cloudera

Predictive Maintenance Demo

Samir Gupta, Sales Engineer

Agenda

- Why Cloudera for Data Science?
  - **Open Frameworks** - Apache Spark, Impala, Kudu, Kafka
  - **Enterprise Tools** - Cloudera Data Science Workbench, Apache HUE
  - **Cloud Integration** - automated deployment, object storage, bursting workloads
  - **Comprehensive Security** - auditing, governance, access control, encryption
- Demo - Predictive Maintenance on Cloudera
  - **Business Challenge**
  - **Technical Architecture**
  - **Data Profiling and Model Development**
  - **Real-time Ingest and Scoring**
  - **Visualization and Analytics**

© Cloudera, Inc. All rights reserved.

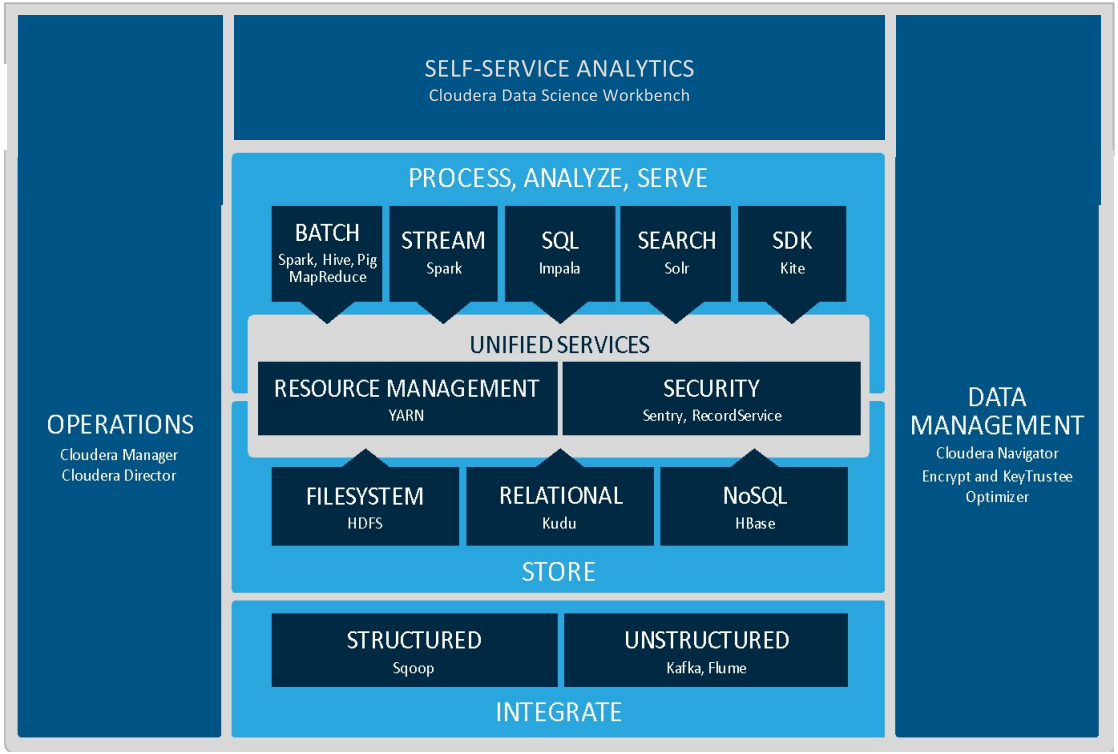
Cloudera Enterprise

Hadoop is a new kind of data platform.

- One place for unlimited data
- Unified data access

Cloudera makes it:

