ML Infra @ Dropbox
Overview

Tsahi Glik
Sep 12, 2019

ML @ dropbox

Our signal sources:

Files

Multi-exabyte data

File Metadata

Trillions

User interactions

Billions / day

ML @ dropbox

ML Impact at Dropbox:

- Smart Sync
- Content Suggestions
- Team Activity Ranking
- Search Ranking
- OCR

And many more …

ML Platform

Challenges:

- **Huge** data sources that are **isolated** in various system across production
- Multiple **privacy** levels of data
- **Custom work** and build dedicated services for each new use case
- **Manual** training which is **hard to reproduce**
- **Wide variety** of development processes and ML frameworks

ML Platform

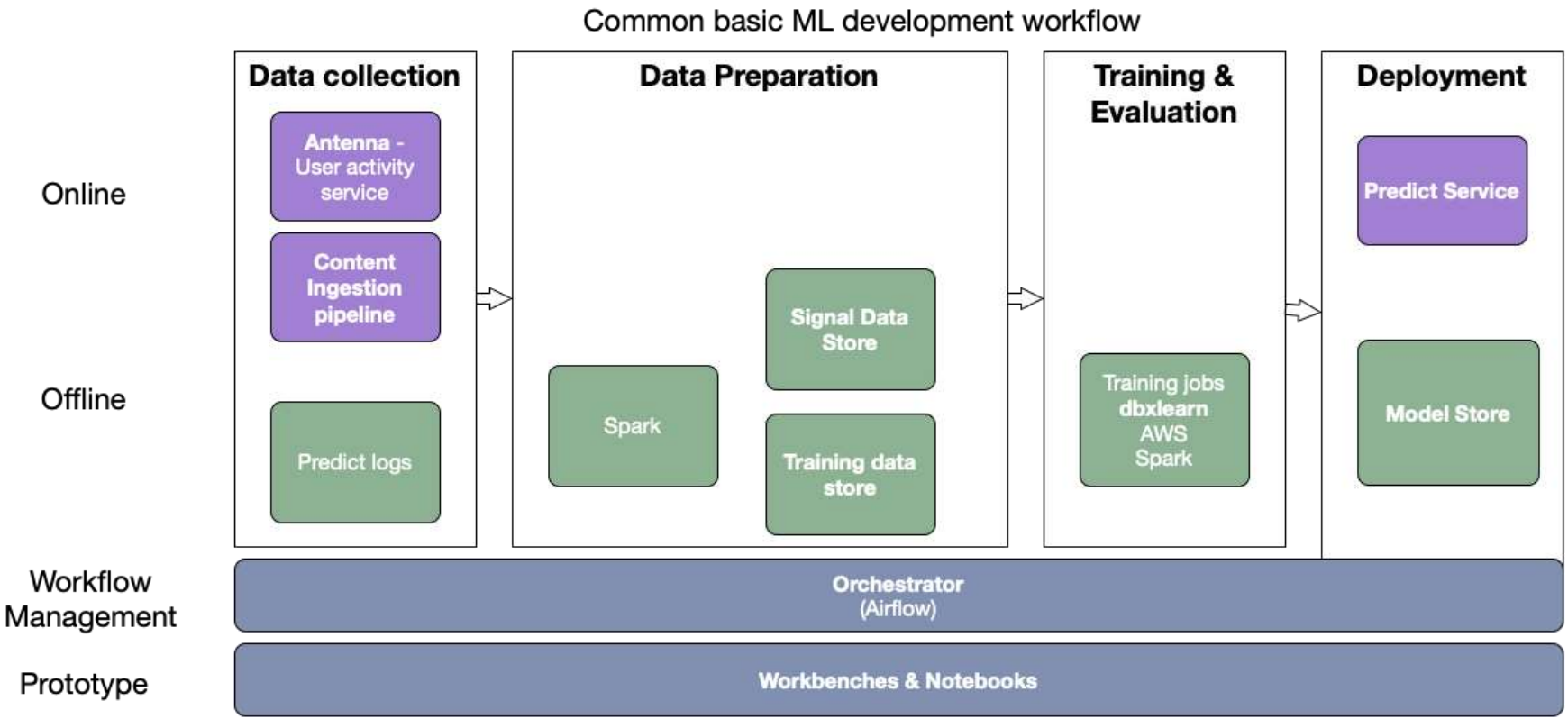
Mission:

Accelerate intelligent product development at Dropbox

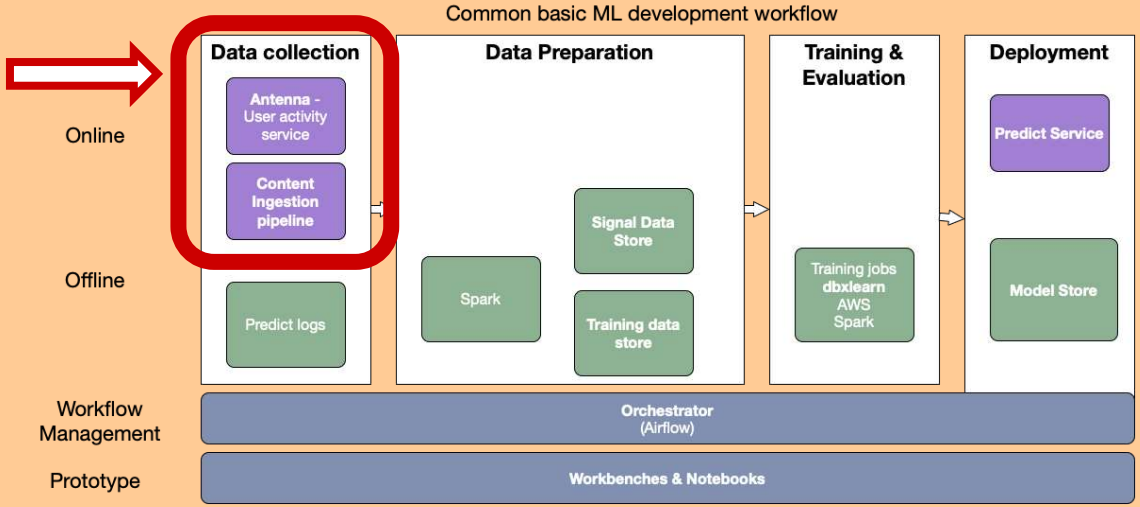
By:

