VIETTEL SOLUTIONS

VIETTEL DATA PLATFORM CATALOG

Nền tảng dữ liệu lớn Viettel

Nền tảng của chúng tôi cung cấp

- License phần mềm theo năm hoặc trọn đời. Hỗ trợ kỹ thuật 1 năm chính hãng hoặc mua gói hỗ trợ theo năm.
- Có khả năng triển khai on-premise và môi trường đám mây (cloud)
- Hỗ trợ các hệ điều hành phổ biến Windows, Linux, Unix
- Có khả năng cài đặt, kiểm tra hệ thống, môi trường các máy chủ phân tán tự động trên một công cụ toàn trình giao diện web
- Tự động áp dụng các cấu hình nâng cao hiệu năng hệ thống tự động trong quá trình cài đặt.
- Khả năng mở rộng máy chủ không giới hạn

Năng lực xử lý

657 tỉ
Lượt bản ghi xử lý mỗi ngày

73+ tỉ
Bản ghi dữ liệu đồng bộ mới hằng ngày

800.000 record/s
Lưu lượng dữ liệu xử lý trung bình hằng ngày

2.3 triệu record/s
Lưu lượng dữ liệu xử lý tại thời gian cao điểm (peak time)

7 TB
Dữ liệu đồng bộ mới hằng ngày

“TÍNH NĂNG

LƯU TRỮ DỮ LIỆU PHÂN TÁN

- Cung cấp khả năng lưu trữ dữ liệu lớn (> 1000TB) và lưu trữ phân tán trên các máy chủ
- Các loại dữ liệu: cấu trúc, bán cấu trúc, phi cấu trúc.
- Tích hợp với các công nghệ xử lý dữ liệu lớn(Apache Spark, Apache MapReduce, Apache Hive, Apache HBase)
- Loại phần cứng lưu trữ: SSD, HDD
- Chịu lỗi và hỗ trợ tính toán song song thông qua nhân bản dữ liệu, đảm bảo tính sẵn sàng cao của dữ liệu
- Hỗ trợ các thuật toán nén dữ liệu phổ biến: Gzip, Snappy, LZO
- Hỗ trợ phân vùng dữ liệu, phân dữ liệu thành các vùng để tăng tốc độ truy cập dữ liệu
- Truy cập, hỗ trợ các giao thức truy cập dữ liệu phổ biến: HDFS, S3
- Sao lưu và khôi phục dữ liệu tự động, định kỳ theo lịch được lập
- Đảm bảo dữ liệu được bảo vệ an toàn trong quá trình lưu trữ và truyền tải.

Stack công nghệ

hadoop HDFS

APACHE HBASE

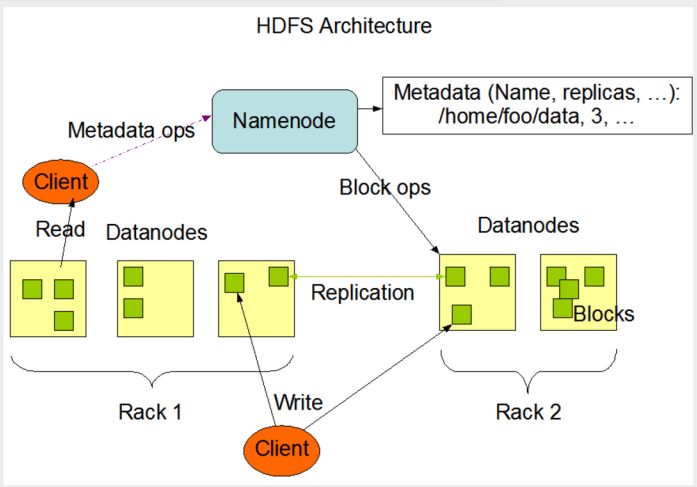
HDFS: hệ thống quản lý tệp phân tán giúp lưu trữ lượng dữ liệu khổng lồ trên các máy chủ khác nhau:

- Kiến trúc Master-Slave:
  - **NameNode**: Quản lý metadata của hệ thống file.
  - **DataNode**: Lưu trữ dữ liệu thực tế dưới dạng block.
  - Cơ chế tự động phân phối dữ liệu giữa các **DataNode**
- Cơ chế HA(High Availability):
  - **Zookeeper Quorum**
  - Replicate dữ liệu trên nhiều **DataNode**
- Mã hóa dữ liệu lưu trữ(Encryption at Rest): HDFS Transparent Encryption
- Mã hóa dữ liệu trên đường truyền(Encryption in Transit): TLS/SSL
- Kiểm soát truy cập bằng **POSIX Permissions, Access Control Lists (ACLs), Apache Ranger**
- Khả năng mở rộng theo chiều dọc; mở rộng theo chiều ngang không giới hạn.

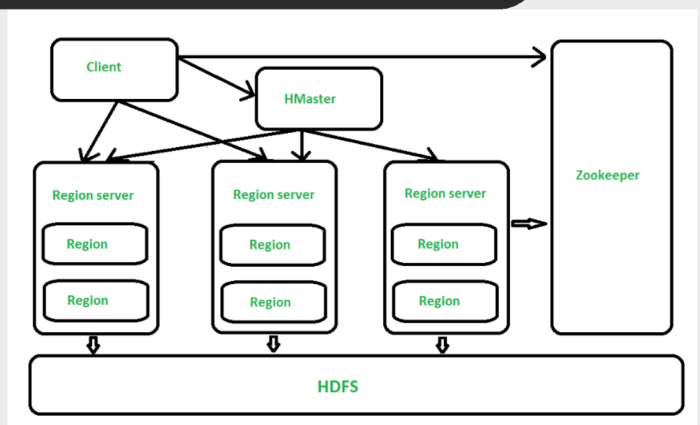
Apache HBase : Cơ sở dữ liệu NoSQL phân tán, column-oriented, xây dựng trên HDFS

- Lưu trữ và truy vấn ngẫu nhiên (random read/write)
- Truy vấn độ trễ thấp

Kiến trúc công nghệ HDFS



Kiến trúc công nghệ HBase

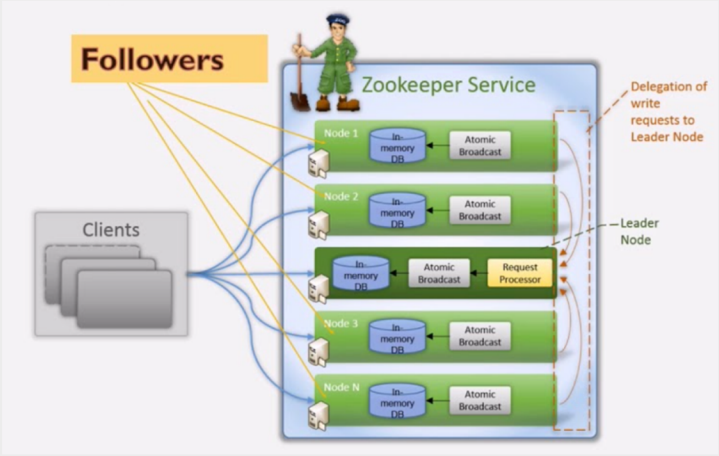


“TÍNH NĂNG

QUẢN LÝ VÀ ĐIỀU PHỐI TÀI NGUYÊN PHÂN TÁN

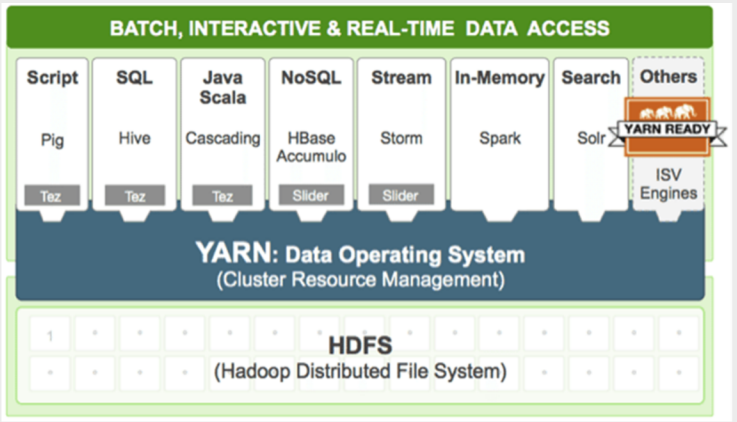
Viettel Data Platform sử dụng Framework YARN làm công cụ quản lý tài nguyên của toàn bộ hệ thống. Ngoài ra Viettel Data Platform sử dụng Zookeeper để cung cấp các cơ chế đồng bộ hóa giữa các máy chủ và dịch vụ trong môi trường phân tán, đảm bảo tính nhất quán và phục hồi lỗi của các thành phần trong nền tảng.

Zookeeper: ZooKeeper là một hệ thống quản lý và đồng bộ hóa phân tán, được sử dụng rộng rãi trong các hệ thống phân tán để quản lý cấu hình, đồng bộ hóa trạng thái và cung cấp dịch vụ điều phối cho các ứng dụng phân tán.



