CLOUDERA

# CLOUDERA EDGE TO AI PLATFORM OVERVIEW

Philip Morch - Account Executive Nordics
Johannes Muselaers - Solutions Engineer Nordics

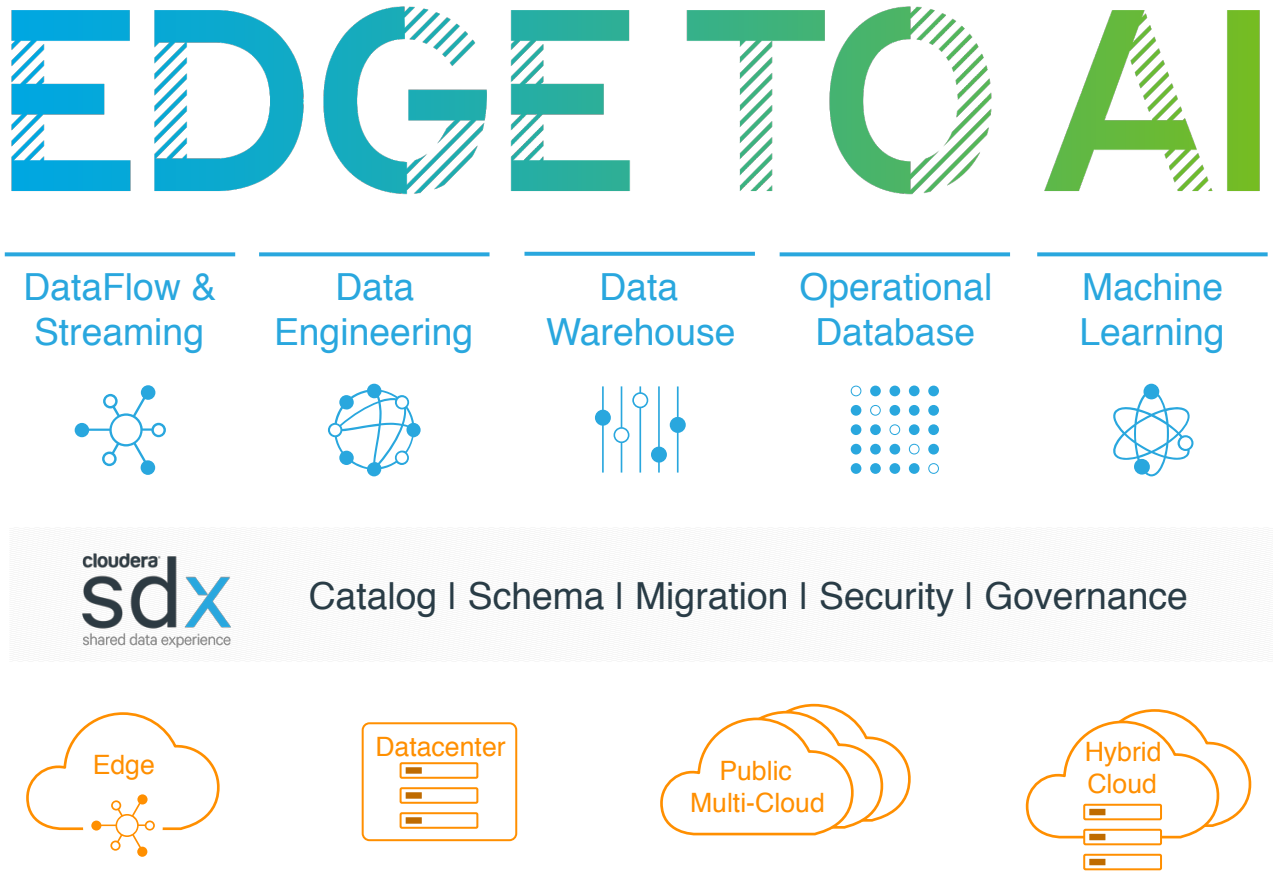
# PDF of the presentation

github.com/jmuselaers/presentations/Edge2AI20190927.pdf

# CLOUDERA ENTERPRISE + HORTONWORKS PLATFORM

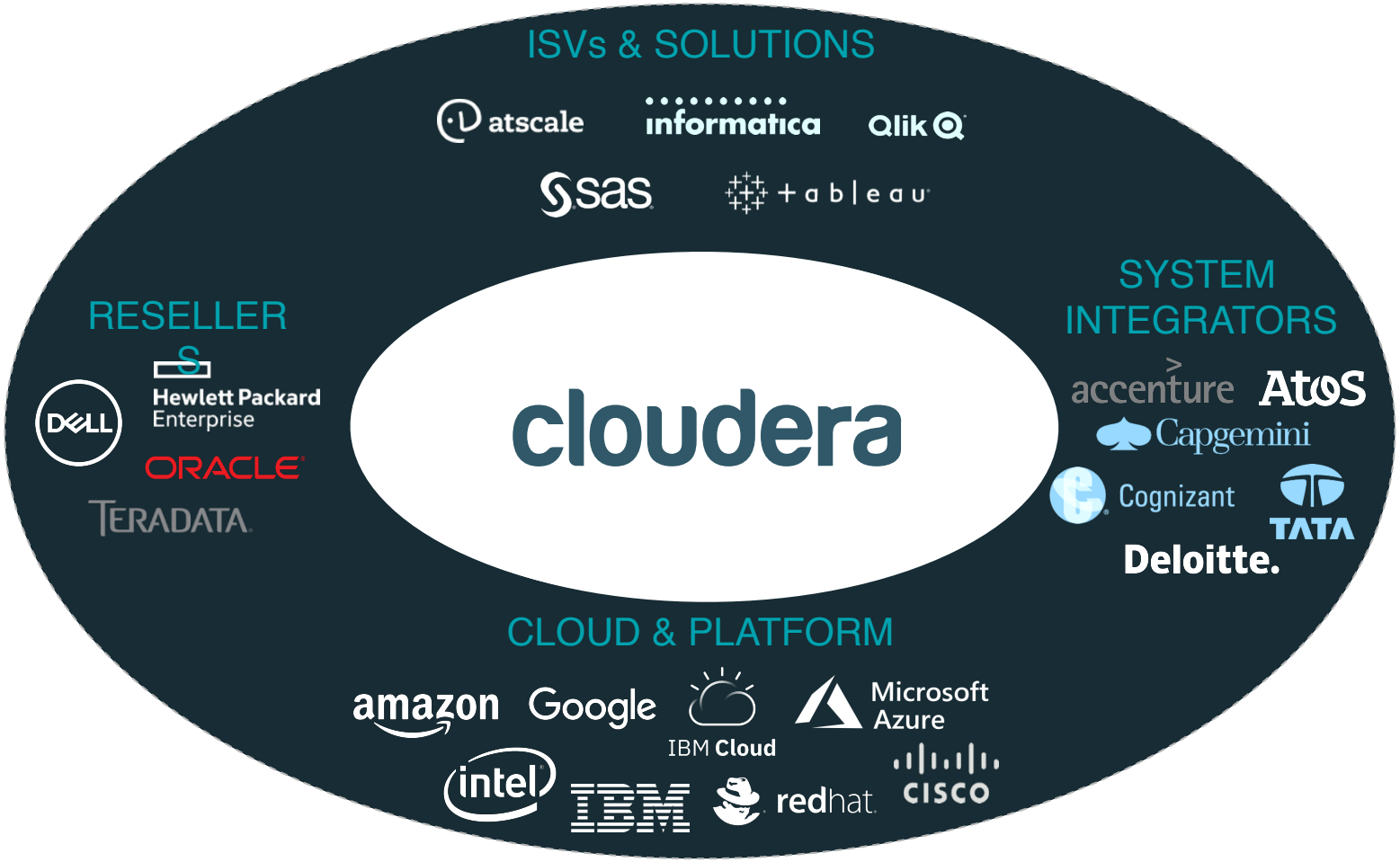
Available today, your first step towards a true EDC

