Apache Atlas: Governance for your data

Madhan Neethiraj
Director - Engineering, Atlas PMC

Suma Shivaprasad
Staff Engineer, Atlas PMC

Disclaimer

- *This document may contain product features and technology directions that are under development, may be under development in the future or may ultimately not be developed.*
- *Project capabilities are based on information that is publicly available within the Apache Software Foundation project websites ("Apache"). Progress of the project capabilities can be tracked from inception to release through Apache, however, technical feasibility, market demand, user feedback and the overarching Apache Software Foundation community development process can all effect timing and final delivery.*
- *This document's description of these features and technology directions does not represent a contractual commitment, promise or obligation from Hortonworks to deliver these features in any generally available product.*
- *Product features and technology directions are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.*
- *Since this document contains an outline of general product development plans, customers should not rely upon it when making purchasing decisions.*

Agenda

- Introduction
- Apache Atlas
  - Overview
  - Data Provenance/Lineage
  - Classification
  - Metadata Catalog Search
  - Architecture
- Demo
- Roadmap
- Q & A

HORTONWORKS

Apache Atlas: Introduction

Provides metadata-driven core foundational governance services for Hadoop and enterprise data ecosystem

Data Lineage/Provenance

- Captures data lineage across components

Data Classification

- Supports classification of data assets using tags – PII, PHI, PCI, EXPIRES_ON, CLAIMS, LIFE_INSURANCE

Metadata Catalog Search

- Free text search on metadata
- Advanced search using DSL

Integrations

- OOB real-time metadata and lineage ingestion with Hive, Sqoop, Storm/Kafka
- APIs for custom metadata ingestion
- Apache Ranger integration for classification based security

Metadata Repository

- Flexible metamodel to capture technical, business, operational metadata
- Out-of-box models for Hive, Storm, Sqoop, HDFS, Kafka, HBase
- APIs to register custom models

HORTONWORKS

Background: DGI Community becomes Apache Atlas

Dec 2014
DGI group Kickoff

May 2015
Apache Atlas Incubation

Aug 2016
HDP 2.5/ Apache 0.7 Foundation Release

Apr 2017
HDP 2.6/ Apache 0.8 Release

- #Committers – 35
- Code contributors from
  - Hortonworks, IBM, Aetna, Merck, Target

Apache 0.8/HDP 2.6

- Simplified Search UI
- Simplified APIs
- Classification-based security for HDFS, Kafka, HBase
- Knox SSO
- Performance/scalability improvements

Apache 0.7.1/HDP 2.5.3

- High availability support
- LDAP Authentication/Authorization
- Classification based security for Hive
- UI Redesign

** DGI: Data Governance Initiative*

HORTONWORKS

Apache Atlas : Lineage

Lineage

- Where does this data originate from (source/provenance)?
- **Upstream path**: Path through all data assets and processes leading up to current data asset

Impact

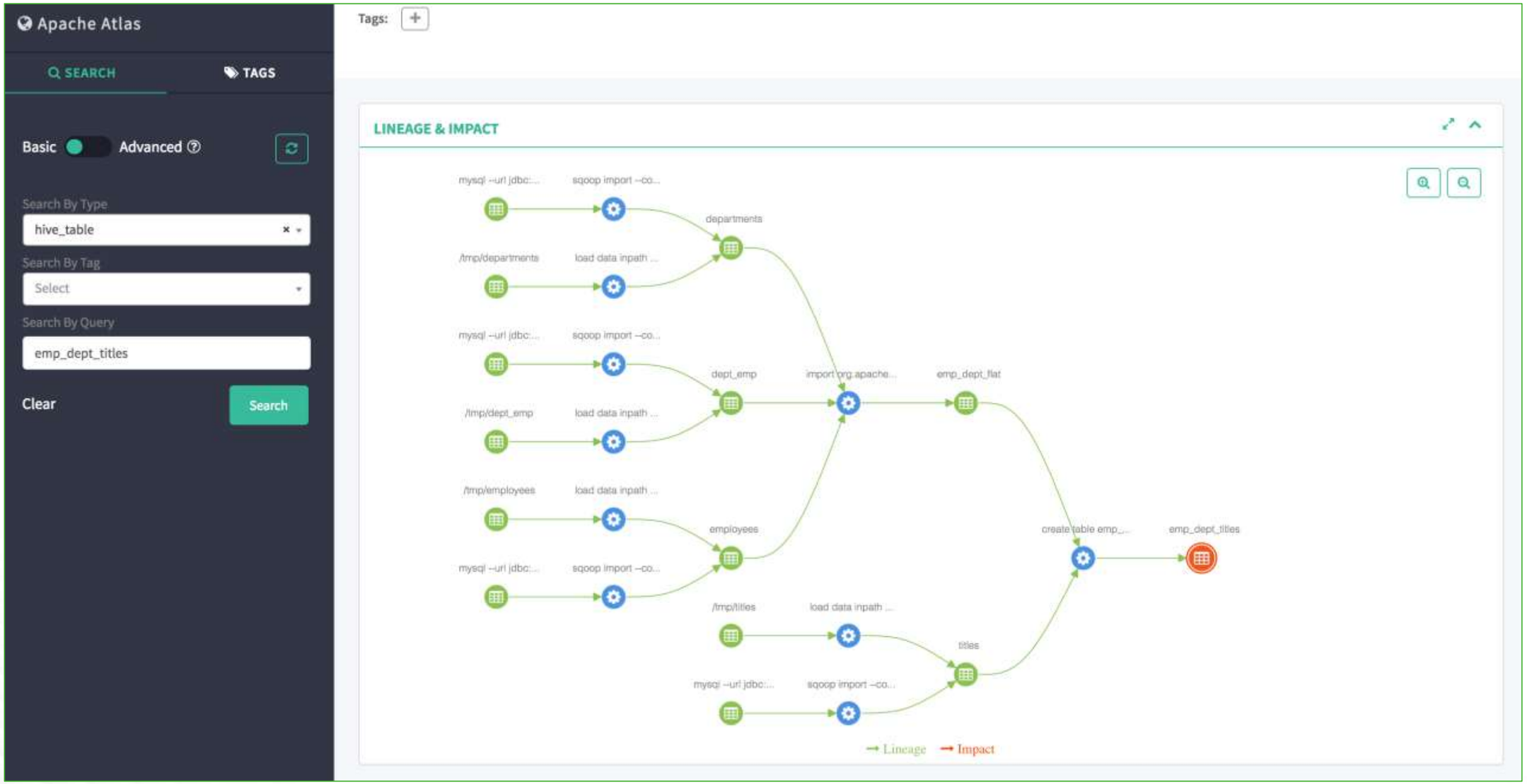
- How is this data being used ?
- What other data assets (derivative/dependent) does this impact?
- **Downstream path**: Path through all data assets and processes leading out of current data asset

