**BÁO CÁO ĐỀ XUẤT**

**DỰ ÁN**

**XÂY DỰNG KHO DỮ LIỆU DÙNG CHUNG VÀ CÁC CƠ SỞ DỮ LIỆU CHUYÊN MÔN**

**Năm 2024**

**MỤC LỤC**

[PHẦN A. CÁC THÔNG TIN CHUNG 3](#_Toc170571917)

[I. Căn cứ pháp lý 3](#_Toc170571918)

[II. Tên dự án 3](#_Toc170571919)

[III. Chủ đầu tư 3](#_Toc170571920)

[IV. Địa điểm thực hiện 3](#_Toc170571921)

[V. Thời gian thực hiện 3](#_Toc170571922)

[VI. Đơn vị tư vấn 3](#_Toc170571923)

[VII. Loại nguồn vốn 4](#_Toc170571924)

[PHẦN B. SỰ CẦN THIẾT THỰC HIỆN 5](#_Toc170571925)

[I. Cập nhật hiện trạng 5](#_Toc170571926)

[II. Sự cần thiết đầu tư 7](#_Toc170571927)

[III. Đánh giá sự phù hợp với quy hoạch có liên quan theo quy định của pháp luật về quy hoạch 7](#_Toc170571928)

[IV. Phân tích, xác định mục tiêu, nhiệm vụ, kết quả đầu ra của dự án; phân tích, lựa chọn quy mô hợp lý; xác định phân kỳ đầu tư; lựa chọn hình thức đầu tư 8](#_Toc170571929)

[1. Mục tiêu đầu tư 8](#_Toc170571930)

[2. Kết quả của dự án 8](#_Toc170571931)

[3. Phân tích, lựa chọn quy mô 9](#_Toc170571932)

[4. Hình thức đầu tư: 10](#_Toc170571933)

[V. Phân tích các điều kiện tự nhiên, điều kiện kinh tế - kỹ thuật, lựa chọn địa điểm đầu tư 10](#_Toc170571934)

[1. Các điều kiện tự nhiên, điều kiện kinh tế - kỹ thuật 10](#_Toc170571935)

[2. Lựa chọn địa điểm đầu tư: 10](#_Toc170571936)

[PHẦN C. NỘI DUNG CHÍNH CỦA BÁO CÁO 11](#_Toc170571937)

[I. Đánh giá sự tuân thủ Khung Kiến trúc Chính phủ điện tử Việt Nam, Kiến trúc Chính phủ điện tử cấp bộ 11](#_Toc170571938)

[II. Phân tích, lựa chọn phương án công nghệ, kỹ thuật, thiết bị 11](#_Toc170571939)

[1. Tiêu chí lựa chọn hệ thống quản lý tài nguyên máy 12](#_Toc170571940)

[2. Phân tích lựa chọn công nghệ quản lý hệ thống và tài nguyên máy 13](#_Toc170571941)

[3. Tiêu chí lựa chọn giải pháp công nghệ 18](#_Toc170571942)

[4. Phân tích lựa chọn công nghệ cho các khối chức năng 20](#_Toc170571943)

[5. Khối hạ tầng 68](#_Toc170571944)

[PHẦN D. DỰ TOÁN 70](#_Toc170571945)

[I. Tổng mức đầu tư: 70](#_Toc170571946)

[II. Tóm tắt dự toán các hạng mục chính 71](#_Toc170571947)

# CÁC THÔNG TIN CHUNG

## Căn cứ pháp lý

Nghị định số 64/2007/NĐ-CP ngày 10/4/2007 của Chính phủ về ứng dụng công nghệ thông tin trong hoạt động của cơ quan nhà nước.

Nghị định số 73/2019/NĐ-CP ngày 05/09/2019 của Chính phủ về việc Quy định quản lý đầu tư ứng dụng công nghệ thông tin sử dụng nguồn vốn ngân sách nhà nước.

Nghị định số 40/2020/NĐ-CP ngày 06/4/2020 của Chính phủ về Quy định chi tiết thi hành một số điều của Luật Đầu tư công.

Nghị định số 85/2016/NĐ-CP ngày 01/7/2016 của Chính phủ Về bảo đảm an toàn hệ thống thông tin theo cấp độ.

Nghị định số 47/2020/NĐ – CP ngày 09 tháng 04 năm 2020 của Chính phủ về việc quản lý, kết nối và chia sẻ dữ liệu số của cơ quan nhà nước.

Quyết định số 749/QĐ-TTg của Thủ tướng Chính phủ: Phê duyệt "Chương trình Chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2030".

Thông tư số 39/2017/TT-BTTTT ngày 15/12/2017 của Bộ Thông tin và Truyền thông về việc Ban hành Danh mục tiêu chuẩn kỹ thuật về ứng dụng công nghệ thông tin trong cơ quan nhà nước.

Quyết định số 2323/QĐ-BTTTT ngày 31/12/2019 của Bộ Thông tin và Truyền thông về việc Ban hành Khung Kiến trúc Chính phủ điện tử Việt Nam, phiên bản 2.0.

Bổ xung thêm sau khi có quyết định duyệt chủ chương

## Tên dự án

Xây dựng kho dữ liệu dùng chung và các cơ sở dữ liệu chuyên môn

## Chủ đầu tư

Vụ tin học – Văn phòng Quốc hội

## Địa điểm thực hiện

Trung tâm thông tin…

## Thời gian thực hiện

Năm 2024 - 2025.

## Đơn vị tư vấn

Viện Công nghệ thông tin

Địa chỉ: 18 Hoàng Quốc Việt - Phường Nghĩa Đô - Quận Cầu Giấy - Hà Nội

## Loại nguồn vốn

Ngân sách nhà nước.

# SỰ CẦN THIẾT THỰC HIỆN

## Cập nhật hiện trạng

#### Hiện trạng về hạ tầng kỹ thuật

#### Hiện trạng trung tâm tích hợp dữ liệu

Trung tâm tích hợp dữ liệu của Quốc Hội được đặt tại tầng ….. có diện tích 100 m2, trung tâm được sử dụng để dùng chung cho các cơ đơn vị trong Quốc Hội

Trung tâm tích hợp dữ liệu gồm các thành phần:

* Hệ thống máy chủ ảo hóa
* Hệ thống lưu trữ dữ liệu
* Hệ thống thiết bị mạng và anh ninh bảo mật

Hệ thống hạ tầng cho trung tâm dữ liệu.

* Hạ tầng của trung tâm tích hợp dữ liệu bao gồm:
* Hệ thống sàn nâng;
* Hệ thống tủ rack: 07 tủ rack;
* Hệ thống điện, UPS: bao gồm 01 tủ phân phối và 03 UPS.
* Hệ thống điều hòa: 02 điều hòa âm trần điện 01 phase, 01 điều hòa treo tường có công suất 18.000 BTU.
* Hệ thống phòng cháy, chữa cháy
* ....

Khảo sát viết thêm sau

#### Hiện trạng hệ thống máy chủ, lưu trữ:

* Bao gồm 07 máy chủ (03 máy chủ Dell EMC PowerEdge R760; 03 máy chủ HPE ProLiant DL360 Gen 10; 01 máy chủ HPE ProLiant DL20 Gen9); 02 thiết bị lưu trữ Dell EMC NX440, 48 TB Raw. Các máy chủ ứng dụng triển khai dựa trên công nghệ ảo hoá cơ bản đáp ứng được yêu cầu cho các ứng dụng ở giai đoạn 2015 – 2023….

Khảo sát viết thêm sau

#### Hiện trạng ứng dụng phần mềm

Danh sách các phần mềm đang được triển khai tại Quốc Hội:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TT** | **Tên phần mềm ứng dụng** | **Mô tả** | **Ngôn ngữ lập trình/CSDL** | **Giao tiếp người dùng (Web/App...)** | **Năm đưa vào sử dụng** | **Phạm vi triển khai, đối tượng người dùng** |
| 1 | Phần mềm 1 |  | Visual C/Java/C++ | Phần mềm chuyên ngành nội bộ | 2018 |  |
| 2 | Phần mềm 2 |  | Java | Phần mềm chuyên ngành nội bộ | 2020 |  |
| 3 | Phần mềm …. |  |  | Phần mềm chuyên ngành nội bộ | 2020 |  |

Khảo sát viết thêm sau

#### Hiện trạng dữ liệu

Quy mô dữ liệu được khảo sát như sau:

* Số nguồn dữ liệu lưu trữ và chia sẻ: 66 nguồn dữ liệu (đã khảo sát 6/28 đơn vị).
* Dung lượng dữ liệu: khoảng 2.5TB/5 sở/năm (chưa tính Tài nguyên và Môi trường).
* 6/6 đơn vị có nhu cầu tham gia chia sẻ thông tin và tham gia Kho dữ liệu mở.

Trong đó:

Các nguồn dữ liệu được chia sẻ:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Lĩnh vực** | **Số nguồn dữ liệu chia sẻ** | **Dữ liệu quan trọng** |
| 1 | Quốc hội | 6 | CSDL Quốc hội |
| 2 | Văn bản | 16 | Văn bản pháp luật |
| 3 | … | 4 | Dữ liệu khác |

Khảo sát viết thêm sau

1. **Đánh giá hiện trạng chung**

Hiện tại, các hoạt động chuyên môn của các Đơn vị đang được tổ chức trên nhiều phần mềm phân tán/phần mềm theo ngành dọc (chưa có cơ chế chia sẻ ngang với Quốc hội); tuy nhiên, các dữ liệu này đã được số hóa và sẵn sàng (dưới dạng CSDL hoặc file excel).

Khảo sát viết thêm sau

Bên cạnh đó, khả năng liên thông dữ liệu của các phần mềm/CSDL này đã và đang được triển khai (trích xuất định dạng file flat hoặc chia sẻ công cụ kết nối trung gian (API - Application Programming Interface).

Như vậy, các dữ liệu đầu vào cho Kho dữ liệu dùng chung tập trung của Quốc hội đã đảm bảo và sẵn sàng, cũng như các phần mềm đang được hoàn thiện phục vụ cho việc chia sẻ dữ liệu, liên thông vào Kho dữ liệu dùng chung tập trung của văn phòng chính phủ.

Khảo sát viết thêm sau

## Sự cần thiết đầu tư

Kho dữ liệu dùng chung tập trung của Quốc hội là một cấu phần vô cùng quan trọng trong hệ thống báo cáo quản trị thông minh (BI). Đây là kho tổng hợp nhiều dạng dữ liệu khác nhau để từ đó chuẩn hóa và cung cấp thông tin cho hệ thống BI tổng hợp dữ liệu, xử lý thành báo cáo, dự đoán những biến động trong tương lai và huấn luyện các mô hình AI hỗ trợ các cán bộ trong nghiệp vụ.

Hiện nay, các ngành từ Trung ương đến địa phương thực hiện triển khai nhiều hệ thống quản lý chuyên ngành để phục vụ công tác chuyên môn và công tác chỉ đạo điều hành của ngành nên lượng thông tin của một đối tượng có thể đến từ nhiều nguồn khác nhau, dẫn đến việc dữ liệu không đồng nhất, bên cạnh đó khối lượng dữ liệu sẽ tăng nhanh theo thời gian. Điều này dẫn đến quá tải và giảm đáng kể khả năng truy xuất dữ liệu.

Vì vậy cần phải thực hiện dự án này để kết nối, tổng hợp thông tin dữ liệu trên các hệ thống cơ sở dữ liệu để lưu trữ dữ liệu nghiệp vụ, đang là cấp bách trên từng cập bộ ban ngành

## Đánh giá sự phù hợp với quy hoạch có liên quan theo quy định của pháp luật về quy hoạch

Năm 2021, thực hiện đầu tư nâng cấp hạ tầng cho Trung tâm tích hợp dữ liệu của Quốc hội sẽ đảm bảo hạ tầng cho việc triển khai dự án này. Bên cạnh đó, Văn phòng quốc hội cũng đang nghiên cứu, đề xuất thêm giải pháp thuê hạ tầng ở các nhà cung cấp dịch vụ để đảm bảo và tăng cường thêm khả năng xử lý các hệ thống và đặc biệt là trong vấn đề đảm bảo an toàn thông tin cho các hệ thống…

Khảo sát viết thêm sau

## Phân tích, xác định mục tiêu, nhiệm vụ, kết quả đầu ra của dự án; phân tích, lựa chọn quy mô hợp lý; xác định phân kỳ đầu tư; lựa chọn hình thức đầu tư

### Mục tiêu đầu tư

Chuẩn hóa việc khai thác dữ liệu tích hợp từ các nguồn dữ liệu chuyên ngành của các hệ thống thông tin thuộc các Đơn vị, phòng ban. Dữ liệu được lưu trữ và quản lý tại kho dữ liệu dùng chung, để cung cấp các báo cáo thống kê, phân tích dự báo, mô phỏng thông tin kinh tế xã hội của tỉnh, phục vụ cán bộ công chức trong công tác chuyên môn nghiệp vụ và phục vụ lãnh đạo trong công tác chỉ đạo điều hành.

Dữ liệu sau khi xử lý, lưu trữ hình thành kho dữ liệu dùng chung của Quốc hội, các hệ thống thông tin khác có thể thông qua trục LGSP và NDXP để kết nối và khai thác sử dụng. Đây là nền tảng hướng đến mục tiêu xây dựng chính quyền điện tử và tiến tới chính quyền số.

Kho dữ liệu dùng chung tập trung cho phép người dùng cấp quản lý ra quyết định tốt hơn thông qua việc tổng hợp và phân tích dữ liệu thông qua tương tác với nhiều dạng dữ liệu khác nhau. Ngoài ra, Kho dữ liệu dùng chung tập trung các hệ thống phần mềm, dữ liệu máy học (machine learning) của Quốc hội, được dùng cho báo cáo, data mining, AI và phân tích thống kê. Kho dữ liệu dùng chung tập trung có khả năng lưu trữ tới hàng trăm Terabyte thậm chí hàng Petabyte. Kho dữ liệu dùng chung tập trung được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau, sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước.

### Kết quả của dự án

* Chuẩn hóa các dữ liệu đến từ nhiều nguồn, cho phép lưu trữ các dữ liệu thô và dữ liệu đã được xử lý, được chia ra bởi theo các bộ tiêu chí, nghiệp vụ, phân quyền.
* Xây dựng, hoàn thiện Nền tảng Lưu trữ dữ liệu dùng cho khai thác, phân tích trên Kho dữ liệu dung chung tập trung trên nền tảng dữ liệu lớn, đảm bảo hiệu năng khai thác, chất lượng dữ liệu và tính an toàn, bảo mật cho dữ liệu.
* Xây dựng, hoàn thiện phần mềm thu thập, liên kết dữ liệu đảm bảo khả năng xử lý dữ liệu đa nguồn, đa cấu trúc và định kỳ thu thập dữ liệu mới, dữ liệu thay đổi theo cấu hình quy định sẵn. Các tác vụ thu thập, liên kết dữ liệu vận hành ổn định trong thời gian dài.
* Xây dựng, hoàn thiện phần mềm phân tích biểu diễn dữ liệu cung cấp các công cụ truy vấn và khai thác dữ liệu theo nhu cầu, xây dựng các báo cáo trực quan sinh động và phù hợp về thẩm mỹ. Phần mềm cho phép phân quyền cho cá nhân, tổ chức bất kỳ để khai thác các dữ liệu đó.
* Xây dựng, hoàn thiện các mô hình AI cung cấp các dịch vụ tổng hợp, khai thác dữ liệu theo yêu cầu.
* Xây dựng, hoàn thiện hệ thống quản trị tài nguyên, hệ thống các phần mềm, hệ thống backup…
* Xây dựng, hoàn thiện phần mềm nền tảng giao diện dữ liệu cho phép chia sẻ dữ liệu thông qua Cổng dữ liệu mở và các API chia sẻ qua hệ thống quản trị API. Dữ liệu chia sẻ đảm bảo an toàn thông tin và an toàn dữ liệu theo nguyên tắc cung cấp dữ liệu mở của cơ quan nhà nước - Nghị định số 47/2020/NĐ-CP ngày 09/4/2020 của Chính phủ về quản lý, kết nối và chia sẻ dữ liệu số của cơ quan nhà nước.
* Dữ liệu sau khi xử lý và lưu trữ có thể kết nối với hệ thống trục liên thông (LGSP), làm cơ sở tích hợp và kết nối dữ liệu vào hệ thống NDXP của Quốc gia.

### Phân tích, lựa chọn quy mô

Trong quá trình triển khai dự án, Đơn vị tư vấn sẽ tham mưu cùng Văn phòng quốc hội thành lập Tổ triển khai dự án (bao gồm cả các đơn vị, địa phương liên quan) để phối hợp trong quá trình đưa ra các yêu cầu về việc xử lý dữ liệu và họ cũng sẽ thực hiện sử dụng, đánh giá để kho dữ liệu đáp ứng yêu cầu và càng hoàn thiện hơn.

* Xử lý dữ liệu từ kho dữ liệu: xử lý định hướng các dữ liệu phục vụ báo cáo thống kê, cung cấp thông tin có trong kho dữ liệu dùng chung tập trung của tỉnh.
* Lưu trữ dữ liệu đã xử lý: quản lý và lưu trữ các dữ liệu sau khi xử lý nhằm phục vụ các bước tiếp theo.
* Khai phá dữ liệu: hỗ trợ các tính năng giúp khai phá các dữ liệu đã xử lý và lưu trữ phục vụ các mục đích khác nhau.
* Trực quan dữ liệu: hỗ trợ các mô hình, biểu đồ trong việc trực quan hóa các số liệu.
* Mô phỏng dự báo dữ liệu theo mô hình: định nghĩa, quản lý, thực thi các mô hình mô phỏng, dự báo.
* Xây dựng mô hình ngôn ngữ lớn phục vụ dữ liệu chuyên ngành (AI, Machine Learning).
* Dịch vụ cung cấp dữ liệu: hình thành, quản lý các dịch vụ khai thác dữ liệu dựa trên dữ liệu đã khai phá.
* Quản trị xử lý và hình thành dịch vụ dữ liệu: cung ấp các tính năng cho phép kết nối trực tích hợp sử dụng các dịch vụ dữ liệu; quản trị phân quyền người sử dụng; cá nhân hóa thông tin; quản lý danh mục đặc thù; tìm kiếm thông tin; thông báo, cảnh báo; quản lý truy vết; … liên quan phân hệ xử lý và hình thành dịch vụ dữ liệu.

### Hình thức đầu tư:

Đầu tư mới.

## Phân tích các điều kiện tự nhiên, điều kiện kinh tế - kỹ thuật, lựa chọn địa điểm đầu tư

### Các điều kiện tự nhiên, điều kiện kinh tế - kỹ thuật

#### Điều khiên tự nhiên

#### Điều kiện kinh tế - kỹ thuật

Dự án nhằm xây dựng kho dữ liệu dùng chung của tỉnh, đáp ứng nhu cầu lưu trữ, tổng hợp và xử lý thông tin hướng tới xây dựng nền hành chính hiện đại.

Khảo sát viết thêm sau

### Lựa chọn địa điểm đầu tư:

#### Trung tâm Tích hợp dữ liệu VPQH

# NỘI DUNG CHÍNH CỦA BÁO CÁO

## Đánh giá sự tuân thủ Khung Kiến trúc Chính phủ điện tử Việt Nam, Kiến trúc Chính phủ điện tử cấp bộ

Dự án đảm bảo tuân thủ Khung Kiến trúc Chính phủ điện tử Việt Nam (phiên bản 2.0) theo Quyết định số 2323/QĐ-BTTT ngày 31/12/2019 của Bộ Thông tin và truyền thông, Các văn bản ….

Khảo sát viết thêm sau

## Phân tích, lựa chọn phương án công nghệ, kỹ thuật, thiết bị

Mô hình kho dữ liệu dùng chung tập trung của tỉnh được tham chiếu đến mô hình Nền tảng đô thị thông minh – căn cứ theo Công văn số 4176/ BTTTT-THH của Bộ Thông tin và Truyền thông về việc hướng dẫn triển khai thí điểm dịch vụ đô thị thông minh cho các Tỉnh/Thành phố trực thuộc trung ương….

Khảo sát viết thêm sau

Việc thiết kế và quản lý kho dữ liệu cần được cân nhắc đến các điều kiện quản lý hệ thống, khả năng kiểm soát và phân bổ tài nguyên linh hoạt, có giá thành thấp và độ an toàn cao. Ngoài ra, trong những trường hợp xảy ra các vấn đề ảnh hưởng đến hạ tầng, cần phải đảm bảo tính sẵn sàng cao, khả năng co giãn, có cơ chế dư thừa để có tránh thời gian chết hệ thống quá lâu, ảnh hưởng đến việc sử dụng dữ liệu.

Kiến trúc kho dữ liệu dùng chung tập trung của Trung ương sẽ kế thừa từ khung tham chiếu ICT phát triển đô thị thông minh, cũng như tham khảo hướng dẫn nền tảng đô thị thông minh do Bộ Thông tin và Truyền thông ban hành. Giải pháp công nghệ của Kho dữ liệu dùng chung tập trung của Quốc hội sẽ phải đáp ứng tiêu chí kỹ thuật của các nhóm chức năng sau:

**Khối chức năng thu thập dữ liệu:** cho phép thu thập đa nguồn dữ liệu, từ các cơ sở dữ liệu có cấu trúc như SQL, CSV đến các loại chuẩn dữ liệu phi và bán cấu trúc đều được thu thập sang CSDL dùng chung dưới dạng thô, không được chỉnh sửa. Tất các các dữ liệu sẽ được lưu trong kho dữ liệu dạng Object Storage như Minio hoặc chuẩn S3. Ngoài ra, các kho dữ liệu sẽ được quản lý theo các bộ tiêu chí được đề ra và có khả năng cho phép chia sẻ dữ liệu.

**Khối chức năng xử lý dữ liệu:** xử lý các dữ liệu thô được thu thập một cách kịp thời, chính xác. Các chức năng xử lý dữ liệu bao gồm các bước chuẩn hóa các chuẩn dữ liệu, chuẩn lưu trữ và đưa vào các kho lưu trữ tuân thủ theo nghiệp vụ đề ra. Các loại dữ liệu theo các dạng cấu trúc hoặc bán cấu trúc, cần phải chuyển đổi sang các dạng có khả năng lưu trữ nén tốt, và tốc độ truy vấn cao. Ngoài ra, để giảm thiểu thời gian xử lý dữ liệu, các chức năng trên nên sử dụng triệt để tài nguyên của hệ thống máy chủ, cho phép chạy nhiều chức năng cùng một lúc. Để tránh việc xảy ra các trường hợp dữ liệu trùng hoặc các dữ liệu bị xung đột, ảnh hưởng đến độ đáng tin cậy của dữ liệu, cần có một bộ công cụ để giám sát và nếu như cần thiết, có thể thực hiện khắc phục dữ liệu trực tiếp. Tất cả các dữ liệu được chuyển đổi cần phải có các cơ chế để sao lưu, không những để tránh việc mất dữ liệu đã được làm sạch, mà còn có thể sử dụng để giảm tải trong việc truy vấn dữ liệu.