YARN: đảm nhận vai trò quản lý các tài nguyên của hệ thống lưu trữ liệu kết hợp với việc chạy các ứng dụng phân tích.

- Resource Manager: Là thành phần master trong YARN, tiếp nhận các yêu cầu thực thi từ người dùng, điều phối tài nguyên của hệ thống cho các tác vụ thực thi và quản lý, giám sát các ứng dụng phân tán, có khả năng cung cấp:
  - Pre-warmed Containers – Giữ sẵn tài nguyên để job không mất thời gian khởi tạo.
  - Long-running Services – Giữ dịch vụ chạy liên tục để giảm latency.
- Node Manager: Là thành phần thực thi ứng dụng.
- Timeline Server: Dịch vụ chịu trách nhiệm thu thập và lưu trữ thông tin về trạng thái, hiệu suất, và lịch sử của các ứng dụng đang chạy trong cluster. Dịch vụ hỗ trợ người dùng và quản trị viên trong việc phân tích hiệu suất hoặc xử lý sự cố
- History Server: Dịch vụ lưu trữ thông tin về các ứng dụng đã hoàn thành
- DNS Registry: Dịch vụ đóng vai trò như một dịch vụ đăng ký và định vị các thành phần trong cụm, giúp các thành phần như ResourceManager, NodeManager, và Timeline Server có thể định vị nhau một cách linh động



“TÍNH NĂNG

THU THẬP VÀ CHIA SẺ DỮ LIỆU

- Giao diện quản lý truy cập, xác thực người dùng truy cập vào giao diện điều khiển và thiết kế các luồng thu thập, chia sẻ dữ liệu
- Cung cấp hệ thống dashboard tổng hợp thông tin các luồng thu thập, chia sẻ dữ liệu
- Quản lý các luồng đồng bộ dữ liệu (thêm, sửa, xoá)
- Giám sát theo dõi trạng thái của quá trình thu thập dữ liệu
- Thiết kế và quản lý các mẫu luồng đồng bộ dữ liệu để tái sử dụng
- Khả năng tiền xử lý dữ liệu, biến đổi dữ liệu, chuẩn hoá dữ liệu
- Kết nối đa dạng đến các nguồn dữ liệu bên ngoài thông qua các giao thức: API, FTP, STMP, Kafka,...
- Thu thập dữ liệu thời gian thực với độ trễ thấp
- Thu thập dữ liệu theo lô với dung lượng lớn
- Lập lịch tự động thu thập dữ liệu

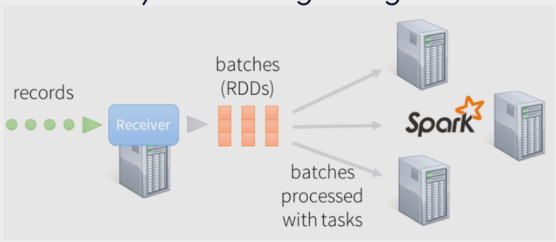
Kafka: Là một hệ thống truyền tin (messaging system) mạnh mẽ và có khả năng mở rộng cao, được thiết kế để xử lý và lưu trữ các luồng dữ liệu trong thời gian thực. Kiến trúc gồm 4 thành phần:

- Broker: Là server Kafka thực thi việc lưu trữ, nhận và phục vụ dữ liệu. Một cụm Kafka gồm nhiều broker, phối hợp với nhau để chia sẻ và nhân bản dữ liệu.
- Zookeeper: Quản lý metadata của cluster (thông tin broker, topic, partition, leader election). Đảm bảo tính nhất quán và chịu lỗi.
- Producer: gửi (publish) các bản ghi (records) vào một hoặc nhiều topic trên Kafka
- Consumer: Đọc (subscribe) và tiêu thụ (consume) các bản ghi từ topic

Nifi : Apache NiFi là một nền tảng tích hợp dữ liệu mạnh mẽ và linh hoạt, được thiết kế để tự động hóa luồng thu thập, chia sẻ dữ liệu giữa các hệ thống khác nhau.

- Web Server
- Flow Controller
- Provenance Repository
- Content Repository
- FlowFile Repository
- NiFi Registry

Spark Streaming : được thiết kế để xử lý luồng dữ liệu theo thời gian thực. Spark Streaming hoạt động dựa trên mô hình micro-batching, trong đó dữ liệu thời gian thực được chia thành các lô nhỏ (micro-batches) với khoảng thời gian cấu hình (batch interval).



CÔNG NGHỆ SỬ DỤNG

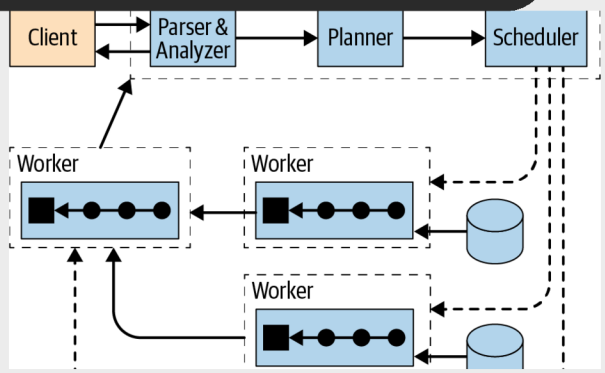
XỬ LÝ DỮ LIỆU PHÂN TÁN

- Cung cấp môi trường làm việc cho các kỹ sư dữ liệu:
  - **Trino:** SQL query engine, Ad-hoc query
  - **Apache Hive & Spark:**
    - Xử lý, tổng hợp và phân tích dữ liệu
    - Lưu trữ, xuất dữ liệu
- Xây dựng, quản lý luồng công việc:
  - **Apache Airflow:** Quản lý job, kiểm tra và cảnh báo lỗi các luồng job
- Cung cấp môi trường làm việc cho nhà khoa học dữ liệu, chuyên viên phân tích dữ liệu:
  - **MLFlow:** Quản lý vòng đời các mô hình Machine Learning
  - **Apache Zeppelin:** web-based notebook
- Xử lý các luồng dữ liệu có tải lên tới 2 triệu message/s, hoạt động 24/7 đảm bảo tính đúng đắn, đầy đủ dữ liệu lên đến 99.9% với KPI sai số không quá 0,01%

Trino: công cụ SQL mạnh mẽ, truy vấn dữ liệu lớn tốc độ near-realtime từ nhiều nguồn khác nhau trong phân tích, BI, và Data Lakehouse.

- Truy vấn dữ liệu phân tán từ nhiều nguồn (Có thể tận dụng các máy chủ xử lý dữ liệu để trả về kết quả truy vấn nhanh chóng):
  - Data Lake: HDFS, S3, ADLS, Data Warehouse: Hive, Iceberg, Delta Lake
  - Database: MySQL, PostgreSQL, Oracle, SQL Server
  - Streaming: Kafka
  - NoSQL: MongoDB, Cassandra
- Federated Query
- Tích hợp dễ dàng với: Superset, Tableau, Power BI
- Hỗ trợ khám phá dữ liệu, tìm kiếm dữ liệu dựa trên mô tả, dựa trên siêu dữ liệu của dữ liệu

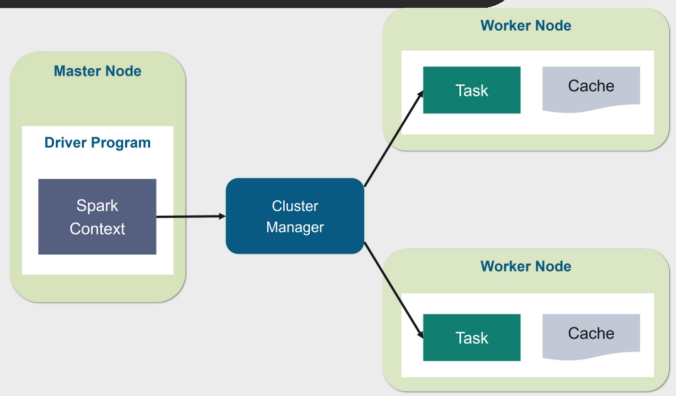
Kiến trúc công nghệ Trino



Apache Spark: công cụ xử lý dữ liệu phân tán In-Memory:

- Kiến trúc Master-Slave:
  - Master (Driver) chịu trách nhiệm phân tích, lập kế hoạch và phân phối công việc.
  - Workers (Các Executors) nằm trên nhiều máy chủ, thực hiện công việc nhận task, thực thi và trả kết quả.
  - Cluster Manager (YARN,..) cấp phát tài nguyên cho cả master và workers.
- Sử dụng DAG (Directed Acyclic Graph) để ghi lại toàn bộ các bước xử lý, khôi phục lại kết quả tính toán khi có lỗi xảy ra
- Tích hợp với các thư viện học máy(Spark MLLib)
- Xử lý theo lô (SparkSQL)
- Xử lý near-realtime(Spark Streaming)
- Hỗ trợ đa ngôn ngữ: Python, Java, Scala, R, SQL

