CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

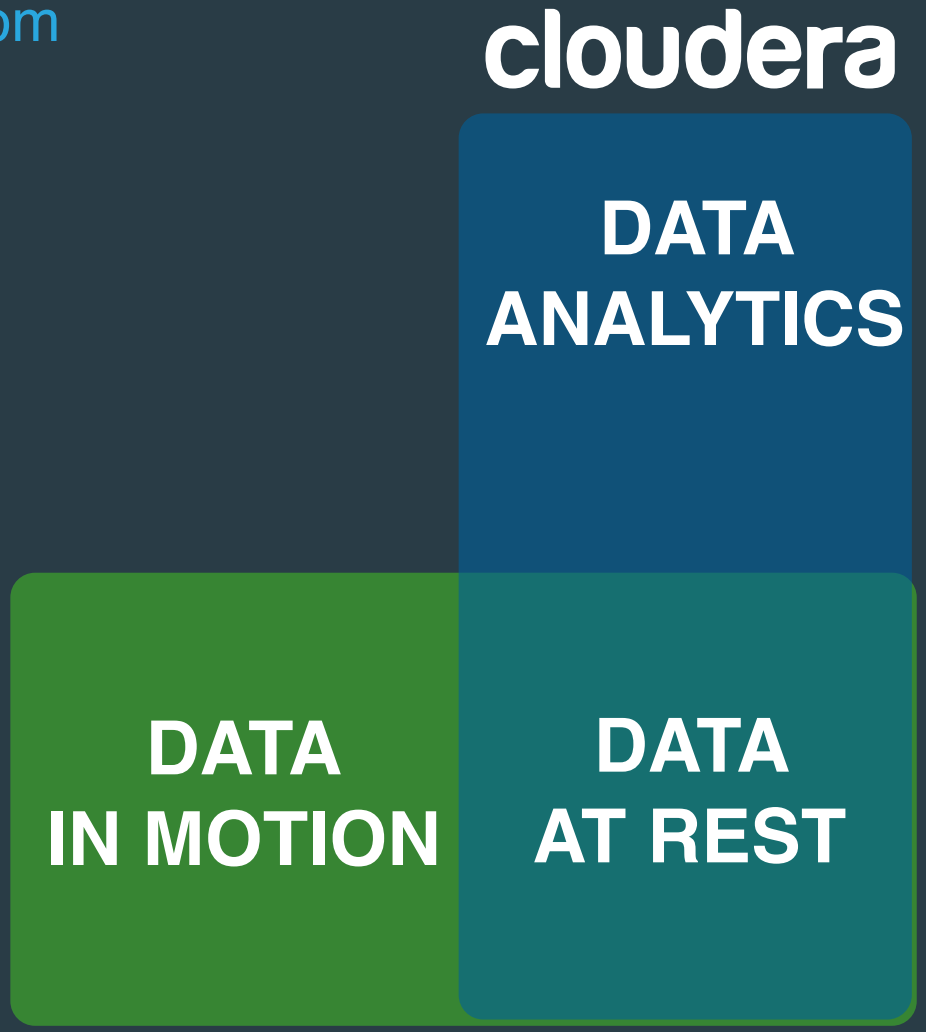
Edge2AI

Johannes Muselaers I Solutions Engineer Nordics
jmuselaers@cloudera.com
+46725881091

PDF of presentation:
https://github.com/jmuselaers/HI-IS

COMBINED STRENGTH

Where we come from



HORTONWORKS

NEW COMPANY

Cloudera at a glance

2000+

Customers across all markets

Meet our data heroes

3000+

Solution and service partners

Find a partner

3000+

Employees doing business in 28 countries

Join us

NEW COMPANY LOGO

OPPORTUNITY

"AI, IoT and Augmented Reality are about to reinvent how industries utilize data—and could drive an era of productivity growth in our cities, farms, factories and hospitals."

Morgan Stanley Research

REALITY

91%

ORGANIZATIONS STRUGGLE TO REACH DATA MATURITY

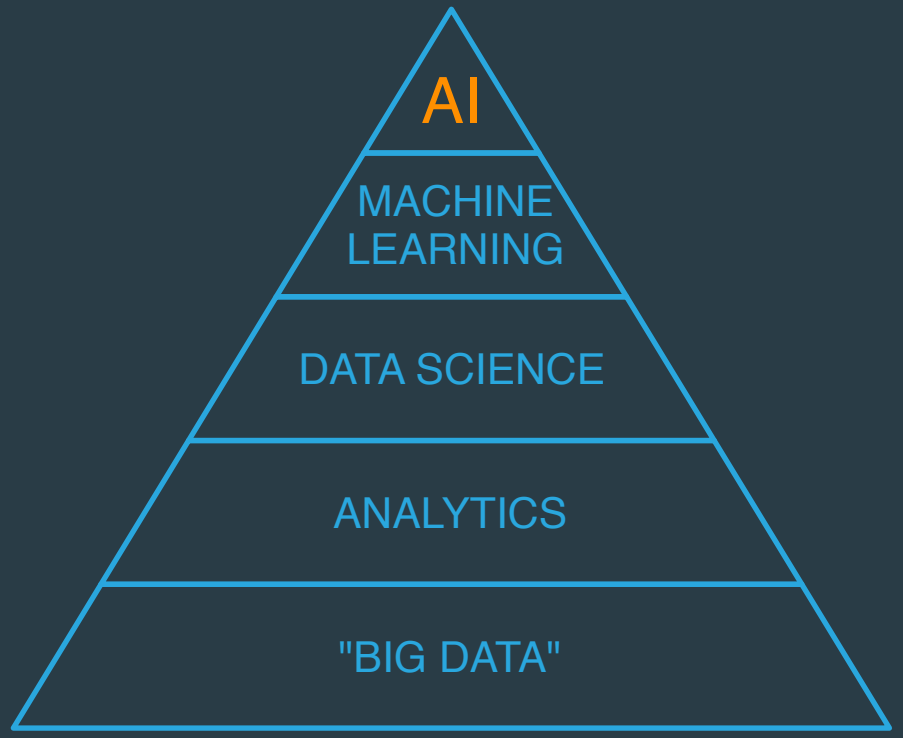
Gartner

DIGITAL TRANSFORMATION TOP PRIORITY

- Enterprises become data driven

HIERARCHY OF NEEDS FOR THE DATA-DRIVEN ENTERPRISE

The “AI Ladder”



DATA SCIENCE EXPERTS BEST PRACTICE

- You can’t do ML without doing data science
- You can’t do data science without analytics
- You can’t do analytics without Big Data
- They are all dependent on each other, and must operate on the same platform

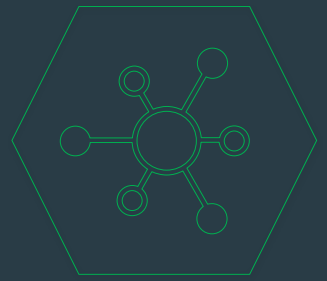
CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

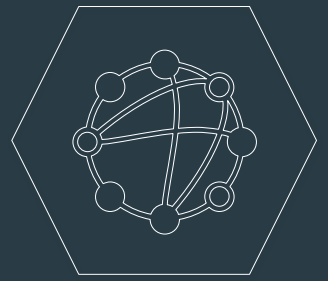
Edge2AI

WHERE WE ARE TODAY: DATA FROM ANYWHERE, FROM THE EDGE TO AI

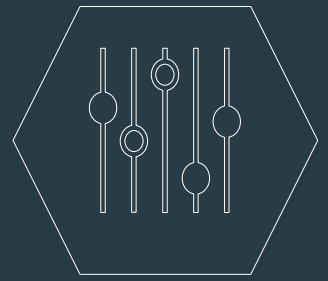
IoT, Ingest & Streaming



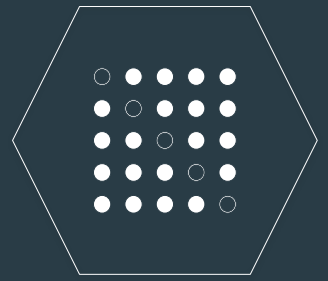
Data Engineering



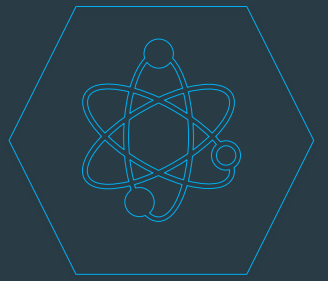
Data Warehouse



Operational Database



AI, ML & Data Science



Edge I Private I Public I Multi I Hybrid

WHAT'S NEEDED FOR THE FUTURE DATA ARCHITECTURE?

AGILE INTERACTION MODEL

- Each data application runs on its own virtual private cluster (while sharing the data)
- Strong isolation of tenants and workloads (mitigate noisy neighbors)
- Independent upgrade cycles for components associated with each application

EFFICIENT/FLEXIBLE COMPUTE MODEL

- Run on virtualized & elastic compute infrastructure (primarily Kubernetes based)
- Scale compute separate from storage (given a solid network backbone architecture)

UNIFIED SECURITY, MANAGEMENT & METADATA

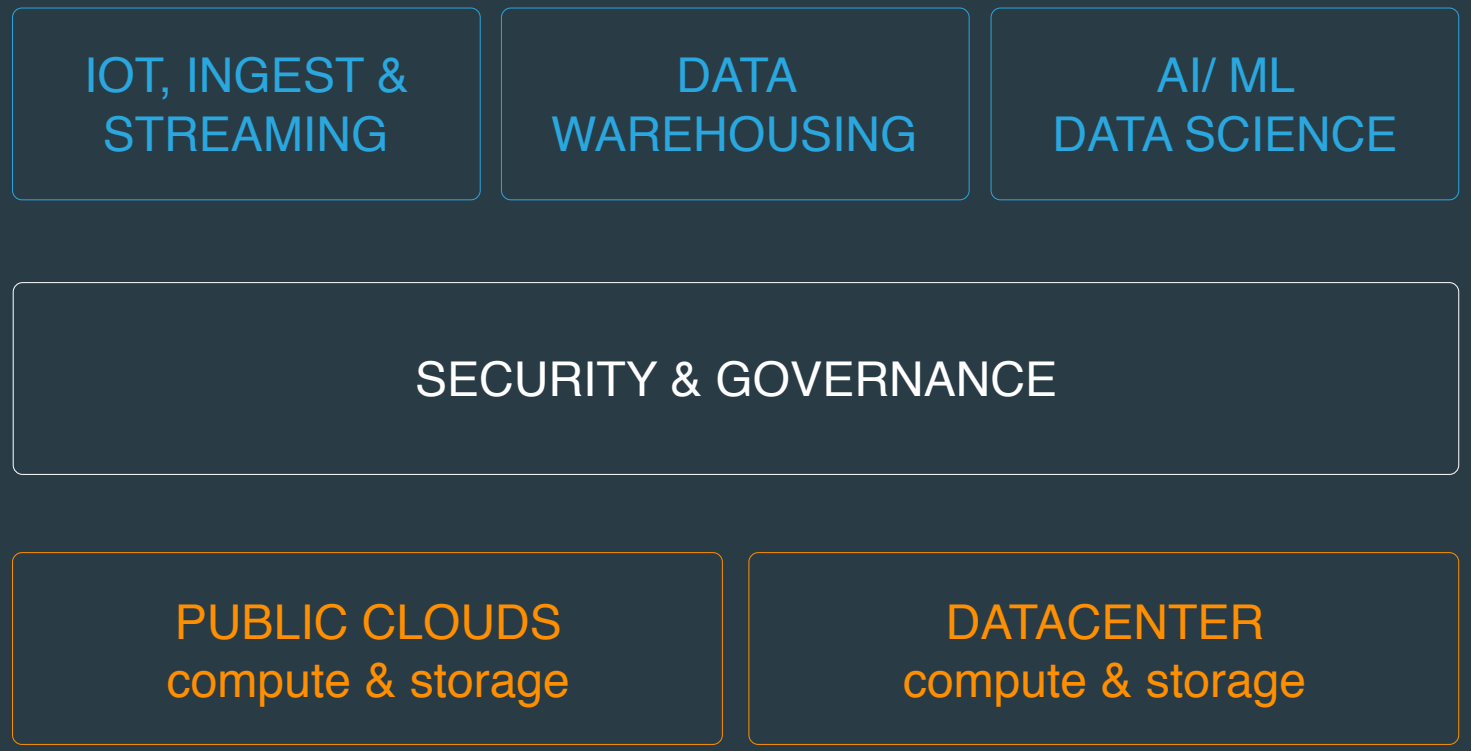
- Single pane of glass to manage 1000s of applications with a common data substrate
- Automation to drive dozens of application patterns and lifecycles
- Common security, governance, compliance, and metadata catalog

