SAMSUNG

Samsung Innovation Campus

Khoá học Big Data

Together for Tomorrow!
Enabling People
Education for Future Generations

Chương 9.

An ninh và kiểm soát truy cập

Khoá học Big Data

Mô tả chương

Mục tiêu:

✓ Bảo mật là một lĩnh vực quan trọng đối với CNTT. Vì cụm Hadoop cũng lưu trữ và xử lý rất nhiều dữ liệu nên dữ liệu quan trọng phải được lưu giữ an toàn. Các khái niệm và phương pháp chính để bảo mật Hadoop được giải thích.

- Đầu tiên chúng ta sẽ tìm hiểu khái niệm bảo mật cụm Hadoop và tại sao bảo mật lại quan trọng trong Hadoop.
- Hiểu các khái niệm về quyền HDFS, ACL, kiểm lâm, v.v. kiểm soát truy cập dữ liệu bất hợp pháp và hiểu giao thức mã hóa lưu trữ dữ liệu và bảo mật đường truyền.
- Chúng ta sẽ tìm hiểu kiến thức cơ bản về bảo mật bao gồm các khái niệm về xác thực, ủy quyền và mã hóa.
- Kerberos hỗ trợ sử dụng an toàn bằng cách tích hợp với nhiều hệ sinh thái trong Hadoop bằng các phương thức xác thực mạnh mẽ. Hiểu các thành phần cơ bản của Kerberos và Thuật ngữ chính
- Cuối cùng, cách Kerberos hoạt động cũng như cách ứng dụng khách và dịch vụ Hadoop được xác thực sẽ được giải thích.

Nội dung:

1. Truy cập an toàn vào dịch vụ cụm
2. Truy cập an toàn vào dữ liệu cụm

Bài 1.

Truy cập an toàn vào các dịch vụ cụm

An ninh và Kiểm soát truy cập

Bài 1.

Truy cập an toàn vào các dịch vụ cụm

Bảo mật cụm Hadoop

- Bảo mật cụm Hadoop chủ yếu liên quan đến việc giới hạn những người có thể truy cập dữ liệu được bảo vệ, bằng cách giới hạn những người có thể gửi công việc trên cụm
- Ba yếu tố chính

Xác thực

Ủy quyền

Mã hóa

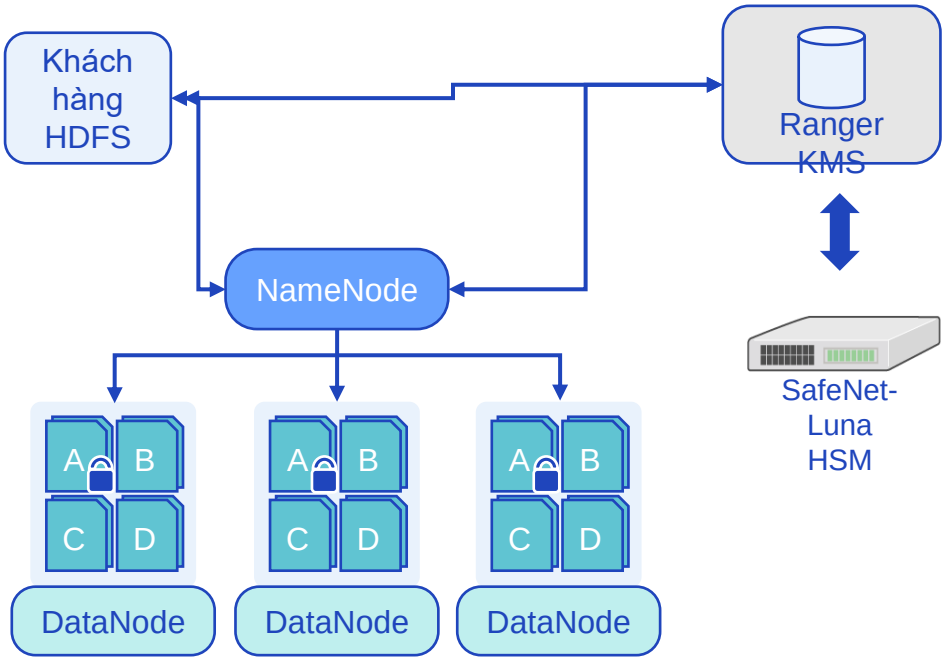
- Kerberos được sử dụng để xác thực và các dịch vụ bảo mật như kiểm lâm được sử dụng để ủy quyền.

Tổng quan cơ bản

- Chu vi
  - Xác thực người dùng mạnh
  - Cách ly mạng, các nút cạnh
  - tường lửa
- Truy cập
  - Kiểm soát ủy quyền
  - Truy cập chi tiết vào các tệp HDFS, đối tượng Hive/Impala với Ranger
- Dữ liệu
  - Mã hóa tại phần còn lại
  - Lưu lượng HTTP được mã hóa (Transport Layer Security -TLS)

Mã hóa dữ liệu minh bạch trong HDFS

- Mã hóa có chọn lọc các tệp/thư mục có liên quan
- Ngăn chặn quyền truy cập của quản trị viên có ý định vào dữ liệu nhạy cảm
- Kiểm soát truy cập chi tiết
- Ứng dụng trực quan xuyên suốt mà không có thay đổi
- Ranger KMS tích hợp với HSM bên ngoài



Các loại xác thực Hadoop

- Các Mục này yêu cầu xác thực mạnh trong Hadoop
  - Người dùng Hadoop
  - Dịch vụ Hadoop
  - Bảng điều khiển dựa trên HTTP
  - Bảo vệ dữ liệu khi đang di chuyển
- Kerberos tăng cường bảo mật
  - Tùy chọn-Cung cấp xác thực mạnh mẽ cho cả máy khách và máy chủ
- Kết thúc các cuộc gọi API Hadoop trong một cái bắt tay SASL
- Ứng dụng có thể được thực hiện với tài khoản riêng của người gửi

quyền HDFS

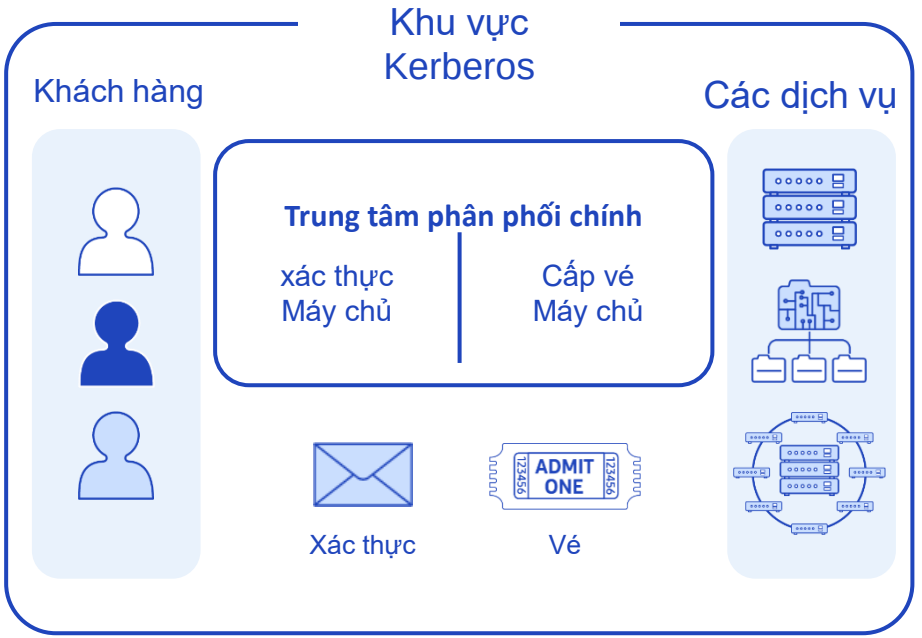
- Quyền HDFS chủ yếu là POSIX
  - Hãy nhớ rằng hdfs là siêu người dùng HDFS, không phải gốc
  - Bit thực thi thư mục là sticky bit
- ACL kiểu POSIX được hỗ trợ
  - Tuy nhiên, nó bị tắt theo mặc định (dfs.namenode.acls.enabled)
  - Bạn có thể thêm người dùng, nhóm và các quyền khác cũng như áp dụng mặt nạ mặc định
  - ACL được sử dụng tốt nhất để điều chỉnh quyền của tệp
- Ủy quyền chi tiết
  - Quyền HDFS và ACL
  - Quyền truy cập tệp cho người dùng-nhóm-người khác có thể quá đơn giản

