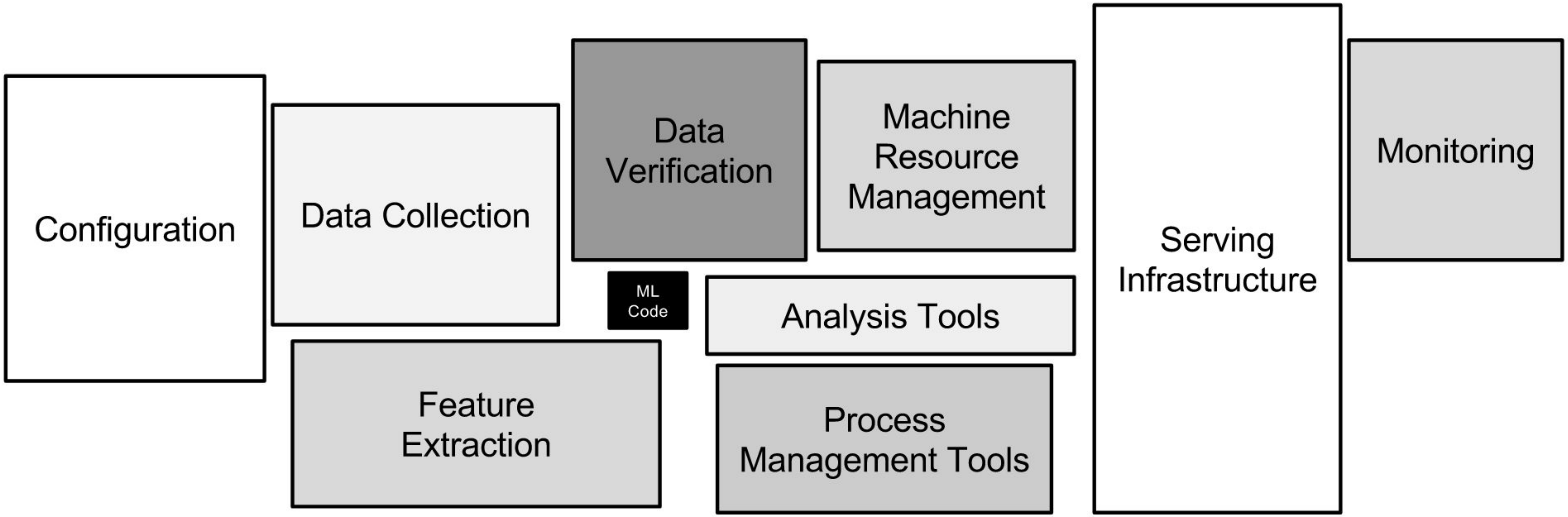

BOSCH

DEEPBACH: A STEERABLE MODEL FOR BACH CHORALES GENERATION

What Data Scientist should be doing…



What Data Scientist are doing…



Sculley, D., Holt, G., Golovin, D. et al. Hidden Technical Debt in Machine Learning Systems

Challenge: Persona(s)

The Rise of the *DataOps Engineer*

Combines two key skills:

- Data science
- Distributed systems engineering

The equivalent of *DevOps* for *Data Science*

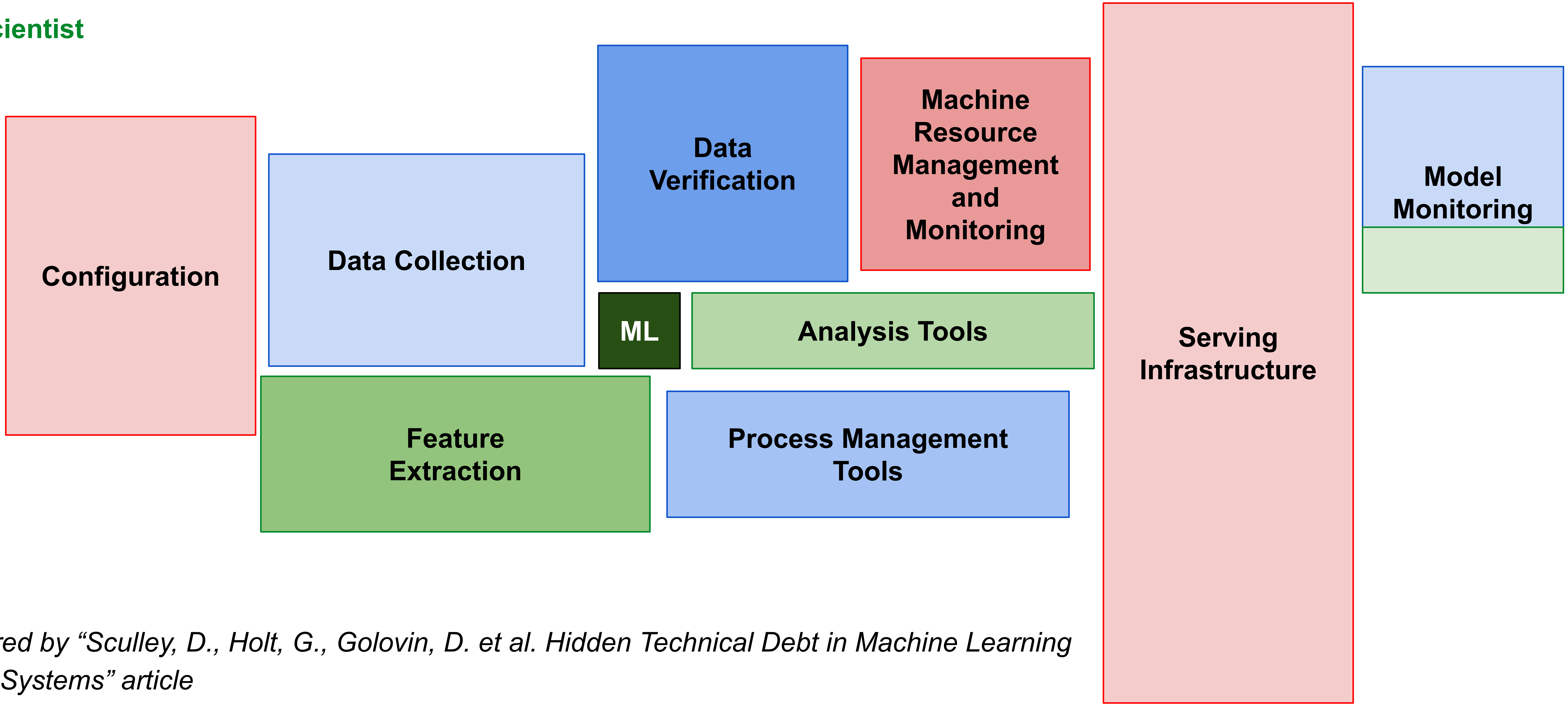
- **Build** automation software to run machine learning systems
- **Operate** systems so they're available, scalable, and performant
- **Evangelize** tools and best practices among data scientists

