CLOUDERA

Cloudera Navigator

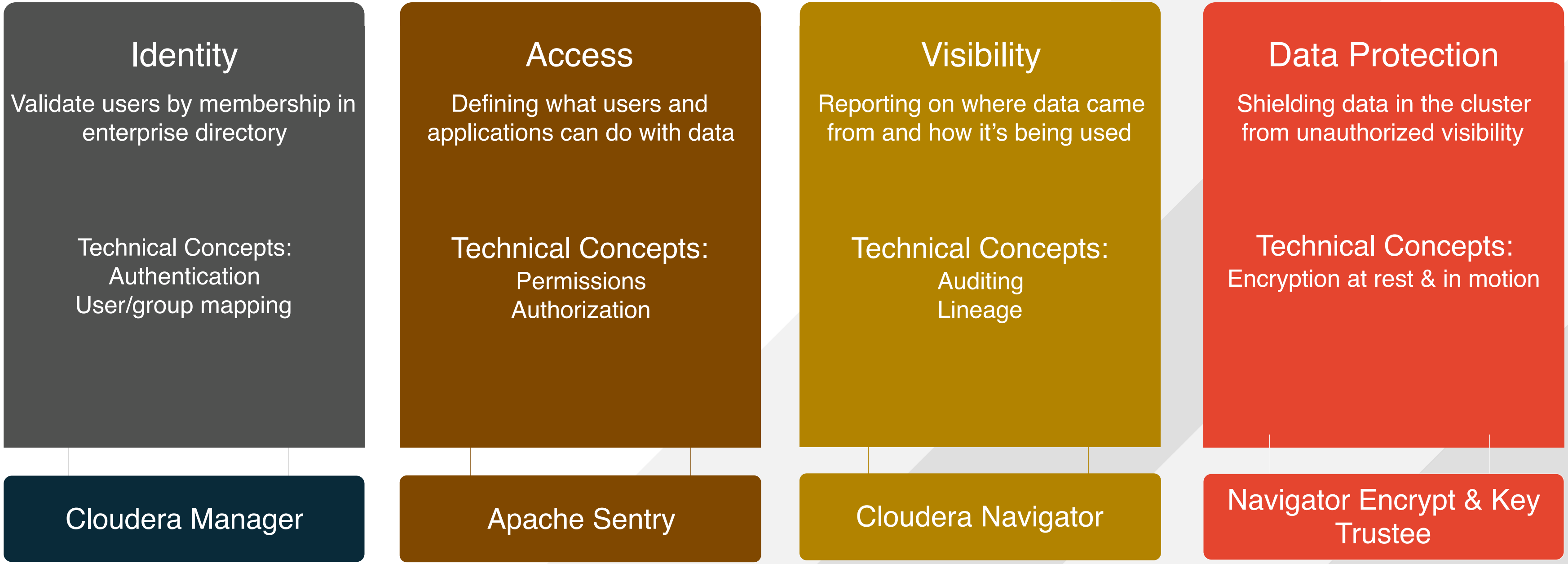
Enterprise Data Governance

Johannes Muselaers
Solutions Engineer Nordics
jmuselaers@cloudera.com
+46 72 588 1091

PDF of this presentation:

https://github.com/jmuselaers/Presentations/Cloudera Navigator.pdf

Cloudera Enterprise-Grade Security and Governance



Identity

Validate users by membership in enterprise directory

Technical Concepts:
Authentication
User/group mapping

Cloudera Manager

Access

Defining what users and applications can do with data

Technical Concepts:
Permissions
Authorization

Apache Sentry

Visibility

Reporting on where data came from and how it's being used

Technical Concepts:
Auditing
Lineage

Cloudera Navigator

Data Protection

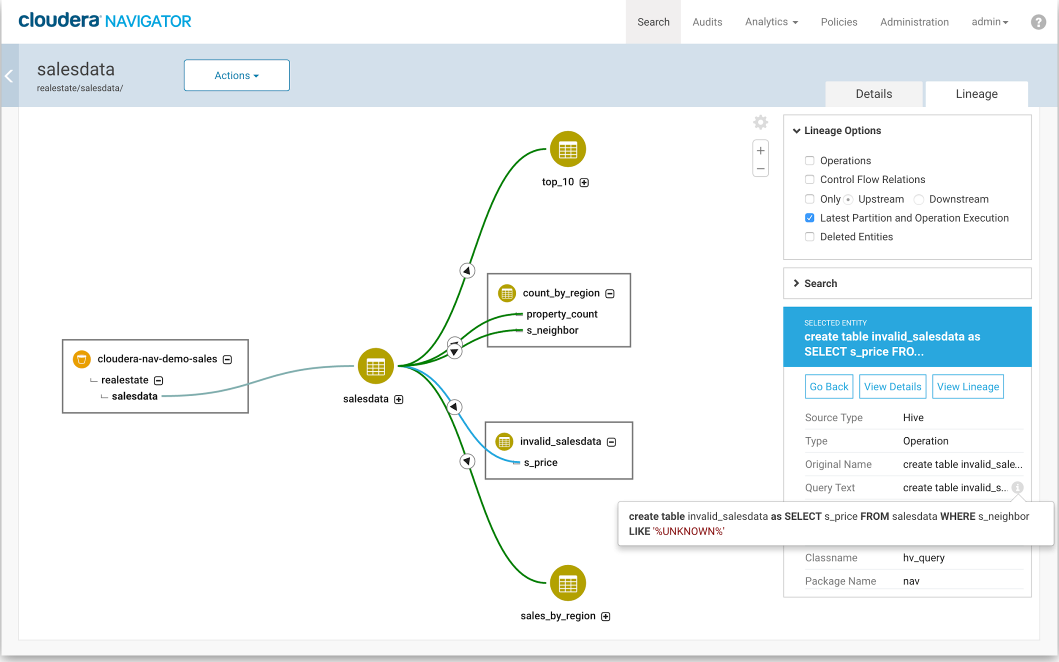
Shielding data in the cluster from unauthorized visibility

Technical Concepts:
Encryption at rest & in motion

Navigator Encrypt & Key Trustee

CLOUDERA NAVIGATOR

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



Lineage Options

- ☐ 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...
Classname	hv_query
Package Name	nav

