CLOUDERA

CDPDc SDX Overview/Demo

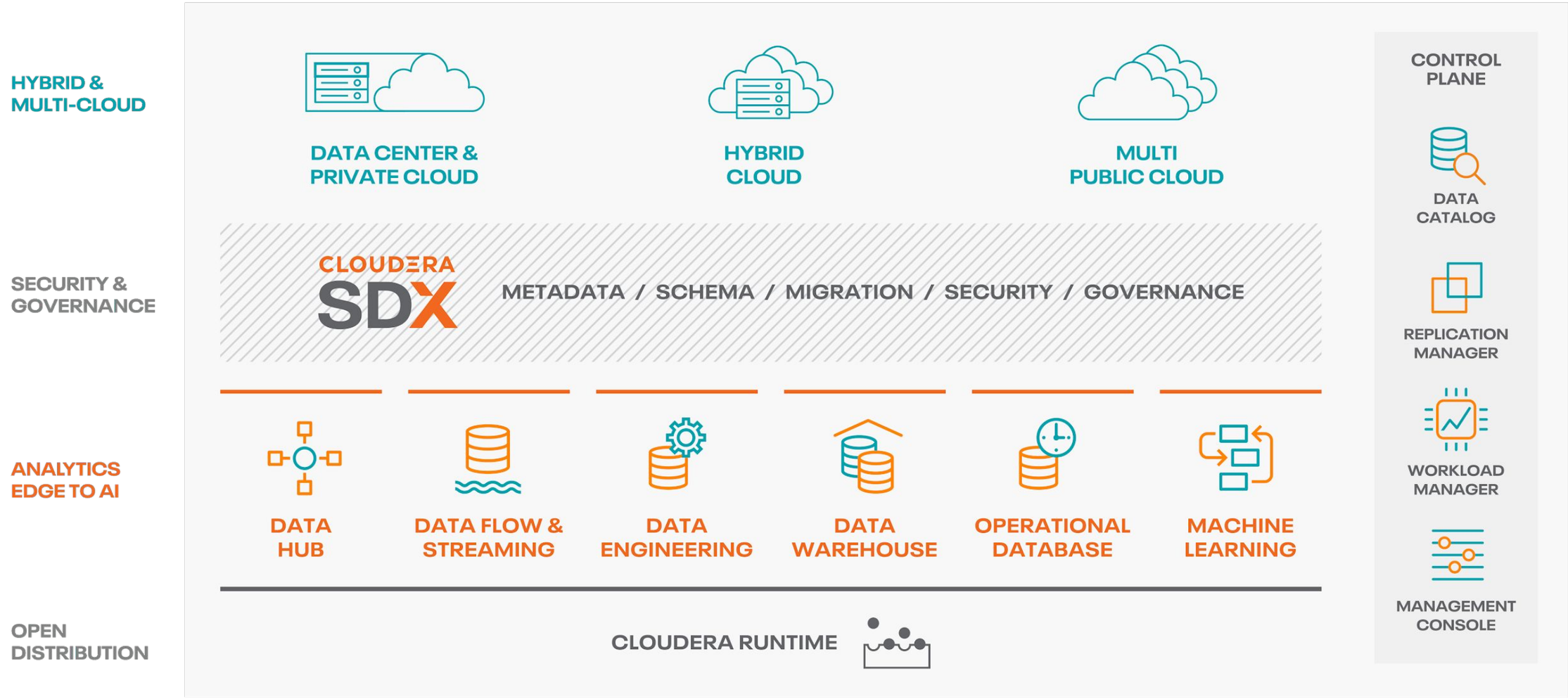
Ali Bajwa | Director Partner Solution Engineer

Agenda

- CDP Refresher
- CDP Security And Governance
- CDPDc SDX 101
  - Firewall
  - Authentication/Encryption
  - Authorization/Audit/Admin
  - Data Governance
- Demo/Lab
- Next Steps

CDP Refresher

CLOUDERA DATA PLATFORM



CDP Data Center

Take the two best open-source data analytics platforms, fuse them together, add new capabilities, and we get CDP Data Center.