Kiến trúc công nghệ Apache Spark



“TÍNH NĂNG

XỬ LÝ DỮ LIỆU PHÂN TÁN

- Cung cấp môi trường làm việc cho các kỹ sư dữ liệu:
  - **Trino**: SQL query engine, Ad-hoc query
  - **Apache Hive & Spark**:
    - Xử lý, tổng hợp và phân tích dữ liệu
    - Lưu trữ, xuất dữ liệu
- Xây dựng, quản lý luồng công việc:
  - **Apache Airflow**: Quản lý job, kiểm tra và cảnh báo lỗi các luồng job
- Cung cấp môi trường làm việc cho nhà khoa học dữ liệu, chuyên viên phân tích dữ liệu:
  - **MLFlow:** Quản lý vòng đời các mô hình Machine Learning
  - **Apache Zeppelin:** web-based notebook
- Xử lý các luồng dữ liệu có tải lên tới 2 triệu message/s, hoạt động 24/7 đảm bảo tính đúng đắn, đầy đủ dữ liệu lên đến 99.9% với KPI sai số không quá 0,01%

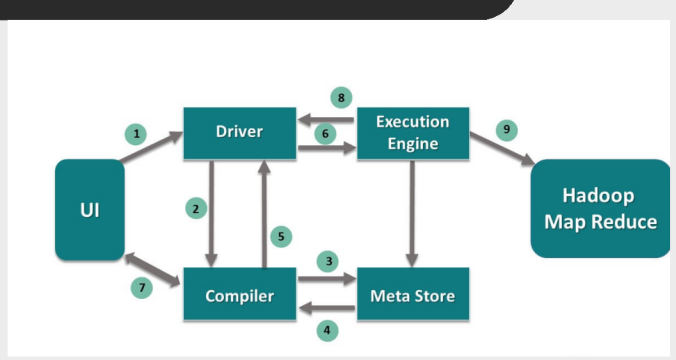
Apache Hive: công cụ xử lý dữ liệu lớn theo lô(batch-data):

- Sử fụng ngôn ngữ truy vấn giống SQL – HiveQL
- Phù hợp với các tập dữ liệu có quy mô terabyte đến petabyte.
- Hỗ trợ phân vùng (Partition) & phân mảnh (Bucketing)

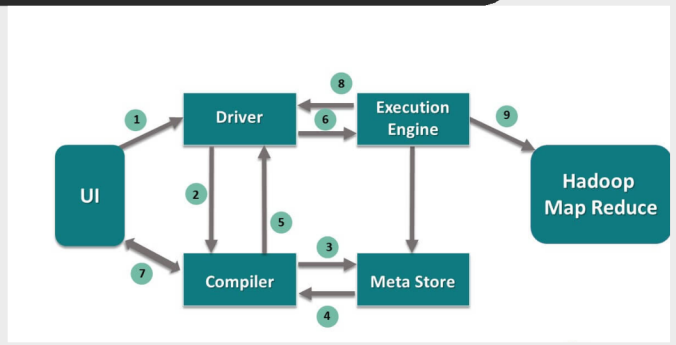
Apache Airflow: công cụ lập lịch, quản lý luồng công việc

- Các thành phần chính:
  - **Scheduler:** Đọc DAG, lên lịch các task, đưa các task cần chạy vào Message Broker
  - **Web Server:** Giao diện UI để quản lý, theo dõi DAG, trigger, xem log, ...
  - **Message Broker:** Hàng đợi (RabbitMQ hoặc Redis) lưu các message đại diện cho task chờ thực thi
  - **Celery Workers:** Nhóm worker chạy song song các tác vụ, lắng nghe broker, lấy task về thực thi, cảnh báo bất thường và gửi kết quả về backend
  - **Result Backend:** Lưu trạng thái và kết quả của task (Redis, database như MySQL/Postgres, hoặc S3...)
- Khả năng xây dựng luồng xử lý dữ liệu phức tạp, gồm nhiều tác vụ thực hiện tuần tự hoặc song song thông qua ngôn ngữ lập trình: Python, Java, Scala
- Tính năng retry_on_failure, on_failure_callback(gửi email, cảnh báo, ghi log,..)
- Có khả năng truyền biến thời gian khi chạy lại DAG

Kiến trúc công nghệ Apache Hive



Kiến trúc công nghệ Apache Airflow



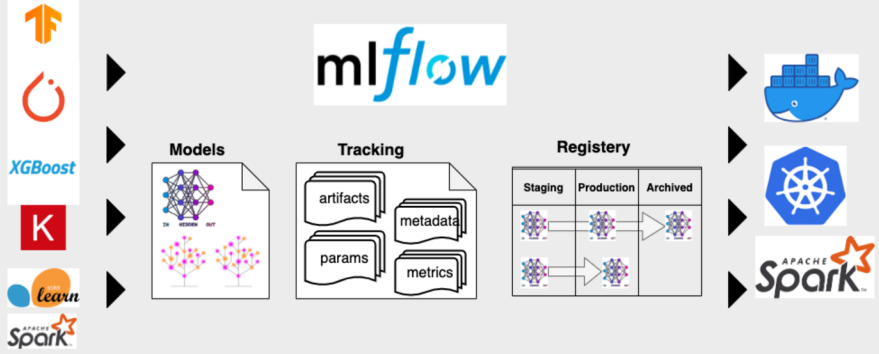
“TÍNH NĂNG

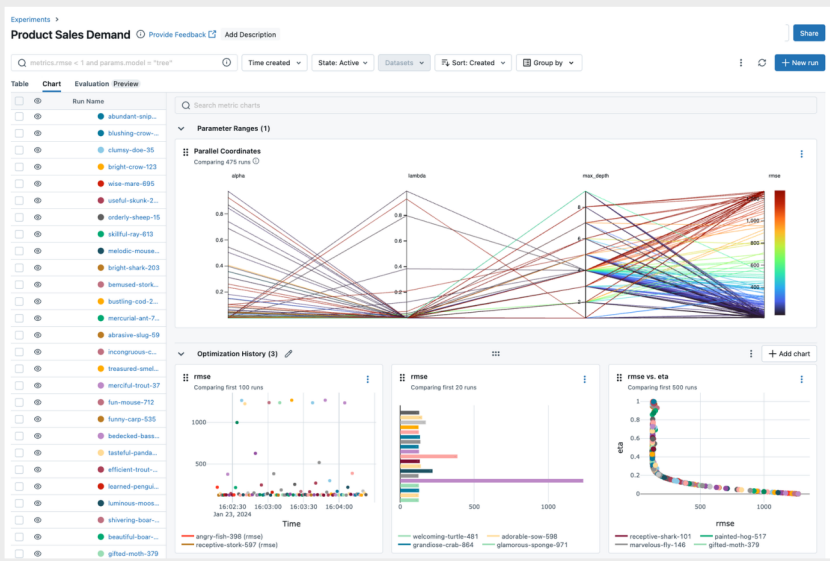
XỬ LÝ DỮ LIỆU PHÂN TÁN

- Cung cấp môi trường làm việc cho các kỹ sư dữ liệu:
  - **Trino**: SQL query engine, Ad-hoc query
  - **Apache Hive & Spark**:
    - Xử lý, tổng hợp và phân tích dữ liệu
    - Lưu trữ, xuất dữ liệu
- Xây dựng, quản lý luồng công việc:
  - **Apache Airflow**: Quản lý job, kiểm tra và cảnh báo lỗi các luồng job
- Cung cấp môi trường làm việc cho nhà khoa học dữ liệu, chuyên viên phân tích dữ liệu:
  - **MLFlow**: Quản lý vòng đời các mô hình Machine Learning
  - **Apache Zeppelin:** web-based notebook
- Xử lý các luồng dữ liệu có tải lên tới 2 triệu message/s, hoạt động 24/7 đảm bảo tính đúng đắn, đầy đủ dữ liệu lên đến 99.9% với KPI sai số không quá 0,01%

MLFlow: công cụ quản lý toàn bộ vòng đời của các mô hình Machine Learning

- **MLflow Tracking**: Ghi lại và theo dõi các thí nghiệm (experiments), bao gồm:
  - Parameters (tham số hyperparameter)
  - Metrics (độ chính xác, độ lỗi...)
  - Artifacts (mô hình đã huấn luyện, đồ thị, log...)
- **MLflow Models**
  - Quản lý và đóng gói mô hình đã huấn luyện theo nhiều định dạng (Flavors) như:
    - Python-function
    - TensorFlow
    - Scikit-learn
    - PyTorch
    - ONNX
  - Triển khai mô hình đã được đào tạo trên một máy chủ hoặc cluster và cung cấp API để gửi dữ liệu đầu vào và nhận kết quả từ mô hình học máy
- **MLflow Projects**
  - Đóng gói code, môi trường và dependencies thành project (MLproject file).
  - Cho phép chạy lại experiments
- **MLflow Model Registry**
  - Lưu trữ trung tâm cho các mô hình đã đóng gói.
  - Hỗ trợ versioning, staging, production, và archiving.





