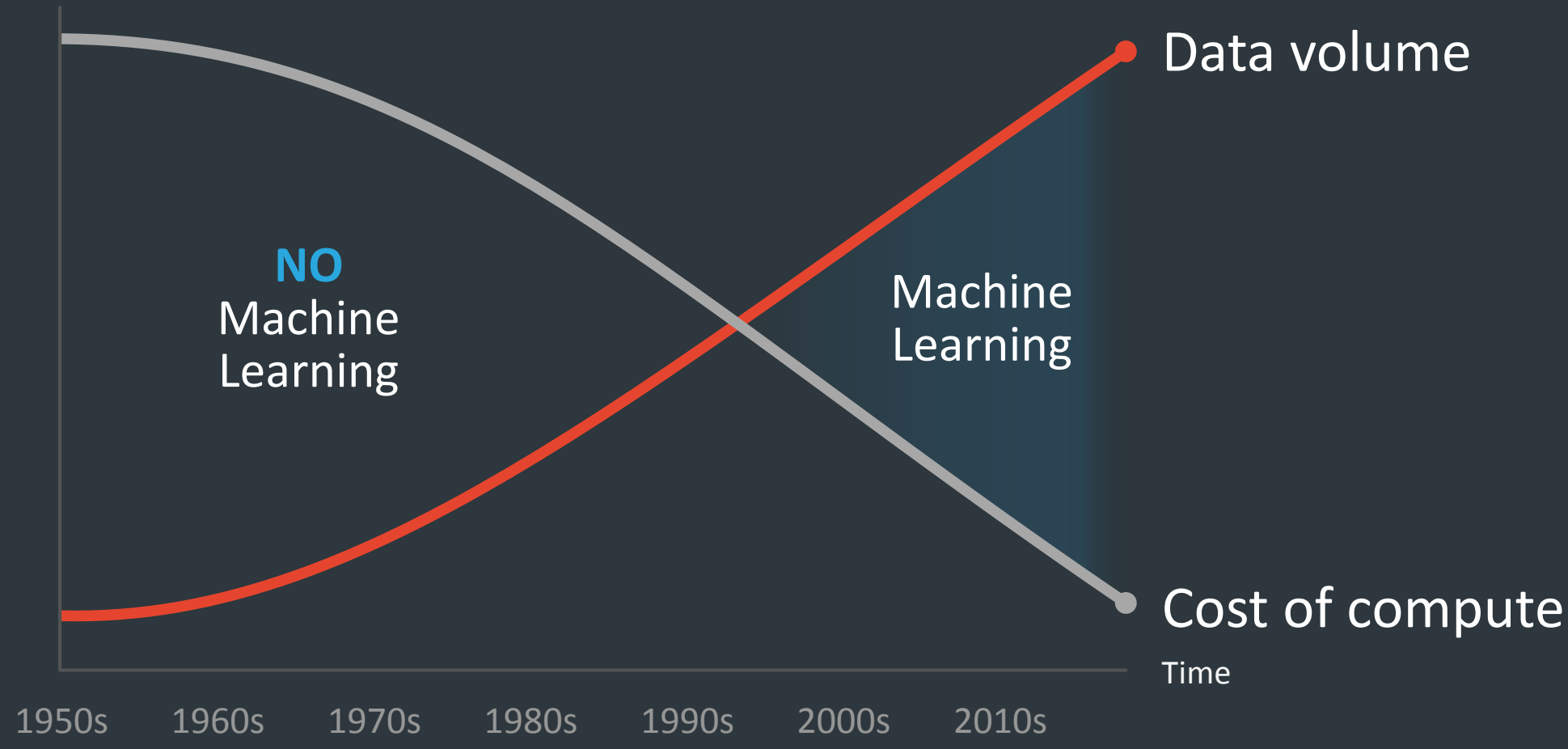
cloudera

# Cloudera Data Science and Machine Learning

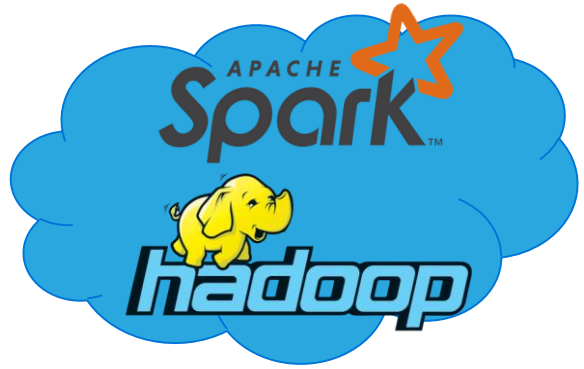
Robin Harrison, Account Executive
David Kemp, Systems Engineer

# This is the age of machine learning.



# Machine learning presents a multitude of opportunities

| Data | Analytics | Deployment |
| --- | --- | --- |


Data has never been more plentiful

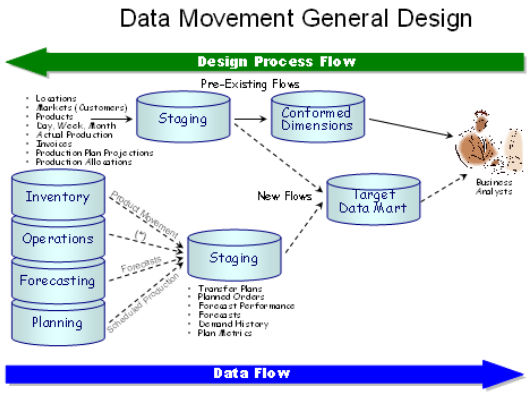
Open source data science and machine learning libraries are rapidly evolving

Flexible commodity storage and compute make scalable production machine learning affordable

cloudera

# But there are practical challenges

| Data | Analytics | Deployment |
|---|---|---|




Data volumes are increasing and it needs to move across multiple different systems

Teams have different, conflicting requests for languages & libraries

Most data science done at small scale, individually, and is difficult to replicate

Very few models reach production

cloudera

# What is Machine Learning and Data Science

- Machine Learning and Data Science:  algorithms and methods that extract useful insights and patterns from data.

- These insights can drive profits, find outliers, cluster like items, predict future issues and insights, cut losses, classify different groupings and many other tasks.

# What does a Data Scientist Do?


cloudera

# Types of data science

## Exploratory
(discover and quantify opportunities)



## Operational
(deploy production systems)

- **Team:** Data scientists and analysts
- **Goal:** Understand data, develop and improve models, share insights

- **Data:** New and changing; often sampled
- **Environment:** Local machine, sandbox cluster
- **Tools:** R, Python, SAS/SPSS, SQL; notebooks; data wrangling/discovery tools, …
- **End State:** Reports, dashboards, PDF, MS Office

- **Team:** Data engineers, developers, SREs
- **Goal:** Build and maintain applications, improve model performance, manage models in production

- **Data:** Known data; full scale
- **Environment:** Production clusters
- **Tools:** Java/Scala, C++; IDEs; continuous integration, source control, …
- **End State:** Online/production applications

cloudera

# Our goal: Open data science at enterprise scale

## Help more data scientists use the power of Cloudera

Use a powerful, familiar environment with direct access to Cloudera data and compute

Data Scientist
Data Engineer

## Make it easy and secure to add new users, use cases

Offer secure self-service analytics and a faster path to production on common, affordable infrastructure