- Analytics from the Edge to AI
- Common security & governance
- Public, hybrid & datacenter
- Powered by open source



# 3000+ PARTNER ECOSYSTEM

Strategic partnerships for expanded reach, integrated solutions and richer technical and industry expertise



# LEADING IN TOP INDUSTRIES

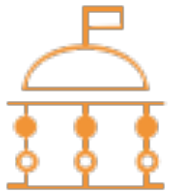
8/10
TOP GLOBAL

BANKING

10/10
TOP GLOBAL

TELCO

40+
GOVERNMENT CUSTOMERS



PUBLIC

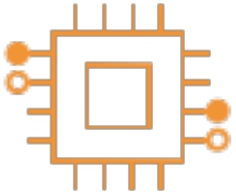
10/10
TOP GLOBAL

AUTOMOTIVE

9/10
TOP GLOBAL

PHARMA

8/10
TOP GLOBAL



TECHNOLOGY

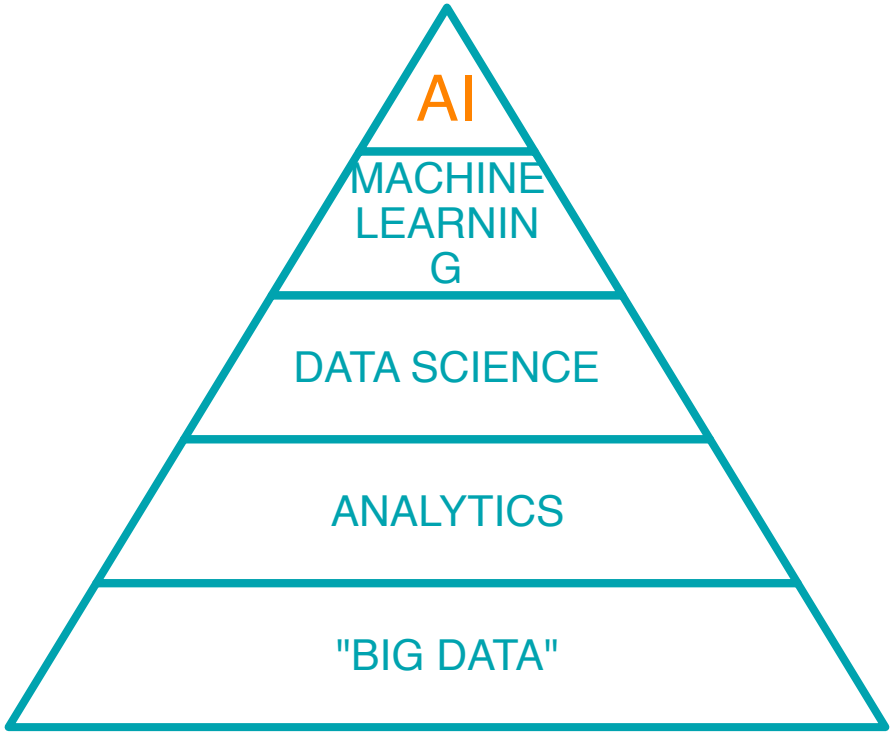
# KEY DIFFERENTIATORS

## What I want you to remember after today

- Powered by Open Source
- Analytics from the Edge to AI
- Common security and governance
- Public, Hybrid & Data Center

