FEATURE STORE SUMMIT 2024

DATA FOR AI:
REAL-TIME, BATCH, AND LLMS

HOPSWORKS

Introducing The AI Lakehouse

Raymond Cunningham, Ph.D.
Engineering Director
Hopsworks
ray.cunningham@hopsworks.ai

Organized by HOPSWORKS

HOPSWORKS

The Story of the

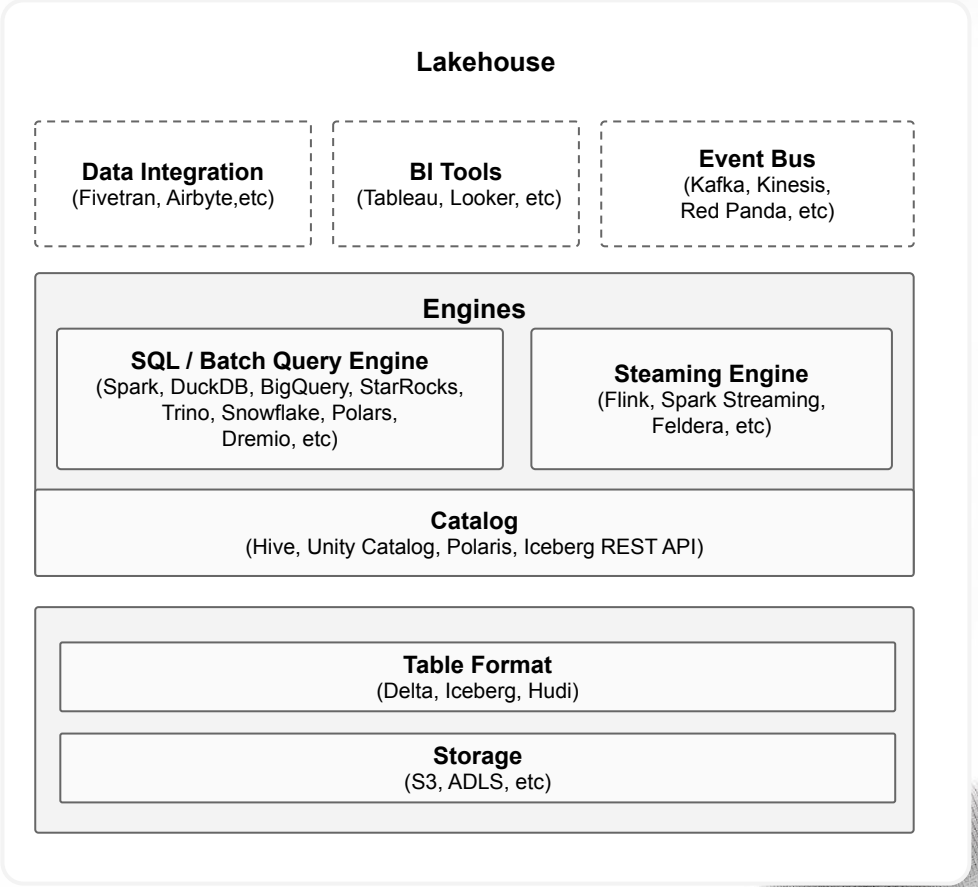
AI Lakehouse

is the Story of the

Offline Feature Store

HOPSWORKS

Just as the cloud revolutionized Enterprise computing by separating storage and compute, **the Lakehouse is revolutionizing Enterprise data** by separating data from its query engines.



Hopsworks Offline Store

Hudi on Hops
Initial support for Hudi on Hopsworks

Offline Store
Apache Hive Only

Offline Store
Apache Hudi becomes default Offline Store

Pluggable Offline Store
Choose Data Store for External Feature Groups

Offline Store is the Lakehouse
Hudi, Delta, (Iceberg)

Lakehouse

Delta Lake
early release by Databricks.

Iceberg spec and incubation as Apache Project

CIDR paper by Databricks introducing the Lakehouse Architecture.