- Scalable access to data for offline and online
- Ensures sensitive data is protected and accessed only in approved ways
- Easy model deployment & experimentation
- Automate workflows
- Standardize the process, frameworks and tools

Platform Architecture



Online Data Collection

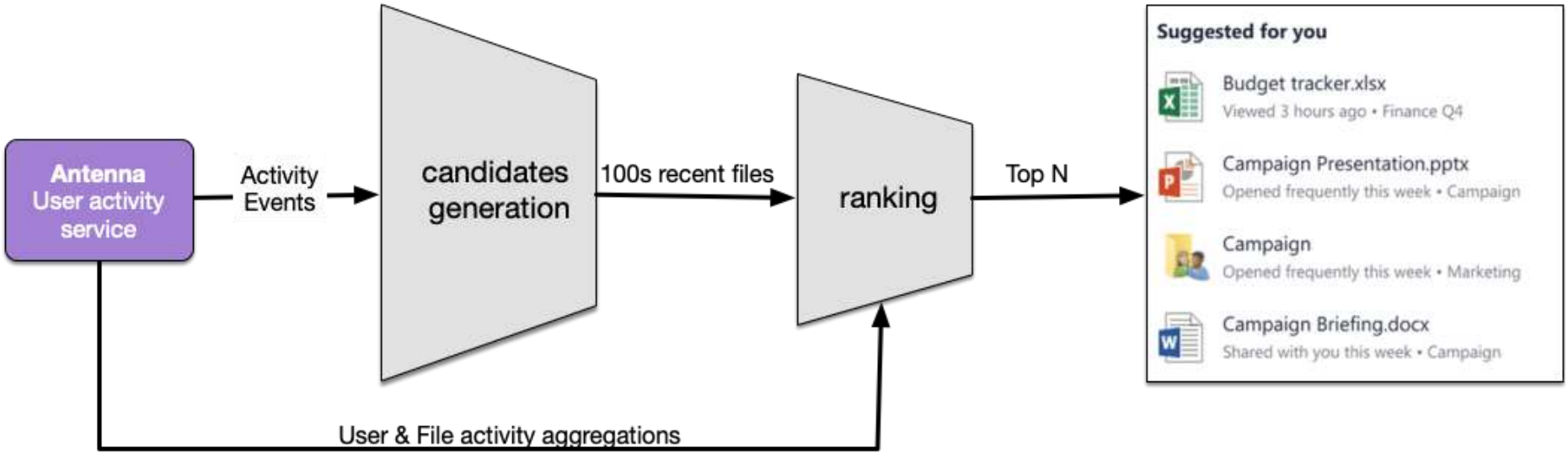


Antenna

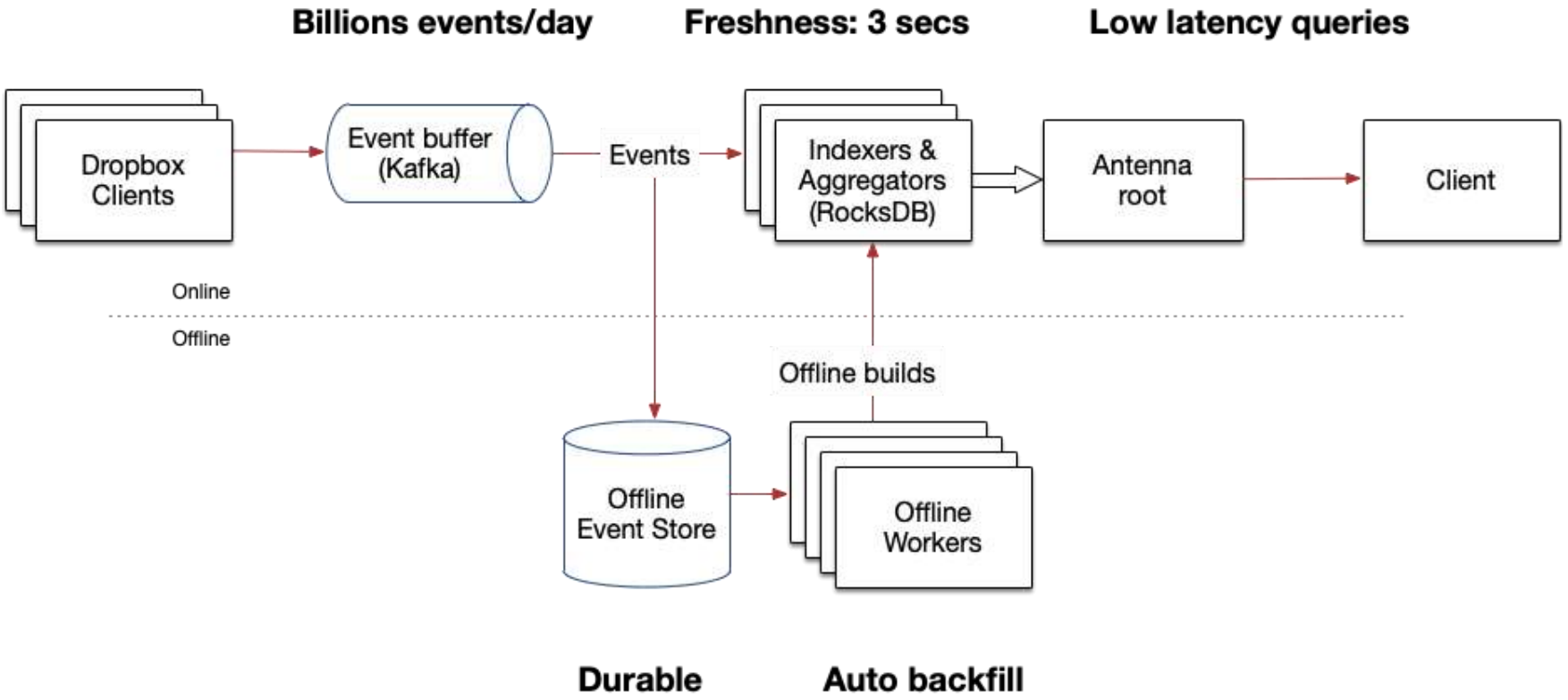
What is Antenna?

