

# 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



# 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

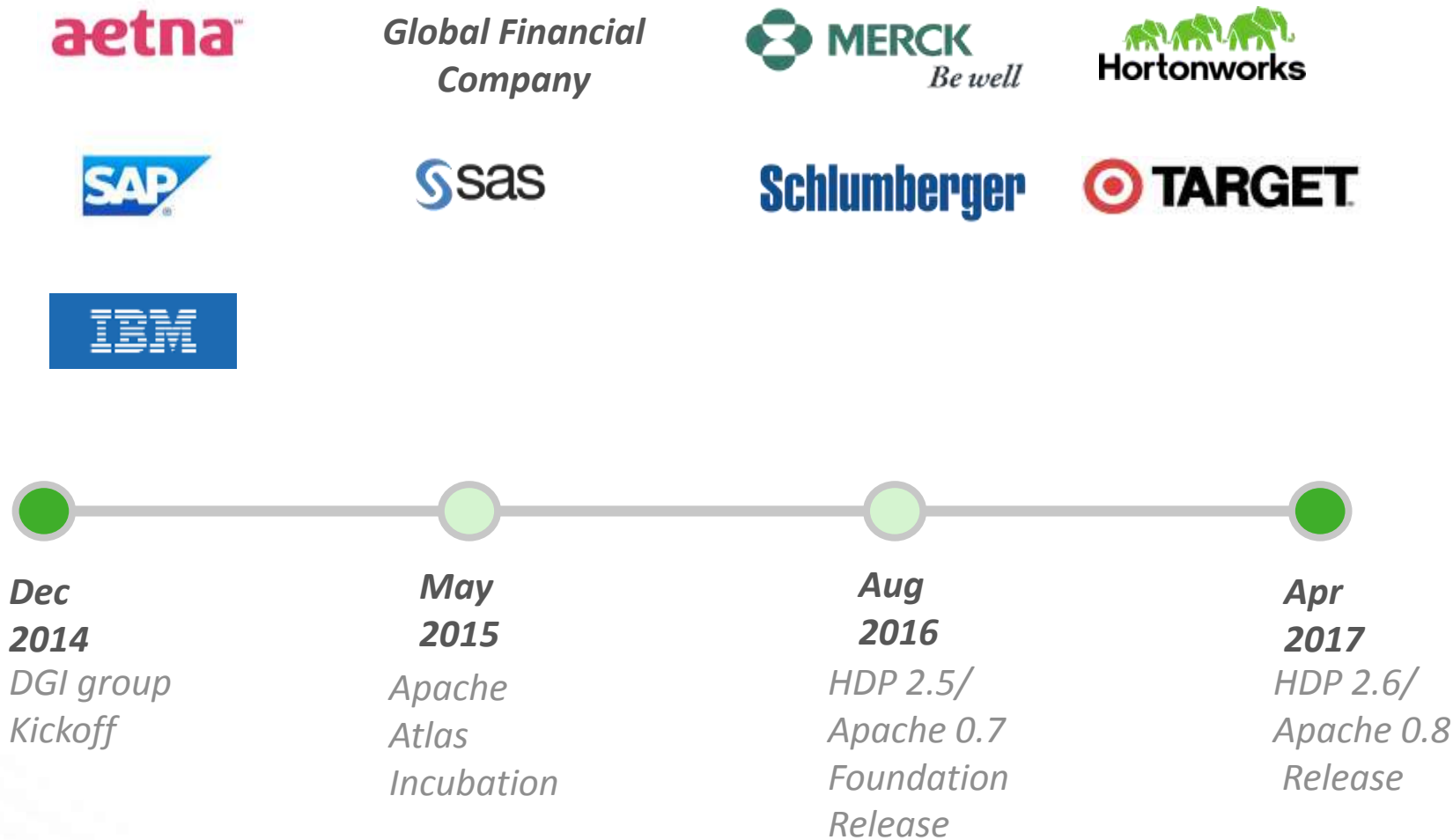
## Integrations

- 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

# Background: DGI Community becomes Apache Atlas



- #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

**HORTONWORKS**

Community Connection

294 Questions	536 Answers
------------------	----------------

\* DGI: Data Governance Initiative

# 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

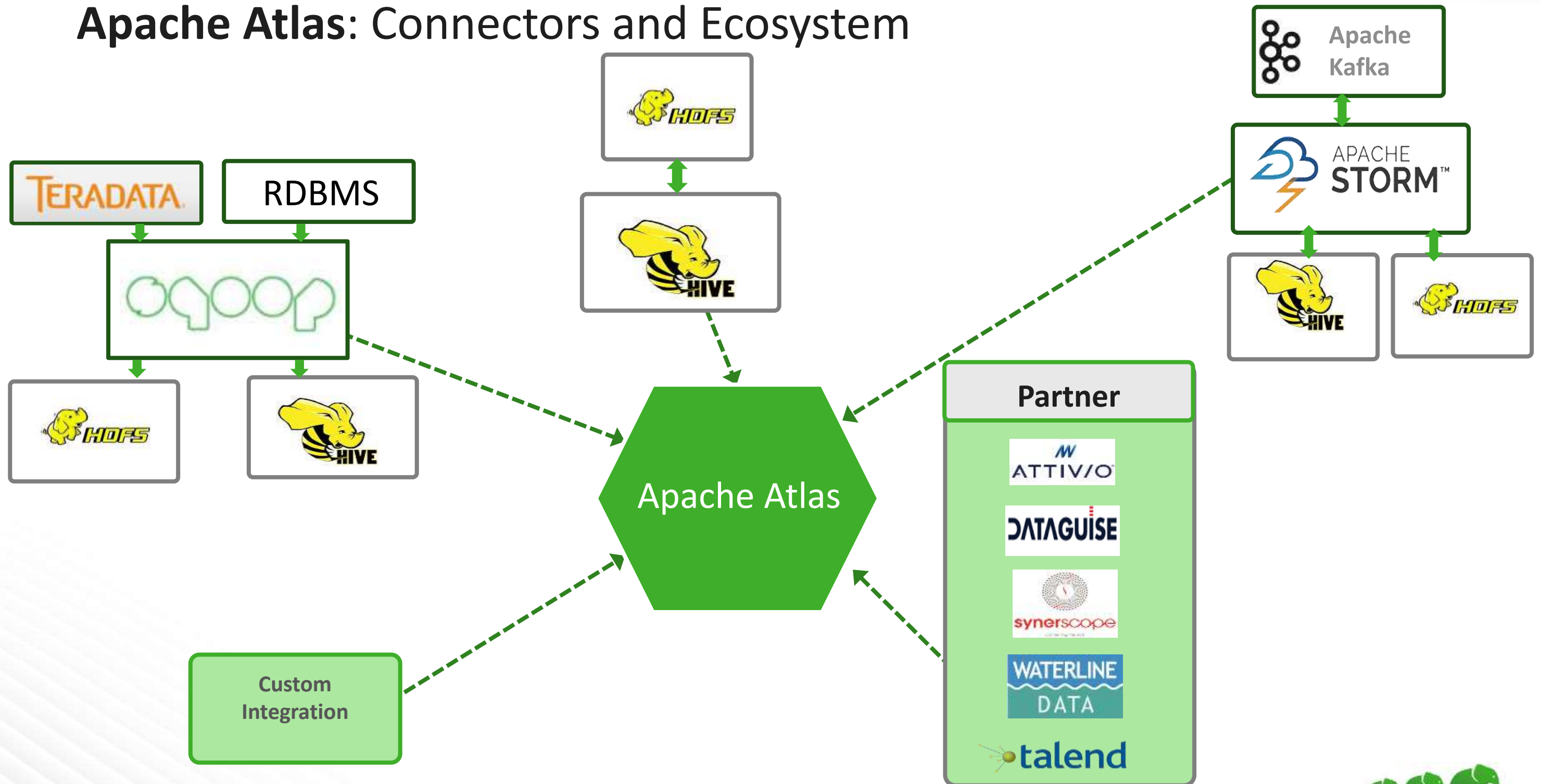
## Used for forensics

- Impact analysis
- Auditing and Compliance

# Apache Atlas: Lineage and Impact



# Apache Atlas: Connectors and Ecosystem





# Apache Atlas: Classification

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

The screenshot displays the Apache Atlas web interface. On the left is a dark sidebar with a search bar and a list of tags. The 'EXPIRES\_ON' tag is selected and highlighted in green. To the right, the main content area shows the details for the 'EXPIRES\_ON' tag. It includes a section for 'Attributes' with 'expiry\_date' and