- **Fast** for business
- **Easy** to manage
- **Secure** without compromise



Deployment Flexibility

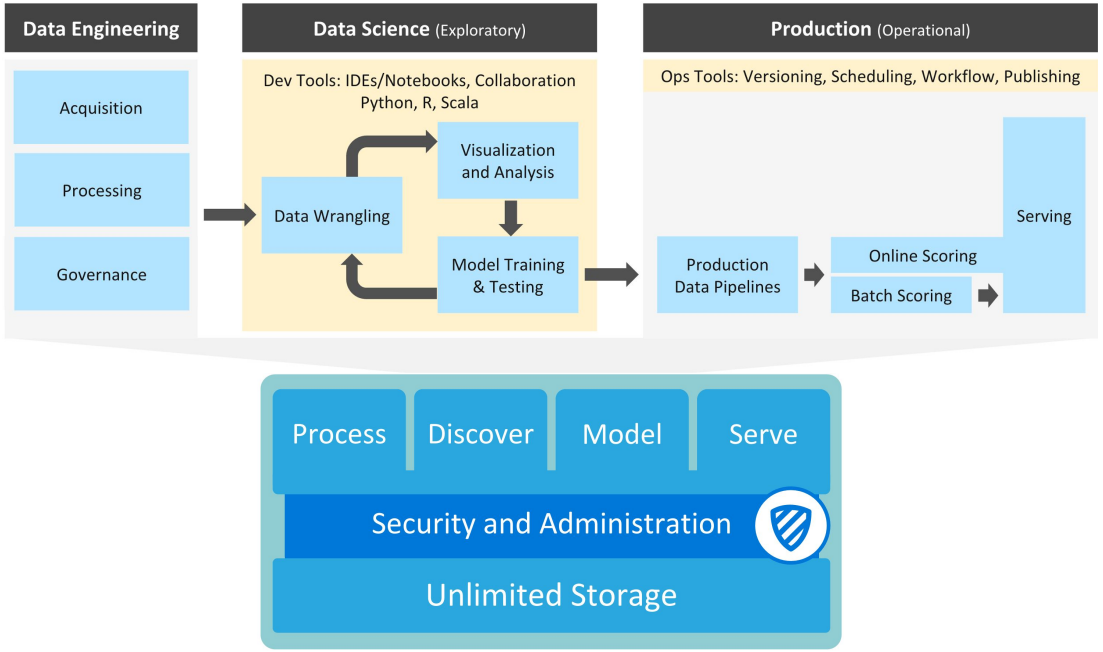
On-Premises
Appliances
Engineered Systems

Public Cloud
Private Cloud
Hybrid Cloud

cloudera

Cloudera Data Science Workbench

Typical Data Science Workflows



The Challenge

Data Scientist Needs: Exploration

IT Needs: Adoption & Compliance

Goal: Bring more data scientists users to Hadoop

cloudera

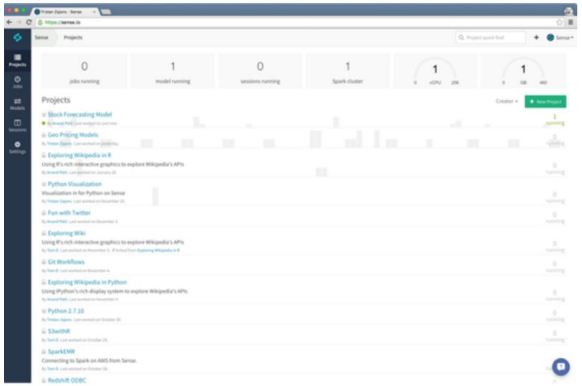
© Cloudera, Inc. All rights reserved.

Cloudera Data Science Workbench

Scalable, Secure, Self-Service Environment for Data Scientists

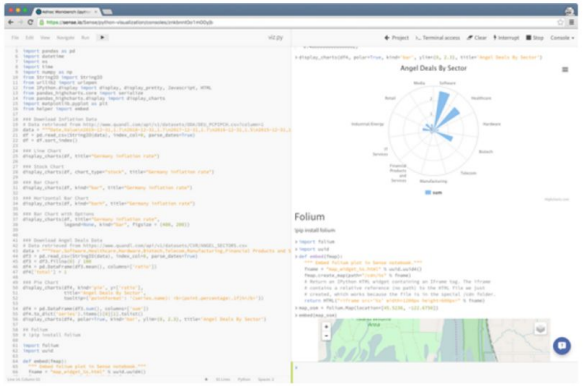
Projects

Collaborative hub for enterprise data science with isolated environments, secure collaboration, and simple dependency management.



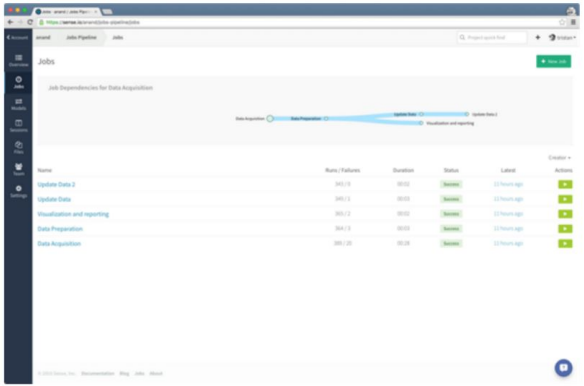
Workbench

Integrated development environment for Python, R, and Scala with support for Spark 2 and connectivity to secured Hadoop clusters.



Jobs

Lightweight job and pipeline system for data science workloads that supports real-time monitoring, results tracking, and email alerting.



cloudera

© Cloudera, Inc. All rights reserved.

