Getting ready for GDPR and CCPA

Securing and governing hybrid, cloud,
and on-premises big data deployments

Strata
DATA CONFERENCE

Your Speakers

- **Lars George,** Principal Solutions Architect, Okera
- **Ifi Derekli,** Senior Solutions Engineer, Cloudera
- **Mark Donsky,** Senior Director of Products, Okera
- **Michael Ernest,** Senior Solutions Architect, Okera

Format

- Five sections
- Each section:
  - Introduces a security concept
  - Shows how to enable
  - Demonstrates the function
- **Please hold questions until the end of each section**
- Short break in the middle

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

Agenda

- Introduction – Lars
  - Authentication – Lars
  - Authorization – Ifi
  - Wire Encryption – Michael
  - Encryption-at-rest – Michael
  - Data Governance & Emerging Regulation – Mark
- Final Thoughts – Mark

Introduction

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 ransomware

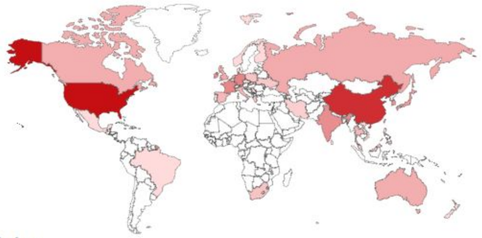
SHODAN | cloudera | Explore | Enterprise Access | Contact Us

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>
```

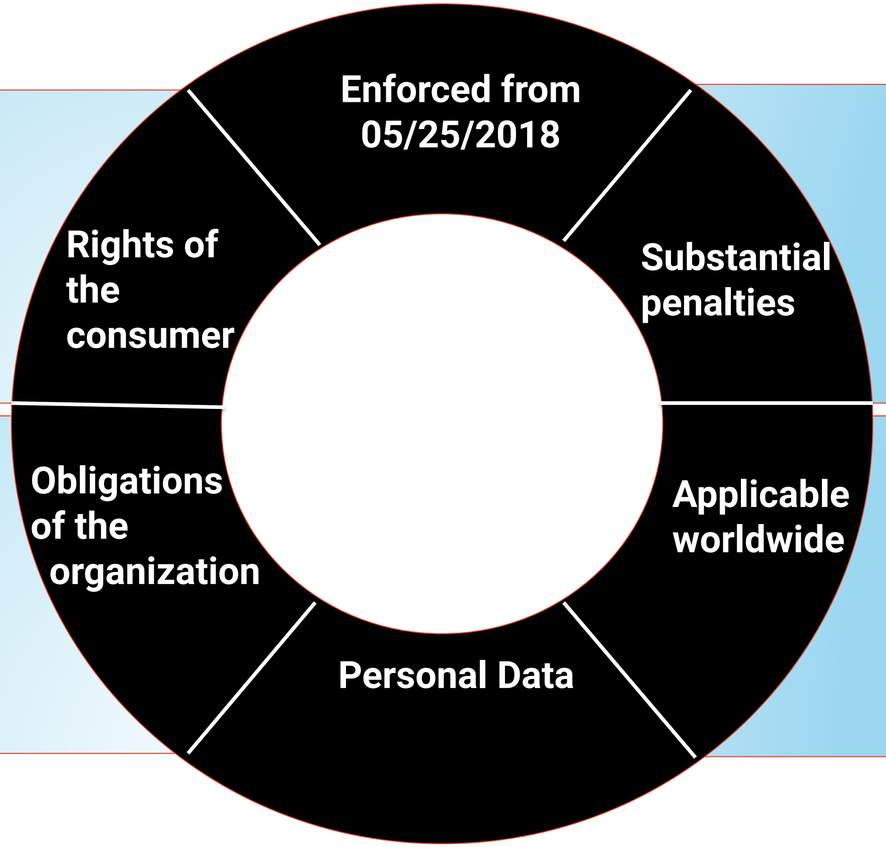
Basic Networking Checks

- 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 using {AWS, Azure, GCE}, check networking configuration

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

Questions?

Strata
DATA CONFERENCE

Authentication

Lars George
Principal Solution Architect
Okera

Strata
DATA CONFERENCE

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

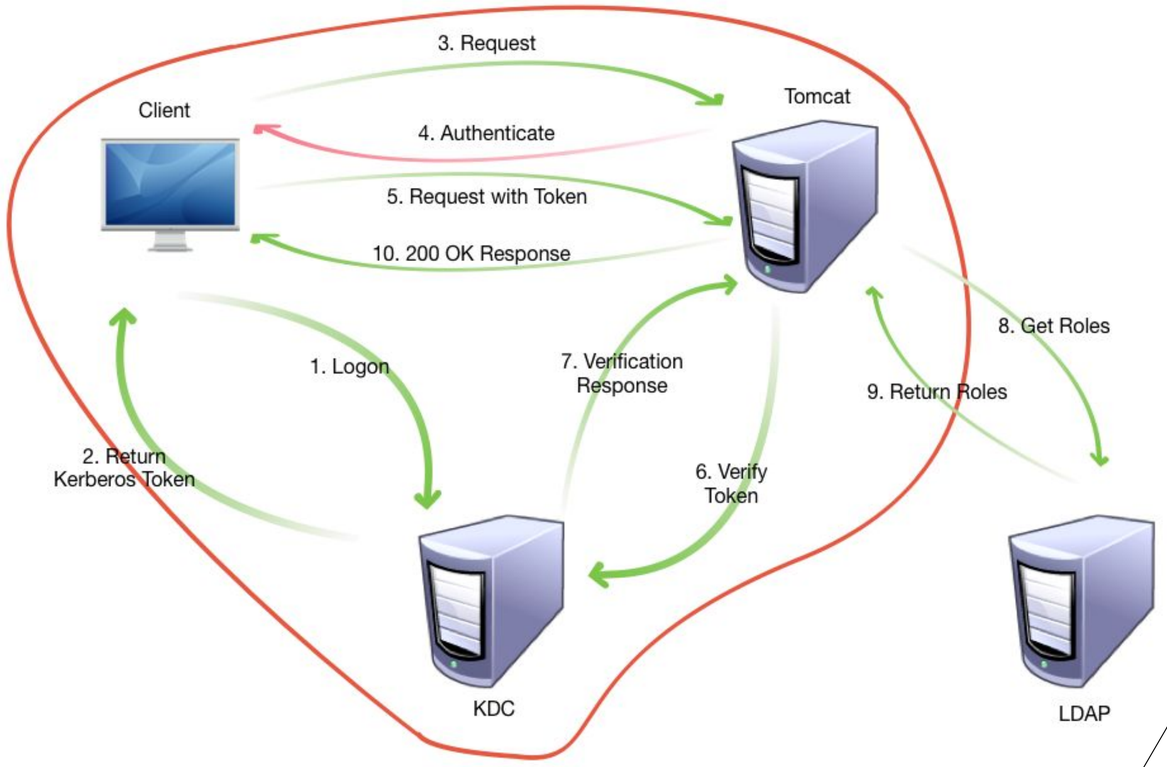
- In Hadoop, identity largely associates with **username**
- Using a common source of identity is paramount

Identity Sources

- Individual Linux servers use `/etc/passwd` and `/etc/group`
  - Not scalable and prone to **errors**
- Kerberos Key Distribution Center (KDC) Principals
  - Only stores users, no group or other related details
- LDAP + Kerberos
  - 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
- Other notable technologies
  - OpenID Connect (OIDC), SAML
  - GSS-API, SASL, SPNEGO
  - OAuth 2.0, JWT

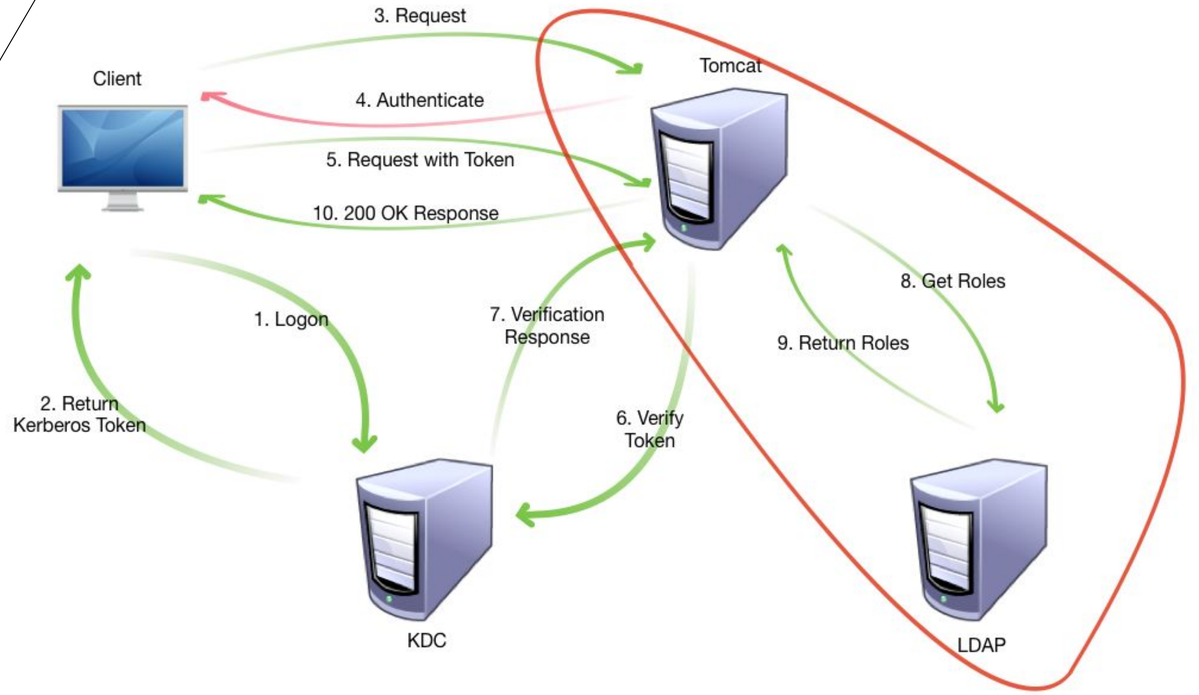


Authentication

Kerberos

LDAP

Authorization

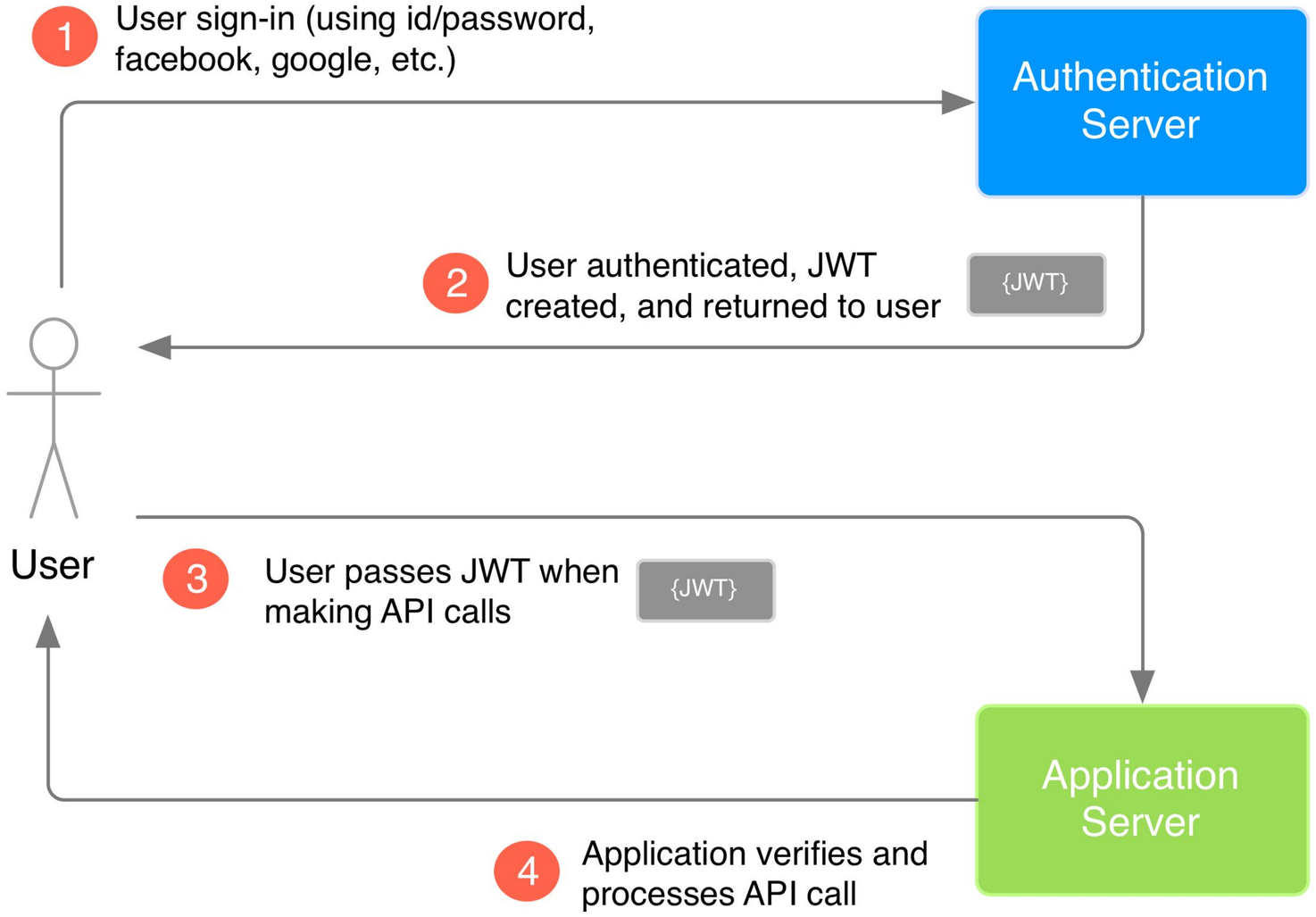


Source: https://dzone.com/articles/introduction-to-kerberos-for-managers

Tokens/Tickets vs Direct Authentication

- At scale, clusters can act as a distributed denial-of-service (DDOS) attack
- Better to authenticate once and receive a token for subsequent access to services
- Renew tokens based on corporate policies
- Common for intra-network (Kerberos tickets with SASL) and extranet authentication (SAML, OAuth with JWT)

- BUT… tokens have longer lifetime. Requires matching authorization!



Source: https://medium.com/vandium-software/5-easy-steps-to-understanding-json-web-tokens-jwt-1164c0adfcec

Hadoop's Default "Authentication"

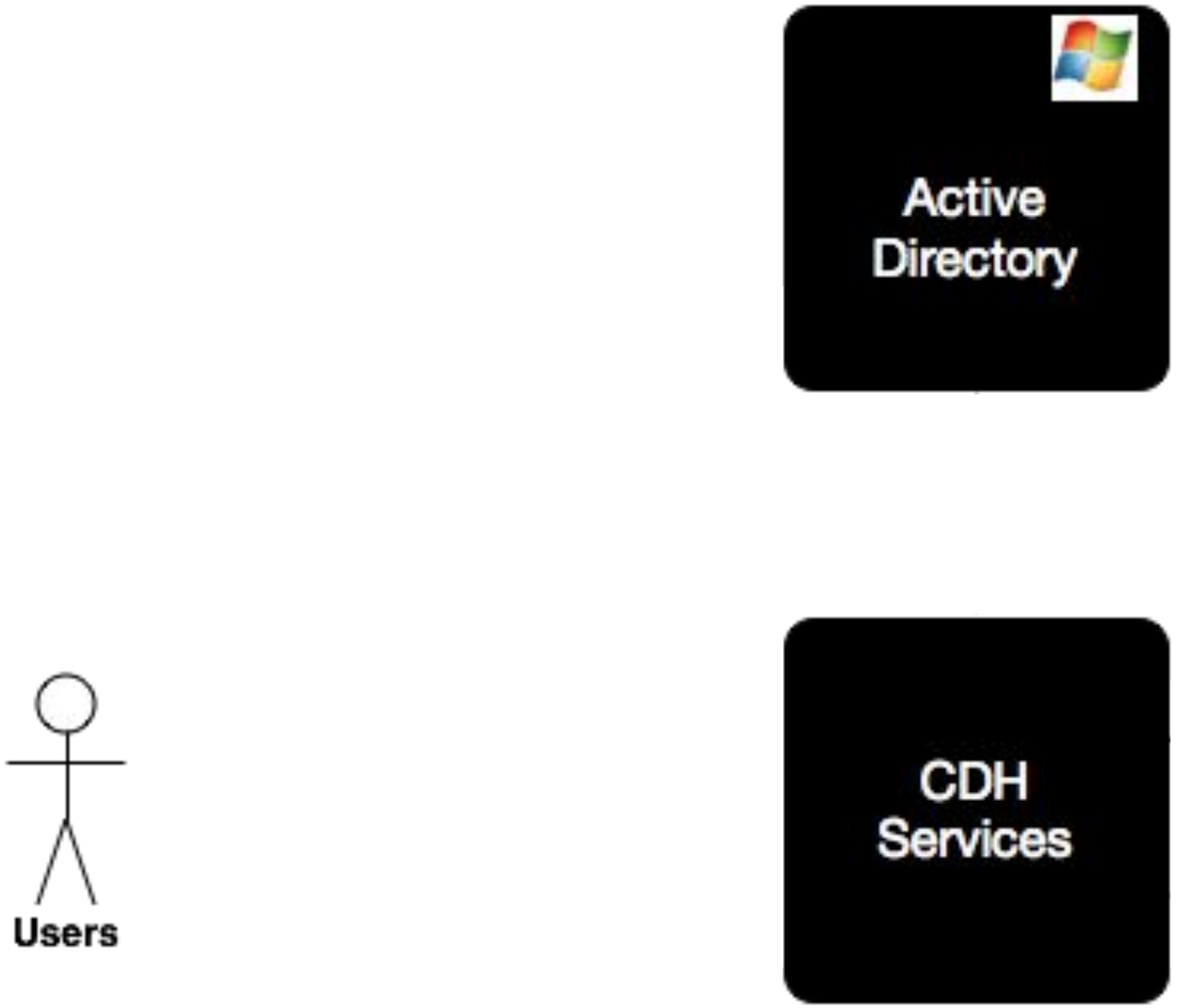
- Out of the box, Hadoop "authenticates" users by simply believing whatever username you tell it you are
- This includes telling Hadoop you are the **`hdfs`** user, a **superuser**!

