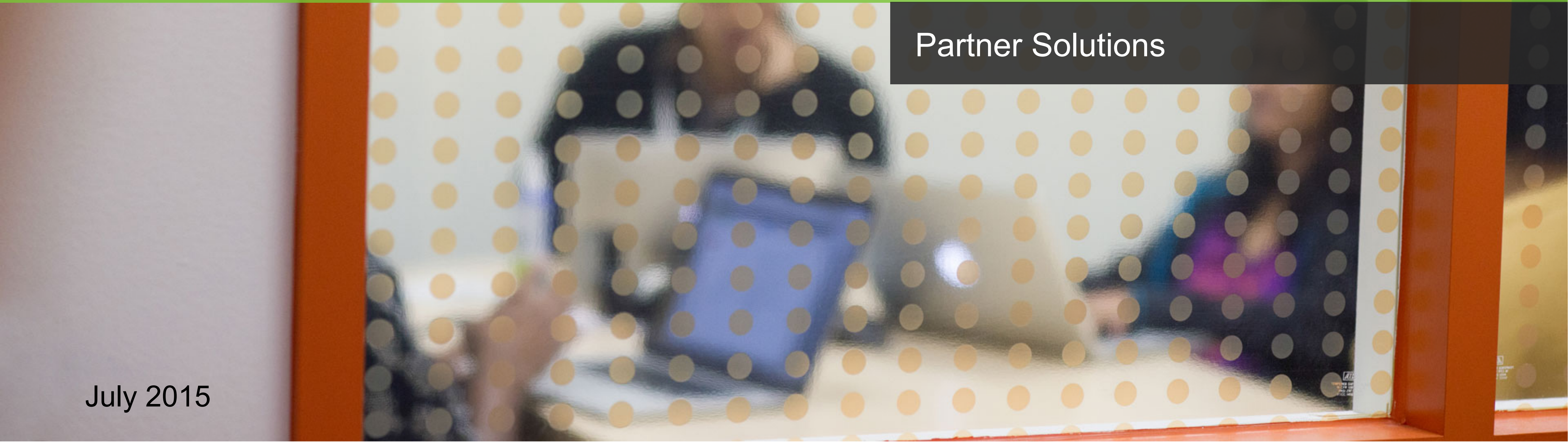
Apache Atlas: Data Governance

Partner Solutions



July 2015

Agenda

Overview

- Enterprise Goals
- Data Governance Initative

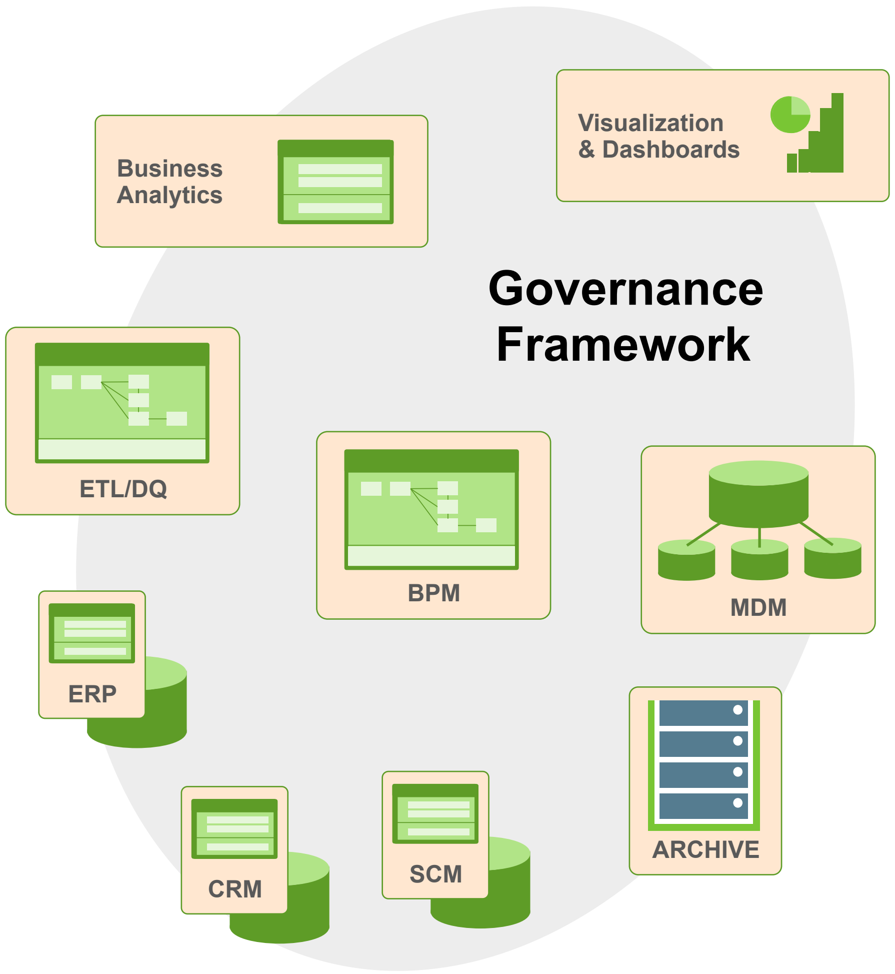
Atlas

- Feature tour
- Roadmap
- UI Tour

Demo

- Example: Sqoop
- Walk through step
- Search Tables / Tags

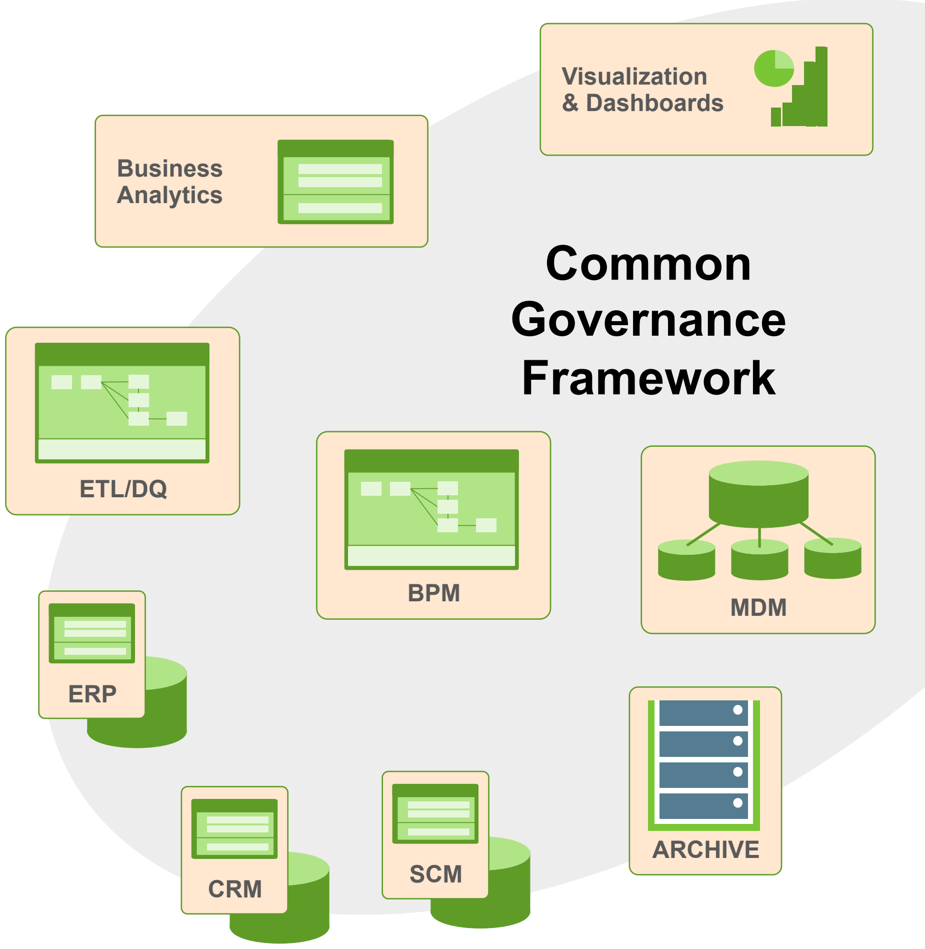
Enterprise Data Governance Goals



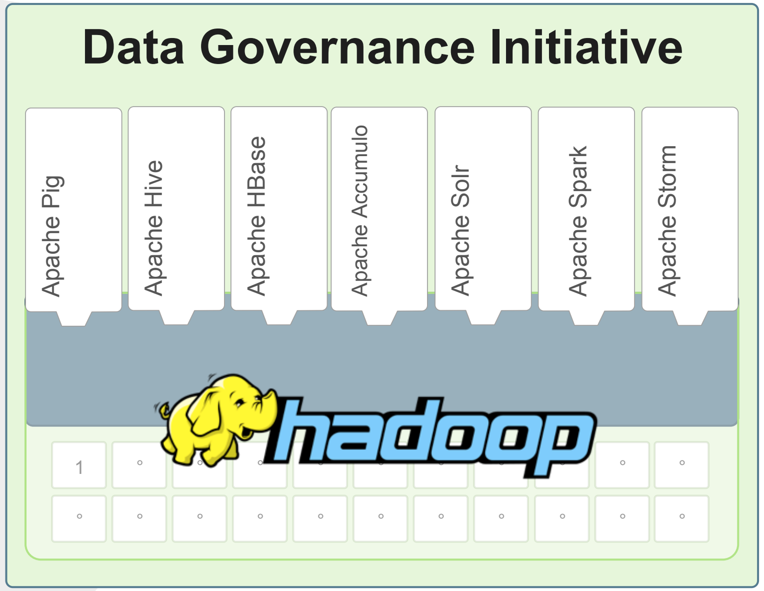
GOAL: Provide a common approach to data governance across all systems and data within the organization

- **Transparent**
  Governance standards & protocols must be clearly defined and available to all
- **Reproducible**
  Recreate the relevant data landscape at a point in time
- **Auditable**
  All relevant events and assets but be traceable with appropriate historical lineage
- **Consistent**
