**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

-----🙞🙜🕮🙞🙜-----

****

**KHÓA LUẬN TỐT NGHIỆP**

**NĂM HỌC 2022-2023**

**Đề tài:**

**XÂY DỰNG HỒ DỮ LIỆU (DATA LAKE) CHO**

**TẬP ĐOÀN BƯU CHÍNH VIỄN THÔNG VIỆT NAM**

|  |  |
| --- | --- |
| **Sinh viên thực hiện:** | **Bùi Thu Hà** |
| **Lớp:** | **K22HTTTA** |
| **Mã sinh viên:** | **22A4040011** |
| **Giảng viên hướng dẫn:** | **Ths.NCS Ngô Thùy Linh** |

**Hà Nội, tháng 05 năm 2023**

**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

-----🙞🙜🕮🙞🙜-----

****

**KHÓA LUẬN TỐT NGHIỆP**

**NĂM HỌC 2022-2023**

**Đề tài:**

**XÂY DỰNG HỒ DỮ LIỆU (DATA LAKE) CHO**

**TẬP ĐOÀN BƯU CHÍNH VIỄN THÔNG VIỆT NAM**

|  |  |
| --- | --- |
| **Sinh viên thực hiện:** | **Bùi Thu Hà** |
| **Lớp:** | **K22HTTTA** |
| **Mã sinh viên:** | **22A4040011** |
| **Giảng viên hướng dẫn:** | **Ths.NCS Ngô Thùy Linh** |

**Hà Nội, tháng 05 năm 2023**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

**NHẬN XÉT CỦA GVHD**

**LỜI CẢM ƠN**

Nhà văn người Mỹ William Arthur Ward đã từng nói “Cảm thấy biết ơn và không thể hiện điều đó giống như gói quà mà không trao”. Lời đầu tiên trong bài khóa luận này, em xin được gửi những lời cảm ơn đến thầy cô, nhà trường, đơn vị thực thực tập và gia đình, những người đã giúp đỡ em rất nhiều trong quá trình hoàn thiện bài khóa luận này. Trước hết, em xin gửi lời cảm ơn tới Ths.NCS Ngô Thùy Linh, giảng viên đã trực tiếp hướng dẫn em thực hiện bài khóa luận này. Cô đã dành nhiều thời gian và tâm huyết để hướng dẫn, hỗ trợ, chỉ dạy tận tình cho em từ khi lựa chọn đề tài khóa luận đến khi hoàn thành bài khóa luận tốt nghiệp. Cô không chỉ giúp đỡ em trong việc xây dựng nội dung bài luận mà còn hướng dẫn em cách trình bày bài một cách khoa học và hàn lâm nhất. Cùng với đó, em cũng xin được thể hiện lời biết ơn sâu sắc tới toàn thể các thầy, cô là giảng viên của khoa Hệ thống thông tin quản lý Học viện Ngân hàng. Các thầy, cô là những người đã truyền tải kiến thức cũng như các kỹ năng về chuyên môn và cuộc sống cho tất cả sinh viên, đó chính là những nền tảng vững chắc để em có thể hoàn thiện bài khóa luận tốt nghiệp của mình.

Bên cạnh đó, em cũng xin gửi lời cảm ơn chân thành nhất đến Công ty CP Phát Triển Công Nghệ Viễn Thông Tin Học Sun Việt (SVTech) nói chung và phòng D&A nói riêng đã tạo điều kiện cho em có cơ hội được thực tập để học hỏi và tích lũy thêm kiến thức cũng như kinh nghiệm làm việc. Từ lúc được gia nhập vào công ty, anh chị đã tạo cơ hội cho em được tiếp xúc và làm việc với nhiều công việc mới, trong suốt qua trình hoàn thành bài khóa luận anh chị cũng đã đưa ra nhiều lời khuyên và góp ý, đó chính là nguồn ý tưởng, tài liệu và động lực để em có thể hoàn thiện bài luận này.

Ngoài ra, em cũng xin gửi lời cảm ơn tới mọi người trong gia đình và bạn bè đã là luôn đồng hành cùng em trong suốt qua trình học tập từ trước đến nay. Mọi người chính là nguồn động lực chính để em vượt qua được mọi khó khăn và đi đến ngày hôm nay.

Do hạn chế về mặt thời gian cũng như kiến thức và kinh nghiệm nên đề tài vẫn còn nhiều thiếu sót nên em rất mong nhận được những lời góp ý, nhận xét từ phía thầy cô. Em xin chân thành cảm ơn!

**LỜI CAM ĐOAN**

Em xin cam đoan toàn bộ nội dung trong bài khóa luận tốt nghiệp này là sản phầm nghiên cứu, tìm hiểu của cá nhân em trong quá trình được học tập và thực tập tại Công ty CP Phát Triển Công Nghệ Viễn Thông Tin Học Sun Việt (SVTech). Trong bài luận này, bên cạnh những nội dung là kiến thức của bản thân em cũng đã tham khảo một số nguồn tài liệu công khai, đảm bảo uy tín. Tất cả các tài liệu tham khảo trong bài đều được trích dẫn một cách hợp pháp.

Em xin chịu trách nhiệm hoàn toàn và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

**Sinh viên thực hiện**

**Bùi Thu Hà**

**DANH MỤC VIẾT TẮT**

|  |  |
| --- | --- |
| **Từ viết tắt** | **Diễn giải** |
| AI | Artificial Intelligence – Trí tuệ nhân tạo |
| CP4D | Cloud Pak for Data |
| CSDL | Cơ sở dữ liệu |
| CSHT | Cơ sở hạ tầng |
| DB | Database – Cơ sở dữ liệu |
| DL | Deep Learning |
| DWH | Data Warehouse |
| ESS | Enterprise Storage Server |
| ICOS | IBM Cloud Object Storage |
| LDAP | Lightweight Directory Access Protocol |
| ML | Machine Learning |
| OLAP | Online Analysis Processing – Xử lý phân tích trực tuyến |
| OLTP | Online Transaction Processing – Xử lý giao dịch trực tuyến |
| WKC | Watson Knowledge Catalog |

**DANH MỤC BẢNG**

[Bảng 1: Điểm khác biệt giữa Hồ dữ liệu và Kho dữ liệu 17](#_Toc135479731)

[Bảng 2: Các thành phần cần xác định khi triển khai quản trị dữ liệu 54](#_Toc135479732)

[Bảng 3: Chi tiết số lượng thiết bị lắp đặt theo Rack 55](#_Toc135479733)

[Bảng 4: Chi tiết các sản phẩm được sử dụng trong hồ dữ liệu 64](#_Toc135479734)

[Bảng 5: Thông tin nguồn dữ liệu 65](#_Toc135479735)

[Bảng 6: Các giai đoạn biến đổi được DataStage hỗ trợ 74](#_Toc135479736)

[Bảng 7: Các groups trong hệ thống LDAP 90](#_Toc135479737)

**DANH MỤC HÌNH**

[Hình 1: Các mốc phát triển của Tập đoàn VNPT từ 2016 đến 2018 (VNPT) 6](#_Toc135479740)

[Hình 2: Các mốc phát triển của Tập đoàn VNPT từ 2012 đến 2015 (VNPT) 7](#_Toc135479741)

[Hình 3: Các mốc phát triển của Tập đoàn VNPT từ 2009 đến 2011 (VNPT) 8](#_Toc135479742)

[Hình 4: Các mốc phát triển của Tập đoàn VNPT từ 2006 đến 2008 (VNPT) 9](#_Toc135479743)

[Hình 5: Các mốc phát triển của Tập đoàn VNPT từ 1995 đến 1997 (VNPT) 10](#_Toc135479744)

[Hình 6: Sơ đồ tổ chức bộ máy của Tập đoàn VNPT 11](#_Toc135479745)

[Hình 7: Kiến trúc tham chiếu cho AI Data Lake 22](#_Toc135479746)

[Hình 8: Kiến trúc phân cấp của IBM cho AI Data Lake 26](#_Toc135479747)

[Hình 9: Thách thức của Data Ingestion 27](#_Toc135479748)

[Hình 10: Quy trình lập kế hoạch triển khai hồ dữ liệu 30](#_Toc135479749)

[Hình 11: Quy trình xây dựng Data Lake 36](#_Toc135479750)

[Hình 12: Luồng hoạt động Change Data Capture 38](#_Toc135479751)

[Hình 13: Thứ tự các loại phân tích theo độ khó và giá trị trong phân tích dữ liệu 40](#_Toc135479752)

[Hình 14: Kiến trúc IBM Information Server 43](#_Toc135479753)

[Hình 15: Kiến trúc quản lý các dịch vụ với SingleStore 48](#_Toc135479754)

[Hình 16: Node trong SingleStore 49](#_Toc135479755)

[Hình 17: Cấu tạo Universal Storage 50](#_Toc135479756)

[Hình 18: Đặc điểm của SingleStore 50](#_Toc135479757)

[Hình 19: Sơ đồ mạng của hệ thống Data Lake 57](#_Toc135479758)

[Hình 20: Mô hình Openshift cho triển khai cụm Cloud Pak for Data 59](#_Toc135479759)

[Hình 21: Kiến trúc thành phần chức năng của nền tảng Big Data 60](#_Toc135479760)

[Hình 22: Kiến trúc hệ thống hồ dữ liệu tương ứng với giải pháp IBM 61](#_Toc135479761)

[Hình 23: Kiến trúc tổng quan về hồ dữ liệu cho tập đoàn VNPT 63](#_Toc135479762)

[Hình 24: Tổng quan luồng dữ liệu 67](#_Toc135479763)

[Hình 25: Tạo pipeline nhập dữ liệu streaming 70](#_Toc135479764)

[Hình 26: Pipeline running 71](#_Toc135479765)

[Hình 27: Tạo connection mới 73](#_Toc135479766)

[Hình 28: Tạo kết nối tới SingleStore 74](#_Toc135479767)

[Hình 29: Tạo luồng nhập dữ liệu từ nguồn tới đích với DataStage 77](#_Toc135479768)

[Hình 30: Thiết lập lịch chạy cho job 78](#_Toc135479769)

[Hình 31: Lịch chạy của job được tạo thành công 78](#_Toc135479770)

[Hình 32: Tổng quan các chức năng ở tầng tính toán/xử lý và lưu trữ 79](#_Toc135479771)

[Hình 33: Dữ liệu lưu trữ tại ESS 80](#_Toc135479772)

[Hình 34: Dữ liệu lưu trữ tại ICOS 80](#_Toc135479773)

[Hình 35: SingleStore kết hợp OLTP và OLAP 81](#_Toc135479774)

[Hình 36: Mô tả Universal Storage 82](#_Toc135479775)

[Hình 37: Tỷ lệ nén dữ liệu thực tế của SingleStore với các bảng dữ liệu của VNPT 83](#_Toc135479776)

[Hình 38: Thông tin các node trong SingleStore 83](#_Toc135479777)

[Hình 39: PowerBI kết nối tới SingleStore 84](#_Toc135479778)

[Hình 40: Thông tin instance Analytics Engine Powered by Apache Spark 85](#_Toc135479779)

[Hình 41: Submit job thành công 86](#_Toc135479780)

[Hình 42: Advanced feature được enable 86](#_Toc135479781)

[Hình 43: Deployment space được tạo gán với instance 87](#_Toc135479782)

[Hình 44: Thông tin các App khi job được submit thành công 88](#_Toc135479783)

[Hình 45: Spark History Server lưu log của mỗi App 88](#_Toc135479784)

[Hình 46: Storage Volume lưu event log của App 88](#_Toc135479785)

[Hình 47: Mô hình tích hợp LDAP với hệ thống hồ dữ liệu 89](#_Toc135479786)

[Hình 48: Tích hợp hệ thống LDAP với hệ thống OpenShift 91](#_Toc135479787)

[Hình 49: Tích hợp hệ thống Storage 92](#_Toc135479788)

[Hình 50: Kiến trúc multi-tenancy cho VNPT 93](#_Toc135479789)

[Hình 51: Các mức thiết lập theo định hướng multi-tenancy của SingleStore 94](#_Toc135479790)

[Hình 52: Chính sách và các quy tắc bảo vệ dữ liệu 96](#_Toc135479791)

[Hình 53: Dữ liệu sau khi áp dụng chính sách và quy tắc bảo vệ 96](#_Toc135479792)

[Hình 54: Dữ liệu được đưa vào vùng ingestion 97](#_Toc135479793)

[Hình 55: Dữ liệu được ingest near real time 97](#_Toc135479794)

[Hình 56: Ingest dữ liệu dạng batch 98](#_Toc135479795)

[Hình 57: Tổng số lỗi khi ingest dữ liệu với 4.514.794.555 bản ghi 98](#_Toc135479796)

[Hình 58: Kết quả truy vấn 99](#_Toc135479797)

**MỤC LỤC**

[**LỜI MỞ ĐẦU** 1](#_Toc134979560)

[1. Lý do chọn đề tài 1](#_Toc134979561)

[2. Mục đích nghiên cứu 2](#_Toc134979562)

[3. Đối tượng và phạm vi nghiên cứu 2](#_Toc134979563)

[4. Phương pháp nghiên cứu 2](#_Toc134979564)

[5. Bố cục đề tài 2](#_Toc134979565)

[**CHƯƠNG 1: KHÁI QUÁT VỀ TẬP ĐOÀN** 4](#_Toc134979566)

[**BƯU CHÍNH VIỄN THÔNG VIỆT NAM VNPT** 4](#_Toc134979567)

[**VÀ BÀI TOÁN GIẢI QUYẾT** 4](#_Toc134979568)

[1.1. Giới thiệu chung Tập đoàn Bưu chính Viễn thông Việt Nam VNPT 4](#_Toc134979569)

[1.2. Tầm nhìn, sứ mệnh, giá trị cốt lõi 4](#_Toc134979570)

[1.3. Lịch sử hình thành và phát triển 5](#_Toc134979571)

[1.4. Lĩnh vực hoạt động 10](#_Toc134979572)

[1.5. Sơ đồ tổ chức 10](#_Toc134979573)

[1.6. Mô tả bài toán 11](#_Toc134979574)

[1.6.1. Thực trạng 11](#_Toc134979575)

[1.6.2. Thách thức 12](#_Toc134979576)

[1.6.3. Giá trị mang lại của bài toán 13](#_Toc134979577)

[1.6.4. Hướng giải quyết 13](#_Toc134979578)

[1.7. Kết luận chương 1 15](#_Toc134979579)

[**CHƯƠNG 2: TỔNG QUAN LÝ THUYẾT DATA LAKE** 16](#_Toc134979580)

[**VÀ CÁC CÔNG CỤ SỬ DỤNG** 16](#_Toc134979581)

[2.1. Khái niệm, kiến trúc và đặc điểm của Data Lake 16](#_Toc134979582)

[2.2. Sự khác biệt giữa Data Lake và Data Warehouse 17](#_Toc134979583)

[2.3. Lợi ích của Data Lake 20](#_Toc134979584)

[2.4. Một số thách thức khi triển khai Data Lake 21](#_Toc134979585)

[2.5. Kiến trúc tham chiếu cho AI Data Lake 22](#_Toc134979586)

[2.6. Kiến trúc phân cấp của IBM cho AI Data Lake 25](#_Toc134979587)

[2.7. Quy trình lập kế hoạch và xây dựng Data Lake 30](#_Toc134979588)

[2.7.1. Quy trình lập kế hoạch triển khai Data Lake 30](#_Toc134979589)

[2.7.2. Quy trình xây dựng Data Lake 36](#_Toc134979590)

[2.8. Một số công cụ sử dụng 41](#_Toc134979591)

[2.8.1. IBM Information Server 41](#_Toc134979592)

[2.8.2. IBM Elastic Storage System 45](#_Toc134979593)

[2.8.3. IBM Cloud Object Storage 46](#_Toc134979594)

[2.8.4. CP4D – Walston Knowledge Catalog 46](#_Toc134979595)

[2.8.5. CP4D – Analytics Engine 47](#_Toc134979596)

[2.8.6. CP4D – Virtualization 47](#_Toc134979597)

[2.8.7. SingleStore 48](#_Toc134979598)

[2.9. Kết luận chương 2 51](#_Toc134979599)

[**CHƯƠNG 3: LẬP KẾ HOẠCH VÀ XÂY DỰNG DATA LAKE** 52](#_Toc134979600)

[**CHO TẬP ĐOÀN VNPT** 52](#_Toc134979601)

[3.1. Lập kế hoạch xây dựng Data Lake 52](#_Toc134979602)

[3.1.1. Xác định thách thức 52](#_Toc134979603)

[3.1.2. Phát triển và triển khai giải pháp chiến lược 52](#_Toc134979604)

[3.1.3. Dự đoán sự tăng trưởng của dữ liệu 55](#_Toc134979605)

[3.1.4. Lên kế hoạch về CSHT 55](#_Toc134979606)

[3.1.5. Xác định chiến lược hoạt động 62](#_Toc134979607)

[3.2. Triển khai xây dựng Data Lake cho tập đoàn VNPT 62](#_Toc134979608)

[3.2.1. Kiến trúc tổng quan 63](#_Toc134979609)

[3.2.2. Xác định nguồn dữ liệu và người dùng hệ thống 65](#_Toc134979610)

[3.2.3. Chiến lược thu thập dữ liệu 67](#_Toc134979611)

[3.2.4. Lưu trữ dữ liệu 78](#_Toc134979612)

[3.2.5. Phân tích dữ liệu 80](#_Toc134979613)

[3.2.6. Quản lý và quản trị dữ liệu 89](#_Toc134979614)

[3.3. Kết quả đạt được sau khi triển khai 96](#_Toc134979615)

[3.4. Kết luận chương 3 99](#_Toc134979616)

[**KẾT LUẬN** 101](#_Toc134979617)

[**TÀI LIỆU THAM KHẢO** 103](#_Toc134979618)

# **LỜI MỞ ĐẦU**

1. Lý do chọn đề tài

Tập đoàn Bưu chính Viễn thông Việt Nam (VNPT) là một trong những tập đoàn viễn thông lớn nhất tại Việt Nam với nhiều công ty thành viên và hoạt động trong nhiều lĩnh vực khác nhau, từ viễn thông, viễn thông quân đội, viễn thông vệ tinh, internet, truyền thông đa phương tiện và dịch vụ giá trị gia tăng. Với sự số hóa mạnh mẽ hiện nay, dữ liệu đang trở thành tài sản vô cùng quý giá đối với các doanh nghiệp. Việc sử dụng và phân tích dữ liệu trở thành yếu tố quan trọng giúp các doanh nghiệp đưa ra quyết định kinh doanh chính xác hơn. Đối với tập đoàn VNPT cũng vậy, với lượng dữ liệu lớn và đa dạng từ các hoạt động kinh doanh của mình, VNPT đã sớm nhận ra nếu biết tận dụng đúng cách thì vô vàn cơ hội sẽ được mở ra cho tập đoàn. Bên cạnh những cơ hội, với lượng dữ liệu rất lớn, tốc độ tăng trưởng nhanh và đa dạng, VNPT cũng gặp phải nhiều thách thức trong việc quản lý, lưu trữ và sử dụng dữ liệu. Do đó, tập đoàn đang rất cần một giải pháp đáp ứng được: lưu trữ lượng dữ liệu khổng lồ, đảm bảo hiệu suất và dễ dàng mở rộng; thống nhất các nguồn dữ liệu và tránh tình trạng thông tin silo; phục vụ cả OLAP và OLTP; đảm bảo tính bảo mật cao.

Một trong những xu hướng hiện nay trên thế giới để đáp ứng những nhu cầu về lưu trữ, quản lý, sử dụng dữ liệu lớn đó chính là hồ dữ liệu. Trên thế giới đã có nhiều doanh nghiệp trong lĩnh vực viễn thông đã ứng dụng hồ dữ liệu và đem lại những kết quả vượt trội như: AT&T[[1]](#footnote-1) với giải pháp Data Lake đã giúp họ bảo vệ 182 triệu khách hàng không dây khỏi bị tổn hại (giảm tới 80% các cuộc tấn công gian lận), tiết kiệm hàng triệu đô la (Hopkins); Data Lake giúp Vodafone[[2]](#footnote-2): cập nhật mạng toàn châu Âu nhanh hơn, năng động hơn và hiệu quả hơn; tám tỷ nguồn dữ liệu mạng mỗi ngày giúp cải thiện các quyết định liên quan đến mạng; tăng cường sử dụng lập kế hoạch thông minh để triển khai mạng rẻ hơn (Vodafone, 2022). Dựa vào những kết quả tích cực của việc sử dụng hồ dữ liệu hiện nay của các doanh nghiệp trên thế giới. Có thể thấy xây dựng một hệ thống hồ dữ liệu sẽ giúp tập đoàn quản lý dữ liệu một cách hiệu quả hơn, từ đó đưa ra những quyết định kinh doanh chính xác hơn. Hồ dữ liệu cũng cho phép tập đoàn lưu trữ các dữ liệu có tính chất khác nhau từ các nguồn khác nhau và thực hiện phân tích dữ liệu trên một nền tảng chung. Ngoài ra, việc xây dựng một hệ thống hồ dữ liệu cho VNPT còn đặc biệt quan trọng trong bối cảnh mà Việt Nam đang trở thành một trong những điểm đến hấp dẫn đối với các nhà đầu tư nước ngoài, đồng thời cũng đang phát triển chủ đạo các khu công nghiệp thông minh, điều này đòi hỏi các doanh nghiệp phải tận dụng tối đa dữ liệu để tăng cường hiệu quả và cạnh tranh trong thị trường. Việc xây dựng hồ dữ liệu cho tập đoàn VNPT sẽ giúp tập đoàn tiên phong trong việc áp dụng công nghệ tiên tiến, từ đó củng cố thế vị của tập đoàn trong lĩnh vực viễn thông và kinh doanh điện tử. Từ những nhu cầu cũng như thực trạng trên, đề tài “Xây dựng Hồ dữ liệu (Data Lake) cho Tập đoàn Bưu chính viễn thông Việt Nam” được lựa chọn.

