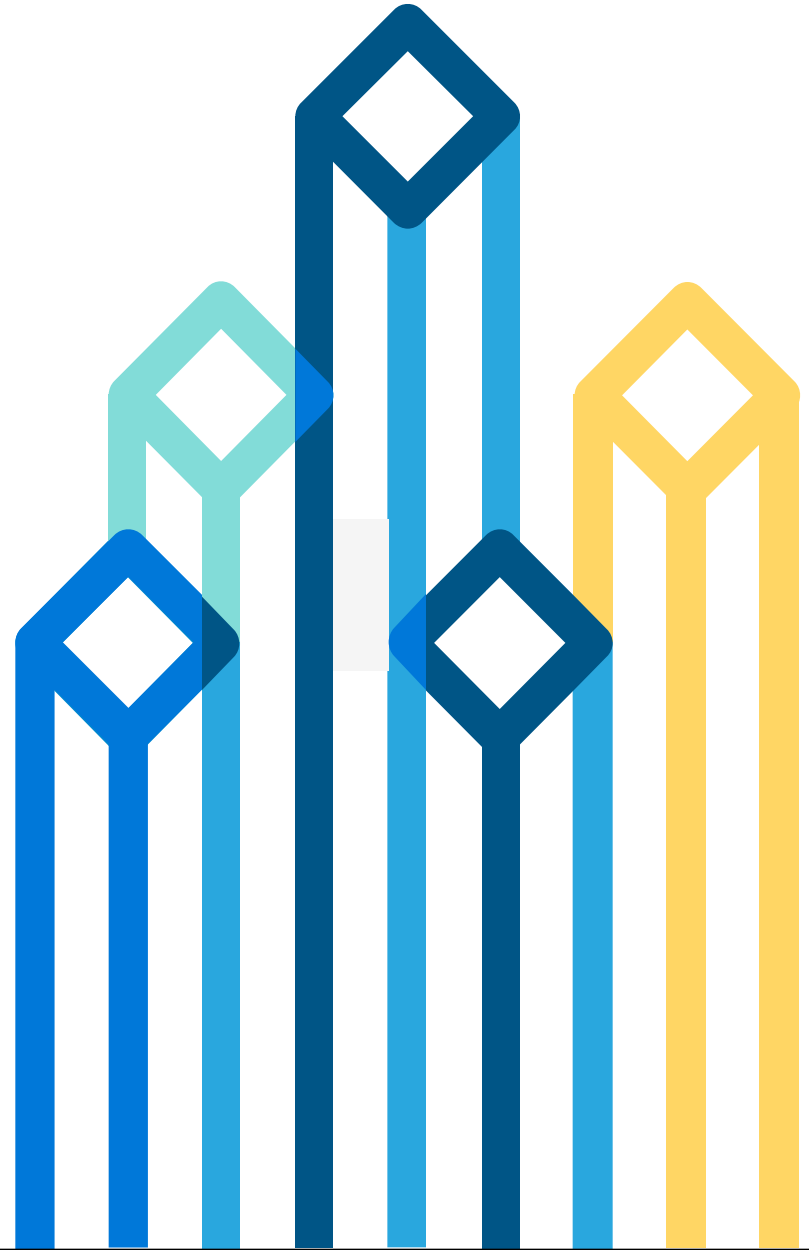
cloudera

Building the Enterprise Data Lake with Cloudera & Cisco

Prepared by :
Marilyn Tan, Country Manager Singapore
Xue Daming, Senior Systems Engineer



Digital Transformation with Data

cloudera

DATA is Transforming the World!

CONNECTED WORLD

INTERNET OF THINGS
INDUSTRY 4.0

- Smart Things / Devices
- New Use Cases
- New Architectures
- By 2020: 20.8b devices
- IoT: $1.7trillion in 2020

MORE DATA

DATA AS
4th PRODUCTION FACTOR

- New Analytics
- Machine Learning & AI
- New data sources
- Data Virtualization
- Data Science

SMART APPs

THE NEW UX

- Connected Experience
- Ubiquitous computing: Everywhere & on every device (Voice, VR, AR, mobile, Wearables)

DIGITAL DEMOCRACY & SECURITY

RISE OF OPEN SOURCE

- Digital Sharing Economy: Open Data & Algorithms
- Enterprise ready Open Source (e.g. Apache)
- Digital (distributed) Trust (esp. Blockchain)

cloudera

The 9 year Cloudera journey …

2008
CLOUDERA FOUNDED BY MIKE OLSON, AMR AWADALLAH & JEFF HAMMERBACHER, CHRISTOPHE BISCIGLIA JOINED BY DOUG CUTTING (2009)