**Khối chức năng quản lý, huấn luyện và sử dụng mô hình AI:** có khả năng cho phép sử dụng các tệp dữ liệu được chuyển đổi, để phục vụ cho việc sử dụng các Model AI. Hệ thống máy chủ cần được tích hợp để sử dụng, giám sát các tài nguyên GPU và phải đảm bảo là các tài nguyên GPU này sẽ chỉ được sử dụng cho môi trường huấn luyện AI.

**Khối chức năng quản lý dữ liệu:** có các cơ chế phục vụ việc quản lý dữ liệu, bao gồm việc giám sát cấu trúc, cơ chế xử lý, phân quyền và giám sát các audit log thao tác dữ liệu.

**Khối chức năng trực quan dữ liệu:** cho phép các chức năng có tính chất trình diễn dữ liệu như Dashboard, báo cáo hoặc AI có thể truy vấn được dữ liệu đã được chuyển đổi. Có thể sử dụng SQL lên dữ liệu đã được chuyển đổi trong các chức năng phân tích dữ liệu. Có các tính năng cho phép cấu hình các mô hình trực quan, dashboard và xây dựng các mô hình báo cáo chi tiết. Có thể phần quyền người dùng, giới hạn việc truy cập các mô hình dựa theo yêu cầu nghiệp vụ của từng cán bộ.

**Khối chức năng quản lý chia sẻ dữ liệu dùng chung:** xây dựng một bộ API truy vấn dữ liệu trong kho, có khả năng phân quyền và quản lý quyền truy vấn thông tin. Đảm bảo an toàn thông tin trong việc truy vấn dữ liệu, có các biện pháp xác thực đơn vị truy vấn và lưu log truy vấn để có thể đối soát.

Ngoài ra hệ thống cần có **Khối chức năng trực quan quản trị, cấp phát tài nguyên hệ thống:** cho phép các chức năng quản lý, cấp phát tài nguyên của hệ thống. Các hệ thống tài nguyên có báo cáo trực quan tình tình trạng sử dụng.

### Tiêu chí lựa chọn hệ thống quản lý tài nguyên máy

Với sự phát triển mạnh mẽ trong việc quản lý hạ tầng hiện nay, có rất nhiều lựa chọn về kiểu kiến trúc hạ tầng như on-premise, VM, hoặc Cloud. Mỗi nền tảng đều có những ưu và nhược điểm khác nhau và sẽ ảnh hưởng đến cách vận hành hệ thống. Ngoài ra, Do đặc thù dữ liệu của Quốc hội, việc xây dựng kho dữ liệu sẽ được đặt tại các địa điểm trực thuộc quản lý của đơn vị.

### Phân tích lựa chọn công nghệ quản lý hệ thống và tài nguyên máy

#### Cơ sở hạ tầng bare-metal

Bare-metal, hay còn được biết đến là các máy chủ vật lý, là một thuật ngữ chuyên dụng để chỉ một máy tính với cấu hình có công suất lớn hơn so với các máy tính thông thường, hoạt động với một hệ điều hành riêng và có các tính năng đa dạng, đặc biệt như kết nối các hệ thống máy chủ, thiết bị khác với nhau hoặc là nơi lưu trữ dữ liệu của đơn vị. Một số các đặc điểm của cơ sở hạ tầng này như sau:

* Toàn quyền sử dụng tài nguyên có trong máy, hiệu suất hoạt động cao.
* Có tính an toàn bảo mật cao.

Tuy nhiên, việc sử dụng các máy chủ vật lý cũng có các nhược điểm:

* Không tận dụng được toàn bộ tài nguyên của máy.
* Tính sẵn sàng thấp, trong trường hợp hệ thống có vấn đề phải chờ được xử lý mới có thể sử dụng lại.
* Chi phí mở rộng cao.

#### Cơ sở hạ tầng Private Cloud-server

Máy chủ ảo hay còn được gọi là Cloud Server hoặc máy chủ điện toán đám mây, đây là một dịch vụ máy chủ riêng được xây dựng trên hạ tầng máy chủ vật lý chuyên dụng và triển khai trên nền tảng công nghệ ảo hoá điện toán đám mây. Cloud Server cũng sẽ bao gồm các thông số cấu hình cơ bản như CPU, RAM, dung lượng ổ cứng và địa chỉ IP tương tư như một máy chủ vật lý chuyên dụng.

Cơ chế Private Cloud ở đây được hiểu là lưu trữ đám mây riêng, được quản lý bởi doanh nghiệp hoặc đơn vị. Thông thường mạng máy chủ cho lưu trữ đám mây riêng sẽ không rộng lớn như mạng máy chủ công cộng. Điều này đặc biệt đúng nếu bạn đang sử dụng mạng đám mây riêng của mình trong nội bộ. Một mạng lưu trữ đám mây riêng thông thường sẽ hoạt động theo cụm. Bạn sẽ có nhiều máy chủ khác nhau cùng chạy các phiên bản website cùng một lúc và các tài nguyên có thể được lấy ra từ nhiều nguồn khác nhau khi cần.

Private Cloud thừa hưởng các lợi thế của kiến trúc Cloud, bao gồm:

* Tận dụng triệt để tài nguyên máy.
* Dễ dàng cấu hình tài nguyên phân bổ.
* Có tính sẵn sàng cao, hệ thống hoạt động bình thường trong trường hợp một máy trong cụm không sử dụng được.
* Có thể co giãn phân tài nguyên dựa theo mức độ sử dụng
* Có thể mở rộng tài nguyên hệ thống mà không ảnh hưởng hệ thống đang chạy

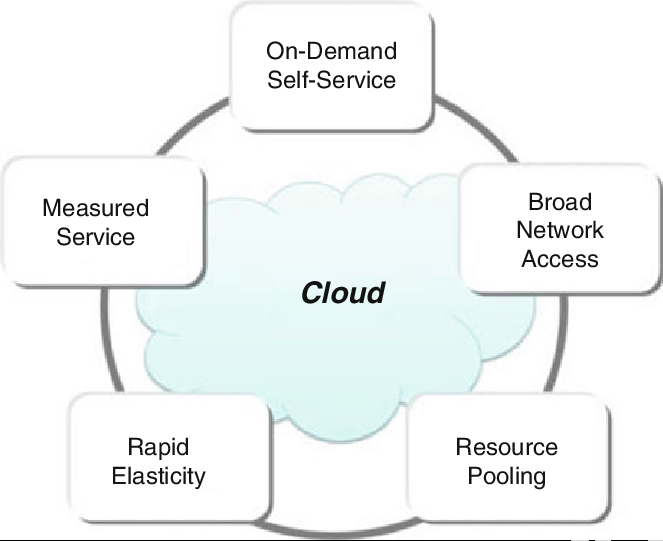
Ngoài ra Private Cloud có thể được xây dựng ngay trên cơ sở hạ tầng Bare-metal có sẵn, không những đảm bảo về mức độ bảo mật do thừa hưởng hệ thống bảo mật có sẵn của hệ thống, mà đơn vị có thể kiểm soát được chi phí và nguồn lực bảo trì có sẵn của hệ thống máy vật lý.

Tuy nhiên, một số hạn chế khi sử dụng Private Cloud bao gồm:

* Đòi hỏi nhân sự chuyên môn cao khi vận hành Private Cloud
* Chỉ có thể sử dụng được tài nguyên dựa trên máy vật lý có sẵn
* Tương tự như các máy vật lý, chi phí đầu tư phần cứng cho Private Cloud đắt

#### Lựa chọn giải pháp quản lý hệ thống

Để tận dụng được toàn bộ hệ thống tài nguyên bare-metal có sẵn, và lợi dụng được sự linh hoạt trong việc phân bổ linh hoạt của Private Cloud, cần phải cân nhắc đến việc lựa chọn công cụ thích hợp để thỏa mãn các chức năng cần có của một Cloud.



1. Năm tính năng chính của Cloud

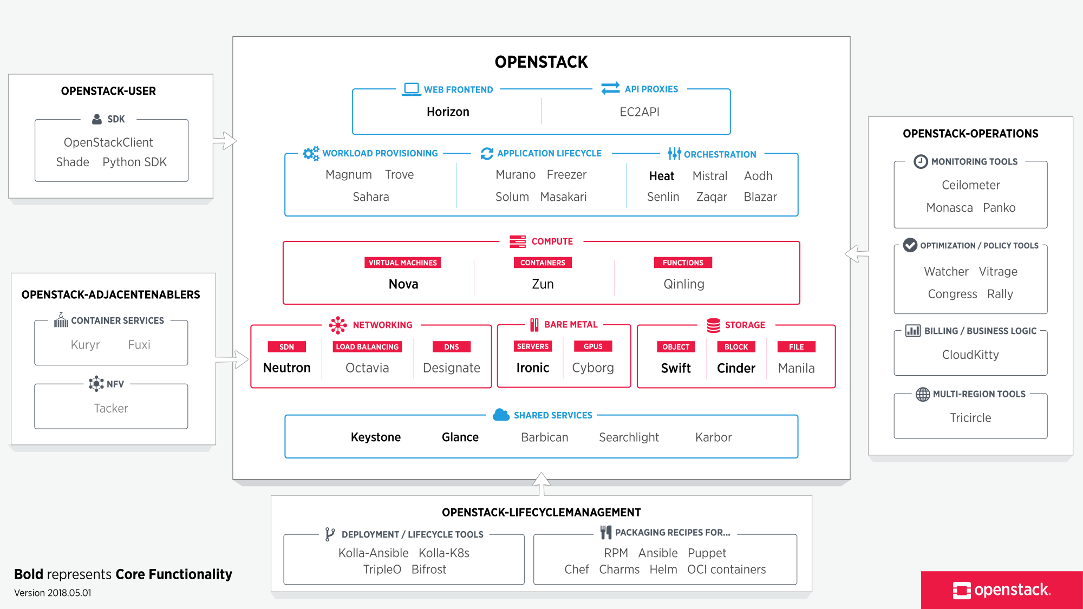
Các chức năng này bao gồm:

* Tự phục vụ: Quản trị viên có thể chủ động tự phân bổ, giám sát và quản lý tài nguyên khi cần.
* Chia sẻ tài nguyên: Các tài nguyên IT (v.d. mạng, máy chủ, lưu trữ, ứng dụng và dịch vụ) được chia sẻ trên nhiều ứng dụng và không bị ràng buộc về mặt hệ thống. Có thể cấu hình cho nhiều ứng dụng sử dụng trên cùng một máy vật lý.
* Độ linh hoạt mở rộng: Các dịch vụ điện toán nên có các tài nguyên IT có thể mở rộng hoặc thu hẹp nhanh và theo yêu cầu. Người dùng khi cần có thể tự động bổ sung thêm tài nguyên ngay lập tức và thu hẹp lại khi nhu cầu giảm.
* Các dịch vụ được thống kê: Việc sử dụng tài nguyên được theo dõi cho từng ứng dụng và người dùng. Công việc này cần thiết cho nhiều yếu tố như giám sát hiệu quả tài nguyên.
* Các dịch vụ trên cloud có thể cho phép được truy cập thông qua mạng và trên nhiều loại ứng dụng khác nhau.

Hiện nay trên thị trường, hầu hết các giải pháp private cloud đều được các hãng lớn như AWS, Cisco, Google, Microsoft hỗ trợ, nhưng đều là các giải pháp host ở trên máy của bên cung cấp, không thuộc quyền quản lý của người sử dụng. Chính vì vậy, cần phải có giải pháp phần mềm để cấu hình Private Cloud trên máy được quản lý của người dùng. Dưới đây là mô tả tổng quan của giải pháp Private Cloud cho OpenStack và OpenShift.

##### OpenStack

OpenStack là giải pháp mã nguồn mở cho tính toán đám mây. Giải pháp này hầu như được triển khai dưới dạng Infrastructure-as-a-service (IAAS), cho phép quản lý và sử dụng các tài nguyên dùng chung giữa các máy vật lý. Nền tảng sử dụng các chức năng được liên kết chặt chẽ với nhau, không giới hạn phần cứng đến từ nhà cung cấp nào, trên máy vật lý đặt trên hạ tầng của người dùng. Người dùng có thể quản lý thông qua dashboard được cung cấp sẵn, các công cụ gõ lệnh hoặc thông qua các RestAPI.OpenStack hiện đang được sử dụng ở trong các viện nghiên cứu hàng đầu trên thế giới như NASA hay CERN.



1. Sơ đồ các dịch vụ trên OpenStack

Dưới đây là một vài những tính năng của OpenStack:

* Hỗ trợ kho lưu trữ dạng S3 với Swift.
* Hỗ trợ lưu trữ dạng Block .
* Công cụ quản lý điện toán Nova cho phép cấu hình và phân bổ tài nguyên máy.
* Công cụ quản lý IP và mạng Neutron.
* Công cụ điều phối tài nguyên và hệ thống như Heat.
* Công cụ xử lý dữ liệu Sahara, tận dụng công nghệ và xử dụng tài nguyên của Hadoop và Spark.

Lợi thế của OpenStack:

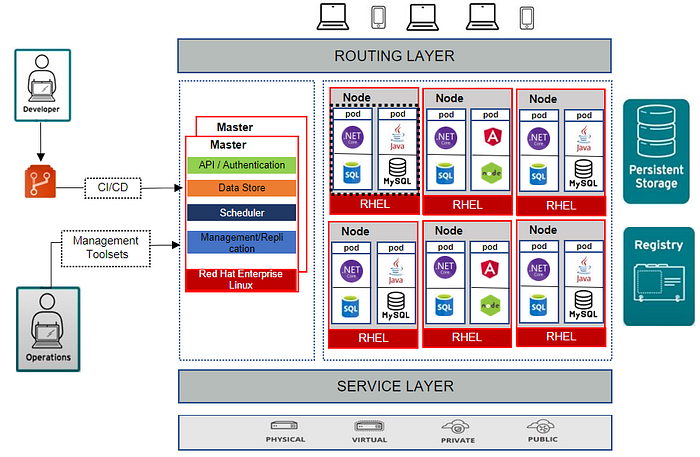
* OpenStack là một dự án mã nguồn mở, không có chi phí License đi kèm.
* OpenStack được thiết kế để dễ dàng mở rộng, cho phép bổ sung thêm tài nguyên vào máy ảo khi cần. Ngoài ra hệ thống được thiết kế để có thể dễ dàng bổ sung thêm tài nguyên vào trong cụm mà không bị ảnh hưởng đến các tiến trình đang chạy.
* Có thể tinh chỉnh hệ thống phụ thuộc theo nhu cầu sử dụng.
* Cồng đồng phát triển lớn, có thể dễ dàng tìm kiếm và sử dụng các tài liệu phát triển và các ứng dụng bổ trợ hệ thống.
* Có các phương pháp bảo mật cài đặt sẵn trong hệ thống, bao gồm quản lý và nhận diện tài khoản, mã hóa và cách ly đường truyền.

Tuy có những lợi thế như trên, hệ thống có một vài điểm bất lợi sau:

* OpenStack có rất nhiều cấu phần, có thể sẽ gây khó khăn trong việc triển khai.
* Cần có mỗi đội ngũ chuyên gia để quản lý hiệu quả.

##### Openshift

Openshift là một nhóm sản phẩm phục vụ triển khai các dịch vụ theo dạng quản lý container được phát triển bởi RedHat. Đây là nhóm sản phẩm phục vụ cho cơ chế PaaS của Cloud (Platform as a Service), tức là người dùng sẽ chỉ cần tập trung vào việc phát triển phần mềm và triển khai hệ thống ứng dụng thông qua các container.



1. Sơ đồ hoạt động của một cụm OpenShift

Dưới đây là một vài đặc điểm lợi thế của Openshift:

* Hoạt động dưới dạng PaaS, Openshift hỗ trợ việc triển khai và quản lý ứng dụng.
* Có thể cấu hình cho phép tăng số lượng ứng dụng để cân bằng tải.
* Được hỗ trợ sử dụng và cập nhật bởi hãng RedHat.
* Hỗ trợ việc xây dựng các lộ trình tích hợp và phát triển, hỗ trợ các hoạt động DevOps.
* Có hệ thống bảo mật có sẵn, dễ dàng quản trị các thông tin đường truyền vào ra hệ thống.
* Có nhiều tính năng phục vụ quản lý hệ thống, ứng dụng và người dùng.

Nhược điểm:

* OpenShift có thể sẽ rất khó để quản lý, nếu như thiếu các chuyên gia có kinh nghiệm về DevOps.
* Là sản phẩm thương mại của RedHat, người dùng sẽ phải trả tiền License nếu như cần sử dụng các chức năng quan trọng của RedShift.

### Tiêu chí lựa chọn giải pháp công nghệ

Theo sự phát triển và xu hướng công nghệ hiện nay, có nhiều lựa chọn về công nghệ phần mềm cho việc phát triển các hệ thống thông tin lớn. Tuy nhiên có thể chia thành các nền tảng mã nguồn mở (như Minio, Apache Airflow,Nifi, Spark...) và nền tảng mã nguồn đóng (như PowerBI, Tableau).

Mỗi nền tảng đều có những ưu nhược điểm khác nhau, nên để đánh giá được thế mạnh của các hệ thống thông tin được xây dựng dựa trên 02 dòng nền tảng công nghệ trên cần nhìn vào khả năng ứng dụng của chúng và các ưu nhược điểm của từng công nghệ. Các tiêu chí đánh giá bao gồm như sau:

* Sự hỗ trợ từ cộng đồng.
* Sự hỗ trợ từ nhà cung cấp nền tảng.
* Có triển khai được trên nhiều môi trường hay không?
* Các công cụ hỗ trợ để phát triển ứng dụng có nhiều hay không?
* Có dễ dàng lập trình và cài được trên nhiều môi trường hay không?
* Chi phí đầu tư nhân lực để phát triển có lớn không?
* Công nghệ có dễ dàng bảo trì, sửa đổi hay không?
* Công nghệ có bảo mật hay không?
* Tốc độ của ứng dụng có cao hay không?
* Ứng dụng xây dựng có phù hợp và xử lý được một hệ thống thông tin phức tạp không?
* Ứng dụng có dễ dàng kết nối với các hệ thống phần mềm đã có sẵn hay không?
* Có phù hợp với các tài nguyên sẵn có hay không (các hệ thống sẵn có, nhân lực quản trị, bản quyền phần mềm...)?

Vì vậy, việc lựa chọn công nghệ nào cần được đánh giá một cách toàn diện dựa trên các tiêu chí nêu trên.

| **Tiêu chí** | **Nguồn mở** | **Nguồn đóng** |
| --- | --- | --- |
| Sự hỗ trợ từ cộng đồng? | Đáp ứng tốt hơn | Đáp ứng kém |
| Sự hỗ trợ từ nhà cung cấp nền tảng? | Đáp ứng tốt hơn | Đáp ứng kém |
| Có triển khai được trên nhiều môi trường hay không? | Đáp ứng | Không đáp ứng |
| Các công cụ hỗ trợ để phát triển ứng dụng có nhiều hay không? | Đáp ứng | Đáp ứng |
| Có dễ dàng lập trình và cài được trên nhiều môi trường hay không? | Đáp ứng | Không đáp ứng |
| Chi phí đầu tư nhân lực để phát triển có lớn không? | Chi phí vừa phải | Chi phí lớn |
| Công nghệ có dễ dàng bảo trì, sửa đổi hay không? | Đáp ứng | Đáp ứng kém |
| Công nghệ có bảo mật hay không? | Đáp ứng | Đáp ứng |
| Tốc độ của ứng dụng có cao hay không? | Đáp ứng | Đáp ứng |
| Ứng dụng xây dựng có phù hợp và xử lý được một hệ thống thông tin phức tạp không? | Đáp ứng | Đáp ứng |
| Ứng dụng có dễ dàng kết nối với các hệ thống phần mềm đã có sẵn hay không? | Đáp ứng | Đáp ứng |
| Có phù hợp với các tài nguyên sẵn có hay không (các hệ thống sẵn có, nhân lực quản trị, bản quyền phần mềm, ...)? | Đáp ứng | Đáp ứng |

Việc lựa chọn công nghệ để xây dựng hệ thống phải đáp ứng các yêu cầu chung sau đây:

* Giải pháp công nghệ phải mang lại sự thành công của dự án, điều đó có nghĩa là hệ thống phải hoạt động tốt đáp ứng mục tiêu đề ra.
* Giải pháp công nghệ lựa chọn phải đảm bảo hệ thống hoạt động ổn định, liên tục, lâu dài và tính bảo mật cao.
* Đáp ứng khả năng tăng trưởng dữ liệu và thay đổi mở rộng nâng cấp trong tương lai.
* Chi phí bản quyền ít tốn kém.

Qua các tiêu chí trên, việc lựa chọn công nghệ triển khai Kho dữ liệu dùng chung tập trung cho Văn Phòng Quốc hội sẽ ưu tiên lựa chọn các công nghệ mã nguồn mở để tiết kiệm chi phí bản quyền. Các công nghệ mã nguồn mở ban đầu có thể vẫn còn thô sơ và chưa đầy đủ các tính năng để đáp ứng mục tiêu đề ra, trong phạm vi dự án, các công nghệ này sẽ được nâng cấp và phát triển mới để tiến đến đáp ứng các mục tiêu của dự án, đảm bảo hệ thống hoạt động lâu dài và có khả năng mở rộng tốt trong tương lai.

### Phân tích lựa chọn công nghệ cho các khối chức năng

#### Khối chức năng thu thập liên kết

Cung cấp các cơ chế để thu thập dữ liệu từ các hệ thống thu thập dữ liệu hoặc liên kết với các dịch vụ cung cấp dữ liệu từ bên ngoài. Lớp chức năng thu nhận/liên kết cung cấp các chức năng cơ bản như sau:

* Tích hợp thông tin từ những nguồn dữ liệu khác nhau.
* Cung cấp thông tin tới khối các chức năng dữ liệu. Thông tin này có khả năng định dạng phù hợp với việc xử lý trí tuệ nhân tạo.
* Chức năng thu nhận/liên kết tách biệt với mạng thông tin và điều khiển.

