



(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN
GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ**

(11)



2-0004359

(51) **G06F 16/10; G06F 8/60**
2023.01

(13) **Y**

(21) 2-2024-00210

(22) 02/04/2024

(45) 27/10/2025 451

(43) 25/06/2024 435

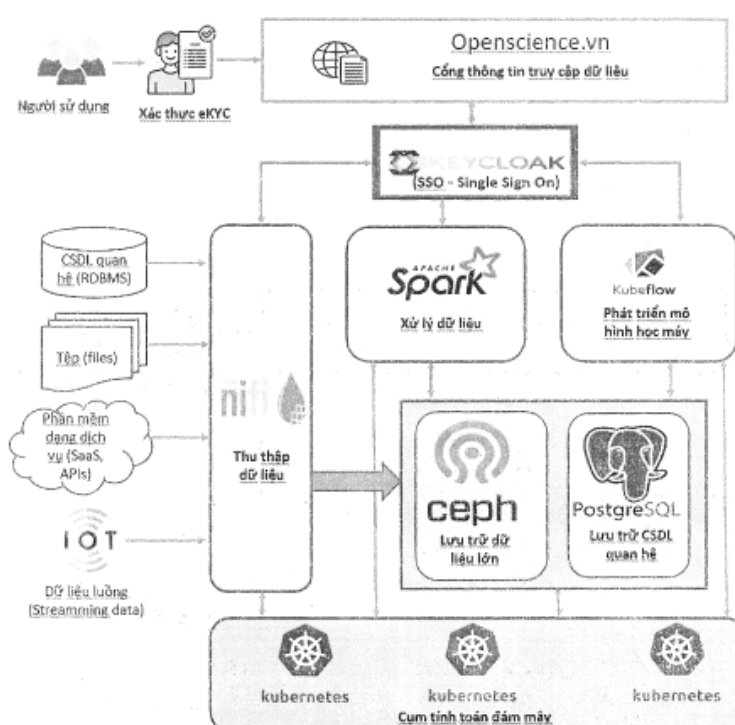
(73) Viện Công nghệ thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam (VN)
18 Hoàng Quốc Việt, Cầu Giấy, Hà Nội, Việt Nam

(72) Nguyễn Long Giang (VN); Đinh Văn Dũng (VN); Võ Sỹ Nam (VN); Phùng Công
Định (VN).

(54) **HỆ THỐNG QUẢN LÝ VÀ CHIA SẺ DỮ LIỆU DÙNG CHUNG MÀ TÍCH HỢP
CÁC NỀN TẢNG XỬ LÝ, PHÂN TÍCH DỮ LIỆU LỚN VÀ PHÁT TRIỂN MÔ
HÌNH HỌC MÁY**

(21) 2-2024-00210

(57) Giải pháp hữu ích đề cập đến hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển học mô hình máy bao gồm: mô đun giao diện người dùng được tạo cấu hình để thực hiện định danh và xác thực người dùng điện tử, phân cấp truy cập và tạo cơ chế đăng nhập một lần cho người dùng; mô đun thu thập dữ liệu được tạo cấu hình để thu thập dữ liệu và tích hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau, các nguồn dữ liệu bao gồm: tệp dữ liệu, cơ sở dữ liệu quan hệ, hệ thống phần mềm dạng dịch vụ, dữ liệu luồng từ các thiết bị truyền dữ liệu theo thời gian thực; mô đun lưu trữ trên nền tảng điện toán đám mây được tạo cấu hình để lưu trữ dữ liệu lớn sau khi được chuyển đổi từ mô đun thu thập dữ liệu; mô đun xử lý dữ liệu được tạo cấu hình để xử lý dữ liệu được lưu trữ trong mô đun lưu trữ hoặc xử lý dữ liệu luồng được thu thập trực tiếp, kết quả xử lý dữ liệu được lưu trữ trong mô đun lưu trữ phục vụ các tác vụ phân tích dữ liệu tiếp theo; mô đun phát triển mô hình học máy được tạo cấu hình để cung cấp các thư viện nền tảng hoặc môi trường cộng tác mà cho phép người dùng quản lý và thực thi quy trình phát triển mô hình học máy, bao gồm xây dựng mô hình, huấn luyện, kiểm thử và quản lý mô hình và tích hợp với các công cụ triển khai phần mềm. Hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển mô hình học máy đáp ứng được yêu cầu về xử lý dữ liệu lớn về các tài liệu khoa học cũng như đáp ứng đầy đủ các tiêu chí về tính chính xác, an toàn và bảo mật.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Giải pháp hữu ích liên quan đến lĩnh vực công nghệ thông tin, cụ thể là đề cập đến hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển mô hình học máy.