```
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

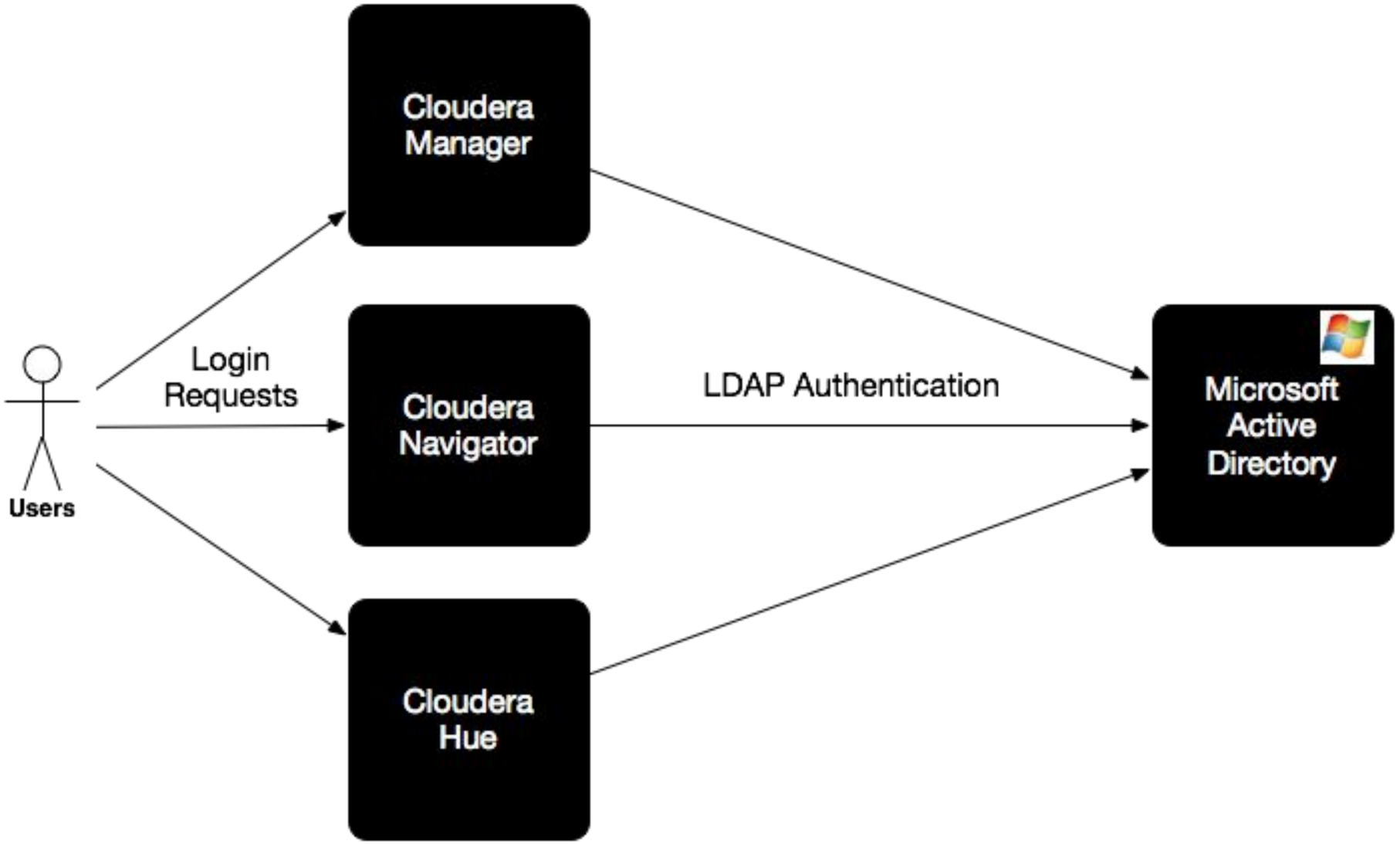
Kerberos Authentication



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
- 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



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

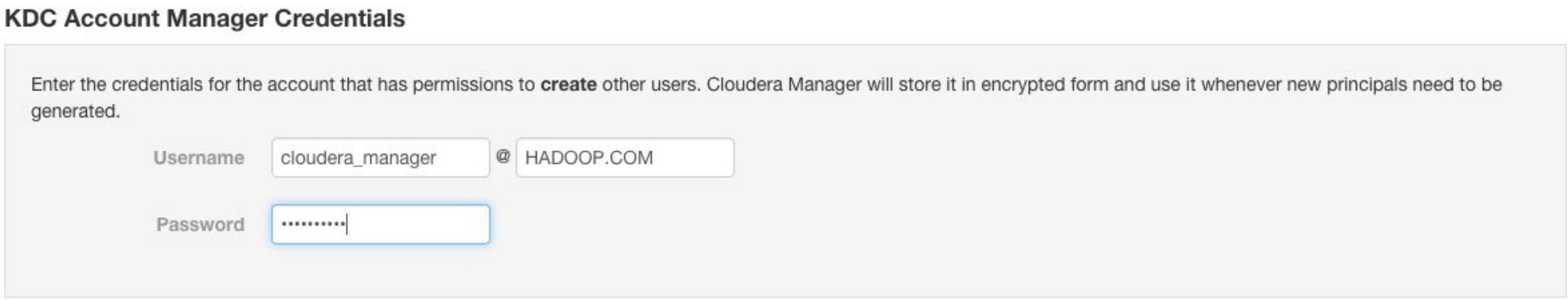
Active Directory Account Prefix

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!

Questions?

Strata
DATA CONFERENCE

Authorization

Ifi Derekli
Senior Solutions Engineer
Cloudera

Authorization - GDPR

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

- **Data minimization**
  - Personal data shall be:
    (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed
  - e.g. Data Scientist should only see masked PII data

- **Integrity and confidentiality**
  - Personal data shall be:
    (f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing [...]
  - e.g. US employee can only see data of US customers
  - e.g. EU Analyst can only see data from EU citizens that have given consent

Authorization - Agenda

- Authorization – Overview
- Configuration Stronger Authorization
- Authorization tools
  - Apache Sentry
  - Apache Ranger
  - Commercial Products (Okera)
- **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)
  - Attribute-based access control (ABAC)
- Many different authorization mechanisms in Hadoop components

Authorization in Hadoop

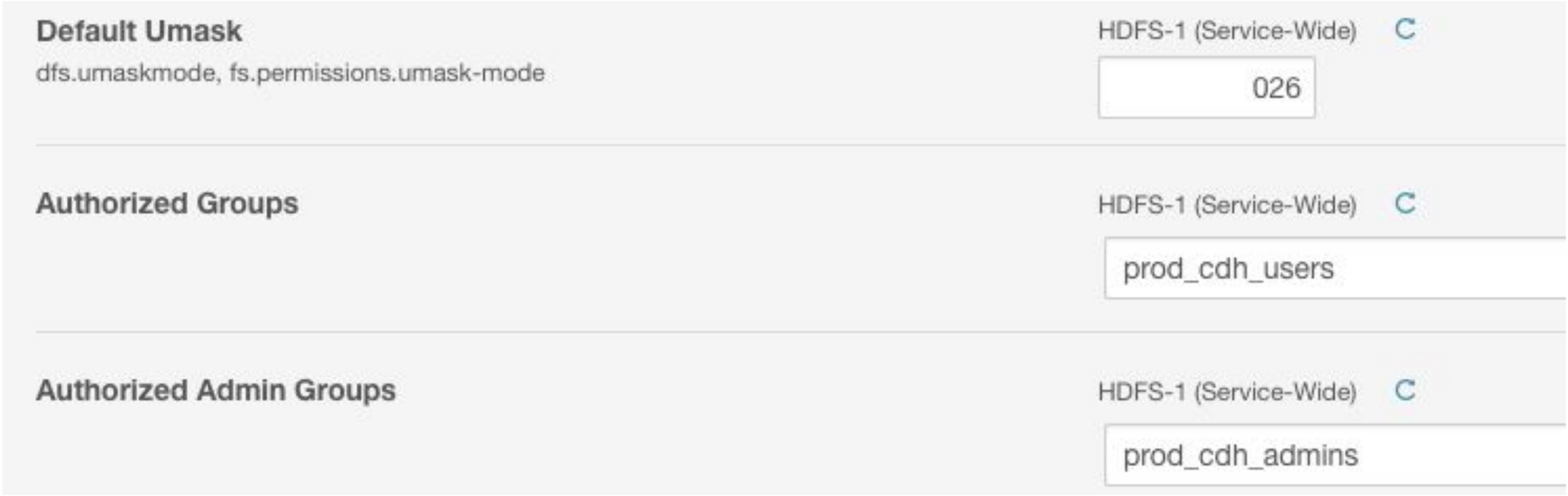
- HDFS file permissions (POSIX ‘rwx rwx rwx’ style)
- Yarn job queue permissions
- Ranger (HDFS / HBase / Hive / Yarn / Solr / Kafka / NiFi / Knox / Atlas)
- Atlas ABAC
- Sentry (HDFS / Hive / Impala / Solr / Kafka)
- Cloudera Manager RBAC
- Cloudera Navigator RBAC
- Hadoop KMS ACLs
- HBase ACLs
- Commercial authorization tools (e.g. Okera)
- 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.)

