Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Getting Ready for GDPR

Securing and governing hybrid cloud and on-prem big data clusters

Your Speakers

- **André Araújo,** Senior Solutions Architect, Cloudera
- **Syed Rafice,** Principal Systems Engineer, Cloudera
- **Mubashir Kazia,** Principal Solutions Architect, Cloudera
- **Mark Donsky,** Director of Products, Cloudera

Format

- Five sections
- Each section:
  - Introduce a security concept
  - Review its relevance with GDPR
  - How to enable
  - Demos
- **Please hold questions until the end of each section**
- Short break in the middle

- Slides are available from http://strataconf.com

Agenda

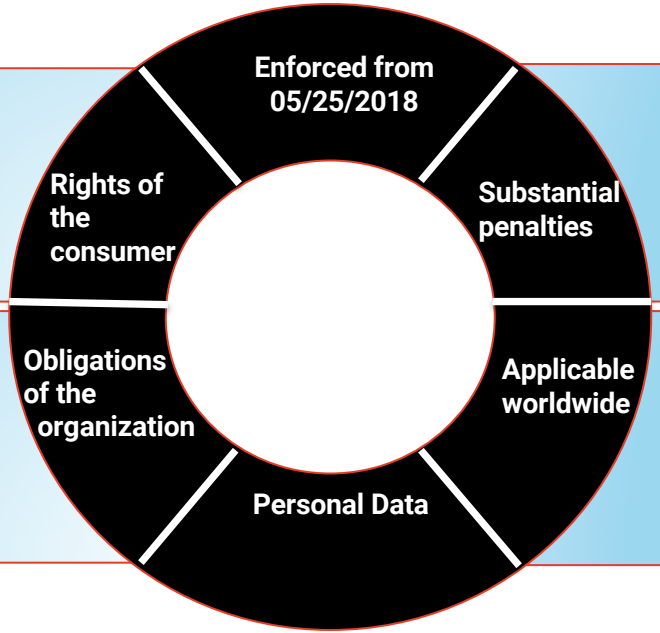
- Prelude: GDPR Overview – Syed
- Authentication – André
- Authorization – André
- Wire Encryption – Syed
- Encryption-at-rest – Mubashir
- Data Governance – Mark
- Final Thoughts – Mark

Prelude: GDPR Overview

General Data Protection Regulation (GDPR)

- Right to be forgotten/erasure
- Right to access information
- Right to data portability
- Right for processing to be restricted

- Heavy fines for violations
- Up to 20M Euros or 4% of the annual global turnover for the preceding financial year



- Across people, process and technology
- Impacts how personal data is collected and used

- Any organization with any users in the EU needs to be compliant.
- Includes companies based outside the EU, processing personal data about EU individuals in connection with the offering any goods or services or monitoring user behavior.
- Includes data processor and data controller

Governance and Compliance Pillars

Identity

Validate users by membership in enterprise directory

Technical Concepts:
Authentication
User/group mapping

Access

Defining what users and applications can do with data

Technical Concepts:
Permissions
Authorization

Visibility

Discovering, curating and reporting on how data is used

Technical Concepts:
Auditing
Lineage
Metadata catalog

Data Protection

Shielding data in the cluster from unauthorized visibility

Technical Concepts:
Encryption at rest & in motion

Don’t Put Your Hadoop Cluster on the Open Internet

- NODATA4U
  - Data wiped out from unsecured Hadoop and CouchDB
- MongoDB ransomware
  - Tens of thousands of unsecured MongoDB instances on the internet
  - The attack: All data deleted or encrypted; ransom note left behind
- NHS ramsonware

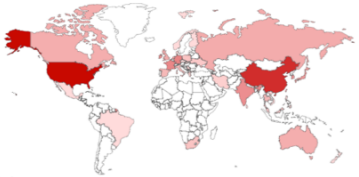
SHODAN | cloudera | Explore | Enterprise Access | Contact Us

Exploits | Maps

TOTAL RESULTS

1,649

TOP COUNTRIES



United States	812
China	438
Germany	50
France	45
India	36

TOP SERVICES

50070	1,039
8086	428
HTTPS	51
Splunk	47
HTTP	20

TOP ORGANIZATIONS

Amazon.com	431
Digital Ocean	123
Hangzhou Alibaba Advertising Co.,Ltd.	111
Microsoft Azure	107
Google Cloud	78

Hadoop Administration

13.56.76.231
ec2-13-56-76-231.us-west-1.compute.amazonaws.com
Amazon.com
Added on 2017-09-25 11:37:34 GMT
United States, San Jose
Details

cloud

3.0 GB	514 Files

Total Blocks	442
Number of Threads	93

```
HTTP/1.1 200 OK
Cache-Control: no-cache
Expires: Mon, 25 Sep 2017 11:33:09 GMT
Date: Mon, 25 Sep 2017 11:33:09 GMT
Pragma: no-cache
Expires: Mon, 25 Sep 2017 11:33:09 GMT
Date: Mon, 25 Sep 2017 11:33:09 GMT
Pragma: no-cache
Content-Type: text/html; charset=utf-8
Expires: Mon, 25 Sep 2017...
```

52.66.40.178

ec2-52-66-40-178.ap-south-1.compute.amazonaws.com
Amazon.com
Added on 2017-09-25 11:26:18 GMT
India, Mumbai
Details

```
HTTP/1.1 200 OK
Date: Mon, 25 Sep 2017 11:26:18 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 63
Connection: keep-alive
Expires: Thu, 01-Jan-1970 00:00:00 GMT
Set-Cookie: CLOUDERA_MANAGER_SESSIONID=e36c82gb2qxi1ru3zqeospyfu;Path=/;HttpOnly
Last-Modified: Fri, 18 Aug 2017 15:22...
```

104.197.227.61

61.227.197.104.bc.googleusercontent.com
Google Cloud
Added on 2017-09-25 11:24:44 GMT
United States, Mountain View
Details

```
HTTP/1.1 200 OK
Content-Type: text/html; charset=iso-8859-1
Transfer-Encoding: chunked
Server: Jetty(6.1.26.cloudera.4)

6000
<html><head><title>Firehose_SERVICE_MONITORING</title></head><body>
```

Basic Networking Checks

- Engage your network admins to plan the network security
- Make sure your IP address isn't an internet-exposed address
  - These are the private IP address ranges:
  - 10.* (10.0/8)
  - 172.16.* - 172.31.* (172.16/12)
  - 192.168.* (192.168/16)
- Use `nmap` from outside your corporate environment
- If in {AWS, Azure, GCE}, check networking configuration

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Questions?

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Authentication

André Araújo

Senior Solutions Architect
Cloudera

Authentication - GDPR

- Broadly underpins most of the GDPR Article 5 Principles

- **Lawfulness, fairness and transparency**
- **Purpose limitation**
- **Data minimization**
- **Accuracy**
- **Storage limitation**
- **Integrity and confidentiality**
- **Accountability**

Authentication - Agenda

- Intro - identity and authentication
- Kerberos and LDAP authentication
- Enabling kerberos and LDAP using Cloudera Manager
- **DEMO:** Actual strong authentication in Hadoop
- Questions

Identity

- Before we can talk about authentication, we must understand **identity**
- An object that uniquely identifies a user (usually)
  - Email account, Windows account, passport, driver's license

- In Hadoop, identity largely means **username**
- Using a common source of identity is paramount

Identity Sources

- Individual Linux servers use /etc/passwd and /etc/group
  - Not scalable and prone to **errors**
- LDAP is the preferred way
  - Integrate at the Linux OS level
    - RedHat SSSD
    - Centrify
  - **All** applications running on the OS can use the same LDAP integration
  - Most enterprises use Active Directory
  - Some enterprises use a Linux-specific LDAP implementation

Identity and Authentication

- So you have an identity database, now what?
- Users and applications must **prove** their identities to each other
- This process is authentication
- Hadoop strong authentication is built around **Kerberos**
- Kerberos is built into Active Directory and this is the most common Hadoop integration

Hadoop Default “Authentication”

- Out of the box, Hadoop “authenticates” users by simply believing whatever username you tell it you are
- This includes telling Hadoop you are the hdfs user, a **superuser**!

```
export HADOOP_USER_NAME=hdfs
```

Kerberos

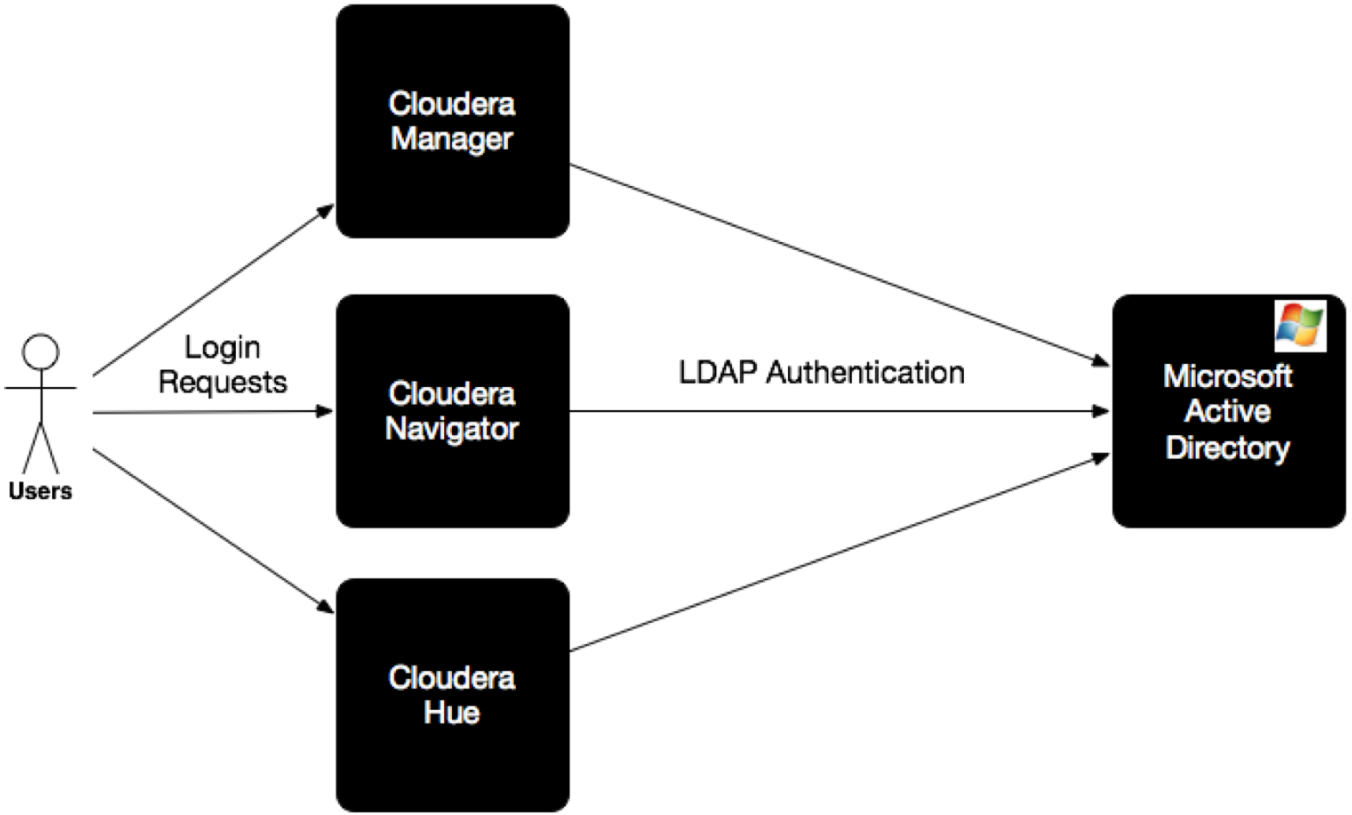
- To enable security in Hadoop, everything starts with Kerberos
- **Every role type of every service has its own unique Kerberos credentials**
- Users must **prove** their identity by obtaining a Kerberos ticket, which is honored by the Hadoop components
- Hadoop components themselves authenticate to each other for intra and inter service communication

Kerberos Authentication

LDAP and SAML

- Beyond just Kerberos, other components such as web consoles and JDBC/ODBC endpoints can authenticate users differently
- **LDAP** authentication is supported for Hive, Impala, Solr, and web-based UIs
- **SAML** (SSO) authentication is supported for Cloudera Manager, Navigator, and Hue
- Some components support both Kerberos and LDAP authentication at the same time
- Generally speaking, LDAP is a much easier authentication mechanism to use for external applications – No Kerberos software and configuration required!
- **…just make sure wire encryption is also enabled to protect passwords**

Web UI LDAP Authentication



Impala Dual-mode Authentication

Enabling Kerberos

- Setting up Kerberos for your cluster is no longer a daunting task
- Cloudera Manager and Apache Ambari provide wizards to automate the provisioning of service accounts and the associated keytabs
- Both MIT Kerberos and Active Directory are supported Kerberos KDC types
- Again, most enterprises use Active Directory so let's see what we need to set it up!