Tình trạng kỹ thuật của giải pháp hữu ích

Với sự phát triển mạnh mẽ của cuộc cách mạng công nghiệp lần thứ 4, việc xây dựng và phát triển các nền tảng xử lý, phân tích dữ liệu lớn và các nền tảng phục vụ phát triển các mô hình học máy là vấn đề đang được các nhà nghiên cứu, ứng dụng, các tập đoàn công nghệ lớn quan tâm. Cho đến nay, có rất nhiều các nền tảng phục vụ phát triển các mô hình học máy (machine learning platform) đã được sử dụng rộng rãi trong cộng đồng như KNIME Analytics, Amazon SageMaker, Alteryx Analytics, H2O.ai, Microsoft Azure Machine Learning Studio, Google Cloud AI, AutoML... Song song với nó, nhiều nền tảng phân tích dữ liệu lớn cũng được các tập đoàn công nghệ lớn phát triển như Apache Hadoop, Apache Spark, Google Cloud BigQuery, Amazon EMR, Microsoft Azure HDInsight, Databricks. Tại Việt Nam, một số tập đoàn đã xây dựng các nền tảng phân tích dữ liệu lớn như nền tảng Bkav Big Data Platform của BKAV cho phép thu thập, lưu trữ, xử lý và phân tích dữ liệu lớn (<https://www.bkav.com.vn/big-data-platform>). Tuy nhiên, các giải pháp tích hợp cả hai nền tảng: 1) nền tảng phục vụ thu thập, lưu trữ, xử lý, phân tích dữ liệu lớn; 2) Nền tảng phục vụ phát triển mô hình ML nhằm triển khai đầy đủ một tiến trình (pipeline) từ thu thập, lưu trữ, xử lý dữ liệu lớn và phát triển mô hình học máy còn hạn chế và có tiềm năng ứng dụng hiệu quả.

Việc tích hợp các nền tảng phục vụ xử lý, phân tích dữ liệu lớn và phát triển học máy đã và đang được thực hiện. Tuy nhiên, việc tích hợp các nền tảng đồng thời phục vụ cả hai nhiệm vụ nêu trên trong một khung kiến trúc thống nhất chưa được thực hiện.

Bản chất kỹ thuật của giải pháp hữu ích

Giải pháp hữu ích đề xuất hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển mô hình học máy, hệ thống bao gồm:

mô đun giao diện người dùng được tạo cấu hình để thực hiện định danh và xác thực người dùng điện tử (eKYC), phân cấp truy cập và tạo cơ chế đăng nhập một lần (Single Sign On - SSO) cho người dùng, trong đó, mô đun giao diện người dùng còn tạo giao diện vùng làm việc (workspace) cho người dùng mà người dùng có thể sử dụng dữ liệu được chia sẻ từ nhiều nguồn, sử dụng và phát triển các mô hình học máy để phân tích dữ liệu;

mô đun thu thập dữ liệu được tạo cấu hình để thu thập dữ liệu và tích hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau, các nguồn dữ liệu bao gồm: tệp dữ liệu (files), cơ sở dữ liệu quan hệ (RDBMS), hệ thống phần mềm dạng dịch vụ (API), dữ liệu luồng từ các hệ thống IoT hay các thiết bị truyền dữ liệu theo thời gian thực, trong đó, mô đun thu thập dữ liệu còn thực hiện chuyển đổi dữ liệu từ các nguồn dữ liệu sang nền tảng lưu trữ dữ liệu mà vẫn giữ nguyên nội dung và định dạng của dữ liệu được chuyển đổi;

mô đun lưu trữ trên nền tảng điện toán đám mây được tạo cấu hình để lưu trữ dữ liệu lớn sau khi được chuyển đổi từ mô đun thu thập dữ liệu, trong đó dữ liệu được lưu trữ theo đối tượng (Object storage), khối (Block storage) và tệp dữ liệu (File storage) trong một nền tảng đơn nhất;

mô đun xử lý dữ liệu được tạo cấu hình để xử lý dữ liệu được lưu trữ trong mô đun lưu trữ hoặc xử lý dữ liệu luồng được thu thập trực tiếp, kết quả xử lý dữ liệu được lưu trữ trong mô đun lưu trữ phục vụ các tác vụ phân tích dữ liệu tiếp theo, trong đó, mô đun xử lý dữ liệu sử dụng khả năng ghi vào bộ nhớ đệm nằm trong bộ nhớ và thực thi truy vấn tối ưu hóa nhằm giúp truy vấn phân tích nhanh dữ liệu có kích thước bất kỳ;

mô đun phát triển mô hình học máy được tạo cấu hình để cung cấp các thư viện nền tảng hoặc môi trường cộng tác mà cho phép người dùng quản lý và thực thi quy trình phát triển mô hình học máy, bao gồm xây dựng mô hình, huấn luyện, kiểm thử và quản lý mô hình và tích hợp với các công cụ triển khai phần mềm.

Hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển mô hình học máy đáp ứng được yêu cầu về xử lý dữ liệu lớn về các tài liệu khoa học cũng như đáp ứng đầy đủ các tiêu chí về tính chính xác, an toàn và bảo mật.

Mô tả vắn tắt các hình vẽ

Hình 1 là hình vẽ minh họa sơ đồ kiến trúc hệ thống quản lý và chia sẻ dữ liệu dùng chung được tích hợp các nền tảng khác.

Hình 2 là hình vẽ minh họa sơ đồ kiến trúc ứng dụng của hệ thống quản lý và chia sẻ dữ liệu dùng chung được tích hợp các nền tảng khác.

Mô tả chi tiết giải pháp hữu ích

Các mô tả dưới đây về hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển mô hình học máy sẽ được giải thích theo các hình vẽ đính kèm, từ đó giải pháp hữu ích sẽ được hiểu một cách rõ ràng và rành mạch. Các mô tả dưới đây có đưa ra những ví dụ khi thực hiện giải pháp hữu ích này và không giới hạn phạm vi kỹ thuật của giải pháp hữu ích.

Tham chiếu tới hình 1, hệ thống theo giải pháp hữu ích bao gồm các mô đun sau:

Mô đun thu thập, tích hợp dữ liệu từ nhiều nguồn (Data Ingestion) được tạo cấu hình để thu thập dữ liệu và tích hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau, các nguồn dữ liệu bao gồm: tệp dữ liệu (files), cơ sở dữ liệu quan hệ (RDBMS), hệ thống phần mềm dạng dịch vụ (API), dữ liệu luồng từ các hệ thống IoT hay các thiết bị truyền dữ liệu theo thời gian thực, trong đó, mô đun thu thập dữ liệu còn thực hiện chuyển đổi dữ liệu từ các nguồn dữ liệu sang nền tảng lưu trữ dữ liệu mà vẫn giữ nguyên nội dung và định dạng của dữ liệu được chuyển đổi. Theo đó, dữ liệu được thu thập là dữ liệu thô và được thu thập theo lô từ các nguồn tệp dữ liệu (files), cơ sở dữ liệu quan hệ (RDBMS), hệ thống phần mềm dạng dịch vụ (API) và dữ liệu được thu thập là dữ liệu theo luồng gồm dữ liệu từ các thiết bị IoT hay các thiết bị truyền dữ liệu theo thời gian thực. Theo một phương án của giải pháp hữu ích, mô đun thu thập, tích hợp dữ liệu có thể được xây dựng dựa trên nền tảng Apache NiFi,

đây là một trong những giải pháp mã nguồn mở phổ biến cho phép kết nối với nhiều nguồn dữ liệu khác nhau và đưa dữ liệu vào nền tảng lưu trữ dữ liệu. NiFi sử dụng kiến trúc có thể cho phép tạo các trình kết nối mới bằng Java. Quá trình tích hợp dữ liệu bao gồm một số tác vụ chính như: kết nối với các nguồn dữ liệu thời gian thực (streaming data) hoặc nguồn dữ liệu dạng gói (batch data); chuyển đổi dữ liệu từ các nguồn dữ liệu sang nền tảng lưu trữ dữ liệu mà vẫn giữ nguyên nội dung và định dạng của dữ liệu (việc bảo toàn dữ liệu này rất quan trọng đối với các dữ liệu được xử lý lại sau này); ghi lại các số liệu thống kê và trạng thái của dữ liệu sau đó lưu thông tin vào khối lưu trữ dữ liệu. NiFi có thể tích hợp dữ liệu từ nhiều nguồn khác nhau, bao gồm: cơ sở dữ liệu quan hệ (RDBMS), tệp (files), phần mềm dạng dịch vụ (SaaS, APIs) và dữ liệu luồng từ các hệ thống IoT. Ngoài ra, NiFi cho phép cấu hình để thu thập dữ liệu với dung lượng lưu trữ lớn. Do đó, NiFi được xem là một trong những nền tảng thu thập dữ liệu thô tiên tiến nhất hiện nay với nhiều tính năng và hiệu quả vượt trội.