Division of Labor

System Admin/ DevOps

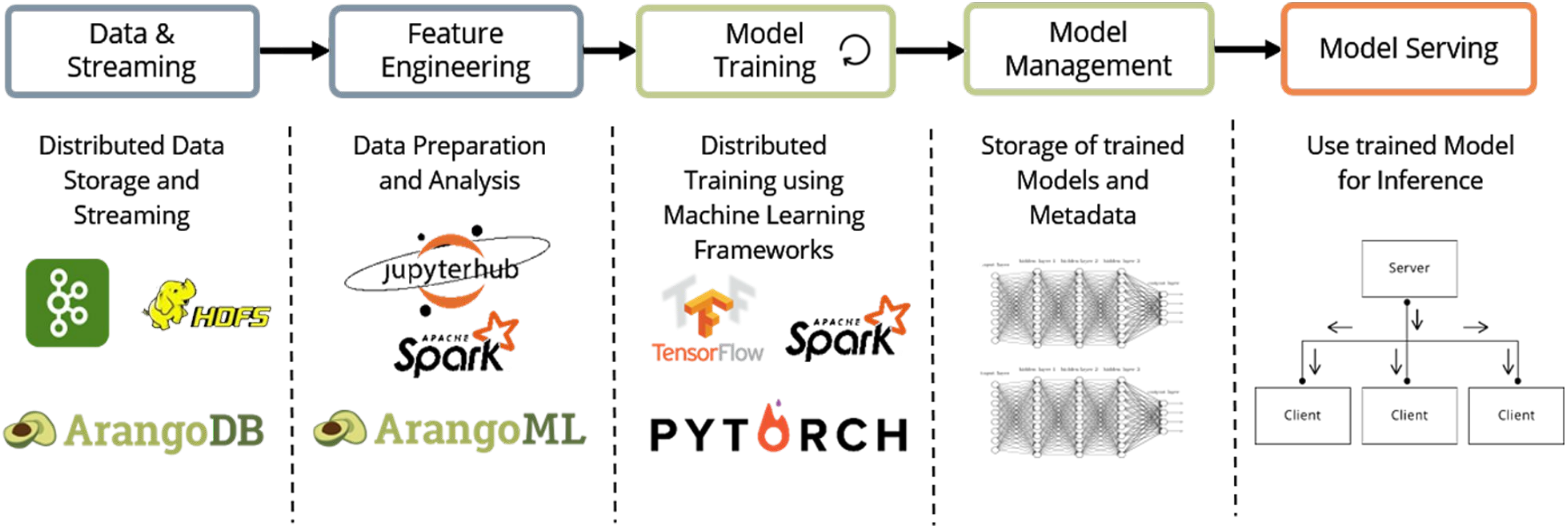
Data Engineer/DataOps

Data Scientist

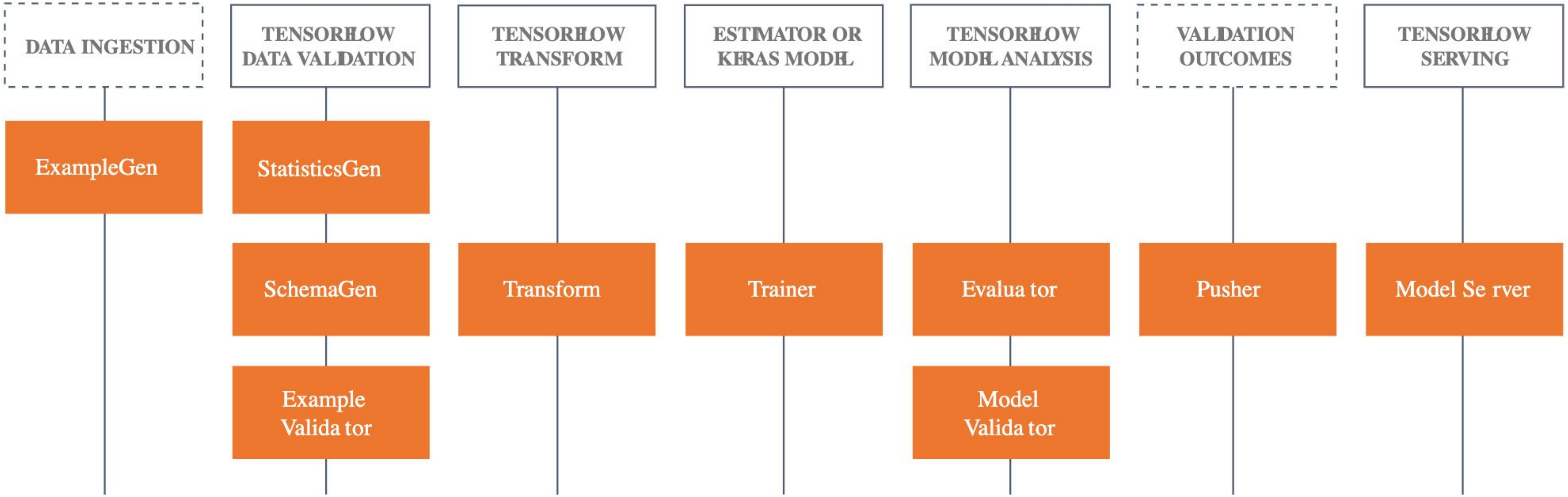


Inspired by "Sculley, D., Holt, G., Golovin, D. et al. Hidden Technical Debt in Machine Learning Systems" article