Xác thực mạnh

- Hadoop có thể sử dụng Kerberos để cung cấp xác thực mạnh hơn cho bảo mật
  - Trình nền Hadoop có thể sử dụng điều này để xác thực tất cả các RPC
- Kerberos
  - Linux hỗ trợ MIT Kerberos nguyên bản
- Kerberos là mô hình xác thực duy nhất mà Hadoop hỗ trợ
  - Móc dịch vụ mã hóa trong quá cảnh có sẵn
  - Xác thực trình duyệt được hỗ trợ bởi HTTP SPNEGO
- Tích hợp LDAP/Active Directory được hỗ trợ
  - Áp dụng cơ sở dữ liệu người dùng hiện có cho cụm Hadoop là một câu hỏi phổ biến

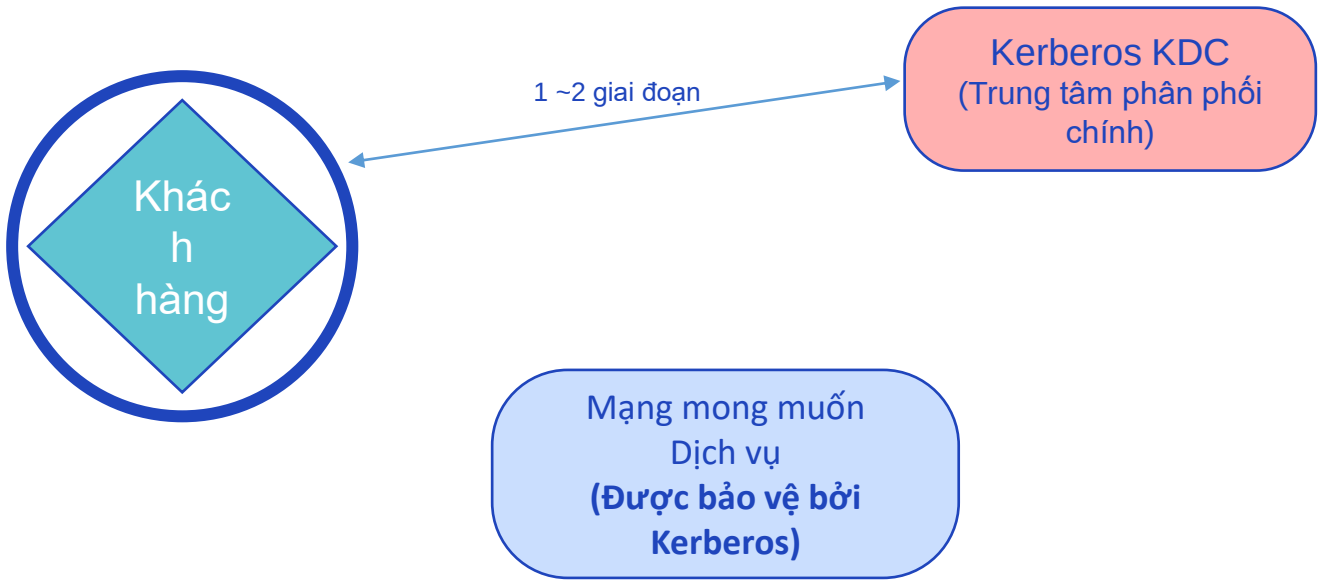
Thuật ngữ Kerberos

- Khu vực
  - Mạng sử dụng Kerberos, bao gồm một hoặc nhiều máy chủ được gọi là KDC và một số lượng lớn máy khách tiềm năng.
- Vé
  - Một bộ thông tin xác thực điện tử tạm thời xác minh danh tính của khách hàng đối với một dịch vụ cụ thể. Còn được gọi là thông tin đăng nhập
- Hiệu trưởng (hoặc tên hiệu trưởng)
  - Tên chính là tên duy nhất của người dùng hoặc dịch vụ được phép xác thực bằng Kerberos.
- keytab (hoặc bàn phím)
  - Một tệp bao gồm danh sách hiệu trưởng không được mã hóa và khóa của họ



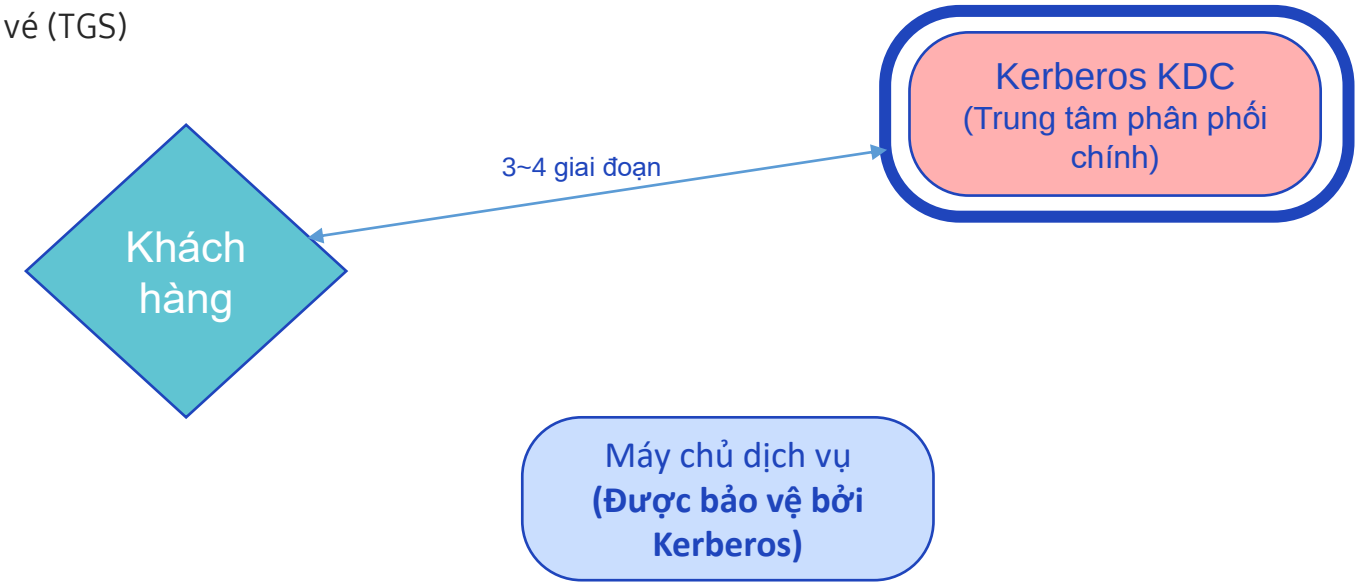
Các thành phần chính trong Kerberos (1/3)