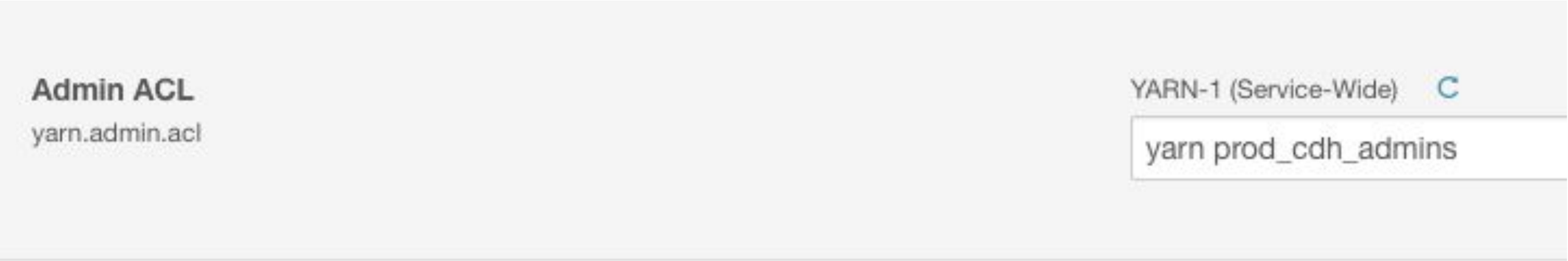
Configuring HDFS Authorization

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



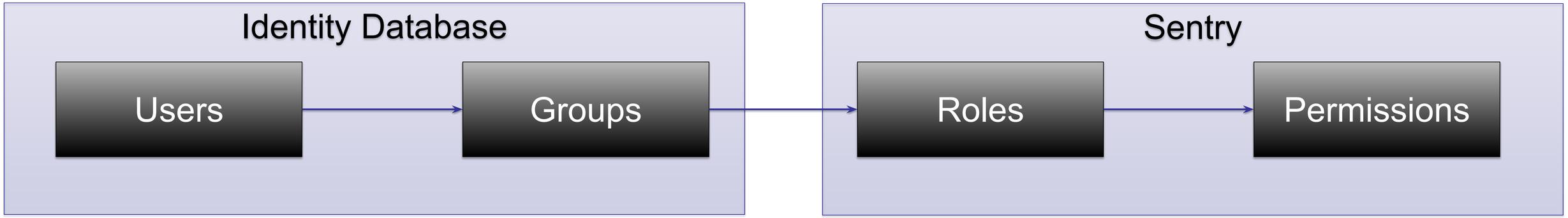
Configuring Yarn Authorization

- Setup the YARN admin ACL

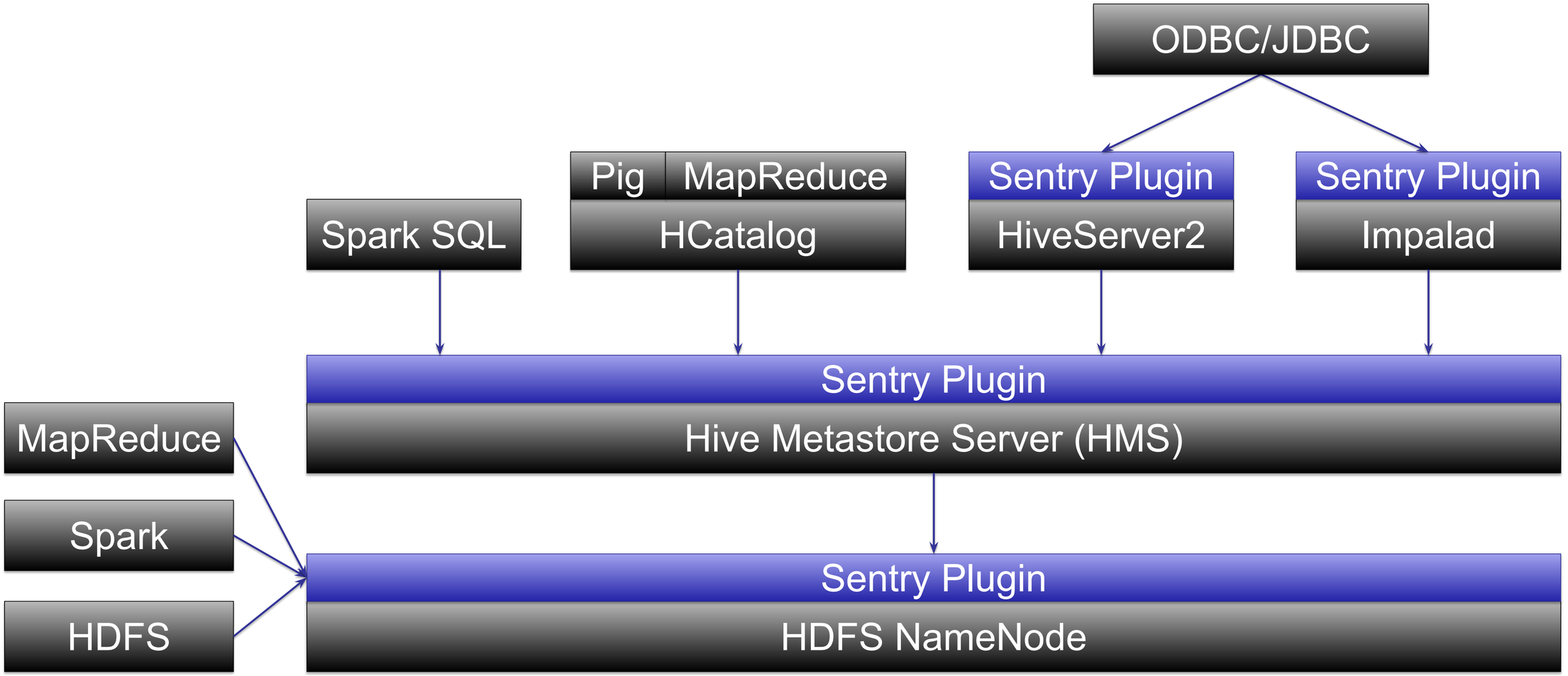


Apache Sentry

- Provides centralized RBAC for several components
  - **Hive / Impala:** database, table, view, column
  - **HDFS:** file, folder (auto-sync with hive/impala)
  - **Solr:** collection, document, index
  - **Kafka:** cluster, topic, consumer group

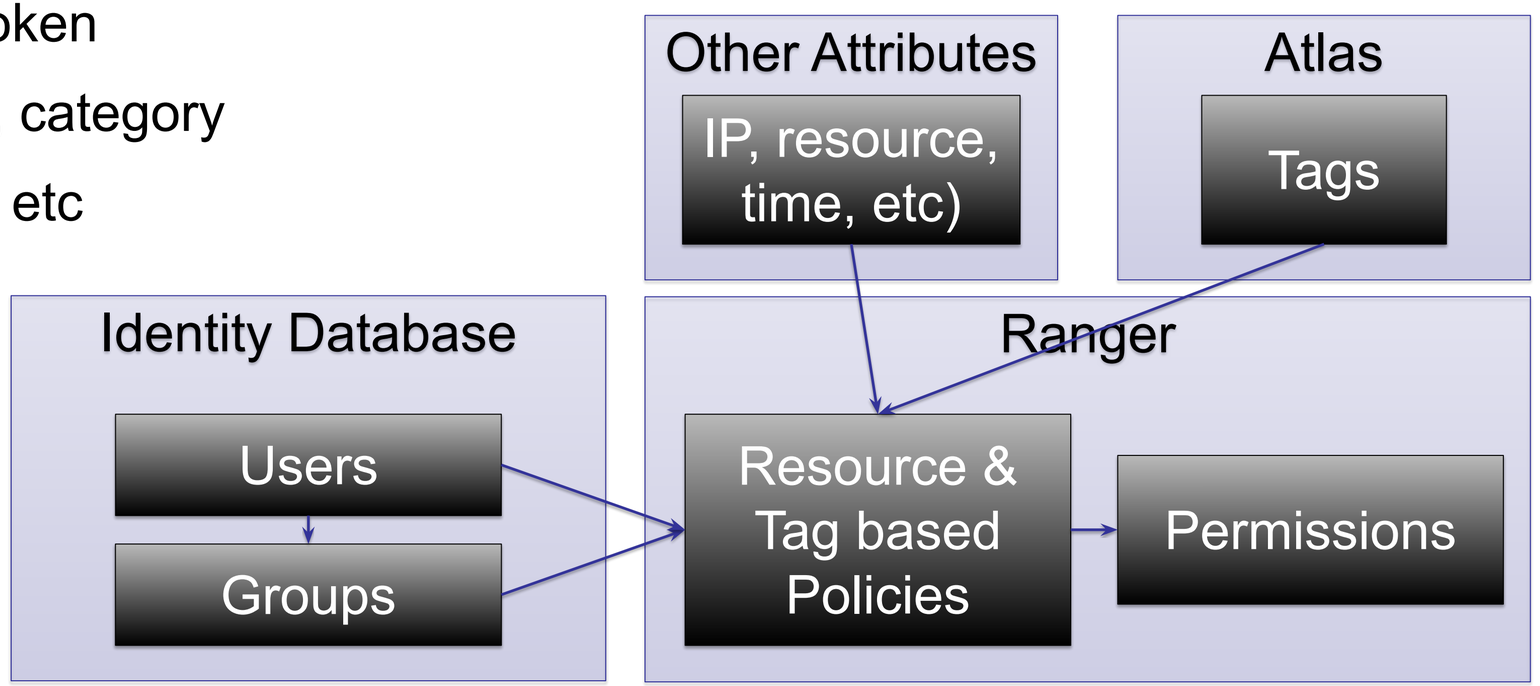


Apache Sentry (Cont.)

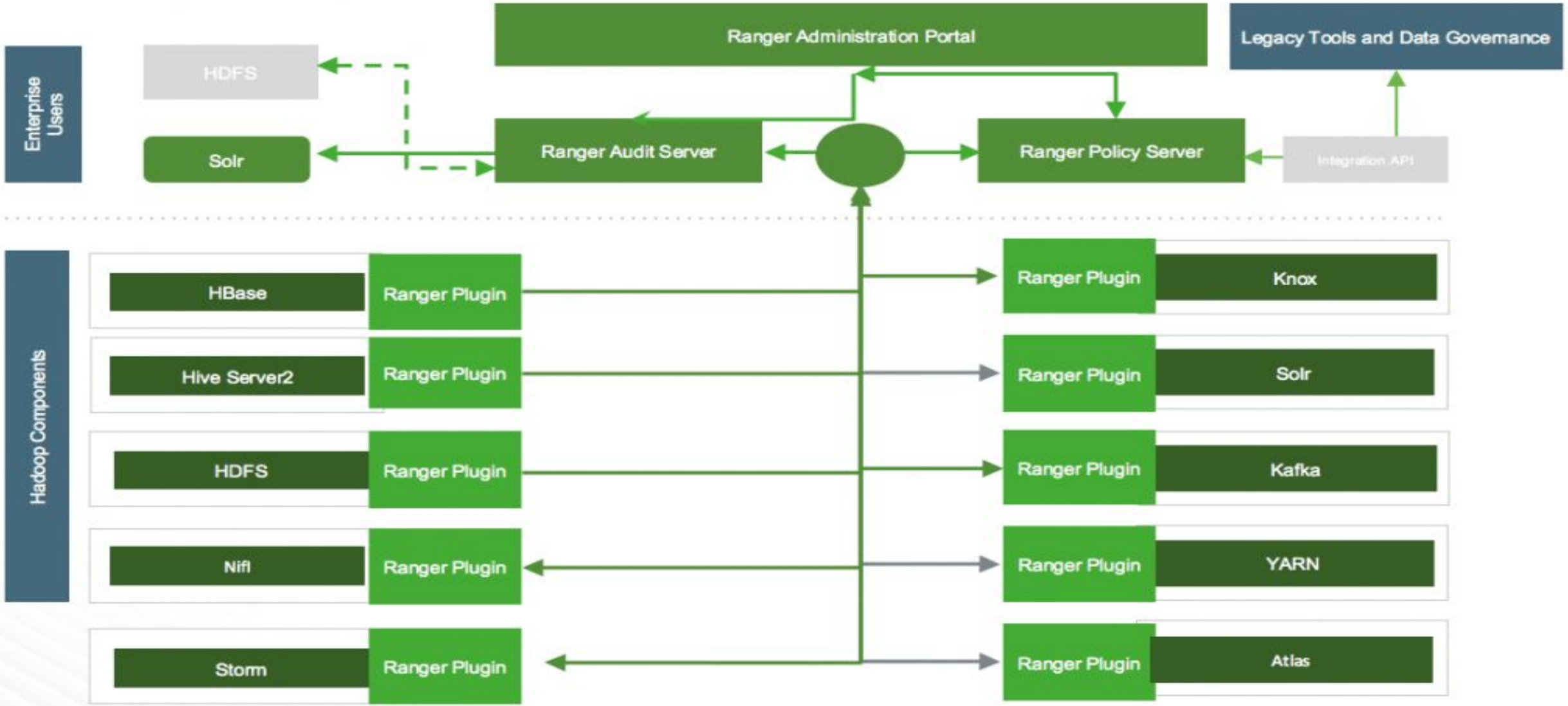


Apache Ranger

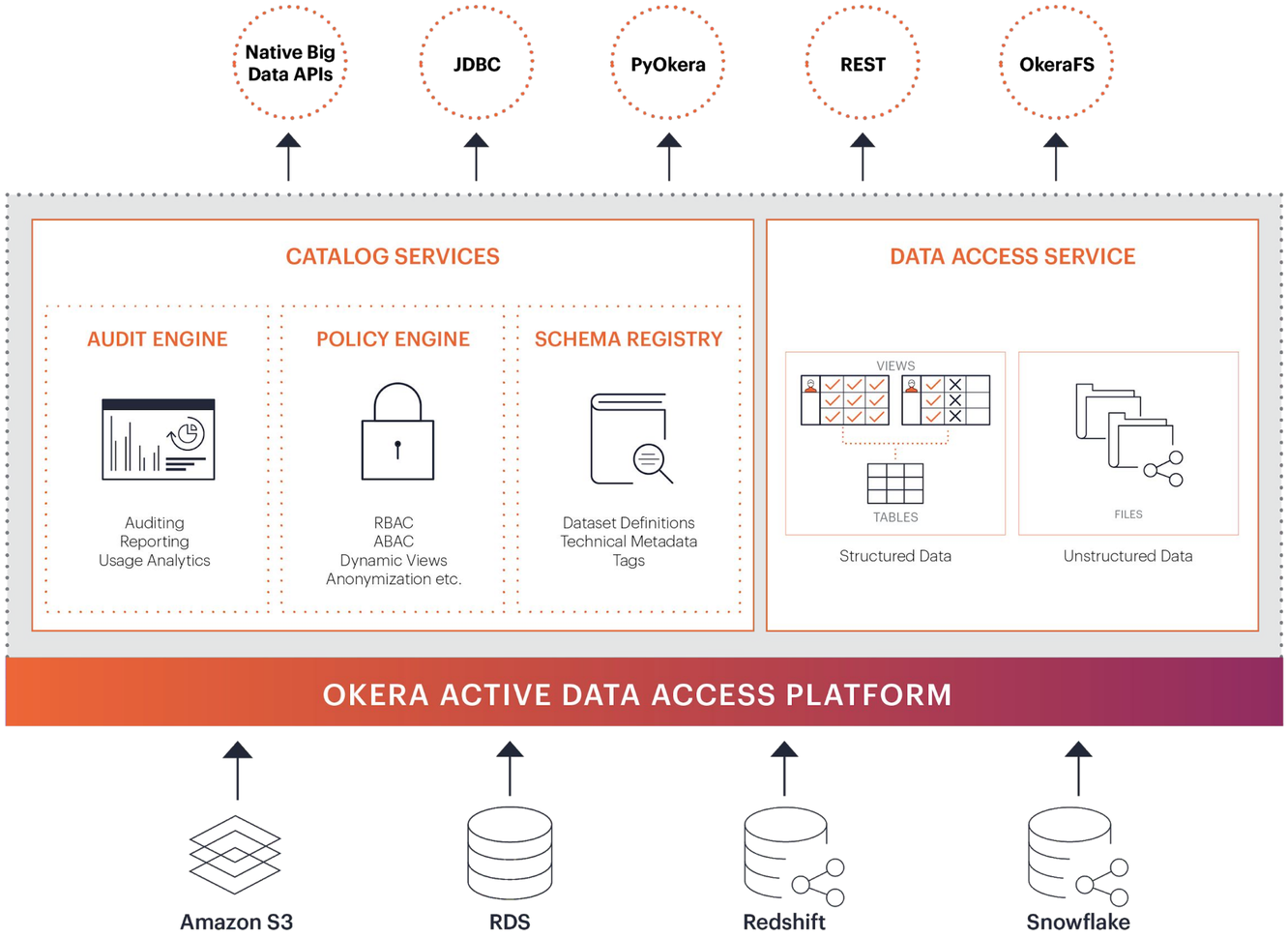
- Provides centralized ABAC for several components
  - **HBase:** table, column-family, column
  - **Hive:** database, table, view, column, udf, row, masking
  - **Solr:** collection, document, index
  - **Kafka:** cluster, topic, delegation token
  - **Atlas:** service, entity, relationship, category
  - **NiFi:** flow, controller, provenance, etc
  - **HDFS:** file, folder
  - **YARN:** queue
- Extensible
- Consistent with NIST 800-162



Apache Ranger (Cont.)



