

Multi Tenant Security Architecture for Big Data Systems

Suresh Yadagotti Jayaram
Sr. IT Technical Architect

What is Big Data

“Big Data refers to datasets whose size and/or structure is beyond the ability of traditional software tools or database systems to store, process, and analyze within reasonable timeframes”

HADOOP is a computing environment built on top of a distributed clustered file system (HDFS) that was designed specifically for large scale data operations (e.g. MapReduce)

Reasons for securing data in Big Data systems

Contains Sensitive
Data

- Teams go from a POC to deploying a production cluster, and with it petabytes of data.
 - Contains sensitive cardholder and other customer or corporate data that must be protected

Subject to Regulatory
Compliance

Compliance to PCI
DSS, FISMA, HIPAA,
federal/state laws to
protect PII

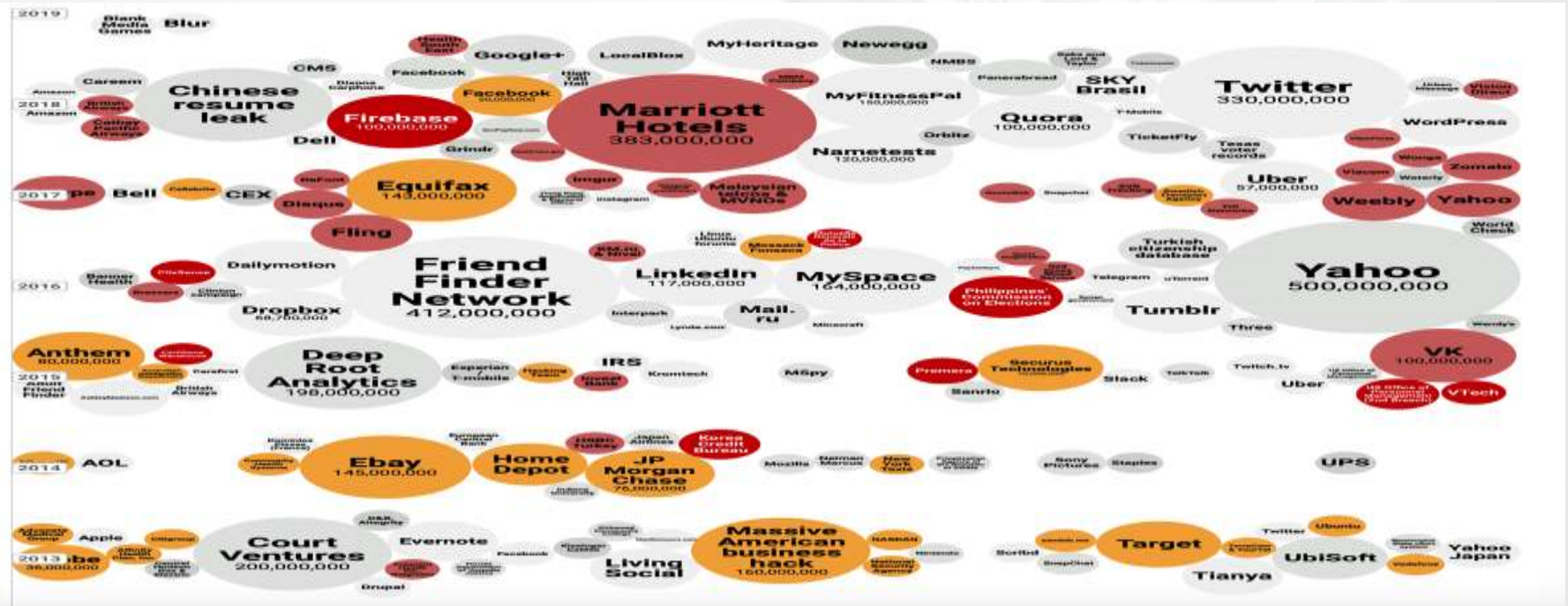
Business
Enablement

- Usage was restricted to non-sensitive data
 - Allow access to restricted datasets with Security

es & Hacks

hed. Healthcare, Retail, Federal Govt., Financial
companies etc.

