From building to using feature stores for ML systems

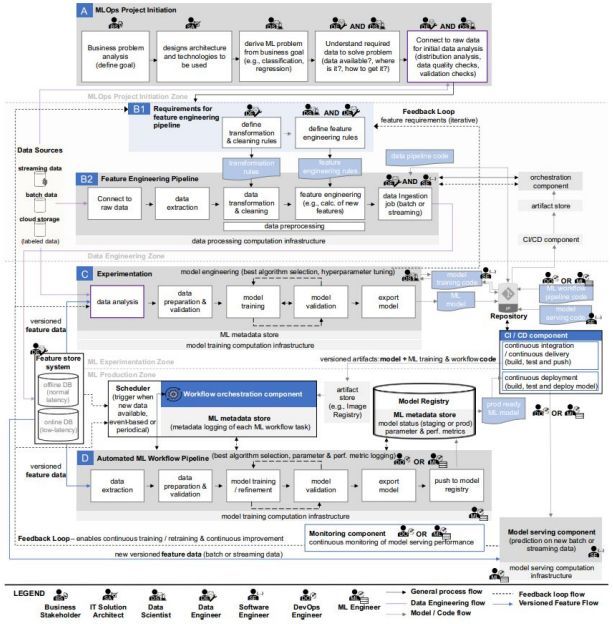
Fabio Buso
Head of Engineering

Hopsworks

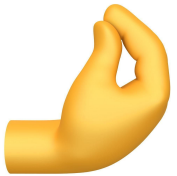
FEATURE STORE SUMMIT 2023

Organized by HOPSWORKS

MLOps has turned into a spaghetti monster



=



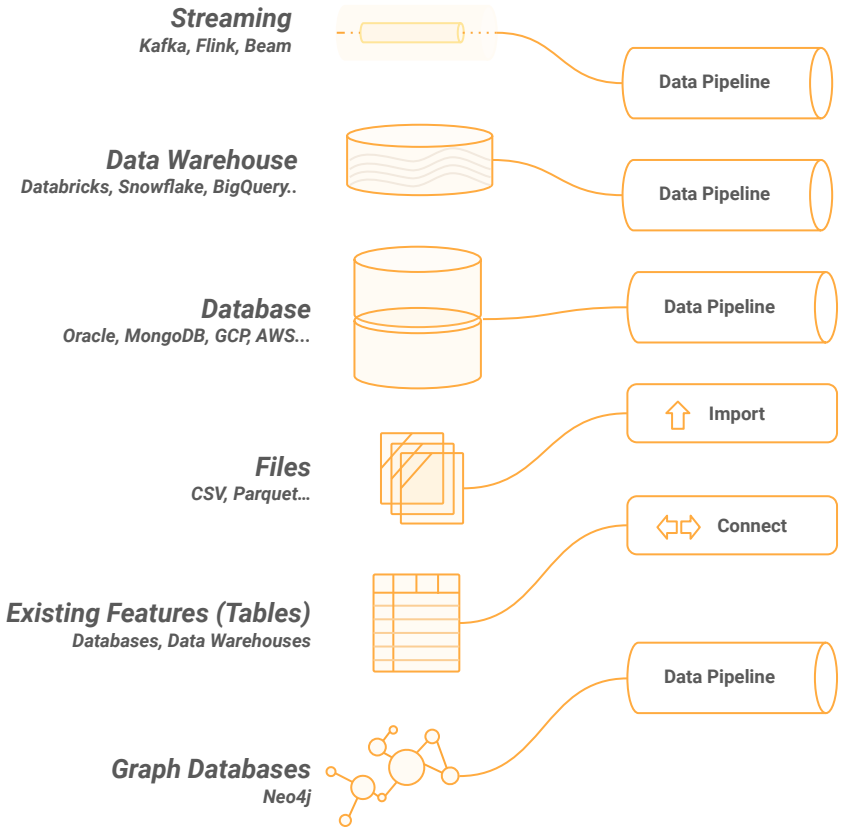
HOPSWORKS

Img from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10081336

Deploying models to production is a data challenge

Data challenges

1. **Variety of data sources**
2. Need for a variety of frameworks
3. Disconnect between experimentation / training / production
4. Custom one-off pipelines to make data available in real time

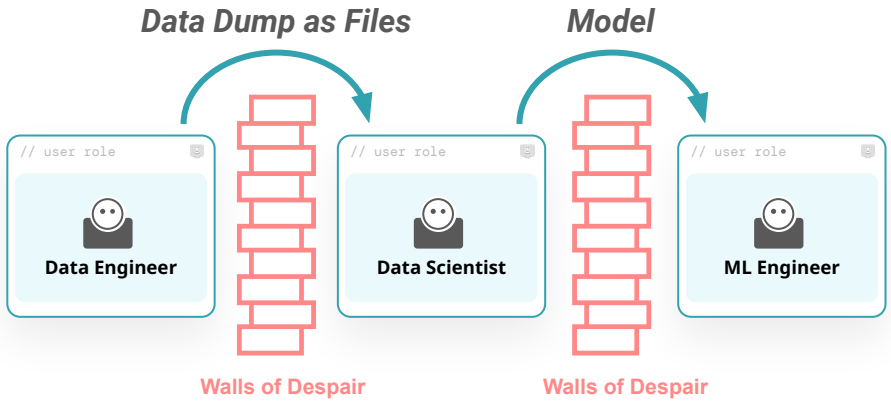


Data challenges

1. **Variety of data sources**
2. **Need for a variety of frameworks**
3. Disconnect between experimentation / training / production
4. Custom one-off pipelines to make data available in real time

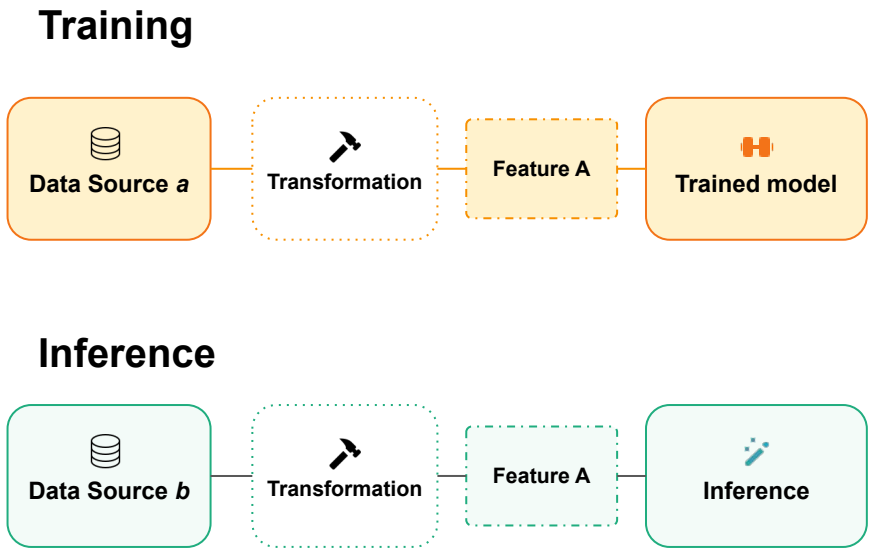
Data challenges

1. **Variety of data sources**
2. **Need for a variety of frameworks**
3. **Disconnect between experimentation / training / production**
4. Custom one-off pipelines to make data available in real time