EASILY MOVE DATA & APPS BETWEEN ON PREMISES AND CLOUD

- Migration tools to easily move data and workloads
- Automatically synchronize metadata/security policies (and unified compliance tools)

THE ANSWER IS AN ENTERPRISE DATA CLOUD ARCHITECTURE

- Multi-function and multi-tenant analytics
- Hybrid and multi-cloud (migration/synchronization)
- Unified security and governance across workloads and cloud deployments on the same data
- Open platform



OUR ANSWER: THE CLOUDERA DATA PLATFORM (CDP)

Unified Control Plane

Management Console

Identity | Orchestration | Management | Operations

Manage all self-serve experiences and Distro-X clusters from a single pane of glass

Analytic Experiences

Data Flow & Streaming	Data Engineering	Data Warehouse	Operational Database	Machine Learning
DF-X	DE-X	DW-X	OD-X	ML-X

Reduce time and effort to deploy common application types with 5 new self-serve experiences

Open source distribution

DISTRO-X

Upgrade to a new distro and a new service to manage virtual private clusters (Distro-X)

Data Anywhere

cloudera sdx shared data experience

Catalog | Schema | Migration | Security | Governance

Create a shared data experience while maintaining safety, security, and compliance

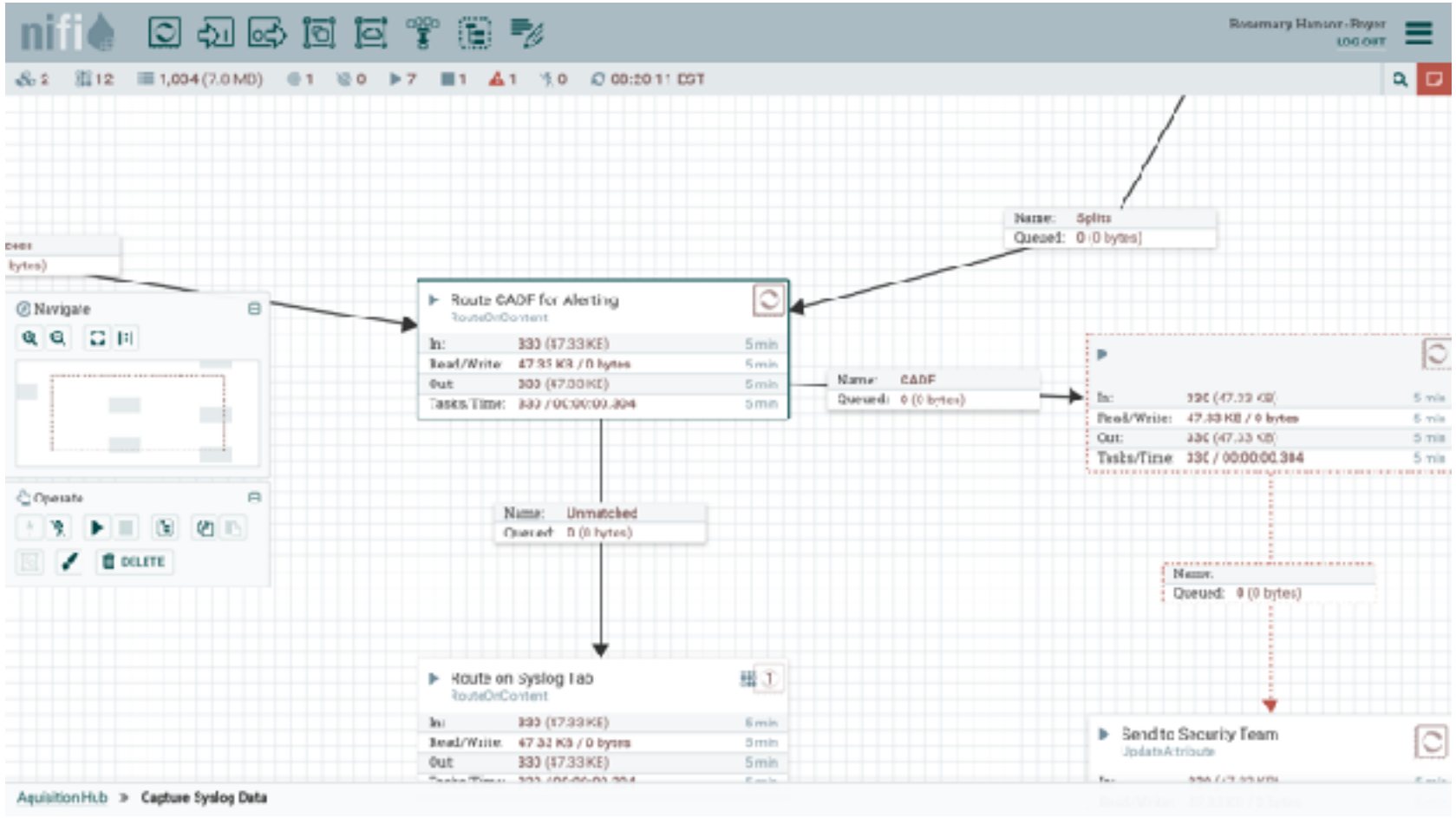
Any Infrastructure

Edge | Private Cloud | Public Multi-Cloud | Hybrid Cloud

For existing or new on-prem infrastructure
For multiple public clouds
For hybrid deployment

FLOW MANAGEMENT

- Web-based user interface
- Highly configurable
- Data provenance
- Designed for extensibility
- Secure
- NiFi Registry
  - DevOps support
  - FDLC
  - Versioning
  - Deployment



CLOUDERA DATA SCIENCE WORKBENCH

Accelerate Machine Learning from Research to Production

For data scientists

- **Experiment faster**
  Use R, Python, or Scala with on-demand compute and secure CDH data access
- **Work together**
  Share reproducible research with your whole team
- **Deploy with confidence**
  Get to production repeatably and 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**
  On-premises or in the cloud

CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

Data Science Workbench

CDSW

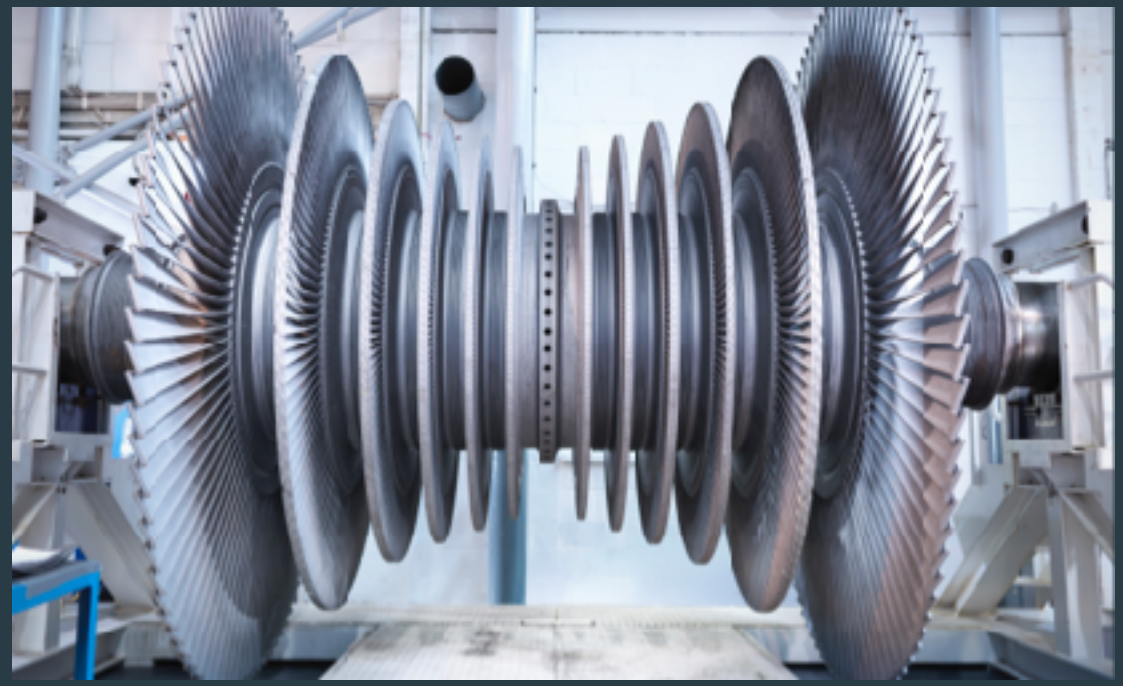
MACHINE LEARNING AT CLOUDERA

Our philosophy

We empower our customers to run their business on data with an open platform:

- Your data
- Open algorithms
- Running anywhere

We accelerate enterprise data science.



What is Cloudera Data Science Workbench (CDSW)

Supports your workloads in an enterprise secure way

Accelerate data science from exploration to production using R, Python, Spark and more

For data scientists

Open data science, your way.

Use R, Python, or Scala with your favorite libraries and frameworks

No need to sample.

Directly access data in secure Hadoop clusters through Apache Spark and Apache Impala

Reproducible, collaborative research.

Share insights with your whole team

For IT professionals

Bring analysis to the data.

Give your data science team the freedom to work how they want, when they want

Secure by default.

Stay compliant with out-of-the-box support for full Hadoop security

Flexible deployment.

Run on-premises or in the cloud

THE CHALLENGE

Balance these needs

DATA SCIENCIST

- Access to granular data
- Flexibility
  - Preferred open source tools
- Elastic provisioning
  - Compute
  - Storage
- Reproducible research
- Path to production

VS.

