

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

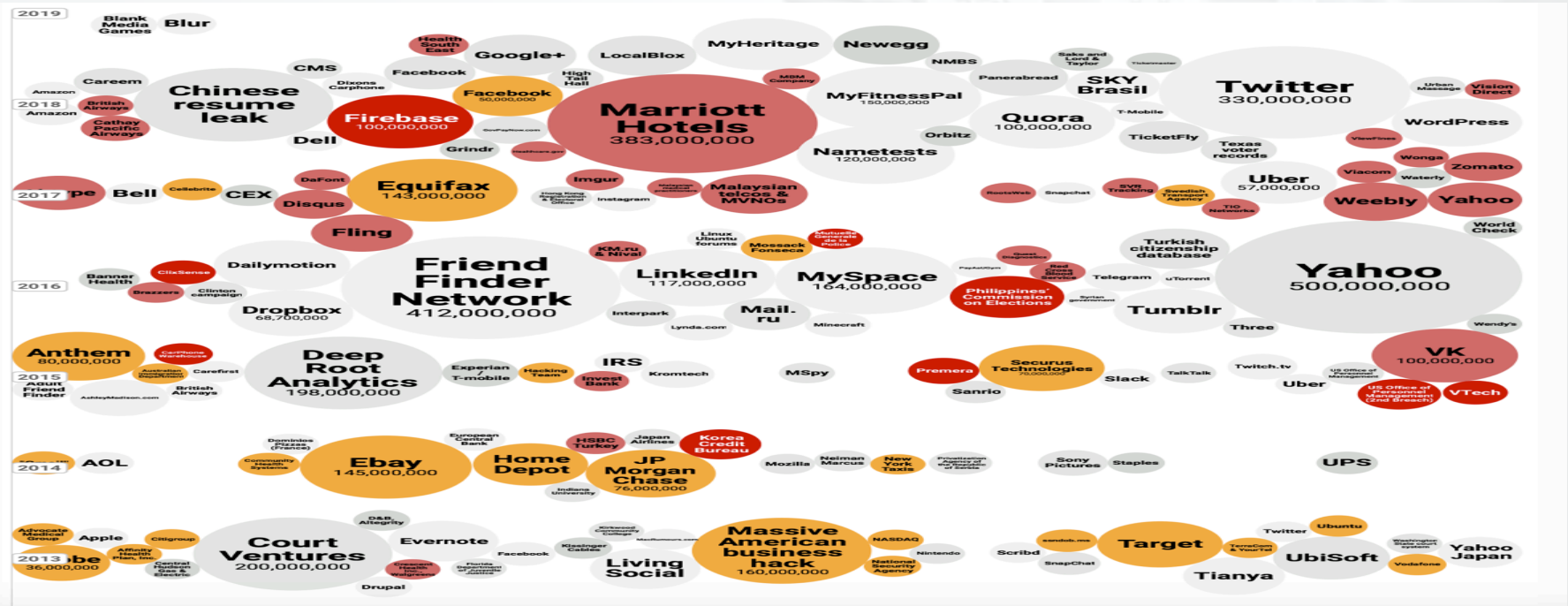
Compliance to PCI
DSS, FISMA, HIPAA,
federal/state laws to
protect PII

