CLOUDERA

MACHINE LEARNING IN THE ENTERPRISE

Joon Kim | Solutions Engineer
Vincent Fortier | Senior Solutions Engineer

NDA

DISCLAIMER

The information in this document is proprietary to Cloudera. No part of this document may be reproduced, copied or transmitted in any form for any purpose without the express prior written permission of Cloudera.

This document is a preliminary version and not subject to your license agreement or any other agreement with Cloudera. This document contains only intended strategies, developments and functionalities of Cloudera products and is not intended to be binding upon Cloudera to any particular course of business, product strategy and/or development. Please note that this document is subject to change and may be changed by Cloudera at any time without notice.

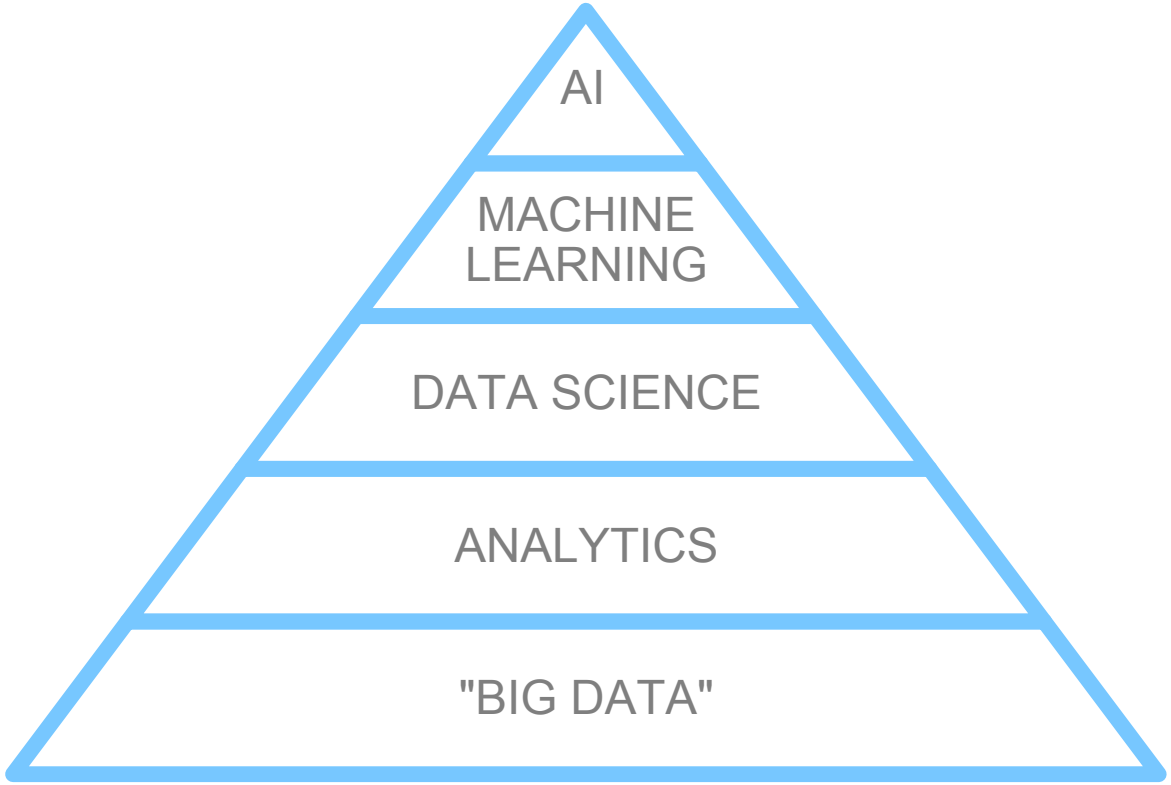
Cloudera assumes no responsibility for errors or omissions in this document. Cloudera does not warrant the accuracy or completeness of the information, text, graphics, links or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose or non-infringement.

Cloudera shall have no liability for damages of any kind including without limitation direct, special, indirect or consequential damages that may result from the use of these materials. The limitation shall not apply in cases of gross negligence.

CLOUDERA

MACHINE LEARNING PRODUCT OVERVIEW

Vincent Fortier | Senior Solutions Engineer


AI
MACHINE LEARNING
DATA SCIENCE
ANALYTICS
"BIG DATA"

THE PLATFORM FOR INDUSTRIALIZED AI

End-to-end machine learning infrastructure for teams at scale

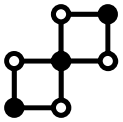
DEVELOP

Make **teams** more productive

Explore shared data

Open, extensible workspaces

Collaborate with peers



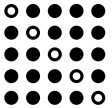
TRAIN

Scale **resources** efficiently

Train models

Tune parameters

Track performance



DEPLOY

Manage **data, pipelines, models**

Automate pipelines

Deploy models anywhere

Monitor performance

MANAGE

Run anywhere with a common **architecture**

Instantly enable new teams

Manage access and resources

Scale cost with usage

Enable the AI-first enterprise by making data teams more productive

1

Self-service

Give data science teams easy access to the diverse data, compute and open source tools they need.

2

Workflows

Enable modern collaborative and reproducible software development practices, and multiple paths to production.

3

Anywhere

Deliver the same experience for public cloud, private cloud, or traditional data management environments.

Why this is hard

Balance these needs

DATA SCIENCE

- Granular data
- Diverse open source tools
- Elastic resources
  - CPU, GPU, memory, disk
- Track everything
- Drive, measure impact

VS.

IT / DATA MANAGEMENT

- Don’t move data
- Maintain security, governance
- Enforce standards
- Isolate noisy neighbors
- Control costs
- Maintain portability

The typical solution

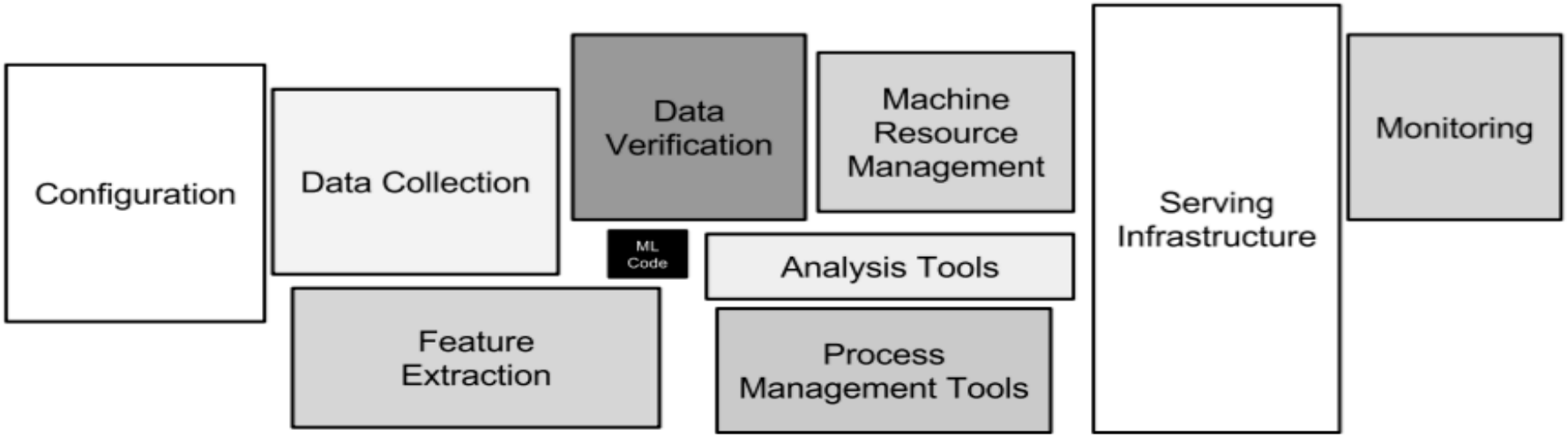
"If I can't use my favorite tools,
or get the compute I need, I'll…"

- Copy data to my laptop
- Copy data to a data science appliance
- Copy data to a cloud service

Why this is a problem:

- Complicates security
- Breaks data governance
- Adds latency to process
- Makes collaboration more difficult
- Complicates model management and deployment
- Creates infrastructure silos