Data Integration

The Most Complete (1300+) Partner Ecosystem for Hadoop

Partner Integrations

ZOOMDATA
informatica
TRIFACTA
Datameer
talend
Spotfire
Qlik
tableau
Microsoft
REVOLUTION ANALYTICS
rapidminer
data iku
DataRobot
sas

Acquisition → Transformation → Modeling → Profiling → Visualization → Machine Learning

StreamSets
kafka
Spark Streaming
cloudera IMPALA
APACHE Spark
HUE
jupyter
RStudio
ANACONDA
matplotlib
APACHE Spark
TensorFlow
H_2O

Open Source Integrations

© Cloudera, Inc. All rights reserved.

Cloudera Enterprise-Grade Security

Governing Access to, and Management of, All Data-at-Rest and Data-in-Motion

Table Stakes for Big Data and Native to Cloudera Enterprise

- Cloudera Manager and Navigator automate protections for Hadoop and related projects
- Perimeter security
- Role-based access control
- The only complete policy-based management of sensitive data
- Data lineage and discoverability

AUTHENTICATION

Guarding access to the system, its data, and its various systems

LDAP
Kerberos RPC

AUTHORIZATION

Controlling who or what has access to a resource or service

POSIX Permissions
Apache Sentry

PROTECTION

Encryption for data at rest or in motion with full key management

Cloudera Navigator:
Encrypt & Key Trustee

AUDIT

Capture a complete and immutable record of all activity

Cloudera Navigator
SIEM Tools

cloudera

© Cloudera, Inc. All rights reserved.

Cloudera Management

End-to-End Cluster Operations, Monitoring, and Management

Cloudera Manager

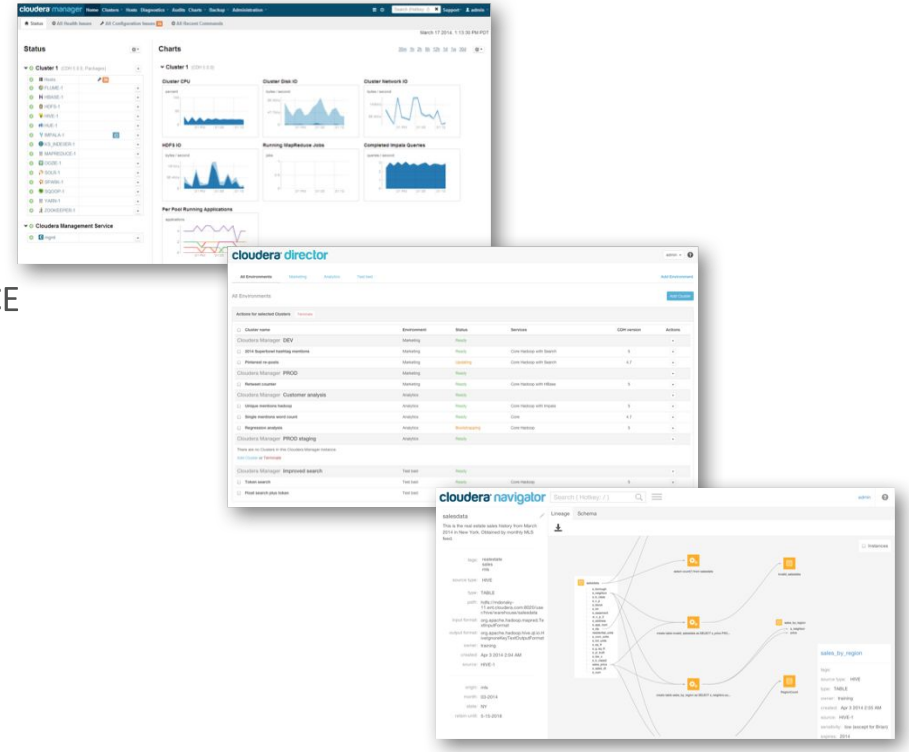
- Unified configuration, management and monitoring
- Online installation and rolling upgrades
- Direct connection to Cloudera Support

Cloudera Director

- Cluster lifecycle management on AWS, Azure and GCE
- Extensible cloud blueprints
- Multi-cluster health visibility
- Usage reporting for billing models

Cloudera Navigator

- Complete auditing & lineage of every data flow
- Metadata Tagging and Discovery
- Data lifecycle Management



cloudera

© Cloudera, Inc. All rights reserved.