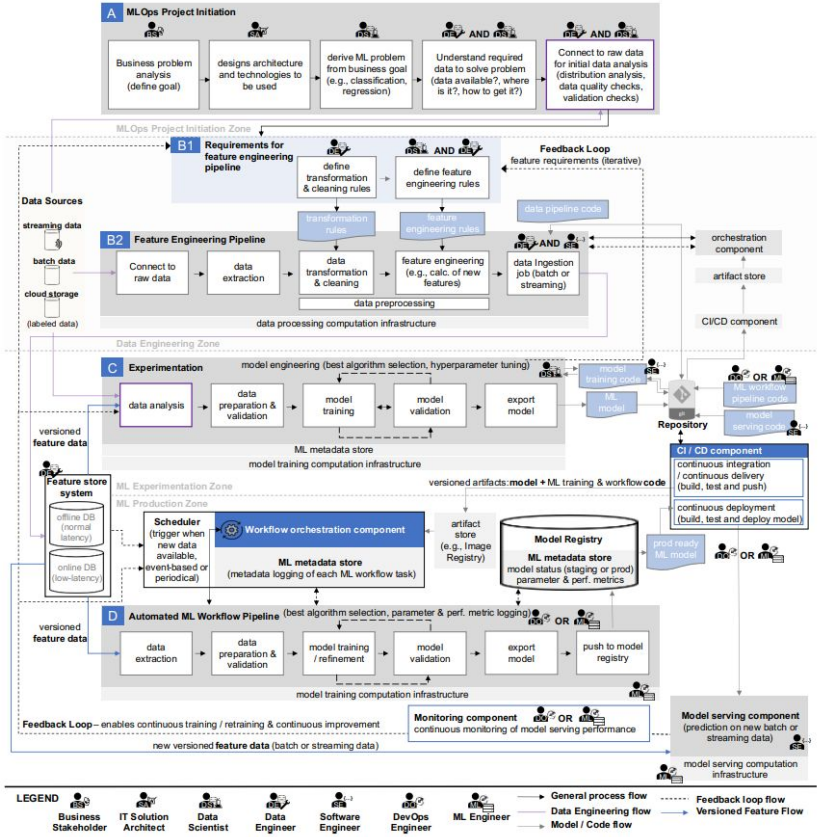
Data challenges

1. Variety of data sources
2. Need for a variety of frameworks
3. Disconnect between experimentation / training / production
4. Custom one-off pipelines to make data available in real time



