

Privacidad de datos en el Data Lake.

5^{to} FORO

en Seguridad de la Información

RETOS Y SOLUCIONES

PARA LA PRIVACIDAD EN UN MUNDO CONECTADO

Sobre mí

Iván Herrero Bartolomé

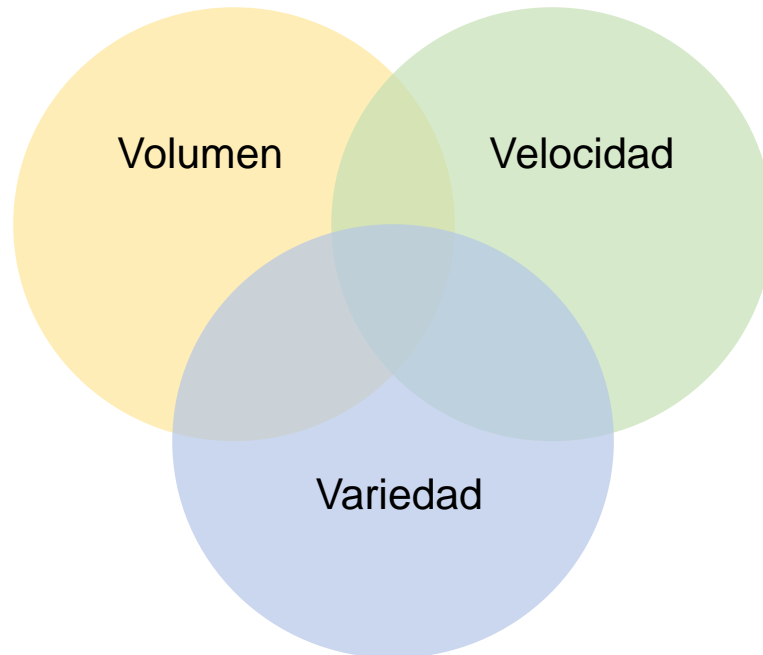
Gerente de Data & Analytics en **everis Colombia**.



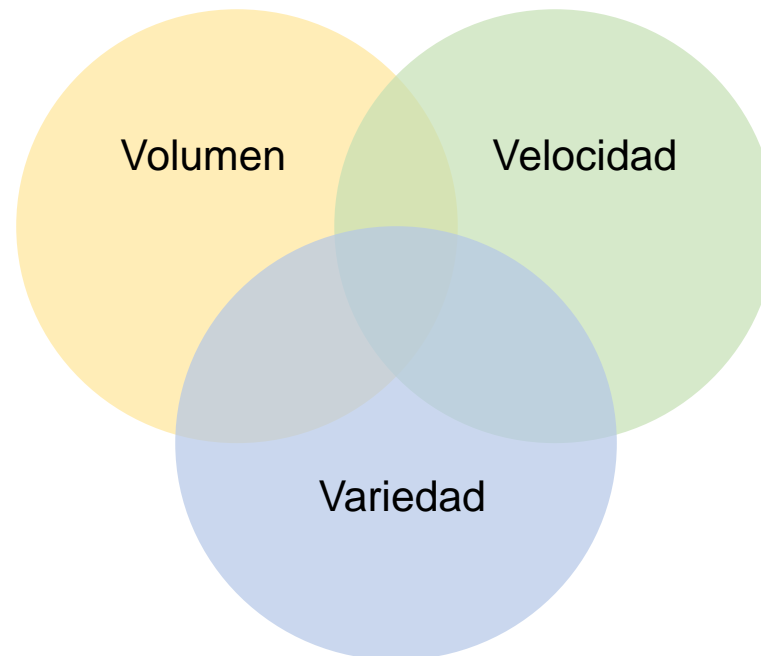
- Ingeniero de Telecomunicaciones.
- Programa de Desarrollo Directivo.
- Maestría en BI y Big Data.
- Especialización en Visualización de datos.

Introducción

La evolución de las 3 V's



La evolución de las 3 V's

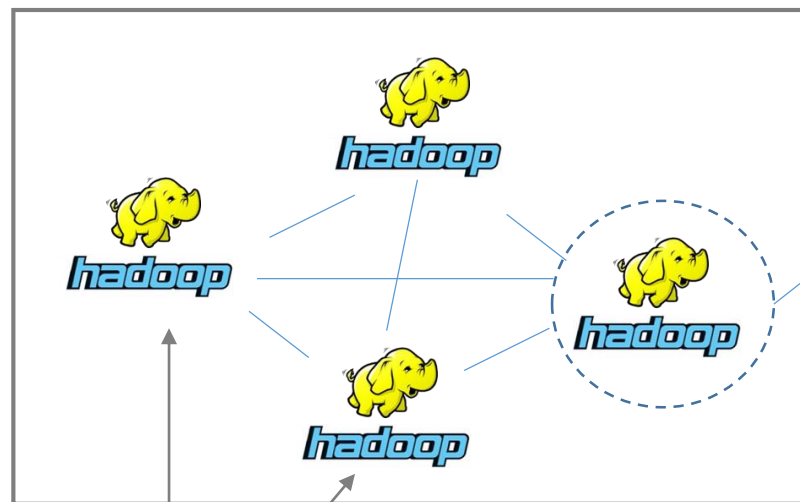


VERACIDAD

VALOR

Complejidad del entorno

CLUSTER HADOOP



10x – 100x nodos.