2009
CDH: FIRST COMMERICAL APACHE HADOOP DISTRIBUTION & CLOUDERA MANAGER

2011
CLOUDERA REACHES 100 PRODUCTION CUSTOMERS

2011
CLOUDERA UNIVERSITY EXPANDS TO 140 COUNTRIES
SUPPORT IMPLEMENTS FOLLOW THE SUN MODEL

2012
CLOUDERA ENTERPRISE 4 THE STANDARD FOR HADOOP IN THE ENTERPRISE

2012
CLOUDERA CONNECT REACHES 300 PARTNERS ACROSS SI, HARDEWARE, AND SOFTWARE PARTNERS

2013
CLOUDERA EXPANDS BEYOND MR AND HBASE, INTRODUCING IMPALA, SOLR AND SPARK

2014
CLOUDERA INTRODUCES THE ENTERPRISE DATA HUB AND CLOUDERA ENTERPRISE 5

2014
CLOUDERA FOCUSSES ON SECURITY, AND GOVERNANCE WITH NAVIGATOR 2 AND CLOUD WITH DIRECTOR

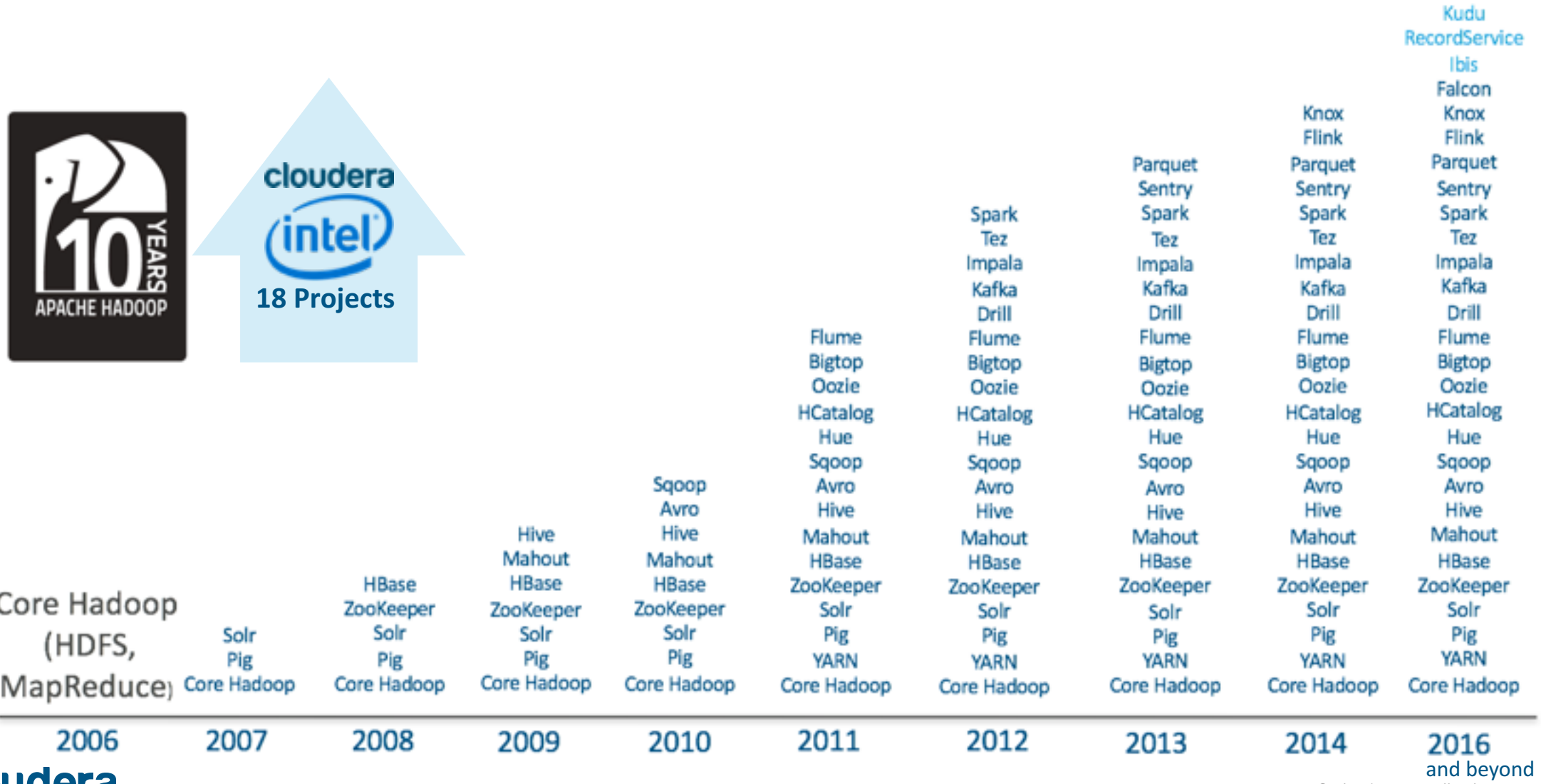
2015
CLOUDERA INCLUDES KAFKA, KUDU AND RECORD SERVICE WITHIN CLOUDERA ENTERPRISE

2016
NAVIGATOR OPTIMIZER GENERAL AVAILABILITY, IMROVED CLOUD COVERAGE WITH AWS, AZURE AND GCP

2017…
CLOUDERA ACQUIRED FAST FORWARD LABS, ANNOUNCED PaaS ALTUS, DATA SCIENCE WORKBENCH, SHARED DATA EXPERIENCE (SDX) AND MORE TO COME!