Machine Learning Pipeline

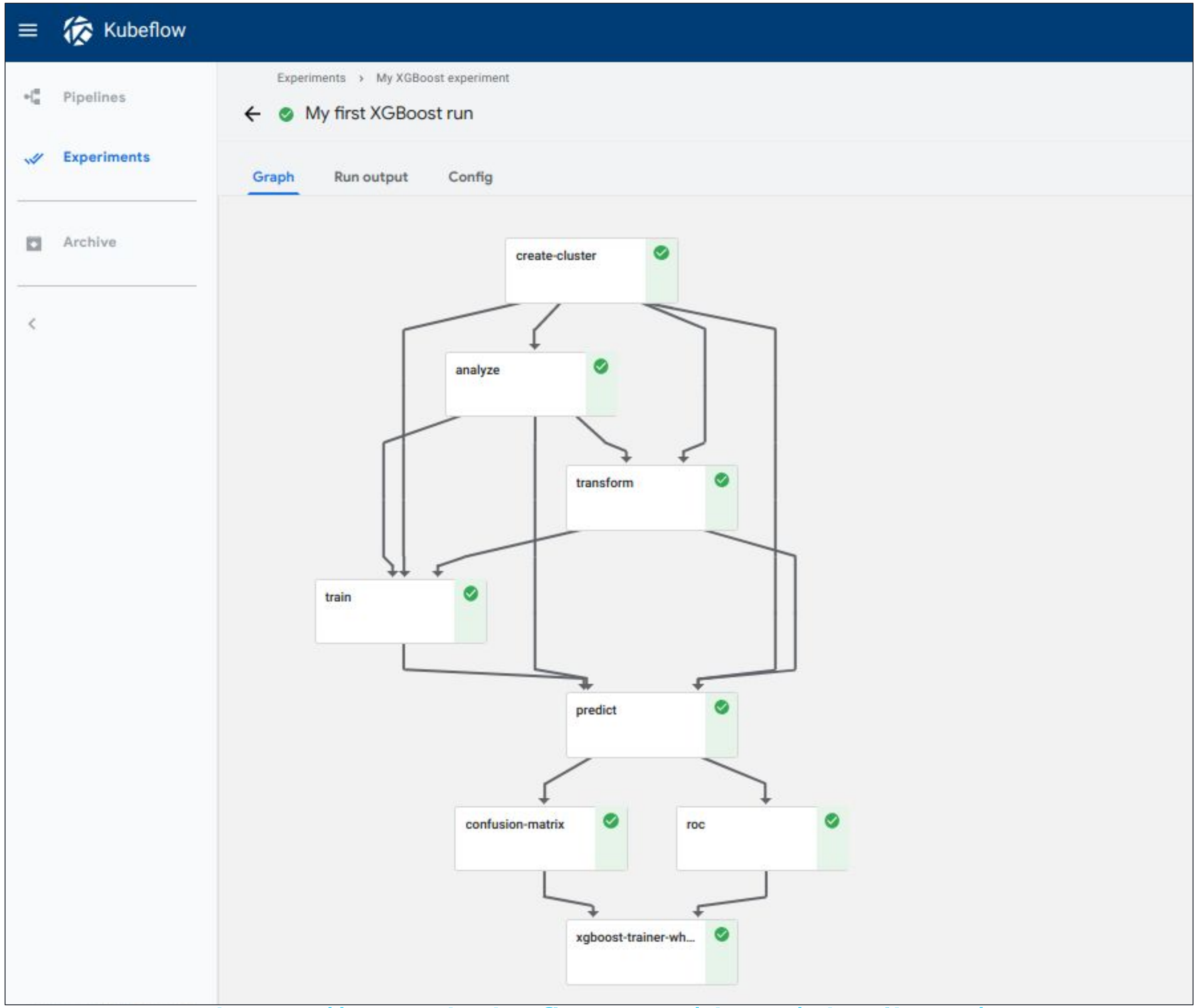


TensorFlow Extended



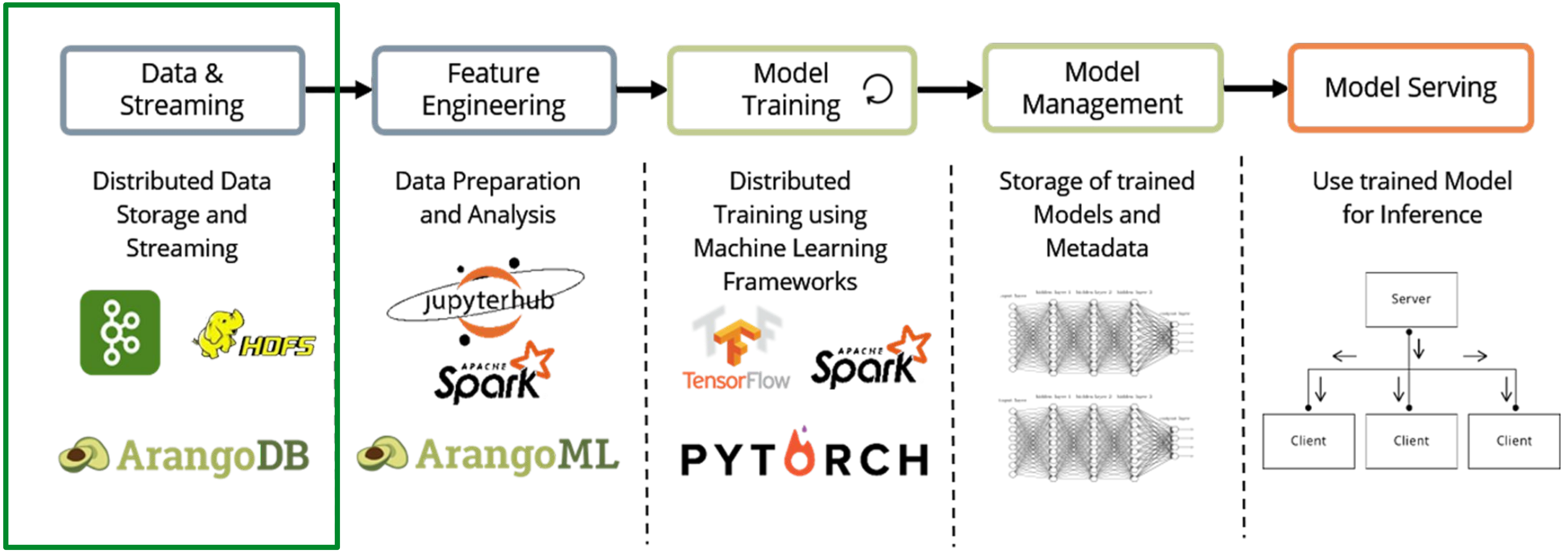
https://www.tensorflow.org/tfx/guide

Kubeflow Pipelines

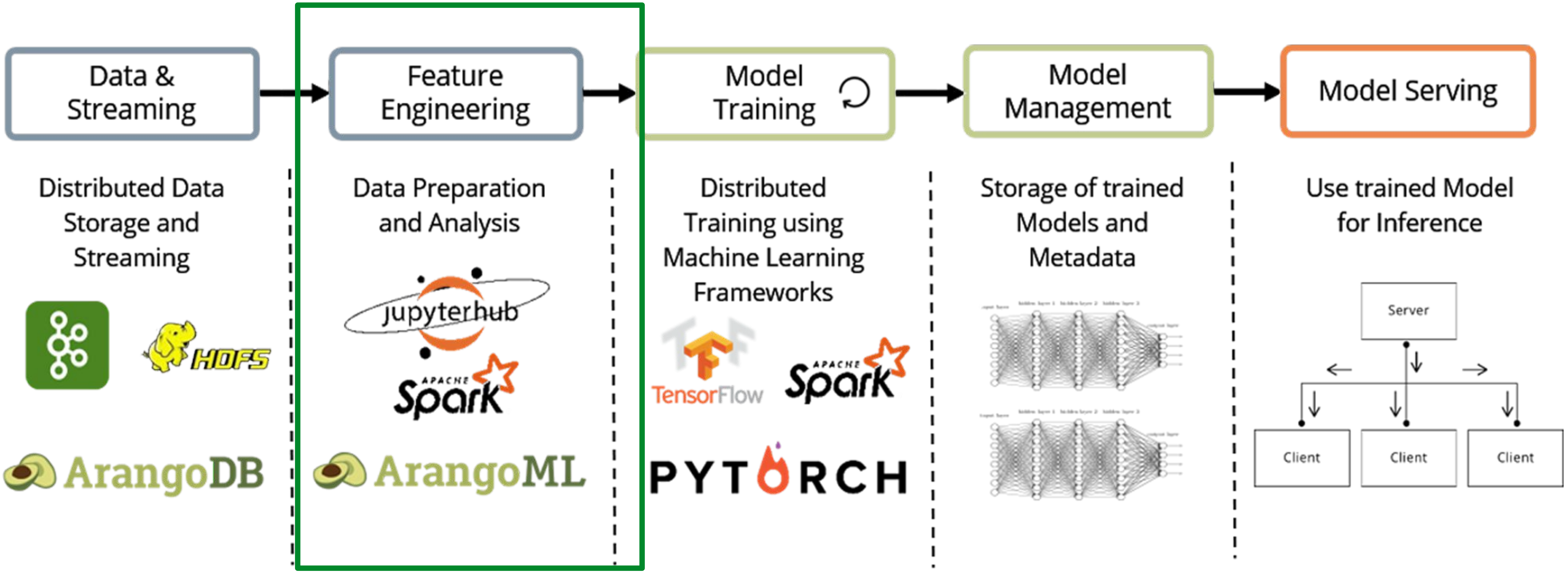


https://www.kubeflow.org/docs/pipelines/

Databases I


Data &
Streaming
Feature
Engineering
Model
Training
Model
Management
Model Serving
Distributed Data
Storage and
Streaming
Data Preparation
and Analysis
Distributed
Training using
Machine Learning
Frameworks
Storage of trained
Models and
Metadata
Use trained Model
for Inference
HDFS
ArangoDB
jupyterhub
APACHE
Spark
ArangoML
TensorFlow
APACHE
Spark
PYTORCH
Server
Client
Client
Client

Databases II



Graphs and Machine Learning

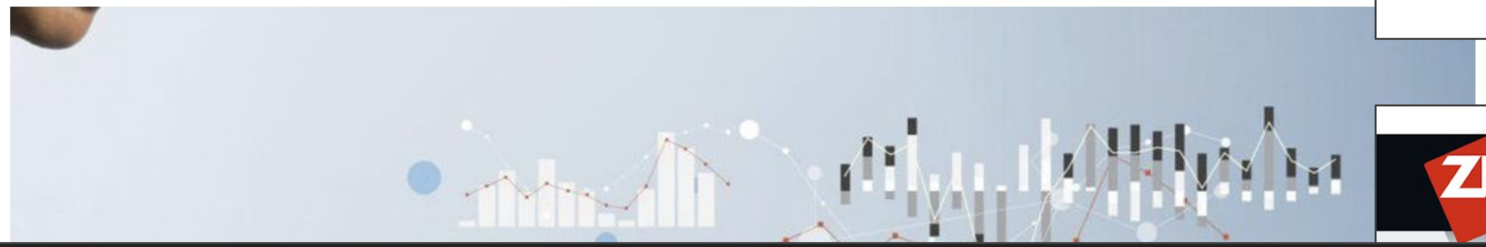
Forbes | Billionaires | Innovation | Leadership | Money | Business | Small Business | Lifestyle | Bran

8,817 views | Jun 26, 2019, 01:00am

Knowledge Graphs And Machine Learning -- The Future Of AI Analytics?

Bernard Marr Contributor
Enterprise & Cloud

The unprecedented explosion in the amount of information we are generating ar collecting, thanks to the arrival of the internet and the always-online society, po all the incredible advances we see today in the field of artificial intelligence (AI) Big Data.



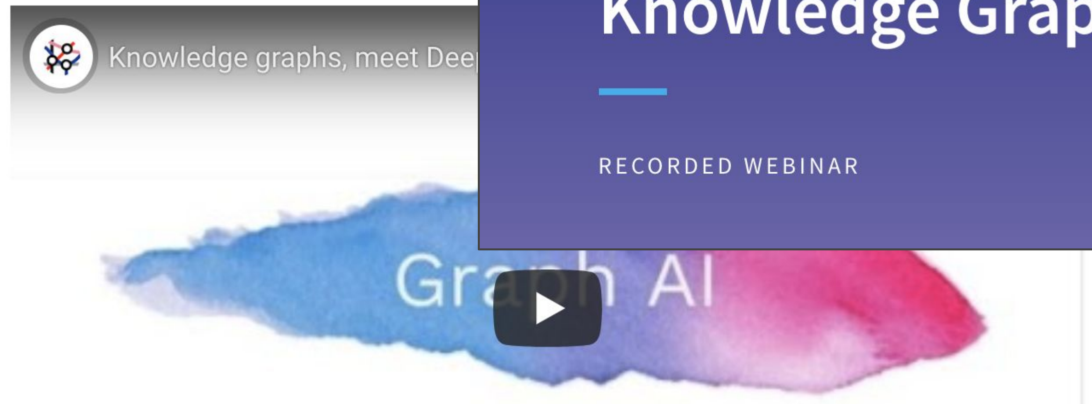
M | OCTAVIAN

Deep Learning with Knowledge Graphs

Andrew Jefferson Follow
Nov 13, 2018 · 8 min read

Last week I gave a talk at Connected Data London on the approach that we have developed at Octavian to use neural networks to perform tasks on knowledge graphs.

Here's the recording of the talk, f



Semantic AI: Bringing Machine Learning and Knowledge Graphs Together

RECORDED WEBINAR

arXiv.org > cs > arXiv:1908.06556

Search...
Help | Advanced

Computer Science > Computation and Language

Transfer in Deep Reinforcement Learning using Knowledge Graphs

Prithviraj Ammanabrolu, Mark O. Riedl

(Submitted on 19 Aug 2019)

Text adventure games, in which players must make sense of the world through text descriptions and declare actions through text descriptions, provide a stepping stone toward grounding action in language. Prior work has demonstrated that using a knowledge graph as a state representation and question-answering to pre-train a deep Q-network facilitates faster control policy transfer. In this paper, we explore the use of knowledge graphs as a representation for domain knowledge transfer for training text-adventure playing reinforcement learning agents. Our methods are tested across multiple computer generated and human authored games, varying in domain and complexity, and demonstrate that our transfer learning methods let us learn a higher-quality control policy faster.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: arXiv:1908.06556 [cs.CL]
(or arXiv:1908.06556v1 [cs.CL] for this version)

EDITION: EU

ZDNet | CENTRAL EUROPE | MIDDLE EAST | SCANDINAVIA | AFRICA | UK | ITALY | SPAIN | MORE | NEWSLE

MUST READ: Innovation: How to get your great ideas approved

Salesforce Research: Knowledge graphs and machine learning to power Einstein

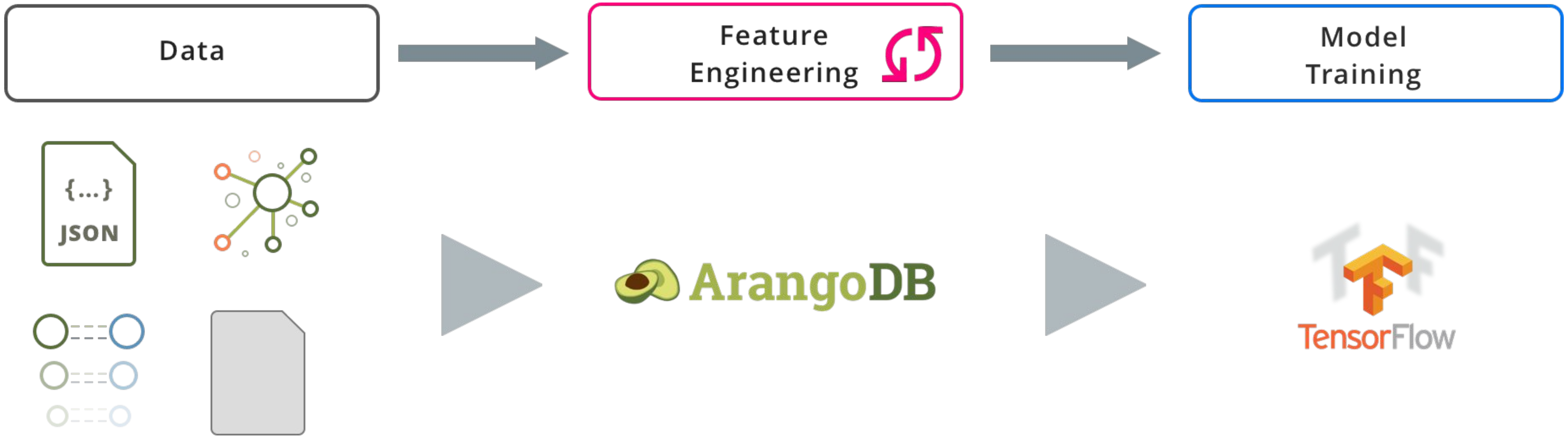
Explainable AI in real life could mean Einstein not just answering your questions, but also providing justification. Advancing the state of the art in natural language processing is done on the intersection of graphs and machine learning.