Feature Store is an answer to these challenges

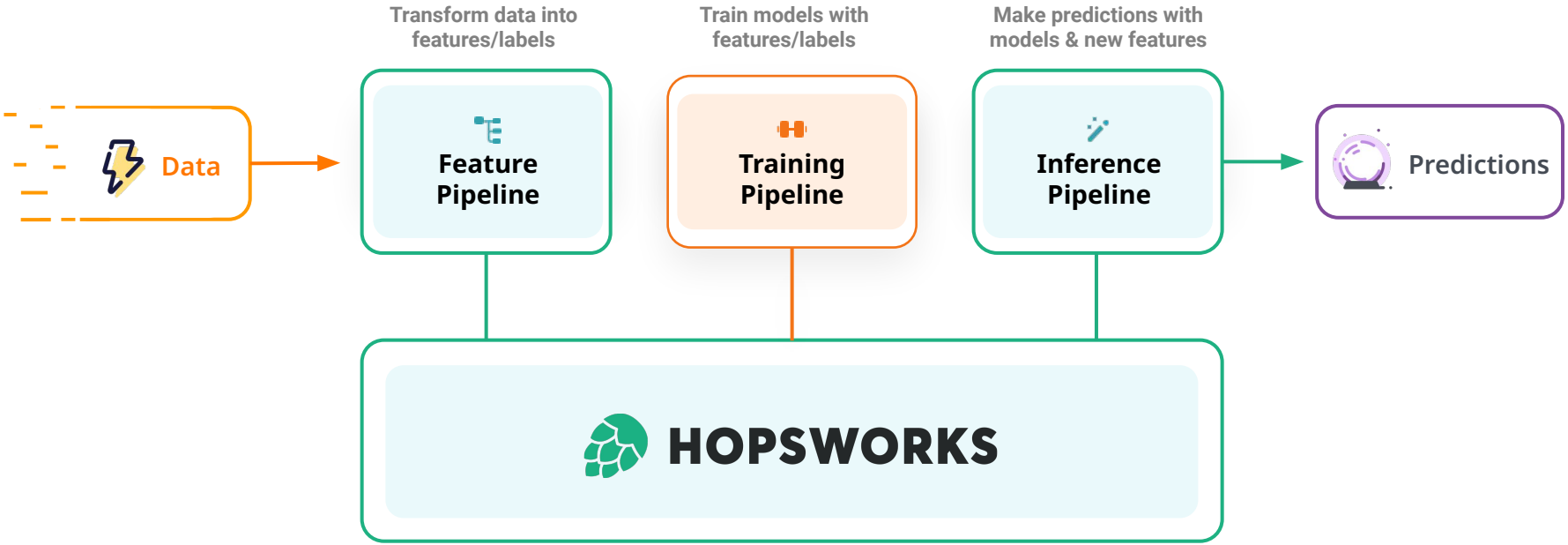
But tools only get you so far



Img from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10081336

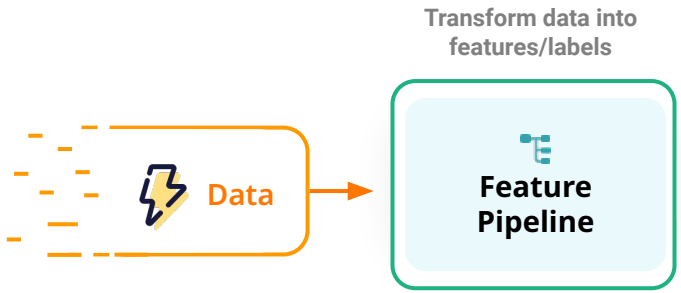
FTI PIPELINES

Make the pipelines independent, but also let them share outputs

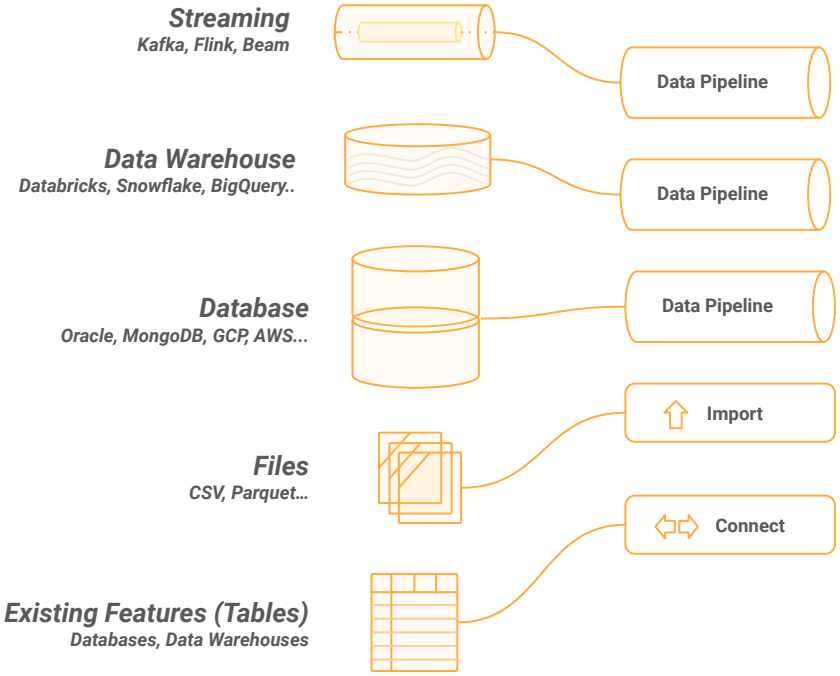


Feature Pipeline

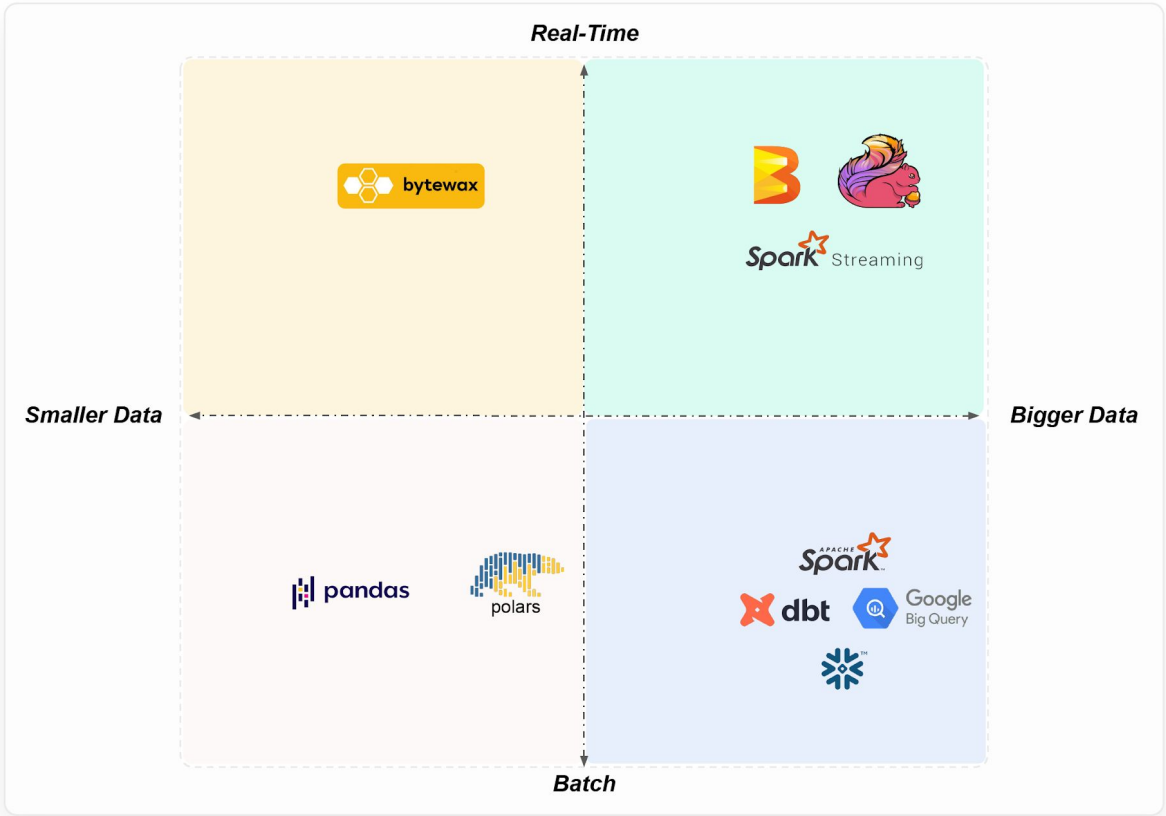
Expectation



Reality



Pick the best framework for your feature pipelines



Python API

df = 

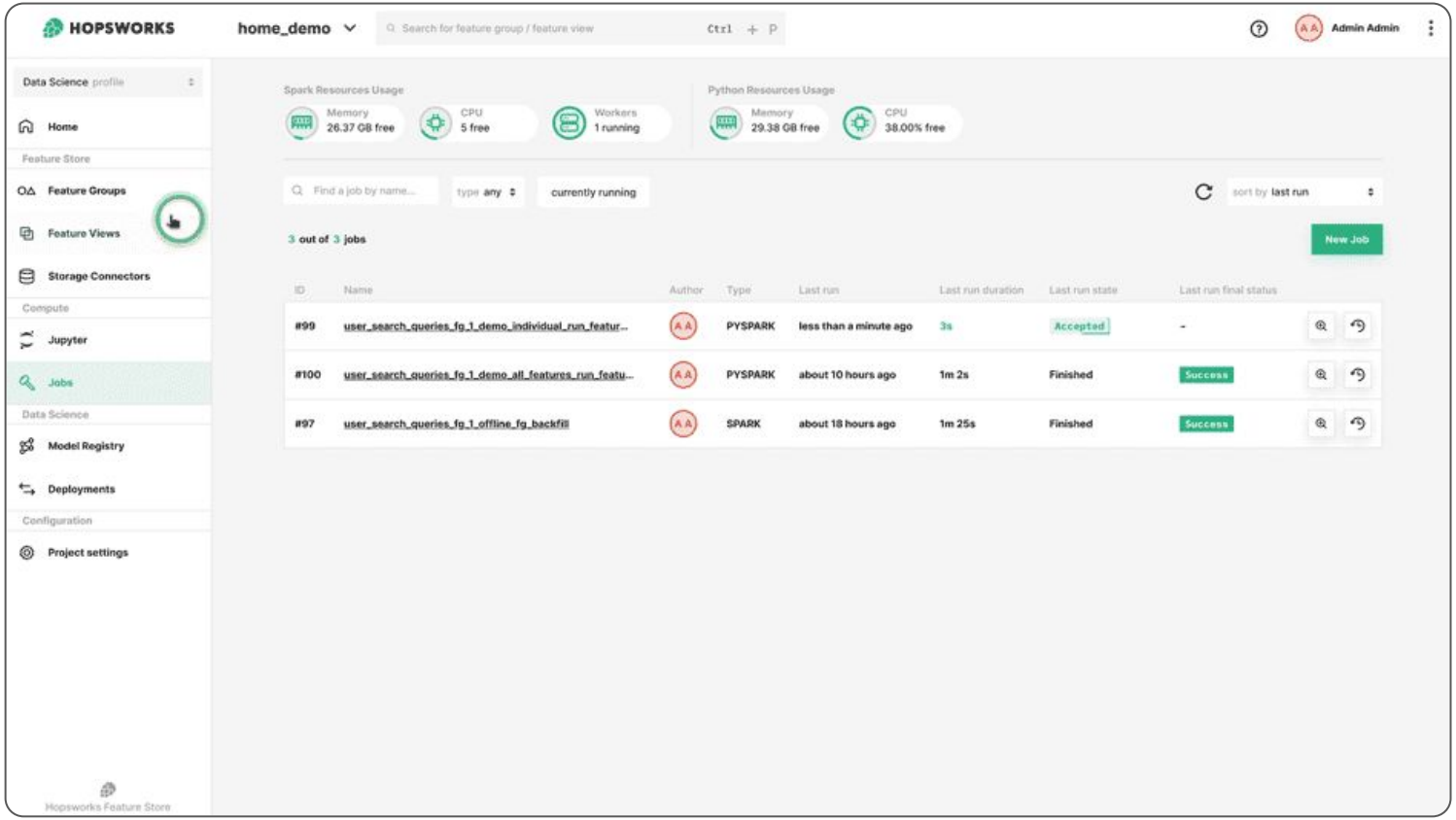

```
fg = fs.get_or_create_feature_group(name="query_terms_yearly",
                                    version=1,
                                    description="Count of search term by year",
                                    primary_key=['year', 'search_term'],