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

Trusted for production

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

Compliance-grade

- The only Hadoop distribution to pass PCI audit

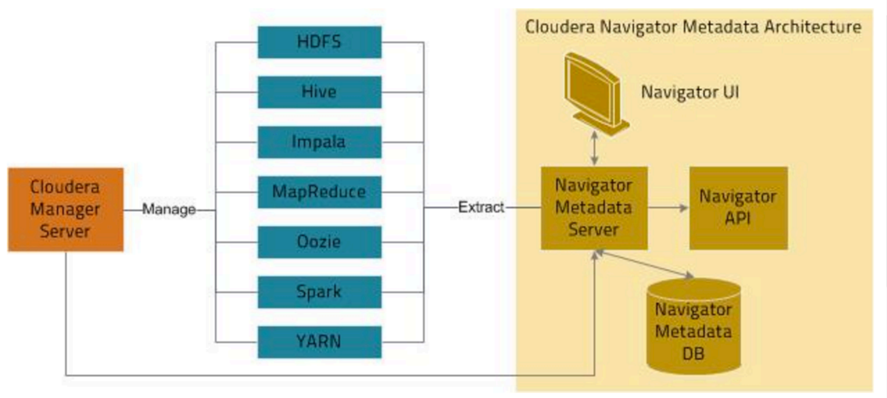
Open and interoperable

- Integrated with leading partner solutions

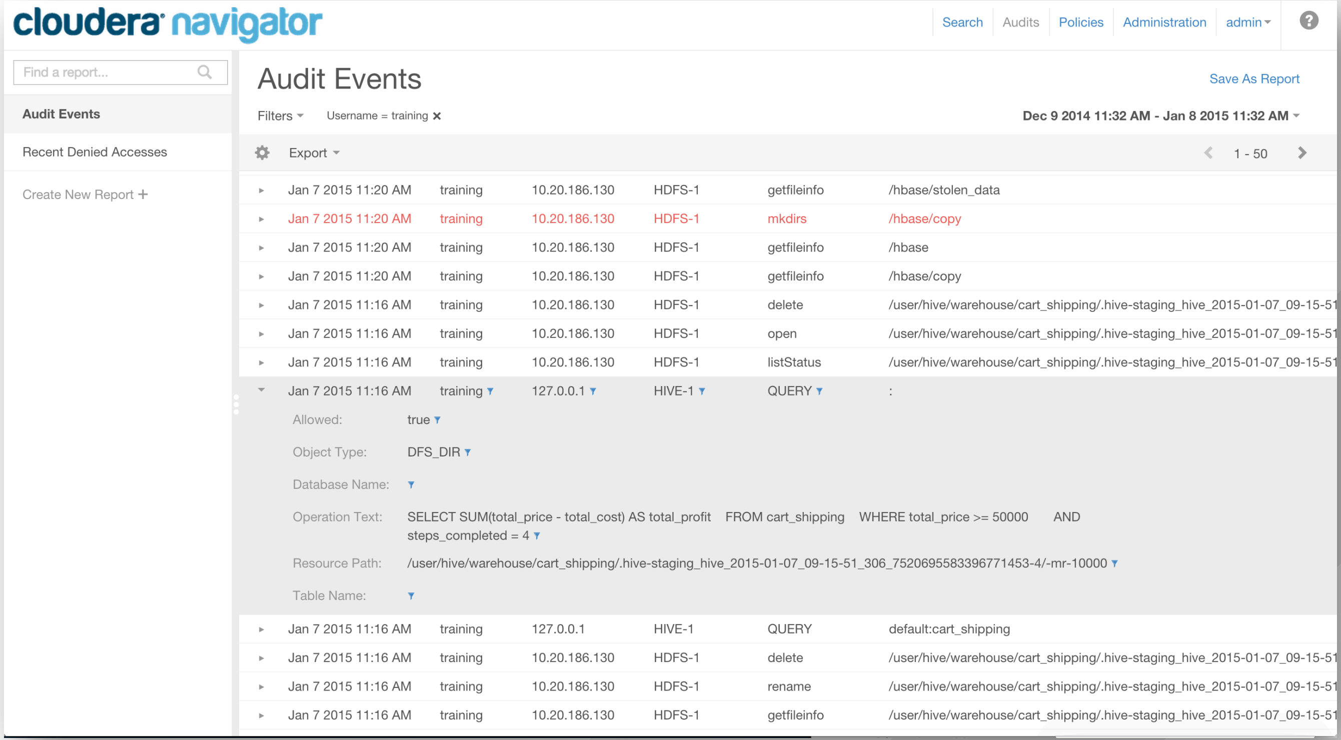
CLOUDERA

Cloudera Navigator

- **Audit & Access Control**
  - Maintain full audit history across Hue, Cloudera Manager, HDFS, Impala, HBase & Solr
  - Ensuring appropriate permissions and reporting on data access for compliance
- **Discovery & Exploration**
  - Finding out what data is available and what it looks like
  - Tag datasets with metadata for pooling, knowledge sharing and common definitions
- **Lineage**
  - Tracing data back to its original source
  - Download and consume lineage data
- **Policy Engine**
  - automated classification on arrival
- **SDK** for integration with 3rd party tooling



Auditing



Audit Events

Username = training ✕

Dec 9 2014 11:32 AM - Jan 8 2015 11:32 AM

Jan 7 2015 11:20 AM	training	10.20.186.130	HDFS-1	getfileinfo	/hbase/stolen_data
Jan 7 2015 11:20 AM	training	10.20.186.130	HDFS-1	mkdirs	/hbase/copy
Jan 7 2015 11:20 AM	training	10.20.186.130	HDFS-1	getfileinfo	/hbase
Jan 7 2015 11:20 AM	training	10.20.186.130	HDFS-1	getfileinfo	/hbase/copy
Jan 7 2015 11:16 AM	training	10.20.186.130	HDFS-1	delete	/user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51
Jan 7 2015 11:16 AM	training	10.20.186.130	HDFS-1	open	/user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51
Jan 7 2015 11:16 AM	training	10.20.186.130	HDFS-1	listStatus	/user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51
Jan 7 2015 11:16 AM	training	127.0.0.1	HIVE-1	QUERY	:

Allowed: true

Object Type: DFS_DIR

Database Name:

Operation Text: SELECT SUM(total_price - total_cost) AS total_profit FROM cart_shipping WHERE total_price >= 50000 AND steps_completed = 4