1. Mục đích nghiên cứu

Mục đích chính của đề tài là phân tích và đưa ra kế hoạch triển khai xây dựng cụ thể Hồ dữ liệu (Data Lake) cho Tập đoàn Bưu chính viễn thông Việt Nam. Để có thể thực hiện mục tiêu chính trên, đề tài thực hiện các mục tiêu cụ thể như sau:

* Nghiên cứu thực trạng của doanh nghiệp và các lý thuyết về Data Lake
* Xây dựng quy trình lập kế hoạch và triển khai xây dựng Data Lake
* Xây dựng giải pháp Data Lake cho tập đoàn dựa trên giải pháp của IBM

Từ những mục tiêu trên, nghiên cứu này mong muốn đưa ra giải pháp và hướng phát triển Data Lake phù hợp với tập đoàn viễn thông, giúp tăng cường khả năng quản lý và khai thác dữ liệu của tập đoàn, đồng thời nâng cao năng lực cạnh tranh và đáp ứng nhu cầu của thị trường. Từ đó, cũng làm cơ sở, nền tảng cho việc triển khai xây dựng Data Lake cho các doanh nghiệp khác.

1. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: các thành phần trong giải pháp hồ dữ liệu của IBM được xây dựng tại Tập đoàn Bưu chính viễn thông Việt Nam và các quy trình, nghiệp vụ liên quan đến việc lưu trữ, quản lý và sử dụng dữ liệu tại VNPT.

Phạm vi nghiên cứu: đề tài được thực hiện nghiên cứu việc xây dựng hồ dữ liệu tại Tập đoàn Bưu chính viễn thông Việt Nam trong thời gian từ tháng 8 năm 2022 đến tháng 5 năm 2023.

1. Phương pháp nghiên cứu

* Phương pháp phân tích – tổng hợp
* Phương pháp so sánh
* Phương pháp liệt kê
* Phương pháp thống kê số liệu
* Phương pháp nghiên cứu thực nghiệm trực tiếp trên giải pháp hồ dữ liệu của IBM

1. Bố cục đề tài

Bài khóa luận có bố cục gồm mở đầu, nội dung chính và kết luận. Trong phần nội dung chính có các nội dung như sau:

Chương 1: Khái quát về tập đoàn bưu chính viễn thông Việt Nam VNPT và bài toán giải quyết.

Chương 2: Tổng quan lý thuyết Data Lake và các công cụ sử dụng.

Chương 3: Lập kế hoạch và xây dựng Data Lake cho tập đoàn VNPT.

# **CHƯƠNG 1: KHÁI QUÁT VỀ TẬP ĐOÀN**

# **BƯU CHÍNH VIỄN THÔNG VIỆT NAM VNPT**

# **VÀ BÀI TOÁN GIẢI QUYẾT**

1. Giới thiệu chung Tập đoàn Bưu chính Viễn thông Việt Nam VNPT

Tập đoàn Bưu chính Viễn thông Việt Nam với tên giao dịch quốc tế Vietnam Posts and Telecommunications Group (viết tắt: VNPT) là một tập đoàn kinh tế nhà nước chủ yếu đầu tư, sản xuất, kinh doanh trong lĩnh vực viễn thông và công nghệ thông tin tại Việt Nam. “Tập đoàn được thành lập vào ngày 30/04/1995 và đến năm 2006 được chính thức thay đổi thành mô hình tập đoàn thay cho mô hình cũ là Tổng công ty theo quyết định số 06/2006/QĐ-TTg của Thủ tướng Chính Phủ.”

Trụ sở chính tọa lạc tại: Tòa nhà VNPT, số 57 phố Huỳnh Thúc Kháng, phường Láng Hạ, quận Đống Đa, thành phố Hà Nội.

Điện thoại: 84-24 3 774 1091

Fax: 84-24 3 774 1093

Website: [www.vnpt.comvn](http://www.vnpt.comvn)

Email: [vanphong@vnpt.vn](mailto:vanphong@vnpt.vn)

VNPT hiện nay là một doanh nghiệp có hạ tầng mạng lưới vững mạnh với các mảng viễn thông quốc tế, mạng băng thông cố định, mạng thông tin di động, hệ thống vệ tinh Vinasat. “Đối với mạng viên thông quốc tế, tập đoàn có 5 tuyến cáp quang biển quốc tế là SMW-3, AAG, Faster, APG và AAE-1 kéo dài từ châu Á đến châu Phi và châu Âu. Mạng băng thông cố định với dịch vụ truy nhập Internet băng rộng ADSL, xDSL tốc độ lên tới 15 Mbps trên khắp 63/63 tỉnh thành. Mạng thông tin di động có trên 75.000 trạm thu phát sóng 2G,3G,4G,5G, phủ sóng 63/63 tỉnh thành trên cả nước. Riêng với hệ thống vệ tinh Vinasat: VNPT là doanh nghiệp viễn thông duy nhất tại Việt Nam được Nhà nước và Chính phủ tin tưởng giao trọng trách bao gồm đầu tư và quản lý hệ thống vệ tinh viễn thông của Việt Nam.” (Cổng thông tin điện tử Ủy ban quản lý vốn nhà nước tại doanh nghiệp)

1. Tầm nhìn, sứ mệnh, giá trị cốt lõi

Là một doanh nghiệp viễn thông và công nghệ thông tin, với xu thế trong thời đại mới toàn cầu hóa và phát triển không ngừng của cuộc cách mạng công nhiệp 4.0. Tập đoàn luôn giữ vững phong thế đi đầu trong việc chuyển đổi số từ nhà cung cấp dịch vụ viễn thông truyền thống sang nhà cung cấp dịch vụ số Digital service provider (viết tắt: DSP). Với mục tiêu trở thành Tập đoàn kinh tế năng động, trọng tâm phát triển các sản phẩm, dịch vụ và giải pháp Information Communication Technology (ICT) sáng tạo, mang tính đột phá để kiến tạo nên những giá trị đích thực cho cuộc sống và góp phần vào công cuộc xây dựng, phát triển nền kinh tế vững mạnh của Việt Nam.

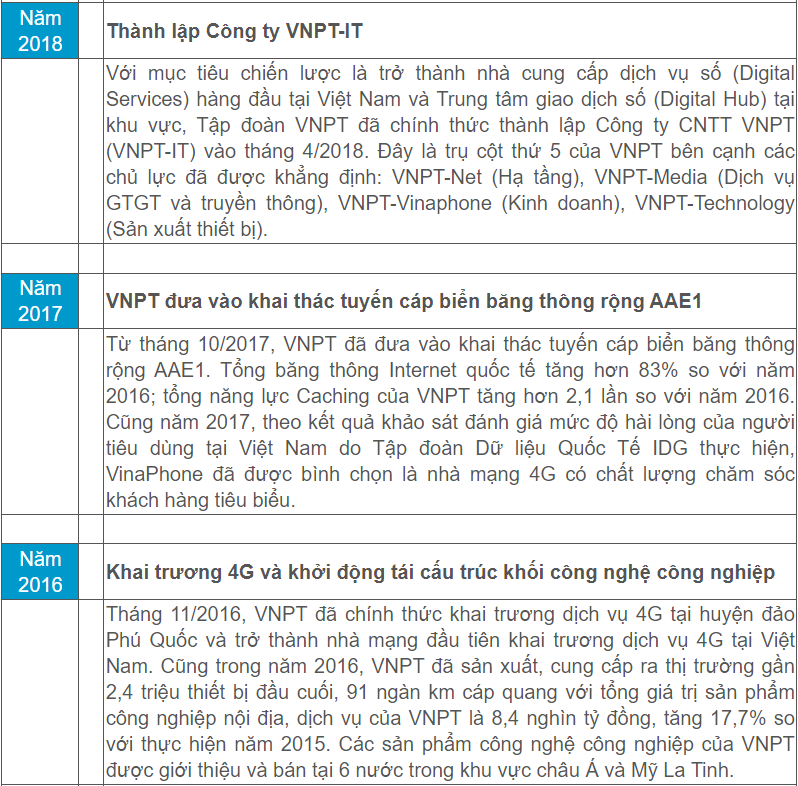
Chính vì vậy, tập đoàn đặt ra sứ mệnh:

* Cung cấp cho khách hàng và đối tác các sản phẩm, dịch vụ Viễn thông - CNTT - Truyền thông và Dịch vụ số chất lượng, đột phá, sáng tạo.
* Trở thành trung tâm số (Digital Hub) của khu vực châu Á.
* Tôn vinh và đánh giá giá trị đích thực của người lao động.
* Tiên phong trong các hoạt động về cộng đồng.

Cùng với những sứ mệnh đó là tầm nhìn chiến lược: (1) Trở thành nhà cung cấp dịch vụ số hàng đầu Việt Nam vào năm 2025, (2) trở thành Trung tâm số (Digital Hub) của châu Á vào năm 2030, (3) Trở thành lựa chọn số 1 của khách hàng sử dụng sản phẩm, dịch vụ Công nghệ thông tin - Viễn thông (ICT) tại thị trường. (VNPT)

1. Lịch sử hình thành và phát triển

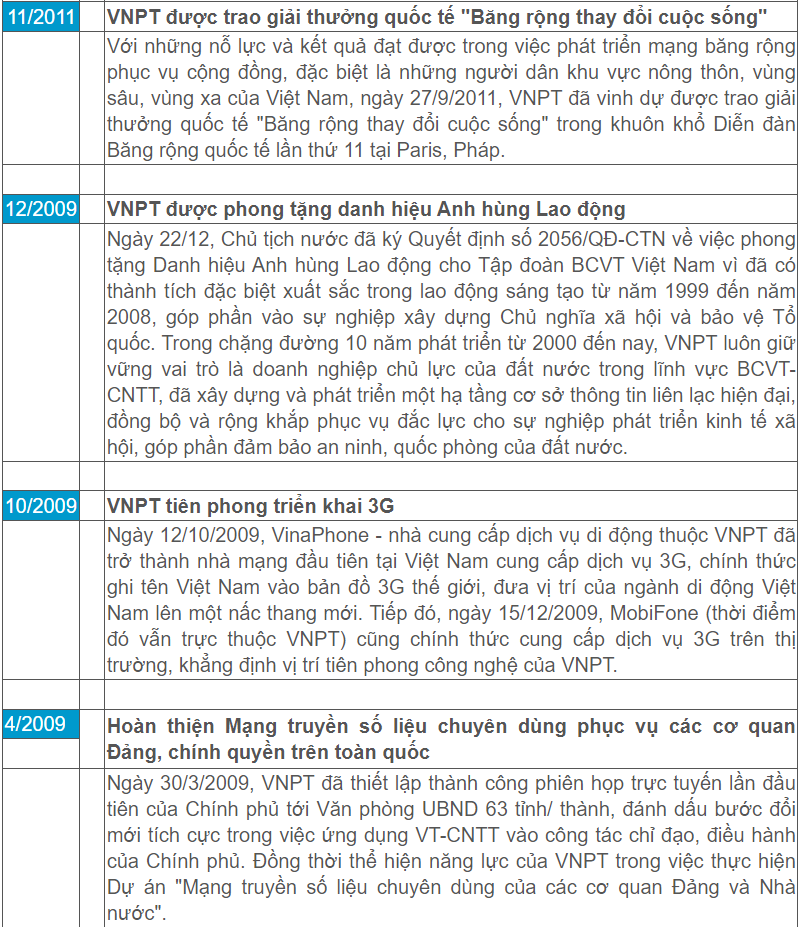
Thành lập từ năm 1995, đến nay tập đoàn VNPT đã có lịch sử hình thành và phát triển 28 năm với những dấu mốc phát triển đáng nhớ.



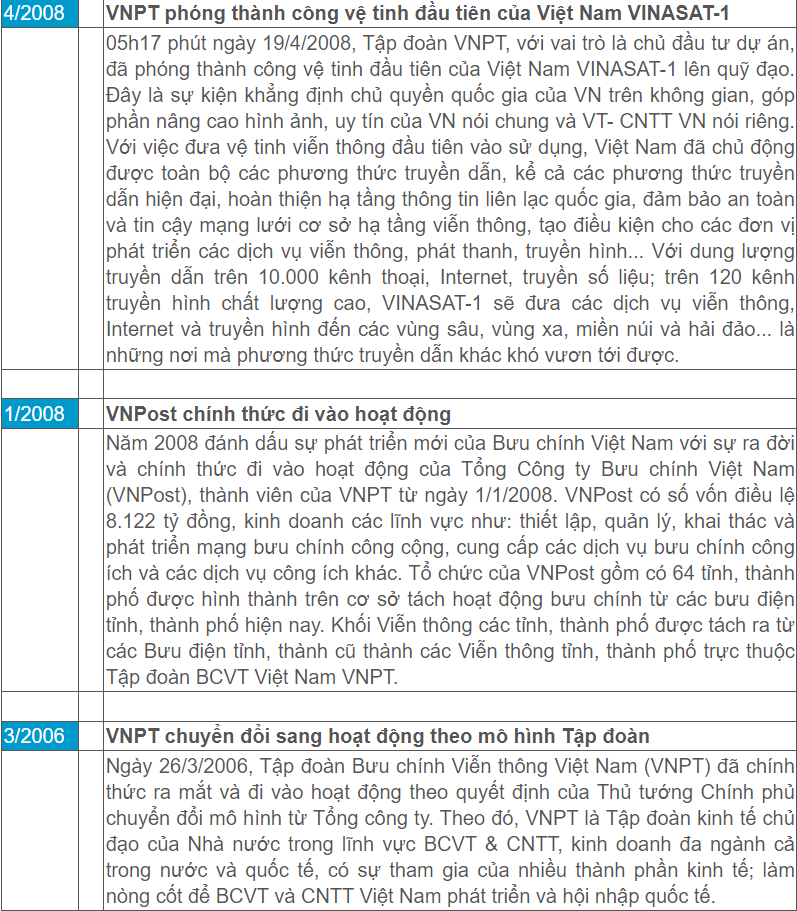
Hình 1: Các mốc phát triển của Tập đoàn VNPT từ 2016 đến 2018 (VNPT)



Hình 2: Các mốc phát triển của Tập đoàn VNPT từ 2012 đến 2015 (VNPT)



Hình 3: Các mốc phát triển của Tập đoàn VNPT từ 2009 đến 2011 (VNPT)



Hình 4: Các mốc phát triển của Tập đoàn VNPT từ 2006 đến 2008 (VNPT)



Hình 5: Các mốc phát triển của Tập đoàn VNPT từ 1995 đến 1997 (VNPT)

1. Lĩnh vực hoạt động

Tập đoàn VNPT có lĩnh vực hoạt động chủ yếu trong lĩnh vực viễn thông, công nghệ thông tin, phát thanh, truyền hình và truyền thông đa phương tiện. Theo thông tin từ website chính thức của VNPT, có 5 lĩnh vực hoạt động chính như sau:

* Kinh doanh dịch vụ và sản phẩm viễn thông, công nghệ thông tin;
* Kinh doanh dịch vụ và sản phẩm truyền thông;
* Kinh doanh dịch vụ quảng cáo, tổ chức sự kiện;
* Sản xuất, kinh doanh xuất nhập khẩu, cung ứng vật tư, thiết bị viễn thông, công nghệ thông tin và truyền thông;
* Khảo sát, tư vấn, thiết kế, lắp đặt, bảo dưỡng các công trình viễn thông, công nghệ thông tin và truyền thông. (VNPT, 2021)

1. Sơ đồ tổ chức

“Theo Nghị định số 25/2016/NĐ - CP do Chính phủ ban hành ngày 6 tháng 4 năm 2016 về phê duyệt điều lệ tổ chức và hoạt động của Tập đoàn Bưu chính Viễn thông Việt Nam (viết tắt là “VNPT”), VNPT là công ty mẹ trong Tập đoàn Bưu chính Viễn thông Việt Nam do Nhà nước sở hữu 100% vốn điều lệ. Công ty mẹ - Tập đoàn Bưu chính Viễn thông Việt Nam có các đơn vị trực thuộc, các công ty con như sau:

*Các đơn vị trực thuộc* bao gồm: 63 đơn vị viễn thông tỉnh, thành phố; tổng công ty hạ tầng mạng (VNPT – NET); chi nhánh Tập đoàn Bưu chính Viễn thông Việt Nam tại Hoa Kỳ; ban Quản lý Dự án Cáp quang biển, trung tâm Nghiên cứu và Phát triển; trung tâm Bồi dưỡng Nghiệp vụ III; công ty Công nghệ Thông tin VNPT (VNPT – IT); bệnh viện Bưu điện và bệnh viện Đa khoa Bưu điện.

*Các công ty con là các công ty TNHH một thành viên mà Tập đoàn sở hữu 100% vốn:* Tổng Công ty Dịch vụ viễn thông (VNPT – Vinaphone); tổng Công ty Truyền thông (VNPT – Media); công ty TNHH Một thành viên Cáp quang.” (VNPT, 2021)

Sơ đồ tổ chức bộ máy của tập đoàn như sau:



Hình 6: Sơ đồ tổ chức bộ máy của Tập đoàn VNPT

*Nguồn: Cổng thông tin điện tử Ủy ban quản lý vốn Nhà nước tại doanh nghiệp*

1. Mô tả bài toán
2. Thực trạng

Việc sử dụng điện thoại thông minh, internet, lưu lượng truy cập ngang hàng, trò chuyện trên mạng xã hội và IoT (Internet of Things) góp phần vào việc tăng khối lượng dữ liệu và mức tiêu thụ băng thông một cách nhanh chóng. Mô hình và nhu cầu sử dụng mạng internet thay đổi hàng ngày, theo ước tính mới nhất từ Statista[[3]](#footnote-3) có khoảng 328.77 million terabytes lượng dữ liệu được người tiêu dùng tạo ra mỗi ngày và hình thành sự bùng nổ về thông tin. Người tiêu dùng cũng là người cần các dịch vụ tương tác với độ trễ thấp. Các doanh nghiệp viễn thông đã thu thập được một lượng lớn dữ liệu của người tiêu dùng mỗi ngày dựa trên lượng thuê bao lớn của họ được kết nối với mạng mỗi ngày. Bằng cách mở rộng tiếng nói của khách hàng vào doanh nghiệp, họ đang ngày càng thu thập nhiều dữ liệu.

Theo báo cáo của trung tâm dữ liệu tại tập đoàn, hiện nay lượng dữ liệu của VNPT lên tới 3.5 PB, được xuất phát từ các nguồn như: Telco Data (VNPT-NET/Vinaphone?IT), Media, VNPT provinces, DPI, IT applications vnEdu/eGov/, … Lượng dữ liệu này được phục vụ việc xây dựng các dashboard, report và cả cho việc ứng dụng ML, DL, AI, v.v. Từ đó, giúp cho các giám đốc điều hành, các bộ phận theo dõi hoạt động kinh doanh, biến động của thị trường theo thời gian thực và hơn thế nữa là dự đoán được những xu hướng trong tương lai, cải thiện hiệu quả kinh doanh của doanh nghiệp. Tuy nhiên với lượng dữ liệu khổng lồ như vậy, doanh nghiệp đang chưa có một hệ thống duy nhất để đồng bộ, quản lý dữ liệu. Theo những thông tin thu thập được của dự án, VNPT đang gặp phải một số vấn đề như:

Dữ liệu phân tán và phức tạp: với việc kinh doanh rộng lớn trong nhiều lĩnh vực, VNPT đang sở hữu một lượng lớn dữ liệu phân tán và phức tạp từ nhiều nguồn khác nhau. Việc quản lý và truy cập dữ liệu này chưa được tập trung trên toàn tập đoàn, vì vậy mang đến sự phức tạp và khó khăn trong công tác quản lý và sử dụng.

Khó khăn trong việc tích hợp dữ liệu: Do dữ liệu phân tán và lưu trữ trong các hệ thống không liên kết, việc tích hợp dữ liệu từ các nguồn khác nhau trở nên khó khăn. Điều này dẫn đến việc tồn tại các bản sao dữ liệu, không đồng nhất và khó kiểm soát.

Khả năng phân tích và khai thác dữ liệu hạn chế: Do dữ liệu phân tán và không được tập trung, công ty thường gặp khó khăn trong việc khai thác toàn bộ tiềm năng, giá trị từ dữ liệu của mình. Các phân tích và báo cáo dựa trên dữ liệu phân tán tốn thời gian, trung bình để tạo một report, dashboard mất từ 5 đến 6 tiếng do phải tổng hợp dữ liệu từ các nguồn, biến đổi, xử lý và xuất ra thành report, dashboard. Tại VNPT Media, trong DWH hiện tại dữ liệu được lưu trữ chỉ với tối đa cho 30 ngày. Chính vì vậy, chất lượng dữ liệu hiện tại chưa đáp ứng được nhu cầu về tốc độ, độ chính xác và tính sẵn sàng.

Hiện tại, hệ thống Hadoop cũng đang rơi vào tình trạng như: thiếu storage, xử lý, quản trị, multi tenant (CPU, Memory, Storage), các ETL tool, việc mở rộng cũng gặp khó khăn.

1. Thách thức

Hiệu suất và sự chậm trễ của các dữ liệu thu thập được hiện nay đồng nghĩa với việc doanh nghiệp đang phải tốn nhiều thời gian để có được những thông tin chi tiết mới về các giao dịch, lượt truy cập, sử dụng dịch vụ của khách hàng. Những sự chậm trễ này ảnh hưởng xấu đến việc ra mắt sản phẩm, chiến dịch tiếp thị, quảng cáo và nhiều hoạt động kinh doanh khác. Ngoài ra, lượng dữ liệu được tạo ra mỗi ngày rất lớn dẫn đến việc xử lý và hiển thị dữ liệu bị chậm trễ. Tiếp đến, dữ liệu thu thập từ các nguồn khác nhau có thể không đồng nhất về định dạng, chất lượng và độ chính xác. Điều này có thể dẫn đến việc khó khăn trong việc sử dụng dữ liệu. Dữ liệu có thể chứa những thông tin nhạy cảm và quan trọng, như thông tin cá nhân khách hàng và giao dịch. Việc bảo vệ dữ liệu trước các mối đe dọa bảo mật như hack, tấn công mạng, hoặc lỗi phần mềm là một thách thức lớn.

1. Giá trị mang lại của bài toán

Nhu cầu nhức nhối nhất hiện nay của VNPT là cần tải nhanh một lượng dữ liệu lớn để làm các báo cáo, phân tích, hỗ trợ ra quyết định, xây dựng và triển khai các thuật toán ML, DL, AI dự đoán xu hướng, cải thiện các hoạt động kinh doanh. Xung quanh đó là những yêu cầu về lưu trữ, quản trị, bảo mật dữ liệu. Có thể liệt kê những giá trị mang lại đó như sau:

Tích hợp và quản lý dữ liệu: Data Lake giúp tập đoàn viễn thông có thể tích hợp và quản lý các nguồn dữ liệu khác nhau từ các hệ thống khác nhau một cách dễ dàng và hiệu quả. Các nguồn dữ liệu này có thể bao gồm các hệ thống thu thập dữ liệu, hệ thống CRM, hệ thống billing và các hệ thống khác.

