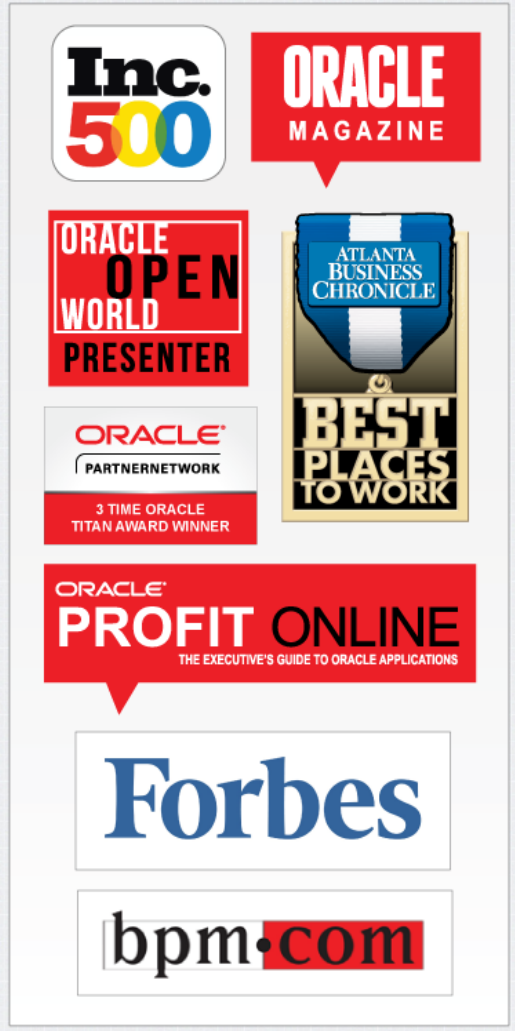
Use Cases for Governing Hadoop

About BIAS Corporation

Who We Are…

- Founded in 2000
- Distinguished Oracle Leader
  - Technology Momentum Award
  - Portal Blazer Award
  - Titan Awards
  - Excellence in Innovation Award
- Management Team is Ex-Oracle
- 250 U.S. employees & contractors, 100 India employees, average with 10+ years of Oracle experience
- Inc.500|5000 Fastest Growing Private Company in the U.S. for the 7th Time
- Voted Best Place to work in Atlanta for 3rd year
- Top 10 Healthiest Workplace in Atlanta Business Chronicle
- 33 Oracle Specializations spanning the entire stack



About the Speaker

Kenton Troy Davis

Senior Director & Enterprise Architect, BIAS Corporation

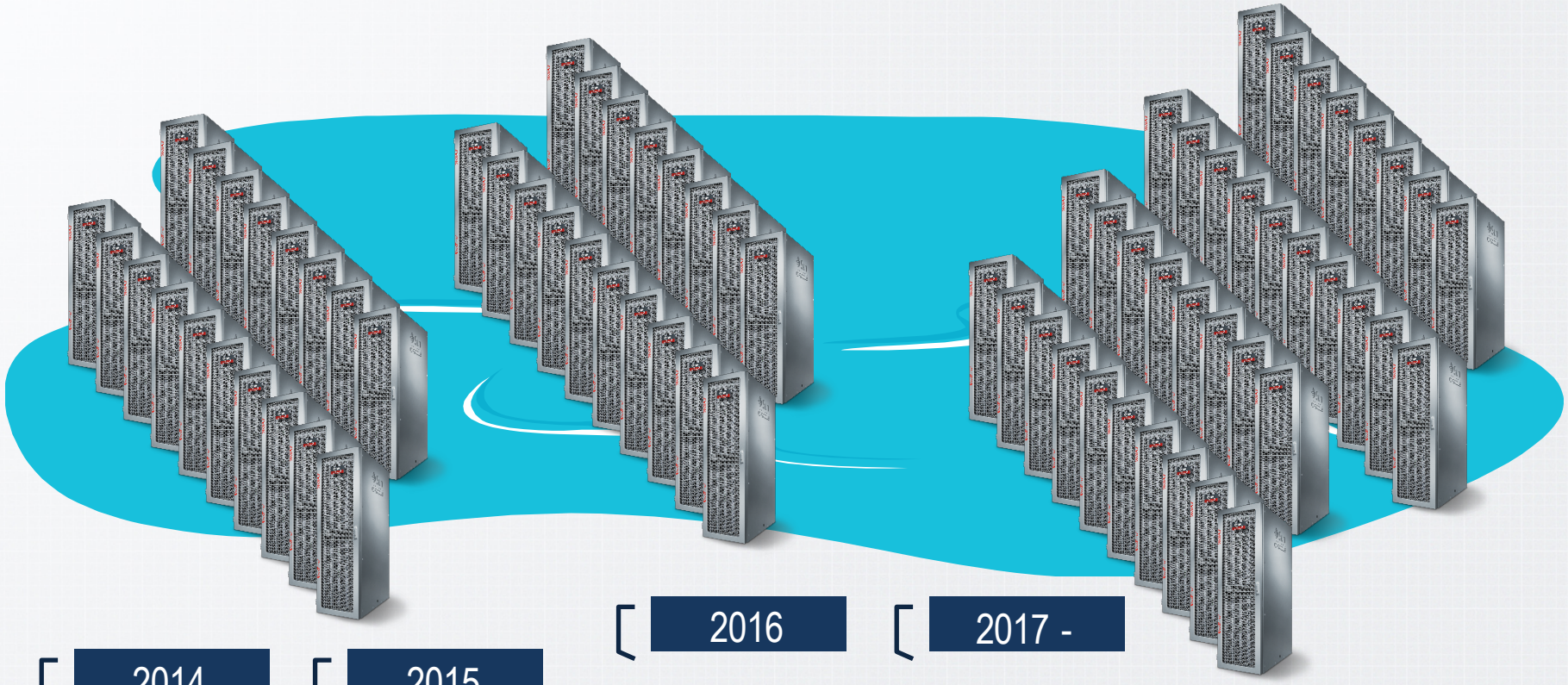
- Patented work in database security and IoT
- Oracle Alumnus
- Statistician before being labeled a Data Scientist