- Kerberos có ba phần: máy khách, máy chủ và bên thứ ba đáng tin cậy (KDC) làm trung gian giữa chúng
- Máy khách là phần mềm muốn truy cập vào dịch vụ Hadoop
  - Lệnh beeline là một ví dụ về ứng dụng khách



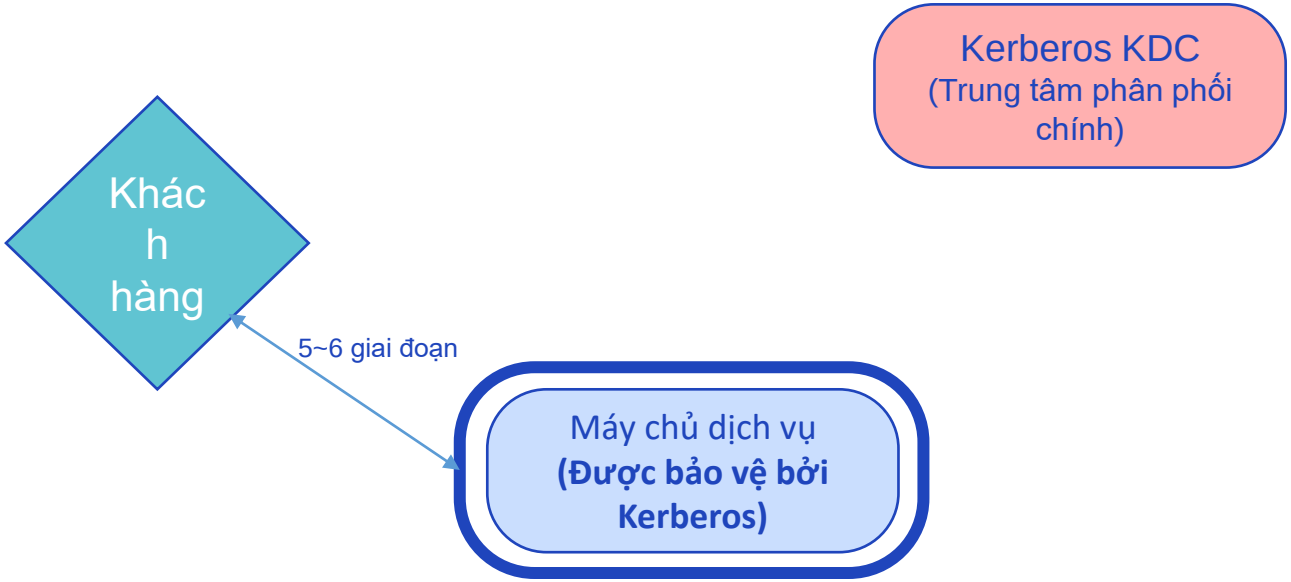
Các thành phần chính trong Kerberos (2/3)

- Máy chủ Kerberos (KDC) xác thực và ủy quyền cho máy khách
- KDC bao gồm hai dịch vụ
  - Dịch vụ xác thực (AS)
  - Dịch vụ cấp vé (TGS)

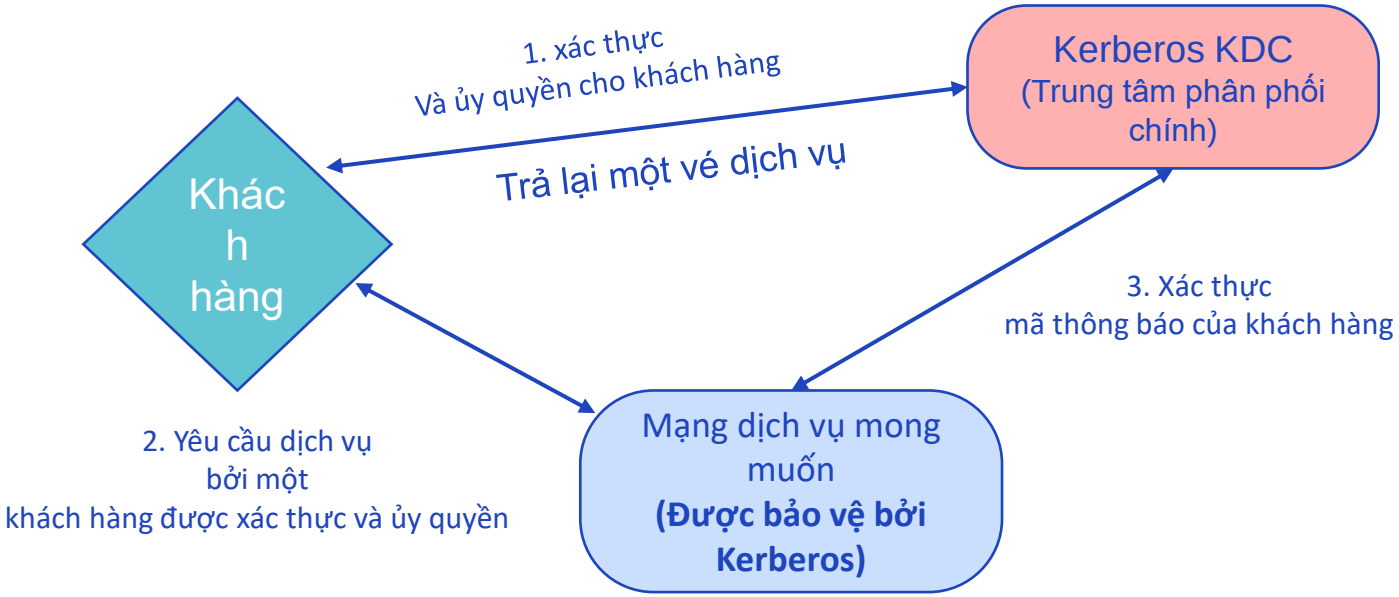


Các thành phần chính trong Kerberos (3/3)

- Đây là dịch vụ Hadoop mà khách hàng muốn truy cập
  - Đây là daemon dịch vụ như Hiveserver 2