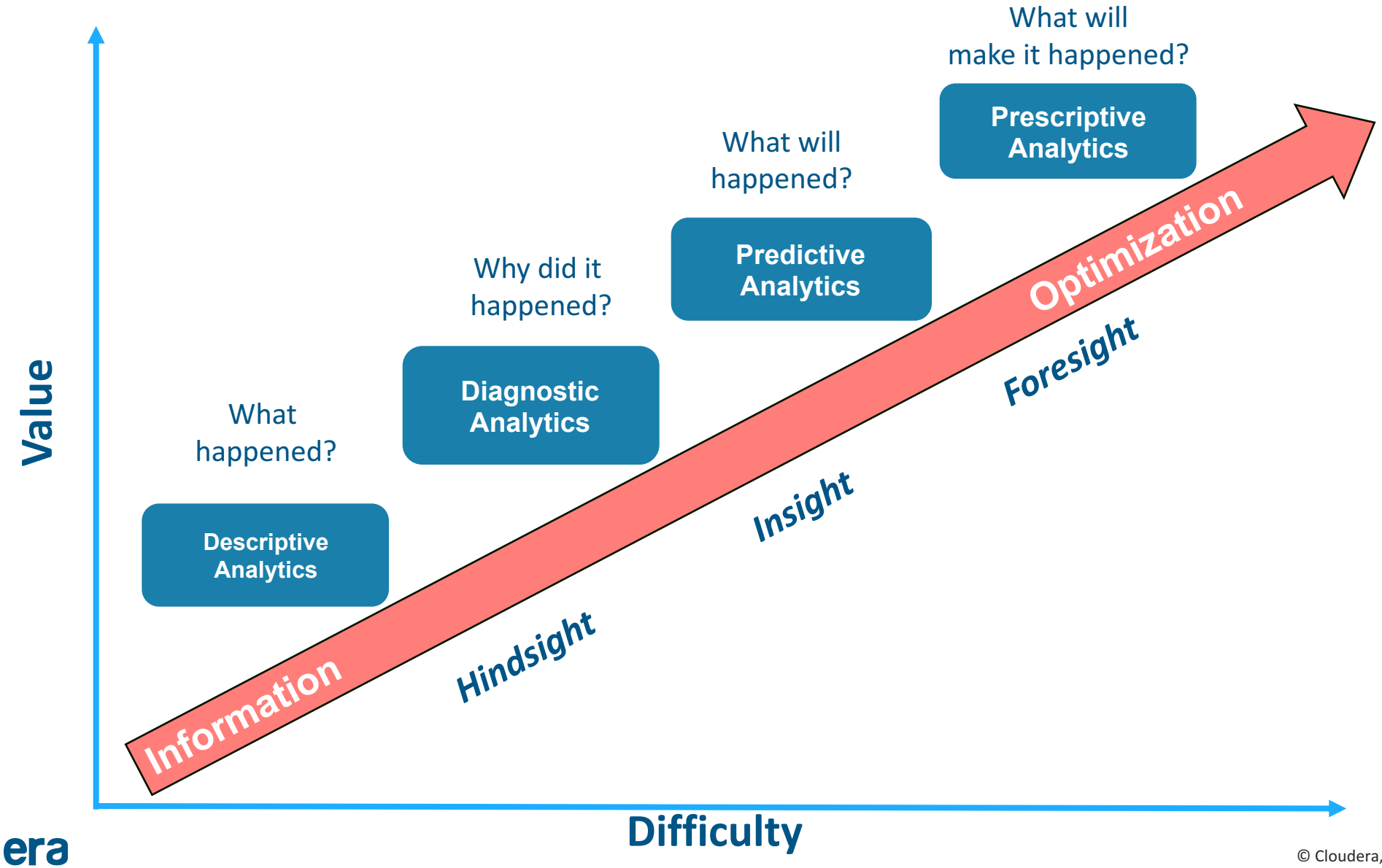
cloudera

What Happen Next: A Decade of Hadoop



cloudera

Gartner Analytics Ascendancy Model

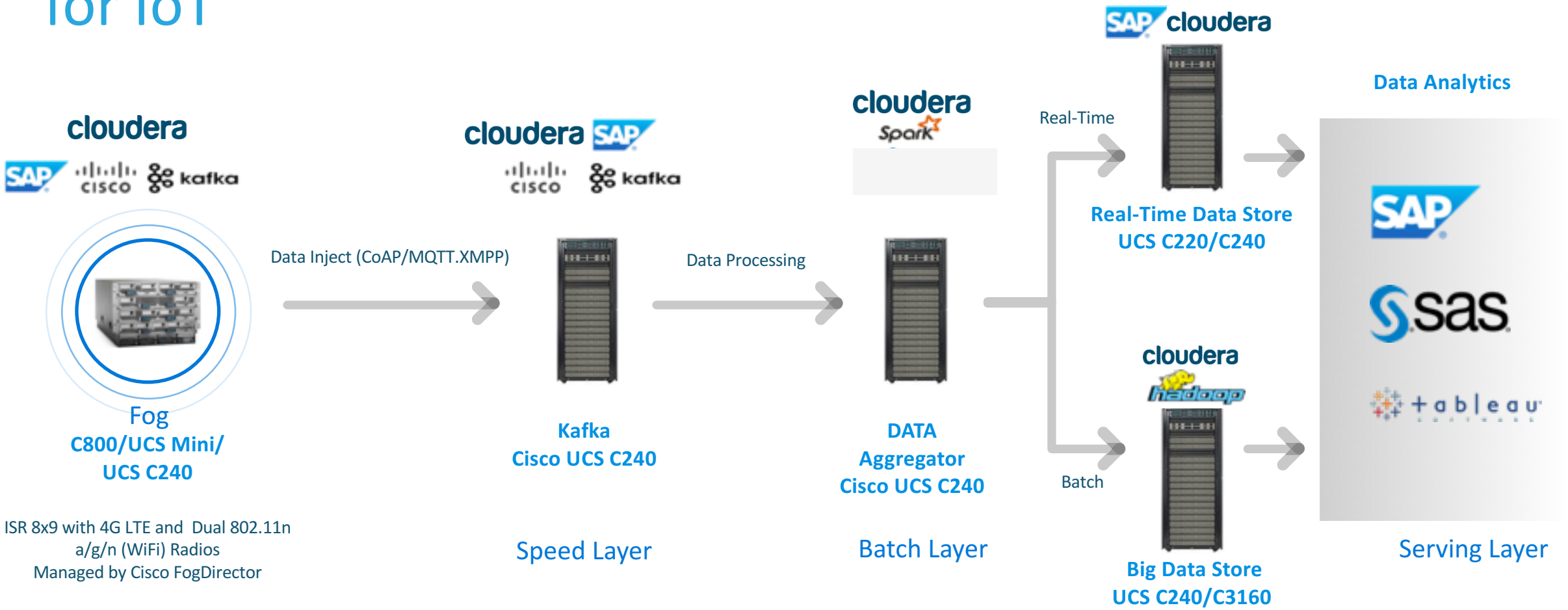


cloudera

Cloudera & Cisco Enterprise Data Lake Innovation

cloudera

Cisco UCS Integrated Infrastructure with Cloudera for IoT



Cisco UCS at all layers, fully validated architectures with all major players

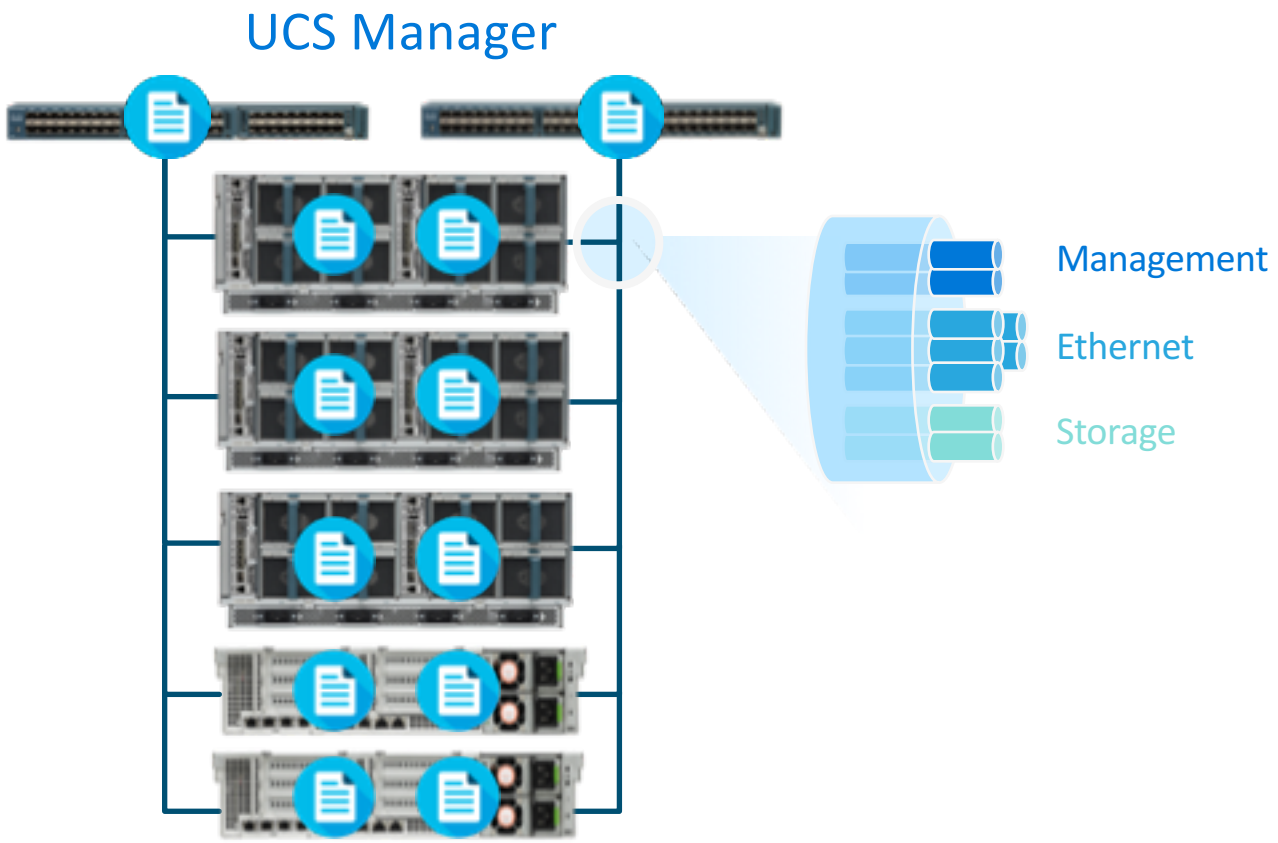
cloudera

Fabric Centric Design

High Performance
40 GB/s Ethernet; 320 GB/s per Chassis

Unified Fabric
Single Cable for Network, Storage, and Management Traffic

Easy to Scale
Single Point of Management: Add Cables for Bandwidth vs. Fabric Type



cloudera

Management Simplicity

Big Data: Management Consistency

Hundreds of Servers

Thousands of management points

Simplified Scalability

Easily Scale your infrastructure from few servers to thousands of servers with a fully Integrated Infrastructure