Used for forensics

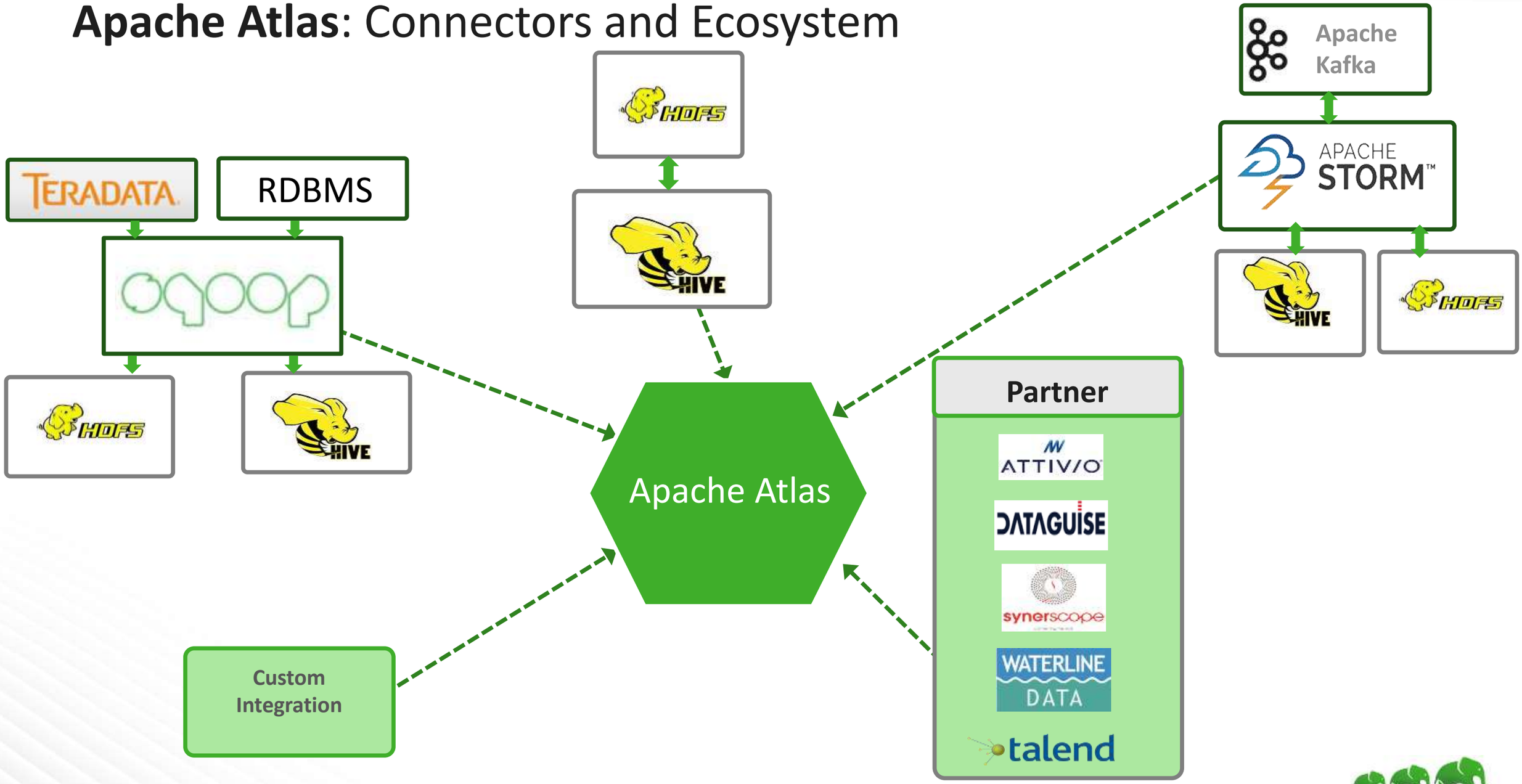
- Impact analysis
- Auditing and Compliance