Mô đun lưu trữ trên nền tảng điện toán đám mây được tạo cấu hình để lưu trữ dữ liệu lớn sau khi được chuyển đổi từ mô đun thu thập dữ liệu, trong đó dữ liệu được lưu trữ theo dạng cấu trúc như lưu trữ đối tượng (Object storage), lưu trữ khối (Block storage) và lưu trữ tệp dữ liệu (File storage) trong một nền tảng đơn nhất. Trong đó, hệ thống lưu trữ tệp chủ yếu được sử dụng để lưu trữ dữ liệu nghiên cứu khoa học, hệ thống lưu trữ đối tượng được sử dụng để lưu trữ dữ liệu nghiên cứu khoa học có giá trị ở định dạng đối tượng. Theo một phương án của giải pháp hữu ích, mô đun lưu trữ có thể được xây dựng dựa trên nền tảng Ceph, đây là nền tảng mã nguồn mở để xây dựng hạ tầng lưu trữ (storage) phân tán, ổn định, độ tin cậy và hiệu năng cao, dễ dàng mở rộng. Với hệ thống lưu trữ được điều khiển bằng phần mềm, Ceph cung cấp giải pháp lưu trữ theo đối tượng (Object storage), khối (Block storage) và tệp dữ liệu (File storage) trong một nền tảng đơn nhất. Ceph chạy trên nền tảng điện toán đám mây (cloud) với các thiết bị phần cứng ổn định và tiên tiến nhất, giúp tiết kiệm chi phí và sử dụng dễ dàng. Cho đến nay, Ceph là công nghệ lưu trữ tiên tiến nhất với kiến trúc lưu trữ hồ dữ liệu (Loakehouse).

Mô đun xử lý dữ liệu được tạo cấu hình để xử lý dữ liệu được lưu trữ trong mô đun lưu trữ hoặc xử lý dữ liệu luồng được thu thập trực tiếp, kết quả xử lý dữ liệu được lưu trữ

trong mô đun lưu trữ phục vụ các tác vụ phân tích dữ liệu tiếp theo, trong đó, mô đun xử lý dữ liệu sử dụng khả năng ghi vào bộ nhớ đệm nằm trong bộ nhớ và thực thi truy vấn tối ưu hóa nhằm giúp truy vấn phân tích nhanh dữ liệu có kích thước bất kỳ. Theo một phương án của giải pháp hữu ích, mô đun xử lý dữ liệu có thể được xây dựng dựa trên nền tảng Apache Spark, đây là một hệ thống xử lý phân tán mã nguồn mở được sử dụng cho các khối lượng công việc xử lý dữ liệu lớn. Hệ thống này sử dụng khả năng ghi vào bộ nhớ đệm nằm trong bộ nhớ và thực thi truy vấn tối ưu hóa nhằm giúp truy vấn phân tích nhanh dữ liệu có kích thước bất kỳ. Apache Spark cung cấp các API phát triển bằng ngôn ngữ Java, Scala, Python và R và hỗ trợ tái sử dụng mã trên nhiều khối lượng công việc, chẳng hạn như xử lý dữ liệu theo lô, truy vấn tương tác, xử lý dữ liệu theo luồng.

Mô đun phát triển mô hình học máy được tạo cấu hình để cung cấp các thư viện nền tảng hoặc môi trường cộng tác mà cho phép người dùng quản lý và thực thi quy trình phát triển mô hình học máy, bao gồm xây dựng mô hình, huấn luyện, kiểm thử và quản lý mô hình và tích hợp với các công cụ triển khai phần mềm. Các mô hình học máy sau khi được phát triển có thể được chia sẻ và áp dụng vào việc phân tích dữ liệu, hoặc sử dụng nguồn dữ liệu có sẵn để huấn luyện cho các mô hình học máy khác. Theo một phương án của giải pháp hữu ích, mô đun phát triển mô hình học máy có thể dựa trên nền tảng Kubeflow, là nền tảng mã nguồn mở nhằm triển khai các mô hình học máy (Machine learning – ML). Nó được phát triển trên nền tảng Kubernetes giúp tự động hoá việc triển khai, mở rộng quy mô và quản lý các ứng dụng ML. Bằng cách tận dụng sức mạnh của Kubernetes, Kubeflow cho phép các nhà khoa học dữ liệu xây dựng, triển khai và quản lý quy trình phát triển ML một cách dễ dàng, hiệu quả, từ đó đẩy nhanh quá trình phát triển và ứng dụng các mô hình ML vào các bài toán thực tế. Kubeflow bao gồm các tính năng như cung cấp nền tảng hợp nhất với các công cụ phát triển ML như Tensorflow, PyTorch... linh hoạt trong kiến trúc mô đun, tự động hoá phân bổ tài nguyên và tích hợp với các công cụ triển khai liên tục cho phép người dùng tự động hoá vòng đời ML, giúp giảm thời gian và công sức cần thiết để triển khai các mô hình ML. Do đó, Kubeflow được xem là nền tảng phát triển mô hình ML tiên tiến nhất hiện nay.