Okera Active Data Access Platform



An active management layer that makes data lakes accessible to multiple workers

- Provides consistent, airtight protection with fine-grained access policies
- Supports all leading analytics tools and data formats
- Introduces no delay or additional overhead

Post Configuration

- HDFS setup with a better umask and service level authorization
- YARN setup with restrictive admin ACLs
- Hive, Impala, Solr, Kafka, etc setup with access control

- **DEMO:** No more default authorization holes!
  - US Analyst can only access US data, masked PII
  - EU HR can only access EU data that have given consent, including PII
  - US intern can only access US data, masked PII, from VPN, for two months

Authorization - Summary

- HDFS file permissions (POSIX ‘rwx rwx rwx’ style)
- Yarn job queue permissions
- Ranger (HDFS / HBase / Hive / Yarn / Solr / Kafka / NiFi / Knox / Atlas)
- Atlas ABAC
- Sentry (HDFS / Hive / Impala / Solr / Kafka)
- Cloudera Manager RBAC
- Cloudera Navigator RBAC
- Hadoop KMS ACLs
- HBase ACLs
- Commercial authorization tools (e.g. Okera)
- etc.

Questions

Encryption of Data in Transit

Michael Ernest
Solution Architect
Okera

Strata
DATA CONFERENCE

Encryption in Transit - GDPR

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

- **Integrity and confidentiality**

Agenda

- Why encrypting data in transit matters
- Key Technologies used with Hadoop
  - **S**imple **A**uthentication & **S**ecurity **L**ayer (SASL)
  - **T**ransport **L**ayer **S**ecurity (TLS)
- For each technology:
  - Without it, network snoopers can see data in transit
  - How it works
  - How to enable it
  - How to demonstrate it’s working

Why Encrypt Data in Transit?

- Firewalls and other perimeter defenses mitigate some risk
  - But some attacks originate inside the network
- Data passing on the wire isn’t protected by authentication or authorization controls
- Industry/regulatory standards for protecting transmitted, sensitive data

Example

- Transfer data into a cluster
- Simple file transfer: “hadoop fs –put”
- A snooper can see file content in the clear

Two Encryption Technologies

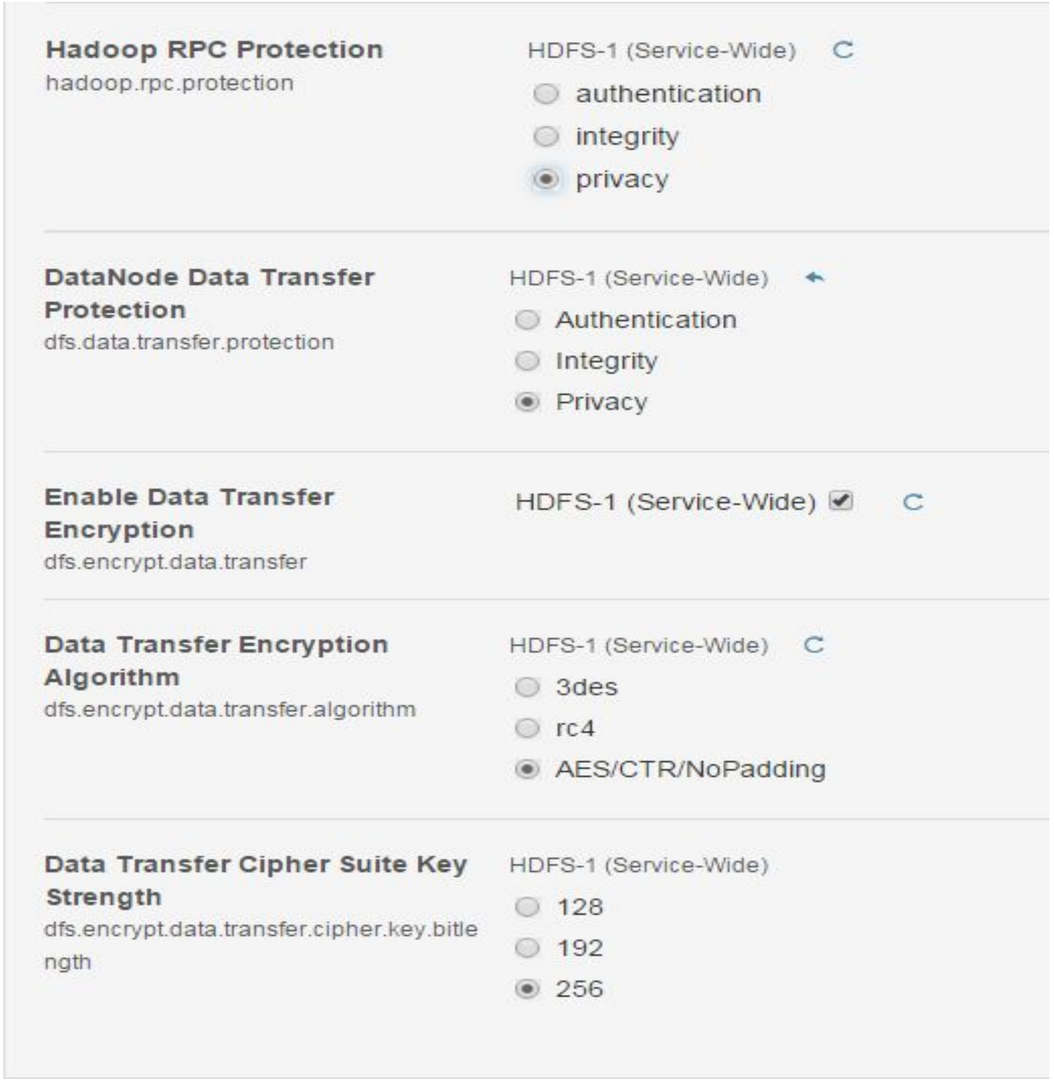
- SASL “confidentiality” or “privacy” mode
  - Encryption on RPC
  - Encryption on block data transfers
  - Encryption on web consoles (except HttpFS and KMS)
- TLS – Transport Layer Security
  - Used for everything else

SASL Defined

- A framework for negotiating authentication 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
- Enabling them requires a property change and restart.

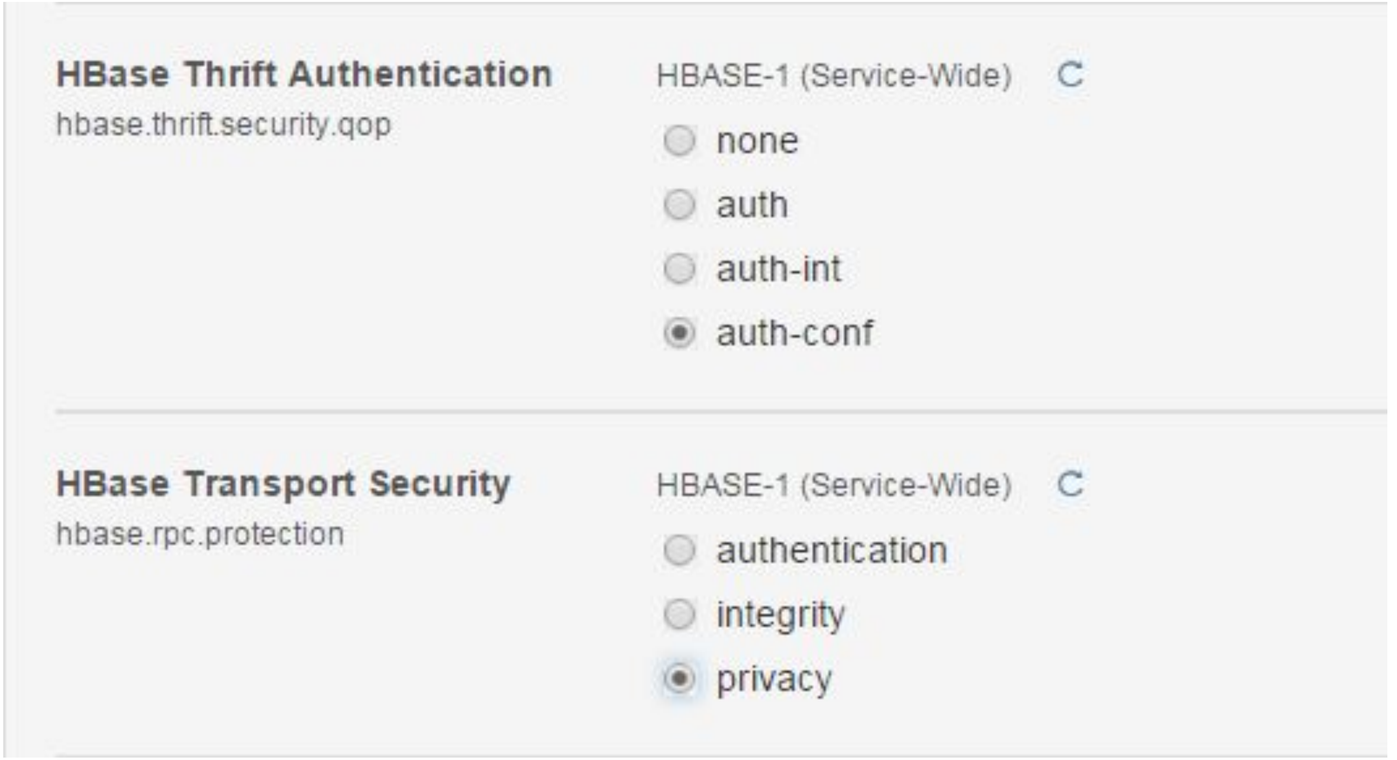
SASL Encryption - HDFS

- Kerberos manages the authentication
- For HDFS
  - Hadoop RPC Protection
  - Datanode Data Transfer Protection
  - Enable Data Transfer Encryption
  - Data Transfer Encryption Algorithm
  - Data Transfer Cipher Suite Key Strength



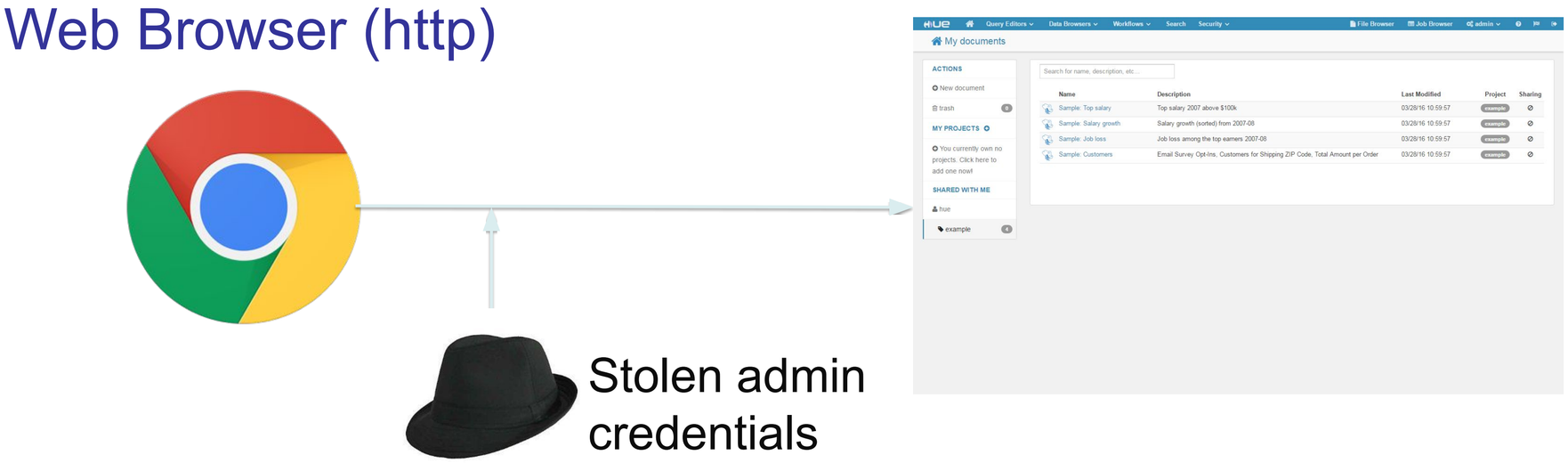
SASL Encryption - HBase

- HBase
  - HBase Thrift Authentication
  - HBase Transport Security



TLS

- Transport Layer Security
  - The successor to SSL – Secure Sockets Layer
  - We often say “SSL” where TLS is actually used.
  - TLS supports HTTPS-configured websites



TLS - Certificates

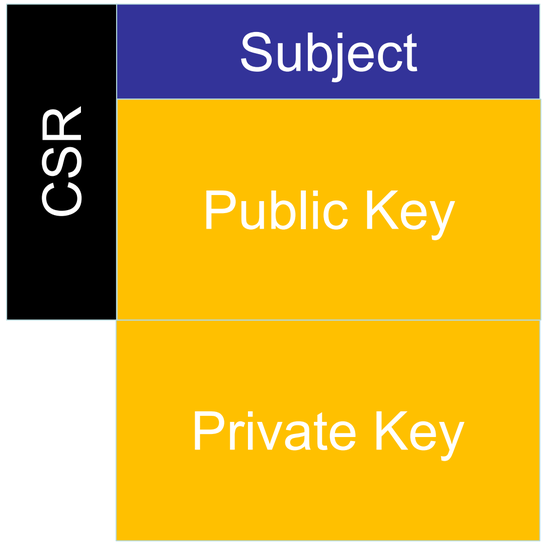
- TLS uses X.509 certificates to authenticate the bearer
- Hadoop best practice: a unique certificate on each cluster node
- Certificates:
  - Cryptographically prove the bearer's identity
  - The certificate's signer (issuer) "vouches" for the bearer.
  - Content includes: subject identity, issuer identity, valid period
  - Many other attributes as well, such as "Extended Key Usage"
  - Let's inspect an https site certificate

TLS – Certificate Authorities

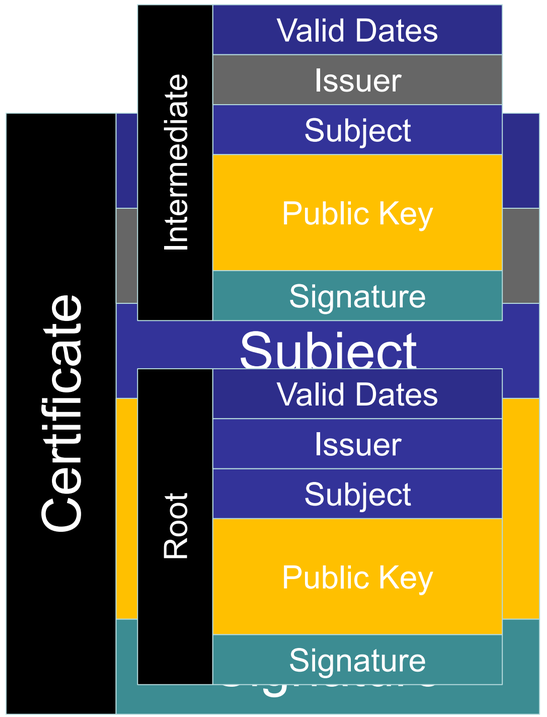
- You can generate & sign your own certificate
  - Useful for testing: fast and cheap
- Internal Certificate Authority
  - Some department everyone at a company trusts
  - Active Directory Certificate Services is widely used
  - To make it work, clients must also trust it
  - Useful for enterprise deployments: good-enough, cheap
- Public Certificate Authority
  - Widely-known and trusted: VeriSign, GeoTrust, Symantec, etc.
  - Useful for internet-based applications such as web retail
  - Strong, in some cases fast