# HIERARCHY OF NEEDS FOR THE DATA-DRIVEN ENTERPRISE

## The “AI Ladder”

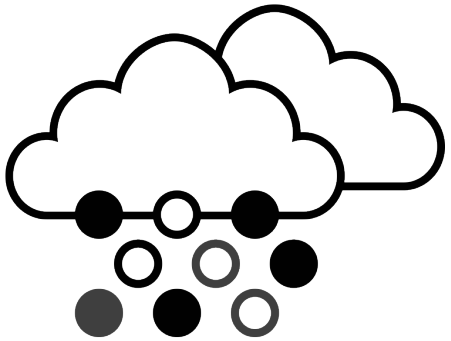



### DATA SCIENCE EXPERTS BEST PRACTICE

- 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



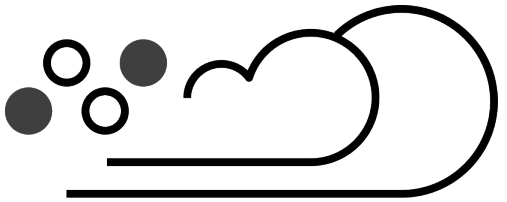
Any Cloud

EDGE 2AI

Multi-Function

CLOUDERA sdx shared data experience

Secure & Governed

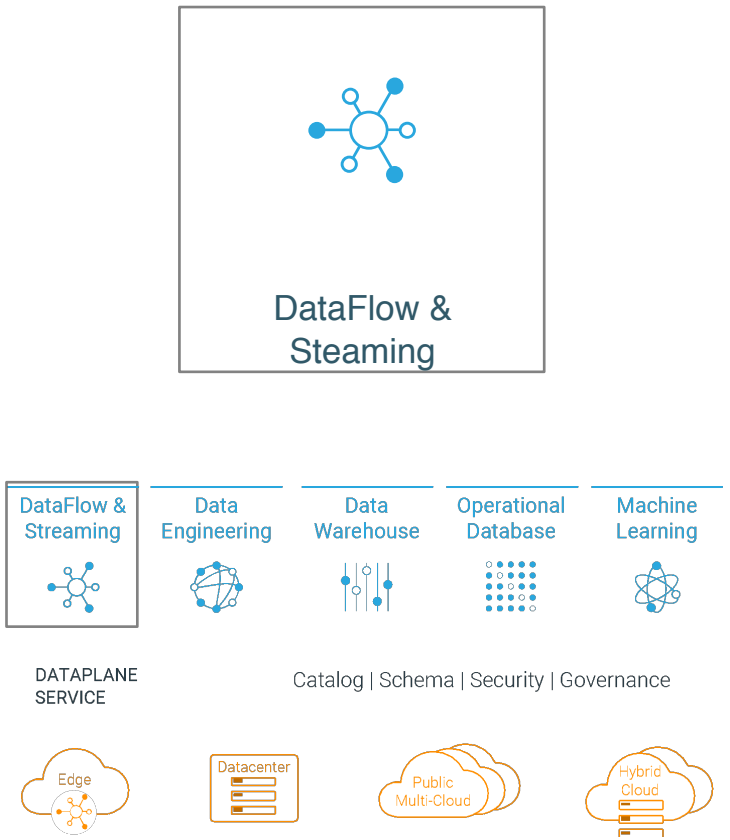


Open

# MULTI-FUNCTION ANALYTICS

# MANAGE DATA-IN-MOTION FROM EDGE-TO-ENTERPRISE

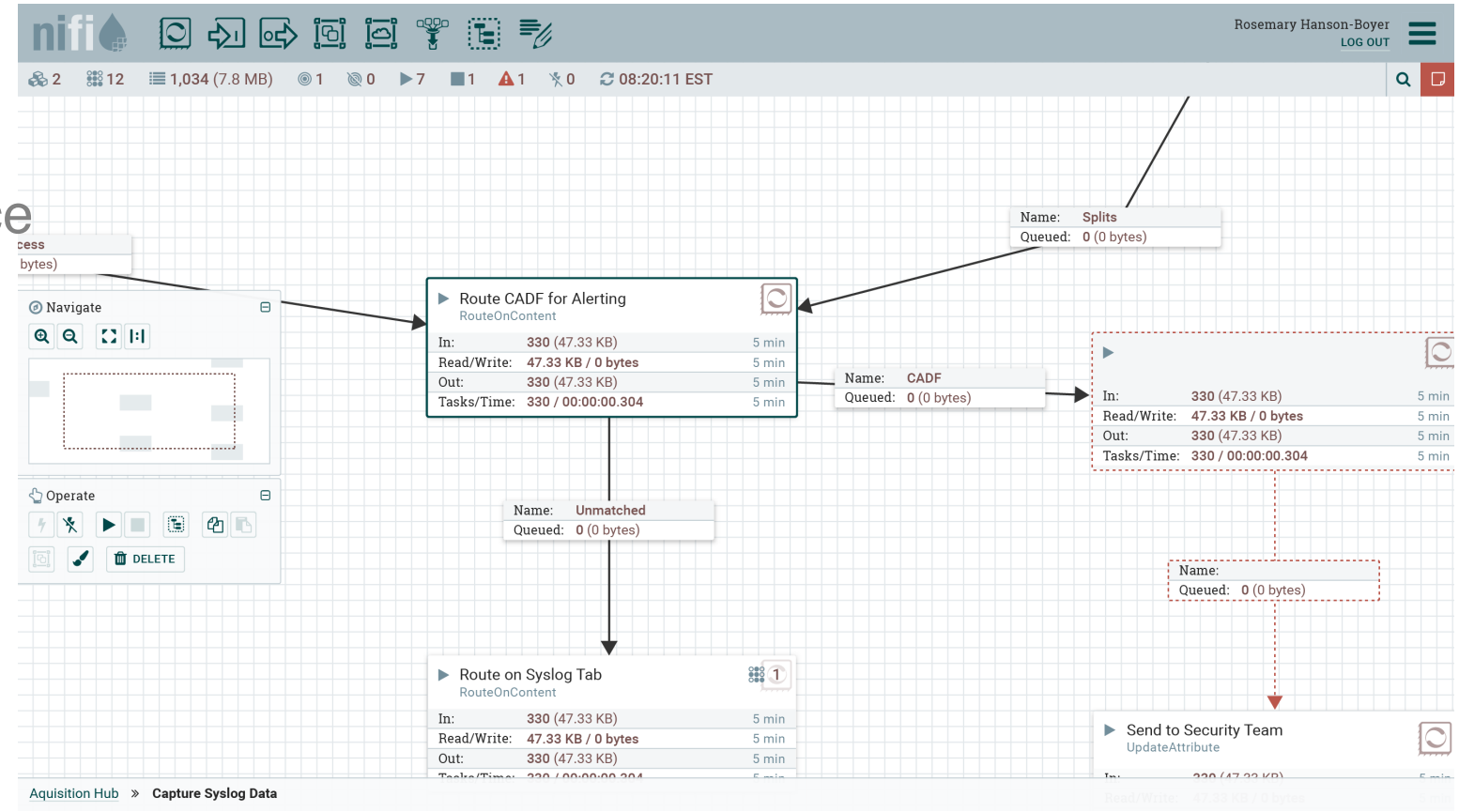
## Cloudera DataFlow - Collect, Curate and Analyze Data-in-Motion



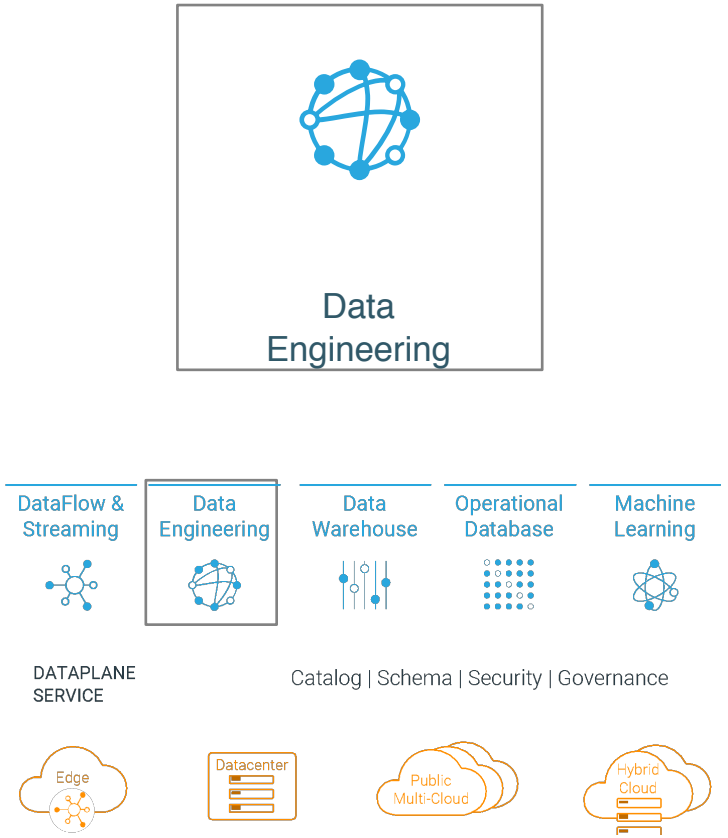

- Edge-to-enterprise streaming data platform for management, security and governance of real-time streaming data
- Edge data collection, processing and content routing of sensor data from edge devices
- Continuous data ingestion from any streaming source or IoT device
- Ease-of-use in building sophisticated data flows with drag-and-drop user interface
- Real-time stream processing and content syndication at the scale of millions of messages per second
- Predictive and prescriptive analytics from streaming analytics engines to gain actionable intelligence

# 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

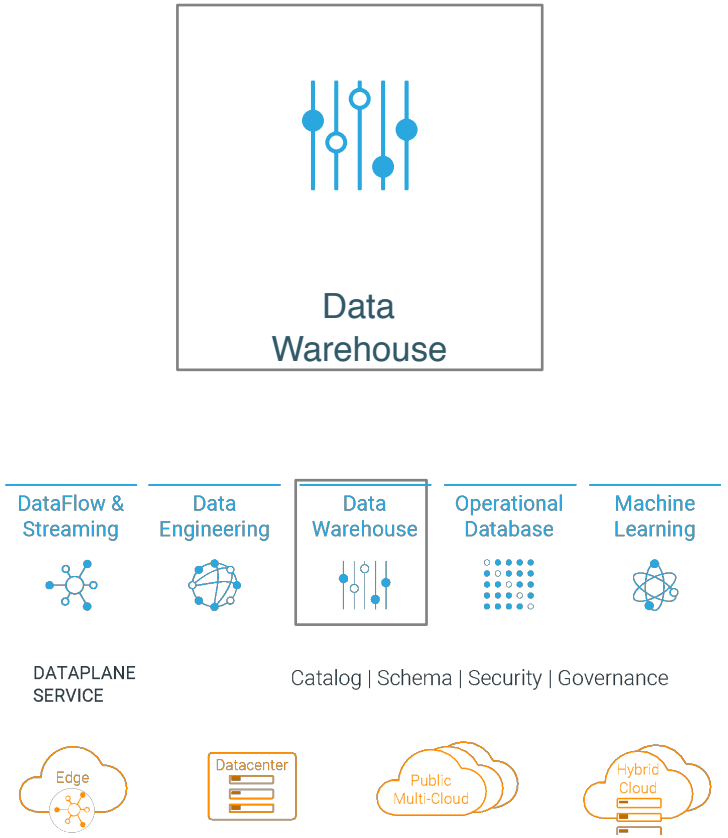


# STREAMLINE AND SIMPLIFY BIG DATA PROCESSING




- Integrated batch and streaming
- Powerful full-text search
- Familiar user languages – SQL, Java, Python & Scala
- Fault-tolerant and high-performance processing of continuous streams of data