                                    partition_key=['year'],
                                    online_enabled=True,
                                    expectation_suite=expectation_suite
                                    )

fg.insert(df)
```

Link to docs;
https://docs.hopsworks.ai/latest/user_guides/fs/feature_group/create

Feature monitoring (NEW)

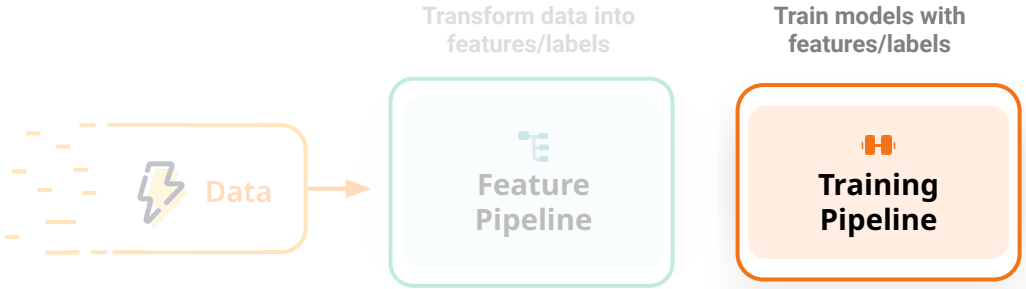
HOPSWORKS



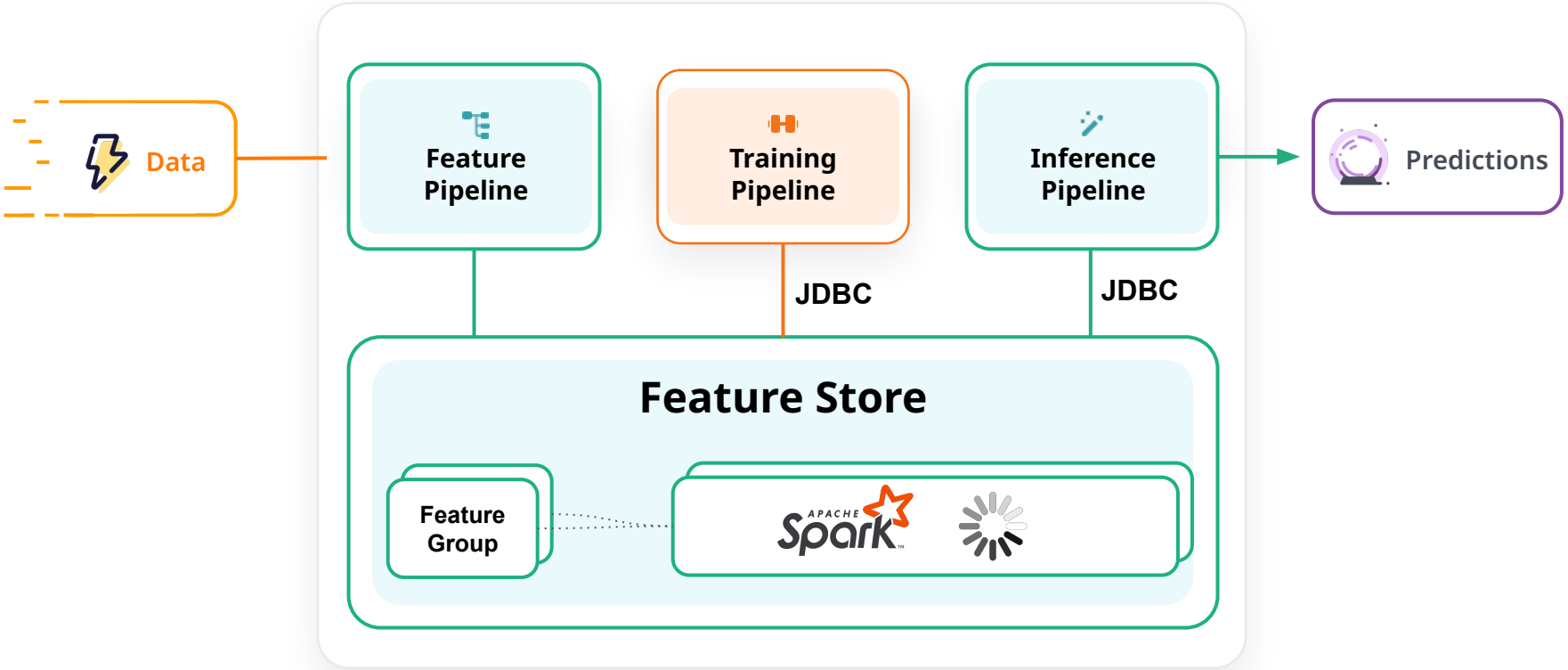
Feature monitoring (NEW)

```
user_search_queries_fg.enable_feature_monitoring(
    name="brand_name_monitoring"
    feature_name="has_brand_name",
    job_frequency="DAILY",
).with_detection_window(
    time_offset="1d",
    row_percentage=0.1,
).with_reference_window(
    specific_value=td.statistics.feature.mean
).compare_on(
    metric="mean",
    relative=True,
    threshold=0.5,
).save()
```