HORTONWORKS

Apache Atlas: Lineage and Impact



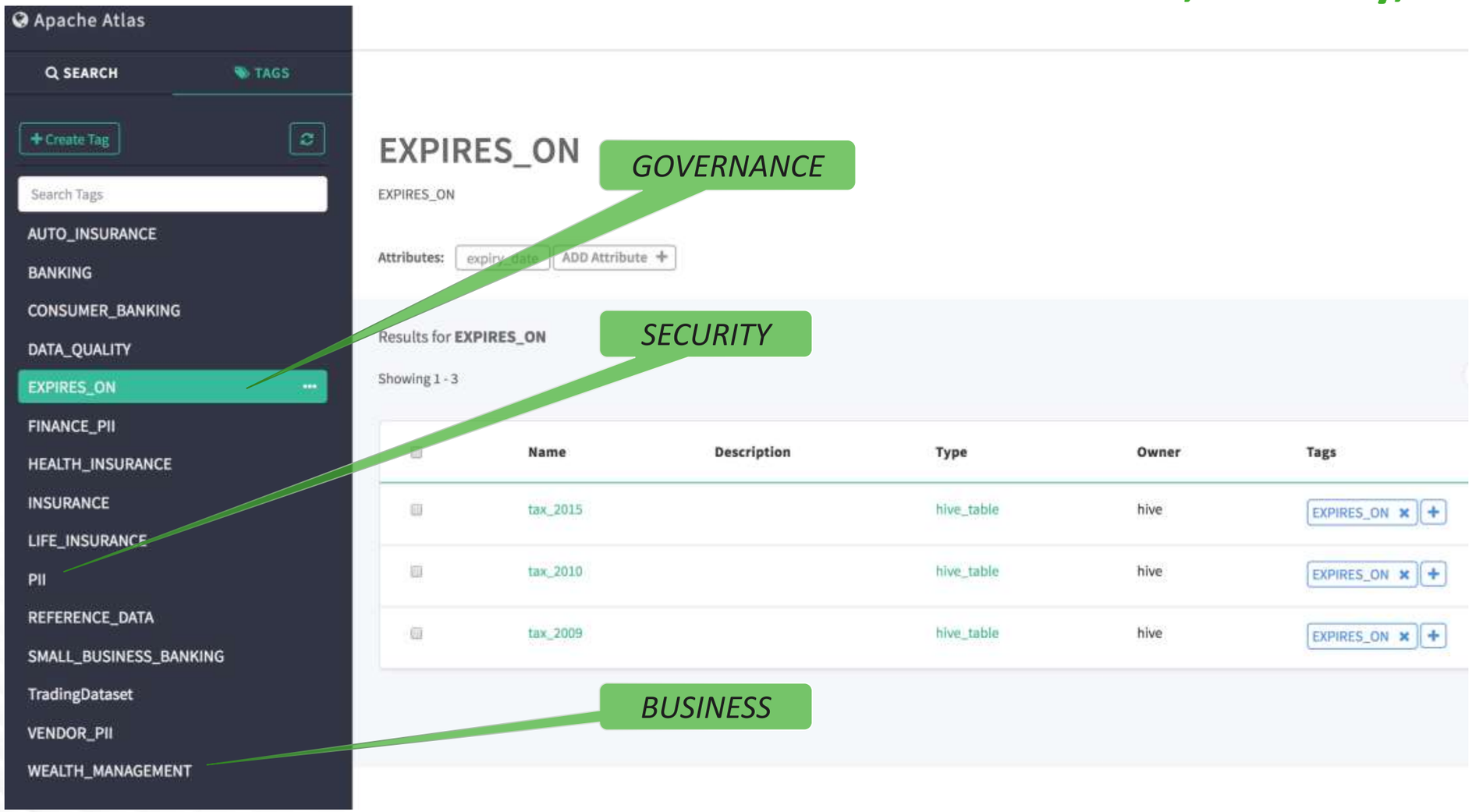
Apache Atlas: Connectors and Ecosystem



HORTONWORKS

Apache Atlas: Classification

- Categorize and curate data assets for easier discovery
- Associate context with data assets – Governance, Security, Business, …

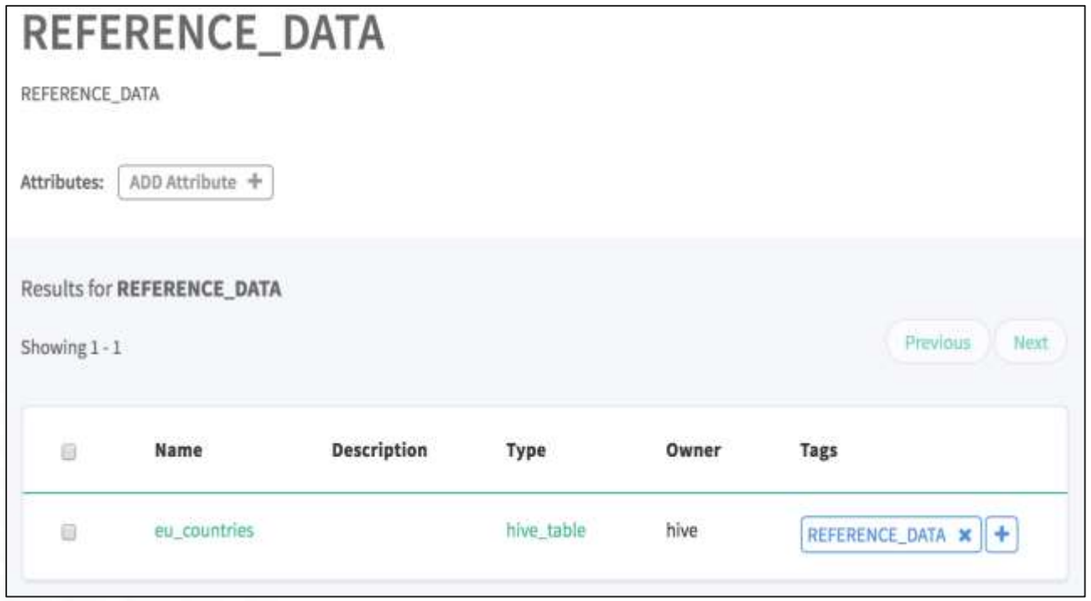


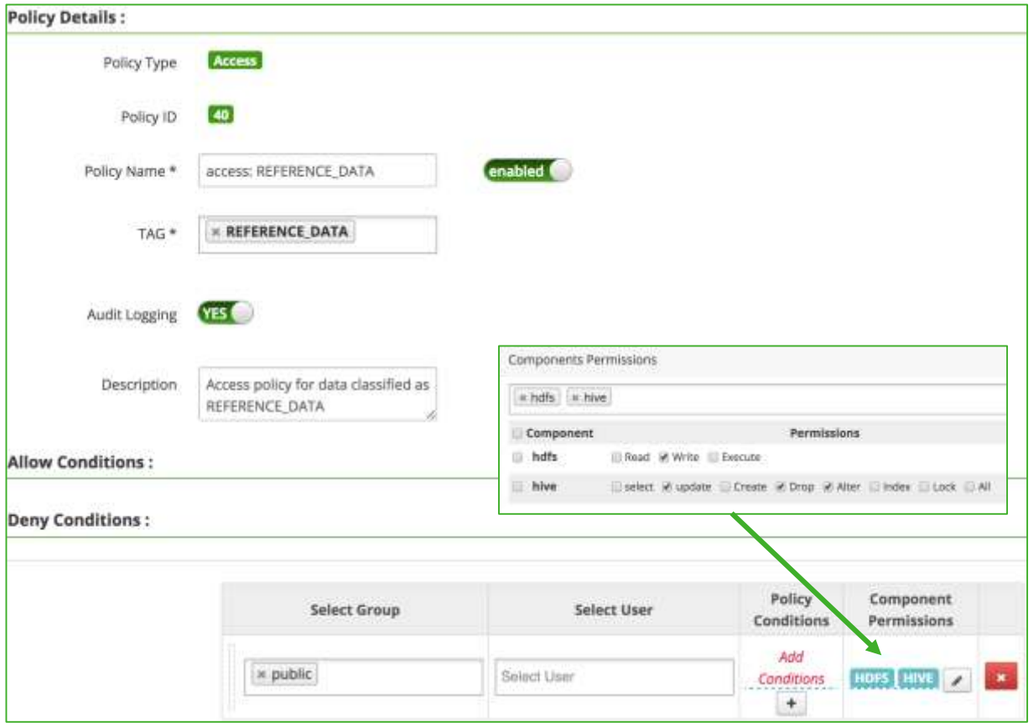
HORTONWORKS

Apache Atlas Classification : usecase – REFERENCE_DATA

Security policy enforcement for denying updates on immutable data assets

- **REFERENCE_DATA** classification associated with immutable hive_table **eu_countries**
- Apache Ranger policies block updates on the table for all users except admins