Enterprise Data Hub

+

HDP Enterprise Plus

+ New Features =

CDP Data Center

- ✓ First CDP product on-premises
- ✓ 20+ components
- ✓ Highly customizable

A NEW OPEN SOURCE DISTRIBUTION FOR BETTER CAPABILITY

Cloudera Runtime 7.0 – created from the best of CDH and HDP

Adopt superior technologies

~~Ambari~~	Cloudera Manager
~~Sentry~~	Ranger
~~Cloudera Director~~	Cloudbreak
~~Hive on Spark~~	Hive on Tez

Merge overlapping technologies

- Navigator + Atlas + DSS
- WXM + DAS
- BDR + DLM
- Hue + DAS Lite

Keep complementary technologies

Impala	Hive LLAP w/HWC
Parquet	ORC
Kudu	Hive ACID + Druid
CDSW	Zeppelin
NiFi	Phoenix
Knox	Livy

Upgrade shared technologies

Hadoop 3.1	Spark 2.4
Hive 3.1	Oozie 5.1
HBase 2.2	Accumulo 1.9
Kafka 2.1	Sqoop 1.4
Solr 7.4	Zookeeper 3.4
Pig 0.17	

© Cloudera, Inc. All rights reserved.

CDP-Dc: Three Key Selling Points

1. Whether you are coming from CDH or HDP, there is something new for you
2. Unified distribution means more features and bug fixes, sooner
3. Prerequisite for CDP containerized applications

Selling Point #1: Something new for everyone

If you are coming from CDH...	
Ranger	• Dynamic row filtering • Dynamic column masking • Attribute-based access control • SparkSQL fine-grained access control
Hive 3	• ACID support - simplify development with transaction guarantees • Comprehensive ANSI SQL 2016 coverage
Hive on Tez	• Better ETL performance

If you are coming from HDP...	
Cloudera Manager	• Manage multiple clusters • Automated wire encryption setup • Fine-grained RBAC for administrators • Streamlined maintenance workflows
Impala	• Better fit for Data Mart migration use cases (interactive, BI style queries)
Hue	• Built-in SQL editor
Kudu	• Better performance for fast changing / updateable data

For all of you...
• **Virtual Private Clusters** - Separate Compute & Storage
• **Atlas 2.0** - Advanced Data Discovery, Spark support
• **Ozone Object Storage** (Tech Preview), HDFS Erasure Coding

NEW FEATURES IN CDP DATA CENTER

New features for CDH 6 customers	
Ranger 2.0	• Dynamic row filtering & column masking • Attribute-based access control • SparkSQL fine-grained access control
Atlas 2.0	• Advanced data discovery • Improved performance and scalability
Hive 3	• Hive-on-Tez for better ETL performance • ACID transactions
Ozone (Preview)	• 10x scalability of HDFS
Knox*	• Gateway-based SSO
Druid*	• Low-latency DataMart for real-time and aggregate data
Spark on Docker *	• Simplified dependency management

New features for HDP 3 customers	
Cloudera Manager	• Virtual private clusters • Automated wire encryption setup • Fine-grained RBAC for administrators • Streamlined maintenance workflows
Atlas 2.0	• Advanced data lineage • Faceted search
Solr 7	• Relevance-based text search over unstructured data (text, pdf, .jpg, ...)
Impala	• Better fit for Data Mart migration use cases (interactive, BI style queries)
Hue	• Built-in SQL editor
Kudu	• Better performance for fast changing / updateable data
Better at-rest Encryption	• Key Trustee Server, NavEncrypt*