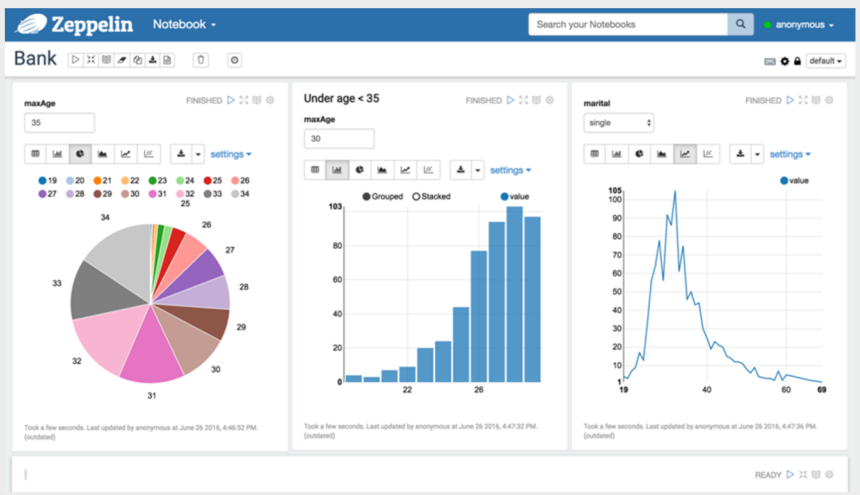
“TÍNH NĂNG

XỬ LÝ DỮ LIỆU PHÂN TÁN

- Cung cấp môi trường làm việc cho các kỹ sư dữ liệu:
  - **Trino**: SQL query engine, Ad-hoc query
  - **Apache Hive & Spark**:
    - Xử lý, tổng hợp và phân tích dữ liệu
    - Lưu trữ, xuất dữ liệu
- Xây dựng, quản lý luồng công việc:
  - **Apache Airflow**: Quản lý job, kiểm tra và cảnh báo lỗi các luồng job
- Cung cấp môi trường làm việc cho nhà khoa học dữ liệu, chuyên viên phân tích dữ liệu:
  - **MLFlow**: Quản lý vòng đời các mô hình Machine Learning
  - **Apache Zeppelin:** web-based notebook
- Xử lý các luồng dữ liệu có tải lên tới 2 triệu message/s, hoạt động 24/7 đảm bảo tính đúng đắn, đầy đủ dữ liệu lên đến 99.9% với KPI sai số không quá 0,01%

Apache Zeppelin: cung cấp các trình thông dịch đa ngôn ngữ, hỗ trợ nhiều người cùng làm việc trên cùng một báo cáo hay phân tích:

- Khả năng tích hợp với nhiều công cụ học máy thông qua trình thông dịch, cho phép:
  - Lựa chọn thuật toán học máy (MLlib, TensorFlow, Scikit-learn, H2O.ai)
  - Tiền xử lý dữ liệu (Spark, Hive, NiFi)
  - Huấn luyện mô hình (Spark MLlib, trên YARN)
  - Tinh chỉnh mô hình (Hyperparameter tuning, Grid Search)
  - Đánh giá mô hình (Cross-validation, Metrics)
- Hỗ trợ xuất báo cáo, biểu đồ
- Hỗ trợ làm việc nhóm, người dùng có thể:
  - Cùng chỉnh sửa notebook.
  - Gắn comment, markdown.
  - Trình bày kết quả dễ hiểu thông qua biểu đồ.



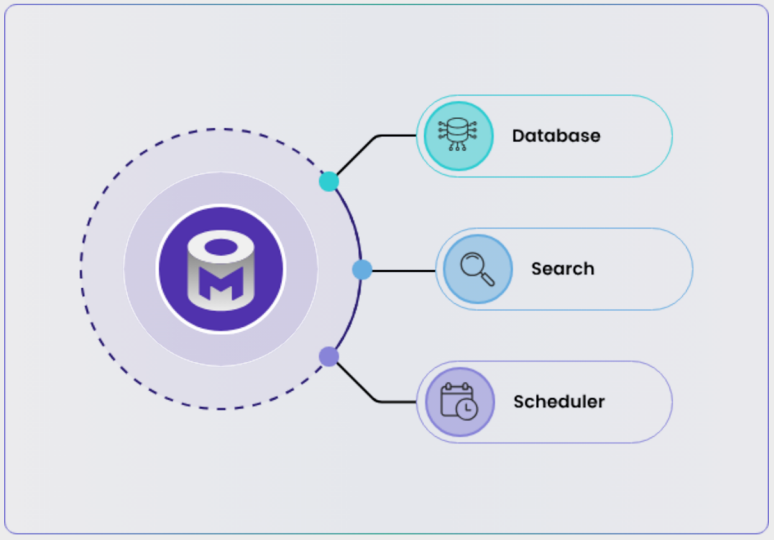
“TÍNH NĂNG

QUẢN TRỊ SIÊU DỮ LIỆU VÀ CHẤT LƯỢNG DỮ LIỆU

- Khả năng thu thập thông tin về các siêu dữ liệu gắn với dữ liệu, định dạng dữ liệu, tính đầy đủ của dữ liệu, tần suất cập nhật của dữ liệu
- Phân tích chất lượng dữ liệu của dữ liệu nguồn: tính toàn vẹn dữ liệu, thiếu dữ liệu, trùng lặp dữ liệu
- Nắm bắt siêu dữ liệu về quá trình xử lý dữ liệu để kiểm soát và tra cứu nguồn gốc dữ liệu
- Khả năng lưu dấu vết các hoạt động thêm, xoá, sửa dữ liệu do người dùng thực hiện

OpenMetadata: nền tảng quản lý metadata mạnh mẽ, bao gồm các thành phần

- **Metadata Ingestion:** Kết nối & thu thập metadata tự động từ nhiều nguồn: databases (MySQL, Postgres...), data lake (S3, HDFS), data warehouse (Snowflake, BigQuery...), BI tools (Tableau, PowerBI), pipelines (Airflow, dbt)...
- **Metadata Store:** Cơ sở dữ liệu trung tâm lưu trữ metadata (schemas, bảng, cột, dashboards, pipelines...)
- **Data Catalog:** Giao diện web cho phép người dùng tìm kiếm, duyệt, và xem chi tiết metadata
- **Lineage:** Hiển thị đồ thị lineage (dòng chảy dữ liệu) giữa các đối tượng (tables → jobs → dashboards)
- **Data Quality & Profiling:**Tích hợp kiểm tra chất lượng dữ liệu, profiling (số lượng null, phân phối giá trị...)
- **Glossary & Business Metadata:** Xây dựng business glossary, gắn nhãn (tag), định nghĩa business terms cho metadata
- **Access Control & Security:** Phân quyền chi tiết trên metadata, tích hợp với LDAP/SSO/Kerberos
- **APIs & SDKs:** Cung cấp REST API và client SDK (Python, Java) để tự động hóa và tích hợp sâu với hệ thống
- **Notifications & Lineage Alerts:** Cấu hình cảnh báo khi metadata thay đổi hoặc vi phạm quality rules



“TÍNH NĂNG

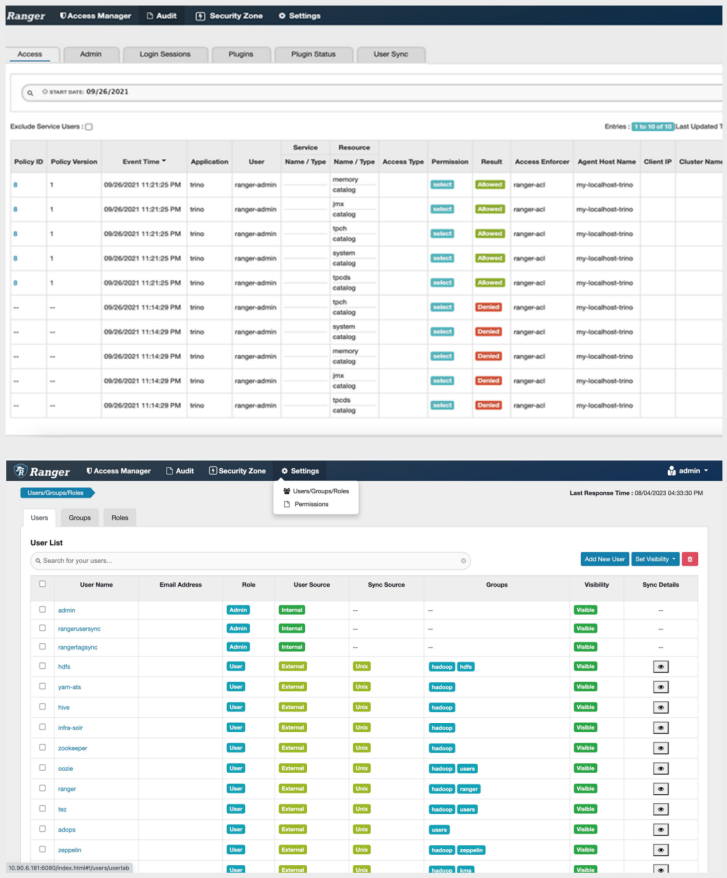
BẢO MẬT

- **Các dịch vụ được sử dụng trong Viettel Data Platform bao gồm : Ranger, Kerberos, Knox**