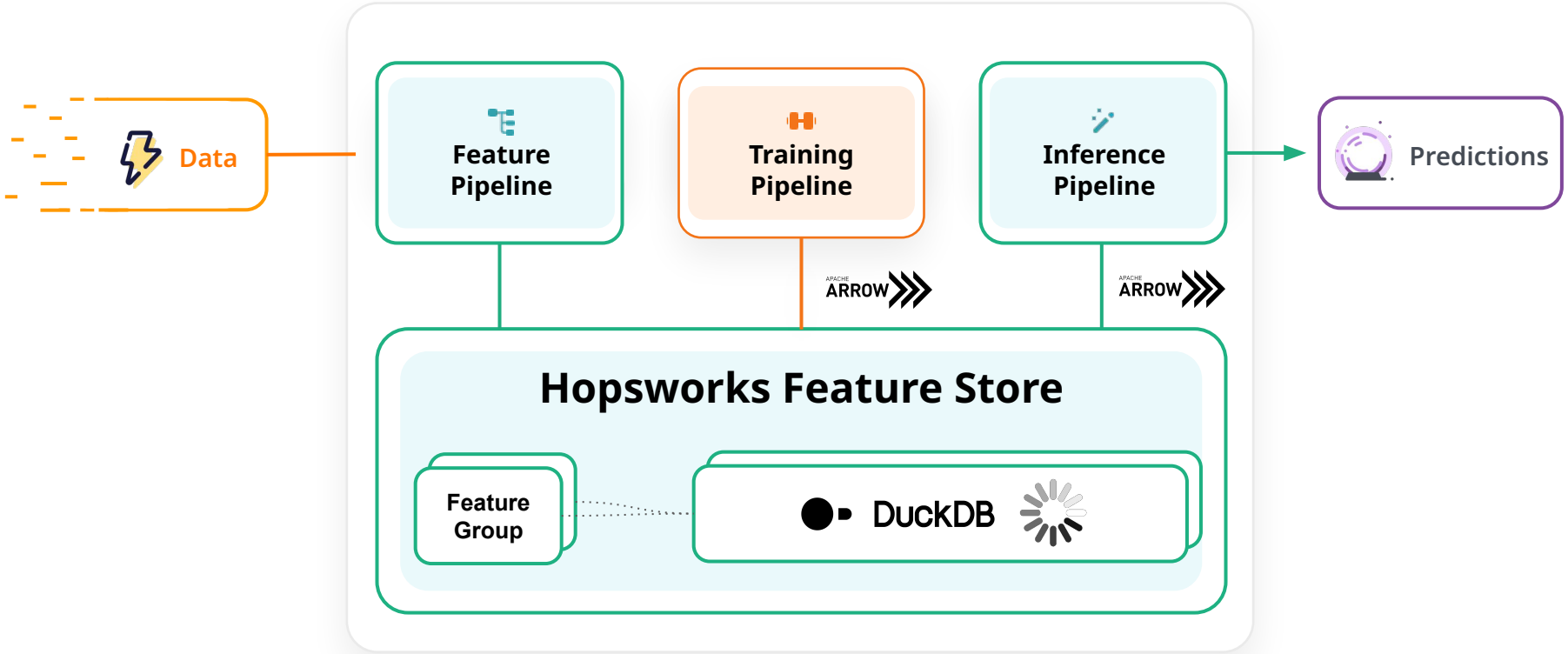
Training Pipeline

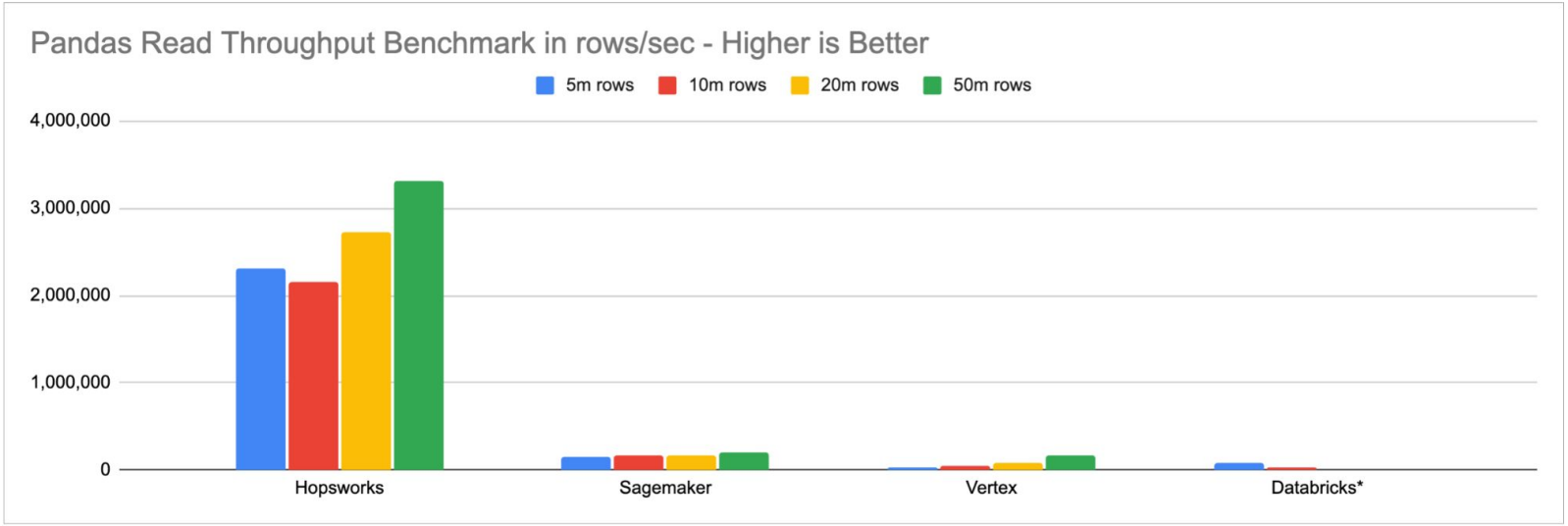

Transform data into features/labels
Train models with features/labels
Data
Feature Pipeline
Training Pipeline