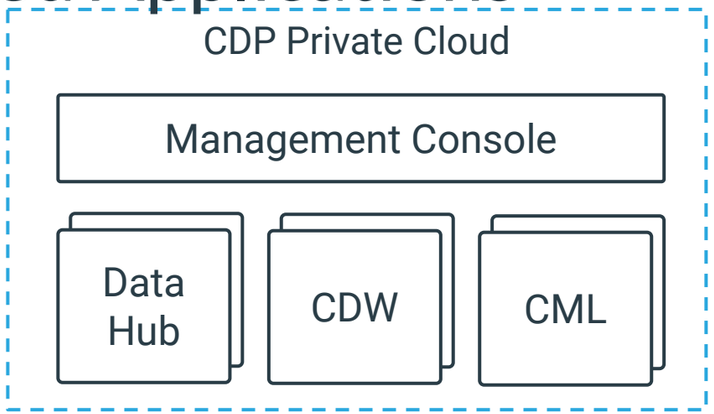
** In future release*

Selling Point #2 - Unified distribution means faster innovation

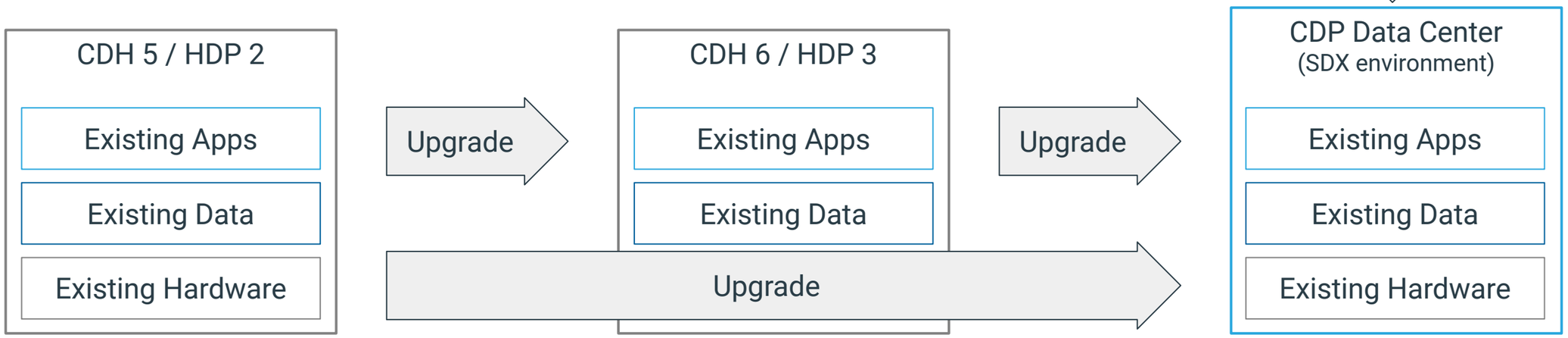
- Concentrated development means more features and bug fixes for our CDP customers, sooner
- Shared development model with public cloud means every on-prem release has been battle tested for months with real workloads

Selling Point #3 - Foundation for Containerized Applications

- CDP Data Center provides the stateful elements for new wave of containerized compute applications
  - Storage
  - Table Schema
  - Authentication & Authorization
  - Governance



- Plan your path to CDP-DC now, expand to new experiences later



CDP DATA CENTER ROADMAP

CDP Data Center 7.0 (Released)		Roadmap
• Cloudera Manager 7.0 • Hadoop 3.1 • Spark 2.4 • Hive 3.1 • Impala 3.2 • Oozie 5.1 • Hue 4.5 • Ranger 2.0 • Atlas 2.0 • Solr 7.4 • Tez 0.9	• HBase 2.2 • Phoenix 5.0 • Kudu 1.11 • Sqoop 1.4.7 • Parquet 1.10 • Avro 1.8 • ORC 1.5 • Zookeeper 3.5 • Kafka 2.3 • Key Trustee Server 7 • Ozone (Tech Preview)	• Livy • Druid • Ranger KMS • Key HSM • Navigator Encrypt • Zeppelin • Knox • Accumulo

CDP DATA CENTER - FOCUS AREAS

Top Priorities

- In-place upgrades from
  - CDH 5.13 - 5.16
  - HDP 2.6.5
  - CDP DC 7.0
- “Unity” feature completion
  - Knox, Zeppelin, Livy, Ranger KMS, Navigator Encrypt, Key HSM
  - Kafka and friends (SR, SMM, SRM)
  - Storage connectors (S3, ADLSv2)
  - Impala column masking
  - Kudu / ranger integration
- Expanded support matrix
  - Java: JDK 11
  - DB: Oracle DB, MySQL, Maria DB
- Ozone Beta 1

Next Priorities

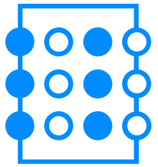
- In-place upgrade from CDH 6
- In-place upgrade from HDP 3
- Solr 8
- Druid
- Accumulo
- IBM Power
- RHEL / CENTOS 8
- Ubuntu 18
- SLES 12 SP 4

© Cloudera, Inc. All rights reserved.

SDX: Under the Hood : 101

CHALLENGES

Security & Governance



Sharing data across workloads

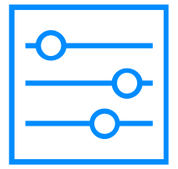
- Requires multiple copies of data need to be created
- Each with its own set of data context

Burdensome admin effort

- Multiple clusters = multiple places to administer

One missing permission in one copy of the data can lead to significant financial and reputation risk