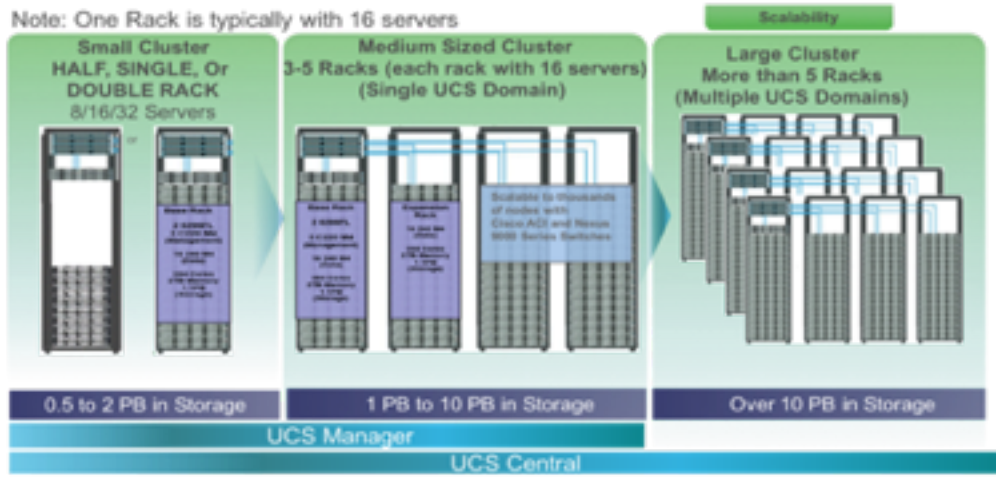
Centralized Management

Service Profiles for Servers

- Manage all servers centrally

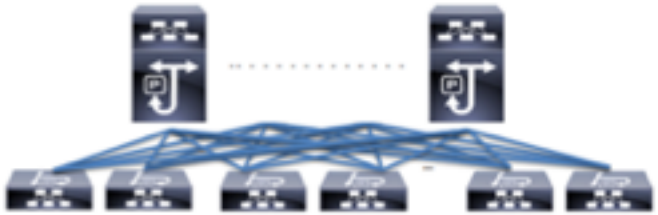
Application Profiles for Network

- Manage all network centrally



UCS Service Profile

Cisco ACI Application Profile



cloudera

The enterprise platform for machine learning

PATTERN RECOGNITION

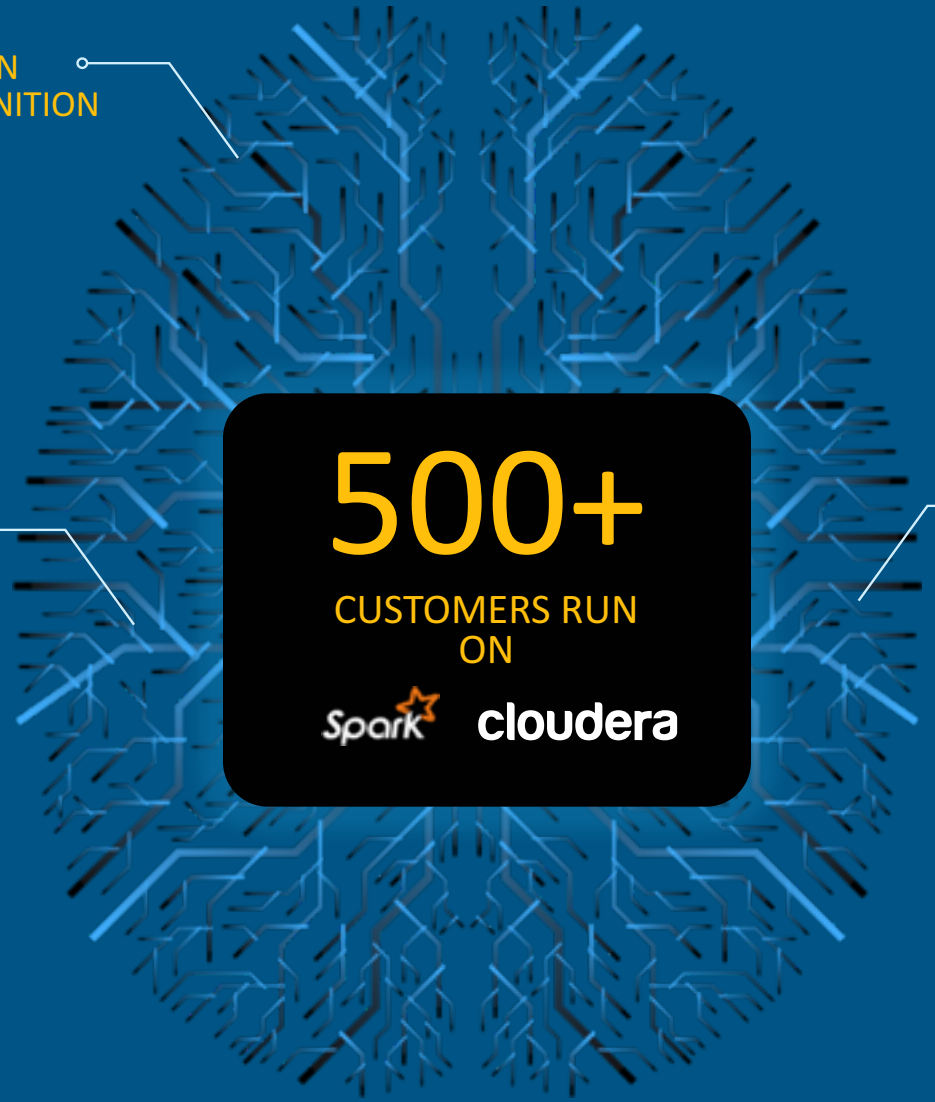
DRIVE CUSTOMER INSIGHTS

Market segmentation
Customer 360
Next best offer
Churn analysis & prevention

DETECTION

PROTECT BUSINESS

Cybersecurity
Fraud
Anti-money laundering
Risk modeling & assessment
SPAM detection