By George Anadiotis for Big on Data | March 18, 2019 -- 16:49 GMT (16:49 GMT) | Topic: Artificial Intelligence

Feature Engineering

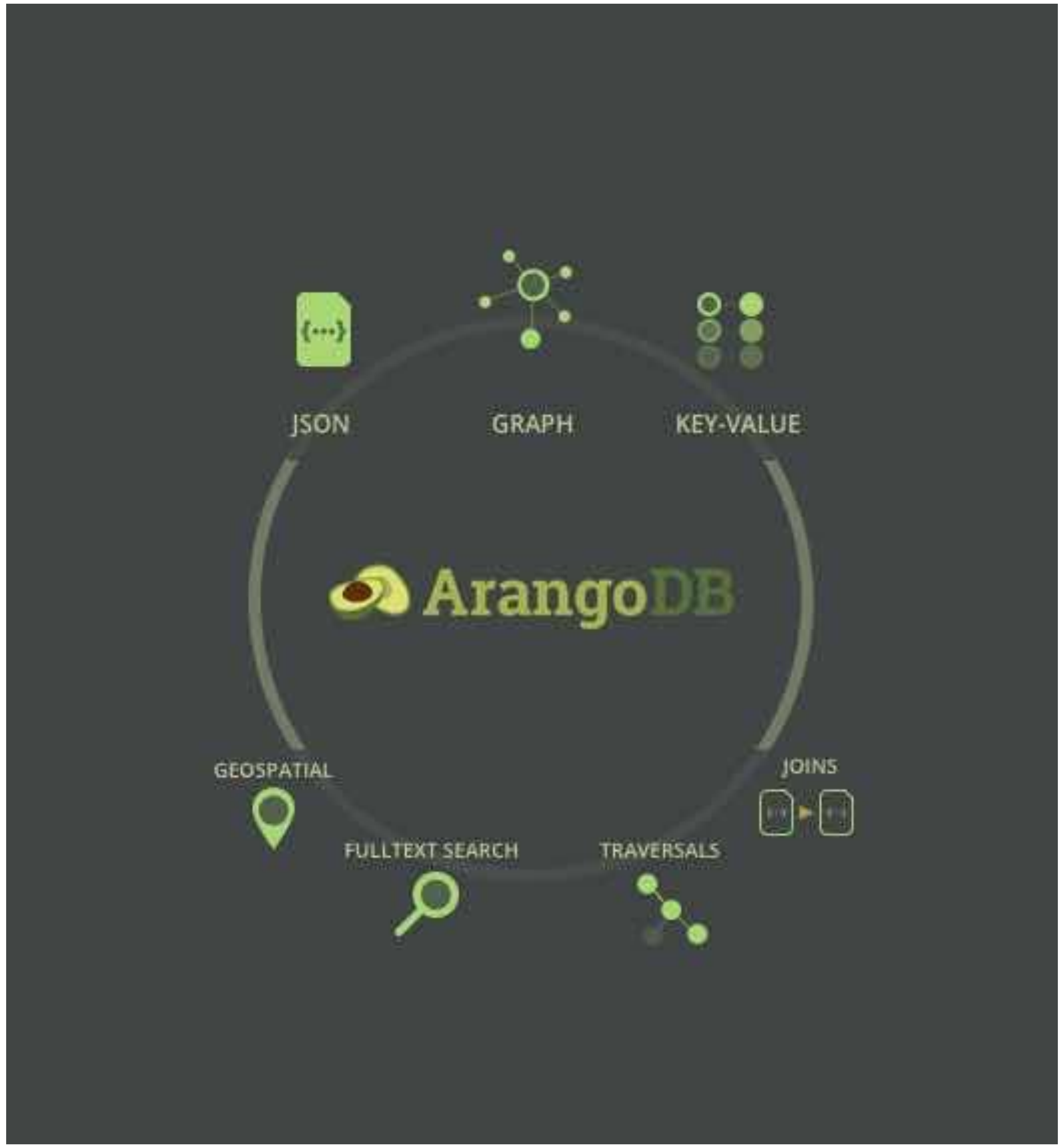
Director	Number Movies
George_S_Fleming	10
….	…..

Feature Engineering



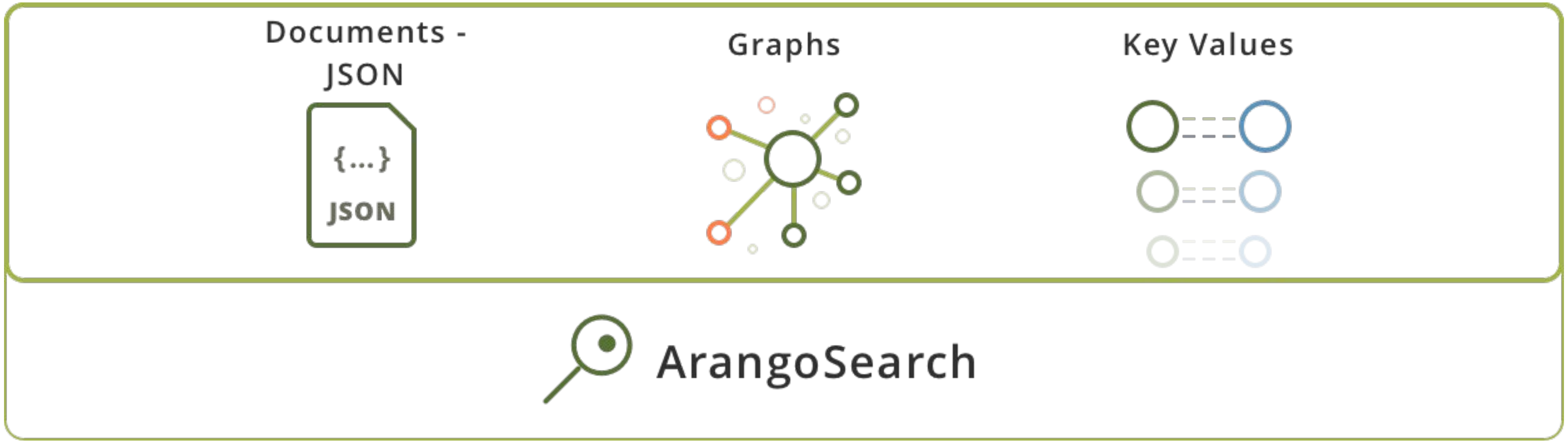
ArangoDB

- Native Multi Model Database
    - Stores, K/V, Documents & Graphs
- Distributed
    - Graphs can span multiple nodes
- AQL - SQL-like multi-model query language
- ACID Transactions including Multi Collection Transactions

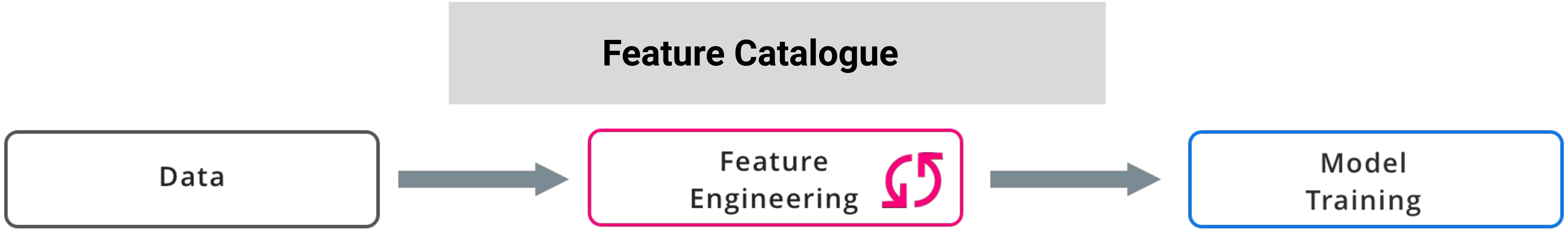


Multi-Model?