BIAS
Business Intelligent Application Solutions

Big Data Growth

SUNTRUST

Projected to be Largest Oracle Big Data Appliance Implementation at a Bank



2014	2015	2016	2017 -
18 Node POC	66 Nodes	192 Nodes	2-3x growth currently projected

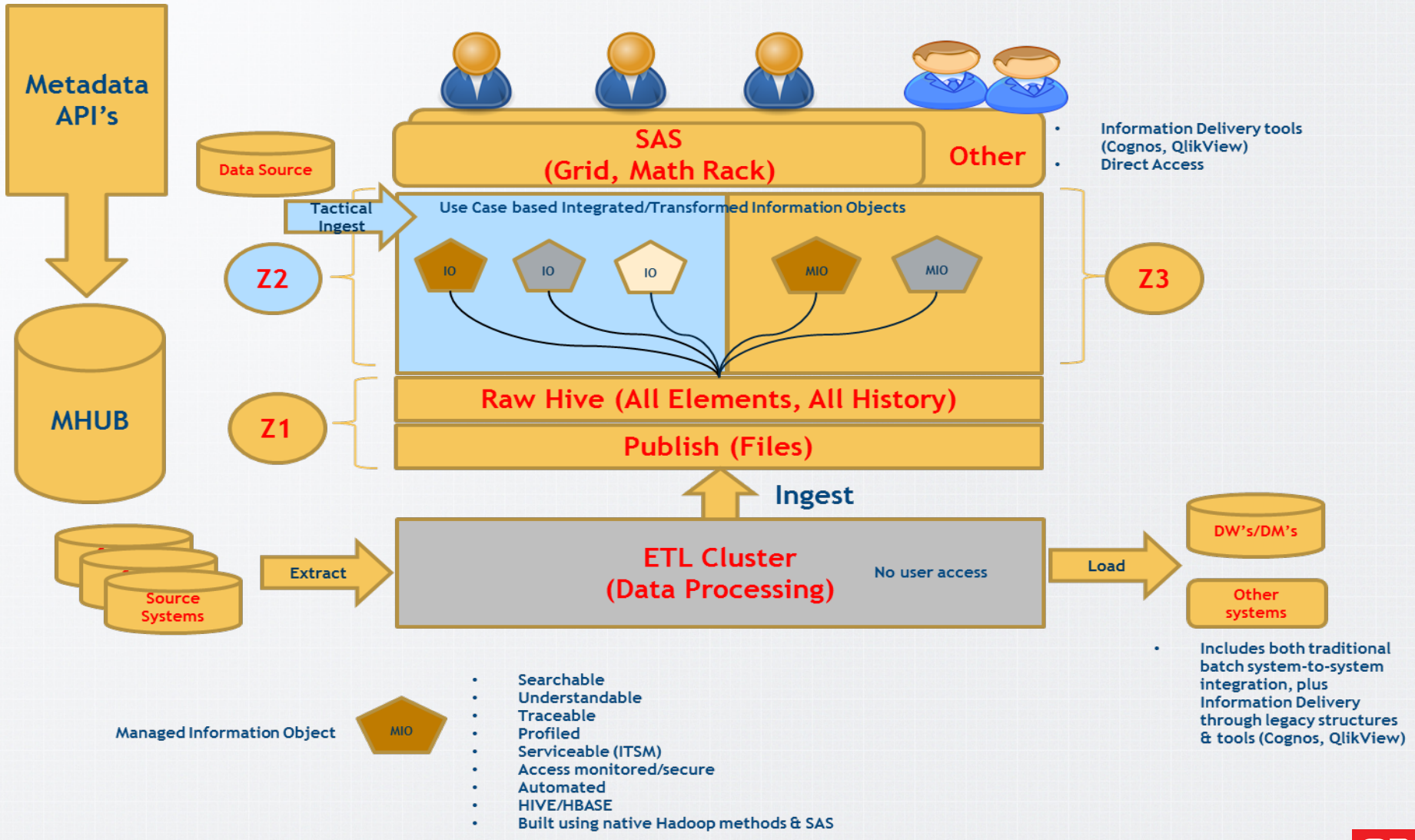
Some real-world use cases at the Bank

- Increase the data points used to profile a customer (WCV 360)
- Use analytics to derive real time offers (RTO) for customers banking at branches
- Minimize the data sprawl and establish a single source of truth (SSOT)
- Replace legacy data warehouses used for transactional inputs
- Establish better management around how data is being consumed and by whom
- Achieve all of the above with a scalable, lower-cost platform that aggregates storage

Storage Projections

Assumed consistent growth, Uncompressed estimates, Not including HDFS replication	2015	2016	2017	2018+
Social Media	23.00	23.00	23.00	23.00
IT Operational Data	11.50	11.50	11.50	11.50
Documentation, Images, Cheques Images (ECM)	57.50	57.50	57.50	57.50
Third Party Data Sources (700 Sources); Reference/ Bureau Quarterly	50.60	50.60	50.60	50.60
Bureau	8.05	8.05	8.05	8.05
Total Volume (TB)	**142.57**	**323.94**	**505.31**	**695.75**