- User activity service
- Provides various ways to query activity events
- Support aggregations for simple summaries and histograms of activity data

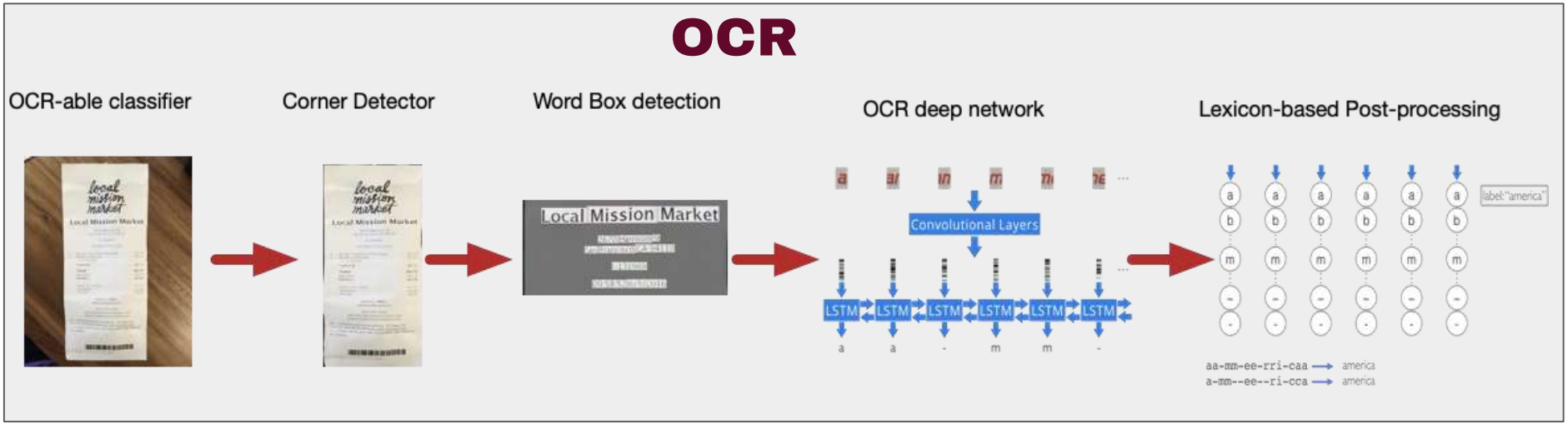
Example usage of Antenna



Antenna Architecture

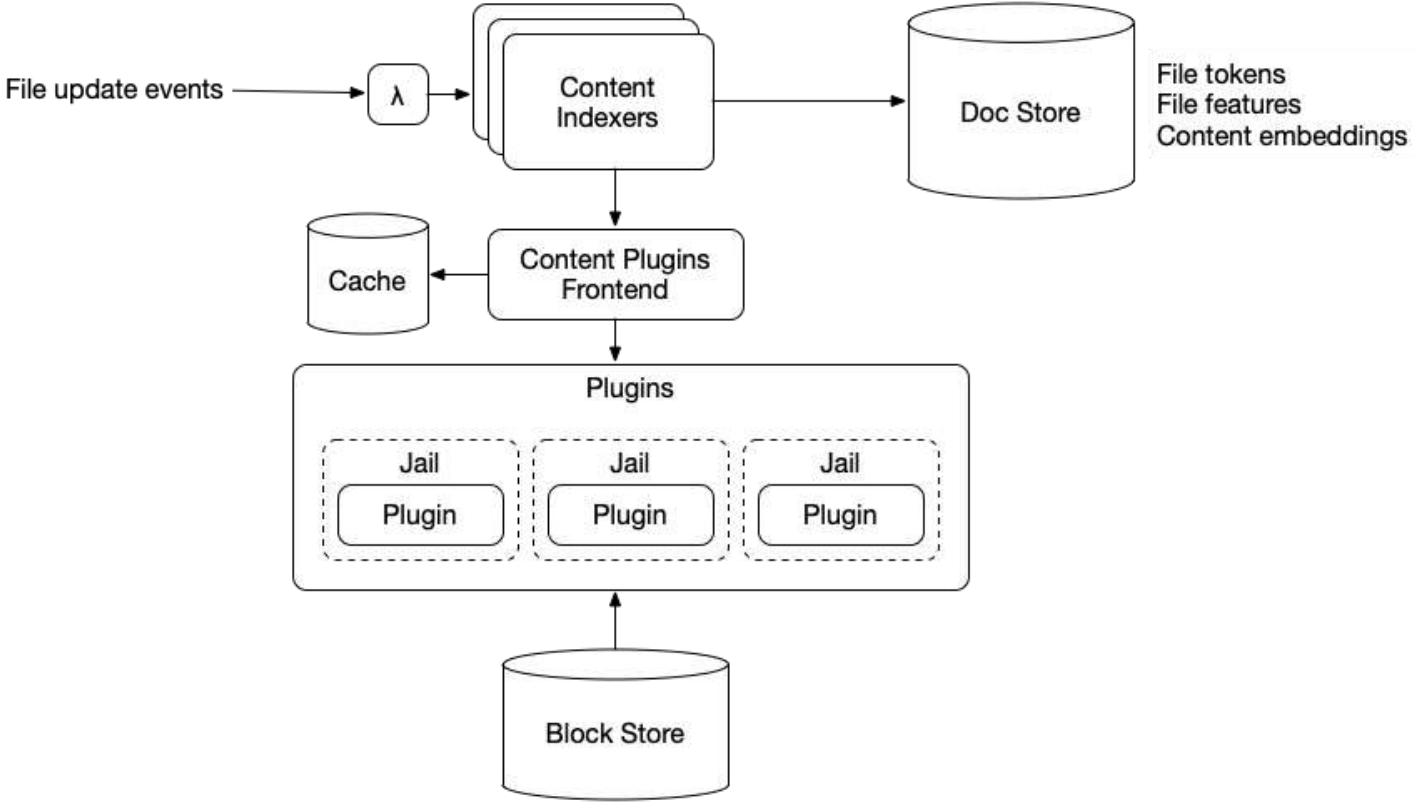


Content Ingestion pipeline

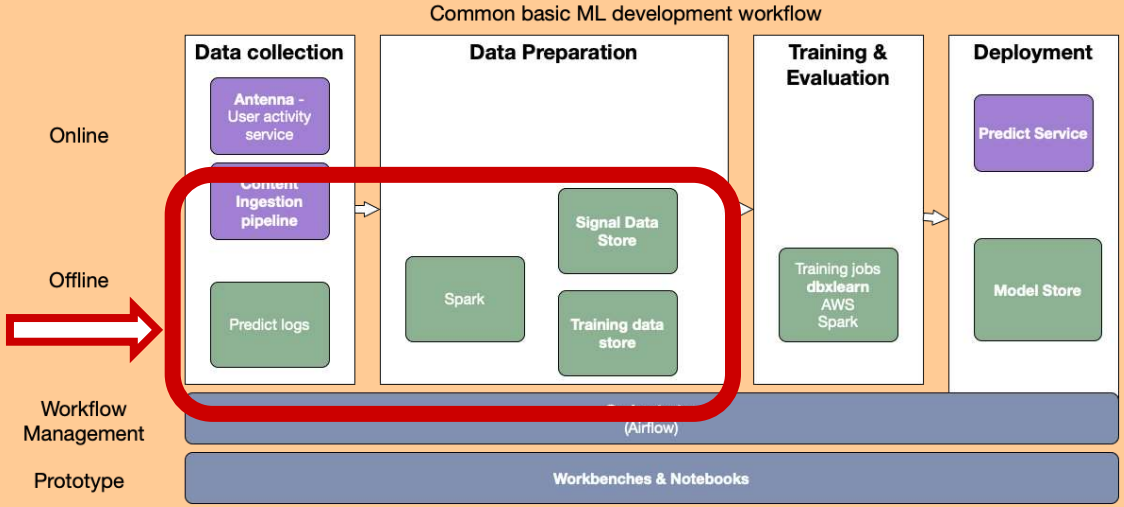


Read more in our blog post:

