Enabling Secure Hadoop Environments

Fred Koopmans
Sr. Director of Product Management

Data Drives Government:
Key Data Trends in Public Sector

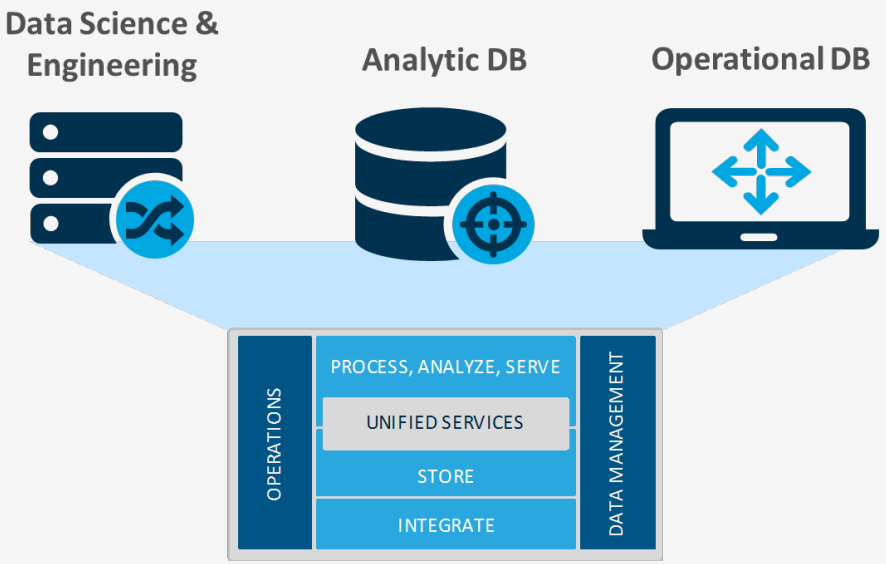
Mike Olson | Co-founder & Chief Strategy Officer

The future of government is data management

What's your strategy?

cloudera

Cloudera's Enterprise Data Hub makes it possible



Bring all your data together

Bring all your knowledge workers together

Bring all your data applications together

Run anywhere

cloudera

But, an EDH can also make a juicy target

All data in one place?

Provide everyone access to one platform?

How do I ensure security?

How do I maintain security as the platform grows?

cloudera

4 focus areas for securing your Hadoop environment

Perimeter

Guarding access to the cluster itself

Technical Concepts:
Authentication
Network isolation

Access

Defining what users and applications can do with data

Technical Concepts:
Permissions
Authorization

Visibility

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

Technical Concepts:
Auditing
Lineage

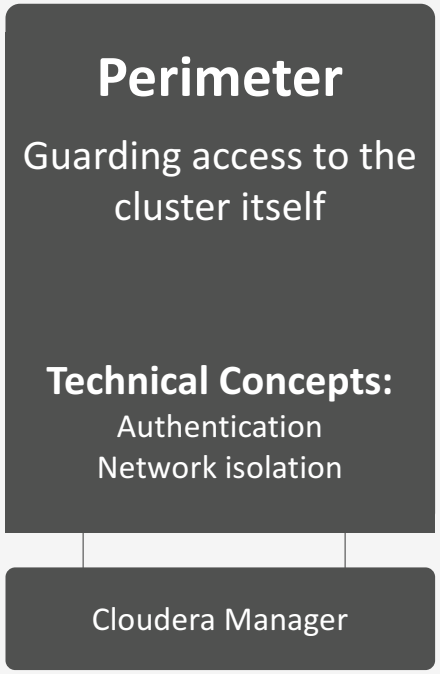
Data

Protecting data in the cluster from unauthorized visibility

Technical Concepts:
Encryption, Key management,
Data masking

cloudera

Perimeter Security Requirements



Preserve user choice of Hadoop service

Conform to centrally managed authentication policies

Implement with existing standard systems

cloudera

Authentication

CDH components

- Kerberos authentication
- Automation provided by Cloudera Manager to leverage Active Directory