Hidden technical debt in machine learning systems

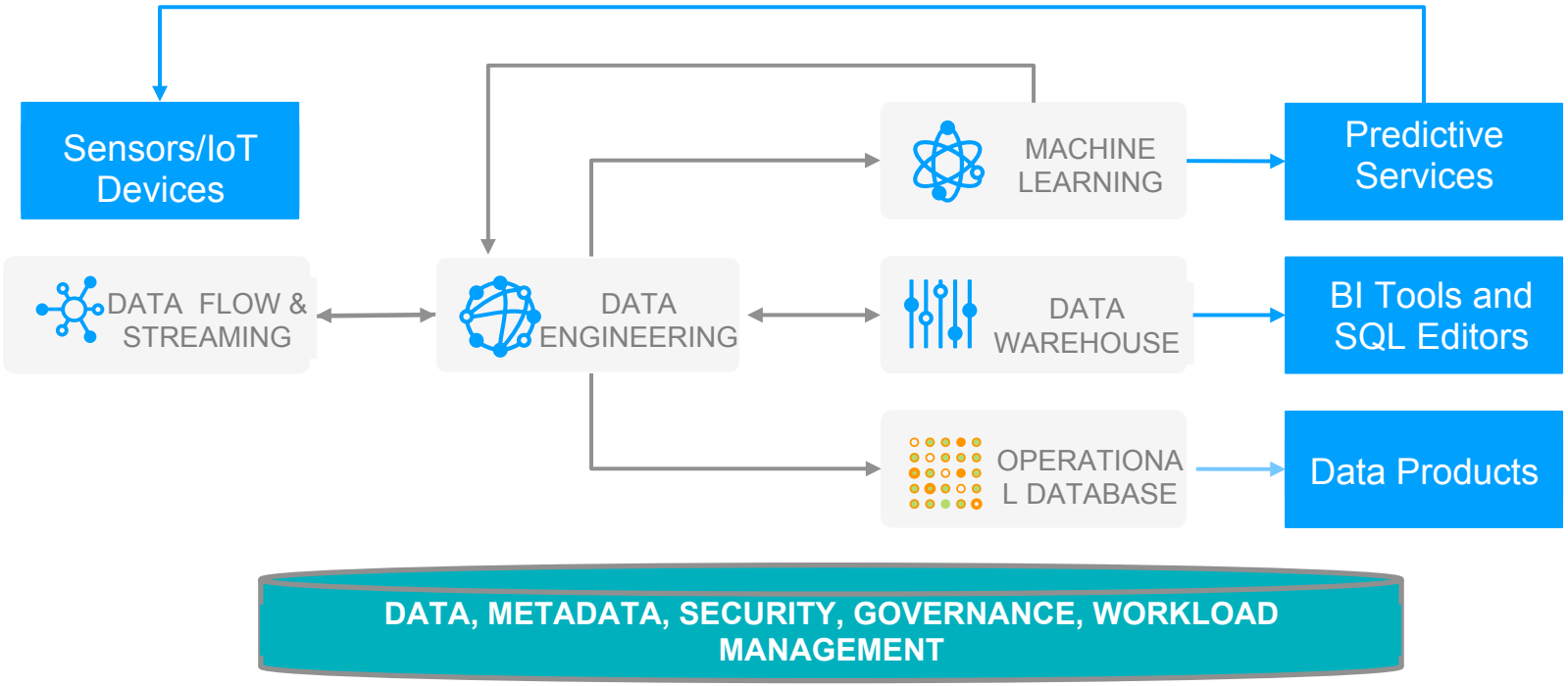


Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

CLOUDERA
MACHINE LEARNING

WHAT INDUSTRIALIZED MACHINE LEARNING LOOKS LIKE



MACHINE LEARNING IS BUILT ON DATA MANAGEMENT

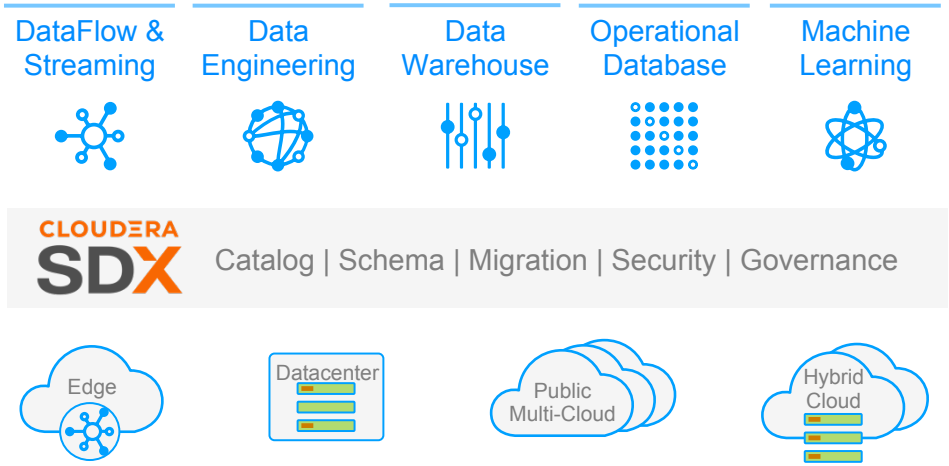
We deliver an Enterprise Data Cloud for any data, anywhere, from the edge to AI

EDGE TO AI

Enterprise grade
Secure, performant and compliant

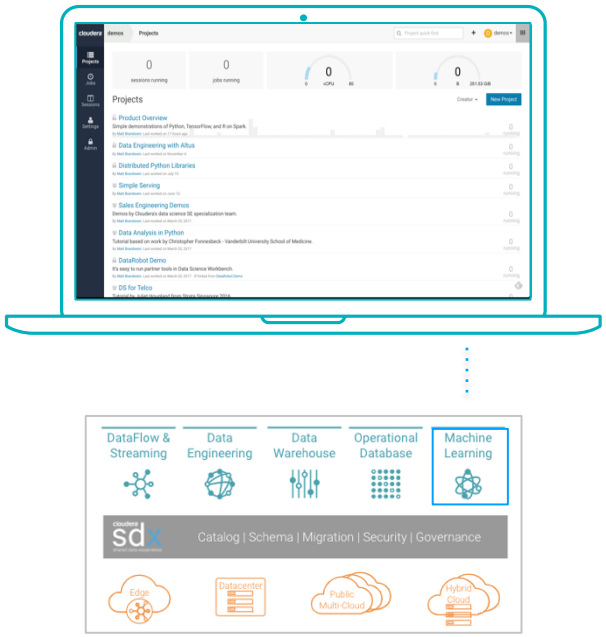
Scalable
Elastic, cost-effective and lower TCO

Runs anywhere
Public cloud, on-premises, multi, hybrid



CLOUDERA MACHINE LEARNING (CDSW TODAY)

Accelerate machine learning from research to production



For data scientists

- **Experiment faster**
  Use R, Python, or Scala with on-demand compute and secure CDH/HDP data access
- **Work together**
  Share reproducible research with your whole team
- **Deploy with confidence**
  Get to production consistently without recoding

For IT professionals

- **Bring data science to the data**
  Give your data science team more freedom while reducing the risk and cost of silos
- **Secure by default**
  Leverage common security and governance across workloads
- **Run anywhere**
  Cloud-native or on-premise (With CDSW)

MACHINE LEARNING FOR CLOUDERA DATA PLATFORM

Cloud-native enterprise machine learning as-a-service

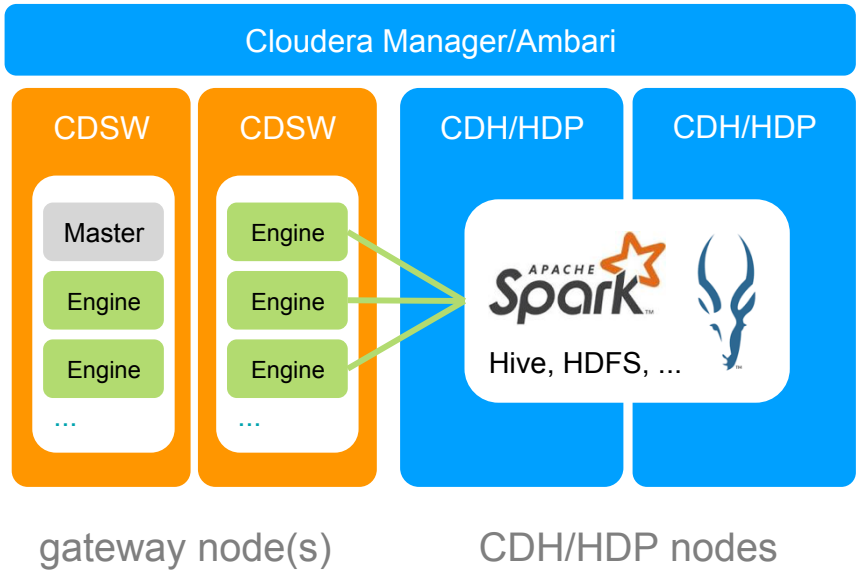
- **DS/ML from research to production**
  - Purpose built for Seamless Data Engineer and Data Science workflows
- **Seamless scale-out experience**
  - Rapid provisioning and elastic autoscaling
  - Automatic dependency management
  - Distributed CPU and GPU model training
- **Managed service on CDP**
  - No (Spark) clusters to manage
  - Containerized multi-cloud portability
  - Private cloud option with CDP-Private

CDSW IN ACTION

A MODERN DATA SCIENCE ARCHITECTURE

Containerized environments with scalable, on-demand compute

- Built with Docker and Kubernetes
  - Isolated, reproducible user environments
- Supports both big and small data
  - Local Python, R, Scala runtimes
  - Schedule & share GPU resources
  - Run Spark, Impala, and other CDH services
- Secure and governed by default
  - Easy, audited access to Kerberized clusters
  - Leverages SDX platform services
- Deployed with Cloudera Manager

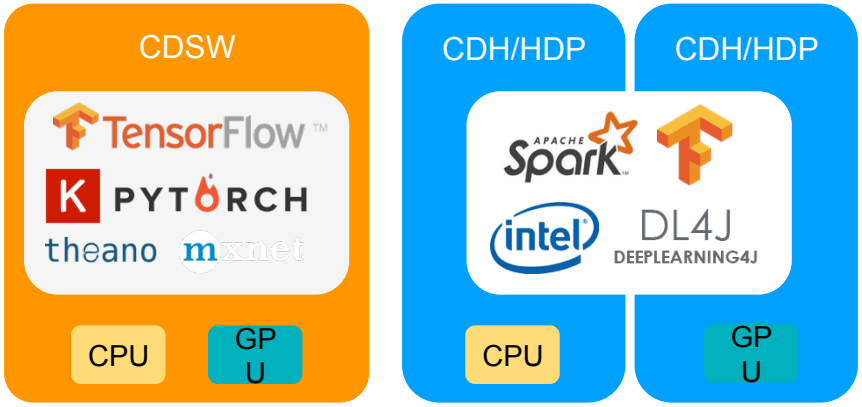


ACCELERATED DEEP LEARNING WITH GPUS

Multi-tenant GPU support on-premises or cloud

"Our data scientists want GPUs, but we need multi-tenancy. If they go to the cloud on their own, it's expensive and we lose governance."