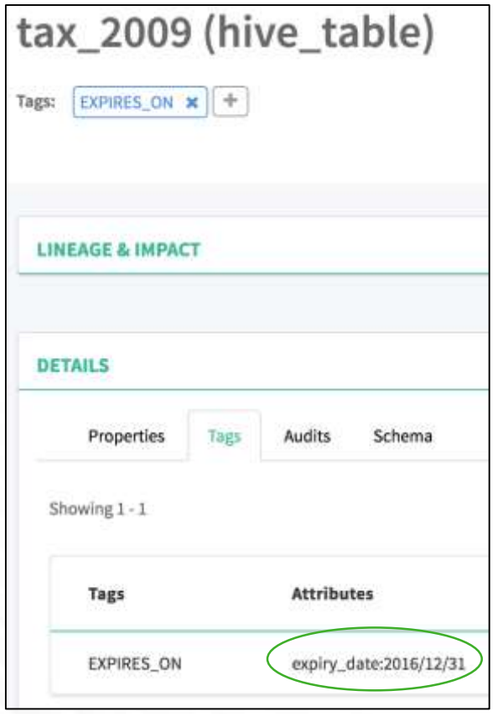


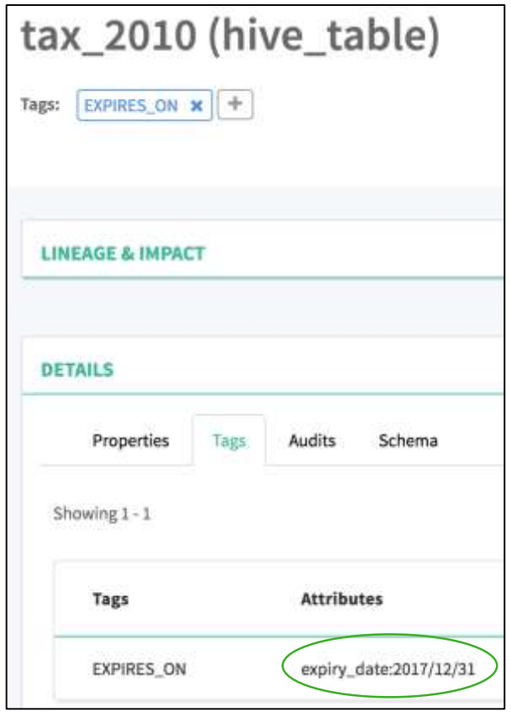


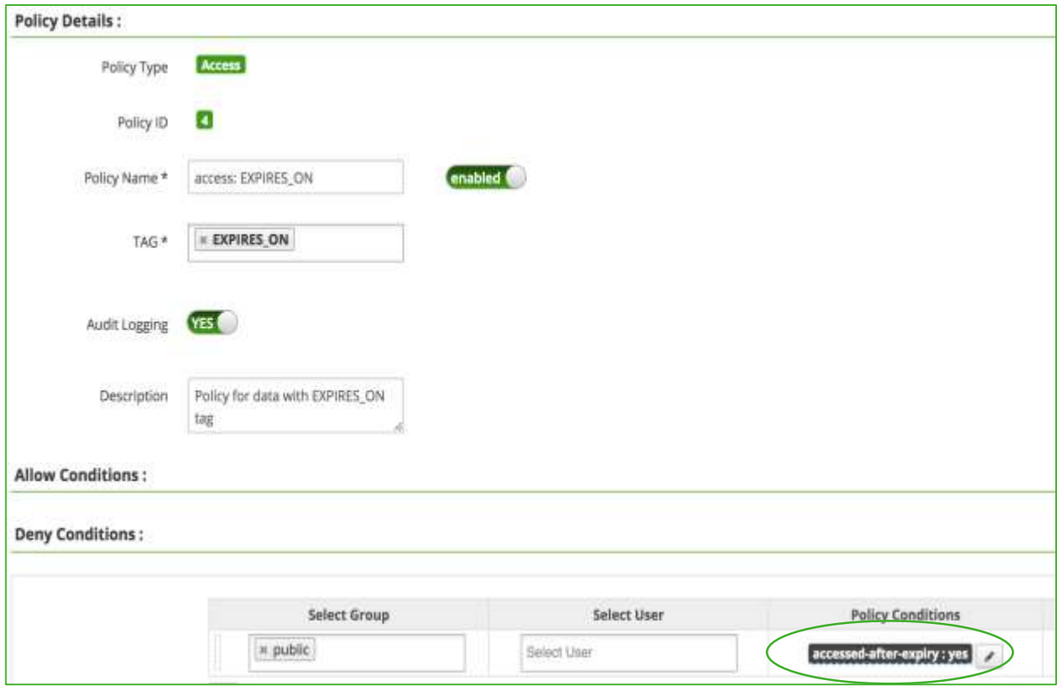
Apache Atlas Classification : usecase – access expiry

Data expiration

- **EXPIRES_ON** classification with attribute `expiry_date`
- tax_2009 table tagged with EXPIRES_ON(expiry_date=2016/12/31)
- tax_2010 table tagged with EXPIRES_ON(expiry_date=2017/12/31)
- Apache Ranger policies use the attribute to block access after expiry date



