CLOUDERA

CLOUDERA NAVIGATOR

Philip Morch	Account Executive
Johannes Muselaers	Solution Engineer
Philippe Lanckvrind	Solution Engineer

WHAT MAKES BIG DATA GOVERNANCE DIFFERENT?

Governing big data requires governing petabytes of diverse types of data

No one application will solve every big data governance problem

Applications are shifting to the cloud, and data governance must still be applied consistently

Self-service data discovery is mandatory for big data

DATA MANAGEMENT IS THE FOUNDATION OF ADOPTION

Governance & Compliance

Track, understand and protect access to data

- Am I prepared for an audit?
- Who's accessing sensitive data?
- What are they doing with the data?
- Is sensitive data governed and protected?

Curation

Manage and organize data assets at Hadoop scale

- How can I identify and classify sensitive data in accordance with regulation?
- How can I organize and classify data for business users?
- How can I efficiently make data available to business users?

Self-Service Discovery

Effortlessly find and trust the data that matters most

- How can I find explore data sets on my own?
- Can I trust what I find?
- How do I use what I find?
- How do I find and use related data sets?

Stewardship

Boost user productivity and cluster performance

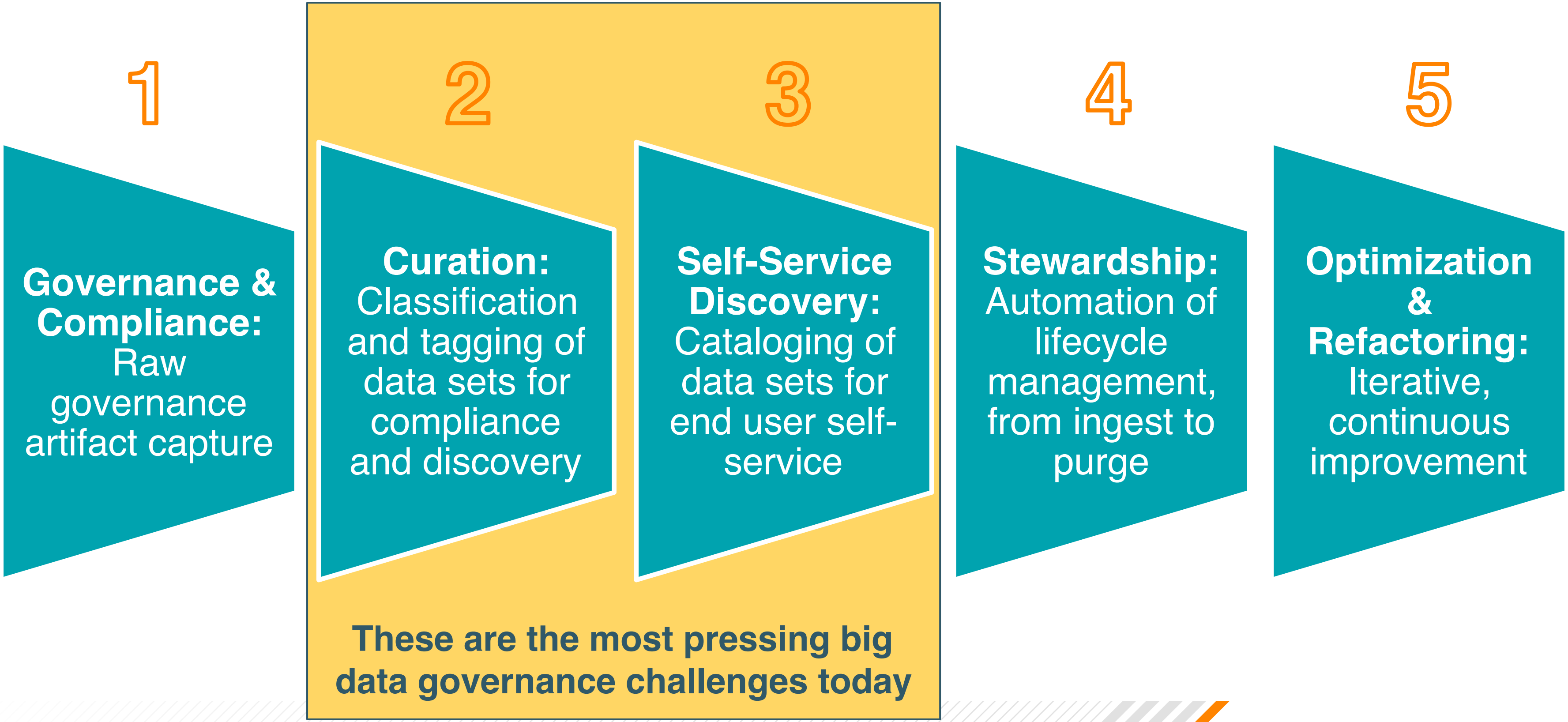
- How can I efficiently manage data lifecycle, from ingest to purge?
- How can I optimize my data models to support common access patterns?
- How can I migrate workloads to Hadoop risk-free?

BIG DATA GOVERNANCE FOUNDATION

- Centralized audits
- Unified data catalog
- Comprehensive lineage
- Data policies

CLOUDERA

THE CURRENT STATE OF BIG DATA GOVERNANCE



1

Governance & Compliance: Raw governance artifact capture

2

Curation: Classification and tagging of data sets for compliance and discovery

3

Self-Service Discovery: Cataloging of data sets for end user self-service

4

Stewardship: Automation of lifecycle management, from ingest to purge

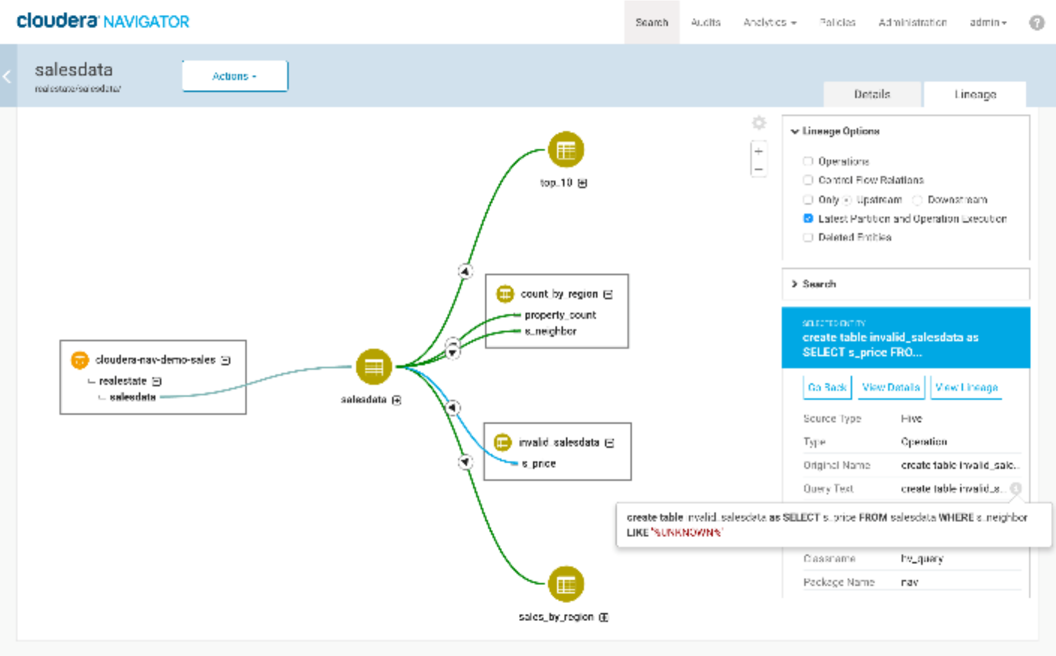
5

Optimization & Refactoring: Iterative, continuous improvement

These are the most pressing big data governance challenges today

CLOUDERA NAVIGATOR

Governance, stewardship, and discovery for big data built on machine learning and analytics