Need for speed

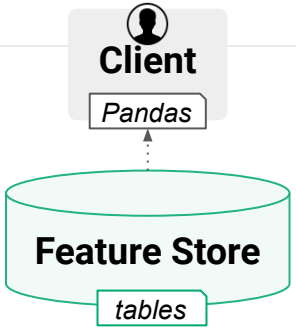


Need for speed (Hopsworks 3.3)





***Databricks failed at 20m, 50m rows**

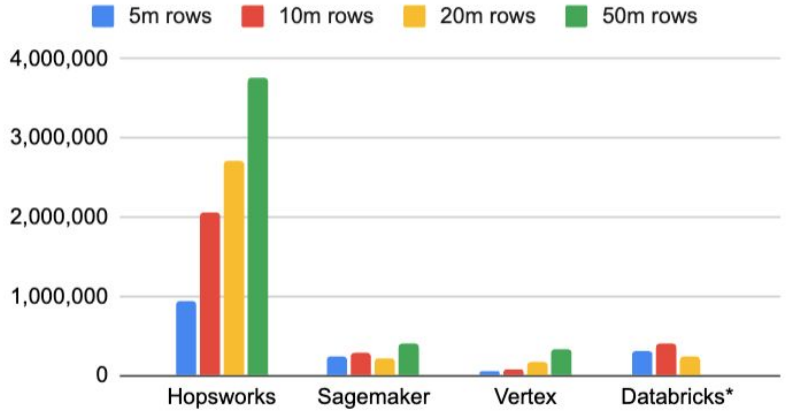


Pandas Read (rows/secs)	5m rows	10m rows	20m rows	50m rows
Hopsworks	2,314,815	2,155,172	2,724,796	3,313,453
Sagemaker	155,328	170,358	167,364	202,053
Vertex	38,011	54,672	77,289	172,247
Databricks*	85,807	27,666	*	*

HOPSWORKS

www.featurestore.org/benchmarks

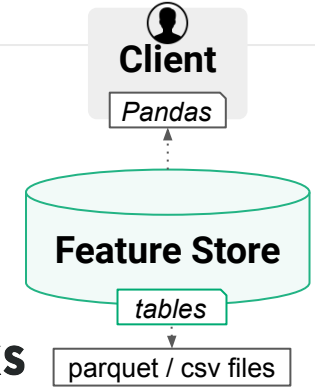
Training Data Parquet Write in rows/sec - Higher is Better



Parquet Write (rows/secs) . *Databricks failed at 50m rows.

Parquet Write (rows/secs)	5m rows	10m rows	20m rows	50m rows
Hopsworks	952,381	2,057,613	2,724,796	3,770,739
Sagemaker	243,427	280,505	223,389	395,163
Vertex	54,831	87,665	161,186	332,094
Databricks*	308,642	403,714	237,897	*