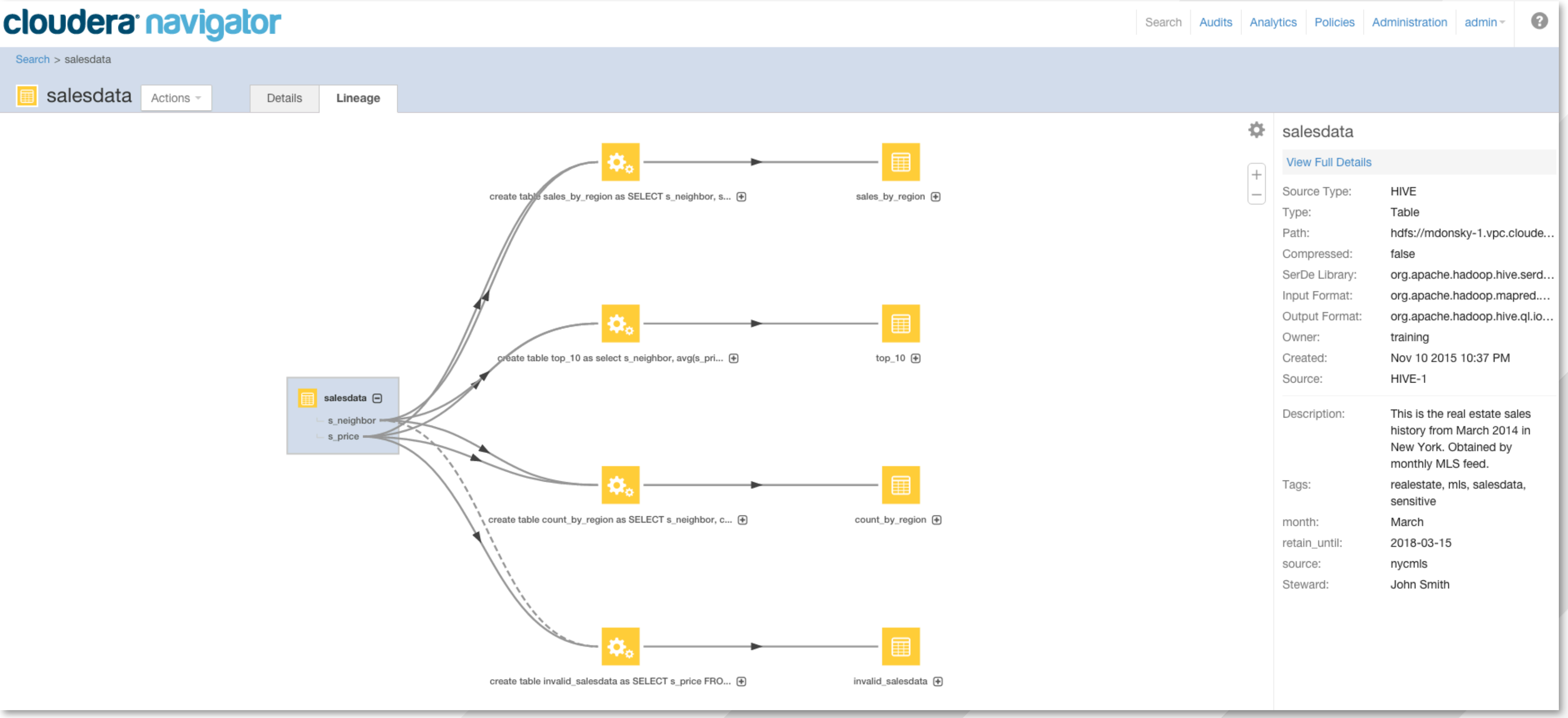
Resource Path: /user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51_306_7520695583396771453-4/-mr-10000

Table Name:

Jan 7 2015 11:16 AM	training	127.0.0.1	HIVE-1	QUERY	default:cart_shipping
Jan 7 2015 11:16 AM	training	10.20.186.130	HDFS-1	delete	/user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51
Jan 7 2015 11:16 AM	training	10.20.186.130	HDFS-1	rename	/user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51
Jan 7 2015 11:16 AM	training	10.20.186.130	HDFS-1	getfileinfo	/user/hive/warehouse/cart_shipping/.hive-staging_hive_2015-01-07_09-15-51

- Who (IP), What (operation), When (timestamp) across the stack
- Any HDFS file access
- Denied logins to Hue, CM, Navigator
- Query text from Impala, Hive, HBase, Solr, Sentry

Lineage

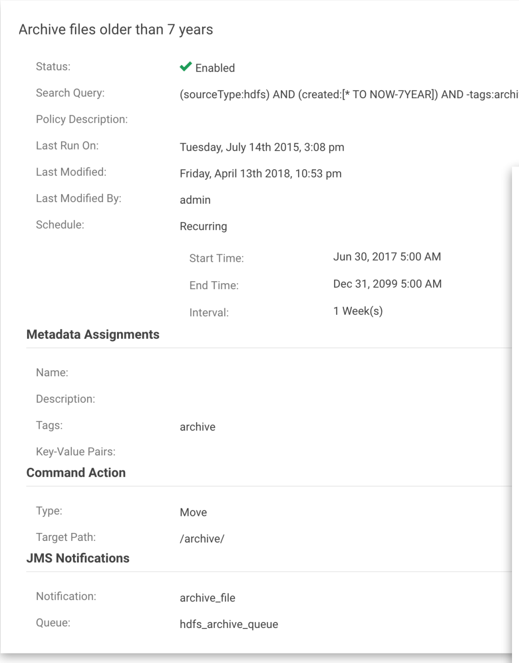

cloudera navigator
Search
Audits
Analytics
Policies
Administration
admin
Search > salesdata
salesdata
Actions
Details
Lineage
create table sales_by_region as SELECT s_neighbor, s...
sales_by_region
create table top_10 as select s_neighbor, avg(s_pri...
top_10
salesdata
s_neighbor
s_price
create table count_by_region as SELECT s_neighbor, c...
count_by_region
create table invalid_salesdata as SELECT s_price FRO...
invalid_salesdata
View Full Details

salesdata

Source Type:	HIVE
Type:	Table
Path:	hdfs://mdonsky-1.vpc.cloude...
Compressed:	false
SerDe Library:	org.apache.hadoop.hive.serd...
Input Format:	org.apache.hadoop.mapred....
Output Format:	org.apache.hadoop.hive.ql.io...
Owner:	training
Created:	Nov 10 2015 10:37 PM
Source:	HIVE-1
Description:	This is the real estate sales history from March 2014 in New York. Obtained by monthly MLS feed.
Tags:	realestate, mls, salesdata, sensitive
month:	March
retain_until:	2018-03-15
source:	nycmls
Steward:	John Smith

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

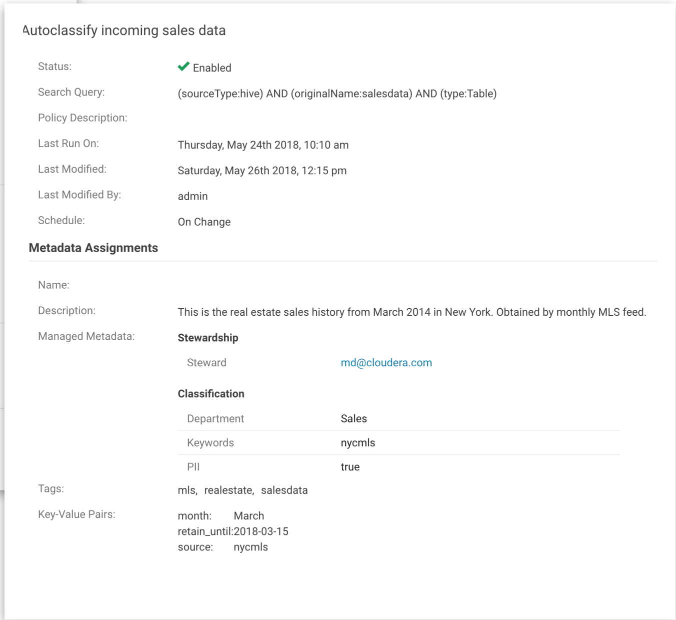
Type: Move

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, 10:10 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