□ Operations
□ Control Flow Relations
□ Only ● Upstream ○ Downstream
☑ Latest Partition and Operation Execution
□ Deleted Entities

SELECTED ENTITY
create table invalid_salesdata as SELECT s_price FRO...

Source Type Hive
Type Operation
Original Name create table invalid_sale...
Query Text create table invalid_s...

create table invalid_salesdata as SELECT s_price FROM salesdata WHERE s_neighbor LIKE '%UNKNOWN%'

Classname hv_query
Package Name nav

Trusted for production

- Deployed by hundreds of customers across multiple industries
- Over four years in production

Compliance-grade

- The only Hadoop distribution to pass PCI audit

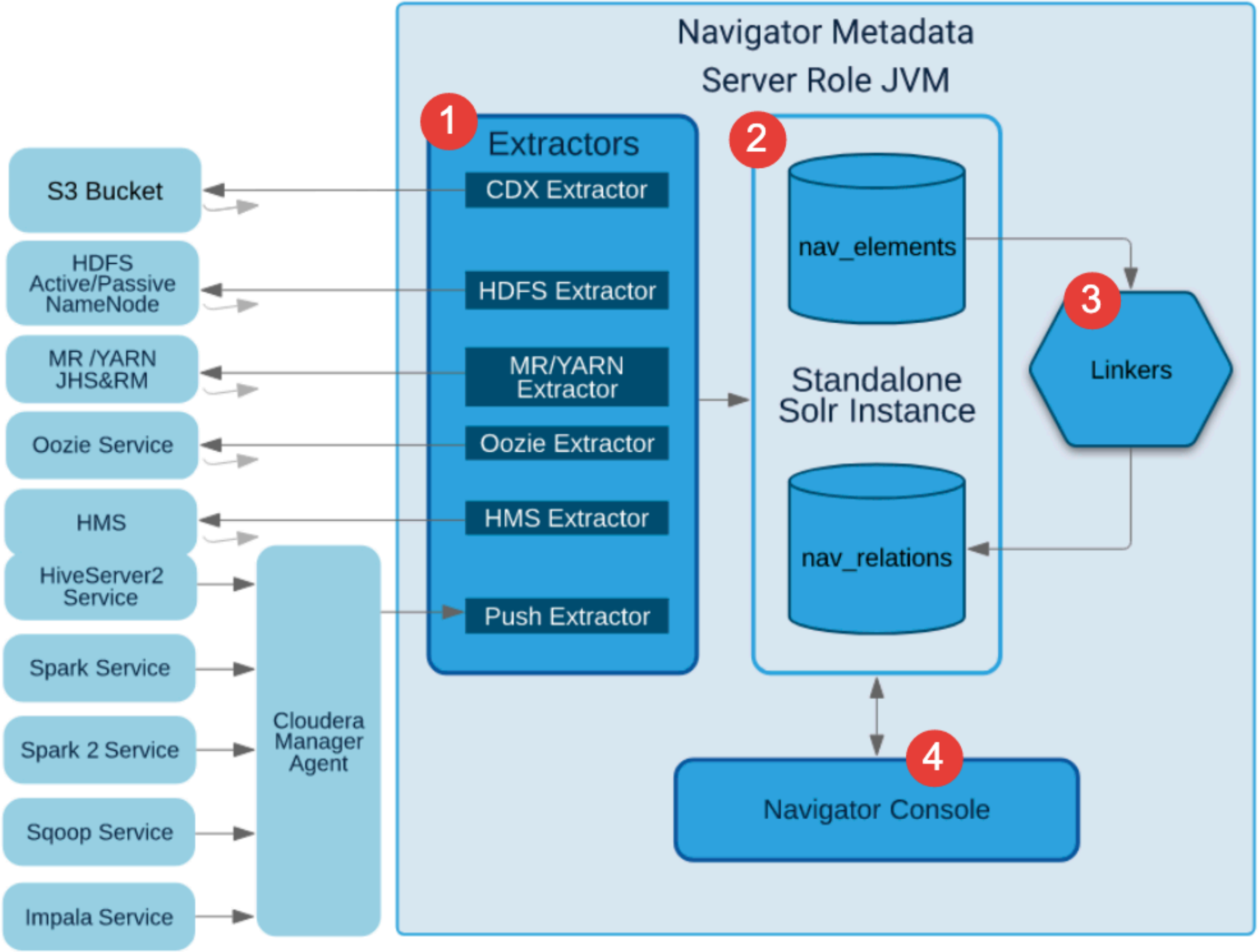
Open and interoperable

- Integrated with leading partner solutions

CLOUDERA

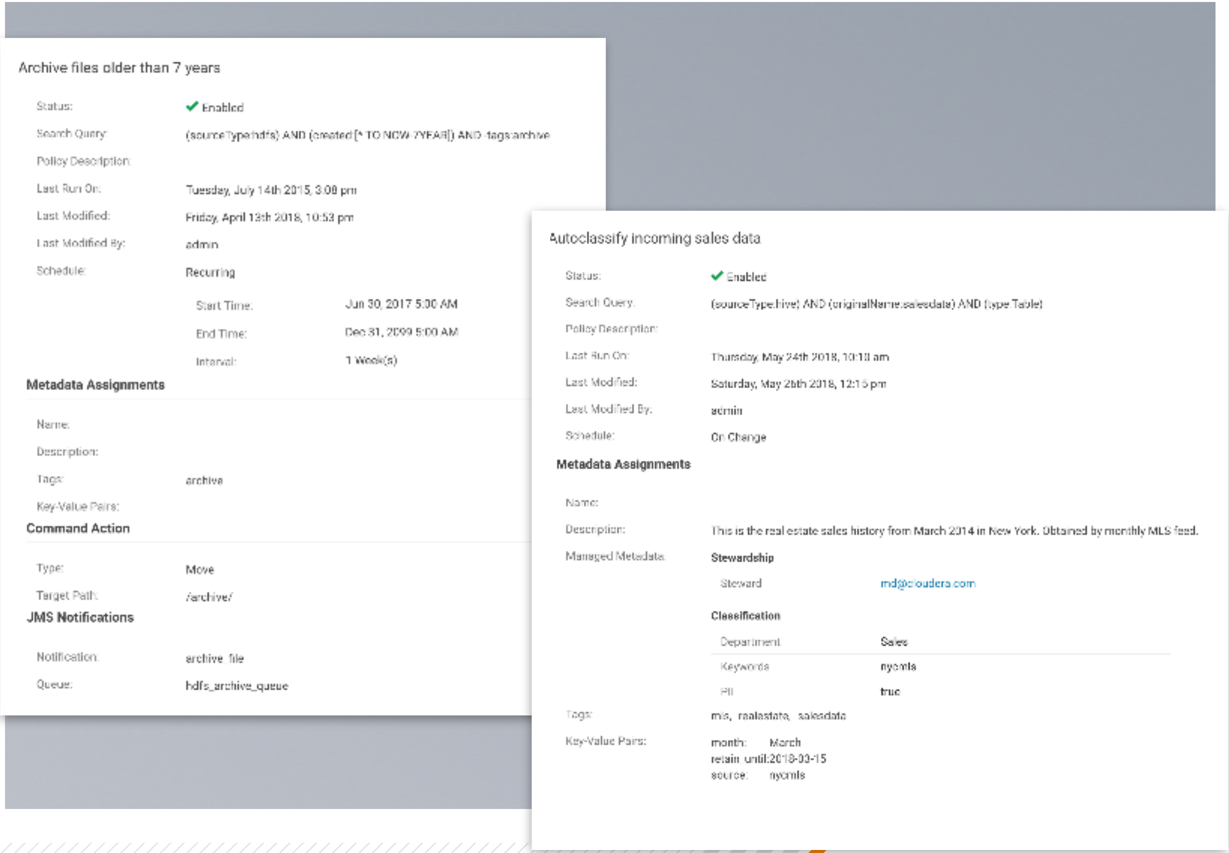
TURNKEY COMPLIANCE-GRADE **GOVERNANCE**

- Automatically collect column-level lineage and audit logs
- Effortlessly publish to enterprise governance frameworks, such as Informatica and IBM InfoSphere



AUTOMATED DATA **CURATION**

- Set up policies that classify data sets automatically upon ingest
- Add business glossary definitions and profiling information for faster self-service
- Automatically trigger data preparation, profiling, and data quality activities



Archive files older than 7 years

Status: ✓ Enabled
Search Query: (sourceType:hdfs) AND (created:[* TO NOW-7YEAR]) AND -tags:archive
Policy Description:
Last Run On: Tuesday, July 14th 2015, 3:08 pm
Last Modified: Friday, April 13th 2018, 10:53 pm