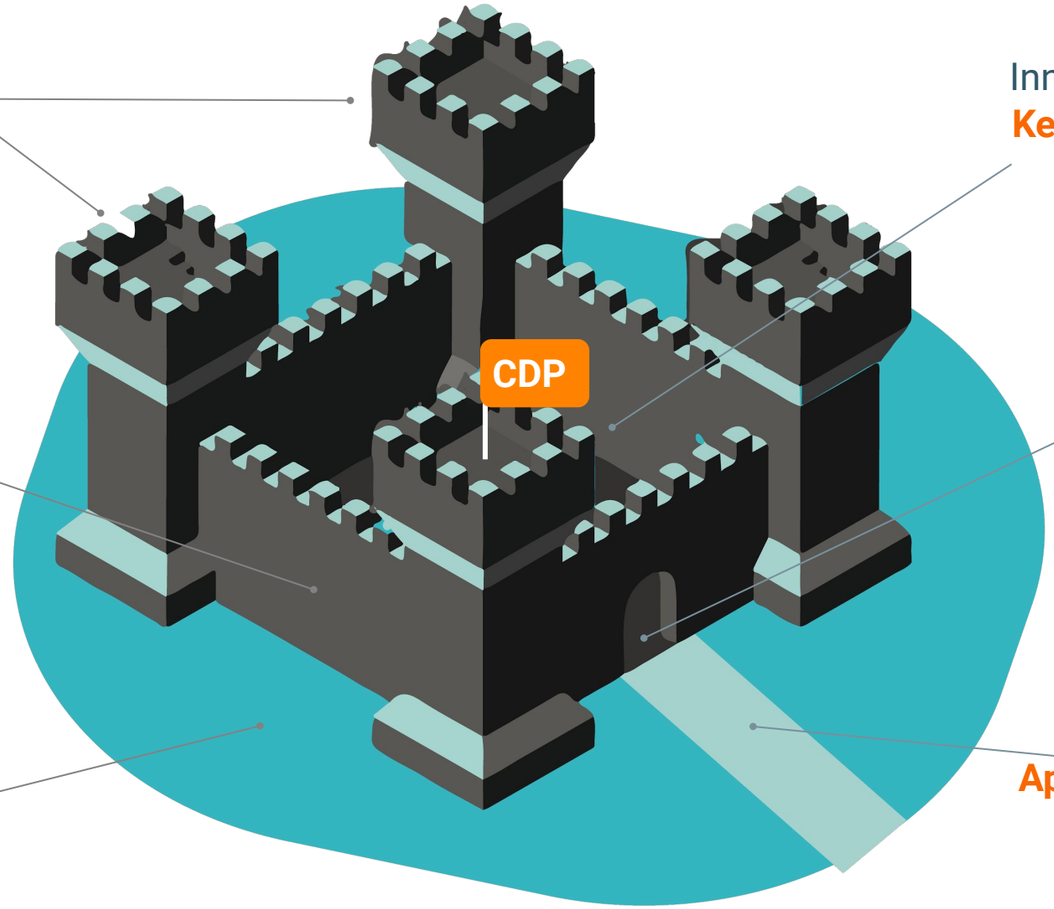
Difficult to share data safely for new analyses



Heavy new regulation such as GDPR makes the challenges even greater

© Cloudera, Inc. All rights reserved.

CDP Security Landscape



Watch Towers
Apache Ranger (HDP/CDP)
~~**Apache Sentry (CDH)**~~

Inner Walls
Kerberos

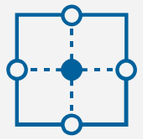
High Walls
HDFS Encryption
Apache Ranger KMS (HDP)
Navigator KMS/KTS (CDH)
Nav Encrypt (CDH)

User Identity Verification
LDAP/AD

Moat
Firewall

Limited Entry Points
Apache Knox (HDP/CDP)

COMPREHENSIVE APPROACH TO SECURITY



Identity & Perimeter

Validate users in enterprise directory

Technical Concepts:
Authentication
User/group mapping

Kerberos,
Apache Knox

Access

Defining what users and applications can do with data

Technical Concepts:
Permissions
Authorization

Apache Ranger

Visibility

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

Technical Concepts:
Auditing
Lineage

Apache Atlas

Data Protection

Shielding data in the cluster from unauthorized visibility

Technical Concepts:
Encryption, Key Management

SSL/TLS, HDFS TDE, Ranger (KMS, Masking, Filtering)

80% of large customers leverage our capabilities across all 4 pillars, to address use cases that include sensitive and regulated data.

© Cloudera, Inc. All rights reserved.

APACHE KNOX

ENHANCED PERIMETER SECURITY

- Protect cluster by hiding network details (reduce port exposure)
- Web App vulnerability filter

API GATEWAY

- Extend API access
- Kerberos encapsulation
- Single SSL Cert
- Multi-cluster support

CENTRALIZED ACCESS

- No more SPNEGO for REST/HTTP Clients
- Central REST API auditing
- Service Level Authorization
- Alternative to SSH edge Node

ENTERPRISE AUTHENTICATION

- LDAP/AD integration
- SSO
- Multi-protocol support (Oauth, SAML, OpenID Connect) via pac4

© Cloudera, Inc. All rights reserved.