- Extend CDSW to deep learning
- Schedule & share GPU resources
- Train on GPUs, deploy on CPUs
- Works on-premises or cloud



CHALLENGE: REPRODUCIBLE, AUDITABLE MODEL TRAINING

How do you know what model is better? How can you repeat a result?

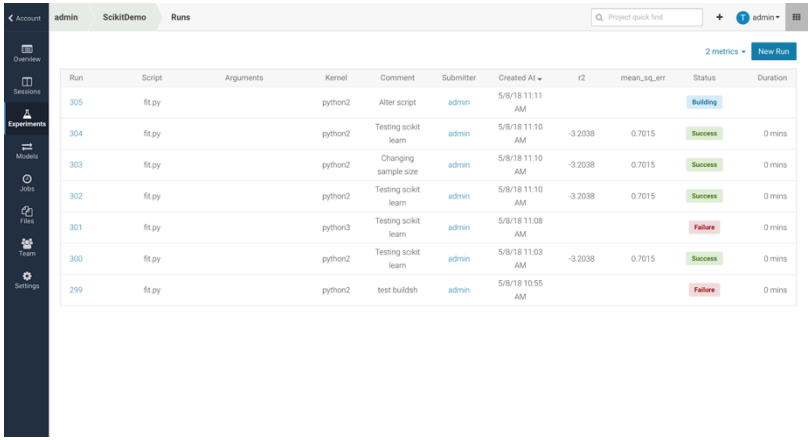
- Model development is iterative
  - Try different data, features, libraries, algorithms, hyperparameters, etc.
- Reproducing a model means you need
  - Training data
  - Data/feature pipeline code
  - Model training code + dependencies
  - Runtime environment (CPU, GPU, memory, …)
  - Any results or performance metrics
- This is a lot to keep track of!

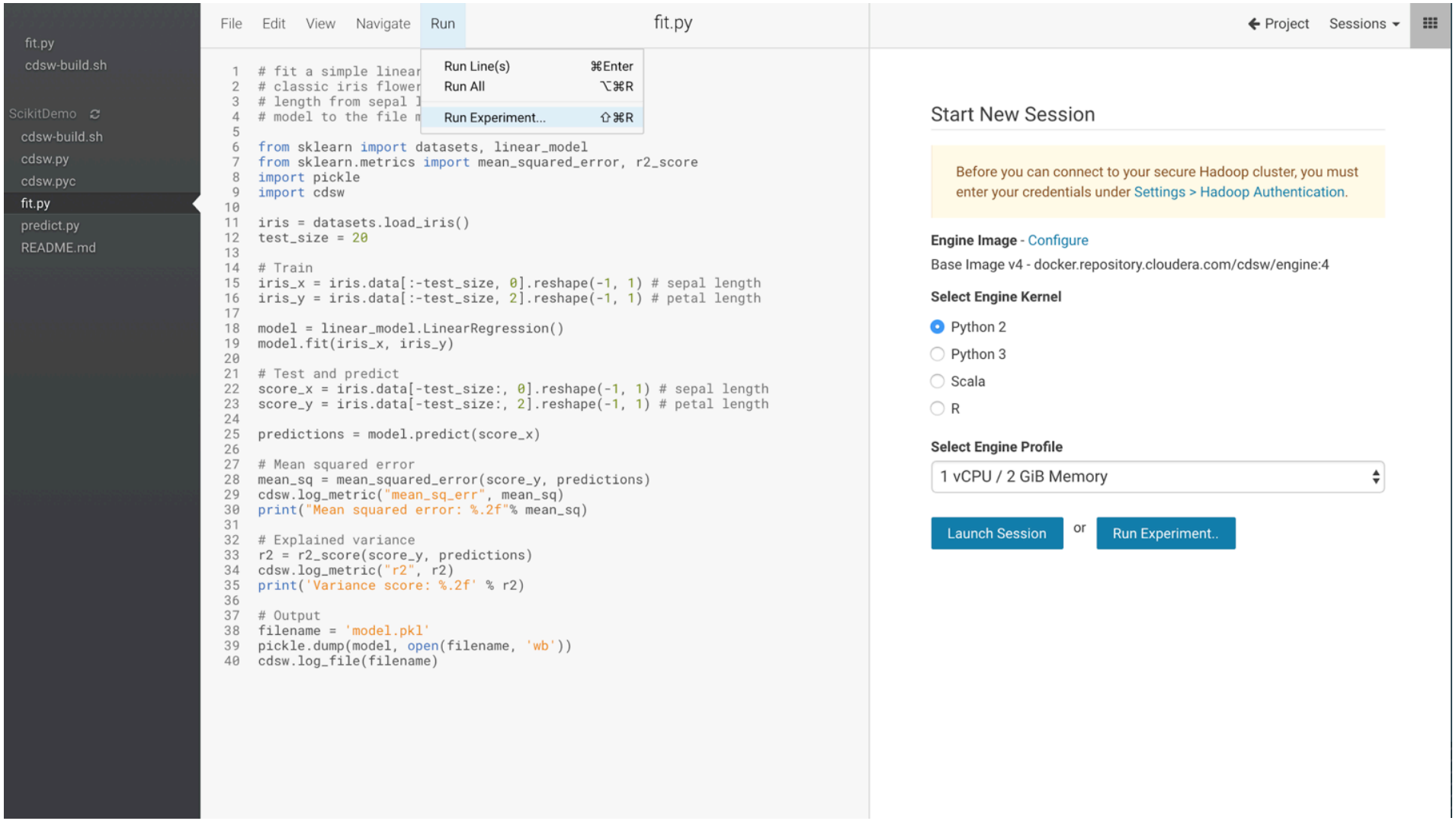
EXPERIMENTS

Versioned model training runs for evaluation and reproducibility

Data scientists can now...

- Create a snapshot of model code, dependencies, and configuration necessary to train the model
- Build and execute the training run in an isolated container
- Track specified model metrics, performance, and model artifacts
- Inspect, compare, or deploy prior models




fit.py
cdsw-build.sh
ScikitDemo
cdsw-build.sh
cdsw.py
cdsw.pyc
fit.py
predict.py
README.md
File
Edit
View
Navigate
Run
fit.py
Run Line(s) ⌘Enter
Run All ⌥⌘R
Run Experiment... ⇧⌘R
# fit a simple linear
# classic iris flower
# length from sepal l
# model to the file m
from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score
import pickle
import cdsw
iris = datasets.load_iris()
test_size = 20
# Train
iris_x = iris.data[:-test_size, 0].reshape(-1, 1) # sepal length
iris_y = iris.data[:-test_size, 2].reshape(-1, 1) # petal length
model = linear_model.LinearRegression()
model.fit(iris_x, iris_y)
# Test and predict
score_x = iris.data[-test_size:, 0].reshape(-1, 1) # sepal length
score_y = iris.data[-test_size:, 2].reshape(-1, 1) # petal length
predictions = model.predict(score_x)
# Mean squared error
mean_sq = mean_squared_error(score_y, predictions)
cdsw.log_metric("mean_sq_err", mean_sq)
print("Mean squared error: %.2f"% mean_sq)
# Explained variance
r2 = r2_score(score_y, predictions)
cdsw.log_metric("r2", r2)
print('Variance score: %.2f' % r2)
# Output
filename = 'model.pkl'
pickle.dump(model, open(filename, 'wb'))
cdsw.log_file(filename)
Project
Sessions
Start New Session
Before you can connect to your secure Hadoop cluster, you must enter your credentials under Settings > Hadoop Authentication.
Engine Image - Configure
Base Image v4 - docker.repository.cloudera.com/cdsw/engine:4
Select Engine Kernel
Python 2
Python 3
Scala
R
Select Engine Profile
1 vCPU / 2 GiB Memory
Launch Session
or
Run Experiment..

CHALLENGE: GETTING TO PRODUCTION

So you've got a trained model. Now what?