Last Modified By: admin
Schedule: Recurring
Start Time: Jun 30, 2017 5:00 AM
End Time: Dec 31, 2099 5:00 AM
Interval: 1 Week(s)

Metadata Assignments

Name:
Description:
Tags: archive
Key-Value Pairs:

Command Action

Type: Move
Target Path: /archive/

JMS Notifications

Notification: archive_file
Queue: hdfs_archive_queue

Autoclassify incoming sales data

Status: ✓ Enabled
Search Query: (sourceType:hive) AND (originalName:salesdata) AND (type:Table)
Policy Description:
Last Run On: Thursday, May 24th 2018, 10:18 am
Last Modified: Saturday, May 26th 2018, 12:15 pm
Last Modified By: admin
Schedule: On Change

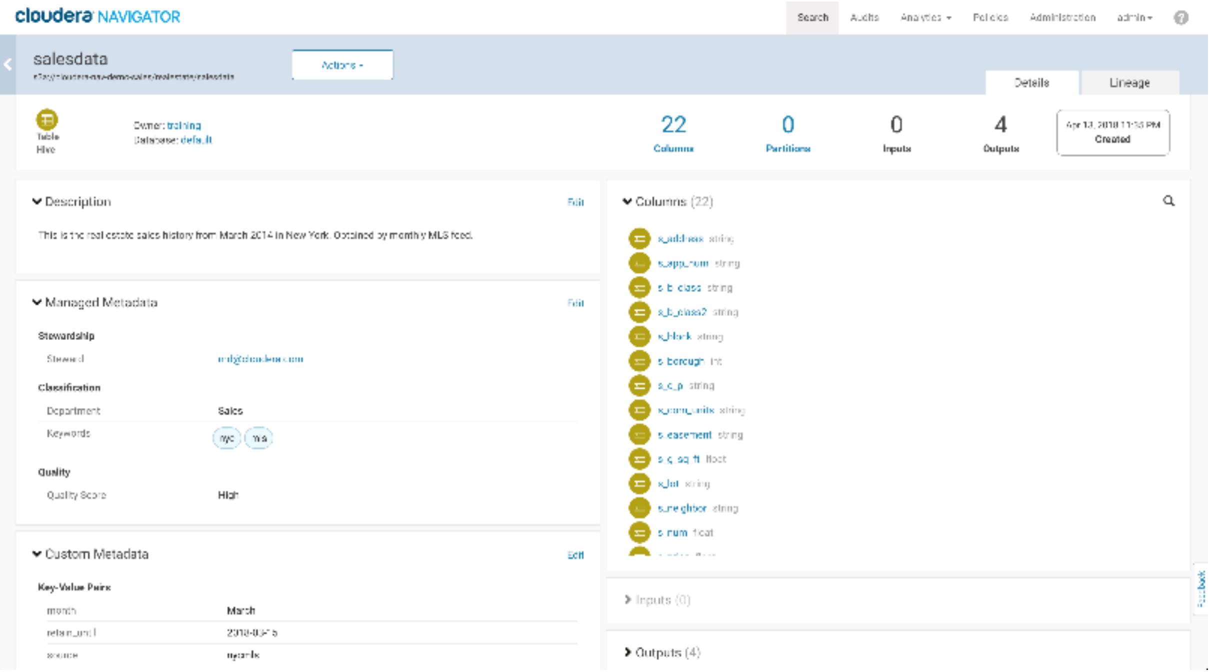
Metadata Assignments

Name:
Description: This is the real estate sales history from March 2014 in New York. Obtained by monthly MLS feed.
Managed Metadata: **Stewardship**
Steward md@cloudera.com
Classification
Department Sales
Keywords nycmls
PII true
Tags: mls, realestate, salesdata
Key-Value Pairs: month: March
retain_until:2018-03-15
source: nycmls

CLOUDERA

OUT-OF-THE-BOX SELF-SERVICE **DISCOVERY**

- Let business users collaborate on and classify data sets while leveraging centrally-curated classifications
- Leverage deep analytics on historical usage to empower users to find, trust, and use data sets on their own



salesdata
[illegible]

Owner: training
Database: default

22 Columns | 0 Partitions | 0 Inputs | 4 Outputs | Apr 13, 2018 11:35 PM Created

Description

This is the real estate sales history from March 2014 in New York. Obtained by monthly MLS feed.