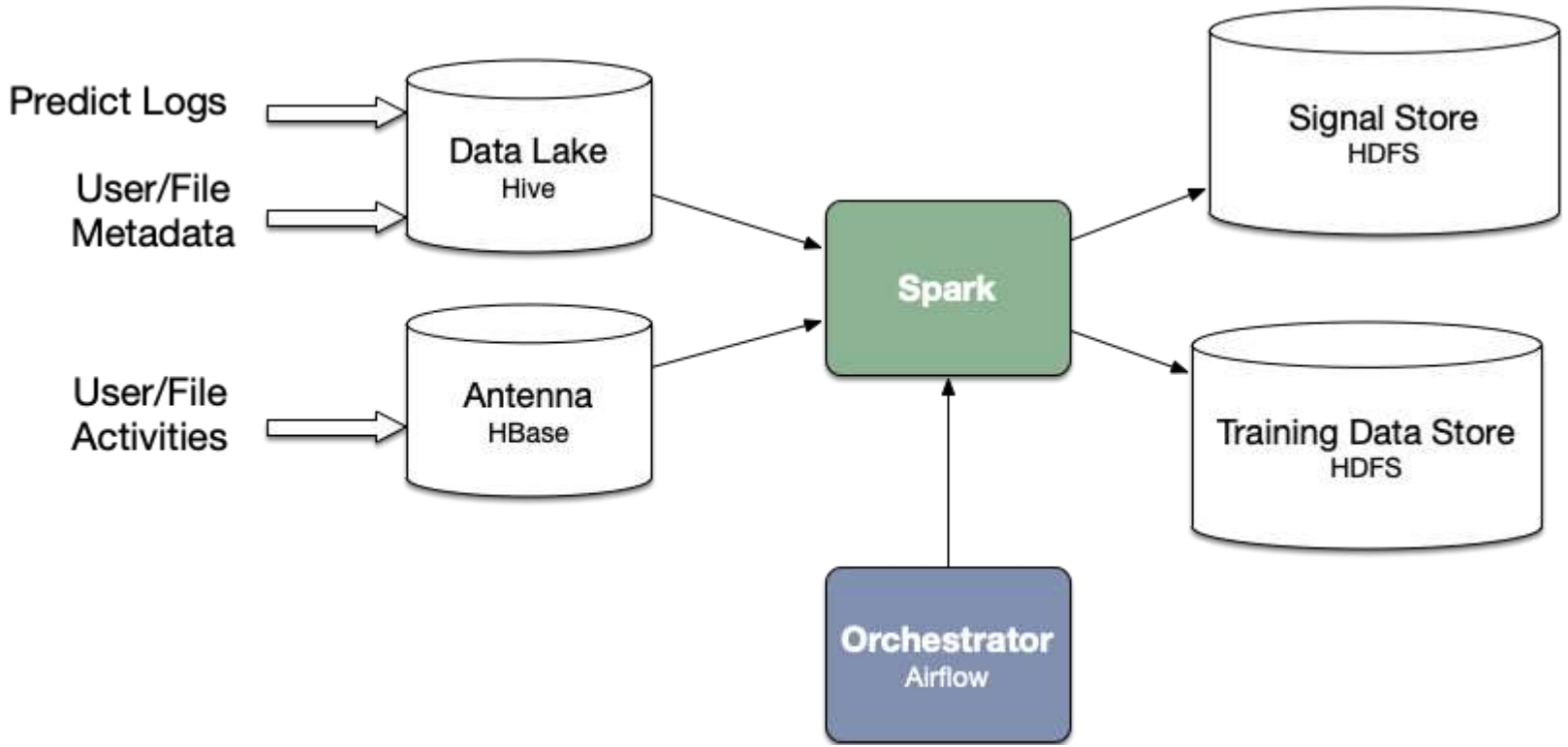
Content Ingestion Architecture



Offline Data Preparation

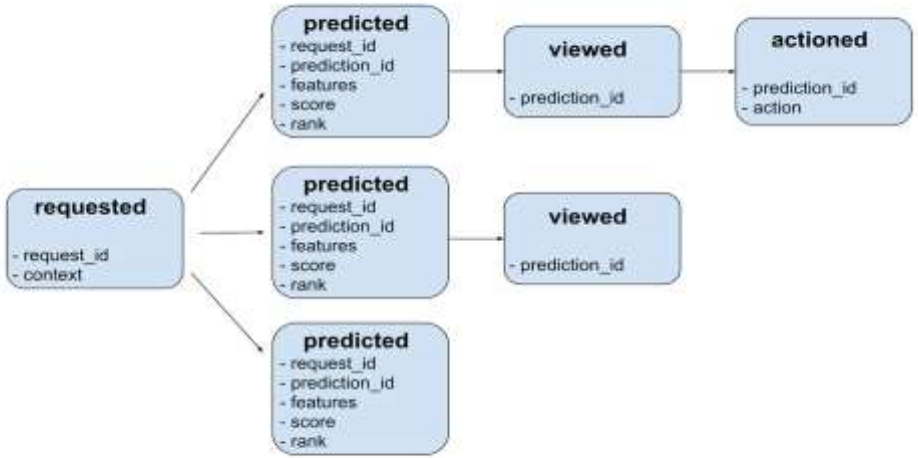


Data Preparation - ETL pipeline

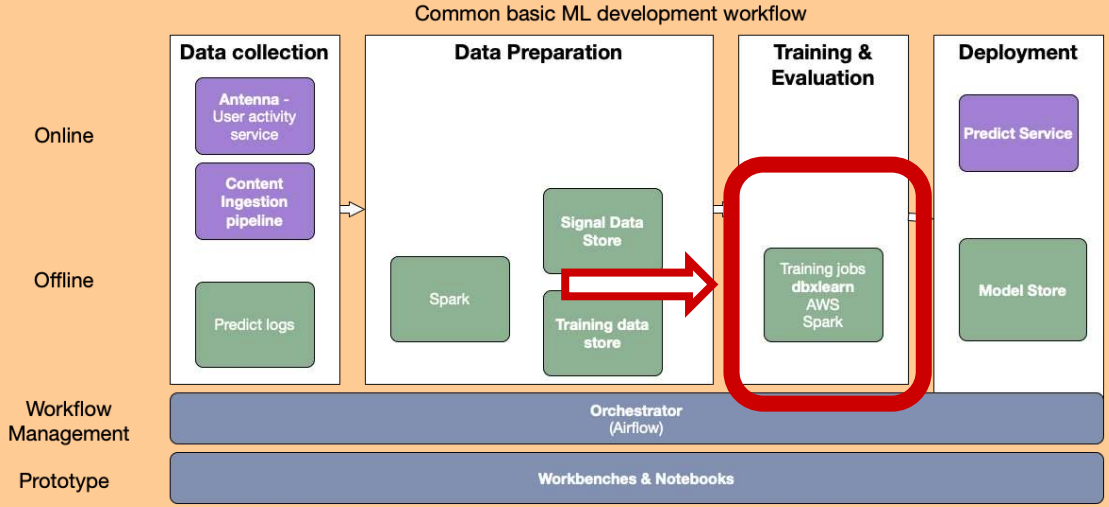