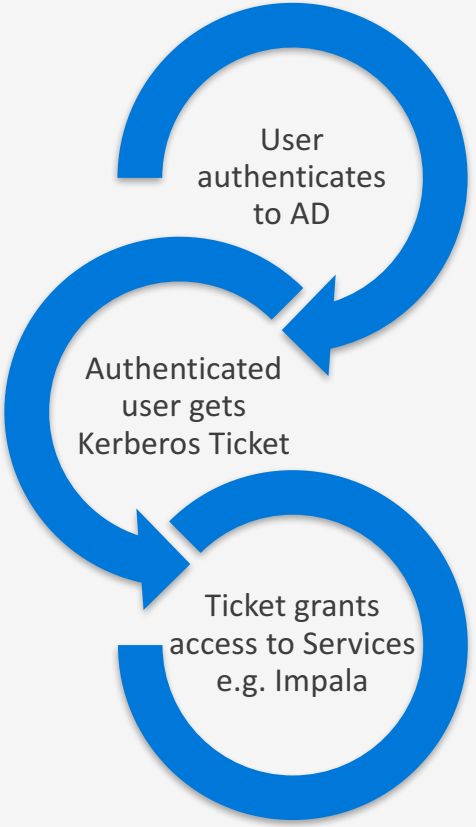
Web UIs

- LDAP and SAML authentication

SQL Access

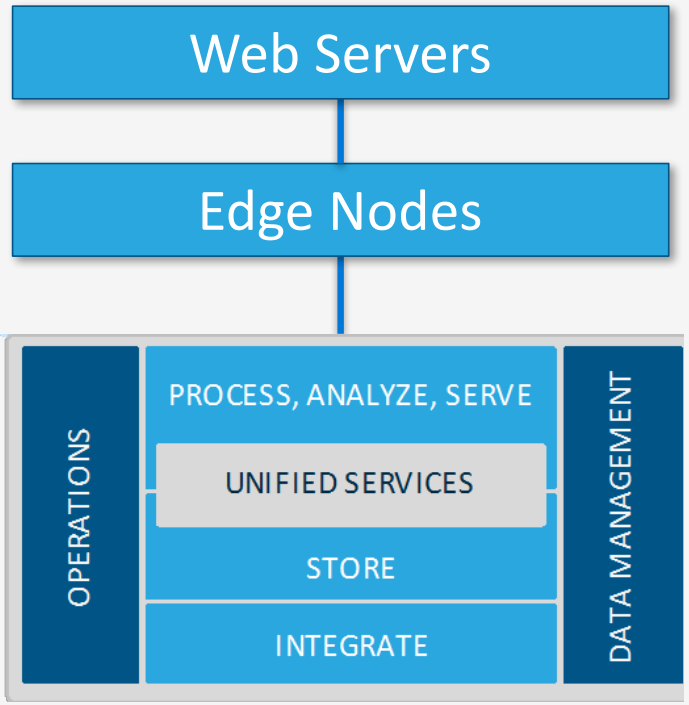
- LDAP and Kerberos authentication

User [ssmith]
Password[*****]



cloudera

Network Isolation



Most users only permitted access to gateway services running on cluster periphery