- But data scientists want to rapidly expose candidate models to serve predictions
- Development and production are very different
  - Owners: Data Scientists vs. Data Engineers
  - Languages: Python/R vs. Java/Scala/C++
  - Policy Controls: Approved code, packages, etc.
  - Vocabulary: Data Science vs. DevOps
- Data scientists do not often have the skills (or entitlements) to deploy models

MODELS

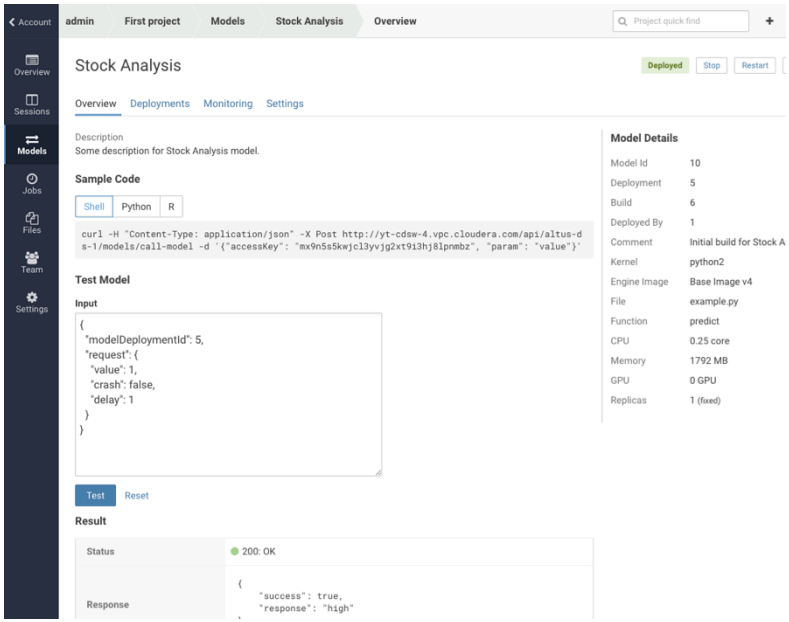
Machine learning models as one-click microservices (REST APIs)

1. Choose file, e.g. `score.py`
2. Choose function, e.g. `forecast`

```
f = open('model.pk', 'rb')
model = pickle.load(f)
def forecast(data):
    return model.predict(data)
```

3. Choose resources
4. Deploy!

Running model containers also have access to CDH for data lookups.



NEW IN CDSW 1.6

DATA SCIENTIST EDITOR PREFERENCES

One size does not fit all

Software engineering backgrounds

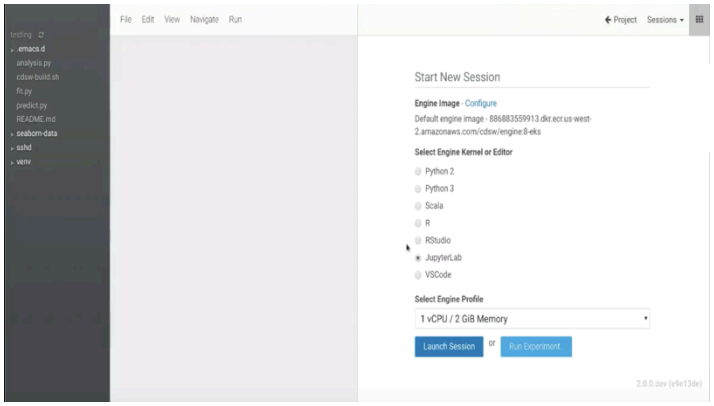
- Mostly favor IDEs e.g., PyCharm
  - Richness of features
  - Familiarity and personal preference

Other code-first data scientists

- Mostly favor Notebooks and RStudio
  - Interactivity of Notebooks
  - Familiarity and personal preference

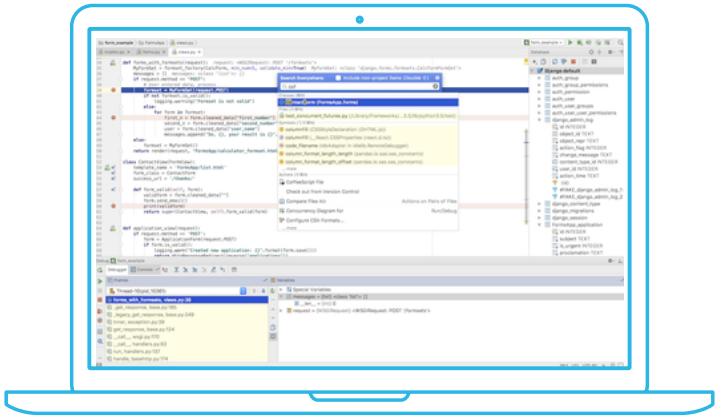
CDSW 1.6 THIRD PARTY EDITOR SUPPORT

Browser-based editors



- Popular editors (RStudio, JupyterLab)
- shipped as a docker image with CDSW
- Third-party editors are enabled within CDSW

Local editors



CDSW (Remote)

- Code sync with CDSW via Git
- Remote execution in CDSW

Lab time!

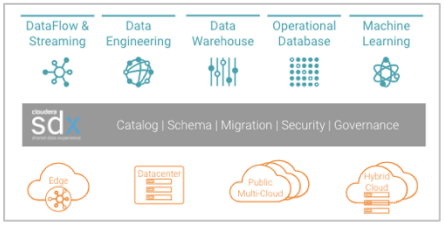
CLOUDERA

ML Services

Joon Kim, Solutions Engineer

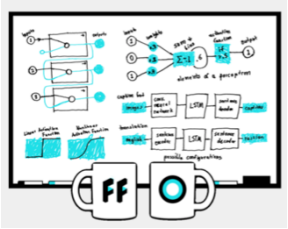
MACHINE LEARNING AT CLOUDERA

Our approach



Open **platform** to build, train, and deploy many scalable ML applications

Comprehensive data science **tools** to accelerate team productivity

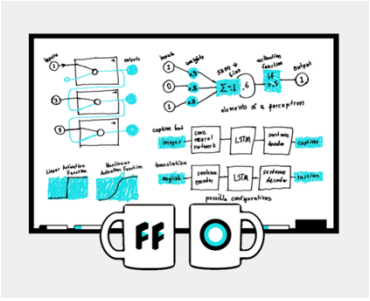


Expert guidance & services to fast track value & scale

CLOUDERA'S ML SERVICES

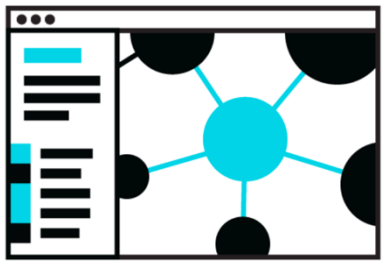
Expert guidance and technical services to accelerate value and scale

ML STRATEGY ENGAGEMENT



Delivers a **strategy prescription** focused on ML/AI

ML APPLICATION DEVELOPMENT



Evaluates and, if feasible, delivers an **ML application** (model, code, docs)

DEPLOYMENT, SCALE & MLOPS

Facilitates **industrialization** of an ML application

ADVISING & RESEARCH

Ongoing **ML expert advising** plus access to CFFL **research reports & prototypes**.

CLOUDERA'S ML SERVICES FOR THE JOURNEY TO INDUSTRIALIZED AI

STRATEGY / FOUNDATION → DEVELOP → DEPLOY → SCALE

PLATFORM & TOOLS

PLATFORM & TOOLS
EDC + CDSW/CML

EXPERT GUIDANCE

ML STRATEGY ENGAGEMENT
Strategy prescription for use cases, organizational and process design

ON DEMAND CDSW TRAINING

ADVISING & RESEARCH
White glove concierge service including ML expert & SW/platform advising plus access to CFFL research reports and prototypes.

RESIDENT DATA SCIENTIST
CFFL expert onsite with your teams 1 week/month

SERVICES FOR PRODUCTIO N & SCALE

ML PLATFORM ENABLEMENT
Deploy and enable CDSW in your environment per best practices

ML APP DEVELOPMENT
Identify, build, and document an ML app in your environment

ML APP DEPLOY
Deploy existing ML app to an existing Cloudera production environment

ML APP SCALE
Increase data and/or performance for existing production ML app

ML OPERATIONS
Apply operational tooling, automated processes, and best practices to maintain production ML models

EXPERT GUIDANCE

STRATEGY ENGAGEMENT

Strategy & roadmap to capitalize on ML opportunity for your business

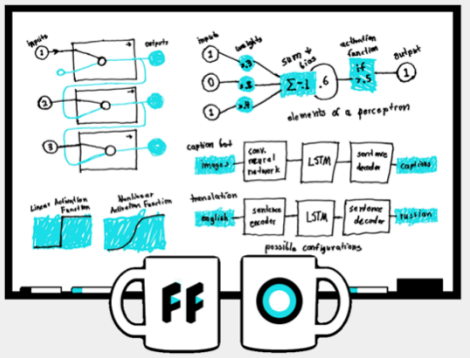
STRATEGY / FOUNDATION

Engagement

- Pre-onsite conversations to align on goals and resources (1-4 weeks)
- 2 Days w/ 2 CFFL team members on site info gathering with diverse stakeholders
- Follow up and delivery of a dynamic and effective plan (3-5 weeks)

Deliverables

- Machine learning strategy and tactical roadmap
- Optimized for your business opportunities and technical capabilities
- Recommendations across technology, product, people and process, metrics



CFFL ADVISING & RESEARCH

DEVELOP | DEPLOY | SCALE

White Glove Concierge Service

GOOD WHEN

Customers want ongoing guidance and assistance covering the breadth of Cloudera's software and services for ML.