Active Directory Prerequisites

- At least one AD domain controller is setup with LDAPS
- An AD account for Cloudera Manager
- A **dedicated OU** in your desired AD domain
- An account that has **create/modify/delete** user privileges on this OU
- This is **not** a domain admin / administrative account!
- While not required, AD **group policies** can be used to further restrict the accounts
- Install **openldap-clients** on the CM server host, **krb5-workstation** on every host

- From here, use the wizard!

Cloudera Manager Kerberos Wizard

Before using the wizard, please ensure that you have performed the following steps:

Set up a working KDC. Cloudera Manager supports MIT KDC and Active Directory.

- [x] Yes, I've set up a working KDC.

The KDC should be configured to have non-zero ticket lifetime and renewal lifetime. CDH will not work properly if tickets are not renewable.

- [x] Yes, I've checked that the KDC allows renewable tickets.

OpenLdap client libraries should be installed on the Cloudera Manager Server host if you want to use Active Directory. Also, Kerberos client libraries should be installed on ALL hosts.

- [x] Yes, I've installed the client libraries.

Cloudera Manager needs an account that has permissions to create other accounts in the KDC.

- [x] Yes, I've created a proper account for Cloudera Manager.

KDC Information

Specify information about the KDC. The properties below are used by Cloudera Manager to generate principals for CDH daemons running on the cluster.

KDC Type

- ○ MIT KDC
- ◉ Active Directory

KDC Server Host
kdc

ad.hadoop.com

Kerberos Security Realm
default_realm

HADOOP.COM

Kerberos Encryption Types

aes256-cts

aes128-cts

rc4-hmac

Active Directory Suffix

ou=hadoop,DC=hadoop,DC=com

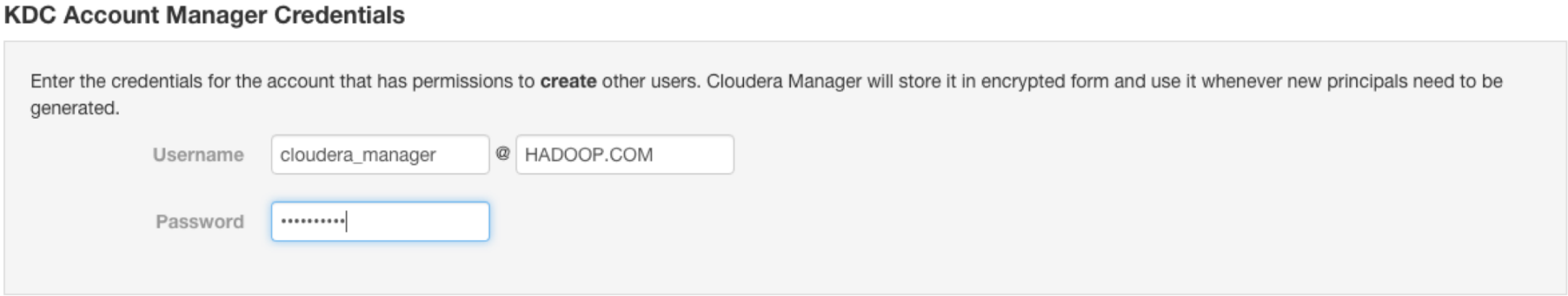
Active Directory Account Prefix

cdh_

Active Directory Domain Controller Override

my-ad-dc1.hadoop.com

Cloudera Manager Kerberos Wizard



Click through the remaining steps

Setting up LDAP Authentication

- CM -> Administration -> Settings
  - Click on category “External Authentication”
- Cloudera Management Services -> Configuration
  - Click on category “External Authentication”
- Hue / Impala / Hive / Solr -> Configuration
  - Search for “LDAP”

Post-Configuration

- Kerberos authentication is enabled
- LDAP authentication is enabled

- **DEMO**: No more fake authentication!

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Questions?

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Authorization

André Araújo

Senior Solutions Architect
Cloudera

Authorization - GDPR

- Broadly underpins **two** of the GDPR Article 5 Principles

- **Data minimization**
- **Integrity and confidentiality**

Authorization - Agenda

- Authorization – Overview
- Configuration Stronger Authorization
- Apache Sentry
- Record Service
- **DEMO:** Strong Authorization
- Questions

Authorization - Overview

- Authorization dictates what a user is permitted to do
- Happens after a user has authenticated to establish identity
- Authorization policies in Hadoop are typically based on:
  - Who the user is and what groups they belong to
  - Role-based access control (RBAC)
- Many different authorization mechanisms in Hadoop components

Authorization in Hadoop

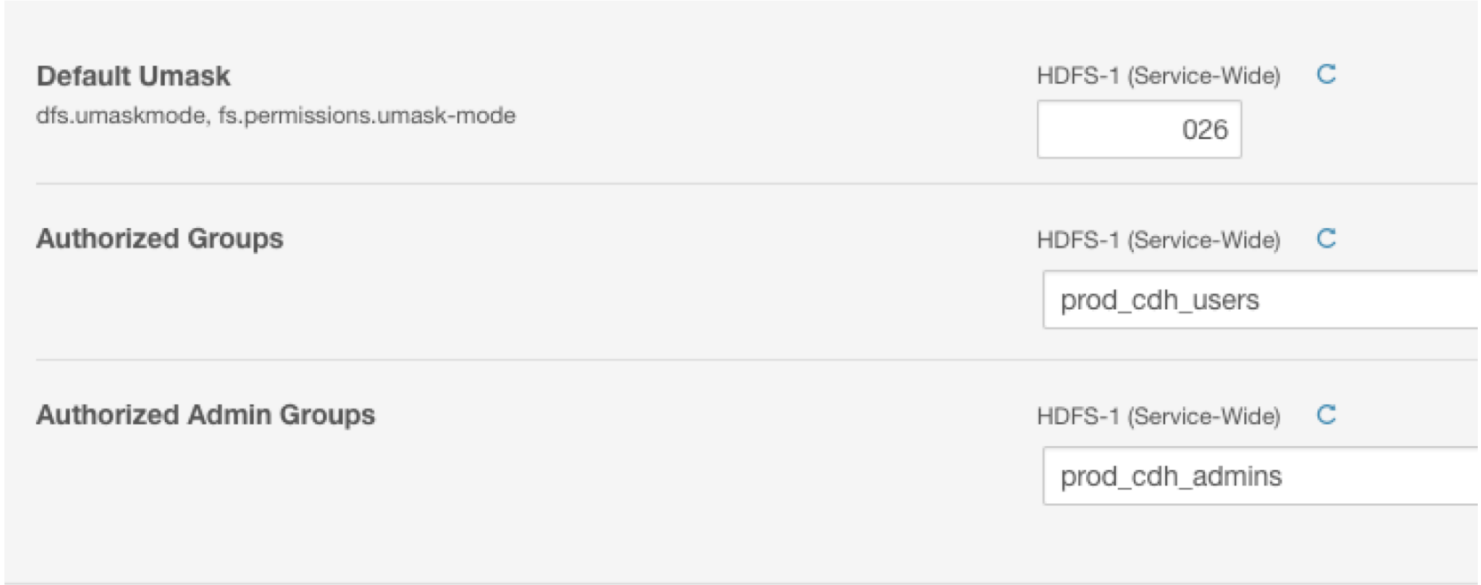
- HDFS file permissions (POSIX ‘rwx rwx rwx’ style)
- Yarn job queue permissions
- Sentry (Hive / Impala / Solr / Kafka)
- Cloudera Manager RBAC
- Cloudera Navigator RBAC
- Hue groups
- Hadoop KMS ACLs
- HBase ACLs
- etc.

Default Authorization Examples

- HDFS
  - Default umask is 022, making all new files world readable
  - Any authenticated user can execute hadoop shell commands
- YARN
  - Any authenticated user can submit and kill jobs for any queue
- Hive metastore
  - Any authenticated user can modify the metastore (CREATE/DROP/ALTER/etc.)

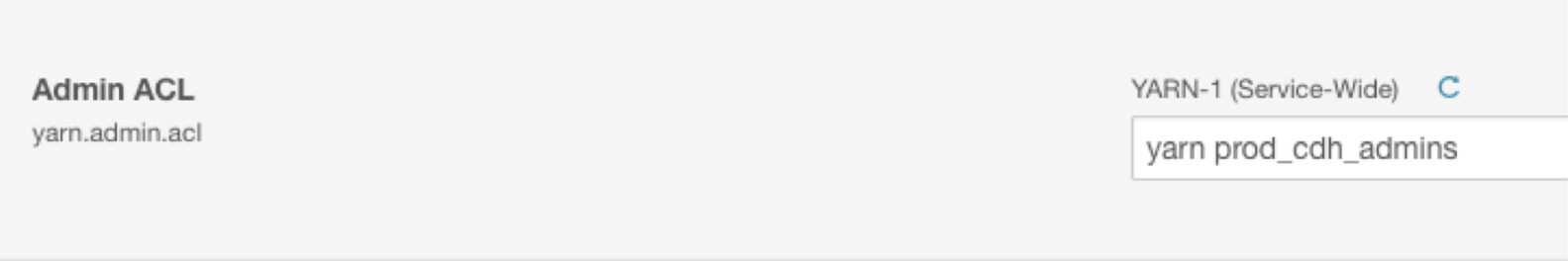
Configuring HDFS Authorization

- Set default umask to 026
- Setup hadoop-policy.xml (Service Level Authorization)



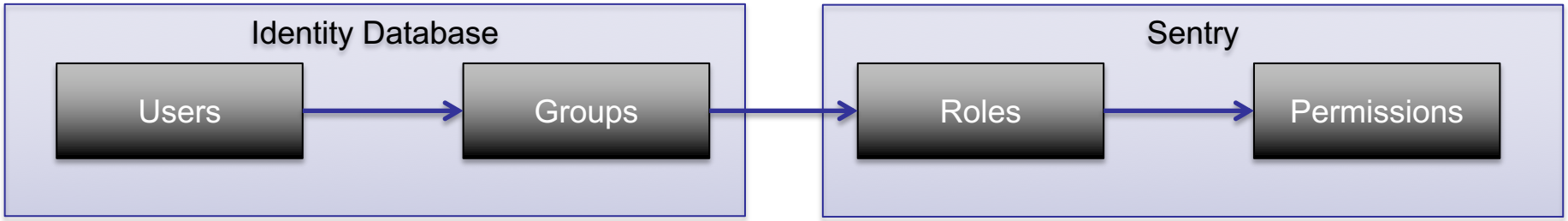
Configuring Yarn Authorization

- Setup the YARN admin ACL

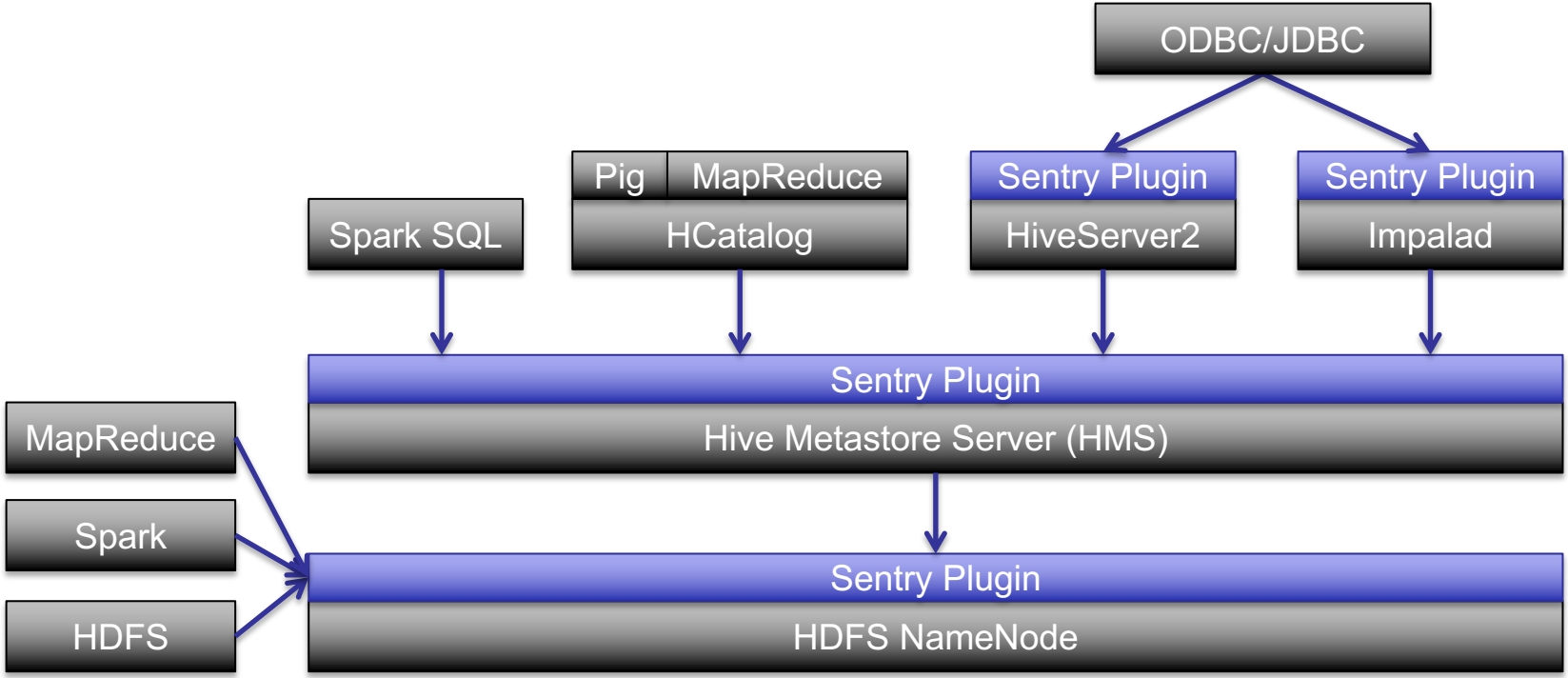


Apache Sentry

- Provides centralized RBAC for several components
  - **Hive / Impala:** Databases, tables, views, columns
  - **Solr:** Collections, documents, indexes
  - **Kafka:** Cluster, topic, consumer group



Apache Sentry (Cont.)



