cloudera

Data Science and Analytics
on BIG and FAST Data in Oil & Gas

Samir Gupta, Solution Engineer | **Cloudera**
sgupta@cloudera.com

Problem:

*Less than **1%** of data from sensors in global oilfields is made available to key decision makers.*

Opportunity:

Enhanced data-driven exploration & drilling advantages could help oil and gas companies:

- *Reduce **maintenance costs** by up to **25%***
- *Reduce **capital expenditures** by up to **18%***
- *Increase **revenues** up to **4%***

cloudera

What if we could capture and analyze ***real-time*** data for ***all*** sensors in the field?

cloudera

Cloudera in the Oil & Gas Industry

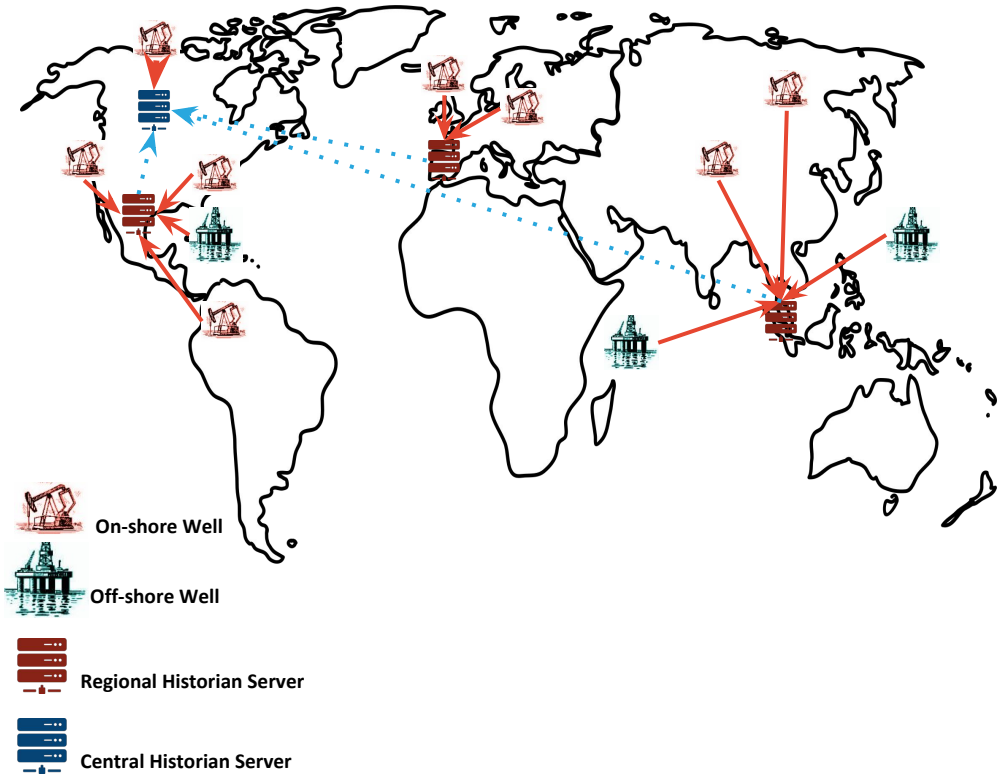
Customers

Partners

Use Cases

- Reservoir Level Analytics
- Asset Predictive Maintenance
- Supply Chain Optimization
- Usage & Demand Monitoring
- Production Prediction
- Regulation Compliance
- Seismic Model Generation