  Compliance practices must be consistent

Data Governance Initiative for Hadoop



Common Governance Framework



Data Governance Initiative

TWO Requirements

1. Hadoop must snap in to the existing frameworks and be a good citizen
2. Hadoop must also provide governance within its own stack of technologies

A group of companies dedicated to meeting these requirements in the open

aetna TARGET MERCK Be well Hortonworks

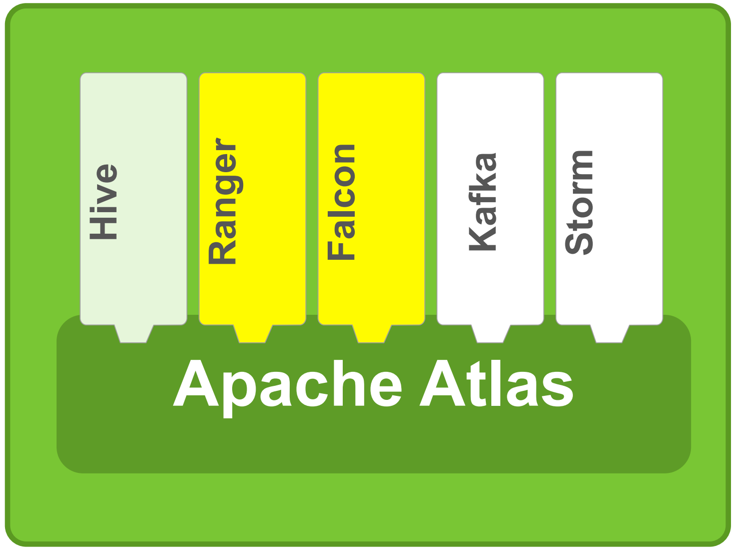
SAP *Major Bank* Schlumberger sas

Hortonworks

Apache Atlas Overview

We Do Hadoop

Apache Atlas Vision



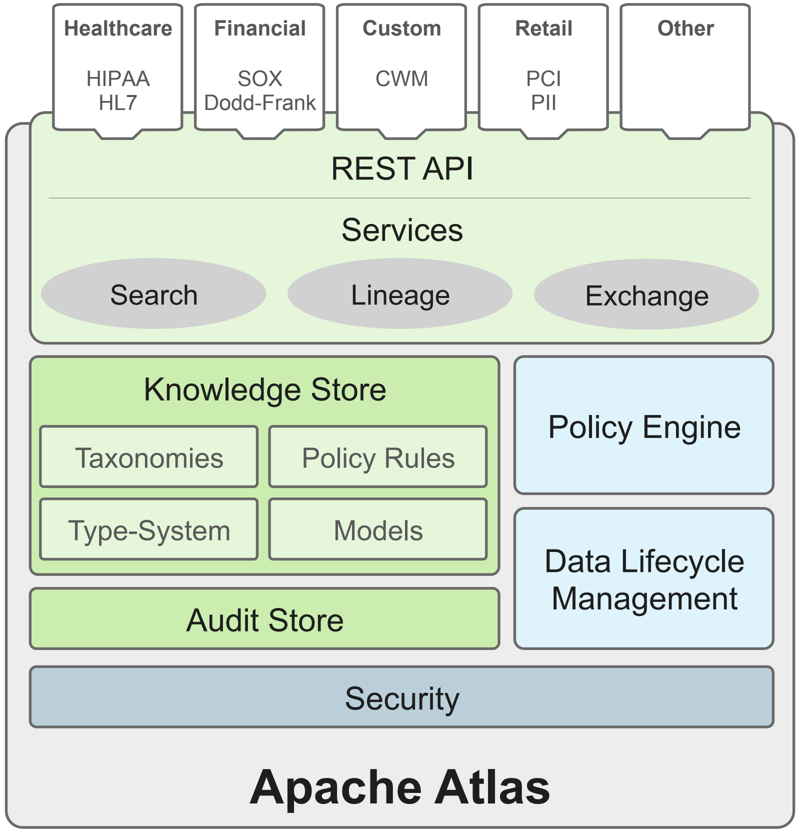
Metadata Services

- Flexible Knowledge Store
- Business Catalog / Operational Data
- Search & **Proscriptive Lineage**
- Centralized location for all metadata within HDP
- Interface point for Metadata Exchange with platforms outside of HDP.