an 'ADD Attribute +' button. Below this, a section titled 'Results for EXPIRES\_ON' shows 'Showing 1 - 3' items. A table lists these items with columns for Name, Description, Type, Owner, and Tags. Three green callout boxes with arrows point to specific elements: 'GOVERNANCE' points to the tag name 'EXPIRES\_ON', 'SECURITY' points to the 'tax\_2015' entry in the table, and 'BUSINESS' points to the 'WEALTH\_MANAGEMENT' tag in the sidebar.

Apache Atlas

Q SEARCH TAGS

+ Create Tag

Search Tags

AUTO\_INSURANCE

BANKING

CONSUMER\_BANKING

DATA\_QUALITY

**EXPIRES\_ON**

FINANCE\_PII

HEALTH\_INSURANCE

INSURANCE

LIFE\_INSURANCE

PII

REFERENCE\_DATA

SMALL\_BUSINESS\_BANKING

TradingDataset

VENDOR\_PII

WEALTH\_MANAGEMENT

**EXPIRES\_ON**

EXPIRES\_ON

Attributes: expiry\_date ADD Attribute +

Results for **EXPIRES\_ON**

Showing 1 - 3

	Name	Description	Type	Owner	Tags
	tax_2015		hive_table	hive	EXPIRES_ON x +
	tax_2010		hive_table	hive	EXPIRES_ON x +
	tax_2009		hive_table	hive	EXPIRES_ON x +

**GOVERNANCE**

**SECURITY**

**BUSINESS**

# 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

**REFERENCE\_DATA**


REFERENCE\_DATA

Attributes:

Results for **REFERENCE\_DATA**

Showing 1 - 1

Previous Next

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

**Policy Details :**

Policy Type: **Access**

Policy ID: **40**

Policy Name \*:

TAG \*:

Audit Logging: **YES**

Description:

**Components Permissions**

Component	Permissions
<input type="checkbox"/> hdfs	<input type="checkbox"/> Read <input checked="" type="checkbox"/> Write <input type="checkbox"/> Execute
<input type="checkbox"/> hive	<input type="checkbox"/> select <input checked="" type="checkbox"/> update <input type="checkbox"/> Create <input checked="" type="checkbox"/> Drop <input checked="" type="checkbox"/> Alter <input type="checkbox"/> Index <input type="checkbox"/> Lock <input type="checkbox"/> All

**Allow Conditions :**

**Deny Conditions :**

Select Group	Select User	Policy Conditions	Component Permissions
<input type="text" value="x public"/>	<input type="text" value="Select User"/>	<input type="button" value="Add Conditions"/> <input type="button" value="+"/>	<input type="button" value="HDFS"/> <input type="button" value="HIVE"/> <input type="button" value="x"/>