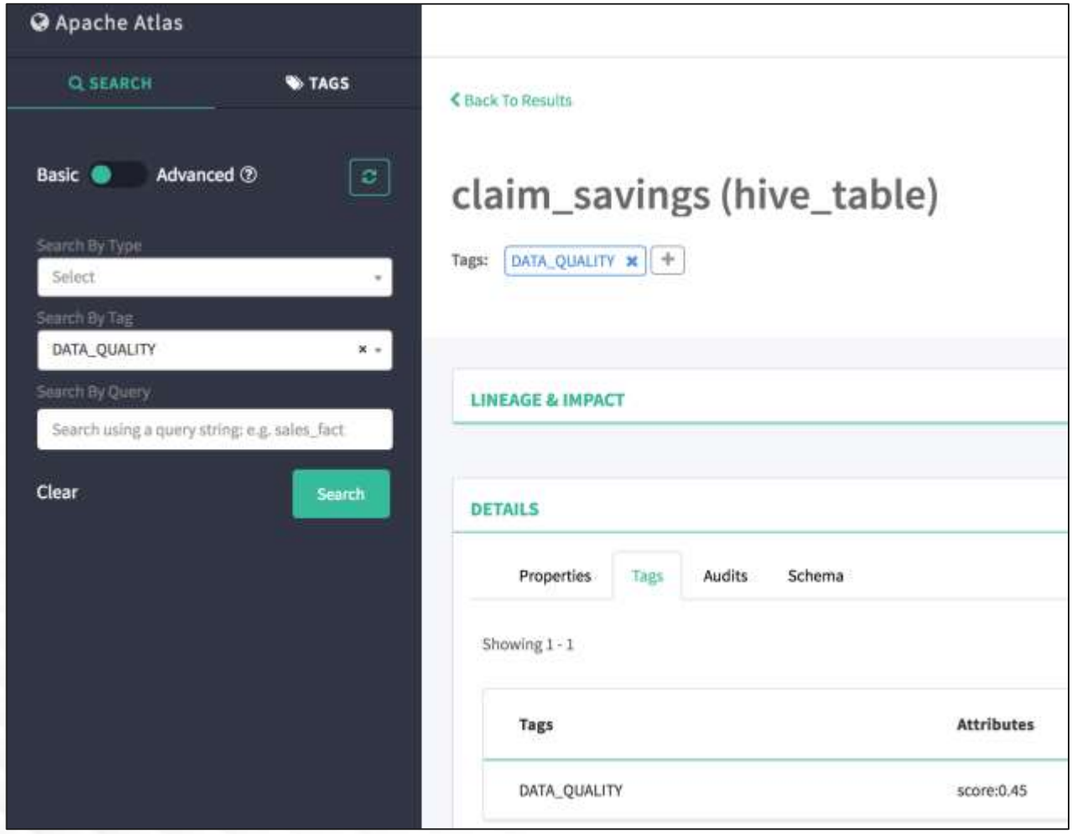


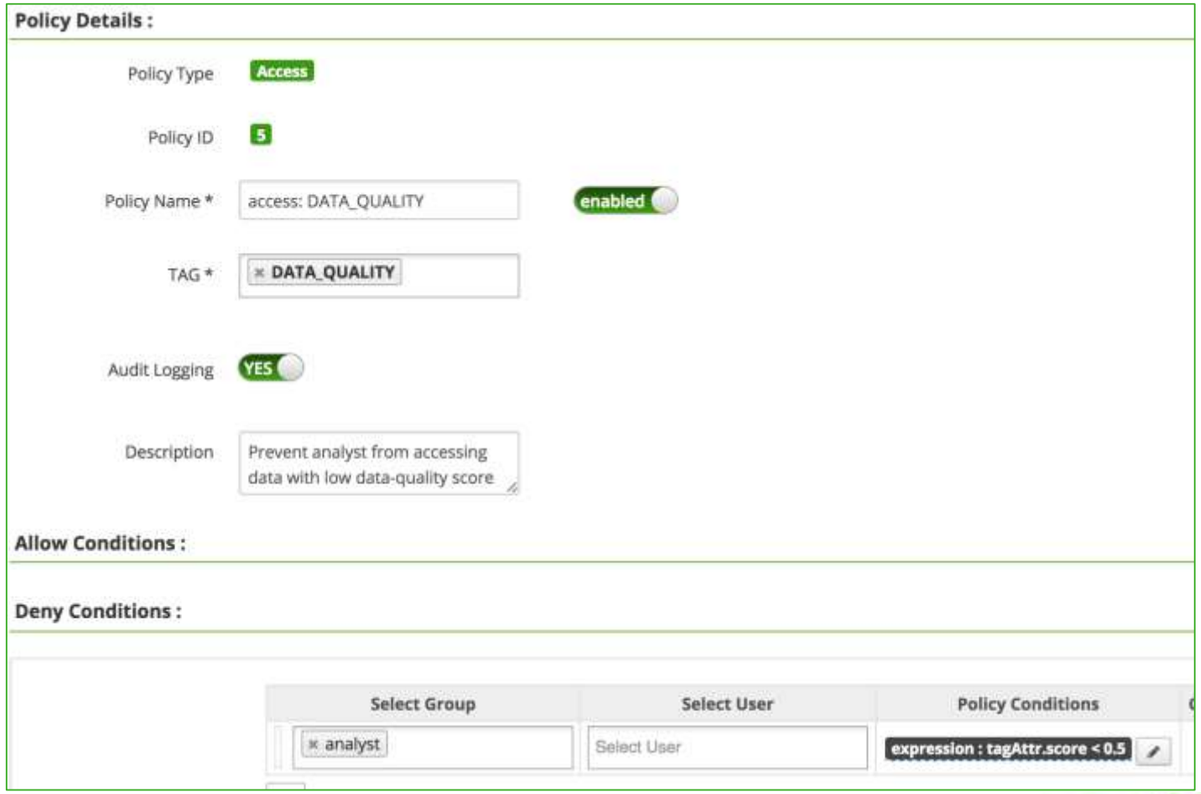


Apache Atlas Classification: usecase – attribute based authorization