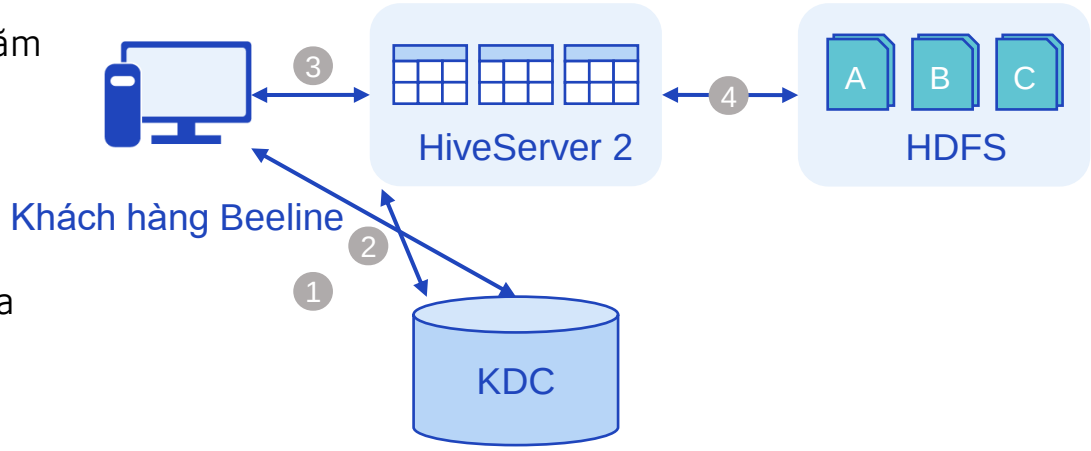


Cách Kerberos hoạt động



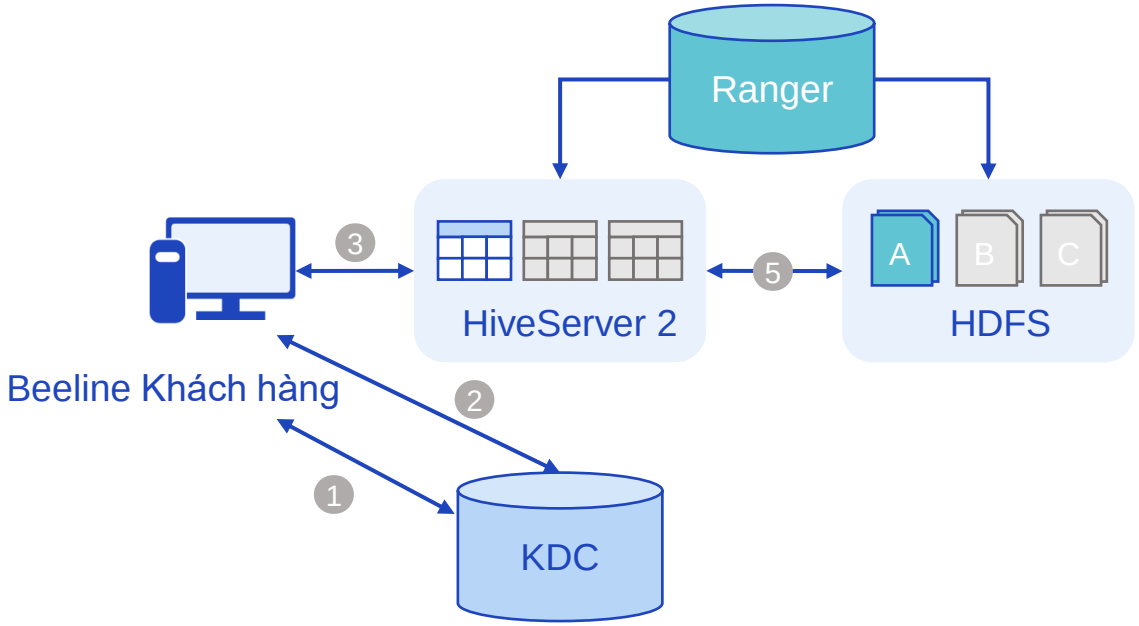
Sử dụng Hive thông qua Beeline Kerberos

- Khách hàng
  - Yêu cầu TGT (Vé cấp vé)
  - Nhận TGT
  - Khách hàng giải mã nó bằng mật khẩu băm
  - Gửi TGT và nhận Vé dịch vụ
  - Gửi truy vấn
- KDC
  - Phát hành vé dịch vụ cho Khách hàng dựa trên vé cấp Vé do máy chủ xác thực cấp
  - Gửi một vé dịch vụ
- HiveServer2
  - Hive nhận vé dịch vụ NN và tạo MapReduce bằng Vé dịch vụ NN



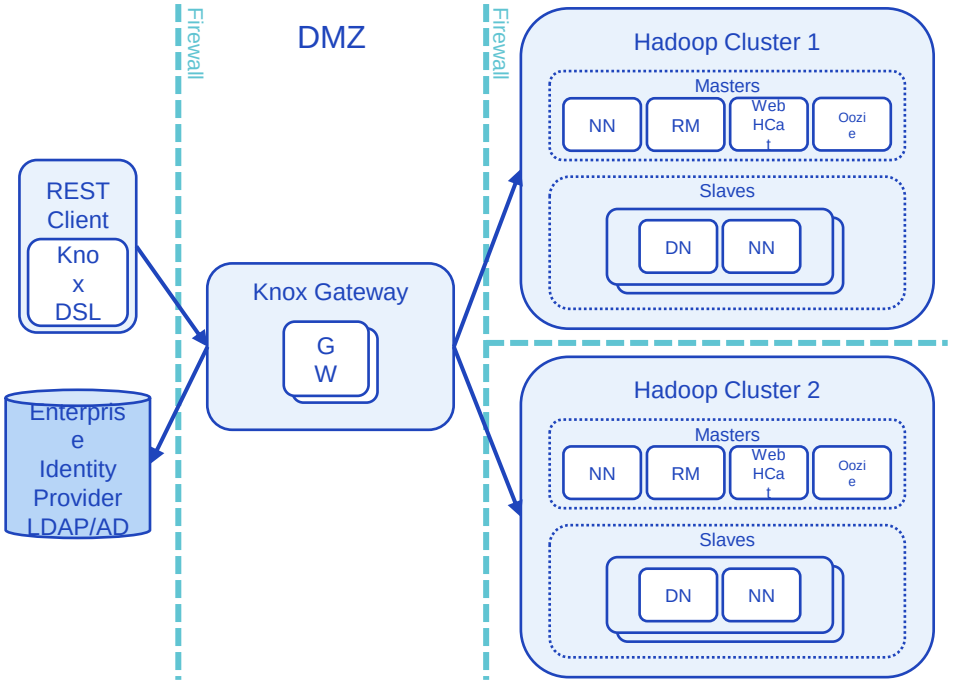
Thêm ủy quyền với Ranger

- Apache Ranger để cung cấp quản lý và quản lý bảo mật tập trung
- Để tạo và cập nhật chính sách trong cơ sở dữ liệu chính sách



Tính năng bảo mật

- Xác thực
  - Hỗ trợ Kerberos
  - Kerberos là một tiêu chuẩn ngành được sử dụng để xác thực người dùng và tài nguyên trong các cụm Hadoop
  - Bảo mật vành đai bởi Apache Knox
- Ủy quyền
  - Kiểm soát truy cập chi tiết - HDFS, HBase và Hive
  - Kiểm soát truy cập dựa trên vai trò
  - Hỗ trợ cấp cột
  - Hỗ trợ quyền – tạo, thả, lập chỉ mục, người dùng
- Mã hóa
  - Đảm bảo chỉ những người dùng được xác minh mới có thể truy cập dữ liệu
  - Hệ thống tệp hệ điều hành, HDFS, hỗ trợ mã hóa cấp độ mạng



Bài 2

Truy cập an toàn vào dữ liệu cụm

Apache Ranger (1/2)