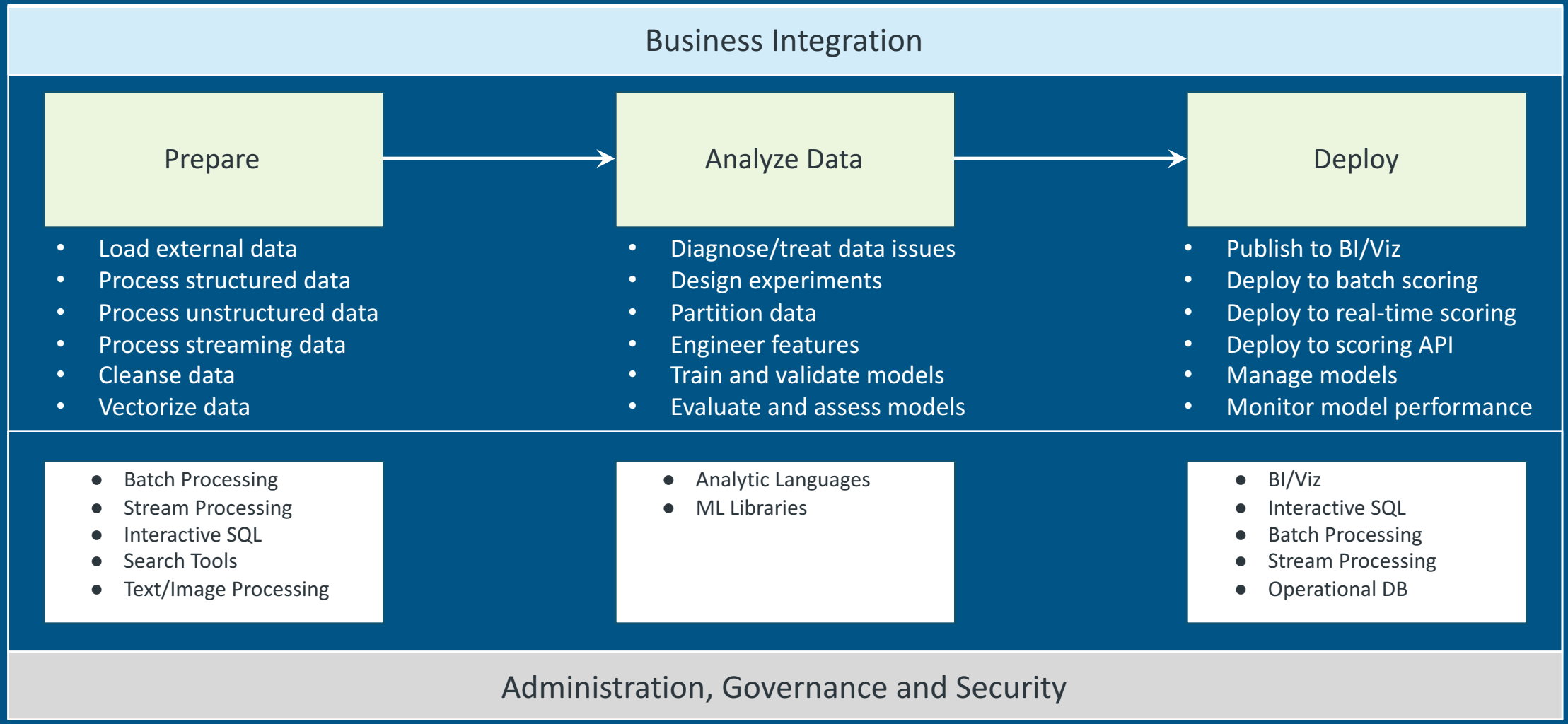
PREDICTION

CONNECT PRODUCTS & SERVICES (IoT)

Predictive maintenance
Genomics & personalized medicine
Predicting and preventing disease
Natural language

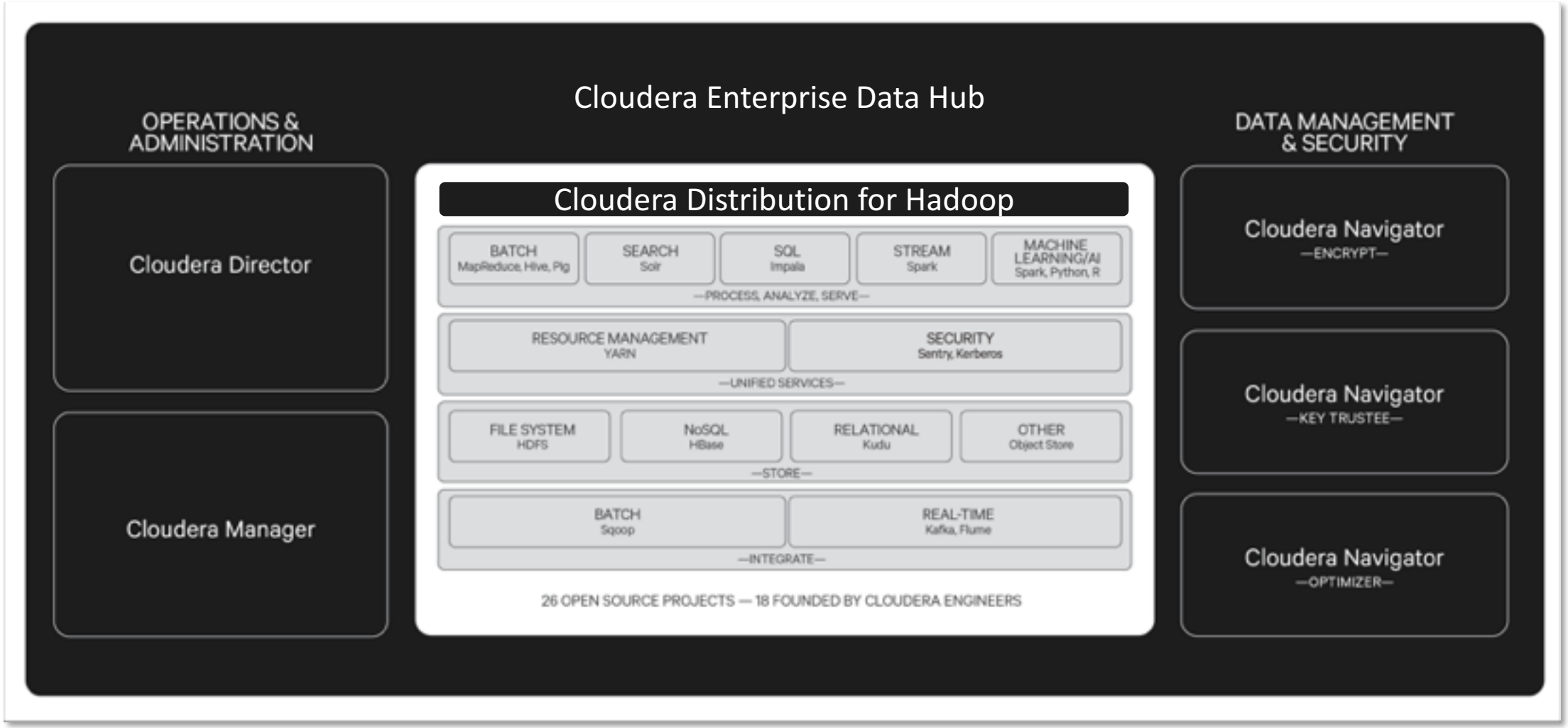
cloudera

Machine learning requires a complete stack.