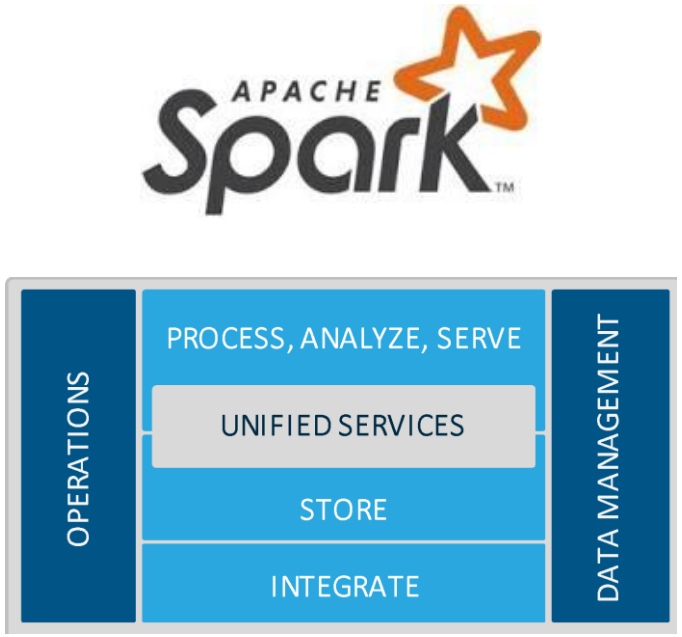
Enterprise Architect
Hadoop Admin

cloudera

# Balancing the needs of data scientists and IT

## Data Scientists
explore, experiment, collaborate


## IT
drive adoption, maintain compliance



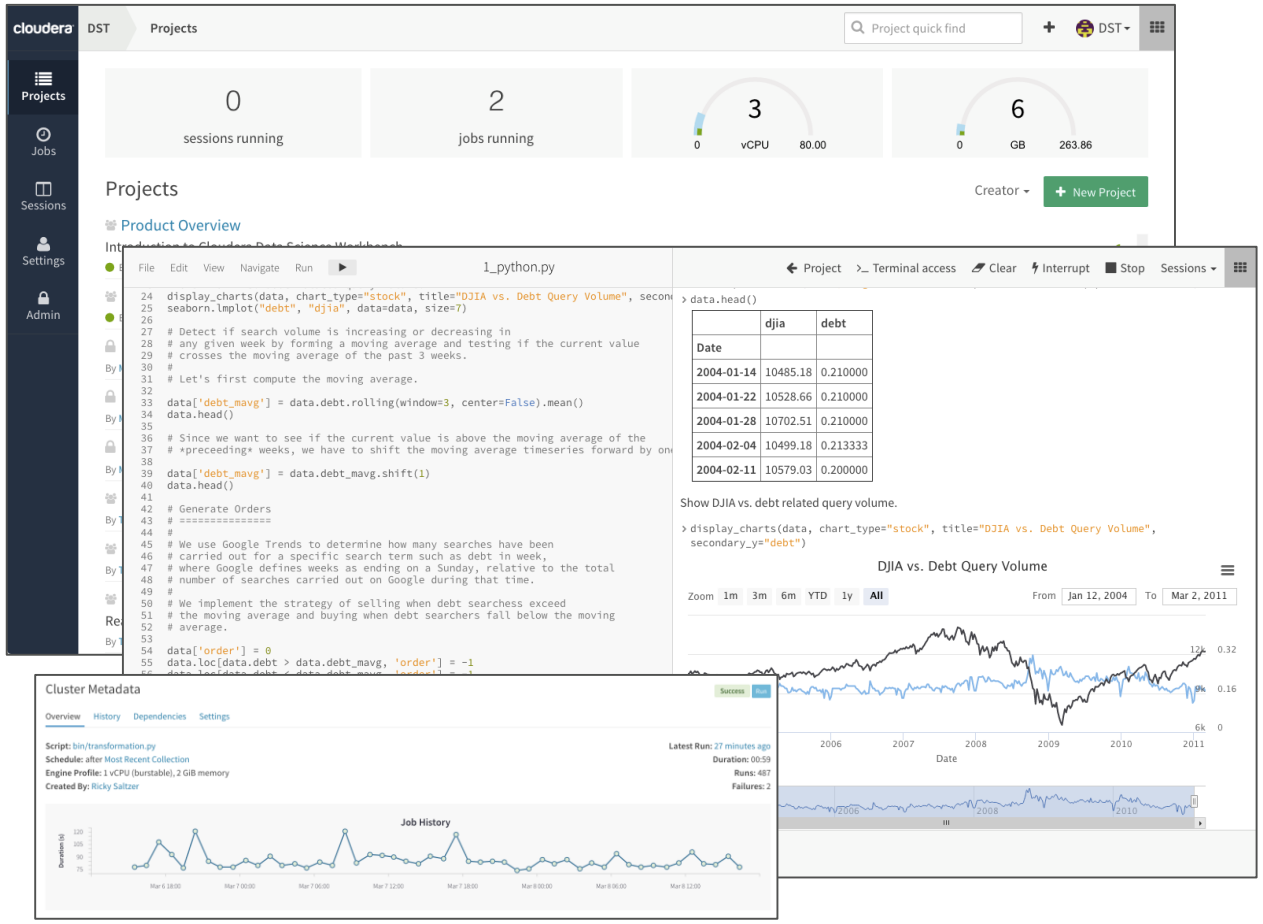

cloudera

# Cloudera Data Science Workbench

Self-service data science for the enterprise

Accelerates data science from development to production with:

- Secure self-service data access
- On-demand compute
- Support for Python, R, and Scala
- Project dependency isolation for multiple library versions