Only admins permitted access to full cluster

cloudera

Access Security Requirements

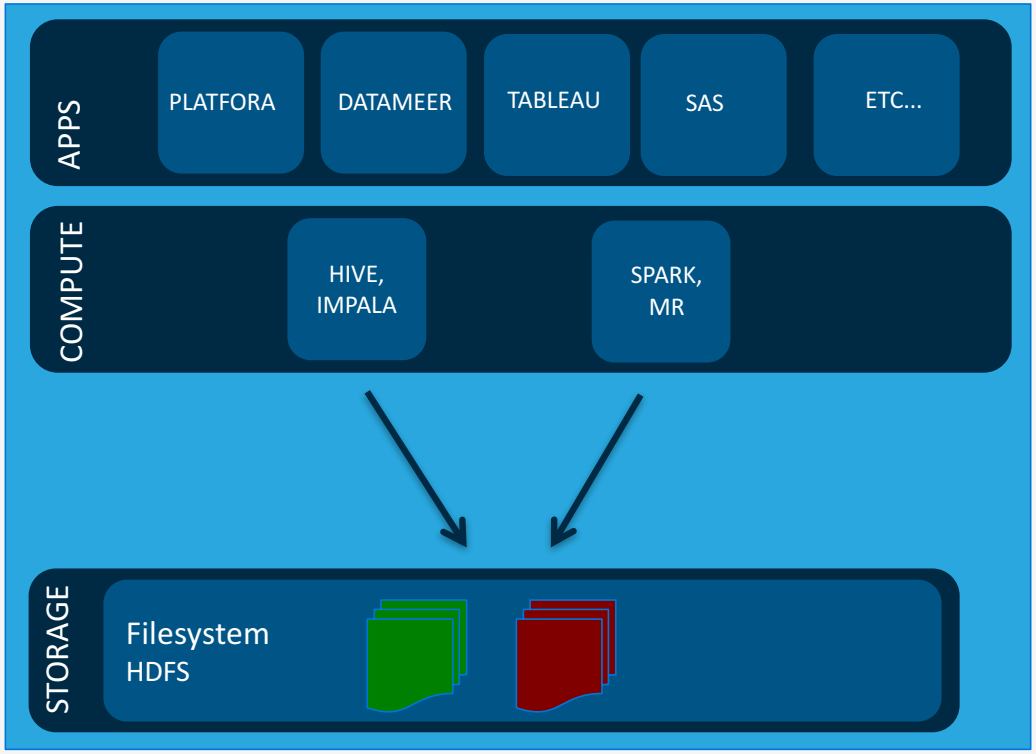
Keep only one logical copy of data

Create only one permissions rule for all applications and all compute frameworks

Enforce permissions at column and row level granularity

cloudera

Early days of Hadoop: Storage permissions only



- Simple "All or Nothing" permissions for each file/table

But...

- Tables often contain 10s – 100s of columns
- Not all users are allowed to see all columns and rows

cloudera

Use cases for fine-grained access control

Columns

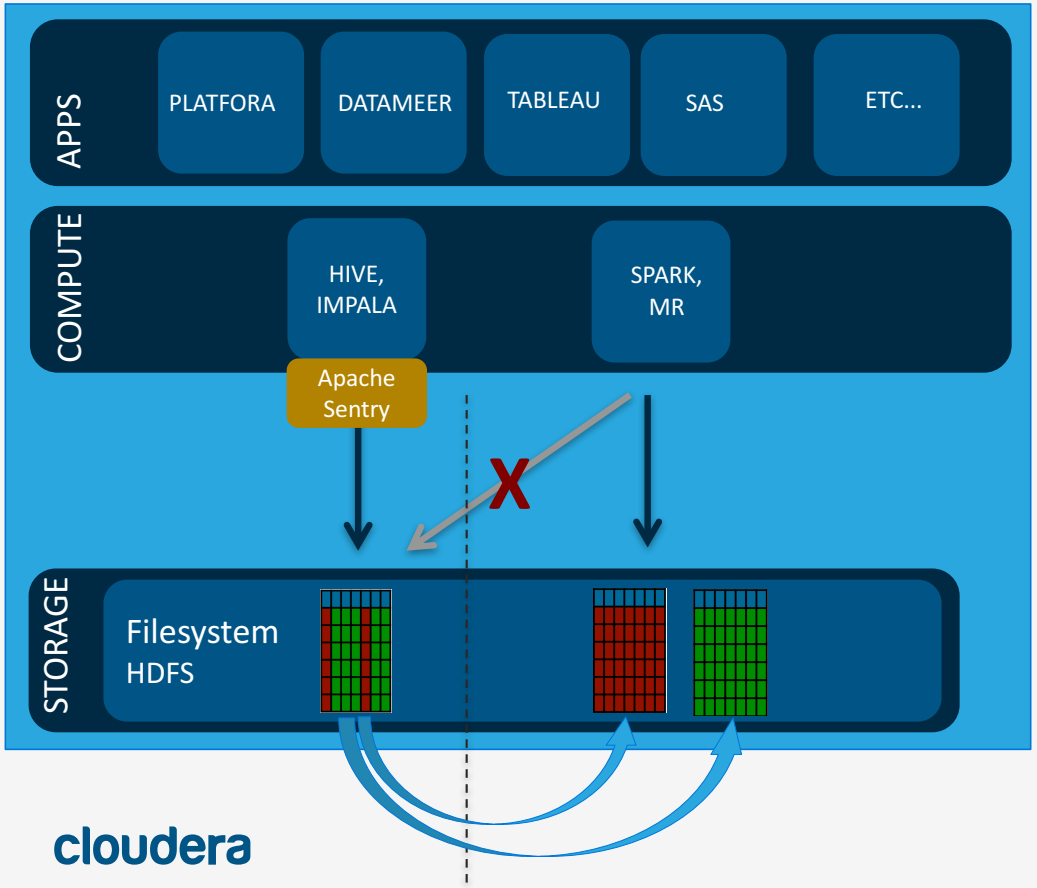
- Different user groups need access to different columns (ex: social security numbers)