Signing a Certificate

You



Certificate Authority



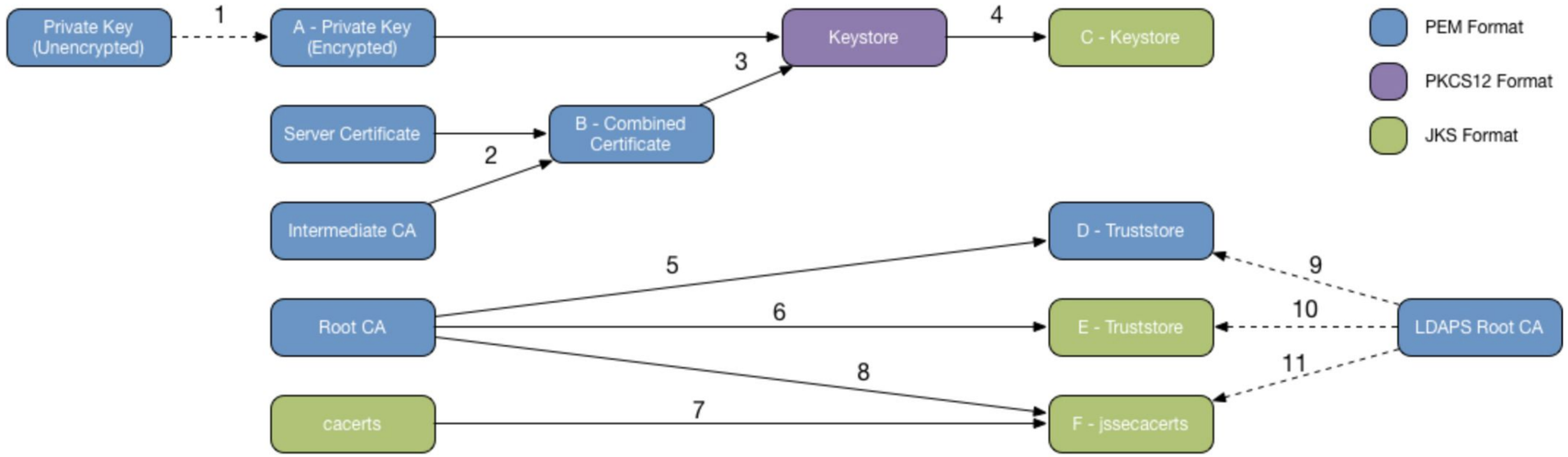
TLS – Certificate File Formats & Storage

- Two formats Hadoop cluster services need to store certificates and keys
- Privacy Enhanced Email (PEM)
  - Designed for use with text-based transports (e.g., HTTP servers)
  - Base64-encoded certificate data
- Java KeyStore (JKS)
  - Designed for use with JVM-based applications
  - The JVM keeps its own list of trusted CAs (for when it acts as a client)
- Each cluster node needs keys/certificates kept in both formats

TLS – Key Stores and Trust Stores

- Keystore
  - Used when serving a TLS-based client request
  - JKS: Contains private keys and host certificate; passphrase-protected
  - PEM: Usually contains one certificate file, one private key file (passphrase-protected)
- Truststore
  - Used when requesting a service over TLS
  - Contains CA certificates that the client trusts
  - JKS: Password-protected, used only as an integrity check
  - PEM: Same idea, no password
  - One system-wide store for both PEM and JKS formats

TLS – Key Stores and Trust Stores



Certificate Preparation Steps:

1) (If needed) openssl rsa -in cert.key -out `hostname -f`.key -aes256 -passout pass:somepass
2) cat cert.pem int-CA.pem > `hostname -f`.pem
3) openssl pkcs12 -export -inkey `hostname -f`.key -passin pass:somepass -in `hostname -f`.pem -out `hostname -f`.pfx -passout pass:somepass
4) keytool -importkeystore -srcstoretype PKCS12 -srckeystore `hostname -f`.pfx -srcstorepass somepass -destkeystore `hostname -f`.jks -deststorepass somepass
5) cp root-CA.pem truststore.pem
6) keytool -importcert -file root-CA.pem -keystore truststore.jks -storepass changeit -alias root-CA
7) cp $JAVA_HOME/jre/lib/security/cacerts $JAVA_HOME/jre/lib/security/jssecacerts
8) keytool -importcert -file root-CA.pem -keystore $JAVA_HOME/jre/lib/security/jssecacerts -storepass changeit -alias root-CA
9) (If needed) cat ldaps-CA.pem >> truststore.pem
10) (If needed) keytool -importcert -file ldaps-CA.pem -keystore truststore.jks -storepass changeit -alias ldaps-CA
11) (If needed) keytool -importcert -file ldaps-CA.pem -keystore $JAVA_HOME/jre/lib/security/jssecacerts -storepass changeit -alias ldaps-CA

Final Objects Used By Cloudera EDH:

A - The private key for the server certificate, in PEM format
B - The server certificate and intermediate CA certificate, in PEM format
C - The server's keystore in JKS format, containing the private key, certificate, and intermediate CA certificate
D - The truststore in PEM format, containing the root CA certificate, and the LDAPS root CA certificate (if different from root CA)
E - The truststore in JKS format, containing the root CA certificate, and the LDAPS root CA certificate (if different from root CA)
F - The jssecacerts truststore in JKS format, containing all the bundled trusted certs, the root CA certificate, and the LDAPS root CA certificate (if different from root CA)

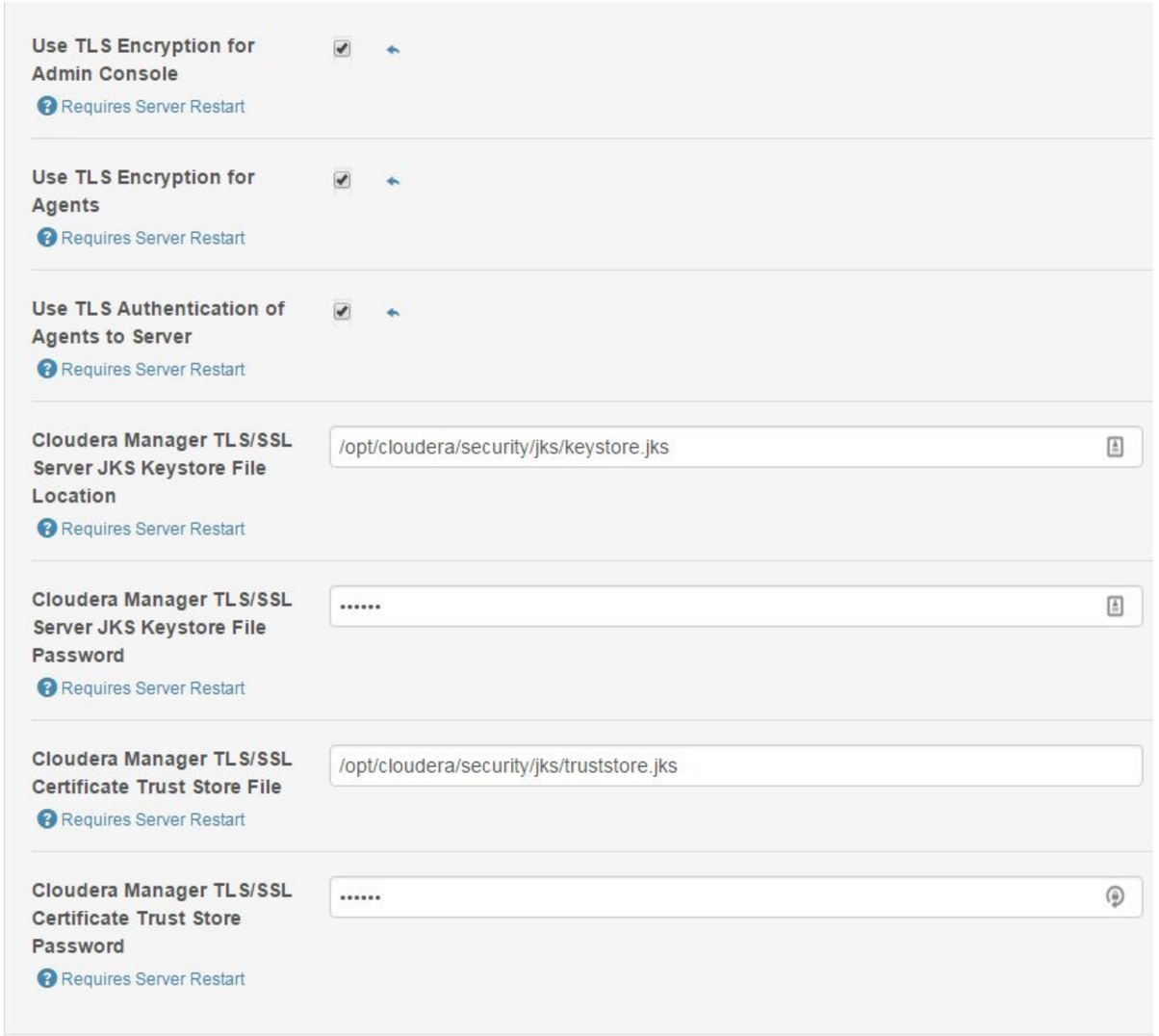
TLS – Securing Cloudera Manager

- CM Web UI - 

- CM Agent -> CM Server communication – three steps to enabling TLS
  - Encrypting but without certificate verification. Akin to clicking on 

  - CM agents verify the CM server's certificate (similar to a web browser)
  - CM server verifies CM agents, known as *mutual authentication*. Each side ensures it's talking to a cluster member
    - This means every node has to have a keystore
  - Used here because agents send (and may request) sensitive operational metadata
    - Consider Kerberos keytabs. You may want TLS in CM before you integrate Kerberos!

Cloudera Manager TLS



CM Agent Settings

- Agent config location: **`/etc/cloudera-scm-agent/config.ini`**

`use_tls`=1 Enable privacy

`verify_cert_file`= full path to CA certificate.pem file One-way

`client_key_file`= full path to private key.pem file

`client_keypw_file`= full path to file containing password for key

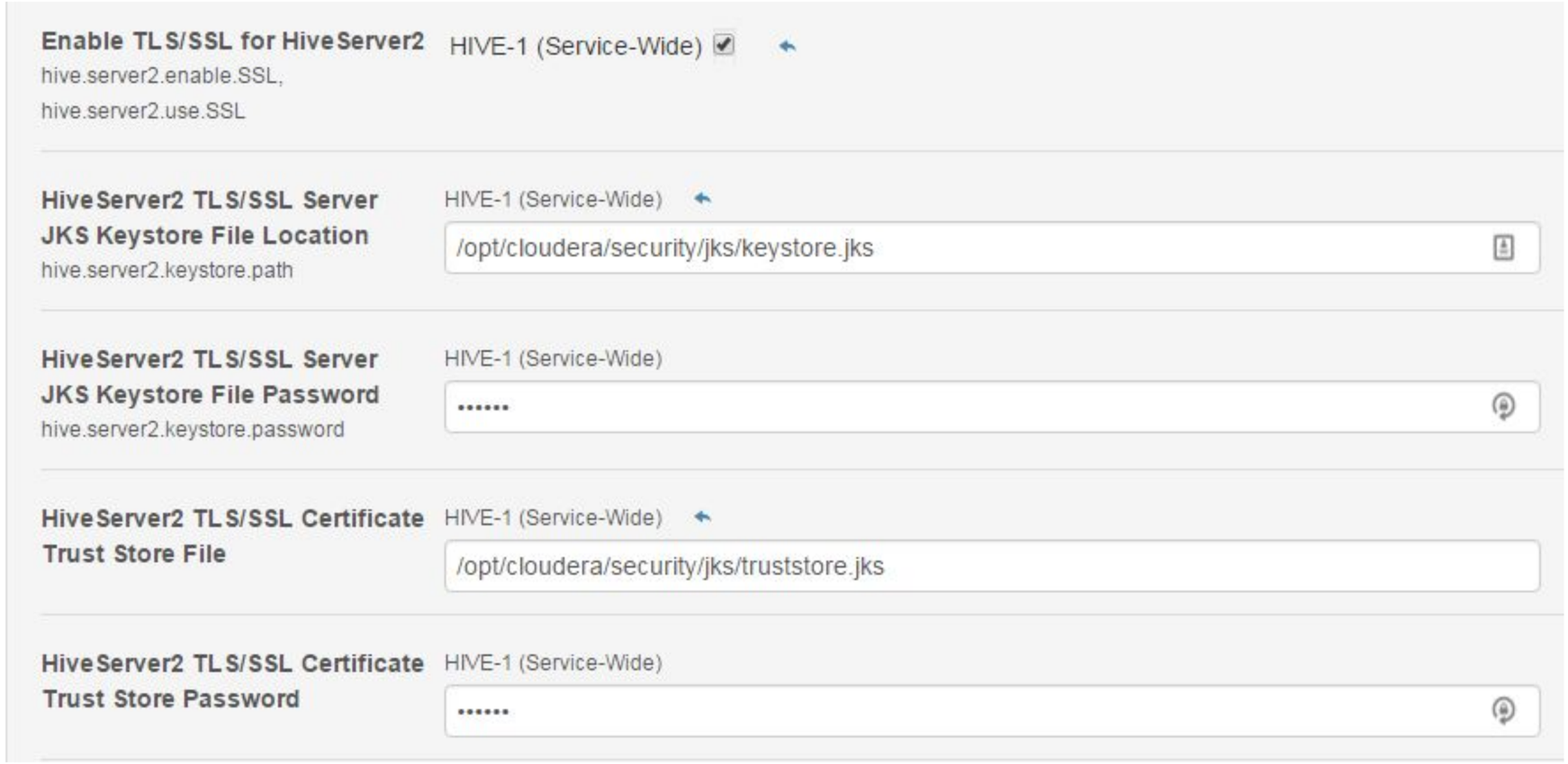
`client_cert_file`= full path to certificate.pem file

Mutual

TLS for CM-Managed Services

- CM expects all certificate-based files to share one location on all machines
  - e.g., /opt/cloudera/security
- Then for each cluster service (HDFS, Hive, Hue, HBase, Impala, …)
  - Find “TLS” in the service’s Configuration tab
  - Check to enable; restart
  - Identify location for keystore and truststore, provide passwords

Hive Example

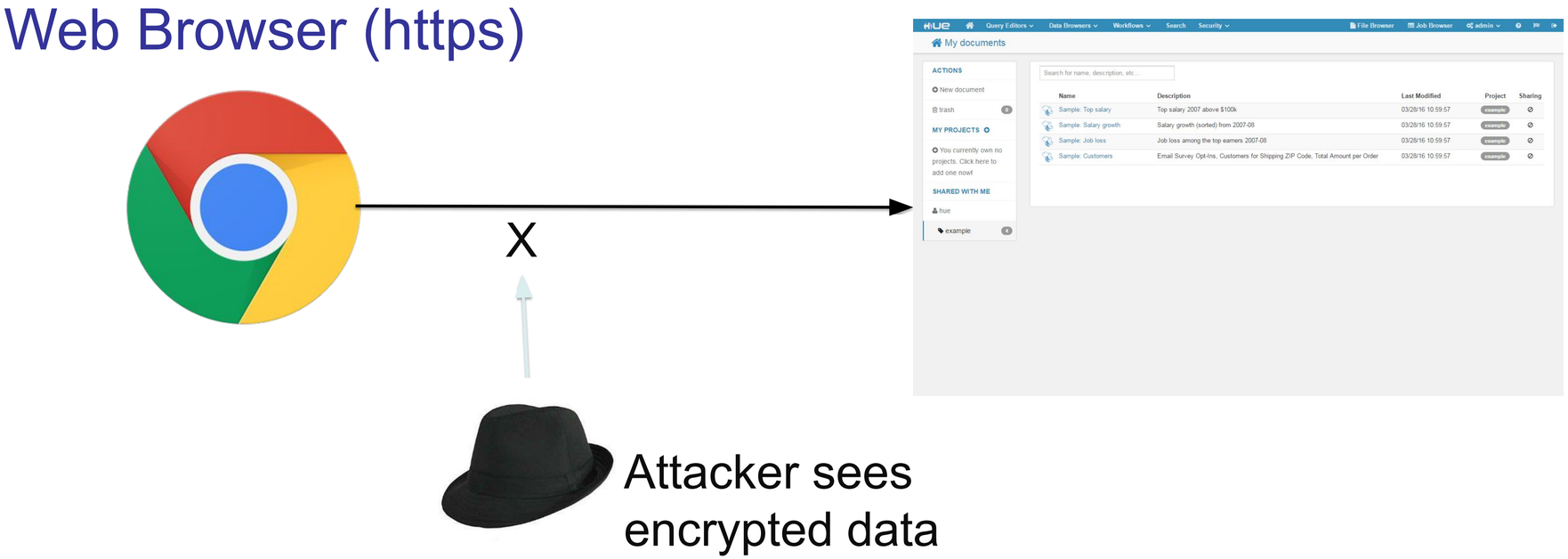


TLS - Troubleshooting

- 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 session shows 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

- Information as it passes from point to point is vulnerable to snooping
- Hadoop uses SASL & TLS for privacy & encryption
- Enabling SASL is straightforward
- Enabling TLS requires certificates for every cluster node

Questions?