Tăng tốc độ xử lý dữ liệu: Với Data Lake, có thể xử lý lượng dữ liệu lớn và phức tạp một cách nhanh chóng và hiệu quả hơn. Data Lake cho phép lưu trữ và truy xuất các tập dữ liệu lớn mà không cần phải trích xuất dữ liệu và xây dựng các kho dữ liệu truyền thống.

Tăng tính bảo mật dữ liệu: Cung cấp một lớp bảo mật dữ liệu để bảo vệ các tập dữ liệu quan trọng của tập đoàn viễn thông. Đồng thời, cho phép thiết lập các quyền truy cập dữ liệu và giám sát các hoạt động truy cập dữ liệu.

Đưa ra quyết định chính xác hơn: Cho phép tập đoàn sử dụng các tập dữ liệu với khối lượng lớn để đưa ra các phân tích và đánh giá, từ đó đưa ra quyết định chính xác hơn. VNPT có thể sử dụng các công cụ phân tích dữ liệu để tìm ra mẫu và xu hướng trong dữ liệu và từ đó đưa ra các quyết định kinh doanh quan trọng.

Tiết kiệm chi phí và thời gian: Triển khai Data Lake giúp tiết kiệm chi phí và thời gian trong việc lưu trữ và xử lý dữ liệu. Với Data Lake, không cần phải xây dựng các kho dữ liệu truyền thống và có thể sử dụng các công cụ và dịch vụ lưu trữ dữ liệu đám mây.

1. Hướng giải quyết

Để giải quyết bài toán trên cần xây dựng một hồ dữ liệu, nơi tập trung mọi nguồn dữ liệu của tập đoàn. Hiện nay, có rất nhiều các giải pháp hồ dữ liệu, có thể kể đến các giải pháp của các nhà cung cấp lớn như: Amazon Web Service, Microsoft Azure, Google Cloud, Oracle, IBM, v.v. Sau khi tham chiếu tới các giải pháp hồ dữ liệu, công ty quyết định lựa chọn bộ giải pháp Cloud Pak for Data (CP4D) của hãng IBM và cơ sở dữ liệu SingleStore. Hệ thống lưu trữ sử dụng IBM Elastic Storage System (ESS) cho hot data và IBM Cloud Object Storage (ICOS) cho cold data. Open source Kafka làm hệ thống message pub/sub truyền dữ liệu streaming. Cùng với đó là Hadoop cluster tối ưu hóa băng thông.

CP4D (Cloud Pak for Data) là một giải pháp của IBM, được thiết kế để cung cấp nền tảng dữ liệu đáng tin cậy, linh hoạt và an toàn cho các tổ chức. CP4D kết hợp các công nghệ và công cụ tiên tiến để giúp người dùng trích xuất thông tin từ dữ liệu phức tạp và xử lý các tác vụ phân tích dữ liệu.

Các tính năng chính của CP4D bao gồm:

*Quản lý dữ liệu*: CP4D cung cấp một giải pháp quản lý dữ liệu toàn diện, bao gồm các tính năng như lưu trữ dữ liệu, quản lý và tích hợp dữ liệu, và quản lý khách hàng.

*Tích hợp công cụ*: CP4D tích hợp nhiều công cụ phân tích dữ liệu như Watson Studio, Watson Machine Learning, Watson Assistant, Watson Discovery, v.v. Giúp người dùng trích xuất thông tin từ những bộ dữ liệu phức tạp qua đó phát triển và cải tiến các mô hình dự đoán.

*Tích hợp giải pháp*: CP4D cung cấp tính năng tích hợp giải pháp để giúp người dùng tích hợp các ứng dụng và dịch vụ khác vào giải pháp của mình.

*An toàn và bảo mật*: CP4D cung cấp các tính năng an toàn và bảo mật cao như xác thực, quản lý quyền truy cập, mã hóa dữ liệu, ...

*Quản lý khai thác*: CP4D cung cấp các tính năng quản lý khai thác để hỗ trợ người dùng có thể quản lý và theo dõi các công việc khai thác dữ liệu.

*Đa nền tảng*: CP4D có thể triển khai trên các nền tảng cloud hoặc trên on-premise (cơ sở hạ tầng của tổ chức).

Với những tính năng trên, CP4D cho phép các tổ chức có thể tập trung vào việc phân tích dữ liệu và đưa ra các quyết định dựa trên thông tin phân tích để cải thiện hiệu quả kinh doanh. Trong việc xây dựng Data Lake, CP4D được ứng dụng cho hoạt động ingestion, xây dựng model, làm storage, virtualization, visualization, phân tích dữ liệu, quản trị dữ liệu.

SingleStore (trước đây là MemSQL) là một “cơ sở dữ liệu phân tán” (distributed database) hàng đầu trên thị trường hiện nay, được thiết kế để xử lý các ứng dụng dữ liệu tốc độ cao và phức tạp. SingleStore có khả năng xử lý hàng triệu truy vấn mỗi giây và hàng tỷ hàng ngàn kết nối đồng thời, đồng thời cung cấp khả năng phân tích dữ liệu trực tuyến (real-time analytics) và tính toán dữ liệu phân tán (distributed computing).

Các tính năng của SingleStore bao gồm:

*Khả năng mở rộng linh hoạt*: SingleStore có thể mở rộng từ một cụm gồm một máy chủ đơn lẻ đến hàng trăm hay thậm chí là cả nghìn máy chủ trong một cụm.

Tính toán dữ liệu phân tán: SingleStore hỗ trợ tính toán dữ liệu phân tán, cho phép các ứng dụng truy xuất dữ liệu từ nhiều nguồn khác nhau và thực hiện tính toán trên đó.

*Tính toán dữ liệu thời gian thực*: SingleStore hỗ trợ tính toán dữ liệu thời gian thực, hỗ trợ xử lý dữ liệu đến từ các nguồn khác nhau và đưa ra các phản hồi và hành động trong thời gian thực.

*Khả năng phân tích dữ liệu trực tuyến*: SingleStore cung cấp khả năng phân tích dữ liệu trực tuyến, cho phép các ứng dụng thực hiện các thao tác truy xuất và tính toán trực tiếp trên dữ liệu trong khi dữ liệu vẫn đang được ghi vào cơ sở dữ liệu.

*Độ tin cậy cao*: SingleStore có khả năng tự động khôi phục dữ liệu trong trường hợp có lỗi, đồng thời hỗ trợ sao lưu và khôi phục dữ liệu để đảm bảo an toàn cho dữ liệu của khách hàng.

*Tính toán cân bằng tải tự động*: SingleStore có khả năng phân phối tải tự động trên các máy chủ trong cụm, giúp đảm bảo hiệu suất và khả năng phục vụ của cơ sở dữ liệu.

Nhờ những tính năng trên mà SingleStore hỗ trợ được cả OLTP và OLAP trên cùng 1 hệ CSDL và real time analytic. Vì vậy, rất phù hợp cho bài toán tại VNPT.

1. Kết luận chương 1

Chương 1 trình bày về đơn vị triển khai dự án Hồ dữ liệu là Tập đoàn Bưu chính Viễn thông Việt Nam VNPT như giới thiệu chung, tầm nhìn, sứ mệnh, giá trị cốt lõi, lịch sử hình thành và phát triển, lĩnh vực hoạt động cũng như sơ đồ tổ chức. Cùng với đó là thực trạng và thách thức hiện tại mà Tập đoàn đang gặp phải khi khai thác dữ liệu. Qua đó chỉ ra những giá trị mang lại của việc xây dựng hồ dữ liệu và nêu lên phương án giải quyết cho VNPT. Các lý thuyết tổng quan về Data Lake và một số công cụ được sử dụng trong dự án sẽ được trình bày ở chương 2.

# **CHƯƠNG 2: TỔNG QUAN LÝ THUYẾT DATA LAKE**

# **VÀ CÁC CÔNG CỤ SỬ DỤNG**

Khái niệm, kiến trúc và đặc điểm của Data Lake

Trong kỷ nguyên bùng nổ với dữ liệu lớn – Big Data, một thuật ngữ mới mang tên “Data Lake” (hồ dữ liệu) xuất hiện. Mục đích chính của hồ dữ liệu là làm nơi lưu trữ mọi dữ liệu mà doanh nghiệp tạo ra để lưu trữ các thông tin mang lại nhiều giá trị hơn và ở mức độ chi tiết hơn. Thuật ngữ Data Lake lần đầu tiên được đề cập đến là vào tháng 10/2010 bởi Giám đốc điều hành Pentaho – Jame Dixon (Dixons, 2010). Data Lake lưu trữ các dữ liệu của doanh nghiệp từ các nguồn dưới dạng có cấu trúc, bán cấu trúc, không cấu trúc và không có schema (lược đồ) được xác định trước. Chính vì vậy, Data Lake chứa đựng những giá trị tiềm ẩn mà người dùng có thể khai thác. Đó là những dữ liệu ở dạng nguyên bản, chứa các dữ liệu chi tiết mà những giá trị mang lại vẫn là ẩn số.

Để hiểu rõ hơn về kiến trúc Data Lake, tại hội nghị quốc tế lần thứ năm của IEEE về Dữ liệu lớn và Điện toán đám mây 2015 Hassan Alrehamy và Coral Walker đã đưa ra giải thích: “Hồ dữ liệu sử dụng kiến trúc phẳng để lưu trữ dữ liệu ở định dạng thô. Mỗi thực thể dữ liệu trong hồ được liên kết với một mã định danh duy nhất và một tập hợp siêu dữ liệu mở rộng và người tiêu dùng có thể sử dụng các lược đồ được xây dựng có mục đích để truy vấn dữ liệu có liên quan, điều này sẽ tạo ra một tập hợp dữ liệu nhỏ hơn có thể được phân tích để giúp trả lời câu hỏi của người dùng”. Về bản chất, hồ dữ liệu sẽ lưu tất cả dữ liệu trong một doanh nghiệp bất kể loại, định dạng, cấu trúc. Những thông tin dữ liệu mang lại sẽ tùy thuộc vào người sử dụng dữ liệu tại thời điểm họ truy xuất dữ liệu. Khi dữ liệu được truy xuất, người dùng sẽ biến đổi dữ liệu đó tùy theo nhu cầu của mình để có những giá trị cho doanh nghiệp.

Từ khái niệm và kiến trúc của Data Lake có thể thấy được một số đặc điểm của Data Lake như sau:

* Tính chính xác

Hồ dữ liệu lưu trữ dữ liệu của hệ thống kinh doanh. Không giống kho dữ liệu chứa các dữ liệu đã qua xử lý, hồ dữ liệu lưu trữ dữ liệu thô, không thể sửa đổi định dạng, lược đồ và nội dung. Hồ dữ liệu lưu trữ nguyên trạng dữ liệu kinh doanh của doanh nghiệp. Dữ liệu được lưu trữ có thể bao gồm dữ liệu ở bất kỳ định dạng nào và thuộc bất kỳ loại nào.

* Tính linh hoạt

Hồ dữ liệu lưu trữ dữ liệu không cần lược đồ, điều này có nghĩa là nó có thể thích ứng với những thay đổi kinh doanh không thể đoán trước. Bạn có thể thiết kế lược đồ dữ liệu trong bất kỳ giai đoạn nào nếu cần, vì vậy toàn bộ cơ sở hạ tầng tạo ra dữ liệu đáp ứng nhu cầu kinh doanh của bạn. Tính chính xác và linh hoạt có quan hệ chặt chẽ với nhau. Vì những thay đổi trong kinh doanh là không thể đoán trước nên ta luôn có thể giữ nguyên trạng dữ liệu và xử lý dữ liệu khi cần. Do đó, một hồ dữ liệu phù hợp hơn cho các doanh nghiệp đổi mới và doanh nghiệp có sự thay đổi, tăng trưởng kinh doanh nhanh chóng.

* Khả năng quản lý

Hồ dữ liệu cung cấp khả năng quản lý dữ liệu toàn diện. Do tính chính xác và tính linh hoạt, hồ dữ liệu lưu trữ ít nhất hai loại dữ liệu: dữ liệu thô và dữ liệu đã xử lý. Dữ liệu được lưu trữ liên tục tích lũy và tăng lên. Điều này phát sinh ra yêu cầu về khả năng quản lý dữ liệu mạnh mẽ, bao gồm nguồn dữ liệu, kết nối dữ liệu, định dạng dữ liệu và lược đồ dữ liệu. Lược đồ dữ liệu bao gồm CSDL và các bảng, cột và hàng có liên quan. Hồ dữ liệu cung cấp lưu trữ tập trung cho dữ liệu của một doanh nghiệp hoặc tổ chức, điều này đòi hỏi khả năng quản lý quyền.

* Truy xuất nguồn gốc dữ liệu

Hồ dữ liệu lưu trữ mọi dữ liệu của một tổ chức hoặc doanh nghiệp và quản lý chúng trong suốt vòng đời của nó, từ định nghĩa dữ liệu, truy cập và lưu trữ đến xử lý, phân tích và ứng dụng. Hồ dữ liệu đảm bảo rằng mỗi bản ghi dữ liệu đều có thể theo dõi được thông qua các quy trình truy cập, lưu trữ, xử lý và sử dụng. (Alibaba Clouds, 2020)

Sự khác biệt giữa Data Lake và Data Warehouse

Hồ dữ liệu và kho dữ liệu đều là hệ thống lưu trữ dữ liệu lớn được sử dụng bởi các nhà khoa học dữ liệu, kỹ sư dữ liệu và nhà phân tích kinh doanh. Nhưng trong khi kho dữ liệu được thiết kế để truy vấn và phân tích, thì hồ dữ liệu có nhiều nguồn với dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc cùng được tích hợp.

Hai hệ thống lưu trữ phục vụ các mục đích khác nhau, do đó mỗi hệ thống sẽ có các vai trò công việc khác nhau. Đối với một số công ty, hồ dữ liệu hoạt động tốt nhất, đặc biệt là những công ty được hưởng lợi từ dữ liệu thô cho ứng dụng học máy. Đối với những người dùng khác, kho dữ liệu lại phù hợp hơn, bởi các nhà phân tích kinh doanh của cần giải mã các phân tích trong một hệ thống có cấu trúc.

Dưới đây là bảng so sánh điểm khác biệt giữa hồ dữ liệu và kho dữ liệu được tổng hợp:

Bảng 1: Điểm khác biệt giữa Hồ dữ liệu và Kho dữ liệu

*Nguồn: Athira Nambiar, Divyansh Mundra (2022)*

|  |  |  |
| --- | --- | --- |
| Điểm so sánh | Hồ dữ liệu | Kho dữ liệu |
| Loại dữ liệu | Dữ liệu thô (ở tất cả các định dạng và bất kể cấu trúc) | Đã xử lý (dữ liệu được lưu trữ theo số liệu và thuộc tính) |
| Mục đích sử dụng | Có thể chưa được xác định | Đã được xác định trước |
| Quy trình | Trích xuất, tải và biến đổi (ELT) | Trích xuất, biến đổi và tải (ETL) |
| Thời điểm xác định lược đồ | Sau khi lưu trữ dữ liệu, để cung cấp sự linh hoạt và thu thập dữ liệu dễ dàng | Trước khi lưu trữ dữ liệu để bảo mật và có hiệu suất cao |
| Người dùng | Nhà khoa học dữ liệu những người cần phân tích chuyên sâu và công cụ để hiểu dữ liệu | Những chuyên gia trong kinh doanh, người cần dữ liệu cho việc hoạt động của doanh nghiệp |
| Khả năng tiếp cận | Tiếp cận dễ dàng và dễ thay đổi | Khó thực hiện các việc thay đổi |
| Lịch sử | Tương đối mới với dữ liệu lớn | Đã xuất hiện trong nhiều thập kỷ |

Một, hồ dữ liệu lưu trữ dữ liệu thô ở mọi định dạng và cấu trúc còn kho dữ liệu lưu trữ dữ liệu đã được xử lý. Dữ liệu thô là dữ liệu chưa được xử lý cho một mục đích nào đó. Có đây lẽ sự khác biệt lớn nhất giữa hồ dữ liệu và kho dữ liệu. Hồ dữ liệu chủ yếu lưu trữ dữ liệu thô, chưa xử lý, trong khi kho dữ liệu lưu trữ dữ liệu đã xử lý và tinh chỉnh. Do đó, các hồ dữ liệu thường yêu cầu dung lượng lưu trữ lớn hơn nhiều so với kho dữ liệu. Ngoài ra, dữ liệu thô, chưa qua xử lý có thể điều chỉnh được, có thể được phân tích nhanh chóng cho bất kỳ mục đích nào và rất lý tưởng cho việc ứng dụng học máy. Tuy nhiên, rủi ro của tất cả dữ liệu thô đó là các hồ dữ liệu đôi khi khiến hồ dữ liệu trở thành đầm lầy dữ liệu nếu không có các biện pháp quản trị dữ liệu và quản lý dữ liệu phù hợp. Kho dữ liệu, bằng cách chỉ lưu trữ dữ liệu đã xử lý, tiết kiệm không gian lưu trữ đắt tiền bằng cách không duy trì dữ liệu có thể không bao giờ được sử dụng. Ngoài ra, dữ liệu được xử lý giúp nhiều đối tượng có thể hiểu nội dung hơn.

Hai, mục đích sử dụng của hồ dữ liệu có thể chưa được xác định còn kho dữ liệu đã được xác định trước. Dữ liệu thô chảy vào hồ dữ liệu, đôi khi đã được xác định để sử dụng cho các công việc cụ thể trong tương lai và đôi khi chỉ để sẵn sàng cho bất kỳ ý định nào xuất hiện. Dữ liệu đã xử lý là dữ liệu thô đã được đưa vào sử dụng cụ thể. Vì kho dữ liệu chỉ chứa dữ liệu đã được xử lý, nên tất cả dữ liệu trong kho dữ liệu được xác định để sử dụng cho một mục đích cụ thể trong tổ chức. Điều này có nghĩa là không gian lưu trữ không bị lãng phí đối với dữ liệu có thể không bao giờ được sử dụng.

Ba, với hồ dữ liệu dữ liệu được trích xuất và tải vào hồ sau đó tùy theo mục đích sử dụng sẽ được biến đổi, còn kho dữ liệu được trích xuất, biến đổi sau đó mới tải vào kho. Cũng bởi mục đích sử dụng của hồ dữ liệu có thể chưa được xác định trước nên sẽ cần lưu ở định dạng ban đầu sau đó mới biến đổi tùy theo yêu cầu sử dụng. Với kho dữ liệu thì việc dùng dữ liệu cho những công việc, mục đích nào đã được xác định từ trước, do vậy cần biến đổi trước khi cho vào kho để giảm thiểu sự lãng phí khi phải cung cấp nơi lưu trữ cho những dữ liệu không cần thiết.

Bốn, hồ dữ liệu xác định lược đồ sau khi dữ liệu được lưu trữ còn kho dữ liệu được xác định lược đồ trước khi dữ liệu được lưu trữ. Tương tự như điểm khác biệt thứ ba, hồ dữ liệu sẽ phục vụ cho những mục đích chưa được xác định trước, vì vậy cần đảm bảo dữ liệu được thu thập, truy cập dễ dàng và nhanh chóng. Với kho dữ liệu do đã xác định trước việc sử dụng nên lược đồ cũng được xác định trước khi lưu trữ để không chỉ đảm bảo dữ liệu là phù hợp với các mục đích mà còn có hiệu suất cao và đảm bảo tính bảo mật.

Năm, hồ dữ liệu thường được dùng bởi các nhà khoa học dữ liệu còn kho dữ liệu được dùng bởi các nhà quản lý doanh nghiệp. Hồ dữ liệu thường gây khó khăn khi dùng với những người không quen với dữ liệu chưa được xử lý. Dữ liệu thô, phi cấu trúc thường yêu cầu một nhà khoa học dữ liệu và các công cụ chuyên dụng để hiểu và sử dụng dữ liệu đó cho bất kỳ mục đích sử dụng kinh doanh cụ thể nào. Dữ liệu đã xử lý được sử dụng trong biểu đồ, bảng tính, bảng, v.v. để hầu hết, nếu không muốn nói là tất cả, nhân viên trong công ty có thể đọc được dữ liệu đó. Dữ liệu đã xử lý được lưu trữ trong kho dữ liệu, chỉ yêu cầu người dùng phải quen thuộc với chủ đề được trình bày. Vì vậy nó rất phù hợp cho những nhà quản lý doanh nghiệp, người cần sử dụng dữ liệu để vận hành doanh nghiệp hiệu quả.

Sáu, hồ dữ liệu dễ tiếp cận và thay đổi còn với kho dữ liệu việc thay đổi khó khăn hơn. Khả năng truy cập và dễ sử dụng đề cập đến việc sử dụng toàn bộ vùng lưu trữ dữ liệu, chứ không phải dữ liệu bên trong chúng. Kiến trúc hồ dữ liệu không có cấu trúc và do đó dễ truy cập và dễ thay đổi. Ngoài ra, bất kỳ thay đổi nào được thực hiện đối với dữ liệu đều có thể được thực hiện nhanh chóng vì các hồ dữ liệu có rất ít hạn chế. Kho dữ liệu, theo thiết kế, có cấu trúc hơn. Một lợi ích chính của kiến ​​trúc kho dữ liệu là việc xử lý và cấu trúc dữ liệu giúp dữ liệu dễ giải mã hơn, những hạn chế về cấu trúc khiến kho dữ liệu khó thao tác và tốn kém.