Data Preparation - Predict logger

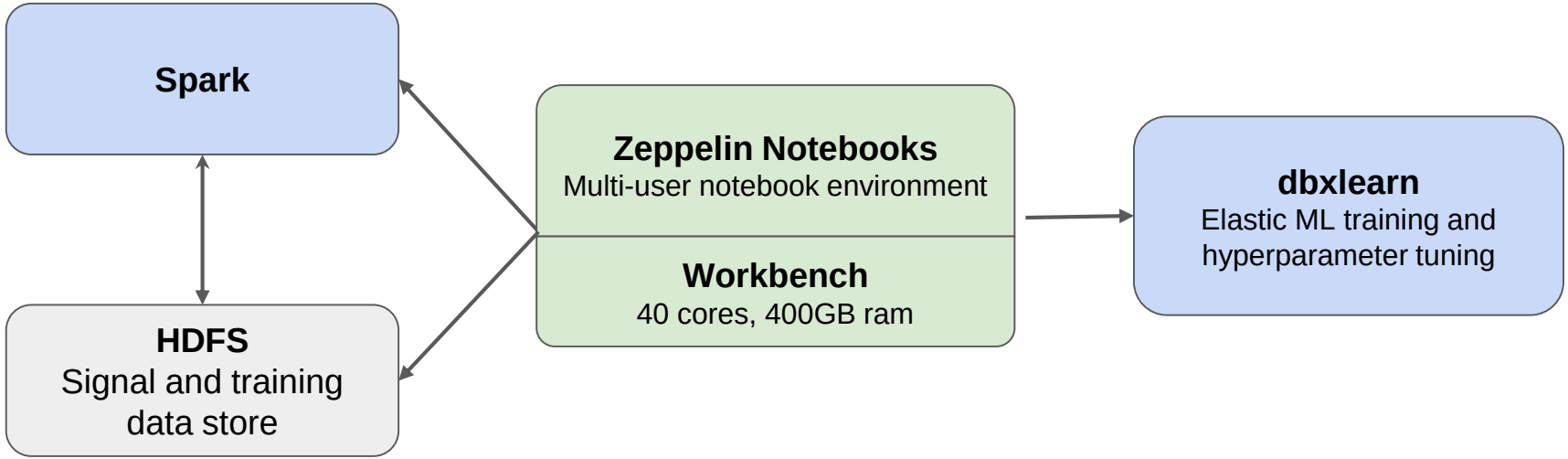
- Converting raw logs into labeled datasets
- Logging partial information from different services at different times
- Eliminate discrepancies between online and offline