Ranger: là một dịch vụ quản lý an ninh và kiểm soát truy cập trong các hệ thống dữ liệu lớn. Ranger cung cấp khả năng quản lý quyền chi tiết và chính sách bảo mật tập trung cho nhiều dịch vụ dữ liệu như Hadoop, Hive, Kafka, HBase, và nhiều dịch vụ khác. Các tính năng của Ranger:

Các tính năng của Ranger:

- **Centralized Authorization:** Quản lý các chính sách truy cập (access control policies) từ một giao diện tập trung, thay vì cấu hình phân tán ở từng service.
- **Fine-Grained Access Control:** Kiểm soát truy cập chi tiết theo cấp độ: database, table, column, row, topic (Kafka), file path (HDFS)...
- **Role-Based Access Control (RBAC):** Phân quyền theo vai trò người dùng (Role), giúp dễ dàng gán nhiều quyền cho nhóm người dùng cùng loại.
- Xác thực người dùng qua LDAP hoặc Active Directory
- **Attribute-Based Access Control (ABAC):** Cho phép tạo rule truy cập dựa trên thuộc tính của người dùng (user attributes).
- **Policy Conditions:** Chính sách có thể điều kiện theo IP, thời gian, nhóm người dùng, v.v.
- **Data masking:** Ẩn dữ liệu nhạy cảm (data masking) và lọc dòng dữ liệu dựa theo người dùng.
- **Audit Logging:** Ghi log chi tiết tất cả hành động truy cập.
- Giao diện quản lý trực quan

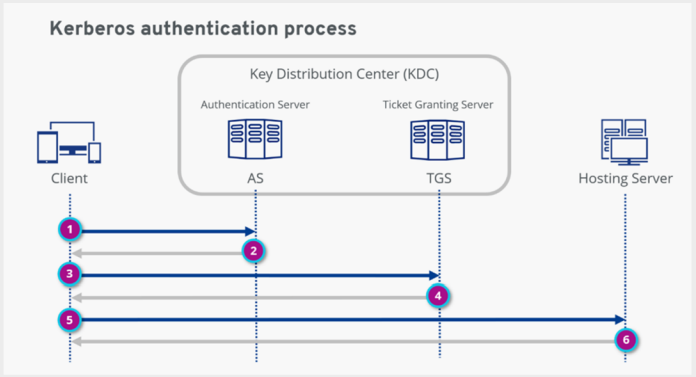


“TÍNH NĂNG

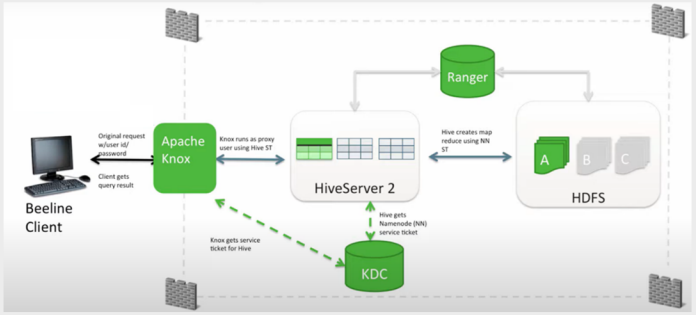
BẢO MẬT

- **Các dịch vụ được sử dụng trong Viettel Data Platform bao gồm : Ranger, Kerberos, Knox**

Kerberos: Kerberos là một giao thức xác thực qua mạng được phát triển tại Viện Công nghệ Massachusetts (MIT) và hiện là một phần quan trọng trong các hệ thống bảo mật mạng. Kerberos sử dụng cơ chế mã hóa đối xứng và một hệ thống dựa trên ticket để đảm bảo rằng danh tính của người dùng hoặc dịch vụ được xác thực trước khi truy cập tài nguyên. Điểm nổi bật của Kerberos là khả năng bảo vệ chống lại các cuộc tấn công nghe lén và giả mạo, nhờ việc hạn chế truyền mật khẩu qua mạng



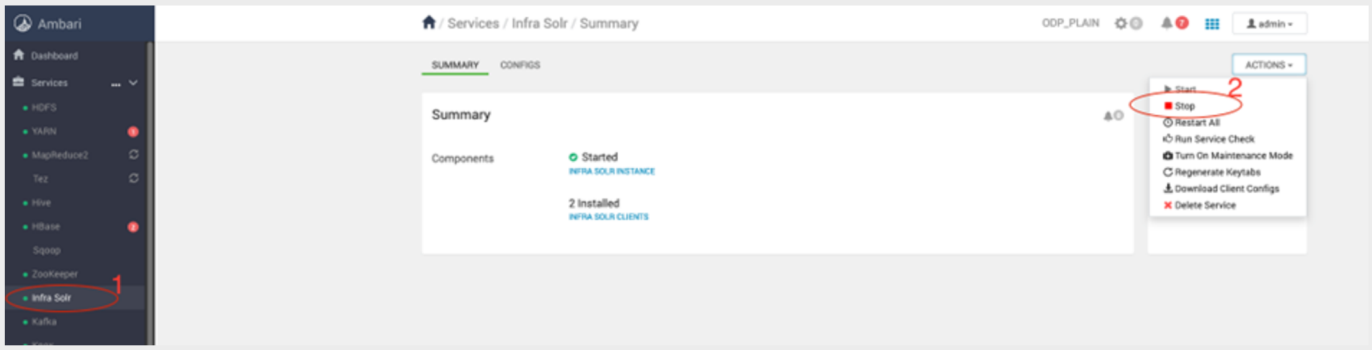
Knox: Apache Knox là một cổng dịch vụ (gateway) được thiết kế để bảo mật và đơn giản hóa truy cập vào các dịch vụ dữ liệu lớn trong Viettel Data Platform. Knox cung cấp một điểm truy cập duy nhất, giúp bảo vệ các API của hệ sinh thái Viettel Data Platform thông qua các cơ chế xác thực, ủy quyền, và ghi nhật ký tập trung. Việc tích hợp với Kerberos giúp Knox đảm bảo rằng chỉ các yêu cầu hợp lệ mới được phép truy cập vào tài nguyên



“TÍNH NĂNG

QUẢN TRỊ HỆ THỐNG

- Cung cấp khả năng tương tác start, stop với các thành phần ứng dụng.

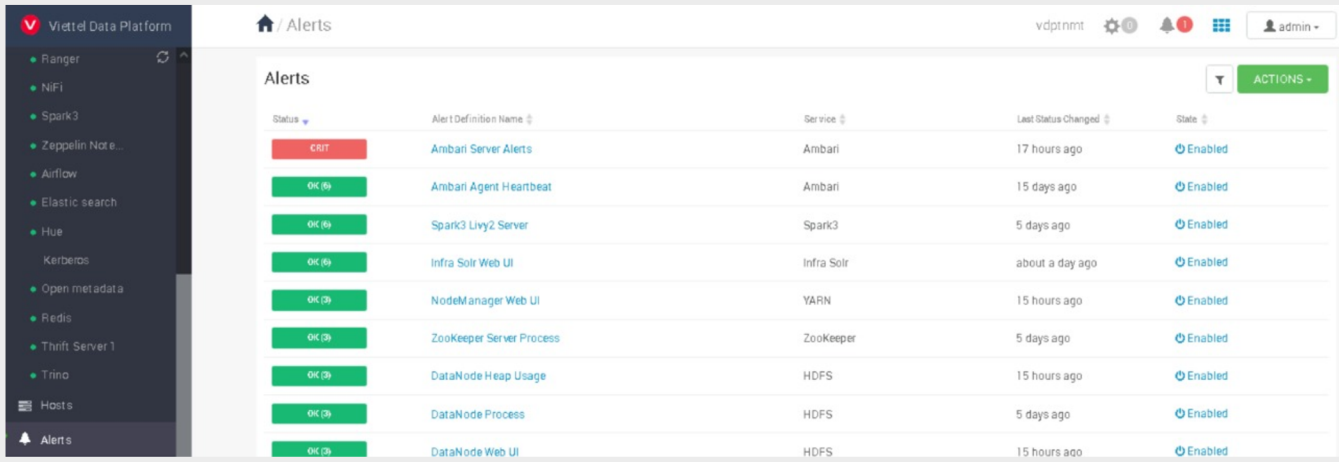


- Quản trị log ứng dụng trong quá trình vận hành hệ thống, trực quan hóa dữ liệu giám sát qua Ambari Log Search