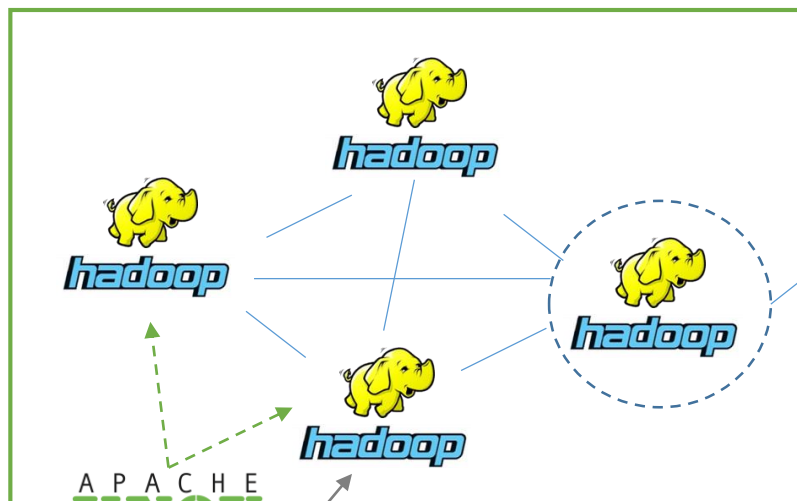
100x usuarios /
10x perfiles.



10x componentes /
100x activos.

Punto único de acceso

CLUSTER HADOOP



10x – 100x nodos.



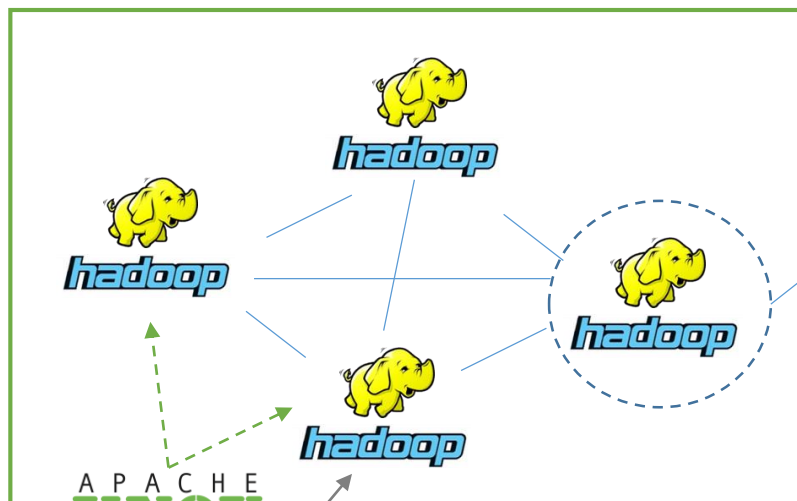
100x usuarios /
10x perfiles.



10x componentes /
100x activos.

Gestión centralizada de la seguridad

CLUSTER HADOOP



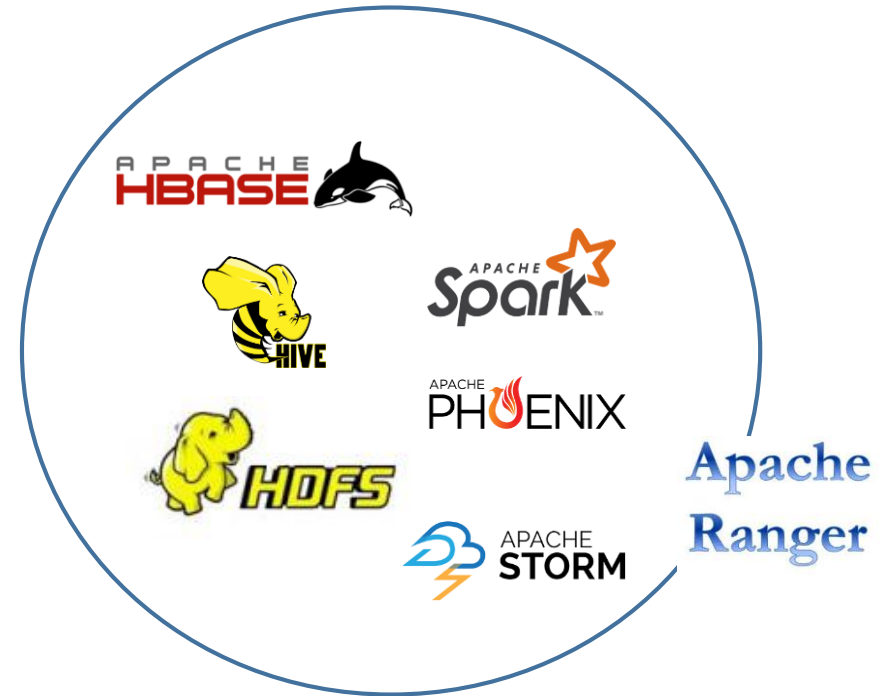
10x – 100x nodos.

100x usuarios /
10x perfiles.



10x componentes /
100x activos.

Gestión centralizada de la seguridad



10x componentes /
100x activos.