TOPICS COVERED

- CFFL Research, Technology & Best Practices
- Cloudera platform + CDSW / CML
- Cloudera Professional Services including Platform Enablement & ML Ops

DELIVERABLES

- 4 hours/month remote advising
- Quarterly Research Reports and Prototypes
- Weekly Newsletter
- Ad-hoc blogs, code artifacts, helper libraries, prototypes and demos (delivered via CFFL Experiments portal)

SPOC Cloudera team member to connect customer's ML specific questions or problems to the right person or content within the entire ML portfolio

CFFL RESIDENT DATA SCIENTIST

DEVELOP | DEPLOY | SCALE

The customer's own data nerd best friend

GOOD WHEN

Customer has:

- Complex and diverse technical environment

- Nascent ML development

- A desire to put it all together

SAMPLE ACTIVITIES

On-site, flexible ML/AI support:

- Feature engineering
- ML model migration
- Knowledge transfer

ML Readiness and Operations Support

DELIVERABLES

- 1 week/mo on site

- Quarterly ML technical and process recommendations

- 480 Hours of on-site and remote time

- Quarterly ML Report

Connected customer ML ecosystem

SERVICES FOR PRODUCTION AND SCALE

ML PLATFORM ENABLEMENT

STRATEGY / FOUNDATION

Deploy and enable CDSW in customer's environment per best practices

Phase 1

Basic Setup [OPTIONAL]

- CDSW installation, testing and configuration

Phase 2

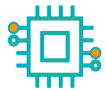
Platform Enablement

- Dependency management/ custom engines (customer-specific approach)
- Platform administration (monitoring, alerting, disaster recovery plan)
- Enable use of GPU; connectivity to other data sources through ODBC

Phase 3

End-user Access

- Knowledge transfer with data scientists tailored to customer's ecosystem
- Migrate existing projects and support effective use of the tool



- Platform team enabled to deliver an effective data science environment

- Data scientists enabled to use the tool and platform in customer's unique ecosystem

ML APPLICATION DEVELOPMENT

The fast track for a new machine learning capability

DEVELOP

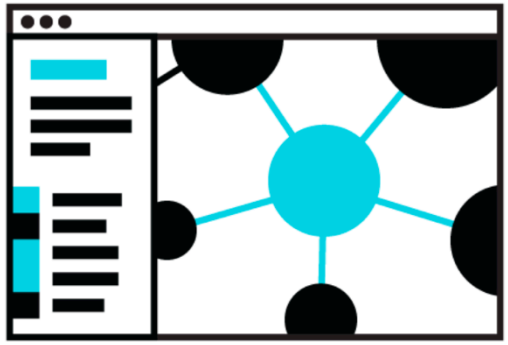
Engagement

Close collaboration with your teams across sequential development phases (2-6 months total):

- Exploration phase (~2 weeks)
- Algorithmic excellence phase (a few weeks to a couple of months)
- Operationalization phase (a few weeks to a couple of months)

Deliverables

- Working Prototype delivered in your environment
- Code & extensive documentation
- Optional ongoing support for production implementation
- Phased development off-ramps for feasibility concerns



ML APP DEPLOY

DEPLOY

Deploy an existing ML app to an existing Cloudera production environment

GOOD WHEN

Customer has a prototype/experiment in non-production environment and wants it implemented with the same functionality in a different environment.

NOTE: **Not a massively parallel distributed app**

SAMPLE ACTIVITIES

- Mapping technical capability to use case requirements
- Architecture design and implementation
- Ensure production-readiness of code
- Port code and/or model to production environment
- Blueprint for model and code maintenance going forward

DELIVERABLES

- Services hours
- Architecture design document
- Model maintenance recommendations document
- Code assets

- Existing application deployed in the customer's environment

ML APP SCALE

SCALE

Adapt an existing ML app for performance at scale

GOOD WHEN

Customer has a production application and wants to increase performance using any combination of distributed ML and HW/SW optimization for parallelization.

SAMPLE ACTIVITIES

- Mapping technical capability to use case requirements
- Architecture design and implementation
- As required, any or all of:
    - Code optimization
    - Parallelization
    - Reimplementation within alternate language/framework

DELIVERABLES

- Services hours
- Architecture design document
- Model maintenance recommendations document
- Code assets

- Customer's existing application adapted to perform at scale

ML OPERATIONS

DEPLOY SCALE

Apply tools, processes and best practices for maintaining models in production

GOOD WHEN

Customer needs help with best practices, processes and tooling for maintaining models in production, either in context of specific projects/ teams or wider enterprise data science community best practices.

SAMPLE ACTIVITIES

Implement best practices for:

- maintainable iterative models
- model monitoring
- MLDLC (inc. deployment/ promotion)
- DS/ML collaboration
- workflow monitoring/ scheduling
- model factories

DELIVERABLES

- Services hours
- Documented customer ecosystem assessment and recommendations

(to be updated quarterly for long-running engagements)

- Model ops strategy/framework in place and/or

- Successful model ops approach implemented for one or more customer use cases

CML
- ML PaaS on Cloud

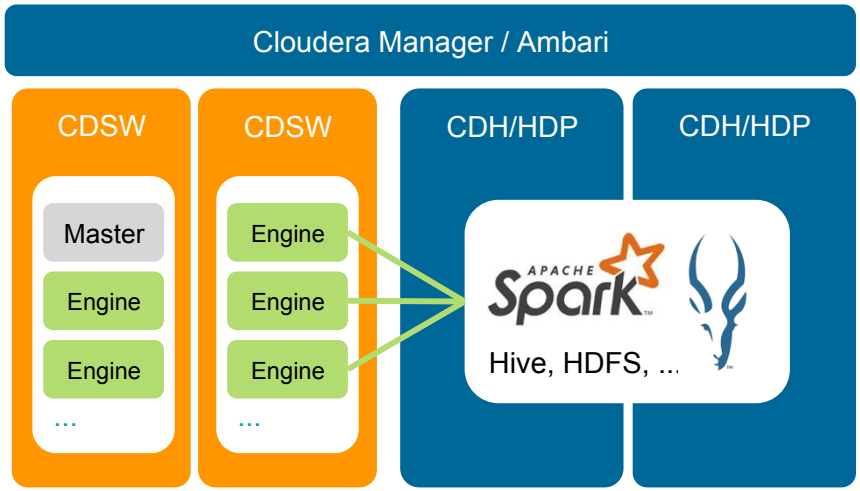
Machine Learning on the Cloudera Data Platform (CDP)

Cloud-native enterprise machine learning service

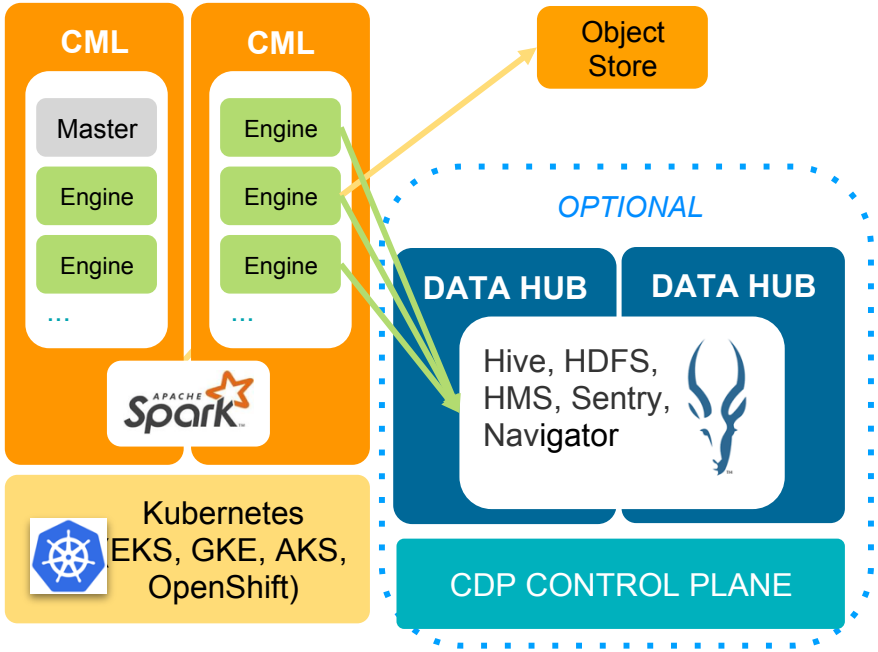
- **DS/ML from research to production**
  - Same end-to-end workflows as CDSW
- **Seamless scale-out experience**
  - Rapid provisioning and elastic autoscaling
  - Automatic dependency management
  - Distributed CPU and GPU model training
- **Managed service on CDP**
  - No (Spark) clusters to manage
  - Containerized multi-cloud portability
  - Private cloud option with CDP-Private