Offline Training & Evaluation



Prototyping

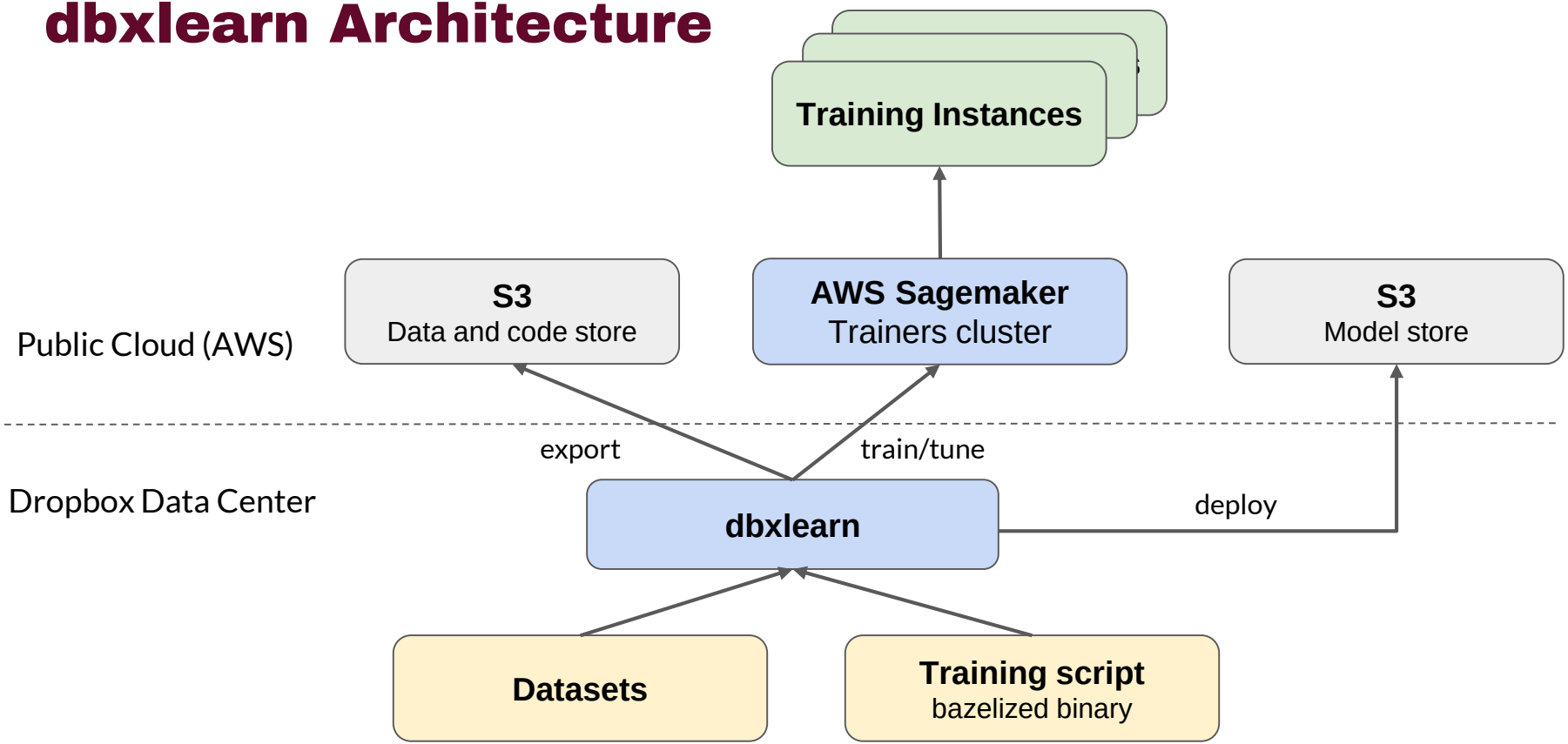


dbxlearn

What is dbxlearn?

- dbxlearn provides an easy way to use computing at scale for training
- Core problems dbxlearn is addressing:
  - Elasticity
  - Standard way to train on different hw configurations (GPU, TPU) on different cloud platforms.
- Hybrid cloud architecture - Interface with private cluster and well as public clouds
- Currently integrated with AWS and use SageMaker

dbxlearn Architecture



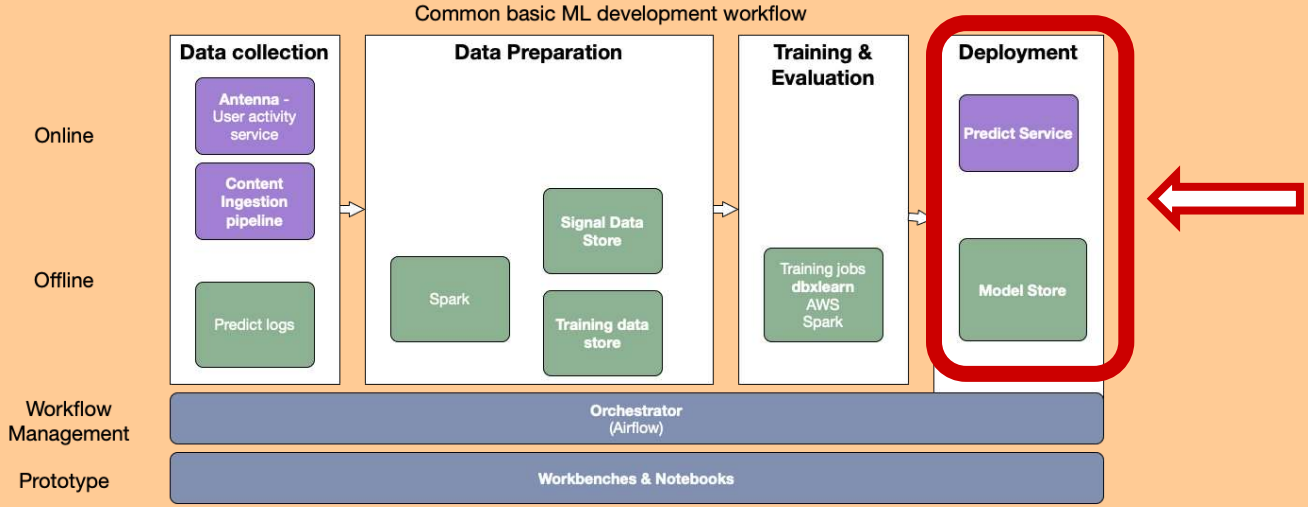
dbxlearn workflow

```
$ dbxlearn train --py-binary <script>
  --train_uri <...> --validation_uri <...> [--local]

$ dbxlearn tune --py-binary <script> --train_uri <...> --validation_uri <...>

$ dbxlearn query --tuning_job_id <id> print_top_summary
```

