



Cloudera Machine Learning (CML)

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter



POWERED BY **CLOUDERA**
SDX

Security | Governance | Lineage | Management | Automation

CONSISTENT SECURITY AND GOVERNANCE

Built for multi-functional analytics anywhere



- **Data Catalog:** a comprehensive catalog of all data sets, spanning on-premises, cloud object stores, structured, unstructured, and semi-structured
- **Schema:** automatic capture and storage of any and all schema and metadata definitions as they are used and created by platform workloads
- **Replication:** deliver data as well as data policies there where the enterprise needs to work, with complete consistency and security
- **Security:** role-based access control applied consistently across the platform. Includes full stack encryption and key management
- **Governance:** enterprise-grade auditing, lineage, and governance capabilities applied across the platform with rich extensibility for partner integrations

CML



THE ENTERPRISE DATA CLOUD COMPANY

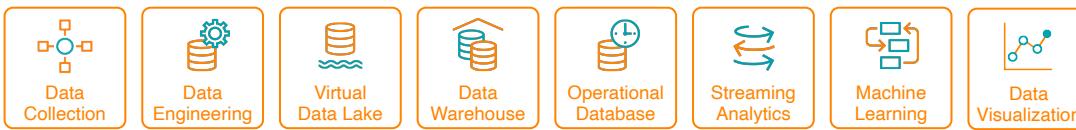
A HYBRID DATA CLOUD *AND* ANALYTICS FOR THE DATA LIFECYCLE

Storage & compute separation
Containers
SDX

Delivered as data
architectures for key
workload patterns

Real-time
Batch
Structured
Unstructured

**Data
Sources**



Analysts
Engineers
Scientists
Developers

**Data
Users**

CML is a data service for creating and maintaining ML projects from code to production. Elasticity and collaboration are enabled through a **containerized architecture** and **shared workspaces**.

Functionality is offered to conduct experiments, promote and govern models, and serve applications.

Some Examples of AI and ML

Logistic Regression

- Predicting customer churn for subscription-based services from content providers

Neural Networks

- Image and obstruction detection in self-driving cars

Survival Analysis

- Predictive maintenance with IoT analytics in the manufacturing industry

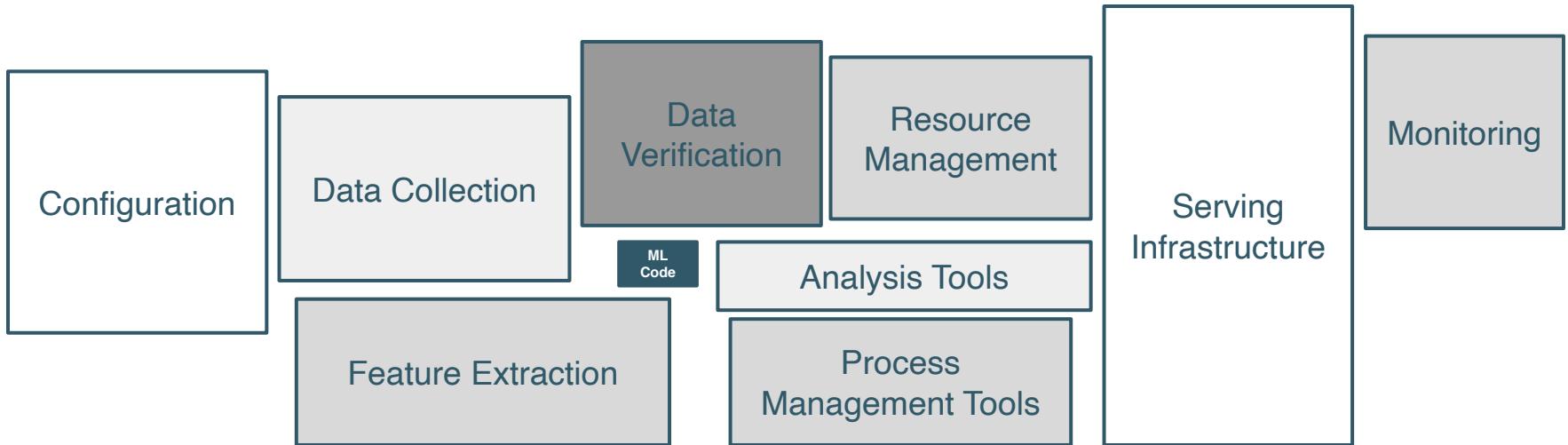
Clustering

- Customer segmentation in the retail industry for targeted, marketing campaigns

NLP

- Sentiment analysis to react quickly to call center and support issues

HIDDEN TECHNICAL DEBT IN MACHINE LEARNING SYSTEMS



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

MOVING FROM EXPLORATION TO OPERATIONALIZATION

Production machine learning at scale



FROM THE LAB...