DevOps/IT

- Security
- Governance
- Standards
- Low maintenance
- Low cost
- Self-service access

THE TYPICAL SOLUTION

"If I can't use my favorite tools, 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

CLOUDERA DATA SCIENCE WORKBENCH

Accelerate Machine Learning from Research to Production

For data scientists

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

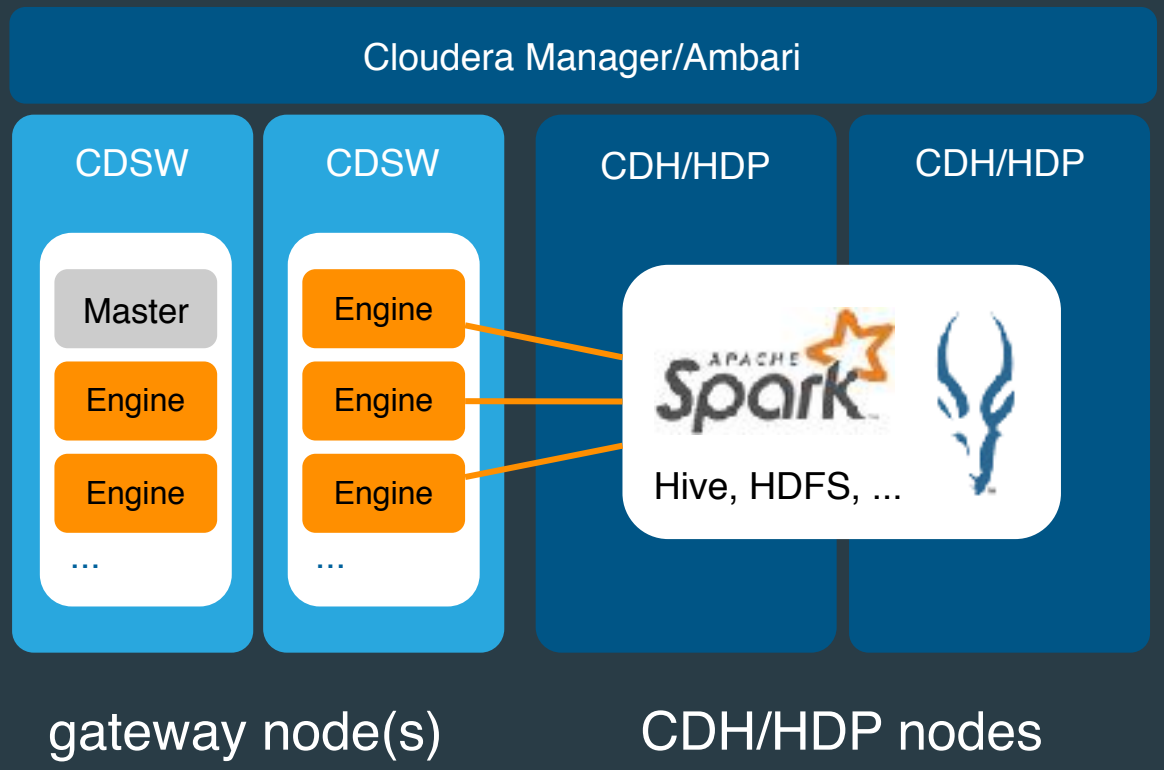
For IT professionals

- **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**
  On-premises or in the cloud

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/Ambari

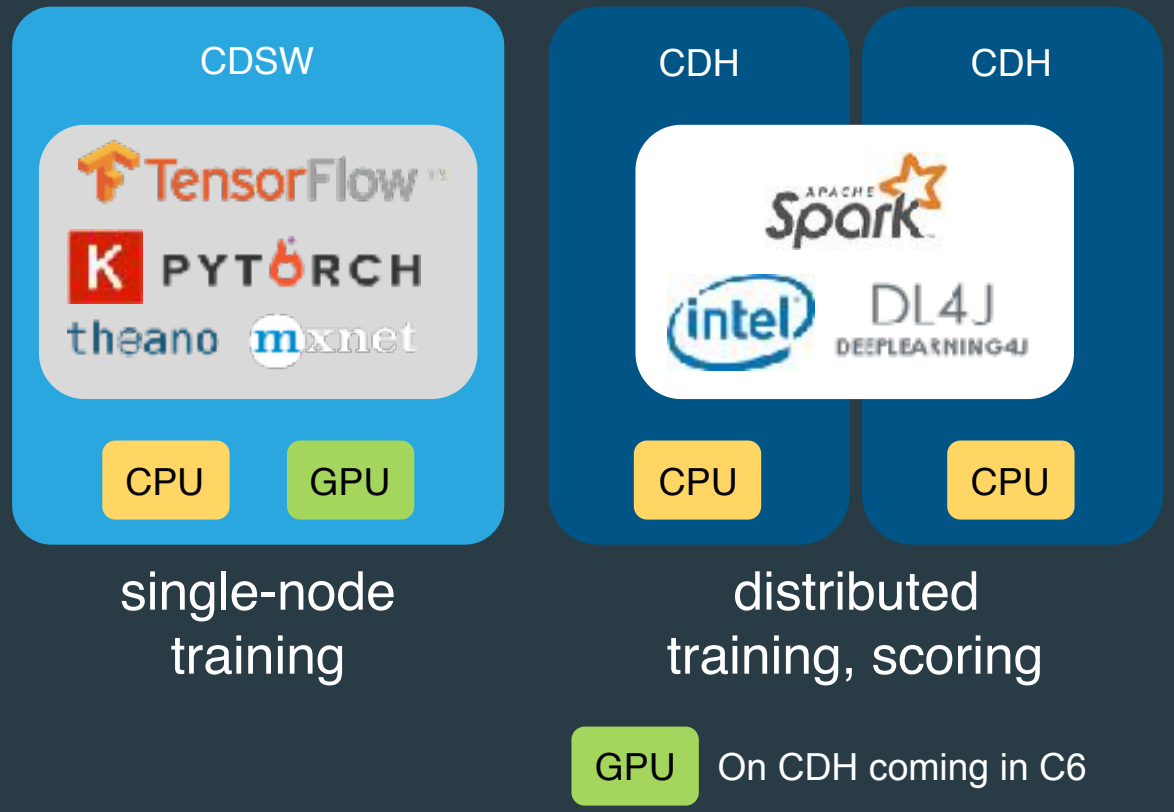


ACCELERATED DEEP LEARNING WITH GPUS

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



WHAT DATA SCIENCE TEAMS DO

PREPARE DATA

Ingest data at scale.

Store and secure data.

Clean and transform data for analysis.

BUILD MODELS

Explore data and build predictive models, offline.

Evaluate and tune models.

Develop and deliver a modeling pipeline.

DEPLOY MODELS

Test, verify, and approve model for deployment.

Create and maintain batch/ stream pipelines, embedded models, APIs.

Update models in production.

CLOUDERA DATA SCIENCE WORKBENCH

Accelerate and simplify machine learning from research to production

ANALYZE DATA

- Explore data securely and share insights with the team

TRAIN MODELS

- Run, track, and compare reproducible experiments

DEPLOY APIs

- Deploy and monitor models as APIs to serve predictions

MANAGE SHARED RESOURCES

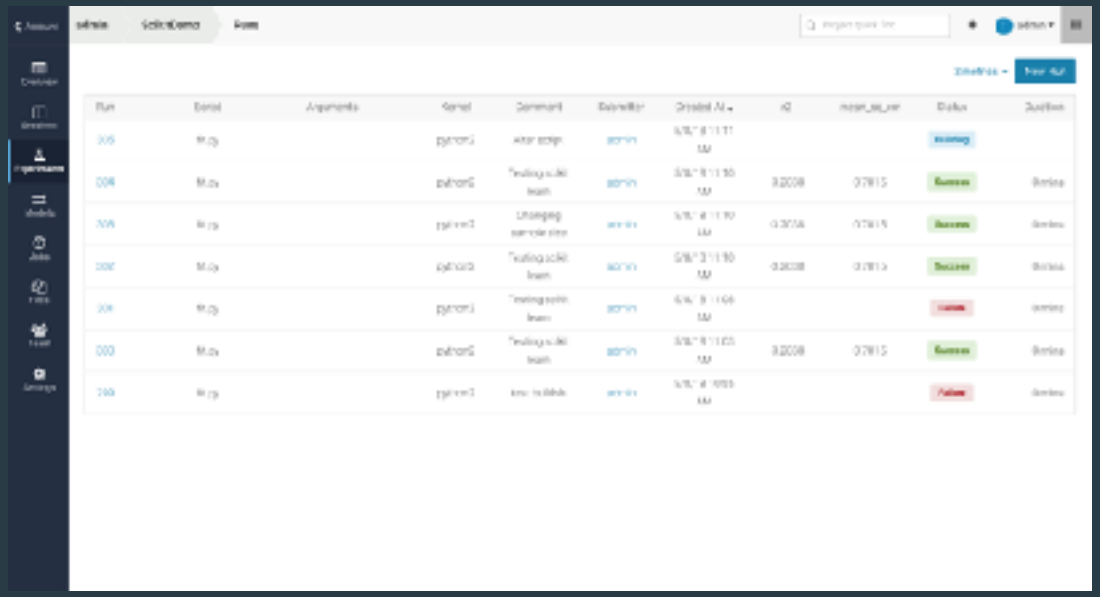
- Provide a secure, collaborative, self-service platform for your data science teams

INTRODUCING 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



INTRODUCING MODELS

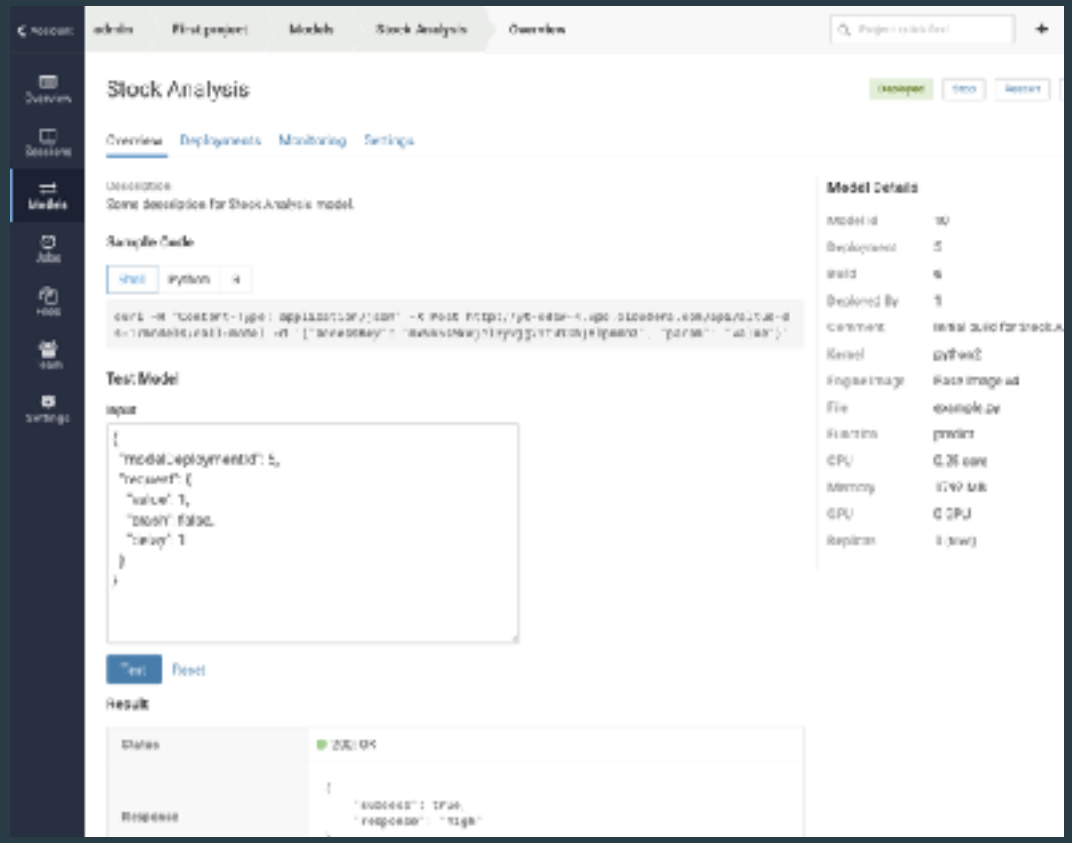
Machine learning models as one-click microservices (REST APIs)