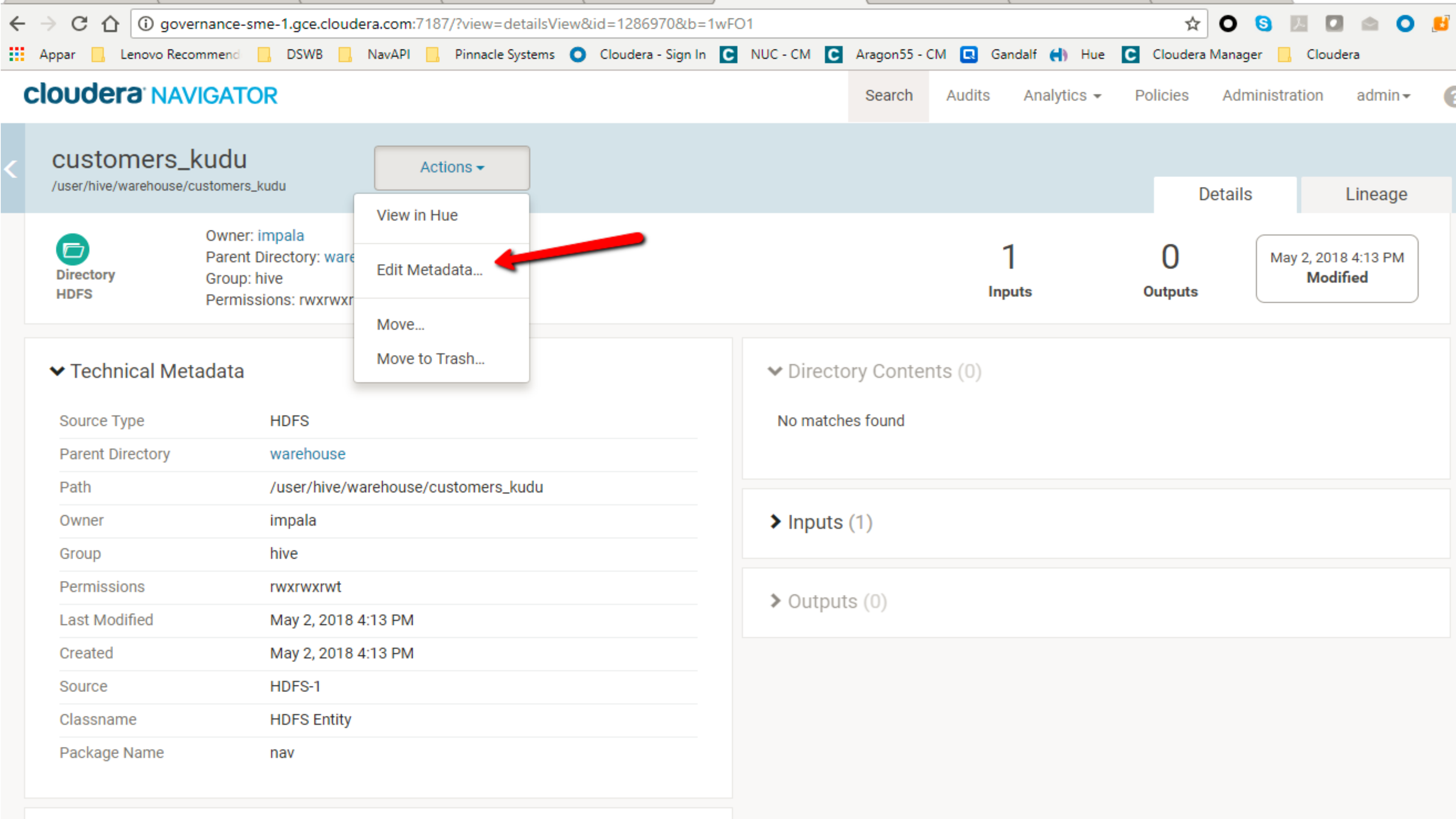
Department Sales

Keywords nycmls

PII true

Tags: mls, realestate, salesdata

Key-Value Pairs:
month: March
retain_until:2018-03-15
source: nycmls



customers_kudu

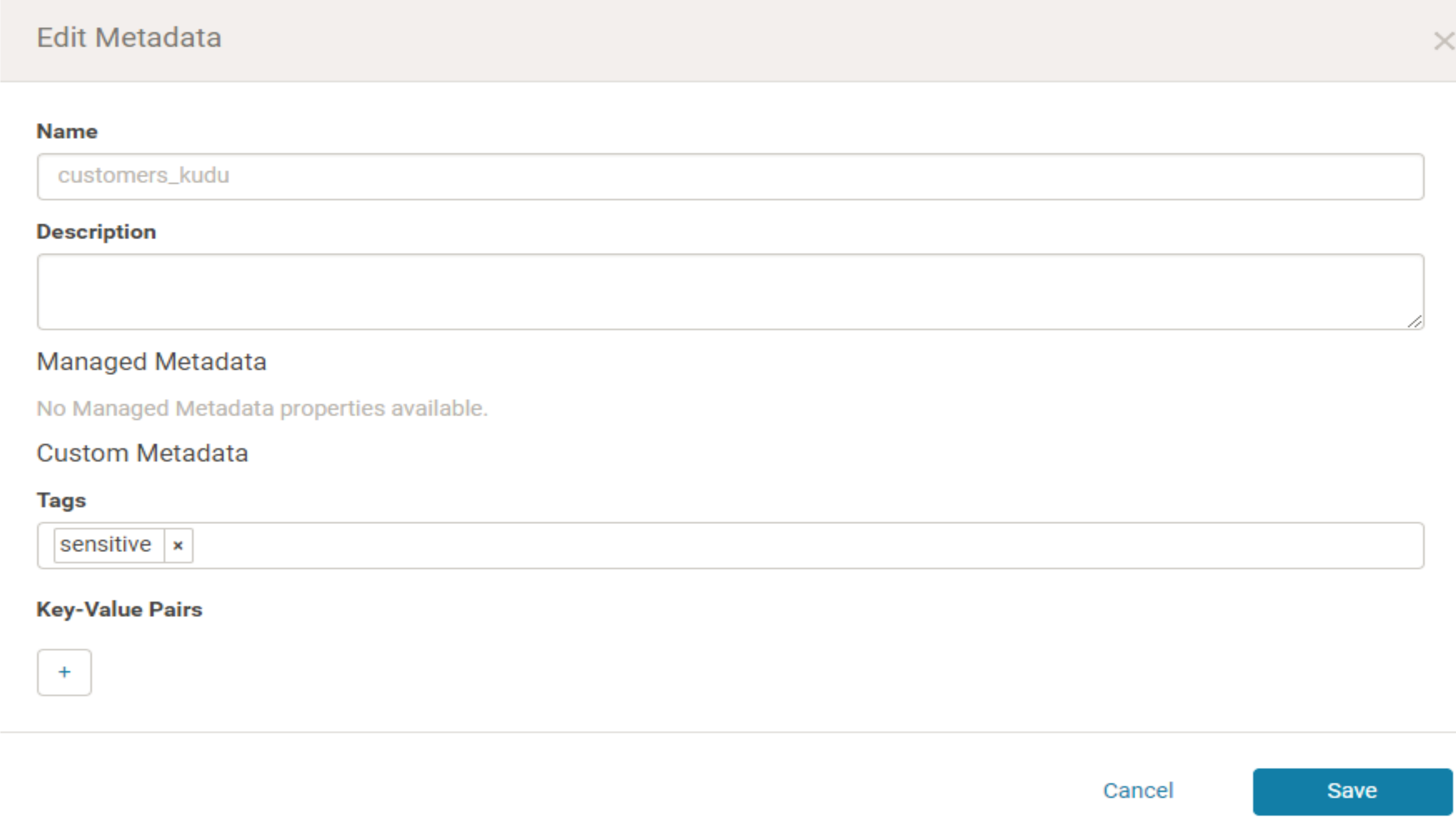
/user/hive/warehouse/customers_kudu

Owner: impala
Parent Directory: ware
Group: hive
Permissions: rwxrwxr

Technical Metadata

Source Type	HDFS
Parent Directory	warehouse
Path	/user/hive/warehouse/customers_kudu
Owner	impala
Group	hive
Permissions	rwxrwxrwt
Last Modified	May 2, 2018 4:13 PM