- Workflow automation, version control, collaboration and sharing




cloudera

# With Cloudera Data Science Workbench…

## Data scientists can:

- Use R, Python, or Scala from a web browser, with no desktop footprint
- Install any library or framework within isolated project environments
- Directly access data in secure clusters with Spark and Impala
- Share insights with their team for reproducible, collaborative research
- Automate and monitor data pipelines using built-in job scheduling

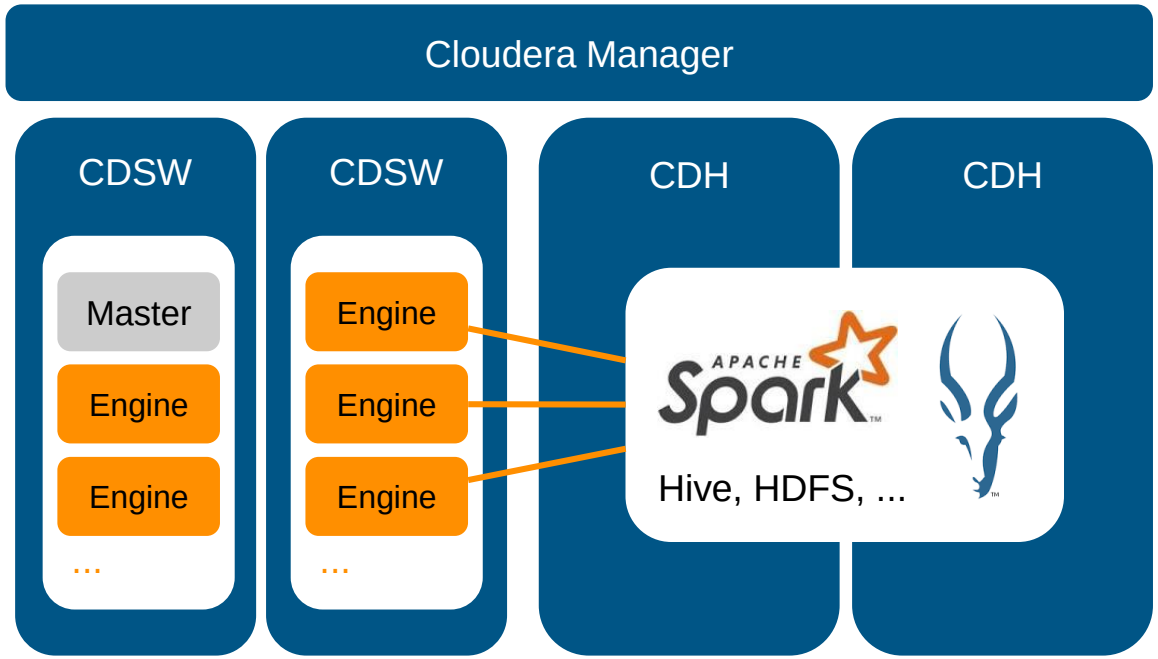
## IT can:

- Give their data science team the freedom to work how they want, when they want
- Stay compliant with out-of-the-box support for full platform security, especially Kerberos
- Run on-premises or in the cloud, wherever data is managed

cloudera

# A modern data science architecture

- Built on Docker and Kubernetes
- Runs on dedicated gateway nodes
- User sessions run in isolated "engine" containers which:
  - Host Kerberos-authenticated Python/R/Scala runtimes
  - Interact with Spark via YARN client mode (Driver runs in container, workers on CDH)
- Single-cluster only (for now)




cloudera

LOWER BUSINESS RISKS

CYBERSECURITY
» THREAT DETECTION
» DATA SECURITY
» MACHINE LEARNING

FireEye™

## Uncovering Zero-Day Attacks and Stopping Advanced Persistent Threats More Quickly

- Helps threat hunters obtain responses to queries magnitudes faster
- Provides access to a wider range of data that wasn't accessible before
- Increases researcher productivity by 60 percent

cloudera

IMPROVE PRODUCT & SERVICE EFFICIENCY

MANUFACTURING

» PREDICTIVE ANALYTICS

» PROCESS IMPROVEMENT

» PRODUCT INNOVATION

SIKORSKY
A LOCKHEED MARTIN COMPANY

## Improving Flight Safety with Rapid, Data-Driven Decision Support

- Uncovers patterns in aircraft performance and parts that can help Sikorsky engineers improve flight safety and optimize aircraft operations
- Extend useful life of key components
- Helps prevent unscheduled maintenance and better prioritize repairs

cloudera

CounterTack

## PROBLEM

Needed scalable system for real-time endpoint threat detection and response

- System couldn't handle growing number of endpoints
- No real-time processing
- Limited operational resources

## SOLUTION

Replatformed to deliver actionable security intelligence to users

- Support deployments with **>100,000 endpoints**
- Threat detection and **response in minutes** vs months
- Cloudera Predictive Support **anticipates issues before they occur** & direct connection to the experts

cloudera

CASE STUDY

DATA-DRIVEN PRODUCTS

AEROSPACE
» SPACECRAFT TELEMETRY
» REMOTE MONITORING
» PREDICTIVE MAINTENANCE

# Aerospace – Spacecraft Telemetry

## Advanced analytics on streaming data to reduce human space mission risks

### Challenge:

- Over **2 TB/ hour** of telemetry test data streaming in from over 1200 sensors in test environment

### Solution:

- Cloudera cluster supporting high rate of data ingest – up to **~300MB/sec**
- Advanced analytics run on the streaming data to check for issues or determine patterns and reduce risk