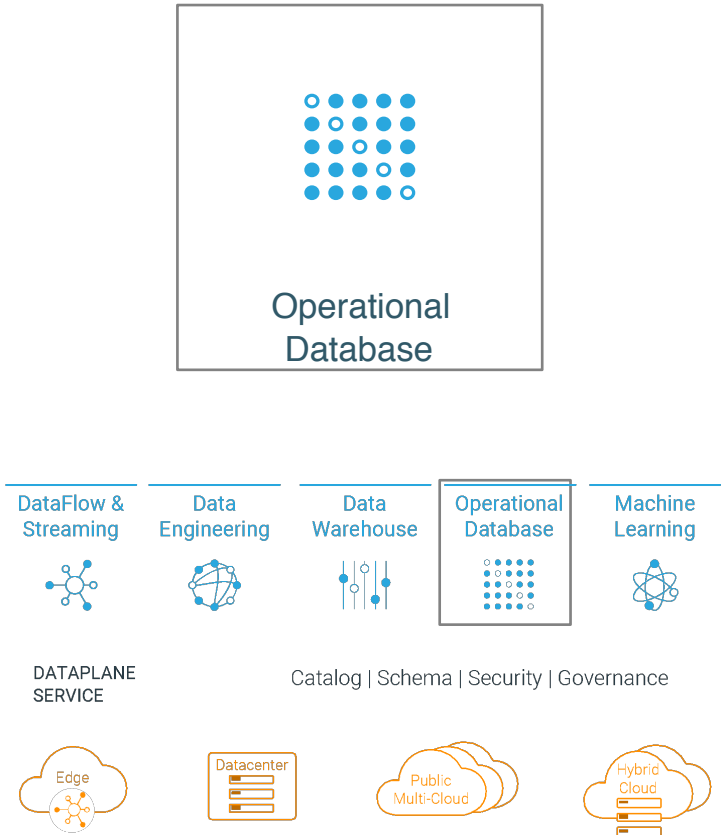
# UNLEASH HIGH-PERFORMANCE SQL ANALYTICS




**Modern Enterprise-Grade Data Warehouse with:**

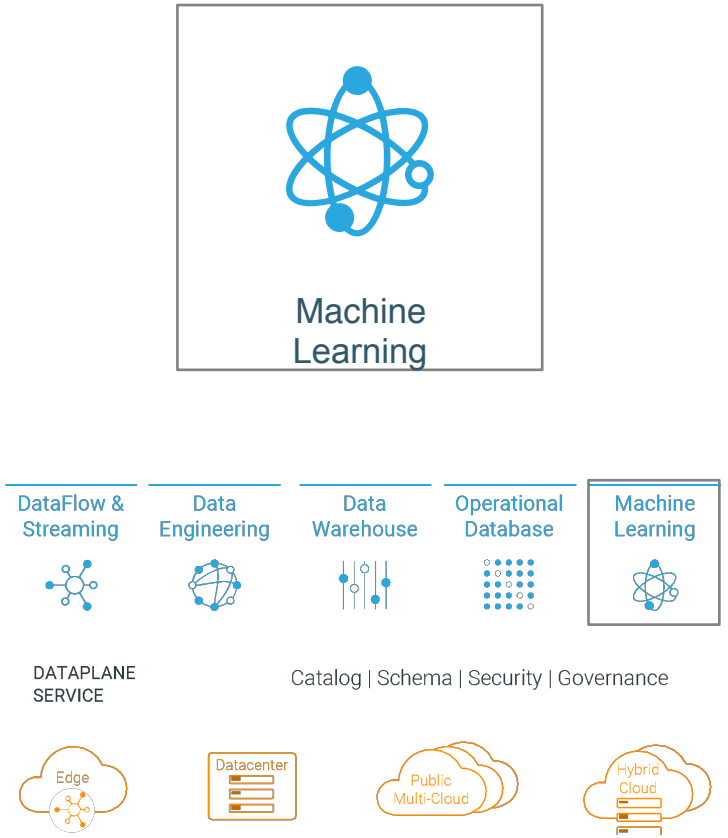
- Self-service high-performance SQL for exploration and discovery
- Flexibility to iterate with more data and more use cases
- Go beyond SQL for shared data with open standard tools for machine log,  IoT and text search
- Data warehouse optimization of your existing databases
- Cost-effectively scale with on-premises and hybrid multi-cloud