Rows

- Different user groups need access to different records (ex: by security clearance level)

cloudera

Few years ago: Storage permissions + SQL Auth.



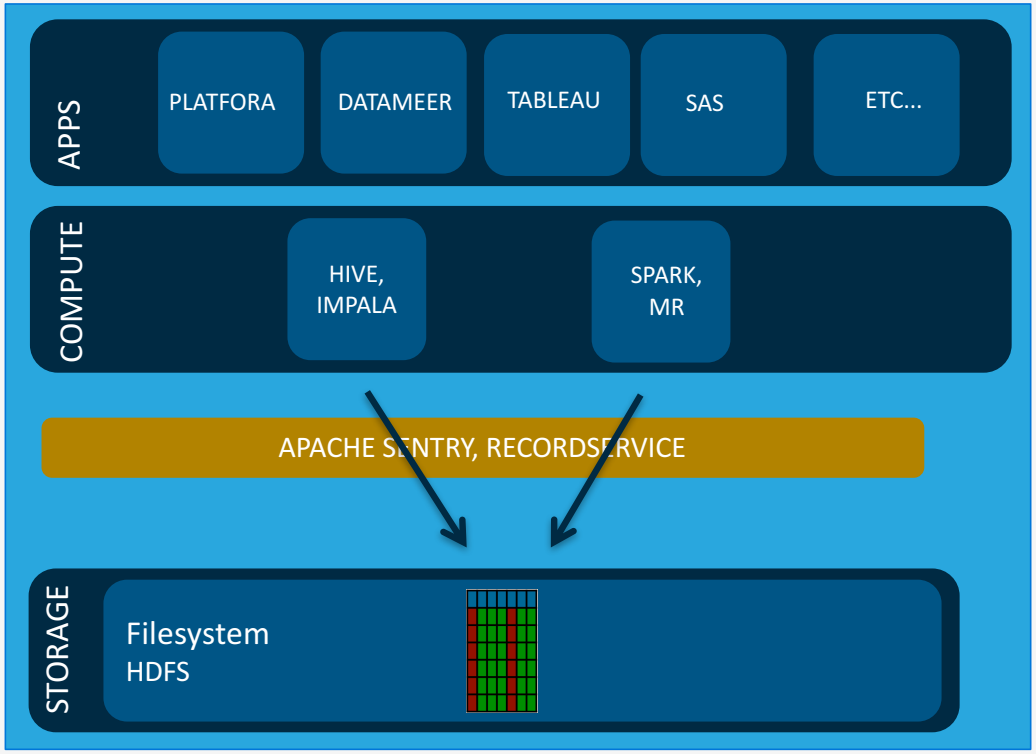
- Adds column and row-level permissions

But...