# 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

**tax\_2009 (hive\_table)**

Tags: EXPIRES\_ON +

**LINEAGE & IMPACT**

**DETAILS**

Properties Tags Audits Schema

Showing 1 - 1

Tags	Attributes
EXPIRES_ON	expiry_date:2016/12/31

**tax\_2010 (hive\_table)**

Tags: EXPIRES\_ON +

**LINEAGE & IMPACT**

**DETAILS**

Properties Tags Audits Schema

Showing 1 - 1

Tags	Attributes
EXPIRES_ON	expiry_date:2017/12/31

**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

**Allow Conditions :**

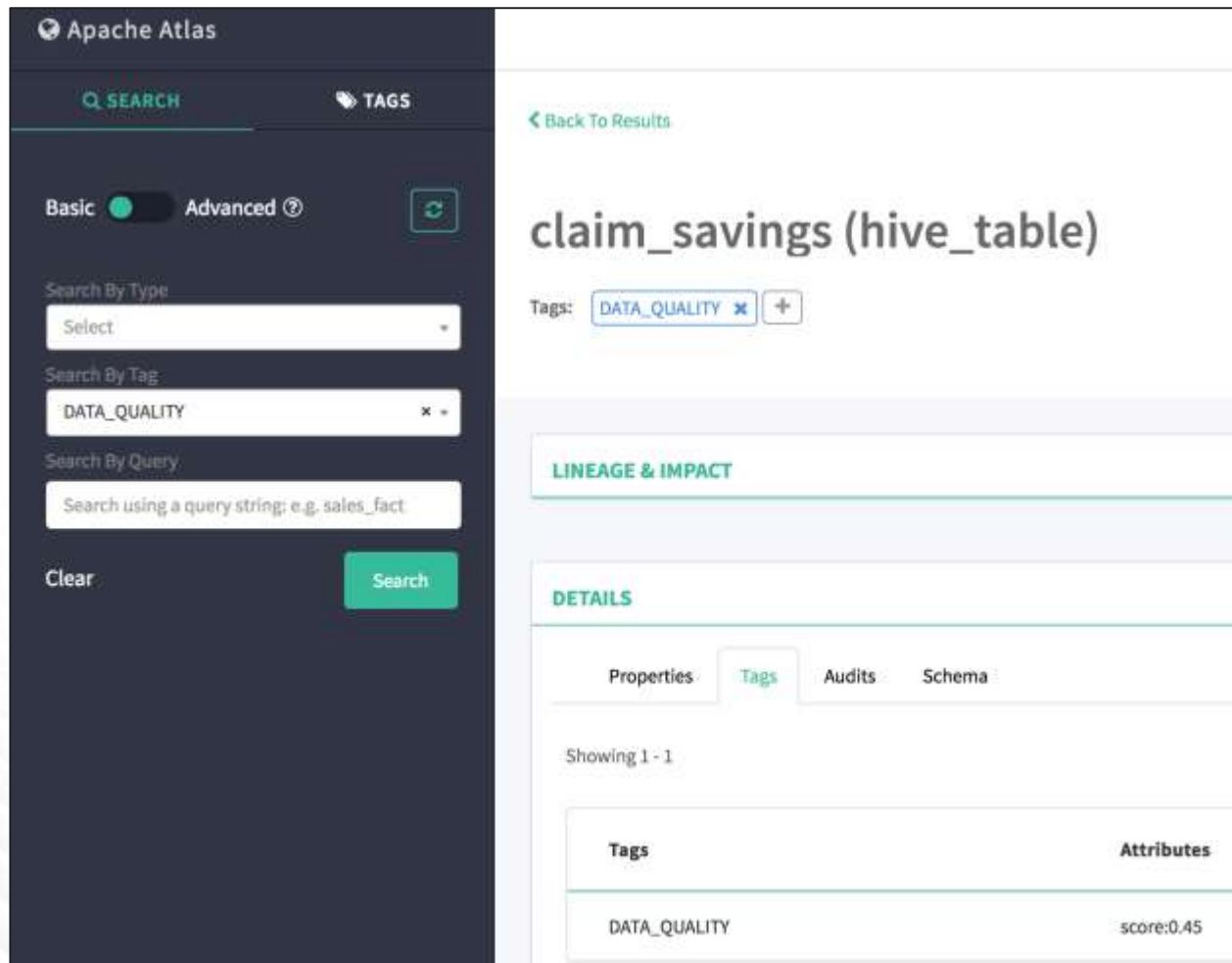
**Deny Conditions :**

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

# Apache Atlas Classification: usecase – attribute based authorization

## Data quality

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



Apache Atlas

SEARCH TAGS

Basic ☒ Advanced ☐

Search By Type: Select

Search By Tag: DATA\_QUALITY

Search By Query: Search using a query string; e.g. sales\_fact

Clear Search

claim\_savings (hive\_table)