Data Lake Architecture at the Bank



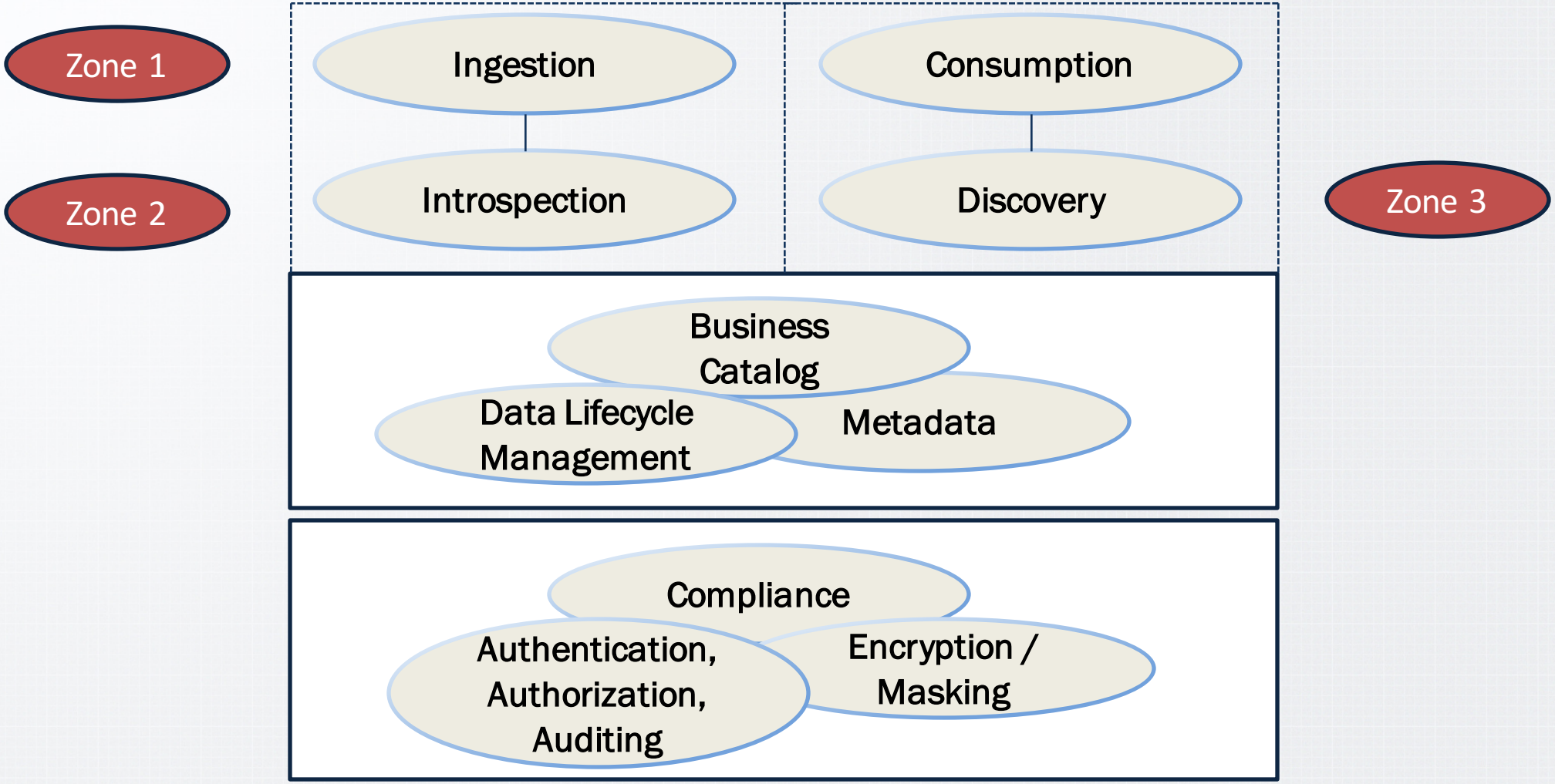
Data Wrangling Challenge

- What happens after the POCs actually work?
- What happens when internal adoption of Hadoop occurs faster than anticipated?

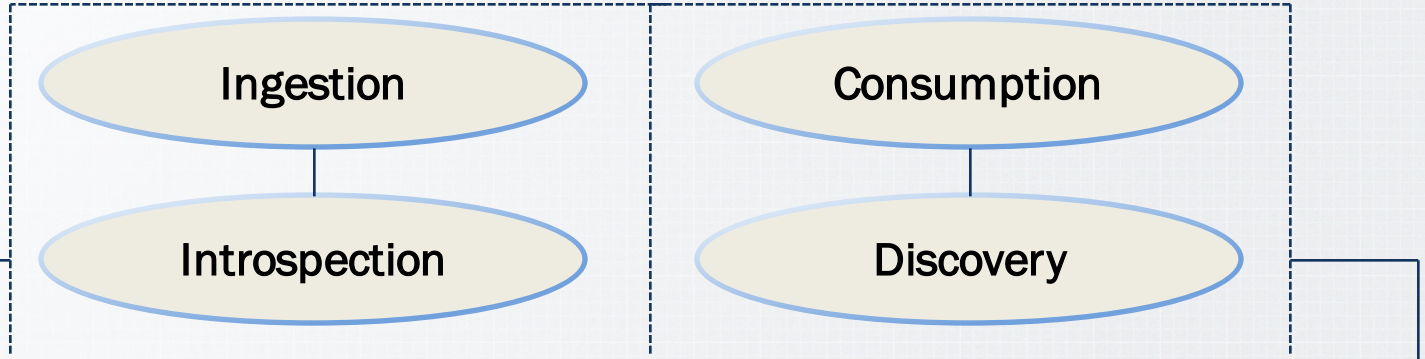
Prevent the Data Lake from becoming a Data Swamp