- Creates duplicate data, duplicate permissions rules to support Spark and MR

cloudera

Up Next: Apache Sentry + RecordService* working together



- Column and Row-level Permissions
- One copy of data
- One set of permissions

*in beta

cloudera

Fine-grained access control without Sentry & RecordService*

- Split the HDFS permissions original file
- Use to limit access

Date/time	Accnt #	SSN	Asset	Trade	Country
09:33:11 16-Feb-2015	0234837823	238-23-9876	AZP	Sell	US
11:33:01 16-Feb-2015	3947848494	329-44-9847	TBT	Buy	EU
14:12:34 16-Feb-2015	4848367383	123-56-2345	IDI	Sell	UK
09:22:03 16-Feb-2015	3485739384	585-11-2345	ICBD	Buy	US
11:55:33 16-Feb-2015	3847598390	234-11-8765	FWQ	Buy	US
10:22:55 16-Feb-2015	8765432176	344-22-9876	UAD	Buy	UK
13:45:24 16-Feb-2015	3456789012	412-22-8765	NZMA	Sell	EU
09:03:44 16-Feb-2015	4857389329	123-44-5678	TMV	Buy	US
15:55:55 16-Feb-2015	4756983234	234-76-9274	DRW	Buy	UK

Date/time	Accnt #	SSN	Asset	Trade	Country
09:33:11 16-Feb-2015	0234837823	238-23-9876	AZP	Sell	US
09:22:03 16-Feb-2015	3485739384	585-11-2345	ICBD	Buy	US
11:55:33 16-Feb-2015	3847598390	234-11-8765	FWQ	Buy	US
09:03:44 16-Feb-2015	4857389329	123-44-5678	TMV	Buy	US

Date/time	Accnt #	SSN	Asset	Trade	Country
11:33:01 16-Feb-2015	3947848494	329-44-9847	TBT	Buy	EU
13:45:24 16-Feb-2015	3456789012	412-22-8765	NZMA	Sell	EU

Date/time	Accnt #	SSN	Asset	Trade	Country
14:12:34 16-Feb-2015	4848367383	123-56-2345	IDI	Sell	UK
10:22:55 16-Feb-2015	8765432176	344-22-9876	UAD	Buy	UK
15:55:55 16-Feb-2015	4756983234	234-76-9274	DRW	Buy	UK

cloudera

Fine-grained access control with Sentry & RecordService*

- Sentry: Define permissions at the table, column and row levels
- Sentry + RecordService: Enforce these across all access paths

Single HDFS file:

Column-Level Controls

Date/time	Accnt #	SSN	Asset	Trade	Country
09:33:11 16-Feb-2015	0234837823	238-23-9876	AZP	Sell	US
11:33:01 16-Feb-2015	3947848494	329-44-9847	TBT	Buy	EU
14:12:34 16-Feb-2015	4848367383	123-56-2345	IDI	Sell	EU
09:22:03 16-Feb-2015	3485739384	585-11-2345	ICBD	Buy	US
11:55:33 16-Feb-2015	3847598390	234-11-8765	FWQ	Buy	US
10:22:55 16-Feb-2015	8765432176	344-22-9876	UAD	Buy	EU
13:45:24 16-Feb-2015	3456789012	412-22-8765	NZMA	Sell	EU

Row-Level Controls

Hive, Impala, MR, Spark, Pig

What U.S. Brokers See

Column-Level Controls

Date/time	Accnt #	SSN	Asset	Trade	Country
09:33:11 16-Feb-2015	0234837823	XXX-XX-	AZP	Sell	US
09:22:03 16-Feb-2015	3485739384	XXX-XX-	ICBD	Buy	US
11:55:33 16-Feb-2015	3847598390	XXX-XX-	FWQ	Buy	US