TO THE FACTORY

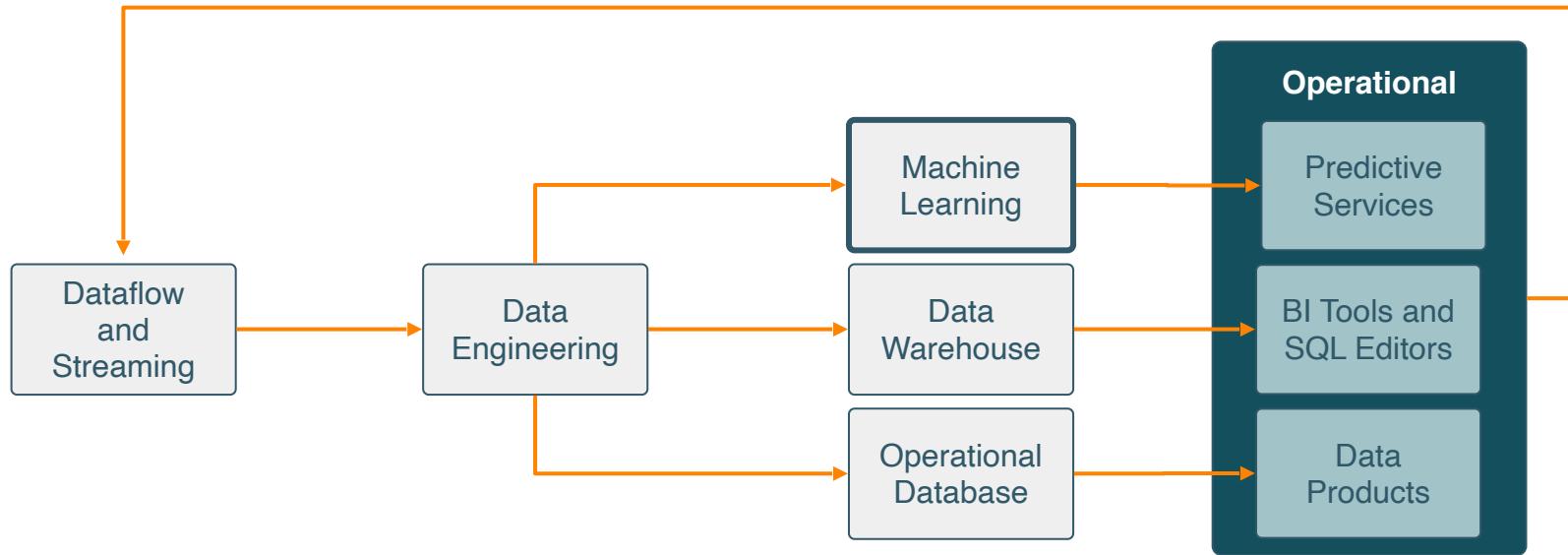
35%

Making it to production

Currently only 35% of organizations indicate that analytical models are fully deployed in production and are often challenged in the “last mile” of the complex and iterative ML workflow