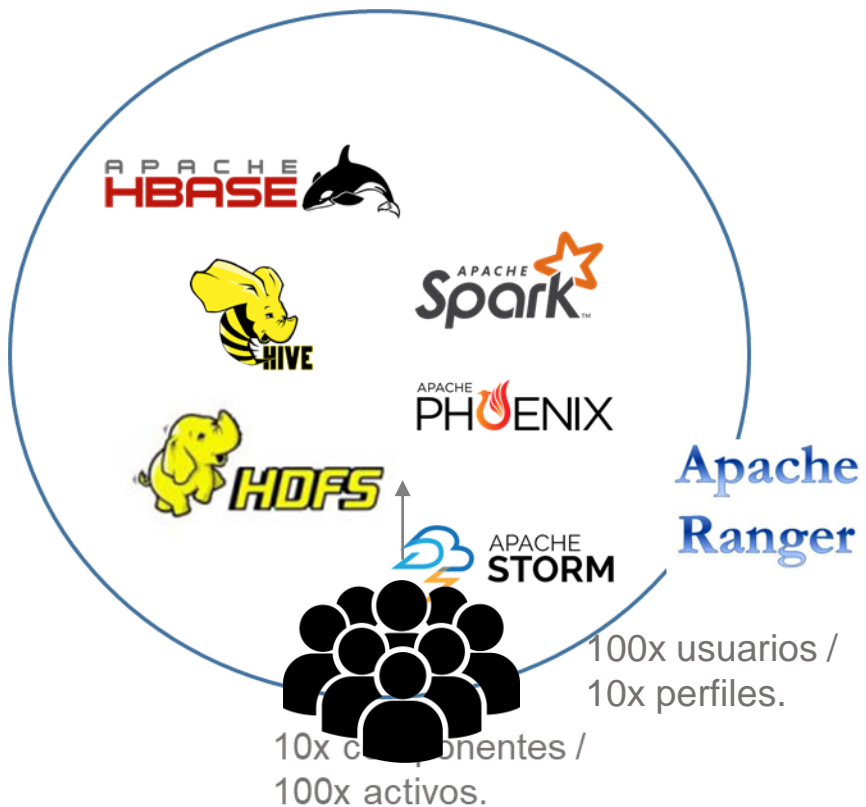
Gestión centralizada de la seguridad



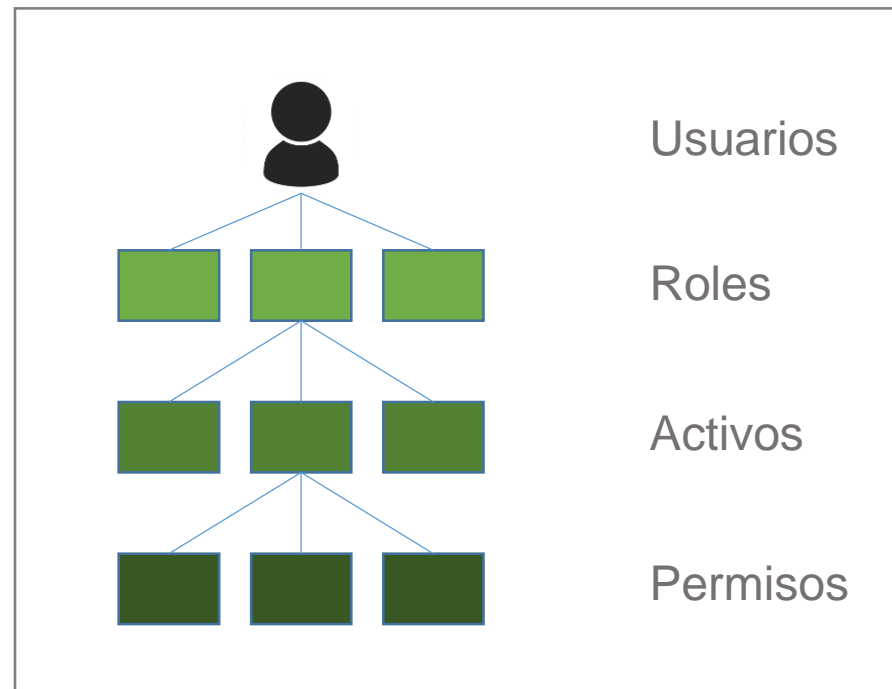
10x componentes /
100x activos.

- Objetivo: proveer seguridad al ecosistema Hadoop.
- Gestión centralizada de la seguridad mediante UI.
- Métodos de autorización estándar para todos los componentes.
- Logs de auditoría centralizados para todos los componentes.