Mô đun giao diện người dùng được tạo cấu hình để thực hiện định danh và xác thực người dùng điện tử (eKYC), phân cấp truy cập và tạo cơ chế đăng nhập một lần (Single Sign On - SSO) cho người dùng, trong đó, mô đun giao diện người dùng còn tạo giao diện vùng làm việc (workspace) cho người dùng mà người dùng có thể sử dụng dữ liệu được chia sẻ từ nhiều nguồn, sử dụng và phát triển các mô hình học máy để phân tích dữ liệu. Trong đó, vùng làm việc có nhiều danh mục nội dung mà người dùng có thể truy cập và thực hiện các tác vụ như: xem, chỉnh sửa, cập nhật, bình luận, tải xuống ... Để quản lý danh mục nội dung này hệ thống có thể lưu trữ danh mục nội dung được gửi hoặc truyền lên trong mô đun lưu trữ dữ liệu.

Giải pháp xác thực người dùng trực tuyến qua eKYC cho phép thực hiện “Xác minh người dùng điện tử” trực tuyến qua hệ thống. Quá trình eKYC bao gồm hình ảnh chụp chứng minh nhân dân, thông tin cá nhân và công nghệ nhận dạng khuôn mặt hoặc dấu vân tay để xác minh danh tính của người dùng. Hình thức eKYC có thể thay thế và đáp ứng đầy đủ các tiêu chí về tính chính xác, an toàn và bảo mật so với cách thức xác thực truyền thống mà các hệ thống quản trị dữ liệu đang sử dụng hiện nay.

Các bước thực hiện xác thực eKYC bao gồm:

(1) Cung cấp thông tin.

Để bắt đầu quá trình eKYC, người sử dụng cần đăng ký tài khoản và thực hiện xác thực tài khoản qua email. Sau khi xác thực tài khoản qua email, người sử dụng có đủ điều kiện tham gia vào quy trình eKYC một cách nhanh chóng và thuận tiện từ hệ thống được tích hợp.

(2) Xác minh giấy tờ

Người sử dụng cung cấp và chụp ảnh giấy tờ tùy thân như CMND/CCCD, hộ chiếu... Công nghệ OCR giúp nhận diện thông tin từ hình ảnh giấy tờ với độ chính xác cao gần như tuyệt đối.

(3) Xác thực khuôn mặt

Người sử dụng sẽ được yêu cầu chụp ảnh khuôn mặt hoặc quay video để xác minh. Hệ thống tích hợp sẽ so sánh khuôn mặt của người sử dụng với hình ảnh trên giấy tờ tùy thân, đảm bảo tính chính xác và ngăn chặn hiện tượng giả mạo.

(4) Trả kết quả xác thực

Hệ thống tích hợp đánh giá và trả về kết quả xác thực. Nếu thành công, người sử dụng được cấp quyền truy cập vào hệ thống Openscience.vn và các nền tảng quản trị và phân tích dữ liệu bên trong. Nếu không, người dùng cần xác thực lại.

Giải pháp đăng nhập một lần SSO

Hệ thống cho phép các lớp người sử dụng truy cập vào các nền tảng tích hợp bên trong của hệ thống để thực hiện một chu trình (pipeline) phân tích dữ liệu.. Mỗi nền tảng mã nguồn mở như Nifi, CEPH, Spark, Kubeflow...đều có danh sách các người sử dụng riêng với cơ chế phân quyền riêng. Người dùng có thể truy cập thẳng vào các nền tảng tích hợp bên trong của hệ thống. Giải pháp SSO sử dụng nền tảng nguồn mở Keycloak để quản lý danh sách người sử dụng và quản lý truy cập của tất cả các nền tảng của hệ thống. Keycloak có vai trò đồng bộ hoá danh sách người sử dụng của hệ thống Openscience.vn với danh sách người sử dụng của các nền tảng Nifi, Kubeflow, Spark.