Demo: Business Challenge

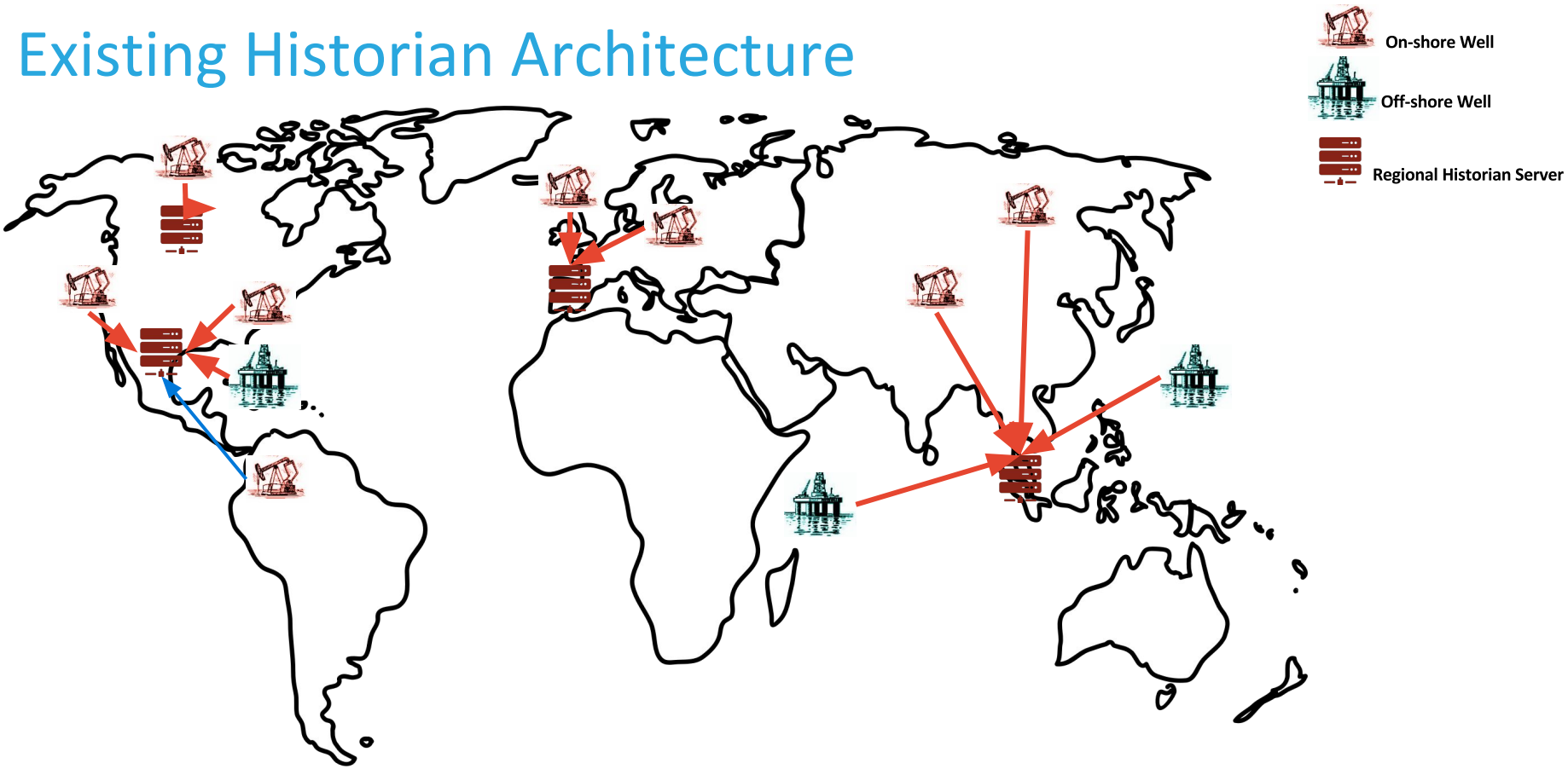
Can we reduce our monthly maintenance costs?

- **We have an asset that is regularly maintained:**
  - Program Maintenance - routine checks of all sensors and asset condition - **$**
  - Preventive Maintenance - part replacement as a result of inspection - **$$**
  - Corrective Maintenance - unplanned shutdown and parts replacement - **$$$$$**
- **We want to understand:**
  - How efficient is our program maintenance schedule?
  - How can we reduce corrective maintenance spend?
- **We have the following data:**
  - Daily Maintenance logs - unstructured/text-based
  - Monthly maintenance costs - structured, low-volume
  - Sensor measurements - structured (time-value records), high-volume

cloudera

© Cloudera, Inc. All rights reserved.

Existing Historian Architecture



cloudera

© Cloudera, Inc. All rights reserved.