ArangoDB



Feature Catalogue

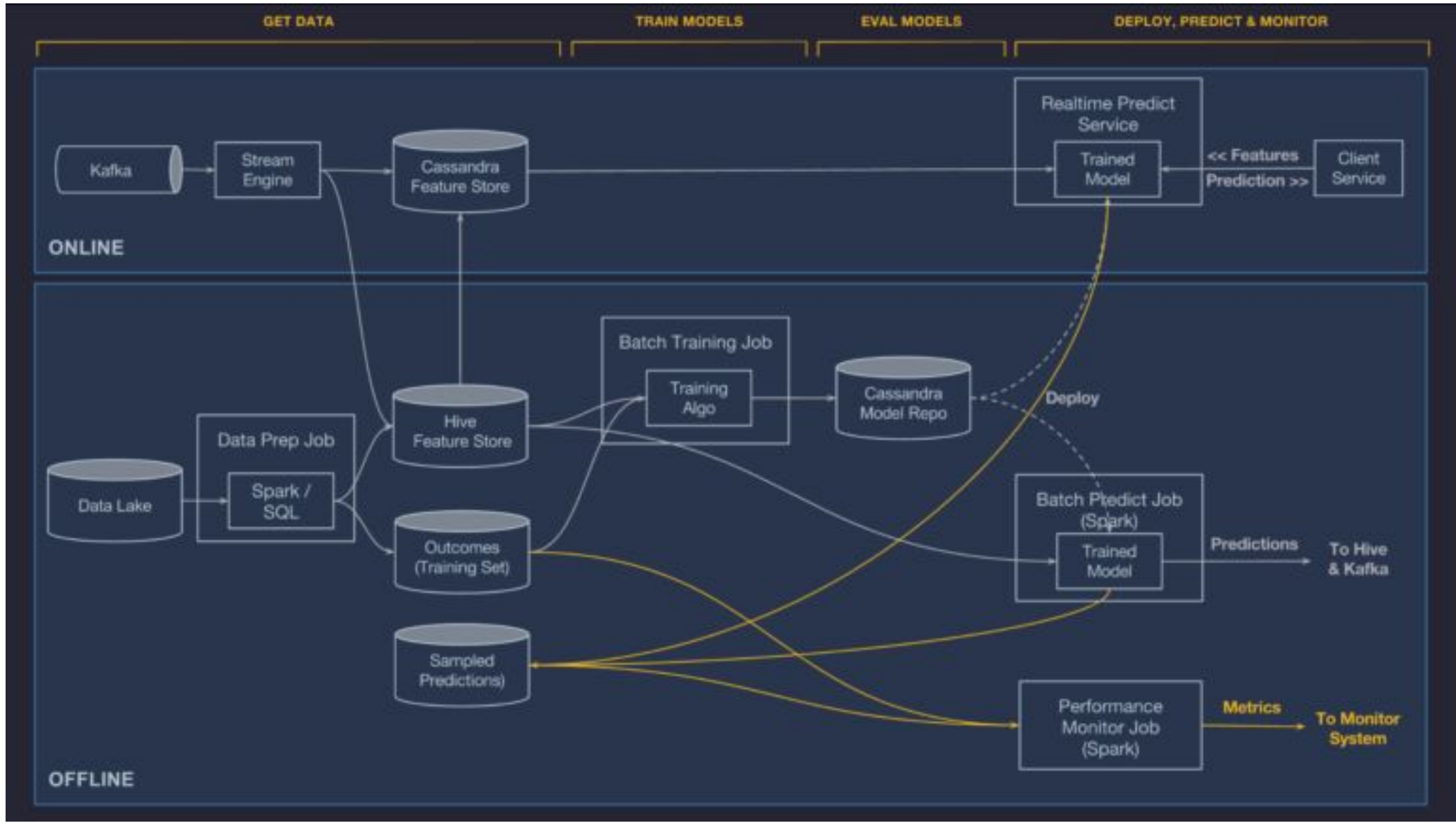


- Feature Catalogue ≈ Preprocessing Cache + Discovery
  - Uber Michelangelo
  - Logical Clocks
  - Kubeflow FEAST

Uber Michelangelo

“..there were no systems in place to build reliable, uniform, and reproducible pipelines for creating and managing training and prediction data at scale.”

- **Feature store**

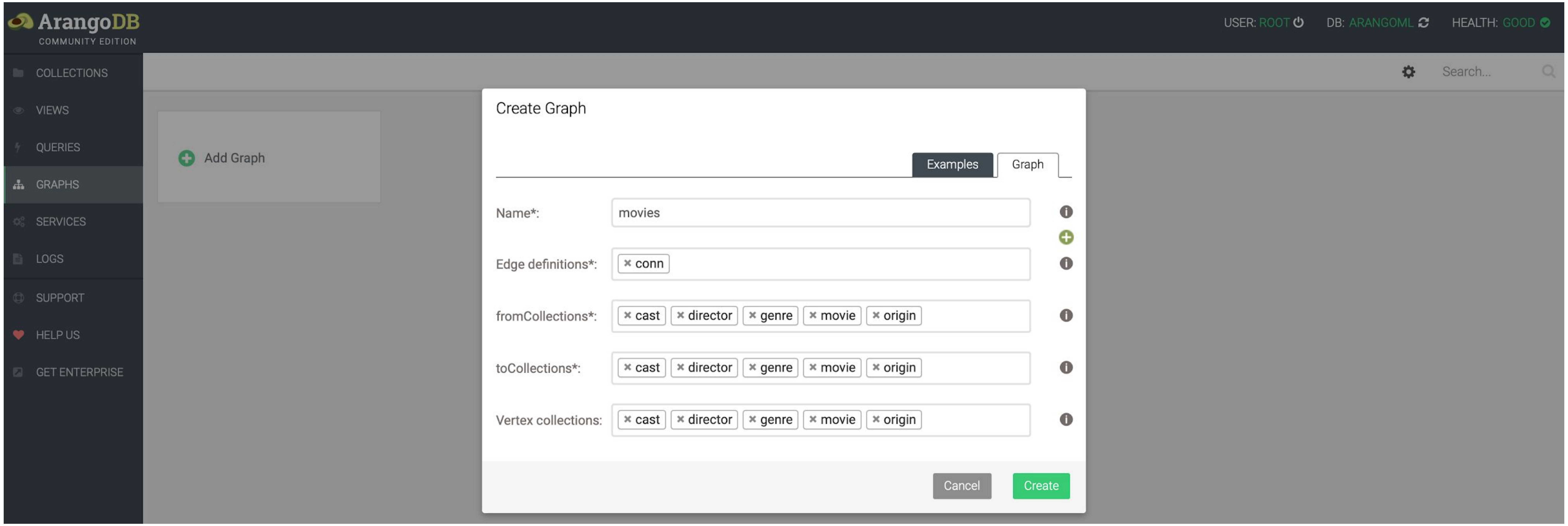


https://eng.uber.com/michelangelo/

Feature Store

https://techblog.appnexus.com/lessons-learned-from-building-scalable-machine-learning-pipelines-822acb3412ad

Multi-Model ML Demo



https://github.com/arangoml/knowlegegraph-demo

https://mybinder.org/v2/gh/arangoml/knowlegegr
aph-demo/master?filepath=movie_data_graph.ipyn

What is next?

The Graph Neural Network Model

Publisher: IEEE

5 Author(s) Franco Scarselli ; Marco Gori ; Ah Chung Tsoi ; Markus Hagenbuchner ; Gabriele Monfardini **View All Authors**

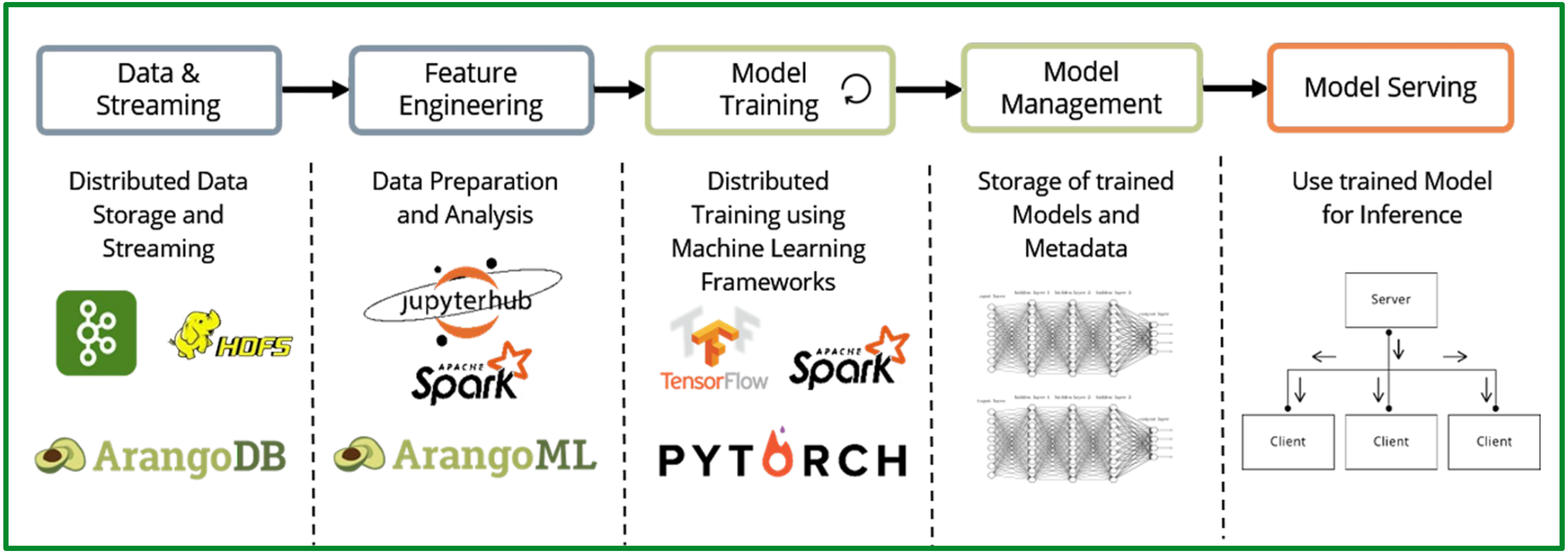
Abstract

Document Sections

Abstract:
Many underlying relationships among data in several areas of science and engineering, e.g., computer vision, molecular chemistry, molecular biology, pattern recognition, and data mining, can be represented in terms of graphs. In this paper, we propose a new neural network model, called graph neural network (GNN) model, that extends existing neural network methods for processing the data represented in graph domains. This GNN model, which can directly process most of the practically useful types of graphs, e.g., acyclic, cyclic, directed, and undirected, implements a function tau(G,n) isin IR^m that maps a graph G and one of its nodes *n* into an *m*-dimensional Euclidean space. A supervised learning algorithm is derived to estimate the parameters of the proposed GNN model. The computational cost of the proposed algorithm is also considered. Some experimental results are shown to validate the proposed learning algorithm, and to demonstrate its generalization capabilities.

https://ieeexplore.ieee.org/abstract/document/4700287

Databases III



Challenges

The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

Nicholas Carlini[1,2] Chang Liu[2] Úlfar Erlingsson[1] Jernej Kos[3] Dawn Song[2]