Bảy, hồ dữ liệu mới được xuất hiện gần đây, gắn với công nghệ dữ liệu lớn còn DWH đã xuất hiện từ những năm 1980. Khi công nghệ dữ liệu lớn xuất hiện, với nhiều đặc điểm và đòi hỏi theo đó là những yêu cầu về việc lưu trữ dữ liệu thì thuật ngữ hồ dữ liệu mới bắt đầu xuất hiện. Lần đầu tiên vào năm 2011 bởi bởi Giám đốc điều hành Pentaho – Jame Dixon. Với kho dữ liệu, các khái niệm đầu tiên đã có vào khoảng năm 1960 sau đó được sử dụng vào cho các hệ hỗ trợ ra quyết định lần đầu vào năm 1983 bởi Teradata. (Talend)

Lợi ích của Data Lake

*Thứ nhất*, khi triển khai Data Lake giúp lưu trữ tập trung dữ liệu của tổ chức không giới hạn. Điều này vừa giúp cho doanh nghiệp có thể truy cập vào các loại dữ liệu ở các nguồn khác nhau theo thời gian, mà vẫn đảm bảo tất cả các đơn vị sử dụng đều có được những cái nhìn thống nhất về dữ liệu, tránh tình trạng thông tin silo. Cùng với đó là việc ứng dụng sức mạnh của trí tuệ nhân tạo để khai phá, phân tích dữ liệu dễ dàng hơn nhờ có được lượng dữ liệu thống nhất từ tất cả các nguồn.

*Thứ hai*, cho phép phân tích, khai thác dữ liệu một cách linh hoạt và nhanh chóng. Bởi hồ dữ liệu lưu trữ dữ liệu không có lược đồ (schema), được nhập theo thời gian thực. Việc này đồng nghĩa với việc nó có thể thích ứng với những thay đổi kinh doanh không thể đoán trước. Người dùng có thể thiết kế lược đồ dữ liệu trong bất kỳ giai đoạn nào nếu cần, vì vậy toàn bộ cơ sở hạ tầng tạo ra dữ liệu đáp ứng nhu cầu kinh doanh. Từ đó, giúp thu thập được những thông tin mới nhất, góp phần hỗ trợ cho việc ra quyết định và mang lại giá trị cho doanh nghiệp. Những giá trị đó có thể là những cái nhìn 360 độ về khách hàng, thông tin mới nhất về xu hướng tài chính, phát hiện gian lận tự động, v.v

*Thứ ba*, Data Lake chứa những dữ liệu có giá trị tiềm ẩn. Giá trị và những hiểu biết nhận được từ Data Lake có thể là những ẩn số và thay đổi dựa trên những phương thức khai thác khác nhau. Data Lake thậm chí có thể sẽ lưu trữ những dữ liệu mà hiện tại chưa được sử dụng nhưng lại có thể mang đến những giá trị tích cực trong tương lai mà hiện tại chúng ta chưa nhận ra hoặc xác định.

*Thứ tư*, lưu trữ dữ liệu ở bất kỳ định dạng nào từ có cấu trúc, bán cấu trúc, không cấu trúc. Khi nhập dữ liệu, Data Lake không yêu cầu phải có mô hình dữ liệu. Người dùng có thể lưu trữ dữ liệu với kỳ định dạng nào và phương tiện nào: RDBMS, Cơ sở dữ liệu NoSQL, File Systems, Time Series Databases, v.v. Dữ liệu có thể được tải vào ở định dạng ban đầu như: log, CSV, XML, parquet, v.v. mà không cần bất kỳ sự biến đổi nào. Vì vậy, giúp Data Lake rẻ hơn so với Data Warehouse, thuận tiện cho các nhà khoa học dữ liệu sử dụng.

*Thứ năm*, là nền tảng vững chắc cho việc ứng dụng học máy và trí tuệ nhân tạo. Data Lake lưu trữ tập trung dữ liệu thô từ tất cả các nguồn với độ trễ thấp, dễ dàng truy vấn và phân tích bằng các công cụ khác nhau. Vì vậy, dữ liệu có thể được kết hợp để huấn luyện và triển khai các mô hình học máy, trí tuệ nhân tạo nhằm phân tích, dự đoán, đào tạo phục vụ cho các hoạt động kinh doanh của doanh nghiệp.

*Thứ sáu*, quản trị dữ liệu thuận tiện. Data Lake cho phép quản trị dữ liệu tập trung, giúp quản lý chất lượng dữ liệu, kiểm soát truy cập và tuân thủ dễ dàng hơn.

Một số thách thức khi triển khai Data Lake

Mặc dù hồ dữ liệu mang đến nhiều lợi ích cho doanh nghiệp, tuy nhiên nó cũng mang lại nhiều thách thức khi triển khai. Các thách thức thường xảy ra bao gồm các thách thức về mặt công nghệ, quy trình và cả nhân lực.

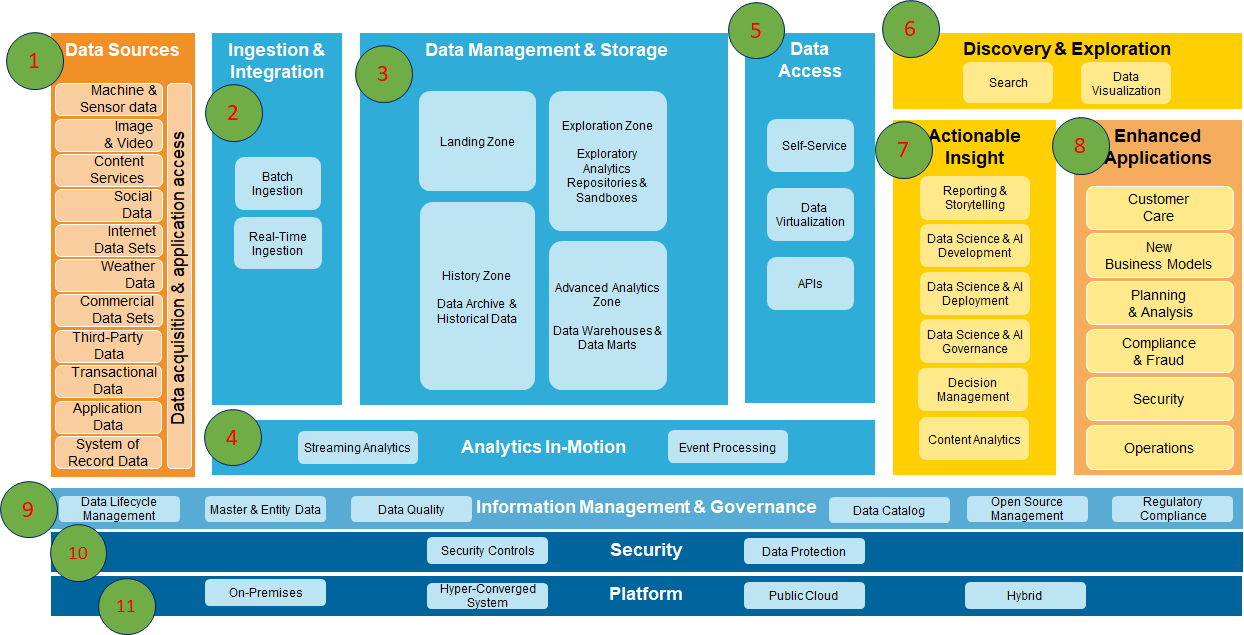
Trước tiên về mặt *công nghệ*, hồ dữ liệu có thể lưu trữ một lượng lớn dữ liệu. Do đó, cần xây dựng các chính sách quản lý dữ liệu phù hợp. Nếu không, hồ dữ liệu rất có khả năng trở thành đầm lầy dữ liệu và không sử dụng được. Doanh nghiệp cần kiểm soát quá trình nhập dữ liệu, theo dõi để cập nhật, xóa những dữ liệu không còn cần thiết. Bằng cách này, tránh lãng phí tài nguyên lưu trữ dữ liệu, tăng hiệu suất sử dụng. Trong một số trường hợp, dữ liệu nhạy cảm, cần bảo mật cũng có trong hồ dữ liệu. Thách thức lớn nhất lúc này là lưu trữ mã vẫn phải đảm bảo thực hiện nghiêm ngặt các biện pháp bảo mật và giám sát.

Về *quy trình*, sự phối hợp về quản trị dữ liệu, bảo mật và sử dụng dữ liệu là một thách thức to lớn. Các chính sách và tiêu chuẩn bảo mật dữ liệu không được áp dụng đúng cách như quy trình quản trị có thể gây ra sự cố về quyền truy cập vào dữ liệu được bảo vệ bởi các quy định về quyền riêng tư và các loại dữ liệu nhạy cảm khác. Mặc dù hồ dữ liệu là một nguồn dữ liệu mở với mọi người, nhưng cần có các biện pháp kiểm soát truy cập và bảo mật để đảm bảo an toàn thông tin. Các bộ phận khác nhau có thể có các quy tắc kinh doanh khác nhau cho dữ liệu tương tự, điều này có thể dẫn đến việc xung đột khi sử dụng dữ liệu và làm các phân tích giảm độ chính xác. Vì vậy, xác định rõ chương trình quản trị dữ liệu, bảo mật và sử dụng dữ liệu sẽ có thể giảm các vấn đề phát sinh khi nhiều đơn vị sử dụng một hồ dữ liệu. Nếu một tổ chức có nhiều hồ dữ liệu, thì mỗi hồ dữ liệu phải được đưa vào quy trình quản trị và được chỉ định người quản lý dữ liệu.

Cuối cùng là *nhân sự*, hồ dữ liệu cho phép mọi người trong doanh nghiệp truy cập. Tuy nhiên, việc sử dụng nó trong thực tế lại đòi hỏi một số yêu cầu gây khó khăn cho người dùng không có kỹ thuật bởi nó lưu trữ cả những dữ liệu không cấu trúc. Với cả những người dùng có kiến thức về công nghệ, việc sử dụng cũng sẽ phải tuân tủ các quy tắc nhất định do sự liên kết giữa các thành phần trong hồ.

Kiến trúc tham chiếu cho AI Data Lake

Theo IBM, kiến trúc tham chiếu cho AI Data Lake như sau:



Hình 7: Kiến trúc tham chiếu cho AI Data Lake

*Nguồn: IBM*

Thành phần đầu tiên là thành phần *Data Sources*. Đây là nguồn dữ liệu đầu vào cho hồ dữ liệu. Đó có thể bao gồm dữ liệu có cấu trúc, dữ liệu phi cấu trúc, dữ liệu multi-media và các dữ liệu thu thập từ các mạng xã hội hay từ các mạng lưới cảm biến, các thiết bị số ngoại biên[[4]](#footnote-4) (edge devices).

Khối *Ingestion & Integration* bao gồm các công cụ thu nhận và tích hợp dữ liệu; các công cụ này cho phép thu nhận dữ liệu theo 2 hình thức là xử lý theo lô hoặc xử lý trực tiếp theo thời gian thực. Thu thập dữ liệu: Đây là quá trình thu thập các nguồn dữ liệu từ các hệ thống khác nhau, bao gồm các cơ sở dữ liệu, tệp tin, dữ liệu từ các thiết bị IoT và các nguồn dữ liệu trực tuyến khác. Tích hợp dữ liệu: Sau khi dữ liệu đã được thu thập, quá trình tích hợp dữ liệu sẽ được thực hiện để đưa dữ liệu vào hồ dữ liệu. Các công cụ và kỹ thuật tích hợp dữ liệu, có thể bao gồm Apache Kafka, Apache Flume, Apache NiFi, AWS Glue và các nhiều công cụ, kỹ thuật khác, được ứng dụng để đưa dữ liệu vào hồ dữ liệu. Trong khối Ingestion & Integration, việc đảm bảo tính toàn vẹn và đáng tin cậy của dữ liệu là rất quan trọng. Vì vậy, các tiêu chuẩn và quy trình kiểm tra dữ liệu, kiểm soát chất lượng dữ liệu và quản lý quá trình đưa dữ liệu vào hồ dữ liệu được sử dụng để đảm bảo rằng dữ liệu được xử lý và tích hợp đúng cách. Việc làm này sẽ giúp đảm bảo tính đáng tin cậy và chất lượng của các phân tích và báo cáo dữ liệu sau này.

Khối *Data Management & Storage*, gồm các công cụ quản lý và lưu trữ dữ liệu, đây là phần cơ bản nhất của các kho dữ liệu hay hồ dữ liệu. Nó bao gồm các vùng chức năng:

* Vùng tiếp nhận dữ liệu là nơi dữ liệu được lưu trữ tạm thời để chờ được phân tích, kiểm tra cũng như xử lý trước khi được lưu trữ.
* Vùng các kho lưu trữ & sandbox dữ liệu phục khai và thăm dò dữ liệu với các công cụ phân tích, các mô hình dữ liệu báo cáo.
* Vùng phân tích nâng cao với các công cụ, kỹ thuật tiên tiến được sử dụng để khai phá dữ liệu, học máy và phân tích dữ liệu lớn.
* Vùng lịch sử dữ liệu sẽ lưu giữ và tổ chức dữ liệu để lưu trữ lâu dài.

Khối *Analytics In-Motion*, các công cụ và kỹ thuật hỗ trợ phân tích dữ liệu trực tiếp ngay trong quá trình truyền nhận dữ liệu từ các nguồn dữ liệu đa dạng như thiết bị IoT, máy chủ, cơ sở dữ liệu, hệ thống tài chính, v.v. Nhờ khả năng xử lý dữ liệu trực tiếp, phát hiện các sự kiện quan trọng ngay khi chúng xảy ra mà người dùng có thể phán đoán, xác định và từ đó triển khai các hành động cũng như quyết định dựa trên dữ liệu mới nhất và chính xác nhất. Ngoài ra, khối Analytics In-Motion còn cung cấp khả năng phân tích dữ liệu lớn (big data analytics) để xử lý các tập dữ liệu lớn và phức tạp.

Khối *Data Access*, cung cấp các khả năng truy cập vào dữ liệu thông qua cổng thông tin dữ liệu (Self-service portal), các công cụ ảo hoá dữ liệu (Data Virtualization) hay các API dữ liệu. Khối này cũng cung cấp các kỹ thuật để quản lý dữ liệu, bao gồm quản lý quyền truy cập và phân quyền, đảm bảo rằng chỉ những người được ủy quyền mới có thể truy cập vào các dữ liệu nhạy cảm. Không chỉ vậy, còn cung cấp các công cụ để xử lý và biến đổi dữ liệu bao gồm việc tối ưu hóa hiệu suất và đảm bảo tính toàn vẹn của dữ liệu.

Khối *Discovery & Exploration*, khai thác và thăm dò dữ liệu thông qua các công cụ tìm kiếm, trực quan hoá dữ liệu. Cụ thể, khối này cung cấp các công cụ để tìm kiếm và phân tích các tập dữ liệu khác nhau từ các nguồn dữ liệu khác nhau. Nó cung cấp các tính năng khám phá dữ liệu để tìm kiếm thông tin, tìm kiếm các xu hướng, tương quan và mối liên hệ giữa các dữ liệu khác nhau. Cung cấp các công cụ để tạo ra các bản đồ dữ liệu, các biểu đồ và báo cáo để giúp người dùng hiểu rõ hơn về các thông tin được khám phá và khai thác từ các nguồn dữ liệu khác nhau.

Khối *Actionable Insight*, khai thác, phân tích dữ liệu chuyên sâu, đây là phần quan trọng khi đem đến các năng lực và ứng dụng AI vào phân tích dữ liệu, cùng với kỹ thuật của khoa học dữ liệu để hiểu dữ liệu một cách sâu sắc, đa chiều. Từ đó đưa ra các kết quả của việc phân tích, đề xuất và dự đoán về các xu hướng và tương lai của các dữ liệu đó. Nó cũng cung cấp các công cụ và giải pháp để phân tích mối liên hệ giữa các dữ liệu, để giúp nhà quản lý hiểu rõ hơn về các yếu tố ảnh hưởng đến doanh nghiệp của họ và đưa ra các lời khuyên để giúp doanh nghiệp tăng cường hoạt động, tối ưu hóa các quy trình.

Khối *Enhanced Applications*, các lớp ứng dụng được hưởng lợi, được tăng cường về năng lực từ việc tận dụng được dữ liệu một cách hiệu quả. Nói cách khác, khối này chứa các ứng dụng được sử dụng những lợi ích của dữ liệu từ hồ dữ liệu. Các ứng dụng này có thể được phát triển trên nhiều nền tảng khác nhau và được tích hợp với các công nghệ khác nhau, từ đó tạo ra các giải pháp tùy chỉnh để đáp ứng nhu cầu cụ thể của doanh nghiệp.

Khối *Information Management & Governance*, khối chức năng quan trọng giúp định nghĩa rõ về mô hình vận hành của toàn bộ hệ thống, tác nhân tham gia, vai trò, trách nhiệm và quyền hạn của các tác nhân trong việc khai thác, sử dụng dữ liệu và tham gia vào quản lý vòng đời dữ liệu. Các thành phần của khối này bao gồm:

* Data Governance: Cung cấp các chính sách và quy trình để đảm bảo rằng dữ liệu được quản lý một cách hiệu quả và đúng cách.
* Data Quality: Cung cấp các công cụ để kiểm tra và đánh giá chất lượng dữ liệu và đảm bảo tính toàn vẹn của dữ liệu.
* Master Data Management: Cung cấp các công cụ để quản lý các bản ghi dữ liệu chính của doanh nghiệp, đảm bảo tính nhất quán và chính xác của thông tin.
* Metadata Management: Cung cấp các công cụ để quản lý thông tin về dữ liệu, bao gồm các mô tả, thuộc tính và quan hệ giữa các bộ dữ liệu.
* Data Lineage: Cung cấp các công cụ để theo dõi quá trình di chuyển dữ liệu từ nguồn tới đích, đảm bảo tính toàn vẹn và đáng tin cậy của thông tin.

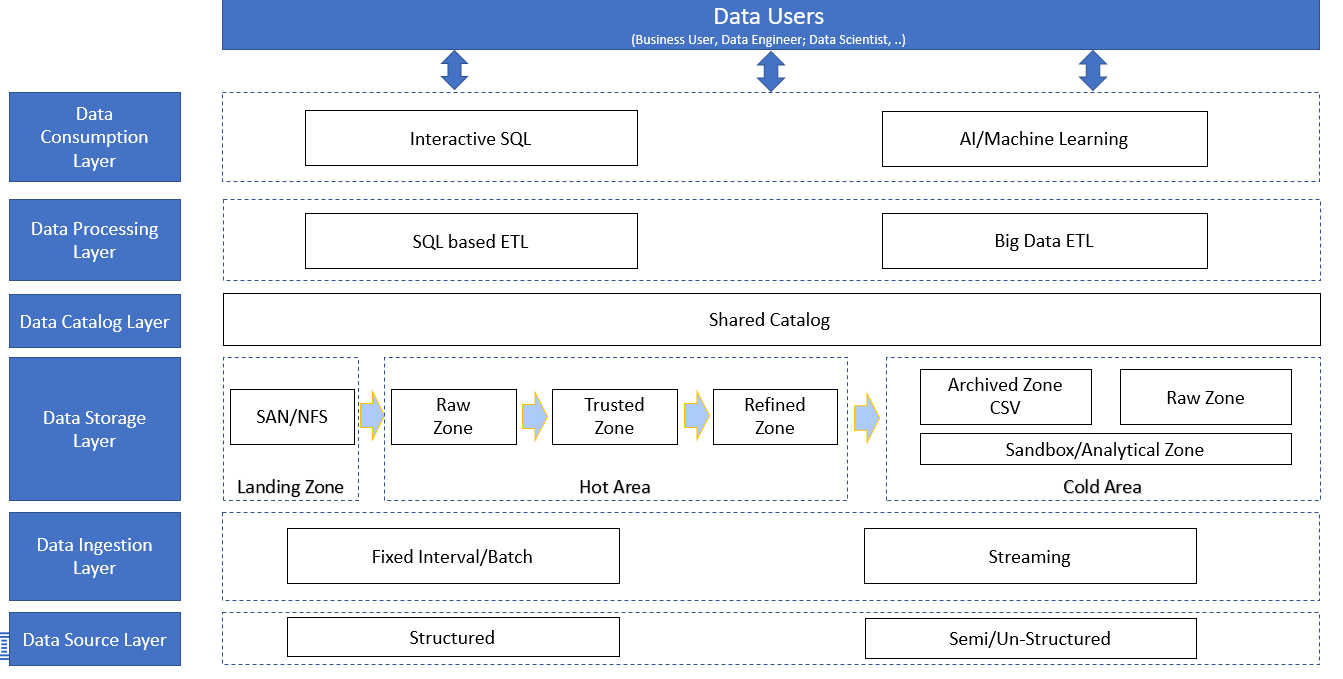
Khối *Security*, Khối chức năng về an toàn an ninh thông tin khi làm việc với các hệ thống dữ liệu: các giải pháp bảo mật phục vụ cho vận hành, điều khiển hệ thống; và các giải pháp liên quan để bảo vệ dữ liệu. Các công cụ bảo mật của khối này bao gồm:

* Authentication: Cung cấp các phương thức xác thực người dùng để đảm bảo rằng chỉ người dùng được ủy quyền mới có thể truy cập vào dữ liệu.
* Authorization: Cung cấp các quyền truy cập và phân quyền để đảm bảo rằng người dùng chỉ có thể truy cập vào các dữ liệu mà họ được ủy quyền.
* Encryption: Cung cấp các công cụ để mã hóa dữ liệu, bảo vệ dữ liệu khỏi các mối đe dọa bảo mật và truy cập trái phép.
* Monitoring and Logging: Cung cấp các công cụ phục vụ việc theo dõi các hoạt động hệ thống, xác định các hoạt động bất thường và phát hiện các mối đe dọa bảo mật.
* Threat Detection and Prevention: Cung cấp các công cụ để phát hiện và ngăn chặn các mối đe dọa bảo mật, bao gồm phát hiện xâm nhập, tấn công từ chối dịch vụ (DDoS), mã độc và phần mềm độc hại.

Khối *Platform*, xác định các nền tảng công nghệ phù hợp để triển khai các hệ thống về dữ liệu, có thể triển khai theo phương án truyền thống tại các trung tâm dữ liệu (on-premise); Các hệ thống siêu hội tụ (Hyper-Converged System); các hệ thống trên đám mây (đám mây công cộng, đám mây lai …).