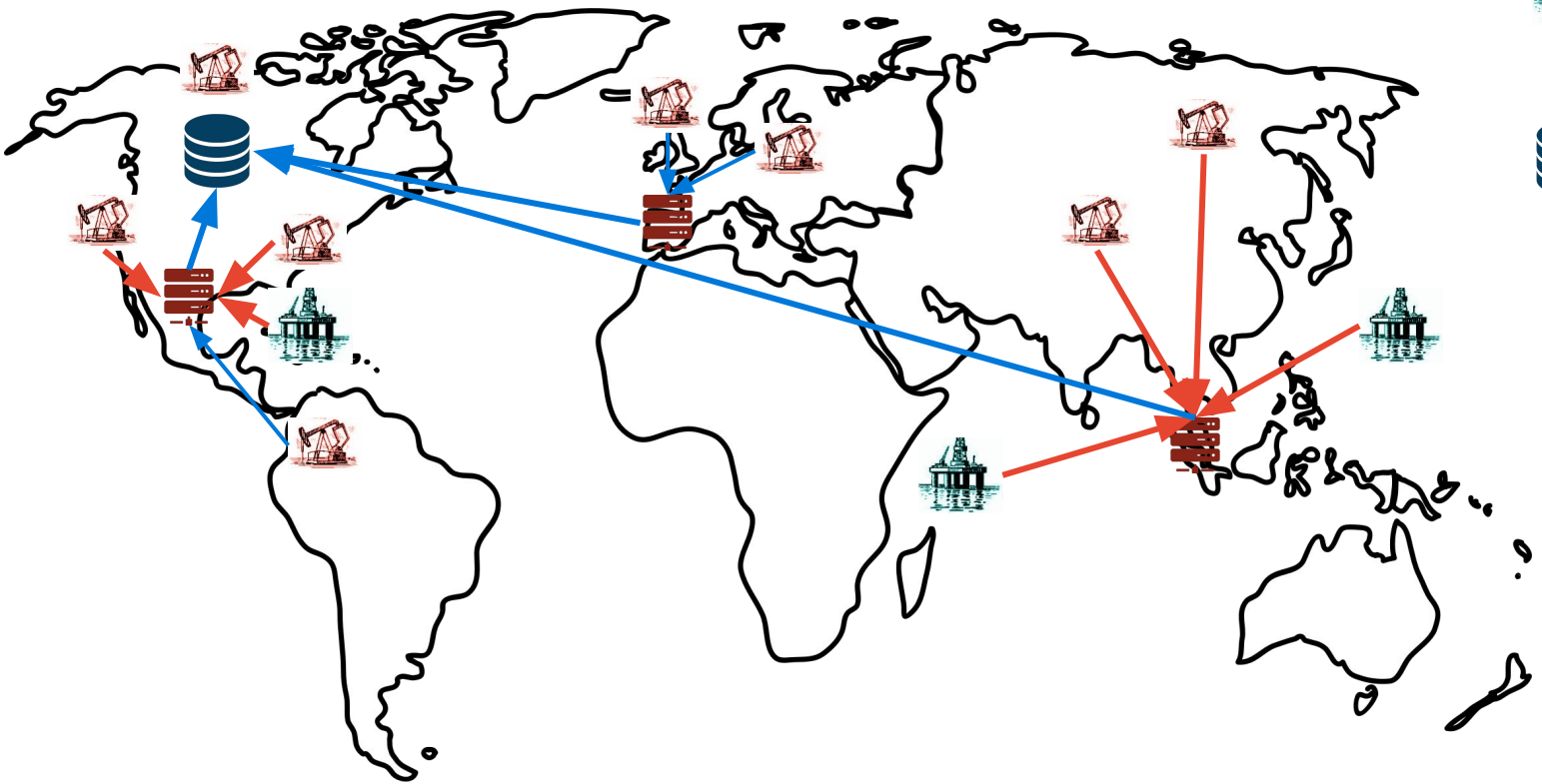
Challenges with Existing Architecture

- Expensive
- Proprietary technology
  - Limited analytics capabilities
  - No inherent relationships between data and tags
  - Difficult to get data out in a useful format - ie. by individual tag ID
  - Ancient historian tooling to analyze the data
- No real-time access to the data
  - Raw data needs processing to be easily visualized
- Not well integrated with the data management ecosystem
  - Need to analyze in Excel, which as a 1M record limit (<.001% of data)

cloudera

© Cloudera, Inc. All rights reserved.

Modern Historian Architecture



On-shore Well

Off-shore Well

Regional Historian Server

Cloudera Hadoop

cloudera

© Cloudera, Inc. All rights reserved.