Configuring Sentry

- Cloudera Manager -> Add Service -> Sentry
- Hive
  - Set Sentry service
  - Disable HiveServer2 impersonation
- Impala
  - Set Sentry Service
- HDFS
  - Enable Sentry HDFS Synchronization
  - Enable extended ACLs
  - Specify path prefixes

Post Configuration

- HDFS setup with a better umask and service level authorization
- YARN setup with restrictive admin ACLs
- Hive, Impala, and HDFS setup with Sentry integration
  - `create role hive_admins;`
  - `grant role hive_admins to group hive_admins;`
  - `grant all on server server1 to role hive_admins;`
  - `create role hadoop_users;`
  - `grant role hadoop_users to group hadoop_users;`
  - `grant select,insert on database test to role hadoop_users;`

- **DEMO:** No more default authorization holes!

Authorization - Summary

- HDFS file permissions (POSIX ‘rwx rwx rwx’ style)
- Yarn job queue permissions
- Sentry (Hive / Impala / Solr / Kafka)
- Cloudera Manager RBAC
- Cloudera Navigator RBAC
- Hue groups
- Hadoop KMS ACLs
- HBase ACLs
- etc.

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Questions

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Encryption of Data in Transit

Syed Rafice

Principal System Engineer
Cloudera

Encryption in Transit - GDPR

- Broadly underpins **one** of the GDPR Article 5 Principles
- **Integrity and confidentiality**

Agenda

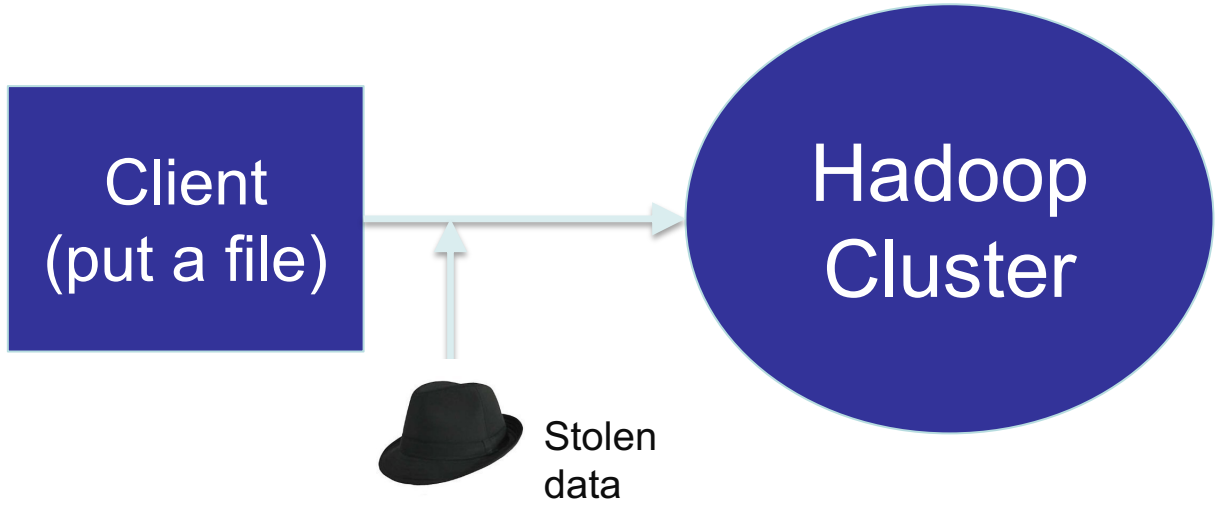
- Why encryption of data on the wire is important
- Technologies used in Hadoop
  - SASL “Privacy”
  - TLS
- For each:
  - Demo without
  - Discussion
  - Enabling in Cloudera Manager
  - Demo with it enabled

Why Encrypt Data in Transit?

- Networking configuration (firewalls) can mitigate some risk
- Attackers may already be inside your network
- Data and credentials (usernames and passwords) have to go into and out of the cluster
- Regulations around transmitting sensitive information

Example

- Transfer data into a cluster
- Simple file transfer: “hadoop fs –put”
- Attacker sees file contents go over the wire



Two Encryption Technologies

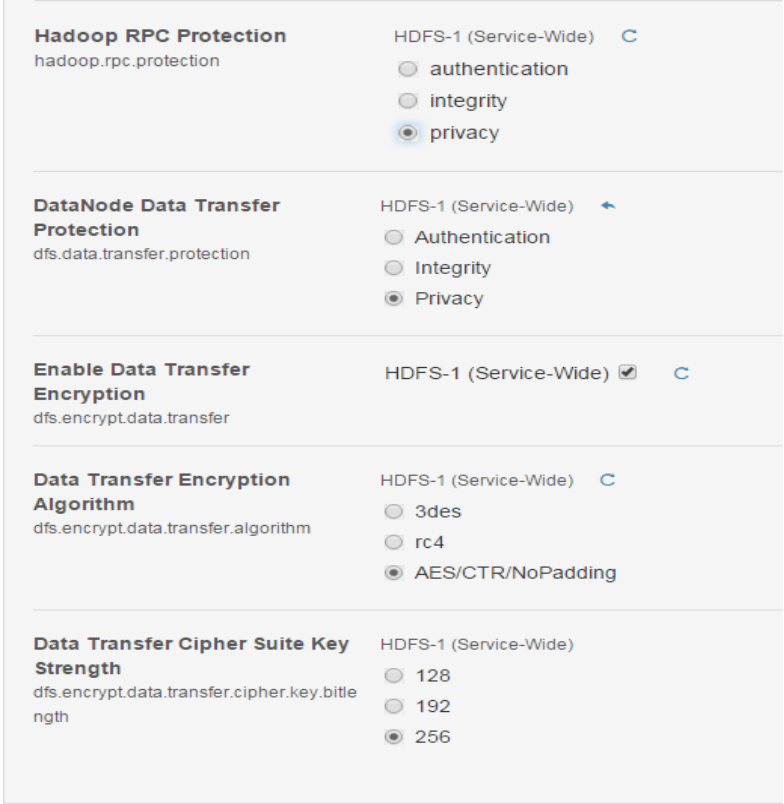
- SASL “confidentiality” or “privacy” mode
  - Protects core hadoop
- TLS – Transport Layer Security
  - Used for “everything else”

SASL

- Simple Authentication and Security Layer
- Not a protocol, but a framework for passing authentication steps between a client and server
- Pluggable with different authentication types
  - GSS-API for Kerberos (Generic Security Services)
- Can provide transport security
  - “auth-int” – integrity protection: signed message digests
  - “auth-conf” – confidentiality: encryption

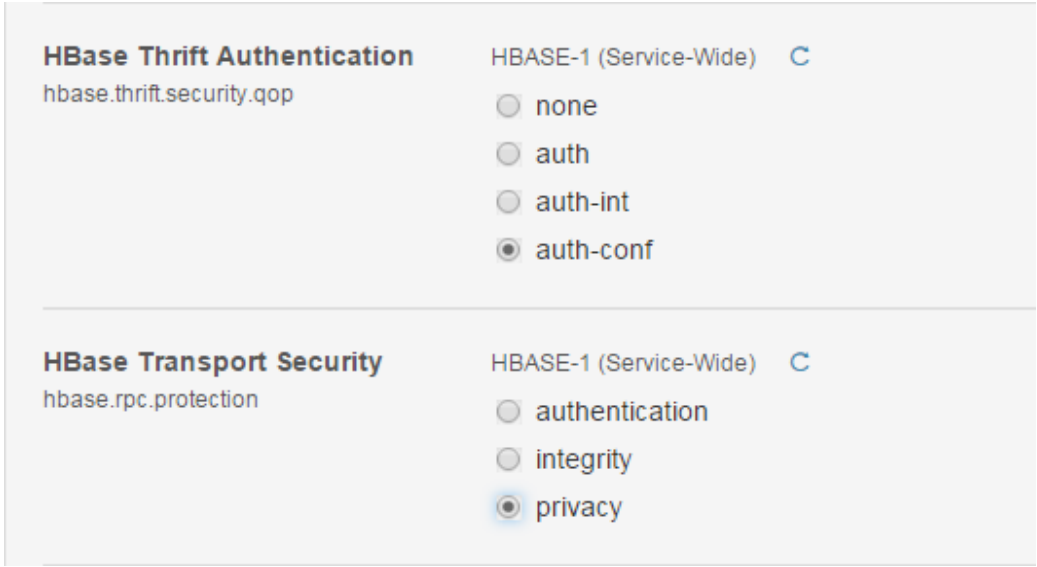
SASL Encryption - Setup

- First, enable Kerberos
- HDFS:
  - Hadoop RPC Protection
  - Datanode Data Transfer Protection
  - Enable Data Transfer Encryption
  - Data Transfer Encryption Algorithm
  - Data Transfer Cipher Suite Key Strength



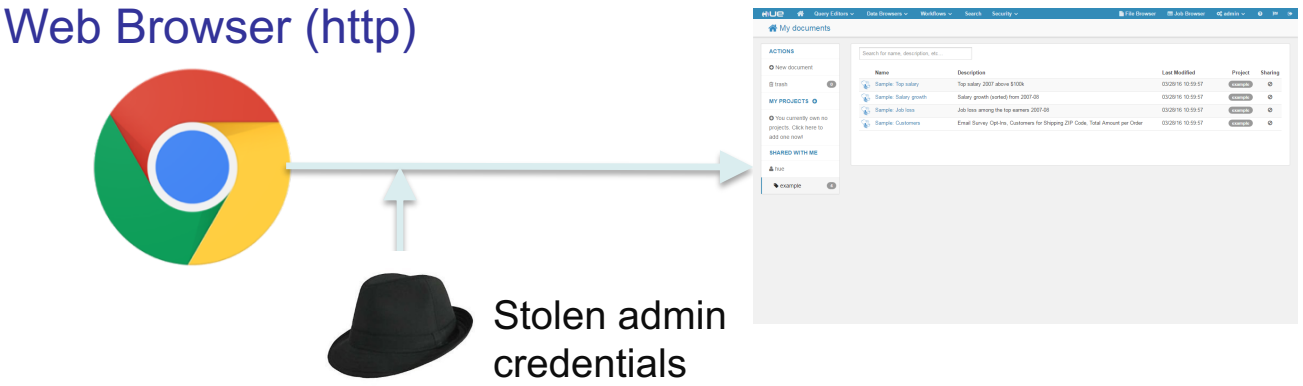
SASL Encryption - Setup

- Hbase
  - HBase Thrift Authentication
  - Hbase Transport Security



TLS

- Transport Layer Security
  - The successor to SSL – Secure Sockets Layer
  - The term SSL was deprecated 15 years ago, but we still use it
  - TLS is what's behind https:// web pages