Business
Enablement

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

Data Breaches & Hacks

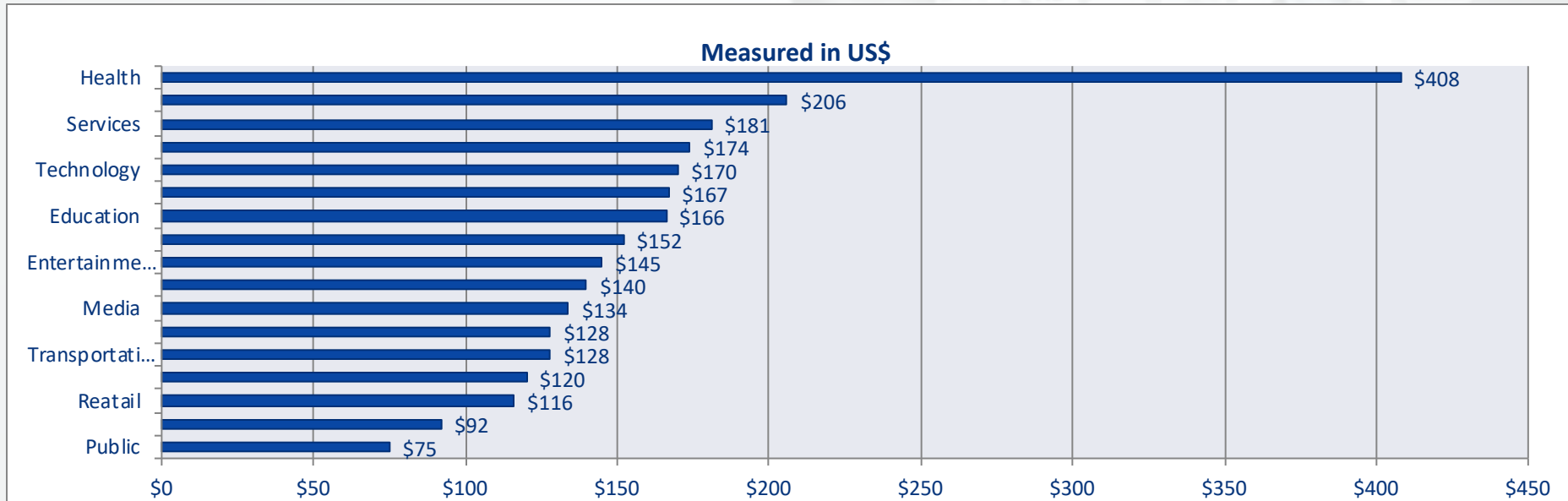
Different kinds of PII, financial data, and IP breached. Healthcare, Retail, Federal Govt., Financial Institutions, Tech companies etc.