Challenges With Existing Historian Architectures



Expensive

Proprietary historian 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

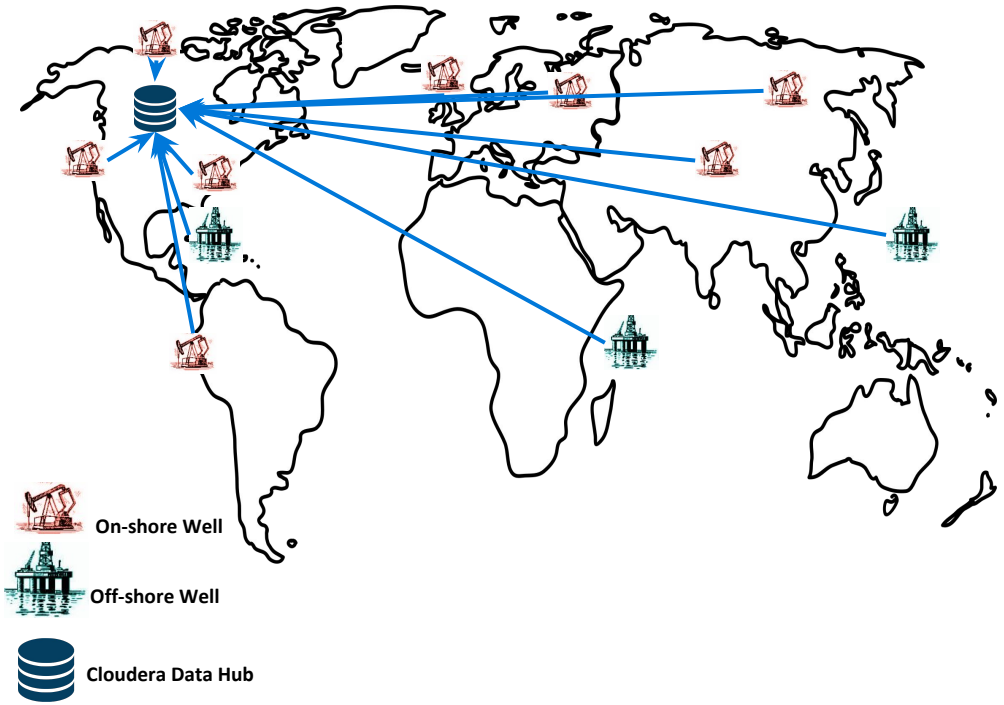
- Raw data needs processing to be easily visualized

Not well integrated with the data management ecosystem

- Need to analyze in Excel, which as a 1M record limit (<.001% of data)

cloudera

Benefits of a Modern & Open Architecture



Cost Effective

- Low-cost archive for all data points

Open technology

- Advanced analytics capabilities - machine learning
- Easily create relationships between raw data and tag information - SQL-based joining
- Integrate with any analytic tooling - use existing BI & visualization tools
- Ability to combine sensor data with other data sources (weather, maintenance, etc.)