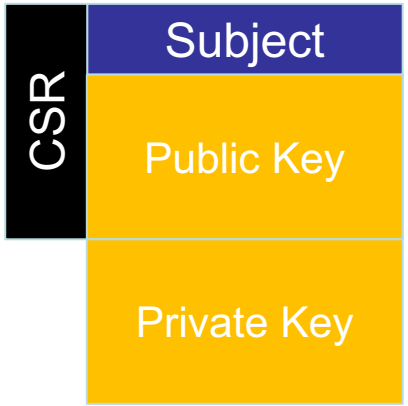
TLS - Certificates

- TLS relies on certificates for authentication
- You'll need one certificate per machine
- Certificates:
  - Cryptographically prove that you are who you say you are
  - Are issued by a "Certificate Authority" (CA)
  - Have a "subject", an "issuer" and a "validity period"
  - Many other attributes, like "Extended Key Usage"
  - Let's look at an https site

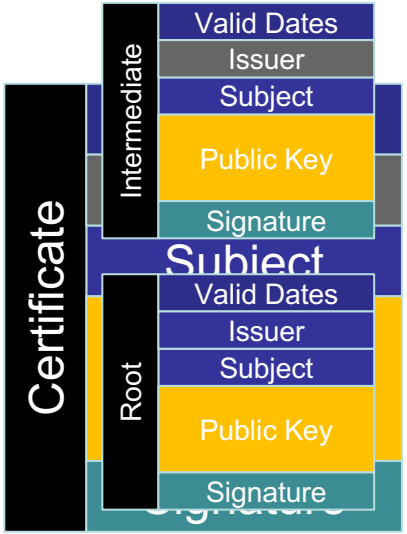
TLS – Certificate Authorities

- “Homemade” CA using openssl
  - Suitable for test/dev clusters only
- Internal Certificate Authority
  - A CA that is trusted widely inside your organization, but not outside
  - Commonly created with Active Directory Certificate Services
  - Web browsers need to trust it as well
- External Certificate Authority
  - A widely known CA like VeriSign, GeoTrust, Symantec, etc
  - Costs $$$ per certificate

You



Certificate Authority



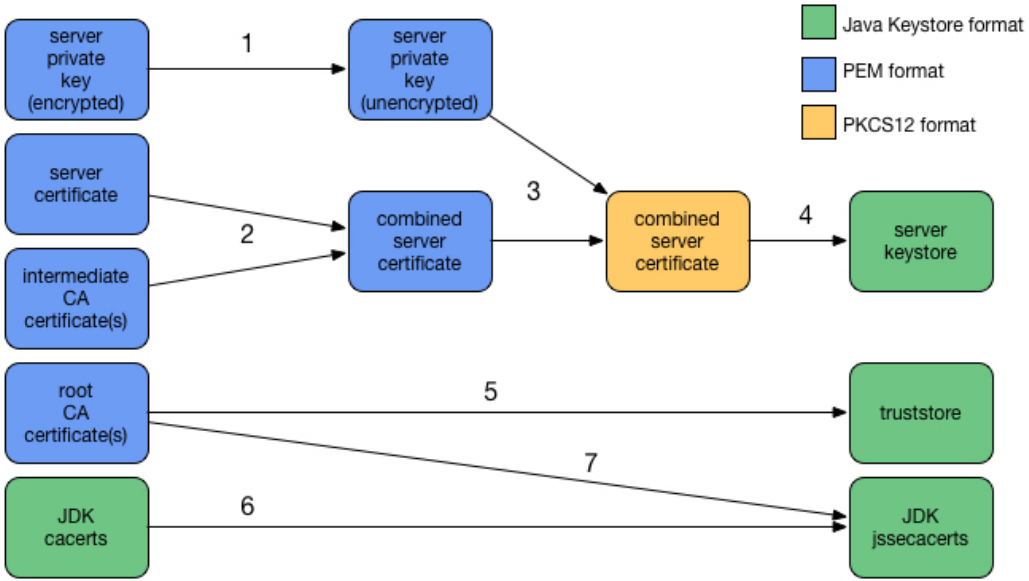
TLS – Certificate File Formats

- Two different formats for storing certificates and keys
- PEM
  - “Privacy Enhanced Mail” (yes, really)
  - Used by openssl; programs written in python and C++
- JKS
  - Java KeyStore
  - Used by programs written in Java
- The Hadoop ecosystem uses both
- Therefore you must translate private keys and certificates into both formats

TLS – Key Stores and Trust Stores

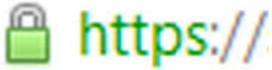
- Keystore
  - Used by the server side of a TLS client-server connection
  - JKS: Contains private keys and the hosts's certificate; Password protected
  - PEM: typically one certificate file and one password-protected private key file
- Truststore
  - Used by the client side of a TLS client-server connection
  - Contains certificates that the client trusts: the Certificate Authorities
  - JKS: Password protected, but only for an integrity check
  - PEM: Same concept, but no password
  - There is a system-wide certificate store for both PEM and JKS formats.

TLS – Key Stores and Trust Stores



1 - openssl rsa -in `hostname -f`.key.temp -out `hostname -f`.key
2 - cat server.pem int-CA.pem > `hostname -f`.pem
3 - openssl pkcs12 -export -in `hostname -f`.pem -inkey `hostname -f`.key -out `hostname -f`.pfx
4 - keytool -importkeystore -srcstoretype PKCS12 -srckeystore `hostname -f`.pfx -destkeystore `hostname -f`.jks
5 - keytool -importcert -file root-CA.pem -alias root-CA -keystore truststore.jks
6 - cp $JAVA_HOME/jre/lib/security/cacerts $JAVA_HOME/jre/lib/security/jssecacerts
7 - keytool -importcert -file root-CA.pem -alias root-CA -keystore $JAVA_HOME/jre/lib/security/jssecacerts

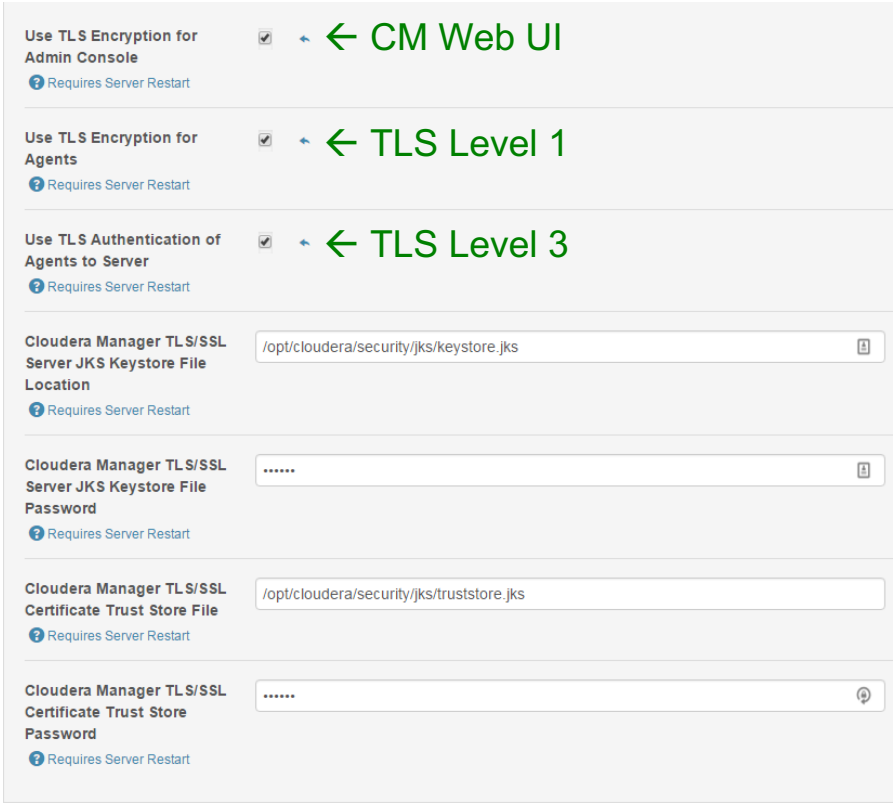
TLS – Securing Cloudera Manager

- CM Web UI - 

- CM Agent -> CM Server communication – 3 “Levels” of TLS use
  - Level 1: Encrypted but no certificate verification. Akin to clicking on 

  - Level 2: Agent verifies the server’s certificate
  - Level 3: Agent and Server verify each other’s certificate. This is called TLS mutual authentication: each side is confident that it’s talking to the other
    - Note: TLS level 3 requires that certificates are suitable for both “TLS Web Server Authentication” and “TLS Web Client Authentication”
  - Very Sensitive Information goes over this channel
  - Like Kerberos Keytabs. Therefore, set up TLS in CM first before Kerberos

Cloudera Manager TLS



The CM Agent Settings

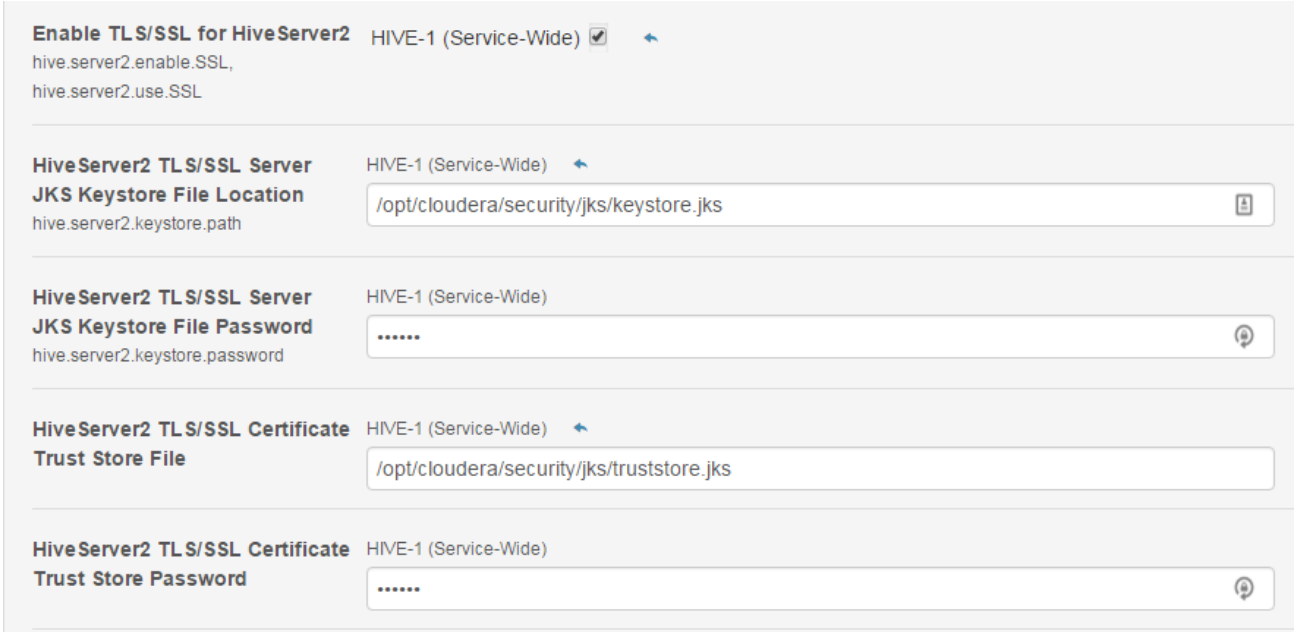
- Agent **/etc/cloudera-scm-agent/config.ini**

```
use_tls=1                ← TLS Level 1
verify_cert_file=  full path to CA certificate.pem file   ← TLS Level 2
client_key_file=   full path to private key.pem file      ┐
client_keypw_file= full path to file containing password for key  ├ TLS Level 3
client_cert_file=  full path to certificate.pem file      ┘
```

TLS for CM-Managed Services

- CM requires that all files (jks and pem) are in the same location on each machine
- For each service (HDFS, Hue, Hbase, Hive, Impala, …)
  - Search the configuration for “TLS”
  - Check the “enable” boxes
  - Provide keystore, truststore, and passwords

Hive Example

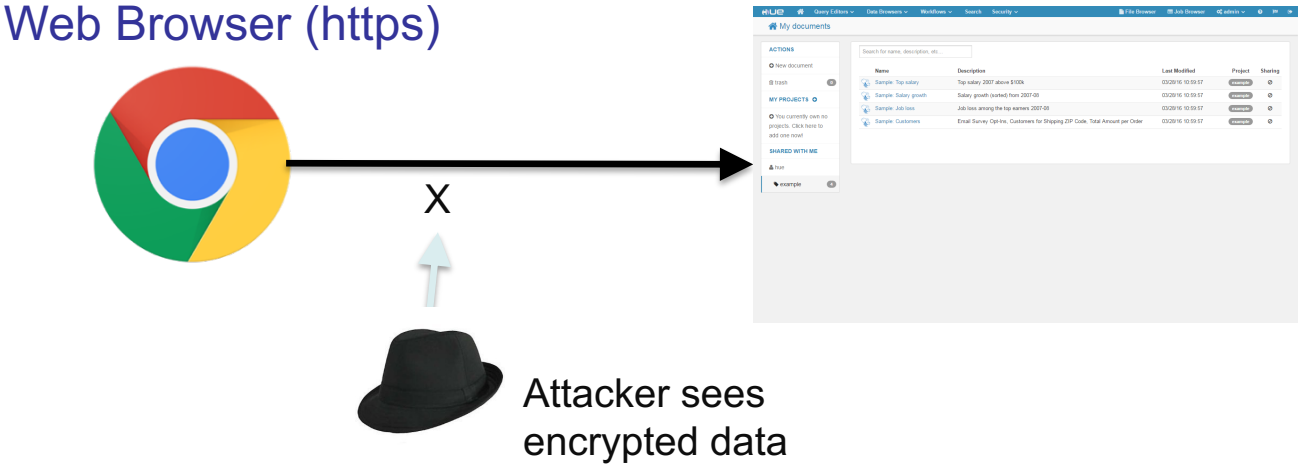


TLS - Troubleshooting

- To examine certificates
  - `openssl x509 -in <cert>.pem -noout -text`
  - `keytool -list -v -keystore <keystore>.jks`
- To attempt a TLS connection as a client
  - `openssl s_client -connect <host>:<port>`
  - This tells you all sorts of interesting TLS things

Example - TLS

- Someone attacks an https connection to Hue
- Note that this is only one example, TLS protects many, many things in hadoop



Conclusions

- You need to encrypt information on the wire
- Technologies used are SASL encryption and TLS
- TLS requires certificate setup

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Questions?