APACHE RANGER CAPABILITIES

1

Authorization

- Centralized platform to define and manage security policies consistently across Hadoop ecosystem
- **HDFS, Hive, HBase, YARN, Kafka, Solr, Storm, Knox, NiFi, Atlas, Impala***
- Extensible Architecture with custom policy conditions & context enrichers
- Easy to add new component types for authorization

2

Key Management

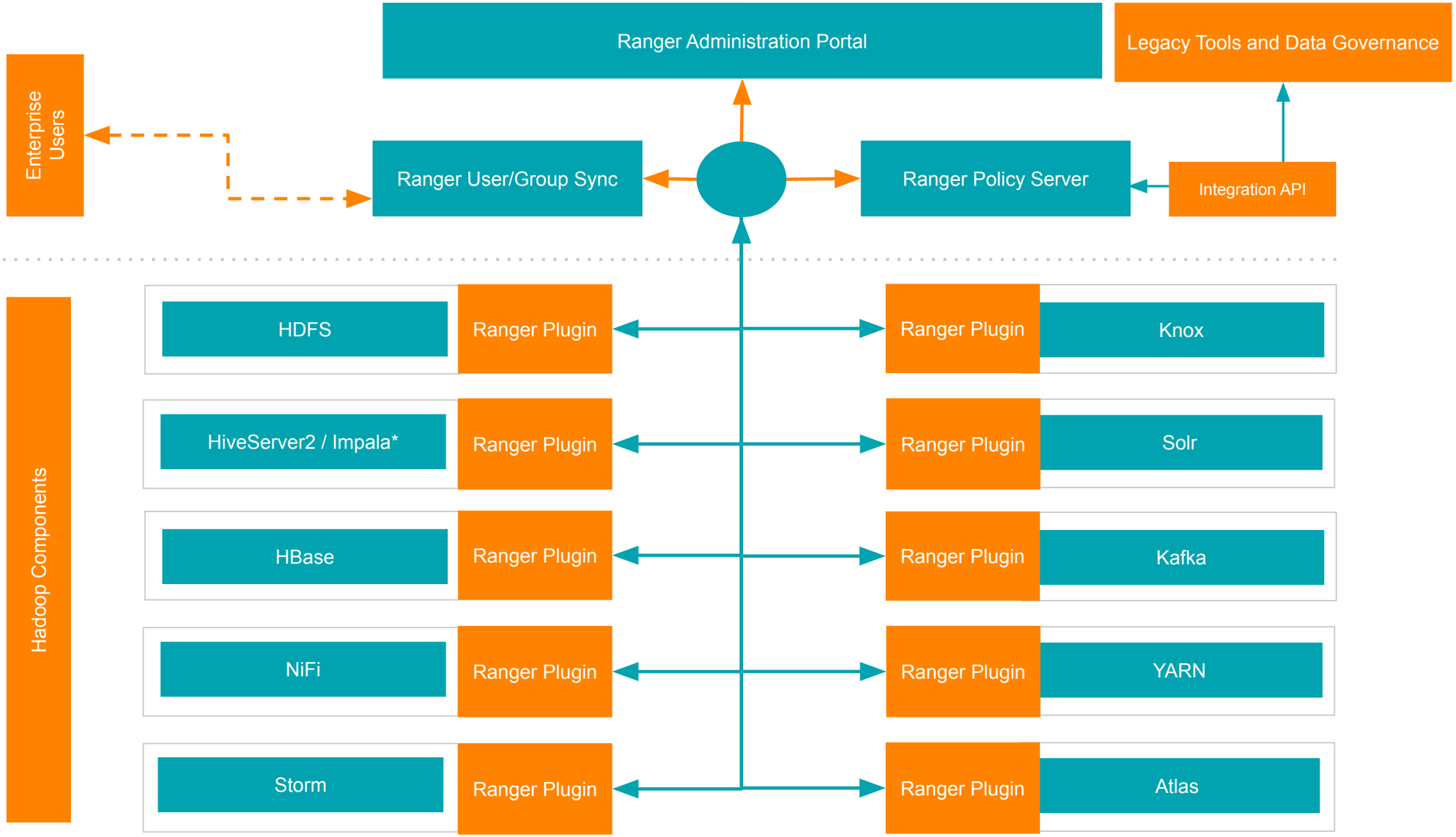
- Store and manage encryption key policies & lifecycle
- Support HDFS Transparent Data Encryption
- Integration with HSM
- **Safenet (LUNA, KeySecure)**