Real-time access to the data

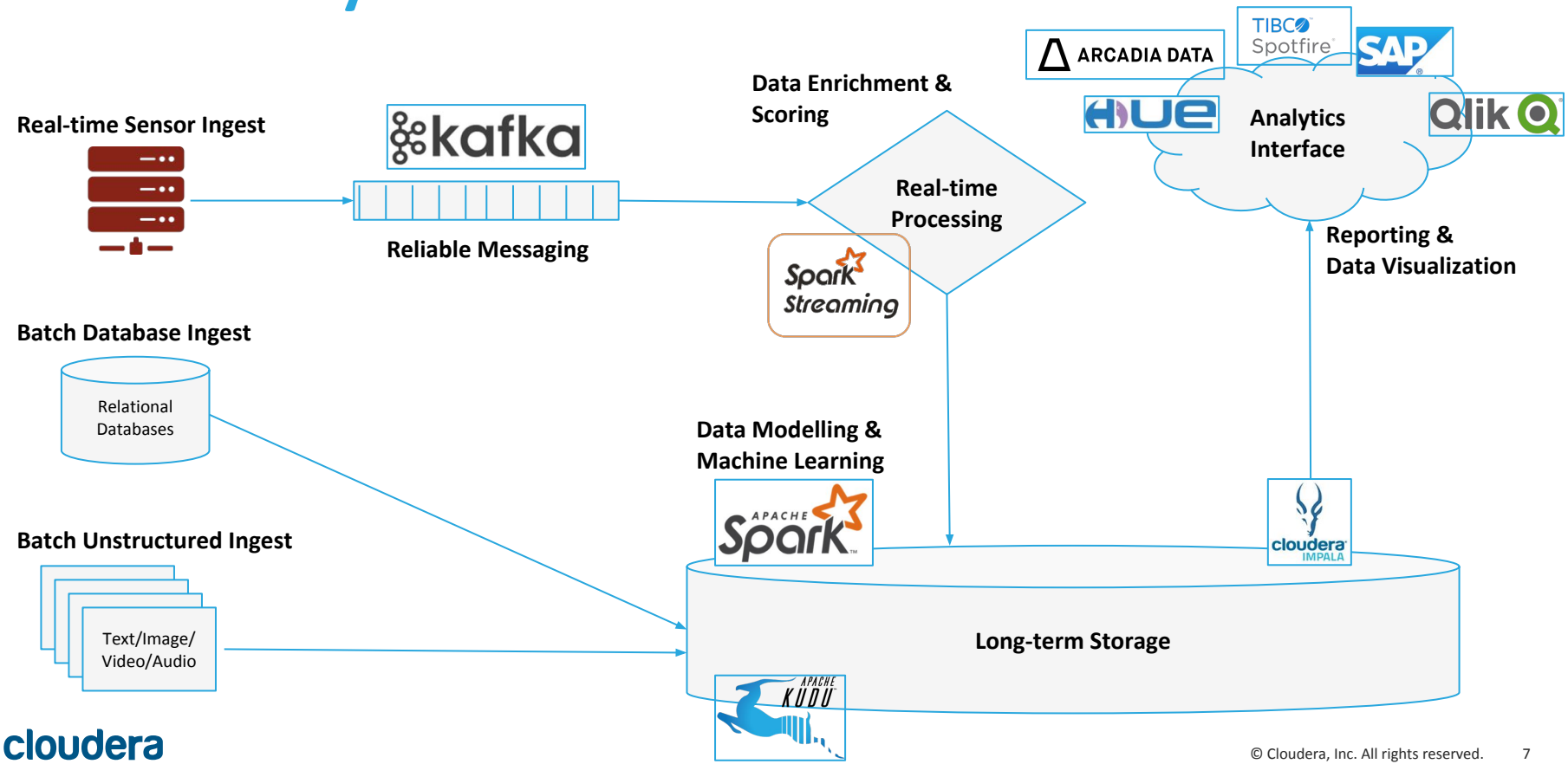
- Real-time processing of data

Analyze all sensor data, instead of a small subset

- 100B+ rows vs. 1M

cloudera

Sensor Analytics: Technical Architecture

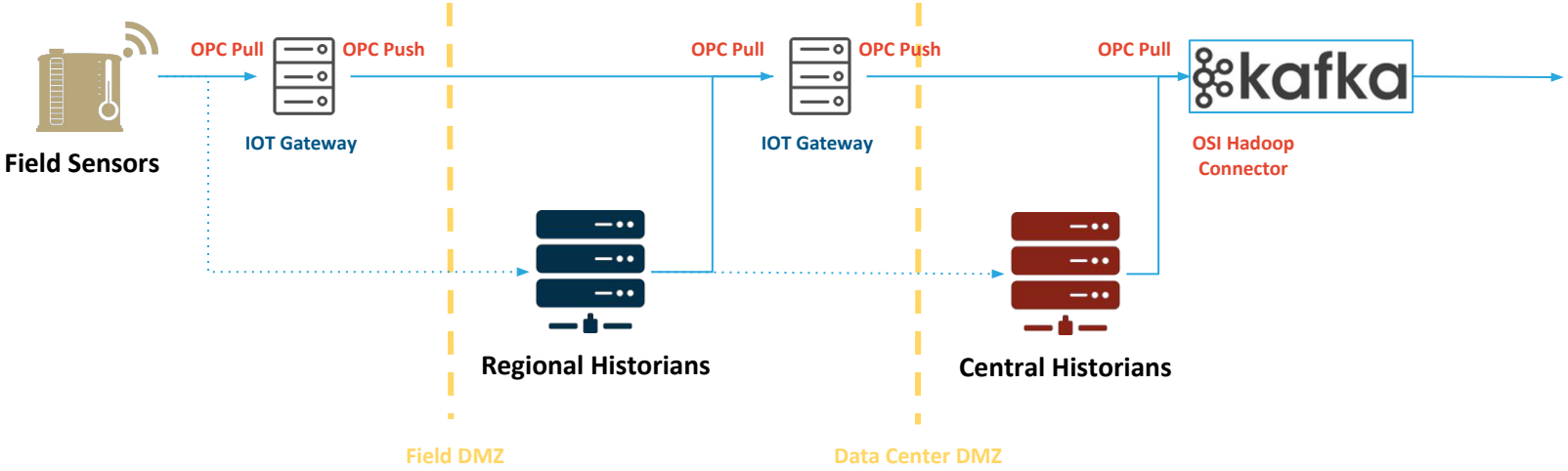


cloudera

