Ranger Demo on CDP DataCenter

Sunita Sharma | | Senior Solution Architect

Between January and September 2019 there were over 7.9 billion data records exposed — a 33% increase from the same time in 2018! : by 2020-data-breaches

Let's build secured data pipeline together

Agenda

- CDP (Cloudera Data Platform) Overview
- CDP Data Center with Best of Security And Governance
- CDP Data Center Shared Data Xperience
- Bank Demo with Ranger
- Setting up EMR with Ranger
- Key Takeaways

CLOUDERA DATA PLATFORM Overview

CONTROL

DATA CATALOG

REPLICATION MANAGER

WORKLOAD

MANAGER

MANAGEMENT CONSOLE

HYBRID & MULTI-CLOUD

SECURITY &
GOVERNANCE

ANALYTICS EDGE TO AI

OPEN DISTRIBUTION



Existing CHALLENGES With Current Model w/o SDX

Security & Governance



Sharing data across workloads

- Requires
 multiple copies
 of data need to
 be created
- Each with its own set of data context



Burdensome admin effort

Multiple
clusters =
multiple places
to administer



One missing permission in one copy of the data can lead to significant financial and reputation risk

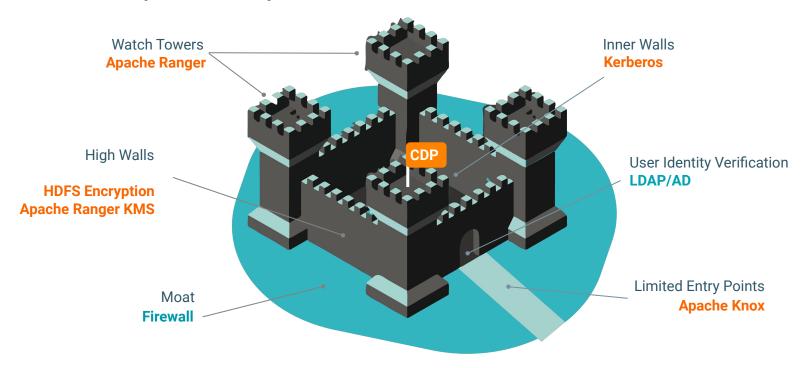


Difficult to share data safely for new analyses



Heavy new regulation such as GDPR makes the challenges even greater

CDP Security Landscape



COMPREHENSIVE APPROACH TO SECURITY



Identity & Perimeter

Validate users in enterprise directory

Technical Concepts:

Authentication User/group mapping

Kerberos, Apache Knox



Access

Defining what users and applications can do with data

Technical Concepts:

Permissions Authorization

Apache Ranger



Visibility

Reporting on where data came from and how it's being used

Technical Concepts:

Auditing Lineage

Apache Atlas



Data Protection

Shielding data in the cluster from unauthorized visibility

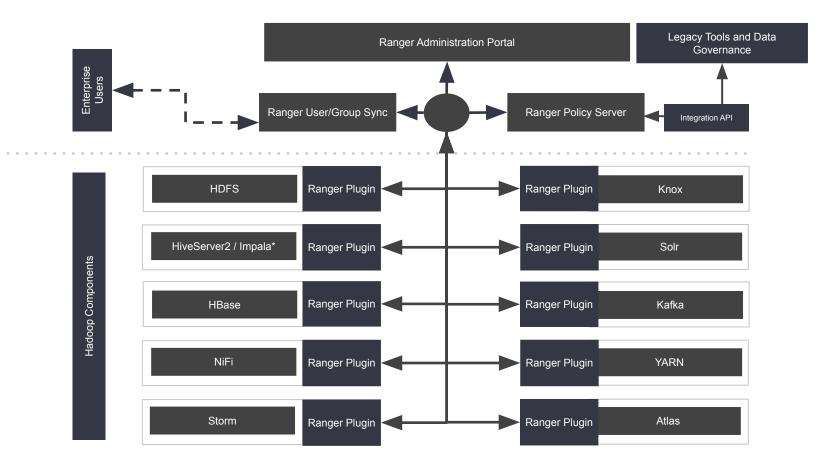
Technical Concepts:

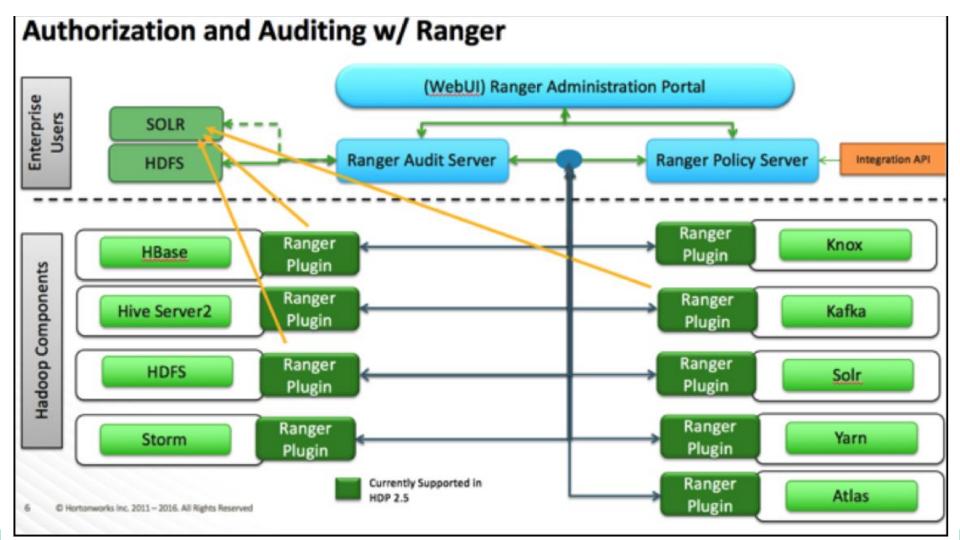
Encryption, Key Management

SSL/TLS, HDFS TDE, Ranger (KMS, Masking, Filtering)

80% of large customers leverage our capabilities across all 4 pillars, to address use cases that include sensitive and regulated data.