- Encourage consumers to collaborate via a shared data catalog
- Focus even more on data cleansing and preparation - HDFS schema-on-read encourages naïve ingestion
- Glue the Apache ecosystem and vendor tools together by linking governance to enterprise security
- 'Operationalize' all of the above

Data Governance Whiteboard



Data Governance – Introspection and Discovery



- Automate the parsing of unstructured and semi-structured feeds
- Infer data types
- Detect sensitive data (e.g. PII, PCI, HIPAA)
- Enable data stewards to interact with and adjust the process

TRIFACTA

DATAGUISE

- Search with faceted navigation
- Enable sandbox, ad-hoc queries
- Customize dashboards and reports

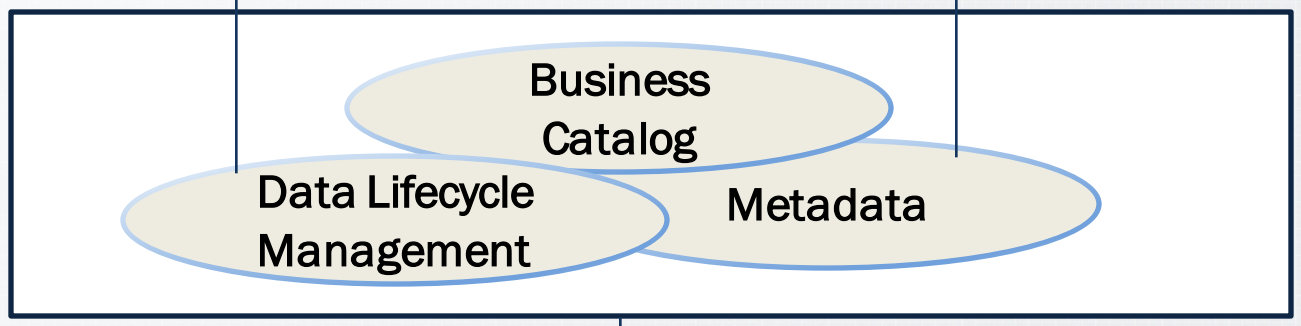
cloudera

Oracle Big Data Discovery

Data Governance – Smart Data Cataloging

- Track lineage
- Audit and report
- Maintain chain of custody

- Tag PII and PAN data
- Base tags on resource, location, or time



Better to use graph databases here

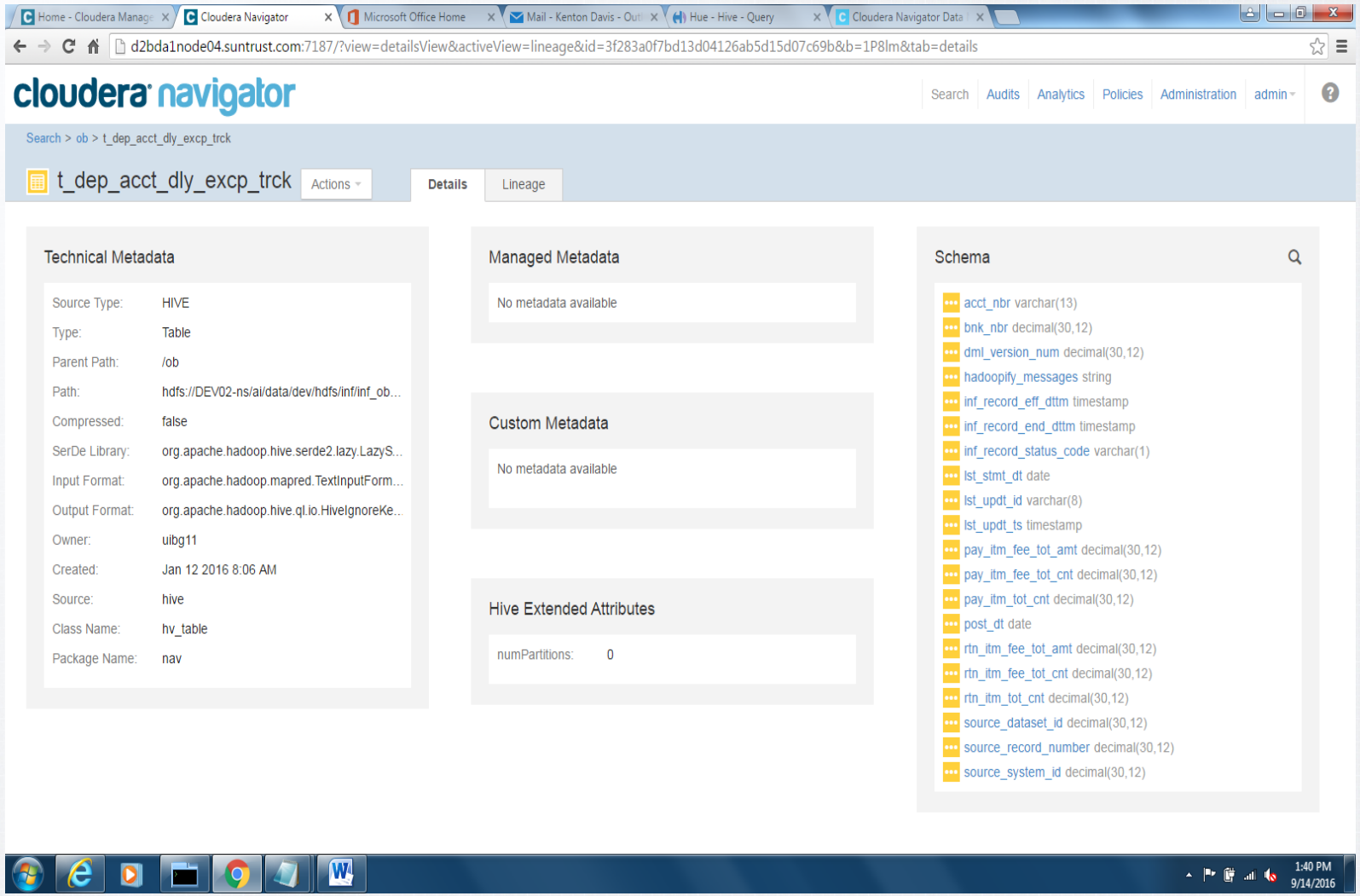
Aggregated search is crucial

Tags should trigger encryption, masking, and access rules

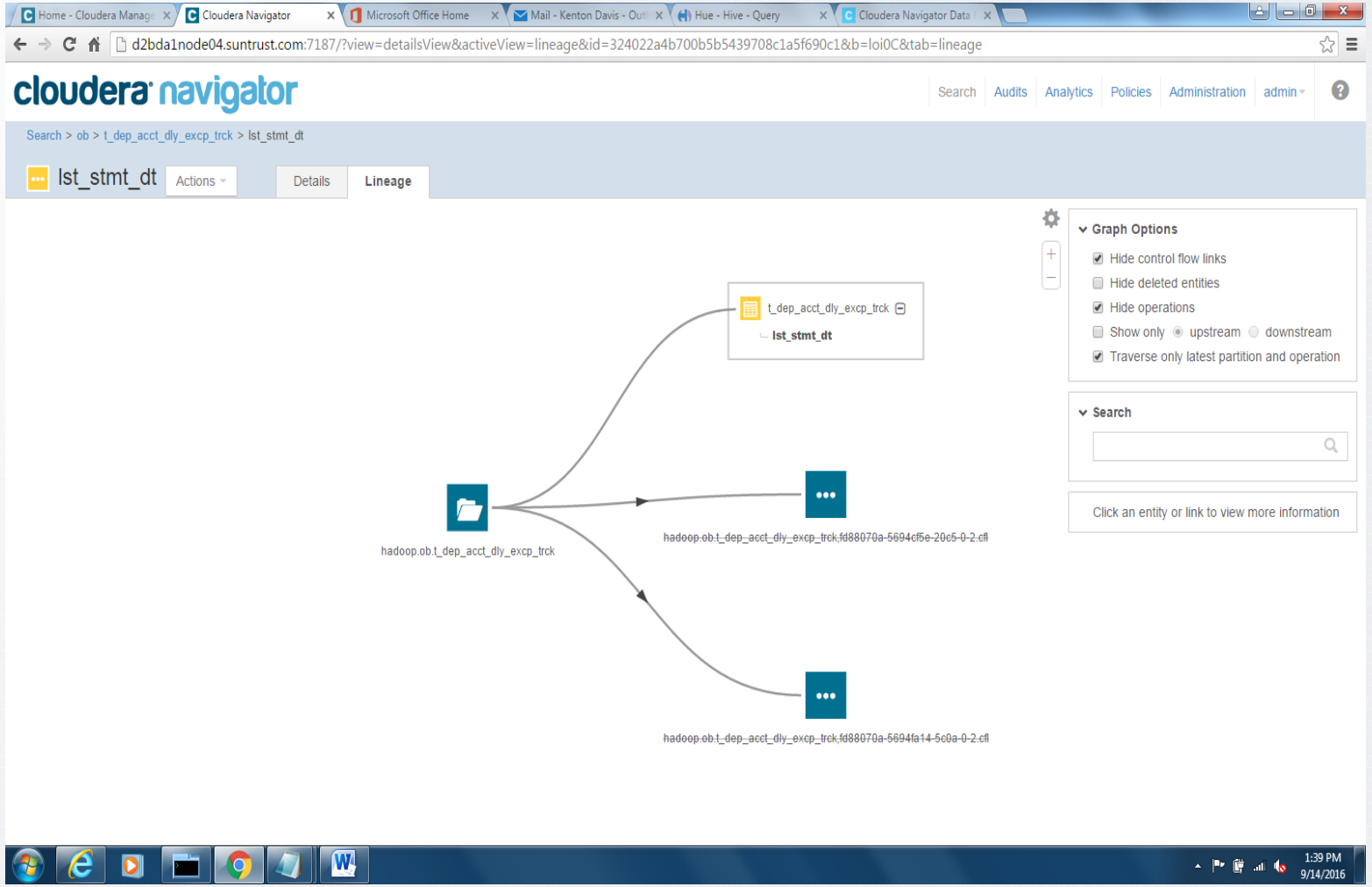
Time-based usage tracking important for subpoenas and SOX