Created	May 2, 2018 4:13 PM
Source	HDFS-1
Classname	HDFS Entity
Package Name	nav

No matches found



Edit Metadata

Name

customers_kudu

Description

Managed Metadata

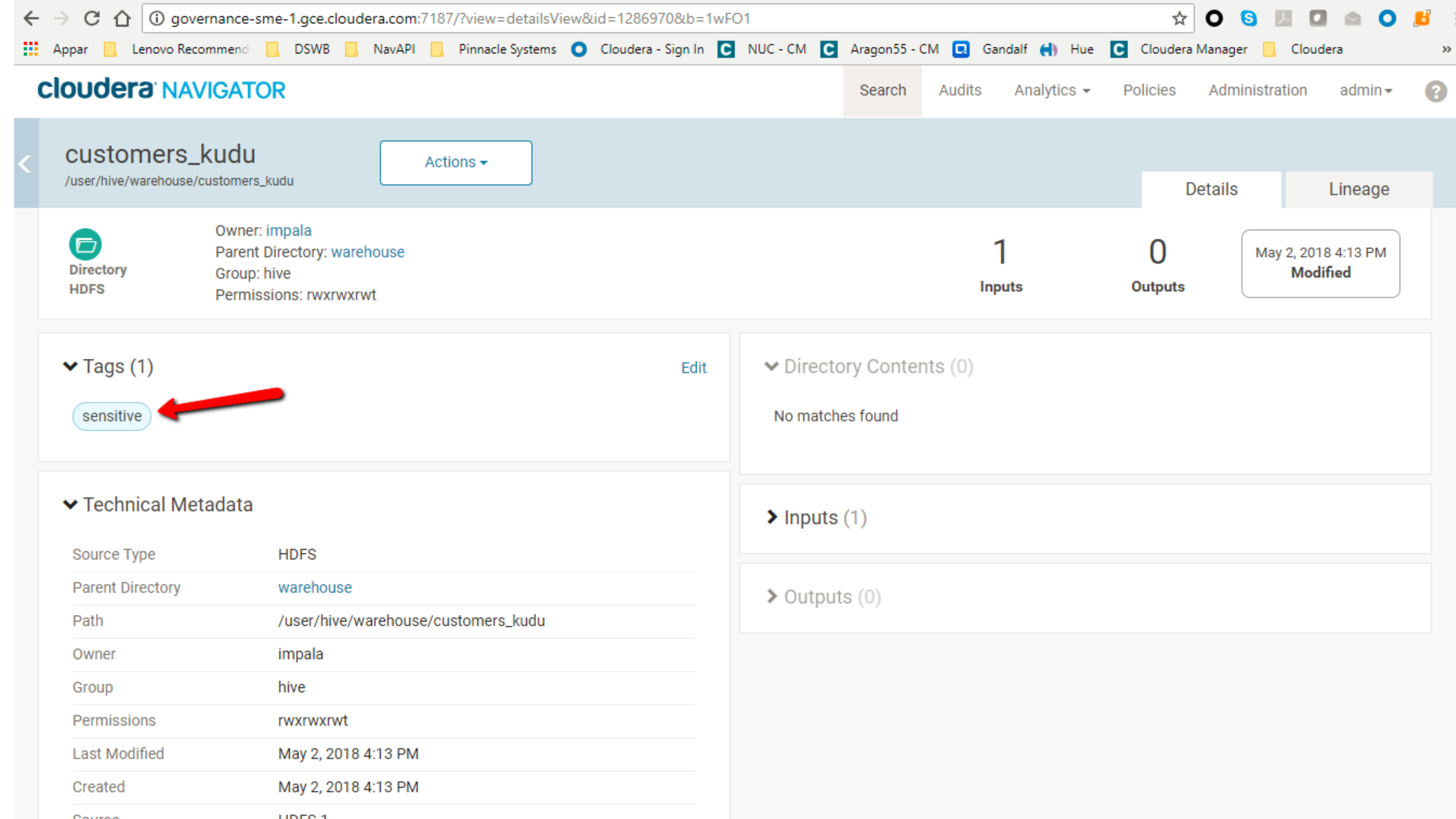
No Managed Metadata properties available.

Custom Metadata

Tags

sensitive

Key-Value Pairs



customers_kudu

/user/hive/warehouse/customers_kudu

Directory
HDFS

Owner: impala
Parent Directory: warehouse
Group: hive
Permissions: rwxrwxrwt

1 Inputs

0 Outputs

May 2, 2018 4:13 PM
Modified

Tags (1)

sensitive

Directory Contents (0)