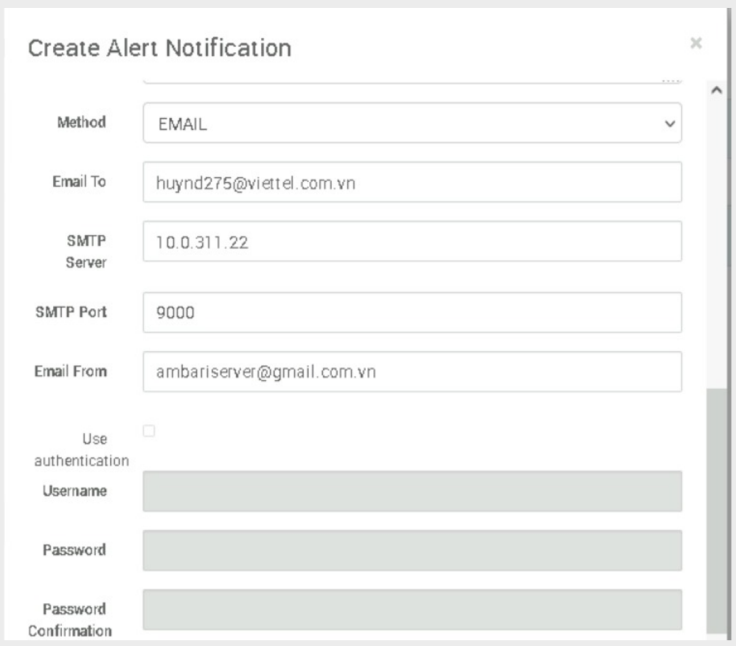
“TÍNH NĂNG

QUẢN TRỊ HỆ THỐNG

- Khả năng tùy biến cảnh báo hệ thống



- Hỗ trợ các kênh thông báo thông tin cảnh báo: SMS, Email



- Khả năng tích hợp với các phần mềm trong nền tảng kho dữ liệu
- Khả năng thu thập và lưu trữ thông tin hoạt động của hệ thống

“TÍNH NĂNG

QUẢN TRỊ HỆ THỐNG

- Cung cấp khả năng phân quyền cho người vận hành, nhóm người vận hành theo vai trò, mức độ xử lý công việc.

Permissions	Cluster User	Service Operator	Service Administrator	Cluster Operator	Cluster Administrator	Ambari Administrator
View metrics	✅	✅	✅	✅	✅	✅
View status information	✅	✅	✅	✅	✅	✅
View configurations	✅	✅	✅	✅	✅	✅
Compare configurations	✅	✅	✅	✅	✅	✅
View service alerts	✅	✅	✅	✅	✅	✅
Start, stop, or restart service		✅	✅	✅	✅	✅
Decommission or recommission		✅	✅	✅	✅	✅
Run service checks		✅	✅	✅	✅	✅
Turn maintenance mode on or off		✅	✅	✅	✅	✅
Perform service-specific tasks		✅	✅	✅	✅	✅
Modify configurations			✅	✅	✅	✅
Manage configuration groups			✅	✅	✅	✅
Move to another host			✅	✅	✅	✅
Enable HA			✅	✅	✅	✅
Enable or disable service alerts			✅	✅	✅	✅
Add service to cluster					✅	✅

Danh sách quyền tương tác với service theo vai trò

Permissions	Cluster User	Service Operator	Service Administrator	Cluster Operator	Cluster Administrator	Ambari Administrator
View metrics	✅	✅	✅	✅	✅	✅
View status information	✅	✅	✅	✅	✅	✅
View configuration	✅	✅	✅	✅	✅	✅
Turn maintenance mode on or off				✅	✅	✅
Install components				✅	✅	✅
Add or delete hosts				✅	✅	✅

Danh sách quyền tương tác với các host trong cụm theo vai trò

Permissions	Cluster User	Service Operator	Service Administrator	Cluster Operator	Cluster Administrator	Ambari Administrator
View metrics	✅	✅	✅	✅	✅	✅
View status information	✅	✅	✅	✅	✅	✅
View configuration	✅	✅	✅	✅	✅	✅
View stack version details	✅	✅	✅	✅	✅	✅
View alerts	✅	✅	✅	✅	✅	✅
Enable or disable alerts					✅	✅
Enable or disable Kerberos					✅	✅
Upgrade or downgrade stack					✅	✅

Danh sách quyền tương tác với service theo vai trò

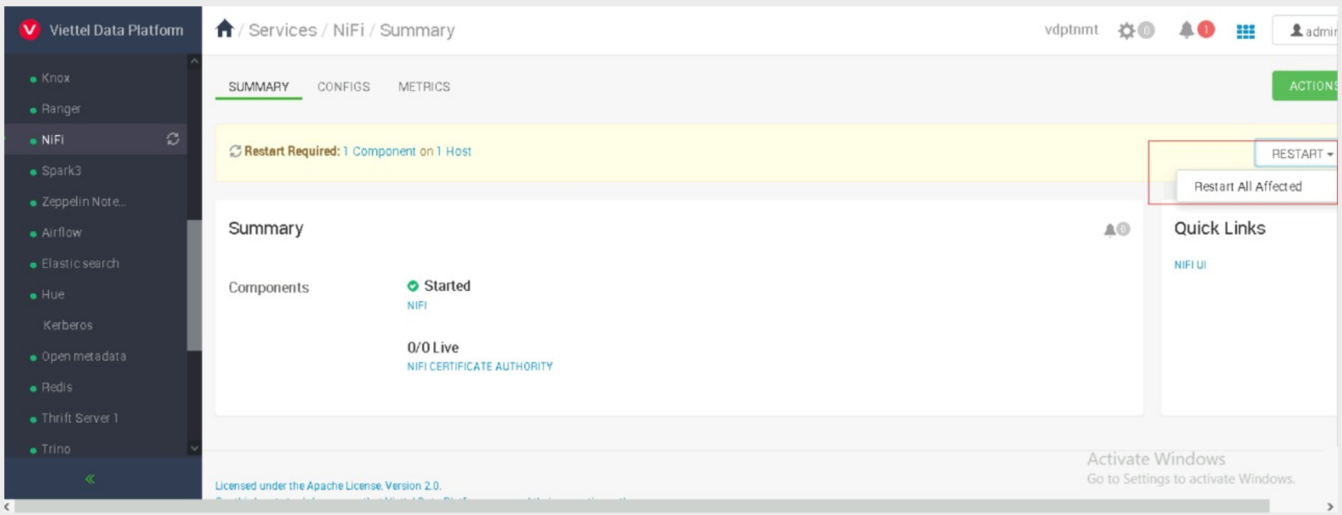
Permissions	Cluster User	Service Operator	Service Administrator	Cluster Operator	Cluster Administrator	Ambari Administrator
Create new clusters						✅
Set service users and groups						✅
Rename clusters						✅
Manage users						✅
Manage groups						✅
Manage Ambari Views						✅
Assign permission and roles						✅
Manage stack versions						✅
Edit stack repository URLs						✅

Danh sách quyền tương tác với giao diện quản trị theo vai trò

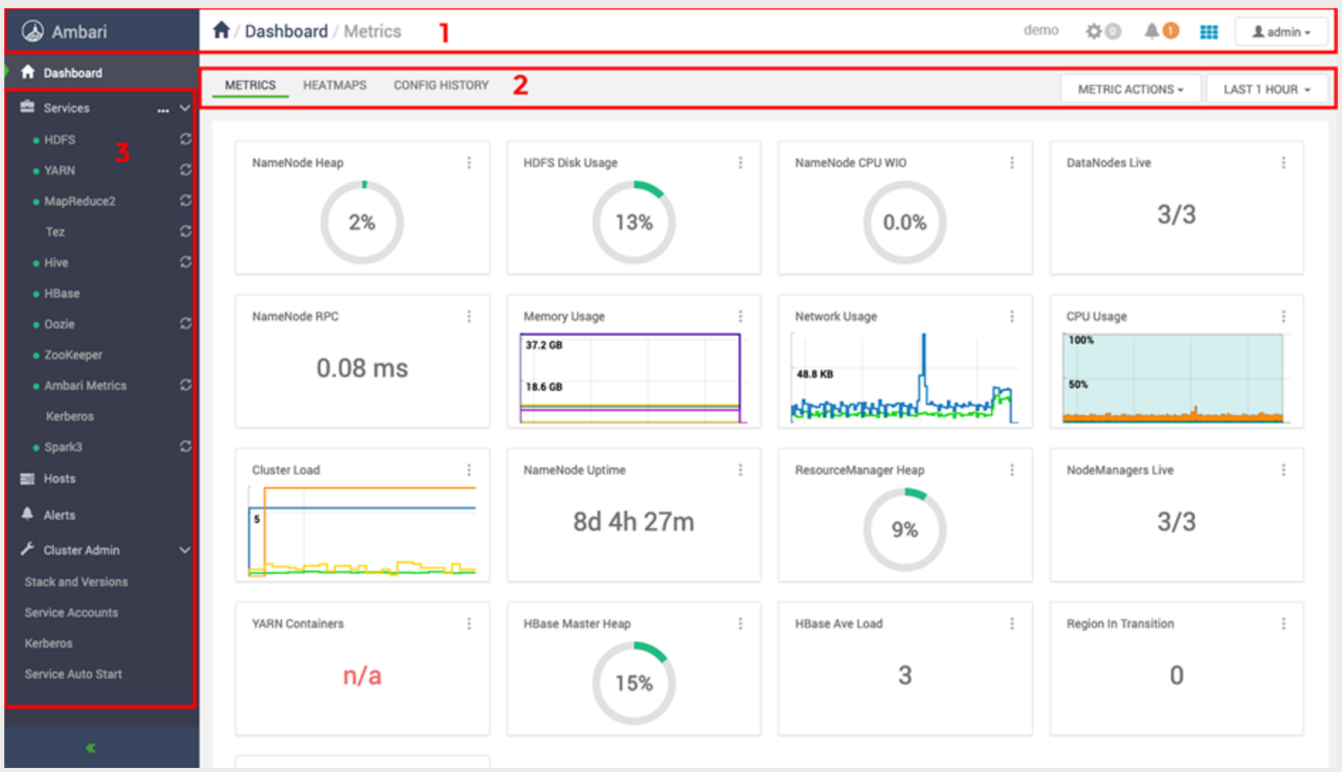
“TÍNH NĂNG

QUẢN TRỊ HỆ THỐNG

- Cung cấp khả năng thay đổi cấu hình ứng dụng, áp dụng cho toàn bộ ứng dụng liên quan trên tập máy chủ lớn bằng cách thay đổi cấu hình trên giao diện quản lý Ambari và cập nhật cho toàn bộ ứng dụng bằng tính năng Restart All Affected Components