SPARK ON KUBERNETES

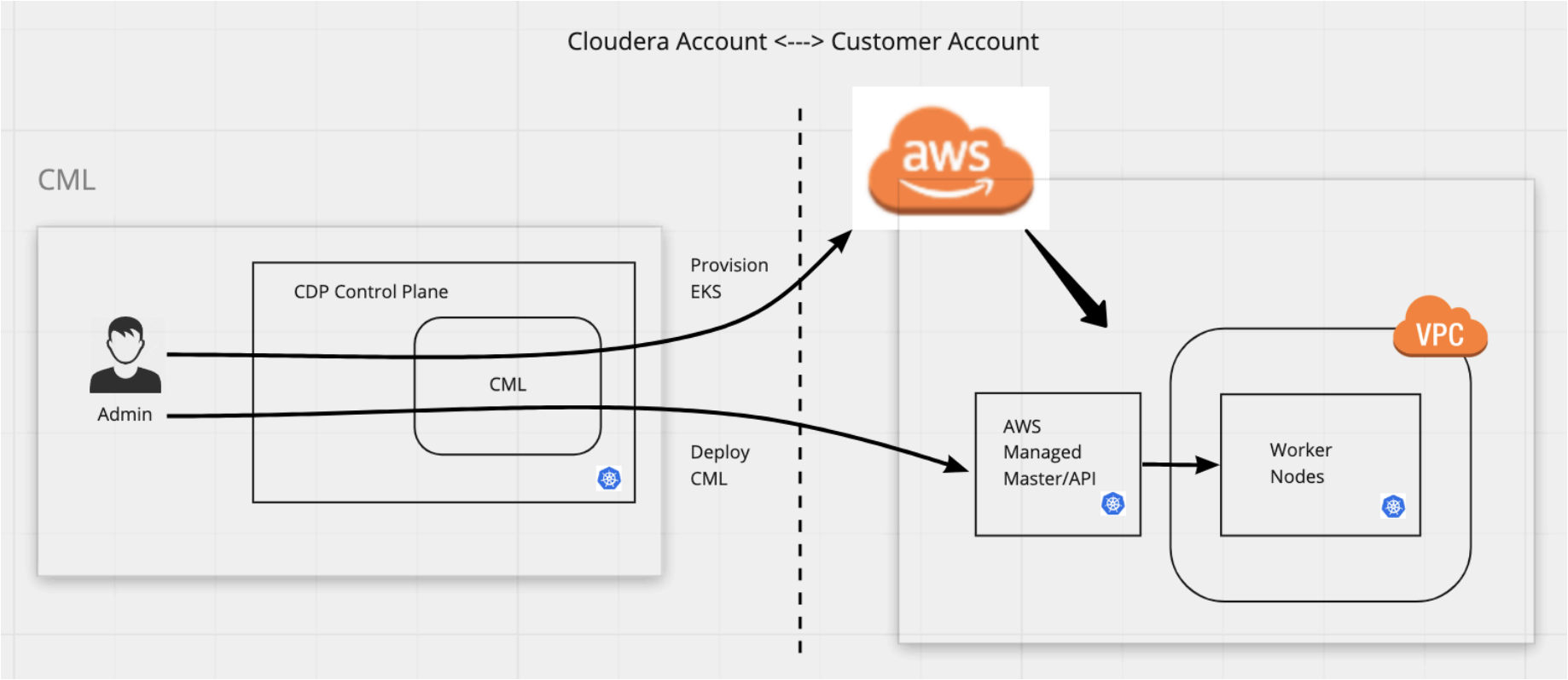
EXPERIMENTATION WITH SCALE



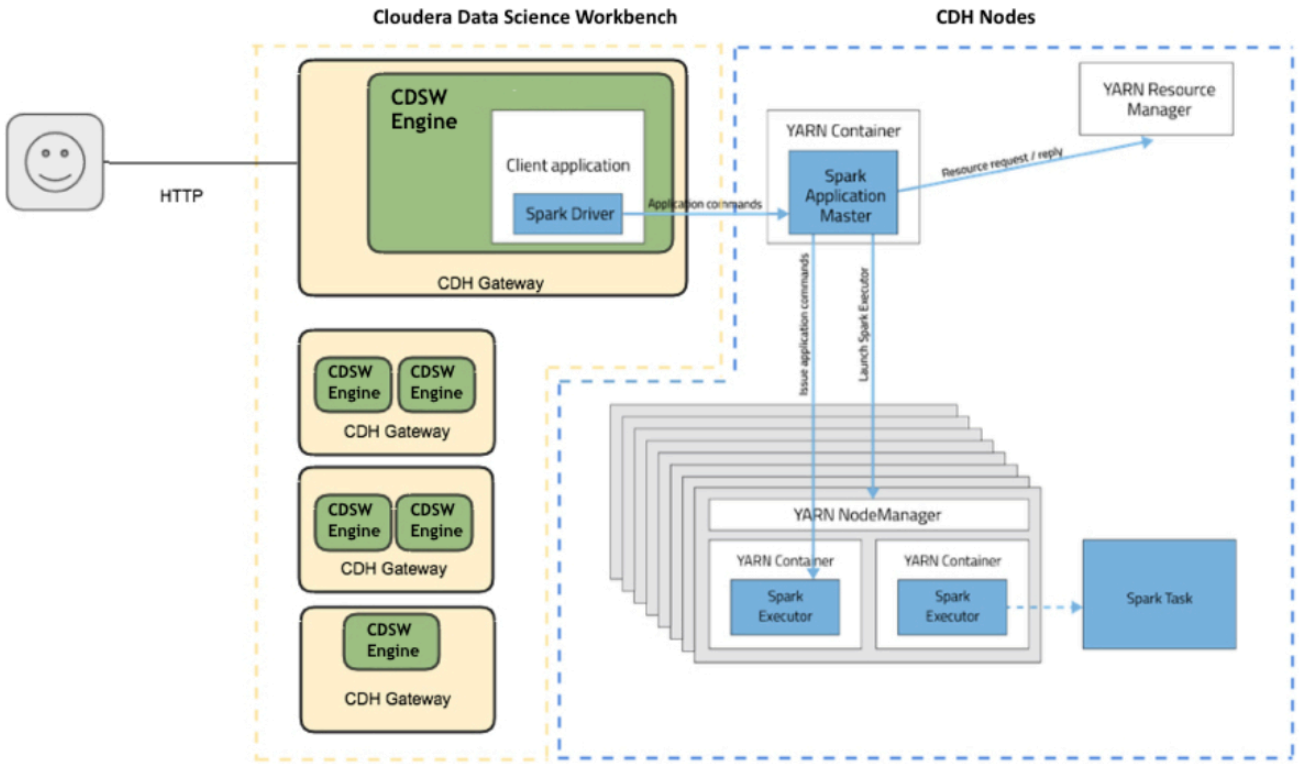
Requires and extends CDH/HDP, pushing distributed compute to the cluster



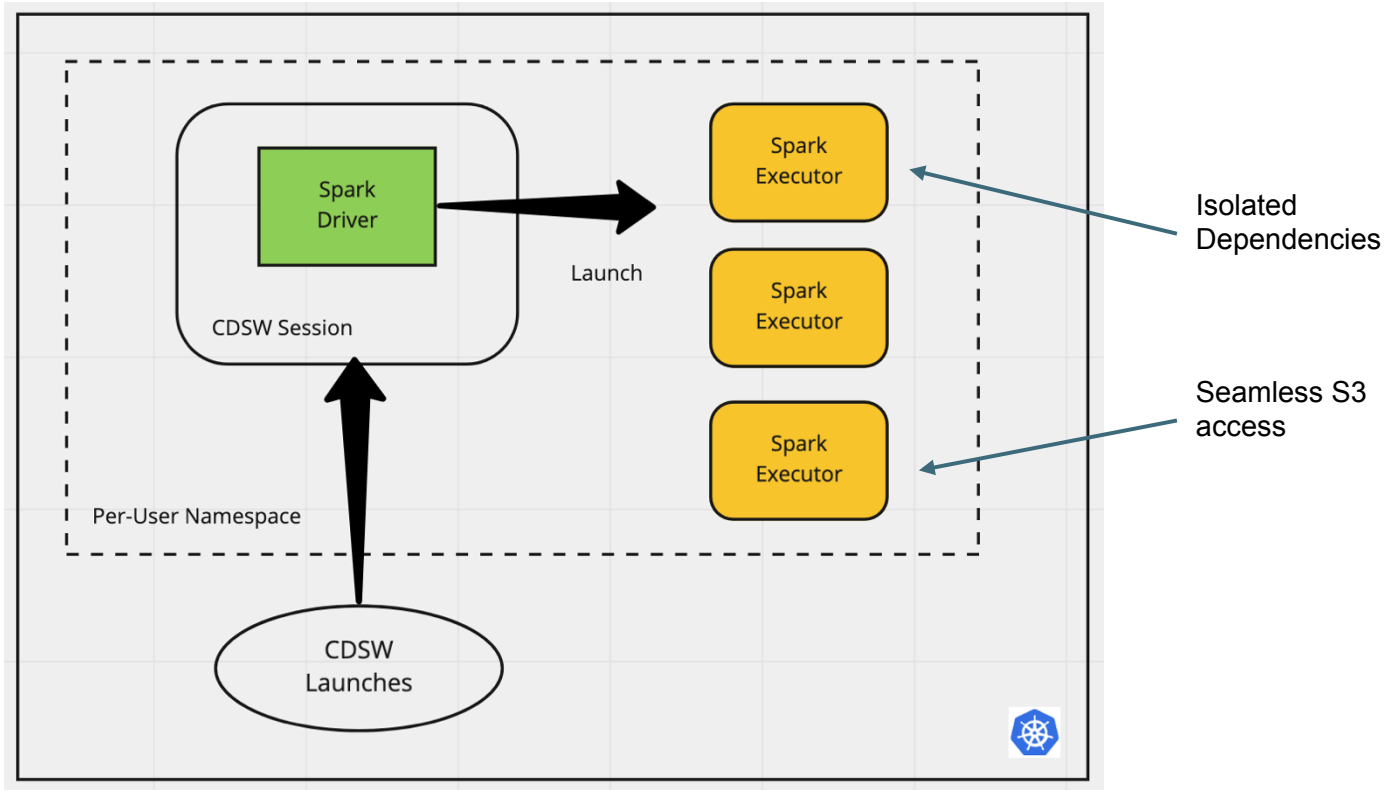
Self-contained and manages own distributed compute; can optionally use CDH/HDP