Dremio, Trino, Presto add support for Iceberg

Growth of Iceberg

Snowflake support for Iceberg

AWS Glue adds Iceberg support (Hudi supported since 2020)

Lakehouse becomes primary architecture for analytical workloads

Databricks supports Iceberg (Tabular acq)

Unity Catalog and **Polaris Catalog** open sourced

The Offline Feature Store is now a Lakehouse*

***The offline store is open,** so it cannot be a *feature platform* that limits the compute engines you use.

***The offline store is a Lakehouse,** so the feature store is not *virtual*.

HOPSWORKS

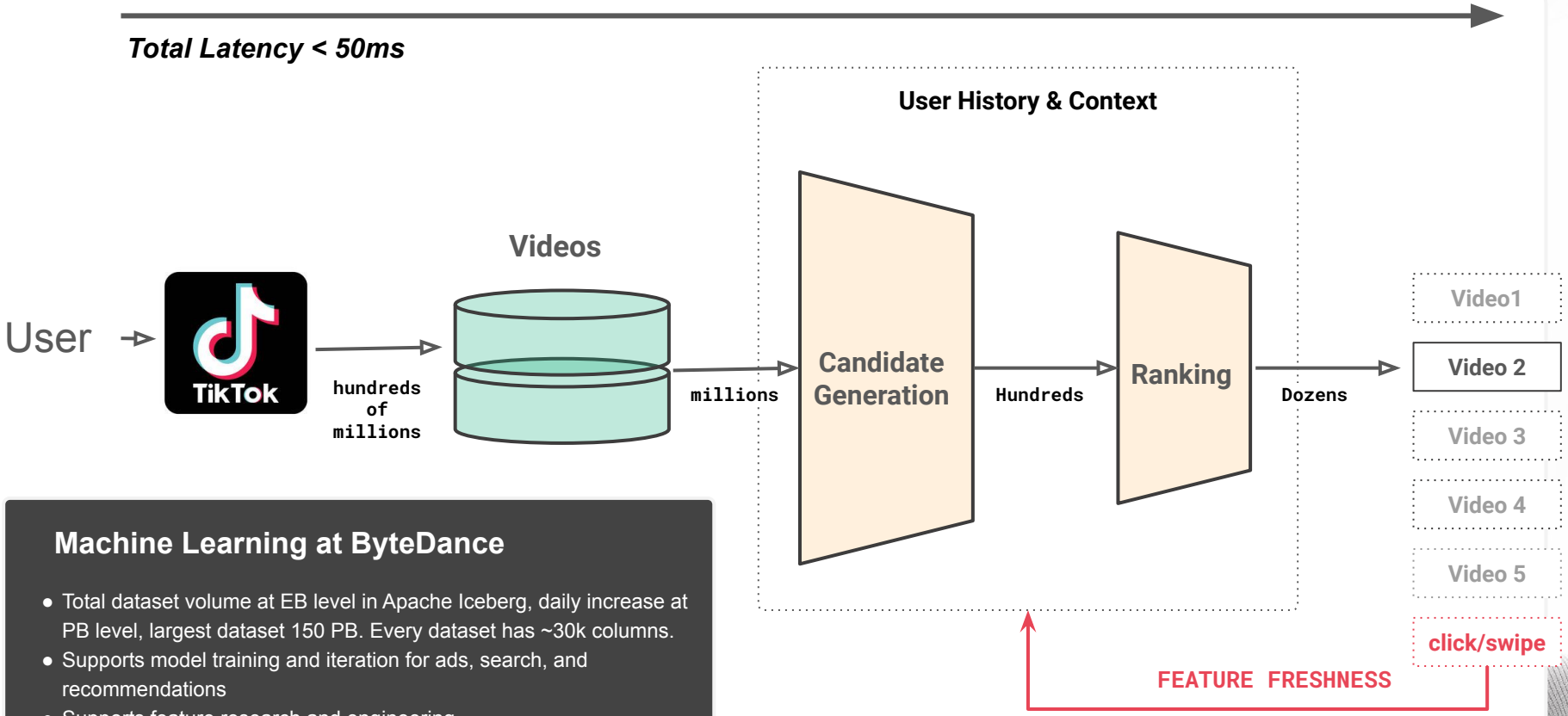
The Online Store and Vector Index extend the Lakehouse to become the AI Lakehouse