Data quality

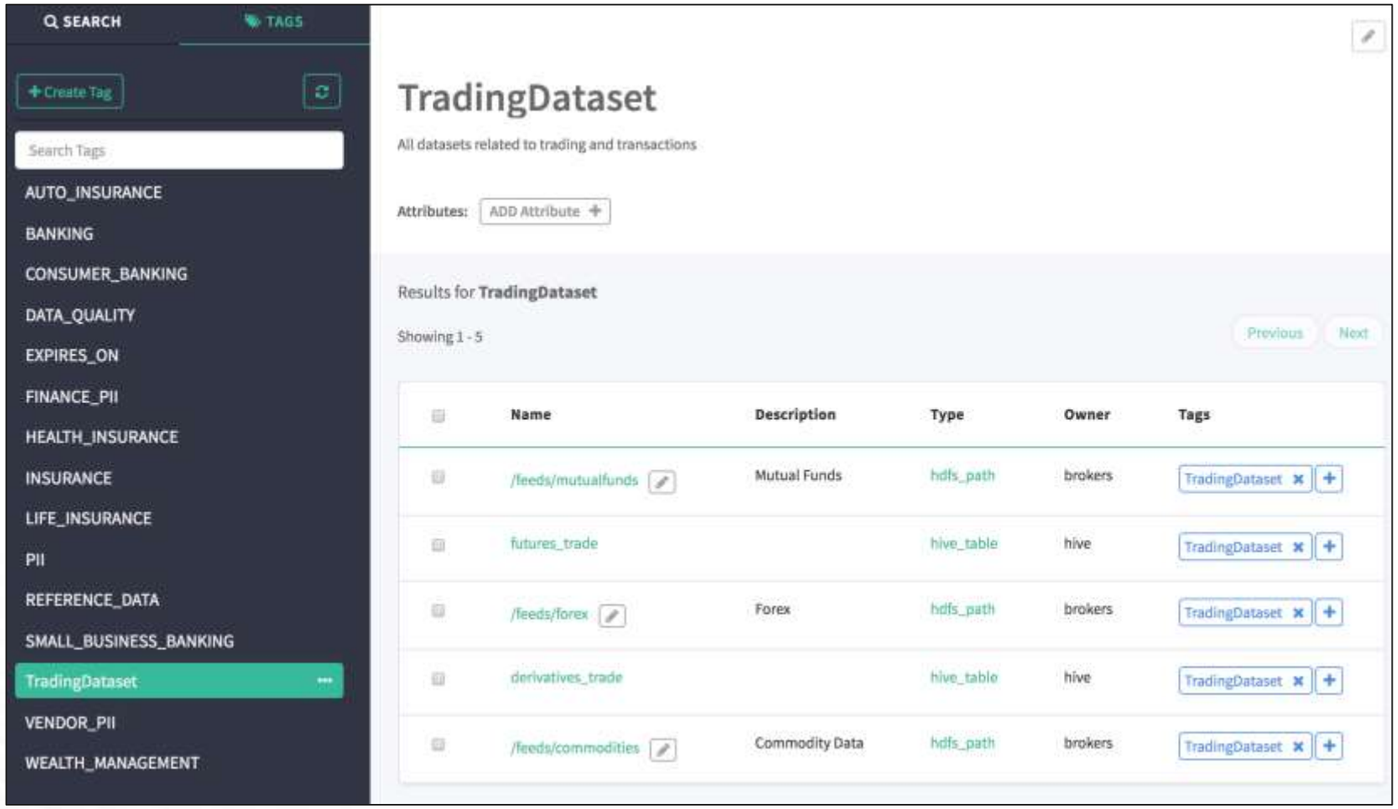
- Deny access to **analysts** group based on data quality threshold