Managed Metadata

Stewardship

Steward	[illegible]@cloudera.com

Classification

Department	Sales
Keywords	nyc ma

Quality

Quality Score	High

Custom Metadata

Key-Value Pairs

month	March
retain_until	2018-03-15
source	nycmls

Columns (22)

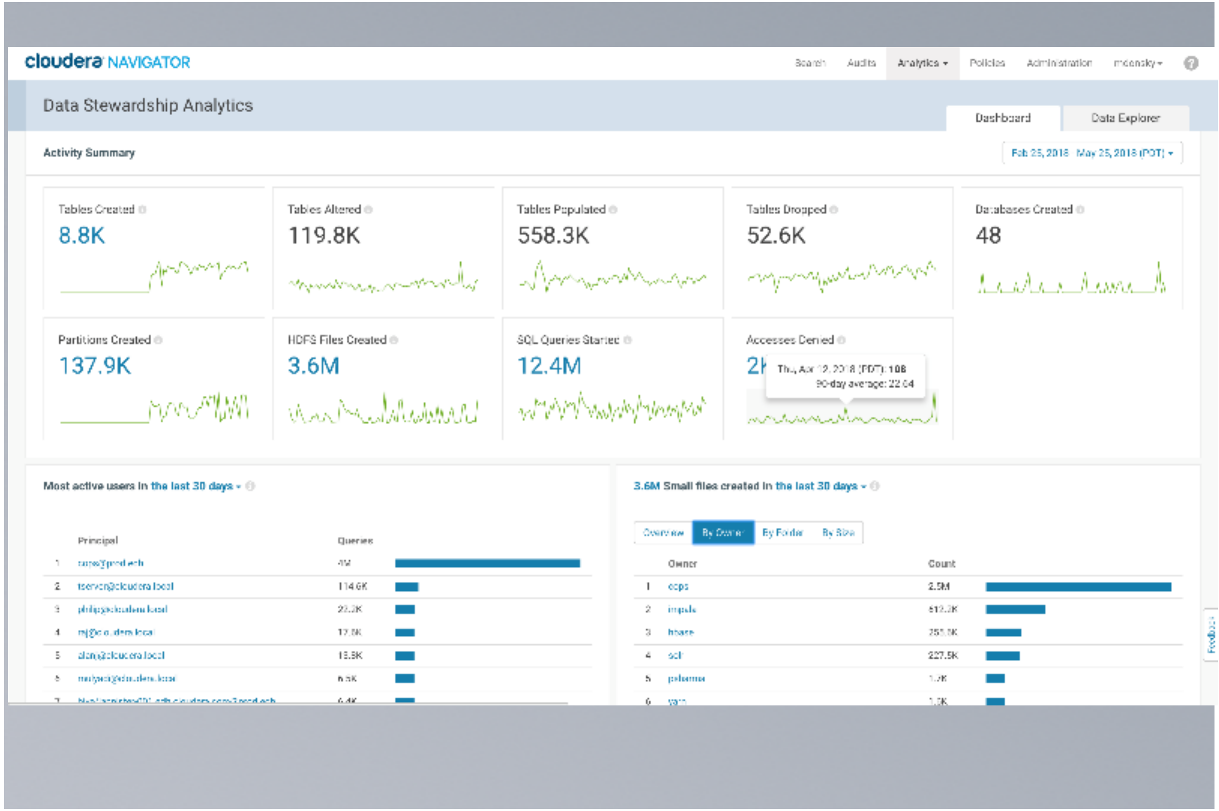
- s_address string
- s_app_num string
- s_b_class string
- s_b_class2 string
- s_block string
- s_borough int
- s_c_p string
- s_com_units string
- s_easement string
- s_g_sq_ft float
- s_lot string
- s_neighbor string
- s_num float

Inputs (0)

Outputs (4)

STEWARDSHIP BUILT ON MACHINE LEARNING AND ANALYTICS

At-a-glance stewardship metrics



Data Stewardship Analytics

Activity Summary

Feb 25, 2018 - May 25, 2018 (PDT) ▾

Tables Created 8.8K	Tables Altered 119.8K	Tables Populated 558.3K	Tables Dropped 52.6K	Databases Created 48
Partitions Created 137.9K	HDFS Files Created 3.6M	SQL Queries Started 12.4M	Accesses Denied 2K	

Thu, Apr 12, 2018 (PDT): 108
90-day average: 22.04

Most active users in the last 30 days ▾

	Principal	Queries
1	cops@prod.ech	4M
2	tserver@cloudera.local	114.6K
3	philip@cloudera.local	22.2K
4	mj@cloudera.local	17.5K
5	alan@cloudera.local	15.3K
6	mdyad@cloudera.local	6.5K
7	[illegible]	6.4K

3.6M Small files created in the last 30 days ▾

Overview | By Owner | By Folder | By Size

	Owner	Count
1	cops	2.5M
2	impala	512.2K
3	hbase	255.5K
4	sol	227.5K
5	paloma	1.7K
6	yarn	1.5K

CLOUDERA

Cloudera Navigator
Walk Through

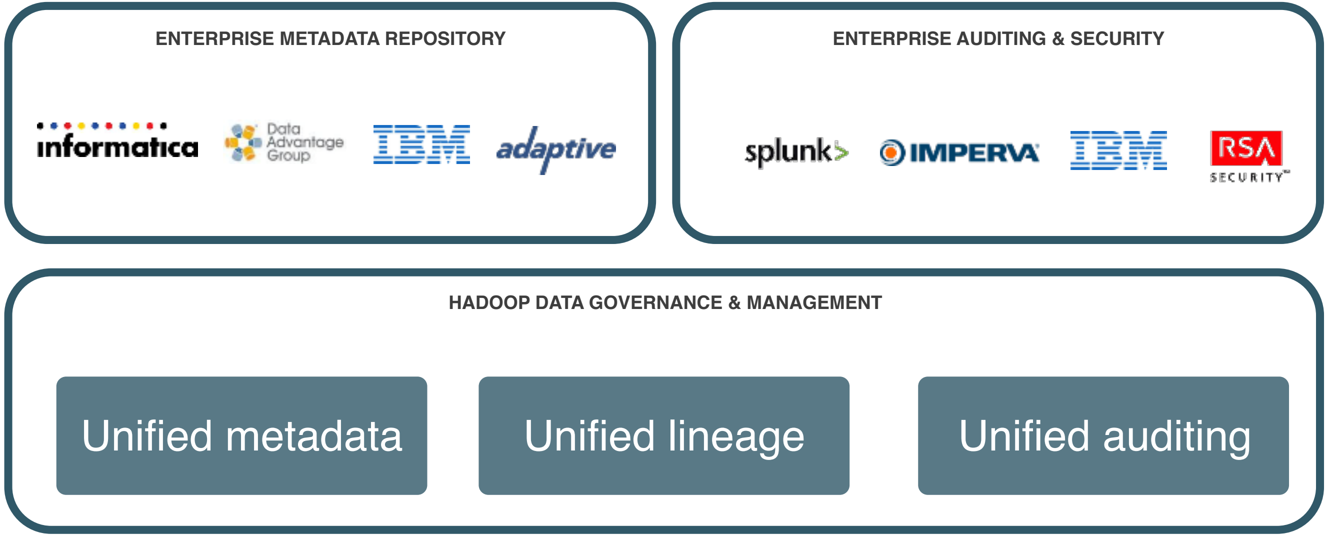
Use Cases & Best Practices

USE CASES: COMPLIANCE

Compliance

Track, understand and protect access to data

- Am I prepared for an audit?
- Who’s accessing sensitive data?
- What are they doing with the data?
- Is sensitive data governed and protected?