HOPSWORKS

The tale of

2 AI Lakehouse Architectures

HOPSWORKS // TikTok's AI Recommender needs Fresh Features

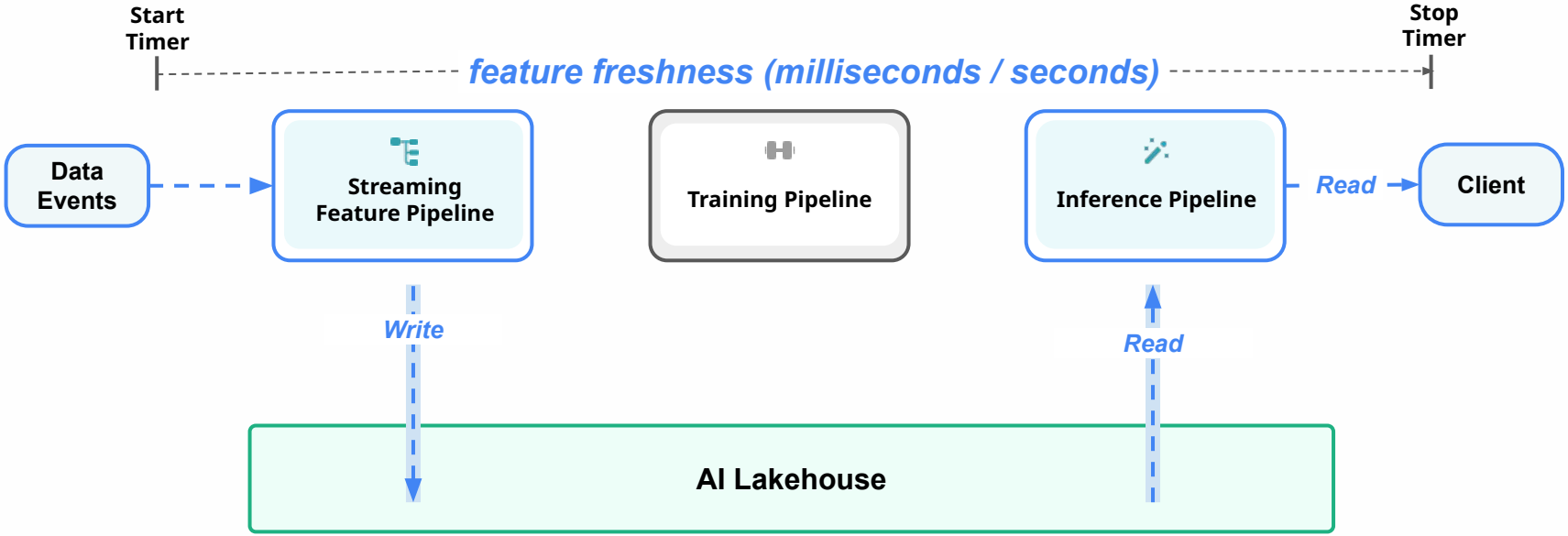


Machine Learning at ByteDance

- Total dataset volume at EB level in Apache Iceberg, daily increase at PB level, largest dataset 150 PB. Every dataset has ~30k columns.
- Supports model training and iteration for ads, search, and recommendations
- Supports feature research and engineering.