# INSTANT INSIGHTS WITH CLOUDERA OPERATIONAL DATABASE



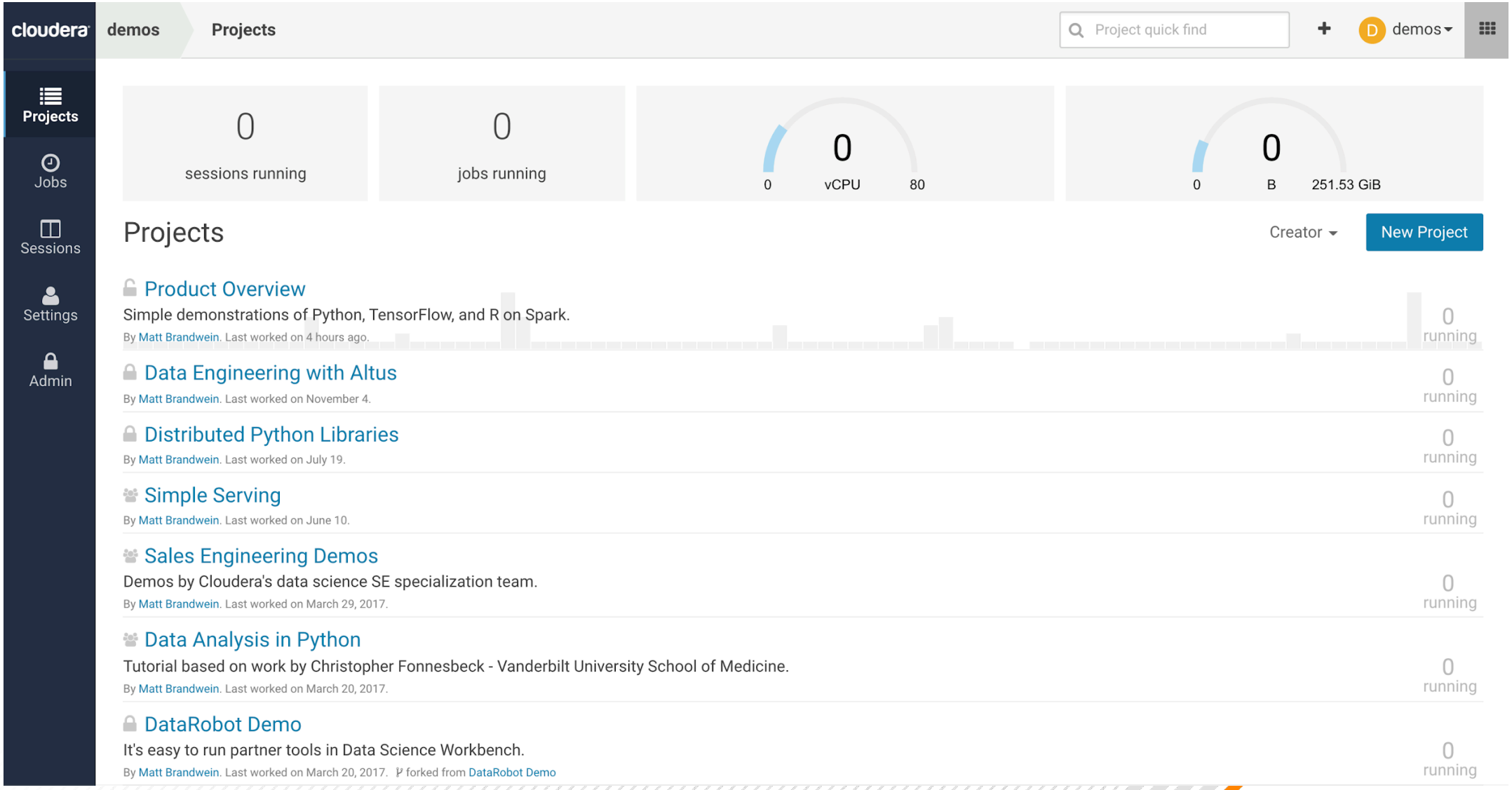

- Build and score models on operational data for prevention, optimization, prescription, and prediction.
- Inject real-time data and analysis into decision points across organization.
- Achieve operational excellence by reducing TCO, improving efficiency, eliminating threats.
- Leverage IoT to evolve your business model and operations for greater efficiencies
- Address far-reaching effects of shift in consumer expectations by enabling holistic view of your business.

# ACCELERATE ENTERPRISE DATA SCIENCE

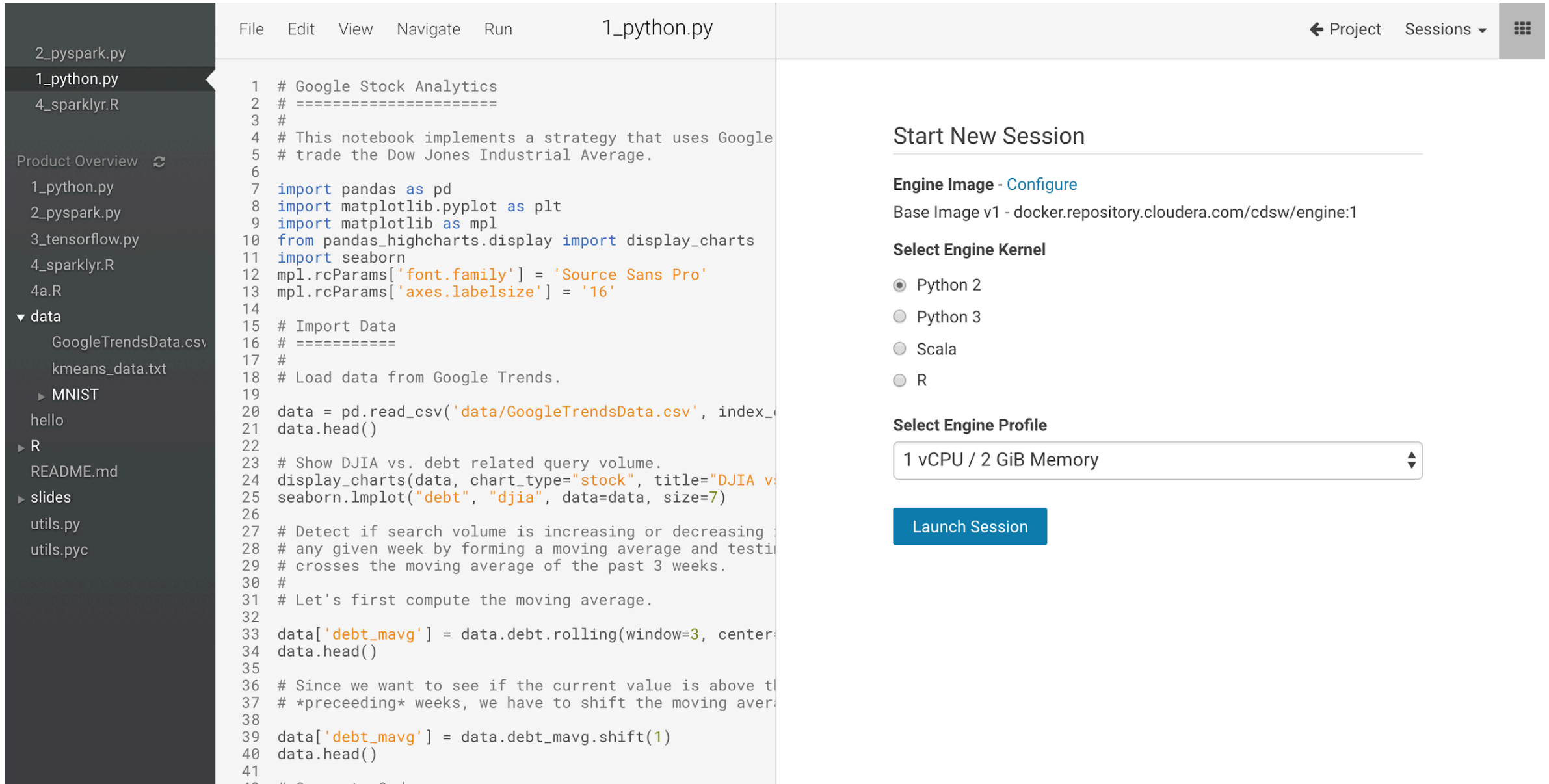



- Unified platform for data engineering and data science eliminates silos, speeds time to value -- on premises or in the cloud
- Secure, self-service access to enterprise data in corporate clusters, or data anywhere, makes data scientists more productive
- Elastic on-demand provisioning delivers the computing power data scientists need for the most demanding analysis, including GPU accelerated Deep Learning training
- Python, R, and Scala support brings the flexibility, innovation, and value of open source machine learning to the data
- Containerized environments provide collaboration, sharing, reproducibility, and project isolation for eliminating dependency management concerns