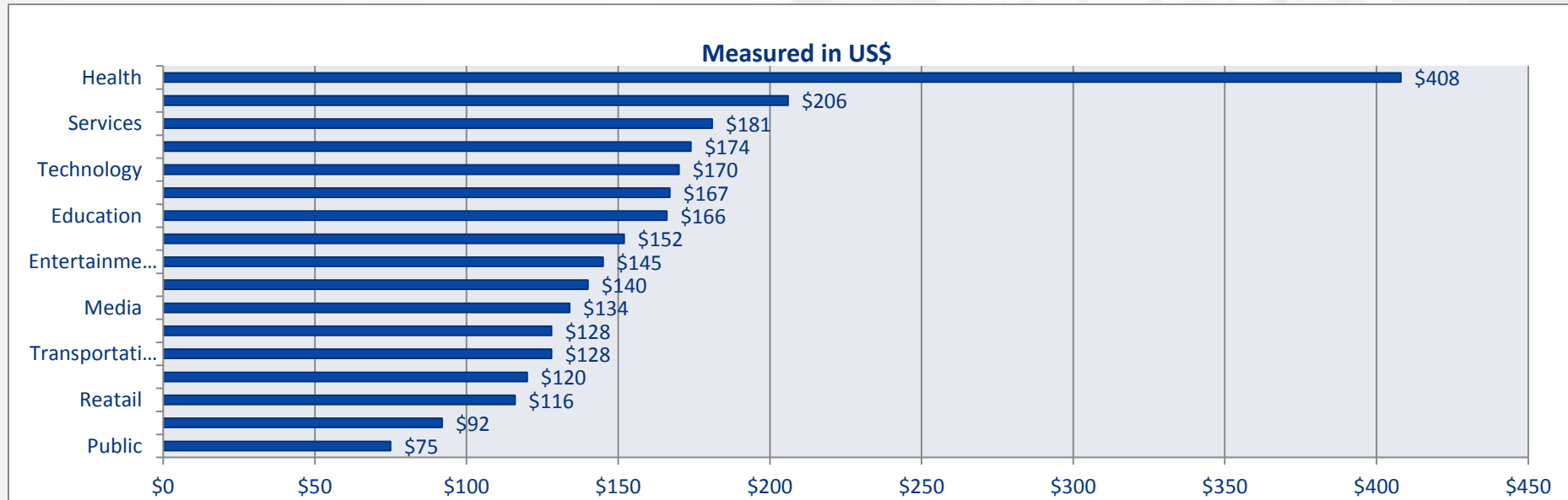
hed. Healthcare, Retail, Federal Govt., Financial companies etc.