Các bước thực hiện giải pháp SSO bao gồm:

i) Xây dựng giải pháp SSO

Để xây dựng giải pháp SSO trong Openscience.vn sử dụng Keycloak, các bước thực hiện như sau:

- (1) Cài đặt hệ thống mã nguồn mở Keycloak trên cụm K8S
- (2) Cấu hình, kết nối Keycloak với hệ thống Openscience.vn và kết nối với các nền tảng Nifi, Spark, Kubeflow
- (3) Đồng bộ danh sách người sử dụng Openscience.vn với Keycloak (khi thực hiện đăng ký người sử dụng)
- (4) Đồng bộ danh sách người sử dụng của Keycloak với người sử dụng của các nền tảng Nifi, Spark, Kubeflow

- (5) Xây dựng module tạo vùng làm việc (workspace) đối với mỗi người sử dụng trong Kubeflow khi người sử dụng đăng nhập Openscience.vn
- (6) Xây dựng module tạo vùng làm việc (workspace) đối với mỗi người sử dụng trong Nifi khi người sử dụng đăng nhập Openscience.vn

ii) Thử nghiệm giải pháp SSO

- (1) Tạo và phân quyền cho 1 người sử dụng trên Openscience.vn
- (2) Đăng nhập hệ thống qua Openscience.vn
- (3) Truy cập vào Nifi thực hiện nhận dữ liệu. Kiểm tra tài khoản và vùng làm việc (workspace) của người sử dụng trên Nifi
- (4) Truy cập vào Spark thực hiện xử lý dữ liệu. Kiểm tra tài khoản của người sử dụng trên Spark
- (5) Truy cập vào Kubeflow thực hiện chạy mô hình học máy. Kiểm tra tài khoản và vùng làm việc của người sử dụng trên Kubeflow
- (6) Thoát khỏi hệ thống và vào tài khoản khác để kiểm tra lại.

Tham chiếu tới hình 2 là sơ đồ kiến trúc ứng dụng của hệ thống quản lý và chia sẻ dữ liệu dùng chung được tích hợp các nền tảng khác. Kiến trúc ứng dụng hệ thống được mô tả ở hình 2, bao gồm các lớp:

Lớp người dùng: bao gồm nhóm quản trị và nhóm khai thác hệ thống. Nhóm quản trị hệ thống có nhiệm vụ quản trị, phân quyền cho người sử dụng và phê duyệt, hiển thị các bộ dữ liệu trên hệ thống. Nhóm khai thác hệ thống bao gồm các nhà khoa học dữ liệu, cán bộ nghiên cứu, giảng viên, sinh viên đại học thuộc lĩnh vực khoa học dữ liệu, trí tuệ nhân tạo và các lĩnh vực liên quan.

Lớp giao diện: là cổng thông tin khai thác hệ thống. Lớp giao diện cho phép các lớp người dùng truy cập và khai thác các nền tảng bên trong hệ thống với cơ chế đăng nhập một lần SSO, bao gồm các khối chức năng: quản trị hệ thống, nạp dữ liệu, tải dữ liệu, chia sẻ dữ liệu theo phân quyền và truy cập vào các nền tảng bên trong của hệ thống.

Lớp ứng dụng nền tảng: bao gồm các nền tảng thập dữ liệu, xử lý dữ liệu, quản trị, lưu trữ dữ liệu, quản trị đám mây và phát triển mô hình học máy.

Ngoài ra, hệ thống Openscience.vn còn bao gồm các lớp:

- Lớp lưu trữ dữ liệu với: lưu trữ dữ liệu lớn bằng công nghệ CEPH và lưu trữ cơ sở dữ liệu quan hệ với hệ quản trị PostgreSQL.
- Lớp hạ tầng vật lý: bao gồm hạ tầng đám mây (Cloud), hạ tầng đường truyền (Internet), hạ tầng tính toán (CPU, GPU) và hạ tầng lưu trữ vật lý (máy chủ, thiết bị lưu trữ).

Yêu cầu bảo hộ

1. Hệ thống quản lý và chia sẻ dữ liệu dùng chung mà tích hợp các nền tảng xử lý, phân tích dữ liệu lớn và phát triển học mô hình máy bao gồm:

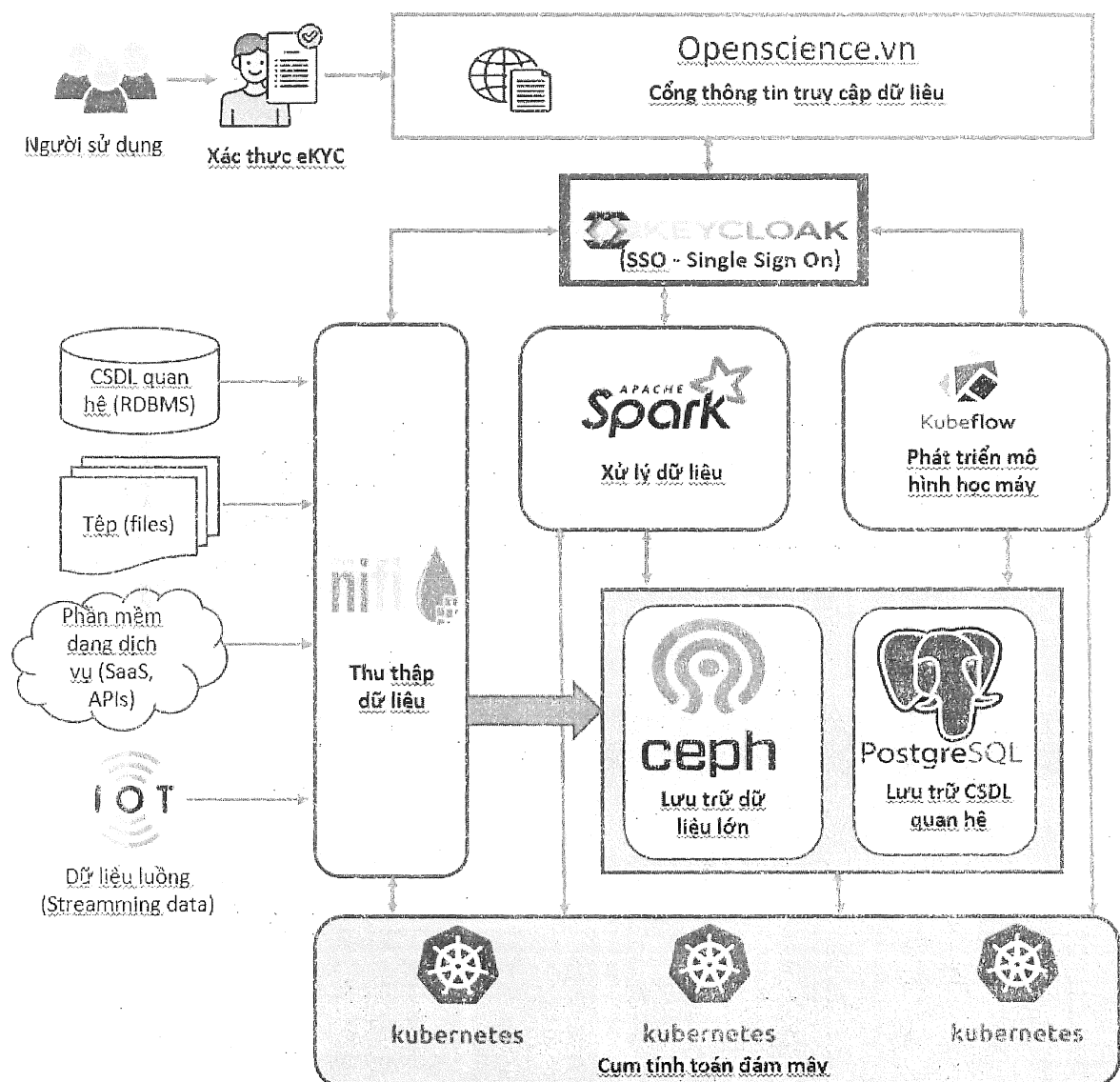
mô đun giao diện người dùng được tạo cấu hình để thực hiện định danh và xác thực người dùng điện tử (eKYC), phân cấp truy cập và tạo cơ chế đăng nhập một lần (Single Sign On - SSO) cho người dùng, trong đó, mô đun giao diện người dùng còn tạo giao diện vùng làm việc (workspace) cho người dùng mà người dùng có thể sử dụng dữ liệu được chia sẻ từ nhiều nguồn, sử dụng và phát triển các mô hình học máy để phân tích dữ liệu;

mô đun thu thập dữ liệu được tạo cấu hình để thu thập dữ liệu và tích hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau, các nguồn dữ liệu bao gồm: tệp dữ liệu (files), cơ sở dữ liệu quan hệ (RDBMS), hệ thống phần mềm dạng dịch vụ (API), dữ liệu luồng từ các hệ thống IoT hay các thiết bị truyền dữ liệu theo thời gian thực, trong đó, mô đun thu thập dữ liệu còn thực hiện chuyển đổi dữ liệu từ các nguồn dữ liệu sang nền tảng lưu trữ dữ liệu mà vẫn giữ nguyên nội dung và định dạng của dữ liệu được chuyển đổi;