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)
```

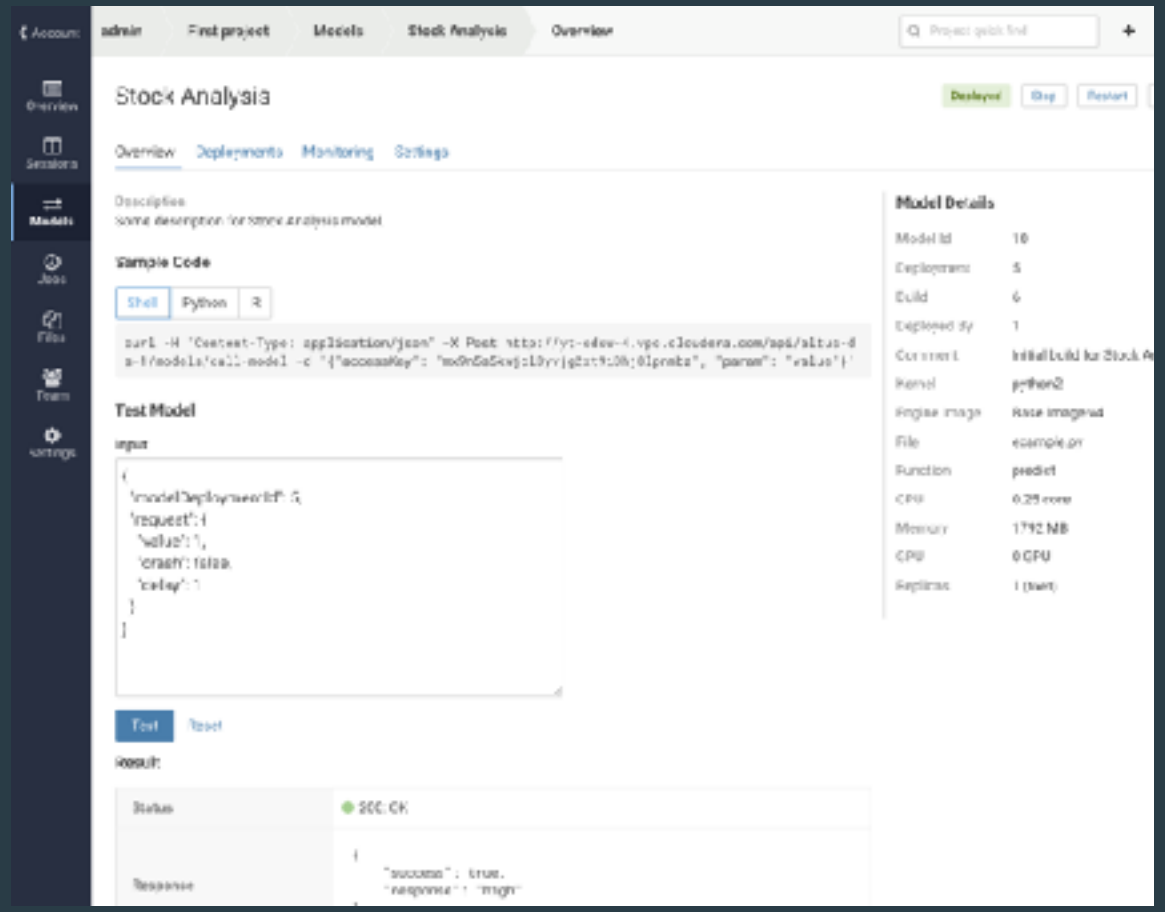
3. Choose resources
4. Deploy!

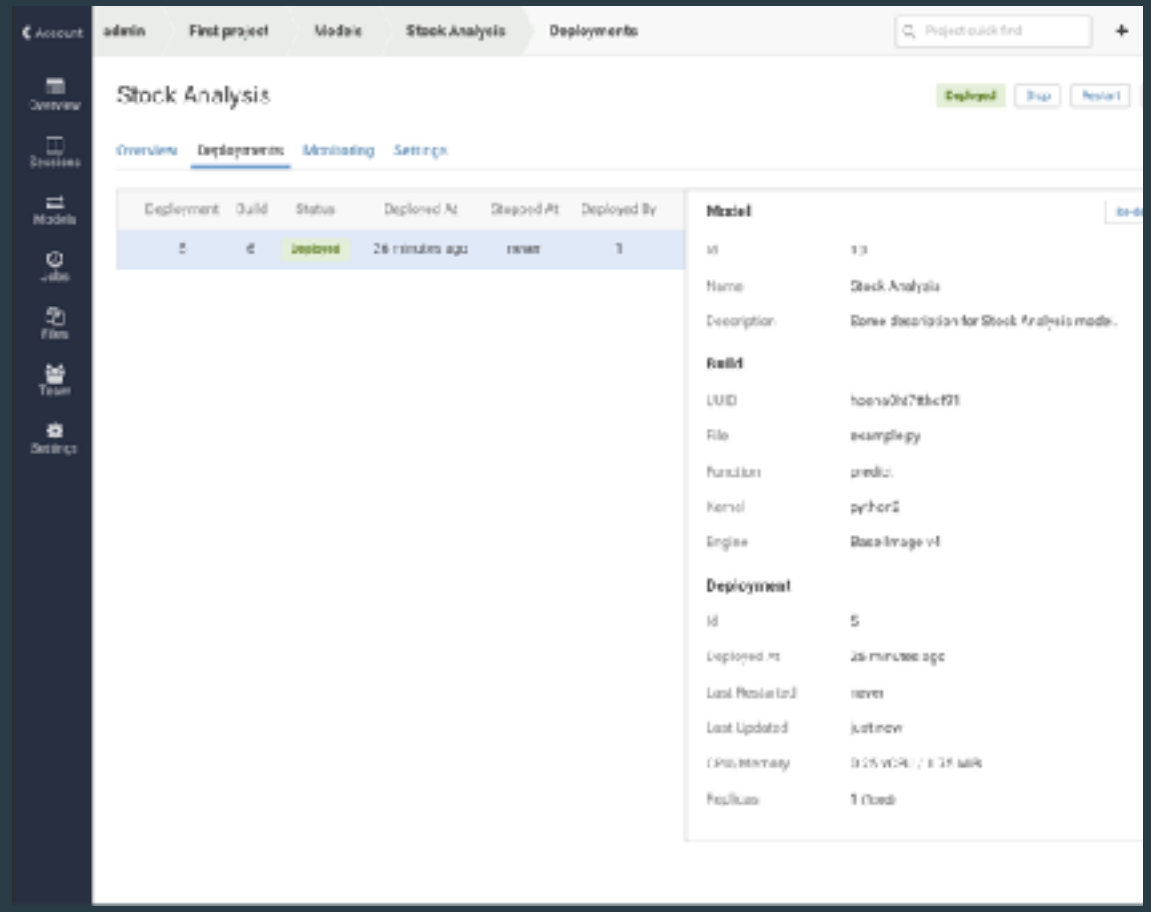
Running model containers also have access to CDH for data lookups.

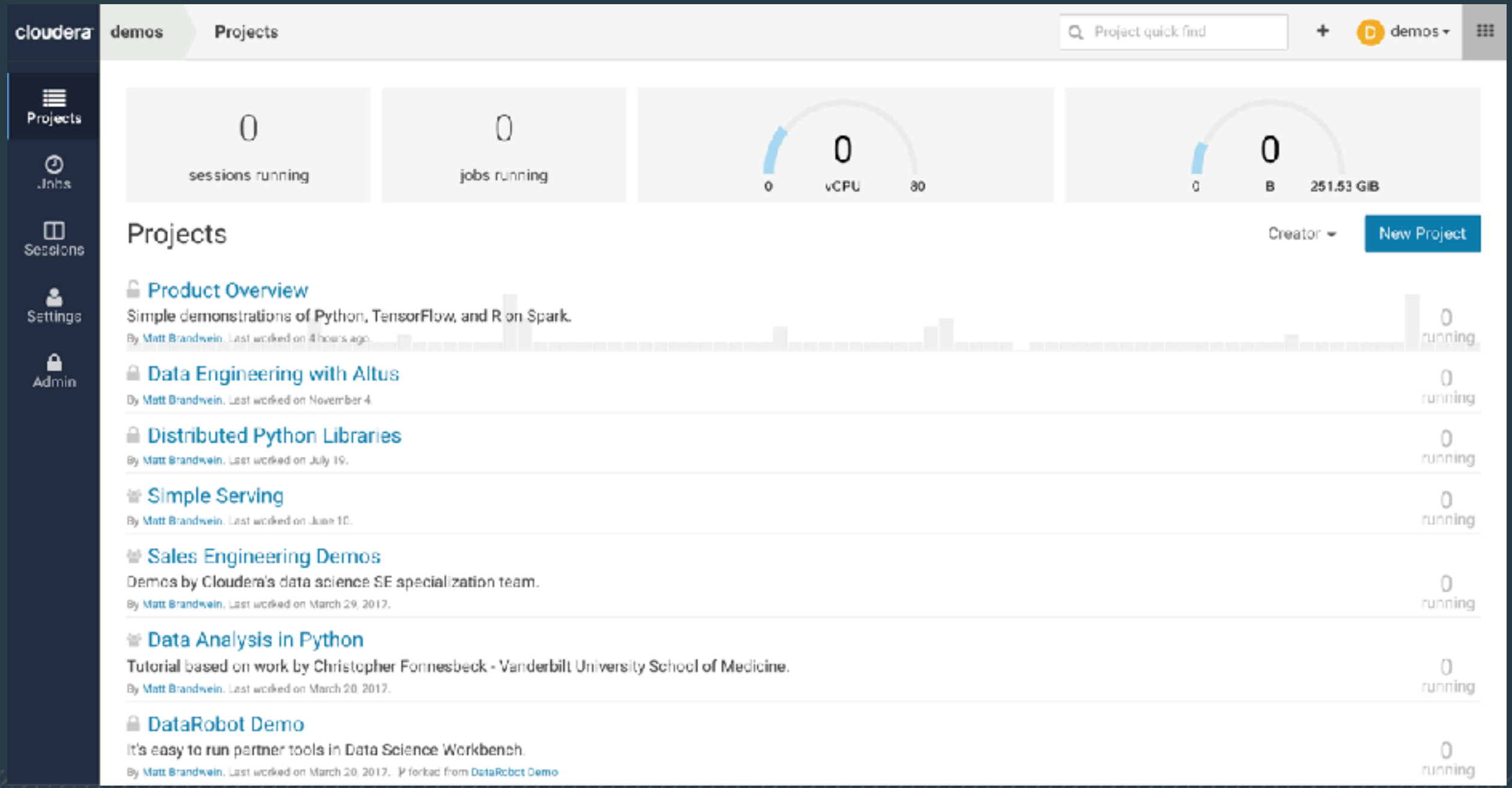


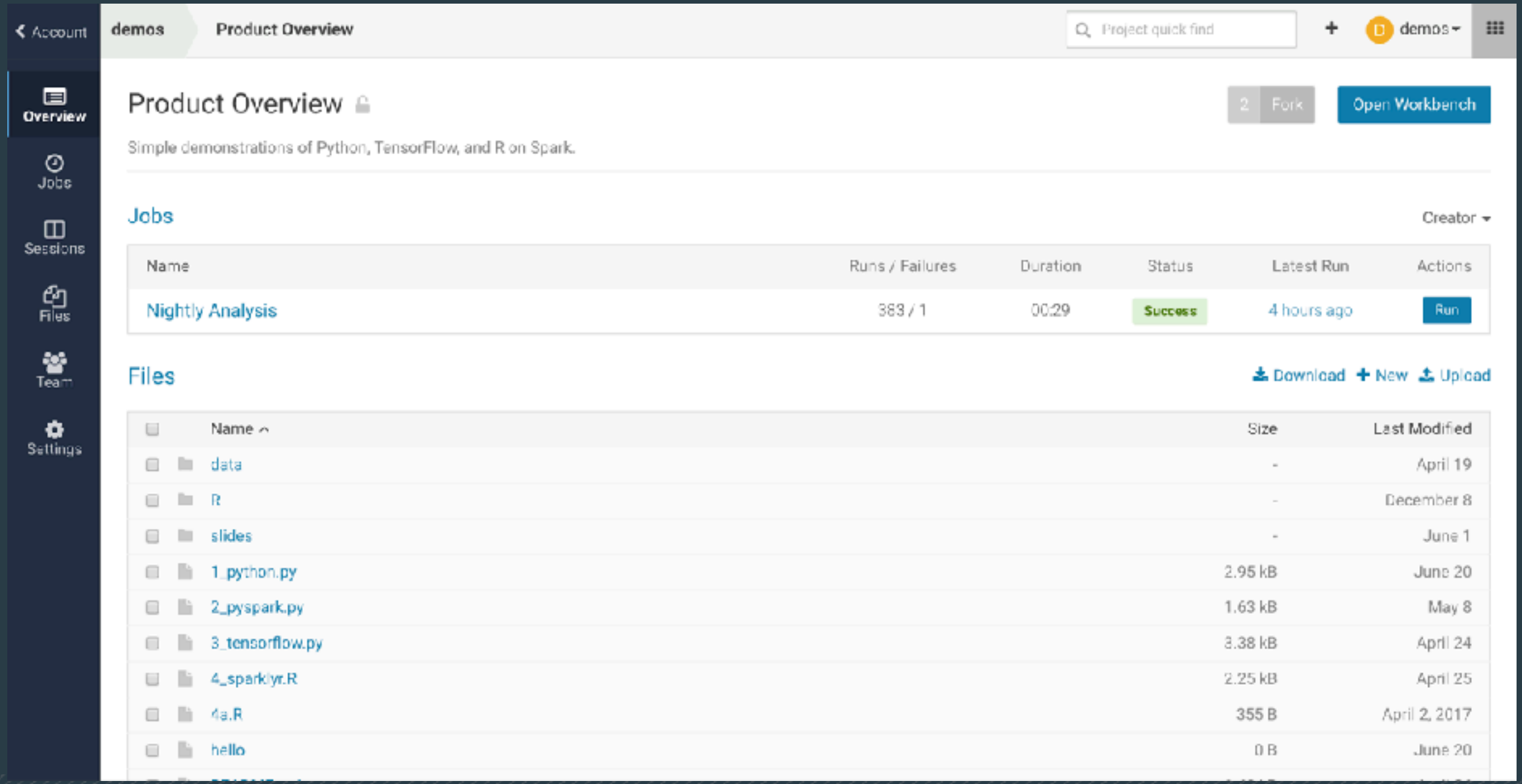
MODEL MANAGEMENT

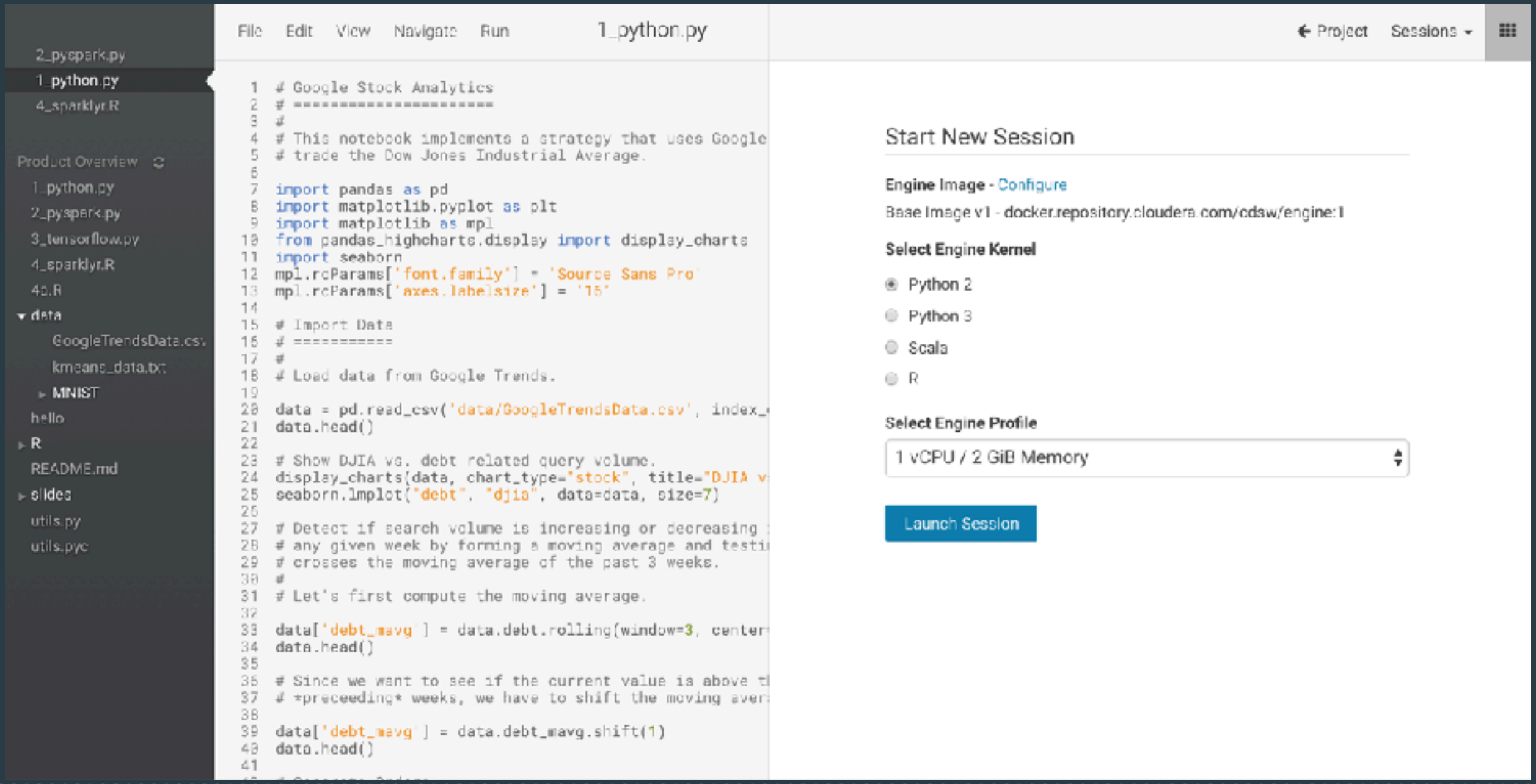
View, test, monitor, and update models by team or project

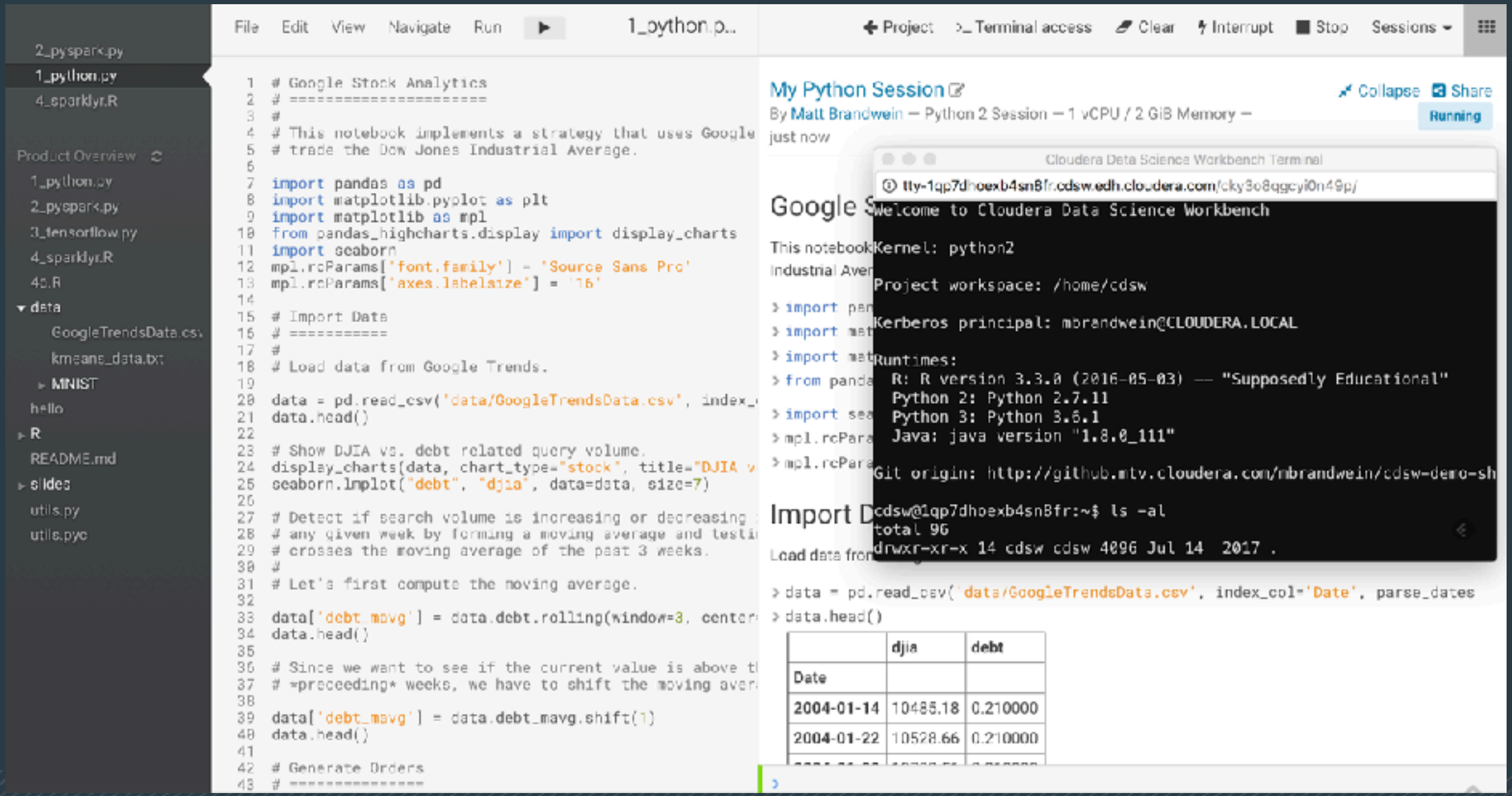












CLOUDERA

THE ENTERPRISE DATA CLOUD COMPANY

Cloudera Data Flow

CDF

MARKET OPPORTUNITIES

Streaming
~$1.65 B

Data Science
~$180 B

Big Data
~$210 B

Cloud
~$410 B

IoT
~$1.2 T

IOT MARKET

24.9B	By 2024 more than 24.9 Billion IoT connections will be established
$70B	An estimated $70 billion will be spent by global manufacturers on IoT solutions in 2020
646M	An estimated 646 million healthcare devices (excluding fitness trackers and wearable devices) will be connected by 2020
78%	An estimated 78% of cars shipped globally will be built with hardware that connects to the internet by 2020
50%	50% of decision-makers in IT, services, utilities, and manufacturing have either deployed IoT, or will deploy it in the next 12-24 months

PROBLEMS IN THE MARKET – PAIN THE CUSTOMER EXPERIENCES

Data movement

Continuous data ingestion

Streaming ETL

Streaming analytics

COMMON USE CASES

Data Movement
Optimize resource utilization by moving data between data centers or between on-premises infrastructure and cloud infrastructure