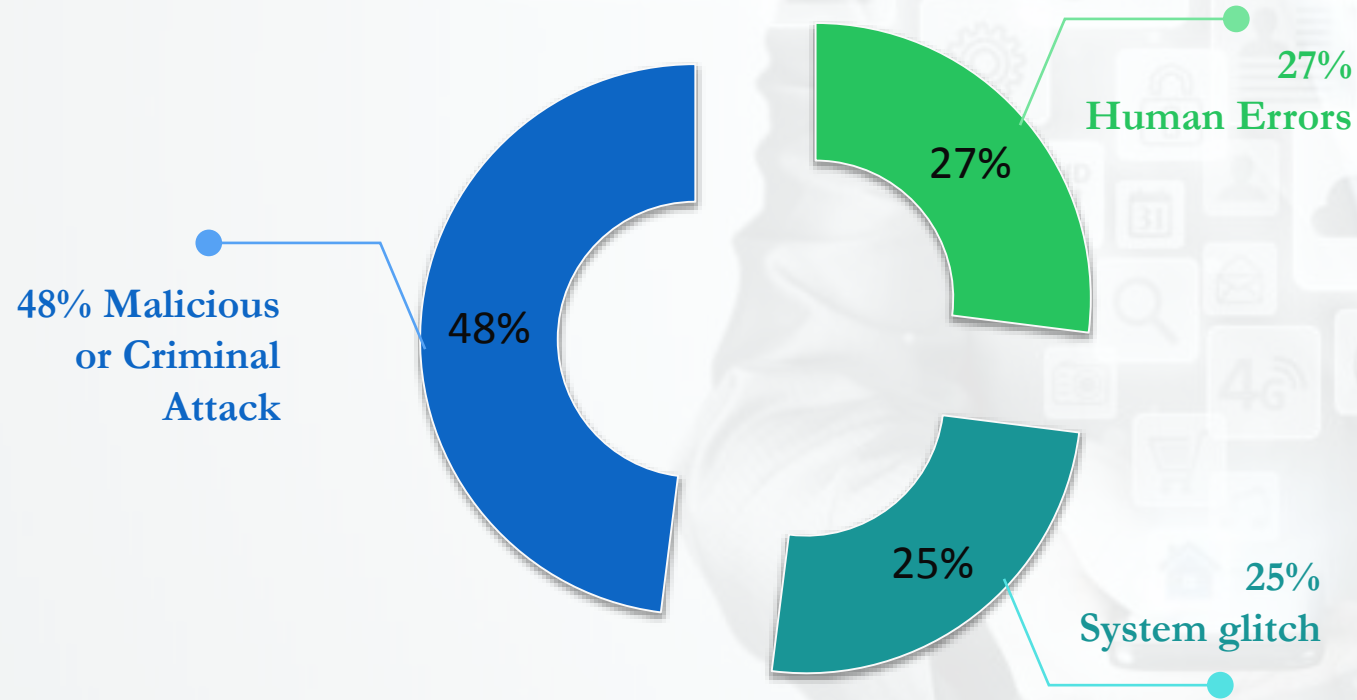
Per capita cost – Industry Sector

Certain industries have higher data breach costs. compares 2018 year's per capita costs for the consolidated sample by industry classification.

As can be seen, heavily regulated industries such as healthcare and financial organizations have a per capita data breach cost substantially higher than the overall mean.



Root Causes



Goals of an Attacker

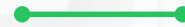
01



The primary goal is to obtain sensitive data that sits in Organization Databases

02

This could include different kinds of regulated data (e.g. Payment data, Health data) or other personally identifiable data (PII)



03

Other attacks could include attacks attempting to destroy or modify data or prevent availability of this platform.