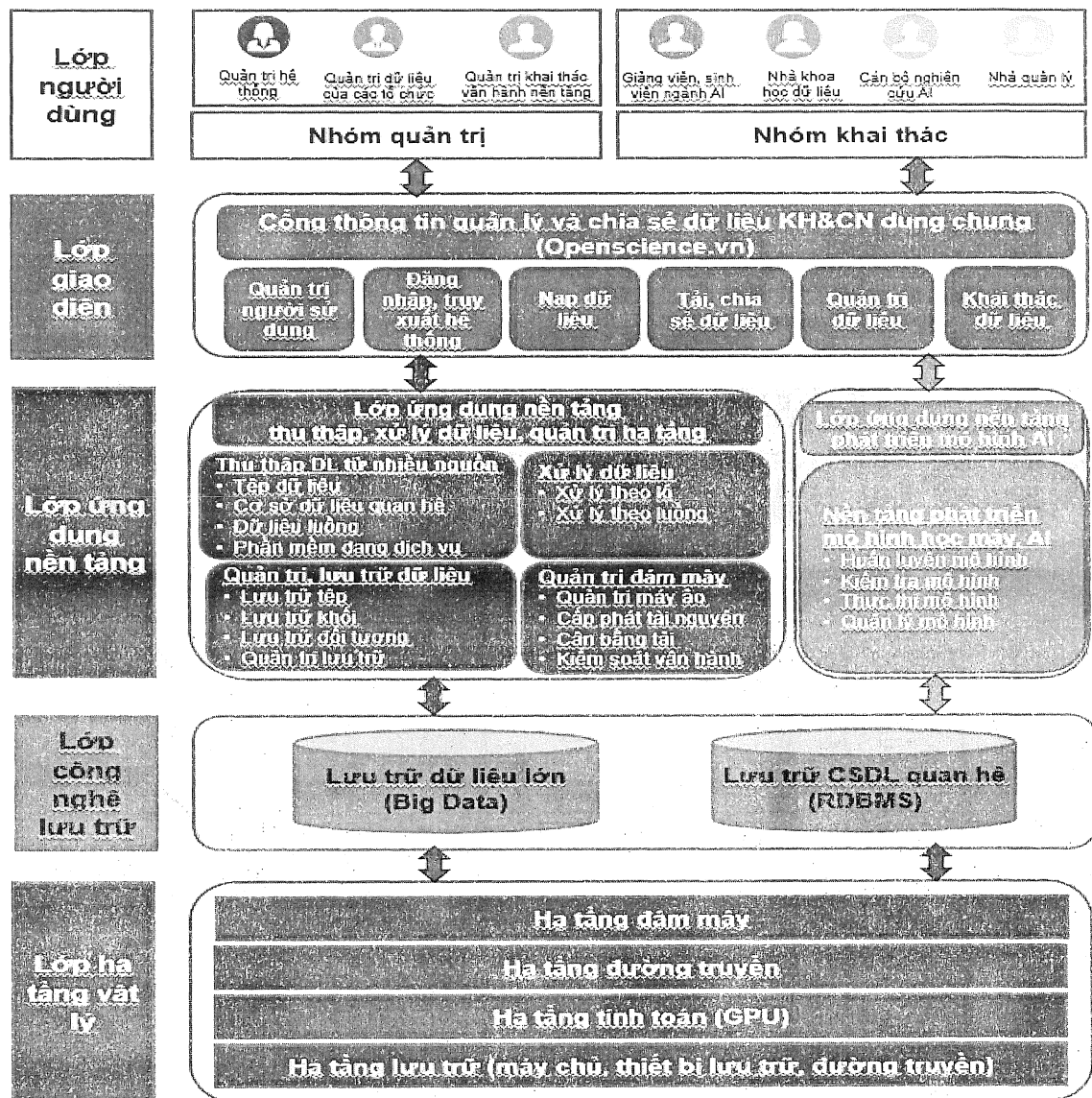
mô đun lưu trữ trên nền tảng điện toán đám mây được tạo cấu hình để lưu trữ dữ liệu lớn sau khi được chuyển đổi từ mô đun thu thập dữ liệu, trong đó dữ liệu được lưu trữ theo đối tượng (Object storage), khối (Block storage) và tệp dữ liệu (File storage) trong một nền tảng đơn nhất;

mô đun xử lý dữ liệu được tạo cấu hình để xử lý dữ liệu được lưu trữ trong mô đun lưu trữ hoặc xử lý dữ liệu luồng được thu thập trực tiếp, kết quả xử lý dữ liệu được lưu trữ trong mô đun lưu trữ phục vụ các tác vụ phân tích dữ liệu tiếp theo, trong đó, mô đun xử lý dữ liệu sử dụng khả năng ghi vào bộ nhớ đệm nằm trong bộ nhớ và thực thi truy vấn tối ưu hóa nhằm giúp truy vấn phân tích nhanh dữ liệu có kích thước bất kỳ;

mô đun phát triển mô hình học máy được tạo cấu hình để cung cấp các thư viện nền tảng hoặc môi trường cộng tác mà cho phép người dùng quản lý và thực thi quy trình phát triển mô hình học máy, bao gồm xây dựng mô hình, huấn luyện, kiểm thử và quản lý mô hình và tích hợp với các công cụ triển khai phần mềm.



Hình 1



Hình 2