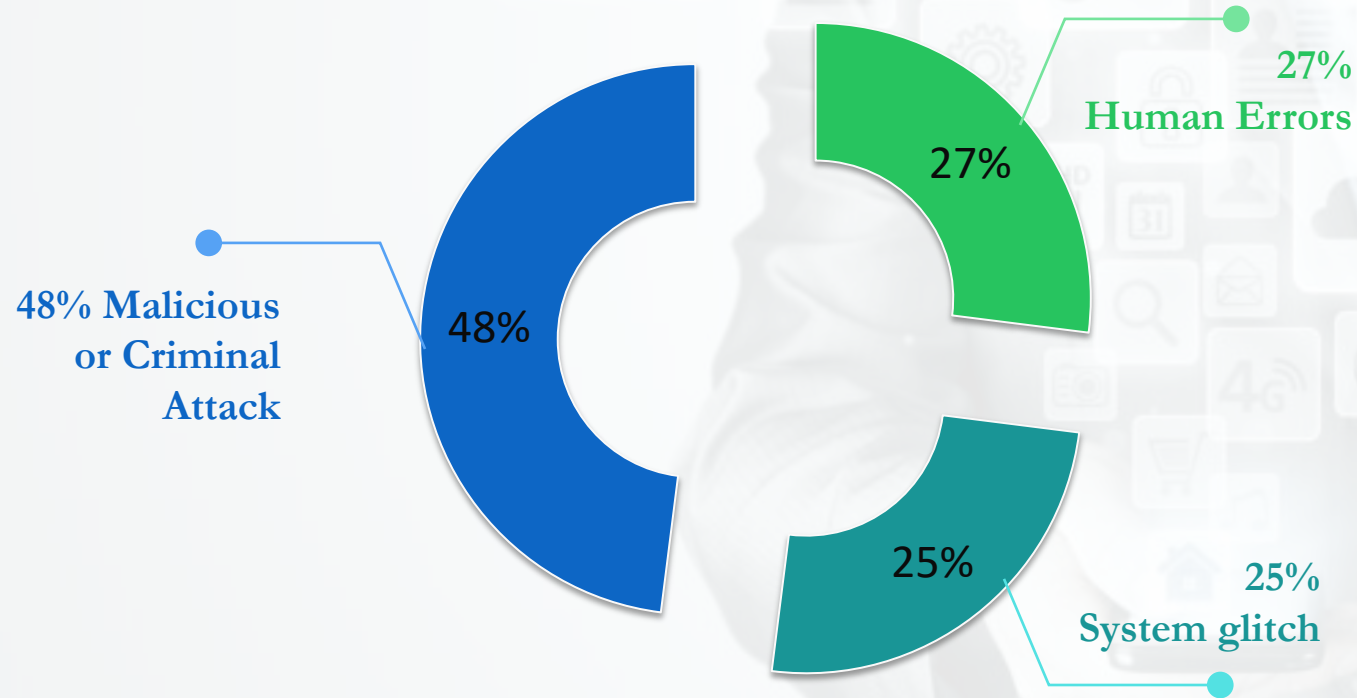
Per capita cost – Industry Sector

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



02



03

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

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

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

Threats – Counter measures

Attacker attempts to gain privileges to access data

Unauthorized access

- Authentication
- Authorization
- Auditing

Network Based Attacks

- Transport Layer Security
- SASL Encryption

Types of Threats

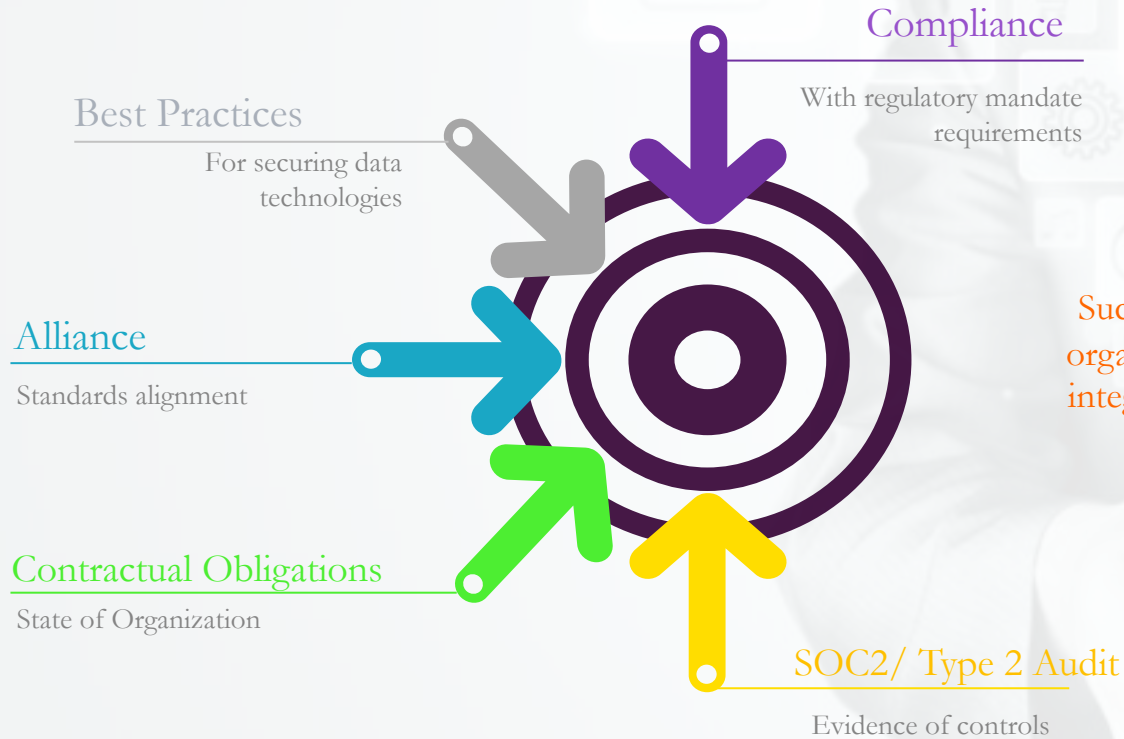
Host Level Data at Rest Attacks

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

Infrastructure Security

- Automation
- SELinux

Security Objectives

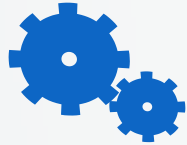


“It’s all about the data.”

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 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

Domain Level OBJECTIVES



Enterprise



Tenant Level

</> Domain Level Data will...