Tóm lại, quy trình dữ liệu di chuyển trong hồ dữ liệu từ khi được thu thập đến khi đưa vào sử dụng như sau: Dữ liệu đến từ nhiều nơi, không chỉ bao gồm cơ sở dữ liệu, mà có thể bao gồm bên thứ ba và các dữ liệu từ bên ngoài. Dữ liệu thô được nhập vào theo đợt hoặc theo thời gian thực và được sắp xếp theo cấu trúc thư mục hợp lý. Sau đó dữ liệu tiếp tục được lưu trữ theo dạng Object Storage (Object Storage – Lưu trữ đối tượng là “kiến ​​trúc lưu trữ dữ liệu để lưu trữ dữ liệu phi cấu trúc, phân chia dữ liệu thành các đơn vị –đối tượng và lưu trữ chúng trong môi trường dữ liệu phẳng có cấu trúc. Mỗi đối tượng bao gồm dữ liệu, siêu dữ liệu và một mã định danh duy nhất mà các ứng dụng có thể sử dụng để truy cập và truy xuất dễ dàng” (Google Cloud)). Dữ liệu được lưu trữ có thể được sử dụng luôn cho việc phân tích hoặc tiếp tục đưa vào biến đổi, chuẩn bị cho những dự đoán, phân tích chuyên sâu. Dữ liệu muốn được sử dụng dễ dàng, bảo mật và hiệu quả cần đảm bảo có các vùng bảo vệ (quản lý xác thực danh tính và truy cập, kiểm toán, dịch vụ quản lý bảo mật) và quản trị dữ liệu.

Kiến trúc phân cấp của IBM cho AI Data Lake



Hình 8: Kiến trúc phân cấp của IBM cho AI Data Lake

*Nguồn: IBM*

1. Data sources: hệ thống hồ dữ liệu thu thập và xử lý dữ liệu từ nhiều nguồn khác nhau

- Nguồn dữ liệu có cấu trúc dạng batch (theo lô) vào khung thời gian cố định ví dụ như: các ứng dụng LOB, hệ thống tính cước, CDR, CRM sẽ xuất ra, v.v.

- Nguồn dữ liệu bán cấu trúc, không cấu trúc và dưới dạng stream liên tục như: ứng dụng web, sensor, mạng xã hội, thiết bị di động, video stream.

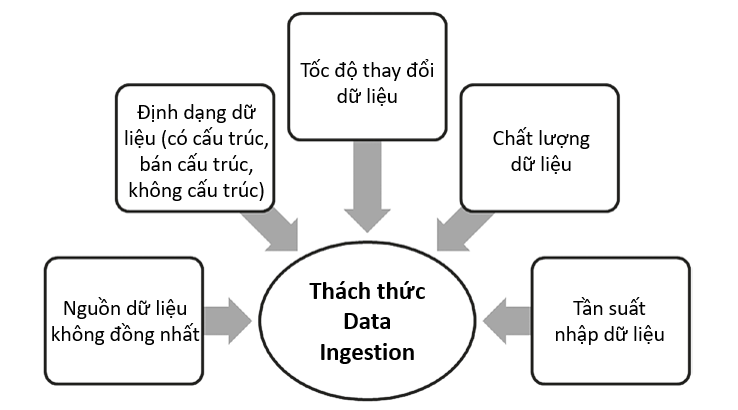
1. Ingestion Layer: tầng này bao gồm hai thành phần là Data Collector (thu thập dữ liệu) và Data Integrator (tích hợp dữ liệu) có nhiệm vụ tương ứng là thu thập dữ liệu và đẩy vào tầng lưu trữ.

Các nguồn dữ liệu từ trong, từ ngoài được tầng này hỗ trợ kết nối với nhau bằng nhiều giao thức. Tầng này có khả năng đẩy dữ liệu dạng theo đợt (batch) cũng như dạng thời gian thực (streaming) vào hệ thống kho dữ liệu hoặc hồ dữ liệu.

Theo như tên gọi, chế độ thời gian thực có nghĩa là các thay đổi được áp dụng cho kho dữ liệu ngay khi chúng xảy ra, trong khi quá trình nhập theo đợt áp dụng các thay đổi theo đợt. Tuy nhiên, một điểm rất cần chú ý khi đề cập đến chế độ thời gian thực đó là thời gian thực có độ trễ riêng giữa sự kiện thay đổi và ứng dụng. Vì lý do này, thời gian thực có thể được hiểu một cách chính xác là gần thời gian thực. Có yếu tố xác định kiểu nhập dữ liệu là tốc độ thay đổi dữ liệu tại nguồn và khối lượng thay đổi. Tốc độ thay đổi là thước đo các thay đổi xảy ra mỗi giờ. Đối với chế độ nhập theo thời gian thực, hệ thống thu thập dữ liệu thay đổi (Change Data Capture – CDC) đã đáp ứng được những nhu cầu thiết yếu. Khung CDC đọc các thay đổi từ nhật ký giao dịch được sao chép trong kho dữ liệu. Độ trễ giữa thời điểm Capture Process ghi nhận thay đổi và tích hợp là rất nhỏ. Các nhà cung cấp phần mềm hàng đầu như Oracle, Pentaho, IBM đều cung cấp các công cụ tích hợp dữ liệu trong thời gian thực.

Trong chế độ nhập theo đợt, các thay đổi được ghi lại và duy trì sau mỗi khoảng thời gian nhất định, sau đó áp dụng cho hồ dữ liệu theo từng khối. Độ trễ dữ liệu là khoảng thời gian giữa công việc thu thập và tích hợp.

Dưới đây là một số thách thức tại tầng Ingestion.



Hình 9: Thách thức của Data Ingestion

*Nguồn: Saurabh Gupta, Venkata Giri (2018)*

1. Data storage Layer

Hồ dữ liệu là nơi lưu trữ dữ liệu tập trung cho toàn bộ dữ liệu của doanh nghiệp. Nó cung cấp nhiều định dạng lưu trữ dữ liệu khác nhau: có cấu trúc, bán cấu trúc, không cấu trúc. Cung cấp khả năng lưu trữ phân tầng giúp giảm thiểu chi phí và có khả năng mở rộng một cách tự động để lưu trữ lên tới hàng Exabytes. Thông thường, hồ dữ liệu sẽ được chia thành các vùng (zone) lưu trữ gồm: landing, hot, cold tùy vào dữ liệu có mức độ sẵn sàng như thế nào. Hồ dữ liệu cho phép phân tích các tập dữ liệu khác nhau theo nhiều phương thức, thậm chí là xử lý AI/ML hay dữ liệu lớn.

*Landing Zone* là một vùng lưu trữ trong hệ thống hồ dữ liệu, được sử dụng để tiếp nhận các nguồn dữ liệu mới và chưa được xử lý. Landing Zone thường là vùng đầu tiên của hệ thống Data Lake, nơi sẽ lưu trữ dữ liệu mới trước khi chúng được xử lý, di chuyển tới các vùng lưu trữ khác của hệ thống. Mục đích của Landing zone là để chứa tất cả các dữ liệu mới được đưa vào hệ thống, bao gồm dữ liệu từ những nguồn bên ngoài, các bản sao của dữ liệu từ các hệ thống khác, các tệp tin dữ liệu tạm thời, và các định dạng dữ liệu khác nhau. Dữ liệu trong Landing Zone thường không được sử dụng trực tiếp cho các mục đích phân tích hoặc hệ thống, mà được xử lý trước khi được di chuyển vào các vùng lưu trữ khác của hệ thống. Các công việc thường thực hiện trong Landing Zone bao gồm: Tiếp nhận và lưu trữ dữ liệu mới từ các nguồn khác nhau; kiểm tra tính toàn vẹn và độ chính xác của dữ liệu; xác định định dạng dữ liệu và đưa ra quyết định về việc di chuyển dữ liệu vào các vùng lưu trữ khác của hệ thống; xử lý dữ liệu để loại bỏ các lỗi hoặc nhiễu trước khi di chuyển vào các vùng lưu trữ khác của hệ thống. Landing Zone giúp giảm thiểu rủi ro mất mát dữ liệu, đồng thời tăng tính linh hoạt và khả năng mở rộng của hệ thống Data Lake. Các dữ liệu mới có thể được đưa vào hệ thống một cách nhanh chóng và an toàn, giúp đảm bảo rằng dữ liệu trong hệ thống Data Lake luôn cập nhật và đầy đủ nhất.

*Hot Area* là một phần của tầng lưu trữ Data Lake được sử dụng để lưu trữ các dữ liệu quan trọng và được truy cập thường xuyên. Đây là nơi lưu trữ dữ liệu có tính tương tác cao và thường được sử dụng để hỗ trợ các ứng dụng phân tích thời gian thực hoặc các “truy vấn phức tạp” trên dữ liệu. Vì các tính chất của Hot Area, nó thường được triển khai trên các “công nghệ lưu trữ” có khả năng xử lý tốt và hiệu quả với các tập dữ liệu lớn. Trong Hot Area bao gồm Raw Zone, Trusted Zone, Refined Zone. Các thành phần này được mô tả như sau:

* Raw Zone là nơi lưu trữ dữ liệu dưới dạng thô mà chưa được xử lý hay biến đổi, dữ liệu tại đây rất đa dạng bao gồm dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc, giúp đảm bảo tính linh hoạt, đáp ứng nhu cầu của nhiều ứng dụng phân tích khác nhau và có khả năng mở rộng.
* Trusted Zone nơi mà dữ liệu đã được xử lý và biến đổi để đảm bảo tính chính xác, đáng tin cậy và dễ sử dụng. Sau khi dữ liệu được thu thập trong Raw Zone, nó sẽ được chuyển đến Trusted Zone để tiếp tục được xử lý. Dữ liệu trong Trusted Zone đã được chuẩn hóa và có cấu trúc hơn với dữ liệu trong Raw Zone, giúp cho việc phân tích và sử dụng dữ liệu dễ dàng hơn. Cùng với đó, dữ liệu cũng được xử lý và kiểm tra tính chính xác để đảm bảo rằng dữ liệu có thể được sử dụng trong các ứng dụng phân tích mà không gặp vấn đề về tính toàn vẹn dữ liệu. Không chỉ vậy, dữ liệu trong Trusted Zone đã được liên kết và đồng nhất, giúp cho việc tạo các bản báo cáo hoặc phân tích trực quan về dữ liệu dễ dàng hơn.
* Refined Zone là nơi dữ liệu đã được biến đổi và làm sạch để sử dụng cho các mục đích phân tích và thực hiện các công việc xử lý dữ liệu phức tạp hơn. Khi dữ liệu được xử lý và kiểm tra tính chính xác ở Trusted Zone, nó sẽ được chuyển đến Refined Zone để tiếp tục làm sạch và chuẩn hóa.

*Cold Area*, vùng lưu trữ cũng như quản lý dữ liệu với khoảng thời gian dài nhưng chi phí lưu trữ thấp. Đây là nơi lưu trữ cho các dữ liệu lịch sử hoặc dữ liệu ít được sử dụng. Dữ liệu trong Cold Area được giữ lại cho mục đích dự báo, phân tích dữ liệu lịch sử và tìm kiếm thông tin trong tương lai. Trong Cold Area có các thành phần như:

* Archived Zone CSV: nơi mà dữ liệu lưu với dạng file CSV và được giữ lại cho mục đích lưu trữ dữ liệu lịch sử với chi phí thấp.
* Raw Zone: dữ liệu được lưu trữ ở định dạng gốc, chưa qua bất kỳ quá trình xử lý hay biến đổi nào.
* Sandbox/Analytical Zone nơi mà các chuyên gia và nhà phân tích dữ liệu có thể truy xuất và làm việc với dữ liệu đã được xử lý hoặc biến đổi từ Raw Zone.

1. Data catalog Layer

Một hồ dữ liệu yêu cầu quản trị dữ liệu. Nếu không có tổ chức, việc tìm kiếm dữ liệu là một quá trình hỗn loạn, không hiệu quả và tốn thời gian. Và nếu quá nhiều thời gian trôi qua mà không có tổ chức và quản trị rõ ràng, giá trị của một hồ dữ liệu có thể sụp đổ dưới dữ liệu tích lũy của chính nó (IBM, 2021). Khối lượng dữ liệu chảy vào kho lưu trữ ngày càng lớn hơn mỗi ngày. Để đảm bảo tính hiệu quả và chính xác, một hình thức quản trị là cần thiết để tạo ra trật tự giữa những luồng dữ liệu chảy vào một cách hỗn loạn này. Để tránh tình trạng hồ dữ liệu trở thành một đầm lầy dữ liệu không thể không cần đến danh mục dữ liệu. Tầng Data catalog cung cấp một nền tảng quản lý danh mục dữ liệu được hỗ trợ bởi kiến trúc quản trị dữ liệu. Một danh mục dữ liệu kết nối người dùng và dữ liệu và hiểu người dùng cần dữ liệu nào. Kiến trúc quản trị dữ liệu đảm bảo việc truy cập tới dữ liệu và chất lượng dữ liệu tuân thủ các quy tắc, quy chuẩn của tổ chức.

1. Data processing Layer

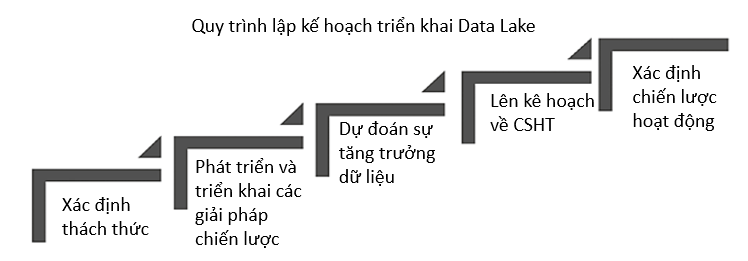
Có nhiệm vụ chuyển đổi dữ liệu sang dạng sử dụng được bằng các bước như: xác thực, làm sạch, chuẩn hóa, biến đổi và làm giàu dữ liệu. Processing Layer giải quyết nhiệm vụ trên nhờ có các thành phần được có khả năng thực hiện nhiều loại chuyển đổi như: SQL, xử lý dữ liệu lớn và ETL gần thời gian thực. Tầng này có khả năng mở rộng quy mô một với số tiền bỏ ra thấp để xử lý dữ liệu với khối lượng lớn và cung cấp các thành phần đáp ứng schema-on-write, schema-on-read, partitioned datasets và các định dạng dữ liệu đa dạng. Tầng này có thể truy cập các giao diện lưu trữ dữ liệu hợp nhất và danh mục chung, do đó truy cập tất cả dữ liệu và siêu dữ liệu trong hồ dữ liệu. Điều này có những lợi ích sau: Tránh dư thừa dữ liệu, di chuyển dữ liệu không cần thiết và trùng lặp mã ETL có thể xảy ra khi xử lý một hồ dữ liệu và kho dữ liệu riêng rẽ, rút giảm thời gian đưa dịch vụ ra thị trường.

1. Data consumption Layer

Tầng sử dụng dữ liệu có thể được dùng mọi dữ liệu lưu trữ ở hồ dữ liệu bằng một giao diện hồ dữ liệu đồng nhất. Lúc này, người dùng không cần phải truy cập vào từng nơi lưu trữ như kho dữ liệu và hồ dữ liệu để có thể truy cập vào toàn bộ dữ liệu trong kho lưu trữ dữ liệu. Khả năng tích hợp tự nhiên giữa kho dữ liệu và hồ dữ liệu giúp cho người dùng có thể linh hoạt thực hiện các công việc sau: Lưu trữ lên tới hàng exabyte dữ liệu có cấu trúc và phi cấu trúc trong hồ dữ liệu một cách tối ưu nhất về chi phí; các dạng dữ liệu có cấu trúc được quản lý, mô hình hóa và phù hợp cao được lưu ở vùng dữ liệu hot của kho dữ liệu.

Quy trình lập kế hoạch triển khai và xây dựng Data Lake

1. Quy trình lập kế hoạch triển khai Data Lake



Hình 10: Quy trình lập kế hoạch triển khai hồ dữ liệu

*Nguồn: Saurabh Gupta, Venkata Giri (2018)*

1. Xác định thách thức

Việc xác định thách thức khi triển khai Data Lake là rất quan trọng vì nó giúp các chuyên gia phân tích đánh giá mức độ phức tạp của quá trình triển khai và đưa ra các giải pháp để giải quyết các thách thức này. Nếu không đưa ra các giải pháp thích hợp cho các thách thức này, quá trình triển khai có thể gặp phải các vấn đề liên quan đến dữ liệu, kiến trúc, quản lý và an ninh. Việc xác định các thách thức còn giúp đảm bảo rằng các mục tiêu và yêu cầu của tổ chức được đáp ứng một cách hiệu quả. Bằng cách đưa ra các giải pháp thích hợp cho các thách thức, tổ chức có thể đảm bảo rằng họ đang sử dụng Data Lake một cách tối ưu nhất để thu thập, xử lý và phân tích dữ liệu của mình. Điều này có thể giúp tổ chức cải thiện hiệu suất và đưa ra các quyết định thông minh dựa trên dữ liệu. Dưới đây là một số các phân tích cần xác định:

*Nhu cầu kinh doanh:* Để xác định các thách thức trong quá trình triển khai Data Lake, cần phải hiểu rõ các nhu cầu và yêu cầu kinh doanh của tập đoàn và đảm bảo rằng Data Lake được thiết kế một cách hợp lý để đáp ứng được các vấn đề này.

*Xác định thách thức về dữ liệu:* Một trong những thách thức lớn nhất khi triển khai Data Lake là quản lý dữ liệu, bao gồm việc lưu trữ, xử lý, truy xuất và bảo mật. Các chuyên gia cần phải xác định các thách thức liên quan đến các quy trình này, bao gồm cách thức thu thập dữ liệu, tích hợp dữ liệu, chuẩn hóa và phân tích dữ liệu.

*Khả năng kỹ thuật:* Cần xác định khả năng kỹ thuật của tập đoàn, bao gồm khả năng lưu trữ và xử lý dữ liệu, cũng như các yêu cầu về bảo mật và tính khả dụng.

*Chi phí:* Triển khai Data Lake có thể đòi hỏi các chi phí đáng kể, bao gồm chi phí cho phần cứng, phần mềm và nhân lực. Cần xác định các chi phí này trước để đảm bảo rằng ngân sách được phân bổ hợp lý.

*Thời gian:* Triển khai Data Lake cũng có thể đòi hỏi thời gian đáng kể, từ việc xây dựng hạ tầng đến tích hợp dữ liệu sau đó là triển khai các thành phần trong giải pháp. Cần xác định thời gian và lập kế hoạch triển khai để đảm bảo sự thành công của dự án.

*Quản lý dữ liệu:* Data Lake có thể cho phép lưu trữ các tập dữ liệu lớn và phức tạp, vì vậy cần phải xác định các quy trình và chính sách để quản lý và bảo vệ dữ liệu.

*Nhân lực:* nhân lực cần có kỹ năng và kiến thức về hạ tầng lưu trữ dữ liệu, các công cụ và kỹ thuật phân tích dữ liệu. Cần xác định các kỹ năng này và cung cấp đào tạo cho nhân viên.

2. Phát triển và triển khai giải pháp chiến lược

Khi lập kế hoạch triển khai hồ dữ liệu, cần phát triển và triển khai các giải pháp chiến lược bởi bốn lý do chính. Thứ nhất, đảm bảo tính nhất quán và hiệu quả trong quản lý dữ liệu: Khi xây dựng hồ dữ liệu, ta phải đảm bảo rằng dữ liệu được lưu trữ và quản lý một cách nhất quán và hiệu quả. Nếu không có giải pháp chiến lược thích hợp, dữ liệu có thể bị lưu trữ ở nhiều nơi khác nhau, gây ra khó khăn trong việc truy xuất, xử lý và quản lý dữ liệu. Thứ hai, Tối ưu hóa việc sử dụng tài nguyên: Giải pháp chiến lược giúp đảm bảo rằng các tài nguyên (như lưu trữ, xử lý) được sử dụng một cách tối ưu nhất, tránh lãng phí và tiết kiệm chi phí. Thứ ba, đảm bảo tính bảo mật và tuân thủ quy định: Giải pháp chiến lược cũng giúp đảm bảo an toàn thông tin và tuân thủ mọi quy phạm pháp luật liên quan đến quản lý dữ liệu. Điều này đặc biệt quan trọng đối với các công ty và tổ chức hoạt động trong các ngành yêu cầu tuân thủ các quy định về bảo mật dữ liệu như viễn thông, ngân hàng, bảo hiểm hay chính phủ. Thứ tư, hỗ trợ quá trình phân tích và ra quyết định: Giải pháp chiến lược còn giúp cải thiện quá trình phân tích dữ liệu và đưa ra quyết định thông minh. Thông qua việc tổ chức dữ liệu đúng theo phương thức mà các nhà phân tích và người ra quyết định có thể dễ dàng truy xuất và hiểu được, giải pháp này giúp cải thiện khả năng đưa ra quyết định dựa trên dữ liệu và tăng hiệu quả trong kinh doanh. Dưới đây là một số điều cần làm để có một giải pháp chiến lược hiệu quả.

Xác định mục tiêu và yêu cầu kinh doanh: Trước khi triển khai Data Lake, cần phải xác định mục tiêu và yêu cầu kinh doanh cụ thể. Điều này giúp đảm bảo rằng giải pháp được thiết kế để đáp ứng nhu cầu kinh doanh của tập đoàn.

Xác định kiến trúc và thiết kế Data Lake: Sau khi xác định mục tiêu kinh doanh, cần phải thiết kế kiến trúc và thiết kế Data Lake để đáp ứng các yêu cầu kinh doanh. Việc này bao gồm các quyết định về nền tảng lưu trữ, tích hợp dữ liệu và quản lý dữ liệu.

Tích hợp dữ liệu: Sau khi thiết kế Data Lake, cần phải tích hợp dữ liệu vào Data Lake. Việc này bao gồm xác định nguồn dữ liệu, thiết lập các quy trình tích hợp và chuyển đổi dữ liệu để đảm bảo rằng dữ liệu được tích hợp và chuẩn hóa đúng cách.

Xây dựng phương án lưu trữ dữ liệu: Dữ liệu sau khi được tích hợp cần có không gian lưu trữ để phục vụ cho các mục đích sau này. Dựa trên tính thường xuyên truy cập và giá trị của dữ liệu mà dữ liệu được chia thành hai loại là Hot data và Cold data. Hot data là những dữ liệu có tính thường xuyên truy cập cao, được lưu trữ và xử lý trong các hệ thống có thể đáp ứng yêu cầu xử lý tốc độ cao và có thể truy cập nhanh. Cold data là những dữ liệu ít được truy cập hoặc không được truy cập trong một thời gian dài. Cần xác định phương án lưu trữ đối với từng loại dữ liệu trên để có thể khai thác một cách hiệu quả. Thông thường, Hot data được lưu trữ trong các hệ thống lưu trữ như hệ thống quản lý cơ sở dữ liệu, hệ thống lưu trữ flash hoặc bộ nhớ trong để đảm bảo tốc độ truy xuất nhanh. Với Cold data được lưu trữ trong các hệ thống lưu trữ có chi phí thấp hơn, chẳng hạn như hệ thống lưu trữ đám mây hoặc hệ thống lưu trữ đĩa cứng, với mục đích tiết kiệm chi phí lưu trữ và bảo quản dữ liệu. Tuy nhiên, khi cần truy cập đến Cold data, quá trình truy xuất và xử lý thường yêu cầu nhiều thời gian hơn và tốc độ truy xuất chậm hơn so với Hot data.