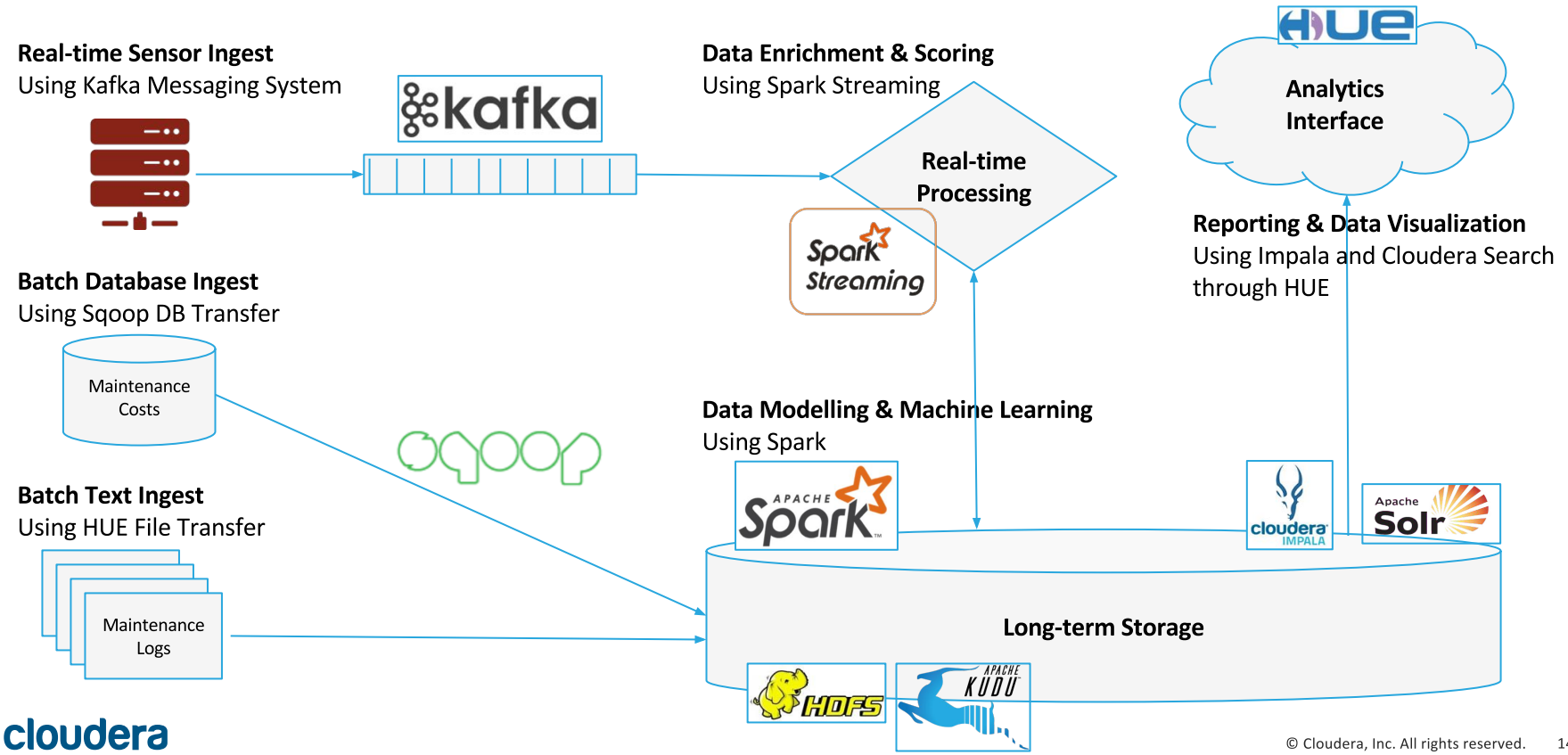
Benefits with Modern Architecture

- Cost Effective
- Open technology
  - Advanced analytics capabilities
  - Easily create relationships between raw data and tag information
  - Integrate with any analytic tooling
- Real-time access to the data
  - Real-time processing of data
- Analyze all sensor data, instead of a small subset
  - 100B+ rows vs. 1M

cloudera

© Cloudera, Inc. All rights reserved.