- Control access and/or capabilities 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



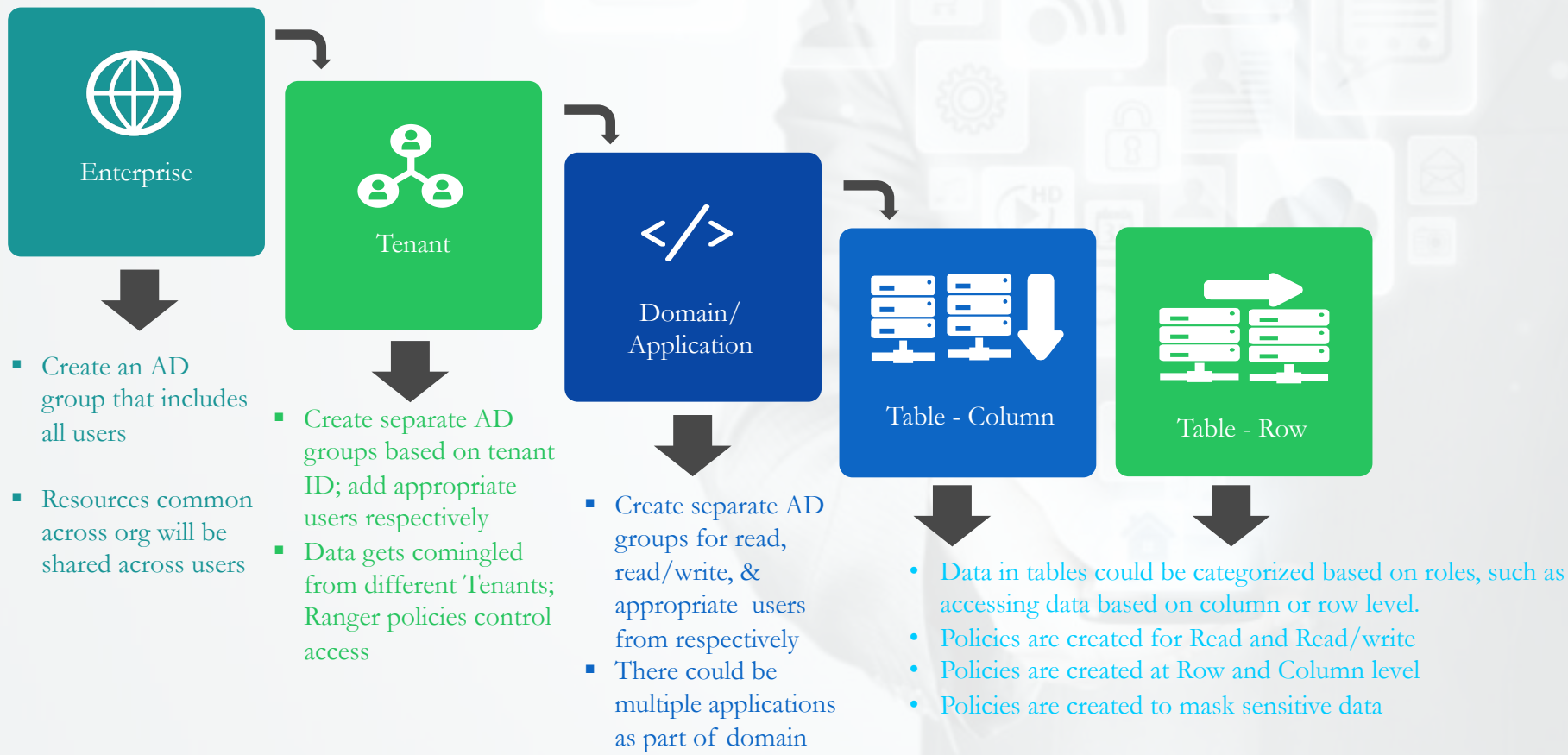
Domain



Database/Application Level Data will ...

- 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 image displays the Ranger Access Manager interface, which is used for managing authorization policies across Hadoop components. The interface is divided into two main sections: a top navigation bar and a main content area.

Top Navigation Bar: The bar includes the Ranger logo, tabs for Access Manager, Audit, and Settings, and a user profile icon in the top right corner.