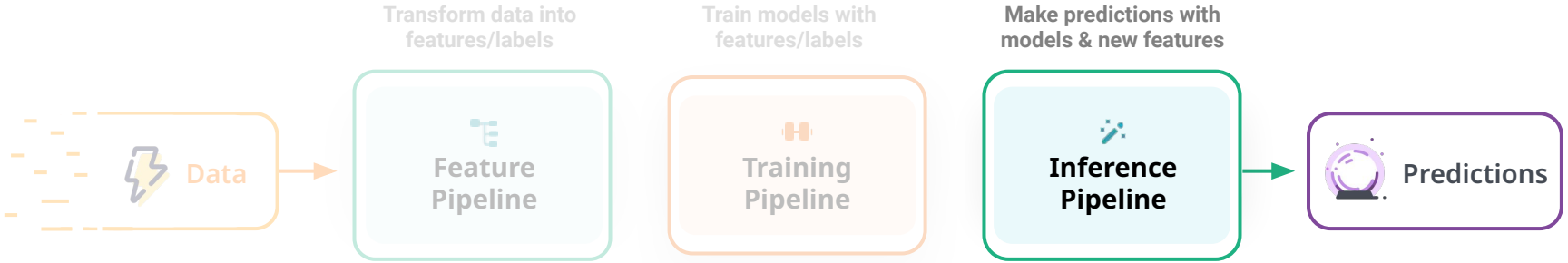
***Databricks failed at 50m rows**

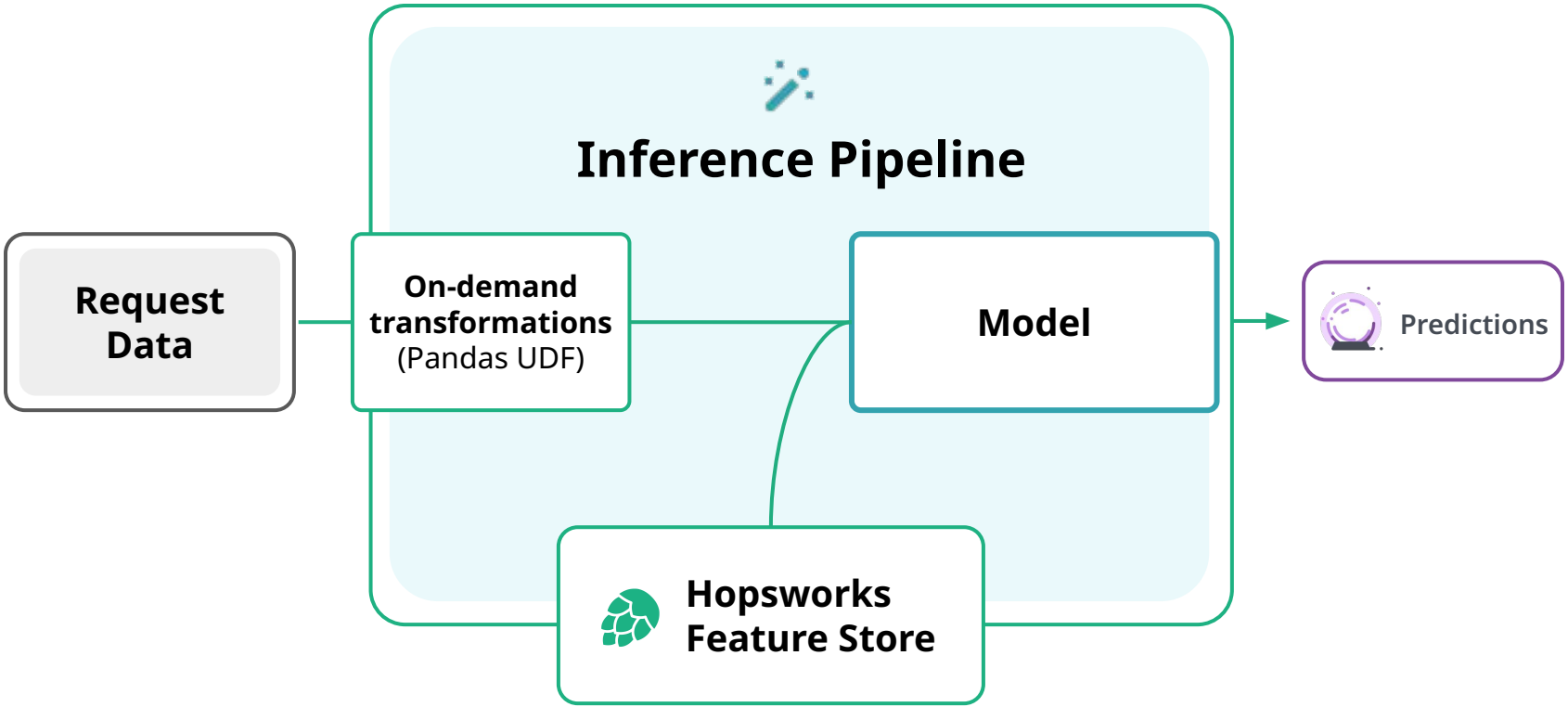


11 times faster than Databricks (20m rows)
11 times faster than Vertex (50m rows)
9.5 times faster than SageMaker (50m rows)

www.featurestore.org/benchmarks

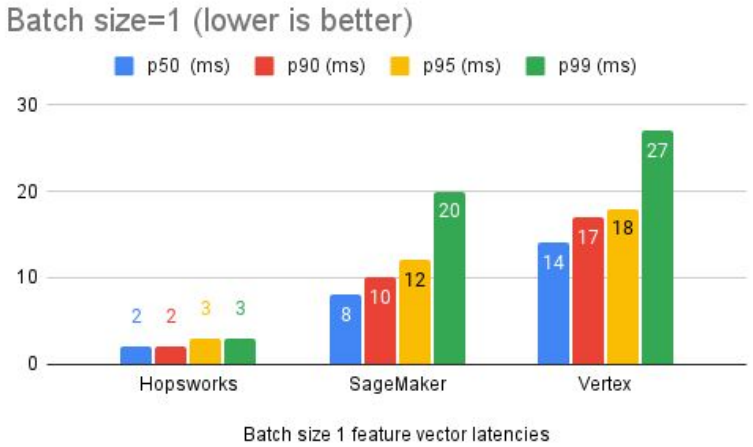
Inference Pipeline

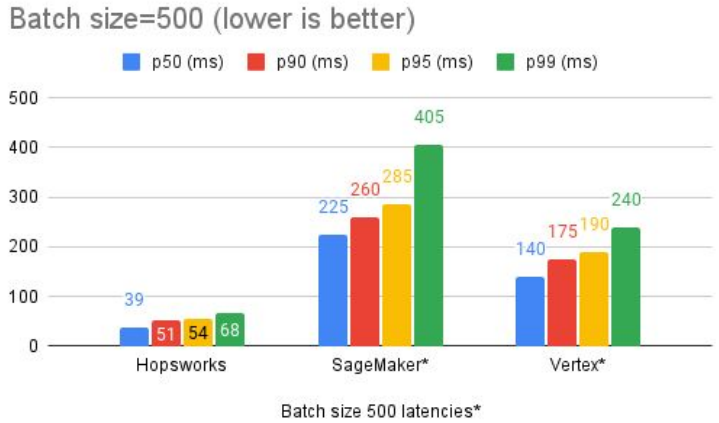

Transform data into features/labels
Train models with features/labels
Make predictions with models & new features
Data
Feature Pipeline
Training Pipeline
Inference Pipeline
Predictions



Use **Pandas UDFs** to keep feature functions consistent **(and performant)** between feature and inference pipeline

Online Feature Store Benchmark Results





*SageMaker and Vertex have a batch size limit of 100 records per request. Therefore for testing batch size 500 we ran 5 sequential requests, each with a batch size of 100.
In Vertex, we did not deserialize the returned features, so it's numbers should be slightly higher

www.featurestore.org/benchmarks

FTI Benefits

Data challenges

1. Variety of data sources
2. Need for a variety of frameworks
3. Disconnect between experimentation / training / production
4. Custom one-off pipelines to make data available in real time

FTI benefits

1. Source Agnostic
2. Framework agnostic
3. Unified architecture for experimentation / training / production
4. Unified architecture for batch and real time

Data challenges → FTI benefits

1. **Variety of data sources** → **Source Agnostic**
2. Need for a variety of frameworks → Framework agnostic
3. Disconnect between experimentation / training / production → Unified architecture for experimentation / training / production
4. Custom one-off pipelines to make data available in real time → Unified architecture for batch and real time

Data challenges → FTI benefits

1. **Variety of data sources**	→	1. **Source Agnostic**
2. **Need for a variety of frameworks**	→	2. **Framework agnostic**
3. Disconnect between experimentation / training / production		3. Unified architecture for experimentation / training / production
4. Custom one-off pipelines to make data available in real time		4. Unified architecture for batch and real time

Data challenges		FTI benefits
1. Variety of data sources	→	1. Source Agnostic
2. Need for a variety of frameworks	→	2. Framework agnostic
3. Disconnect between experimentation / training / production	→	3. Unified architecture for experimentation / training / production
4. Custom one-off pipelines to make data available in real time		4. Unified architecture for batch and real time

Data challenges		FTI benefits
1. Variety of data sources	→	1. Source Agnostic
2. Need for a variety of frameworks	→	2. Framework agnostic
3. Disconnect between experimentation / training / production	→	3. Unified architecture for experimentation / training / production
4. Custom one-off pipelines to make data available in real time	→	4. Unified architecture for batch and real time

HOPSWORKS

Serverless

A free sandbox for everyone

In a year;

Over 3500 Users

What is Serverless?