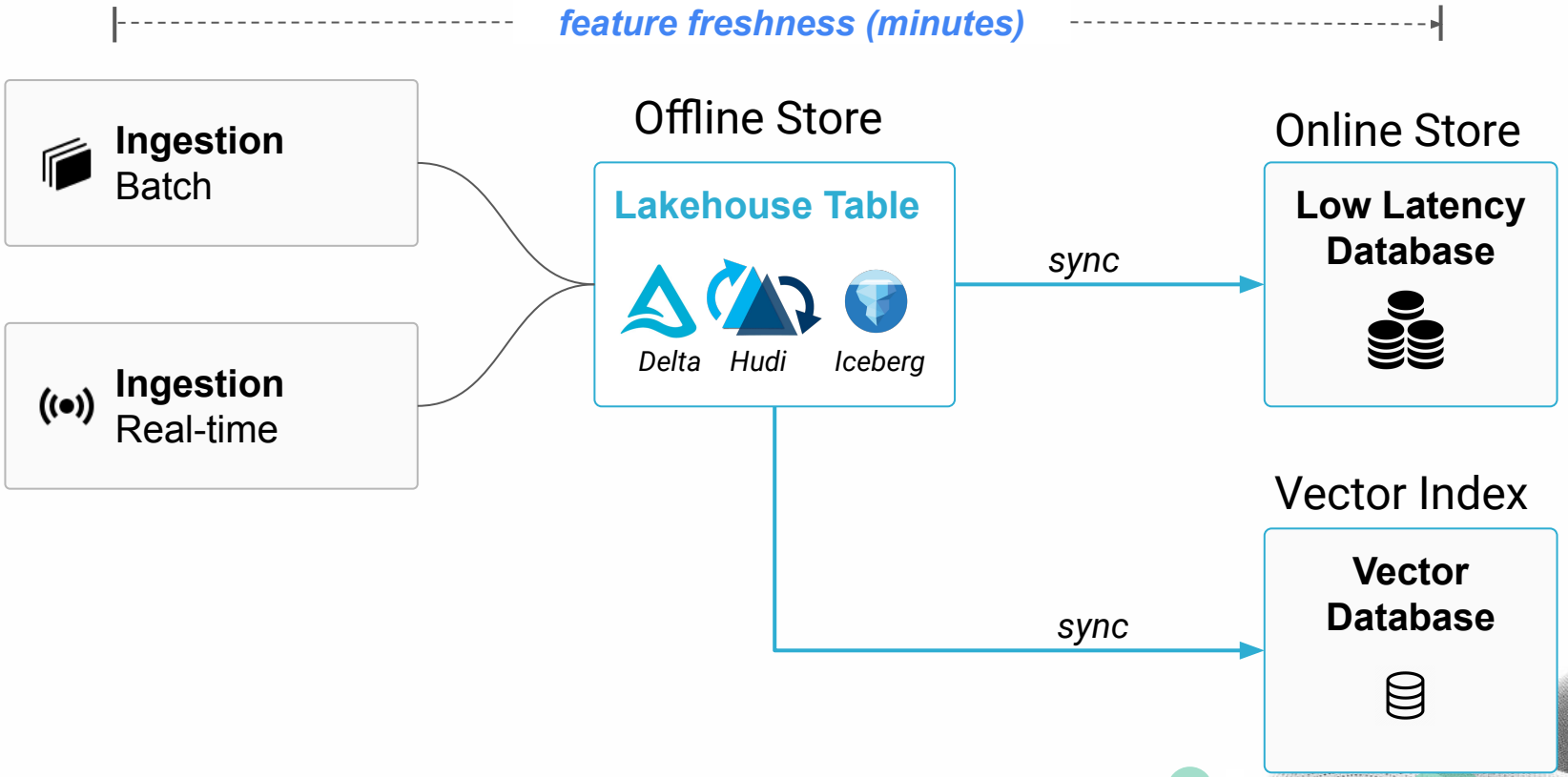
[Exabytes (billions of GBs) of data in ByteDance Iceberg-powered Lakehouse]