Main Content Area:

- Service Manager:** A dropdown menu showing the selected service, "udahdpdev-hive Policies".
- Access Manager:** A section with tabs for Access, Masking, and Row Level Filter.
- List of Policies:** A table listing policies for the "udahdpdev-hive" service. The table has columns for Policy ID, Policy Name, Policy Labels, Status, Audit Logging, Groups, Users, and Action. Policies listed include "all - hiveservice", "all - global", "all - uri", "all - database, table, column", "all - database, udf", "Abinitio_poc", "abinitio_poc_udf", "default", and "n3test".
- Policy Configuration Form:** A detailed view of a policy configuration. It includes fields for Policy Type (Access), Policy Name (Finance Read), Policy Label (Policy Label), database (database), table (table), Hive Column (Hive Column), Description (Ranger Policies for Financial Data - Read Only), and Audit Logging (YES). It also features a "Add Permissions" button and a "Deny Conditions" section.
- Permissions Dialog:** A modal dialog box titled "add/del permissions" that allows users to select or deselect permissions. The permissions listed are: select, update, create, drop, alter, index, lock, read, write, replicate, service admin, temporary udf admin, and select/deselect all.

Row-filter, Column-masking

Ranger
Access Manager
Audit
Settings

Service Manager
udahdpdev.hive Policies
Edit Policy

Edit Policy

Please ensure that users/groups listed in this policy have access to the column via an **Access Policy**. This policy does not implicitly grant access to the column.

Policy Details :

Policy Type

Masking

Policy ID

32

Policy Name *

FinanceReadMask

enabled

Normal

Policy Label

Policy Label

Hive Database *

u_testing

Hive Table *

u_general_information

Hive Column *

u_address

Description

Masking Policies for Finance Data

Audit Logging

YES

Select Masking Option

☐ Redact
 ☐ Partial mask: show last 4
 ☐ Partial mask: show first 4
 ☒ Hash
 ☐ Nullify
 ☐ Unmasked (retain original value)
 ☐ Date: show only year
 ☐ Custom

Mask Conditions :

Select Group	Select User	Access Types
u_users	Select User	select

Save

Cancel

Delete

Ranger

Access Manager

Audit

Settings

Service Manager

udafdpdpv_line Policies

Edit Policy

Edit Policy

Please ensure that users/groups listed in this policy have access to the table via an Access Policy. This policy does not implicitly grant access to the table.

Policy Details:

Policy Type

Row Level Filter

Policy ID

03

Policy Name *

FinanceReadRowLevel

enabled

normal

Policy Label

Policy Label

Hive Database *

testdb

Hive Table *

general_information

Description

RowLevelPolicies for Finance Data

Audit Logging

YES

Add Validity Period

Row Filter Conditions:

Select Group	Select User	Access Types		
<div>SELECT</div> <div>IN USERS</div>	<div>SELECT User</div>	<div>SELECT</div>	<div>row-level</div>	<div></div>
<div>+</div>				