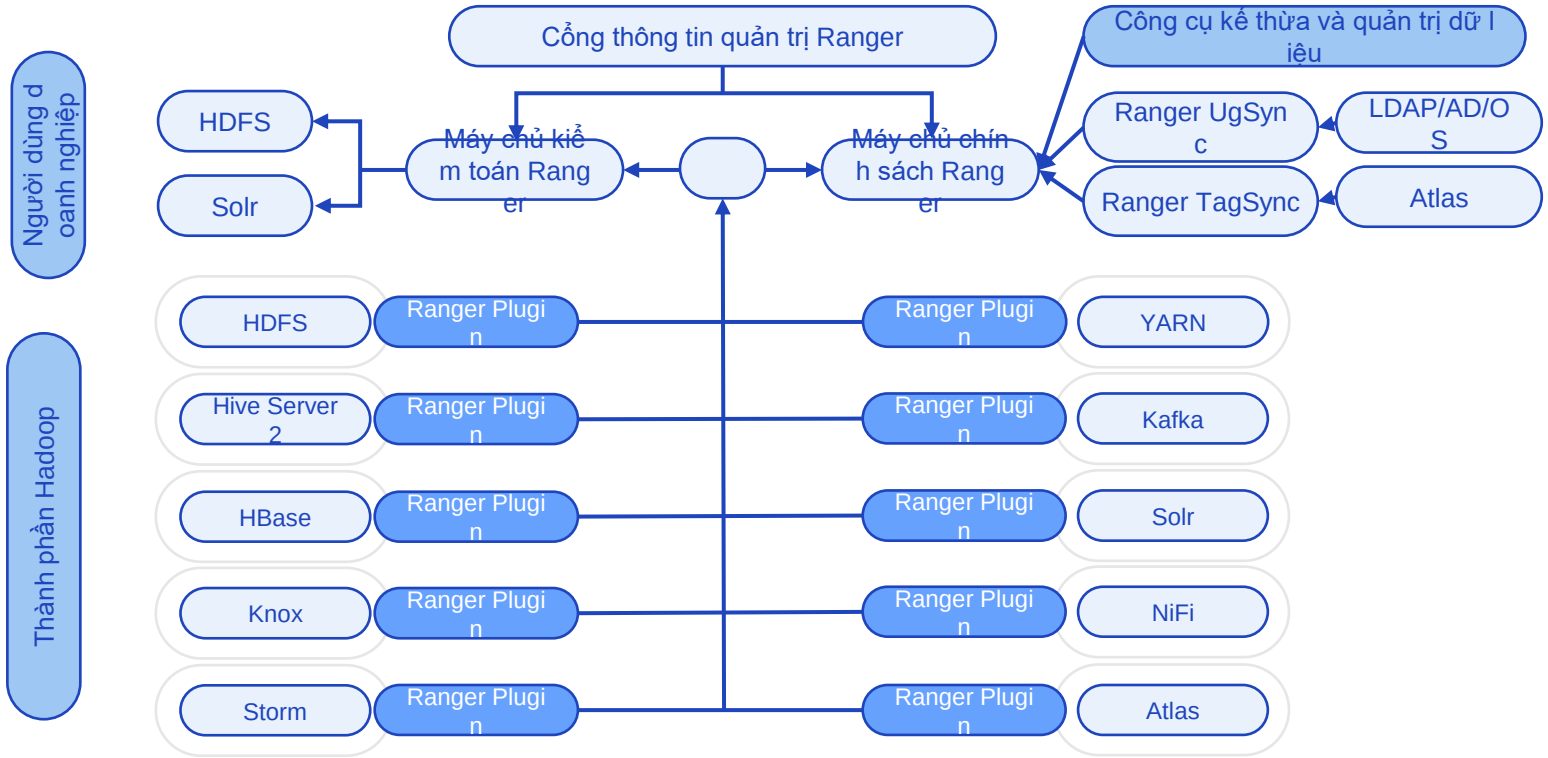
- Apache Ranger là một ứng dụng mã nguồn mở để xác định, quản lý và điều hành các chính sách bảo mật
- Cục An ninh Trung ương
  - Cung cấp một giao diện duy nhất cho các nhà quản lý bảo mật
  - Quản lý chính sách bảo mật tập trung
  - Đảm bảo phạm vi phủ sóng nhất quán trên ngăn xếp Hadoop
- Có thể cắm và có thể dễ dàng mở rộng sang bất kỳ nguồn dữ liệu nào bằng cách sử dụng định nghĩa dựa trên dịch vụ

Apache Ranger (2/2)

Ủy quyền	KMS	Kiểm toán
Nền tảng tập trung để xác định, quản lý và quản lý các chính sách bảo mật một cách nhất quán trên các thành phần Hadoop • HDFS, Hive, HBase, YARN, Kafka, Solr, Storm, Knox, NiFi, Atlas • Kiến trúc mở rộng • Điều kiện chính sách tùy chỉnh, trình làm giàu ngữ cảnh người dùng • Dễ dàng thêm các loại thành phần mới để ủy quyền	• Lưu trữ và quản lý khóa mã hóa • Hỗ trợ mã hóa dữ liệu minh bạch HDFS • Tích hợp với HSM • Safenet LUNA	• Vị trí kiểm tra trung tâm cho tất cả các yêu cầu truy cập • Hỗ trợ nhiều nguồn mục tiêu (HDFS, Solr, v.v.) • Giao diện truy vấn trực quan thời gian thực

Kiến trúc Ranger Apache

- Ủy quyền và Kiểm tra với Ranger

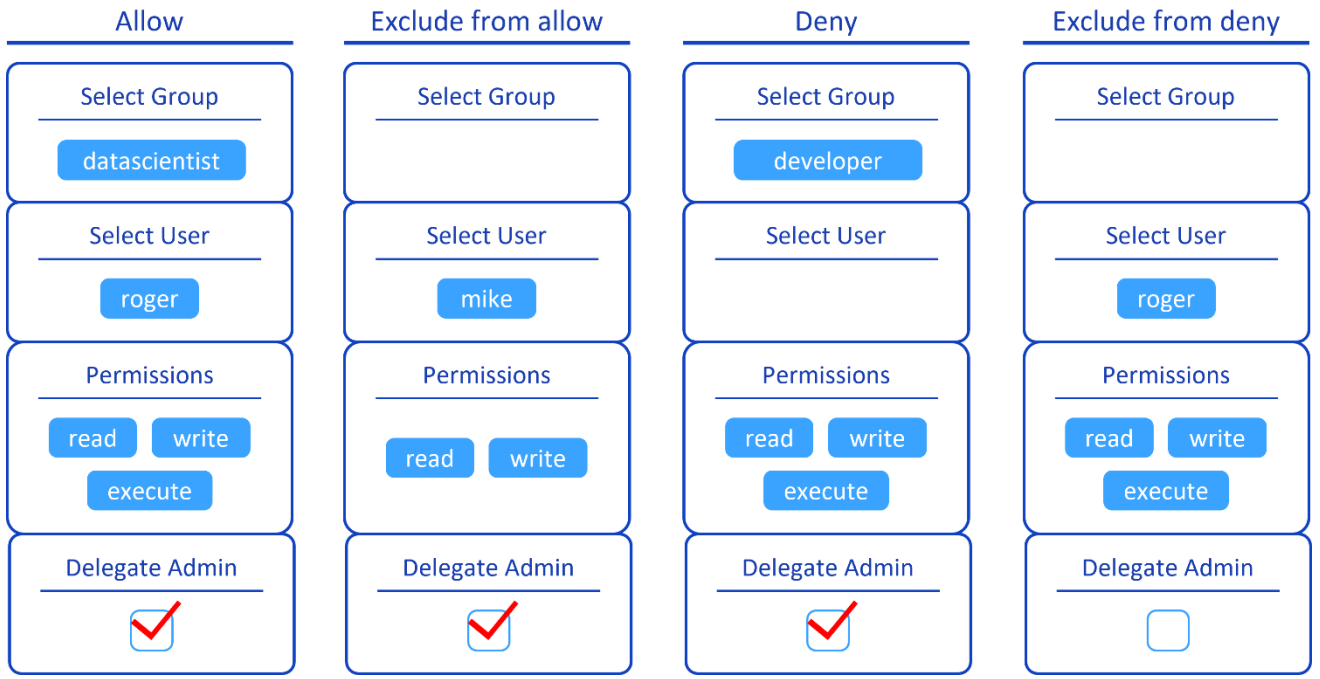


Thành phần Ranger

- Cổng thông tin và máy chủ chính sách Ranger
  - Giao diện trung tâm để quản lý bảo mật
  - Tạo và cập nhật các chính sách được lưu trữ trong cơ sở dữ liệu chính sách
  - Các plugin trong mỗi thành phần thăm dò các chính sách này một cách thường xuyên
- Ranger plugin
  - Dung lượng nhẹ chương trình Java
  - Nhúng vào quy trình của từng thành phần cụm
  - Plugin lấy chính sách từ máy chủ trung tâm và lưu cục bộ vào một tệp
- Đồng bộ hóa nhóm người dùng
  - Tiện ích đồng bộ hóa người dùng để lấy người dùng và nhóm từ Unix, LDAPAD
  - Được lưu trong cổng Ranger và được sử dụng để xác định chính sách