Xây dựng các ứng dụng phân tích dữ liệu: Dữ liệu đã được lưu trữ và sẵn sàng được khai thác, cần phải xây dựng các ứng dụng phân tích dữ liệu để đáp ứng nhu cầu kinh doanh. Việc này bao gồm việc xác định các nhu cầu trong việc phân tích dữ liệu, từ đó lực chọn các ứng dụng phù hợp, phát triển các mô hình và triển khai chúng vào các ứng dụng phân tích dữ liệu.

Giải pháp quản trị dữ liệu: từ những yêu cầu và phân tích thông tin thu thập được từ người dùng đưa ra khung quản trị dữ liệu. Khung quản trị dữ liệu bao gồm các chính sách, quy tắc, quy trình, cơ cấu tổ chức và công nghệ được đưa ra như một phần của chương trình quản trị. Nó cũng giải thích rõ ràng những thứ như tuyên bố sứ mệnh cho chương trình, mục tiêu của chương trình và cách đo lường thành công của chương trình, cũng như trách nhiệm ra quyết định và trách nhiệm giải trình đối với các chức năng khác nhau sẽ là một phần của chương trình. Khuôn khổ quản trị của một tổ chức nên được lập thành văn bản và chia sẻ nội bộ, vì vậy mọi người liên quan đều rõ ràng, biết trước chương trình sẽ hoạt động như thế nào. Về mặt công nghệ, phần mềm quản trị dữ liệu có thể được sử dụng để tự động hóa các phương diện của việc quản lý một chương trình quản trị. Mặc dù các công cụ quản trị dữ liệu không phải là thành phần khung bắt buộc, nhưng chúng hỗ trợ quản lý quy trình và chương trình, cộng tác, phát triển chính sách quản trị, tài liệu quy trình, tạo danh mục dữ liệu và các chức năng khác. Chúng cũng có nhiều trường hợp được sử dụng cùng với các công cụ chất lượng dữ liệu, quản lý siêu dữ liệu (metadata) và quản lý dữ liệu chủ (master) để đem lại hiệu quả.

Thử nghiệm và đánh giá: Sau khi triển khai Data Lake và các ứng dụng phân tích dữ liệu, cần phải thử nghiệm và đánh giá giải pháp. Việc này giúp đảm bảo rằng Data Lake và các ứng dụng phân tích dữ liệu đạt được mọi yêu cầu của tổ chức và đáp ứng được các mục tiêu.

Quản lý và bảo trì: Sau khi triển khai Data Lake và các ứng dụng phân tích dữ liệu, cần phải quản lý và bảo trì giải pháp. Việc này bao gồm quản lý và bảo vệ dữ liệu, cập nhật và nâng cấp hạ tầng và các ứng dụng phân tích dữ liệu.

3. Dự đoán sự tăng trưởng của dữ liệu

Khi lập kế hoạch triển khai Data Lake, việc dự đoán sự tăng trưởng của dữ liệu là một bước quan trọng để đảm bảo rằng giải pháp được thiết kế có thể đáp ứng nhu cầu trong tương lai. Dưới đây là một số điểm cần phân tích khi dự đoán sự tăng trưởng của dữ liệu:

*Phân tích xu hướng sử dụng dữ liệu hiện tại:* Để dự đoán sự tăng trưởng của dữ liệu, cần phân tích những xu hướng sử dụng dữ liệu hiện tại và dựa trên đó để ước tính mức độ tăng trưởng trong tương lai.

*Xác định những nguồn dữ liệu tiềm năng:* Việc xác định các nguồn dữ liệu tiềm năng trong tương lai là rất quan trọng để đảm bảo rằng giải pháp có thể đáp ứng được nhu cầu trong tương lai.

*Đánh giá tốc độ tăng trưởng của dữ liệu:* Dựa trên những dữ liệu có sẵn và các nguồn dữ liệu tiềm năng, cần đánh giá tốc độ tăng trưởng của dữ liệu trong một khoảng thời gian nhất định.

*Dự đoán sự tăng trưởng dữ liệu trong tương lai:* Dựa trên các thông tin đã thu thập được, cần dự đoán sự tăng trưởng dữ liệu trong tương lai và tính toán dung lượng lưu trữ cần thiết để đáp ứng nhu cầu.

*Đưa ra phương án mở rộng:* Dựa trên dự đoán sự tăng trưởng của dữ liệu, cần đưa ra phương án mở rộng để đảm bảo rằng giải pháp có thể đáp ứng được nhu cầu trong tương lai một cách hiệu quả.

Việc dự đoán sự tăng trưởng của dữ liệu là một phần quan trọng trong quá trình lập kế hoạch triển khai Data Lake. Nó giúp đảm bảo rằng giải pháp được thiết kế có thể đáp ứng nhu cầu trong tương lai và đưa ra phương án mở rộng khi cần thiết.

4. Lên kế hoạch về cơ sở hạ tầng

Việc lên kế hoạch về cơ sở hạ tầng là một bước rất quan trọng khi lập kế hoạch triển khai Data Lake. Dưới đây là các bước chính:

*Xác định yêu cầu về cơ sở hạ tầng*: Trước khi triển khai Data Lake, cần phải xác định yêu cầu về cơ sở hạ tầng, bao gồm định lượng tài nguyên (bộ nhớ, lưu trữ, xử lý) và độ tin cậy của hệ thống.

*Chọn nền tảng cơ sở hạ tầng:* Sau khi xác định yêu cầu, việc cần làm tiếp theo là chọn nền tảng cơ sở hạ tầng phù hợp, bao gồm cả phần cứng, phần mềm và hệ thống mạng kết nối.

*Thiết kế kiến trúc:* Khi chọn được nền tảng cơ sở hạ tầng, cần phải thiết kế kiến trúc của hệ thống. Điều này bao gồm xác định các khối lượng dữ liệu, các ứng dụng sử dụng dữ liệu, quy trình xử lý dữ liệu, định dạng lưu trữ dữ liệu và phân tích cơ sở dữ liệu.

*Cài đặt và triển khai:* dựa trên kiến trúc đã được thiết kế sẽ tiến hành cài đặt và triển khai hệ thống. Điều này bao gồm cài đặt phần mềm, cấu hình hệ thống, và đảm bảo tính toàn vẹn của dữ liệu.

*Vận hành và bảo trì:* sau khi cài đặt và triển khai hệ thống thành công, có khả năng hệ thống gặp các sự cố trong quá trình vận hành nên công việc vận hành và bảo trì hệ thống phải chắc chắn được triển khai đảm bảo. Điều này bao gồm giám sát hệ thống, xử lý các sự cố và tối ưu hóa hiệu suất của hệ thống.

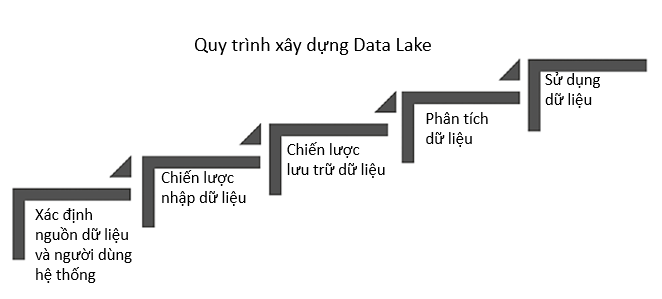
*Nâng cấp và mở rộng:* Cuối cùng, để đàm bảo hệ thống luôn được ở trạng thái tối ưu nhất, đáp ứng được nhiều nhu cầu đa dạng của người dùng sẽ cần phải nâng cấp và mở rộng hệ thống khi có nhu cầu. Việc này có thể bao gồm thêm phần cứng, tăng khả năng lưu trữ hoặc tăng hiệu suất xử lý.

5. Xác định chiến lược hoạt động

Việc xác định chiến lược hoạt động là một bước rất quan trọng trong quá trình lập kế hoạch triển khai Data Lake. Dưới đây là một số bước cần phân tích để xác định chiến lược hoạt động:

* Xác định mục đích và mục tiêu của Data Lake: Việc xác định mục đích và mục tiêu sẽ giúp hiểu rõ hơn về những gì doanh nghiệp cần từ Data Lake và điều gì sẽ giúp đạt được mục tiêu đó. Mỗi tổ chức cần phải xác định rõ mục đích sử dụng dữ liệu của mình là gì, như là phân tích dữ liệu để đưa ra quyết định, hoặc để nghiên cứu thị trường hay cải thiện trải nghiệm khách hàng.
* Phân tích các yêu cầu về dữ liệu: Việc phân tích các yêu cầu về dữ liệu sẽ giúp xác định loại dữ liệu cần sử dụng, dữ liệu nào sẽ được lưu trữ và cách lưu trữ. Từ đó đảm bảo dữ liệu lưu trữ là những dữ liệu có giá trị sử dụng, ở trạng thái sẵn sàng cao và tiết kiệm chi phí.
* Xác định các nguồn dữ liệu: Bạn cần phải xác định các nguồn dữ liệu bạn muốn sử dụng trong Data Lake. Những nguồn dữ liệu này có thể là các hệ thống CRM, ERP hoặc các hệ thống khác đang được sử dụng trong việc kinh doanh.
* Xác định các công nghệ phù hợp: Việc xác định các công nghệ phù hợp với nhu cầu của người dùng là rất quan trọng. Với những nhu cầu và thực trạng của dữ liệu thực tế, cần cân nhắc, xem xét và phân tích các công nghệ lưu trữ dữ liệu hoặc các công nghệ truy xuất dữ liệu và lựa chọn ra công nghệ phù hợp nhất.
* Xây dựng mô hình hoạt động của Data Lake: Sau khi xác định các yêu cầu và các công nghệ phù hợp, cần xác định mô hình hoạt động của Data Lake cho doanh nghiệp. Mô hình này sẽ mô tả cách thức thu thập, lưu trữ, xử lý và truy xuất dữ liệu trong Data Lake.
* Xây dựng kế hoạch triển khai: Cuối cùng, bạn cần phải xây dựng kế hoạch triển khai Data Lake của mình, bao gồm lịch trình triển khai, chi phí và các bước triển khai cụ thể. Kế hoạch triển khai cần phải được thiết kế sau cùng để đảm bảo đáp ứng chính xác nhất mọi nhu cầu.

1. Quy trình xây dựng Data Lake



Hình 11: Quy trình xây dựng Data Lake

*Nguồn: Saurabh Gupta, Venkata Giri (2018)*

1. Xác định nguồn dữ liệu và người dùng hệ thống

Xác định nguồn dữ liệu và người dùng hệ thống là bước đầu tiên và thiết yếu trong việc xây dựng hồ dữ liệu. Bước này bao gồm các công việc liên quan đến việc tìm hiểu thông tin về các nguồn dữ liệu sẽ được thu thập như:

* Các nguồn dữ liệu được lấy từ những hệ thống nào?
* Được truyền trực tiếp hay theo đợt?
* Cấu trúc dữ liệu ra sao?
* Khối lượng dữ liệu hiện nay là bao nhiêu?
* Tốc độ tăng trưởng của dữ liệu v.v.

Cùng với việc xác định nguồn dữ liệu, cần xác định các người dùng:

* Ai là người sử dụng dữ liệu sau khi thu thập?
* Mục đích sử dụng dữ liệu là gì?
* Những ai có quyền truy cập vào dữ liệu và yêu cầu của họ là gì? v.v.

Những thông tin được xác định trên sẽ là cơ sở cho việc triển khai các bước tiếp theo để đảm bảo hồ dữ liệu phục vụ được cho cả yêu cầu trước mắt và trong tương lai.

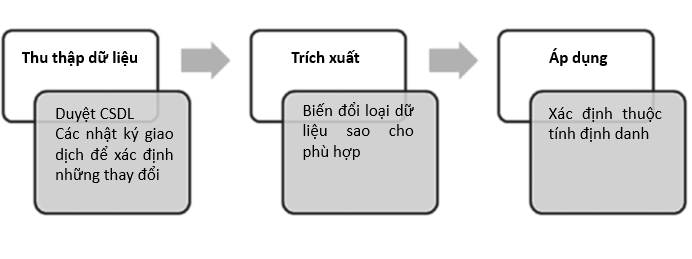
2. Chiến lược nhập dữ liệu (Data Ingestion Strategy)

Việc nhập dữ liệu là việc thu thập dữ liệu từ nhiều nguồn và nhập chúng vào kho lưu trữ. Nhập dữ liệu đảm bảo kết nối an toàn đến các nguồn khác nhau, nắm bắt những thay đổi và ghi lại chúng trong hồ dữ liệu. Việc nhập dữ liệu giữ cho hồ dữ liệu nhất quán với các thay đổi của dữ liệu tại các hệ thống nguồn, từ đó làm cho hồ dữ liệu trở thành một trạm lưu trữ dữ liệu doanh nghiệp duy nhất. Dữ liệu được thu thập vào hồ dữ liệu là rất lớn, nhưng không phải tất cả đều có giá trị. Càng nhiều dữ liệu kém chất lượng xâm nhập vào hồ dữ liệu sẽ làm cho hồ dữ liệu nhanh trở thành đầm lầy dữ liệu. Vì vậy, cần xác định một chiến lược chính xác.

Có ba chế độ nhập dữ liệu như sau:

* Batch ingestion (nhập theo đợt): thu thập dữ liệu và truyền dữ liệu đến đích trong khoảng thời gian như hàng ngày, hàng tuần, hàng tháng, v.v.
* Streaming ingestion (nhập trực tuyến): thu thập và truyền dữ liệu đến đích ngay khi chúng xảy ra, tuy nhiên vẫn có độ trễ riêng nên có thể hiểu đây là kiểu nhập dữ liệu gần thời gian thực.
* Hybrid streaming (truyền trực tuyến kết hợp): sử dụng kết hợp giữa nhập dữ liệu theo đợt và trực tuyến. Trong đó, chỉ định một loại dữ liệu nhất định để truyền trực tuyến và các loại khác cho hàng loạt.

Trong quá trình nhập dữ liệu, dữ liệu trong các hệ thống có thể thay đổi. Để đảm bảo dữ liệu luôn được đồng nhất, nâng cao tính sẵn sàng của dữ liệu nên xử lý các giao dịch ngay khi nó xảy ra. Trong xử lý gần thời gian thực các thay đổi được ghi lại ở độ trễ rất thấp. Bài toán xử lý thời gian thực dựa trên nhật ký được gọi là Change Data capture (CDC). CDC đề cập đến quy trình khai thác nhật ký để chỉ thu thập dữ liệu đã thay đổi (chèn, cập nhật, xóa) khỏi nhật ký giao dịch. CDC thời gian thực hoặc theo lô với khoảng thời gian ngắn phát hiện các sự kiện thay đổi bằng cách quét nhật ký cơ sở dữ liệu khi chúng xảy ra. Với quyền truy cập tối thiểu vào các nguồn doanh nghiệp, CDC không phải chịu tải trên các bảng nguồn; do đó giảm thiểu độ trễ và đảm bảo tính nhất quán giữa hệ thống nguồn và đích. (Saurabh Gupta, 2018)



Hình 12: Luồng hoạt động Change Data Capture

*Nguồn: Saurabh Gupta, Venkata Giri (2018)*

Một câu hỏi phổ biến được đặt ra khi đề cập đến Data Ingestion đó là Data Ingestion khác gì so với ETL? Data Ingestion và ETL là các phần khác nhau của cùng một quy trình làm việc. Trước hết, nhập dữ liệu từ các nguồn, sau đó khi dữ liệu đã sẵn sàng sẽ trích xuất, biến đổi và tải dữ liệu đó bằng cách sử dụng một đường dẫn dữ liệu để di chuyển nó đến một đích khác. Data Ingestion là một thuật ngữ rộng hơn nhiều so với ETL. Data Ingestion đề cập đến quá trình chung nhập dữ liệu từ hàng trăm, hàng nghìn nguồn và chuẩn bị dữ liệu để chuyển đến đích. ETL là một hành động hay công việc rất cụ thể. Tuy nhiên thì việc nhập và truyền dữ liệu rất ít khi không cần sử dụng ETL. Một thuật ngữ khác đang dần được phổ biến hơn ETL đó là ELT (trích xuất và tải dữ liệu vào kho lưu trữ trước khi biến đổi để sử dụng). Các nhóm dữ liệu giảm được lo lắng về chi phí cho các công cụ phân tích, vì giờ đây họ sẽ tải mọi thứ và phân loại sau. Nhưng ngay cả khi chỉ di chuyển dữ liệu, vẫn phải kiểm tra chất lượng và làm sạch dữ liệu. Vì vậy khung nhập dữ liệu ra đời để giải quyết vấn đề trên.

Khung nhập dữ liệu tiêu chuẩn sẽ bao gồm hai thành phần và thu thập dữ liệu và tích hợp dữ liệu. Trong khi bộ thu thập dữ liệu chịu trách nhiệm thu thập hoặc lấy dữ liệu từ nguồn dữ liệu, thì thành phần bộ tích hợp dữ liệu đảm nhiệm việc nhập dữ liệu vào hồ dữ liệu. Dưới đây là một số khía cạnh cần xem xét khi xây dựng khung nhập dữ liệu:

* Dòng dữ liệu – việc xác định danh mục các hệ thống nguồn và hiểu dòng dữ liệu từ lúc bắt đầu đến điểm nhập dữ liệu là một thông tin hữu ích. Điều này giúp những nhà quản trị dữ liệu có thể xem xét theo thời gian để sắp xếp người dùng theo các quy định hiện hành.
* Định dạng dữ liệu – dữ liệu đến ở dạng block hay object.
* Hiệu suất và tốc độ thay đổi dữ liệu – tốc độ thay đổi dữ liệu được định nghĩa là quy mô của những thay đổi xảy ra mỗi giờ. Nó giúp cho việc chọn lựa công cụ nhập thích hợp trong khung nhập dữ liệu. Hiệu suất là kết quả của thông lượng và độ trễ.
* Vị trí và bảo mật dữ liệu: xác định dữ liệu được đặt tại chỗ (on-prem) hay trên đám mây (cloud), khung nhập dữ liệu cần thiết lập một quá trình đảm bảo an toàn để thu thập dữ liệu.
* Kích thước dữ liệu truyền đi: kích thước trung bình và tối đa của block hay object trong một lần nhập là bao nhiêu để xác định cách nén hoặc chia nhỏ tệp dữ liệu.
* Định dạng dữ liệu đích: dữ liệu về hệ thống nguồn cần được tương thích với giải pháp lưu trữ.

3. Chiến lược lưu trữ dữ liệu

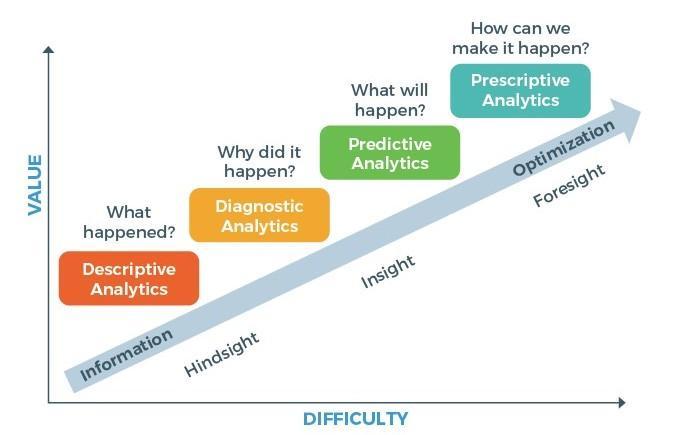
Hồ dữ liệu là nơi lưu trữ dữ liệu tập trung cho toàn bộ dữ liệu của tổ chức. Để dữ liệu được lưu trữ một cách hợp lý nhất cần đáp ứng các yếu tố sau:

* Sử dụng hệ thống phân tán để lưu trữ dữ liệu: Hệ thống phân tán có thể giúp phân tán dữ liệu vào nhiều node khác nhau để đảm bảo tính sẵn sàng và khả năng chịu lỗi cao hơn.
* Tạo các vùng lưu trữ khác nhau cho từng giai đoạn của quy trình xử lý dữ liệu: Các vùng lưu trữ này có thể được chia thành Landing Zone, Hot Area và Cold Area như đã trình bày ở trên. Landing Zone dùng để tiếp nhận dữ liệu mới và đảm bảo tính toàn vẹn của dữ liệu. Hot Area sẽ lưu trữ các dữ liệu quan trọng và được truy cập thường xuyên. Các vùng lưu trữ trong Cold Area được sử dụng để lưu trữ các bản ghi dữ liệu lịch sử, lưu trữ dữ liệu dữ liệu được lưu trữ và quản lý trong thời gian dài với chi phí lưu trữ thấp và các tệp định dạng tĩnh.
* Xác định các kỹ thuật lưu trữ dữ liệu phù hợp: Các kỹ thuật lưu trữ dữ liệu như partitioning (chia dữ liệu thành các phần nhỏ), compression (giảm kích thước của dữ liệu tăng tốc độ truyền tải dữ liệu cùng với đoa là để tiết kiệm không gian lưu trữ) và indexing (truy xuất và tìm kiếm dữ liệu nhanh chóng và hiệu quả hơn) có thể được sử dụng để giảm thiểu chi phí lưu trữ dữ liệu và tối ưu hóa hiệu suất.
* Đảm bảo tính bảo mật và quyền riêng tư của dữ liệu: Khi lưu trữ dữ liệu trong data lake, đảm bảo rằng dữ liệu được bảo vệ và quản lý theo đúng các quy định pháp luật về bảo mật và quyền riêng tư.
* Quản lý metadata: Metadata là các thông tin mô tả về dữ liệu và các thuộc tính của chúng, bao gồm các thông tin như tên, loại dữ liệu, định dạng, kích thước, nguồn dữ liệu, ngày tạo, người tạo, và các thông tin khác liên quan đến dữ liệu. Metadata là một phần quan trọng trong việc quản lý dữ liệu và đảm bảo tính toàn vẹn của dữ liệu trong một hệ thống lưu trữ. Quản lý metadata có thể giúp cho người dùng dễ dàng tìm kiếm, phân tích và sử dụng dữ liệu. Có nhiều các công cụ quản lý metadata có thể được sử dụng để tự động hóa quá trình này. Cần lựa chọn một công cụ phù hợp.