Optimize Log Collection & Analysis
Optimize log analytics solutions by using CDF as a single platform to collect and deliver multiple data sources

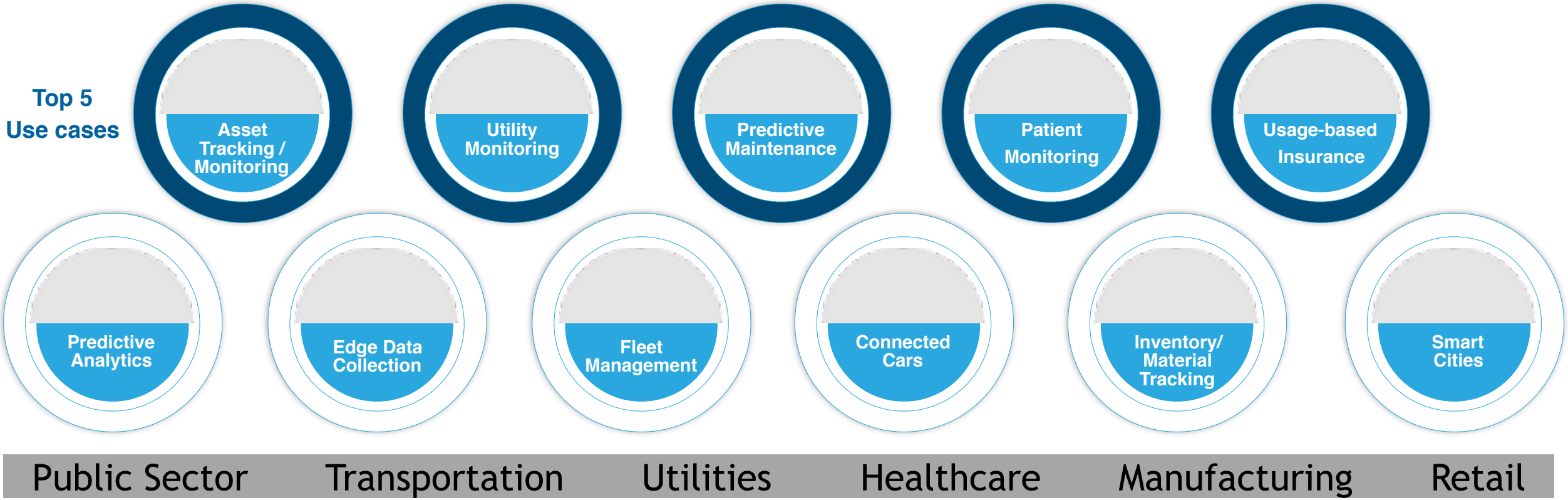
Gain key insights with Streaming Analytics
Accelerate big data ROI by analyzing streaming data for patterns, comparing with ML models and delivering actionable intelligence

Single view / 360° view of customer
Ingest, transform and combine customer data from multiple sources into a single data view / lake

Stream Processing
Combine multiple streams of data in real-time, enrich the data and route it to different end points based on rules

Capture and Analyze IoT Data
Ingest sensor data from IoT devices and stream it for further processing and comprehensive analysis

COMMON IOT USE CASES BY INDUSTRY

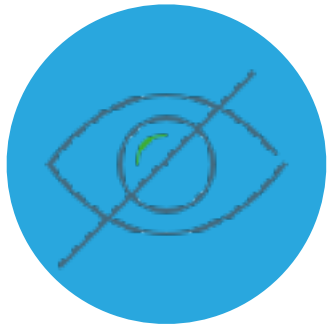


- IoT is a $1.13T market opportunity in 2021.
- Americas - $329B IoT spending. Manufacturing and Transportation are top industries, accounting for 26% of total spending.
- APAC - $500B IoT spending. Manufacturing, Utilities and Transportation are top industries.
- EMEA - $264B IoT spending. Manufacturing is top industry, powered by Industry 4.0 initiatives.
- Worldwide IoT Analytics and Information Management Market = $573M

KEY CUSTOMER CHALLENGES

Data Ingestion: High-volume streaming sources, multiple message formats, diverse protocols and multi-vendor devices creates data ingestion challenges