Metadata will enrich every component

- Hive – Complete lineage, every HiveQL tracked
- Ranger – Tag or Attribute security ABAC
- Falcon – Business Taxonomy

Apache Atlas Capabilities: Overview



Data Classification

- Import or define taxonomy business-oriented annotations for data
- Define, annotate, and automate capture of relationships between data sets and underlying elements including source, target, and derivation processes
- Export metadata to third-party systems

Centralized Auditing

- Capture security access information for every application, process, and interaction with data
- Capture the operational information for execution, steps, and activities

Search & Lineage (Browse)

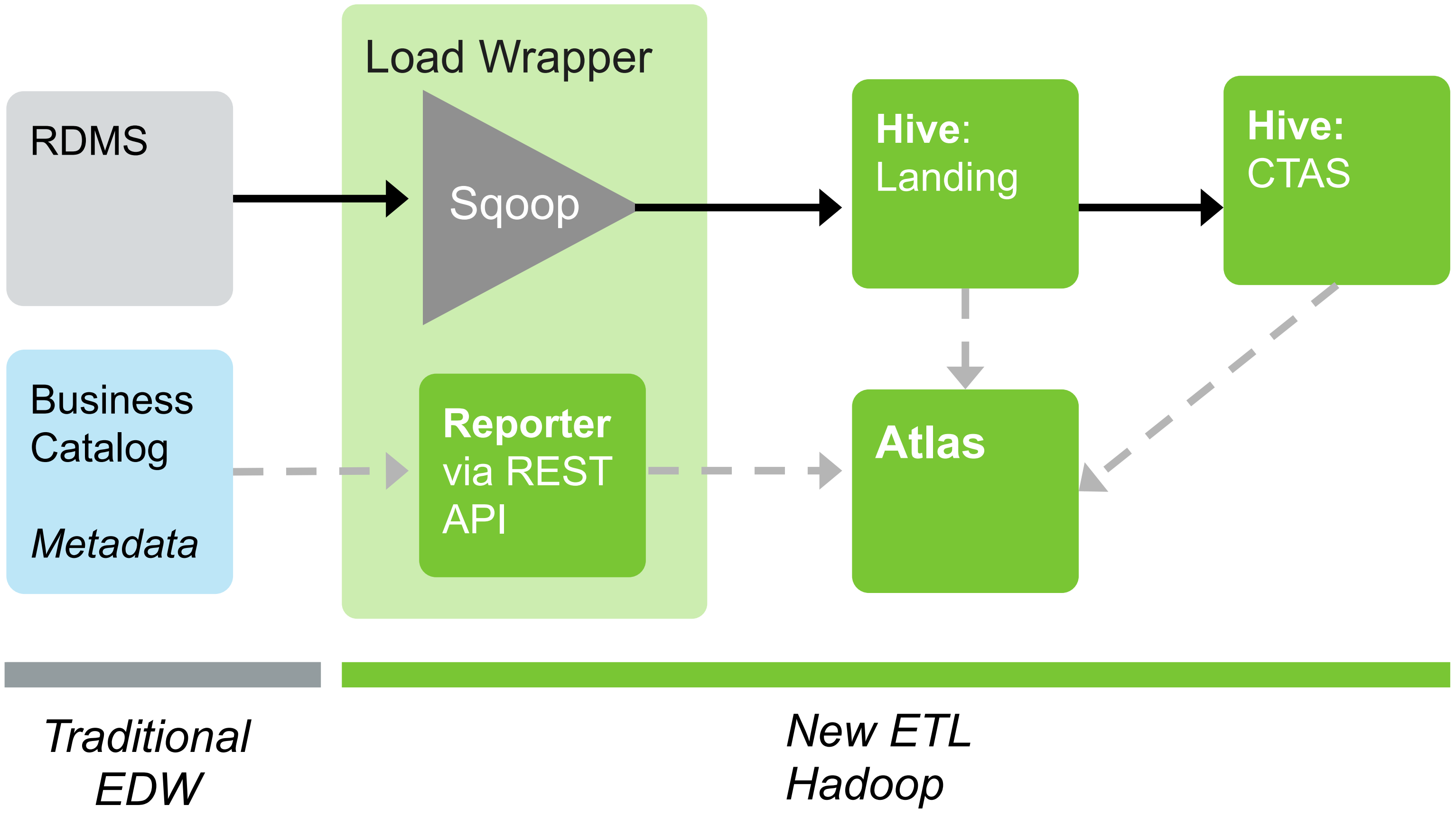
- Pre-defined navigation paths to explore the data classification and audit information
- Text-based search features locates relevant data and audit event across Data Lake quickly and accurately
- Browse visualization of data set lineage allowing users to drill-down into operational, security, and provenance related information

Security & Policy Engine

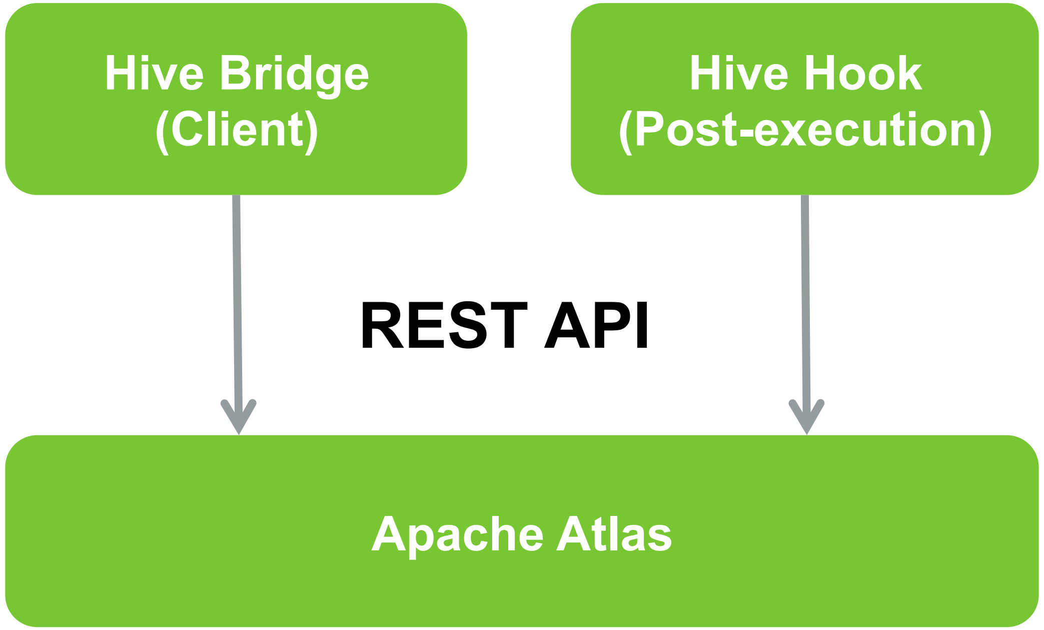
- Rationalize compliance policy at runtime based on data classification schemes
- Advanced definition of policies for preventing data derivation based on classification (i.e. re-identification)

Hortonworks

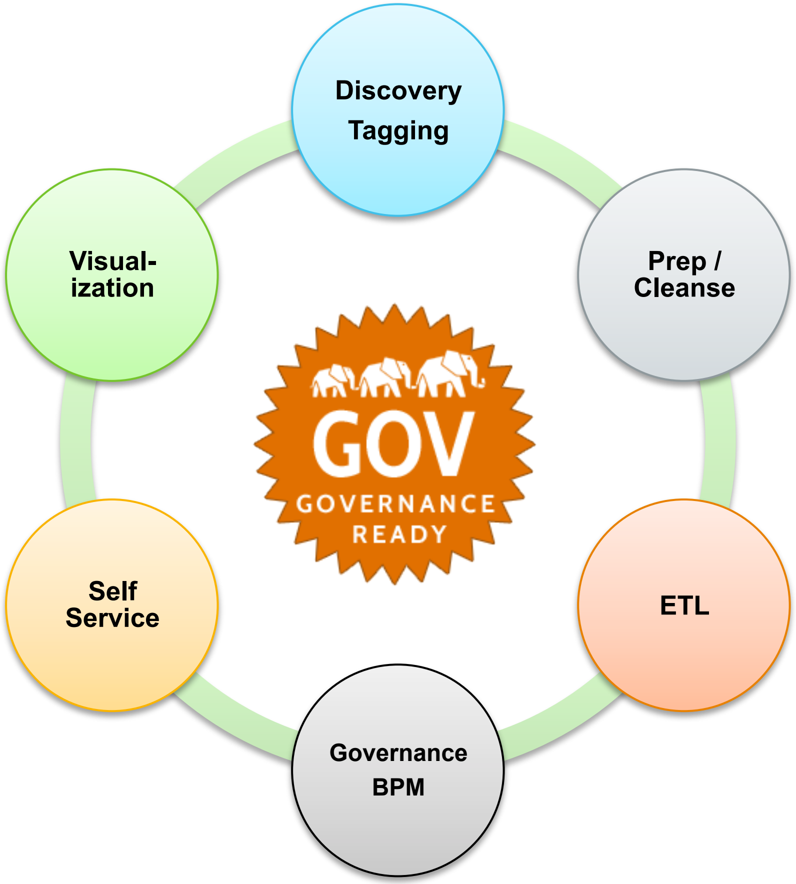
Sample Use Case: ETL Offload



Hive Integration



Governance Ready Certification Program



Curated group of vendor partners to provide rich & complete features

Customers choose features that they want to deploy – a la carte.