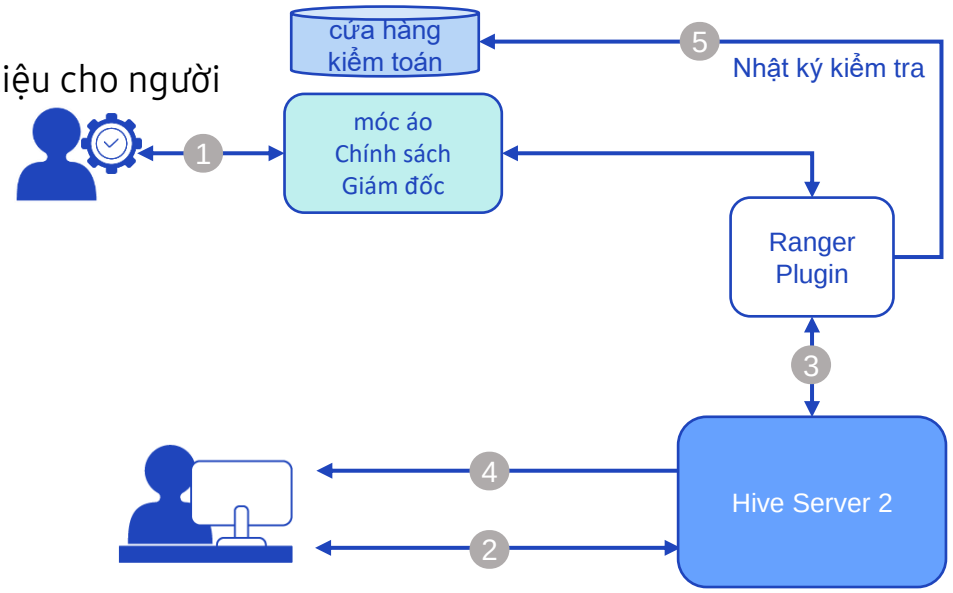
Chính sách Apache Ranger

- Chính sách cung cấp các quy tắc cho phép hoặc từ chối quyền truy cập của người dùng
  - Tất cả các chính sách có thể được áp dụng cho vai trò, nhóm hoặc quy tắc người dùng cụ thể để cho phép hoặc từ chối
- Chính sách dựa trên tài nguyên
  - Xác định ai có thể sử dụng tài nguyên
  - Tạo hoặc chỉnh sửa bằng plugin dịch vụ
- Chính sách dựa trên thẻ
  - Hạn chế quyền truy cập bằng cách sử dụng phân loại và các thuộc tính khác



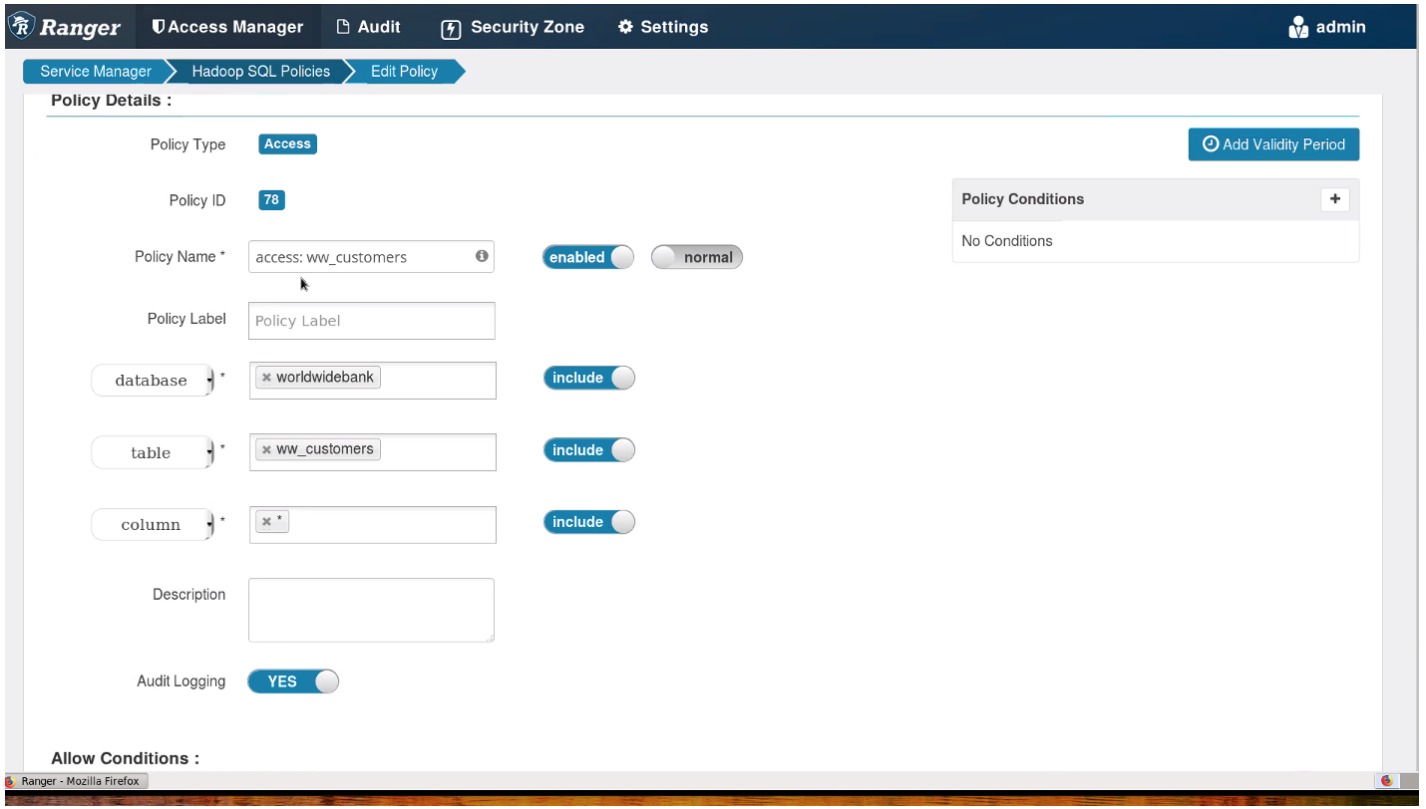
Quy trình làm việc Ranger

- Bước 1: Quản trị viên đặt chính sách cho Hive DB, Bảng, Cột, chế độ xem
- Bước 2: Người dùng truy cập Hive trên công cụ lệnh beeline
- Bước 3: Hive ủy quyền với plugin Ranger
- Bước 4: HiveServer2 cung cấp quyền truy cập dữ liệu cho người dùng
- Bước 5: Để lại nhật ký kiểm tra



Ranger Admin Portal

- Giao diện trung tâm quản trị bảo mật
- Quản trị viên có thể
  - Xác định kho lưu trữ
  - Tạo và cập nhật chính sách
  - Quản lý người dùng/nhóm Ranger
  - Xác định chính sách kiểm toán
  - Xem hoạt động kiểm toán