** IDC's Advanced and Predictive Analytics survey and interviews, n = 400, 2017 – 2019*

... IS MORE THAN JUST ML



**CLOUDERA
SDX**

Model Security

Model Governance

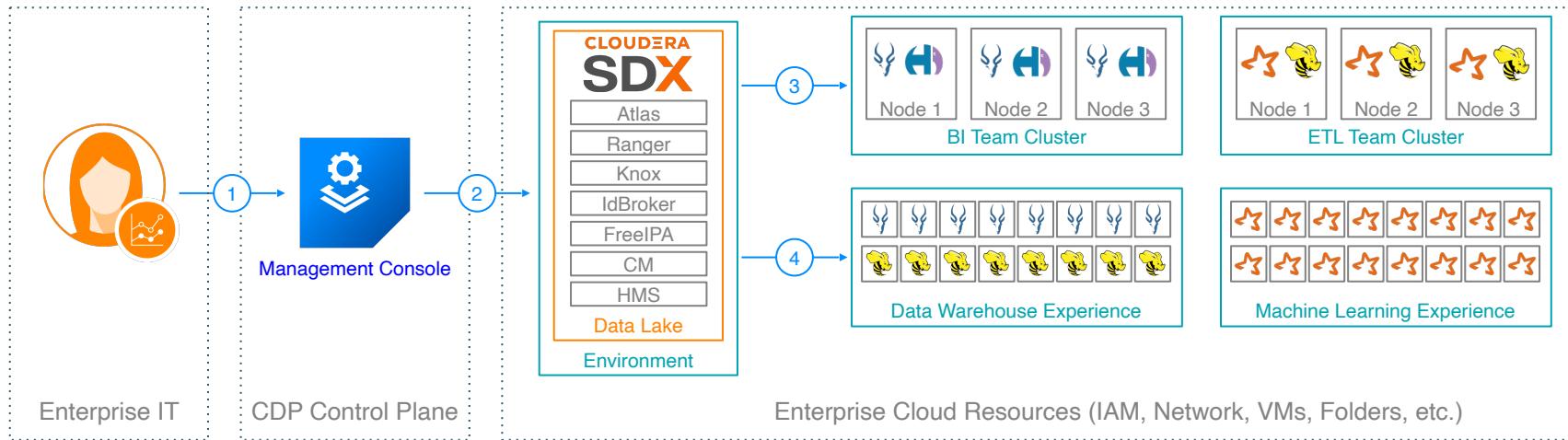
Model Catalog

Feature Store

Data Catalog

EXAMPLE INTEGRATION OF COMPONENTS

Typical user flow



Step 1
User connects to
CDP with their
enterprise identity

Step 2
They create an
environment and
data lake for their
enterprise

Step 3
They create data hub
clusters for traditional
workloads

Step 4
They create access
points for containerized
analytic experiences

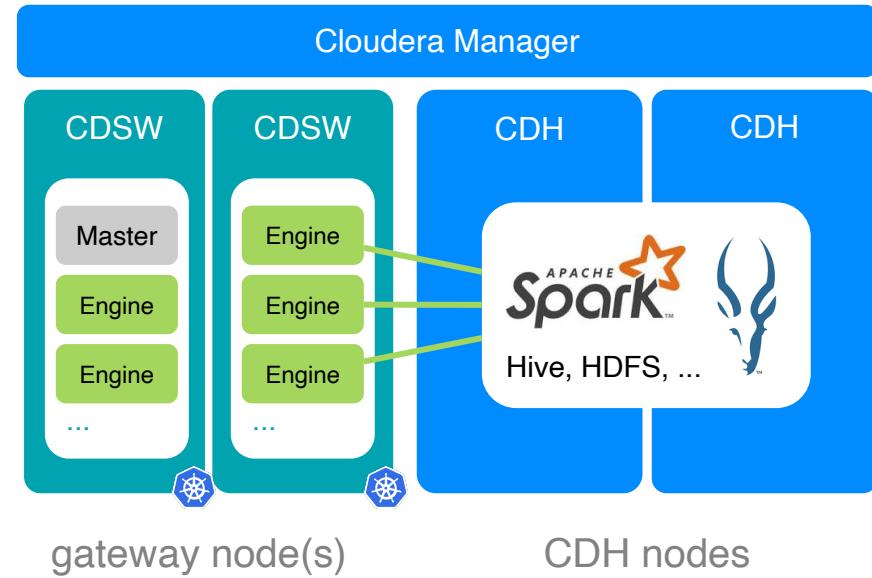
TECHNICAL FEATURES

CDSW IS DESIGNED FOR THE DATACENTER

Containerized environments with scalable compute powered by Kubernetes and CDH

- Easy addition for self-service data science
- Seamless connectivity to existing clusters
- Isolated, reproducible user environments

But... *Trade-offs...*



CML IS CLOUD READY

Cloud-native enterprise machine learning as-a-service

END-TO-END DS/
ML



Accelerate Data
Science from research
to production

SECURE &
GOVERNED



Self-service data &
elastic compute with
guardrails

FLEXIBLE DATA
ENGINEERING



Easy to use,
self-service, on-
demand, elastic spark
for data engineering

KUBERNETES
& CONTAINERS



Adoption as standard
operating environment
for flexibility and agility

CDP Management Console

ML Workspace

Engine Engine ...