cloudera

TECHNOLOGY/ NETWORKING
» CUSTOMER SUPPORT
» NETWORK MONITORING
» IOT: REAL TIME PLANNING

# Connected Product Support

## Juniper Networks Monitors Thousands of In-Field Devices with Cloudera & Zoomdata

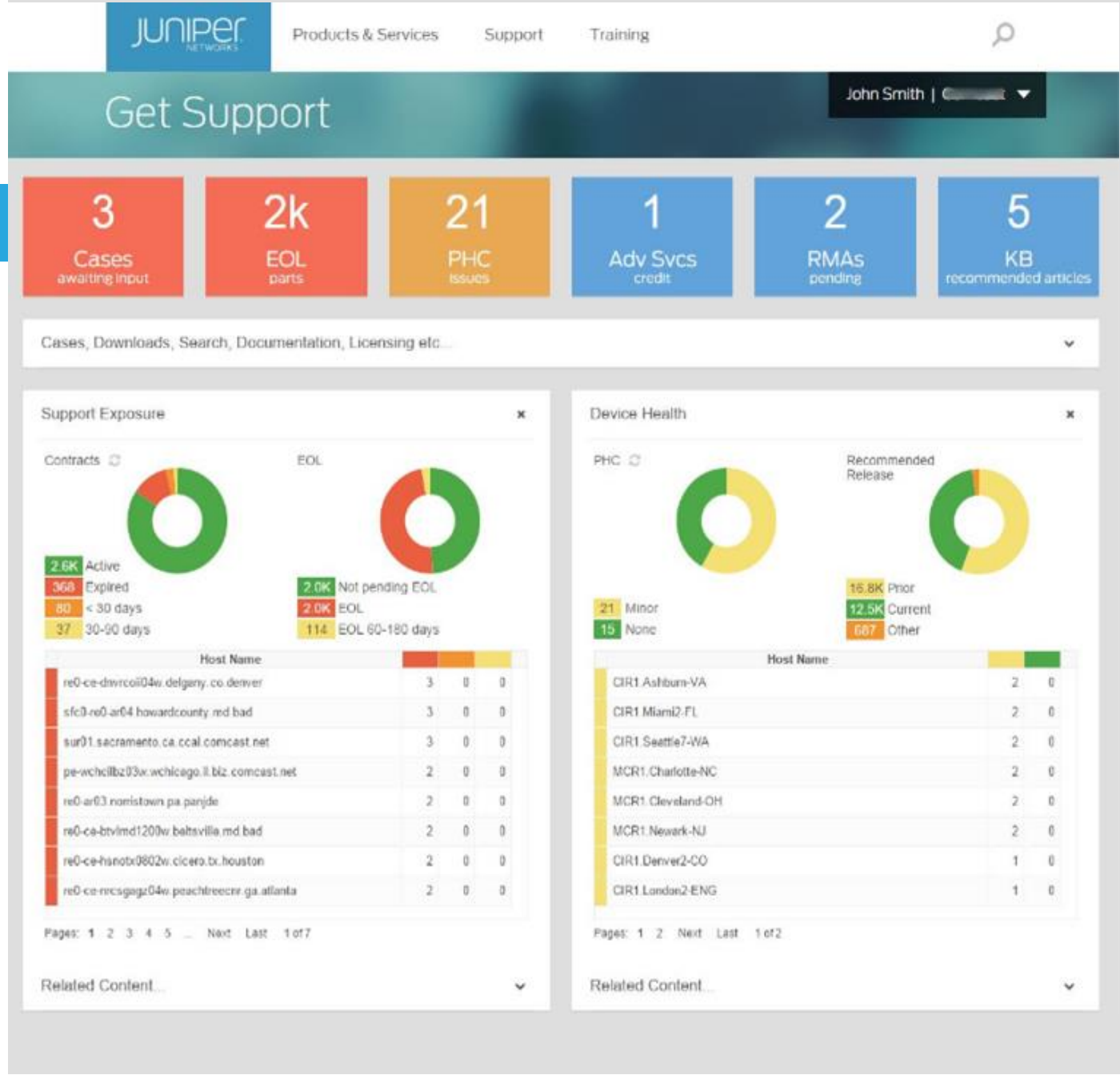
### Challenge:

- Monitor thousands of in-field devices in real time to provide the best support experience

### Solution:

- Cloudera Enterprise + Zoomdata provide every support client with access to both aggregate and detailed view of their devices
- Pinpoint issues at any level: network, device or application

cloudera | ZOOMDATA



cloudera

Thank you!

Robin Harrison, robinharrison@cloudera.com, 703-795-4706
David Kemp, david.kemp@cloduera.com, 703-282-2317