Gestión centralizada de la seguridad



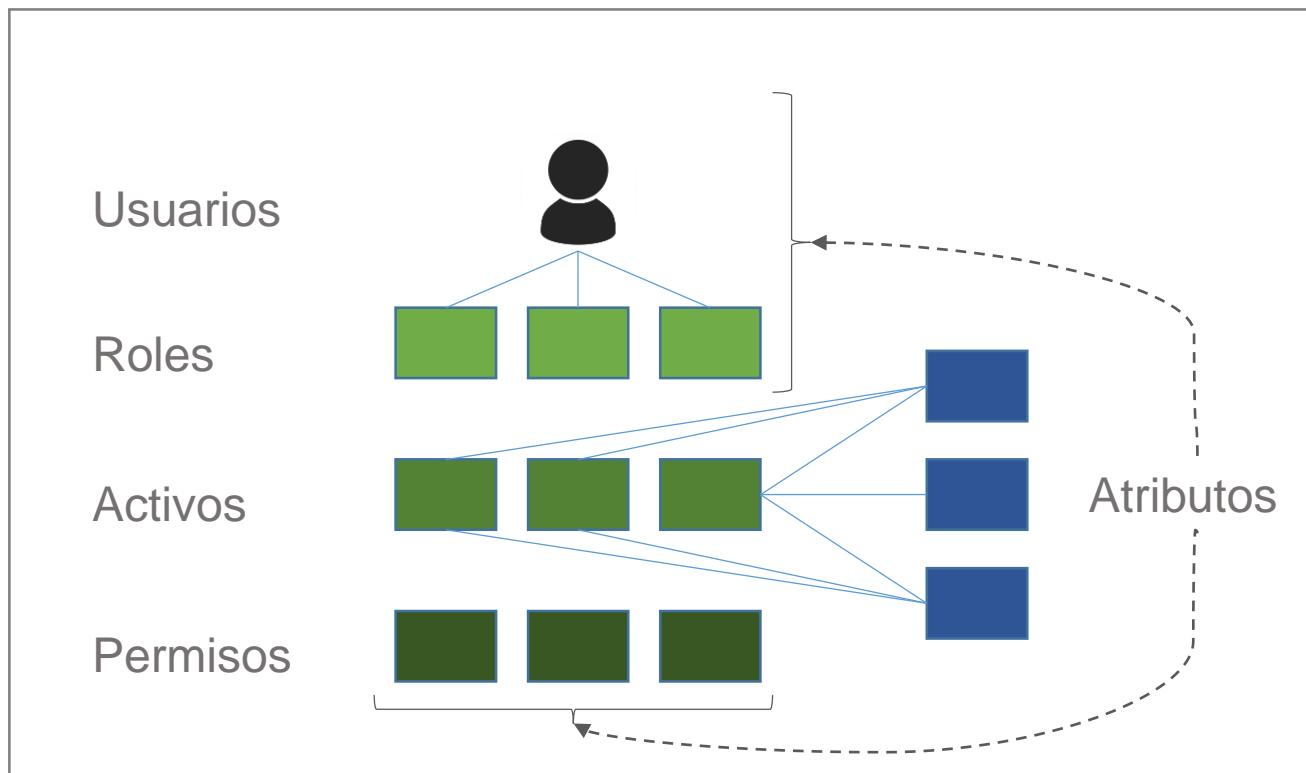
Control de acceso basado en roles



Complejidad → **Errores**

Simplificando el control de acceso

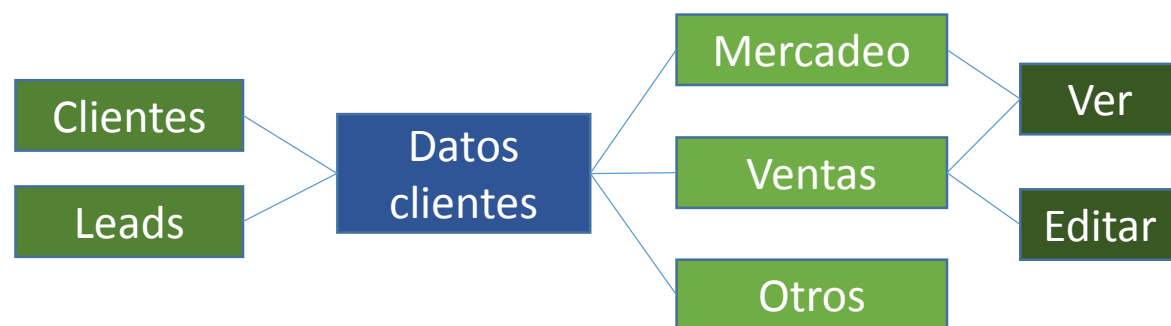
Control de acceso basado en atributos



Simplificando el control de acceso

Control de acceso basado en atributos

Ejemplo



Seguridad en Data Lake

Apache **Atlas**

Apache Ranger

Apache **Atlas**

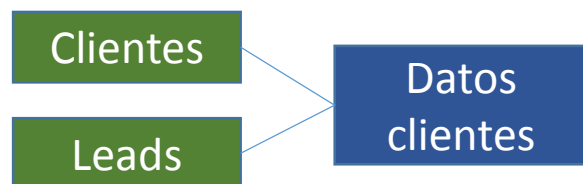
Framework de Gobierno del Dato y Gestión de Metadatos para Hadoop.

- Extenso conjunto de tipos de metadatos por defecto.
- Posibilidad de asignar etiquetas a distintos tipos de activos.
- Las etiquetas pueden incluir atributos.
- Propagación de las etiquetas vía linaje.
- Integración con Apache Ranger.

Seguridad basada en atributos

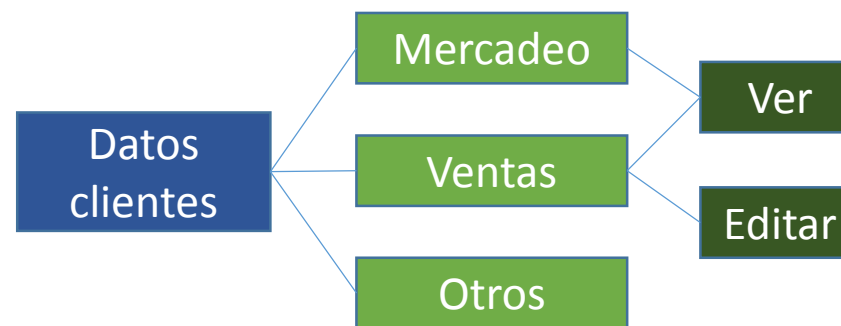
Apache Atlas

Asignación de etiquetas y atributos a los activos (tablas, campos, directorios, archivos, etc.).



Apache Ranger

Configuración de las políticas de control de acceso a nivel de activo o etiqueta.



Una misma etiqueta se puede asociar a activos de distintos componentes de Hadoop, facilitando la consistencia de la seguridad en todo el ciclo de vida del dato.