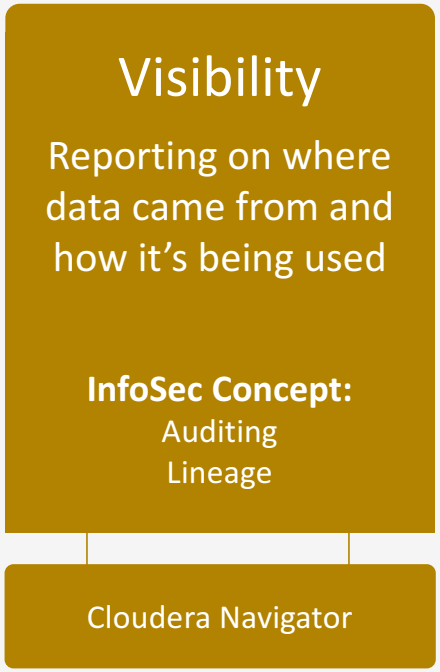
Row-Level Controls

cloudera

Visibility Security Requirements

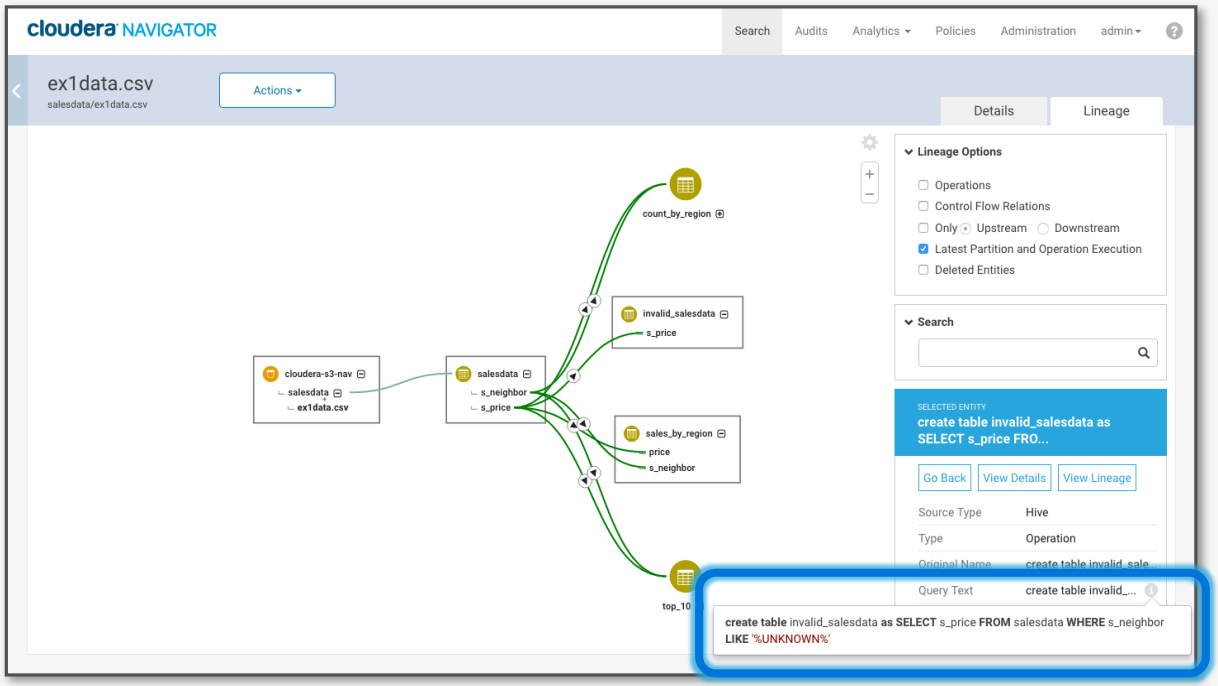
Comply with policies for audit, data classification, and lineage