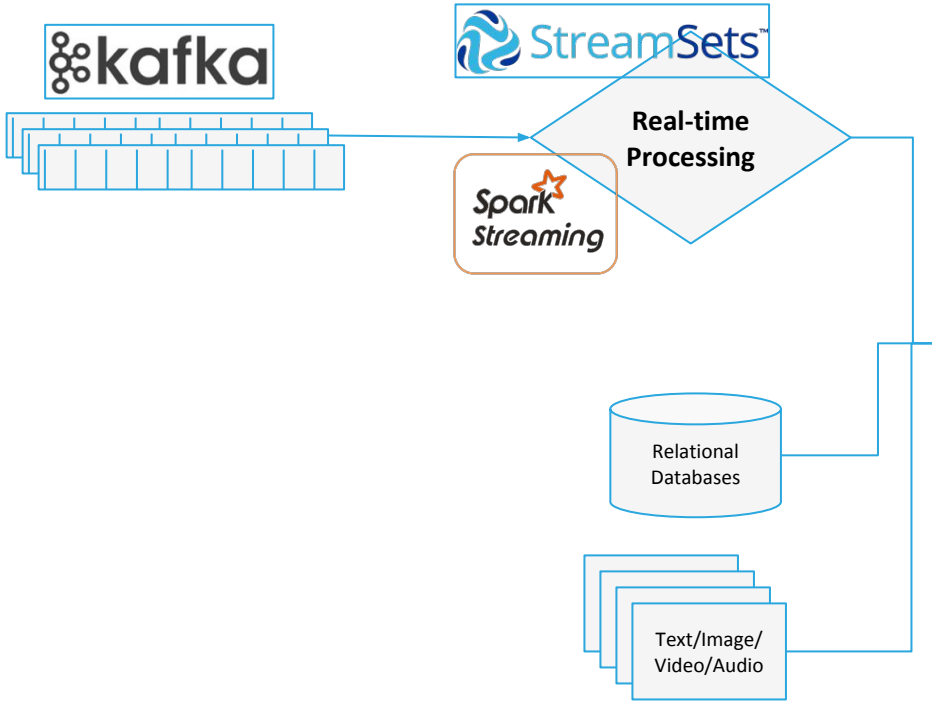
Enabling Technologies: Data Acquisition

IoT Gateway
OPC UA client - land-and-forward
Cloudera Partners: Streamsets, InMation, Microsoft, Intel

Apache Kafka
High throughput, reliable messaging system
Publisher-subscriber model
Unlimited scalability (100M+ records/second)