Seguridad a nivel de registro

Control de acceso basado en las características del usuario (grupo, por ejemplo) y el contexto de la ejecución.

CASOS DE USO

- Permitir a un médico ver información únicamente de sus pacientes.
- Permitir al Gerente Comercial ver únicamente los resultados de su zona.
- Permitir al Director Financiero ver los datos históricos sólo de su filial.

LIMITACIONES

- De momento, sólo disponible para Apache Hive.

Enmascarado dinámico de campos

Protección de datos sensibles en tiempo real mediante ofuscación dinámica.

BENEFICIOS

- No se requieren cambios en la información original.
- Evita la necesidad de generar copias protegidas de los datos sensibles.
- La información sensible nunca sale de la base de datos en el proceso.
- Políticas de enmascarado por defecto, con posibilidad de crear nuevas.
- Las políticas de enmascarado pueden asociarse a etiquetas.
- Las políticas se aplican en tiempo de compilación de la consulta, minimizando el overhead.

Enmascarado dinámico de campos

(Continuación)

CASOS DE USO

- Permitir a los empleados del área comercial ver la información completa de sus clientes.
- Mostrar a los agentes del Contact Center únicamente los 4 últimos dígitos de la cédula de los clientes.
- Devolver null en los campos de información sensible cuando sean consultados por un analista de riesgo de crédito.

LIMITACIONES

- De momento, sólo disponible para Apache Hive.

Enmascarado dinámico de campos



Policy Details :

Policy Type **Masking**

Policy ID **24**

Policy Name *

enabled

Audit Logging **YES**

TAG *

Description

Mask Conditions :

Select Group	Select User	Policy Conditions	Access Types	Select Masking Option	
<input type="text" value="x hr"/>	<input type="text" value="Select User"/>	<i>Add Conditions</i> +	HIVE	HIVE : Unmasked (retain original value)	
<input type="text" value="x analyst"/>	<input type="text" value="Select User"/>	expression : tagAttr.type = 'MRN'	HIVE	HIVE : Nullify	
<input type="text" value="x analyst"/>	<input type="text" value="Select User"/>	expression : tagAttr.type == 'Password'	HIVE	HIVE : Hash	

Restricción temporal de acceso

Establecimiento de políticas de acceso a datos con fecha de vencimiento.

BENEFICIOS

- Permite establecer una fecha límite para la consulta de información histórica a determinados grupos de usuarios.
- Emplea la etiqueta reservada *EXPIRES_ON* junto con el atributo *expiry_date*.
- La política se crea una única vez en Apache Ranger.

Restricción temporal de acceso



tax_2009 (hive_table)

Tags	Attributes
EXPIRES_ON	expiry_date:2016/12/31

tax_2010 (hive_table)

Tags	Attributes
EXPIRES_ON	expiry_date:2017/12/31

Apache
Atlas

Policy Details :

Policy Type **Access**

Policy ID **4**

Policy Name * access: EXPIRES_ON

enabled

TAG * **EXPIRES_ON**

Audit Logging **YES**

Description Policy for data with EXPIRES_ON tag

Apache Ranger

Deny Conditions :

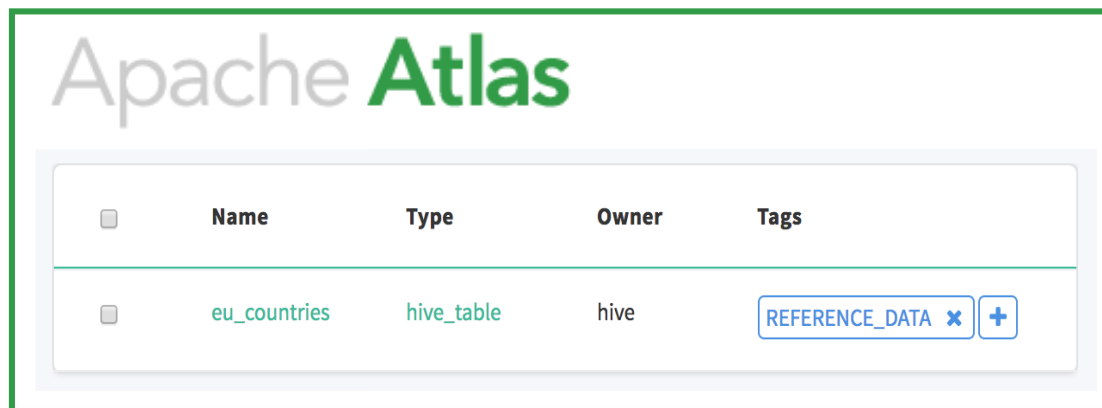
Select Group	Select User	Policy Conditions
public	Select User	accessed-after-expiry : yes

Protección de datos inmutables

Bloqueo de datos, impidiendo su modificación.

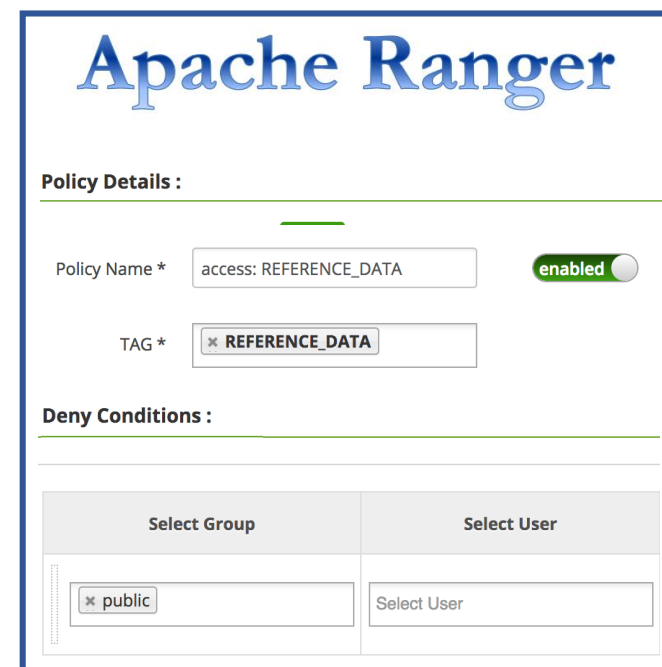
CASOS DE USO

- Protección de tablas paramétricas, evitando la generación de inconsistencias en los datos.