Hiện nay trên thế giới đã có nhiều công cụ hỗ trợ chức năng thu thập liên kết đáp ứng các yêu cầu phức tạp về dữ liệu. Tuy nhiên, với những nhu cầu phức tạp mà cần đến nhiều tri thức ngành, các công cụ đóng gói sẵn đang không đáp ứng được những yêu cầu tùy biến theo nghiệp vụ của người dùng. Ngược lại, các công cụ thiết kế mở có khả năng tùy biến đa dạng hơn nhiều. tuy phải đánh đổi bằng việc giảm mức độ thân thiện với người dùng. Dưới đây là bản đánh giá hai trong số nhiều công cụ thuộc hai trường phái đóng gói sẵn và thiết kế mở, theo thứ tự là Apache Nifi và Apache Airflow.

##### Apache Nifi

Apache nifi được sử dụng để tự động hóa và kiểm soát các luồng dữ liệu giữa các hệ thống. Nó cung cấp cho chúng ta một giao diện trên nền web mà có thể thu thập, xử lý, phân tích dữ liệu.

NiFi được biết đến với khả năng xây dựng luồng chuyển dữ liệu tự động giữa các hệ thống. Đặc biết là hỗ trợ rất nhiều kiểu nguồn và đích khác nhau như: hệ thống file, cơ sở dữ liệu quan hệ, phi quan hệ, …. Ngoài ra Nifi sẽ hỗ trợ các thao tác với dữ liệu như lọc chỉnh sửa, thêm bớt nội dung.

Diagram

Description automatically generated

1. Các thành phần chính của Nifi

Các đặc điểm nổi bật của Nifi:

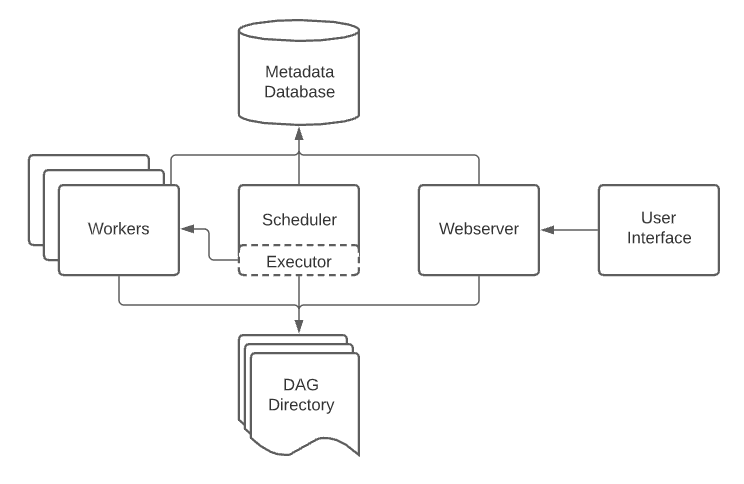
* Khả năng quản lý nguồn dữ liệu:
  + Đảm bảo tính an toàn: Mỗi đơn vị dữ liệu trong luồng sẽ được lưu dưới dạng 1 object là flowfile. Nó sẽ ghi lại thông tin về khối dữ liệu đang được xử lý ở đâu, di chuyển đi đâu… Provenance Repo được sử dụng để lưu các flowfile giúp ta có thể truy vết.
  + Data Buffering: Giải quyết vấn đề tốc độ dữ liệu ghi vào nhanh hơn dữ liệu đọc giữa hai khối xử lý. Dữ liệu này sẽ được lưu ở RAM, nếu quá một ngưỡng thì sẽ được lưu xuống ổ cứng.
  + Thiết lập độ ưu tiên: Trong xử lý dữ liệu có những dữ liệu ta phải xử lý ưu tiên.
  + Đánh đổi tốc độ và khả năng chịu lỗi: Nifi hỗ trợ cài đặt để cân bằng 2 yếu tố này.
* Sử dụng dễ dàng:
  + Nifi hỗ trợ UI cho việc xây dựng luồng dữ liệu.
  + Tính tái sử dụng khi ta có thể lưu được các luồng dữ liệu thành 1 template.
  + Theo dõi trực quan lịch sử.

##### Apache Airflow

Apache Airflow là một open-source project được đội ngũ kỹ sư Airbnb thiết kế từ năm 2014 và được Apache Foundation hỗ trợ từ năm 2016. Apache Airflow là một trong những công nghệ phổ biến hiện nay cho workflow management systems (WMS) được các Data Engineers dùng tới để quản lý quá trình tạo đường dẫn thông tin (data pipeline) trong cơ sở dữ liệu cho việc xử lý và phân tích thông tin. Quá trình tạo data pipeline được Airflow thể hiện bằng directed acyclic graph (DAG) do từng công việc (task) thường phụ thuộc những tasks khác để lấy dữ liệu đã được xử lý.

Cấu trúc Airflow bao gồm 4 thành phần chính:

* Metadata CSDL: là CSDL chính để chứa thông tin về hoạt động của từng task trong hệ thống.
* Scheduler: là hệ thống lập kế hoạch chính để quyết định task nào sẽ được xử lý tiếp theo.
* Executor: là hệ thống message queue để cho Scheduler kết nối với từng bộ phận làm việc (worker) cho việc thi hành từng task.
* Workers: là bộ phận làm việc chính để chạy task được thành lập từ Data Engineer.



1. Cấu trúc Airflow

Airflow có những ưu điểm:

* Nhẹ, tách bạch việc quản lý task và chạy task.
* Nhiều provider, kết nối được tới nhiều data source.
* Core Workflow sử dụng Python, dễ học, dễ sử dụng.
* Giao diện dễ sử dụng.

##### Lựa chọn giải pháp

Apache Nifi có nhiều ưu điểm về giao diện và xử lý luồng dữ liệu liên tục (Streaming), tuy vậy Nifi gặp nhiều khó khăn trong việc xử lý các tác vụ lớn và phức tạp. Nguyên nhân một phần là do giao diện kéo thả tuy giúp đơn giản hóa nghiệp vụ nhưng thường không tái sử dụng được ở các môi trường, bài toán khác nhau. Đối với các tác vụ phức tạp cần nhiều bước truy vấn xử lý, Nifi hầu như không tối ưu.

Ngược lại, Airflow giống như một nền tảng quản lý tác vụ hơn và nó tách biệt rõ ràng giữa quản lý và chạy các tác vụ. Mô hình này cho phép người dùng dễ dàng xây dựng các luồng xử lý lớn, phức tạp mà không gặp khó khăn trong việc quản lý chúng, đặc biệt là với dữ liệu lớn. Khả năng lập lịch của Airflow được tối ưu hóa cho các tác vụ cần phải chạy theo chu kỳ trong thời gian dài. Ngoài ra, mã nguồn mở Apache Airflow cũng là nền tảng có cộng đồng sử dụng lớn hơn so với Nifi, và nền tảng này cũng đã được chứng minh trong thực tế là có tính tương hợp cao với các phần mềm, nền tảng khác.

Qua khảo sát, các dữ liệu của Quốc hội hầu hết không phải dữ liệu cần được cập nhật liên tục, cũng như cần được bóc tách phức tạp để xử lý nghiệp vụ. Trên cơ sơ các đánh giá trên, có thể kết luận, giải pháp thu thập dữ liệu sử dụng ***Apache Airflow*** sẽ phù hợp hơn đối với hệ thống.

Để đảm bảo các yêu cầu trong quá trình sử dụng, phần mềm mã nguồn mở Apache Airflow sẽ phải được phát triển thêm các tính năng sau:

* Tích hợp vào hệ thống quản trị tập trung.
* Bổ sung các tính năng quản lý tác vụ, quản lý công việc đang thực thi.
* Bổ sung các tính năng quản lý cấu hình bao gồm: Biến, thiết kế tác vụ, cấu hình tác vụ, thiết kế các hàm, quản lý các hàm, quản lý kết nối, quản lý file.
* Bổ sung các hàm ETL xây dựng sẵn để có thể sử dụng luôn trong việc xây dựng các luồng xử lý dữ liệu, bao gồm: xử lý dữ liệu, SQL, file, API, MongoDB, Elasticsearch, Spark, Oracle.

Dưới đây là danh sách tính năng của Apache Airflow và các tính năng cần phải phát triển thêm để đáp ứng nhu cầu của dự án:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Chức năng** | **Có sẵn của mã nguồn mở** | **Cần phát triển mới** |
| 1 | Đồ thị không chu trình có hướng (DAGs) | X |  |
| 1.1 | Khởi tạo DAG | X |  |
| 1.2 | Tải/Nạp DAGs | X |  |
| 1.3 | Chạy DAGs | X |  |
| 1.4 | Gán DAG | X |  |
| 1.5 | Đối số mặc định cho DAG | X |  |
| 1.6 | Trang trí DAG | X |  |
| 1.7 | Điều khiển luồng vận hành DAG | X |  |
| 1.8 | Thiết lập DAGs động | X |  |
| 1.9 | Trực quan hóa DAG | X |  |
| 1.1 | Tài liệu cho DAG & tác vụ | X |  |
| 1.11 | Quản lý DAGs phụ thuộc | X |  |
| 1.12 | Nhóm tác vụ và DAGs phụ thuộc | X |  |
| 1.13 | Đóng gói DAGs | X |  |
| 1.14 | Cấu hình phụ thuộc của DAG | X |  |
| 1.15 | Tạm dừng, hủy kích hoạt và xóa DAG | X |  |
| 2 | Chạy DAG | X |  |
| 2.1 | Tình trạng chạy DAG | X |  |
| 2.2 | Thiết lập trước bằng Cron | X |  |
| 2.3 | Chạy lại DAG | X |  |
| 2.4 | Kích hoạt ngoài qua CLI | X |  |
| 3 | Tác vụ | X |  |
| 3.1 | Quan hệ giữa các tác vụ | X |  |
| 3.2 | Phiên bản của tác vụ | X |  |
| 3.3 | Thời hạn cho tác vụ | X |  |
| 3.4 | Thỏa thuận cấp độ dịch vụ (SLA) | X |  |
| 3.5 | Ngoại lệ đặc biệt cho tác vụ | X |  |
| 3.6 | Tác vụ tự vận hành | X |  |
| 3.7 | Cấu hình thực thi | X |  |
| 4 | Hàm | X |  |
| 4.1 | Khuôn mẫu Jinja | X |  |
| 4.2 | Từ khóa thông số dành riêng | X |  |
| 5 | Hàm cảm biến | X |  |
| 5.1 | Luồng tác vụ | X |  |
| 5.2 | Bối cảnh của luồng tác vụ | X |  |
| 5.3 | Ghi nhật ký chạy tác vụ | X |  |
| 5.4 | Truyền các đối tượng tùy ý làm đối số | X |  |
| 5.5 | Hàm cảm biến và API luồng tác vụ | X |  |
| 6 | Thực thi tác vụ | X |  |
| 7 | Quản lý tác vụ |  | X |
| 7.1 | Quản lý Job |  | X |
| 7.2 | Quán lý Job runs |  | X |
| 7.3 | Quán lý Task Instances |  | X |
| 8 | Quản lý cấu hình |  | X |
| 8.1 | Quản lí Variable |  | X |
| 8.2 | Thiết kế Job |  | X |
| 8.3 | Cấu hình Job |  | X |
| 8.4 | Quản lý ETL Operator |  | X |
| 8.5 | Quản lý ETL List Data |  | X |
| 8.6 | Quản lý connection |  | X |
| 8.7 | Quán lý file |  | X |
| 9 | Các hàm ETL |  | X |
| 9.1 | Hàm SQL |  | X |
| 9.1.1 | Hàm chạy câu lệnh SQL |  | X |
| 9.1.2 | Hàm chạy câu lệnh SQL trong file |  | X |
| 9.1.3 | Hàm tạo CSDL |  | X |
| 9.1.4 | Hàm tạo bảng trong CSDL |  | X |
| 9.1.5 | Hàm tạo thủ tục trong CSDL |  | X |
| 9.1.6 | Hàm tạo hàm trong CSDL |  | X |
| 9.1.7 | Hàm gán dữ liệu vào bảng |  | X |
| 9.1.8 | Hàm gán và cập nhật dữ liệu theo khóa |  | X |
| 9.1.9 | Hàm gán và xóa dữ liệu theo khóa |  | X |
| 9.1.10 | Hàm gán tất cả dữ liệu vào bảng |  | X |
| 9.1.11 | Hàm lưu dự phòng danh sách bảng |  | X |
| 9.1.12 | Hàm lưu dự phòng CSDL |  | X |
| 9.2 | Hàm xử lý dữ liệu |  | X |
| 9.2.1 | Hàm chuyển đổi json thành json |  | X |
| 9.2.2 | Hàm chuyển đổi array thành json |  | X |
| 9.2.3 | Hàm chuyển đổi json thành array |  | X |
| 9.2.4 | Hàm chuyển đổi array thành array |  | X |
| 9.2.5 | Hàm chuyển đổi dữ liệu bằng function |  | X |
| 9.2.6 | Hàm chuyển đổi xml thành json |  | X |
| 9.2.7 | Hàm chuyển đổi xml thành array |  | X |
| 9.2.8 | Hàm chuyển đổi txt thành json |  | X |
| 9.2.9 | Hàm chuyển đổi txt thành array |  | X |
| 9.3 | Hàm xử lý file |  | X |
| 9.3.1 | Hàm di chuyển file |  | X |
| 9.3.2 | Hàm copy file |  | X |
| 9.3.3 | Hàm xóa file |  | X |
| 9.3.4 | Hàm mở file csv |  | X |
| 9.3.5 | Hàm mở file excel |  | X |
| 9.3.6 | Hàm mở file parquet |  | X |
| 9.3.7 | Hàm mở file avro |  | X |
| 9.3.8 | Hàm mở file txt |  | X |
| 9.3.9 | Hàm sửa tên file |  | X |
| 9.3.10 | Hàm chuyển file sang server khác |  | X |
| 9.3.11 | Hàm copy file sang server khác |  | X |
| 9.3.12 | Hàm lấy file từ server khác |  | X |
| 9.3.13 | Hàm xóa file từ server khác |  | X |
| 9.4 | Hàm xử lý API |  | X |
| 9.4.1 | Hàm lấy dữ liệu từ API |  | X |
| 9.4.2 | Hàm lấy dữ liệu từ html |  | X |
| 9.4.3 | Hàm lấy dữ liệu từ ajax |  | X |
| 9.4.4 | Hàm lấy dữ liệu từ baseauthen |  | X |
| 9.4.5 | Hàm lấy dữ liệu từ oauth |  | X |
| 9.4.6 | Hàm lấy dữ liệu từ oauth2 |  | X |
| 9.5 | Hàm xử lý MongoDB |  | X |
| 9.5.1 | Hàm lấy dữ liệu MongoDB từ aggregate |  | X |
| 9.5.2 | Hàm lấy dữ liệu MongoDB từ filter |  | X |
| 9.5.3 | Hàm gán dữ liệu (bỏ qua khi key trùng) MongoDB |  | X |
| 9.5.4 | Hàm gán tất cả dữ liệu MongoDB |  | X |
| 9.5.5 | Hàm gán và cập nhật dữ liệu MongoDB |  | X |
| 9.5.6 | Hàm gán và xóa dữ liệu MongoDB |  | X |
| 9.5.7 | Hàm tạo giá trị tham chiếu MongoDB |  | X |
| 9.6 | Hàm xử lý Elasticsearch |  | X |
| 9.6.1 | Hàm lấy dữ liệu Elasticsearch từ filter |  | X |
| 9.6.2 | Hàm gán dữ liệu (bỏ qua khi key trùng) Elasticsearch |  | X |
| 9.6.3 | Hàm gán tất cả dữ liệu Elasticsearch |  | X |
| 9.6.4 | Hàm gán và cập nhật dữ liệu Elasticsearch |  | X |
| 9.6.5 | Hàm gán và xóa dữ liệu Elasticsearch |  | X |
| 9.6.6 | Hàm tạo giá trị tham chiếu Elasticsearch |  | X |
| 9.7 | Hàm xử lý Spark |  | X |
| 9.7.1 | Hàm lấy dữ liệu Spark từ SQL |  | X |
| 9.7.2 | Hàm gán dữ liệu (bỏ qua khi key trùng) Spark |  | X |
| 9.7.3 | Hàm gán tất cả dữ liệu Spark |  | X |
| 9.7.4 | Hàm gán và cập nhật dữ liệu Spark |  | X |
| 9.7.5 | Hàm gán và xóa dữ liệu Spark |  | X |
| 9.7.6 | Hàm tạo bảng Spark |  | X |

#### Khối chức năng lưu, xử lý dữ liệu

Khối chức năng dữ liệu/tri thức hỗ trợ việc xử lý dữ liệu. Dữ liệu đầu vào được tiếp nhận từ các chức năng thu nhận/liên kết và chức năng giao diện. Khối này bao gồm các chức năng, cơ chế cho phép di chuyển dữ liệu, phân tích và xử lý dữ liệu để tạo ra các tập hợp dữ liệu mới hoặc sửa đổi/hoàn thiện dữ liệu đã tồn tại. Khối chức năng này bao gồm các chức năng cơ bản sau:

* Chức năng lưu trữ dữ liệu lớn, dữ liệu thô chưa được xử lý (Data Lake).
* Chức năng lưu trữ dữ liệu tổng hợp, tinh gọn phục vụ báo cáo thống kê (Data Warehouse).
* Chức năng lưu trữ dữ liệu lớn, bao gồm cả dữ liệu chưa và đã được xử lý (Data Lakehouse).
* Chức năng phân tích dữ liệu sử dụng các mô hình học máy để phân tích và đưa ra dự báo dựa trên số liệu.

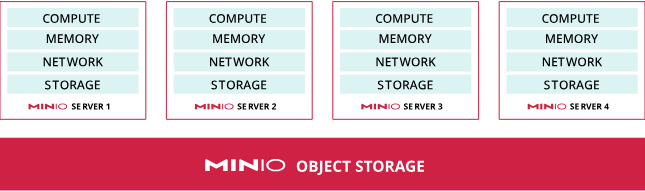
##### Giải pháp lưu trữ dữ liệu lớn (Data Lakehouse)

Data lakehouse là một thuật ngữ được đề ra bởi các kỹ sư ở IBM, với việc tận dụng hiểu quả tốc độ truy vấn của dữ liệu trong Data Warehouse, kèm theo khối lượng lưu trữ và sự đa dạng của dữ liệu, cho phép Lakehouse trở thành kiến trúc lưu trữ dữ liệu lớn của các công ty hiện nay. Không những thế, việc quản lý dữ liệu trở nên đơn giản với việc sử dụng các Metadata của dữ liệu, cho phép trực quan cấu trúc dữ liệu và kiểm soát quyền sử dụng của các bộ khi tương tác với các bộ dữ liệu này.

###### Nền tảng lưu trữ dữ liệu Big Data - Data Lake

Kho dữ liệu dùng chung tập trung của Văn Phòng Quốc Hội sẽ sử dụng công nghệ Minio cho giải pháp lưu trữ và là cốt lõi trong việc sử dụng Data Lakehouse.

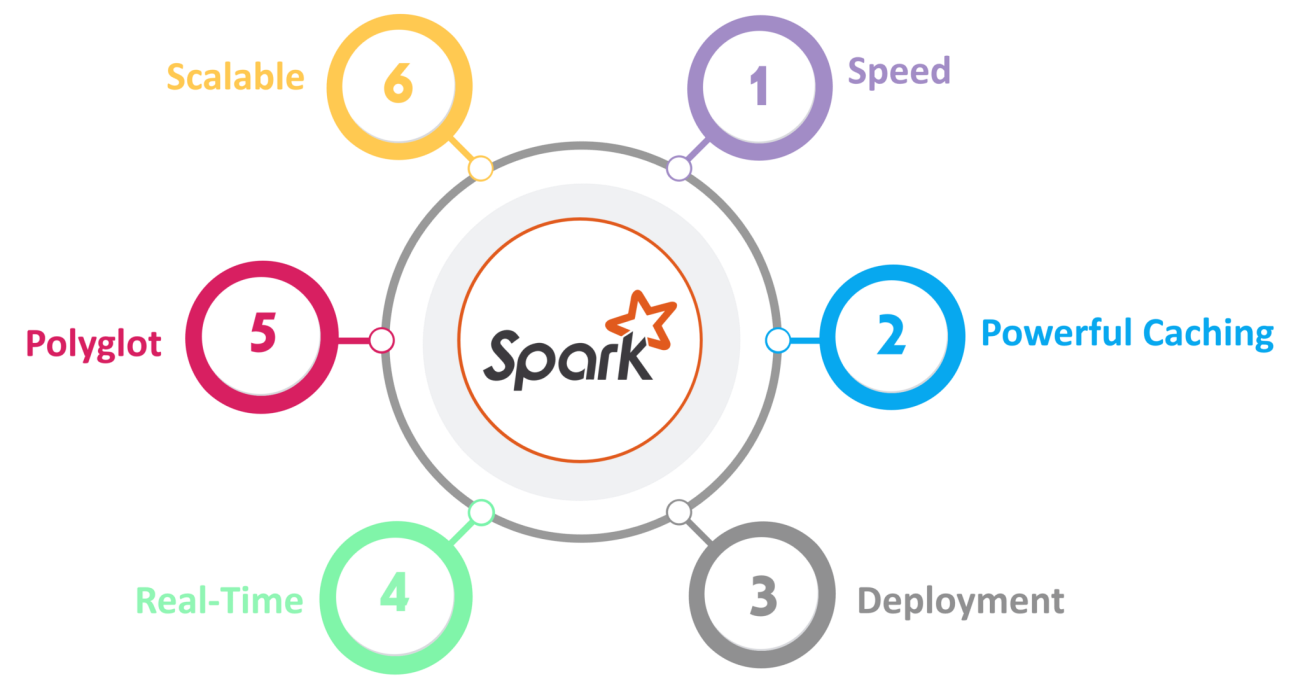
Minio là một kho dữ liệu dạng file flat theo tiêu chuẩn S3 của Amazon, giúp lưu trữ, đọc viết các tệp dữ liệu một các đơn giản. Một trong những tính năng nổi trội nhất của Minio là khả năng đồng bộ dữ liệu thông qua cơ chế Replication, đảm bảo dữ liệu không bị mất trong trường hợp khẩn cấp. Ngoài ra Minio đảm bảo tính toàn vẹn của dữ liệu nhờ cơ chế Versioning, kiểm soát nguồn gốc dữ liệu.