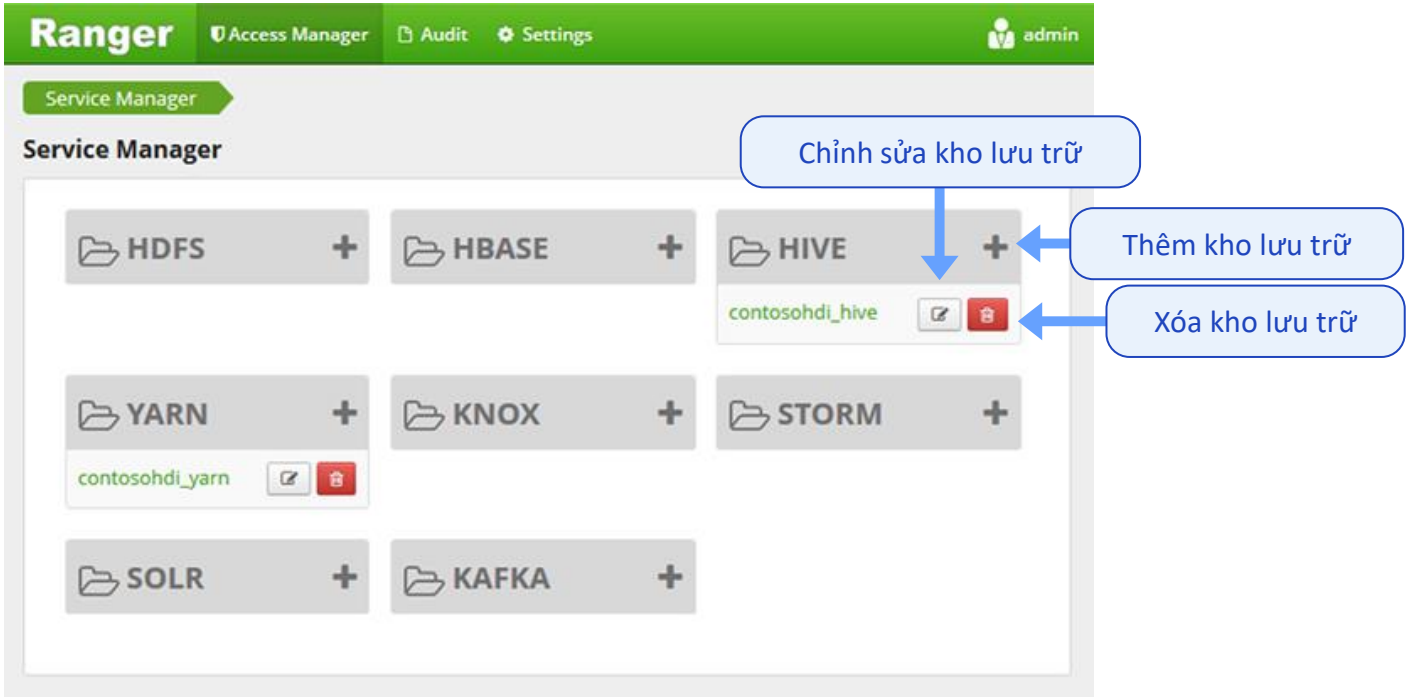


Apache Ranger Plugins

- Ranger plugins
  - HDFS
  - HIVE
  - STORM
  - HBase
- Các bước để kích hoạt plugin
  - Khởi động trình quản lý chính sách
  - Tạo kho lưu trữ plugin trong Trình quản lý chính sách
  - Cài đặt Plugin
  - Khởi động lại dịch vụ plugin

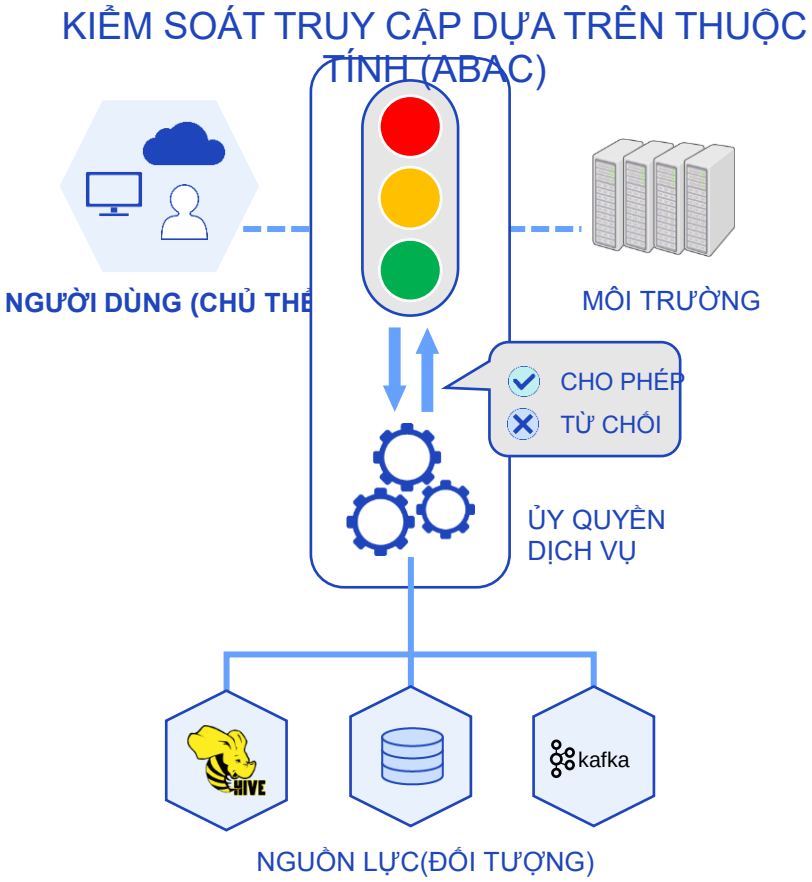
APACHE
STORM

quản lý kho lưu trữ



Ranger – MẪU ABAC

- Sự kết hợp của chủ đề, hành động, tài nguyên và môi trường
- Sử dụng các thuộc tính mô tả: nhóm AD, thẻ hoặc phân loại dựa trên Apache Atlas, vị trí địa lý, v.v.
- Cách tiếp cận của Ranger phù hợp với NIST 800-162
- Tránh phổ biến vai trò và các vấn đề về khả năng quản lý



Apache Atlas là gì?

- Apache Atlas là một nền tảng để quản lý các tiêu chuẩn và dòng dữ liệu.
- Khung quản trị dữ liệu và siêu dữ liệu cho Hadoop
- Chính sách bảo mật dựa trên thẻ động
- Khả năng
  - Danh mục siêu dữ liệu & Tìm kiếm
  - Dòng dõi & Chuỗi hành trình sản phẩm
  - Quản lý/tìm kiếm bằng cách gắn Tag cho bảng/cột
  - Kiểm tra siêu dữ liệu & bảo mật

Tại sao nên sử dụng Apache Atlas? (1/2)