ML Workspace

Engine Engine ...



autoscaling resource groups



Agility

Collaboration

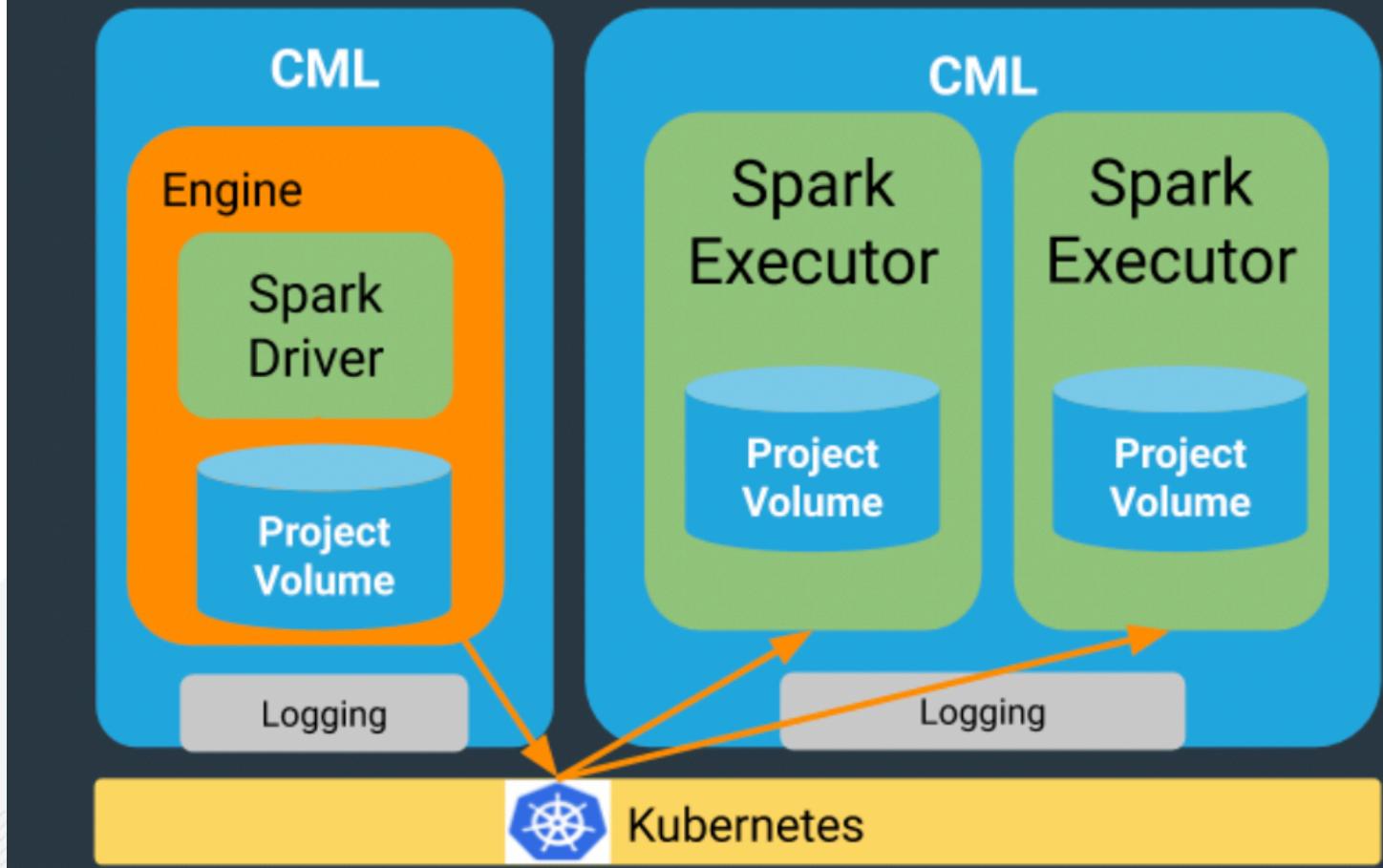
Dependency Management

Auto-scaling

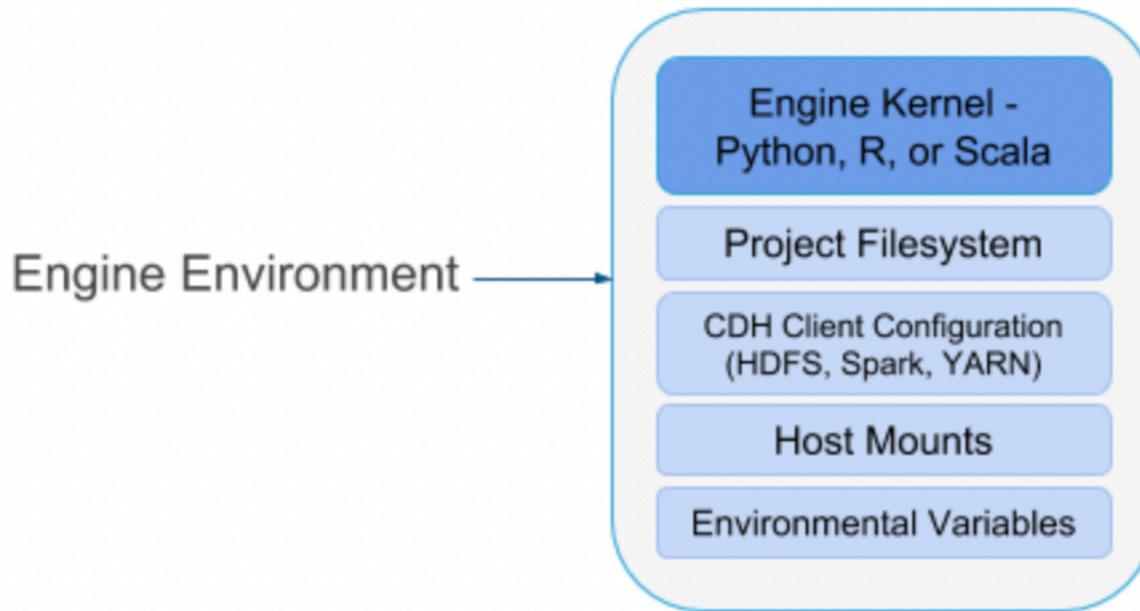
Workload Isolation

Versioning

SPARK on KUBERNETES



CML Session and Jobs use a virtual ‘engine’



KEY COMPONENTS

The screenshot shows the Cloudera Management Console interface. On the left is a dark sidebar with navigation links: Dashboard, Environments (which is selected and highlighted in blue), Data Lakes, User Management, Data Hub Clusters, Data Warehouses, ML Workspaces, and Classic Clusters. The main content area is titled "Environments / List". It displays a table with 10 environments listed. The columns are "Status" (checkbox), "Name", and "Cloud Provider" (aws icon). The environments are:

Status	Name	Cloud Provider
Available	rr-env-burst	aws
Deletion Failed	raman-env1	aws
Available	kat-19	aws
Available	all-se-demo-new	aws
Available	partners-ab-03	aws
Deletion Failed	all-se-demo	aws
Available	sv-wwbank-demo-5	aws
Deletion Failed	cloud-workspace1	aws