# CLOUDERA DATA SCIENCE WORKBENCH CDSW

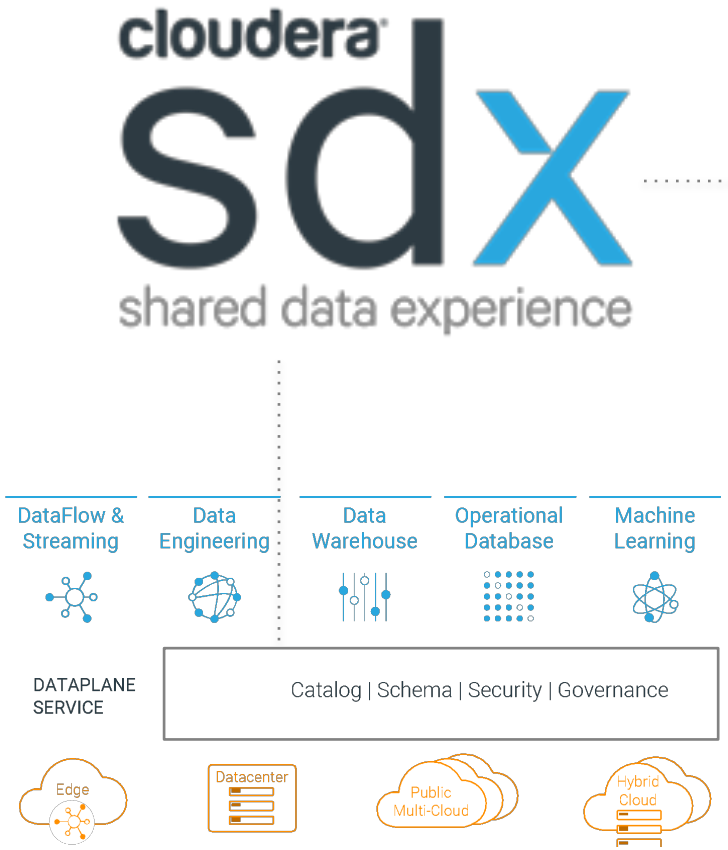


# K8 Docker container



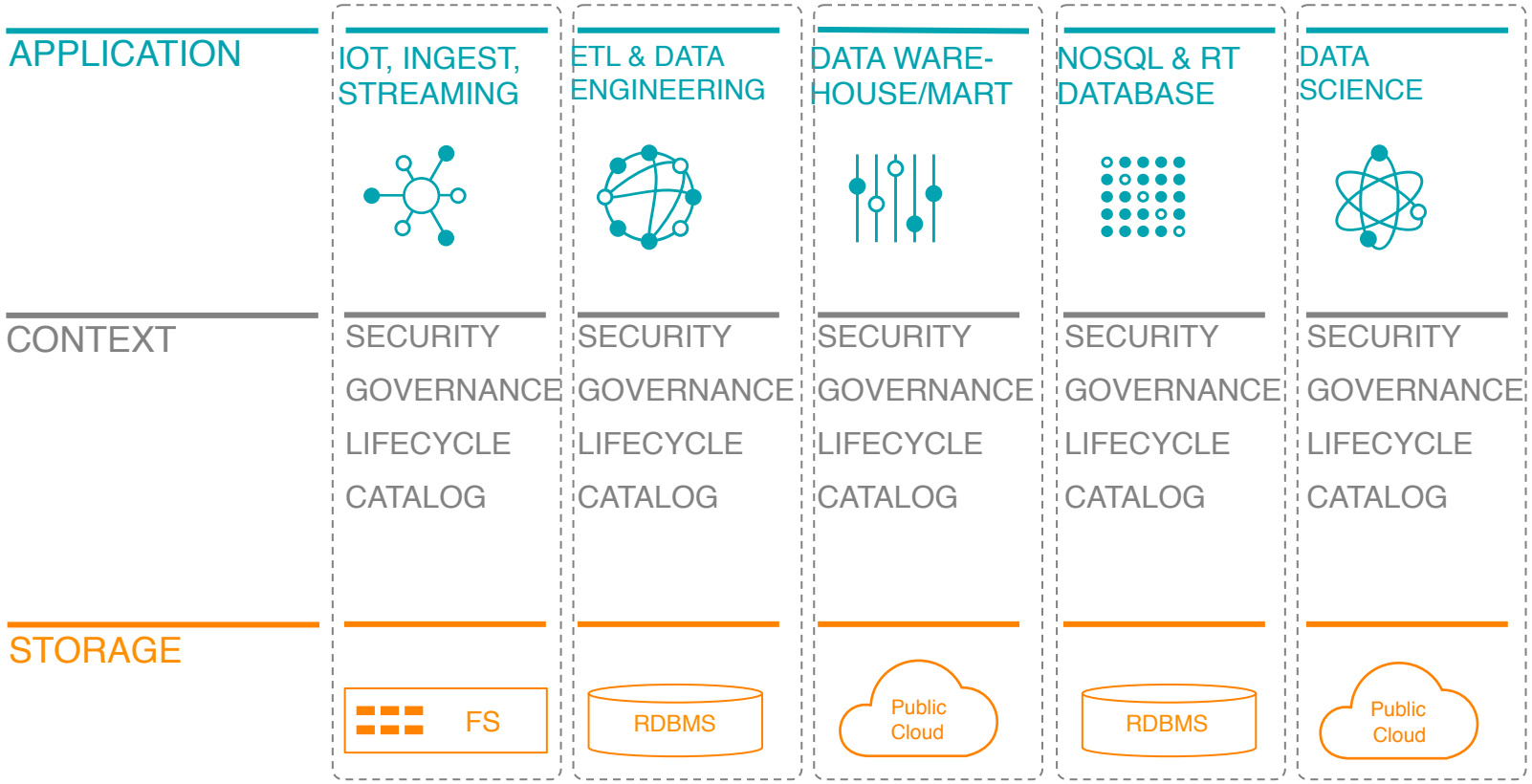
# CONSISTENT SECURITY AND GOVERNANCE

## 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 use and created by platform workloads
- 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

# WHY A CONSISTENT CONTEXT MATTERS



# ENTERPRISE MANAGEMENT AND SECURITY

# CLOUDERA MANAGER

Powerful cluster operation, management and administration, so you can focus on driving insight

## Trusted for production

- Reliability across environments
- Manage at massive scale
- Extensive use across customers

## Refined simplicity

- Intelligent defaults
- Custom visibility, metrics, and alerts
- Unique point-in-time troubleshooting

## Forefront of innovation

- Extensible foundation for seamless third-party integration
- Fastest support for state-of-the-art components
- Continuous usage-driven improvements

# CLOUDERA NAVIGATOR

Integrated data management and governance for your Cloudera platform

Single point of call for all things governance, curation, discovery and stewardship

## Self-service data discovery and analytics

- Unified metadata repository with full-text search and SQL access for effective data discovery and exploration
- Easily find similar and relevant data with business context and classification data set location

## Compliance ready governance and protection

- Complete and automatic data access auditing
- Data origin and impact with end-to-end column-level lineage
- Full data protection with encryption and key management

## Data lifecycle automation

- Automated stewardship and curation with flexible policy engine
- Business continuity with built-in backup and disaster recovery

# CLOUDERA ALTUS DIRECTOR

The fastest, easiest way to run Cloudera Enterprise across cloud environments

## Multi-cloud support

- Ready to go for AWS, GCP and Azure
- Open Cloud Connector framework for limitless extensibility

## Cluster lifecycle support

- Elastic compute and object storage support
- Kerberos bootstrapped and HA enabled
- Customizable, templated cluster configurations
- Any-scale cluster support

## Cluster management

- Multi-region deployment and management with single instance
- Support for AWS spot instances and GCP preemptible
- Automated billing and metering with consumption pricing model
- Tightly coupled with Cloudera Manager for robust monitoring and troubleshooting

# SECURITY & GOVERNANCE CAPABILITIES

The market leader for regulated industries and privacy compliance

All compliance “must haves” for production strength deployments

| Cloudera Enterprise |
| --- |
| ✓ Encryption: All sensitive data encrypted at rest and in motion<br>Cloudera Navigator Encrypt & Key Trustee |
| ✓ Authorization: Access to data is limited by job role<br>Apache Sentry / Cloudera Data Science Workbench |
| ✓ Audit: All user actions must be logged (for forensics use)<br>Cloudera Navigator |
| ✓ Visibility: All data must be classified according to sensitivity<br>Cloudera Navigator |
| ✓ Erasure: Individual records can be updated or deleted (GDPR)<br>Apache Kudu |