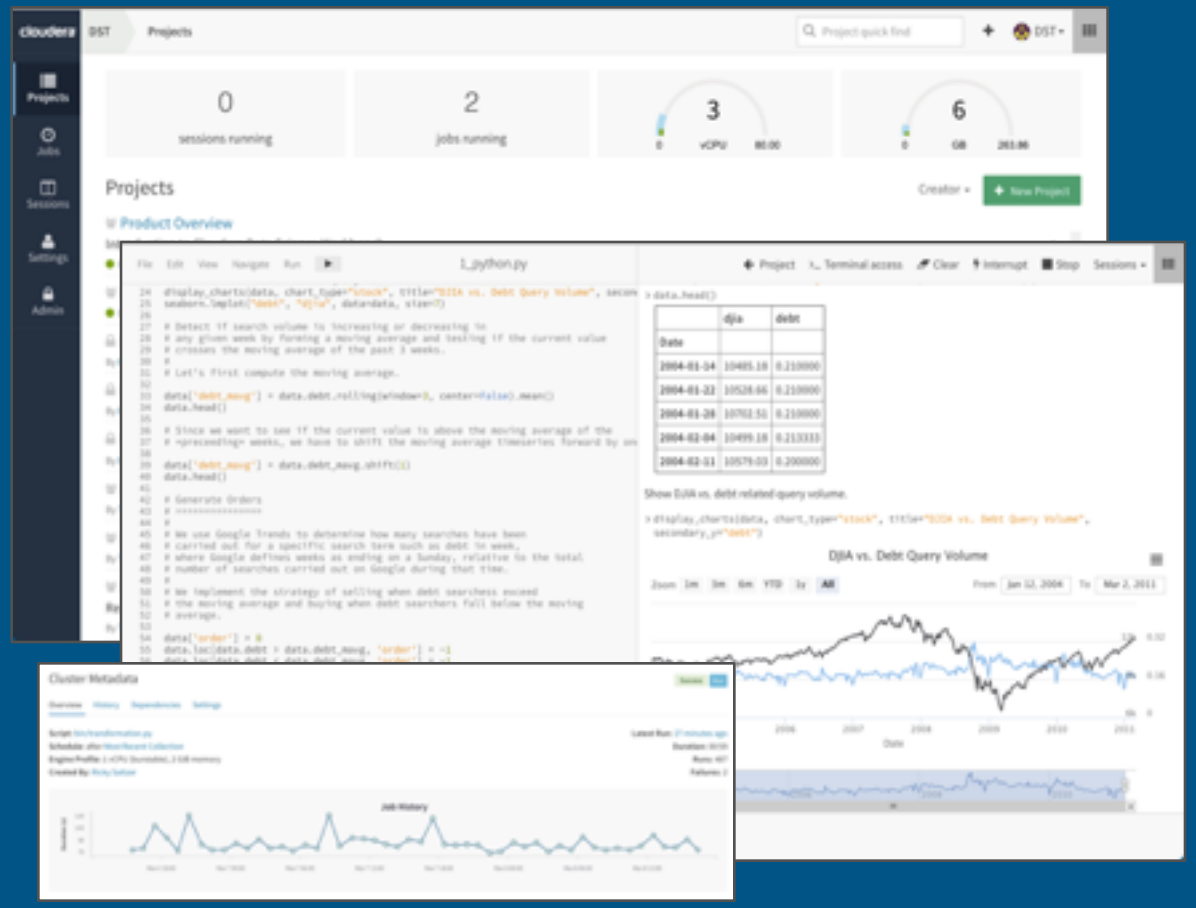
cloudera

A complete, integrated enterprise platform



cloudera

Cloudera Data Science Workbench



- Supports data science end-to-end
- Full access to data
- Secure self-service provisioning
- Containerized environments
- Supports Python, R, and Scala
- Automates:
  - Workflow
  - Version control
  - Collaboration
  - Sharing

cloudera

CDSW Benefits

Data Scientists

- Web browser, no desktop footprint
- Use R, Python, or Scala
- Install any library or framework
- Isolated project environments
- Direct access to data in secure clusters
- Share insights with team
- Reproducible, collaborative research
- Automate and monitor data pipelines
- Built-in job scheduling

IT

- Support self-service data science
- Full platform security
- Kerberos authentication
- Run on-premises or in the cloud

cloudera

Deep learning in Cloudera with Apache Spark

Spark Packages

- Two packages:
  - CaffeOnSpark
  - TensorFlowOnSpark
- Developed by Yahoo
- Python and Scala APIs
- All DL architectures
- Integrated pipeline
- Run on existing clusters
- Training and inference

DL4J

- Open source DL library
- Developed by Skymind
- Built on JVMs
- Supports CPUs and GPUs
- Java, Scala, Python APIs
- Training and inference
- Imports models from:
  - TensorFlow
  - Caffe
  - Torch
  - Theano
- Runs on existing clusters

BigDL

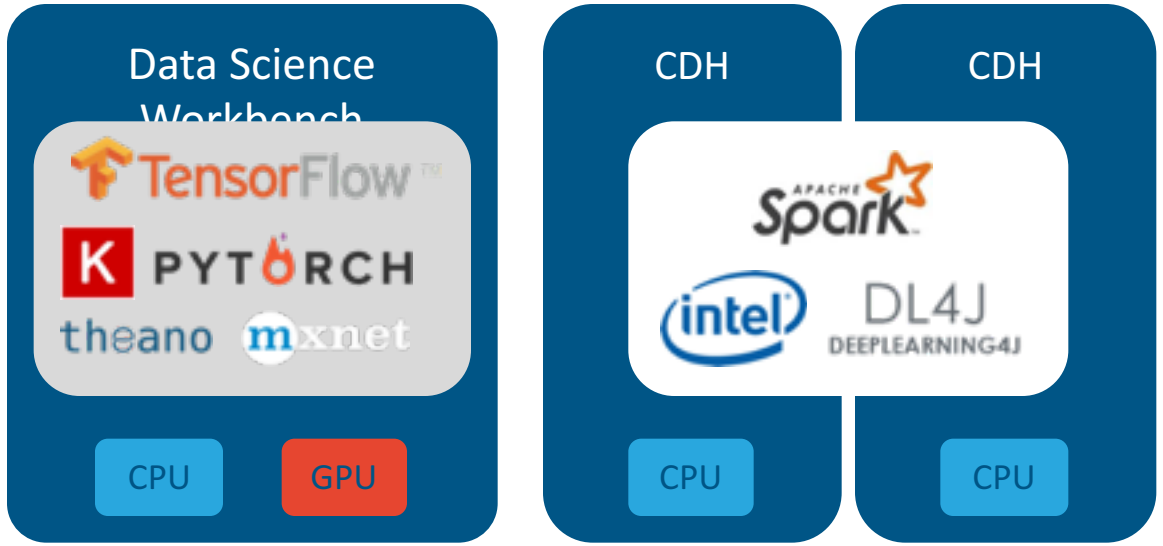
- Deep learning framework
- Developed by Intel
- Supports CPUs only
- Leverages Intel MKL
- Scala, Python APIs
- Imports models from:
  - TensorFlow
  - Caffe
  - Torch
- Runs on existing clusters

New! Accelerated deep learning on-demand with GPUs

"Our data scientists want GPUs, but we can't find a way to deliver multi-tenancy.
If they go to the cloud on their own, it's expensive and we lose governance."