4. Phân tích dữ liệu

Bước này cho phép dữ liệu được biến đổi thành các cấu trúc được tối ưu hóa cho việc phân tích. Các loại phân tích trong phân tích dữ liệu bao gồm:

* Descriptive Analytics (phân tích mô tả): là việc hiểu dữ liệu lịch sử, xác định các vấn đề, mô tả quá khứ và trả lời cho câu hỏi điều gì đã xảy ra.
* Diagnostic Analytics (phân tích chẩn đoán): “phân tích chẩn đoán” là tìm nguyên nhân gốc rễ của các vấn đề, đưa ra được các nhận định mở rộng hơn về đặc điểm của dữ liệu quá khứ.
* Predictive Analytics (phân tích dự đoán): như tên gọi, loại phân tích này giúp dự đoán kết quả trong tương lai.
* Prescriptive analytics (phân tích đề xuất): đề xuất các hành động tiếp theo dựa trên những phân tích dự đoán. (Dilmegani)



Hình 13: Thứ tự các loại phân tích theo độ khó và giá trị trong phân tích dữ liệu

*Nguồn: Prescriptive Analytics: Optimize Business Decisions in 2023 (Cem Dilmegani)*

Đề có thể đáp ứng được những loại phân tích trên, tại bước này cũng cần xác định các công cụ phù hợp, xem xét các loại truy vấn, các mô hình phân tích sẽ cần thiết cho dữ liệu. Điều này có thể liên quan đến việc làm việc với các phòng ban để xác định những yêu cầu, công việc, KPI/chỉ số quan trọng mà các bên liên quan cần biết và điều đó sẽ giúp họ đưa ra quyết định. Thêm nữa là tìm hiểu và lựa chọn các công cụ thích hợp phân tích để đáp ứng được các yêu cầu của mỗi phòng ban.

5. Sử dụng dữ liệu

Đối với việc sử dụng dữ liệu, một số các vấn đề cần quan tâm như:

* Bảo mật dữ liệu: Dữ liệu trong Data Lake thường là các tập tin không cấu trúc và có thể được truy cập từ nhiều nguồn khác nhau. Do đó, cần đảm bảo rằng dữ liệu được bảo mật và chỉ được truy cập bởi những người được ủy quyền.
* Kiểm soát quyền truy cập: Phải xác định và kiểm soát chính xác quyền truy cập cho từng người dùng, đảm bảo rằng chỉ những người có đủ quyền được truy cập vào dữ liệu.
* Quản lý dữ liệu: Dữ liệu trong Data Lake có thể rất lớn và phức tạp. Cần có một quy trình quản lý dữ liệu để đảm bảo tính toàn vẹn của dữ liệu, bao gồm đảm bảo rằng dữ liệu được cập nhật và sao lưu định kỳ.
* Tối ưu hóa truy vấn: Vì dữ liệu trong Data Lake có thể rất lớn, việc truy vấn dữ liệu có thể mất nhiều thời gian và tài nguyên. Cần tối ưu hóa truy vấn để đảm bảo hiệu quả và tốc độ truy vấn.
* Kiểm soát chất lượng dữ liệu: Dữ liệu trong Data Lake có thể đến từ nhiều nguồn khác nhau và có thể không được chuẩn hóa hoặc kiểm soát chất lượng đầy đủ. Cần đảm bảo rằng dữ liệu được chuẩn hóa và kiểm soát chất lượng trước khi sử dụng.
* Khả năng mở rộng: Data Lake có khả năng mở rộng linh hoạt để xử lý và lưu trữ dữ liệu lớn. Do đó, cần đảm bảo rằng hệ thống có thể mở rộng khi cần thiết và đáp ứng được nhu cầu sử dụng dữ liệu trong tương lai.

Một số công cụ sử dụng

1. IBM Information Server

Là một nền tảng hỗ trợ tích hợp dữ liệu giúp người dùng dễ dàng hiểu, làm sạch, quản lý và biến đổi dữ liệu. IBM Information Server là một nền tảng tích hợp dữ liệu và quản lý dữ liệu do IBM phát triển. Nó cung cấp các công cụ để tự động hóa các quy trình tích hợp, quản lý và chuyển đổi dữ liệu giữa các hệ thống và ứng dụng khác nhau.

Kiến trúc của IBM Information Server bao gồm 5 khối:

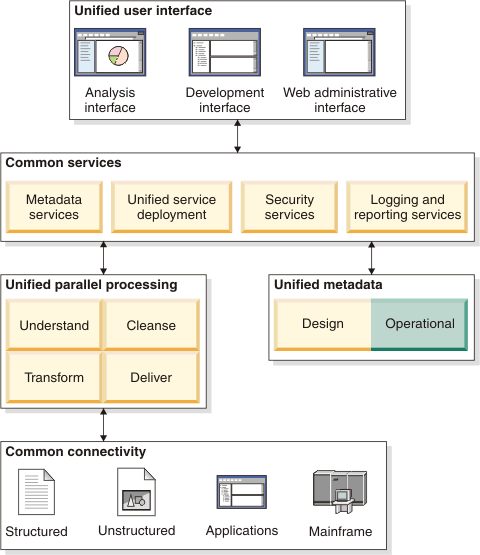
Đầu tiên là khối *Common connectivity*, IBM Information Server kết nối với các nguồn thông tin cho dù chúng có cấu trúc hay không có cấu trúc, trên máy tính lớn hay các ứng dụng. Khả năng kết nối dựa trên siêu dữ liệu được chia sẻ trên các thành phần của bộ ứng dụng và các đối tượng kết nối có thể tái sử dụng.

*Unified parallel processing*, phần lớn công việc của IBM Information Server thực hiện diễn ra trong bộ máy xử lý song song. Công cụ xử lý song song này được thiết kế để mang lại những lợi ích sau: Xử lý song song và truyền dẫn dữ liệu để hoàn thành khối lượng công việc ngày càng tăng trong khoảng thời gian ngắn hơn. Khả năng mở rộng bằng cách thêm phần cứng (ví dụ: bộ xử lý hoặc nút trong lưới) mà không thay đổi thiết kế tích hợp dữ liệu. Cơ sở dữ liệu, tệp và xử lý hàng đợi được tối ưu hóa để xử lý các tệp lớn không thể chứa tất cả trong bộ nhớ cùng một lúc hoặc với số lượng lớn các tệp nhỏ.

*Unified metadata*, InfoSphere Information Server được xây dựng trên cơ sở hạ tầng siêu dữ liệu thống nhất cho phép chia sẻ thông tin giữa các lĩnh vực kinh doanh và kỹ thuật. Cơ sở hạ tầng này giúp giảm thời gian phát triển và cung cấp tài liệu liên tục có thể cải thiện độ tin cậy của thông tin. Tất cả các chức năng của InfoSphere Information Server đều chia sẻ cùng một siêu mô hình, giúp các vai trò và chức năng khác nhau cộng tác dễ dàng hơn. Một kho lưu trữ siêu dữ liệu chung cung cấp khả năng lưu trữ lâu dài cho tất cả các thành phần của bộ InfoSphere Information Server. Tất cả các sản phẩm đều phụ thuộc vào kho lưu trữ để điều hướng, truy vấn và cập nhật siêu dữ liệu. Kho chứa hai loại siêu dữ liệu: Dynamic bao gồm thông tin thời gian thiết kế và Operational bao gồm dữ liệu giám sát hiệu suất, kiểm toán và nhật ký cũng như dữ liệu mẫu lập hồ sơ dữ liệu.

*Common services,* InfoSphere Information Server được xây dựng hoàn toàn trên một tập hợp các dịch vụ chia sẻ mà các tác vụ chính được tập hợp trên cùng nền tảng. Chúng bao gồm các tác vụ quản trị như bảo mật, quản trị người dùng, ghi nhật ký và báo cáo. Các dịch vụ được chia sẻ cho phép các tác vụ này được quản lý và kiểm soát ở một nơi, bất kể thành phần bộ nào đang được sử dụng. Các dịch vụ phổ biến cũng bao gồm các dịch vụ siêu dữ liệu, cung cấp quyền truy cập và phân tích siêu dữ liệu theo định hướng dịch vụ tiêu chuẩn trên nền tảng. Ngoài ra, tầng dịch vụ chung quản lý cách các dịch vụ được triển khai từ bất kỳ chức năng nào của sản phẩm, cho phép các quy tắc làm sạch và chuyển đổi hoặc truy vấn liên kết được xuất bản dưới dạng các dịch vụ được chia sẻ trong SOA, sử dụng cơ chế nhất quán và dễ sử dụng.

*Unified user interface,* giao diện người dùng của InfoSphere Information Server bao gồm các điều khiển trực quan và trải nghiệm người dùng trên các sản phẩm. Các chức năng phổ biến như duyệt danh mục, nhập siêu dữ liệu, truy vấn và duyệt dữ liệu đều hiển thị các dịch vụ chung cơ bản theo cách thống nhất. InfoSphere Information Server cung cấp các giao diện cho công việc phát triển có độ chi tiết cao và cả giao diện chạy trong trình duyệt web cho việc quản trị. Giao diện lập trình ứng dụng (API) hỗ trợ nhiều kiểu giao diện bao gồm yêu cầu trả lời tiêu chuẩn, hướng dịch vụ, hướng sự kiện và gọi tác vụ theo lịch trình. (IBM, 2021)



Hình 14: Kiến trúc IBM Information Server

*Nguồn: IBM*

IBM Information Server bao gồm một loạt các thành phần như sau:

*IBM InfoSphere DataStage*: IBM InfoSphere DataStage là một công cụ tích hợp dữ liệu và ETL (Extract, Transform, Load) được phát triển bởi IBM. Nó cung cấp các tính năng để trích xuất dữ liệu từ nhiều nguồn khác nhau, biến đổi dữ liệu và tải chúng vào các mục tiêu khác nhau, bao gồm cả các hệ thống lưu trữ khác nhau và data warehouse. Các tính năng chính của IBM InfoSphere DataStage bao gồm:

* Hỗ trợ đa nguồn dữ liệu: Có thể trích xuất dữ liệu từ các nguồn khác nhau như database, file, web service, message queue, ...
* Thao tác dữ liệu linh hoạt: Có thể biến đổi dữ liệu bằng các hoạt động như join, sort, filter, ...
* Kiểm soát chất lượng dữ liệu: Cung cấp các tính năng để đảm bảo chất lượng dữ liệu, bao gồm kiểm tra và sửa lỗi dữ liệu. Không chỉ vậy còn có thể tự động quét dữ liệu định kỳ tự động.
* Hỗ trợ nhiều môi trường: Có thể hoạt động trên nhiều hệ điều hành và hệ thống lưu trữ khác nhau.
* Quản lý tiến trình ETL: Cung cấp một giao diện để quản lý và theo dõi tiến trình ETL, bao gồm xử lý lỗi và khôi phục sau sự cố. Người dùng cũng có thể đặt lịch job chạy ETL theo yêu cầu.

*IBM InfoSphere Information Governance Catalog (IGC)*: công cụ quản lý dữ liệu giúp tổ chức phát hiện, truy cập, đánh giá và quản lý thông tin liên quan đến dữ liệu. Nó cho phép tổ chức quản lý và kiểm soát dữ liệu của mình thông qua việc tạo, quản lý, tìm kiếm, định nghĩa và theo dõi metadata. Các tính năng chính của IGC bao gồm:

* Quản lý metadata: IGC cho phép tổ chức tạo, quản lý và theo dõi metadata của dữ liệu trong toàn bộ hệ thống.
* Tìm kiếm metadata: IGC cung cấp khả năng tìm kiếm metadata bằng cách sử dụng các thuộc tính, các thuộc tính con, tên, từ khóa và các tiêu chí khác.
* Hỗ trợ các chuẩn dữ liệu: IGC hỗ trợ các chuẩn dữ liệu chính như ISO 11179, IBM InfoSphere DataStage, IBM InfoSphere Information Server và các chuẩn dữ liệu khác.
* Tương tác với các công cụ khác của IBM: IGC có thể tương tác với các công cụ khác của IBM như IBM InfoSphere Master Data Management và IBM InfoSphere DataStage.
* Theo dõi metadata: IGC cho phép tổ chức theo dõi các thay đổi trong metadata và phát hiện ra các lỗ hổng và sai sót trong quản lý dữ liệu.

*IBM InfoSphere Information Analyzer:* công cụ phân tích dữ liệu giúp đánh giá chất lượng dữ liệu, khám phá quan hệ giữa các dữ liệu và giúp tổ chức hiểu rõ hơn về dữ liệu của mình. Thông qua việc phân tích metadata và các tập tin dữ liệu, IBM InfoSphere Information Analyzer đem đến tính năng tự động nhận biết các lỗi dữ liệu, kiểm tra tính đầy đủ, tính toàn vẹn, tính chính xác của dữ liệu và xác định các mối quan hệ giữa các tập tin dữ liệu khác nhau. Ngoài ra, công cụ này cũng cung cấp khả năng phân tích dữ liệu dựa trên các chuẩn và quy tắc được định nghĩa trước để đảm bảo tính nhất quán của dữ liệu.

*IBM InfoSphere QualityStage:* Được thiết kế để giúp các tổ chức hiểu và cải thiện chất lượng tổng thể tài sản dữ liệu của họ, IBM InfoSphere QualityStage cung cấp các tính năng nâng cao để giúp kiểm tra, sửa chữa, hợp nhất và xác thực dữ liệu không đồng nhất trong quy trình làm việc tích hợp.

*IBM InfoSphere Information Services Director:* IBM Information Server cung cấp một cơ chế hợp nhất để xuất bản và quản lý các dịch vụ Kiến trúc hướng dịch vụ (SOA) được chia sẻ trên các chức năng liên kết, chuyển đổi dữ liệu và chất lượng dữ liệu, cho phép các chuyên gia thông tin dễ dàng triển khai các dịch vụ cho bất kỳ nhiệm vụ tích hợp thông tin nào và quản lý chúng một cách nhất quán. Điều này cho phép các nhà phát triển sử dụng logic tích hợp dữ liệu được xây dựng bằng IBM Information Server và xuất bản nó dưới dạng dịch vụ “luôn bật” – chỉ trong vài phút. Các dịch vụ phổ biến này cũng bao gồm các dịch vụ siêu dữ liệu, cung cấp quyền truy cập và phân tích siêu dữ liệu theo định hướng dịch vụ tiêu chuẩn trên nền tảng.

*IBM InfoSphere FastTrack:* công cụ tự động hóa quy trình tích hợp dữ liệu cho phép các chuyên gia không cần phải là nhà phát triển phần mềm để xây dựng các luồng dữ liệu phức tạp.

*IBM InfoSphere Business Glossary:* IBM Information Server cung cấp một công cụ dựa trên Web cho phép các nhà phân tích kinh doanh và các chuyên gia về chủ đề tạo, quản lý và chia sẻ một hệ thống phân loại và từ vựng doanh nghiệp chung. Bảng thuật ngữ kinh doanh InfoSphere của IBM cho phép người dùng liên kết các thuật ngữ kinh doanh với các thuật ngữ khác do kho lưu trữ siêu dữ liệu quản lý. (Oreilly)

Tất cả các thành phần này được tích hợp chặt chẽ để giúp tổ chức quản lý và tối ưu hóa việc tích hợp, quản lý và chuyển đổi dữ liệu của họ. IBM Information Server có thể được triển khai trên các nền tảng khác nhau và tích hợp với các công cụ và hệ thống khác nhau để giải quyết các nhu cầu tích hợp dữ liệu đa dạng của các tổ chức. Trong dự án xây dựng hồ dữ liệu cho tập đoàn VNPT, thành phần IBM InfoSphere DataStage được ứng dụng cho khối Data Ingestion.

1. IBM Elastic Storage System

IBM Elastic Storage System (ESS) là một giải pháp lưu trữ dữ liệu đám mây tốc độ cao được phát triển bởi IBM. Nó cung cấp một hệ thống lưu trữ hiệu suất cao và scalable cho các ứng dụng có yêu cầu lưu trữ lớn và nhanh như trí tuệ nhân tạo (AI), HPC (High Performance Computing), phân tích dữ liệu lớn (Big Data), các ứng dụng công nghiệp và các ứng dụng web-scale. IBM ESS sử dụng công nghệ Flash và phân phối dữ liệu theo cách phân phối song song, giúp tăng cường khả năng chịu tải, tăng hiệu suất và đáp ứng nhanh với các yêu cầu tải dữ liệu lớn và tốc độ cao. Giải pháp này có thể được triển khai trên các nền tảng khác nhau và được hỗ trợ bởi một loạt các công cụ quản lý và giám sát, giúp đơn giản hóa quản lý hệ thống lưu trữ và cải thiện độ tin cậy. Ngoài ra, IBM ESS còn có tính linh hoạt và có thể tích hợp với các giải pháp lưu trữ khác như IBM Spectrum Scale, IBM Cloud Object Storage và các giải pháp lưu trữ đám mây khác.

1. IBM Cloud Object Storage

IBM Cloud Object Storage là một giải pháp lưu trữ dữ liệu đám mây được phát triển bởi IBM. Nó được thiết kế để lưu trữ và quản lý dữ liệu lớn, có tính linh hoạt cao và hiệu suất cao cho các ứng dụng đòi hỏi lưu trữ dữ liệu lớn. IBM Cloud Object Storage cung cấp một giao diện lập trình ứng dụng (API) đơn giản và dễ sử dụng cho phép các lập trình viên truy cập và lưu trữ dữ liệu trong đám mây. Các dữ liệu được lưu trữ trên các đối tượng, được phân bổ trên nhiều nơi lưu trữ để đảm bảo độ tin cậy và khả năng khôi phục cao. Ngoài ra, IBM Cloud Object Storage còn có các tính năng quản lý dữ liệu như mã hóa dữ liệu, chính sách quản lý dữ liệu, kiểm soát truy cập dữ liệu và quản lý cơ sở dữ liệu để giúp người dùng quản lý và bảo vệ dữ liệu của mình. Giải pháp này cung cấp tính khả dụng cao và tính linh hoạt cao bằng cách tích hợp với các giải pháp đám mây khác như IBM Cloud, Amazon Web Services (AWS), Microsoft Azure và Google Cloud Platform.

1. CP4D – Walston Knowledge Catalog

Trong Cloud Pak for Data, Walston Knowledge Catalog là một công cụ quản lý dữ liệu và metadata giúp tổ chức lưu trữ, tìm kiếm và sử dụng các tài nguyên dữ liệu khác nhau một cách hiệu quả. Cung cấp một nền tảng quản lý danh mục doanh nghiệp an toàn được hỗ trợ bởi khung quản trị dữ liệu. Cung cấp khả năng kiểm soát chi tiết về việc người dùng nào có thể thực hiện các tác vụ nào nhờ sự kết hợp giữa vai trò và quyền của người dùng cũng như kiểm soát những hành động người dùng có thể thực hiện.

Với Walston Knowledge Catalog, người dùng có thể tạo ra các bộ sưu tập dữ liệu và metadata, mô tả các thuộc tính của chúng và tìm kiếm các tài nguyên này theo nhiều cách khác nhau. Công cụ này cũng cung cấp các tính năng để xác định các quyền truy cập và kiểm soát quyền truy cập cho các tài nguyên dữ liệu. Bên cạnh đó, Walston Knowledge Catalog được tích hợp với các công cụ khác trong Cloud Pak for Data để cung cấp cho người dùng một trải nghiệm làm việc liền mạch với dữ liệu và metadata từ nhiều nguồn khác nhau. Công cụ này giúp tăng cường khả năng tương tác của người dùng với dữ liệu và metadata, đồng thời tăng tính minh bạch và hiệu quả của quá trình quản lý dữ liệu. (Simon Cambridge, Lakshmana Ekambaram, Stephen D. Gawtry, Vasfi Gucer, Audrey Holloman, Frank Ketelaars, Darren King, Karen Medhat, Mark Moloney, Payal Patel, Neil Patterson, Deepak Rangarao, Mark Simmonds, Malcolm Singh, Tamara Tatian, Henry L.Quach, 2022)

1. CP4D – Analytics Engine

Analytics Engine là một công cụ tích hợp để xử lý và phân tích dữ liệu lớn. Nó cung cấp một môi trường phân tích dữ liệu đa nền tảng, hỗ trợ nhiều ngôn ngữ lập trình như Python, R, Scala và SQL, giúp người dùng có thể triển khai các tác vụ phân tích và tính toán trên dữ liệu lớn một cách nhanh chóng và hiệu quả.

Analytics Engine cung cấp các tính năng như:

Hỗ trợ các công cụ phân tích dữ liệu như Apache Spark, Jupyter, Zeppelin, RStudio, v.v. Tự động điều chỉnh và tối ưu hóa cấu hình cho các nhiệm vụ phân tích dữ liệu. Hỗ trợ môi trường đám mây cho phép các công việc phân tích dữ liệu có thể được thực hiện trên các nguồn dữ liệu đám mây như Amazon S3, IBM Cloud Object Storage và Azure Blob Storage. Cho phép người dùng quản lý các tài nguyên tính toán một cách linh hoạt để đáp ứng nhu cầu phân tích dữ liệu của họ. (Simon Cambridge, Lakshmana Ekambaram, Stephen D. Gawtry, Vasfi Gucer, Audrey Holloman, Frank Ketelaars, Darren King, Karen Medhat, Mark Moloney, Payal Patel, Neil Patterson, Deepak Rangarao, Mark Simmonds, Malcolm Singh, Tamara Tatian, Henry L.Quach, 2022)

1. CP4D – Virtualization

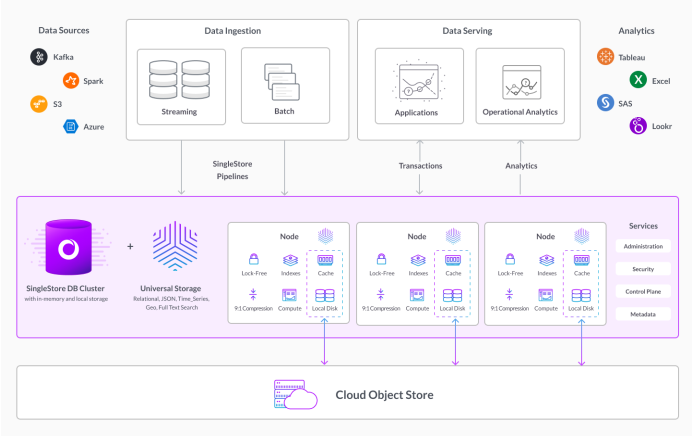
Virtualization là một công cụ dùng để truy cập và tích hợp các nguồn dữ liệu khác nhau từ các hệ thống và cơ sở dữ liệu khác nhau một cách linh hoạt và hiệu quả. Cung cấp khả năng truy cập các nguồn dữ liệu với vai trò là một máy chủ ảo và cung cấp cho người dùng một lớp trừu tượng hoá để truy cập các nguồn dữ liệu này bằng các giao thức và định dạng khác nhau.