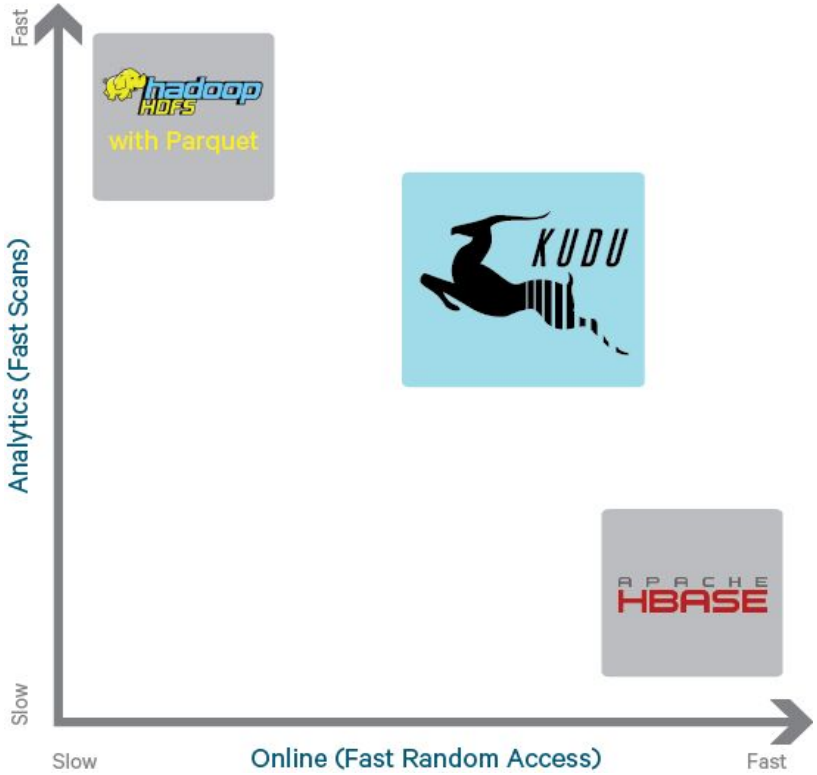
Enabling Technologies: Data Processing & Storage



Apache Spark Streaming
In-memory real-time data processing
Deep integration with Kafka
Sliding windowed aggregations
Real-time scoring

Apache Kudu
Low latency, high-throughput distributed storage engine
Designed for IoT and iterative machine learning workloads
Perform real-time ingest and analytic workloads at the same time

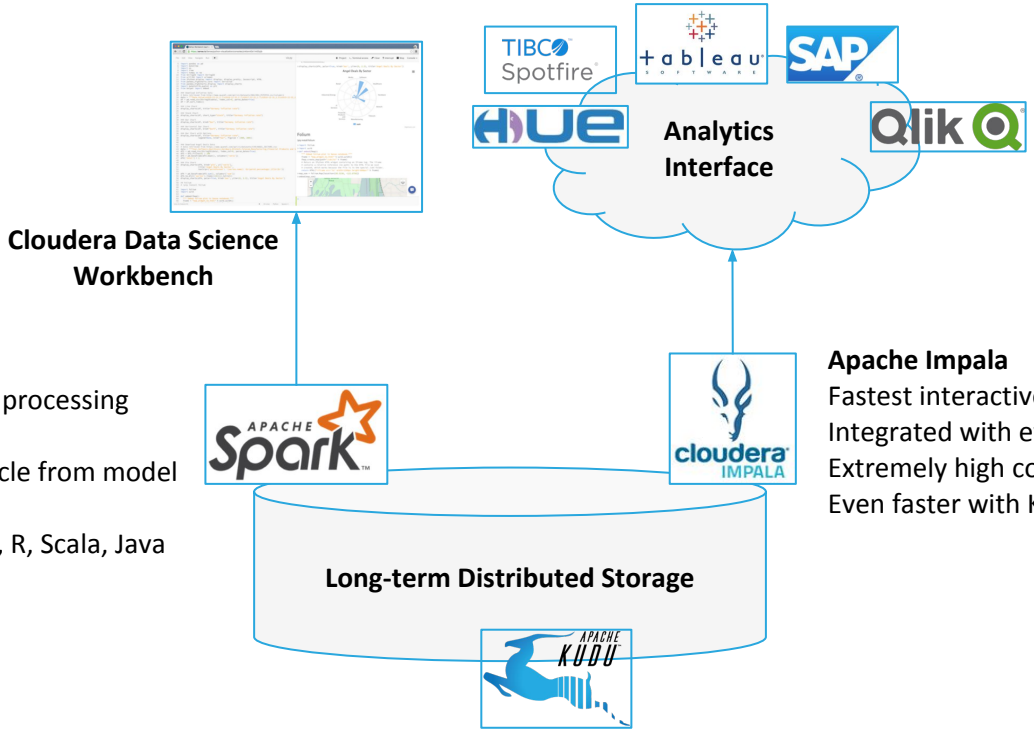
Enable Time Series Analytics with Kudu



- **High throughput** for big scans (columnar storage and replication)
  Goal: Within 2x of Parquet
- **Low-latency** for short accesses (primary key indexes and quorum design)
  Goal: 1ms read/write on SSD
- **Database-like** semantics (initially single-row ACID)
- **Relational data model**
  - SQL query
  - "NoSQL" style scan/insert/update (Java client)

cloudera

Enabling Technologies: Data Science and Analytics



Apache Spark
In-memory, distributed data processing
Machine Learning APIs
Complete data science lifecycle from model development to deployment
Easy development in Python, R, Scala, Java