- Extend existing CDSW benefits to GPU-optimized deep learning tools
- Schedule & share GPU resources
- Train on GPUs, deploy on CPUs
- Works on-premises or cloud

Multi-tenant GPU support on-premises or cloud



single-node training

distributed training, scoring

cloudera

Enterprise Data Lake Architecture

Canonical Ingestion & Spark Streaming Analytics with Cisco Big Data Analytics Platform

- Integrate with Apache Spark Streaming for **real-time analysis** of data
- Write back to Kafka for further processing or to send to an application layer

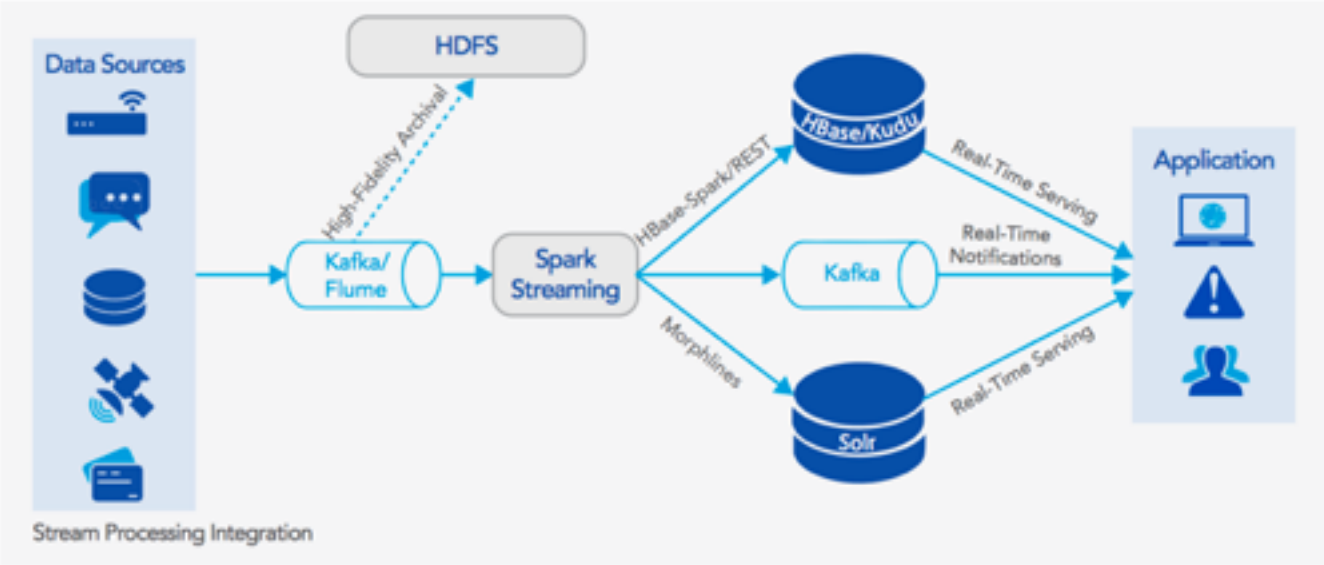
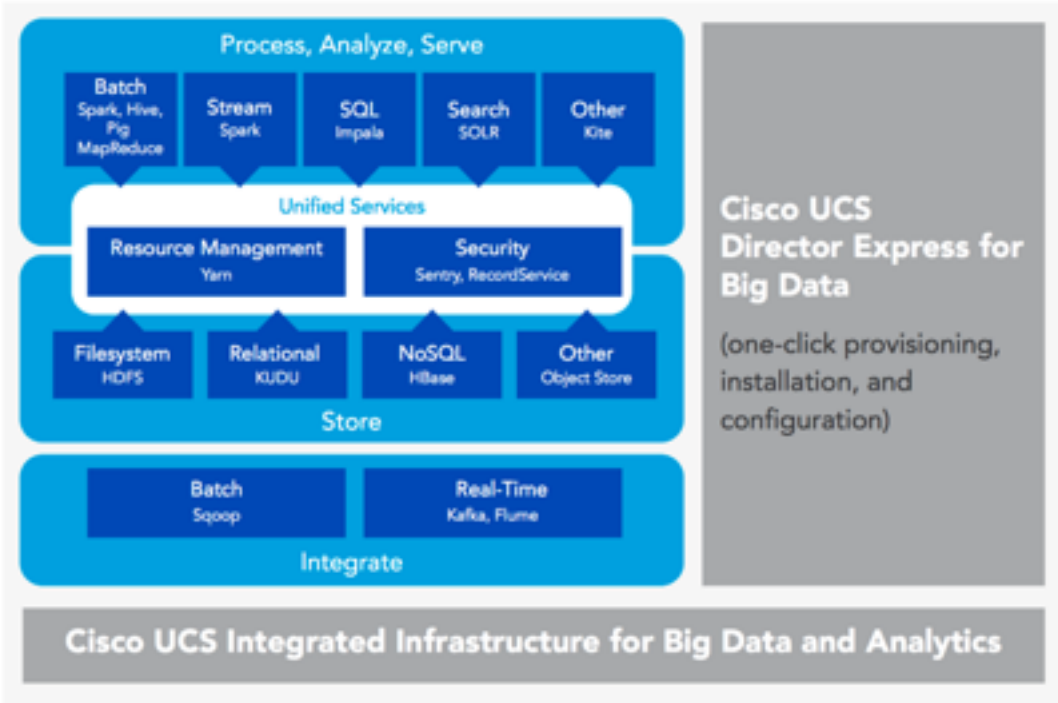


Figure 1. Typical Architecture with Cloudera Enterprise Data Hub



cloudera

Proposed Architecture for Enterprise Data Platform

AI Platform
KOGENTIX
DataRobot
skymind

Advanced Analytics
MicroStrategy
TRIFACTA
TIBCO Spotfire
ARCADIA DATA
SAS

Data Visualization
tableau
GOMDATA
Qlik
SAP Lumira
SAP BW

- Meteorological Data
- Sensors Data
- Historian/SCADA
- Unstructured files
- Social
- Video/Image
- Geospatial

CLOUDERA ENTERPRISE

PROCESS	DISCOVER & ANALYZE	SERVE	GOVERN	OPERATE	DATA SCIENCE
BATCH — MR, SPARK, HIVE, PIG	SQL — IMPALA	ONLINE — HBASE	CLOUDERA NAVIGATOR: Metadata, Policies, Security Audit, Lineage, Encryption, Active Data Optimizer	CLOUDERA MANAGER: Deployment, Configuration, Monitoring dashboard, Proactive health check, SMTP and SNMP Alert, Diagnostic, Audit trials, Secure cluster, BDR, Comprehensive API	CLOUDERA DATA SCIENCE WORKBENCH: Secured Data Science: R, Python, Scala; Hadoop Integration through Spark, Impala Self servicing sandboxing based on Docker
STREAM — SPARK	SEARCH — SOLR			CLOUD: CLOUDERA DIRECTOR — Install Cloudera on Cloud; ALTUS — PaaS Data Engineering Analytics DB	
	MODEL — SPARK				