1. Cấu trúc cụm lưu trữ Minio

Minio cho phép lưu trữ dữ liệu thuộc dạng file flat, như là csv, parquet, mp3, mp4. Với sự linh hoạt, đơn giản khi sử dụng và đa dạng này nên Minio này phù hợp với việc lưu trữ dữ liệu lớn.

###### Xử lý dữ liệu phân tán với Spark



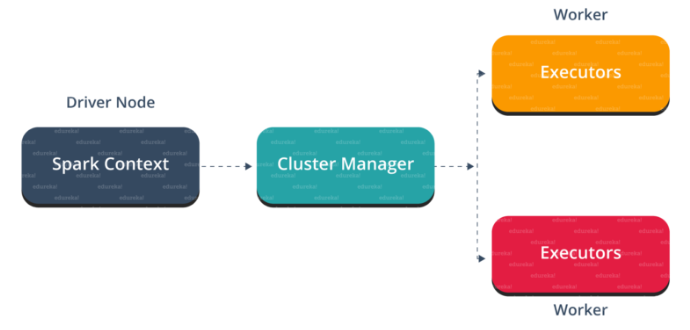
1. Apache Spark

Apache Spark là một open-source cluster computing framework, cho phép xây dựng và phân tích nhanh các mô hình dự đoán. Hơn nữa, nó còn cung cấp khả năng truy xuất toàn bộ dữ liệu cùng lúc, nhờ vậy ta không cần phải lấy mẫu dữ liệu – đòi hỏi bởi các ngôn ngữ lập trình như trên. Thêm vào đó, Spark còn cung cấp tính năng streaming, được dùng để xây dựng các mô hình real-time bằng cách nạp toàn bộ dữ liệu vào bộ nhớ.

Khi ta có một tác vụ nào đó quá lớn mà không thể xử lý trên một laptop hay một server, Spark cho phép ta phân chia tác vụ này thành những phần dễ quản lý hơn. Sau đó, Spark sẽ chạy các tác vụ này trong bộ nhớ, trên các cluster của nhiều server khác nhau để khai thác tốc độ truy xuất nhanh từ RAM.

Spark không có hệ thống file của riêng mình, nó sử dụng hệ thống file khác như: HDFS, Cassandra, S3, …. Spark hỗ trợ nhiều kiểu định dạng file khác nhau (text, csv, json…) đồng thời nó hoàn toàn không phụ thuộc vào bất cứ một hệ thống file nào.

Trong hệ thống Data Lake, Spark sẽ đóng vai trò công cụ chính cho việc phân tích, tổng hợp data và đẩy vào Data Lake, Data Warehouse.



1. Thành phần chính của Apache Spark

Apache Spark gồm có 5 thành phần chính: Spark Core, Spark Streaming, Spark SQL, MLlib và GraphX, trong đó:

* Spark Core là nền tảng cho các thành phần còn lại và các thành phần này muốn khởi chạy được thì đều phải thông qua Spark Core do Spark Core đảm nhận vai trò thực hiện công việc tính toán và xử lý trong bộ nhớ (In-memory computing) đồng thời nó cũng tham chiếu các dữ liệu được lưu trữ tại các hệ thống lưu trữ bên ngoài.
* Spark SQL cung cấp một kiểu data abstraction mới (Schema ADD) nhằm hỗ trợ cho cả kiểu dữ liệu có cấu trúc (structured data) và dữ liệu nửa cấu trúc (semi-structured data – thường là dữ liệu dữ liệu có cấu trúc nhưng không đồng nhất và cấu trúc của dữ liệu phụ thuộc vào chính nội dung của dữ liệu ấy). Spark SQL hỗ trợ DSL (Domain-specific language) để thực hiện các thao tác trên DataFrames bằng ngôn ngữ Scala, Java hoặc Python và nó cũng hỗ trợ cả ngôn ngữ SQL với giao diện command-line và ODBC/JDBC server.
* Spark Streaming được sử dụng để thực hiện việc phân tích stream bằng việc coi stream là các mini-batches và thực hiện kỹ thuật RDD transformation đối với các dữ liệu mini-batches này. Qua đó cho phép các đoạn code được viết cho xử lý batch có thể được tận dụng lại vào trong việc xử lý stream, làm cho việc phát triển lambda architecture được dễ dàng hơn. Tuy nhiên điều này lại tạo ra độ trễ trong xử lý dữ liệu (độ trễ chính bằng mini-batch duration) và do đó nhiều chuyên gia cho rằng Spark Streaming không thực sự là công cụ xử lý streaming giống như Storm hoặc Flink.
* MLlib (Machine Learning Library): MLlib là một nền tảng học máy phân tán bên trên Spark do kiến trúc phân tán dựa trên bộ nhớ. Theo các so sánh benchmark Spark MLlib nhanh hơn 9 lần so với phiên bản chạy trên Hadoop (Apache Mahout).
* GraphX: GraphX là nền tảng xử lý đồ thị dựa trên Spark. Nó cung cấp các Api để diễn tả các tính toán trong đồ thị bằng cách sử dụng Pregel Api.

###### Công cụ quản trị, và theo dõi vận hành hệ thống Data Lakehouse



1. Phần mềm Grafana + Prometheus

Hệ thống Data Lakehouse cần có một bộ công cụ giúp giám sát và quản lý tài nguyên trên các máy vận hành trọng cụm lưu trữ dùng chung. Grafana và Prometheus phù hợp với các yêu cầu trên.

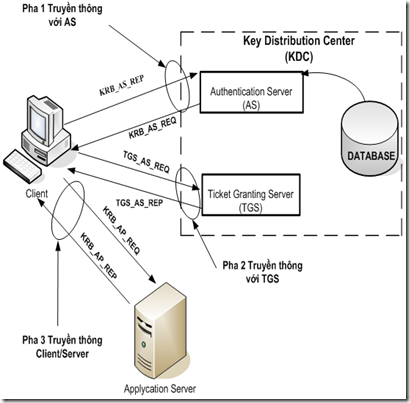
Prometheus có các chức năng sau:

* Thu thập các thông tin metric hệ thống thông qua API cho các dữ liệu realtime.
* Phát sinh cảnh báo theo điều kiện được đặt.

Grafana có các chức năng sau:

* Trực quan các dữ liệu được thu thập từ Prometheus.
* Có thể tích hợp được với các dịch vụ trong Data Lakehouse.

###### Công cụ bảo mật cho mạng các máy chủ

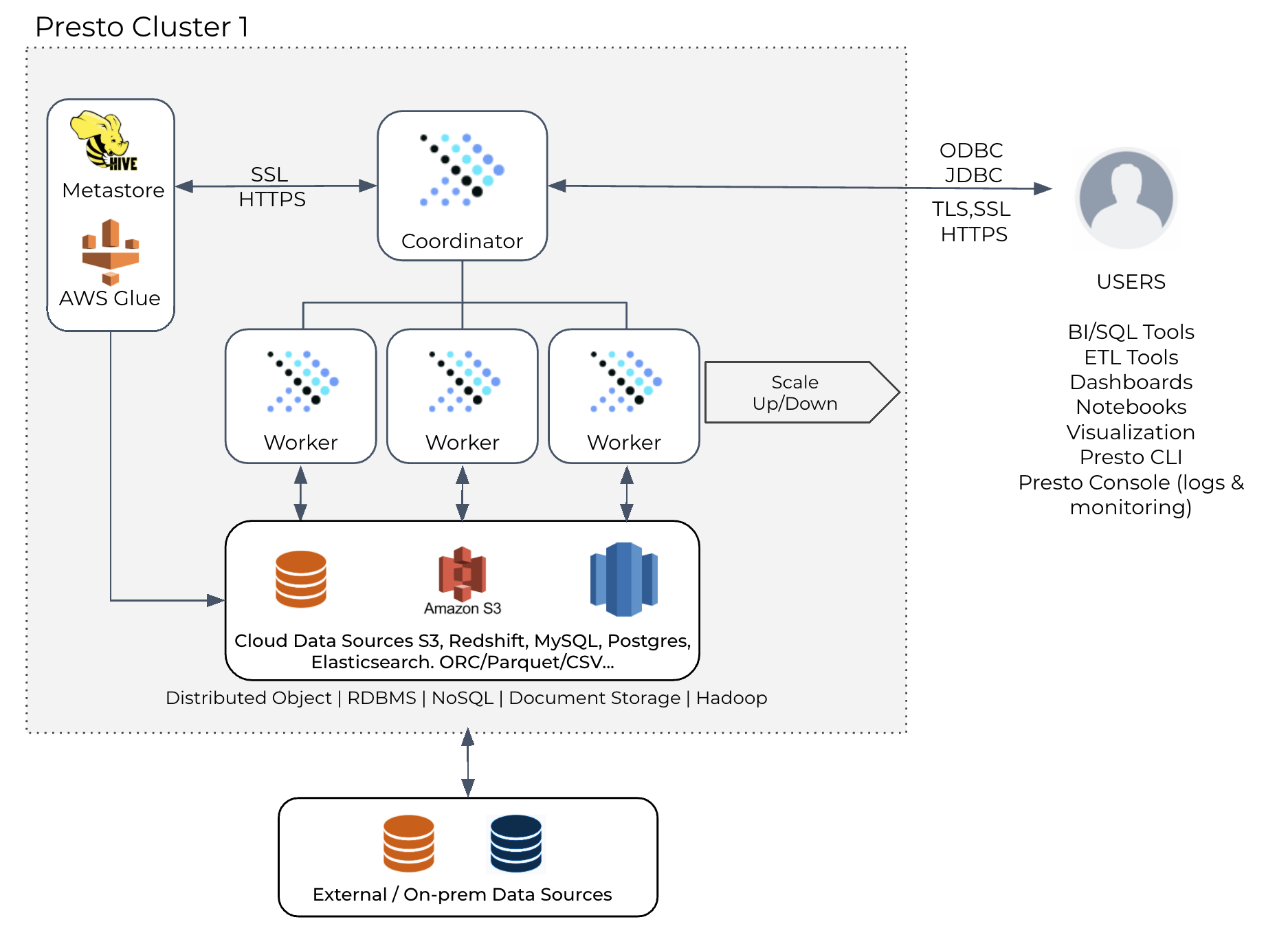


1. giao thức Kerberos

Nền tảng sử dụng giao thức Kerberos để đảm bảo bảo mật cho các services giao tiếp với nhau trong hệ thống Big Data.

Kerberos sẽ cung cấp một máy chủ chứng thực tập trung là Key Distribution Centre - KDC. KDC sẽ cung cấp Ticket cho người dùng hoặc các dịch vụ cho việc xác thực, bảo mật trước khi giao tiếp với các máy chủ, dịch vụ khác. Từ đó đảm bảo an toàn trước việc tấn công vào các máy chủ, dịch vụ trong hệ thống.

###### Công cụ phân quyền người dùng

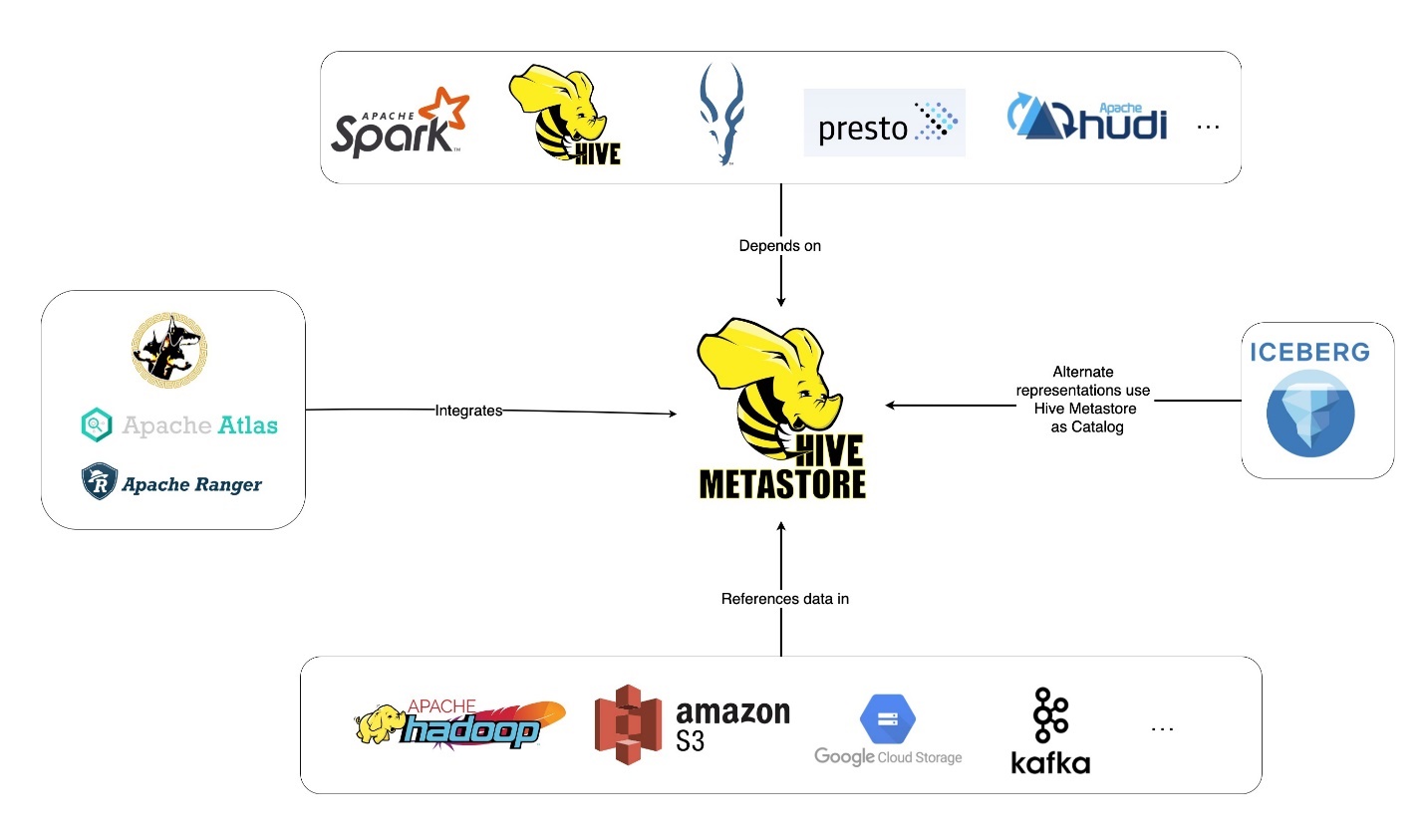


1. Mô hình cụm Presto

Presto là một bộ cơ sở dữ liệu cho phép quản lý và truy vấn nhanh dữ liệu trên nhiều nguồn như trên các CSDL dạng RDBMS như MySQL, Oracle, hoặc NoSQL như MongoDB hoặc trên các file flat như Excel, Json, Parquet. Một trong những lợi thế của Trino cho phép truy vấn dữ liệu phân tán trên nhiều máy, để giảm thiểu thời gian truy vấn. Ngoài ra, đây cũng là một công cụ quản lý quyền truy cập dữ liệu, cho phép cấu hình bảo mật lên từng tài khoản và giới hạn dữ liệu cho phép truy cập.

###### Công cụ quản trị Metadata cho dữ liệu

Metadata trong nền tảng Big Data là dữ liệu lưu lại thông tin mô tả cho dữ liệu khác. Thông qua Metadata, người dùng có thể hiểu được thông tin về dữ liệu, cách dữ liệu được tạo ra, cấu trúc vật lý, logic của dữ liệu cũng như có bức tranh toàn cảnh về những dữ liệu đang có trong hệ thống. Nền tảng Lakehouse sử dụng công cụ Apache Hive là công cụ quản trị Metadata chính trong hệ thống



1. Môi trường Apache Hive

Được thiết kế để sử dụng với Hadoop, Apache Hive là một nền tảng quản trị dữ liệu với khả năng mở rộng và tích hợp dễ dàng, đáp được hiệu quả các yêu cầu về quản trị dữ liệu trong hệ thống Data Lakehouse cũng như tích hợp với hệ sinh thái dữ liệu khác của doanh nghiệp. Những chức năng chính của Apache Hive:

* Cho phép truy vấn dữ liệu các file flat sử dụng SQL.
* Data Classification: Phân loại dữ liệu theo từng nghiệp vụ của doanh nghiệp/Tổ chức.
* Metadata Catalog Search: Hỗ trợ tìm kiếm các loại metadata có trong hệ thống.
* Business Glossary: Xây dựng từ điển nghiệp vụ cho dữ liệu.
* Integrations: Tích hợp vào những công cụ ETL, ELT giúp tự động đẩy và cập nhật dữ liệu Metadata, Data Lineage vào Hive.
* Metadata Repository: Metadata linh hoạt để phù hợp với nhu cầu lưu trữ, cung cấp APIs để thay đổi metadata.

###### Lựa chọn giải pháp

Giải pháp dữ liệu lớn Data Lake sẽ tận dụng tối đa các giải pháp Opensource sẵn có bao gồm: Minio, Apache Spark, Hive, Trino, Apache Ambari, Prometheus và Grafana.

Ngoài các tính năng sẵn có của mã nguồn mở, để đáp ứng các yêu cầu của dự án, giải pháp dữ liệu lớn Data Lake sẽ cần phải phát triển thêm các tính năng sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Chức năng** | **Có sẵn của mã nguồn mở** | **Cần phát triển mới** |
| 1 | Đồ thị không chu trình có hướng (DAGs) | X |  |
| 1.1 | Khởi tạo DAG | X |  |
| 1.2 | Tải/Nạp DAGs | X |  |
| 1.3 | Chạy DAGs | X |  |
| 1.4 | Gán DAG | X |  |
| 1.5 | Đối số mặc định cho DAG | X |  |
| 1.6 | Trang trí DAG | X |  |
| 1.7 | Điều khiển luồng vận hành DAG | X |  |
| 1.8 | Thiết lập DAGs động | X |  |
| 1.9 | Trực quan hóa DAG | X |  |
| 1.1 | Tài liệu cho DAG & tác vụ | X |  |
| 1.11 | Quản lý DAGs phụ thuộc | X |  |
| 1.12 | Nhóm tác vụ và DAGs phụ thuộc | X |  |
| 1.13 | Đóng gói DAGs | X |  |
| 1.14 | Cấu hình phụ thuộc của DAG | X |  |
| 1.15 | Tạm dừng, hủy kích hoạt và xóa DAG | X |  |
| 2 | Chạy DAG | X |  |
| 2.1 | Tình trạng chạy DAG | X |  |
| 2.2 | Thiết lập trước bằng Cron | X |  |
| 2.3 | Chạy lại DAG | X |  |
| 2.4 | Kích hoạt ngoài qua CLI | X |  |
| 3 | Tác vụ | X |  |
| 3.1 | Quan hệ giữa các tác vụ | X |  |
| 3.2 | Phiên bản của tác vụ | X |  |
| 3.3 | Thời hạn cho tác vụ | X |  |
| 3.4 | Thỏa thuận cấp độ dịch vụ (SLA) | X |  |
| 3.5 | Ngoại lệ đặc biệt cho tác vụ | X |  |
| 3.6 | Tác vụ tự vận hành | X |  |
| 3.7 | Cấu hình thực thi | X |  |
| 4 | Hàm | X |  |
| 4.1 | Khuôn mẫu Jinja | X |  |
| 4.2 | Từ khóa thông số dành riêng | X |  |
| 5 | Hàm cảm biến | X |  |
| 5.1 | Luồng tác vụ | X |  |
| 5.2 | Bối cảnh của luồng tác vụ | X |  |
| 5.3 | Ghi nhật ký chạy tác vụ | X |  |
| 5.4 | Truyền các đối tượng tùy ý làm đối số | X |  |
| 5.5 | Hàm cảm biến và API luồng tác vụ | X |  |
| 6 | Thực thi tác vụ | X |  |
| 7 | Quản lý tác vụ |  | X |
| 7.1 | Quản lý Job |  | X |
| 7.2 | Quán lý Job runs |  | X |
| 7.3 | Quán lý Task Instances |  | X |
| 8 | Quản lý cấu hình |  | X |
| 8.1 | Quản lí Variable |  | X |
| 8.2 | Thiết kế Job |  | X |
| 8.3 | Cấu hình Job |  | X |
| 8.4 | Quản lý ETL Operator |  | X |
| 8.5 | Quản lý ETL List Data |  | X |
| 8.6 | Quản lý connection |  | X |
| 8.7 | Quán lý file |  | X |
| 9 | Các hàm ETL |  | X |
| 9.1 | Hàm SQL |  | X |
| 9.1.1 | Hàm chạy câu lệnh SQL |  | X |
| 9.1.2 | Hàm chạy câu lệnh SQL trong file |  | X |
| 9.1.3 | Hàm tạo CSDL |  | X |
| 9.1.4 | Hàm tạo bảng trong CSDL |  | X |
| 9.1.5 | Hàm tạo thủ tục trong CSDL |  | X |
| 9.1.6 | Hàm tạo hàm trong CSDL |  | X |
| 9.1.7 | Hàm gán dữ liệu vào bảng |  | X |
| 9.1.8 | Hàm gán và cập nhật dữ liệu theo khóa |  | X |
| 9.1.9 | Hàm gán và xóa dữ liệu theo khóa |  | X |
| 9.1.10 | Hàm gán tất cả dữ liệu vào bảng |  | X |
| 9.1.11 | Hàm lưu dự phòng danh sách bảng |  | X |
| 9.1.12 | Hàm lưu dự phòng CSDL |  | X |
| 9.2 | Hàm xử lý dữ liệu |  | X |
| 9.2.1 | Hàm chuyển đổi json thành json |  | X |
| 9.2.2 | Hàm chuyển đổi array thành json |  | X |
| 9.2.3 | Hàm chuyển đổi json thành array |  | X |
| 9.2.4 | Hàm chuyển đổi array thành array |  | X |
| 9.2.5 | Hàm chuyển đổi dữ liệu bằng function |  | X |
| 9.2.6 | Hàm chuyển đổi xml thành json |  | X |
| 9.2.7 | Hàm chuyển đổi xml thành array |  | X |
| 9.2.8 | Hàm chuyển đổi txt thành json |  | X |
| 9.2.9 | Hàm chuyển đổi txt thành array |  | X |
| 9.3 | Hàm xử lý file |  | X |
| 9.3.1 | Hàm di chuyển file |  | X |
| 9.3.2 | Hàm copy file |  | X |
| 9.3.3 | Hàm xóa file |  | X |
| 9.3.4 | Hàm mở file csv |  | X |
| 9.3.5 | Hàm mở file excel |  | X |
| 9.3.6 | Hàm mở file parquet |  | X |
| 9.3.7 | Hàm mở file avro |  | X |
| 9.3.8 | Hàm mở file txt |  | X |
| 9.3.9 | Hàm sửa tên file |  | X |
| 9.3.10 | Hàm chuyển file sang server khác |  | X |
| 9.3.11 | Hàm copy file sang server khác |  | X |
| 9.3.12 | Hàm lấy file từ server khác |  | X |
| 9.3.13 | Hàm xóa file từ server khác |  | X |
| 9.4 | Hàm xử lý API |  | X |
| 9.4.1 | Hàm lấy dữ liệu từ API |  | X |
| 9.4.2 | Hàm lấy dữ liệu từ html |  | X |
| 9.4.3 | Hàm lấy dữ liệu từ ajax |  | X |
| 9.4.4 | Hàm lấy dữ liệu từ baseauthen |  | X |
| 9.4.5 | Hàm lấy dữ liệu từ oauth |  | X |
| 9.4.6 | Hàm lấy dữ liệu từ oauth2 |  | X |
| 9.5 | Hàm xử lý MongoDB |  | X |
| 9.5.1 | Hàm lấy dữ liệu MongoDB từ aggregate |  | X |
| 9.5.2 | Hàm lấy dữ liệu MongoDB từ filter |  | X |
| 9.5.3 | Hàm gán dữ liệu (bỏ qua khi key trùng) MongoDB |  | X |
| 9.5.4 | Hàm gán tất cả dữ liệu MongoDB |  | X |
| 9.5.5 | Hàm gán và cập nhật dữ liệu MongoDB |  | X |
| 9.5.6 | Hàm gán và xóa dữ liệu MongoDB |  | X |
| 9.5.7 | Hàm tạo giá trị tham chiếu MongoDB |  | X |
| 9.6 | Hàm xử lý Elasticsearch |  | X |
| 9.6.1 | Hàm lấy dữ liệu Elasticsearch từ filter |  | X |
| 9.6.2 | Hàm gán dữ liệu (bỏ qua khi key trùng) Elasticsearch |  | X |
| 9.6.3 | Hàm gán tất cả dữ liệu Elasticsearch |  | X |
| 9.6.4 | Hàm gán và cập nhật dữ liệu Elasticsearch |  | X |
| 9.6.5 | Hàm gán và xóa dữ liệu Elasticsearch |  | X |
| 9.6.6 | Hàm tạo giá trị tham chiếu Elasticsearch |  | X |
| 9.7 | Hàm xử lý Spark |  | X |
| 9.7.1 | Hàm lấy dữ liệu Spark từ SQL |  | X |
| 9.7.2 | Hàm gán dữ liệu (bỏ qua khi key trùng) Spark |  | X |
| 9.7.3 | Hàm gán tất cả dữ liệu Spark |  | X |
| 9.7.4 | Hàm gán và cập nhật dữ liệu Spark |  | X |
| 9.7.5 | Hàm gán và xóa dữ liệu Spark |  | X |
| 9.7.6 | Hàm tạo bảng Spark |  | X |