Minimize Feature Freshness for Real-Time AI Systems



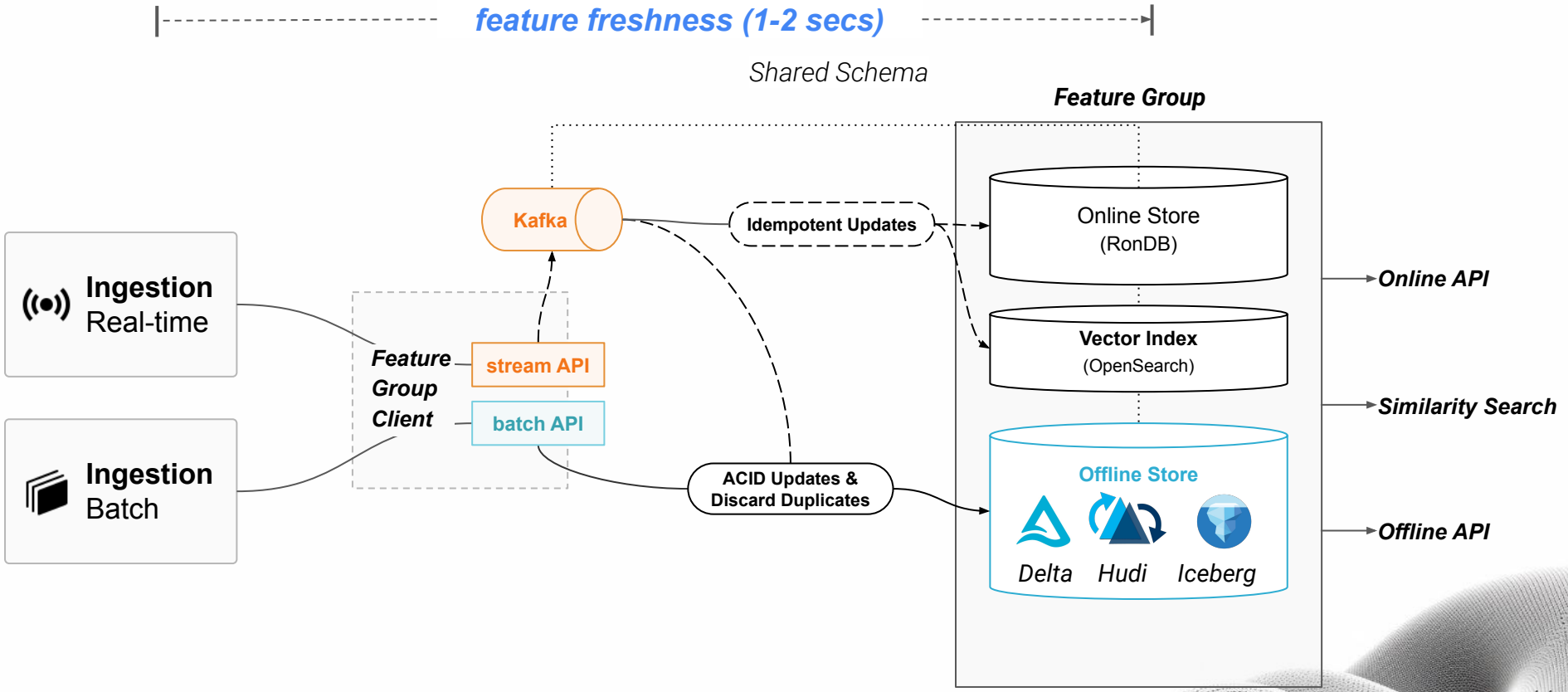
HOPSWORKS The Non Real-Time AI Lakehouse