UNIFIED DATA SERVICES
RESOURCE MANAGEMENT — YARN, Zookeeper
SECURITY — SENTRY

STORE
FILESYSTEM — HDFS
RELATIONAL — KUDU
NoSQL — HBASE
INGEST — SQOOP, FLUME, KAFKA

SAP Partner
DATAVARD GLUE
DATAVARD

SAP
BW HANA
BPC
any SAP NW

Enterprise Data Warehouse

cloudera

Big Data Blueprints: Cisco Validated Designs

Designs Big Data

Cisco® Validated Designs with Cloudera

What you get

- Industry-leading partnerships
- Tested and validated reference architectures to meet performance, capacity, and scale
- Joint engineering lab
- Extensive options for data management (Hadoop, MPP, and NoSQL) to meet your business needs
- Solution bundles optimized for cost of ownership and ease of ordering

Solution designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments.

cloudera

Our Customers' Success Stories

cloudera

CASE STUDY

DATA-DRIVEN PRODUCTS

TRANSPORTATION
» PREDICTIVE MAINTENANCE
» IMPROVED SERVICE
» DATA DRIVEN PRODUCTS

NAVISTAR®

Using Predictive Maintenance to Improve Performance and Reduce Fleet Downtime

- OnCommand Connection is collecting telematics and geolocation data across the fleet
- Reduced maintenance costs to $.03 per mile from $.12-$.15 per mile
- Centralizing data from 13 systems with varying frequency and semantic definitions
- Real-time visibility of 250,000+ trucks in order to improve uptime and vehicle performance

cloudera

CASE STUDY

DATA-DRIVEN PRODUCTS

TRAVEL & TRANSPORTATION
» SMART BUILDINGS
» PREDICTIVE MAINTENANCE
» ADVANCED ANALYTICS

Schiphol Group

Smart Buildings - Preventative Maintenance

Using Sensors & IoT to Improve Passenger Safety and Airport Efficiency

Challenge:

- Improve traveler satisfaction and safety, by reducing downtime for critical operational machinery

Solution:

- Cloudera on Azure to capture, secure, and correlate sensor (IoT) data collected from escalators, elevators, and baggage carousels
- Provide necessary fixes to prevent unplanned downtime

cloudera

CASE STUDY

2016 Data Impact Award Winner
State of Kentucky Department of Transportation

Smart Cities

Enabling the State of Kentucky manage snow and ice events in real time

Challenge:

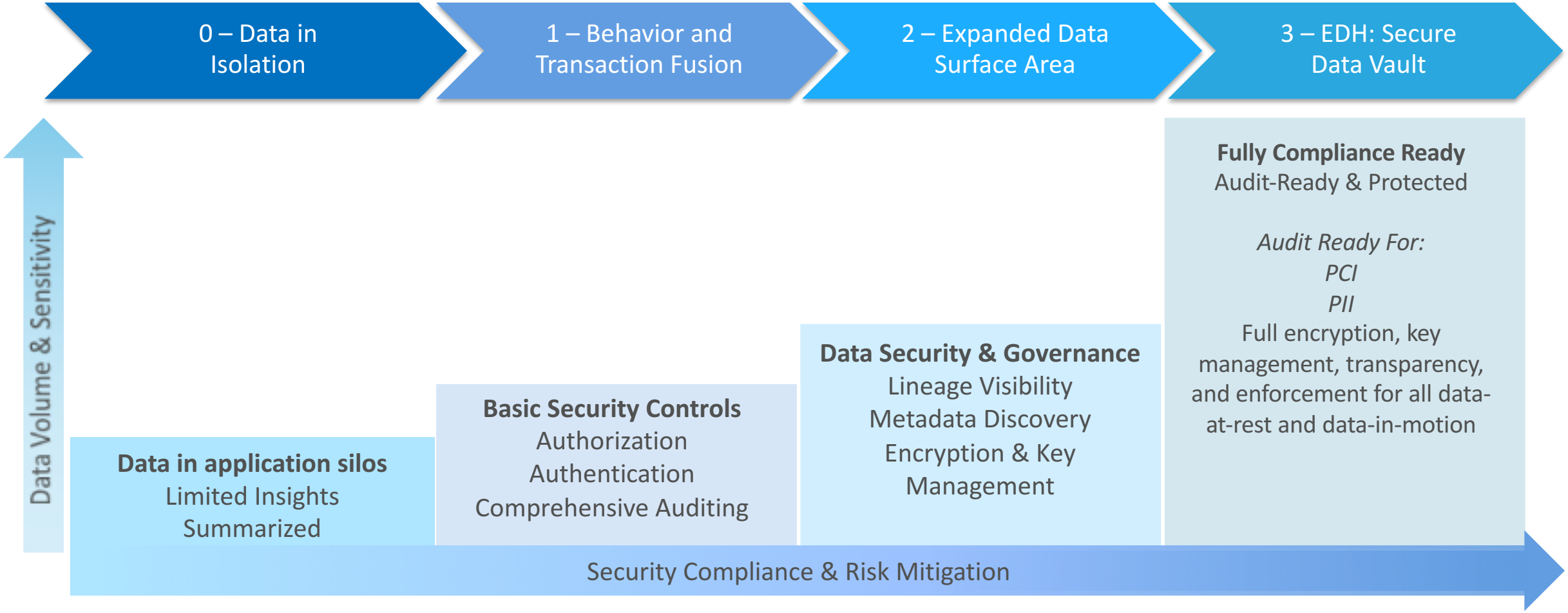
- Needed more efficient approach to inclement weather road management

Solution:

- Real-time weather response system that incorporates real-time data from Waze, HERE, ESRI's GeoEvent processor, and Automatic Vehicle Locations (providing sensor data from salt trucks).
- KYTC aggregates 15-20 million records every day and process more than a million records per second.

cloudera

Data Protection & Governance



cloudera

Why Cloudera

The **Platform** for Next-Generation Analytics

Cloudera Enterprise delivers the capabilities required by the largest enterprises, spanning analytics, security, governance, and management. We make Hadoop fast, easy, and secure.

The **Experience** to Help You Succeed

No one knows Hadoop like Cloudera.

As the first Hadoop company, Cloudera is the world's leading contributor to and provider of enterprise Hadoop, with experience you can rely on to help you succeed.

Open **Innovation**

Our unique hybrid open source strategy enables us to lead the enterprise expansion of the Hadoop ecosystem, driving innovative new capabilities and open standards in the community.

cloudera

cloudera

Thank you

marilyn@cloudera.com | +65 9822 2338
daming@cloudera.com | +65 9368 2316