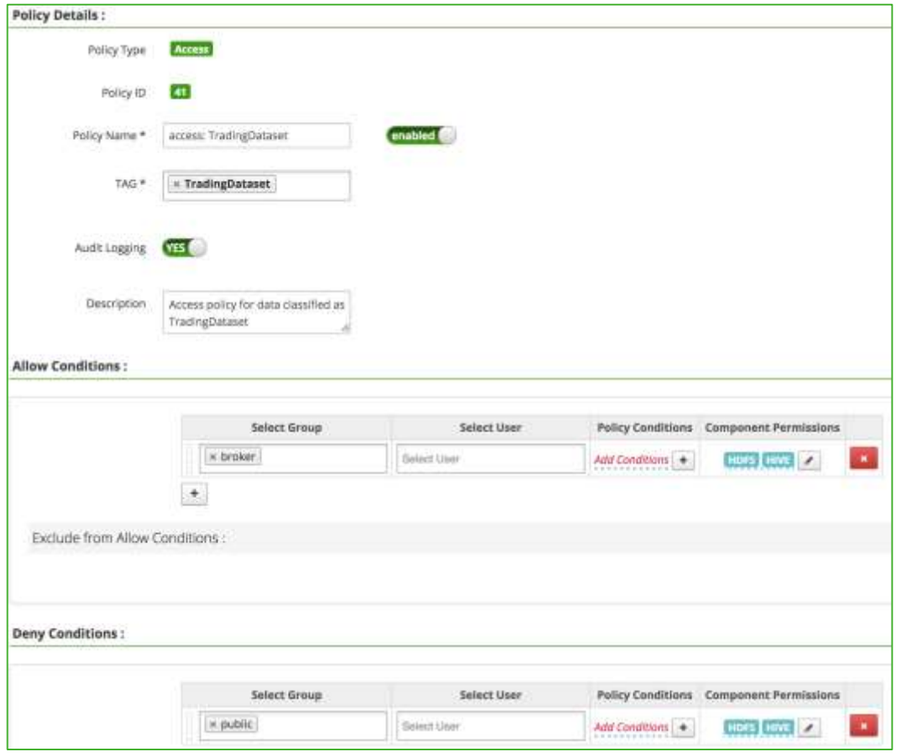




Apache Atlas Classification: usecase – cross component

Classification based security on cross-component data assets



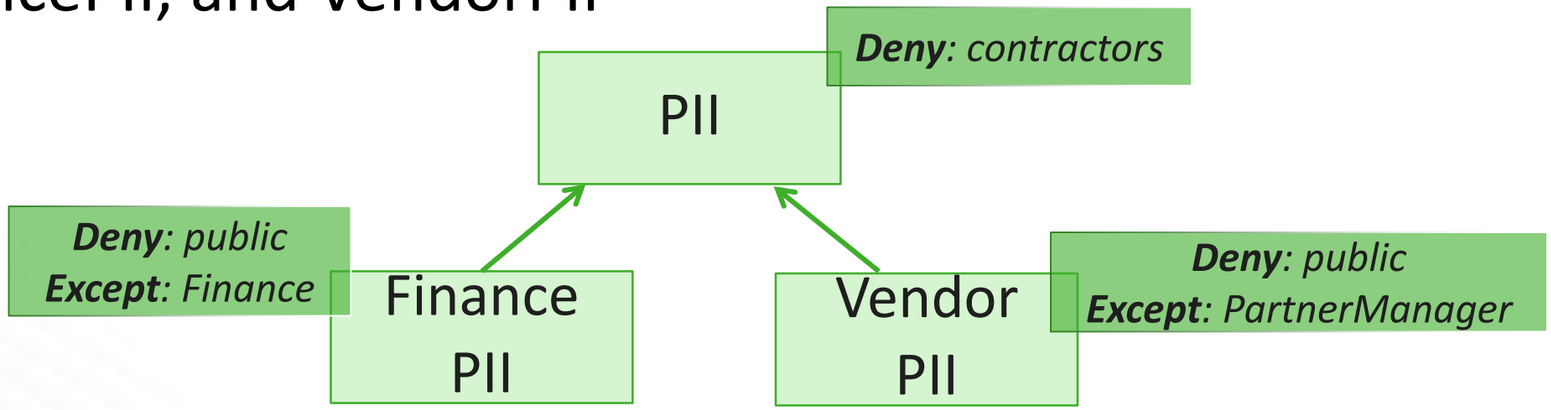


HORTONWORKS

Apache Atlas Classification: usecase - hierarchy

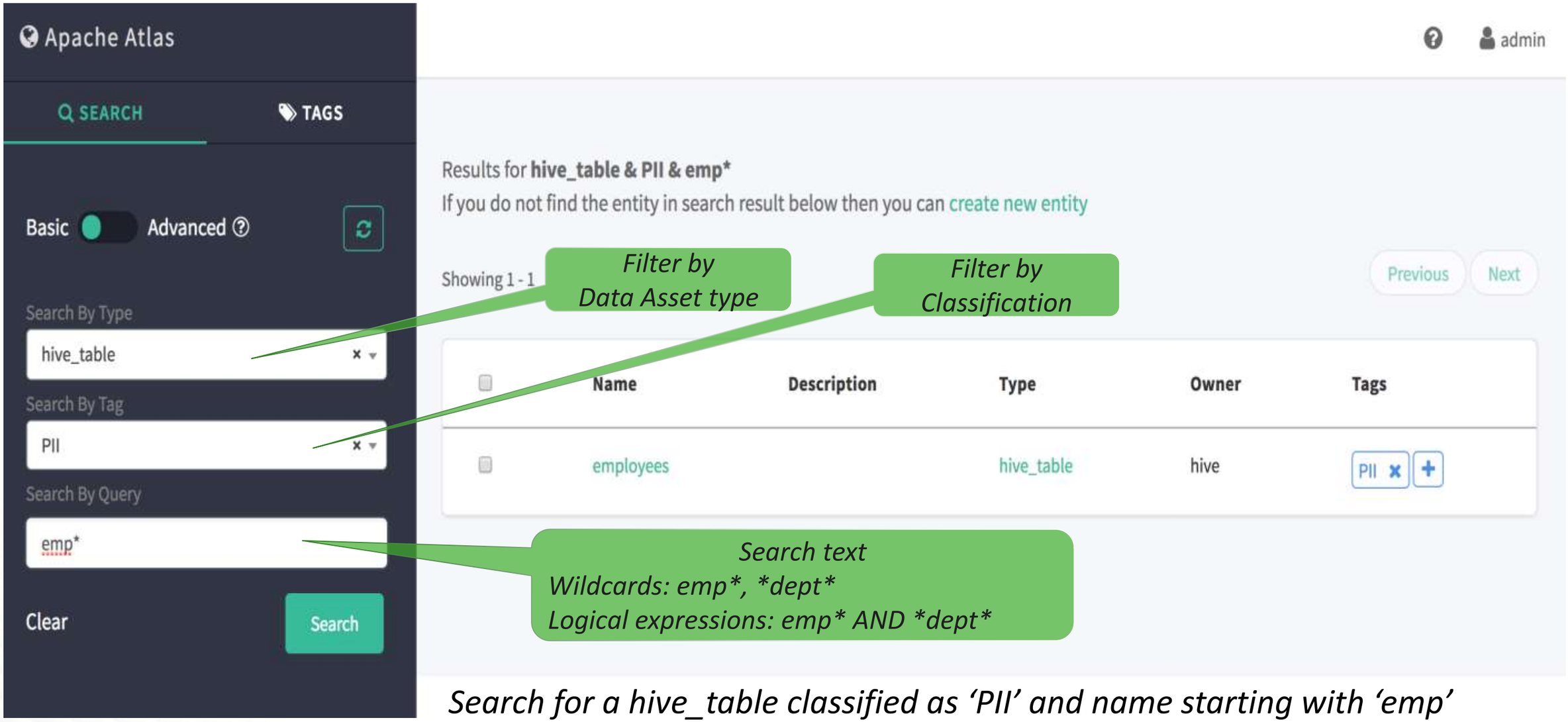
Security policy enforcement based on classification hierarchy

- Data assets classified as **PII** will be denied for all contractors
- Data assets classified as **FinancePII** will be denied for anyone not in Finance group
- Data assets classified as **VendorPII** will be denied for anyone not in PartnerManager group
- ..hence contractors will be denied access to data assets classified as PII, FinancePII, and VendorPII