Tags: DATA\_QUALITY

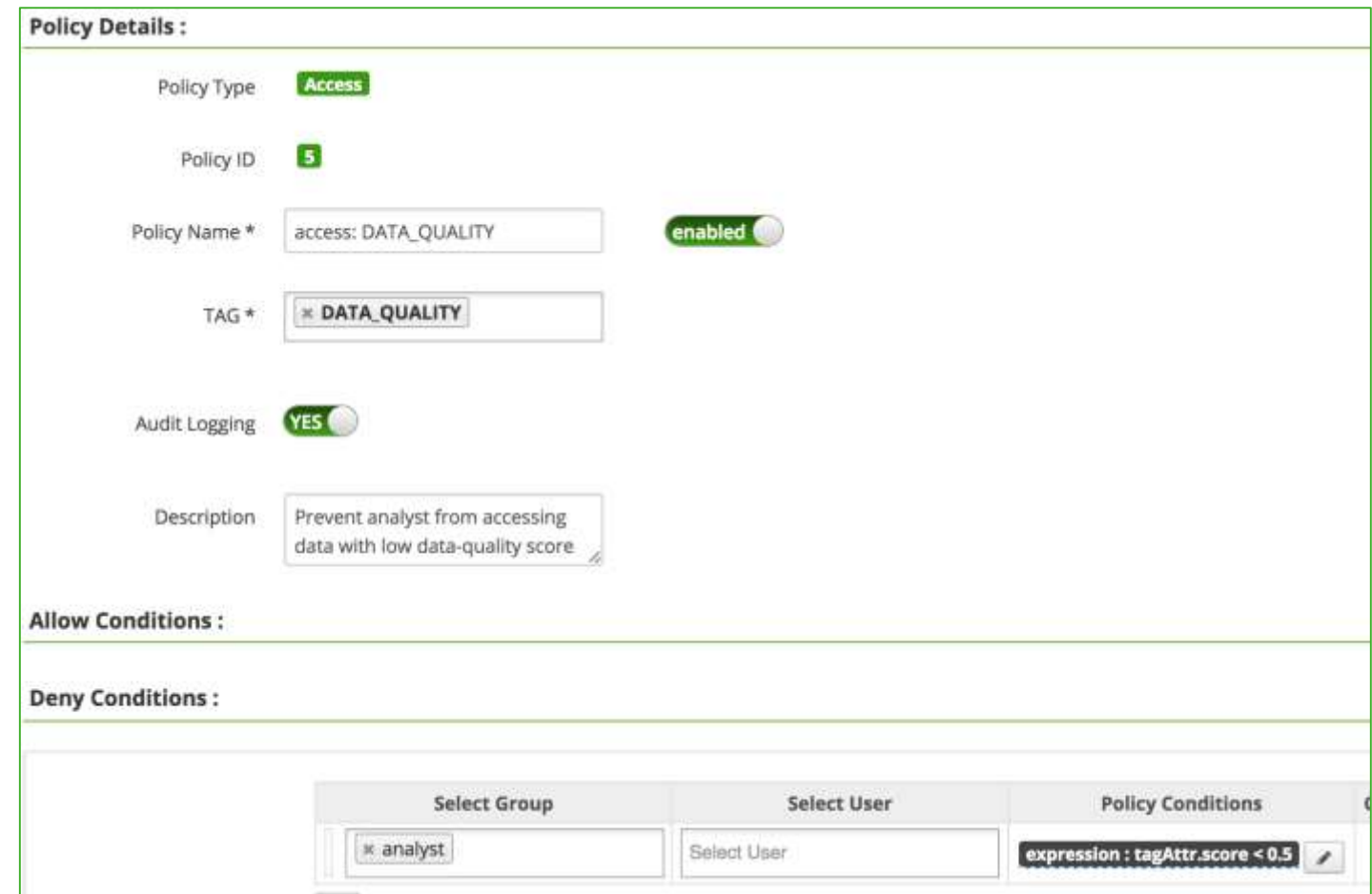
LINEAGE & IMPACT

DETAILS

Properties Tags Audits Schema

Showing 1 - 1

Tags	Attributes
DATA_QUALITY	score:0.45



Policy Details :

Policy Type: Access

Policy ID: 5

Policy Name \*: access: DATA\_QUALITY enabled

TAG \*: DATA\_QUALITY

Audit Logging: YES

Description: Prevent analyst from accessing data with low data-quality score

Allow Conditions :

Deny Conditions :

Select Group	Select User	Policy Conditions
analyst	Select User	expression : tagAttr.score < 0.5



# Apache Atlas Classification: usecase – cross component

## Classification based security on cross-component data assets

**TradingDataset**  
All datasets related to trading and transactions

Attributes:

Results for **TradingDataset**  
Showing 1 - 5

Name	Description	Type	Owner	Tags
/feeds/mutualfunds	Mutual Funds	hdfs_path	brokers	TradingDataset
futures_trade		hive_table	hive	TradingDataset
/feeds/forex	Forex	hdfs_path	brokers	TradingDataset
derivatives_trade		hive_table	hive	TradingDataset
/feeds/commodities	Commodity Data	hdfs_path	brokers	TradingDataset

**Policy Details :**

Policy Type: **Access**

Policy ID: **41**

Policy Name:  **enabled**

TAG:

Audit Logging: **YES**

Description:

**Allow Conditions :**

Select Group	Select User	Policy Conditions	Component Permissions
<input type="text" value="broker"/>	<input type="text" value="Select User"/>	<input type="button" value="Add Conditions +"/>	<input type="button" value="HIDE"/> <input type="button" value="HIDE"/> <input type="button" value="X"/>

Exclude from Allow Conditions :

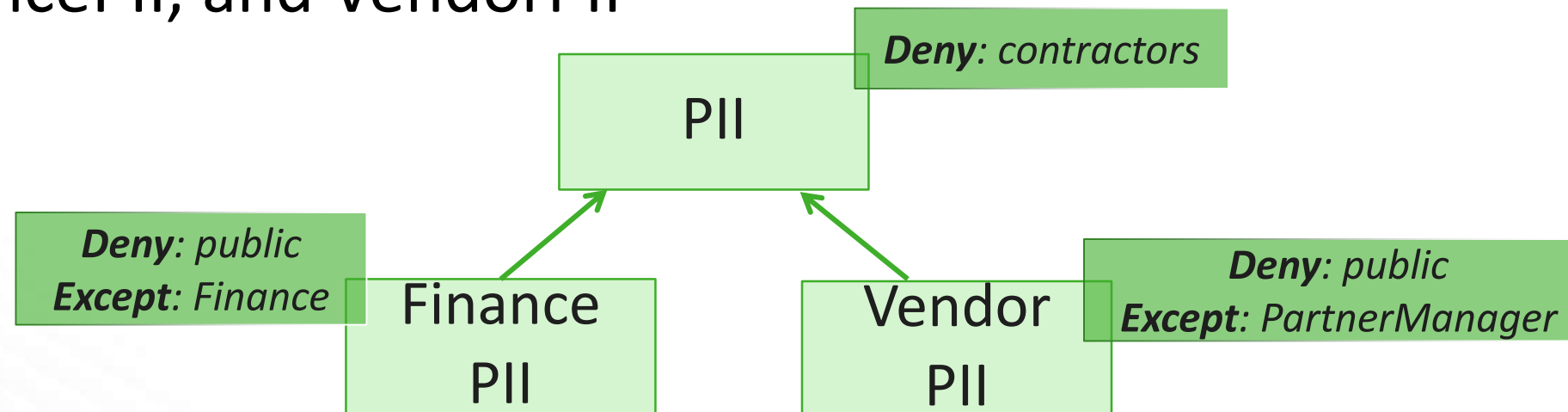
**Deny Conditions :**

Select Group	Select User	Policy Conditions	Component Permissions
<input type="text" value="public"/>	<input type="text" value="Select User"/>	<input type="button" value="Add Conditions +"/>	<input type="button" value="HIDE"/> <input type="button" value="HIDE"/> <input type="button" value="X"/>

# 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



# Metadata Catalog Search : Free Text

Apache Atlas

SEARCH TAGS

Basic ☒ Advanced ?

Search By Type  
hive\_table x

Search By Tag  
PII x

Search By Query  
emp\* x

Clear Search

Results for **hive\_table & PII & emp\***  
If you do not find the entity in search result below then you can [create new entity](#)

Showing 1 - 1

Filter by Data Asset type

Filter by Classification

Name	Description	Type	Owner	Tags
employees		hive_table	hive	PII x +

Search text  
Wildcards: emp\*, \*dept\*  
Logical expressions: emp\* AND \*dept\*

Previous Next

Search for a hive\_table classified as 'PII' and name starting with 'emp'

# Metadata Catalog Search : Advanced

Apache Atlas

Q SEARCH

TAGS

Basic

Advanced ?