Low switching costs !

HDP at core to provide stability and interoperability

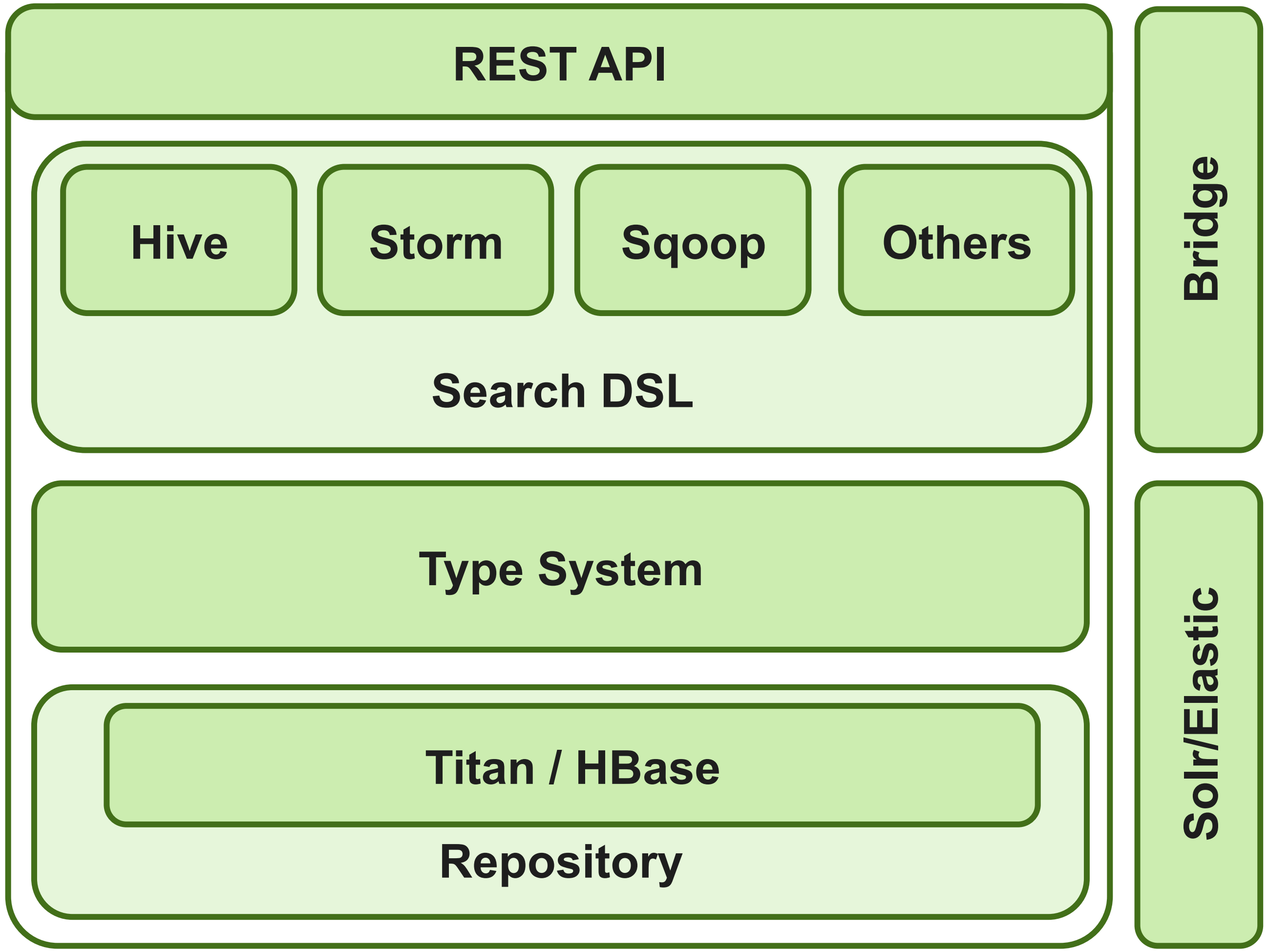
Hortonworks

High Level Roadmap

- **ASF MVP (May)** – Preview Core Metadata Services: Type system, API's, basic UI, Hive connecter
- **HDP 2.3 (July)** - GA Core Metadata Services. Preview Metadata Business Glossary
- **M10 – (Sept)** – Preview ABAC with Ranger integration and Preview Sqoop component connector
- **M20** – Preview Kafka, Storm connectors, Gov Ready Certification program, Preview row level & Column masking.
- **HDP 2.4 (Q4'15)** GA all preview features

Architecture

High Level Architecture



Technology Stack

- Knowledge Store
  - Titan Graph DB
- Pluggable Search Backend
  - Elastic search
  - Solr
- Rules Engine
  - TBD

- Audit Store
  - YARN ATS - Time series DB
- Java 1.7
- Dashboard
  - TBD

APIs: Examples

Admin
GET: /admin/stack
GET: /admin/version

Entity
GET: /entities/definition/{guid}
POST: /entities/submit/{typeName}
GET: /entities/list/{entityType}

Metadata Discovery
GET: /discovery/search/gremlin/{gremlinQuery}
GET: /discovery/search/relationships/{guid}
GET: /discovery/search/fullText?text=<query>
GET: /discovery/getIndexedFields

Rexster
GET: /graph/vertices/{id}
GET: /graph/vertices/properties/{id}
GET: /graph/vertices
GET: /graph/vertices/{id}/{direction}
GET: /graph/edges/{id}

Types
POST: /types/submit/{typeName}
GET: /types/definition/{typeName}
GET: /types/list