Threats

Attacker attempts to gain privileges to access data

Unauthorized access

- Authentication
- Authorization
- Auditing

Network Based Attacks

- Transport Layer Security
- SASL Encryption

Types of Threats

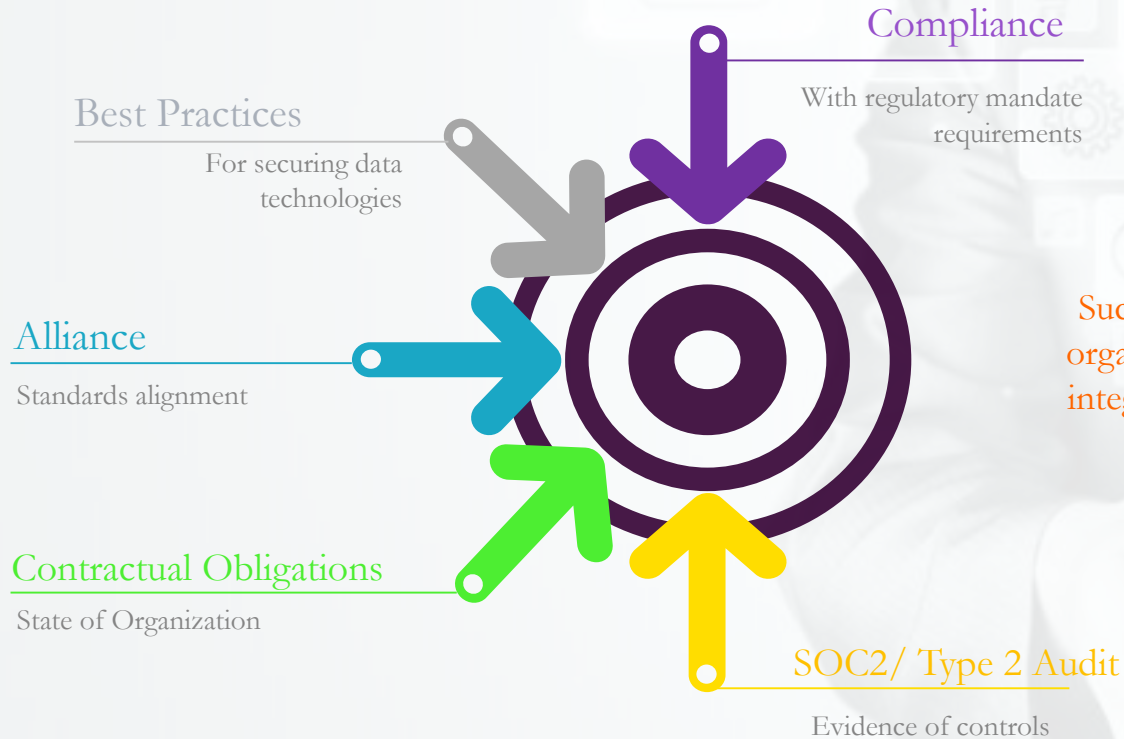
Host Level Data at Rest Attacks

- Application Level
- HDFS level
- File System/Volume level

Infrastructure Security

- Automation
- SELinux

Security Objectives

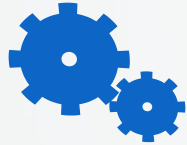


“It’s all about the data.”

Successful implementation of Data Lakes in organizations will demonstrate confidentiality, integrity, and availability across the enterprise.

Achieve Secure Data Enablement

By understanding the key criteria:



GOVERNANCE

- Knowing what the information *is*
- What is the function of the data?



USERS

- *Who* is using the data?
- Who needs what kind of access?



LIFECYCLE

- How does information connect across systems?
- What are retention requirements for the data?



CONTROLS

- Engage early to understand controls complexity
- Know the value & risk factors indicated by the data & solutions.

Data level hierarchy & OBJECTIVES



Enterprise Level

Enterprise is the highest level and any data stored at this level is visible / available for all the tenants (geographical data, code sets, etc.)



Tenant Level

To minimize the impact to the existing legacy systems and home-grown services, we will use the additional attributes like “Tenant ID” and “Data Delimiters” to identify which records belong to which tenant. Members can have multiple records in the same system with different Tenant ID’s in case s/he purchased products from more than one tenant.



Domain Level

Application Layer/Domains to control access and/or capabilities (such as LOB, group, segment, or other data restrictions or classifications) within the tenants they use. Application layer to control what the constituent experiences, what data they can access, and how.



Database/Table

Every data set will include audit attributes such as:

- Who is providing the data? ,
- What data is being collected? ,
- When the data is collected? ,
- Where the data is collected from?
- Why is the data collected ?

Enterprise level objectives



Enterprise Level Data will...

- Be visible & available to ALL tenants
 - Data Classified, labeled, or segregated in a manner that indicates it has been approved for enterprise wide use (classification is TBD) which may include Geographical data, code sets, etc.
 - Data Classified as Public
- Support both internal and external users depending on classification
- Internal users get access through an application Id or directly with User Id

Tenant level OBJECTIVES



Enterprise



Tenant Level Data will...

- Support multiple tenants
- Be segregated logically (tagged, labeled, or container segregated based on tenant ID or data delimiters, not physically where possible based on controls objectives for organizations)
- Be co-mingled; all applications are storing data together with the following defaults:
 - Logical separation when applicable (controlled by Ranger Policies and data object implementation)
 - Default = Applications (Different Log Locations). Services (Ex; Ranger. Same Log locations).
- Use an additional fields: Tenant ID and Data Delimiters
 - This minimizes impact to existing legacy systems and home-grown services
 - Tenant IDs and Data Delimiters will be used in tables to identify which records belong to which tenant and Enterprise Line of Business.
- Use applications to enforce 100% usage of Tenant IDs and Data Delimiters verified through exceptions, audit & recon
- Adhere to the original idea of *Individuation*—each individual should be identified as one individual in the Individuation database, regardless of whether s/he has bought products from more than one tenant.
 - S/he can have multiple records in the same system with different Tenant ID's in case s/he purchased products from more than one tenant.

Domain Level OBJECTIVES



Enterprise



Tenant Level

</> Domain Level Data will...