ENVIRONMENTS



1:1



1:
N

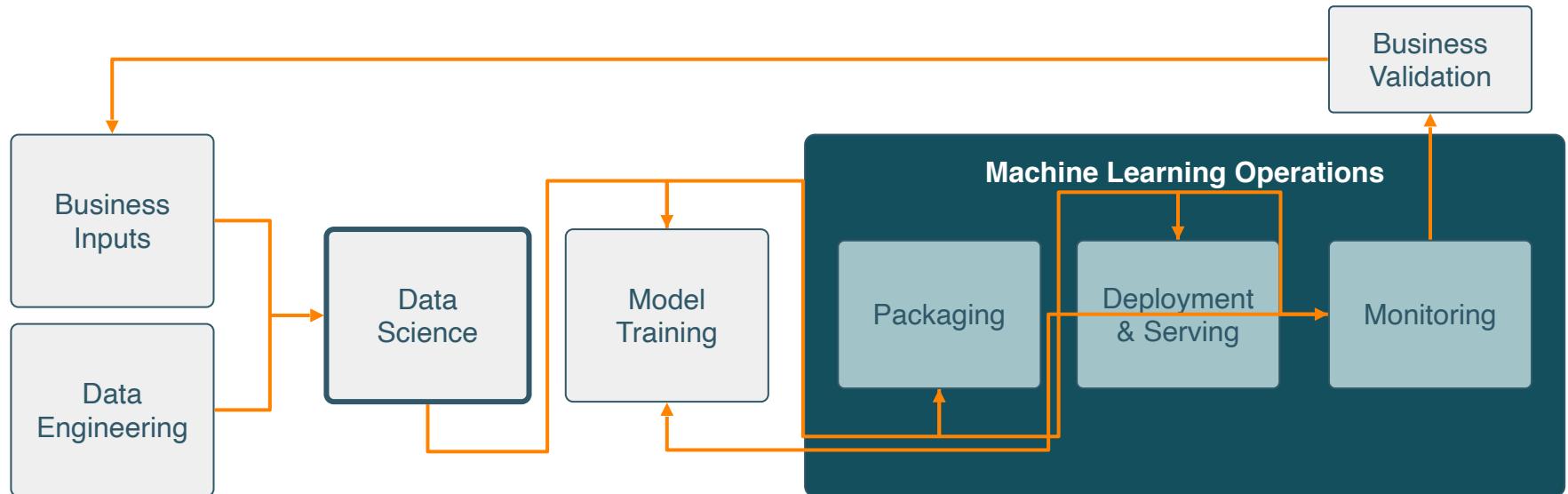


- 1 Template
- 1 Region
- 1 VNET
- Multiple Roles/Folders

- SDX: Atlas, Ranger, Knox, IdBroker, CM
- Associated with groups/users

- DH templates
- ML Env
- DW Database Catalogs/Virtual Compute

MACHINE LEARNING IN PRODUCTION



Business, IT - Data
Engineering



Data Scientists



IT - DevOps



ML Ops /
ML Engineer



Business

CML ENABLEMENT

Accelerate Machine Learning Projects from Research to Production

Explore and Analyze
Data

Deploy Automated
Pipelines

Train and
Evaluate Models

Deploy Models

Workbench
(Interactive)

Jobs
(Batch, non-versioned)

Experiments
(Batch, versioned)

Models
(REST API)

Hyper-parameter tuning

Immutable artifact
Application serving

Demo

Demo: Predict the number of flight cancellations based upon historical data dating back to 2003. Perform time-series analysis to experiment with ARIMA models. Conduct feature extraction and transformations with data persisted in S3.

Demo: ML Lifecycle



- Extract
- Explore
- Visualize
- Transform
- Propose

- Create container with Python dependencies

- Test hyperparameters



REST API



Amazon S3

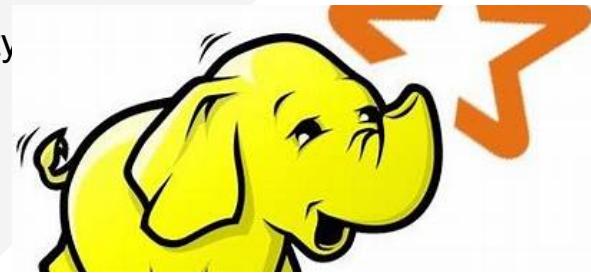


What is a Time Series

- **Time series = a sequence of data points occurring in order as measured over time**
- **Time series prediction = using previous values to predict future values**
- Stationarity = constant mean, constant variance, and no seasonality
- Seasonality = patterns and trends that repeat over time

How to check for stationarity:

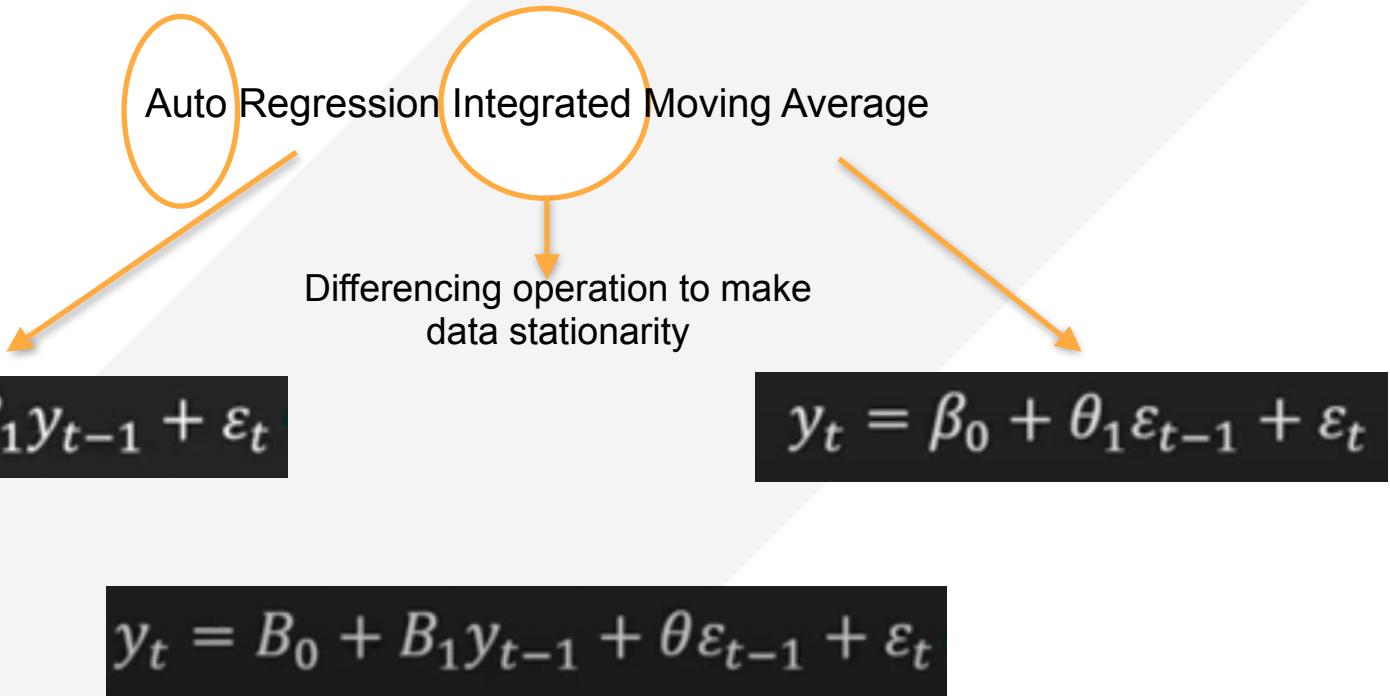
- Plotting
- Local versus global means checks
- Kwiatkowski-Phillips-Schmidt-Shin (KPPS) Test
- Augmented Dickey-Fuller Test (ADG)



Data Engineering exercise

Exponential smoothing $Y = \text{Log}(\text{Exp}(Y))$
Seasonal differencing $Y(t) - Y(t-N)$

What is an ARIMA Model



Converge to the best ARIMA Model

Find the optimal value for the mean and variance
given a number of observed measurements

Balance between maximizing the log likelihood
versus overfitting

L = log likelihood
 K = number of parameters
 N = number of samples used for fitting

Model quality => Lower values of AIC and BIC

$$AIC = 2K - 2L$$

$$BIC = K * \ln(N) - 2L$$

Assess Model Performance

Compare model predictions to actual data

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^T (\hat{y}_t - y_t)^2}{T}}.$$

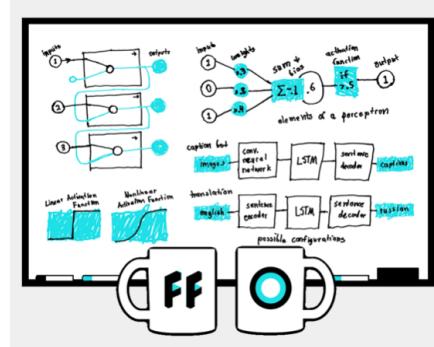
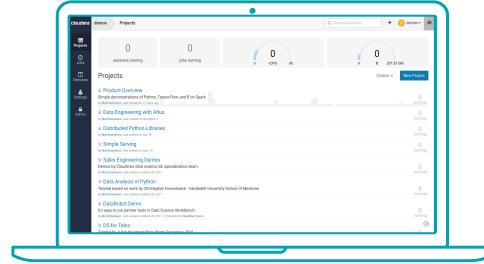
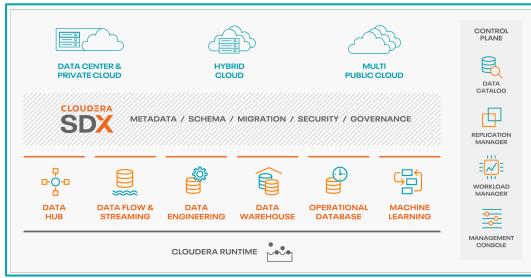
Example Airlines Data