Cloudera Account <---> Customer Account
CML
CDP Control Plane
CML
Admin
Provision
EKS
Deploy
CML
aws
VPC
AWS
Managed
Master/API
Worker
Nodes

CDSW with Spark on YARN



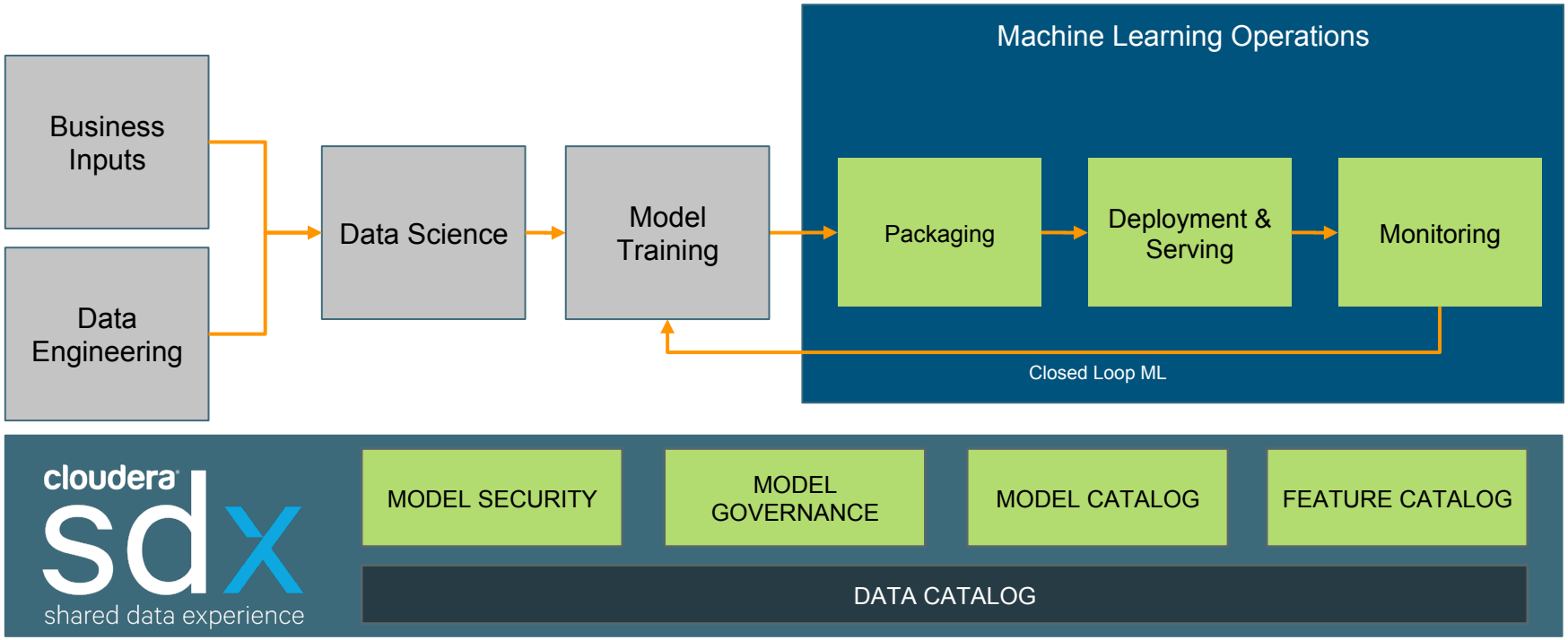
CML - Spark on Kubernetes



ML Production

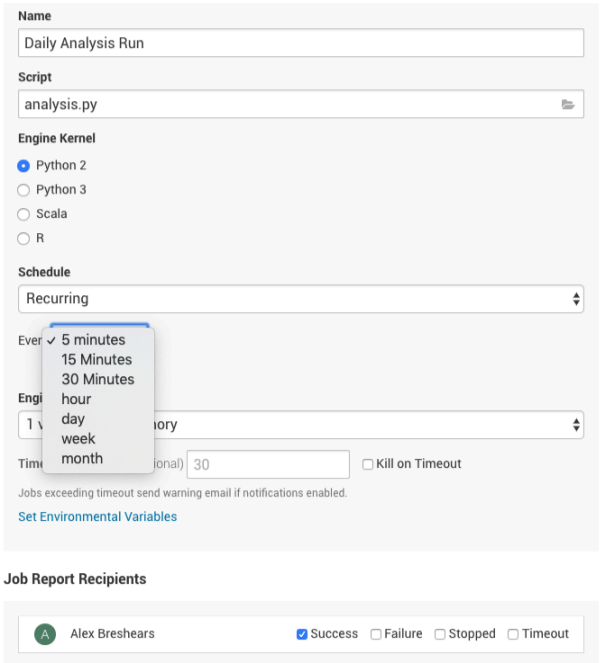
- Product Roadmap

Production ML Workflow



Using CDSW + Hadoop for Serving & Monitoring

Schedule reports & scoring to run on a periodic basis



- Execute arbitrary scripts
- Schedule on a recurring basis
- Create dependencies on other jobs for complex pipelines
- Allow output to be sent via email to recipients

FaaS Models

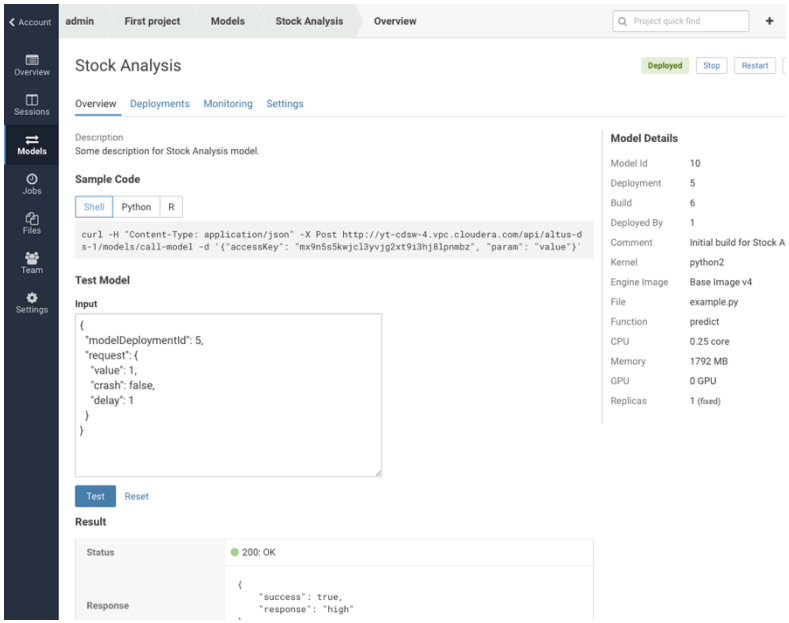
Machine learning models as one-click microservices (REST APIs)

1. Choose file, e.g. `score.py`
2. Choose function, e.g. `forecast`

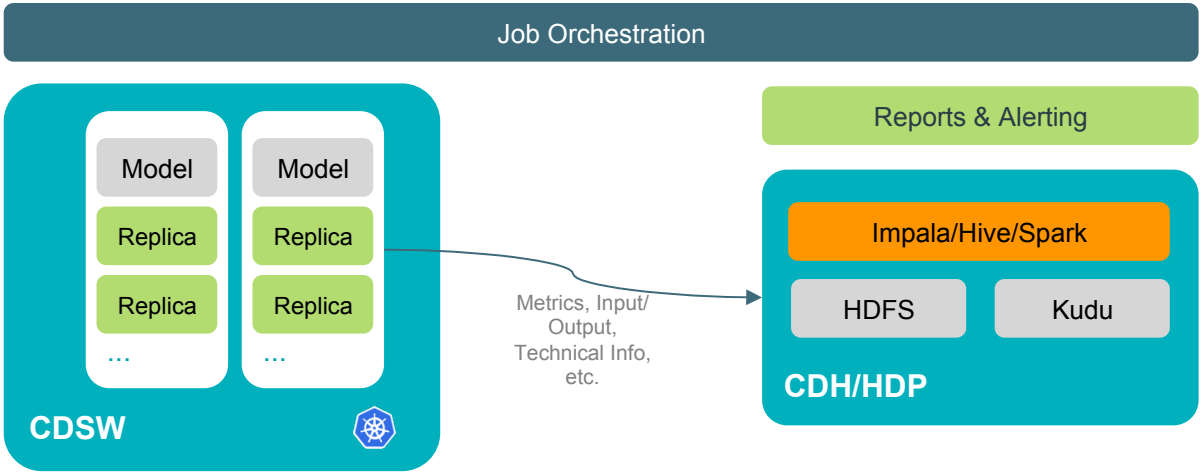
```
f = open('model.pk', 'rb')
model = pickle.load(f)
def forecast(data):
    return model.predict(data)
```

3. Choose resources
4. Deploy!

Running model containers also have access to CDH/HDP for data lookups.