- Control access and/or capabilities (such as LOB, group, segment, etc.) within the tenants they use
- Include application layer that controls what the constituent experiences or what data they may access
 - Also controls *how* the constituent accesses the data

Database Level OBJECTIVES



Enterprise



Tenant Level Data



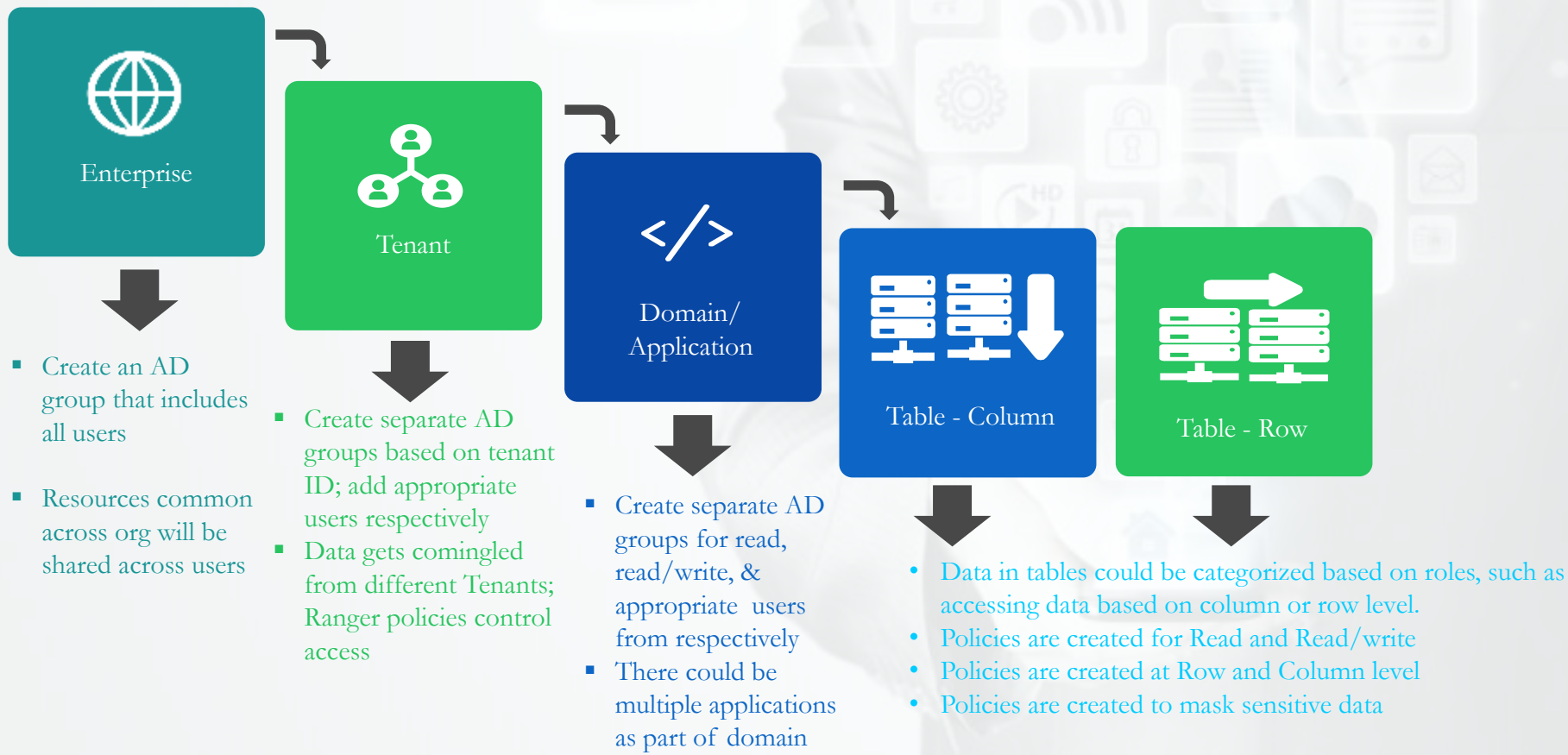
Domain



Database/Application Level Data will ...