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

HDFS Encryption at Rest

Mubashir Kazia

Principal Solutions Architect
Cloudera

Encryption in Rest - GDPR

- Broadly underpins **one** of the GDPR Article 5 Principles
- **Integrity and confidentiality**

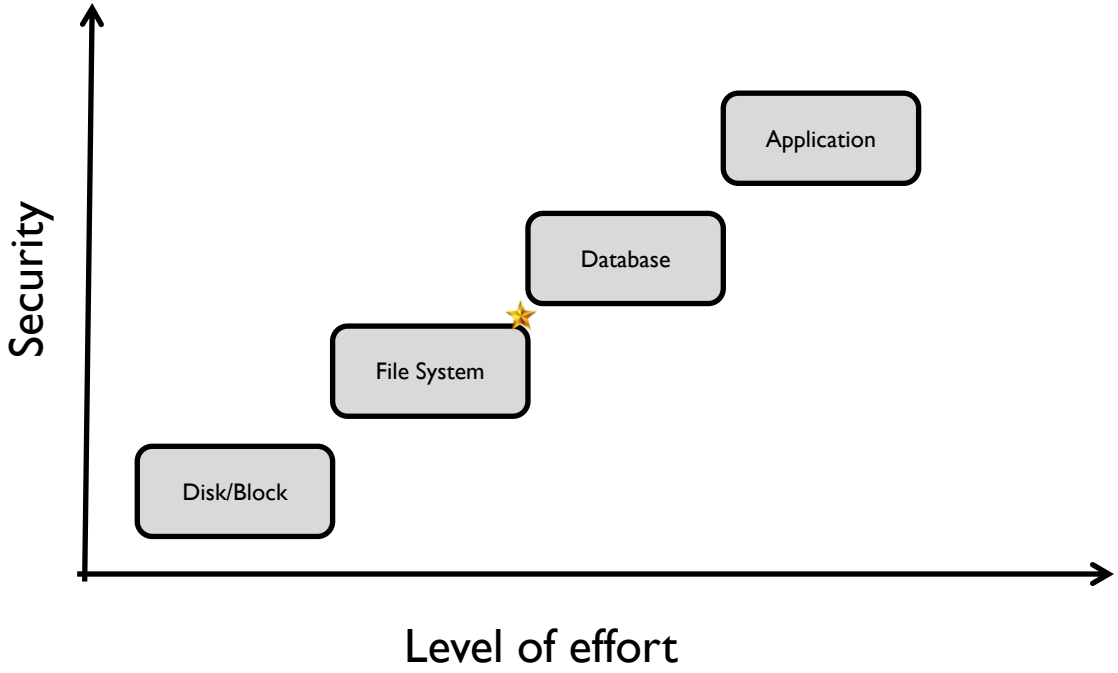
Agenda

- Why Encrypt Data
- HDFS Encryption
- Demo
- Questions

Why store encrypted data?

- Customers often are mandated to protect data at rest
  - GDPR
  - PCI
  - HIPAA
  - National Security
  - Company confidential
- Encryption of data at rest helps mitigate certain security threats
  - Rogue administrators (insider threat)
  - Compromised accounts (masquerade attacks)
  - Lost/stolen hard drives

Options for encrypting data



Architectural Concepts

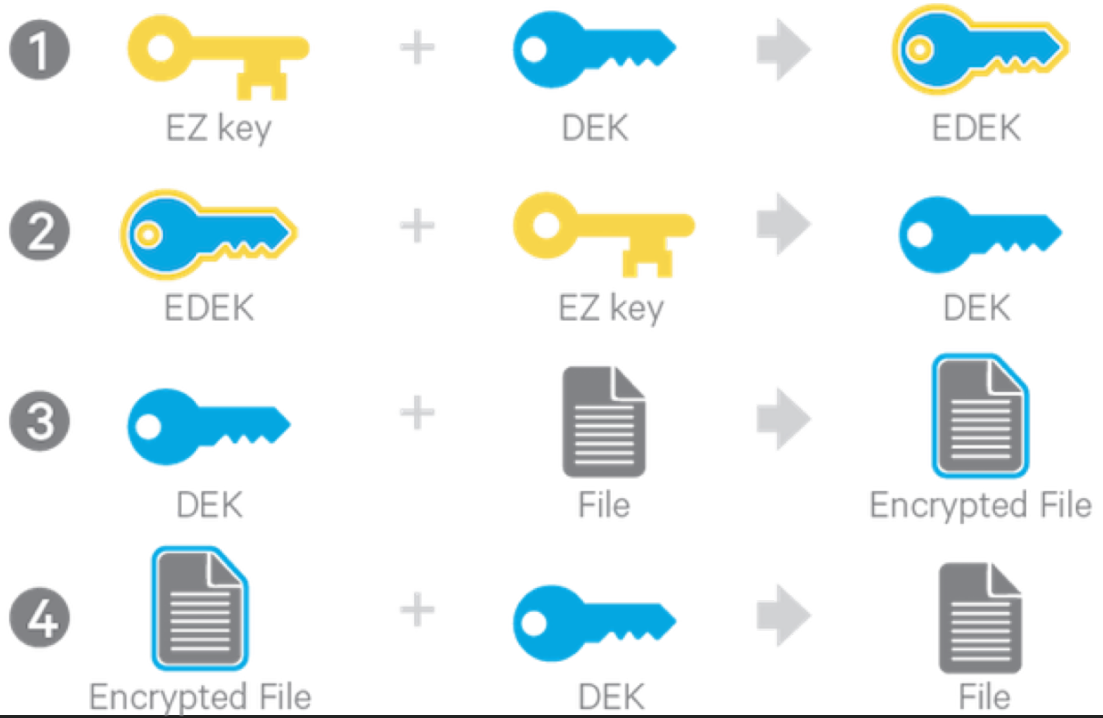
- Encryption Zones
- Keys
- Key Management Server

Encryption Zones

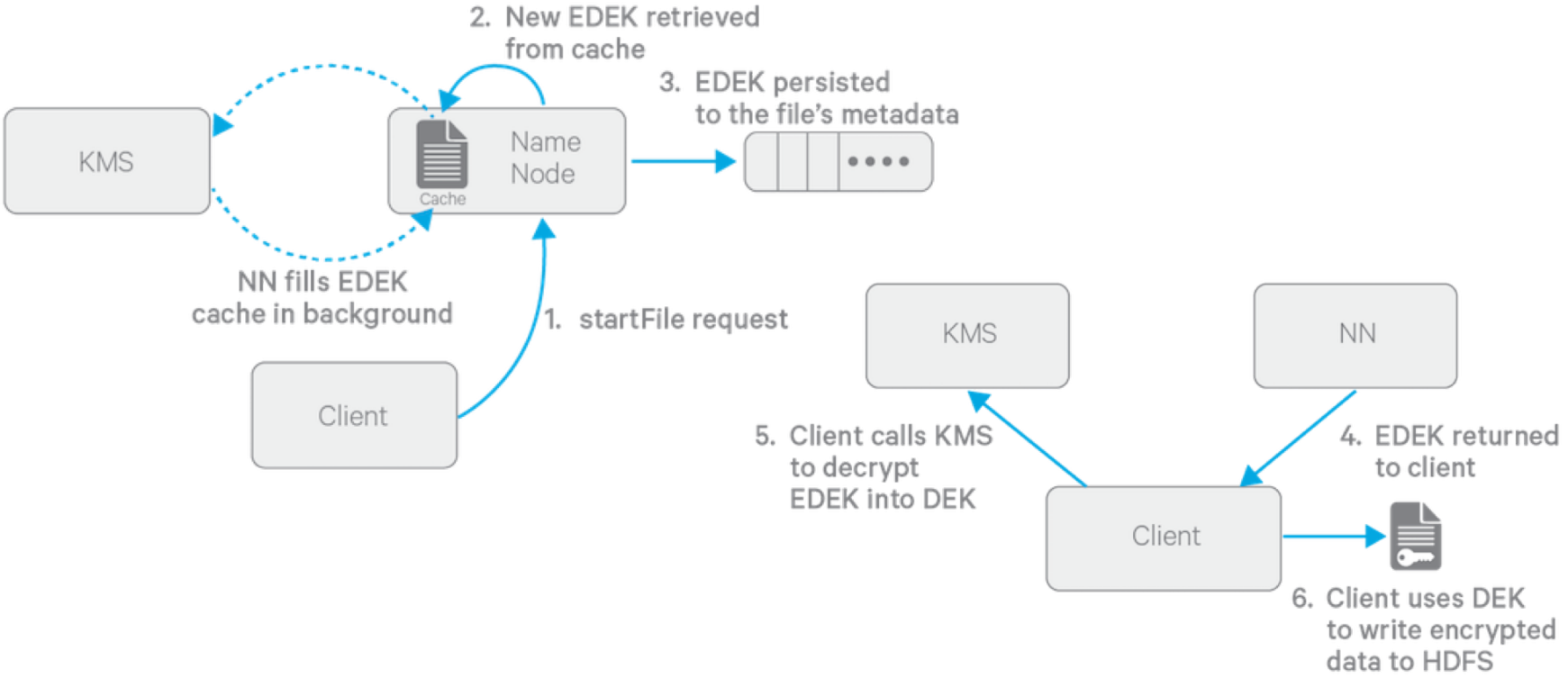
- An HDFS directory in which the contents (including subdirs) are encrypted on write and decrypted on read.
- An EZ begins life as an empty directory
- Rename/Move in/out of an EZ are prohibited

- Encryption is transparent to application with no code changes

EZ Keys, Data Encryption Keys, and Encrypted Data Encryption Keys



Key Handling



Key Management Server (KMS)

- KMS sits between client and key server
  - E.g. Cloudera Navigator Key Trustee
- Provides a unified API and scalability
- REST API
- Does not actually store keys (backend does that), but does cache them
- ACLs on per-key basis

HDFS Encryption Configuration

- `hadoop key create <keyname> -size <keySize>`
- `hdfs dfs –mkdir <path>`
- `hdfs crypto –createZone –keyName <keyname> -path <path>`

KMS Per-User ACL Configuration

- White lists (check for inclusion) and black lists (check for exclusion)
- **etc/hadoop/kms-acls.xml**
  - **hadoop.kms.acl.CREATE**
  - **hadoop.kms.blacklist.CREATE**
  - … **DELETE, ROLLOVER, GET, GET_KEYS, GET_METADATA, GENERATE_EEK, DECRYPT_EEK**
  - **hadoop.kms.acl.<keyname>.<operation>**
  - **MANAGEMENT, GENERATE_EEK, DECRYPT_EEK, READ, ALL**

Best practices

- Enable authentication (Kerberos)
- Enable TLS/SSL
- Use KMS acls to setup KMS roles, blacklist HDFS admins and grant per key access
- Do not use the KMS with default JCEKS backing store
- Use hardware that offers AES-NI instruction set
  - Install openssl-devel so Hadoop can use Openssl crypto codec
- Make sure you have enough entropy on all the nodes
  - Run rngd or haveged

Best practices

- Do not run KMS on master or worker nodes
- Run multiple instances of KMS for high availability and load balancing
- Harden KMS instance and use internal firewall so only KMS and ssh etc. ports are reachable from known subnets
- Make secure backups of KMS

HDFS Encryption - Summary

- Good performance (4-10% hit) with AES-NI
- No mods to existing applications
- Prevents attacks at the filesystem and below
- Data is encrypted all the way to the client
- Key management is independent of HDFS
- Can prevent HDFS admin from accessing secure data

Demo

- Accessing HDFS encrypted data from Linux storage

User	Group	Role
hdfs_admin	cdh_admin	HDFS Admin
kms_admin	key_admin	KMS Admin
alice	key1_decrypt	User with DECRYPT_EEK access to key1
bob	key2_decrypt	User with DECRYPT_EEK access to key2

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Questions?