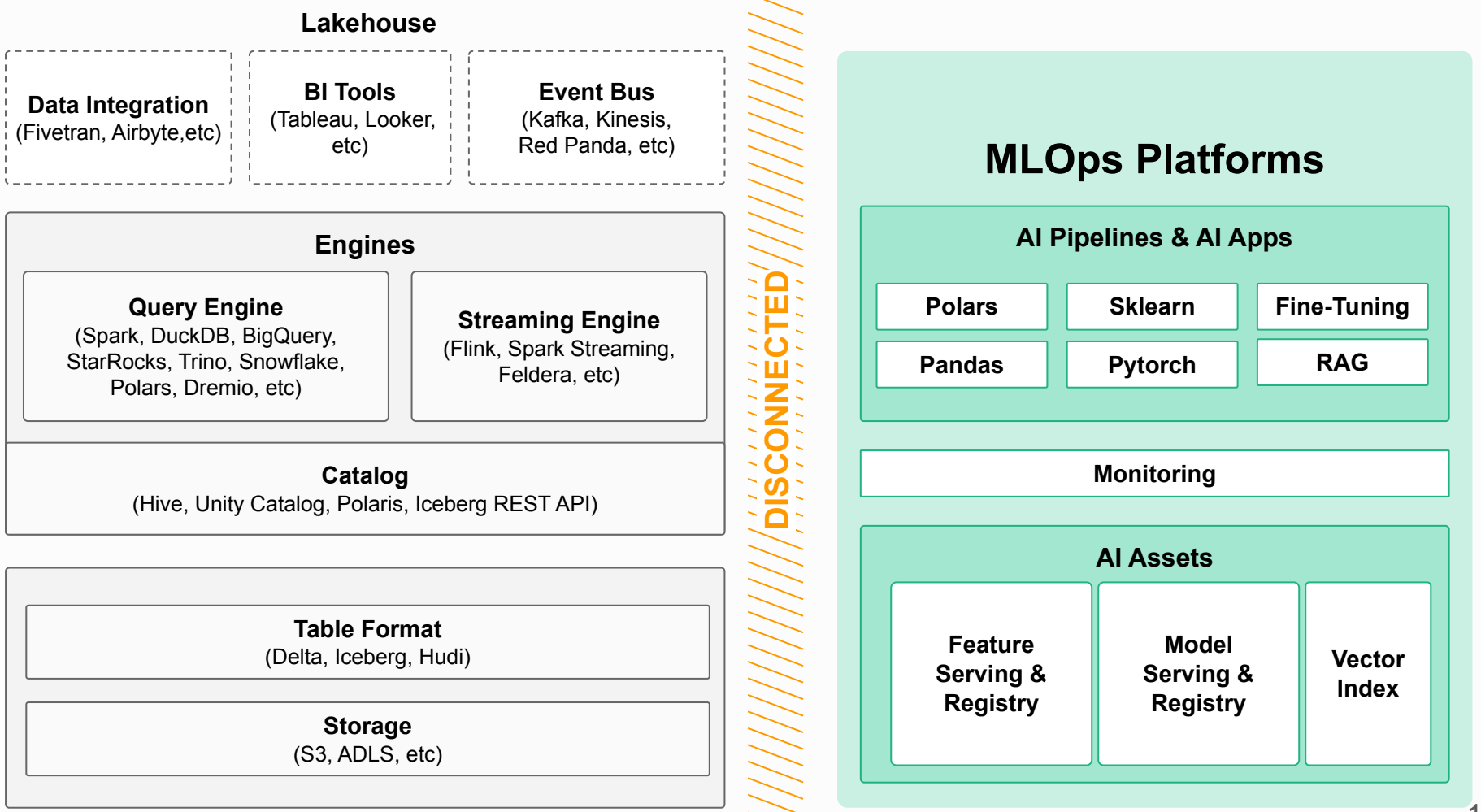
Extend Lakehouse Tables with Real-Time Access and Vector Search