3

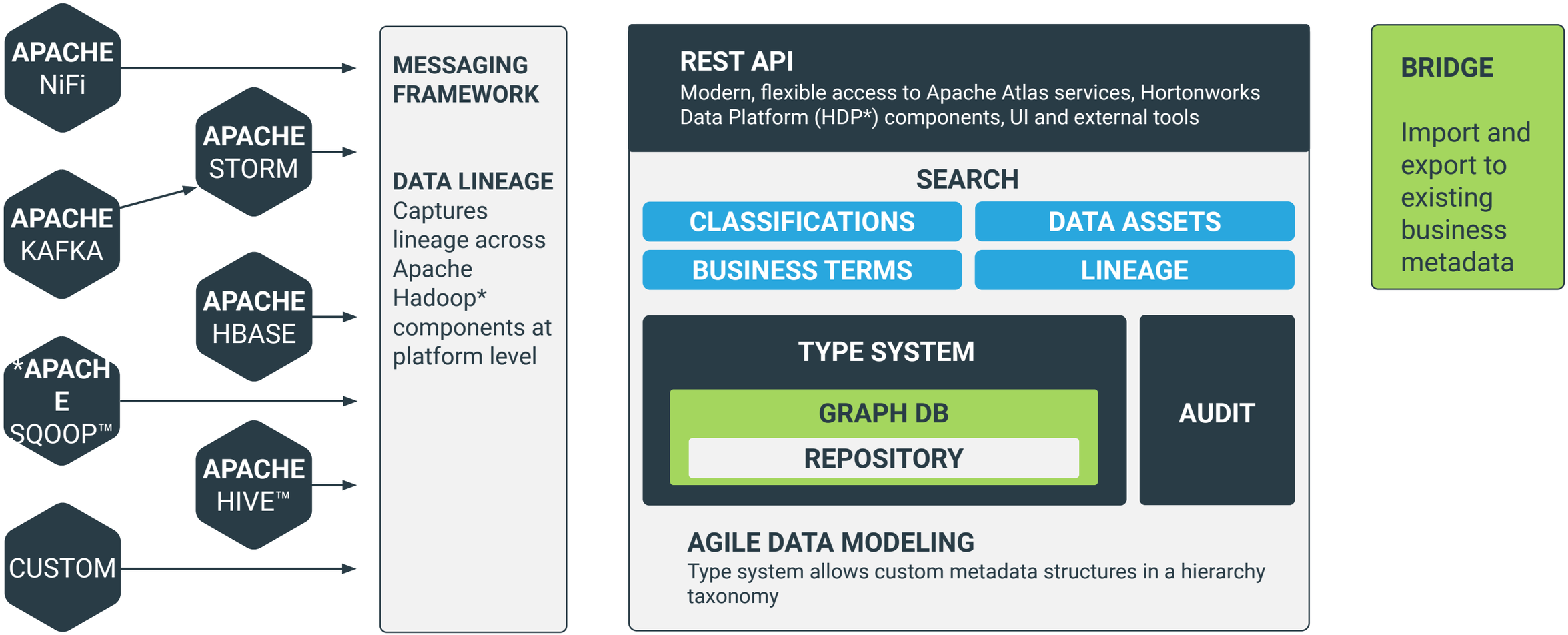
Audits

- **Central audit location for all access requests**
- **Support multiple destination sources (HDFS, Solr, etc.)**
- **Real-time visual query interface**