Search By Type

hive\_table x

Search By Query

where name='employees' and owner='hive'

Clear

Search

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'

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](#)

Showing 1 - 1

Filter by  
Data asset type

Previous

Next

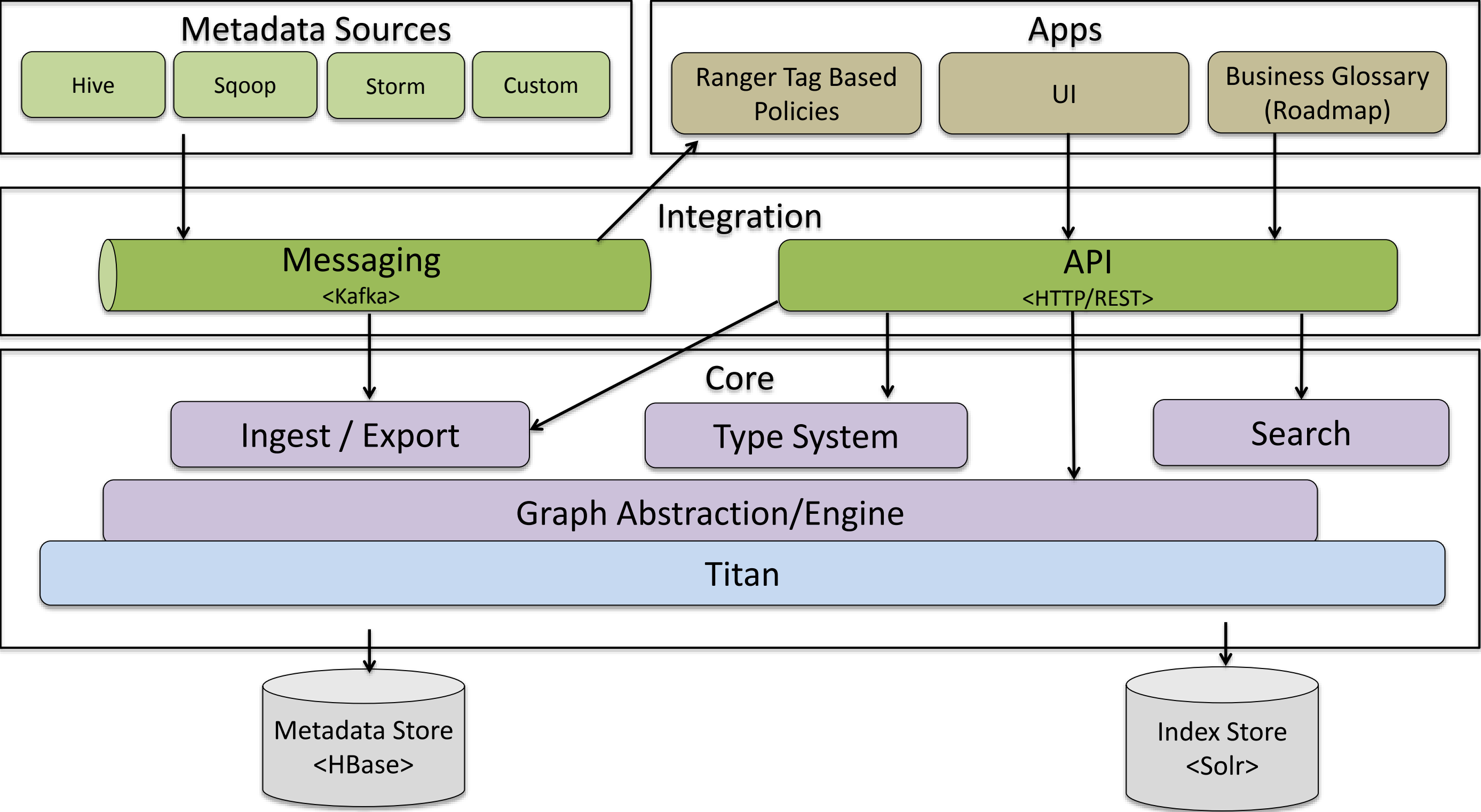
	Name	Description	Type	Owner	Tags
<input type="checkbox"/>	employees		hive_table	hive	PII x +

DSL query string

Search for a *hive\_table* named 'employees' and owner 'hive'



# 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

# Questions

# 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>