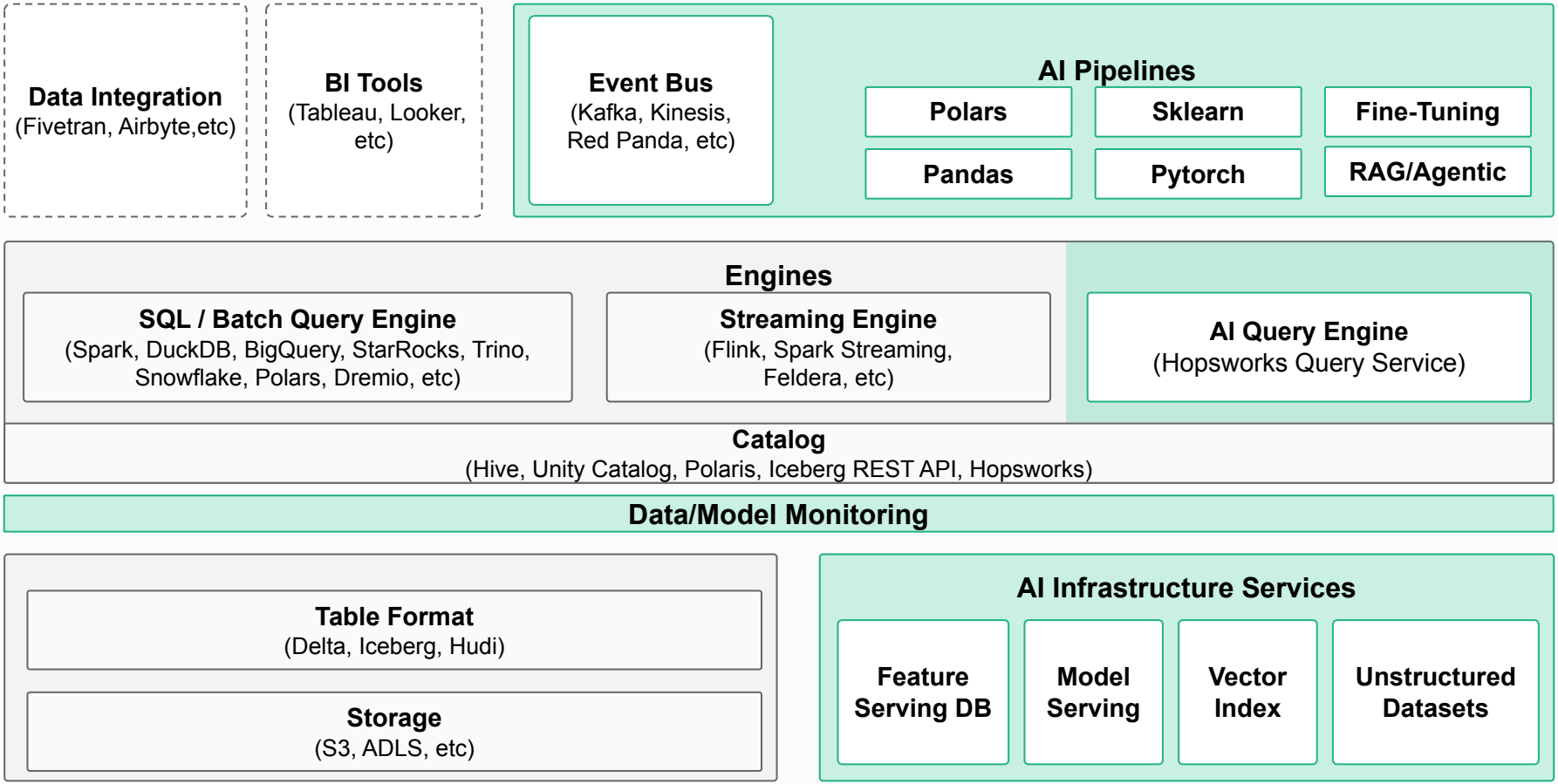
HOPSWORKS Hopsworks: A Real-time AI Lakehouse