WATERLINE DATA

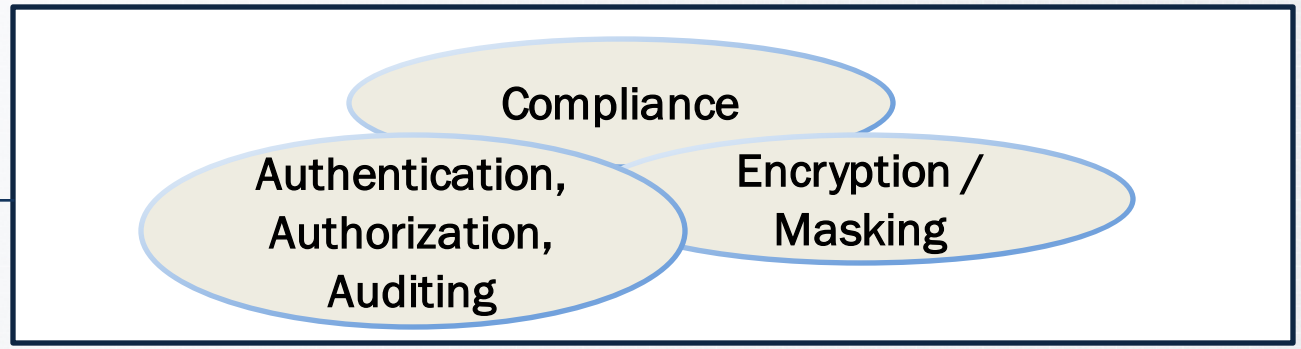
Lineage Tracking Example #1



Lineage Tracking Example #2



Data Governance – Compliance

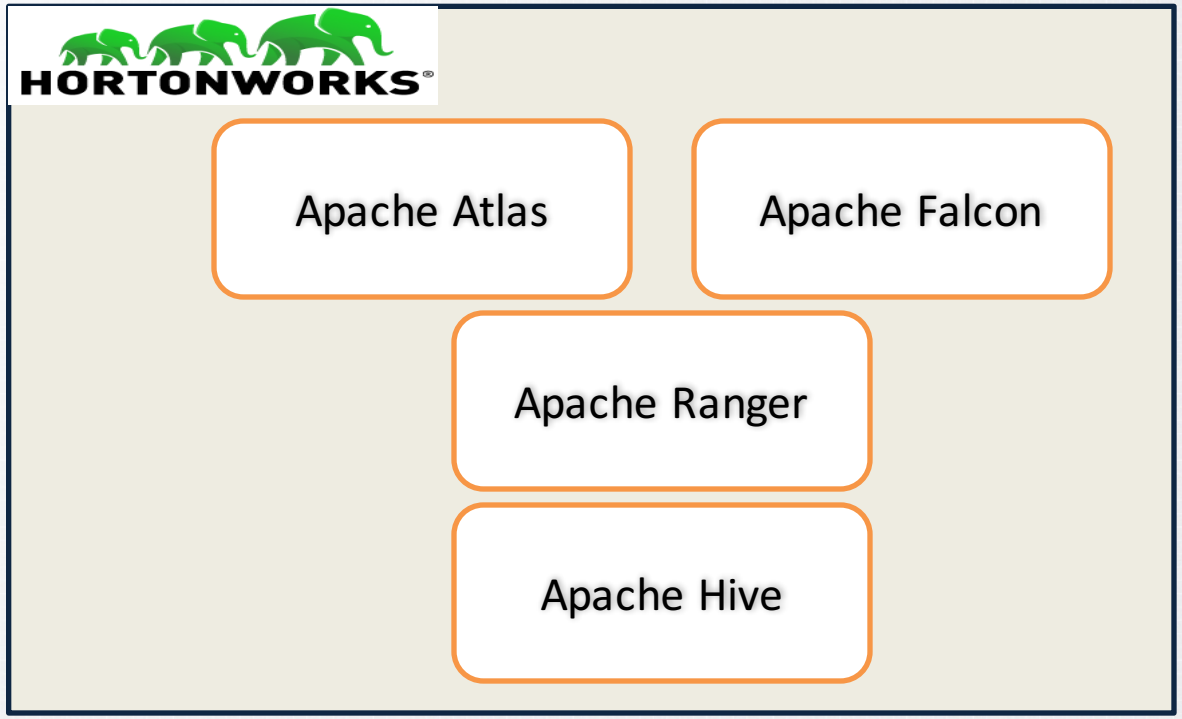


- Develop AAA policies using both tags and resources
- Integrate with Enterprise security
  1. Active Directory and LDAP
  2. Key Management Service (KMS)
- Enforce separation of duties

Data Governance – Challenges

Holistic solutions are still evolving and require plugins to various Hadoop features (e.g. HDFS abstraction is rapidly maturing beyond Hive).

Hortonworks example:



Data Governance – Challenges

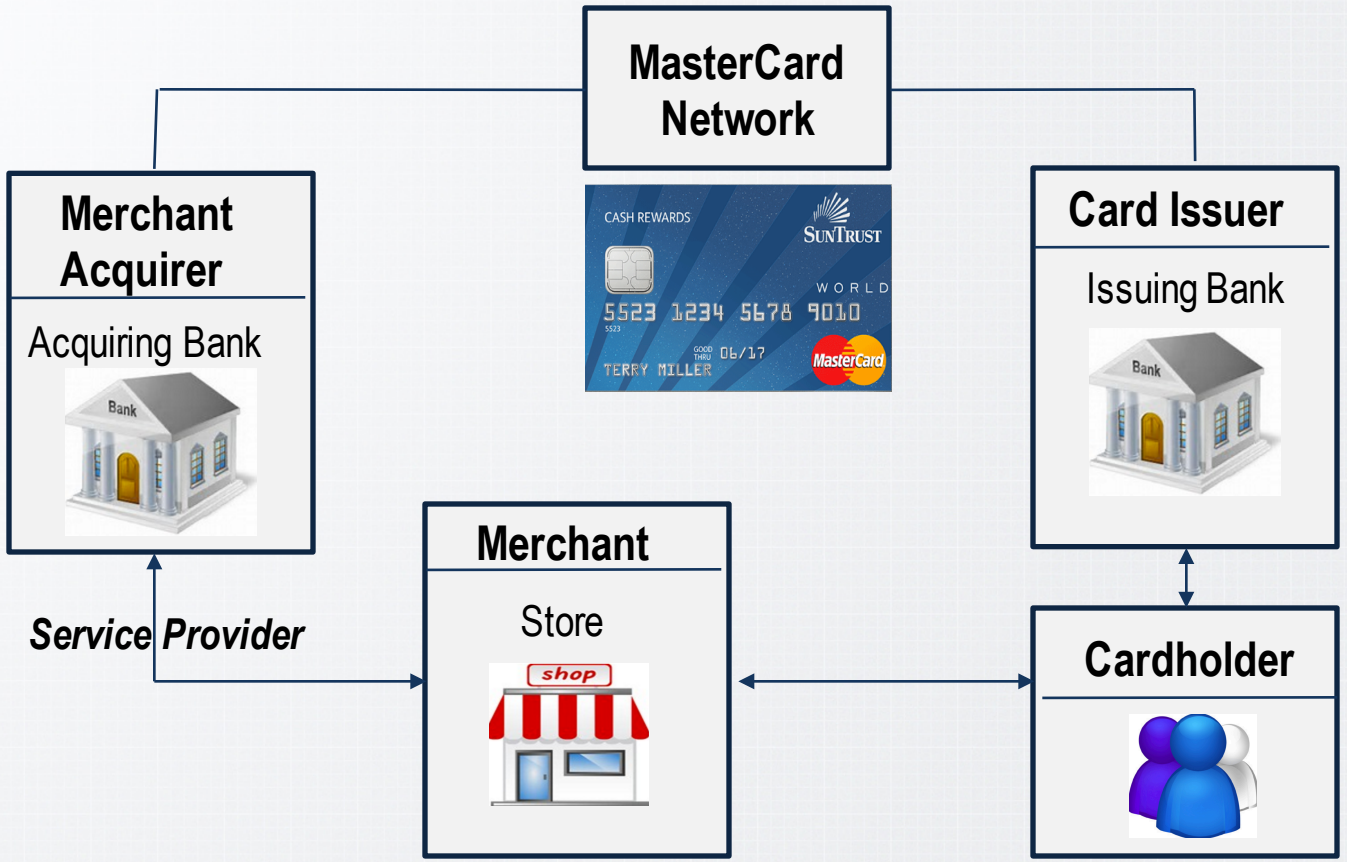
Data Lifecycle Management components become key to taking Hadoop into Production:

- Policies that are responsive to late data handling – tag mutation, rules customization
- Support for rolling upgrades and cleanup
- H/A support via replication
- Lineage tracking that is easily visible to auditors with drilldown and collapse

Metadata creation and ease of use are still evolving:

- Exchange of metadata in many cases requiring custom coding (e.g. REST/JSON)
- Tags against a parent object not following derived objects
- Need to still maintain a business taxonomy

(Appendix) Credit Card Transaction Parties



https://www.suntrust.com/personal-banking/credit-cards

(Appendix) PCI-DSS Data Security Standard V3.2

- Mask the Primary Account Number (PAN) such that at most only the first six digits and the last four digits are displayed.
- If a full unmasked PAN needs to be persisted, then it must be saved in encrypted form at rest.
- Documented procedures must exist for key management processes used for strong cryptography – e.g. for backup, key storage, key rotation *(section 3.6 sub controls)*, key access, etc.
- Principle of least privilege *(section 7)* applies by limiting data access according to which business groups 'need to know'.

(Appendix) Column Masking

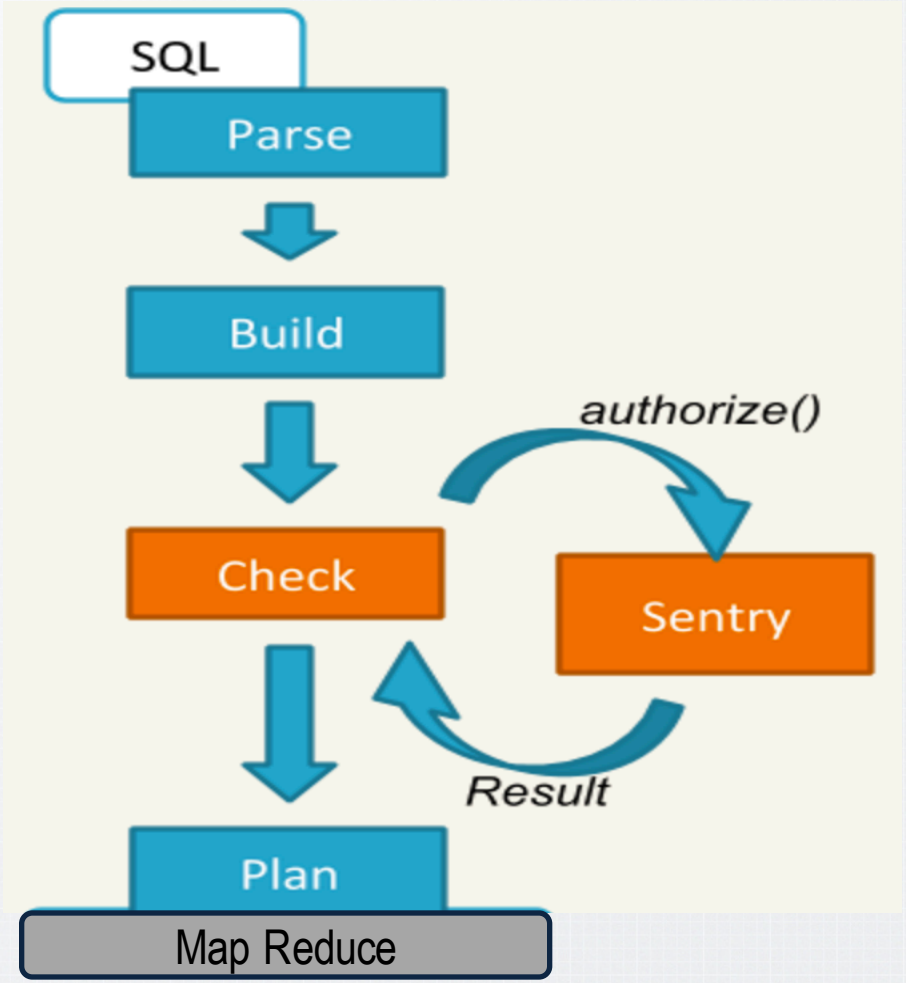
Assign Sentry privileges to view

Java User-Defined Function (UDF)