Enter filter expression

role_code=GENERAL

Save

Cancel

Delete

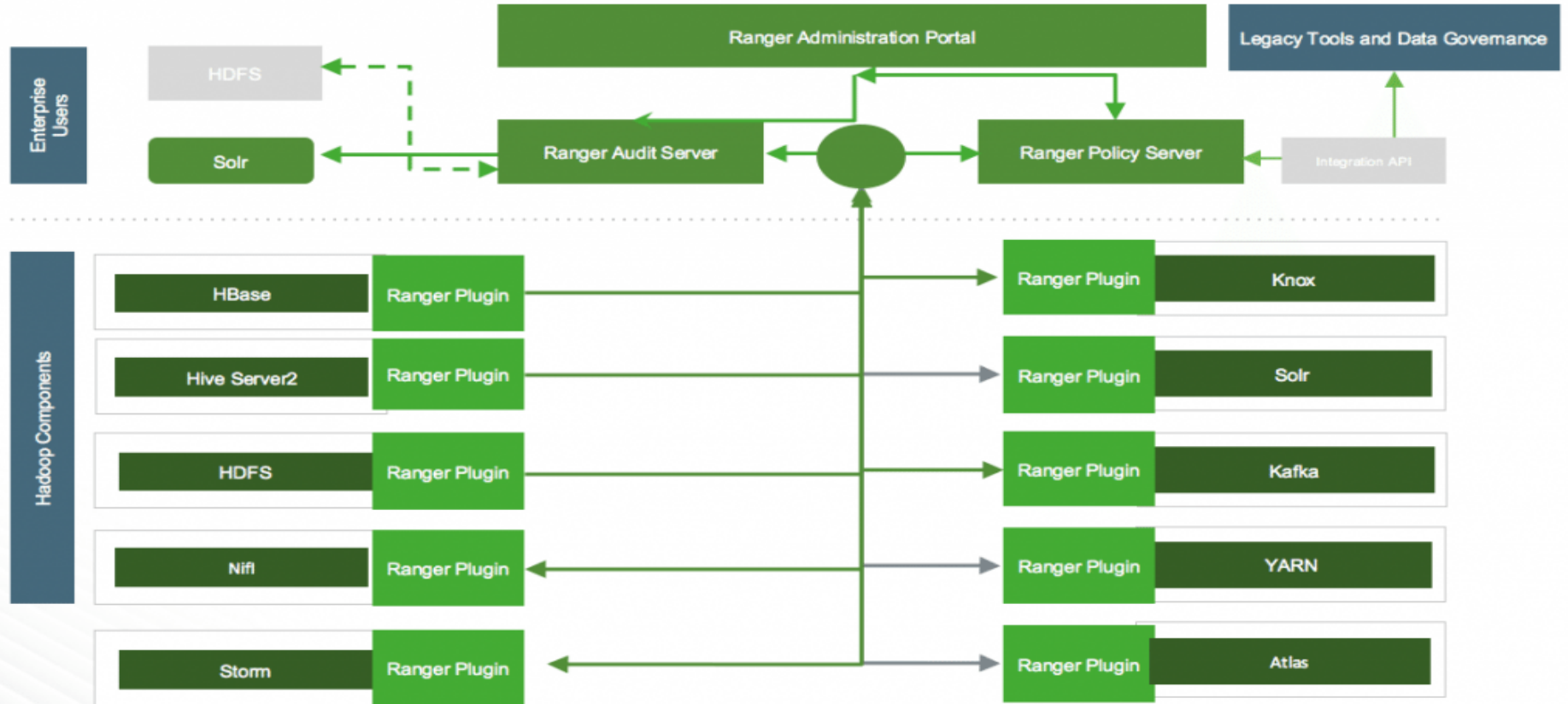
Ranger – Access Audit Logs

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

Ranger Access Manager Audit Settings											
Access Admin Login Sessions Plugins Plugin Status User Sync											
START DATE: 04/22/2019											
Last Updated Time: 04/22/2019 11:08:41 AM											
Policy ID	Event ID	User	Service Name / Type	Resource Name / Type	Access Type	Result	Access Enforcer	Client IP	Cluster Name	Event Count	Tags
---	04/22/2019 11:08:29	udahdpu-ambari-qa	udahdpdev_hadoop_hdfs	/user/udahdpu-ambari-qa/test/delta_0...	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:29	udahdpu-ambari-qa	udahdpdev_hadoop_hdfs	/user/udahdpu-ambari-qa/test/delta_0...	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:29 AM	udahdpu-ambari-qa	udahdpdev_hadoop_hdfs	/user/udahdpu-ambari-qa/test	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:29 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/tmp/hive/udahdpu-hive/553d9f3c-d2...	ALL	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:29 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/warehouse/tablespace/external/hive	READ	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:29 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/tmp/hive/udahdpu-hive/553d9f3c-d2...	WRITE	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:29 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/tmp/hive/udahdpu-hive	WRITE	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:08:28 AM	udahdpu-yarn	udahdpdev_hadoop_hdfs	/ats/active	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:08:25 AM	udahdpu-mapred	udahdpdev_hadoop_hdfs	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:07:28 AM	udahdpu-yarn	udahdpdev_hadoop_hdfs	/ats/active	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:07:25 AM	udahdpu-mapred	udahdpdev_hadoop_hdfs	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:06:44 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/user/udahdpu-hive/yarn/services/lla...	READ	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:06:28 AM	udahdpu-yarn	udahdpdev_hadoop_hdfs	/ats/active	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:06:25 AM	udahdpu-mapred	udahdpdev_hadoop_hdfs	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:06:24 AM	udahdpu-mapred	udahdpdev_hadoop_hdfs	/mr-history/tmp	READ_EXECUTE	Allowed	hadoop-acl	172.18.55.151	udahdpdev	1	---
---	04/22/2019 11:05:29 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/tmp/hive/udahdpu-hive/b0482028-fe...	ALL	Allowed	hadoop-acl	172.18.55.153	udahdpdev	1	---
---	04/22/2019 11:05:29 AM	udahdpu-hive	udahdpdev_hadoop_hdfs	/warehouse/tablespace/external/hive	READ	Allowed	hadoop-acl	172.18.55.153	udahdpdev	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.