# DEPLOY ANYWHERE

# MULTIPLE DEPLOYMENT OPTIONS

DATA ENGINEERING

DATA WAREHOUSE

OPERATIONAL DATABASE

MACHINE LEARNING


On-premises or public cloud infrastructure

DATA ENGINEERING

DATA WAREHOUSE


Managed services

# COMPARISON OF CLOUD OPTIONS

## Breadth of solutions in our Portfolio (not mutually exclusive)

### Altus Services

- Use our **cloud services** for DataEng and Data Warehouse solutions
- Customer has limited / no control of the cluster
- Customer has very prescriptive choices for infrastructure

### CDH (via Director)<br>HDP (via Cloudbreak)

- Use our **cloud provisioning tools** to launch and run on a Cloud Platform
- Customer has full control of the cluster
- Tool provides customer with prescriptive path for the infrastructure

### CDH (via CM)<br>HDP (via Ambari)

- Use our **management tools** directly and run on a Cloud Platform
- Customer has full control of the cluster
- Customer has full control of the infrastructure independent of tool

# RUN IN THE ENVIRONMENT THAT MEETS BUSINESS NEEDS

| | Non-Cloud | Private Cloud | Public Cloud Infrastructure | Managed Services |
|---|---|---|---|---|
| I want to maximize | • Cost-efficiency | • Control, elasticity, and convenience | • Control, elasticity, and convenience | • Agility |
| I want to minimize | • Dependence on unproven technology | • Resource contention between departments | • Dependence on data center floor space | • Dependence on IT and therefore need as simple as possible |
| I want to standardize | • On whatever provides the best ROI | • On a single environment for the entire data center | • On a single cloud provider for all infrastructure needs | • On whatever is easiest to use |
| I want to store my data | • On premises because cheaper and/or more secure | • On premises due to company / government mandate | • In the cloud because easier | • In the cloud because easier |

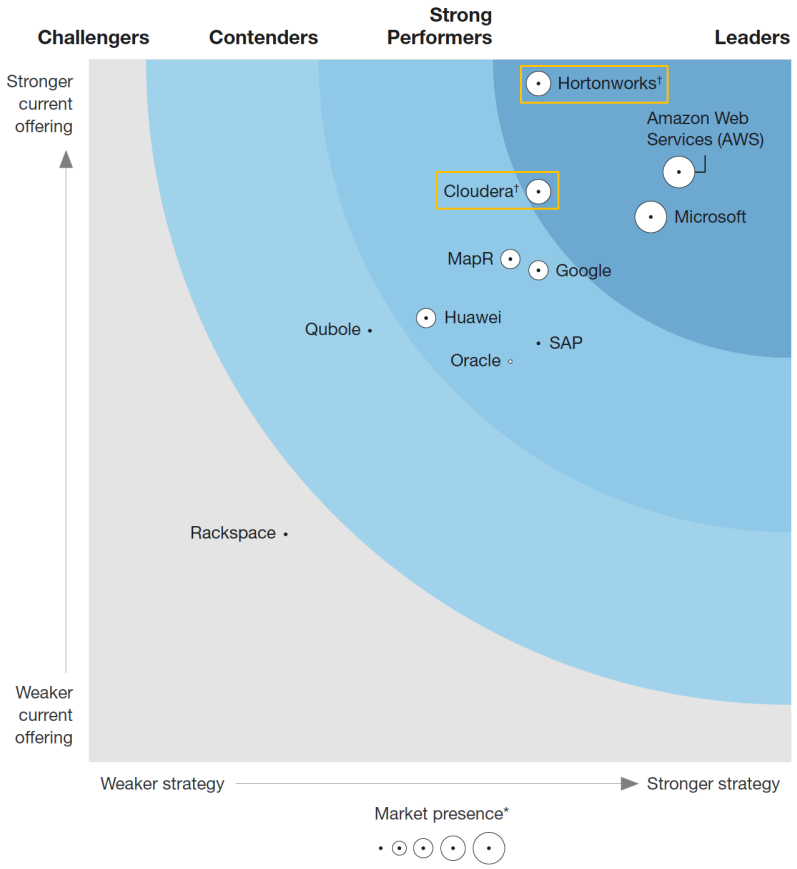
# COMPETITIVE

# HOW WE STACK UP

| | Cloudera | SaaS solutions (Snowflake, Databricks) | Cloud vendors (AWS, Azure, GCP) | Legacy data platforms (Oracle, Teradata) |
|---|---|---|---|---|
| **Multiple analytic functions** | ✔ | | ✔ | ✔ |
| **Multi-cloud** | ✔ | ✔ | | ✔ |
| **Hybrid** | ✔ | | ✔ | |
| **Open** | ✔ | ✔ | | |
| **Consistent security, governance, and catalog applied everywhere** | ✔ | | | ✔ |

# PLATFORM LEADERSHIP

## Forrester Wave – Cloud Hadoop/Spark (HARK) Platforms Q1 2019




- HARK platforms
  - Accelerate getting insight from data
  - Exemplify multi-functional analytics
- Cloudera present in three out of the four leaders
  - Legacy Cloudera, legacy Hortonworks, Hortonworks as part of HDInsight
- Cloudera is the only leader to support multi-cloud

# NEW IN CLOUDERA ENTERPRISE 6.2

# CLOUDERA 6 IS GIANT LEAP FORWARD TO OPEN SOURCE CORE

HADOOP 3.0

HIVE 2.1.1

HBASE 2.1.2

SPARK 2.4

PARQUET 1.9

SOLR 7.4

OOZIE 5.1

SENTRY 2.4

KAFKA 2.1

AVRO 1.8.2

FLUME 1.9

SQOOP 1.4.7

HUE 4.3

CLOUDERA MANAGER 6.2

CLOUDERA NAVIGATOR 6.2

CLOUDERA DIRECTOR 6.2

# CLOUDERA 6.2 RELEASE HIGHLIGHTS

- Management
  - SDX support in Cloudera Manager
  - BDR replication to cloud object store
  - YARN GPU scheduling
- Governance
  - Navigator metadata enhancements
- Security
  - HMS metadata read authorizations
- Search, query, access
  - Improved HUE troubleshooting
  - New Impala guardrail and zero-touch metadata (preview)
  - Hive parallel query compilation and improved Connection Pool Agents configuration
- Platform
  - Ubuntu 18 support
  - Google Cloud Storage, ADLS Gen 2 support

# OPERATIONAL DATABASES

## HBase serial replication

- Updates are now delivered to end points in order they were made
- Prior to 2.1.2, HBase replication was eventually consistent but changed might be received out of order