Feature Store + Model Registry + Model Serving
Same User Experience & Same API
No Infrastructure to Manage
No Time Limit
Free Forever

Community

Predicting Crime in San Francisco
Serverless ML system that classifies the incident category based on its time and location in San Francisco, US.
Source: *https://github.com/Hope-Liang/ID2223Project*

Predicting Electricity Prices in NYC
Prediction service that predicts the daily electricity demand in megawatthours in New York, USA.
Source: *https://github.com/aykhazanchi/id2223-scalable-ml/tree/master/proj*

Electricity Price Prediction for Sweden
Predicting the daily average energy price in Stockholm/SE3 for the upcoming 7 days.
Source: *https://github.com/antonbn/ID2223Project*

Double The Resolution Of Your Image
Doubling pictures' resolution.
Source: *https://github.com/GianlucaRub/Scalable-Machine-Learning-and-Deep-Learning/tree/main/Project*

News Articles For A Specified Sentiment
ML pipeline that predicts the sentiment of and recommends news articles based on their headlines.
Source: *https://github.com/torileatherman/news_articles_sentiment*

serverless-ml.org

Educators

Energy Forecasting

https://github.com/iusztinpaul/energy-forecasting

The Full Stack 7-Steps MLOps Framework
by Paul Iusztin

Real World ML

https://www.realworldml.xyz
https://twitter.com/paulabartabajo_

by Pau Labarta Bajo

Great but…

Not Meant for Enterprise

No SLAs - Shared Infrastructure
- Limited Quotas

Hello, we are using Hopsworks as an option for a serverless feature store as well as your MLOps capabilities. ***We are a small-medium sized companies with expected API calls of less than 20k/month.*** *Can you provide more pricing information for your service?*

"We want to discuss about the increase in quota."

"Hi. ***I would like to know about increasing quotas pricing.*** *I couldn't found on the site. Thanks."*

Hi, We're looking into options that would allow us to produce reusable and consistent features across our data science, analytics and MLOps teams, ***we would like to avoid paying the infrastructure twice****"*

"I am currently working on a personal project and ***would need more than the capability available with free subscription****. Kindly share available subscription plans with me."*

"Our feature store requirements are fairly simple and we'd basically like a better Dev-X over BigQuery"

HOPSWORKS

To Introducing the

Hopsworks SaaS

Enterprise SLAs on a Managed platform, e2e.

FEATURE STORE SUMMIT 2023

What's in the box?

The Feature Store

+ Model Registry & Model Serving

Hopsworks SaaS

Feature Store + Model Registry + Model Serving

Join the Beta Now!

www.hopsworks.ai/saas

From
USD 99 / Month
Beta users:
First Month Free

Enterprise SLAs

Up to 200gb Offline
Up to 5GB Online
5 Model Deployments

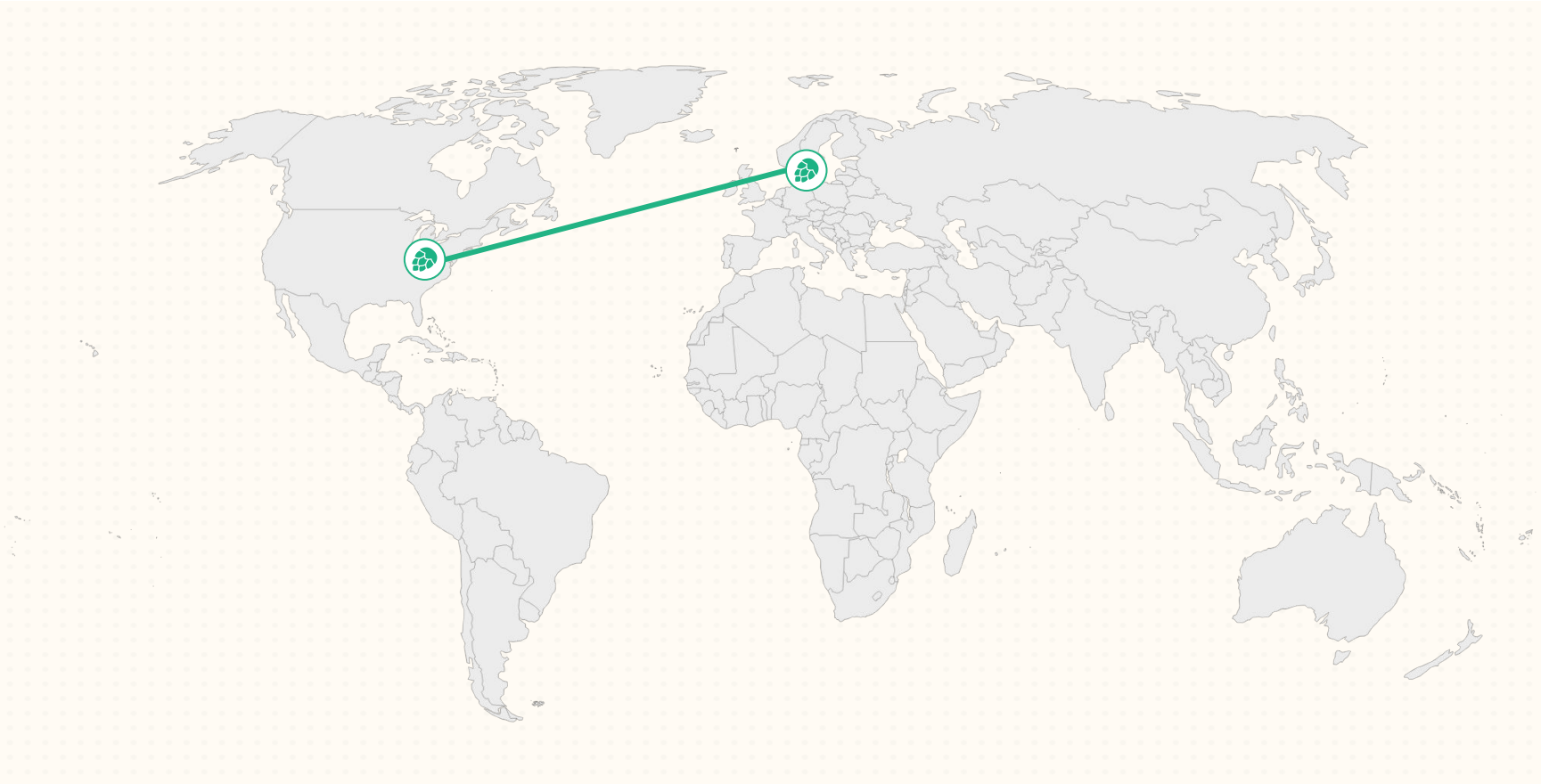
HOPSWORKS

To Introducing the

Multi-Region Availability

FEATURE STORE SUMMIT 2023

Multi-Region Availability



Multi-Region Availability

Thank you!

Follow us on X (Twitter): @hopsworks

Meet us in person: https://hopsworks.ai/events

FEATURE STORE SUMMIT 2023