Apache Atlas

<input type="checkbox"/>	Name	Type	Owner	Tags
<input type="checkbox"/>	eu_countries	hive_table	hive	REFERENCE_DATA <input type="button" value="x"/> <input type="button" value="+"/>



Apache Ranger

Policy Details :

Policy Name * enabled

TAG *

Deny Conditions :

Select Group	Select User
<input type="text" value="public"/>	<input type="text" value="Select User"/>

Condiciones de acceso dinámicas

Establecer políticas de acceso basadas en atributos dinámicos.

CASOS DE USO



- Permitir acceder a la información sensible únicamente desde la IP corporativa.
- Bloquear el acceso a ciertos datos fuera del horario laboral.

User and Group Permissions :

Permissions	Select Group	Select User	Policy Conditions	Permissions	Delegate Admin
	<input type="text" value="hbaseuser"/>	<input type="text" value="Select User"/>	Add Conditions +	Publish	<input type="checkbox"/>

add/edit conditions

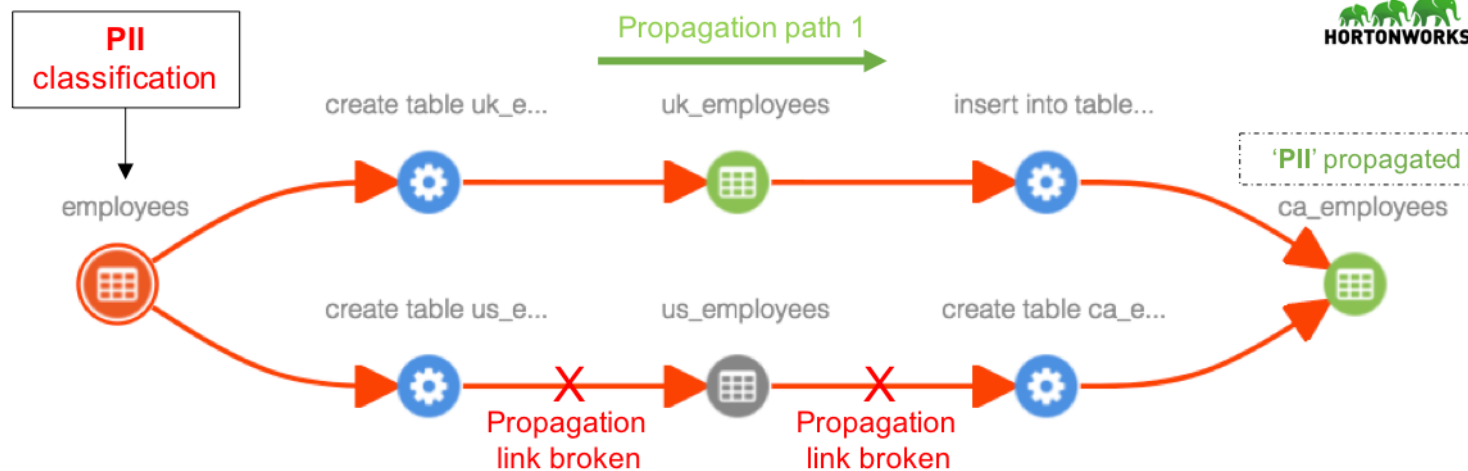
IP Address Range :