RANGER ARCHITECTURE

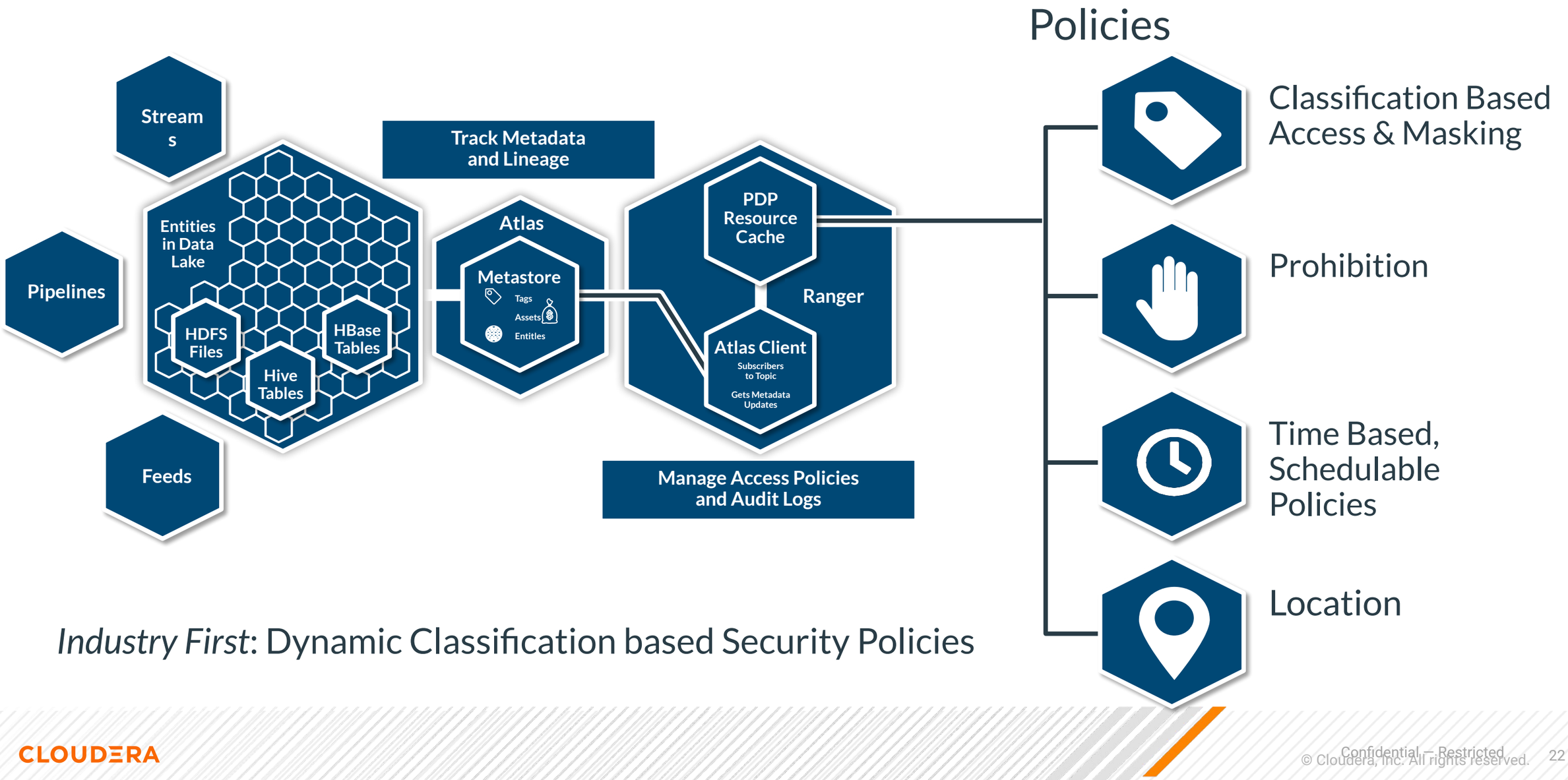


APACHE ATLAS ARCHITECTURE



© Cloudera, Inc. All rights reserved.

CDP – SECURITY & GOVERNANCE



Industry First: Dynamic Classification based Security Policies

DYNAMIC ROW FILTERING & COLUMN MASKING

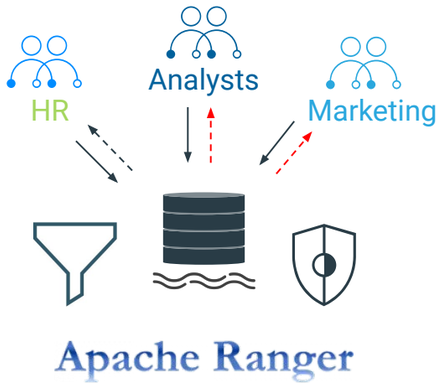
User 1: Joe
Location : US
Group: Analyst

Users from US Analyst group see data for US persons with CC and National ID (SSN) as masked values and MRN is nullified

EU HR Policy Admins can see unmasked but are restricted by row filtering policies to see data for EU persons only

User 2: Ivanna
Location : EU
Group: HR

Country	National ID	CC No	MRN	Name
US	xxxxx3233	4539 xxxx xxxx xxxx	null	John Doe
US	xxxxx7465	5391 xxxx xxxx xxxx	null	Jane Doe



Country	National ID	Name	MRN
Germany	T22000129	Ernie Schwarz	876452830A

Ranger Policy Enforcement

Query Rewritten based on Dynamic Ranger Policies: Filter rows by region & apply relevant column masking

Original Query:

SELECT country, nationalid, name, mrn FROM ww_customers

Original Query:

SELECT country, nationalid, ccnumber, mrn, name FROM ww_customers

Country	National ID	CC No	DOB	MRN	Name	Policy ID
US	232323233	4539067047629850	9/12/1969	8233054331	John Doe	nj23j424
US	333287465	5391304868205600	8/13/1979	3736885376	Jane Doe	cadsd984
Germany	T22000129	4532786256545550	3/5/1963	876452830A	Ernie Schwarz	KK-2345909

© Cloudera, Inc. All rights reserved.