Real-time Insights: Analyzing continuous and rapid inflow (velocity) of streaming data at high volumes creates major challenges for gaining real-time insights



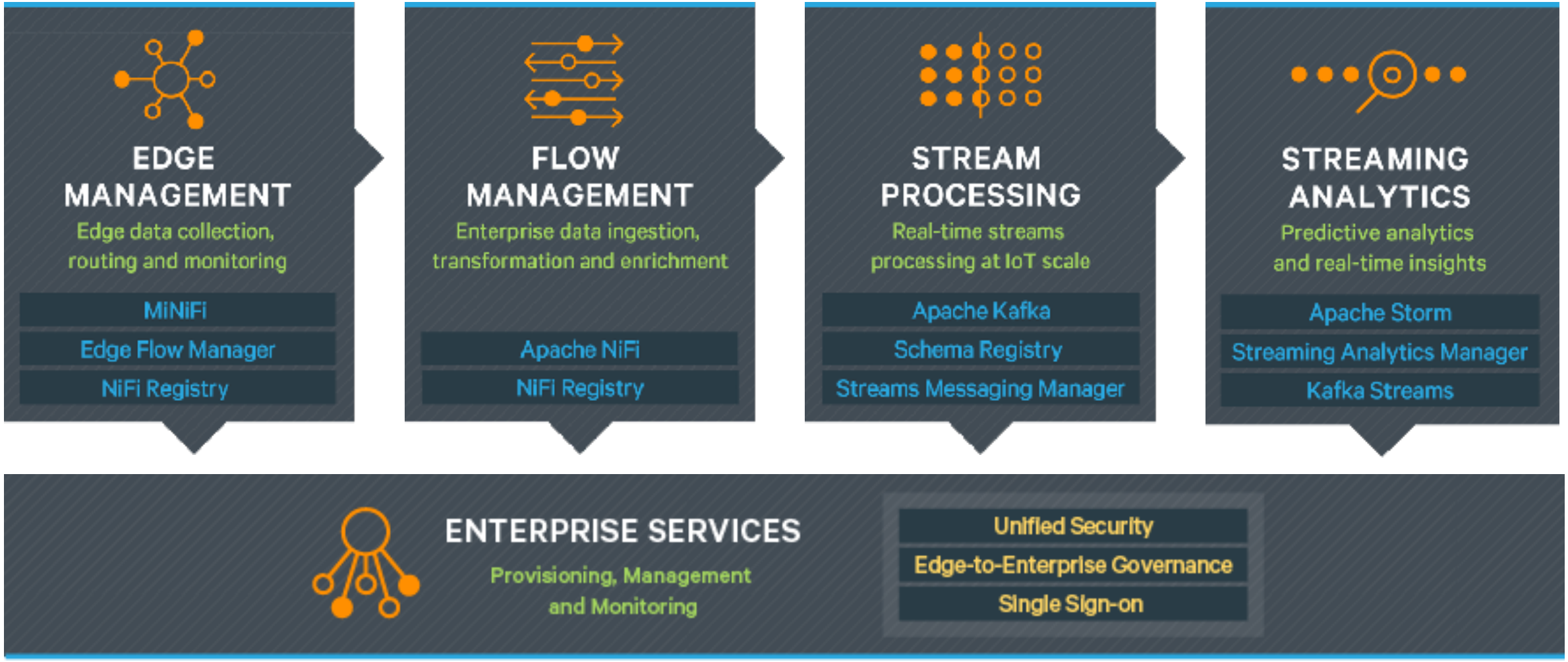
Visibility: Lack visibility of end-to-end streaming data flows, inability to troubleshoot bottlenecks, consumption patterns etc.

PRODUCT OVERVIEW

WHAT IS CLOUDERA DATAFLOW (CDF)?

Cloudera DataFlow (CDF) is a scalable, real-time streaming data platform that collects, curates, and analyzes data so customers gain key insights for immediate actionable intelligence.

CLOUDERA DATAFLOW



CLOUDERA DATAFLOW

WHAT IS CLOUDERA EDGE MANAGEMENT (CEM)?

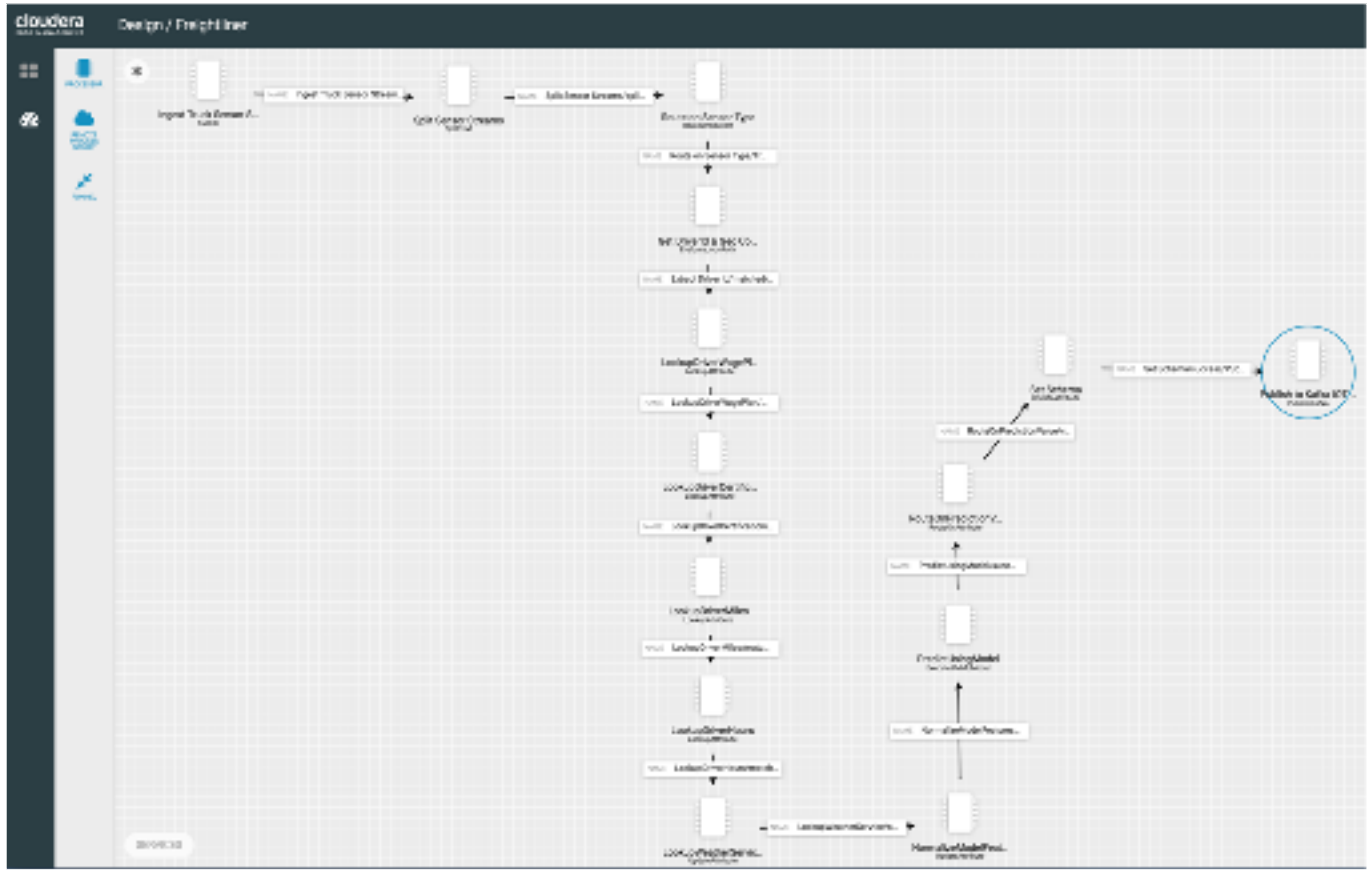
Cloudera Edge Management (CEM) is an edge management solution made up of edge agents and an edge management hub. It manages, controls and monitors edge agents to collect data from edge devices and push intelligence back to the edge. CEM allows you to develop, deploy, run and monitor edge flow apps on thousands of edge devices.

EDGE DATA MANAGEMENT

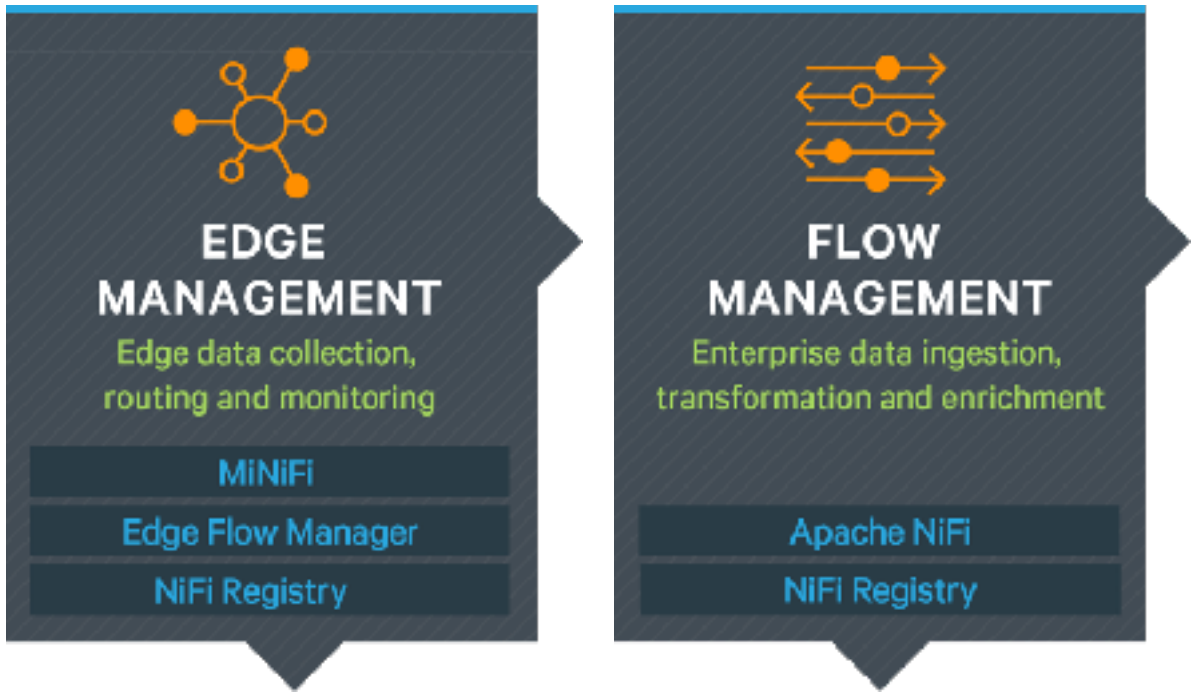
- Edge data collection powered by Apache MiNiFi
- MiNiFi – smaller footprint than NiFi
  - Guaranteed delivery
  - Data buffering
  - Prioritized queuing
  - Flow-specific QoS
  - Data provenance
  - Designed for extension
  - C++ / Java agents
  - TensorFlow support
- Designed for IoT

EDGE FLOW MANAGER

- Edge management hub
- NiFi-like user interface to develop and deploy flow files to the edge