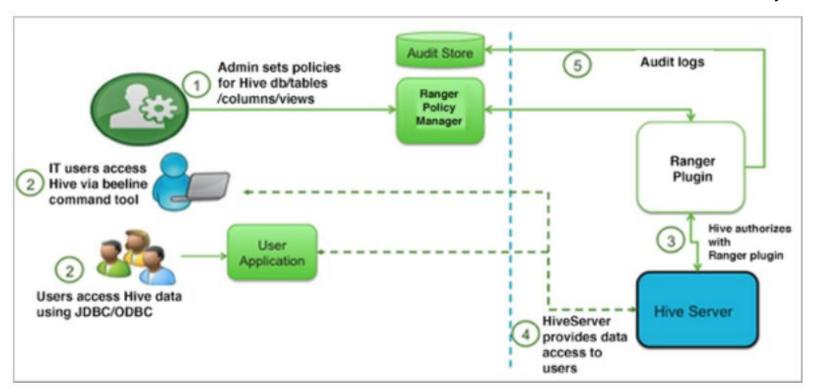
RANGER ARCHITECTURE



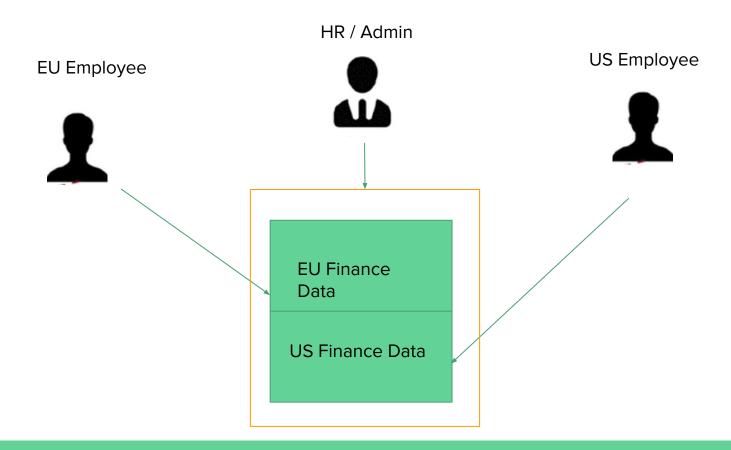


Ranger Authorization for Hive

1 Column level masking2 Row level filtering3 Policy Audits



Bank Demo Scenario

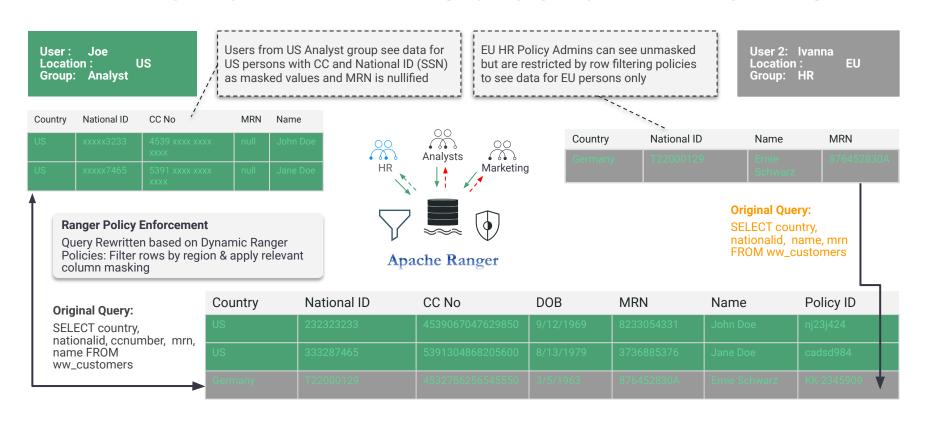


Ranger Set up

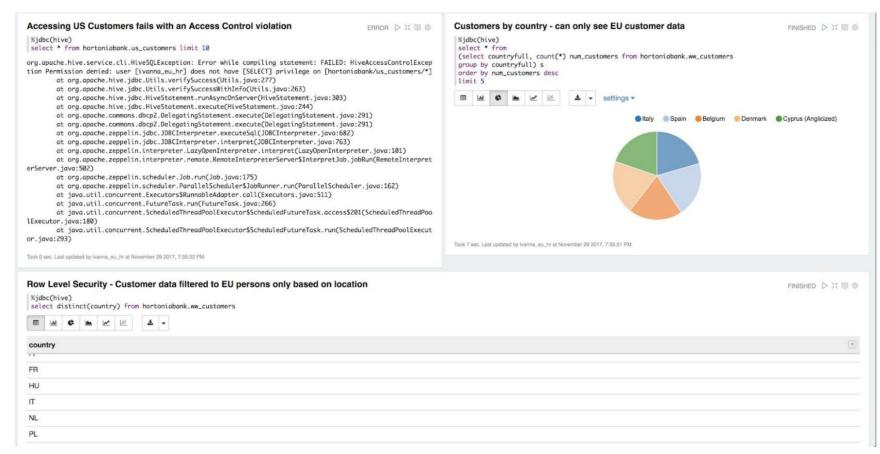
User	Group	Access Privileges
joe_analyst	us_employee	US Data Only, non-sensitive data only, rest masked or forbidden depending on sensitivity
ivanna_eu_hr	eu_employee	EU Data Only (only customers who gave consent), All sensitive data
etl_user	eu_employee	EU Data (all customers), All sensitive data, Update consent/Delete

Data Column Type	Masking	Sample Output
Password	Hash	237672b21819462ff39fcea7d990c3e5
National ID	Last 4 Only	xx-xx-9324
Credit card	First 4 Only	4532xxxxxxxxxxx
Street Address	Static	nnn Xxxxxx Xxxxx
MRN**	NULL	null
Birthdate	Custom	Hide birthday by showing it as 01/01/yyyy
Age	Custom	(Add a random number below 20 to actual age)