Airport.Code	Airport.Name	Time.Label	Time.Month	Time.Month Name	Time.Year	Statistics.# of Delays.Carrier	Statistics.# of I
ATL	Atlanta, GA: Hartsfield-Jackson Atlanta International	2003/06	6	June	2003	1009	
BOS	Boston, MA: Logan International	2003/06	6	June	2003	374	
BWI	Baltimore, MD: Baltimore/Washington International Thurgood Marshall	2003/06	6	June	2003	296	
CLT	Charlotte, NC: Charlotte Douglas International	2003/06	6	June	2003	300	
DCA	Washington, DC: Ronald Reagan Washington National	2003/06	6	June	2003	283	
DEN	Denver, CO: Denver International	2003/06	6	June	2003	516	
DFW	Dallas/Fort Worth, TX: Dallas/Fort Worth International	2003/06	6	June	2003	986	
DTW	Detroit, MI: Detroit Metro Wayne County	2003/06	6	June	2003	376	
EWR	Newark, NJ: Newark Liberty International	2003/06	6	June	2003	322	
FLL	Fort Lauderdale, FL: Fort Lauderdale-Hollywood International	2003/06	6	June	2003	247	
IAD	Washington, DC: Washington Dulles International	2003/06	6	June	2003	320	
IAH	Houston, TX: George Bush Intercontinental/Houston	2003/06	6	June	2003	329	
JFK	New York, NY: John F. Kennedy International	2003/06	6	June	2003	376	
LAS	Las Vegas, NV: McCarran International	2003/06	6	June	2003	511	

Closing Notes

MACHINE LEARNING AT CLOUDERA

Build your enterprise AI Factory

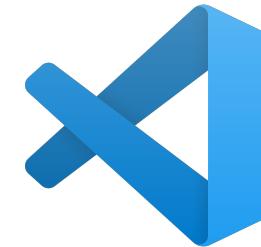


Open **platform** to build,
train, deploy and
operationalize ML
models at scale

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

Expert guidance &
services to fast track
value & scale

Project development can be done in the editor of your choice



PyCharm



Data scientist productivity

CML is open and extensible *by design*

auto-sklearn



TransmogrifAI

automl-gs



App development can be done in the framework of your choice



NEW

APPLIED MACHINE LEARNING PROTOTYPES

- Complete auto-deployed ML projects
- Pre-built algorithms & ML models OOTB
- Auto-deployed predictive apps
- Build and deploy your own AMPs
- Fully built in to CML

The screenshot displays the Cloudera Machine Learning interface. On the left, a dark sidebar menu lists various options: Projects, Sessions, Experiments, Models, Jobs, Applications, Settings, and **Prototype Catalog**, which is highlighted with a red box. Below the sidebar are two buttons: "Get Started" and "Help". The main area is titled "Prototype Catalog" and features a search bar and a "Tags: all" dropdown. It contains several prototype cards, each with a thumbnail, title, and tags. The prototypes shown are:

- Churn Modeling with XGBoost**: EXPLAINABILITY, XGBOOST. Includes a heatmap and a line chart.
- Deep Learning for Image Analysis**: COMPUTER VISION, IMAGES. Includes a heatmap and a grid of image thumbnails.
- Structural Time Series**: TIME SERIES. Includes a line chart.
- Deep Learning for Anomaly Detection**: ANOMALY DETECTION, DEEP NEURAL NETS. Includes a heatmap and a table.
- Fraud Detection**: FRAUD DETECTION, ANOMALY DETECTION. Includes a heatmap and a table.
- A preview section on the right shows a "pedestrian" detection interface with sliders for "How many pedestrians (select a range?)" and "Choose a frame (index)", and a confidence threshold slider.

THANK YOU