##### Giải pháp lưu trữ dữ liệu dùng cho phân tích, báo cáo

Data Warehouse, hay còn gọi là Kho dữ liệu ban đầu được định nghĩa là một tập dữ liệu được dùng để hỗ trợ quy trình ra quyết định. Theo quá trình phát triển, Kho dữ liệu được định nghĩa là một môi trường thông tin (information environment). Dữ liệu được lưu trong Kho dữ liệu không được tạo ra trực tiếp từ người dùng mà được lấy từ các nguồn dữ liệu sẵn có và mục đích là phục vụ tạo ra các báo cáo trực quan.

Mục đích của Kho dữ liệu là phục vụ các yêu cầu phân tích, hoặc khai phá cụ thể được gọi là chủ đề. Ví dụ với chủ đề phân tích nhân sự thì có thể bao gồm các độ đo về doanh thu của từng người, số ngày nghỉ trong tháng, số dự án tham gia trong tháng, theo các chiều phân tích: thời gian, chi nhánh, sản phẩm, v.v.

Công nghệ dùng cho Kho dữ liệu bản chất vẫn là các hệ quản trị CSDL quan hệ, nhưng có cách thức tổ chức lưu chữ và khai thác sử dụng khác hoàn toàn với các CSDL của ứng dụng. Việc lựa chọn công nghệ cho Kho dữ liệu cần quan tâm đến các yếu tố:

* Khả năng mở rộng quy mô.
* Hiệu suất..
* Khả năng bảo trì, tối ưu.
* Tính sẵn sàng của dữ liệu.
* Chi phí.

Dưới đây là bản so sánh một số hệ quản trị CSDL có khả năng đáp ứng các yêu cầu trên của Kho dữ liệu.

###### Hệ quản trị CSDL Oracle của hãng Oracle

Oracle CSDL Enterprise Edition là một Hệ quản trị CSDL dành cho các ứng dụng xử lý giao dịch trực tuyến chất lượng cao và truy vấn dữ liệu lưu trữ mạnh mẽ. Nó cung cấp hầu như tất cả các tính năng cơ bản mà Hệ thống Quản lý CSDL cần sử dụng. Phiên bản Oracle CSDL Enterprise Edition phù hợp cho cơ quan, doanh nghiệp có quy mô lớn và yêu cầu các ứng dụng kho dữ liệu chuyên sâu. Hệ thống này cung cấp khả năng mở rộng trên tất cả các nền tảng phần cứng, và có thể sử dụng để quản lý lượng thông tin rất lớn, với mức bảo mật cao nhất. Oracle CSDL Enterprise Edition có các ưu điểm riêng bảo vệ dữ liệu khỏi mất mát do lỗi của con người, giảm thời gian bảo trì định kỳ, và bao gồm tính năng tự quản lý nhằm giảm bớt chi phí hoạt động. Oracle có hầu như tất cả các tính năng cần thiết của một hệ quản trị CSDL bậc cao như hỗ trợ tất cả môi trường, quản lý tất cả dữ liệu, tích hợp tất cả các thông tin, chạy tất cả các ứng dụng, sẵn sàng trong mọi tình huống, chế độ bảo mật chắc chắn. Điểm hạn chế lớn nhất của Oracle chuyển giao công nghệ phức tạp.

***Ưu Điểm:***

* Hệ thống có sự ổn định, dữ liệu trong sẵn sàng để truy cập.
* Khả năng đáp ứng rất nhanh, từ đó tạo ra hệ thống quản trị dữ liệu có quy mô lớn với tốc độ truy vấn cực nhanh.
* Tính bảo mật cao, giám sát giúp chống xâm nhập trái phép.
* Hỗ trợ đa nền tảng, nó hoạt động trên nhiều nền tảng khác nhau.
* Được hỗ trợ từ nhà phát triển như các sự cố xảy ra sẽ được hãng phát hành tư vấn, hỗ trợ giải quyết.

***Nhược Điểm:***

* Chi phí cho bản quyền cực lớn, khó có cơ hội với các công ty đang trên đà phát triển.
* Chưa được tương thích với các công nghệ, ứng dụng phát triển bởi Microsoft.
* Java là ngôn ngữ khó tiếp cận nên khó khăn trong việc tiếp cận các công cụ thiết kế, lập trình.

###### Hệ quản trị CSDL DB2 của IBM

IBM DB2 là một phần mềm quản lý và tương tác với CSDL RDBMS được phát triển bởi IBM và phát triển độc quyền trên các nền tảng của hãng. Hệ thống cơ sở dữ liệu này có thể cài đặt trên hầu hết các hệ thống và có thể mở rộng theo chiều ngang hoặc chiều dọc. Ngoài ra, Db2 được cấu hình đi kèm với mô hình Machine Learning, cho phép phân tích dữ liệu trực quan hóa truy vấn khi kết hợp với các sản phẩm BI của IBM như IBM Watson, IBM Cognos, IBM SPSS.

***Ưu điểm:***

* Hệ thống dữ liệu ổn định, dễ dàng truy cập
* Có tính bảo mật cao.
* Hỗ trợ đa nền tảng, cho phép cài đặt lên nhiều hệ thống OS khác nhau.
* Dễ dàng tích hợp vào giải pháp BI của IBM.

***Nhược điểm:***

* Cộng đồng phát triển trên Db2 không đông đảo như các CSDL khác, khó tìm tài liệu hướng dẫn.
* Phần mềm có chi phí License cao, mỗi DB Instance cho môi trường Production cần có một License riêng.
* Để tận dụng lợi thế của Db2, đòi hỏi phải sử dụng kèm các giải pháp khác của IBM.

###### Hệ quản trị CSDL SQL-SERVER của hãng Microsoft

SQL Server hay còn gọi là Microsoft SQL Server, viết tắt là MS SQL Server. Đây là một phần mềm được phát triển bởi Microsoft dùng để lưu trữ dữ liệu dựa trên chuẩn RDBMS, và nó cũng là một hệ quản trị cơ sở dữ liệu quan hệ đối tượng (ORDBMS). SQL Server cung cấp đầy đủ công cụ để quản lý, từ giao diện GUI cho đến việc sử dụng ngôn ngữ truy vấn SQL. Ngoài ra điểm mạnh của nó là Microsoft có khá nhiều nền tảng kết hợp hoàn hảo với SQL Server như ASP.NET, C# xây dựng Winform, bởi vì nó hoạt động hoàn toàn độc lập.

***Ưu điểm:***

* Có thể cài nhiều phiên bản MS SQL khác nhau trên cùng một máy tính.
* Duy trì riêng biệt các môi trường sản xuất, phát triển, thử nghiệm.
* Giảm thiểu các vấn đề tạm thời trên cơ sở dữ liệu.
* Tách biệt các đặc quyền bảo mật.
* Duy trì máy chủ dự phòng.

***Nhược điểm:***

* SQL Server chỉ chạy trên hệ điều hành Windows.
* Cần thanh toán phí license để chạy nhiều CSDL (CSDL).

###### Hệ quản trị CSDL MySQL

MySQL là hệ quản trị cơ sở dữ liệu mã nguồn mở phổ biến hàng đầu trên thế giới và đặc biệt được ưa chuộng trong quá trình xây dựng, phát triển ứng dụng. Đây là hệ quản trị cơ sở dữ liệu tốc độ cao, ổn định và dễ sử dụng, có khả năng thay đổi mô hình sử dụng phù hợp với điều kiện công việc khả chuyển. MySQL hoạt động trên nhiều hệ điều hành, cung cấp một hệ thống lớn các hàm tiện ích rất mạnh.

Với tốc độ và tính bảo mật cao, MySQL thích hợp với các ứng dụng có truy cập cơ sở dữ liệu trên internet. MySQL có thể tải miễn phí từ trang chủ với nhiều phiên bản cho các hệ điều hành khác nhau như: phiên bản Win32 cho các hệ điều hành dòng Windows, Linux, Mac OS X, Unix, FreeBSD, NetBSD, Novell NetWare, SGI Irix, Solaris, SunOS...

***Ưu điểm:***

* Sử dụng dễ dàng: MySQL là cơ sở dữ liệu tốc độ cao và ổn định, công cụ này dễ sử dụng và hoạt động trên nhiều hệ điều hành cung cấp hệ thống lớn các hàm tiện ích.
* Tính bảo mật cao: MySQL phù hợp với các ứng dụng có truy cập cơ sở dữ liệu trên internet vì nó sở hữu nhiều tính năng bảo mật, thậm chí là bảo mật cấp cao.
* Đa tính năng: MySQL có thể hỗ trợ hàng loạt các chức năng SQL từ hệ quản trị cơ sở dữ liệu quan hệ trực tiếp và cả gián tiếp.
* Khả năng mở rộng và mạnh mẽ: Công cụ MySQL có khả năng xử lý khối dữ liệu lớn và có thể mở rộng khi cần thiết.
* Tương thích trên nhiều hệ điều hành: MySQL tương thích để chạy trên nhiều hệ điều hành, như Novell NetWare, Windows \* Linux \*, nhiều loại UNIX \* (như Sun \* Solaris \*, AIX và DEC \* UNIX), OS / 2, FreeBSD \*,.... MySQL cũng cung cấp phương tiện mà các máy khách có thể chạy trên cùng một máy tính với máy chủ hoặc trên một máy tính khác (giao tiếp qua mạng cục bộ hoặc Internet).
* Cho phép khôi phục: MySQL cho phép các transaction được khôi phục, cam kết và phục hồi sự cố.

***Nhược điểm:***

* MySQL bị hạn chế dung lượng, cụ thể, khi số bản ghi của người dùng lớn dần, sẽ gây khó khăn cho việc truy xuất dữ liệu, khiến người dùng cần áp dụng nhiều biện pháp để tăng tốc độ chia sẻ dữ liệu như chia tải CSDL ra nhiều server, hoặc tạo cache MySQL.
* So với Microsoft SQL Server hay Oracle, độ bảo mật của MySQL chưa cao bằng. Và quá trình Restore cũng có phần chậm hơn.

###### Hệ quản trị CSDL PostgreSQL

PostgreSQL là hệ quản trị cơ sở dữ liệu quan hệ mã nguồn mở tiên tiến mà mục đích chính là tuân thủ theo chuẩn và khả năng mở rông. PostgreSQL hay còn gọi là Postgree áp dụng chuẩn SQL ANSI/IOS cùng với các phiên bản.

So với các hệ quản trị cơ sử dữ liệu quan hệ khác, PostgreSQL khá khác. Nó hỗ trợ hướng đối tượng mạnh mẽ và chức năng cơ sở dữ liệu quan hệ. Ví dụ như hỗ trợ hoàn toàn cho các giao dịch đáng tin cậy như là Atomicity, Consistency, Isolation, Durability (ACID).

Do có nền tảng công nghệ mạnh mẽ, Postgres có hoàn toàn có khả năng xử lý nhiều tiến trình rất hiệu quả. Sự hỗ trợ đồng thời đạt được mà không cần đọc các khóa nhờ vào sự thực hiện của Multiversion Concurrency Control (MVCC), mà nó cũng đảm bảo việc tuân thủ theo ACID.

***Ưu điểm:***

* PostgreSQL có thể chạy các trang web và ứng dụng web động với LAMP.
* Ghi nhật ký viết trước của PostgreSQL làm cho nó trở thành một cơ sở dữ liệu có khả năng chịu lỗi cao
* Mã nguồn PostgreSQL có sẵn miễn phí theo giấy phép nguồn mở. Điều này cho phép bạn tự do sử dụng, sửa đổi và triển khai nó theo nhu cầu kinh doanh của bạn.
* PostgreSQL hỗ trợ các đối tượng địa lý để bạn có thể sử dụng nó cho các dịch vụ dựa trên vị trí và hệ thống thông tin địa lý.
* PostgreSQL hỗ trợ các đối tượng địa lý để nó có thể được sử dụng làm kho lưu trữ dữ liệu không gian địa lý cho các dịch vụ dựa trên vị trí và hệ thống thông tin địa lý.
* Dễ sử dụng
* Hạn chế việc bảo trì hệ thống
* Cộng đồng sử dụng lớn, tăng mạnh trong những năm gần đây

***Nhược Điểm:***

* Hiệu suất: Đối với các toán tử đơn giản thì PostgreSQL thực hiện kém hiệu quả hơn so với các hệ quản trị cơ sở dữ liệu quan hệ khác như MySQL. Nhưng đối với toán tử phức tạp thì PostgreSQL thực hiện tốt hơn rất nhiều.
* Vì không chịu sự quản lý của bất kỳ tổ chức nào nên PostgreSQL gây khó khăn cho người dùng trong việc tiếp cận đầy đủ tính năng

###### Apache Iceberg

Apache Iceberg là một dạng file mở hiệu suất cao, có cấu trúc bảng vượt trội trong việc xử lý và phân tích khối lượng dữ liệu lớn. Các file Iceberg tạo các metadata, cho phép các công cụ xem chúng như là các bảng có chức năng và sử dụng như là một bảng SQL truyền thống. Iceberg là một dạng file hỗ trợ hoàn toàn cho các giao dịch mang tính chất ACID như các CSDL dạng RDBMS như trên, nhưng cũng linh hoạt cho phép nhiều máy có thể đọc cùng một lúc. Iceberg hoàn toàn tương thích với chuẩn lưu trữ dữ liệu dạng Object như Minio và có thể được truy vấn thông qua các công chj như Spark, Trino, Hive, v.v.

***Ưu điểm:***

* Là một dạng file có format mở, không bị giới hạn bởi các nhà cung cấp
* Có độ nén cao, phù hợp với việc lưu trữ dữ liệu lớn
* Tính ACID tương tự như các CSDL dạng RDBMS
* Được nhiều công cụ truy vấn hỗ trợ như Spark, Trino
* Có tốc độ truy vấn cao, và hỗ trợ sử dụng SQL để truy vấn
* Hoàn toàn tương thích với chuẩn lưu trữ dữ liệu S3
* Cho phép sử dụng các khung xử lý dữ liệu phân tán như Spark chạy trên nhiều máy cùng một lúc

***Nhược điểm:***

* Đòi hỏi cần có kỹ sư tận dụng các công cụ xử lý dữ liệu như Spark, Hive, Presto để tận dụng

###### Lựa chọn giải pháp

Hệ thống xây dựng đòi hỏi lưu trữ khối lượng dữ liệu lớn, ổn định, hiệu năng và không bị giới hạn bởi nhà phát hành, do vậy việc sử dụng Iceberg là hoàn toàn phù hợp và đã được chứng minh trong thực tế. Iceberg là một dạng file bảng mã nguồn mở cho phép phân tích, truy vấn và quản lý dữ liệu trong Data Lakehouse. Iceberg hiện đang được cộng động kỹ sư dữ liệu và các nhà phân tích sử dụng như là tiêu chuẩn SQL mới. Dạng file này được lưu trữ trên Minio, không cần phải quản lý riêng một CSDL cho các dữ liệu dạng bảng, giảm chi phí vận hành.

##### Giải pháp phân tích dữ liệu

Phần mềm phân tích dữ liệu được xây dựng để phân tích dữ liệu lớn từ Kho dữ liệu, dữ liệu sẽ được phần mềm phân tích thành dữ liệu thống kê, phân tích để phục vụ cho mục đích báo cáo từ những dữ liệu này sẽ được hình thành lên các chỉ số, các báo cáo hiển thị trên phần mềm hiển thị trực quan dữ liệu.

* Sử dụng dữ liệu thời gian thực thu được từ các hệ thống khác nhau và dữ liệu lịch sử được ghi lại trong quá khứ, Hệ thống phân tích và tính toán một chỉ số phù hợp, tức là chỉ số chỉ ra số liệu hiện tại có gần đạt được hay vượt quá giá trị mong muốn tính được dựa trên các số liệu dự kiến theo quá khứ.
* Đọc dữ liệu từ các Hệ thống nguồn khác nhau, xử lý và lưu trữ trong cơ sở dữ liệu
* Trích xuất dữ liệu từ hệ thống nguồn,
* Thống nhất các tiêu chuẩn và chất lượng dữ liệu,
* Điều chỉnh dữ liệu để các nguồn riêng biệt có thể được sử dụng cùng nhau
* Phân tích đa chiều, drill down, slice and dice: Có khả năng slice-dice, drill down, drill up, drill through dữ liệu theo các chiều và theo phân cấp dữ liệu.

Phần mềm phân tích dữ liệu sẽ ứng dụng các mô hình học máy bao gồm:

* Scikit-learn là một thư viện hỗ trợ được phát triển cho ngôn ngữ Python giúp các nhà khoa học dữ liệu có thể dễ dàng thực thi những thuật toán phân tích dữ liệu, học máy một cách nhanh chóng. Scikit-learn khác với Spark Engine trong việc ứng dụng các công cụ Machine Learning ở điểm là chạy trên single node, phù hợp với lượng dữ liệu nhỏ, data không có cấu trúc.
* Pandas là thư viện làm việc với dữ liệu có cấu trúc của Python, với đầy đủ những chức năng phân tích dữ liệu, phù hợp với dữ liệu nhỏ.
* TensorFlow: là thư viện do Google phát triển, có phát triển những thuật toán học sâu với lượng dữ liệu lớn mà các công cụ single node như Pandas không thể phân tích được. Điểm mạnh của Tensorflow là xử lý những loại dữ liệu phi cấu trúc với những mô hình mạng Neuron mạnh mẽ như mạng tích chập (CNN), mạng hồi quy (RNN), ...
* XGBoost là một thư viện ứng dụng mô hình Cây quyết định giải quyết các bài toán phân loại dữ liệu. Đầu tiên xây dựng một cây quyết định, sau đó tối ưu dần những thuộc tính để thu được cây có kết quả tốt nhất.
* Maplotlib giúp người dùng có thể biểu diễn dữ liệu thông qua các loại biểu đồ, giúp trực quan hóa dữ liệu thông qua code, giúp xác định trước những template đồ thị có thể vẽ trên công cụ phân tích dữ liệu.

###### Lựa chọn giải pháp

Trên thế giới các nền tảng phục vụ phát triển học máy gồm có các sản phẩm thương mại như IBM Watson, Databricks, AWS EMR và các nền tảng mã nguồn mở như JupyterHub và Zeppelin.

Trong các giải pháp công nghệ này, giải pháp mã nguồn mở được đánh giá nổi bật nhờ những thế mạnh sau:

* Bộ tiện ích mở rộng lớn nhờ cộng đồng hỗ trợ mạnh.
* Miễn phí sử dụng.
* Khả năng cấu hình, cài đặt dễ dàng và linh hoạt.
* Dễ dàng sử dụng các thư viện visualization trực tiếp không cần gọi qua wrapper như Zeppelin.
* Hỗ trợ nhiều người dùng.

Ngoài việc đáp ứng khả năng phân tích dữ liệu và tương thích với các mô hình học máy. Phần mềm phân tích dữ liệu còn cần phải kết nối được đến Chức năng hiển thị trực quan dữ liệu và Khối Chức năng Thống kê liên kết dữ liệu.