Apache Impala
Fastest interactive SQL engine
Integrated with every major BI/analytics tool
Extremely high concurrency
Even faster with Kudu storage

Apache Kudu
Perform real-time ingest and analytic workloads at the same time

cloudera

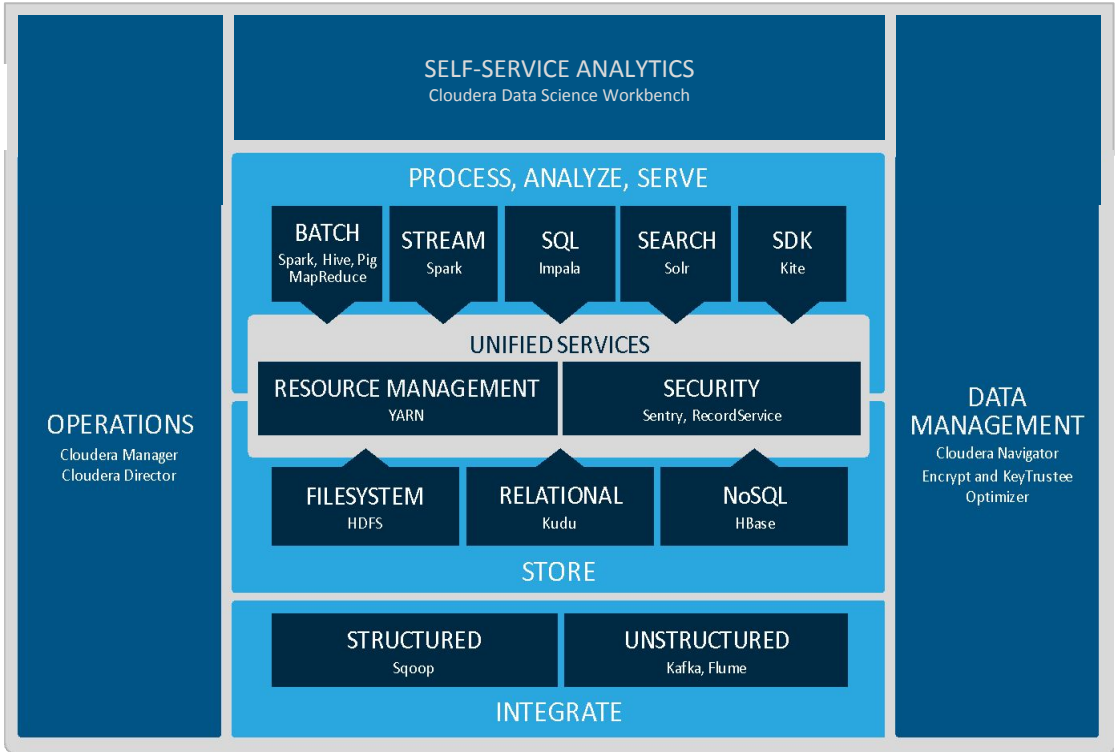
Cloudera Enterprise

Hadoop is a new kind of analytics data platform.