ENTERPRISE METADATA REPOSITORY

ENTERPRISE AUDITING & SECURITY

HADOOP DATA GOVERNANCE & MANAGEMENT

Unified metadata

Unified lineage

Unified auditing

Common use cases:

- Security breach detection
- Data access tracking for PCI compliance
- Audit defense

USE CASES: STEWARDSHIP & CURATION

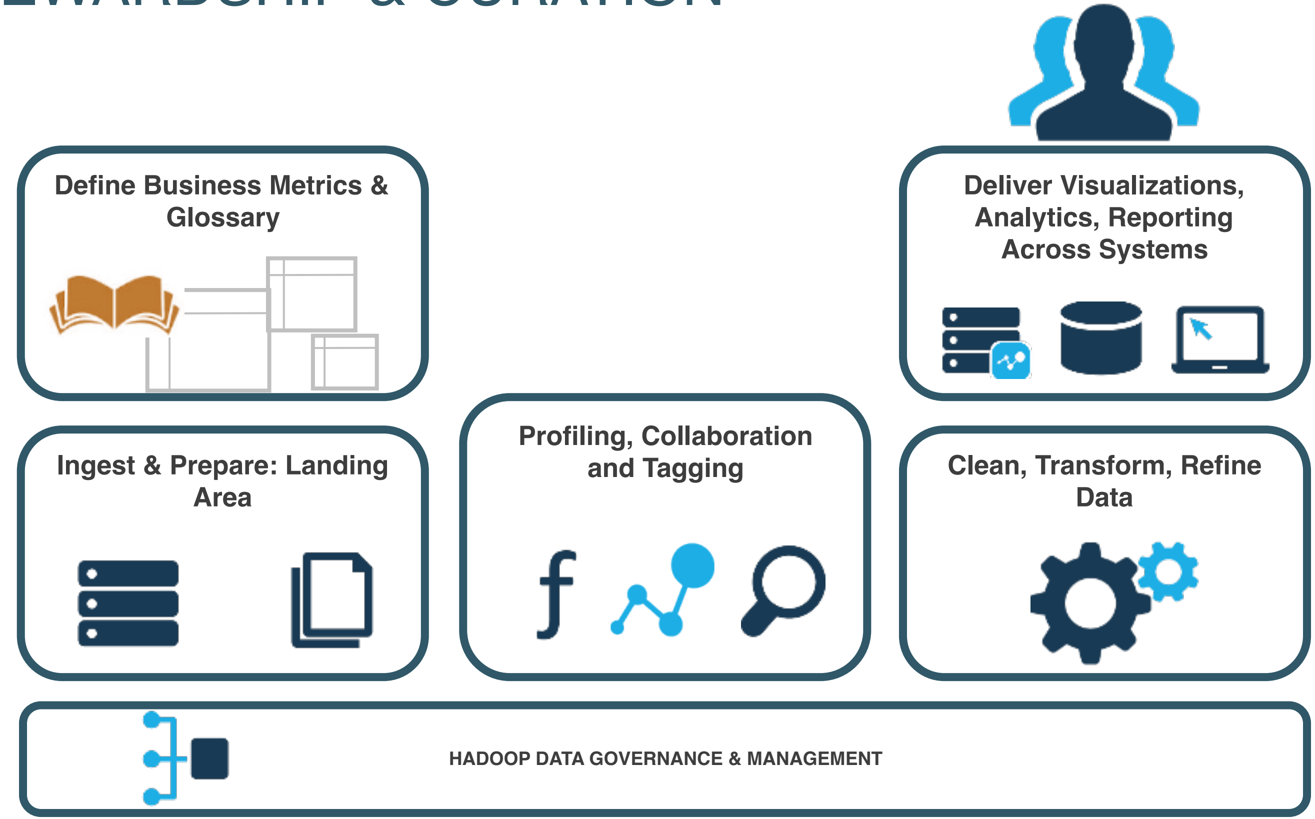
Stewardship, Curation & Discovery

Manage, classify, and use data assets at Hadoop scale

How can I efficiently manage data lifecycle, from ingest to purge?

How can I identify and classify sensitive data in accordance with regulation?

How can end users find, trust, and use data sets on their own?



Define Business Metrics & Glossary

Deliver Visualizations, Analytics, Reporting Across Systems

Ingest & Prepare: Landing Area

Profiling, Collaboration and Tagging

Clean, Transform, Refine Data

HADOOP DATA GOVERNANCE & MANAGEMENT

USE CASES: STEWARDSHIP & CURATION

Stewardship, Curation & Discovery

Manage, classify, and use data assets at Hadoop scale

How can I efficiently manage data lifecycle, from ingest to purge?

How can I identify and classify sensitive data in accordance with regulation?

How can end users find, trust, and use data sets on their own?