Dựa theo các phân tích trên, nền tảng mã nguồn mở JupyterHub là phù hợp nhất với nhu cầu hiện tại. Tuy nhiên, để đảm bảo các yêu cầu sử dụng, mã nguồn mở này cần phải phát triển thêm:

* Tích hợp với hệ thống quản lý tập trung
* Kết hợp với công cụ trực quan hóa dữ liệu để hiển thị kết quả phân tích dữ liệu dựa trên dữ liệu mới nhất một cách trực quan cho người dùng
* Tích hợp sẵn thuật toán thường dùng để phục vụ xây dựng các mô hình học máy:
  + Sử dụng thuật toán linear regression
  + Sử dụng thuật toán logistic regression
  + Sử dụng thuật toán lasso regression
  + Sử dụng thuật toán decision tree
  + Sử dụng thuật toán naive bayes
  + Sử dụng thuật toán random forest
  + Sử dụng thuật toán k nearest neighbours
  + Sử dụng thuật toán kmean clustering
  + Sử dụng thuật toán fuzzy c-mean
  + Sử dụng thuật toán SVM
  + Sử dụng thuật toán LS-SVM
  + Sử dụng thuật toán CNN
  + Sử dụng thuật toán LSTM
* Ngoài các tính năng sẵn có của mã nguồn mở, để đáp ứng các yêu cầu của dự án, các Kỹ sư khoa học dữ liệu sẽ phát triển riêng các mô hình học máy theo phạm vi các bài toán phân tích dữ liệu của dự án

#### Khối chức năng chia sẻ dữ liệu

Nền tảng chia sẻ dữ liệu cho phép mở rộng kết nối từ Kho dữ liệu đến các hệ thống phần mềm thông qua các API. Nền tảng bao gồm hai phân hệ là: Phần mềm quản lý giao diện lập trình ứng dụng (API) và Cổng dữ liệu mở (Open data portal).

##### Phân tích lựa chọn công nghệ phát triển Back-end

Theo sự phát triển và xu hướng công nghệ hiện nay, có nhiều lựa chọn về công nghệ phần mềm cho việc phát triển các hệ thống thông tin lớn.

Tuy nhiên có thể chia thành các nền tảng phát triển mã nguồn mở (như Java, PHP, Nodejs...) và nền tảng phát triển mã nguồn đóng (như ASP.NET, Net core).

Mỗi nền tảng đều có những ưu nhược điểm khác nhau, nên để đánh giá được thế mạnh của các hệ thống thông tin được xây dựng dựa trên 02 dòng nền tảng công nghệ trên cần nhìn vào khả năng ứng dụng của chúng và các ưu nhược điểm của từng công nghệ. Các tiêu chí đánh giá bao gồm như sau:

* Sự hỗ trợ từ cộng đồng.
* Sự hỗ trợ từ nhà cung cấp nền tảng.
* Có triển khai được trên nhiều môi trường hay không?
* Các công cụ hỗ trợ để phát triển ứng dụng có nhiều hay không?
* Có dễ dàng lập trình và cài được trên nhiều môi trường hay không?
* Chi phí đầu tư nhân lực để phát triển có lớn không?
* Công nghệ có dễ dàng bảo trì, sửa đổi hay không?
* Công nghệ có bảo mật hay không?
* Tốc độ của ứng dụng có cao hay không?
* Ứng dụng xây dựng có phù hợp và xử lý được một hệ thống thông tin phức tạp không?
* Ứng dụng có dễ dàng kết nối với các hệ thống phần mềm đã có sẵn hay không?
* Có phù hợp với các tài nguyên sẵn có hay không (các hệ thống sẵn có, nhân lực quản trị, bản quyền phần mềm...)?

Vì vậy, việc lựa chọn công nghệ nào cần được đánh giá một cách toàn diện dựa trên các tiêu chí nêu trên.

| **Tiêu chí** | **Nguồn mở** | **Nguồn đóng** |
| --- | --- | --- |
| Sự hỗ trợ từ cộng đồng? | Đáp ứng tốt hơn | Đáp ứng kém |
| Sự hỗ trợ từ nhà cung cấp nền tảng? | Đáp ứng tốt hơn | Đáp ứng kém |
| Có triển khai được trên nhiều môi trường hay không? | Đáp ứng | Không đáp ứng |
| Các công cụ hỗ trợ để phát triển ứng dụng có nhiều hay không? | Đáp ứng | Đáp ứng |
| Có dễ dàng lập trình và cài được trên nhiều môi trường hay không? | Đáp ứng | Không đáp ứng |
| Chi phí đầu tư nhân lực để phát triển có lớn không? | Chi phí vừa phải | Chi phí lớn |
| Công nghệ có dễ dàng bảo trì, sửa đổi hay không? | Đáp ứng | Đáp ứng kém |
| Công nghệ có bảo mật hay không? | Đáp ứng | Đáp ứng |
| Tốc độ của ứng dụng có cao hay không? | Đáp ứng | Đáp ứng |
| Ứng dụng xây dựng có phù hợp và xử lý được một hệ thống thông tin phức tạp không? | Đáp ứng | Đáp ứng |
| Ứng dụng có dễ dàng kết nối với các hệ thống phần mềm đã có sẵn hay không? | Đáp ứng | Đáp ứng |
| Có phù hợp với các tài nguyên sẵn có hay không (các hệ thống sẵn có, nhân lực quản trị, bản quyền phần mềm, ...)? | Đáp ứng | Đáp ứng |

Những công nghệ lập trình chính được mô tả qua bảng sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **Java** | **ASP.NET, .NET CORE** | **ASP** | **PHP** |
| Tốc độ xử lý nhanh, hiệu quả cao | Tốc độ xử lý nhanh, hiệu quả cao | Tốc độ xử lý chậm | Tốc độ xử lý nhanh, hiệu quả cao |
| Chi phí giá thành thấp (với các phần mềm mã nguồn mở) | Chi phí giá thành cao (do một phần phải mua bản quyền) | Chi phí giá thành trung bình (do một phần phải mua bản quyền) | Chi phí thấp (ngôn ngữ phát triển miễn phí không tốn chi phí mua bản quyền) |
| Thời gian lập trình và triển khai chậm hơn 1 chút | Thời gian lập trình và phát triển nhanh | Thời gian lập trình và phát triển trung bình | Thời gian lập trình và phát triển nhanh |
| Hệ điều hành: WINDOWS, LINUX | Hệ điều hành: WINDOWS | Hệ điều hành: WINDOWS | Hệ điều hành: WINDOWS, LINUX |
| Số nhà cung cấp website ít | Số nhà cung cấp website nhiều, khả năng lựa chọn một trang web phù hợp tốt hơn. | Số nhà cung cấp website ít | Số nhà cung cấp website nhiều, khả năng lựa chọn một trang web phù hợp tốt hơn. |
| Khả năng mở rộng và phát triển dễ dàng, nhanh chóng | Mở rộng và phát triển website khó khăn | Mở rộng và phát triển website khó khan | Mở rộng và phát triền website dễ dàng, nhanh chóng |
| Các công cụ và công nghệ hỗ trợ phong phú, đa dạng. Phát triển web trên nền web 2.0. Sử dụng công nghệ Ajax làm cho quá trình duyệt web nhanh chóng và thân thiện hơn đối với người dùng | Các công cụ và công nghệ hỗ trợ phong phú, đa dạng. Phát triển web trên nền web 2.0. Sử dụng công nghệ Ajax làm cho quá trình duyệt web nhanh chóng và thân thiện hơn đối với người dùng | Công cụ và công nghệ hỗ trợ ít, khó tìm | Các công cụ và công nghệ hỗ trợ phong phú, đa dạng. Phát triển web trên nền web 2.0. Là ngôn ngữ được thiết kế để phát triển webservice, Go cho phép quá trình phát triển và triển khai trở nên nhanh gọn và thân thiện với người dùng |

Việc lựa chọn công nghệ để xây dựng hệ thống phải đáp ứng các yêu cầu chung sau đây:

* Giải pháp công nghệ phải mang lại sự thành công của dự án, điều đó có nghĩa là hệ thống phải hoạt động tốt đáp ứng mục tiêu đề ra.
* Giải pháp công nghệ lựa chọn phải đảm bảo hệ thống hoạt động ổn định, liên tục, lâu dài và tính bảo mật cao.
* Đáp ứng khả năng tăng trưởng dữ liệu và thay đổi mở rộng nâng cấp trong tương lai.
* Chi phí bản quyền ít tốn kém.

##### Phân tích lựa chọn công nghệ phát triển Front-end

Ngày nay, việc lập trình một khối với việc không phân chia rõ lớp back-end và front-end, sử dụng chung một ngôn ngữ lập trình cho cả hai lớp là cách làm đã bộc lộ rất nhiều hạn chế. Các ứng dụng hiện đại, tối ưu ngày nay đều đã phân biệt rõ lớp back-end với lớp front-end. Hai lớp này gọi nhau thông qua API (Web API và RESTful). Cách thức này giúp hệ thống có thể có được giao diện mượt mà, giảm thiểu dung lượng đường truyền (giảm dung lượng gói tin giữa back-end và front-end) hay tái sử dụng front-end và các API phía back-end cho việc phát triển các ứng dụng mobile.

Hiện nay, các công nghệ phát triển front-end ưu việt nhất thuộc về: ReactJS, Angular và VueJS. Khái quát về các công nghệ này như sau:

###### ReactJS

***Ưu điểm:***

* Dễ học, nhờ thiết kế đơn giản, sử dụng JSX (một cú pháp giống như HTML) để tạo khuôn mẫu. Tài liệu hướng dẫn của Facebook rất chi tiết.
* Tốc độ ứng dụng vô cùng ấn tượng. Tất cả là nhờ kỹ thuật DOM ảo của React và tối ưu hóa rendering.
* Hỗ trợ render phía máy chủ rất tốt. Điều này làm cho React trở thành một framework mạnh mẽ cho các ứng dụng tập trung vào nội dung.
* Hỗ trợ tạo ứng dụng Progressive Web App (PWA) nhanh chóng.
* Data-binding một chiều, có nghĩa là ít tác dụng phụ không mong muốn.
* Mô hình Redux: một mô hình quản lý trạng thái ứng dụng cực tốt và rất dễ học.
* React theo trường phái lập trình hàm (Functional Programming), tạo mã dễ kiểm tra và có khả năng tái sử dụng cao.

***Nhược điểm:***

* Thiếu tài liệu hướng dẫn chính chủ: Chính vì sự phát triển siêu nhanh của ReactJS mà tài liệu hướng dẫn chính chủ không theo kịp. Hầu hết các bài viết hướng dẫn trên mạng bị lỗi thời.
* React đang đi theo hướng lập trình hàm, điều này sẽ gây đôi chút khó khăn cho các nhà phát triển đã quen với lập trình hướng đối tượng (OOP).
* Việc trộn giữa templating với logic ứng dụng (JSX) có thể gây nhầm lẫn cho một số nhà phát triển lúc đầu.

###### Angular

***Ưu điểm:***

* Hỗ trợ TypeScript: Angular được tạo ra để được sử dụng cùng với TypeScript.
* Data-binding: tự động đồng bộ hóa dữ liệu giữa thành phần model và view.
* Tài liệu hướng dẫn chi tiết và đầy đủ: Với tài liệu hướng dẫn chi tiết từ chính nhà cung cấp sẽ giúp người mới tìm hiểu Angular nhanh chóng tiếp cận và làm chủ framework này. Nhờ đó mà giảm thời gian training, thảo luận….
* MVVM (Model-View-ViewModel): cho phép chia dự án thành nhiều phần độc lập với nhau. Từ đó sẽ dễ dàng chỉnh sửa, thêm tính năng cũng như bảo trì dự án sau này.
* Dependency Injection: cho phép làm giảm sự phụ thuộc giữa các Object với nhau. Các object càng độc lập với nhau càng tốt, mục đích để sau này nâng cấp, sửa đổi… sẽ hạn chế ảnh hưởng tới các object khác.
* Cấu trúc và kiến trúc của Angular được tạo ra đặc biệt cho khả năng mở rộng dự án lớn.

***Nhược điểm:***

* Angular đa dạng các thành phần/khái niệm như Injectables, Components, Pipes, Modules… Điều này khiến cho việc tìm hiểu trở nên khó khăn hơn một chút so với Reactjs, vốn chỉ có một thành phần duy nhất là Component.
* Theo đánh giá hiệu năng thì Angular hơi chậm hơn React và Vuejs.

###### VueJS

***Ưu điểm:***

* Empowered HTML: điều này có nghĩa là Vue.js có nhiều đặc điểm tương tự với Angular. Chính vì thế nó có thể giúp tối ưu hóa việc xử lý các khối HTML khi sử dụng các thành phần khác nhau.
* Tài liệu hướng dẫn chi tiết: Nhờ có tài liệu chi tiết mà việc học tập trở nên nhanh chóng, tiết kiệm thời gian phát triển ứng dụng chỉ sử dụng kiến thức cơ bản về HTML và JavaScript.
* Khả năng tương thích. Việc chuyển sang sử dụng Vuejs từ các Js framework tương đối nhanh chóng do khá tương đồng với Angular và React về mặt thiết kế và kiến trúc.
* Khả năng tích hợp tốt có thể được sử dụng cho việc xây dựng cả ứng dụng (single-page applications) phức tạp hoặc chỉ một phần của ứng dụng. Điều này cho phép bạn cập nhật, nâng cấp ứng dụng mà không ảnh hưởng quá nhiều tới hệ thống hiện tại.
* Large scaling: Vuejs có thể phát triển templates có khả năng tái sử dụng cao.
* Cho phép triển khai nhanh.

***Nhược điểm:***

* Thiếu hỗ trợ: Vuejs vẫn có thị phần khá nhỏ so với Reactjs hoặc Angular. Điều đó có nghĩa là việc chia sẻ kiến ​​thức trong cộng đồng vẫn còn khá ít.

Chức năng giao diện cho phép mở rộng kết nối từ Kho dữ liệu dùng chung tập trung của tỉnh đến các hệ thống phần mềm, dịch vụ ngoài như nền tảng tích hợp chia sẻ của khối Chính quyền điện tử. Chức năng cụ thể như sau:

* Cho phép các ứng dụng phần mềm dịch vụ kết nối vào nền tảng theo các giao thức khác nhau như websocket, http, soap, .... bằng các cấu trúc đóng gói dữ liệu khác nhau: json, xml.
* Cho phép định tuyến gói tin được gửi đến nền tảng để xác định điểm đến của gói tin dựa vào dữ liệu của gói tin bao gồm header, content; luật định tuyến gói tin được cấu hình động trên phần mềm.
* Cho phép chuyển đổi cấu trúc gói tin, định dạng gói tin theo luật mapping dữ liệu của gói tin nguồn và gói tin đích.
* Cho phép xác thực kết nối thông qua LDAP, token.
* Cho phép các phần mềm, dịch vụ bên ngoài nền tảng khai thác dữ liệu từ khối chức năng dữ liệu chi thức thông qua giao thức web hoặc REST API.
* Cho phép mở rộng hệ thống theo chiều ngang, triển khai nhiều máy chủ khác nhau để đáp ứng lượng kết nối lớn.
* Cho phép lưu vết và gỡ lỗi gói tin trong quá trình chuyển đổi, định tuyến.

Khối chức năng giao diện bao gồm hai phân hệ là: Phần mềm quản lý giao diện lập trình ứng dụng (API) và Cổng dữ liệu mở (Open data portal).

##### Phần mềm quản lý giao diện lập trình ứng dụng (API)

Hệ thống quản lý giao diện lập trình ứng dụng (API) thực hiện các công việc sau:

* Phần mềm được xây dựng để quản lý các kết nối cho các hệ thống dịch vụ, hệ thống phần mềm trong tỉnh, từ đó các dịch vụ phần mềm có thể an toàn trao đổi thông tin nghiệp vụ với nhau theo chuẩn.
* Là một thành phần không thể thiếu trong quá trình xây dựng các hệ thống bảo mật với các kết nối tích hợp giữa các hệ thống được quản lý chặt chẽ, quản lý việc kết nối trao đổi thông tin giữa các phầm mềm, cung cấp các dịch vụ, cơ chế bảo mật, chuẩn tích hợp cho các hệ thống trong một Đơn vị/Tổ chức.
* Quản lý các chuẩn kết nối: Hỗ trợ việc phân quyền truy cập các ứng dụng theo vai trò của từng người sử dụng, phân quyền truy cập các dịch vụ web (Web Service). Phân quyền người dùng quản lý các ứng dụng trong hệ thống.
* Quản lý xác thực truy cập: Cung cấp nền tảng quản lý truy cập mạnh mẽ hiệu quả, quản lý truy cập các dịch vụ trong hệ thống, quản lý truy cập các phần mềm kết nối đến hệ thống. Giám sát truy cập, thống kê truy cập đến các hệ thống.

Giải pháp để xây dựng hệ thống được thiết kế dạng Web-based Application, theo kiến trúc 3 lớp, bao gồm:

* Lớp mô hình: là tầng cung cấp giao diện thao tác với CSDL của ứng dụng. Mọi thao tác và giao dịch với CSDL đều được đặt ở tầng Model.
* Lớp trình diễn: Bao gồm các chương trình hoặc trình duyệt Web tại máy trạm người sử dụng. Lớp trình diễn này cung cấp các giao diện cho phép người sử dụng có thể vận hành hệ thống được. Giao diện người sử dụng của hệ thống sẽ là giao diện Web-based.
* Lớp xử lý nghiệp vụ: Bao gồm các chương trình xử lý nghiệp vụ tại máy chủ. Lớp xử lý nghiệp vụ là các thủ tục, các chương trình để xử lý nghiệp vụ và truy xuất lớp CSDL.
* Ngôn ngữ phát triển:
  + - Backend: Java.
    - Frontend: HTML, Javascript, CSS.
* Hệ quản trị cơ sở dữ liệu: Iceberg.
* Server:
  + - Hệ điều hành mã nguồn mở: Linux.
    - Webserver: Ngnix.

##### Phần mềm cổng dữ liệu mở (open data portal)

Cổng dữ liệu mở được tích hợp thông tin từ các phần mềm/ hệ thống khác, là điểm truy cập tập trung và là cổng thông tin duy nhất thực hiện trao đổi thông tin với các phần mềm/ hệ thống thông tin đồng thời thực hiện cung cấp và trao đổi thông tin với người sử dụng thông qua một phương thức thống nhất trên nền tảng Web ở bất cứ đâu. Người dân có thể truy cập vào hệ thống để tìm kiếm, theo dõi tất cả các thông tin liên quan tới tất cả lĩnh vực của tỉnh như: Kinh tế - Xã hội, Tài nguyên môi trường, Văn hóa, …

Đối với doanh nghiệp có nhu cầu khai thác dữ liệu, có thể sử dụng các tập dữ liệu mở để phân tích số liệu, tìm hiểu các dịch vụ liên quan, nghiên cứu thị trường, … phục vụ hoạt động sản xuất, kinh doanh. Doanh nghiệp có thể sử dụng API cung cấp sẵn để khai thác dữ liệu trên cổng, nhằm xây dựng các tiện ích, ứng dụng công nghệ mới phục vụ chính tại doanh nghiệp.

Giải pháp để xây dựng hệ thống được thiết kế dạng Web-based Application, theo kiến trúc 3 lớp, bao gồm:

* Lớp mô hình: là tầng cung cấp giao diện thao tác với CSDL của ứng dụng. Mọi thao tác và giao dịch với CSDL đều được đặt ở tầng Model.
* Lớp trình diễn: Bao gồm các chương trình hoặc trình duyệt Web tại máy trạm người sử dụng. Lớp trình diễn này cung cấp các giao diện cho phép người sử dụng có thể vận hành hệ thống được. Giao diện người sử dụng của hệ thống sẽ là giao diện Web-based.
* Lớp xử lý nghiệp vụ: Bao gồm các chương trình xử lý nghiệp vụ tại máy chủ. Lớp xử lý nghiệp vụ là các thủ tục, các chương trình để xử lý nghiệp vụ và truy xuất lớp CSDL.
* Ngôn ngữ phát triển:
  + Backend: Python.
  + Frontend: ReactJS.
* Hệ quản trị cơ sở dữ liệu: Iceberg.
* Server:
  + Hệ điều hành mã nguồn mở: Linux
  + Webserver: Ngnix.

#### Chức năng phục vụ trực quan dữ liệu

Việc hiển thị trực quan dữ liệu phục vụ giám sát điều hành sẽ thông qua các Báo cáo trực quan (Dashboard) để hiển thị các thông tin về các lĩnh vực, ngành nghề dưới dạng các chỉ số để theo dõi. Các chỉ số này sẽ được hiển thị dưới dạng các bảng biểu, biểu đồ đa dạng về cách thức hiển thị và màu sắc. Việc sáng tạo nội dung cho các Báo cáo trực quan này phải được quản lý và thực hiện trên một phần mềm trực quan hóa dữ liệu chuyên dụng mà trên thế giới gọi là các phần mềm BI (Bussiness Intelligent). Giải pháp BI sẽ giúp việc thiết kế và khởi tạo báo cáo trực quan một cách linh động, dễ dàng thông qua các công cụ có sẵn mà không phải lập trình lại.

Trên thực tế, hiện nay các giải pháp BI hiện đại đã thể hiện ưu thế vượt trội so với các giải pháp truyền thống, được thể hiện trong đánh giá của hãng tư vấn độc lập Gartner dưới đây:

##### Phần mềm Tabeau BI

Tableau là một công ty phần mềm Mỹ, thành lập năm 2003 bởi 3 nhà khoa học máy tính hàng đầu tại đại học Standford. Từ 2014 trở lại đây, Tableau liên tục dẫn đầu về giải pháp Business Intelligence (BI) trong bảng xếp hạng của Gartner, được Gartner gọi là “Tiêu chuẩn vàng” cho BI hiện đại. Tableau BI vượt trội ở tính năng cho phép kết nối tới nhiều nguồn dữ liệu khác nhau, tốc độ nhanh, khả năng “self-service” và tính trực quan hóa cao.

Tableau BI có mô hình triển khai linh hoạt, cả on-premise, on public clouds (Azure, Amazon...) và on Tableau clouds. Phí bản quyền (license) theo hình thức thuê bao (subscription) và khá rẻ so với các giải pháp BI của SAP, Oracle, IBM. Dự báo nhu cầu về Tableau trên thế giới sẽ bùng nổ trong 3 năm tới, hiện ở Việt Nam cũng đã bắt đầu xuất hiện xu hướng này.