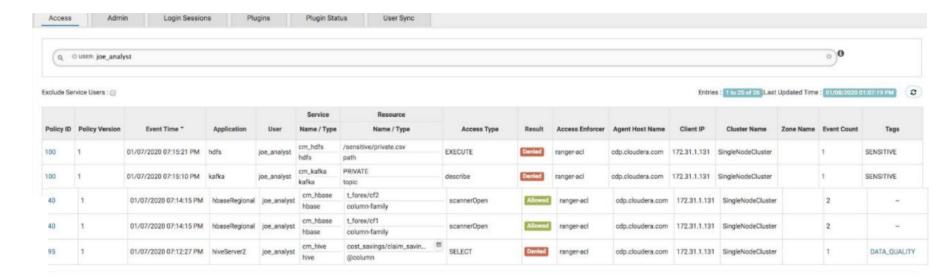
DYNAMIC ROW FILTERING & COLUMN MASKING



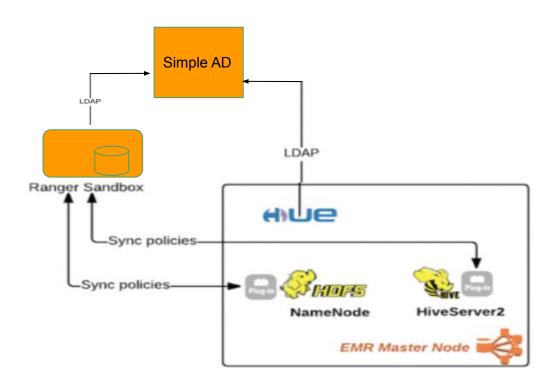
Access for US/EU customer in Notebook



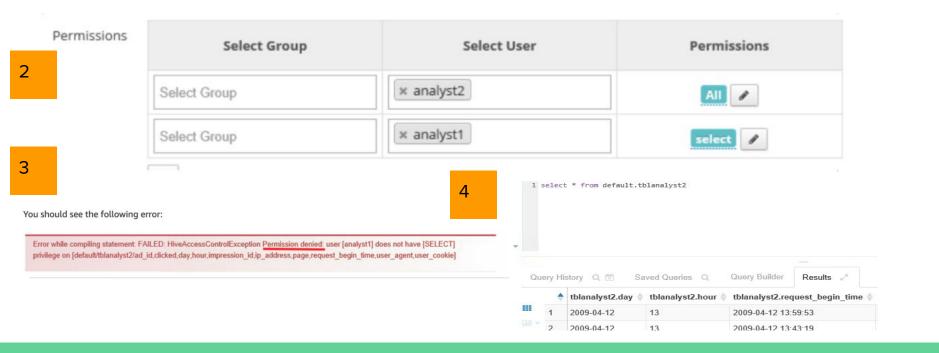
US employee Audit



EMR with Ranger



User	Group	Access Privileges				
Analyst1	analyst	Select Privilege on table tblanalyst1				
Analyst2	admin	Superuser (ALL) on table tblanalyst1 tblanalyst1				



Ranger Audits

	Service					
User	Name / Type	Resource Name	Access Type	Result	Access Enforcer	
	hadoopdev		2510	(Allement)		
analyst1	hdfs	/user/hive/warehouse/tblanalyst	READ	Allowed	hadoop-acl	
analusts	hivedev	dafai il telalanah urt 2 tenda darerin	SELECT	Allowed	ranger ad	
analyst1	hive	hive default/tblanalyst2/code,descrip		Allowed	ranger-acl	

Key Takeaways

CDP brings best of both worlds (HWX and CLDR combined)

Unified Security and Governance using SDX on CDP

Ease of Deployment

Ranger for Data Protection and Audits

Designed to be adopted on Public cloud or Private cloud with ease.

EMR can be set with Ranger Policies to obtain same security audits.

EMR Key Benefits

- Cost Savings (on demand set up)
- AWS Integration (S3,Cloudwatch VPC)
- Deployment (easy deployment)
- Scalability and Flexibility (scale up or down, flexible to use fs HDFS,S3, EMRFS)
- Reliability (auto terminate option)
- Security (IAM Security, VPC, NACL settings)
- Monitoring (Cloudwatch)
- Management Interfaces

Hive on Amazon EMR

Amazon EMR supports Hive Authorization for HDFS but not for EMRFS and Amazon S3.

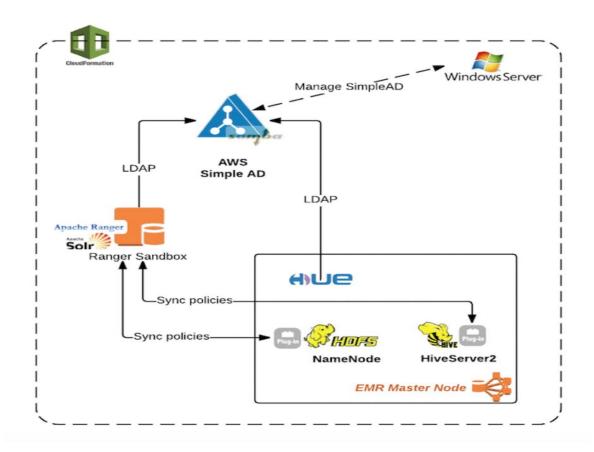
ACID (Atomicity, Consistency, Isolation, Durability) transactions are not supported with Hive data stored in Amazon S3.