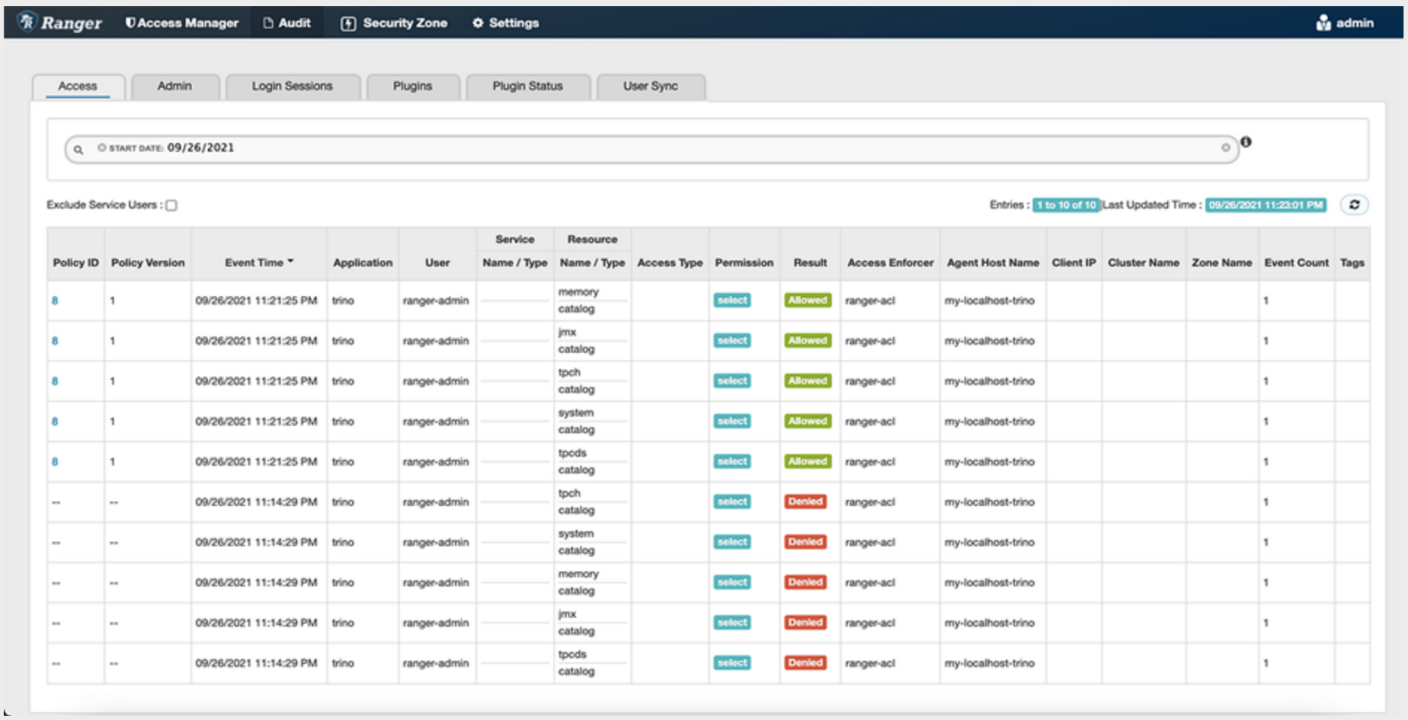
- VDP cung cấp khả năng quản lý, giám sát các tài nguyên hệ thống



“TÍNH NĂNG

QUẢN TRỊ HỆ THỐNG

- Cung cấp khả năng chia sẻ, phân vùng tài nguyên cho các nhóm người dùng khác nhau qua ứng dụng YARN trên giao diện quản lý Ambari. YARN hỗ trợ nhiều phương thức phân vùng tài nguyên, bao gồm:
  - **Queue (Hàng đợi)**: Phân chia tài nguyên theo tổ chức, nhóm người dùng hoặc loại workload.
  - **Node Labels**: Chỉ định tài nguyên cụ thể (CPU, RAM, GPU) cho một nhóm người dùng nhất định.
  - **Resource Pools**: Hạn chế tài nguyên mà mỗi nhóm có thể sử dụng.