Strata
DATA CONFERENCE

HDFS Encryption at Rest

Michael Ernest
Solutions Architect
Okera

Strata
DATA CONFERENCE

Agenda

- Why Encrypt Data
- HDFS Encryption
- Demo
- Questions

Encryption at Rest - GDPR

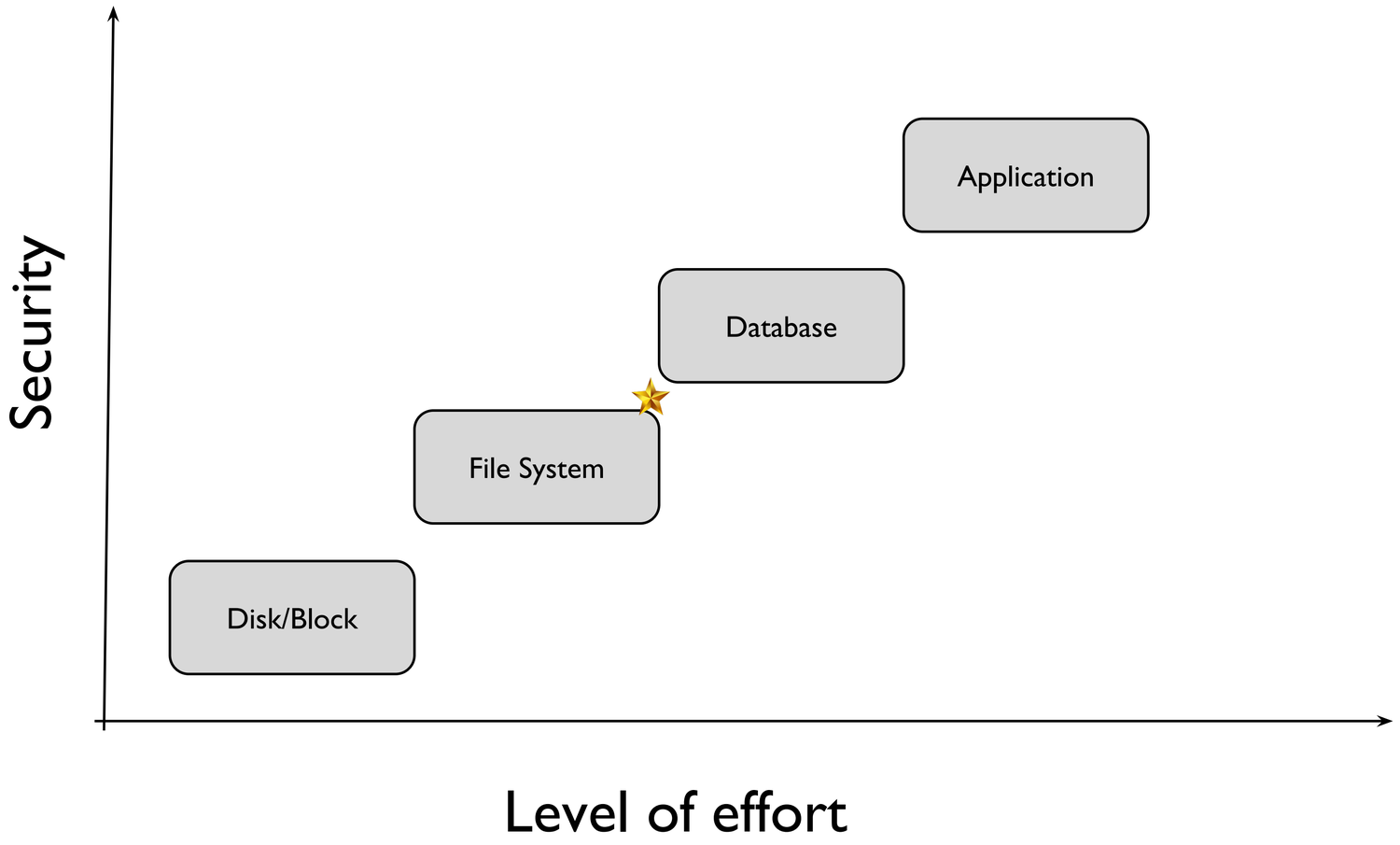
- Broadly underpins **one** of the GDPR Article 5 Principles

- **Integrity and confidentiality**
  - (f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (‘integrity and confidentiality’).

Why encrypt data on disk?

- Many enterprises must comply with
  - GDPR
  - PCI
  - HIPAA
  - National Security
  - Company confidentiality
- Mitigate other security threats
  - Rogue administrators (insider threat)
  - Neglected/compromised user accounts (masquerade attacks)
  - Replaced/lost/stolen hard drives!

Options for encrypting data



Architectural Concepts

- Separate store of encryption Keys
- Key Server
  - External to the cluster
- Key Management Server (KMS)
  - Proxy for the Key Server
  - Part of the cluster
- HDFS Encryption Zone
  - Directory that only stores/retrieves key-encrypted file content
- Encryption/decryption remains transparent to the user
  - No change to the API for putting/getting data

Encryption Zone

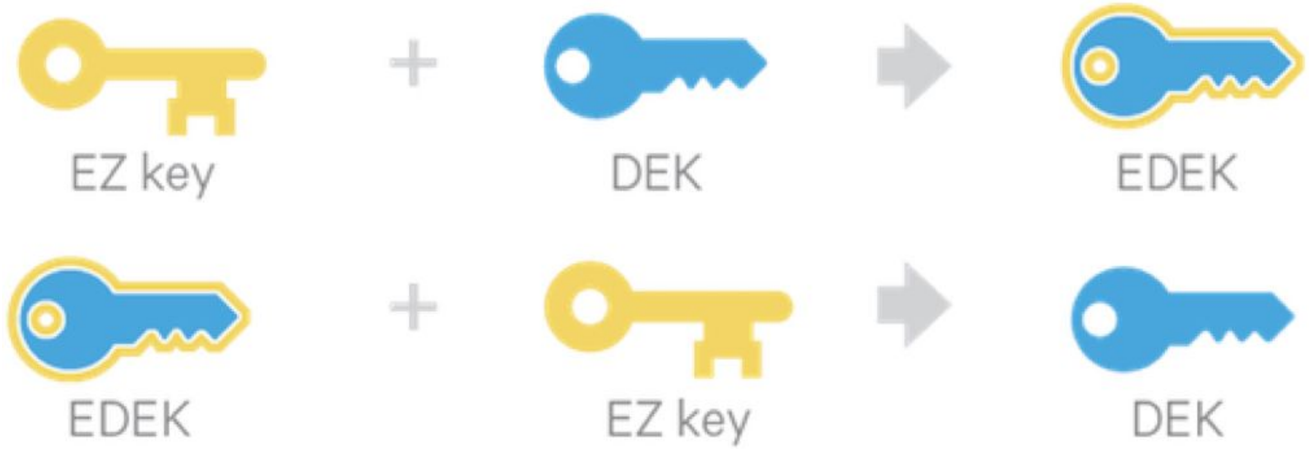
- Is made by binding an encryption key to an empty HDFS directory
- The same key may bind with multiple directories
- Unique keys are made in a zone for each user-file pair

HDFS Encryption Configuration

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

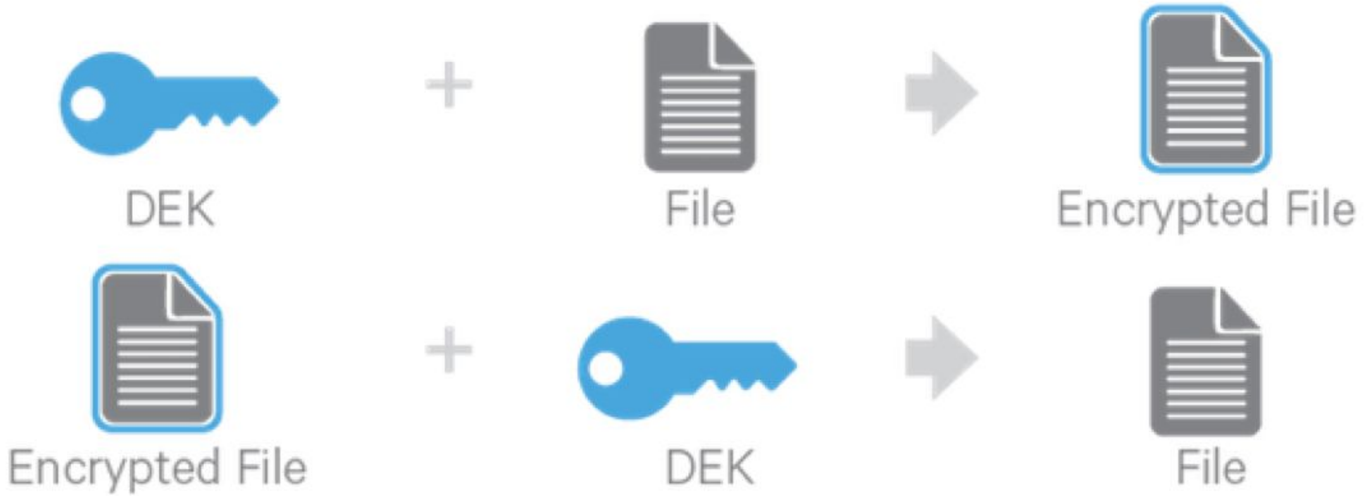
Encryption Zone Keys

- Used to encrypt user/file keys (DEKs)
- Getting an EZ key is governed by KMS ACLs



Data Encryption Keys

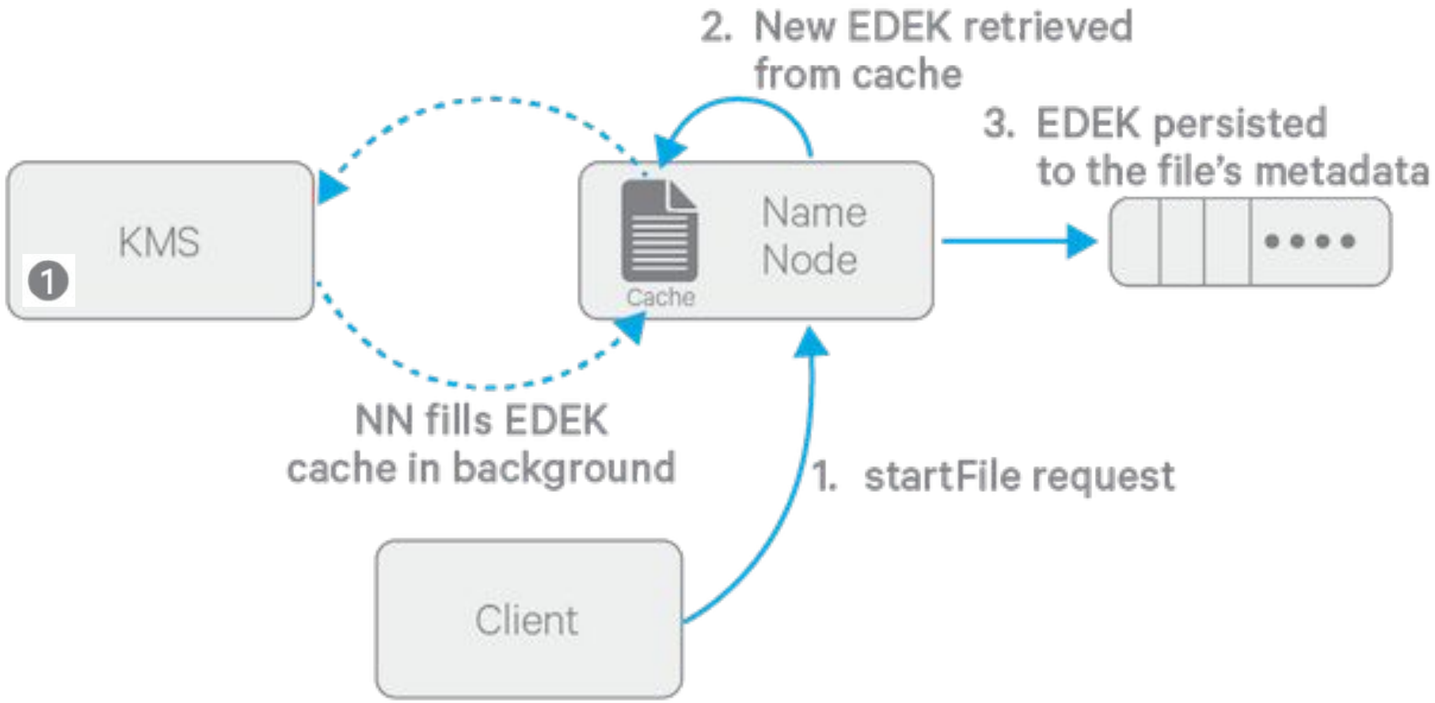
- Encrypts/decrypts file data
- 1 key per file