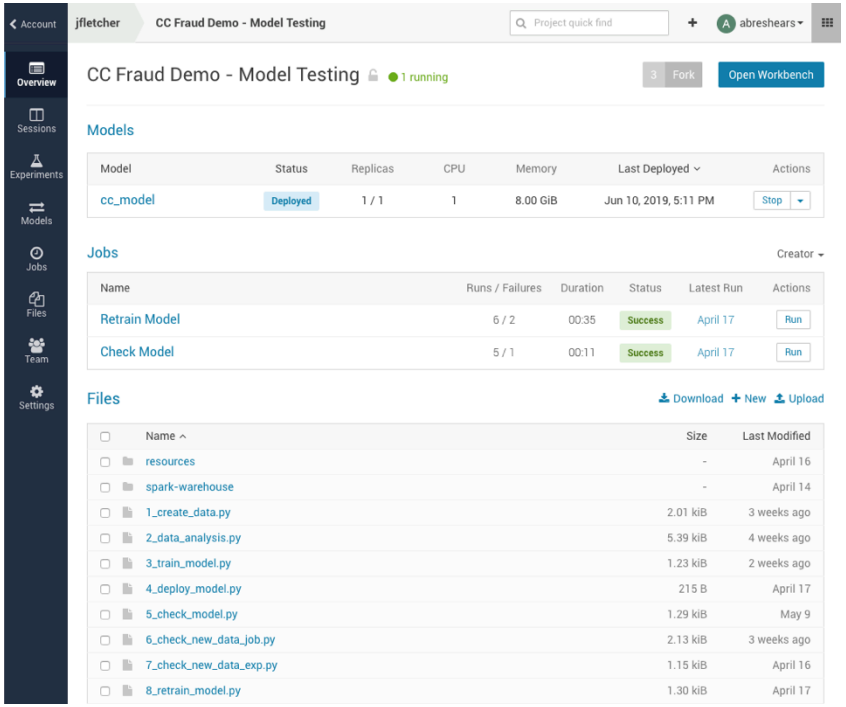


Monitoring architecture using CDSW + CDH/HDP



Schedule reports & monitoring to run on a periodic basis

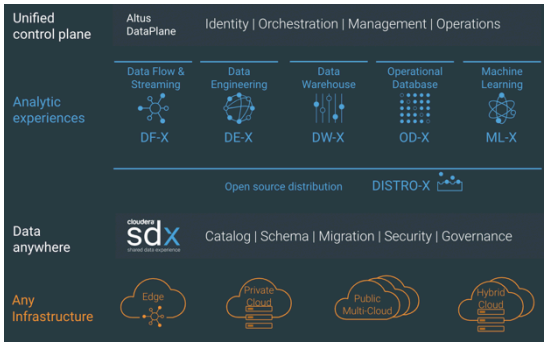
- Check input and output distribution
- Look for drift & accuracy changes
- Add custom thresholds
- Send emails with results
- Use custom code to send to pager systems



Demo:
https://www.cloudera.com/content/dam/www/marketing/resources/webinars/cdsw-office-hours-series-part2.png.landing.html

aka MLOps

Enable data scientists
Teams of developers
Open source
Enterprise ready



On a common platform
End to end workflows
Shared data experience
Cloud portability

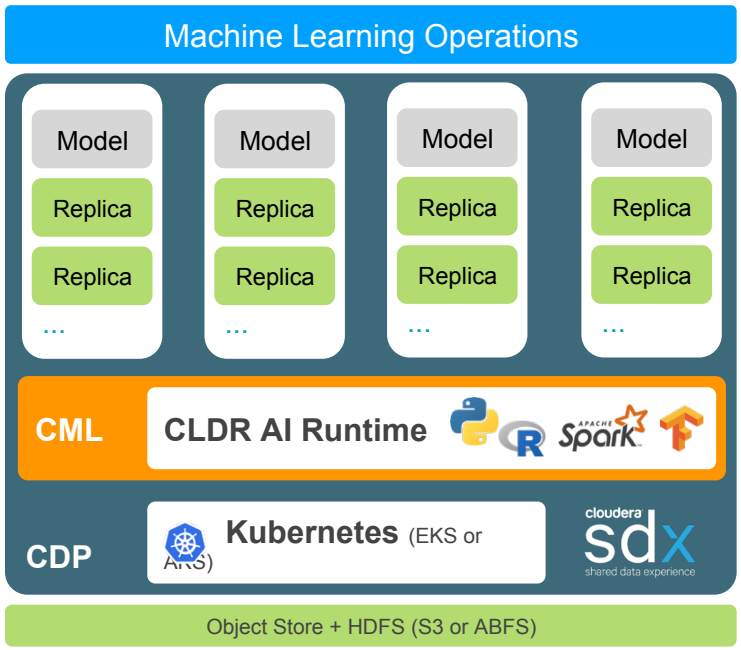
For the real world
Models in production
Model lifecycle management
Autonomous edge

MLOps Product Goals

- Plant technical flags
- Understand and communicate the problems

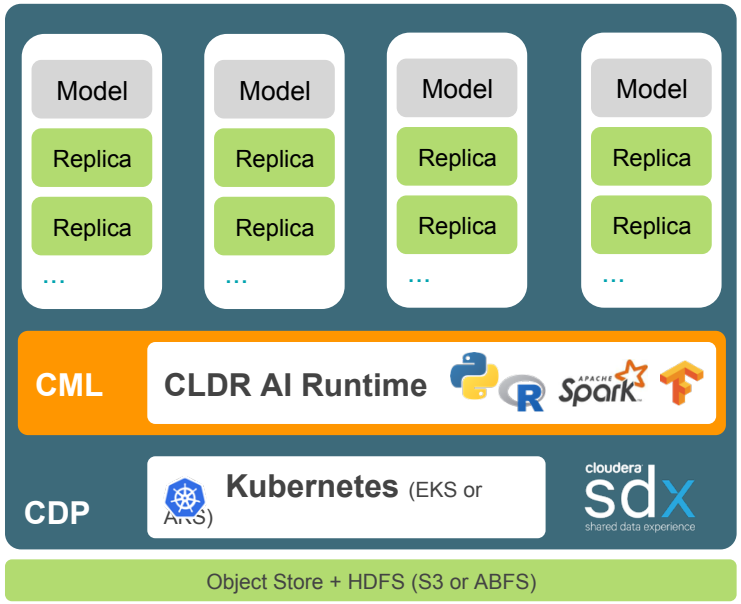
- True value of ML is production
- Production is much stickier

- CDP first: Iterative development
- Get early adoption as proof points



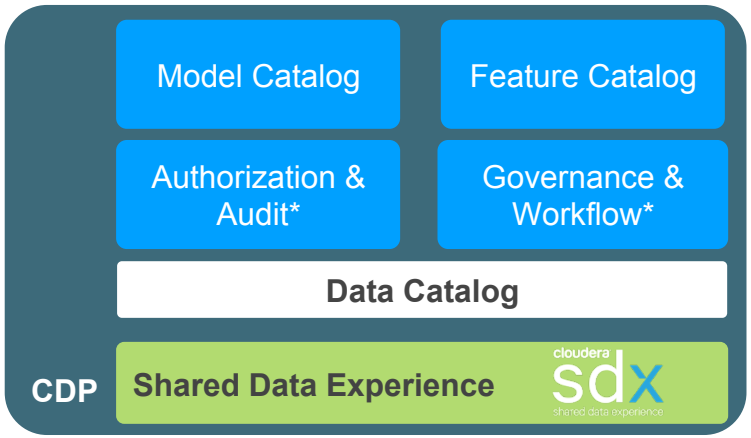
- **Centralized Monitoring Solution**
  - Single pane of glass for all models
  - Alerting and external system integration
- **Track technical metrics**
  - Uptime
  - Status
  - Response time & SLA adherence
- **Track mathematical & functional metrics**
  - E.g. prediction distribution, drift, input distribution
  - Customizable to model
  - Automatic Anomaly Detection

Packaging, deployment, and serving models



- **Automated Deployments**
  - Complete last mile of ML deployments
- **Address Multiple Deployment Patterns**
  - Batch
  - Function as a Service
  - Streaming
  - Edge
- **Enterprise Capabilities**
  - Highly Available
  - Autoscaling
  - Secure by default (Oauth)

- **Centralized Catalog**
  - Track all models along their lifecycle
  - Track all features their lifecycle
  - Understand features and their relationship with the data catalog (i.e. lineage)
- **Authorization & Audit**
  - Protect models*
  - Track access*
- **Governance & Workflow**
  - Approvals*
  - Workflows integrated with Monitoring*



* Post MVP Items

THANK YOU!

CLOUDERA