Strata
DATA CONFERENCE

strataconf.com
#StrataData

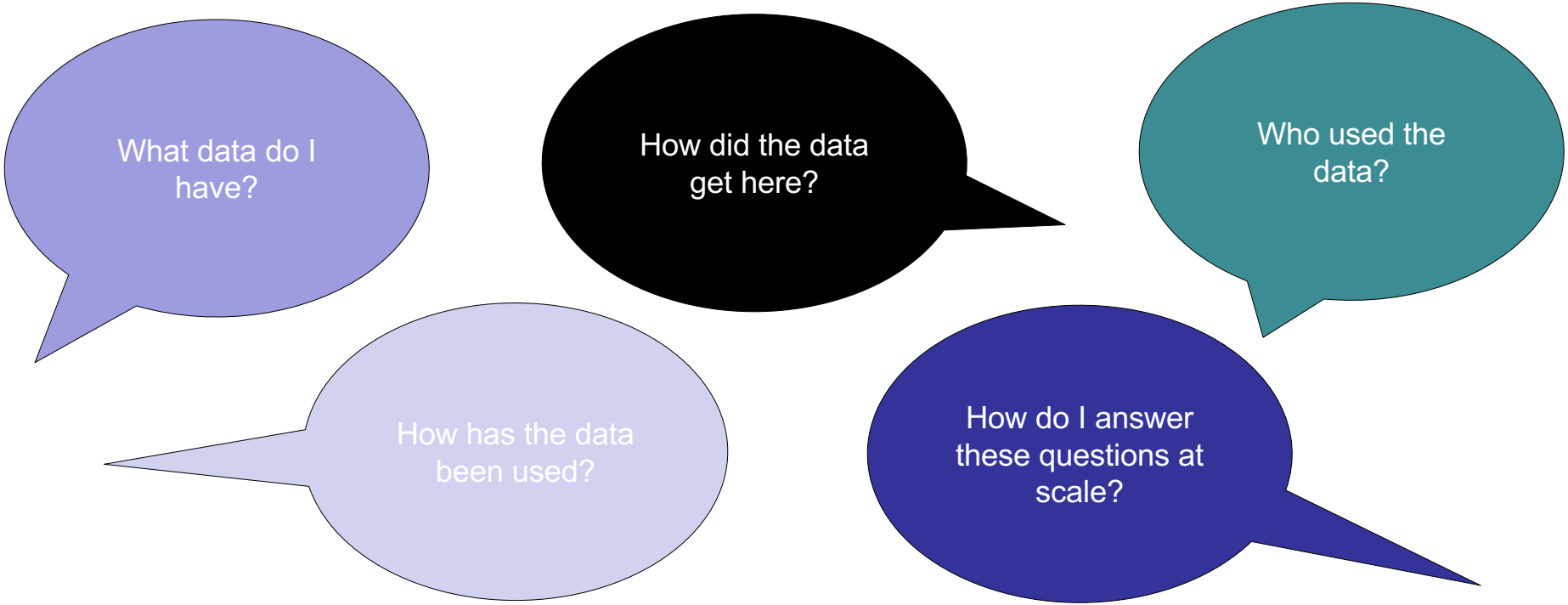
PRESENTED BY
O'REILLY cloudera

Hadoop Data Governance and GDPR

Mark Donsky

Director, Product Management
Cloudera

Data Governance Frequently Asked Questions



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

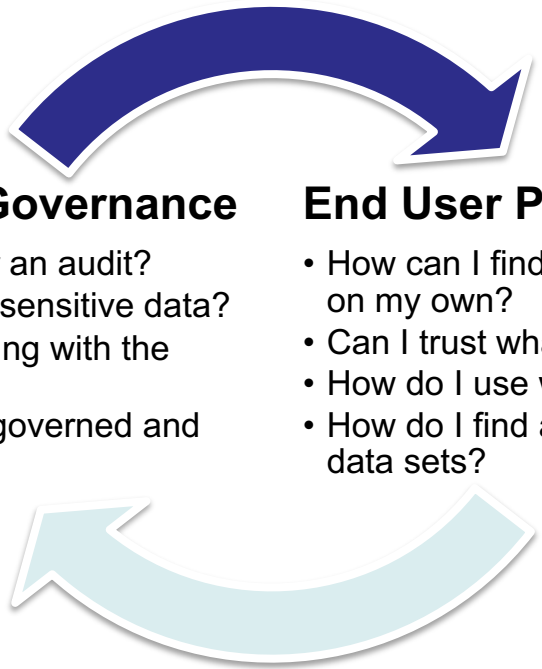
Applications are shifting to the cloud, and data governance must still be applied consistently

Self-service data discovery is mandatory for big data

What are the governance challenges of GDPR?

- **Right to erasure:** enforcement of row-level deletions are challenging with traditional big data storage such as HDFS and block storage
- **Diversity of data:** personal data can be hidden in unstructured data
- **Volume of data:** organizations now must govern orders of magnitude more data
- **Multiple compute engines and lots of users:** many different access points into sensitive data

GDPR compliance must be integrated into everyday workflows



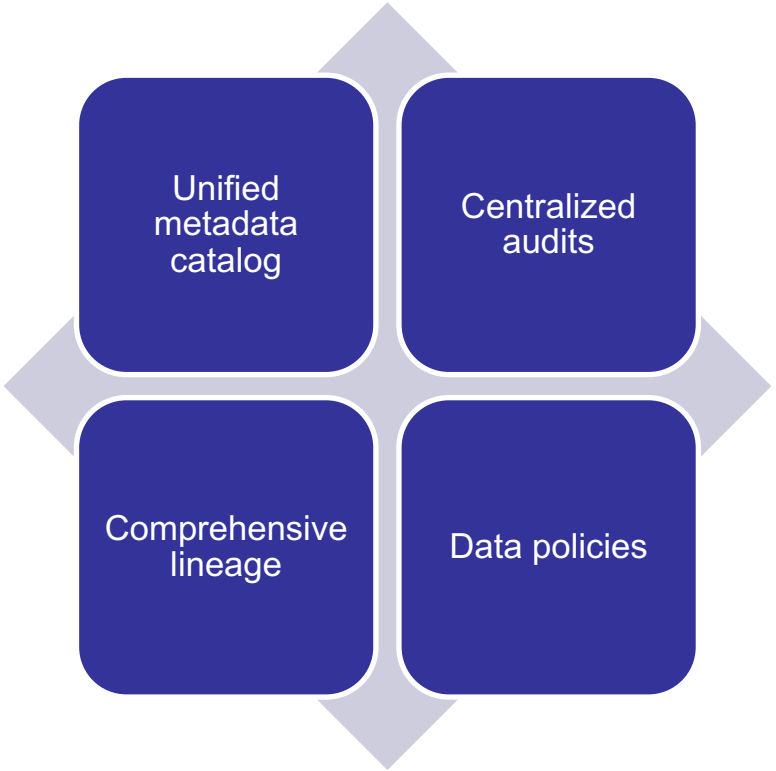
Compliance/Governance

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

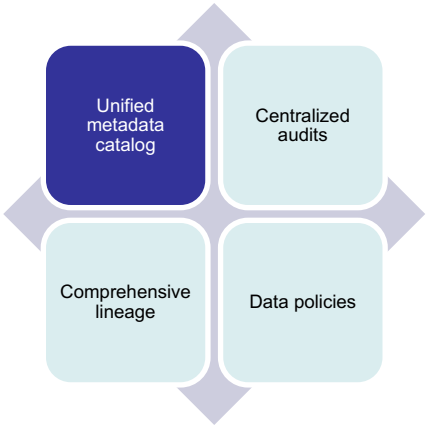
End User Productivity

- 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?

Big Data Governance Requirements for GDPR



Unified Metadata Catalog



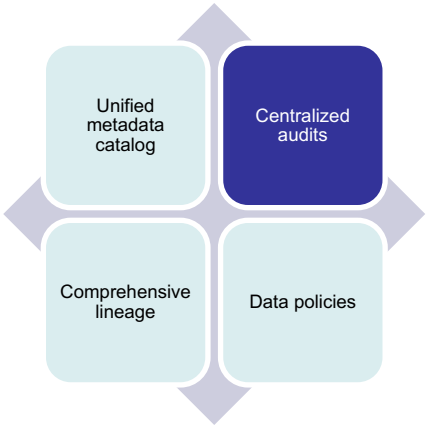
Technical Metadata	All files in directory /sales
	All files with permissions 777
	Anything older than 7 years
	Any not accessed in the past 6 months
Managed Metadata	Sales data from last quarter for the Northeast region
	Protected health information
	Business glossary definitions
	Data sets associated with clinical trial X
Custom Metadata	Tables that I want to share with my colleagues
	Data sets that I want to retrieve later
	Data sets that are organized by my personal classification scheme (e.g., “quality = high”)

Challenges

- Technical metadata in Hadoop is component-specific
- Curated/custom attributes: Hive meta store has comments, and HDFS has extended attributes, but:
  - Not searchable
  - No validation
- Aggregated analytics are not possible
  - How many files are older than two years?

Centralized Audits

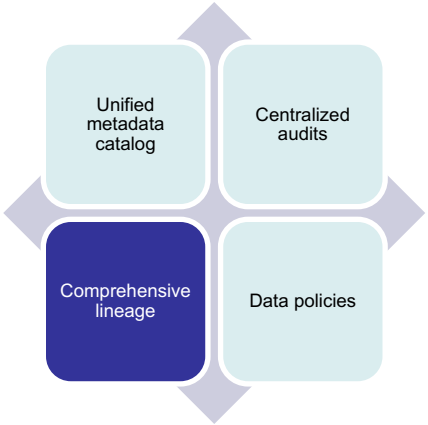
- **Goal:** Collect all audit activity in a single location
  - Redact sensitive data from the audit logs to simplify compliance with regulation
  - Perform holistic searches to identify data breaches quickly
  - Publish securely to enterprise tools

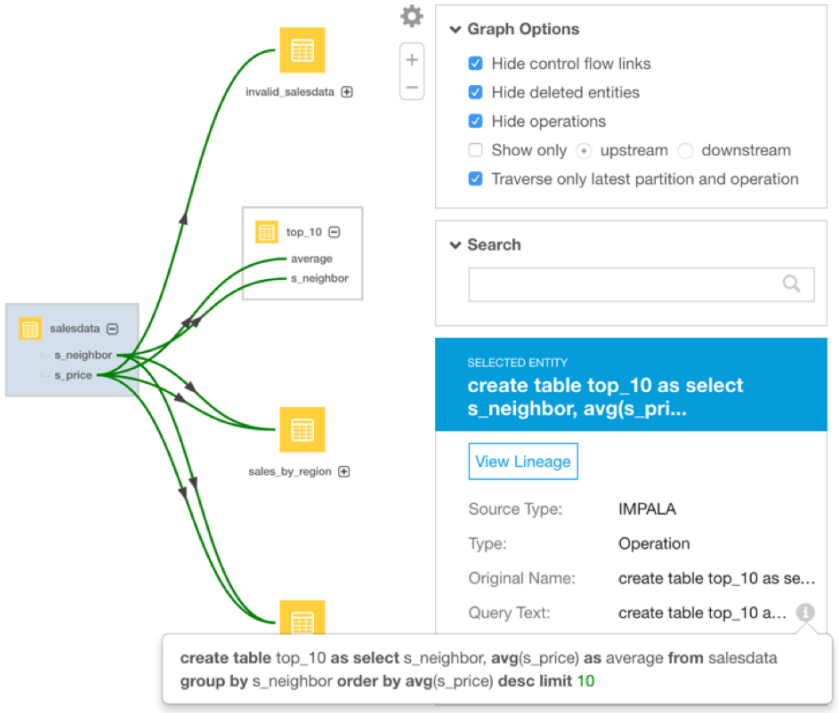


Challenges

- Each component has its own audit log, but:
- Sensitive data may exist in the audit log
  - Select * from transactions where cc_no = “1234 5678 9012 3456”
- It’s difficult to do holistic searches
  - What did user *a* do yesterday?
  - Who accessed file *f*?
- Integration with enterprise SIEM and audit can be complex