Queries that generate a large number of dynamic partitions may fail, and queries that select from tables with many partitions may take longer than expected to execute. For example, a select from 100,000 partitions may take 10 minutes or more.

Amazon EMR 6.0.0 supports the Live Long and Process (LLAP) functionality for Hive. LLAP uses persistent daemons with intelligent in-memory caching to improve query performance compared to the previous default Tez container execution mode.

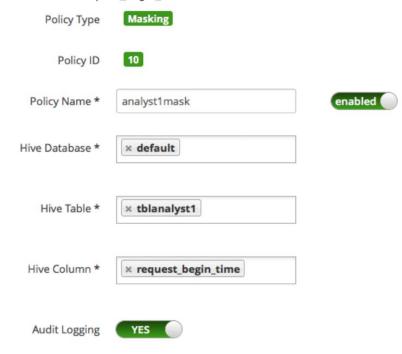
If you need the metastore to persist, you must create an *external metastore* that exists outside the cluster. (with RDS or AWS Glue Data Catalog)

Setting up Ranger on EMR with Windows AD



Column Masking

- · Under the same "hivedev Policies" choose the tab "Masking"
- · Choose "Add New Policy"
- · Create new policy with the following values:
 - Policy Name: analyst1mask
 - Hive Database: default
 - o Hive Table: tblanalyst1
 - Hive Column: request_begin_time



Resources

https://aws.amazon.com/blogs/big-data/using-ldap-via-aws-directory-service-to-access-and-administer-your-hadoop-environment/

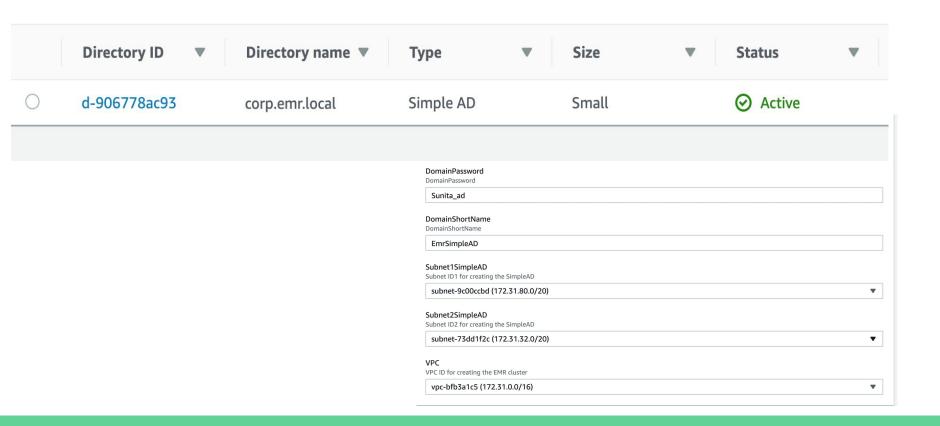
https://aws.amazon.com/blogs/big-data/implementing-authorization-and-auditing-using-apache-ranger-on-amazon-emr/

https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/securing-hive/content/hive_hive_auth_orization_models.html

https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/configuring-ranger-authe-with-unix-ldap-ad/content/adauthentication_settings.html

Back Slides with Use-cases screenshots

Step 1 : Set up Active Directory



Setting up the Ranger server

Now that SimpleAD has been created and the users loaded, you are ready to set up the security policy server (Ranger). This runs on a standard Amazon Linux instance and Ranger is installed and configured on startup.

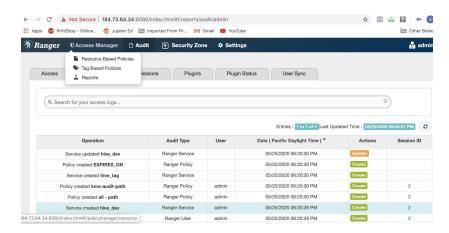
Using the create-ranger-server template, set up this instance. To launch the stack directly through the console, use Launch Stack

. It takes the following parameters:

InstanceType	<instance of="" ranger="" server="" the="" type=""></instance>
KeyName	< EC2 key pair name >
myDirectoryBaseDN	dc=corp,dc=emr,dc=local
myDirectoryBindUser	binduser@corp.emr.local
myDirectoryBindPassword	<user input=""></user>
myDirectoryIPAddress	<one address="" ip="" of="" server="" simplead="" the=""></one>
Subnet	<subnet place="" ranger="" server="" to=""></subnet>
rangerVersion	<version of="" ranger=""> – Choose between 0.6, 0.7 and 1.0 and 2.0</version>
VPC	<vpc></vpc>

CloudFormation output:

IPAddress	<ip address="" of="" ranger="" server="" the=""></ip>
-----------	---





EMR Instance

Summary

Application history

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Connections:

Hue 7, Spark History Server 7, HBase 7, Resource Manager 7 ... (View All)

Master public DNS:

ec2-3-83-137-96.compute-1.amazonaws.com SSH

Tags:

MD = Fremont01, Name = EMR Sample Cluster View All / Edit

Summary

ID: j-O0VESMLDN1T3

Creation date: 2020-05-25 17:46 (UTC-7)

Elapsed time: 1 hour, 26 minutes

After last step Cluster waits

completes:

Termination Off Change

protection:

Configuration details

Release label: emr-5.17.0

Hadoop distribution: Amazon

Applications: Hive 2.3.3, Spark 2.3.1, Hue 4.2.0,

Hbase 1.4.6, Presto 0.206

Log URI: s3://sunita2020/



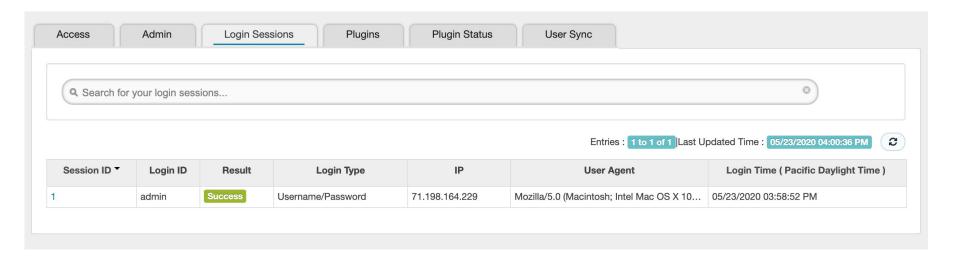
EMRFS consistent Disabled

view:

EMR set up on EC2

Name	aws:elasticm:	Instance ID	Instance Type 🔻	Availability Zone 🔻	Instance State 🔻
EMR Sample Cluster	MASTER	i-02f5de2ead8f1719d	m4.2xlarge	us-east-1c	running
EMR Sample Cluster	CORE	i-047ae5981aa2ebd	m5.2xlarge	us-east-1c	running
RangerServer		i-04d43ab7f7acfd12c	m4.xlarge	us-east-1c	running

Login details Audit from Ranger



	Sync Source	Number Of New		Number Of Modified		
User Name		Users	Groups	Users	Groups	Event Time [▼]
rangerusersync	LDAP/AD	8	0	0	0	05/25/2020 06:04:56 PM

Execute query with Analyst 1

SELECT * FROM default.tblanalyst1

This should return the results as expected. Now, run the following query:

SELECT * FROM default.tblanalyst2

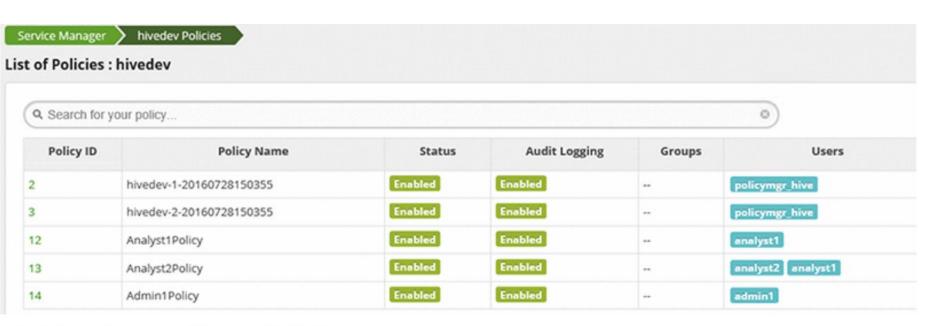
You should see the following error:

Error while compiling statement: FAILED: HiveAccessControlException Permission denied: user [analyst1] does not have [SELECT] privilege on [default/tblanalyst2/ad_id,clicked,day,hour,impression_id,ip_address,page,request_begin_time,user_agent,user_cookie]