- Update and deploy ML model files to the edge agents
- Monitor thousands of edge agents
- Integration with NiFi Registry



CLOUDERA DATAFLOW

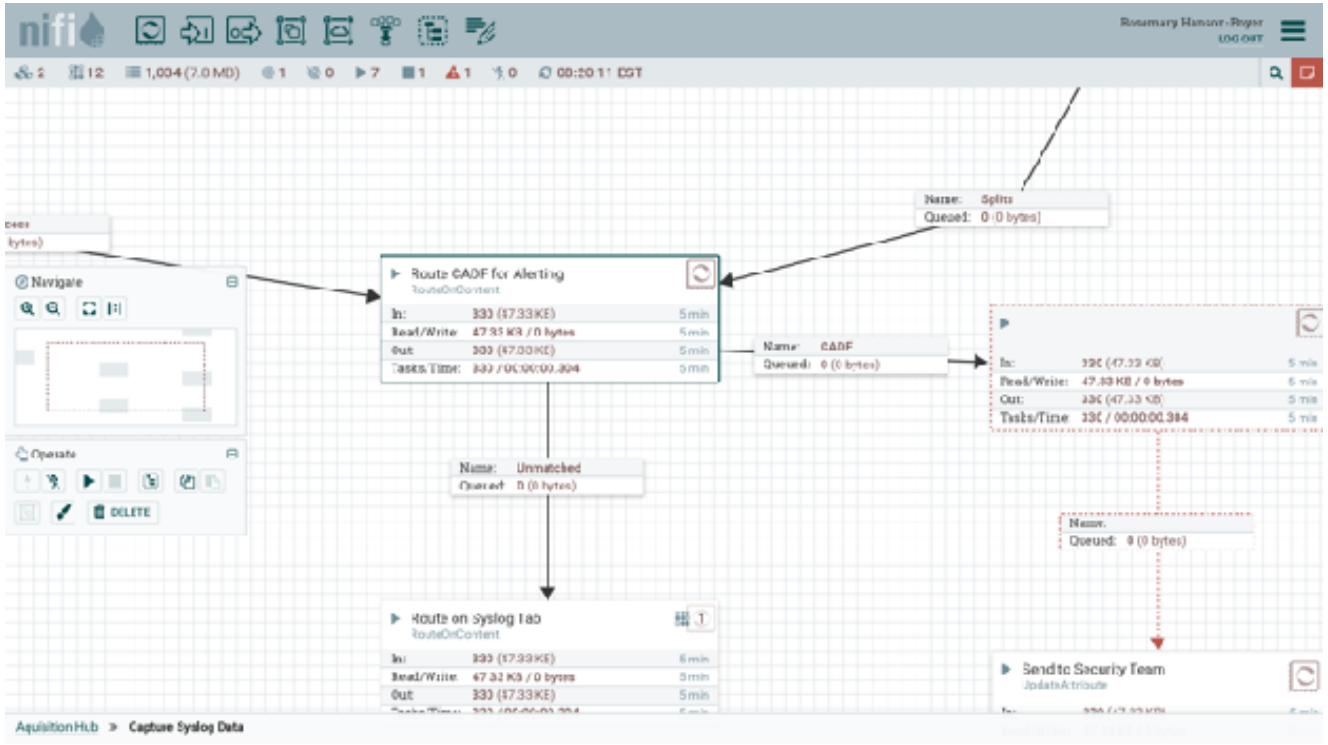


WHAT IS CLOUDERA FLOW MANAGEMENT (CFM)?

Cloudera Flow Management (CFM) is a no-code data ingestion and management solution powered by Apache NiFi. With NiFi's intuitive graphical interface and 300+ processors, CFM delivers highly scalable data movement, transformation and management capabilities to the enterprise. CFM also enables DevOps type development and deployment with its support for NiFi Registry.

FLOW MANAGEMENT

- Web-based user interface
- Highly configurable
- Out-of-the-box data provenance
- Designed for extensibility
- Secure
- NiFi Registry
  - DevOps support
  - FDLC
  - Versioning
  - Deployment



300+ PROCESSORS FOR DEEPER ECOSYSTEM INTEGRATION

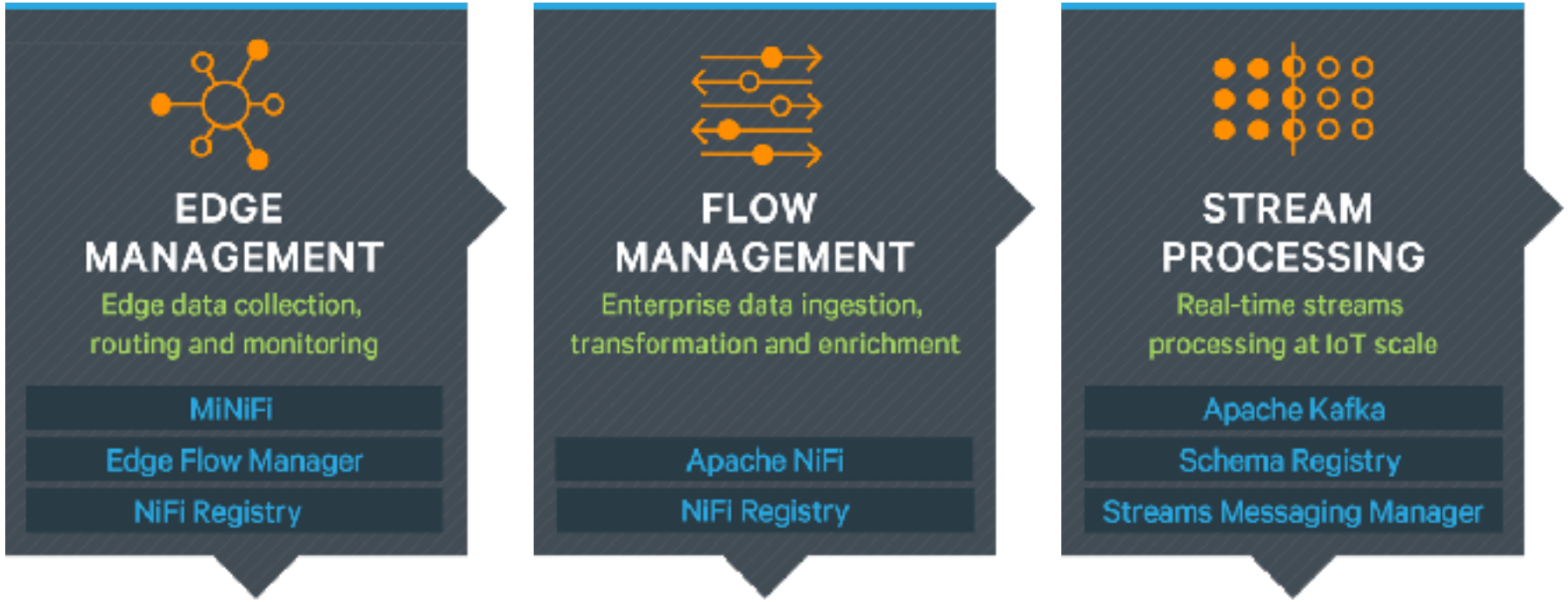
- FTP
- SFTP
- HL7
- UDP
- XML
- HTTP
- WebSocket
- Email
- HTML
- Image
- Syslog
- AMQP

- Hash
- Encrypt
- GeoEnrich
- Merge
- Tail
- Scan
- Extract
- Evaluate
- Replace
- Duplicate
- Execute
- Translate
- Split
- Fetch
- Convert

- Route Text
- Distribute Load
- Route Content
- Generate Table Fetch
- Route Context
- Jolt Transform JSON
- Control Rate
- Prioritized Delivery

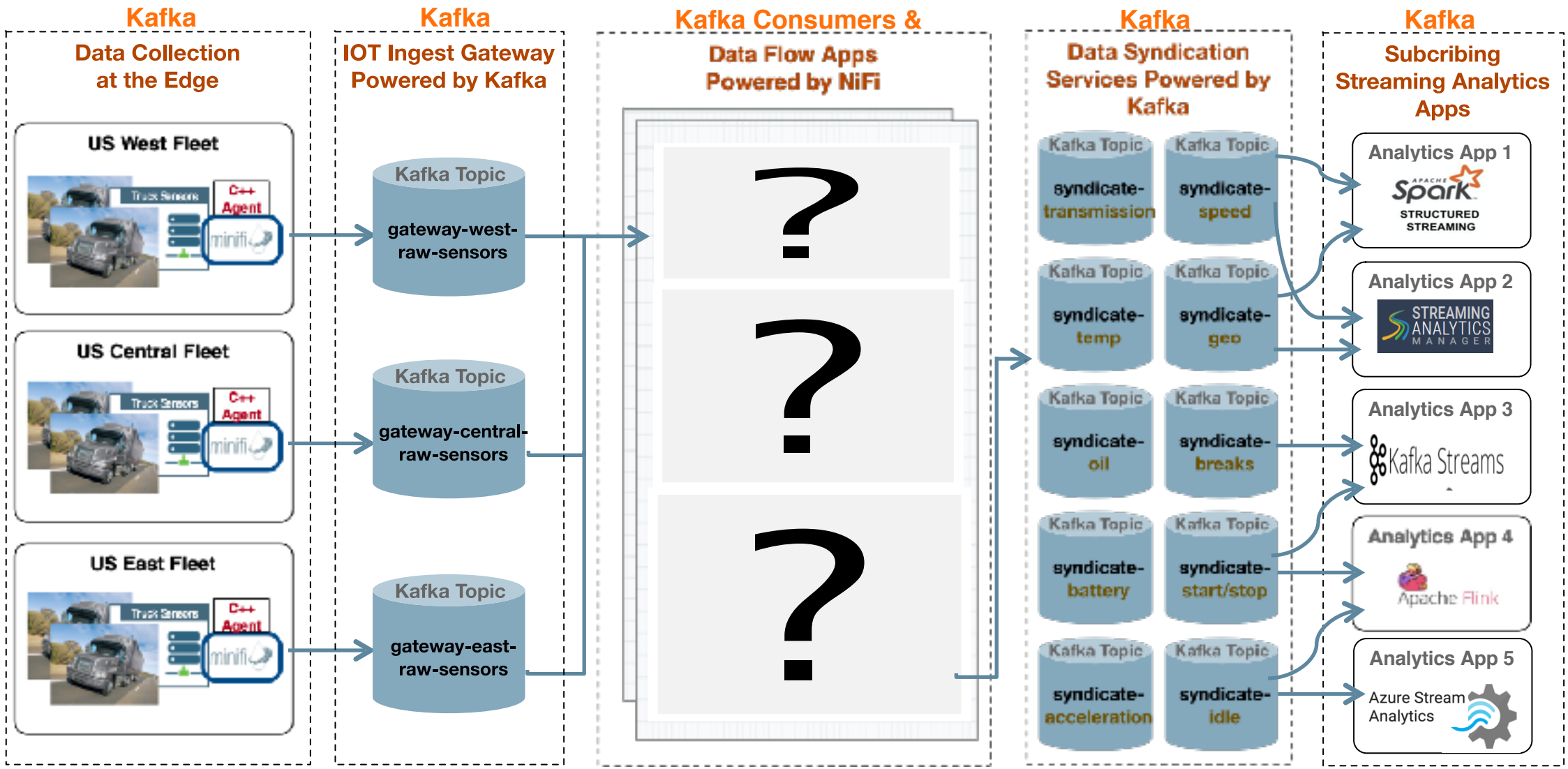
All Apache project logos are trademarks of the ASF and the respective projects.

CLOUDERA DATAFLOW



Streaming Analytics Reference Architecture

Kafka is Everywhere. Critical Component of Streaming