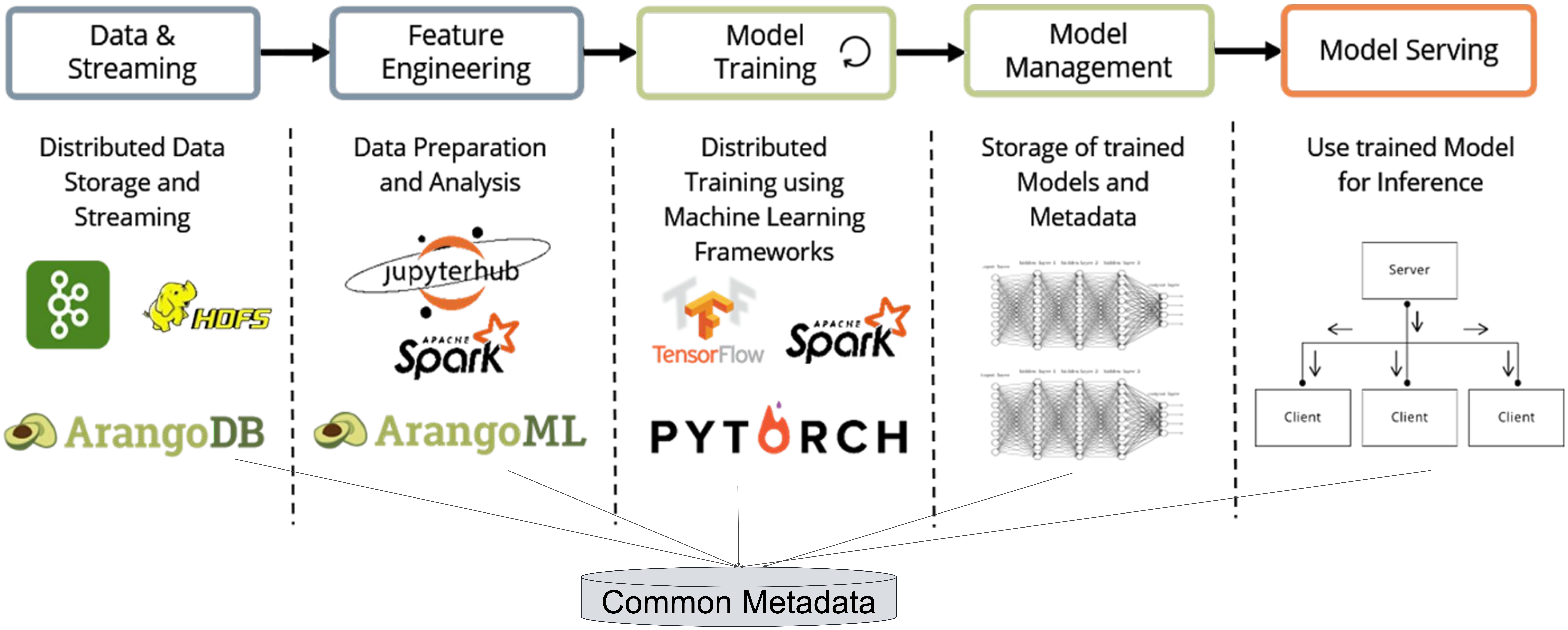
[1]*Google Brain* [2]*University of California, Berkeley* [3]*National University of Singapore*

https://blog.acolyer.org/2019/09/23/the-secret-sharer/

Challenges

- **Understand complete provenance of Model**
    a. Understand Provenance
    b. Complete version history
    c. Audit
- **Find all Models in production derived from dataset x**
- **Compare performance of different model performance**
- **Identify reusable steps**
- **Is my serving data distribution the same as for training data**
- ...

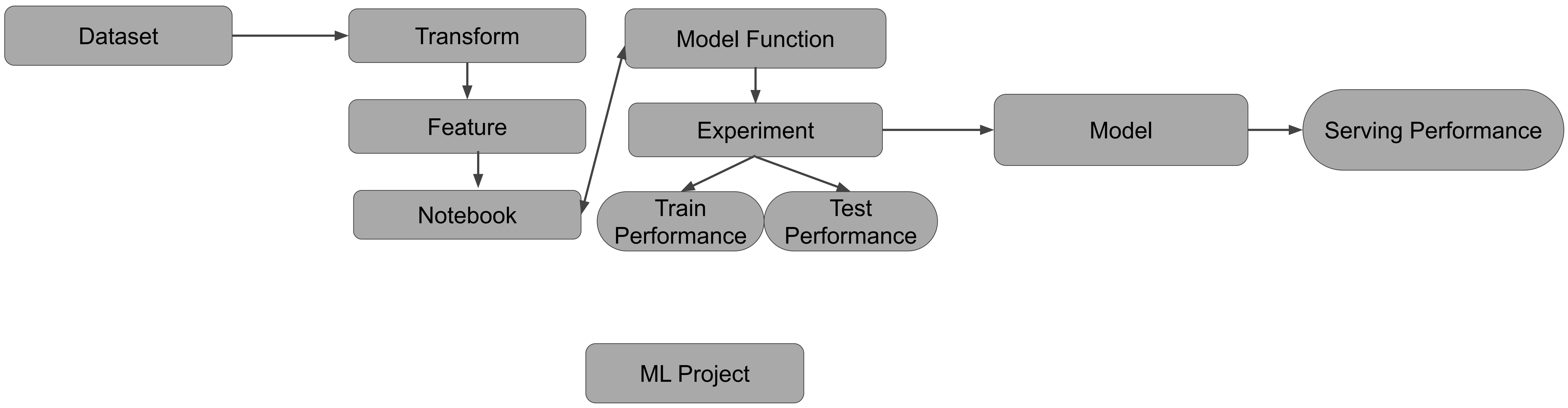
From Data to Metadata....



Metadata?

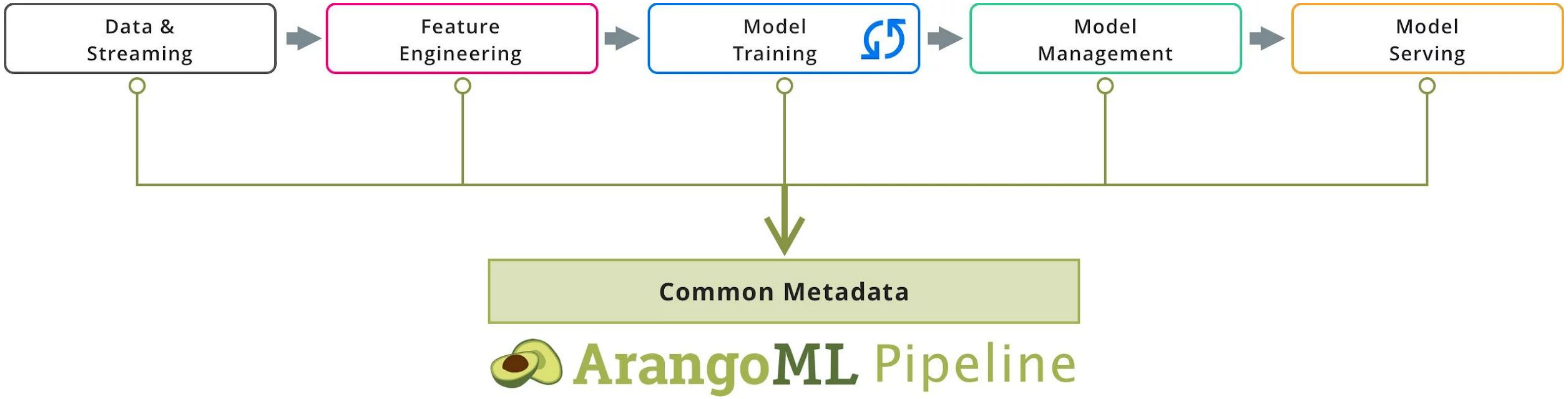
In this context, *metadata* means information about executions (runs), models, datasets, and other artifacts. *Artifacts* are the files and objects that form the inputs and outputs of the components in your ML workflow.

https://www.kubeflow.org/docs/components/misc/metadata/



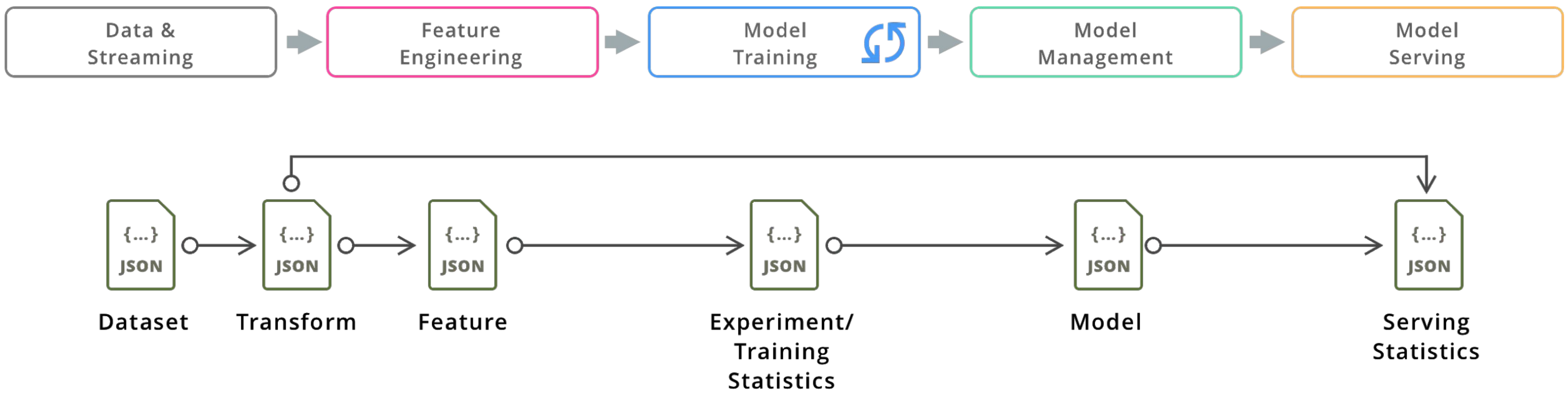
ArangoML Pipeline

"A common extensible metadata layer for ML pipelines which allows Data Scientists and DataOps to manage all information related to their ML pipelines in one place."