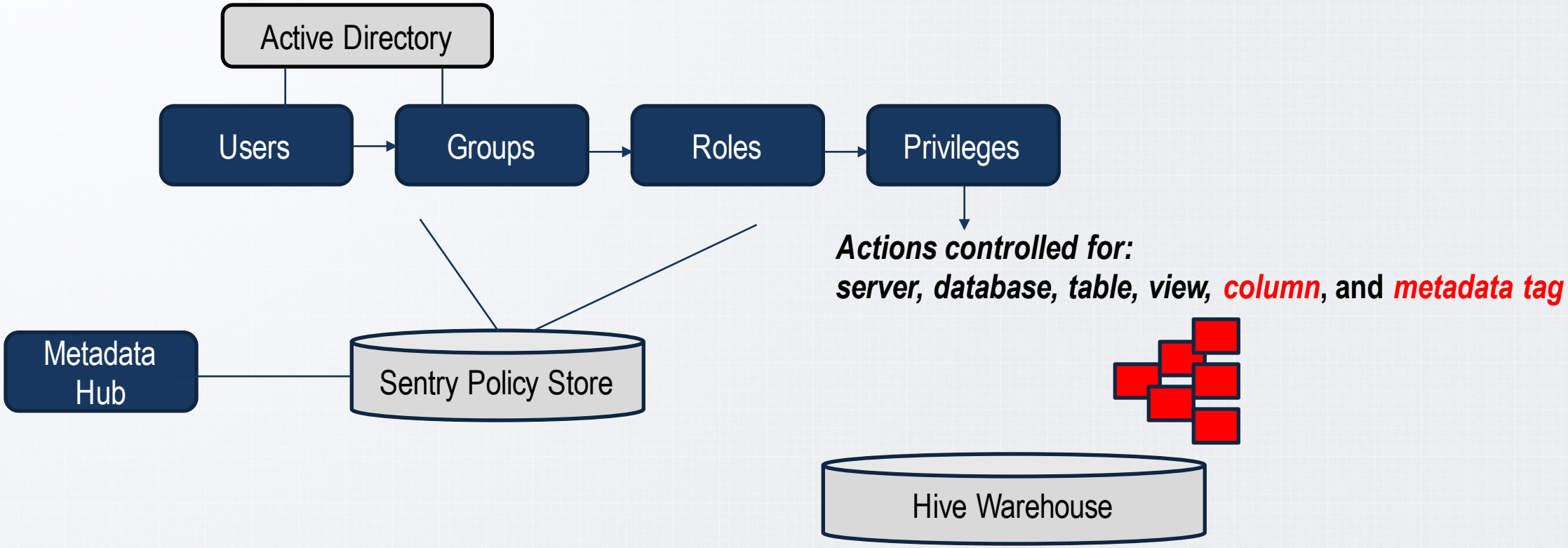
```
USE
ETL_STAGE_{source_hive_database}

CREATE VIEW PII_MASKED_EXAMPLE
as
   SELECT mask_ccn_udf(credit_card_number) as
          ccn, name, balance, region
   FROM
   ETL_STAGE_VIEW_{source_hive_database}.{Table}
   WHERE state = “VA”
```

(Appendix) HiveServer2 Hook



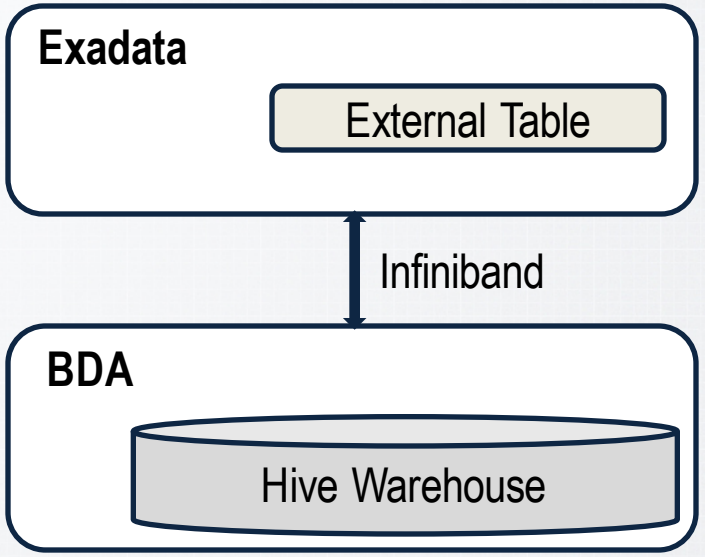
(Appendix) Apache Sentry



(Appendix) Oracle Big Data SQL

Query Franchising

CREATE TABLE … ORGANIZATION EXTERNAL (TYPE oracle_hive);



Data Redaction to transform data on-the-fly (e.g. for credit-card masking)

DBMS_REDACT.ADD_POLICY

Virtual Private Database context predicates for row-level security

DBMS_RLS.ADD_POLICY

HOW TO MOVE TO THE CLOUD

Architect

Migrate

Integrate

Secure

Manage

www.biascorp.com/cloud

BIAS
Business Intelligent Application Solutions

ORACLE Platinum Partner


Q&A

Contact Us

Kenton Davis

Kenton.Davis@biascorp.com

On LinkedIn