Hive Lineage
GET: /bridge/hive/{id}
GET: /bridge/hive
POST: /bridge/hive

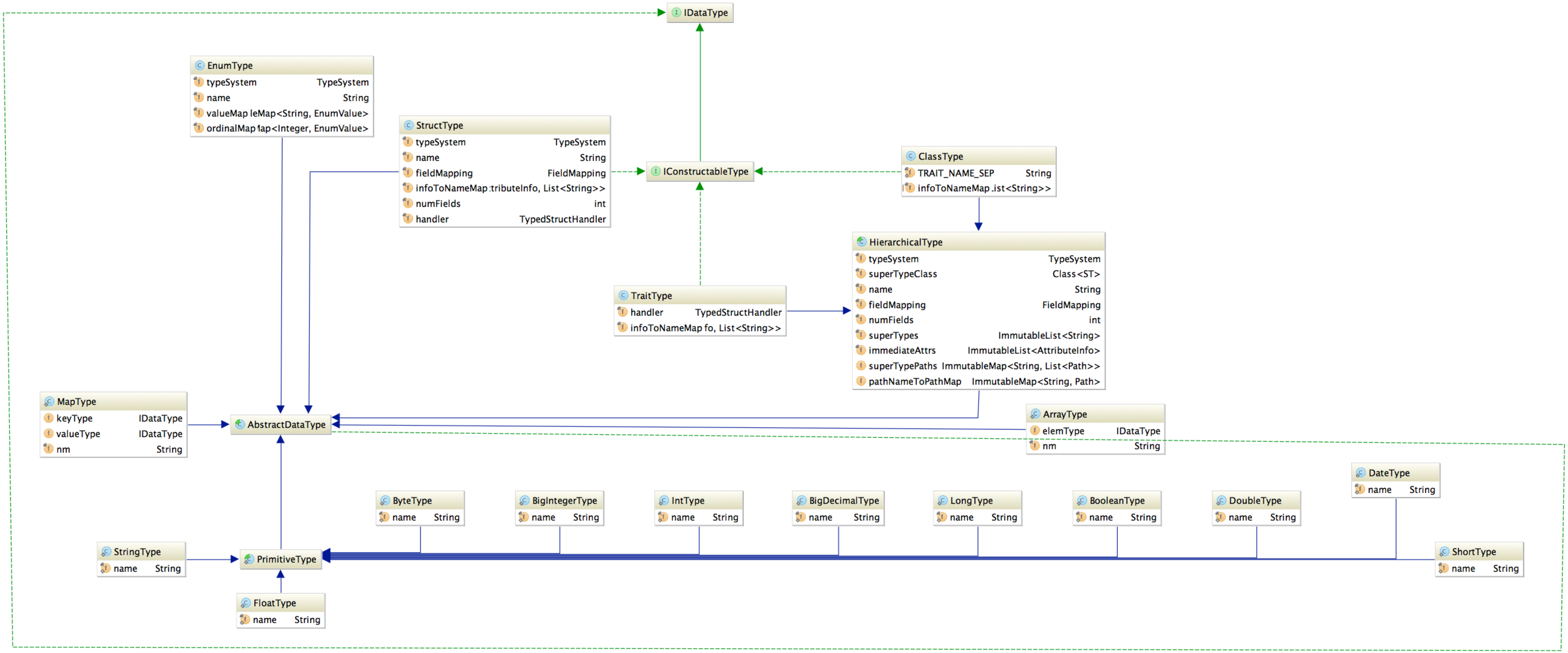
Type System – Overview of Types

- Class
- Struct
- Trait
- Primitives

- Collections
  - Map
  - Array

- Instances (Entity)
- Referenceable

Type System – Data Types



IDataType

EnumType	
typeSystem	TypeSystem
name	String
valueMap	leMap<String, EnumValue>
ordinalMap	łap<Integer, EnumValue>

StructType	
typeSystem	TypeSystem
name	String
fieldMapping	FieldMapping
infoToNameMap	:tributeInfo, List<String>>
numFields	int
handler	TypedStructHandler

IConstructableType

ClassType	
TRAIT_NAME_SEP	String
infoToNameMap	.ist<String>>

HierarchicalType	
typeSystem	TypeSystem
superTypeClass	Class<ST>
name	String
fieldMapping	FieldMapping
numFields	int
superTypes	ImmutableList<String>
immediateAttrs	ImmutableList<AttributeInfo>
superTypePaths	ImmutableMap<String, List<Path>>
pathNameToPathMap	ImmutableMap<String, Path>

TraitType	
handler	TypedStructHandler
infoToNameMap	fo, List<String>>

MapType	
keyType	IDataType
valueType	IDataType
nm	String

AbstractDataType

ArrayType	
elemType	IDataType
nm	String

DateType	
name	String

ByteType	
name	String

BigIntegerType	
name	String

IntType	
name	String

BigDecimalType	
name	String

LongType	
name	String

BooleanType	
name	String

DoubleType	
name	String

StringType	
name	String

PrimitiveType

ShortType	
name	String

FloatType	
name	String

```
_trait("Dimension") {}
_trait("PII") {}
_trait("Metric") {}
_trait("ETL") {}
_trait("JdbcAccess") {}

_class("DB") {
  "name" ~ (string, required,
indexed, unique)
  "owner" ~ (string)
  "createTime" ~ (int)
}

_class("StorageDesc") {
  "inputFormat" ~ (string,
required)
  "outputFormat" ~ (string,
required)
}
```

```
_class("Column") {
  "name" ~ (string, required)
  "dataType" ~ (string, required)
  "sd" ~ ("StorageDesc", required)
}

_class("Table", List()) {
  "name" ~ (string,  required,  indexed)
  "db" ~ ("DB", required)
  "sd" ~ ("StorageDesc", required)
}
```

Repository

- **Graph Database**
  - Titan with storage backed by HBase
- **Types and instances are mapped to the Graph DB**
  - Classes, Structs and Traits map to a vertex
  - Relationships are mapped as edges
- **Search - plugin enabled**
  - Indexing based on type annotations
  - Solr
  - Elastic search

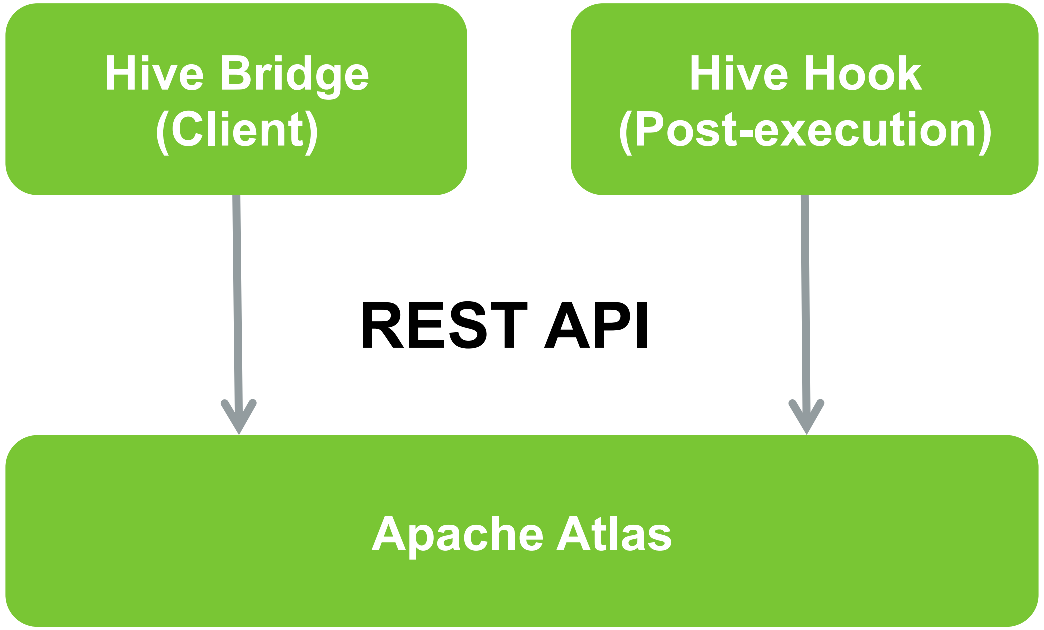
Search

- **DSL with SQL Like Syntax**
  - from $type is $trait where $clause select|has $attributes loop $loopExpression withPath, repeat
- **Examples**
  - from DB
  - DB where name="Reporting" select name, owner
  - DB has name
  - DB is JdbcAccess
  - Column where Column is a PII
  - Table where name="sales_fact", columns
  - Table where name="sales_fact", columns as column select column.name, column.dataType, column.comment
- **Full-text search**

Lineage

- **Uses Search DSL Loop expression**
  - Everything results in search
- **Named Queries**
- **inputs**
  - Table where (name = \"sales_fact_monthly_mv\") as src loop (LoadProcess->outputTable inputTables) as dest select src.name as src_name, dest.name as dest_name withPath
- **outputs**
  - Table where (name = \"sales_fact\") as src loop (LoadProcess->inputTables outputTables) as dest select src.name as src_name, dest.name as dest_name withPath
- **schema**
  - Table where name="sales_fact", columns