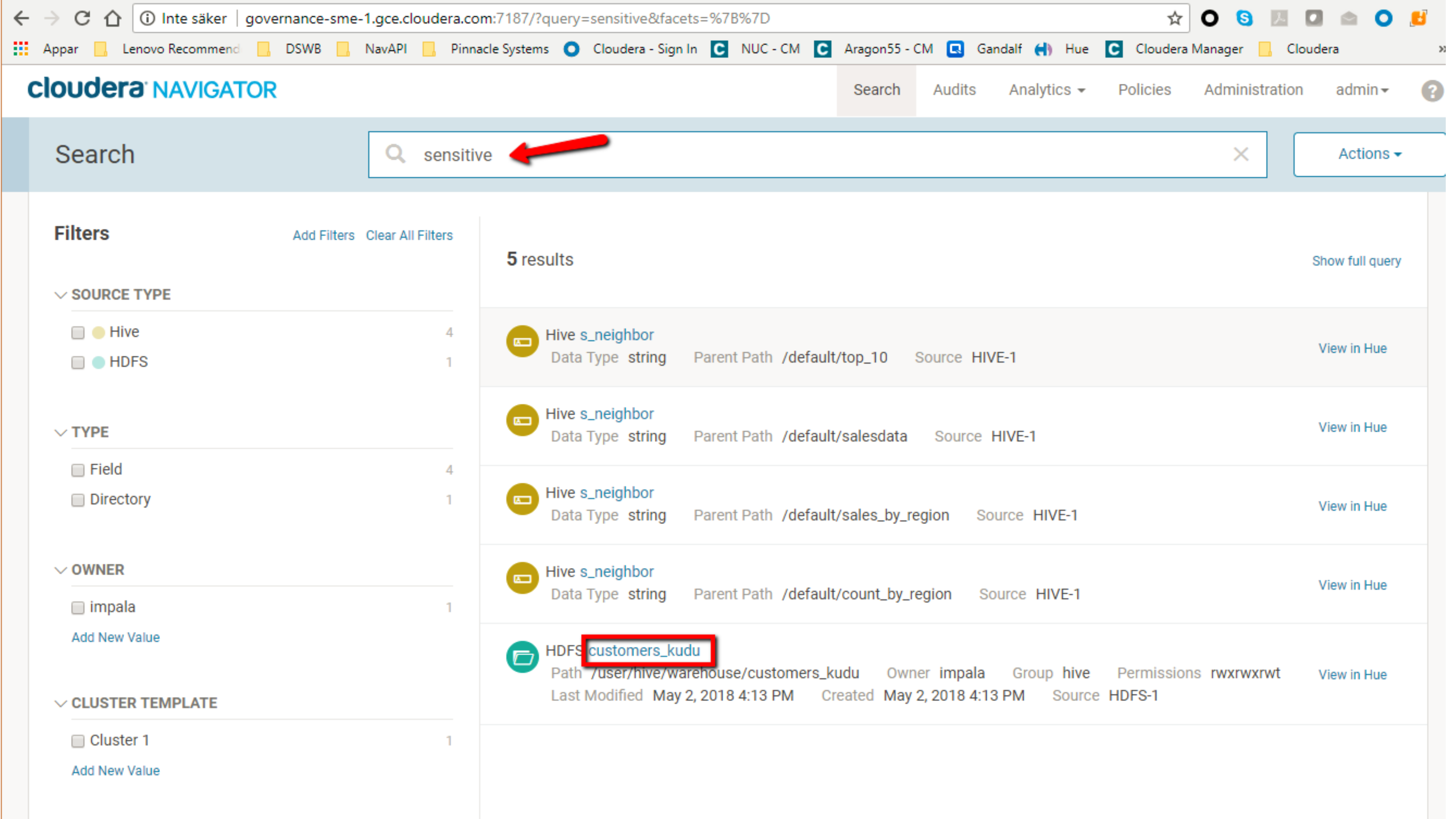
No matches found

Technical Metadata

Source Type	HDFS
Parent Directory	warehouse
Path	/user/hive/warehouse/customers_kudu
Owner	impala
Group	hive
Permissions	rwxrwxrwt
Last Modified	May 2, 2018 4:13 PM
Created	May 2, 2018 4:13 PM
Source	HDFS-1

Inputs (1)

Outputs (0)



Search

sensitive

Filters

SOURCE TYPE

☐ Hive	4
☐ HDFS	1

TYPE

☐ Field	4
☐ Directory	1

OWNER

☐ impala	1

CLUSTER TEMPLATE

☐ Cluster 1	1

5 results

Hive s_neighbor
Data Type string Parent Path /default/top_10 Source HIVE-1
View in Hue

Hive s_neighbor
Data Type string Parent Path /default/salesdata Source HIVE-1
View in Hue

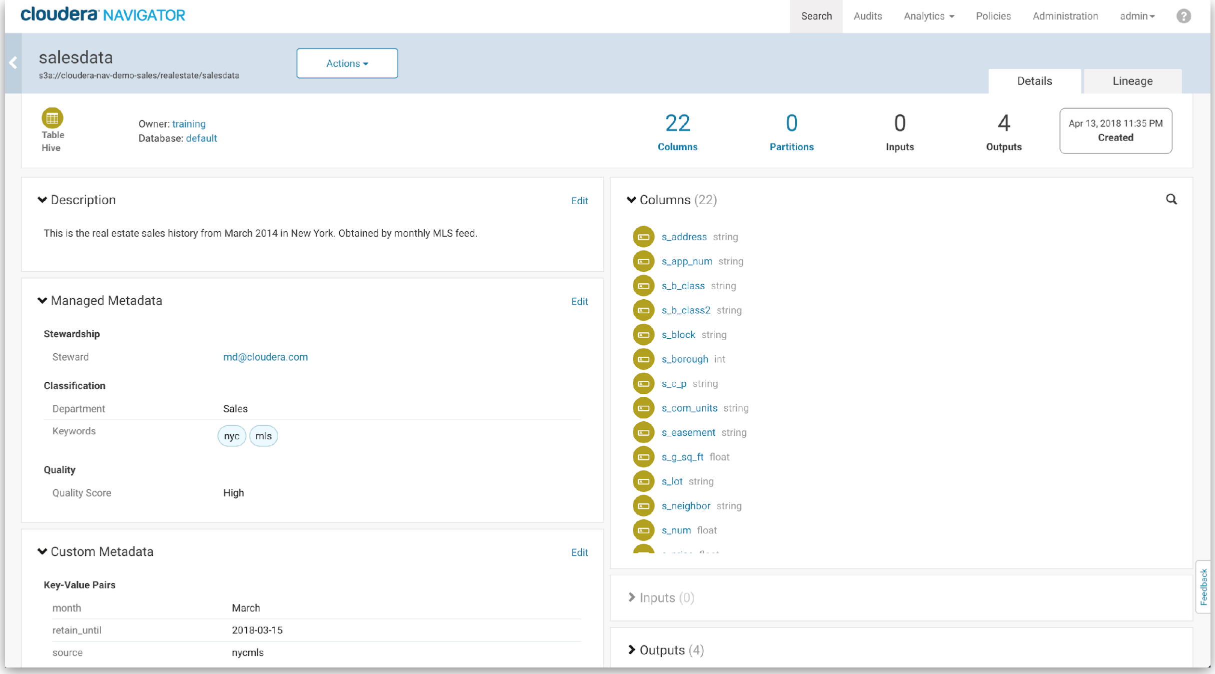
Hive s_neighbor
Data Type string Parent Path /default/sales_by_region Source HIVE-1
View in Hue

Hive s_neighbor
Data Type string Parent Path /default/count_by_region Source HIVE-1
View in Hue

HDFS customers_kudu
Path /user/hive/warehouse/customers_kudu Owner impala Group hive Permissions rwxrwxrwt
Last Modified May 2, 2018 4:13 PM Created May 2, 2018 4:13 PM Source HDFS-1
View in Hue

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
s3a://cloudera-nav-demo-sales/realestate/salesdata

Table
Hive

Owner: training
Database: default

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

Description Edit

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

Managed Metadata Edit

Stewardship

Steward	md@cloudera.com

Classification

Department	Sales
Keywords	nyc mls

Quality

Quality Score	High

Custom Metadata Edit

Key-Value Pairs

month	March
retain_until	2018-03-15
source	nycmls

Columns (22)