- Retain data classifications as they exist today
 - For employee/state/federal employee, etc.
- ePHI attribute classification and inventory
- User Permissions/Authorizations
- Include audit attributes that answer the following questions for *every* dataset:
 - Who provided the data?
 - What data was collected?
 - When was the data collected?
 - From where is the data collected?
 - Why is the data collected ?
- Data activity monitoring - Who accessed, when accessed, where accessed

Data Handling – Tenant, Domain, Application, Database, Table (Row & Column) Level



Five Pillars of Security



1

Administration

Central Management & Consistent Security

2

Authentication

Authenticate Users and System

3

Authorization

Provision Access to Data

4

Data Protection

Protect Data at Rest & in Motion

5

Audit

Maintain a record of Data Access

Ranger – Centralized Administration

Central Management & Consistent security

Single pane of glass for security administration across multiple Hadoop Components for Creating, implement, Manage and Monitor Security Policies

Ranger

Access Manager

Audit

Settings

g5s0

Service Manager

Service Manager

Import

Export

HDFS

+ [icon] [icon]

udahdpdev_hadoop

[icon] [icon] [icon]

HBASE

+ [icon] [icon]

udahdpdev_hbase

[icon] [icon] [icon]

HIVE

+ [icon] [icon]

udahdpdev_hive

[icon] [icon] [icon]

YARN

+ [icon] [icon]

KNOX

+ [icon] [icon]

STORM

+ [icon] [icon]

SOLR

+ [icon] [icon]

KAFKA

+ [icon] [icon]

NIFI

+ [icon] [icon]

ATLAS

+ [icon] [icon]

udahdpdev_atlas

[icon] [icon] [icon]

Ranger – Authorization Policies

Consistent authorization policy structure across Hadoop components

The screenshot displays the Ranger web interface, which is used for managing authorization policies across Hadoop components. The interface is divided into several sections:

- Policy List:** A table showing a list of policies. The columns include Policy ID, Policy Name, Policy Owner, Policy Status, and Policy Type. The table lists several policies, including 'all - all', 'all - all', 'all - all', 'all - all', 'all - all', 'all - all', 'all - all', 'all - all', 'all - all', and 'all - all'.
- Policy Details:** A section for viewing the details of a selected policy. It includes fields for Policy Name, Policy Owner, Policy Status, Policy Type, and Policy Description. It also shows a list of permissions and a list of conditions.
- Policy Configuration:** A section for configuring a policy. It includes fields for Policy Name, Policy Owner, Policy Status, Policy Type, and Policy Description. It also shows a list of permissions and a list of conditions.
- Policy Actions:** A section for performing actions on a policy, such as creating, updating, deleting, and viewing the details of a policy.

The interface is designed to be consistent across different Hadoop components, ensuring that authorization policies are managed in a uniform manner.

Ranger – Row-filter, Column-masking

Ranger Enterprise Manager Audit Settings

Security Manager | Configuration | Audit | Settings

Edit Policy

Please ensure that every group listed in this policy has access to the column via a column policy. This policy does not inherit its group access to the column.

Policy Details:

Policy Type: **Column**

Policy ID: **1**

Policy Name: **PersonnelDataMask**

Policy Label: **Personnel data**

Host Database: **PERSONAL**

Host Table: **PERSONAL_INFORMATION**

Host Column: **ADDRESS**

Description: **Masking Address for Finance Data**

Audit Logging: **ALL**

Mask Conditions:

Masked Group	Masked Data	Access Types
PERSONAL	PERSONAL	SELECT

Select Masking Option

- ☐ None
- ☐ Partial mask: show first & last
- ☐ Partial mask: show first & last
- ☐ Full
- ☐ Nullify
- ☐ Unmasked: return original value
- ☐ None: show only user
- ☐ Custom

Save **Cancel** **Delete**

Ranger Enterprise Manager Audit Settings

Security Manager | Configuration | Audit | Settings

Edit Policy

Please ensure that every group listed in this policy has access to the column via a column policy. This policy does not inherit its group access to the column.

Policy Details:

Policy Type: **Row-Level**

Policy ID: **1**

Policy Name: **PersonnelDataMask**

Policy Label: **Personnel data**

Host Database: **PERSONAL**

Host Table: **PERSONAL_INFORMATION**

Description: **Masking Address for Finance Data**

Audit Logging: **ALL**

New Filter Conditions:

Select Group	Select Data	Access Types
PERSONAL	PERSONAL	SELECT

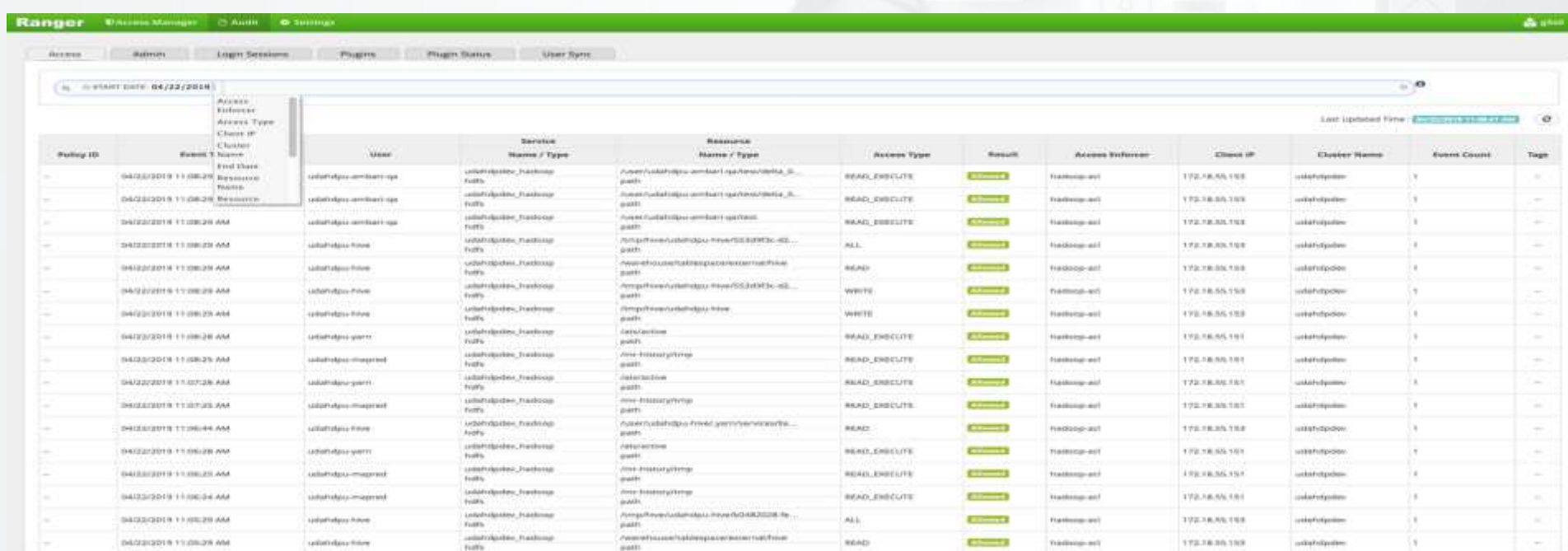
Filter Expression

Save **Cancel** **Delete**

Ranger – Access Audit Logs

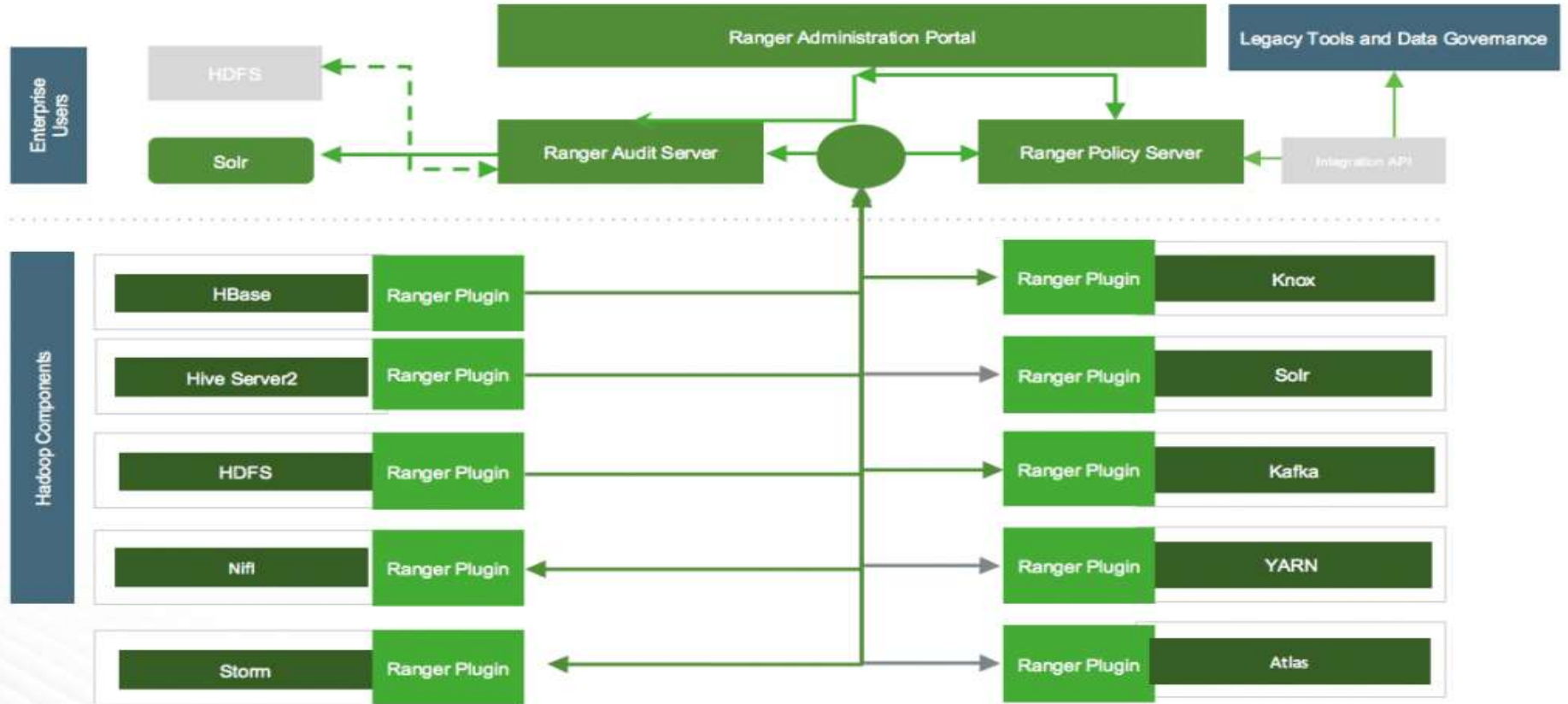
Apache Ranger generates detailed logs of access to protected resources
Audit logs to multiple destinations like HDFS, Solr and Log4j appender

Interactive view of audit logs in Admin console



Policy ID	Event Time	User	Service Name / Type	Resource Name / Type	Access Type	Result	Access Referrer	Client IP	Cluster Name	Event Count	Tags
—	04/22/2018 11:08:20	udshdpdw-ambant-ops	udshdpdw_hadoop	hdfs://udshdpdw-ambant-ops/warehouse/_...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20	udshdpdw-ambant-ops	udshdpdw_hadoop	hdfs://udshdpdw-ambant-ops/warehouse/_...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-ambant-ops	udshdpdw_hadoop	hdfs://udshdpdw-ambant-ops/warehouse/_...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	ALL	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	READ	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	WRITE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	WRITE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-gerrit	udshdpdw_hadoop	hdfs://udshdpdw-gerrit/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:08:20 AM	udshdpdw-magreed	udshdpdw_hadoop	hdfs://udshdpdw-magreed/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:07:28 AM	udshdpdw-gerrit	udshdpdw_hadoop	hdfs://udshdpdw-gerrit/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:07:28 AM	udshdpdw-magreed	udshdpdw_hadoop	hdfs://udshdpdw-magreed/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:06:44 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	READ	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:06:28 AM	udshdpdw-gerrit	udshdpdw_hadoop	hdfs://udshdpdw-gerrit/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:06:28 AM	udshdpdw-magreed	udshdpdw_hadoop	hdfs://udshdpdw-magreed/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:06:24 AM	udshdpdw-magreed	udshdpdw_hadoop	hdfs://udshdpdw-magreed/...	READ_EXECUTE	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:06:20 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	ALL	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—
—	04/22/2018 11:06:20 AM	udshdpdw-riwe	udshdpdw_hadoop	hdfs://udshdpdw-riwe/US33393C-42...	READ	Success	hadoop-ec2	172.16.30.153	udshdpdw	1	—

Ranger – Architecture



Questions

The background of the slide features a hand holding a smartphone. The phone's screen is filled with a grid of various application icons, including a person icon, a gear, a Wi-Fi symbol, a document, a cloud, a magnifying glass, and a globe. The entire image is overlaid with a semi-transparent blue and green gradient, with the word 'Questions' centered in white.