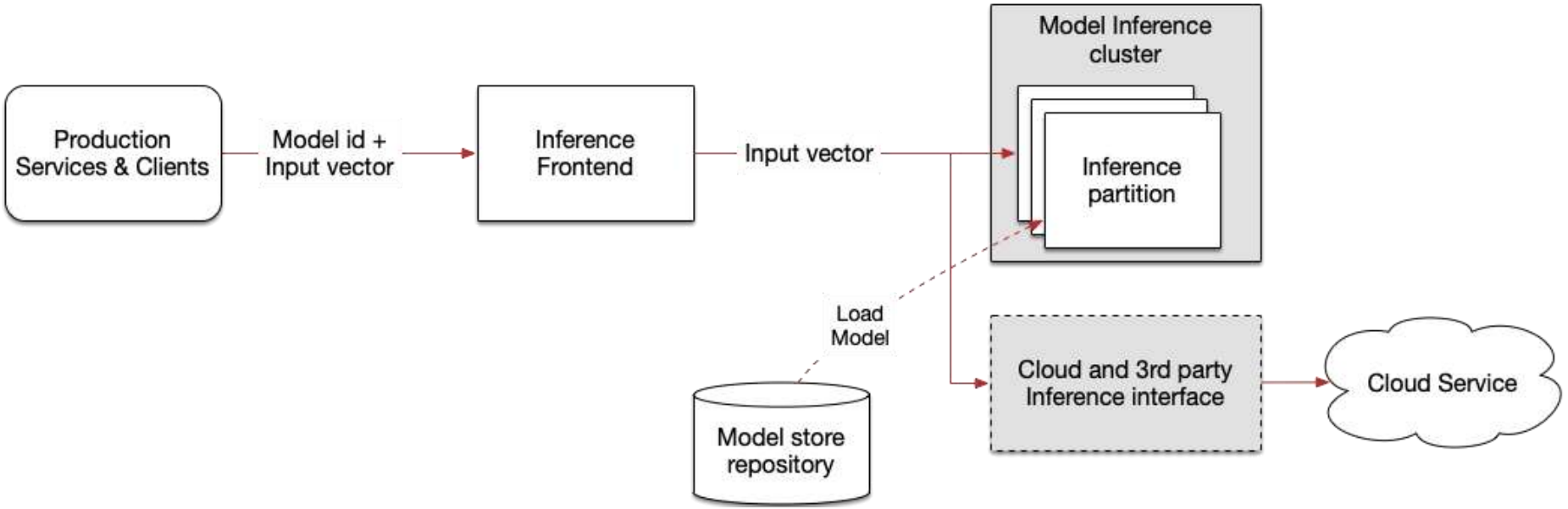
Job Id	Status	layer1_width	layer0_width	learning_rate	dropout_prob	positive_weight_ratio	ROC-AUC	PR-AUC
file-suggestion-tune-1907241740-068-8d31018e	Completed	"32"	"64"	0.0008	0.4	12.1856	0.9245	0.5806
file-suggestion-tune-1907241740-065-4f14af15	Completed	"32"	"64"	0.0009	0.3996	14.6465	0.9245	0.5756
file-suggestion-tune-1907241740-028-de802913	Completed	"32"	"64"	0.0006	0.2556	9.1835	0.9245	0.582
file-suggestion-tune-1907241740-054-0a8e0867	Completed	"32"	"128"	0.0009	0.3918	5.2485	0.9245	0.5844
file-suggestion-tune-1907241740-023-bbddf16e	Completed	"32"	"64"	0.0006	0.3638	12.5284	0.9245	0.5829
file-suggestion-tune-1907241740-060-771e35b2	Completed	"16"	"128"	0.0008	0.2323	16.3034	0.9245	0.5791
file-suggestion-tune-1907241740-050-ee3bfa1e	Completed	"32"	"128"	0.001	0.4	15.9832	0.9244	0.5801
file-suggestion-tune-1907241740-053-1a47cb67	Completed	"16"	"128"	0.0006	0.2974	5.5142	0.9244	0.5842
file-suggestion-tune-1907241740-069-79e79037	Completed	"32"	"64"	0.0008	0.3988	11.9856	0.9244	0.5808
file-suggestion-tune-1907241740-058-1f55d16e	Completed	"32"	"128"	0.0008	0.371	7.4362	0.9244	0.5813

```
$ dbxlearn deploy-model --tuning-job_id <id> <experiment-group>
```