Người dùng có thể xem và hiểu dữ liệu, tạo báo cáo và dashboard nhanh hơn thông qua công nghệ phân tích hình ảnh độc đáo, dễ sử dụng của Tableau. Sử dụng Tableau giúp người dùng khám phá các xu hướng chính và các ngoại lệ mà nếu không sẽ không thể tìm thấy bằng cách sử dụng các giải pháp báo cáo thông minh (BI) truyền thống. Nhờ hoàn toàn chủ động sử dụng và làm chủ công cụ, người dùng làm việc nhanh hơn từ 10 đến 100 lần với Tableau. Họ có thể chia sẻ các dashboard, các báo cáo và đồ họa trên web chỉ với một vài cú nhấp chuột.

Graphical user interface, application

Description automatically generated

1. Mô hình kiến trúc giải pháp Tableau BI

##### Phần mềm Microsoft Power BI

Microsoft Power BI là giải pháp phân tích thông minh cho phép doanh nghiệp lưu trữ dữ liệu của mình và chia sẻ thông tin chi tiết trong toàn tổ chức hoặc nhúng chúng vào ứng dụng hoặc trang web của DN. Power Bi cho phép kết nối với hàng trăm nguồn dữ liệu.

Power BI là một tập hợp các dịch vụ phần mềm, ứng dụng và công cụ kết nối hỗ trợ quá trình phân tích trực quan dữ liệu. Các cấu phần chính bao gồm:

* Power BI Desktop
* Power BI Service
* Power BI Mobile
* Power BI Report Server
* Q&A
* SQL Server Analysis Server
* Power BI Embedded
* Power BI Desktop
* Kết nối tới hàng trăm nguồn dữ liệu

##### Phần mềm Oracle Business intelligent Enterprise Edition (OBIEE)

Oracle Business Intelligence Enterprise Edition (OBIEE) là một sản phẩm BI toàn diện của Oracle. Nó cung cấp đầy đủ khả năng phân tích và lập báo cáo. Ngoài ra OBIEE là sự kết hợp của một bộ các công cụ thông minh hữu ích dựa trên nền tảng kiến trúc thống nhất.

Bên cạnh việc có thể xây dựng báo cáo từ bất cứ nguồn dữ liệu nào, người dùng có thể linh động trong việc thiết lập mẫu báo cáo thông qua thao tác kéo thả đơn giản. Thậm chí giải pháp này còn cung cấp Plugin để tích hợp vào MS Office – công cụ văn phòng thông dụng với mọi người.

Các cấu phần trong OBIEE được mô tả như hình vẽ dưới đây:

A screenshot of a computer

Description automatically generated with medium confidence

1. Các cấu phần trong OBIEE

##### Phần mềm IBM Cognos

IBM Cognos là một sản phẩm BI của IBM. Đây là một công cụ đầy đủ các khả năng cho phép trực quan dữ liệu và xuất báo cáo. Ngoài ra, Cognos cũng có tích hợp mô hình AI của IBM Watson cho phép truy vấn và thiết lập các mô hình dự đoán.

Cognos hỗ trợ rất tốt các đầu kết nối đến từ các nguồn dữ liệu của IBM, ngoài ra cũng có hỗ trợ thu thập dữ liệu từ các dạng file flat và các nguồn CSDL khác nhau. Không những vậy, người dùng có thể thiết lập các mẫu báo cáo thông qua việc kéo thả đơn giản.

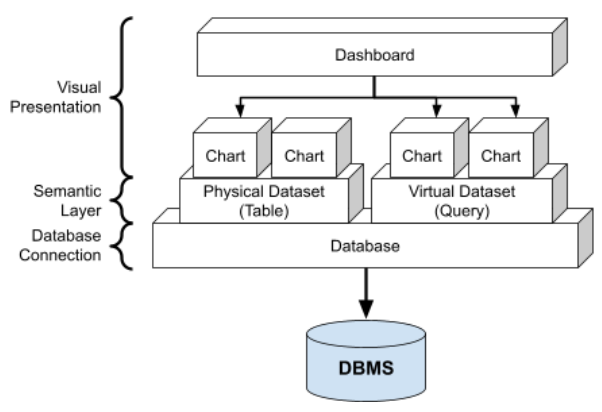
##### Phần mềm Apache Superset

Apache Superset do Maxime Beauchemin (người tạo ra Apache Airflow) khởi động như một dự án hack-a-thon khi anh làm việc tại Airbnb và tham gia chương trình Apache Incubator vào năm 2017. Ngoài Airbnb, dự án còn có sự đóng góp đáng kể từ các công ty công nghệ hàng đầu khác như Lyft và Dropbox. Superset 'tốt nghiệp' chương trình vườn ươm và trở thành dự án cấp cao nhất tại Apache Software Foundation vào năm 2021.

Về cơ bản, các tính năng Superset khá tương đồng các phần mềm phân tích dữ liệu khác, bao gồm:

* Tạo và quản lý dashboard.
* Xác thực với các hệ thống xác thực của doanh nghiệp (OpenID, LDAP, OAuth...).
* Hỗ trợ nhiều loại cơ sở dữ liệu: SQL, Postgres, MySQL...
* Hỗ trợ truy vấn trực tiếp.

Các cấu phần trong Superset được mô tả như hình vẽ dưới đây:



1. Các cấu phần trong Superset

Superset còn khá mới, tuy nhiên đã đáp ứng phần giao diện hoàn chỉnh và cung cấp công cụ trực quan cho người dùng cần làm phân tích dữ liệu, báo cáo, bảng biểu. Superset cũng cho phép trích bảng biểu để nhúng vào các màn hình ứng dụng, gửi email trực tiếp chứa đường dẫn báo cáo, xuất hình ảnh báo cáo để chia sẻ, ... Nhìn chung thì Superset đáp ứng được phần lớn nhu cầu của người làm phân tích dữ liệu và báo cáo.

Superset phù hợp với những người dùng:

* Những người cần công cụ mạnh để có thể làm nhiều loại bảng, query liên tục trên CSDL thì không cần phải dùng thêm công cụ khác.
* Những người đang tìm kiếm một công cụ làm phân tích dữ liệu miễn phí và mạnh mẽ không kém các công cụ nổi tiếng.
* Với khả năng tuỳ chỉnh cực kỳ mạnh, Superset có thể hỗ trợ nhiều loại dữ liệu nguồn và có thể cài đặt thêm nhiều loại biểu đồ chưa có.

##### Lựa chọn giải pháp

Từ kết quả đánh giá, so sánh Tableau BI, Microsoft BI, Oracle OBIEE, SAP BusinessObject, IBM Cognos, Superset có thể rút ra các nhận xét chính sau:

Các sản phẩm BI truyền thống có thế mạnh xử lý dữ liệu khối lượng lớn, và là sự lựa chọn có thể cân nhắc khi hệ thống tác nghiệp chủ yếu của đơn vị là giải pháp ERP tương ứng (Oracle OBIEE với Oracle EBS, SAP BOBJ với SAP ERP).

Các giải pháp BI hiện đại có ưu thế vượt trội về tính năng, đơn giản hơn đáng kể trong triển khai và khai thác sử dụng, đặc biệt là hỗ trợ đặc lực cho người dùng trong việc tự tạo lập, tùy biến và chia sẻ báo cáo. Tổng chi phí sở hữu cũng thấp hơn đáng kể so với các giải pháp truyền thống.

Trong số các giải pháp BI hiện đại, Tableau BI, Microsoft Power BI, Superset có thể coi là những lựa chọn hàng đầu gần như tương đương. Các sản phẩm Tableau BI, Microsoft Power BI có thế mạnh về độ hoàn thiện và chất lượng đã được nhiều khách hàng kiểm chứng. Tuy vậy, các sản phẩm này cũng có những điểm chưa tốt là chi phí cao, không có khả năng mở rộng nâng cấp tính năng và thiếu hỗ trợ trực tiếp từ chính nhà phát triển. Sản phẩm Superset tuy là sản phẩm mới nhưng đã khắc phục tất cả các nhược điểm trên.

Để đảm bảo các yêu cầu sử dụng, mã nguồn mở Apache Superset cần phải phát triển thêm:

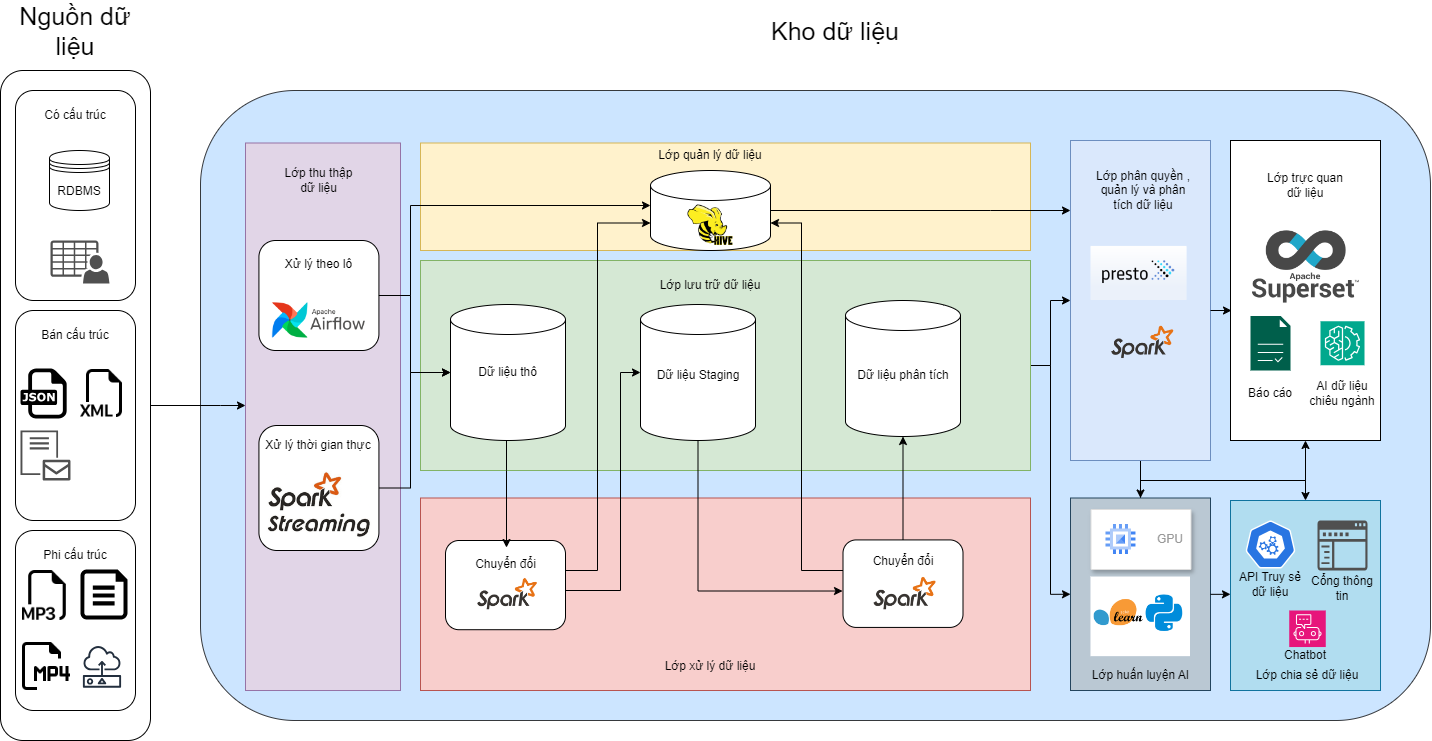
* Tích hợp với hệ thống quản lý tập trung.
* Việt hóa giao diện hệ thống.
* Nâng cấp tính năng phân quyền.
* Nâng cấp giao diện hệ thống.
* Nâng cấp, bổ sung mới các loại biểu đồ.
* Nâng cấp, bổ sung các cách thức kết nối và import dữ liệu.

Dưới đây là danh sách tính năng của Apache Airflow và các tính năng cần phải phát triển thêm hoặc nâng cấp để đáp ứng nhu cầu của dự án:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Chức năng** | **Có sẵn của mã nguồn mở** | **Cần phát triển mới** |
| 1 | Giao diện trực quan để trực quan hóa bộ dữ liệu và tạo bảng điều khiển tương tác | X |  |
| 2 | Đồ họa trực quan đẹp mắt để hiển thị dữ liệu | X |  |
| 3 | Trình tạo trực quan hóa không có mã để trích xuất và trình bày bộ dữ liệu | X |  |
| 4 | IDE SQL để chuẩn bị dữ liệu phục vụ trực quan hóa | X |  |
| 5 | Lớp ngữ nghĩa cho phép các nhà phân tích dữ liệu nhanh chóng thao tác với các biến và thuộc tính | X |  |
| 6 | Hỗ trợ vượt trội cho hầu hết các cơ sở dữ liệu SQL | X |  |
| 7 | Truy vấn và lưu bộ đệm bất đồng bộ trong bộ nhớ | X |  |
| 8 | Mô hình bảo mật có thể mở rộng cho phép cấu hình chi tiết về quyền truy cập | X |  |
| 9 | Tích hợp với các công nghệ xác thực (cơ sở dữ liệu, OpenID, LDAP, OAuth, REMOTE\_NGƯỜI DÙNG, v.v.) | X |  |
| 10 | Khả năng thêm plugin trực quan hóa tùy chỉnh | X |  |
| 11 | API để tùy chỉnh theo chương trình | X |  |
| 12 | Kiến trúc dựa trên đám mây được thiết kế từ đầu để mở rộng quy mô | X |  |
| 13 | Quản lý biểu diễn dữ liệu |  | X |
| 14 | Quản lý Biểu đồ |  | X |
| 15 | Quản lý Báo cáo trực quan |  | X |
| 16 | Quản lý Biểu đồ/Dashboard/Thư mục yêu thích |  | X |
| 17 | Quản lý hoạt động gần đây |  | X |
| 18 | Quản lý lớp chú thích |  | X |
| 19 | Cấu hình Tab trong Báo cáo trực quan |  | X |
| 20 | Cấu hình Hàng trong Báo cáo trực quan |  | X |
| 21 | Cấu hình Cột trong Báo cáo trực quan |  | X |
| 22 | Cấu hình thanh tiêu đề trong Báo cáo trực quan |  | X |
| 23 | Cấu hình Markdown trong Báo cáo trực quan |  | X |
| 24 | Cấu hình Iframe trong Báo cáo trực quan |  | X |
| 25 | Cấu hình bộ lọc trong Báo cáo trực quan |  | X |
| 26 | Cấu hình Phân cách trong Báo cáo trực quan |  | X |
| 27 | Thiết kế báo cáo dạng Biểu đồ |  | X |
| 28 | Thiết kế báo cáo dạng bảng số liệu |  | X |
| 29 | Thiết kế báo cáo dạng Bản đồ |  | X |
| 30 | Quản lý Data |  | X |
| 31 | Quản lý Bộ dữ liệu |  | X |
| 32 | Quản lý CSDL |  | X |
| 33 | Quản lý Truy vấn |  | X |
| 34 | Quản lý thư mục |  | X |
| 35 | Quản lý Thư mục dùng chung |  | X |
| 36 | Quản lý Thư mục cá nhân |  | X |

Ngoài các tính năng trên, để đáp ứng các yêu cầu của dự án, các Kỹ sư dữ liệu sẽ phát triển riêng các báo cáo trực quan cho các lĩnh vực trong phạm vi dự án.

#### Mô hình tổng quan hệ thống kho dữ liệu



1. Mô hình tổng thể kho dùng chung

Dưới đây là mô tả luồng xử lý dữ liệu ở kho:

* Dữ liệu sẽ được thu thập từ các nguồn có cấu trúc, bán và phi cấu trúc đẩy vào trong kho dữ liệu dưới dạng thô.
* Dữ liệu sau đó sẽ được chuẩn hóa và chuyển đổi rồi lưu trữ ở môi trường dữ liệu Staging.
* Phụ thuộc vào nhu cầu nghiệp vụ, dữ liệu sau đó sẽ được chuẩn hóa lưu vào trong kho dữ liệu phân tích.
* Tất cả các dữ liệu khi thu thập ban đầu, chuyển đổi sẽ lưu các cấu trúc metadata vào trong cơ sở dữ liệu trong lớp quản lý dữ liệu.
* Trong quá trình phân tích và truy vấn dữ liệu, lớp phân quyền và quản lý dữ liệu sẽ quản lý quyền truy cập dữ liệu lên đối tượng phù hợp.
* Dữ liệu sau đó sẽ được sử dụng để sử dụng cho các ứng dụng trực quan dữ liệu, chia sẻ cho các đơn vị tìm kiếm và huấn luyện các mô hình AI, phục vụ các chức năng truy vấn chuyên ngành hoặc chatbot.
* Các hệ thống phần mềm chạy trên nền tàng Openstack hoặc Openshift.

#### Danh mục quy chuẩn kỹ thuật, tiêu chuẩn được áp dụng

* + - * 1. *Yêu cầu về chuẩn thông tin dữ liệu*

Hệ thống phải tuân thủ các quy định và chuẩn quốc gia theo các chuyên ngành như:

* Xây dựng hệ thống quản lý siêu dữ liệu (MetaData) và danh mục dữ liệu theo quy định tại Nghị định 73/2017/NĐ-CP.
* Kết nối, tích hợp, chia sẻ dữ liệu với LGSP theo Thông tư 13/2017/TT-BTTTT ngày 15/12/2017 của Bộ Thông tin và Truyền thông.
* Tuân thủ các chuẩn mở quốc tế, các đặc tả mở của các tổ chức như IEEE, OGC OpenGIS, W3C, OASIS, ISO, …
* Các mô hình và chuẩn mở do W3C ban hành (http://www.w3c.org).
* Các mô hình và chuẩn mở do Open Geospatial Consortium đề xuất (http://www.opengis.org).
* Chuẩn mã ký tự chữ Việt do Tổng cục Tiêu chuẩn Đo lường Chất lượng quy định.
* Chuẩn hệ toạ độ quốc gia VN2000 do Bộ Tài nguyên và Môi trường quy định.
* Chuẩn các mã số do Tổng cục Thống kê quy định.
* Danh mục tiêu chuẩn kỹ thuật về ứng dụng công nghệ thông tin trong cơ quan nhà nước được công bố trong Thông tư 39/2017/TT-BTTTT ngày 15/12/2017 của Bộ Thông tin và Truyền thông.
  + - 1. ***Các quy chuẩn, tiêu chuẩn kỹ thuật chính*** 
         1. *Tiêu chuẩn về công nghệ thông tin*
* Truyền siêu văn bản (HTTP v1.1): sử dụng cho việc truyền tải thông tin trên Cổng.
* Truyền thư điện tử (SMTP/MIME): sử dụng khi hệ thống trả lời tự động qua e-mail.
* Dịch vụ Web (SOAP v1.2, WSDL v1.1, UDDI v3): sử dụng trong việc trao đổi dữ liệu với các hệ thống ứng dụng bên ngoài.
* Dịch vụ đồng bộ thời gian (NTP v3): sử dụng trong việc đồng bộ dữ liệu giữa các DC.
* Công văn số 3788 /BTTTT-THH ngày 26/12/2014 của Bộ Thông tin và Truyền thông về việc hướng dẫn liên thông, trao đổi dữ liệu có cấu trúc bằng ngôn ngữ XML giữa các hệ thống thông tin trong cơ quan nhà nước.
* Mô hình hóa đối tượng (UML 2.0): hỗ trợ cho việc phân tích và thiết kế hệ thống.
* Trình diễn bộ kí tự (UTF-8): sử dụng cho trình diễn giao diện người dùng.
* Chuẩn nội dung web (HTML v4.01, HTML v5.1): sử dụng cho hệ thống Cổng.
* Giao diện người dùng (CSS2): sử dụng cho giao diện Cổng.
* Văn bản (.doc, .docx, .xls, .xlsx, .pdf): sử dụng cho các văn bản, tài liệu đính kèm.
* Ảnh đồ họa (JPEG, TIFF, PNG): sử dụng cho các tệp tin ảnh của hệ thống.
* Bộ ký tự và mã hóa cho tiếng Việt (TCVN 6909:2001): sử dụng cho việc hiển thị thông tin tiếng Việt của hệ thống.
* Công văn số 1145/BTTTT-CATTT ngày 03/4/2020 hướng dẫn bộ tiêu chí, chỉ tiêu kỹ thuật để đánh giá và lựa chọn nền tảng điện toán đám mây phục vụ chính phủ điện tử/chính quyền điện tử.
  + - * 1. *Tiêu chuẩn về an toàn thông tin*
* An toàn giao vận SSL v3.0.
* An toàn truyền tệp tin (HTTPS): sử dụng cho việc thiết lập SSL Certificate.
* An toàn giải thuật mã hóa (RSA, SHA-2): sử dụng giải thuật mã hóa công khai để thiết lập các giao dịch an toàn.
* An toàn trao đổi bản tin XML (XML Signature Syntax and Processing): sử dụng cho việc đồng bộ dữ liệu.
* Tiêu chuẩn về quản lý thông tin tài khoản người dùng.
* Tiêu chuẩn OpenID: là một chuẩn mở cho xác thực được quảng bá bởi tổ chức phi lợi OpenID Foundation (OpenID được sử dụng trên internet, và các công ty như Google, WordPress, Yahoo, Paypal... sử dụng OpenID để xác thực người dùng).
* Tiêu chuẩn Oauth 2.0 (OAuth2 là một chuẩn mở để ủy quyền/phân quyền (authorization), OAuth2 cũng là nền tảng của OpenID Connect, nó cung cấp OpenID (xác thực - authentication) ở phía trên của OAuth2 (ủy quyền - authorization) để có một giải pháp bảo mật hoàn chỉnh hơn.
* Thông tư 03/2017/TT-BTTTT ngày 24 tháng 4 năm 2017 của Bộ Thông tin và Truyền thông Quy định chi tiết và hướng dẫn một số điều của nghị định 85/2016/NĐ-CP ngày 01/7/2016 của Chính phủ về bảo đảm an toàn thông tin theo cấp độ.
* Thông tư 31/2017/TT-BTTTT ngày 15 tháng 11 năm 2017 của Bộ Thông tin và Truyền thông Quy định hoạt động giám sát an toàn hệ thống thông tin.
  + - * 1. *Các tiêu chuẩn về trung tâm dữ liệu*
* AS/NZS 3084: Tiêu chuẩn về hệ thống mạng, ống bảo vệ và không gian cho cáp thông tin trong tòa nhà thương mại của DCL.
* AS/NZS 3085.1: Các quy định cơ bản về quản lý hệ thống cáp thông tin của DCL
* TCVN 6160: 1996 phòng cháy chữa cháy.
* TCVN 9250:2012 về Trung tâm dữ liệu - Yêu cầu về hạ tầng kỹ thuật viễn thông.
* Thông tư 39/2017/TT-BTTTT ngày 15/12/2017 của Bộ trưởng Bộ Thông tin và Truyền thông ban hành Danh mục tiêu chuẩn kỹ thuật về ứng dụng công nghệ thông tin trong cơ quan nhà nước.
* Thông tư 03/2013/TT-BTTTT ngày 22/01/2013 của Bộ Thông tin và Truyền thông Thông tư này quy định việc áp dụng tiêu chuẩn, quy chuẩn kỹ thuật đối với trung tâm dữ liệu.
* ANSI/TIA-942 - “Telecommunications Infrastructure Standard for Data Centers” - Published: April 2005.
* ISO/IEC 24764 “Generic Cabling for Data Centres”. reviewed in January by JTC-1/SC 25/WG3.
* CENELEC:
  + EN 50173-5: “Information technology - Generic cabling systems – Part 5: Data Centres”.
  + EN 50174-2 Amendment: Adds Annex on Data Center planning & installation. First draft being circulated in April 20, 2007.