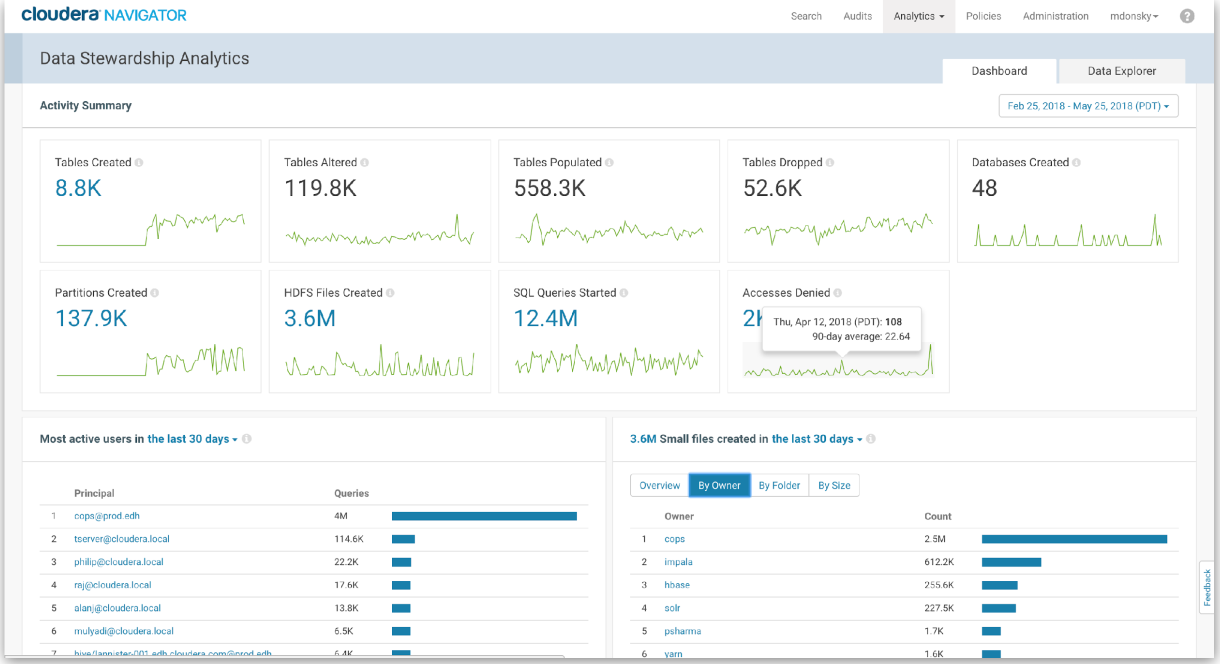
s_address	string
s_app_num	string
s_b_class	string
s_b_class2	string
s_block	string
s_borough	int
s_c_p	string
s_com_units	string
s_easement	string
s_g_sq_ft	float
s_lot	string
s_neighbor	string
s_num	float
s_price	float

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	Tables Altered	Tables Populated	Tables Dropped	Databases Created
8.8K	119.8K	558.3K	52.6K	48

Partitions Created	HDFS Files Created	SQL Queries Started	Accesses Denied
137.9K	3.6M	12.4M	2K

Thu, Apr 12, 2018 (PDT): 108
90-day average: 22.64

Most active users in the last 30 days

	Principal	Queries
1	cops@prod.edh	4M
2	tserver@cloudera.local	114.6K
3	philip@cloudera.local	22.2K
4	raj@cloudera.local	17.6K
5	alanj@cloudera.local	13.8K
6	mulyadi@cloudera.local	6.5K
7	hive/lannister-001.edh.cloudera.com@prod.edh	6.4K

3.6M Small files created in the last 30 days

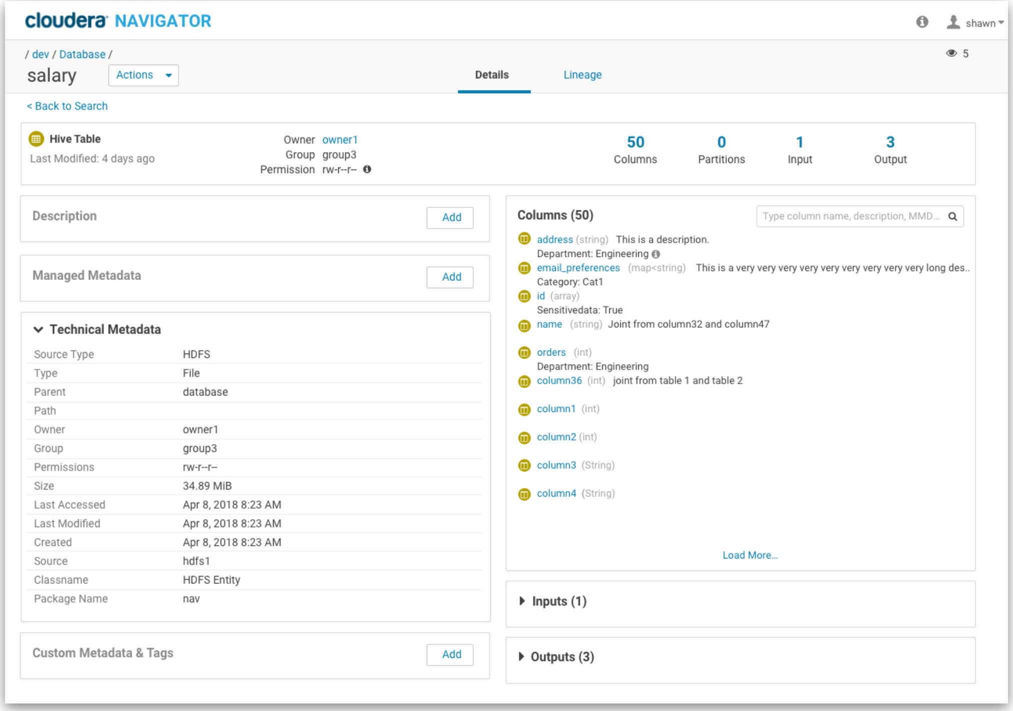
Overview | By Owner | By Folder | By Size

	Owner	Count
1	cops	2.5M
2	impala	612.2K
3	hbase	255.6K
4	solr	227.5K
5	psharma	1.7K
6	yarn	1.6K

USE CASES I SELF-SERVICE DISCOVERY

Find, trust, and use data sets

- **Find** data sets based on business context
- **Trust** data sets with insight based on ironclad governance
- **Use** data sets with analytics on on historical usage patterns



salary

Hive Table

Last Modified: 4 days ago

Owner owner1
Group group3
Permission rw-r--r–

50	0	1	3
Columns	Partitions	Input	Output

Description Add

Managed Metadata Add

Technical Metadata

Source Type	HDFS
Type	File
Parent	database
Path	
Owner	owner1
Group	group3
Permissions	rw-r--r–
Size	34.89 MiB
Last Accessed	Apr 8, 2018 8:23 AM
Last Modified	Apr 8, 2018 8:23 AM
Created	Apr 8, 2018 8:23 AM
Source	hdfs1
Classname	HDFS Entity
Package Name	nav

Custom Metadata & Tags Add

Columns (50)

Type column name, description, MMD...

address (string) This is a description.
Department: Engineering
email_preferences (map<string) This is a very very very very very very very very very long des..
Category: Cat1
id (array)
Sensitivedata: True
name (string) Joint from column32 and column47
orders (int)
Department: Engineering
column36 (int) joint from table 1 and table 2
column1 (int)
column2 (int)
column3 (String)
column4 (String)

Load More...