Demo Scenario

Business Scenario

1. WorldWide – mid-size financial services company (WWBank merged with WWAssurance health insurance services) expanding from US to international markets
2. Employees in EU and US
3. Multiple business units need access to customer data: Analysts, Compliance Admins, HR
4. Customer data is co-mingled as well as isolated
5. Leases data from external data brokers
6. Needs to have rational security policies to provide the right level of access control to customer data across geographies, business functions, and to comply with external regulations (GDPR, PII, HIPAA, EU Privacy etc.)

Setup

1. 2 Customer Tables owned by bank: 50K customer records each with 38 fields (PII, PHI, PCI & non-sensitive data)
    i. us_customers: USA person data only
    ii. ww_customers: multi-language, multi-country, localized person data

2. 1 Reference table: eu_countries (reference table for looking up EU country codes to country mappings – with BRExit etc.)

3. Finance DB: 1 data set leased from a data broker
    i. tax_2015: Data lease is already expired (on Dec 31st 2015)

Ranger Hive Policies Setup

1. Only US employees can see data in **us_customers** table and only from locations within the US (access_us_customers)
2. *US employees* can see only data rows of *US persons* in **ww_customers** table (“filter_ww_customers for consent” + access_ww_customers)
3. *EU employees* can see only data rows of *EU persons who have given consent* in **ww_customers** table (“filter_ww_customers for consent” + access_ww_customers)
4. *HR team* members can see all original *unmasked data* (PCI, PII,….)
5. Super users belonging to etl/DPO groups can see data for all EU customers
6. *Masking*: Analysts can view only masked versions of sensitive data from **ww_customers** table but are prohibited from viewing PII data in the **us_customers** table(s) (All masking policies under Masking Tab of Resource based policies)
7. *Prohibition*: No combination of zip code, insurance, and bloodgroup data are permitted to be joined in any query (prohibition policy)

Ranger Hbase/Kafka Policies

1. Kafka topic PRIVATE and Hbase Table **T_Private** are classified as SENSITIVE
2. Kafka topic FOREX and Hbase Table **T_Forex** do not have any classification (and no sensitive information)
3. ETL User group can publish to both FOREX and PRIVATE Kafka topics (publish)
4. Analysts :
    a. Can access Forex rates table ONLY
    b. Can read from FOREX Kafka topic (consume)
    c. Can't access PRIVATE Kafka topic (consume)
    d. Can't publish to any Kafka topics
5. HR User group:
    a. Can access Forex rates AND PRIVATE Hbase tables
    b. Can read from PRIVATE and FOREX Kafka topics (consume)
    c. Can't publish to any Kafka topics

Personas / Masking Rules

User	Group	Access Privileges
joe_analyst	us_employee	US Data Only, non-sensitive data only, rest masked or forbidden depending on sensitivity
ivanna_eu_hr	eu_employee	EU Data Only (only customers who gave consent), All sensitive data
etl_user	eu_employee	EU Data (all customers), All sensitive data, Update consent/Delete

Data Column	Masking Type	Sample Output
Password**	Hash	237672b21819462ff39fcea7d990c3e5
National ID	Last 4 Only	xx-xx-9324
Credit card	First 4 Only	4532xxxxxxxxxxx
Street Address	Static	nnn Xxxxxx Xxxxx
MRN**	NULL	null
Birthdate	Custom	Hide birthday by showing it as 01/01/yyyy
Age	Custom	(Add a random number below 20 to actual age)

** = masked via tag based masking

Next Steps

CLOUDERA SDX

https://www.cloudera.com/products/sdx.html

CLOUDERA Connect Portal

https://www.cloudera.com/partners/cloudera-connect-partner-program.html

CLOUDERA SDX Webinar

https://mindtickle.app.link/TZynd1Z383

CLOUDERA Demo Center

https://my.cloudera.com/partner-portal/training/demo-center.html

Contacts

Sales IBM/Cloudera (Global)

- Jerry Green jerry.green@us.ibm.com
- Matt Lanagan mlanagan@cloudera.com

Tech Sales IBM/Cloudera (Global)

- Brett Coffman brett.coffman@ibm.com
- Venkatesh Sellappa venky@cloudera.com

Sales IBM/Cloudera (NA/LA)

- Tom Burke thomasburke@us.ibm.com (NA)
- Camilo Esteban Rojas Lopez camilor@co.ibm.com (LA)
- Jeffrey Schmitt jschmitt@cloudera.com

Offering Management

- Nagapriya Tiruthani ntiruth@us.ibm.com
- Kiran Guduguntla kgudugun@us.ibm.com
- Jessica Lee jessicalee@us.ibm.com

Support (Global)

- Michael Pintus pintus@us.ibm.com

Stop by the CLOUDERA booth if you have further questions
Booth #48

Demo

THANK YOU