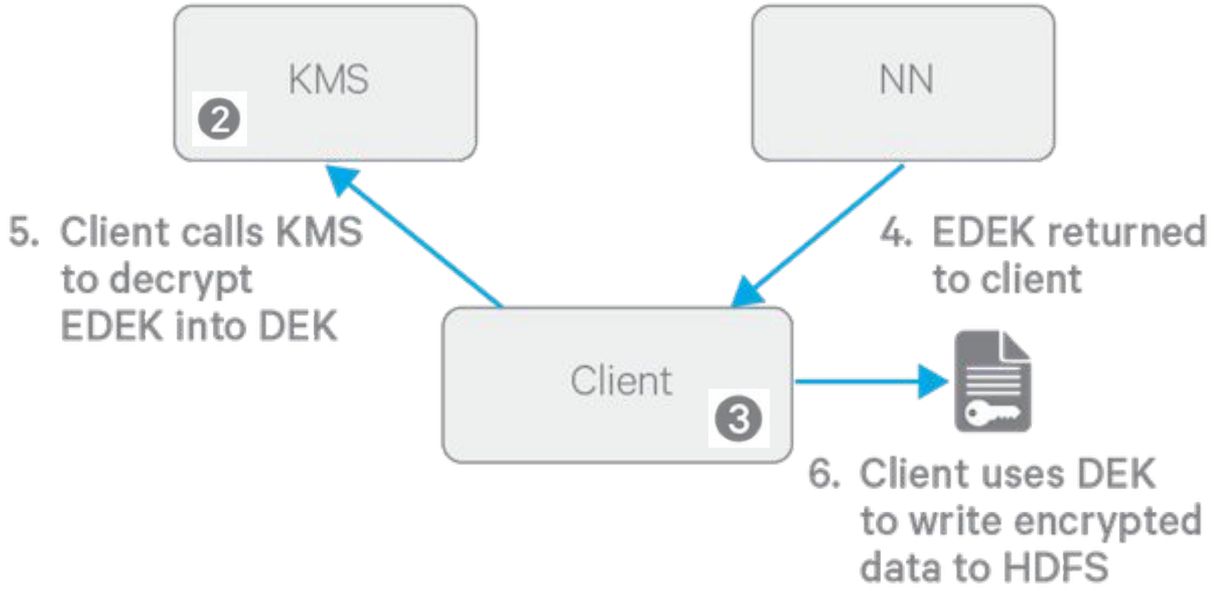
Key Management Server (KMS)

- Client's proxy to the key server
  - E.g. Cloudera Navigator Key Trustee
- Provides a service API and separation of concerns
- Only caches keys
- Access also governed by ACLs (on a per-key basis)

Key Handling



Key Handling



KMS Per-User ACL Configuration

- Use white lists (are you included?) and black lists (are you excluded?)
- Key admins, HDFS superusers, HDFS service user, end users
- **`/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 TLS to protect keytabs in-flight!
- Integrate Kerberos early
- Configure KMS ACLs for KMS roles;
  - Blacklist your HDFS admins -- separation of concerns
  - 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
- Boost entropy on all cluster nodes if necessary
  - Use rngd or haveged

Best practices

- Run KMS on separate nodes outside a Hadoop cluster
- Use multiple KMS instances for high availability, load-balancing
- Harden the KMS instance
  - Use firewall to restrict access to known, trusted subnets
- Make secure backups of KMS configuration!

HDFS Encryption - Summary

- Some performance cost, even with AES-NI (4-10%)
- Requires no modification to Hadoop clients
- Secures data at the filesystem level
- Data remains encrypted from end to end
- Key services are kept separate from HDFS
  - Blacklisting HDFS admins is good practice

Demo

- Accessing HDFS encrypted data from Linux storage

User	Group	Role
hdfs	supergroup	HDFS Admin
keymaster	cm_keyadmins	KMS Admin
carol	keydemo1	User with DECRYPT_EEK access to keydemoA
richard	keydemo2	User with DECRYPT_EEK access to keydemoB

Questions?

Strata
DATA CONFERENCE

Big Data Governance and Emerging Privacy Regulation

Mark Donsky
Senior Director of Products
Okera

Strata
DATA CONFERENCE

Key facts on recent privacy regulation

General Data Protection Regulation (GDPR)

- Adopted on April 14, 2016 and enforceable on May 25, 2018
- Applies to all organizations that handle data from EU data subjects
- Fines of up to €20M or 4% of the prior year's turnover
- Standardizes privacy regulation across the European Union

https://eugdpr.org/

California Consumer Protection Act (CCPA)

- Signed into law on June 28, 2018 and enforceable on January 1, 2020
- Penalties are up to $2500 per violation or up to $7500 per intentional violation
- Clarifications are still being made

https://oag.ca.gov/privacy/ccpa

GDPR vs CCPA: key comparisons

	GDPR	CCPA
Data subjects	Simply refers to “EU data subjects”, some consider this to be EU residents; other consider this to be EU citizens	Applies to California residents
Organizations	All organizations, both public and non-profit	For-profit companies with: (1) gross revenues over $25M, (2) Possesses the personal information of 50,000 or more consumers, households, or devices, or (3) derive at least 50% of revenue from selling consumer information
Rights	● The right to erasure ● The right to access their data ● The right to correct their data ● The right to restrict or object to processing of data (opt-in) ● The right to breach notification within 72 hours of detection	● The right to know what personal information is being collected about them ● The right to know whether their personal information is sold or disclosed and to whom ● The right to say no to the sale of personal information (opt-out) ● The right to access their personal information ● The right to equal service and price, even if they exercise their privacy rights

Common CCPA and GDPR objectives

The right to know: Under both regulations, consumers and individuals are given bolstered transparency rights to access and request information regarding how their personal data is being used and processed.

The right to say no: Both regulations bestow individual rights to limiting the use and sale of personal data, particularly regarding the systematic sale of personal data to third parties, and for limiting analysis/processing beyond the scope of the originally stated purpose.

The right to have data kept securely: While differing in approach, both regulations give consumers and individuals mechanisms for ensuring their personal data is kept with reasonable security standards by the companies they interact with.

The right to data portability: Both regulations grant consumers rights to have their data transferred in a readily usable format between businesses, such as software services, facilitating consumer choice and helping curb the potential for lock-in.

“Businesses need to take a more holistic and less regulation-specific approach to data management and compliance to remain competitively viable.”

Paige Bartley, Senior Analyst, Data Management Data, AI & Analytics, 451 Research

Requirements for holistic privacy readiness

Know what data you have in your data lake

Know how your data is being used

Implement privacy by design

Consent management and right to erasure

Requirements for holistic privacy readiness

Know what data you have in your data lake

- Create a catalog of all data assets
- Tag data sets and columns that contain personal information

Know how your data is being used

- Auditing
- Lineage

Implement privacy by design

- Encrypt data
- Restrict access to data with fine-grained access control
- Pseudonymization
- Anonymization

Consent management and right to erasure

- Implement views that expose only those who have opted in, or hide those who have opted out

Best practices

- Privacy by design
- Pseudonymization and anonymization
- Fine-grained access control
- Consent management and right to erasure

Privacy by design: key principles

- Proactive not reactive; Preventative not Remedial
- Privacy as the Default
- Privacy Embedded into Design
- Full functionality – positive-sum, not zero-sum
- End-to-End Security—Lifecycle Protection
- Visibility and Transparency
- Respect for User Privacy

Key requirements

- End-to-end encryption
- Fine-grained access control
- Comprehensive auditing
- Pseudonymization
- Anonymization
- Visibility by context
- Fine-grained erasure

Privacy by design: pseudonymization and anonymization

Pseudonymization: substitute identifiable data with a reversible, consistent value

Anonymization: destroy the the identifiable data

Pseudonymization is a good practice for privacy, but it does not guarantee anonymity

Dynamic functions in view can implement pseudonymization and anonymization.

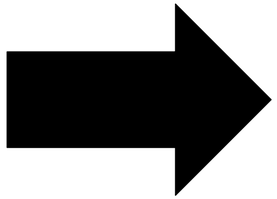
Name	Pseudonymized	Anonymized
Clyde	qOerd	xxxxx
Marco	Loqfh	xxxxx
Les	Mcv	xxxxx
Les	Mcv	xxxxx
Marco	Loqfh	xxxxx
Raul	BhQI	xxxxx
Clyde	qOerd	xxxxx

Privacy by design: fine-grained access control

Master table

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	0234837823	238-23-9876	AZP	Sell	DE
16-Feb-2018	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018	3485739384	585-11-2345	ICBD	Buy	DE
16-Feb-2018	3847598390	234-11-8765	FWQ	Buy	DE
16-Feb-2018	8765432176	344-22-9876	UAD	Buy	FR
16-Feb-2018	3456789012	412-22-8765	NZMA	Sell	FR

Anonymization
Row level filtering



What German brokers see

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	0234837823	xxx-xx-xxxx	AZP	Sell	DE
16-Feb-2018	3485739384	xxx-xx-xxxx	ICBD	Buy	DE
16-Feb-2018	3847598390	xxx-xx-xxxx	FWQ	Buy	DE

What French brokers see

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018	8765432176	344-22-9876	UAD	Buy	FR
16-Feb-2018	3456789012	412-22-8765	NZMA	Sell	FR

Privacy by design: consent management

Consent management

A freely given indication of the data subject's wishes by which he or she signifies agreement to the processing of his or her personal data.

Right to erasure

Individuals have the right to have personal data erased. This is also known as the "right to be forgotten".

GDPR requires opt-in for consent management, whereas CCPA allows opt-out

To implement, you'll need whitelists and blacklists:

- **Whitelists** (opt-in): A list of all record IDs of subjects that have given consent to the use of their data
- **Blacklists** (opt-out): A list of record IDs of subjects that have opted out of the use of their data

Implement consent management by constructing views on top of master tables that join on whitelists or blacklists

Never provide access to master tables to data consumers

Privacy by design: consent management

Master table

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	0234837823	238-23-9876	AZP	Sell	DE
16-Feb-2018	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018	3485739384	585-11-2345	ICBD	Buy	DE
16-Feb-2018	3847598390	234-11-8765	FWQ	Buy	DE
16-Feb-2018	8765432176	344-22-9876	UAD	Buy	FR
16-Feb-2018	3456789012	412-22-8765	NZMA	Sell	FR

Consent table
whitelist

Account ID	Opt-In
0234837823	Yes
3947848494	Yes
4848367383	Yes
3485739384	No
3847598390	Yes
8765432176	Yes
3456789012	No

What marketing analysts see
with global visibility

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	0234837823	238-23-9876	AZP	Sell	DE
16-Feb-2018	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018	3847598390	234-11-8765	FWQ	Buy	DE
16-Feb-2018	8765432176	344-22-9876	UAD	Buy	FR

Privacy by design: right to erasure

Master table

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	0234837823	238-23-9876	AZP	Sell	DE
16-Feb-2018	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018	3485739384	585-11-2345	ICBD	Buy	DE
16-Feb-2018	3847598390	234-11-8765	FWQ	Buy	DE
16-Feb-2018	8765432176	344-22-9876	UAD	Buy	FR
16-Feb-2018	3456789012	412-22-8765	NZMA	Sell	FR

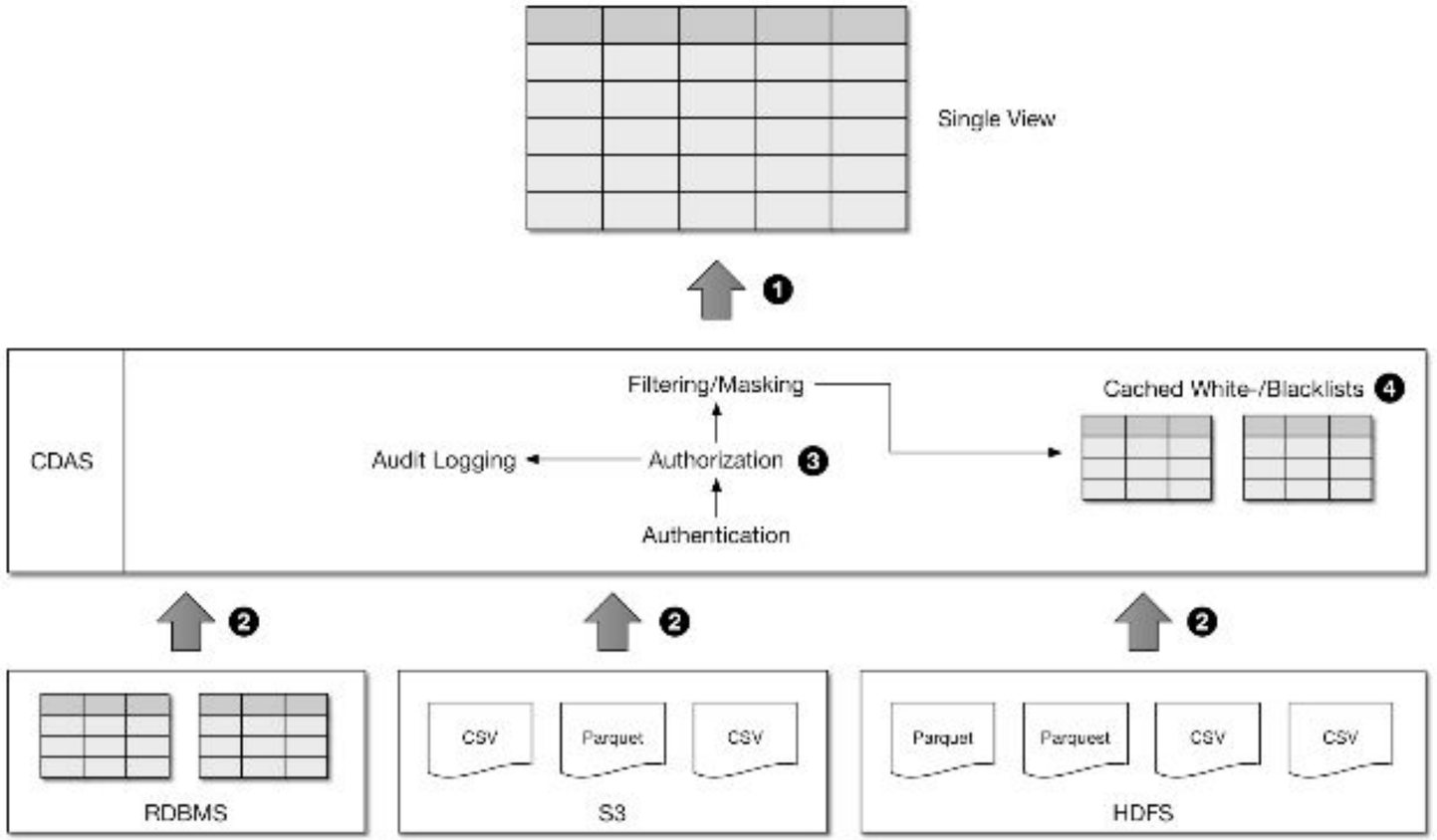
Deletion table
blacklist

Account ID	Deleted
3485739384	Yes
3456789012	Yes
0234837823	Yes

What analysts see
with global visibility

Date	Account ID	National ID	Asset	Trade	Country
16-Feb-2018	3947848494	329-44-9847	TBT	Buy	FR
16-Feb-2018	4848367383	123-56-2345	IDI	Sell	FR
16-Feb-2018	3847598390	234-11-8765	FWQ	Buy	DE
16-Feb-2018	8765432176	344-22-9876	UAD	Buy	FR

Consent management and right to erasure in action



1. Instead of rewriting data every time that a person gives consent or opts out, the data is filtered on access using white lists and/or blacklists
2. Convert every data source into a table structure, regardless of the original format
3. Grant access to databases and datasets in a fine-grained manner