Define Business Metrics & Glossary

informatica
Data Advantage Group
Collibra
adaptive
IBM

Deliver Visualizations, Analytics, Reporting Across Systems

tableau
SAP BusinessObjects
ZOOMDATA
platfora
sas

Profiling, Collaboration and Tagging

informatica
talend
WATERLINE DATA SCIENCE
pentaho
TRIFACTA
IBM
Datameer
Paxata

Ingest & Prepare: Landing Area

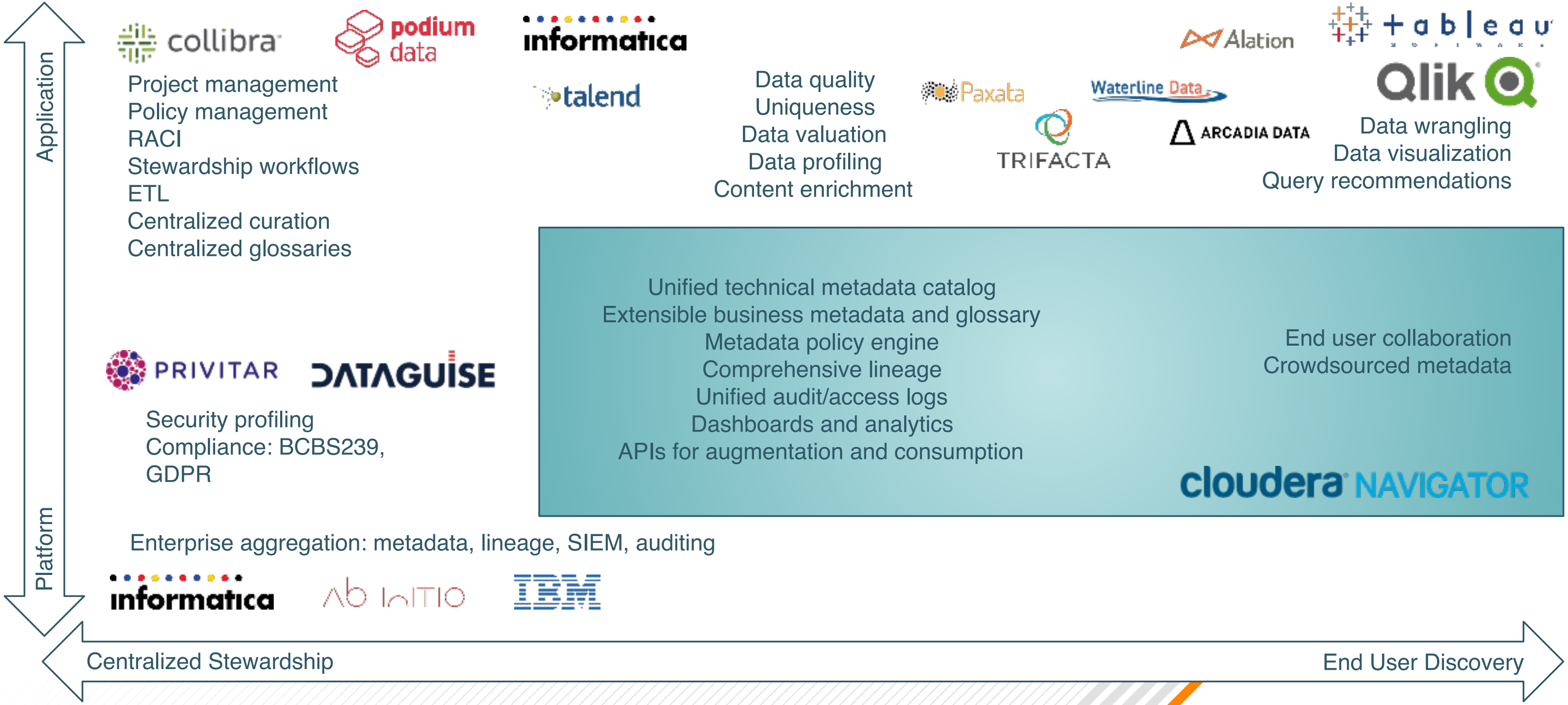
informatica
pentaho
syncsort
IBM
TRIFACTA
talend
Paxata

Clean, Transform, Refine Data

informatica
talend
TRILLIUM SOFTWARE
pentaho
Datameer
TRIFACTA
platfora
IBM

HADOOP DATA GOVERNANCE & MANAGEMENT

CLOUDERA NAVIGATOR'S VAST PARTNER ECOSYSTEM



Project management
Policy management
RACI
Stewardship workflows
ETL
Centralized curation
Centralized glossaries

Data quality
Uniqueness
Data valuation
Data profiling
Content enrichment

Data wrangling
Data visualization
Query recommendations

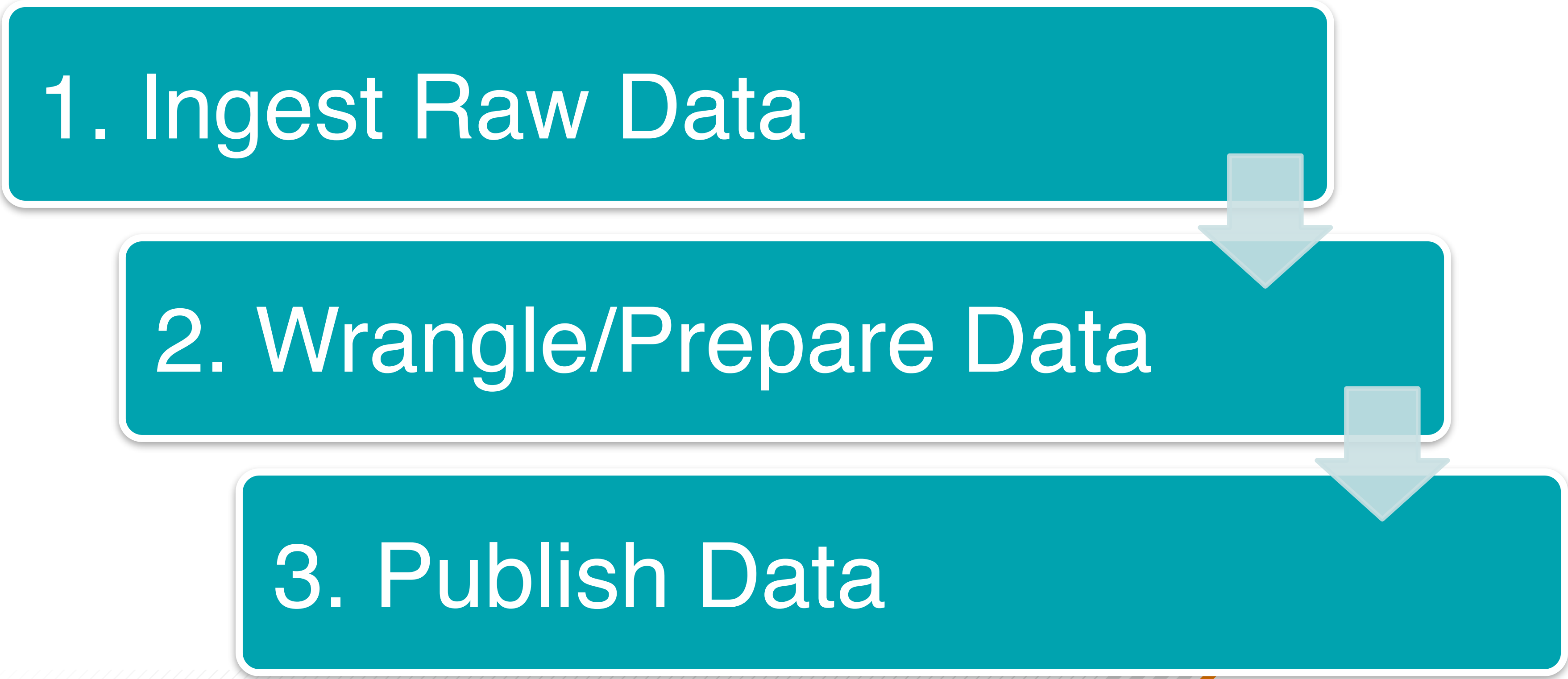
Security profiling
Compliance: BCBS239, GDPR