Comprehensive Lineage





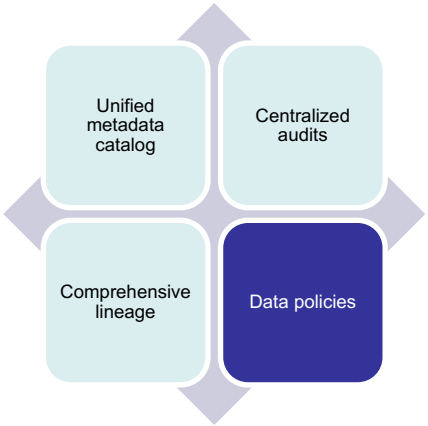
Challenges

- Most uses of lineage require column-level lineage
- Hadoop does not capture lineage in an easily-consumable format
- Lineage must be collected automatically and cover all compute engines
- Third-party tools and custom-built applications need to augment lineage

Data Policies

- **Goal:** Manage and automate the information lifecycle from ingest to purge/cradle to grave, based on the unified metadata catalog

- Once you find data sets, you'll likely need to do something with them
  - GDPR right to erasure
  - Tag every new file that lands in /sales as sales data
  - Send an alert whenever a sensitive data set has permissions 777
  - Purge all files that are older than seven years



Challenges

- Oozie workflows can be difficult to configure
- Event-triggered oozie workflows are limited to very few technical metadata attributes, such as directory path
- Data stewards prefer to define, view, and manage data policies in a metadata-centric fashion

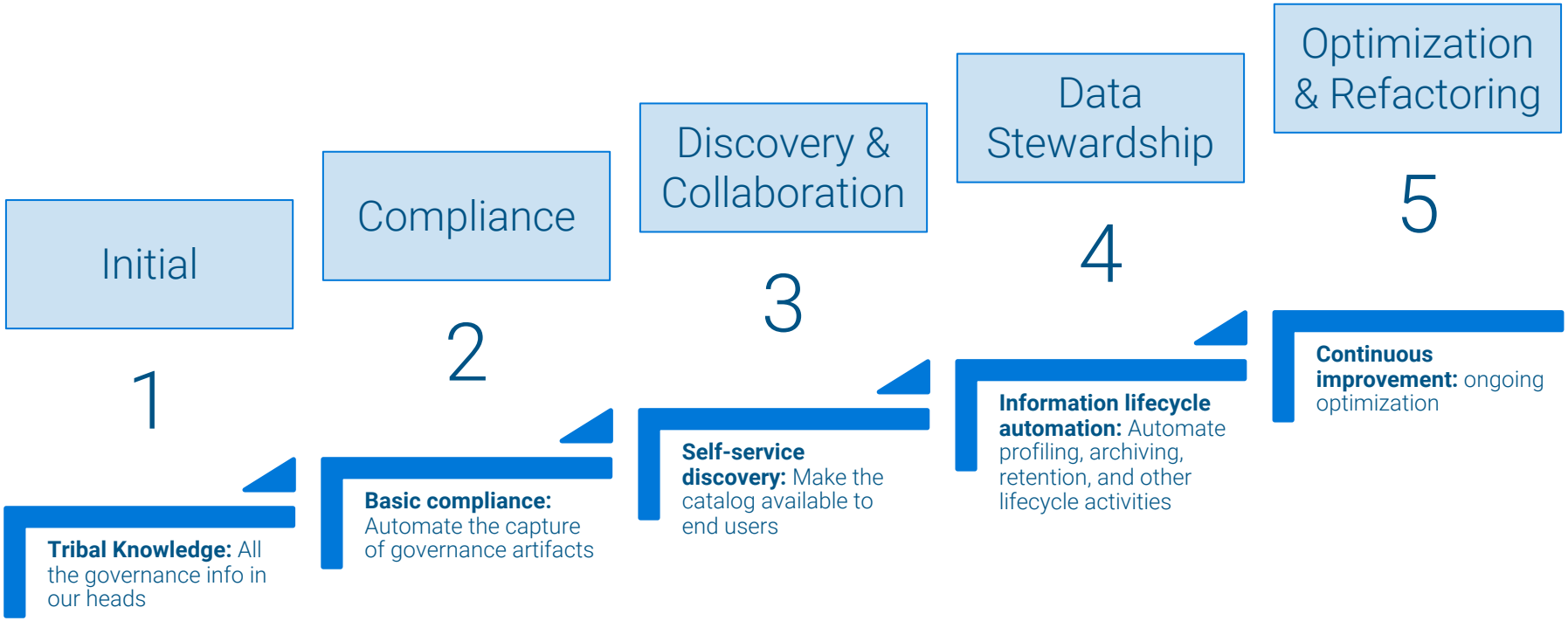
Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

GDPR and Governance Best Practices

Governance Maturity Progression

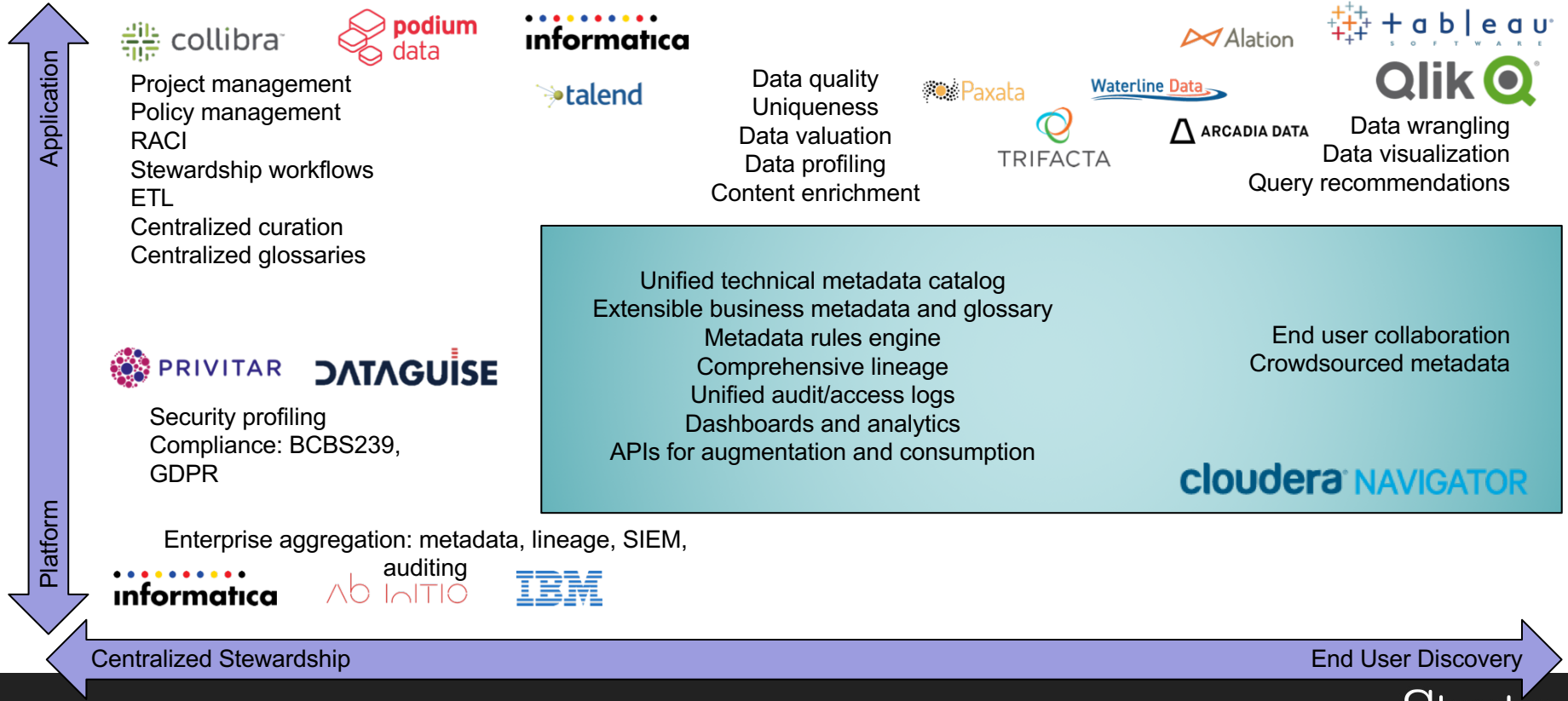


How Cloudera can help with GDPR compliance

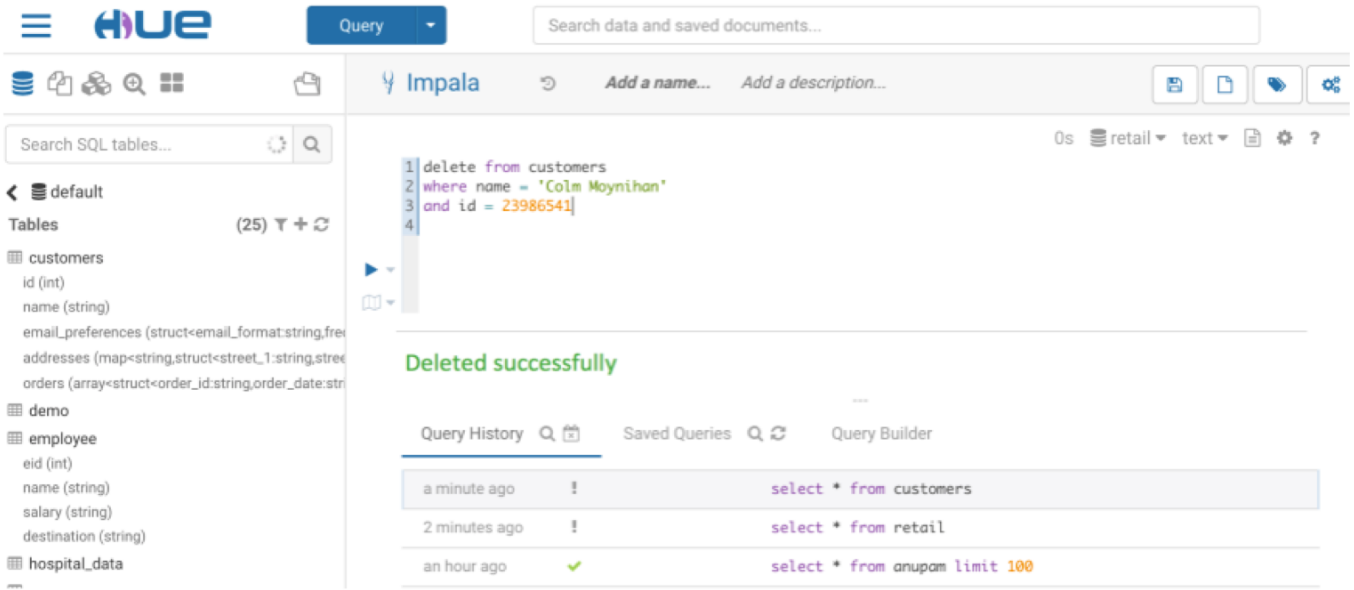
The GDPR principles	Typical customer challenges
Integrity and confidentiality	Applying industry standard IT security controls to prevent unauthorised access.
Accountability	Demonstrating compliance with full audit. Detecting and analyzing breaches, in order to meet the 72 hour reporting requirement
Lawfulness, fairness and transparency	Implementing a way to keep track of personal data.
Purpose limitation	Track consent and data usage while allowing data scientists to mine it using tools of choice
Data minimization	Removing or anonymising data where possible. Preventing unlawful data transfers outside the EU
Accuracy	Finding a low-overhead way to fix data
Storage limitation	Delete individual personal data records in HDFS or Cloud storage, since those file systems are immutable.