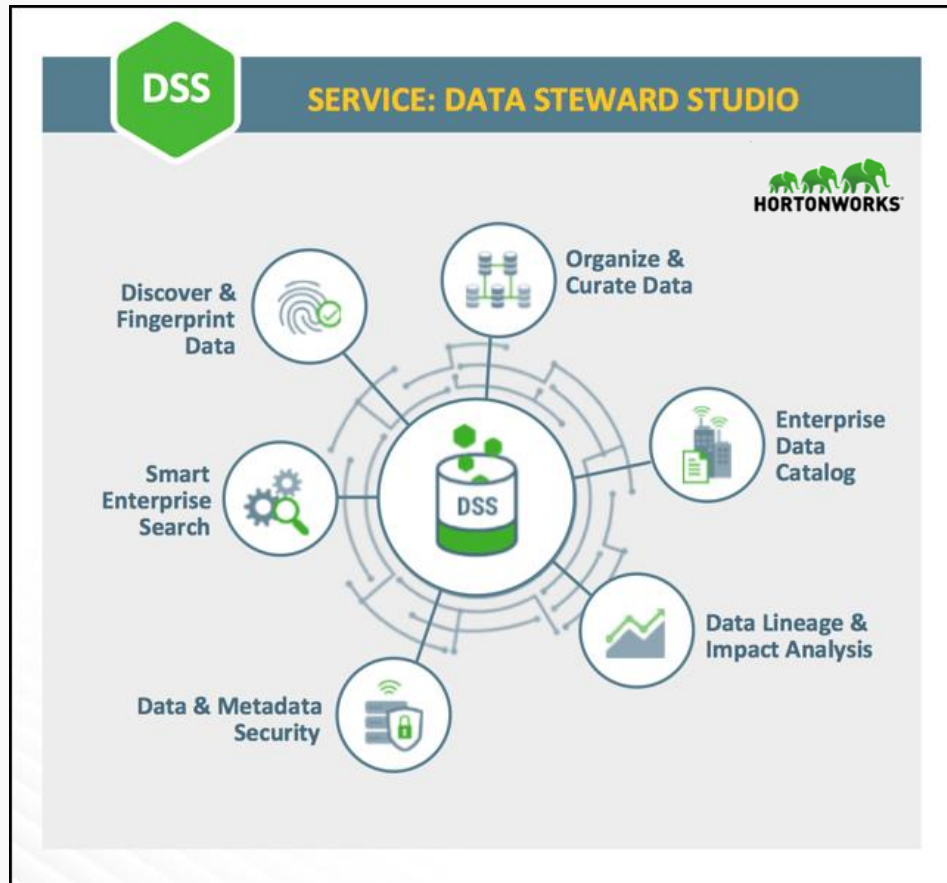
Propagación automática de tags

BENEFICIOS

- Las etiquetas se propagan de forma automática a los activos derivados.
- Desde la interfaz gráfica de Apache Atlas se puede eliminar la propagación.