Hive Integration



Apache Atlas Screens

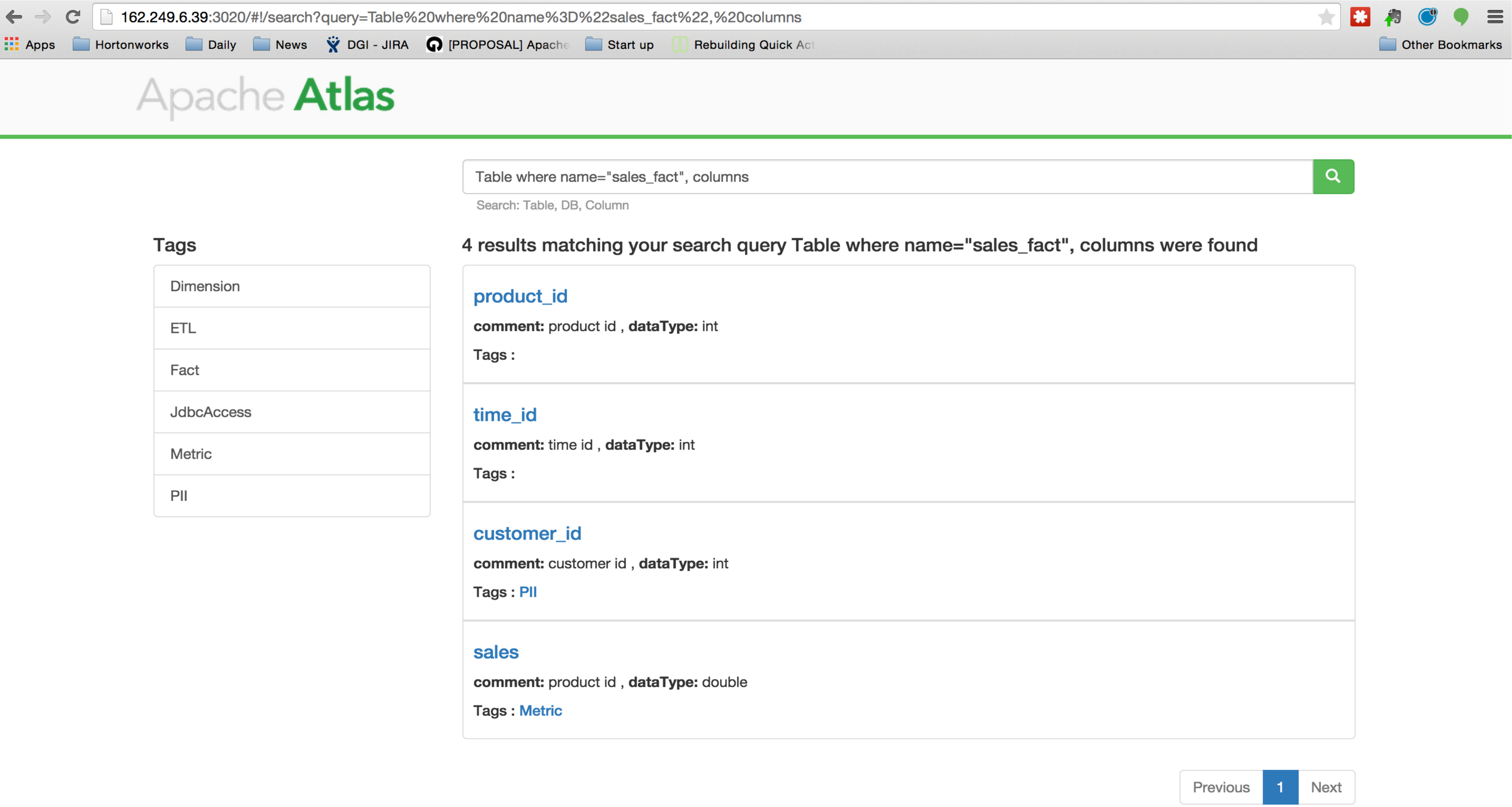


Table where name="sales_fact", columns

Search: Table, DB, Column

Tags

- Dimension
- ETL
- Fact
- JdbcAccess
- Metric
- PII

4 results matching your search query Table where name="sales_fact", columns were found

product_id

comment: product id , **dataType:** int

Tags :

time_id

comment: time id , **dataType:** int

Tags :

customer_id

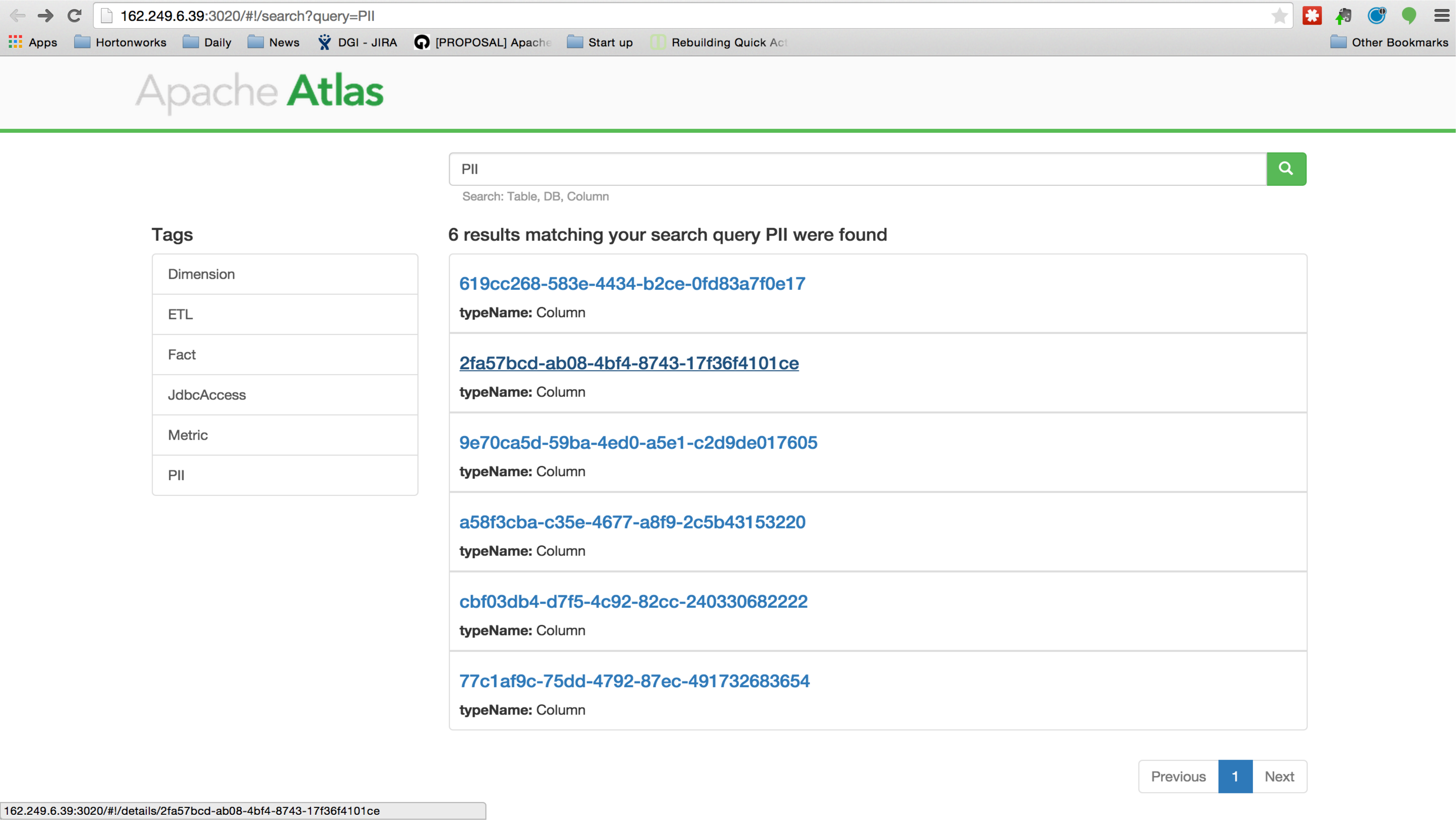
comment: customer id , **dataType:** int