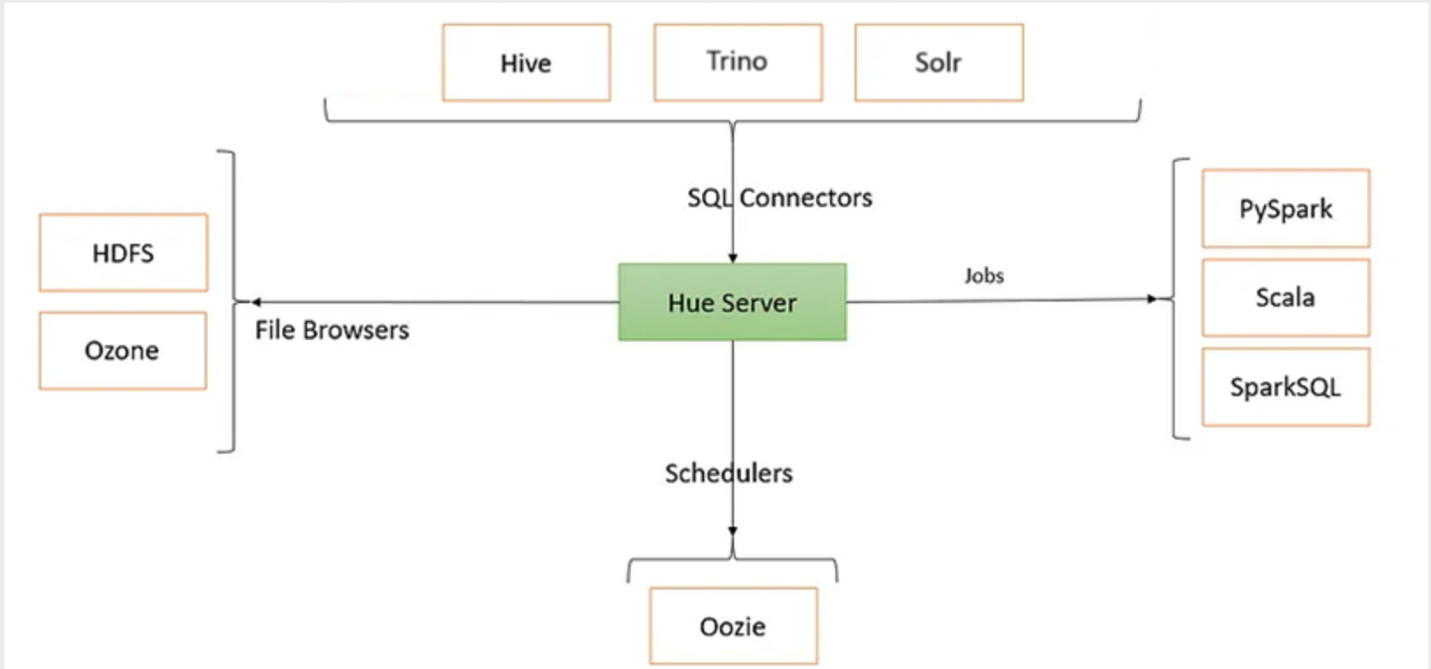


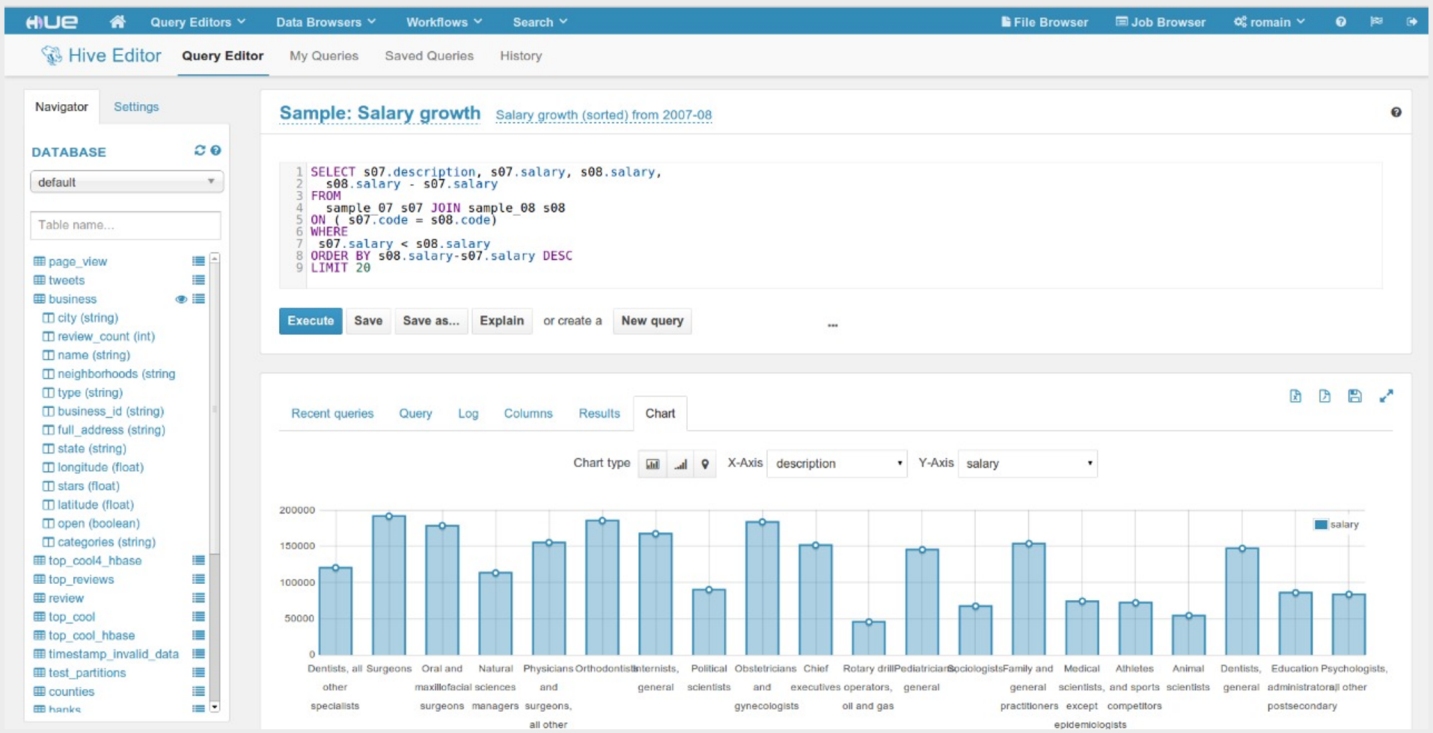
Policy ID	Policy Version	Event Time	Application	User	Service Name / Type	Resource Name / Type	Access Type	Permission	Result	Access Enforcer	Agent Host Name	Client IP	Cluster Name	Zone Name	Event Count	Tags
8	1	09/26/2021 11:21:25 PM	trino	ranger-admin		memory catalog		select	Allowed	ranger-acl	my-localhost-trino				1	
8	1	09/26/2021 11:21:25 PM	trino	ranger-admin		jmx catalog		select	Allowed	ranger-acl	my-localhost-trino				1	
8	1	09/26/2021 11:21:25 PM	trino	ranger-admin		tpch catalog		select	Allowed	ranger-acl	my-localhost-trino				1	
8	1	09/26/2021 11:21:25 PM	trino	ranger-admin		system catalog		select	Allowed	ranger-acl	my-localhost-trino				1	
8	1	09/26/2021 11:21:25 PM	trino	ranger-admin		tpcds catalog		select	Allowed	ranger-acl	my-localhost-trino				1	
--	--	09/26/2021 11:14:29 PM	trino	ranger-admin		tpch catalog		select	Denied	ranger-acl	my-localhost-trino				1	
--	--	09/26/2021 11:14:29 PM	trino	ranger-admin		system catalog		select	Denied	ranger-acl	my-localhost-trino				1	
--	--	09/26/2021 11:14:29 PM	trino	ranger-admin		memory catalog		select	Denied	ranger-acl	my-localhost-trino				1	
--	--	09/26/2021 11:14:29 PM	trino	ranger-admin		jmx catalog		select	Denied	ranger-acl	my-localhost-trino				1	
--	--	09/26/2021 11:14:29 PM	trino	ranger-admin		tpcds catalog		select	Denied	ranger-acl	my-localhost-trino				1	

“TÍNH NĂNG

CÔNG CỤ TƯƠNG TÁC NGƯỜI DÙNG

Hue (Hadoop User Experience): Hue cung cấp một nền tảng trực quan giúp người dùng chạy truy vấn SQL, quản lý dữ liệu trên HDFS, và theo dõi các tác vụ Hadoop mà không cần sử dụng dòng lệnh phức tạp. Hue hỗ trợ tích hợp với nhiều dịch vụ như Hive, Solr, giúp người dùng thực hiện các công việc như phân tích dữ liệu, hoặc quản lý luồng công việc một cách hiệu quả





“TÍNH NĂNG CHUNG

SAO LƯU VÀ KHÔI PHỤC

- Hỗ trợ đa dạng các tần suất sao lưu dữ liệu (hàng ngày, hàng tuần, hàng tháng) và tự động sao lưu theo lịch được lập qua cronjob:

HDFS Snapshot:

- Tạo ảnh chụp nhanh (snapshot) của thư mục trong HDFS mà không ảnh hưởng đến hiệu suất
- Dễ dàng khôi phục dữ liệu về trạng thái trước đó khi gặp sự cố

HDFS DistCp (Distributed Copy):

- Dùng để sao chép dữ liệu giữa các cluster HDP hoặc giữa các thư mục trong HDFS
- Hỗ trợ sao chép dữ liệu song song, tăng tốc độ truyền tải

Hive Export/Import:

- Dùng để sao lưu dữ liệu bảng Hive và khôi phục lại khi cần
- Xuất dữ liệu sang HDFS hoặc hệ thống lưu trữ khác

- VDP hỗ trợ các loại sao lưu dữ liệu (toàn bộ, tăng tiến):

Sao lưu toàn bộ (Full Backup):

- Dùng HDFS DistCp để sao chép dữ liệu toàn bộ: hadoop distcp hdfs://namenode/data hdfs://backup-cluster/data
- Dùng HDFS Snapshot để lưu trữ trạng thái của dữ liệu: hdfs dfs -createSnapshot /data backup_2024_03_23

Sao lưu tăng tiến (Incremental Backup):

- HDFS Snapshot Diff – Xác định & sao lưu chỉ các thay đổi so với snapshot trước đó: hdfs snapshotDiff /data backup_2024_03_22 backup_2024_03_23
- Hive ACID Incremental Backup – Chỉ sao lưu dữ liệu cập nhật trong bảng Hive: EXPORT TABLE sales_data TO 'hdfs://backup/hive/sales_incremental_2024_03_23'

“TÍNH NĂNG CHUNG

SAO LƯU VÀ KHÔI PHỤC

- VDP có chiến lược phòng chống thảm họa bằng các phương pháp:
  - Sử dụng tính năng sao lưu và khôi phục để tạo bản sao dữ liệu.
  - Sử dụng cấu hình cao khả dụng để đảm bảo dịch vụ luôn sẵn có.
  - Sử dụng các công cụ giám sát để phát hiện sự cố.
  - Sử dụng các công cụ bảo mật để bảo vệ dữ liệu khỏi các truy cập trái phép.
- VDP có khả năng kiểm tra tính năng phòng chống thảm họa gồm việc sao lưu dữ liệu và cấu hình, đồng bộ hóa dữ liệu giữa các trung tâm dữ liệu và thiết lập hệ thống dự phòng.
- VDP có khả năng kiểm tra tính năng sao lưu và khôi phục dữ liệu bằng cách sử dụng kịch bản kiểm thử tính năng hoặc sử dụng tính năng so sánh giữa 2 phiên bản dữ liệu để so sánh nội dung snapshot với dữ liệu gốc: hdfs dfs -diff /data/.snapshot/snapshot_2024_03_23 /data

KHẢ NĂNG TRUY CẬP ĐỒNG THỜI

- Khả năng truy cập đồng thời của đội vận hành hệ thống với các cơ chế bảo vệ của Ambari:
  - **Web UI & API**: Cho phép nhiều người dùng giám sát và quản lý cụm cùng lúc.
  - **Phân quyền RBAC**: Giúp nhiều người có thể làm việc đồng thời mà không xung đột quyền.
  - **Cluster State Synchronization**: Đảm bảo mọi thay đổi cấu hình được đồng bộ theo thời gian thực.
  - **Giám sát lịch sử thao tác**: Giúp kiểm soát ai đã thực hiện thay đổi nào để đảm bảo bảo mật

VIETTEL SOLUTIONS

Nền tảng dữ liệu lớn Viettel

- Địa chỉ: Số 1, Trần Hữu Dực, P. Mỹ Đình 2, Q. Nam Từ Liêm, Tp. Hà Nội
- Điện thoại: 096.133.6133
- Email: cskh@viettel-solution.vn