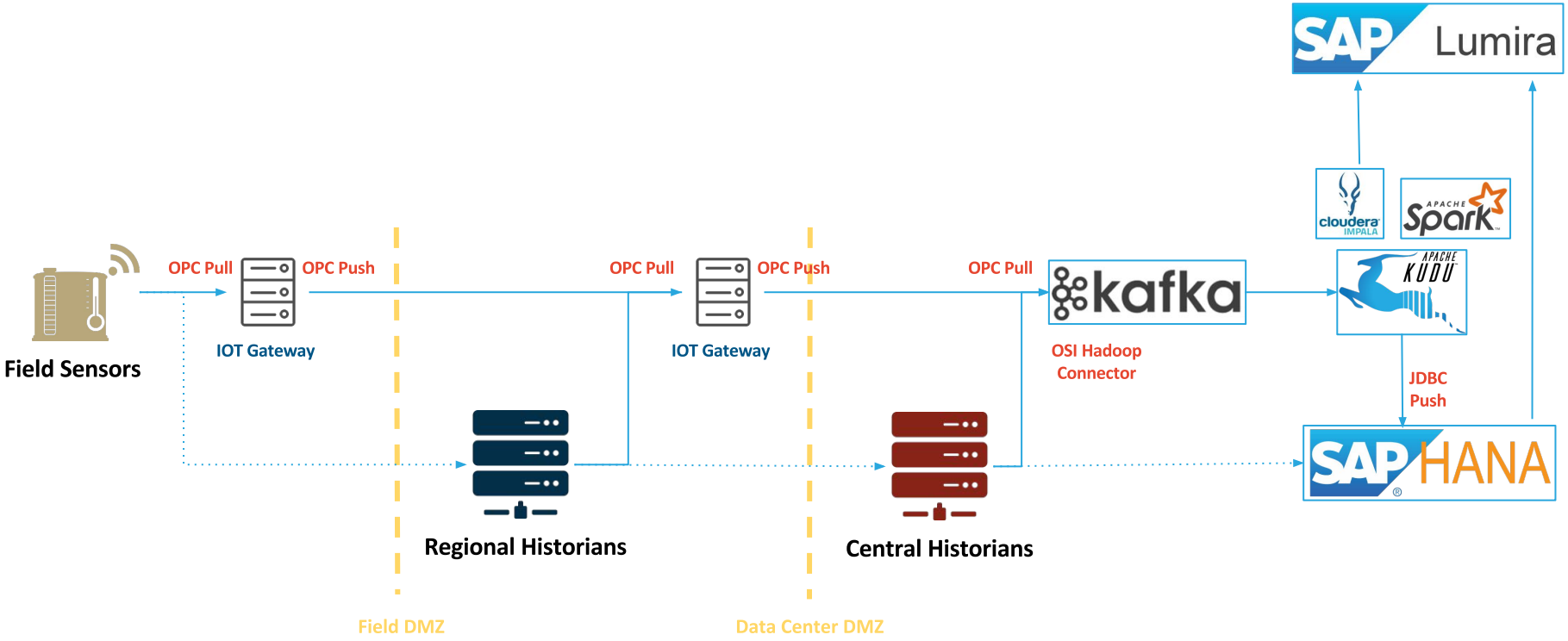
Demo Architecture



cloudera

© Cloudera, Inc. All rights reserved.