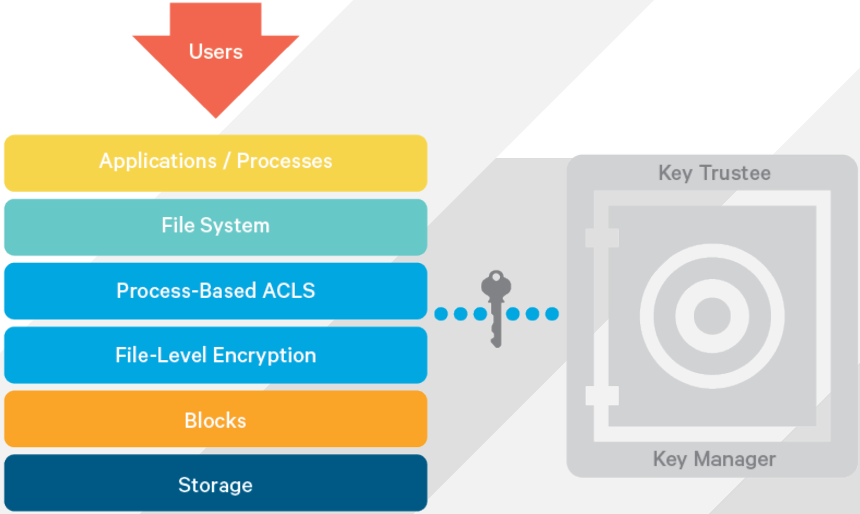
Inputs (1)

Outputs (3)

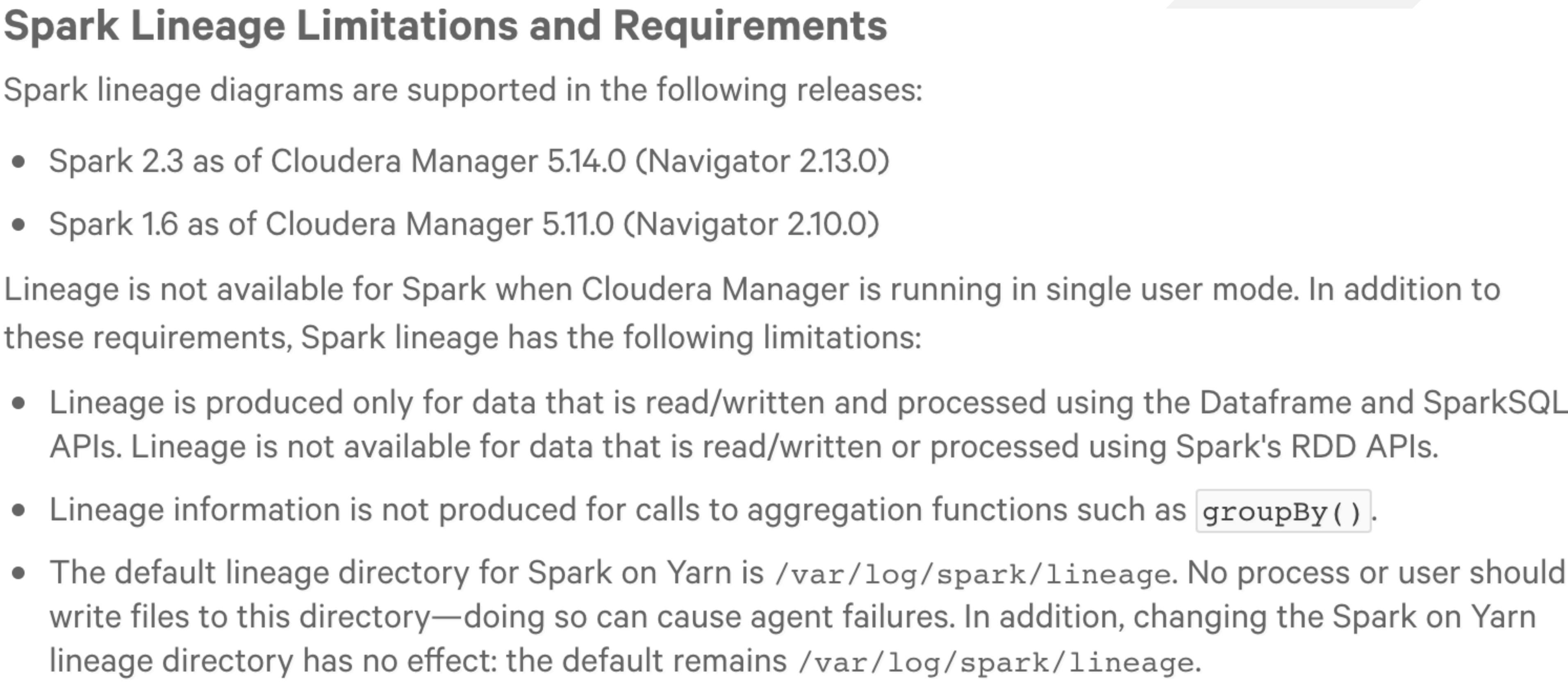
Navigator Encrypt

Transparent layer between application and file system

- Compliance-Ready
- Massively Scalable
- High Performance: Optimized for Intel
- Separation of Duties
- Key Management with Navigator Key Trustee



Spark limitations



Spark Lineage Limitations and Requirements

Spark lineage diagrams are supported in the following releases:

- Spark 2.3 as of Cloudera Manager 5.14.0 (Navigator 2.13.0)
- Spark 1.6 as of Cloudera Manager 5.11.0 (Navigator 2.10.0)

Lineage is not available for Spark when Cloudera Manager is running in single user mode. In addition to these requirements, Spark lineage has the following limitations:

- Lineage is produced only for data that is read/written and processed using the Dataframe and SparkSQL APIs. Lineage is not available for data that is read/written or processed using Spark's RDD APIs.
- Lineage information is not produced for calls to aggregation functions such as `groupBy()`.
- The default lineage directory for Spark on Yarn is `/var/log/spark/lineage`. No process or user should write files to this directory—doing so can cause agent failures. In addition, changing the Spark on Yarn lineage directory has no effect: the default remains `/var/log/spark/lineage`.

Latest releases

https://www.cloudera.com/documentation/enterprise/release-notes/topics/cn_rn_new_features.html

Roadmap

THANK YOU

CLOUDERA

http://sdx-nav-demo-1.gce.cloudera.com:7187/login.html