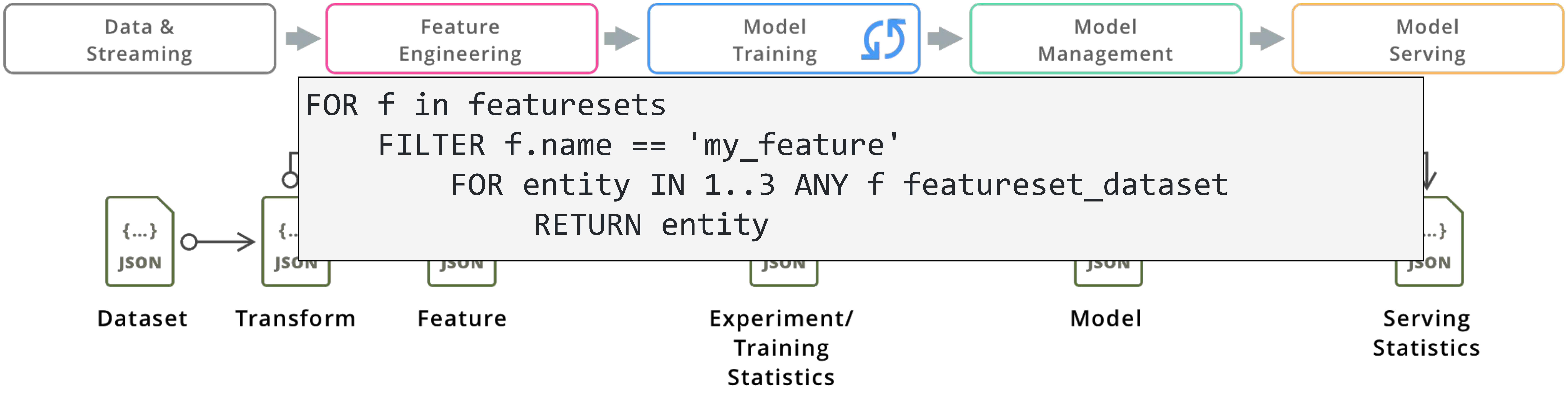


https://www.arangodb.com/2019/09/arangoml-pipeline-common-metadata-layer-machine-learning-pipelines/

Multi-Model Metadata

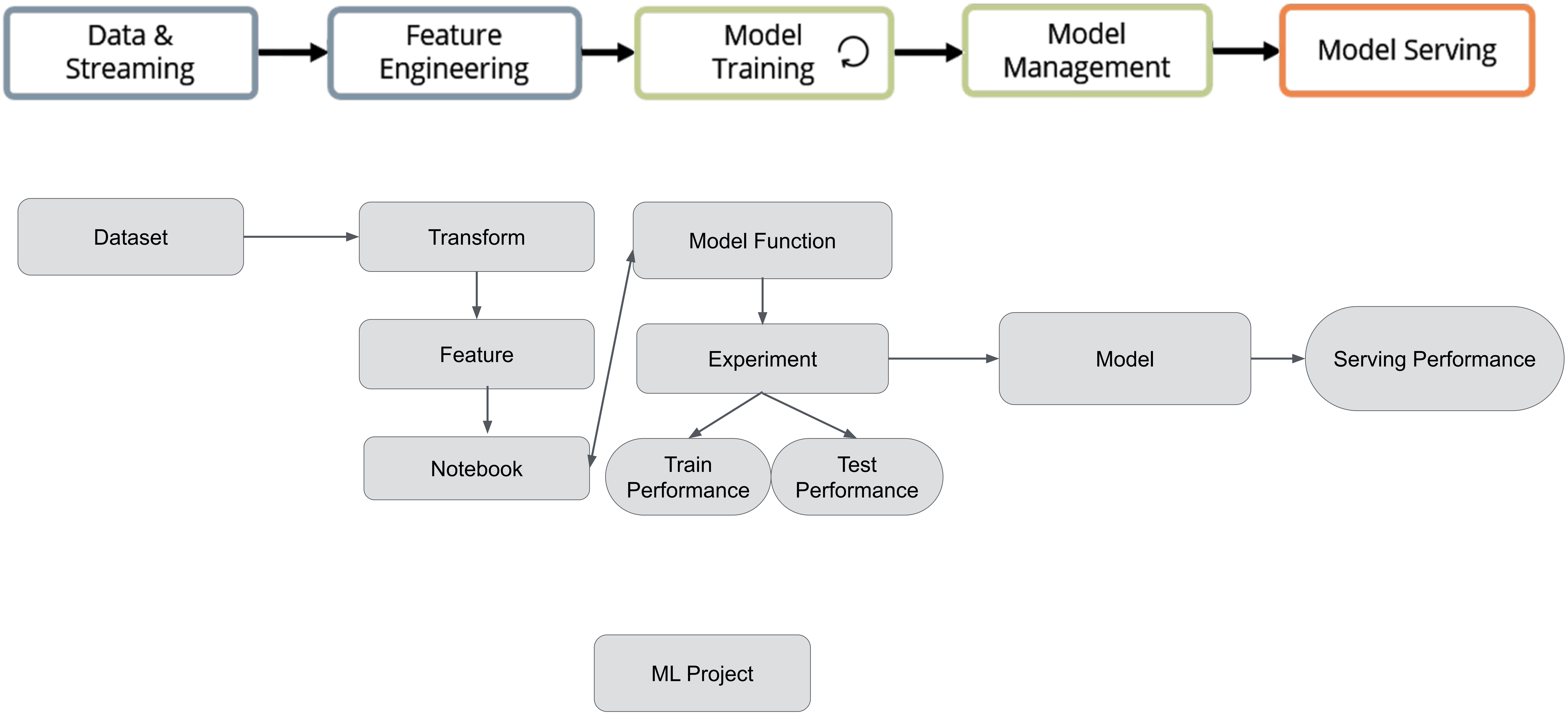


Multi-Model Metadata

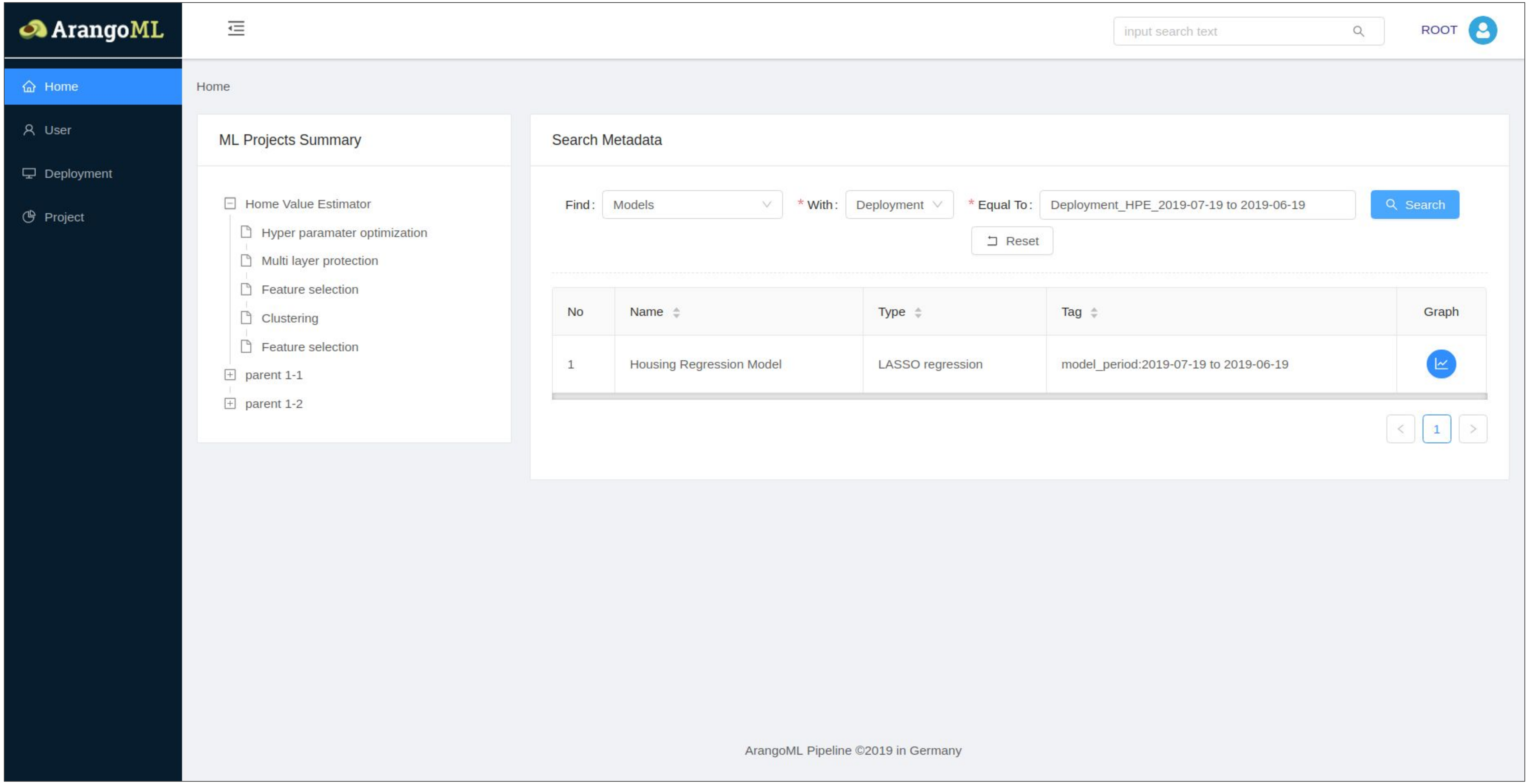



```
FOR f in featuresets
    FILTER f.name == 'my_feature'
        FOR entity IN 1..3 ANY f featureset_dataset
            RETURN entity
```