Unified technical metadata catalog
Extensible business metadata and glossary
Metadata policy engine
Comprehensive lineage
Unified audit/access logs
Dashboards and analytics
APIs for augmentation and consumption

End user collaboration
Crowdsourced metadata

Enterprise aggregation: metadata, lineage, SIEM, auditing

CURATION: ALIGN WITH KEY LIFECYCLE STAGES

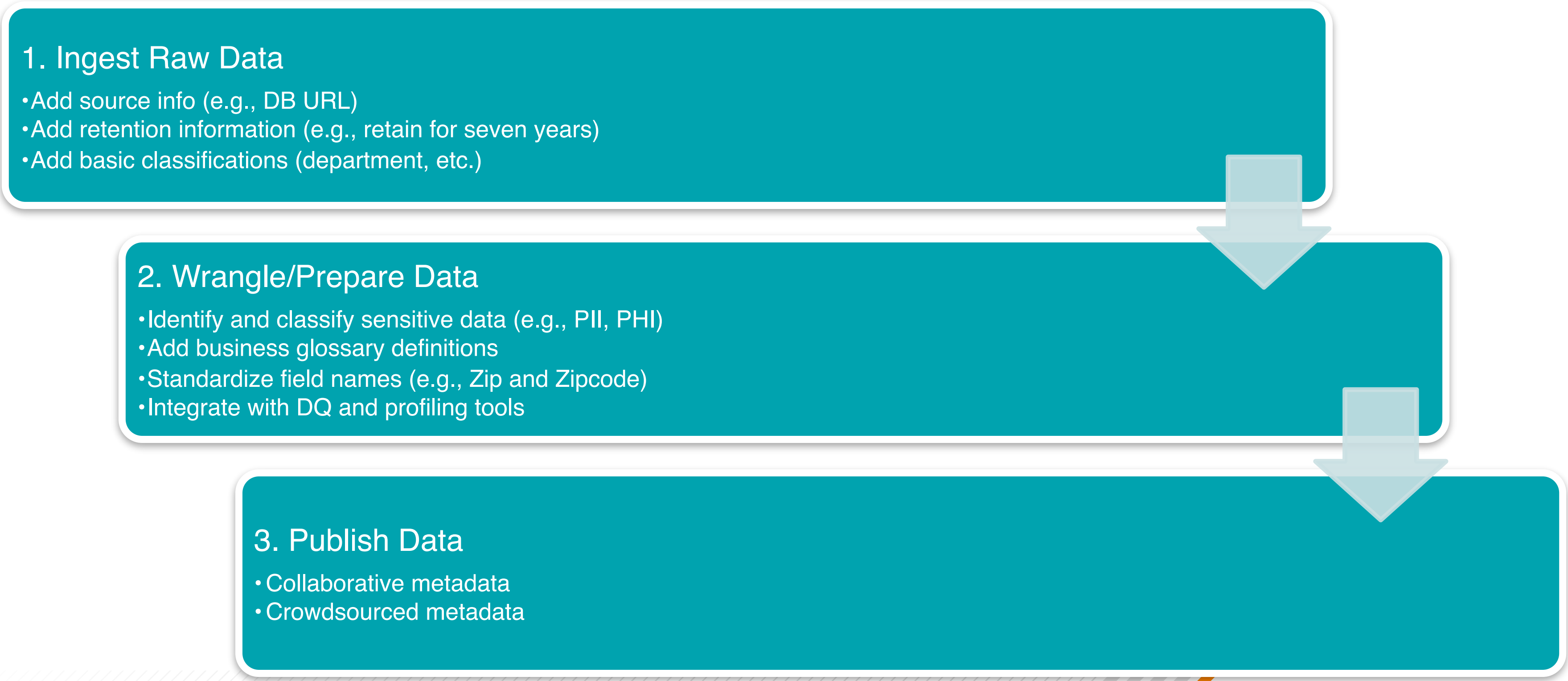


1. Ingest Raw Data

2. Wrangle/Prepare Data

3. Publish Data

ALIGN DATA CURATION WITH KEY LIFECYCLE STAGES



1. Ingest Raw Data

- Add source info (e.g., DB URL)
- Add retention information (e.g., retain for seven years)
- Add basic classifications (department, etc.)

2. Wrangle/Prepare Data

- Identify and classify sensitive data (e.g., PII, PHI)
- Add business glossary definitions
- Standardize field names (e.g., Zip and Zipcode)
- Integrate with DQ and profiling tools

3. Publish Data

- Collaborative metadata
- Crowdsourced metadata

MANAGED METADATA VS CUSTOM METADATA

	Managed Metadata	Custom Metadata
Intended usage	Centrally-curated metadata Sentry ABAC Metadata ABAC	End-user collaboration Data set sharing
Assigned to specific entities (e.g., columns)	✔	
Typed and Validated (e.g., Boolean, Date, Enumeration)	✔	
Editable by data curators	✔	✔
Editable by end users		✔
Viewable by data curators	✔	✔
Viewable by end users	✔	✔

Looking Forward

MERGING OF TECHNOLOGIES

Bringing new capabilities

Authorization

Sentry™

+

Apache Ranger

Governance

cloudera® navigator

+

Apache Atlas

Data fabric

Data Steward Studio

DATA GOVERNANCE: IT'S A TEAM SPORT!