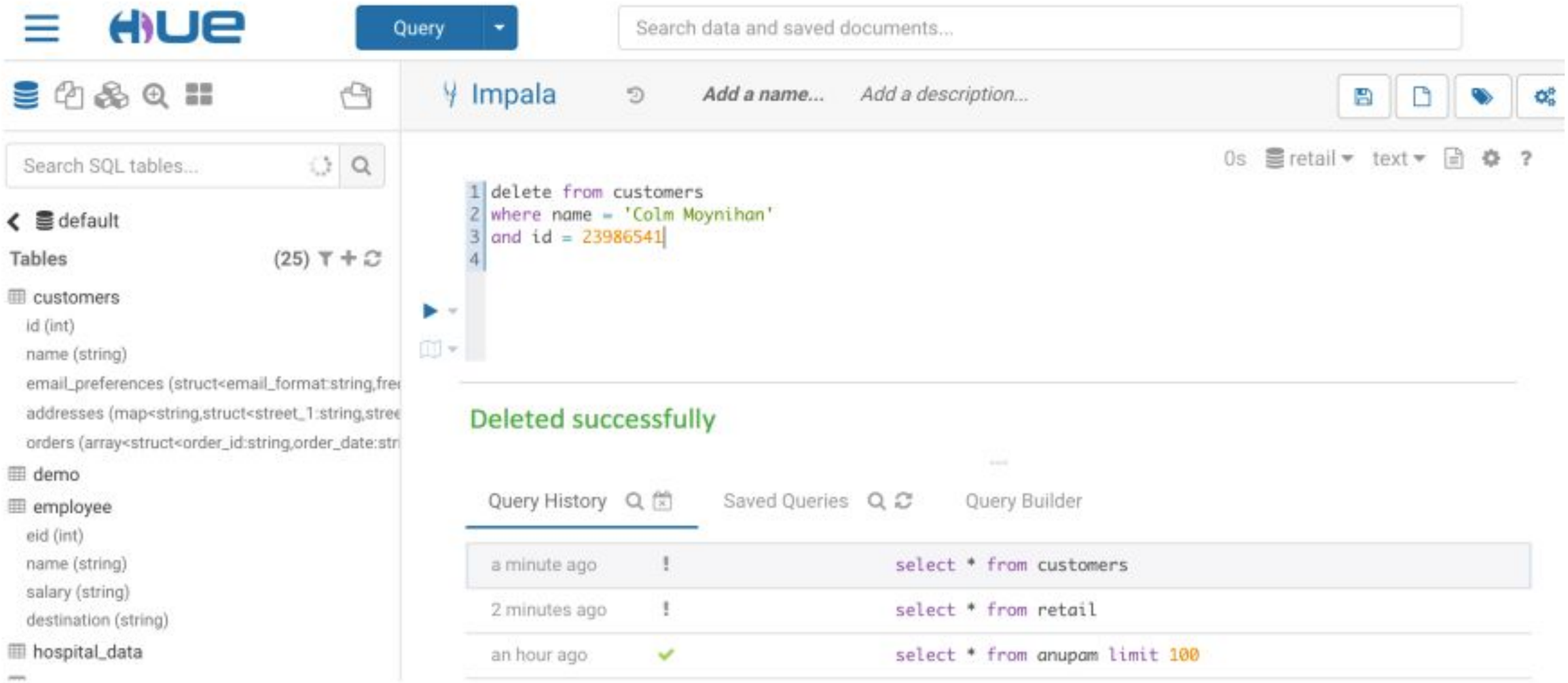
Right to Erasure: HDFS and Cloud (manual)

- Concentrate personal data in a small number of “lookup tables”
- Upon a delete request, add records to a “to be deleted” table
- Execute a periodic batch job to remove “to be deleted” records by rewriting entire files/partitions/tables

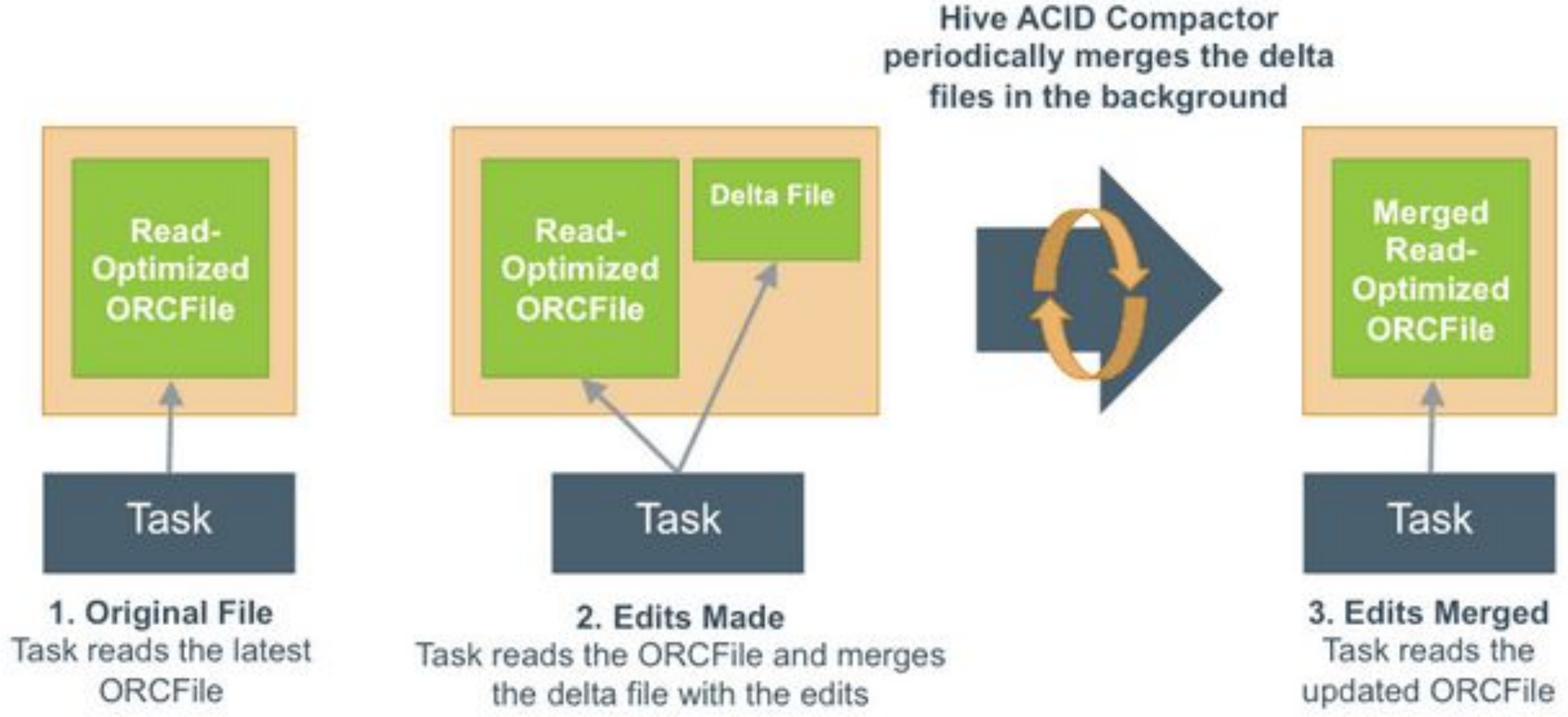
```
--View for users to query
CREATE VIEW merged_view AS SELECT * FROM main_table WHERE id NOT IN (SELECT id FROM
delete_table);

--Periodic merging / rewriting
--First, create merged table
CREATE TABLE main_table_v2 AS SELECT * FROM main_table WHERE id NOT IN (SELECT id
FROM delete_table);
--Second, point the view to the new table
ALTER VIEW merged_view AS SELECT * FROM main_table_v2 WHERE id NOT IN (SELECT id
FROM delete_table);
--Third, clear the delete table
TRUNCATE TABLE delete_table;
```

Right to Erasure: Kudu



Right to Erasure: Hive ACID



Okera's holistic privacy capabilities

Universal policy enforcement across data formats and compute engines

- Define policies once and enforce everywhere (Spark, SparkSQL, Python, EMR, Hive, etc.)

Role-based and Attribute-based access control

- Enrich data sets with and assign access policies on business context instead of technical metadata (grant access to sensitive sales data to Charlie)

Full support for both pseudonymization and anonymization

- Enrich data sets with and assign access policies on business context instead of technical metadata (show last four digits of SSN, replace email address with email@redacted.com)

Dynamic views for right to erasure, consent management, and easy administration

- Simplify view administration with join-based filters and dynamic policies that include Hive UDFs that are evaluated just-in-time (e.g., has_access, has_roles)

Questions

Final Thoughts

Compliance

- We have shown how an various environments can be secured end-to-end
- Is this enough to be compliant?
  - PCI DSS, HIPAA, GDPR, CCPA
  - 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

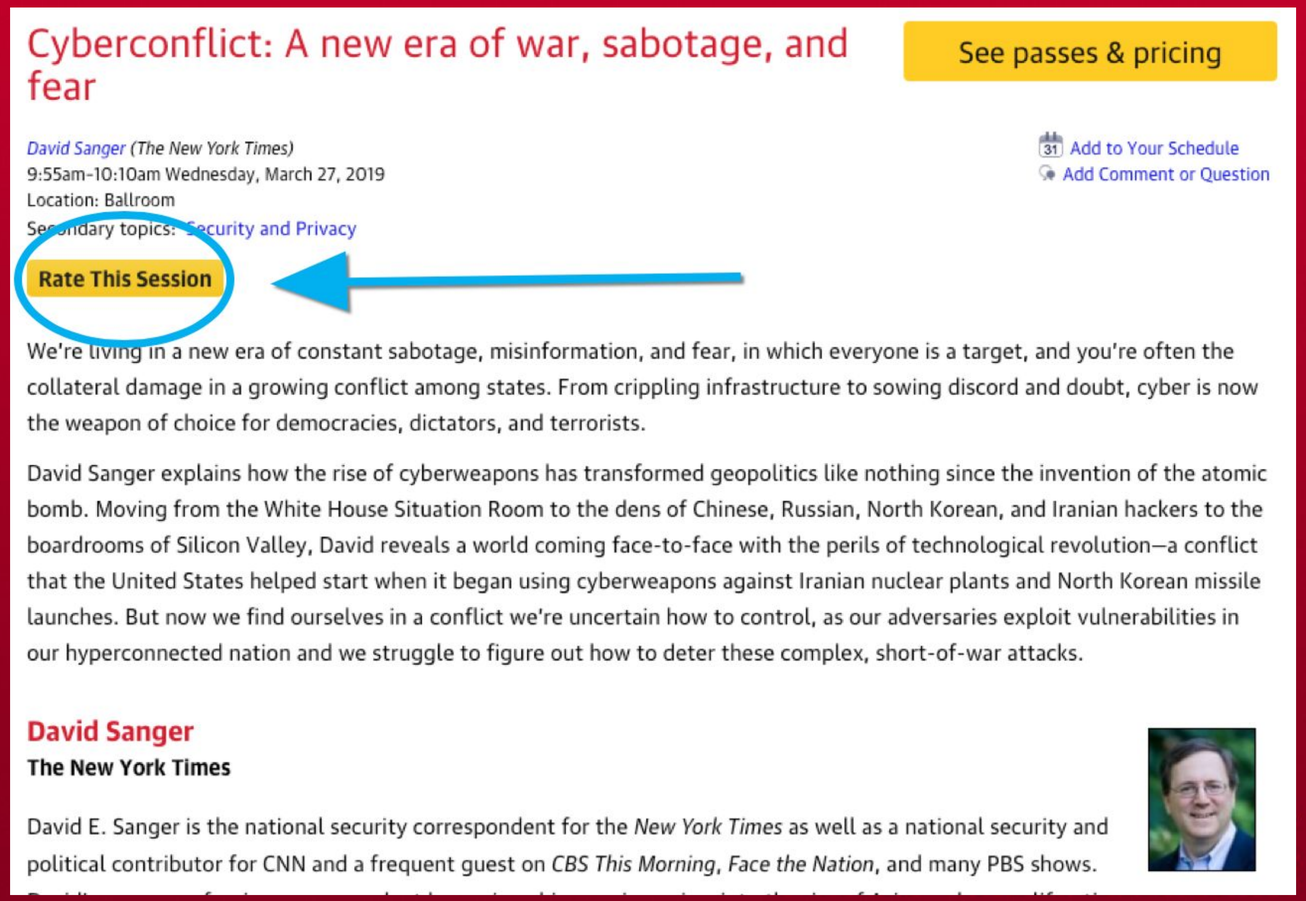
- Many Hadoop deployments occur in the public cloud
- Security considerations presented today all still apply
- Complementary to native cloud security controls

- **Blog posts**
  - http://blog.cloudera.com/blog/2016/05/how-to-deploy-a-secure-enterprise-data-hub-on-aws/
  - https://www.okera.com/blog/solving-gdpr-challenges-with-okera-part-1/

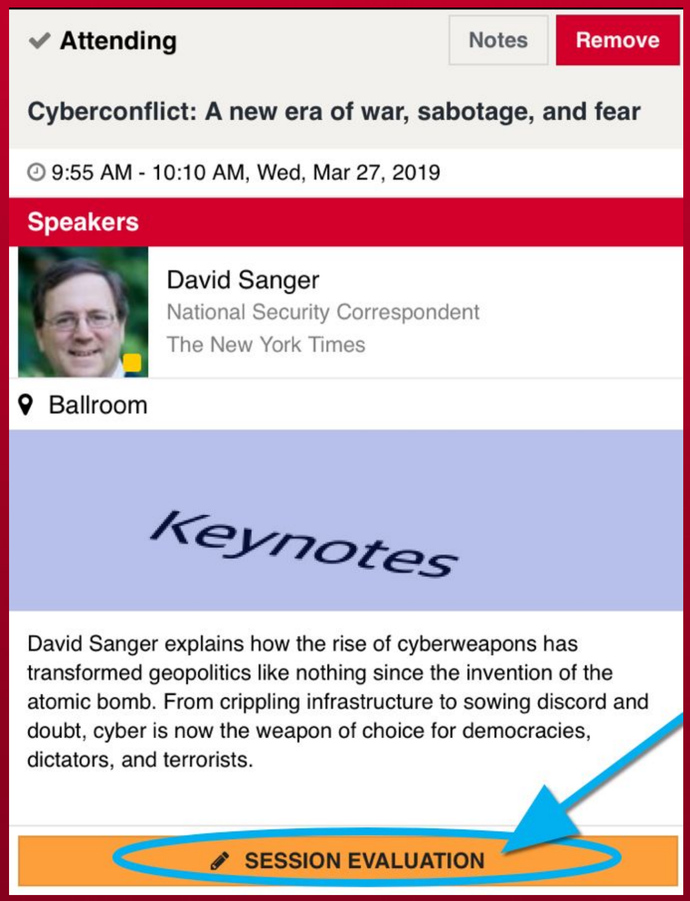
Looking Ahead

- 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!

Rate Today's Session



Session page on conference website



O'Reilly Events App

Final Questions?

Thank you!