Khảo sát viết thêm sau

* + - * 1. *Yêu cầu về an toàn thông tin*

Kho dữ liệu dùng chung tập trung của tỉnh bao gồm 04 hệ thống thành phần (Thu thập, liên kết dữ liệu; phân tích biểu diễn dữ liệu; Nền tảng Lưu trữ dữ liệu dùng cho khai thác, phân tích; Nền tảng giao diện dữ liệu) sẽ được cài đặt, vận hành từ Trung tâm tích hợp dữ liệu của tỉnh nên sẽ được đề xuất, đảm bảo cấp độ theo quy định Nghị định số 85/NĐ-CP.

Khảo sát viết thêm sau

* + - * 1. *Các yêu cầu cần đáp ứng về thời gian xử lý, độ phức tạp xử lý của các phần mềm*

Hệ thống sẽ cung cấp công suất xử lý và dung lượng lưu trữ để hỗ trợ các khối lượng dự kiến, có thể tăng theo thời gian.

Hệ thống cần phải chú ý đến việc truy vấn dữ liệu với tốc độ xử lý nhanh nhất bằng cách phân thành nhiều thao tác nếu chức năng đó có nhiều thao tác truy vấn dữ liệu, hạn chế truy vấn có độ phức tạp cao sẽ dẫn đến hết thời gian (Time out) truy vấn dữ liệu.

Hệ thống sẽ đáp ứng tối thiểu thời gian phản hồi như sau:

* Một đến hai giây đối với các giao dịch không đòi hỏi truy vấn CSDL.
* Hai đến ba giây đối với các giao dịch đòi hỏi ghi vào CSDL hoặc truy vấn CSDL với một liên kết.
* Hai đến bốn giây đối với các giao dịch đòi hỏi truy vấn CSDL đến 5 liên kết.

Có giải pháp đảm bảo được việc nhập dữ liệu đầu vào ngay cả khi đường truyền hoạt động không ổn định.

Hệ thống có khả năng chuyển đổi dữ liệu dễ dàng và chính xác giữa các loại dữ liệu: kiểu ngày sang chuỗi, kiểu chuỗi sang ngày, kiểu số sang chuỗi, kiểu chuỗi (các ký tự số) sang kiểu số, ….

* + - * 1. *Các yêu cầu về cài đặt, hạ tầng, đường truyền, an toàn vận hành, khai thác, sử dụng*

Hệ thống phải đảm bảo cho phép sao lưu dữ liệu định kỳ và đột xuất.

Có các biện pháp phục hồi dữ liệu khi hệ thống gặp sự cố.

Hệ thống cung cấp các báo cáo khai thác, vận hành hệ thống.

Yêu cầu trong quá trình triển khai hệ thống: thiết kế phù hợp với hạ tầng, đường truyền hiện có hoặc được nâng cấp/củng cố đảm bảo hoạt động đầy đủ tính năng và an toàn khi vận hành hệ thống.

#### Hệ thống backup

Backup chính là việc sao chép và lưu trữ lại toàn bộ dữ liệu gốc của máy chủ, CSDL file … sang một thiết bị lưu trữ khác làm dữ liệu dự phòng, đảm bảo rằng có thể phục hồi dữ liệu khi dữ liệu cũ bị mất. Hoạt động backup cần phải được làm thường xuyên để cập nhật thêm những dữ liệu mới sau một quá trình sử dụng.

* + - * 1. *Acronis Cyber Protect Cloud*

Acronis Cyber Protect Cloud Advanced Backup for Workstations là một giải pháp sao lưu và bảo vệ dữ liệu dành cho máy trạm. Đây là một phần mềm mạnh mẽ được cung cấp bởi Acronis, ứng dụng công nghệ tiên tiến để đảm bảo an toàn và khả dụng của dữ liệu trên máy trạm. Dưới đây là một số thông tin về tính năng của Acronis Cyber Protect Cloud Advanced Backup for Workstations:

* Sao lưu toàn diện: Phần mềm cho phép bạn sao lưu toàn bộ hệ điều hành, ứng dụng, tệp tin và thư mục trên máy trạm. Bằng cách tạo các bản sao lưu đầy đủ, nó đảm bảo rằng bạn có thể khôi phục lại máy trạm hoàn toàn sau một sự cố.
* Bảo vệ dữ liệu: Acronis Cyber Protect Cloud Advanced Backup for Workstations cung cấp các tính năng bảo mật để bảo vệ dữ liệu trong quá trình sao lưu và lưu trữ. Các tính năng bao gồm mã hóa dữ liệu, chữ ký số, và cơ chế kiểm tra tính toàn vẹn dữ liệu để đảm bảo rằng dữ liệu của bạn được bảo vệ an toàn.
* Khôi phục linh hoạt: Phần mềm cho phép bạn khôi phục dữ liệu từ các bản sao lưu theo nhiều cách khác nhau. Bạn có thể khôi phục toàn bộ máy trạm, khôi phục các tệp tin và thư mục cụ thể, hoặc khôi phục từng mục tiêu cụ thể như email, danh bạ và ứng dụng.
* Quản lý tập trung: Acronis Cyber Protect Cloud Advanced Backup for Workstations cung cấp giao diện quản lý tập trung để bạn có thể quản lý và giám sát các hoạt động sao lưu của máy trạm. Bạn có thể lên lịch sao lưu tự động, kiểm tra trạng thái sao lưu, xem các báo cáo và cảnh báo, và thực hiện các tác vụ quản lý từ xa.
* Tích hợp đám mây: Phần mềm cho phép sao lưu dữ liệu trực tiếp lên các dịch vụ đám mây phổ biến như Amazon S3, Microsoft Azure và Google Cloud. Điều này giúp bạn lưu trữ dữ liệu sao lưu tại nơi an toàn và dễ dàng khôi phục từ xa.

Acronis Cyber Protect Cloud Advanced Backup for Workstations là một giải pháp đáng tin cậy để sao lưu và bảo vệ dữ liệu trên máy trạm của bạn. Nó cung cấp tính năng linh hoạt, bảo mật và quản lý tập trung để đảm bảo an toàn và khả dụng của dữ liệu trên máy trạm*.*

* + - * 1. *Veeam*

Veeam® Backup & Replication™ giúp doanh nghiệp bảo vệ dữ liệu toàn diện cho tất cả khối lượng công việc như: Virtual, Physical và Cloud. Nó có khả năng sao lưu nhanh chóng, linh hoạt và đáng tin cậy, phục hồi và sao chép tất cả các ứng dụng và dữ liệu.

Đây là một vài tính năng của giải pháp:

* Hỗ trợ sao lưu và phục hồi máy chủ vật lý, máy chủ ảo và máy chủ trên Cloud.
* Hỗ trợ các nền tảng ảo hóa:  VMware, Hyper-V, AHV.
* Hỗ trợ các nền tảng Cloud:  AWS, Azure, Google, IBM.
* Hỗ trợ các ứng dụng SaaS: Microsoft Office 365.
* Hỗ trợ các máy chủ vật lý chạy trên nền tảng Windows, Linux, Mac, UNIX.
* Hỗ trợ các ứng dụng doanh nghiệp như Oracle, SAP, Microsoft.
* Hỗ trợ Kubernetes dạng Native.
* Hỗ trợ sao lưu và phục hồi máy chủ vật lý. Hỗ trợ các máy chủ vật lý chạy trên nền tảng Windows, Linux, Mac, UNIX.
* Có agent backup để cài trên các máy chủ vật lý chạy Windows, Linux cho phép backup theo dạng Entire Machine, Volume-level và File-level. Cho phép tạo đĩa khởi động Custom Recovery Media để có thể phục hồi lại máy chủ.  Có khả năng hỗ trợ khôi phục lại theo các mức: Bare Metal, Volumes, files, folders.
* Có agent backup cài trên các máy chủ IBM AIX để hỗ trợ backup và khôi phục dữ liệu dạng file-level. Cho phép các agent cài trong các AIX LPAR.
* Khả năng backup và phục hồi các ứng dụng (Microsoft Exchange, SharePoint, SQL, Active Directory, Oracle, etc) trong các VM mà không cần cài Agent (Agentless).
* Tính năng điều khiển Backup I/O để tối ưu mạng và hiệu năng của hệ thống backup.
* Tính năng sao lưu bản backup (backup copy).
* Tính năng lập chỉ mục (indexing) và tìm kiếm file trong VM.
* Tính năng backup hoặc sao lưu từ một bản chụp hệ thống lưu trữ (storage snapshots). Khi khôi phục có thể khôi phục theo cả VM, từng file riêng lẻ, hoặc từng items trong ứng dụng.
* Tính năng kiểm tra khả năng khôi phục của bản dữ liệu sao lưu.
* Tính năng tạo bản sao của dữ liệu copy lên cloud hoặc các hệ thống object storage
* Tính năng lưu trữ lâu dài (archiving) lên các hệ thống đám mây hoặc object storage.
* Tính năng tạo ra nhưng phân vùng lưu trữ dữ liệu backup bất biến, không thể xóa sửa trong 1 thời gian nhất định để bảo vệ trước sự tấn công thay đổi dữ liệu của Ransomeware.
* Tính năng đồng bộ VM theo dạng image-base để thực hiện DR mà không cần mua thêm license.
* Hỗ trợ failover và failback.
* Tính năng kiểm tra bản dữ liệu đã được đồng bộ để đảm bảo khi có sự cố thì có thể khôi phục được.
* Tính năng khử lặp dữ liệu và loại bỏ các dữ liệu rác trong swap file.
* Khả năng tích hợp với các hệ thống lưu trữ khử lặp.
* Khả năng khai thác dữ liệu từ các dữ liệu sao lưu cho mục đích kiểm thử.
* Tính năng bảo vệ loại bỏ virus hoặc các phần mềm độc hại ra khỏi dữ liệu sao lưu trước khi khôi.

  3.2 Đề xuất lựa chọn giải pháp công nghệ backup

Phần mềm backup Veeam có khả năng backup linh hoạt, không chỉ backup các máy ảo VM mà con backup các file, database. Phần mềm sao lưu phải hỗ trợ cho phép thay đổi thuật toán data compression để tối ưu cho tốc độ sao lưu hay giảm dung lượng lưu trữ thông qua tăng tỷ lệ nén dữ liệu.

### Khối hạ tầng

#### Máy chủ tính toán (Servers)

Hiệu năng: CPU mạnh mẽ (Intel Xeon, AMD EPYC), dung lượng RAM lớn (từ 256GB trở lên), hỗ trợ ảo hóa.

Lưu trữ: Ổ cứng tốc độ cao (SSD NVMe), hỗ trợ RAID để đảm bảo tính sẵn sàng dữ liệu.

Mạng: Card mạng tốc độ cao (10GbE), hỗ trợ các giao thức mạng tiên tiến.

#### Máy chủ AI machine learning (ML Servers)

* Bộ xử lý (CPU): Các máy chủ AI thường sử dụng CPU mạnh mẽ như Intel Xeon Scalable hoặc AMD EPYC, có nhiều nhân và luồng xử lý để đáp ứng nhu cầu tính toán song song.
* Bộ tăng tốc GPU (GPU Accelerators): Đây là yếu tố quan trọng nhất của máy chủ AI. GPU (Graphics Processing Unit) được thiết kế để xử lý song song một lượng lớn dữ liệu đồ họa, và chúng cũng rất hiệu quả trong việc thực hiện các phép tính AI như ma trận và tensor. Các máy chủ AI thường sử dụng nhiều GPU hiệu năng cao như NVIDIA Tesla.
* Bộ nhớ (Memory): Máy chủ AI cần dung lượng RAM lớn (từ 256GB trở lên) và bộ nhớ băng thông cao (HBM) để xử lý hiệu quả các mô hình AI lớn và tập dữ liệu khổng lồ.
* Lưu trữ (Storage): Máy chủ AI cần hệ thống lưu trữ tốc độ cao như SSD NVMe để đảm bảo tốc độ truy xuất dữ liệu nhanh chóng.
* Mạng (Network): Máy chủ AI cần kết nối mạng tốc độ cao (10GbE) để truyền tải dữ liệu giữa các máy chủ và với hệ thống lưu trữ.

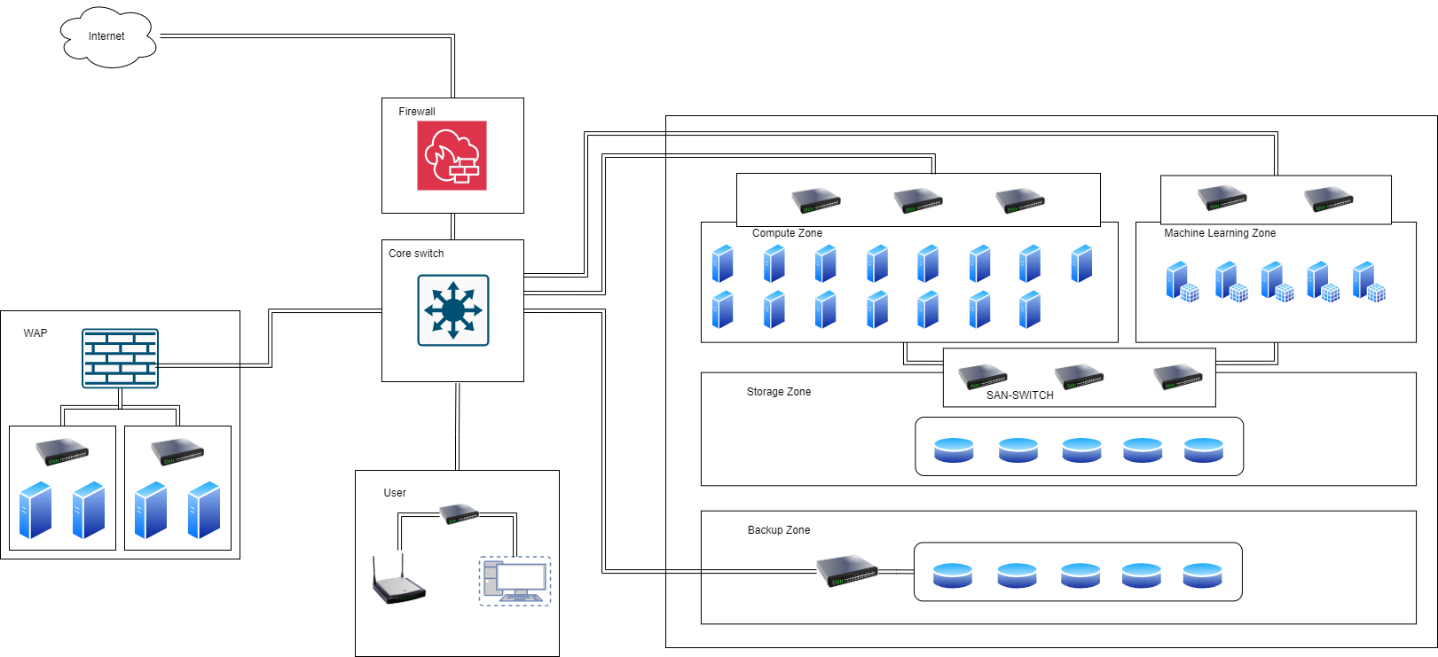
#### Thiết bị lưu trữ (Storage)

* Hệ thống SAN: Dung lượng lớn, tốc độ cao, sử dụng lưu trữ cho các ứng dụng trên máy chủ tính toán và máy chủ AI.
* Hệ thống NAS: Dung lượng lớn, tốc độ không cao, hỗ trợ các giao thức chia sẻ file và block.

#### Thiết bị mạng (Network)

* Switch: Kết nối các máy chủ và thiết bị lưu trữ, hỗ trợ các tính năng quản lý và bảo mật mạng.
* Router: Định tuyến lưu lượng mạng giữa các mạng con và mạng bên ngoài.
* Firewall: Bảo vệ hệ thống khỏi các cuộc tấn công mạng.
* Load balancer: Phân phối lưu lượng truy cập đều giữa các máy chủ, đảm bảo tính sẵn sàng cao.

#### Mô hình vật lý hệ thống



1. Mô hình vật lý hệ thống

Mô hình vật lý được chia thành các khối sau:

* Compute zone: Cụm máy chủ tính toán phục vụ phân tích và xử lý dữ liệu.
* Machine Learning zone: Cụm máy chủ đặc thù cho việc huấn luyện và sử dụng các mô hình AI.
* Storage zone: Cụm máy chủ phục vụ lưu trữ dữ liệu.
* Backup zone: Cụm máy chủ backup dữ liệu.
* WAF: Cụm máy chủ phục vụ các ứng dụng trên mạng ngoài có tường lửa phân vùng mạng ngoài và mạng trong.
* User: Các cụm máy vùng trong cho phép người dùng tương tác với kho dữ liệu.
* Firewall: Máy chủ quản lý tường lửa.
* Core Switch: Switch tổng của hệ thống, cho phép truy cập vào hệ thống kho.

# DỰ TOÁN

## Tổng mức đầu tư:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **Khoản Mục Chi phí** | **Thành tiền chưa có thuế** | **Thuế giá trị gia tăng** | **Thành tiền đã có thuế** |
| **I** | Chi phí xây lắp |  |  |  |
| **II** | Chi phí thiết bị. phần mềm. CSDL. đào tạo | 83.129.845.000 | 4.762.984.500 | **87.892.829.500** |
| **III** | Chi phí quản lý dự án | 1.278.860.909 | 127.886.091 | **1.406.747.000** |
| **IV** | Chi phí tư vấn đầu tư | 2.883.463.547 | 288.346.355 | **3.171.809.902** |
| **V** | Chi phí khác | 3.210.866.292 | 313.379.485 | **3.524.245.777** |
| **VI** | Chi phí dự phòng |  |  |  |
|  | **Tổng cộng** |  |  | **95.995.632.179** |
|  | **Tổng cộng( Làm tròn)** |  |  | **95.995.633.000** |

## Tóm tắt dự toán các hạng mục chính

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Nội dung chi phí** | **Đơn vị tính** | **Số lượng** | **Đơn giá** | **Thành tiền trước thuế** | **Thuế VAT** | **Giá trị thuế GTGT** | **Thành tiền sau thuế** |
| **I** | **CHI PHÍ THIẾT BỊ** |  |  |  | **83.129.845.000** |  | **4.762.984.500** | **87.892.829.500** |
| **1** | **Hạng mục hạ tầng kỹ thuật. CNTT** |  |  |  | **47.128.000.000** | **0%, 10%** | **4.712.800.000** | **51.840.800.000** |
| 1.1 | Máy chủ tính toán | Chiếc | 15 | 950.000.000 | 14.250.000.000 | 10% | 1.425.000.000 | 15.675.000.000 |
| 1.2 | Máy chủ AI | Chiếc | 5 | 1.100.100.000 | 5.500.500.000 | 10% | 550.050.000 | 6.050.550.000 |
| 1.3 | Thiết bi lưu trữ storage | Chiếc | 10 | 950.000.000 | 9.500.000.000 | 10% | 950.000.000 | 10.450.000.000 |
| 1.4 | Thiết bị NAS | Chiếc | 10 | 1.500.000.000 | 15.000.000.000 | 10% | 1.500.000.000 | 16.500.000.000 |
| 1.5 | Thiết bị SAN switch | Chiếc | 5 | 220.000.000 | 1.100.000.000 | 10% | 110.000.000 | 1.210.000.000 |
| 1.6 | Core Switch 10Gbe | Chiếc | 5 | 250.000.000 | 1.250.000.000 | 10% | 125.000.000 | 1.375.000.000 |
| 1.7 | Firewall | Chiếc | 1 | 527.500.000 | 527.500.000 | 10% | 52.750.000 | 580.250.000 |
| **2** | **Hạng mục phần mềm nội bộ, CSDL** |  |  |  | **35.500.000.000** | **0%** | **0** | **35.500.000.000** |
| 2.1 | Khối chức năng thu thập dữ liệu | Phần mềm | 1 | 3.000.000.000 | 3.000.000.000 | 0% | 0 | 3.000.000.000 |
| 2.2 | Khối chức năng xử lý dữ liệu | Phần mềm | 1 | 4.000.000.000 | 4.000.000.000 | 0% | 0 | 4.000.000.000 |
| 2.3 | Khối chức năng quản lý, huấn luyện và sử dụng mô hình AI | Phần mềm | 1 | 8.000.000.000 | 8.000.000.000 | 0% | 0 | 8.000.000.000 |
| 2.4 | Khối chức năng quản lý dữ liệu | Phần mềm | 1 | 5.000.000.000 | 5.000.000.000 | 0% | 0 | 5.000.000.000 |
| 2.5 | Khối chức năng trực quan dữ liệu | Phần mềm | 1 | 5.000.000.000 | 5.000.000.000 | 0% | 0 | 5.000.000.000 |
| 2.6 | Khối chức năng trực quan quản trị. cấp phát tài nguyên hệ thống | Phần mềm | 1 | 6.000.000.000 | 6.000.000.000 | 0% | 0 | 6.000.000.000 |
| 2.7 | Khối chức năng quản lý chia sẻ dữ liệu dùng chung | Phần mềm | 1 | 4.500.000.000 | 4.500.000.000 | 0% | 0 | 4.500.000.000 |
| **3** | **Licence phần mềm backup** | Gói | **2** | **162.500.000** | **162.500.000** | **10%** | **17.684.500** | **194.529.500** |