Sponsor

Champions data governance across enterprise

Data Owner

Accountable for all data generated by an agency

Data Council

Coordinate cross-agency data management activities

Data Steward

Manages business requirements for data sharing

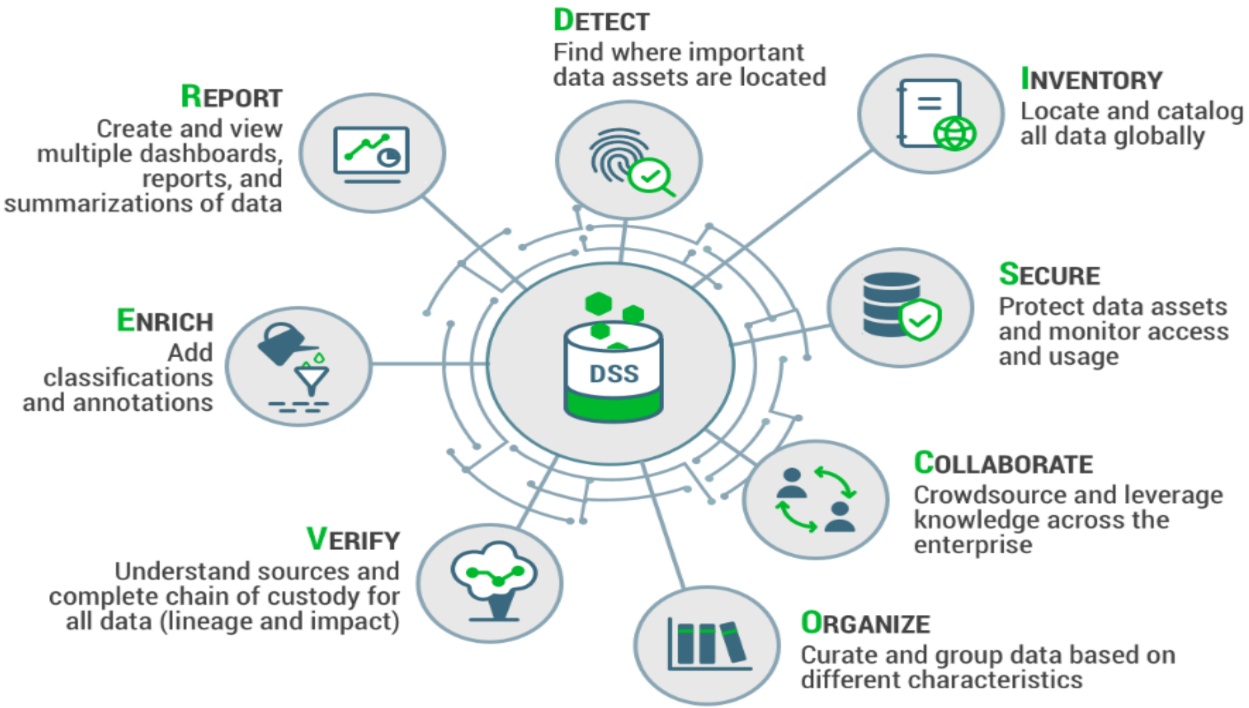
Data Curator

Implements business data requirements

Business Data SME

Supports the Data Steward in data related activities

DSS Discover With Data Steward Studio



DETECT
Find where important data assets are located

INVENTORY
Locate and catalog all data globally

SECURE
Protect data assets and monitor access and usage

COLLABORATE
Crowdsource and leverage knowledge across the enterprise

ORGANIZE
Curate and group data based on different characteristics

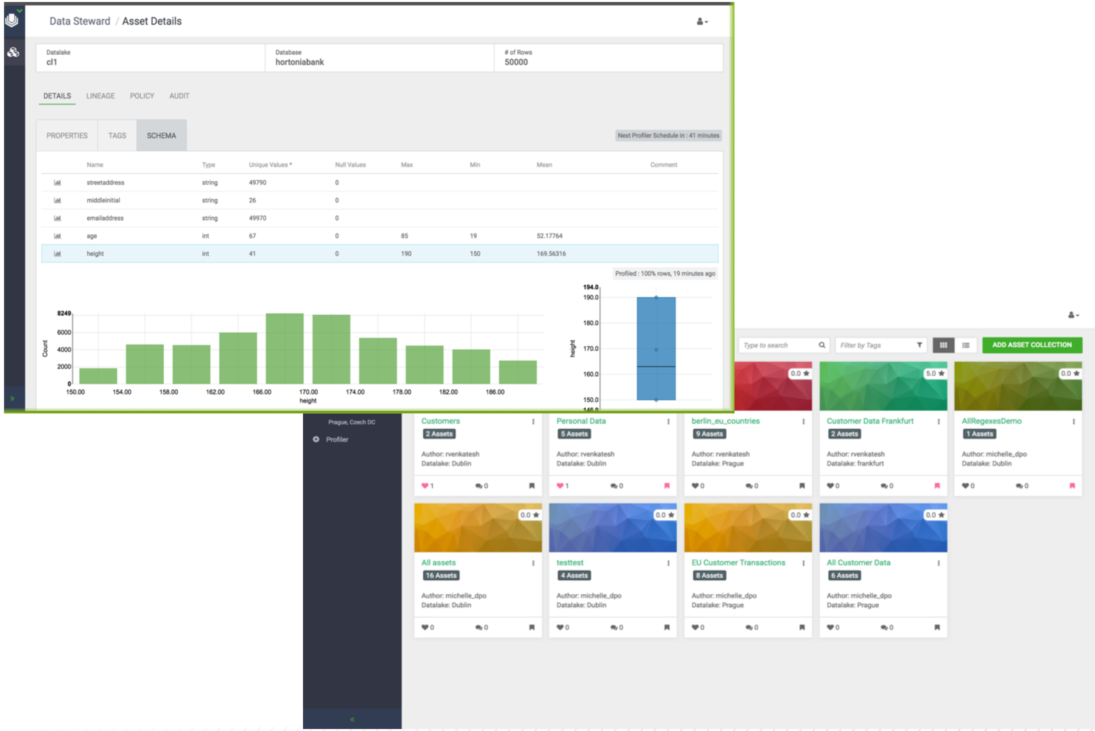
VERIFY
Understand sources and complete chain of custody for all data (lineage and impact)

ENRICH
Add classifications and annotations

REPORT
Create and view multiple dashboards, reports, and summarizations of data

DATA STEWARD STUDIO

DSS provides the “tooling” part of the People, Processes, and Technology required for Hybrid Data Lake Governance



Data Steward / Asset Details

Datalake: cl1

Database: hortoniabank

# of Rows: 50000

Next Profiler Schedule in : 41 minutes

Name	Type	Unique Values *	Null Values	Max	Min	Mean	Comment
streetaddress	string	49790	0				
middleinitial	string	26	0				
emailaddress	string	49970	0				
age	int	67	0	85	19	52.17764	
height	int	41	0	190	150	169.56316	

Profiled : 100% rows, 19 minutes ago

Customers, 2 Assets, Author: rvenkatesh, Datalake: Dublin, 0.0 ★, ♥ 1, 0

Personal Data, 5 Assets, Author: rvenkatesh, Datalake: Dublin, ♥ 1, 0

berlin_eu_countries, 9 Assets, Author: rvenkatesh, Datalake: Prague, 0.0 ★, ♥ 0, 0

Customer Data Frankfurt, 2 Assets, Author: rvenkatesh, Datalake: frankfurt, 5.0 ★, ♥ 0, 0

AllRegexesDemo, 1 Assets, Author: michelle_dpo, Datalake: Dublin, 0.0 ★, ♥ 0, 0

All assets, 16 Assets, Author: michelle_dpo, Datalake: Dublin, 0.0 ★, ♥ 0, 0

testtest, 4 Assets, Author: michelle_dpo, Datalake: Dublin, 0.0 ★, ♥ 0, 0

EU Customer Transactions, 8 Assets, Author: michelle_dpo, Datalake: Prague, 0.0 ★, ♥ 0, 0

All Customer Data, 6 Assets, Author: michelle_dpo, Datalake: Prague, 0.0 ★, ♥ 0, 0

- Profile Data for understanding shape and structure
- Organize and curate data for e.g. by domains they belong to or data usage
- Identify sensitive data
- Collaborate with broader teams on how data needs to be used and by who and provide community ratings for crowdsourcing knowledge
- Monitor ongoing usage, visualize chain of custody and trustworthiness for longer term use, understand data protection

THANK YOU

CLOUDERA