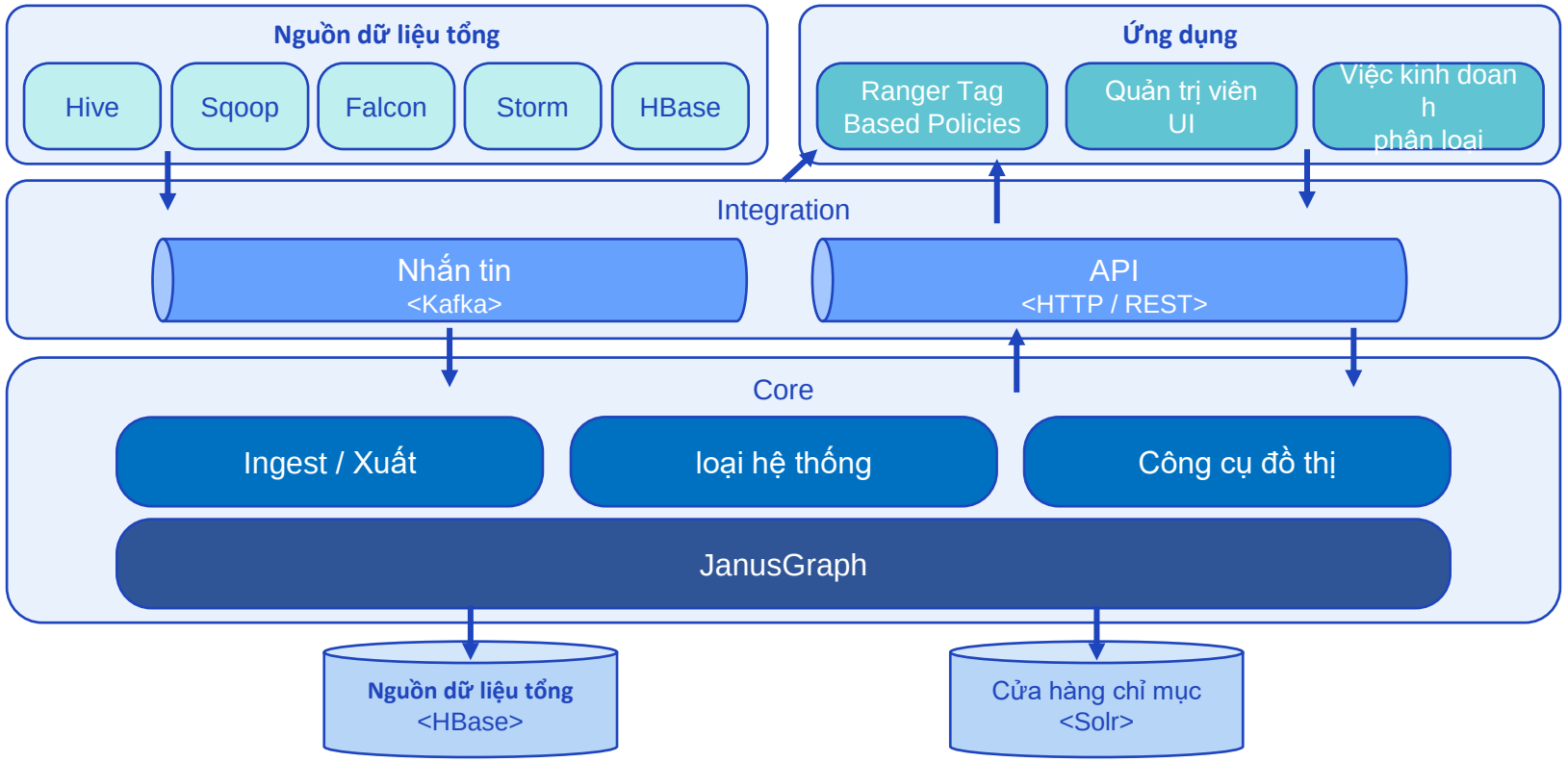
- Các loại siêu dữ liệu và giao diện
  - Nhiều siêu dữ liệu Hadoop và Non-Hadoop có sẵn dưới dạng các loại được xác định trước.
  - Các loại siêu dữ liệu mới có thể được xác định và quản lý.
  - Một loại có thể có các thuộc tính đơn giản, thuộc tính phức tạp hoặc đối tượng tham chiếu.
  - Các thể hiện của các loại được gọi là thực thể thu thập các chi tiết và liên kết của các đối tượng siêu dữ liệu.
- Phân loại
  - Khả năng tự động tạo phân loại - thông tin cá nhân (PII), thông tin hết hạn (EXPIRES_ON), chất lượng dữ liệu (DATA_QUALITY), thông tin nhạy cảm (SENSITIVE)
  - Phân loại bao gồm các thuộc tính - thuộc tính thời hạn sử dụng (expiry date) của thông tin hết hạn (EXPIRES_ON) phân loại
  - Các thực thể có liên quan đến các phân loại khác nhau, dễ dàng tìm kiếm và tăng cường bảo mật.
- truyền thừa
  - Giao diện người dùng trực quan cho phép bạn xem cách dữ liệu được di chuyển và xử lý dưới dạng phả hệ
  - Truy cập và cập nhật phả hệ với API REST

Tại sao nên sử dụng Apache Atlas? (2/2)

- Tìm kiếm và khám phá
  - Giao diện người dùng trực quan cho phép bạn tìm kiếm các đối tượng theo loại, phân loại, giá trị thuộc tính hoặc văn bản miễn phí
  - API REST phong phú cho phép tìm kiếm thậm chí phức tạp
  - Tìm kiếm các đối tượng bằng ngôn ngữ truy vấn giống như SQL - Ngôn ngữ dành riêng cho miền (DSL)
- Bảo mật và che giấu dữ liệu
  - Bảo mật chi tiết của quyền truy cập siêu dữ liệu, cho phép truy cập và kiểm soát các phiên bản đối tượng và các hoạt động phân loại như thêm/sửa đổi/xóa
  - Khi kết hợp với Apache Ranger, việc xử lý ủy quyền/mặt nạ dữ liệu được cung cấp dựa trên quyền truy cập và phân loại dữ liệu được liên kết với các đối tượng trong Apache Atlas.
  - Ví dụ: người dùng có thể truy cập dữ liệu được phân loại là thông tin cá nhân (PII) và thông tin nhạy cảm (SENSITIVE)
- Dịch vụ Meta được hỗ trợ bởi Atlas
  - Hive, HBase, Ranger, Sqoop, Storm/Kafka, v.v.

Kiến trúc bản đồ Apache

- Quy trình làm việc tổng thể trong Apache Atlas



Thành phần Atlas (1/2)

- Lõi
  - Hệ thống loại: Atlas cho phép người dùng xác định các mô hình cho các đối tượng siêu dữ liệu được quản lý
  - Công cụ đồ thị: Cung cấp tính linh hoạt cao và cho phép bạn xử lý hiệu quả các mối quan hệ phong phú giữa các đối tượng siêu dữ liệu
  - Nhập/Xuất: Thành phần ingest cho phép bạn thêm siêu dữ liệu vào Atlas. Thành phần xuất cũng hiển thị các thay đổi siêu dữ liệu do Atlas phát hiện và tăng chúng dưới dạng sự kiện
- Tích hợp
  - API: Tất cả chức năng của Atlas được hiển thị cho người dùng cuối thông qua API REST cho phép tạo, cập nhật và xóa các loại và thực thể
  - Nhắn tin: người dùng có thể chọn tích hợp với Atlas bằng giao diện nhắn tin dựa trên Kafka

Thành phần Atlas (2/2)

- Nguồn dữ liệu tổng
  - Atlas hỗ trợ tích hợp với nhiều nguồn dữ liệu sẵn có
  - Atlas hỗ trợ nhập và quản lý siêu dữ liệu từ các nguồn sau: Hbase, Hive, Sqoop, Storm, Kafka

- Các ứng dụng
  - Atlas Admin UI: một ứng dụng dựa trên web cho phép các nhà quản lý dữ liệu và các nhà khoa học khám phá và chú thích siêu dữ liệu.
  - Chính sách dựa trên thẻ: một giải pháp quản lý bảo mật nâng cao cho hệ sinh thái Hadoop có khả năng tích hợp rộng rãi với nhiều thành phần Hadoop khác nhau

SAMSUNG