Cloudera addresses these challenges:
Kudu: fast erasure of individual records
Cloudera Data Science Workbench: governed data science
Cloudera Navigator: full governance with column-level lineage and rich metadata
Cloudera Navigator Encrypt: comprehensive encryption
Cloudera SDX: consistent data context, including governance and compliance, across all workloads

Data Stewardship and Governance Activities



Kudu: Fast erasure of individual records



Cloudera Data Science Workbench

“Laptop Data Science”

Typical Big Data Environment

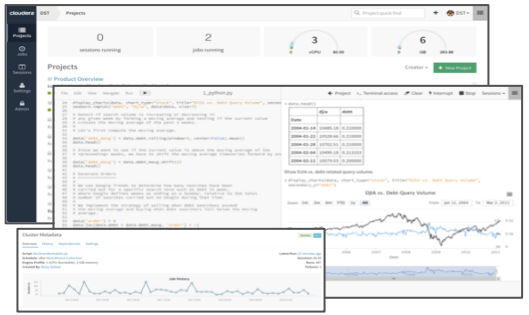
Data scientists pull data to their laptops so they can run their own tools

- Copy personal data to laptops
- Fails GDPR compliance audit
- Potential data breach

Centralized Data Science

cloudera

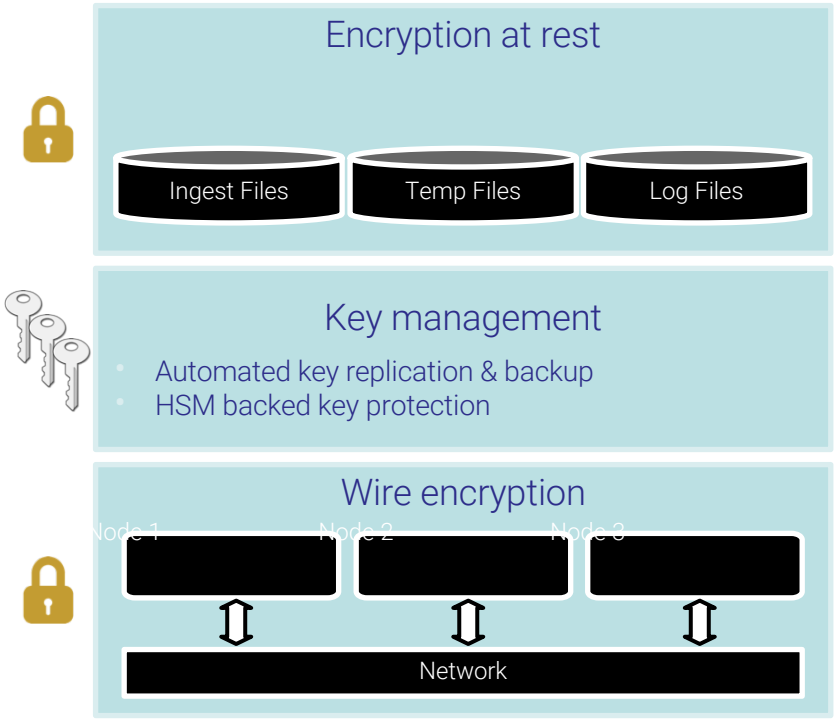
Data Science Workbench

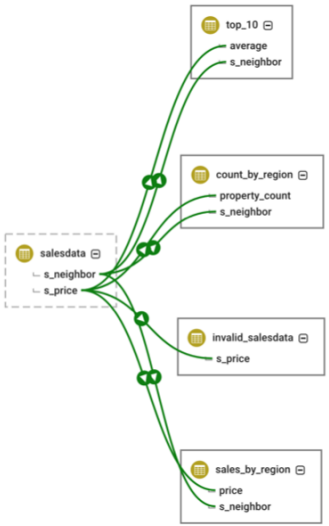


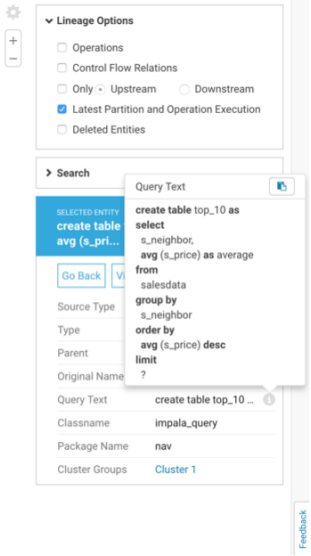
- Personal data remains governed
- Purpose limitation enforced

Cloudera Navigator and Cloudera Navigator Encrypt

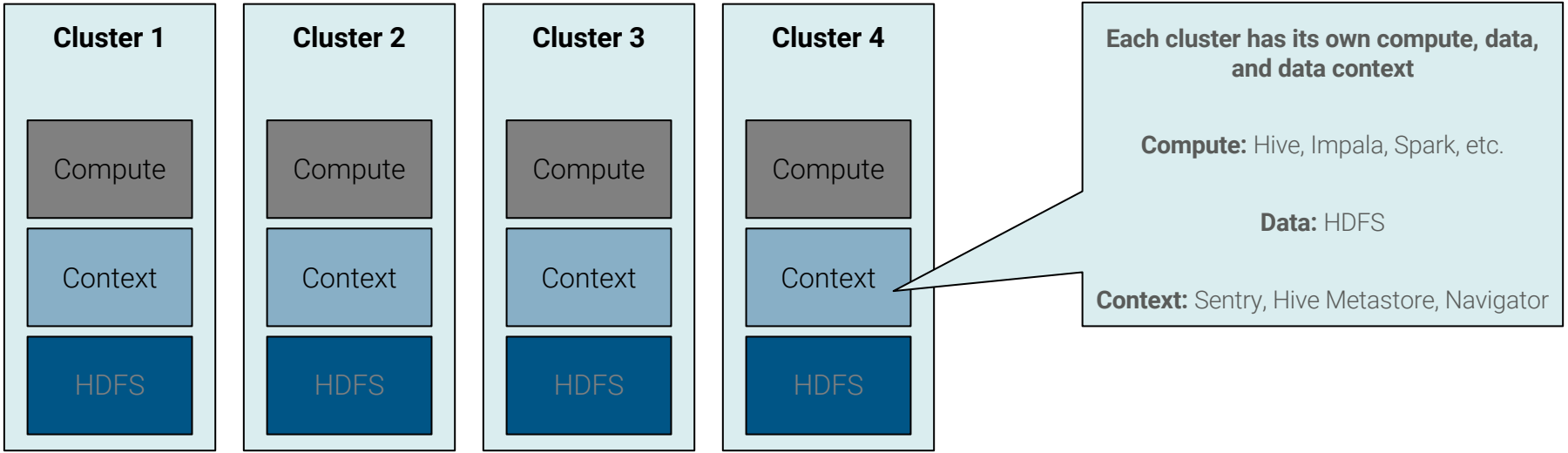
Full-stack encryption and governance



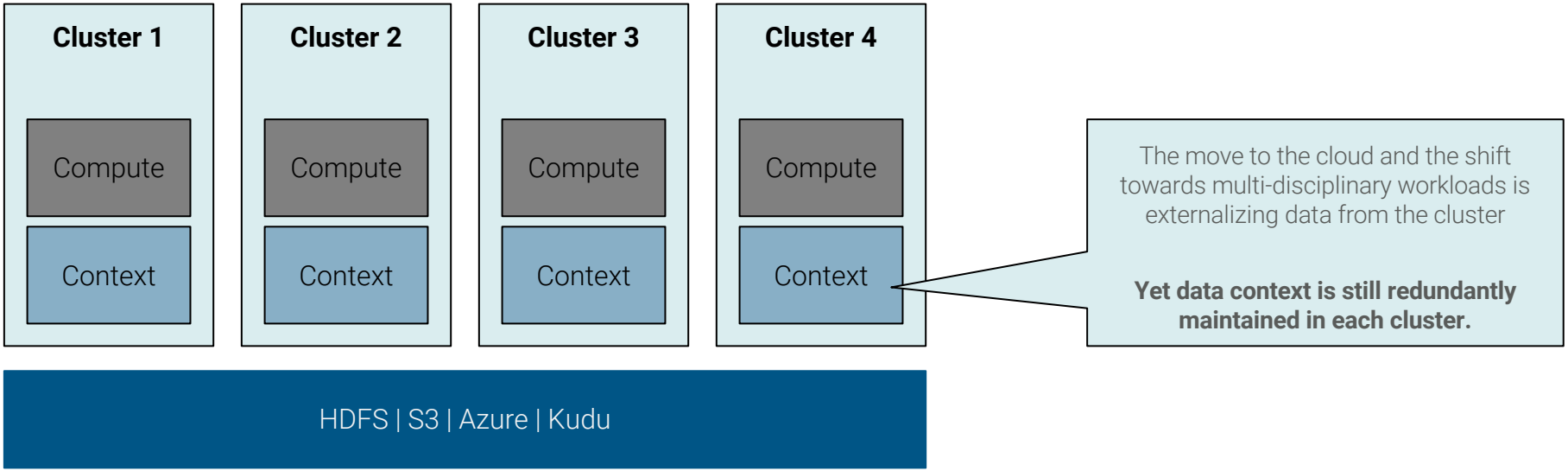




Data context in the early Hadoop years

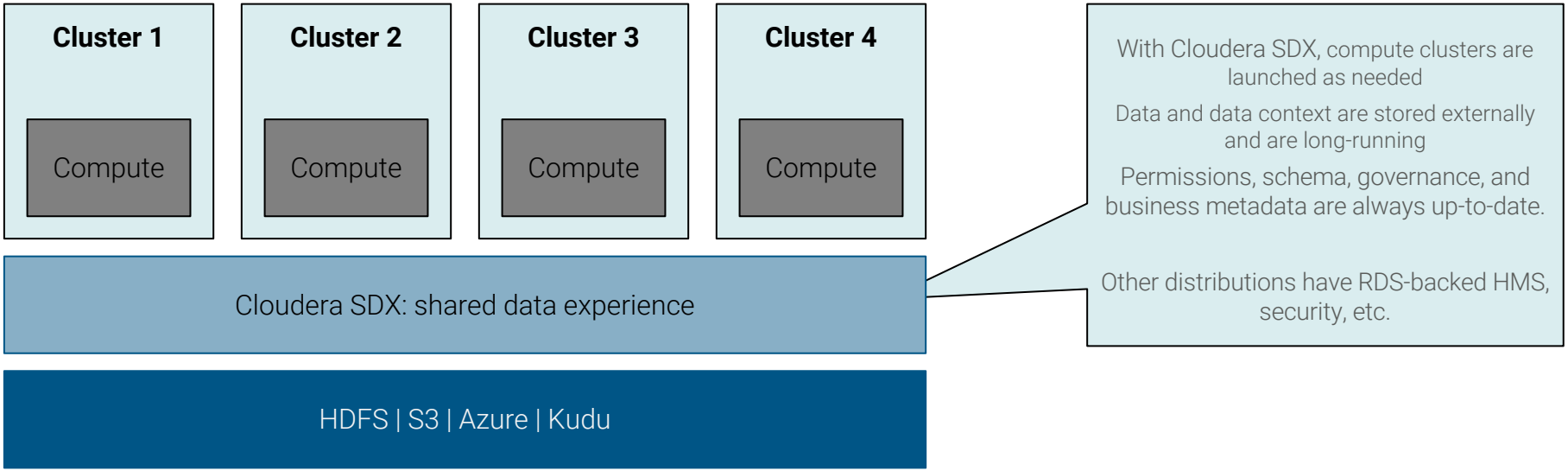


Data context without shared data context
A synchronization nightmare



Data context with Cloudera SDX

Always up-to-date, always in sync



Cloudera SDX solves the problem of redundant data context, which can be a huge pain and security risk for both administrators and users:

- **Best case:** synchronizing data context across clusters is resource-intensive, time-consuming and error-prone
- **Worst case:** access permissions, schemas, and classifications are inconsistent across clusters, thereby frustrating end users and increasing the risk of a security breach

Demo

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Questions

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Final Thoughts

Compliance

- We have shown how an EDH environment can be secured end-to-end
- Is this enough to be compliant?
  - PCI DSS, HIPAA, GDPR
  - Internal compliance – PII data handling
- All of the security features discussed (and others not covered because of time) are enough to cover technical requirements for compliance
- However, compliance also requires additional **people** and **process** requirements
- Cloudera has worked with customers to achieve PCI DSS compliance as well as others – **you can do it too!**

Public Cloud Security

- Many Hadoop deployments occur in the public cloud
- Security considerations presented today all still apply
- Complementary to native cloud security controls

- **Cloudera blog post - How-to: Deploy a secure enterprise data hub on AWS**
- http://blog.cloudera.com/blog/2016/05/how-to-deploy-a-secure-enterprise-data-hub-on-aws/

Looking Ahead

- The Hadoop ecosystem is vast, and it can be a daunting task to secure everything
- Understand that **no system is completely secure**
- However, the proper security controls coupled with regular reviews can **mitigate** your exposure to threats and vulnerabilities
- Pay attention to new components in the stack, as these components often **do not** have the same security features in place
  - Kafka only recently added wire encryption and Kerberos authentication
  - Spark only recently added wire encryption
  - Many enterprises were using both of these in production before those features were available!

Strata
DATA CONFERENCE

strataconf.com
#StrataData

PRESENTED BY
O'REILLY cloudera

Final Questions?

Thank you!