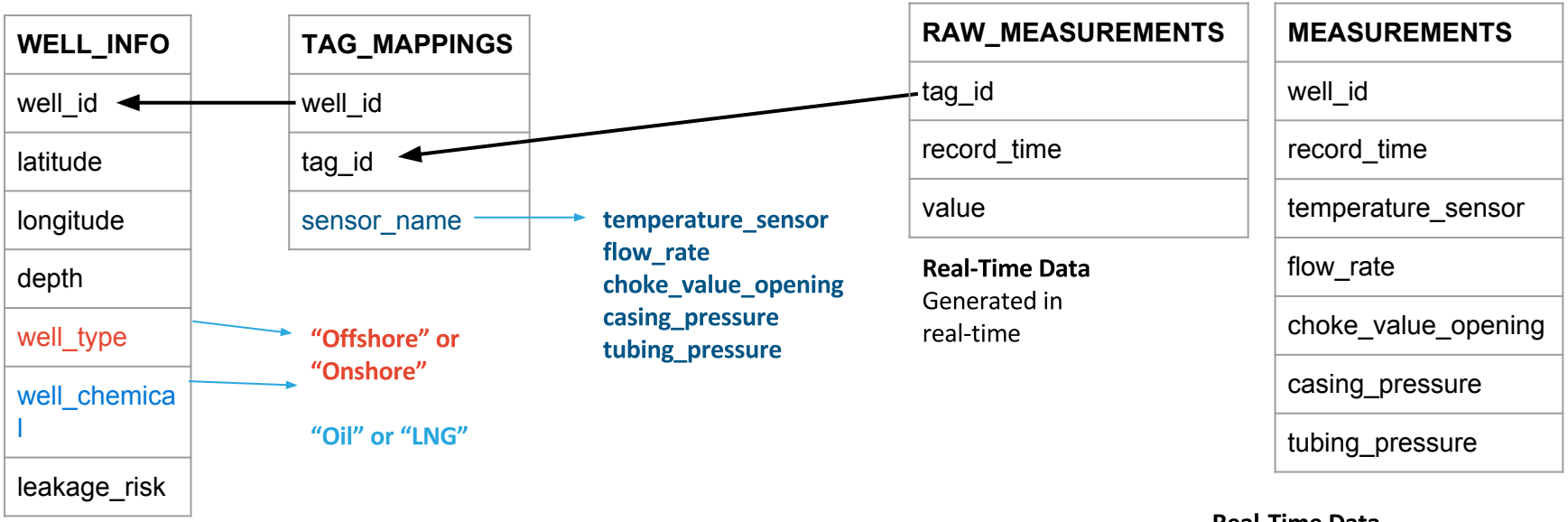
Real-time Ingest of Sensor Data



cloudera

© Cloudera, Inc. All rights reserved.

Demo: Data Model



Static Data
Generated during initial setup, configurable
Stored in Impala using KUDU for fast lookup

Real-Time Data
Generated in real-time

Real-Time Data
Enriched in real-time with Spark Evaluator in Streamsets
Stored in near-real-time into KUDU

cloudera

© Cloudera, Inc. All rights reserved.

Demo: Pivot in Spark

tag_id	record_time	value
1	9:00:00	50
2	9:00:00	12
3	9:00:00	80
1	9:00:30	48
2	9:00:30	14

tag_id	record_time	value	well	sensor_name
1	9:00:00	50	1	temp_sensor
2	9:00:00	12	1	flow_rate
3	9:00:00	80	1	casing_pressure
1	9:00:30	48	1	temp_sensor
2	9:00:30	14	1	flow_rate
3	NULL	NULL	NULL	pressure

well_id	record_time	temp	flow	pressure
1	9:00:00	50	12	80
1	9:00:30	48	15	NULL

PIVOT ON sensor_name

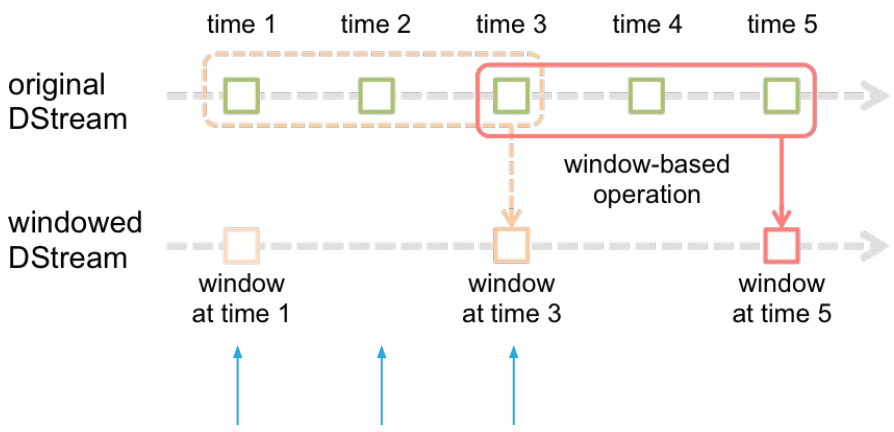
UPSERT INTO KUDU

tag_id	well_id	sensor_name
1	1	temp
2	1	flow
3	1	pressure

RAW_MEASUREMENTS
OUTER JOIN
TAG_MAPPINGS
ON tag_id

© Cloudera, Inc. All rights reserved.

Demo: Spark Streaming Windows



Streaming Interval = 30 seconds

Window Interval = 90 seconds

well_id	record_time	pressure
1	9:00:00	50
2	9:00:00	80
1	9:00:30	52
2	9:00:30	NULL at W1, 81 at W2
1	9:01:00	49
2	9:01:00	76
1	9:01:30	55
2	9:01:30	86
1	9:02:00	50
2	9:02:00	NULL

Window 1

Window 2

Window 3

cloudera

© Cloudera, Inc. All rights reserved.

Demo Steps

1. **Analyze Maintenance Logs**
   a. nGram -> LDA clustering -> Topic Identification
   b. Visualization - monthly maintenance cost vs. types
2. **Analyze Sensor Data**
   a. Visualization - daily sensor values vs. maintenance types
   b. Data profiling
   c. Correlation/Chi-Squared Dependence Test
   d. Model development - linear and random forest regression
3. **Real-time Sensor Data Ingest**
   a. Data processing and cleansing
   b. Real-time data scoring
   c. Automated model improvement
4. **Data Visualization using Impala in HUE**

cloudera

© Cloudera, Inc. All rights reserved.