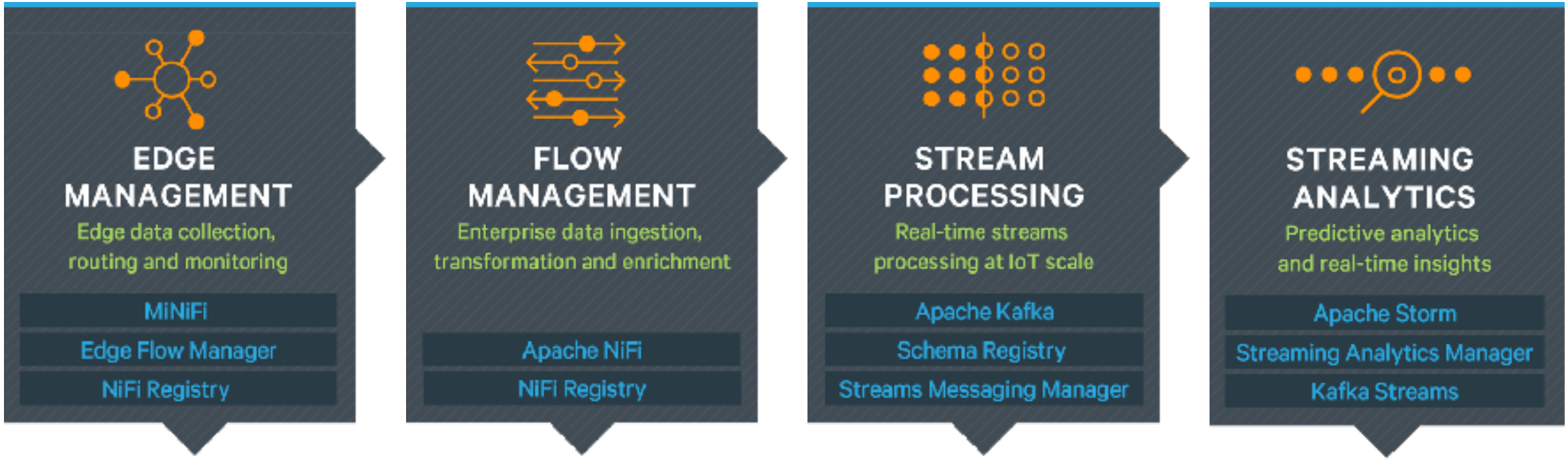


Cloudera Streams Messaging Manager (SMM)

What is SMM?

- Kafka Management and Monitoring tool
- Cure the “Kafka Blindness”
- Single Monitoring Dashboard for all your Kafka Clusters across 4 entities
  - Broker
  - Producer
  - Topic
  - Consumer
- REST as a First Class Citizen

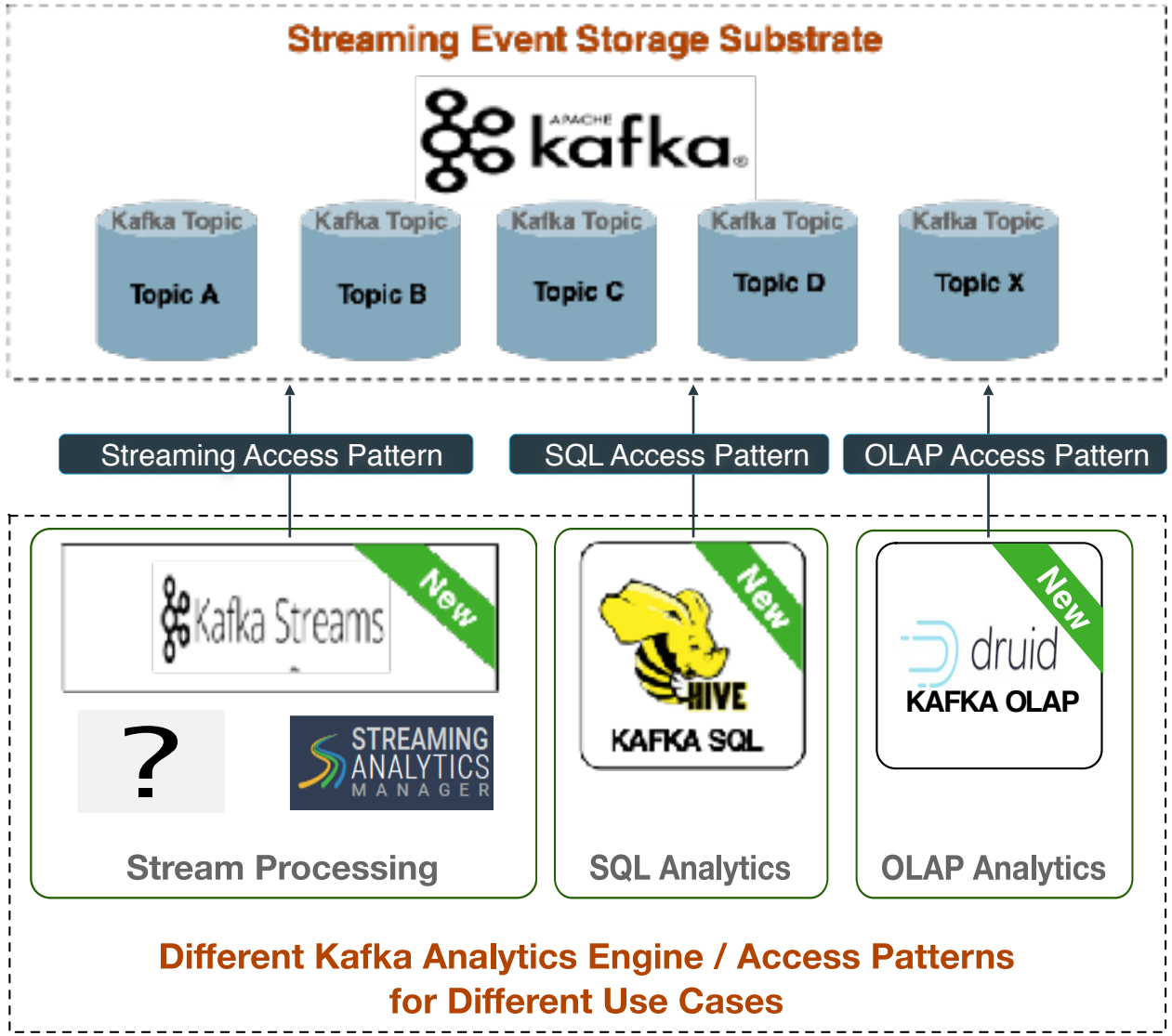
CLOUDERA DATAFLOW



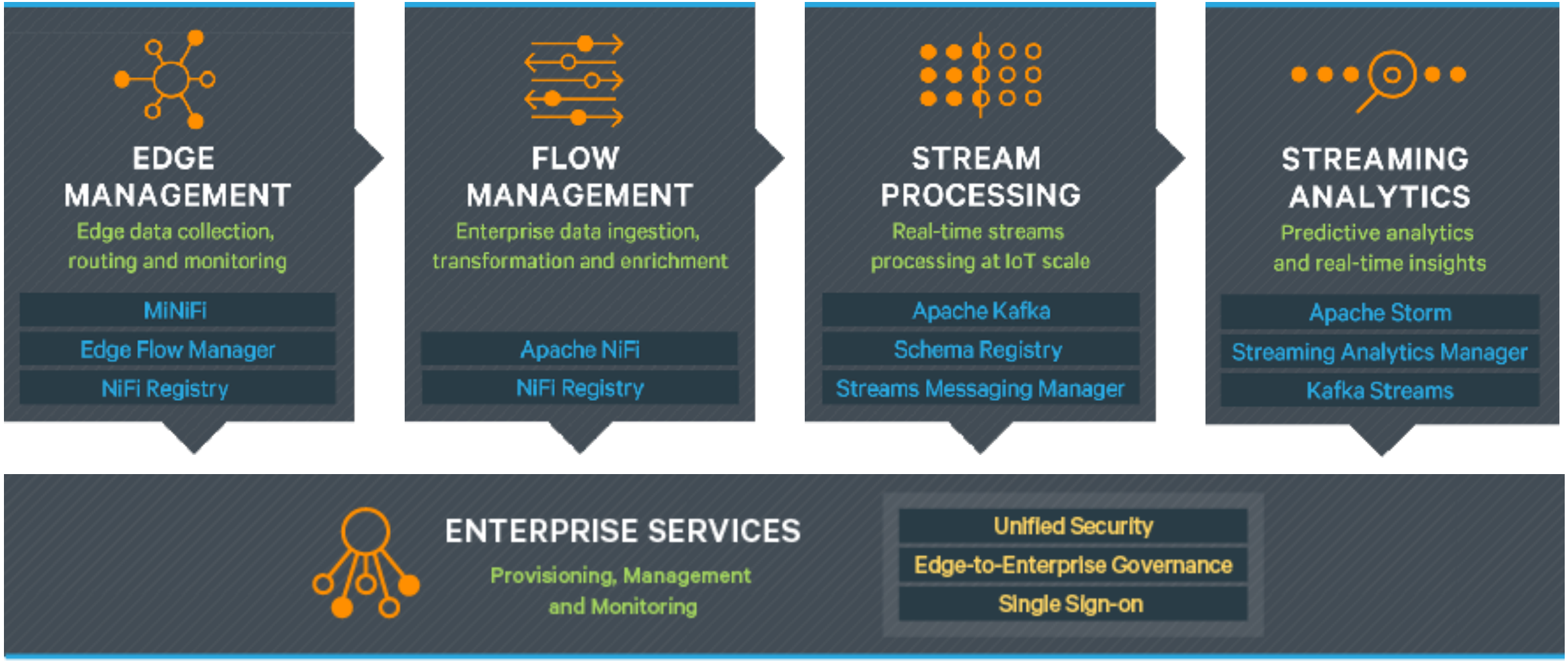
STREAMING ANALYTICS

- Pattern matching
- Predictive and Prescriptive Analytics
- Complex Event Processing
- Continuous & Real-time Insights

3 Kafka Analytics Access Patterns



CLOUDERA DATAFLOW



ENTERPRISE SERVICES

- Provisioning
- Management
- Monitoring
- Unified Security
- Single Sign-on
- Audit
- Compliance
- Edge-to-Enterprise Governance

KEY DIFFERENTIATORS

100% open source technology – Only vendor with this strategy; prevents vendor lock-in

300+ pre-built processors – Only product to offer such comprehensive connectivity from edge to enterprise

3 Streaming analytics engines – Only vendor to offer a choice of three streaming analytics engines to customers for all their streaming architecture needs

Built-in data provenance – Only product in the market to offer out-of-the-box data provenance on data-in-motion

Comprehensive streaming platform – Only big data vendor to offer a comprehensive streaming platform from real-time data ingestion, transformation, routing to descriptive, prescriptive and predictive analytics.

THANK YOU

cloudera