HORTONWORKS

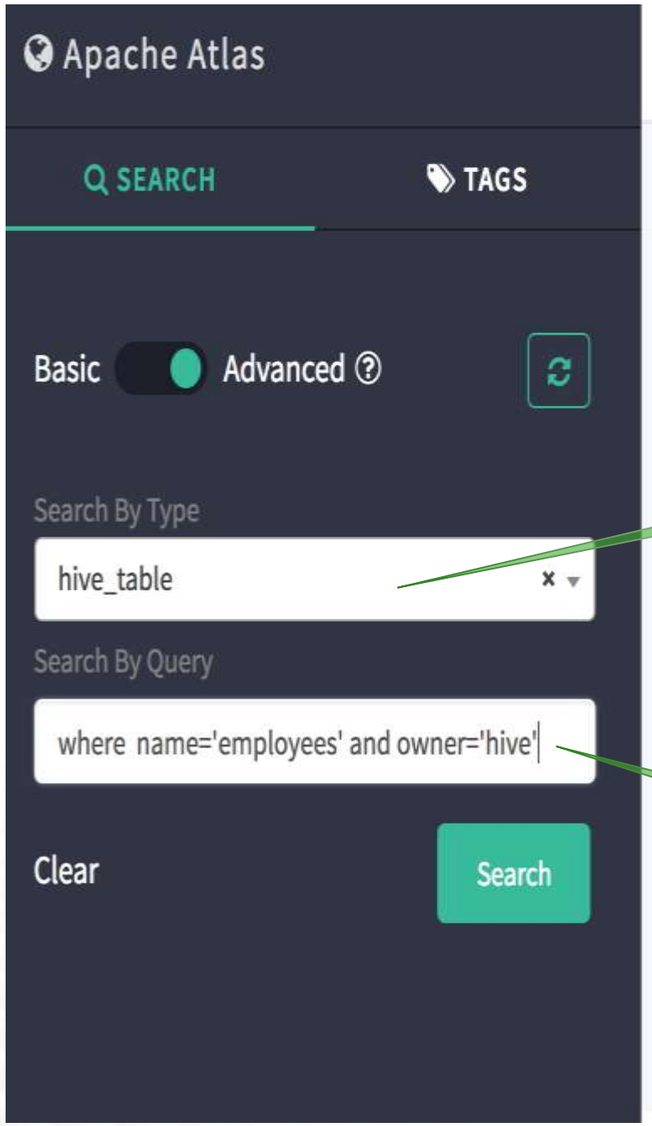
Metadata Catalog Search : Free Text



Search for a hive_table classified as 'PII' and name starting with 'emp'

HORTONWORKS

Metadata Catalog Search : Advanced



DSL search with SQL like syntax

Select columns from *impressions* table in *raw* database

hive_column where table.name='impressions' and table.db.name = 'raw'

admin

Results for **hive_table where name='employees' and owner='hive'**

If you do not find the entity in search result below then you can create new entity

Filter by Data asset type

Showing 1 - 1

Previous Next

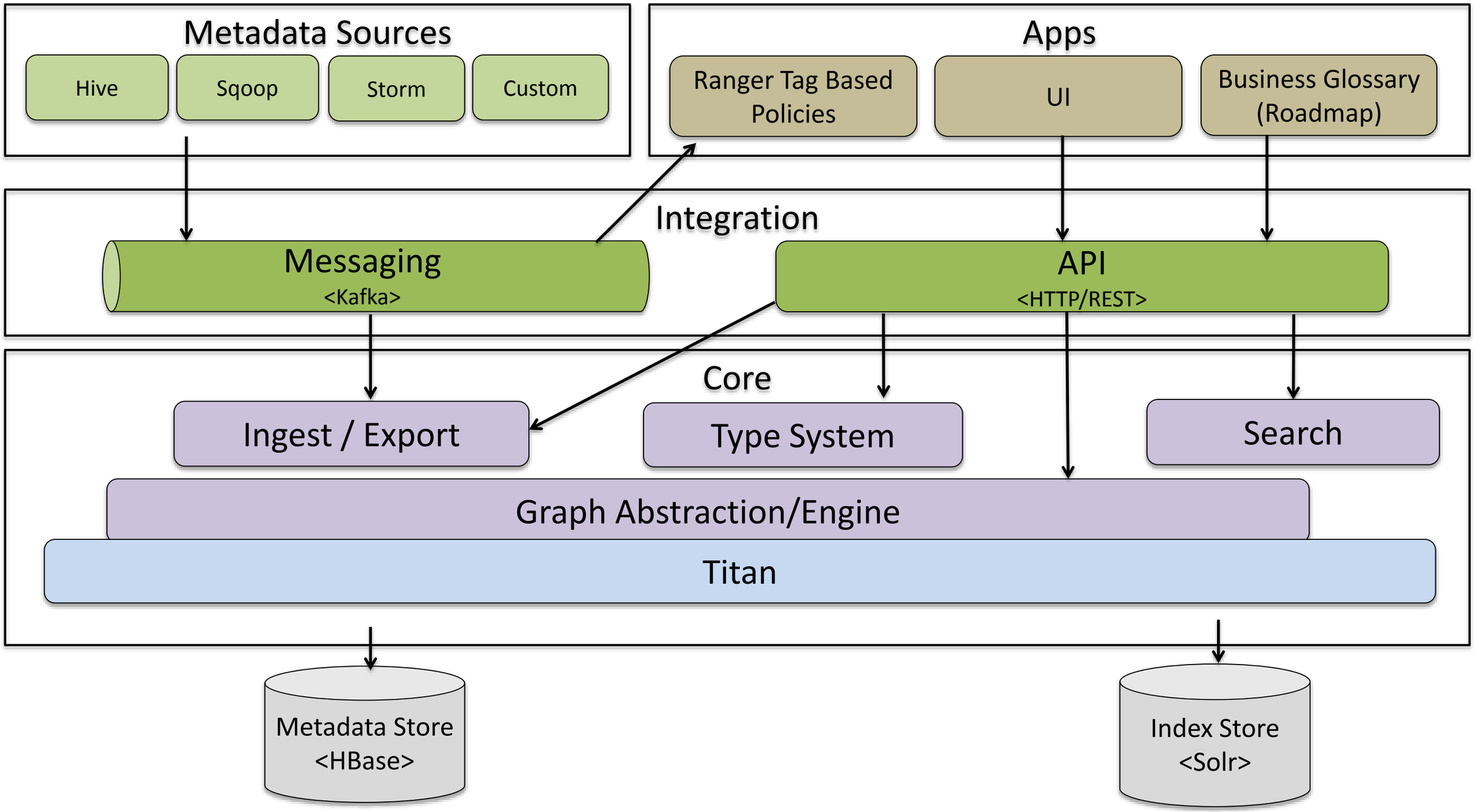
Name	Description	Type	Owner	Tags
employees		hive_table	hive	PII

DSL query string

Search for a hive_table named 'employees' and owner 'hive'

HORTONWORKS

Apache Atlas: Architecture



Apache Atlas : Roadmap

Business Taxonomy/Glossary

- Vocabulary for business users
- Categorized and curated by data stewards
- Association with physical assets like database, tables for effective data classification
- Discover data assets more naturally with glossary terms

Connectors

- Spark, NiFi, HBase

Hive Column-Level Lineage

Search/Filtering on Lineage

Export/Import of Atlas lineage and metadata

HORTONWORKS

Questions

HORTONWORKS

References

- Apache Atlas
  - http://atlas.apache.org
  - http://hortonworks.com/apache/atlas
- Apache Atlas community
  - https://community.hortonworks.com/spaces/64/governance-lifecycle-track.html?topics=Atlas&type=question
- Apache Ranger
  - http://ranger.apache.org
  - http://hortonworks.com/apache/ranger
  - https://cwiki.apache.org/confluence/display/RANGER/Tag+Based+Policies