Tags : PII

sales

comment: product id , **dataType:** double

Tags : Metric



162.249.6.39:3020/#!/search?query=PII

PII

Search: Table, DB, Column

Tags

- Dimension
- ETL
- Fact
- JdbcAccess
- Metric
- PII

6 results matching your search query PII were found

619cc268-583e-4434-b2ce-0fd83a7f0e17

typeName: Column

2fa57bcd-ab08-4bf4-8743-17f36f4101ce

typeName: Column

9e70ca5d-59ba-4ed0-a5e1-c2d9de017605

typeName: Column

a58f3cba-c35e-4677-a8f9-2c5b43153220

typeName: Column

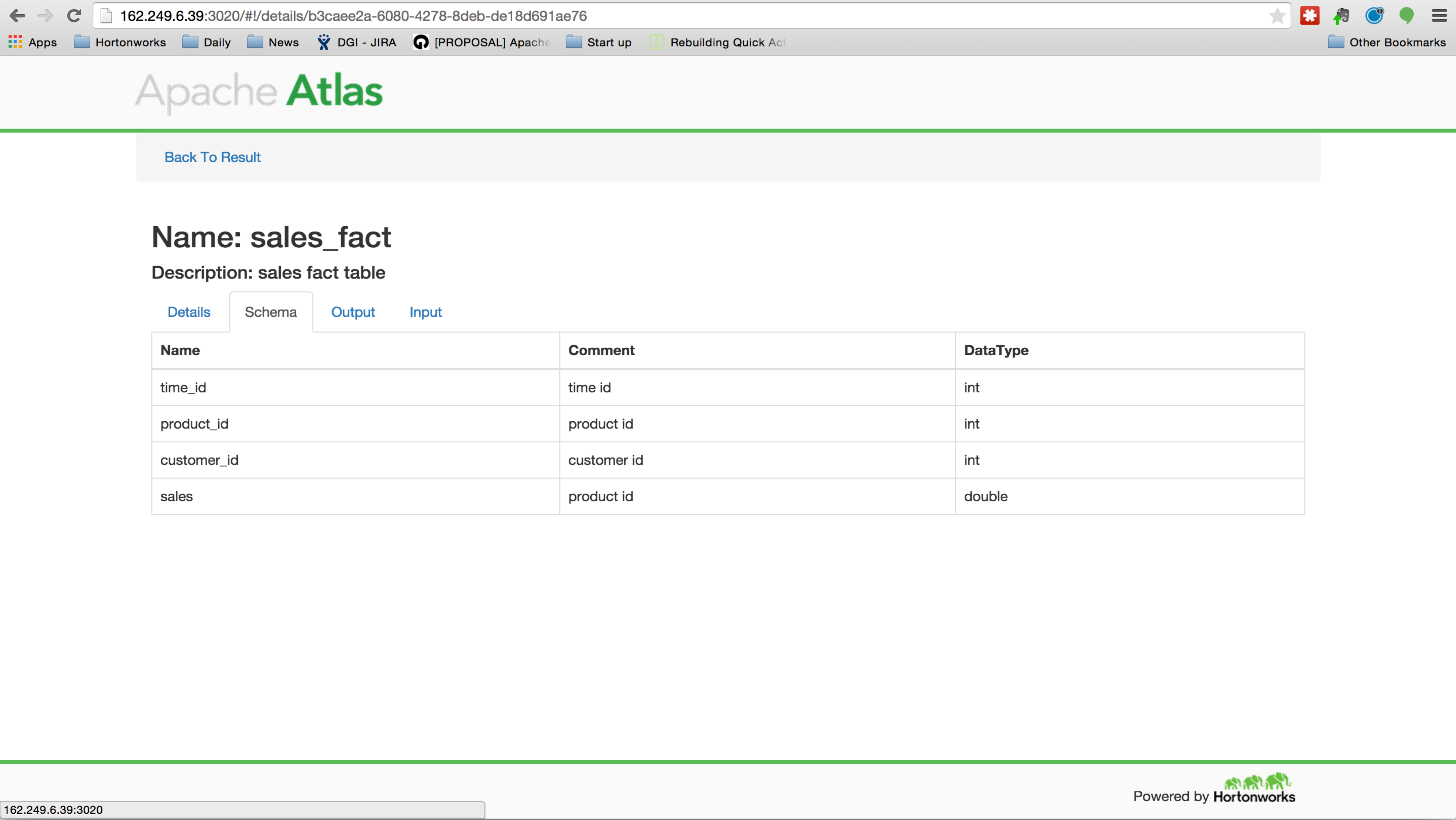
cbf03db4-d7f5-4c92-82cc-240330682222

typeName: Column

77c1af9c-75dd-4792-87ec-491732683654

typeName: Column

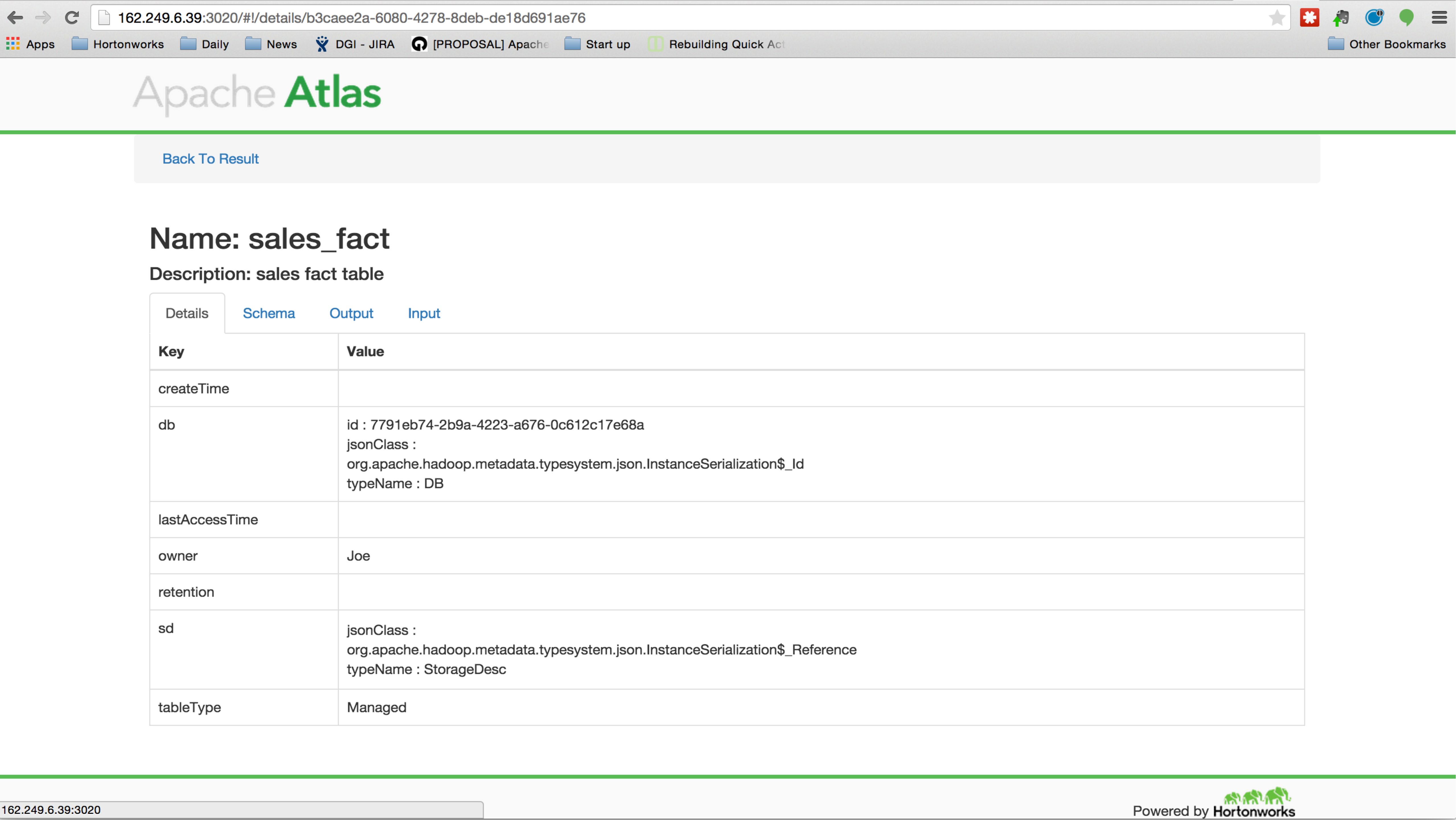
162.249.6.39:3020/#!/details/2fa57bcd-ab08-4bf4-8743-17f36f4101ce