Centralize the audit repository



cloudera

Audit & Lineage



Trusted for production

- 100s of customer deployments of Cloudera Navigator over last 3+ years

Compliance-ready

- Only Hadoop distribution to pass PCI audit

Detailed

- Column and row level access trail

Plays nicely with others

- Integrated with the leading partner solutions

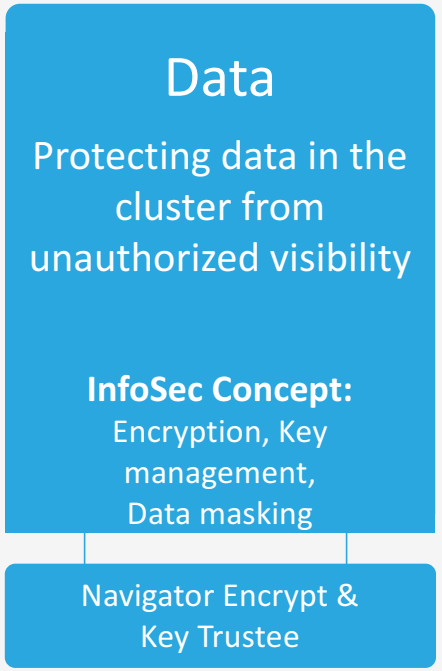
cloudera

Data Security Requirements

Perform analytics on regulated data

Encrypt data, conform to key management policies, protect from root

Integrate with existing HSM as part of key management infrastructure



cloudera

Comprehensive Data Security

Encryption

ALL data on the wire

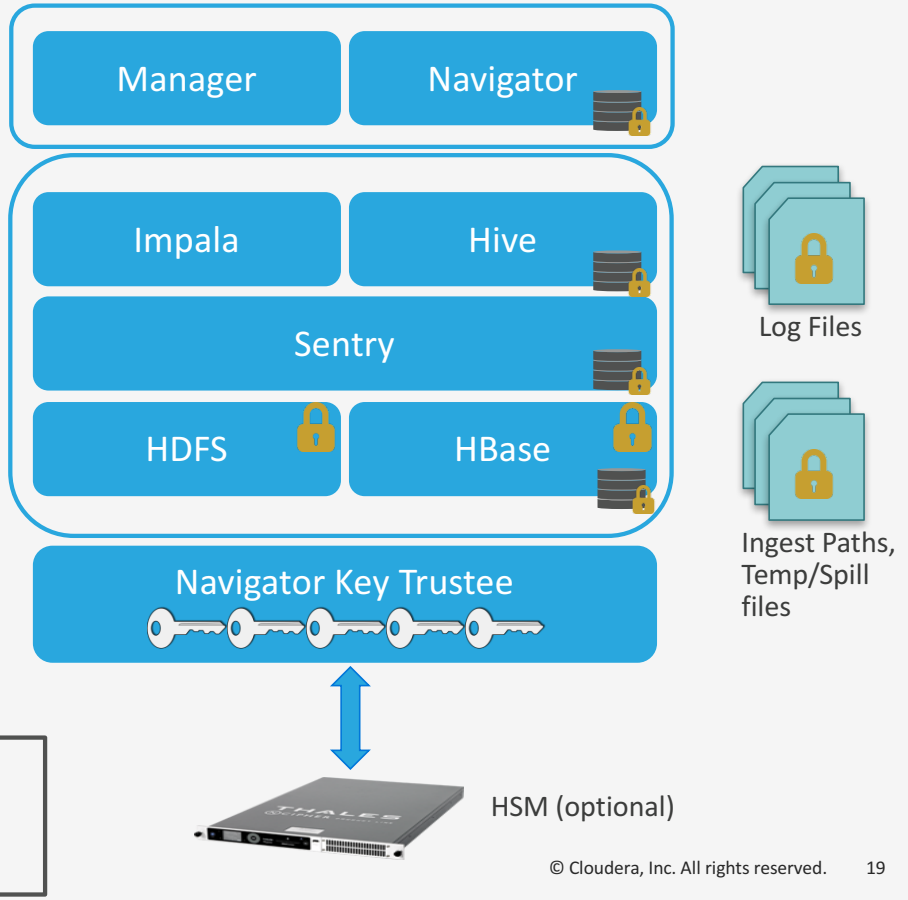
- ALL data at rest: HDFS, HBase, metadata databases, temp files, ingest paths