- One place for unlimited data
- Unified data access

Cloudera makes it:

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



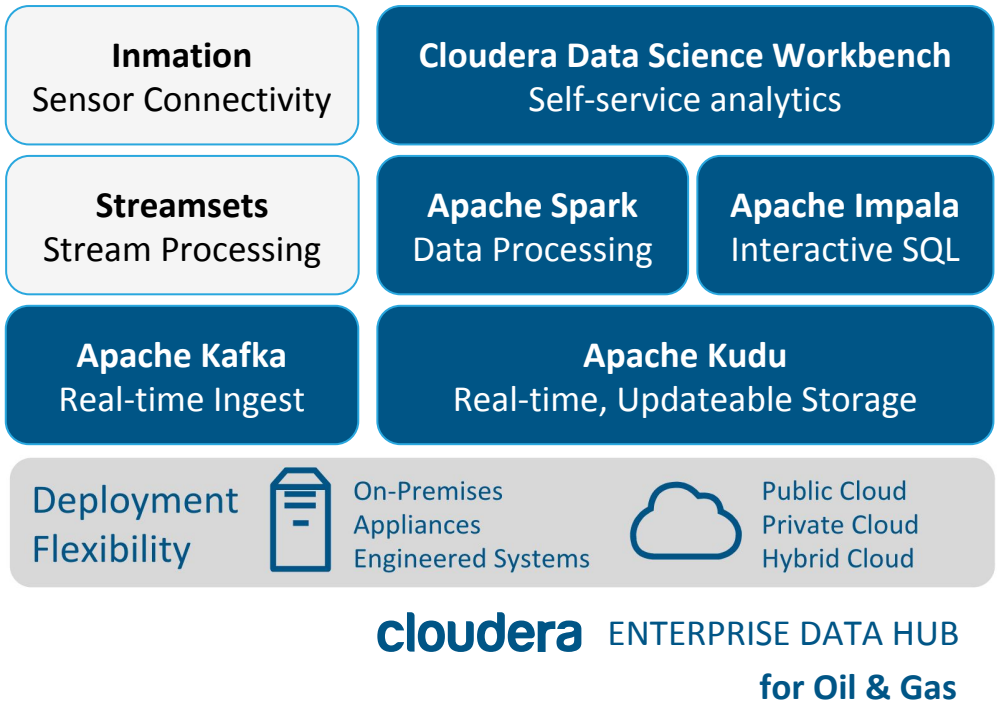
Deployment Flexibility

On-Premises
Appliances
Engineered Systems

Public Cloud
Private Cloud
Hybrid Cloud

cloudera

Cloudera Solution for Real-time Analytics



Features

Pre-built connectivity to all OPC sources (PI, Scada, ...)

Real-time ingestion pipeline for 100k+ tags

Optimized data model and metadata for fast analytics

Machine learning models for predictive analytics

Benefits

Ingest 400K+ sensors directly from your field assets

10-100x cost reduction from existing systems

Access ALL raw SCADA/historian data immediately

Develop & run advanced predictive models

Unlimited scalability in volume & type of data

Complete perimeter, data and audit security

Deploy on-premise or in-cloud using the same stack

cloudera

Sensor Analytics Required Capabilities

Required Capabilities	Cloudera	Other "Big Data" Solutions	Enabling Technology
Performance and Tooling			
Interactive, scalable low-latency SQL querying and BI reporting on sensor data	✓	✗	Apache Impala
Support for random updates/deletes and fast analytical scans in one storage engine	✓	✗	Apache Kudu
Real-time data processing and ingest	✓	✗	Apache Kafka + Spark + Kudu
High performance integration with existing BI/ETL ecosystem	✓	✗	Apache Impala
Tooling to support self-service data science, machine learning and discovery	✓	✗	Data Science Workbench
Security and Governance			
Granular governance, auditing, and lineage of all components down to SQL query level	✓	✗	Cloudera Navigator
Encryption of data at **all** stages to support organization compliance	✓	✗	Navigator Encrypt
Column and record-level authorization	✓	✗	Apache Sentry & RecordService
Cloud Flexibility			
Automated deployment and elastic scaling on all major cloud vendors	✓	✗	Cloudera Director
Easy integration with cloud object storage (S3, AZDL) for data processing and analytics	✓	✗	Cloudera Manager
Enterprise-Grade Management			
Rolling upgrades to support 24/7 operations	✓	✗	Cloudera Manager
Proactive Support and Predictive Issue Analysis	✓	✗	Cloudera Manager
Automated backup and disaster recovery	✓	✗	Cloudera Manager

cloudera

Thank you

sgupta@cloudera.com