ArangoML “Schema”

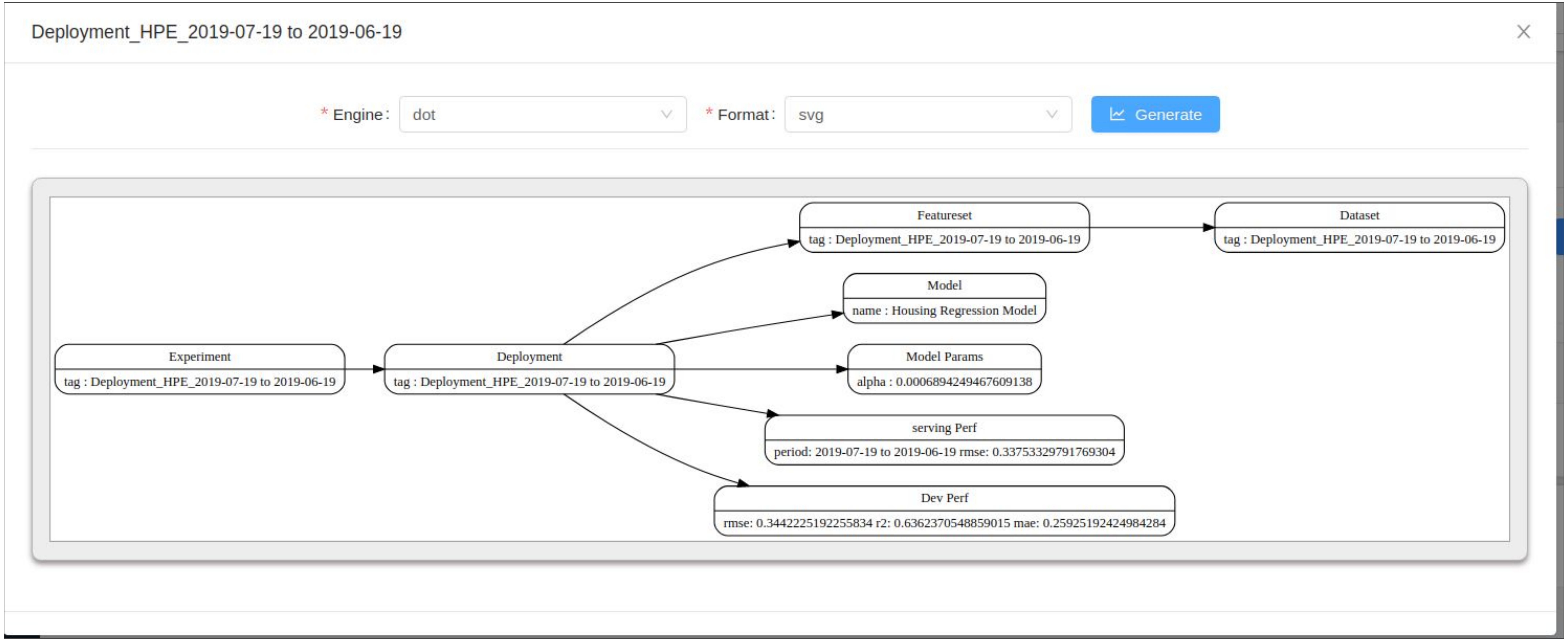


Discover



https://github.com/arangoml/arangopipe

Visualization



https://github.com/arangoml/arangopipe

ArangoML

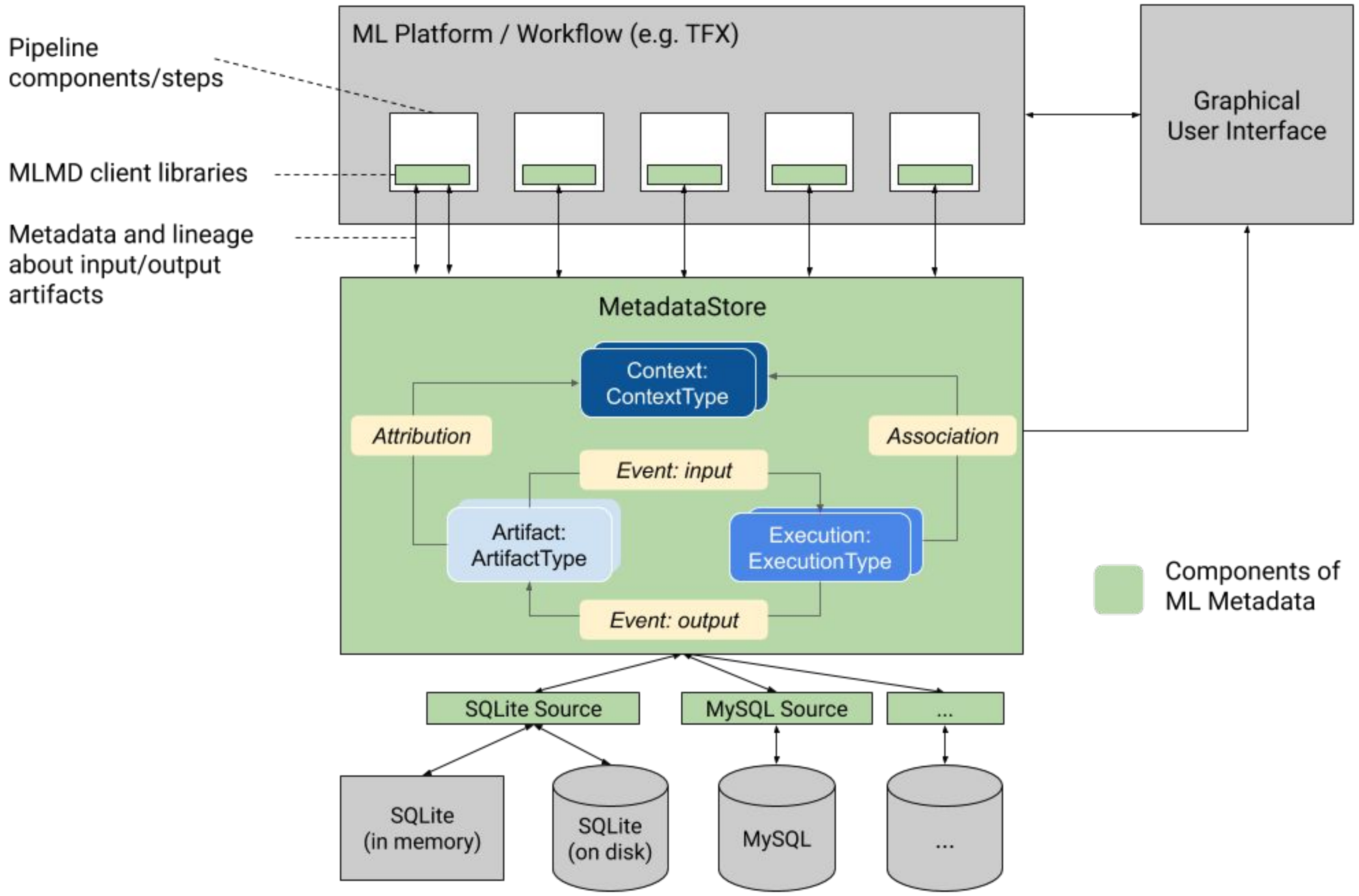
- Python package
- HTTP API
- TFX Integration [coming shortly]

```
from arangopipe.arangopipe_api import ArangoPipe

ap = ArangoPipe(conn_config)
model_info = {"name": "hyper-param-optimization",  "type": "hyper-opt-experiment"}
model_reg = ap.register_model(model_info, project = "Housing_Price_Estimation_Project")
```

https://github.com/arangoml/arangopipe

TFX MLMD



https://www.tensorflow.org/tfx/guide/mlmd

Kubeflow Metadata

Kubeflow

What is Kubeflow? Documentation Blog GitHub v0.6

Documentation
- About
- Getting Started
- Use Cases
- Jupyter Notebooks
- Pipelines
- Fairing
- Kubeflow on AWS
- Kubeflow on Azure
- Kubeflow on GCP
- Components of Kubeflow

Documentation / Components of Kubeflow / Miscellaneous / Metadata

Metadata

Tracking and managing metadata of machine learning workflows in Kubeflow

The goal of the Metadata project is to help Kubeflow users understand and manage their machine learning (ML) workflows by tracking and managing the metadata that the workflows produce.

In this context, *metadata* means information about executions (runs), models, datasets, and other artifacts. *Artifacts* are the files and objects that form the inputs and outputs of the components in your ML workflow.

Alpha version

https://www.kubeflow.org/docs/components/misc/metadata/

Thanks for listening!

ArangoML

- @arangoml
- https://github.com/arangoml/arangopipe
- Demo

  https://github.com/arangoml/arangopipe/blob/master/arangopipe/arangopipe_examples.ipynb

ArangoDB

- @arangodb
- https://www.arangodb.com/
- Demo

  https://github.com/arangoml/knowlegegraph-demo

Building Adaptive Knowledge Graphs

Graphs vs Machine Learning