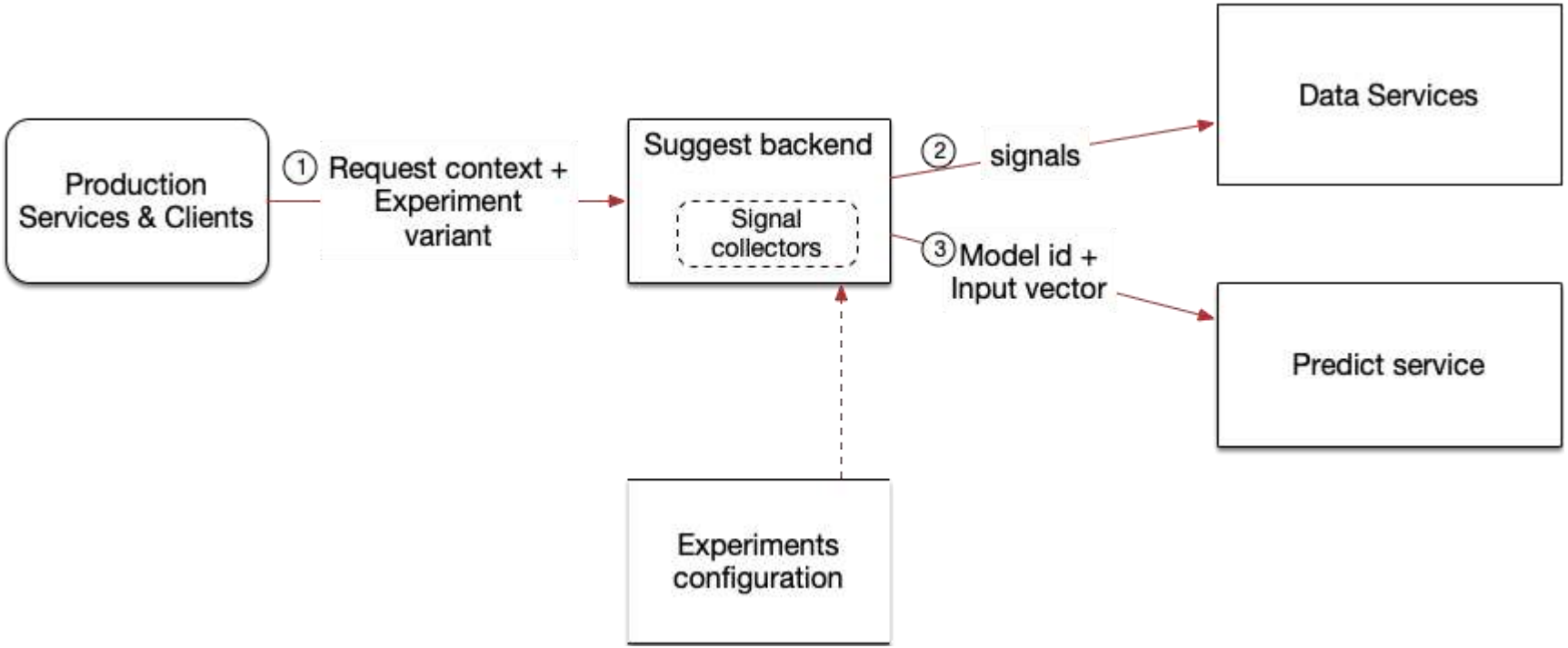
Model Deployment



Predict service



Live experimentation - Suggest backend

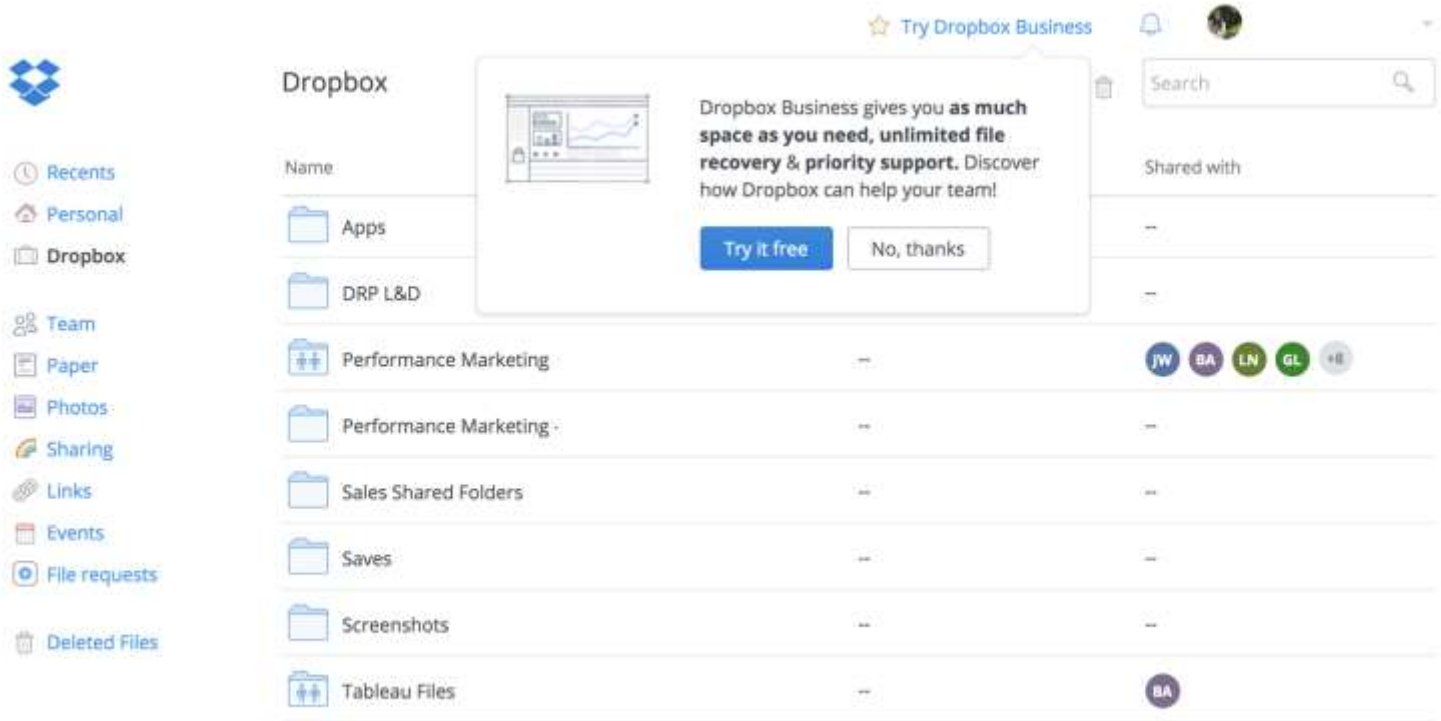


Shadow experimentation - Suggest backend

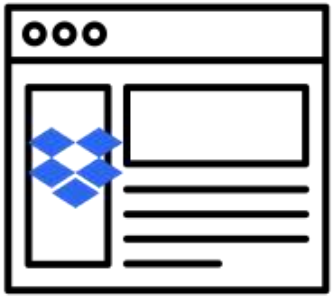
- Send live traffic to shadow cluster with a different experiment variant
- Results are logged for experiment analysis
- Useful to collect labeled datasets using Predict Logger

Example

Campaign Ranker - Using Multi Arm Bandits



Campaign Ranker - Using Multi Arm Bandits



Click

View

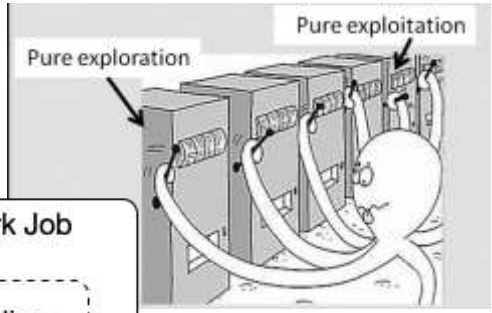
Context + Signals

Predict Logger

New training data

Spark Job

Policy Update



User & Page features

Ranked Campaigns

Predict service

Refresh Model

Model Store

Summary

- **End-to-End platform** that supports all steps in ML development workflow
- **Deep integration** with Dropbox large scale data sources
- **Flexible APIs** to support wide variety of use cases
- **Hybrid cloud architecture** for elasticity and early adoption of new technologies

Next Challenges

- **Better representation of data relations** across multiple systems
- **Democratize ML at dropbox**, extending our tools from ML developers to more engineers

Thank You