Users/ Group in Ranger

User Name	Email Address	Role	User Source	Groups	Visibility
admin		Admin	Internal		Visible
rangerusersync		Admin	Internal		Visible
rangertagsync		Admin	Internal		Visible
awsadmind-9067787ab1		User	External	enterprise admins	Visible
analyst1		User	External	hadoop_analyst	Visible
analyst2		User	External	hadoop_analyst	Visible
administrator		User	External	administrators domain admins group policy creator owners schema admins	Visible
admin1		User	External	hadoop_admin	Visible
binduser		User	External		Visible
krbtgt		User	External	denied rodc password replication group	Visible
guest		User	External	guests	Visible
analyst		User	External		Visible

Ranger policies for User: Analyst

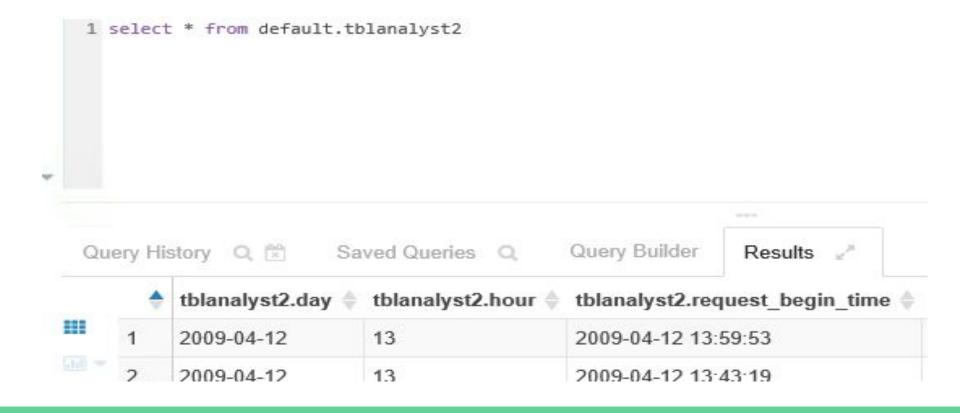


Test Scenario to update the existing policy

Permissions

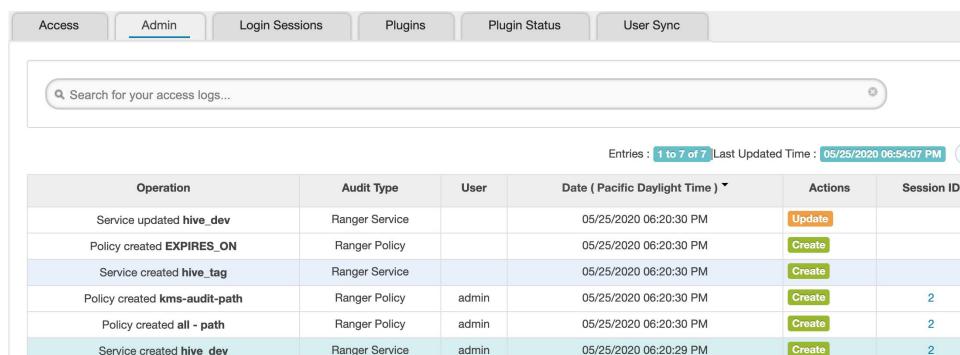


Test Scenario: Retry query with Analyst1



Admin Ranger Policies

User created analyst



admin

Create

05/25/2020 06:20:29 PM

2

Ranger User

Ranger: Users and Permissions

Modules	Groups	Users	Action
Resource Based Policies		admin rangerusersync keyadmin rangertagsync awsadmind-9067787ab1 analyst1 analyst2 administrator admin1 binduser krbtgt guest analyst - Less	
Users/Groups		admin rangerusersync rangertagsync keyadmin	B
Reports		admin rangerusersync keyadmin rangertagsync awsadmind-9067787ab1 analyst1 analyst2 administrator admin1 binduser krbtgt guest analyst - Less	3
Audit		admin rangerusersync rangertagsync keyadmin	B
Key Manager		keyadmin	
Tag Based Policies		admin rangerusersync rangertagsync	B
Security Zone		admin rangerusersync rangertagsync awsadmind-9067787ab1	G.

Scenario Test case :Hadoop user can't move files