## HBase Intel Optane Support

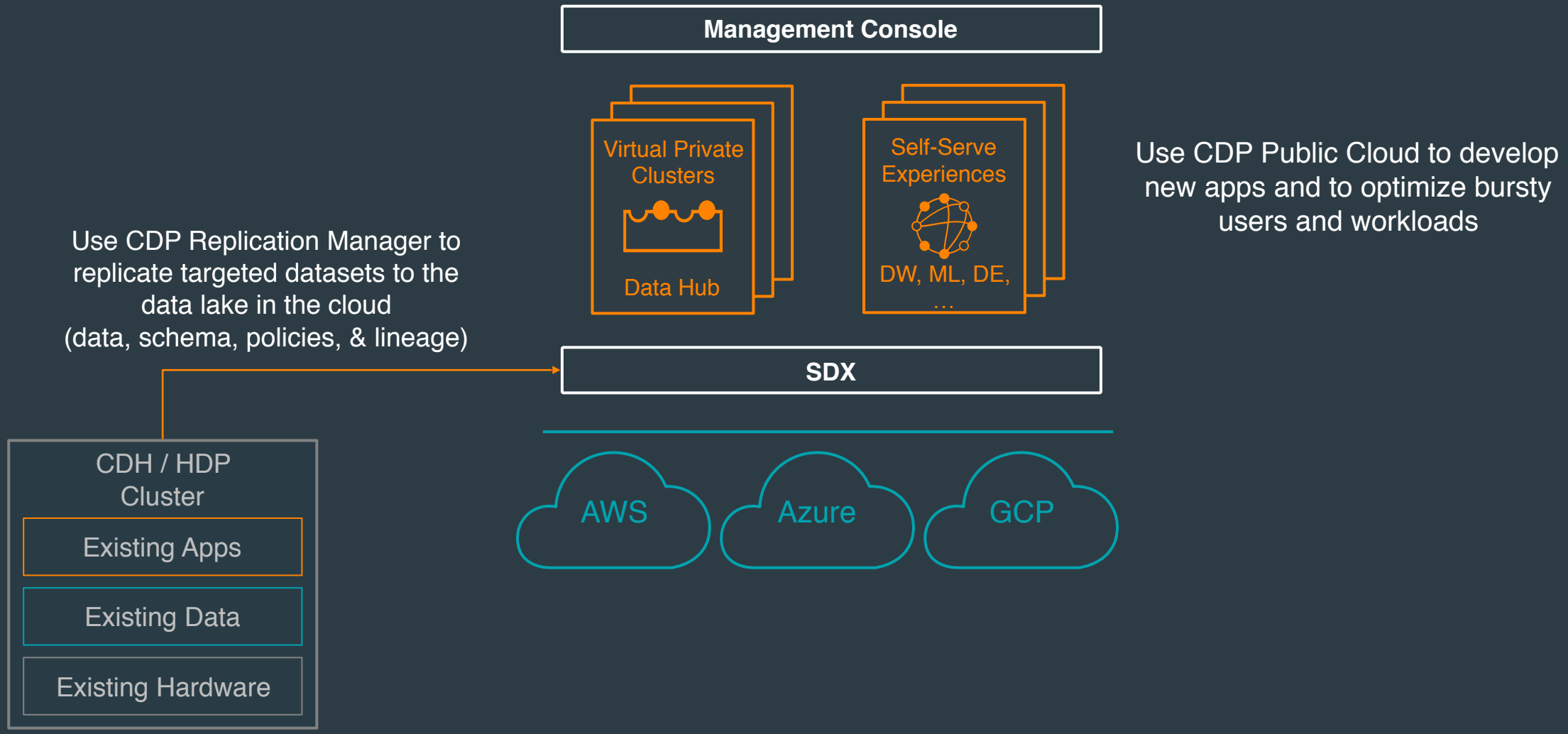
- Intel Optane DC persistent memory support allows for larger BucketCache than DRAM, improving overall performance

## Sqoop Updates

- Support for ADLS Gen 1 and Gen 2 object stores
- Support for DECIMAL type for Parquet import and export
  - Only enabled on new clusters to avoid breaking compatibility on old ones

# CDP ROADMAP
## The road to cloud

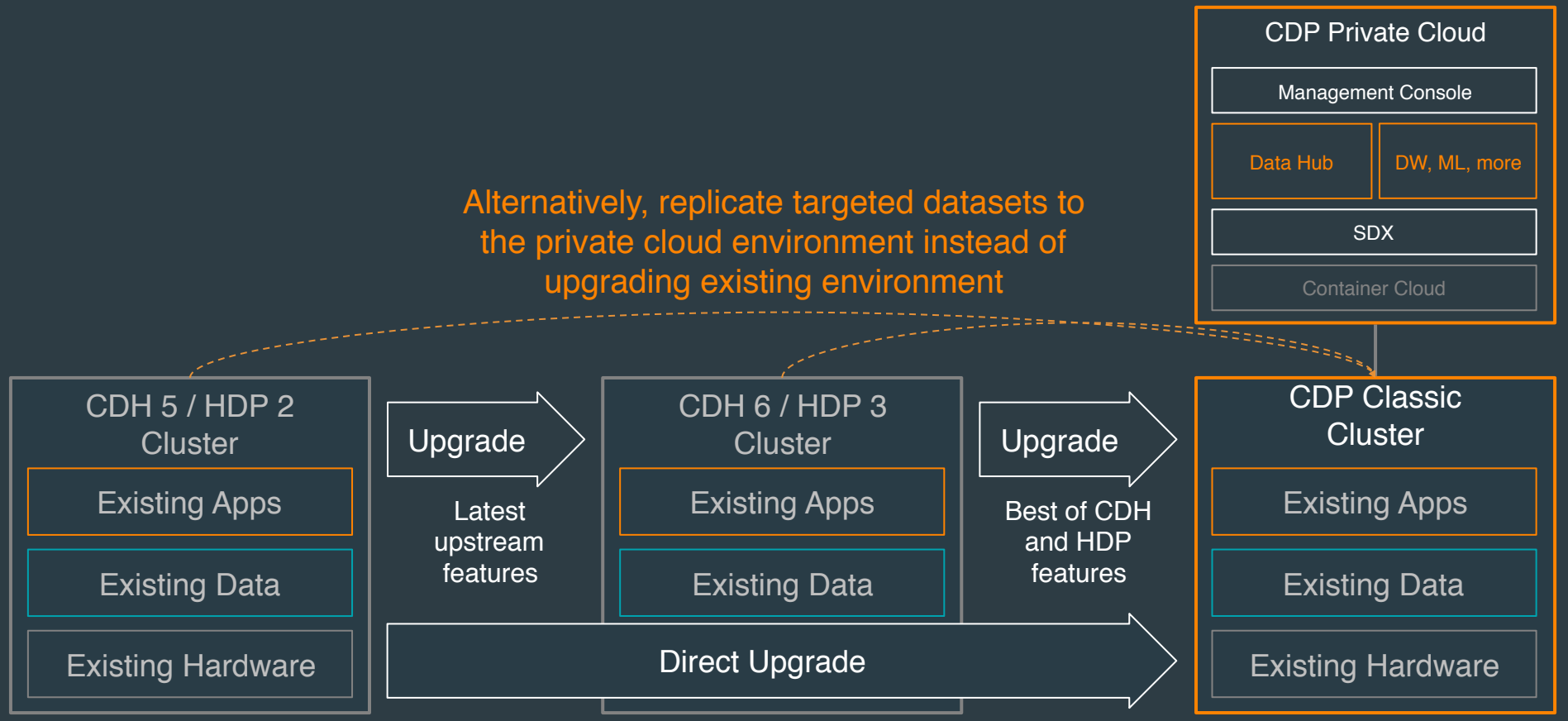
# THE UPGRADE PATH FOR PUBLIC CLOUD




Use CDP Replication Manager to replicate targeted datasets to the data lake in the cloud (data, schema, policies, & lineage)

Use CDP Public Cloud to develop new apps and to optimize bursty users and workloads

# THE UPGRADE PATH FOR PRIVATE CLOUD




Use CDP Private Cloud to develop new apps and to optimize bursty users and workloads

Use CDP Classic for existing apps on existing hardware

# CALL TO ACTION

## Next steps on your Cloudera journey

1. Define you goals and initiatives
2. Explore a demo, proof of value or pilot
3. Deep dives into the technology

CLOUDERA

*the enterprise data cloud company*

# THANK YOU

CLOUDERA