Name: sales_fact

Description: sales fact table

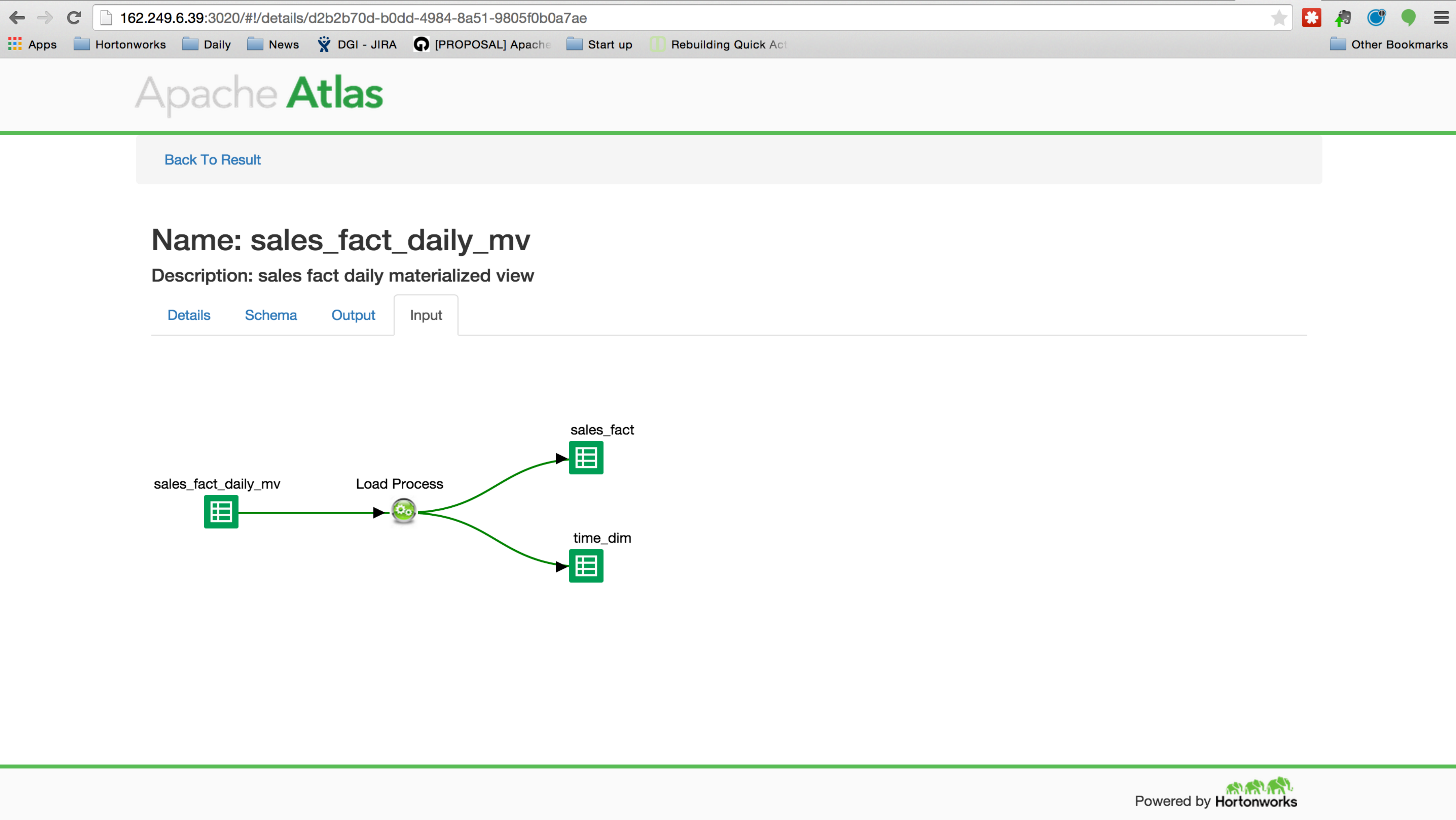
Name	Comment	DataType
time_id	time id	int
product_id	product id	int
customer_id	customer id	int
sales	product id	double



Name: sales_fact

Description: sales fact table

Key	Value
createTime	
db	id : 7791eb74-2b9a-4223-a676-0c612c17e68a jsonClass : org.apache.hadoop.metadata.typesystem.json.InstanceSerialization$_Id typeName : DB
lastAccessTime	
owner	Joe
retention	
sd	jsonClass : org.apache.hadoop.metadata.typesystem.json.InstanceSerialization$_Reference typeName : StorageDesc
tableType	Managed



Name: sales_fact_daily_mv

Description: sales fact daily materialized view

Demo Atlas

Hortonworks

Atlas UI demostration

Search DSL

- Type – DB, Table, Column
- Tag - PII
- Keyword

Results

- Details
- Schema
- Lineage

Coming Features

Ingestion Demo Objective

- **Show Lineage with Sqoop Ingestion of data**
  - Custom process instrumention
- **Use the Hive Hook CTAS Operation**
  - Atlas Follow Lineage
- **Metadata Model in Atlas**
  - The Open Framework
  - Create Custom Types
  - Create Custom Process
- **Sample Codes**

Setup

- **Source System**
  - MySQL Database
    - DRIVERS
    - TIMESHEET
- Destination System
  - Single Node HDP 2.3 (Tech Preview)
    - Apache Atlas

Steps to Create Metadata

- **Create a Atlas Client Instance**
- **Create Type Definitions**
  - Class Types
  - Attributes
  - List the Types
- **Instantiate Entities**
  - - Create Entities (Class Type)
  - - Search the Types
- **Create Process**
  - Create DataSet Type
  - Create Process Type
  - Connect a Process Lineage

Attribute Definition

- **Name**
- **Data Type**
- **Multiplicity**
- **Composite**
- **isIndexable**
- **ReverseAttribute**

Questions and Answers

Atlas Resources

- **HDP 2.3 Preview Sandbox VM:**
  - http://hortonworks.com/hdp/whats-new/

- **Apache Atlas:**
  - http://atlas.incubator.apache.org/
  - http://incubator.apache.org/projects/atlas.html
  - https://git-wip-us.apache.org/repos/asf/incubator-atlas.git

- **Partner Workshops**
  - http://hortonworks.com/partners/learn/

- **More to come with official GA release of HDP 2.3**

Thank you !