Key Management

- Automated key replication & backup
- HSM backed key protection

Data Masking

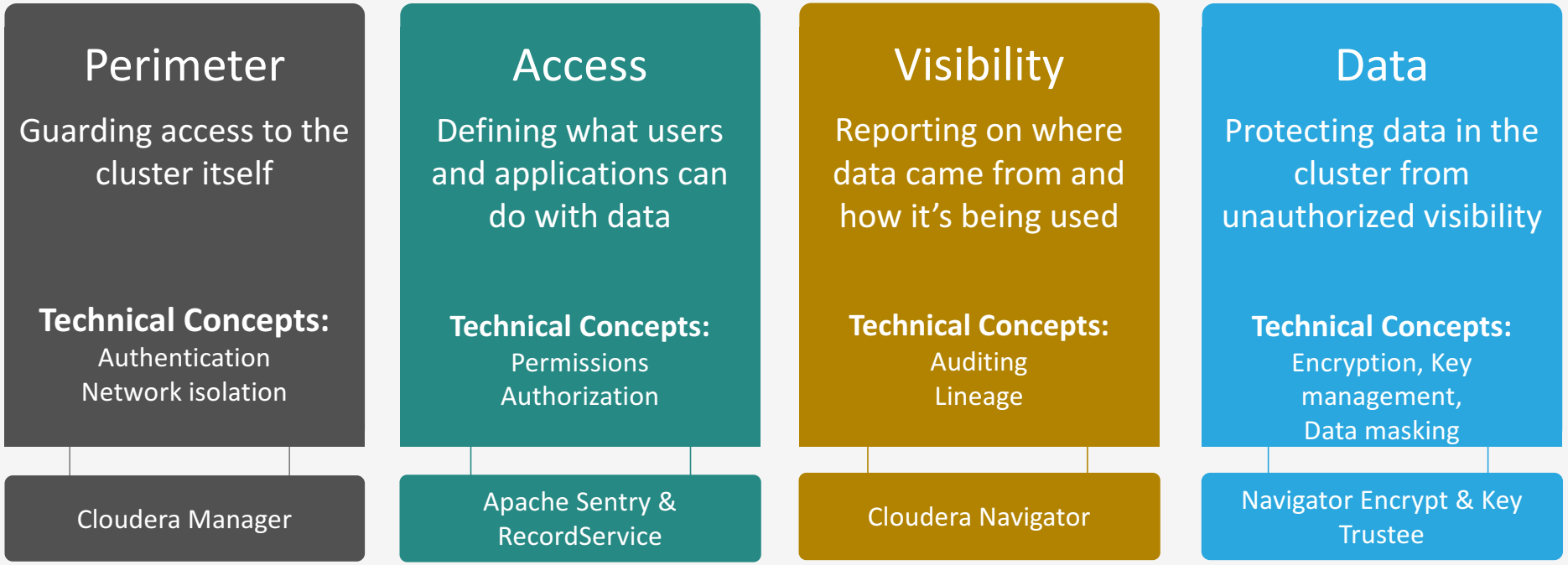
- Sensitive data in logs
- Passwords in config files



Legend

cloudera

Cloudera's comprehensive, compliance-ready security solution



cloudera

Beyond traditional security controls

Automated discovery and tagging of sensitive data

- Automatically scan for protected attribute types
- Automatically apply authorization and encryption policy

“Follow the data” authorization and protection policies

- Leverage lineage data tagging enforce authorization and encryption policy
- Eliminate manual configuration of security for each new table and column

Admins focused on exception handling due to insufficient access

cloudera

cloudera

Thank You

Fred Koopmans