Data Steward Studio

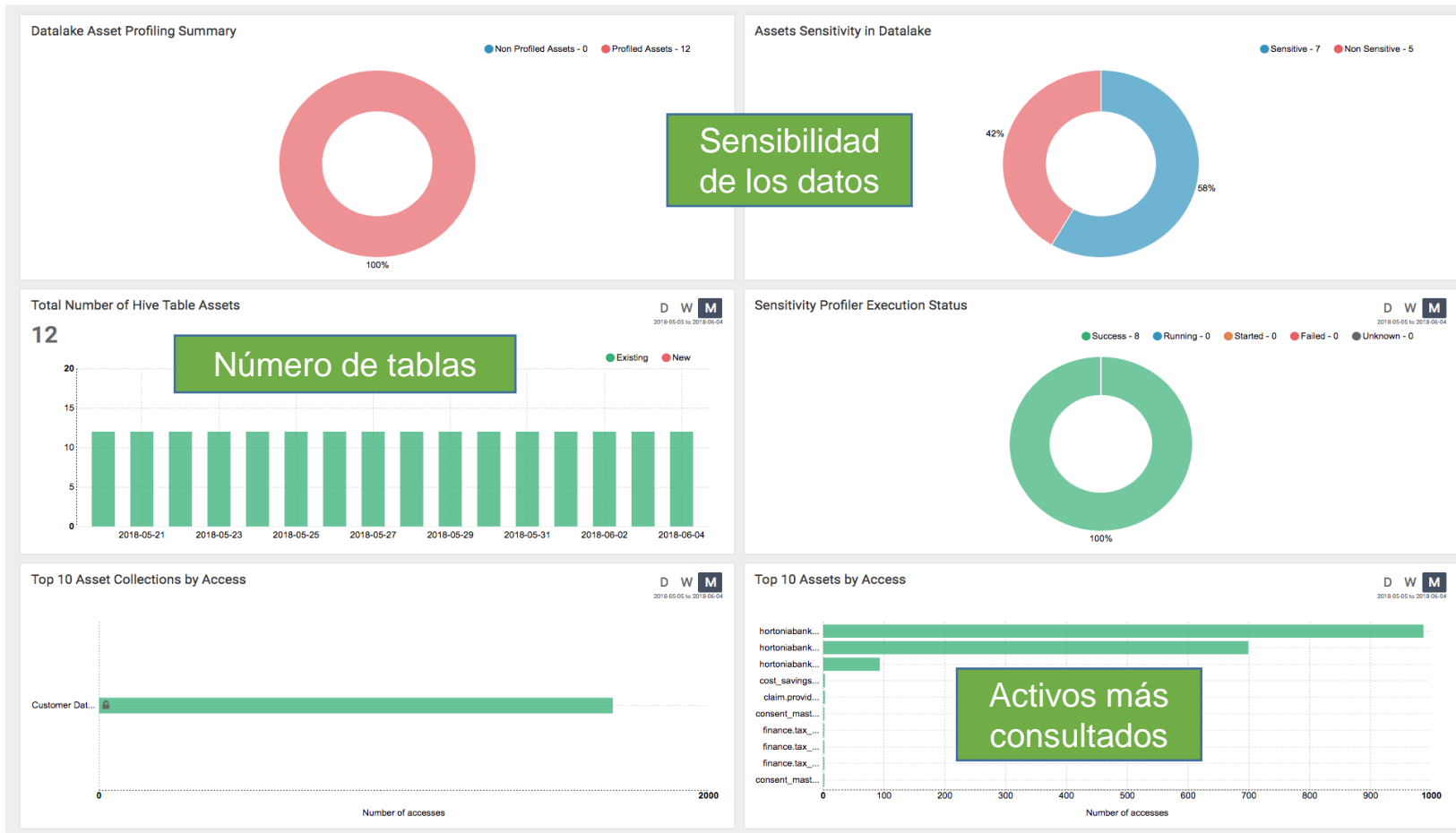


Data Steward Studio

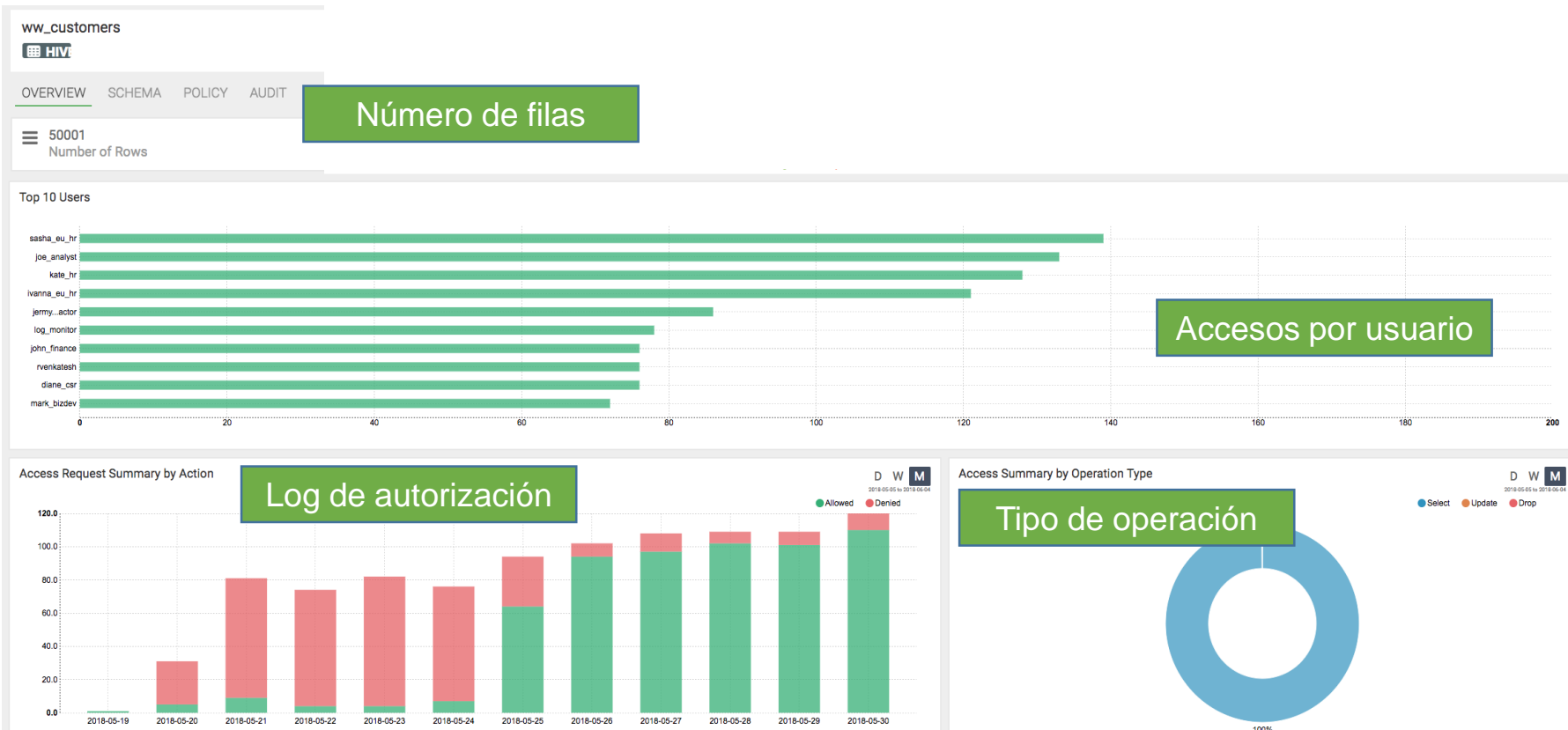
Data Steward Studio (DSS) es una plataforma desarrollada por Hortonworks con el objetivo de proporcionar una consola unificada para el gobierno y la protección de la información en el Data Lake.

- Integrada con Apache Atlas y Apache Ranger.
- Facilita el perfilado y análisis de la información.
- Interfaz única para el análisis de la clasificación de los datos y las políticas de seguridad que les aplican.

Vista resumen del Data Lake



Vista 360 de un activo (tabla de Hive)



Vista resumen de seguridad

us_customers

HIVE

OVERVIEW
SCHEMA
POLICY
AUDIT

Resource Based Policies

Políticas aplicadas directamente al activo

Policy ID	Policy Name	Status	Audit Logging	Group	Users
40	all - database, table, column	ENABLED	ENABLED	public	hive, ambari-qa
42	access: us_customers_table	ENABLED	ENABLED	us_employee, dpo, etl, public	hive
48	mask : nationalid show last 4	ENABLED	ENABLED	analyst	--
49	mask: ccn show first 4	ENABLED	ENABLED	analyst	--
50	mask: hash password	DISABLED	ENABLED	analyst	--
51	mask: redact street address	ENABLED	ENABLED	analyst	--
52	custom mask: randomize age	ENABLED	ENABLED	analyst	--
53	custom mask: retain birth year	ENABLED	ENABLED	analyst	--

Tag Based Policies

Políticas aplicadas por etiquetas

Policy ID	Policy Name	Tags	Status	Audit Logging	Group	Users
15	access: EXPIRES_ON	EXPIRES_ON	ENABLED	ENABLED	public, etl, dpo	--
17	access: PII	PII	ENABLED	ENABLED	hr, etl, dpo, dpadmin, csr, contractor, public, analyst	--
19	mask: PII	PII	ENABLED	ENABLED	hr, analyst	--

Dudas y preguntas

¡GRACIAS!