Fresh Features should stream updates to the Online Store

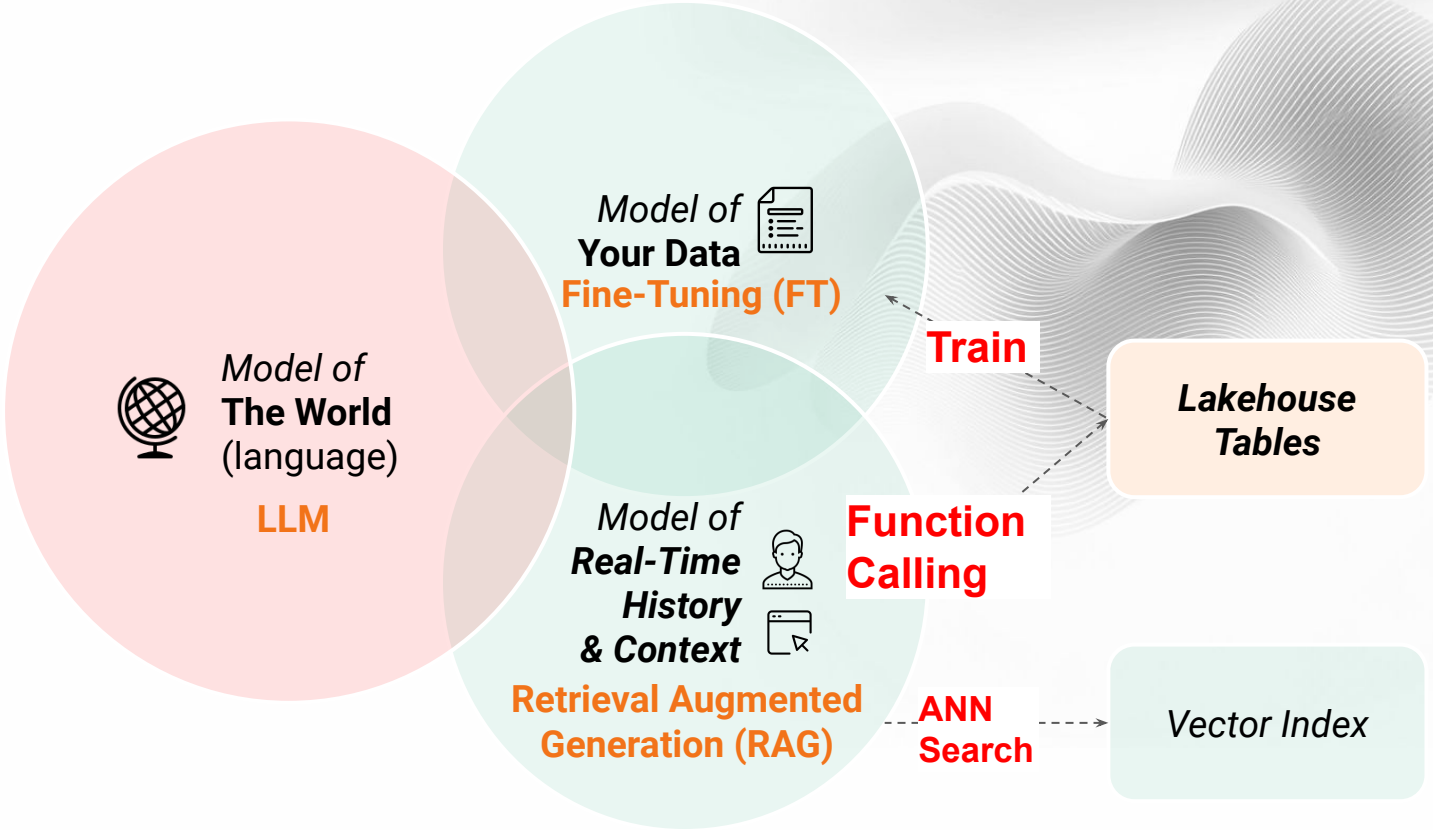



Lakehouse
Data Integration
(Fivetran, Airbyte,etc)
BI Tools
(Tableau, Looker, etc)
Event Bus
(Kafka, Kinesis, Red Panda, etc)
Engines
Query Engine
(Spark, DuckDB, BigQuery, StarRocks, Trino, Snowflake, Polars, Dremio, etc)
Streaming Engine
(Flink, Spark Streaming, Feldera, etc)
Catalog
(Hive, Unity Catalog, Polaris, Iceberg REST API)
Table Format
(Delta, Iceberg, Hudi)
Storage
(S3, ADLS, etc)
DISCONNECTED
MLOps Platforms
AI Pipelines & AI Apps
Polars
Sklearn
Fine-Tuning
Pandas
Pytorch
RAG
Monitoring
AI Assets
Feature Serving & Registry
Model Serving & Registry
Vector Index

The AI Lakehouse with Hopsworks



[Introducing the AI Lakehouse, Blog Post]



Function Calling with LLMs and Lakehouse Tables, PyData London'24

HOPSWORKS

- Join our slack community
- Explore our latest tutorials
- Ask us any questions

public-hopsworks.slack.com