Virtualization cung cấp các tính năng như: Hỗ trợ truy cập đa nguồn dữ liệu từ các cơ sở dữ liệu khác nhau như Oracle, SQL Server, MySQL, PostgreSQL, v.v. Cung cấp các cách tiếp cận đa dạng cho phép người dùng truy cập dữ liệu từ các giao thức khác nhau như ODBC, JDBC, FTP, v.v. Hỗ trợ tính năng truy vấn động, cho phép người dùng tạo các truy vấn dữ liệu phức tạp trên các nguồn dữ liệu khác nhau và kết hợp chúng với nhau. Tích hợp sẵn với các công cụ phân tích dữ liệu như Apache Spark, Jupyter, Zeppelin, v.v. để cho phép người dùng truy cập và phân tích dữ liệu một cách hiệu quả. (Simon Cambridge, Lakshmana Ekambaram, Stephen D. Gawtry, Vasfi Gucer, Audrey Holloman, Frank Ketelaars, Darren King, Karen Medhat, Mark Moloney, Payal Patel, Neil Patterson, Deepak Rangarao, Mark Simmonds, Malcolm Singh, Tamara Tatian, Henry L.Quach, 2022)

1. SingleStore

SingleStore Database là một cơ sở dữ liệu cloud native được xây dựng với tốc độ và quy mô để cung cấp cho các ứng dụng sử dụng lượng lớn dữ liệu. SingleStore kết hợp Hybrid Transactional/Analytical Processing [[5]](#footnote-5)(HTAP), hiệu suất nhập và truy vấn, khả năng mở rộng và phục hồi, tính sẵn sàng cao.

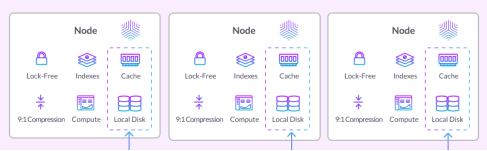
SingleStoreDB là một cơ sở dữ liệu dưới dạng dịch vụ được phân vùng theo chiều ngang, sử dụng bộ nhớ dùng chung dưới dạng blob store [[6]](#footnote-6). SingleStoreDB được tạo từ các Node. Mỗi Node chịu trách nhiệm lưu trữ dữ liệu trong các phân vùng. Mỗi phân vừng là một phân vùng chính có thể đọc và ghi hoặc một bản sao chỉ có thể đọc và phục vụ cho High Availability (HA khả năng sẵn sàng cao).



Hình 15: Kiến trúc quản lý các dịch vụ với SingleStore

*Nguồn: SingleStore (2022)*

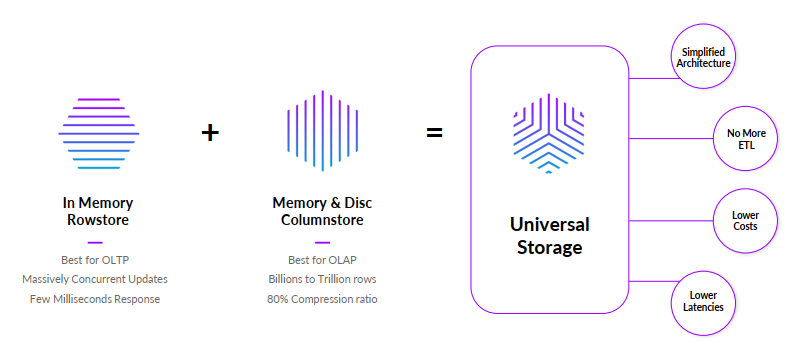
Các Node chịu trách nhiệm lưu trữ dữ liệu trong các phân vùng, di chuyển dữ liệu đến và từ kho lưu trữ đối tượng khi cần, thực hiện truy vấn với các phân vùng cục bộ và trả lại dữ liệu đã thu thập được. Số lượng node có thể được cấu hình bởi kiến trúc sư và quản trị cơ sở dữ liệu, vì chúng rất quan trọng để mở rộng quy mô theo nhu cầu.



Hình 16: Node trong SingleStore

*Nguồn: SingleStore (2022)*

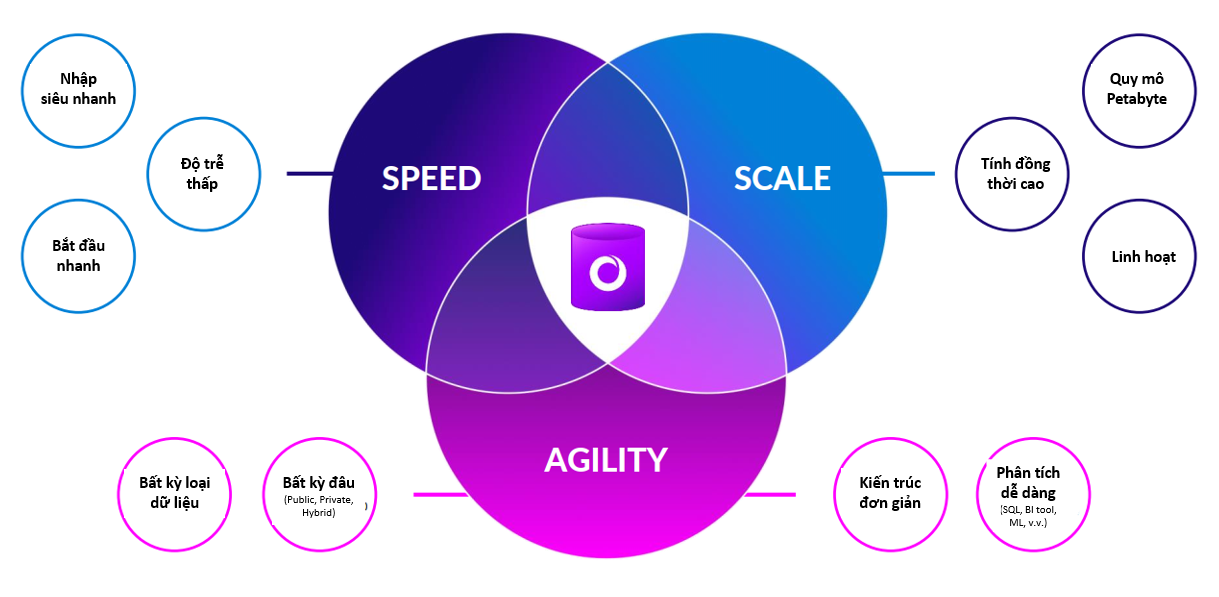
Là một công cụ lưu trữ được xây dựng cho HTAP, các bảng lưu trữ của SingleStoreDB cần hoạt động tốt trong khối lượng công việc lớn. SingleStoreDB ban đầu được thiết kế với hai định dạng lưu trữ bảng riêng biệt: Rowstore dựa trên bộ nhớ và Columnstore dựa trên đĩa. Lưu trữ bảng SingleStoreDB đã phát triển thành một thiết kế thống nhất, là Universal Storage với mục tiêu là cung cấp một loại bảng thống nhất hoạt động tốt cho cả OLTP và OLAP bằng cách kết hợp cả Rowstore và Columnstore. Điều này giúp loại bỏ gánh nặng cho người dùng trong việc chọn bố cục dữ liệu phù hợp với khối lượng công việc cụ thể của họ. Việc có một loại bảng thống nhất cho phép công việc HTAP hoạt động hiệu quả với khối lượng lớn mà không gặp phải sự phức tạp trong việc quản lý (Coates, 2022).



Hình 17: Cấu tạo Universal Storage

*Nguồn: SingleStore (2022)*

Từ kiến trúc của SingleStore, có thể thấy được những đặc điểm nổi bật như sau:



Hình 18: Đặc điểm của SingleStore

*Nguồn: SingleStore (2022)*

Thứ nhất về tốc độ, có khả năng nhập dữ liệu và bắt đầu siêu nhanh (200MB/s) với độ trễ thấp. Bởi SingleStore có khả năng xử lý song song, tối ưu hóa dữ liệu trong bộ nhớ và khả năng mở rộng các nút giúp người dùng có thể chia sẻ tải nhập dữ liệu trên nhiều nút để tăng tốc độ.

Thứ hai về quy mô, có tính đồng thời cao, linh hoạt và quy mô lên tới hàng petabyte. SingleStore hỗ trợ lưu trữ và xử lý song song trên nhiều nút, cho phép người dùng tận dụng tối đa nguồn lực của hệ thống. Các dữ liệu được phân tán trên nhiều nút khác nhau trong mạng lưới và cho phép mở rộng hệ thống bằng cách thêm nút. Không chỉ vậy, hệ thống còn được thiết kế để có thể tự động phân phối và phân bổ dữ liệu cho các nút mới được thêm vào mạng lưới, giúp tăng khả năng mở rộng của hệ thống. Và một nhờ tách biệt giữa lưu trữ và tính toán giúp SingleStore dễ dàng thay đổi quy mô điện toán để đáp ứng nhu cầu của bất kỳ khối lượng công việc nào - cả giao dịch và phân tích - không phụ thuộc vào bộ nhớ.

Thứ ba về khả năng thích nghi, hệ CSDL này có kiến trúc đơn giản, giúp phân tích dễ dàng (với SQL, BI tools hay học máy, v.v.), triển khai trên bất cứ đâu và với bất kỳ loại dữ liệu. SingleStore có thể triển khai tại On-premise, Public Cloud, Private Cloud hay Hybrid Cloud. Cùng với đó là khả năng tương tác với các hệ sinh thái mạnh mẽ.

Kết luận chương 2

Trong chương 2 đã trình bày tổng quan lý thuyết về Data Lake và một số công cụ được sử dụng cho dự án triển khai data lake cho Tập đoàn VNPT. Lý thuyết được trình bày bao gồm khái niệm, kiến trúc, đặc điểm của hồ dữ liệu; sự khác biệt giữa Data Lake và Data Warehouse cũng như lợi ích và thách thức khi triển khai. Cùng với đó là đưa ra kiến trúc tham chiếu và phân cấp cho AI Data Lake dựa trên tài liệu của IBM cung cấp, đây cũng là kiến trúc được áp dụng cho việc triển khai dự án tại VNPT. Tại chương 2 cũng nêu ra quy trình lập kế hoạch và xây dựng hồ dữ liệu, cơ sở cho việc triển khai thực tế sau đó tại chương 3. Cuối cùng là thông tin tổng quan về một số công cụ được sử dụng cho dự án.

# **CHƯƠNG 3: LẬP KẾ HOẠCH VÀ XÂY DỰNG DATA LAKE**

# **CHO TẬP ĐOÀN VNPT**

1. Lập kế hoạch xây dựng Data Lake
2. Xác định thách thức

* Hiện tại, VNPT đang có hệ thống OLTP tuy nhiên chưa đáp ứng được OLAP và dữ liệu được lưu trữ trong thời gian khá ngắn nên cần xây dựng một hệ thống hồ dữ liệu trong đó dữ liệu được sắp xếp, lưu trữ với lịch sử lâu dài và có thể truy vấn, đáp ứng được cả OLTP và OLAP.
* Dữ liệu hiện tại rất lớn nhưng chưa khai thác hết được các giá trị. Vì vậy, cần đáp ứng nhu cầu truy cập và sử dụng các năng lực phân tích dữ liệu toàn hàng cho nhiều đối tượng người dùng khác nhau của VNPT, bao gồm cả các đơn vị kỹ thuật, các nhà khoa học dữ liệu, chuyên gia/kỹ sư dữ phân tích dữ liệu cho tới các đơn vị nghiệp vụ khác. Dữ liệu thu thập được cần có giá trị và đưa ra được insight[[7]](#footnote-7).
* Xây dựng được khung quản trị dữ liệu hiệu quả (phân quyền với từng lớp ứng dụng, CSDL, quản lý truy cập, bảo vệ dữ liệu nhạy cảm, v.v.).
* Khả năng mở rộng của hệ thống, trong tương lai phục cho nhu cầu mở rộng khả năng lưu trữ, tính toán, năng lực phân tích của hệ thống Data Lake.
* Chiến lược xây dựng Data Lake cần được xác định rõ ràng để giúp VNPT có được bức tranh toàn cảnh về kiến trúc của Data Lake đang hướng đến và lộ trình đầu tư phù hợp với các hiệu quả tương ứng để qua đó từng bước triển khai và ứng dụng hồ dữ liệu hiệu quả nhất.

1. Phát triển và triển khai giải pháp chiến lược

Với những yêu cầu và thách thức hiện tại, để đưa ra một phương án giải quyết hoàn thiện và đầy đủ nhất cần kết hợp nhiều các giải pháp khác nhau. Với dự án hồ dữ liệu cho VNPT, công ty lựa chọn các sản phẩm của hãng IBM để tích hợp và cho ra giải pháp hồ dữ liệu như sau:

Trước hết, *ứng dụng kiến trúc tham chiếu và kiến trúc phân lớp của IBM cho AI Data Lake làm kiến trúc tổng quan cho hồ dữ liệu*. Kiến trúc tổng quan sẽ là bức tranh toàn cảnh về hồ dữ liệu hoàn chỉnh trong tương lai cho tập đoàn.

Về *tích hợp các nguồn dữ liệu*, các nguồn dữ liệu có thể khai thác ở thời điểm hiện tại và trong tương lai bao gồm: Telco Data (VNPT-NET/Vinaphone/IT) là những dữ liệu về dịch vụ viễn thông, Data from VNPT-Media dữ liệu của dịch vụ truyền hình và truyền thông trên nền tảng kỹ thuật số, dữ liệu từ các chi nhánh VNPT trên các tỉnh thành, Data from DPI – dữ liệu phản ánh toàn bộ luồng khách hàng sử dụng 3G/4G truy cập web và app, Data from IT applications vnEdu/eGov/… những dữ liệu từ các ứng dụng mà các công ty con của tập đoàn cung cấp như giáo dục, chính phủ và các nguồn dữ liệu mới khác, có thể có trong tương lai. Tất cả các nguồn dữ liệu trên cần được tích hợp vào cùng một kho lưu trữ cần đảm bảo các yếu tố: dữ liệu đưa vào phải có tính nhất quán và có giá trị, đảm bảo tính toàn vẹn và liên tục, tốc độ tích hợp và thời gian phản hồi nhanh (trong vài phút). Vì vậy lựa chọn nền tảng mã nguồn mở Apache Kafka để xây dựng hệ thống xử lý dữ liệu thời gian thực và IBM InfoSphere DataStage để trích xuất dữ liệu từ nhiều nguồn như database, file, web service, message queue, v.v.

*Phương án lưu trữ dữ liệu*, lượng dữ liệu được nhập vào rất lớn và theo tần suất sử dụng cũng như giá trị sử dụng, dữ liệu có hai loại Hot và Cold data. IBM ESS được thiết kế để lưu trữ và xử lý dữ liệu tại chỗ (on-premises). Nó cung cấp một hệ thống tập trung, tốc độ cao, có khả năng mở rộng và đáp ứng yêu cầu của các ứng dụng yêu cầu hiệu suất cao, thường xuyên truy xuất dữ liệu. IBM Cloud Object Storage được thiết kế tối ưu cho việc lưu trữ dữ liệu trên đám mây (cloud). Cung cấp một kho lưu trữ dựa trên đám mây, khả năng mở rộng linh hoạt và giá thành thấp hơn so với lưu trữ on-premises. Đây là lựa chọn tốt cho việc lưu trữ các loại dữ liệu như dữ liệu lịch sử, dữ liệu backup và dữ liệu lưu trữ dài hạn. Vì vậy, với Hot data sử dụng IBM ESS using Spectrum Scale CSI để lưu trữ và Cold data sử dụng IBM Cloud Object Storage (ICOS).

*Ứng dụng phân tích dữ liệu*, dữ liệu trong hồ dữ liệu là dữ liệu lớn nên cần có các công cụ xử lý và phân tích chuyên biệt:

* Để thực hiện các tác vụ phân tích và tính toán trên dữ liệu lớn một cách nhanh chóng và hiệu quả, sử dụng công cụ Analytics Engine. Analytics Engine cung cấp cho môi trường Apache Spark một dịch vụ tách riêng các tầng điện toán và lưu trữ để kiểm soát chi phí và đạt được các phân tích trên quy mô lớn.
* Về các công cụ SQL, sử dụng Big Data RDBMS – SingleStore và Data Virtualization – CP4D Data Virtualization. SingleStore hỗ trợ cả OLTP và OLAP cùng với khả năng nén dữ liệu trung bình đến 80% mà không ảnh hưởng đến hiệu suất xử lý dữ liệu, rất phù hợp với yêu cầu của VNPT. CP4D Data Virtualization cho phép người dùng truy cập các nguồn dữ liệu đóng vai trò là một máy chủ ảo và cung cấp cho người dùng một lớp trừu tượng hoá để truy cập các nguồn dữ liệu này bằng các giao thức và định dạng khác nhau.
* Nền tảng phân tích: CP4D – Walston Studio/AutoAI cung cấp người dùng một môi trường tích hợp để xây dựng, đào tạo và triển khai các mô hình học máy một cách dễ dàng và hiệu quả.

*Khung quản trị dữ liệu:* các chính sách, quy tắc, quy trình, cơ cấu tổ chức và công nghệ được đưa ra như một phần của chương trình quản trị dữ liệu. Nó cũng giải thích rõ ràng những thứ như tuyên bố sứ mệnh, mục tiêu của chương trình và cách đo lường thành công, cũng như trách nhiệm ra quyết định và trách nhiệm giải trình đối với các chức năng khác nhau. Các thành phần cần xác định trong giải pháp:

Bảng 2: Các thành phần cần xác định khi triển khai quản trị dữ liệu

|  |  |
| --- | --- |
| Thành phần | Mô tả |
| Data mapping and classification (ánh xạ và phân loại dữ liệu) | Ánh xạ dữ liệu trong các hệ thống giúp lập tài liệu về nội dung dữ liệu và dòng chảy của dữ liệu trong tổ chức. Sau đó, các bộ dữ liệu khác nhau có thể được phân loại dựa trên các yếu tố như liệu chúng có chứa thông tin cá nhân hoặc dữ liệu nhạy cảm khác hay không. Các phân loại ảnh hưởng đến cách các chính sách quản trị dữ liệu được áp dụng cho các tập dữ liệu riêng lẻ. |
| Business glossary (thuật ngữ kinh doanh) | Bảng thuật ngữ kinh doanh chứa các định nghĩa về thuật ngữ và khái niệm kinh doanh được sử dụng trong một tổ chức. Bằng cách giúp thiết lập một từ vựng chung cho dữ liệu của doanh nghiệp, bảng thuật ngữ kinh doanh có thể hỗ trợ quản trị. |
| Data catalog (danh mục dữ liệu) | Danh mục dữ liệu thu thập siêu dữ liệu từ các hệ thống và sử dụng nó để tạo kho lưu trữ được lập chỉ mục cho các nội dung dữ liệu có sẵn bao gồm thông tin về dòng dữ liệu, chức năng tìm kiếm và công cụ cộng tác. Thông tin về chính sách quản trị dữ liệu và cơ chế tự động thực thi chúng cũng có thể được tích hợp vào danh mục. |

Giải pháp Data Lake cho VNPT sẽ cung cấp phần mềm quản trị dữ liệu CP4D - Watson Knowledge Catalog (WKC), được sử dụng để tự động hóa các khía cạnh của việc quản lý một chương trình quản trị. WKC cung cấp các tính năng tìm kiếm dữ liệu, phân loại dữ liệu, điều khiển quyền truy cập, giám sát và quản lý dữ liệu.

*Thử nghiệm và đánh giá*: xây dựng bộ tài liệu đo kiểm để mô tả giải pháp Data Lake và đánh giá kết quả đạt được. Bộ tài liệu bao gồm các bước thao tác trên các thành phần của giải pháp, các yêu cầu của VNPT, kết quả đạt được, phần đánh giá của VNPT. Cùng với đó là bộ test case cung cấp thông tin kiểm tra hệ thống theo đúng mong muốn của VNPT.

*Quản lý và bảo trì*: phân quyền truy cập dữ liệu trên từng lớp ứng dụng, sao lưu dữ liệu thường xuyên và đưa ra phương án phục hồi khi gặp sự cố, giám sát và theo dõi hoạt động của Data Lake dựa trên những Audit Log gửi về hệ thống, triển khai các buổi đào tạo cho người dùng tại VNPT, nâng cấp và cập nhật hệ thống định kỳ để đảm bảo tính ổn định và tính bảo mật.

1. Dự đoán sự tăng trưởng của dữ liệu

Hiện dữ liệu tại VNPT đang chủ yếu phục vụ cho việc xây dựng các báo cáo, dashboard để hỗ trợ ra quyết định. Nhưng nhiều nguồn dữ liệu chưa được khai thác triệt để hay thậm chí chưa được sử dụng. Ví dụ như những dữ liệu phi cấu trúc: âm thanh, hình ảnh từ những tương tác của người dùng trên các sản phẩm dịch vụ mà tập đoàn cung cấp. Trong tương lai sẽ cần sử dụng các dữ liệu trên cho các model ML, DL, AI và các hoạt động khác nhằm cải thiện việc kinh doanh.

Tổng dữ liệu điện tử hiện nay của VNPT vào khoảng 3.5PB được chia thành 2 nhóm: Hot data với khoảng 500TB và Cold data là 3PB. Riêng với dữ liệu từ hệ thống DPI, VNPT cho biết đã có 2.7TB dữ liệu được truyền vào mỗi ngày. Tính cả những nguồn dữ liệu có cấu trúc và không có cấu trúc khác ước tính trong 5 năm lượng dữ liệu sẽ tăng trưởng khoảng 10% mỗi năm.

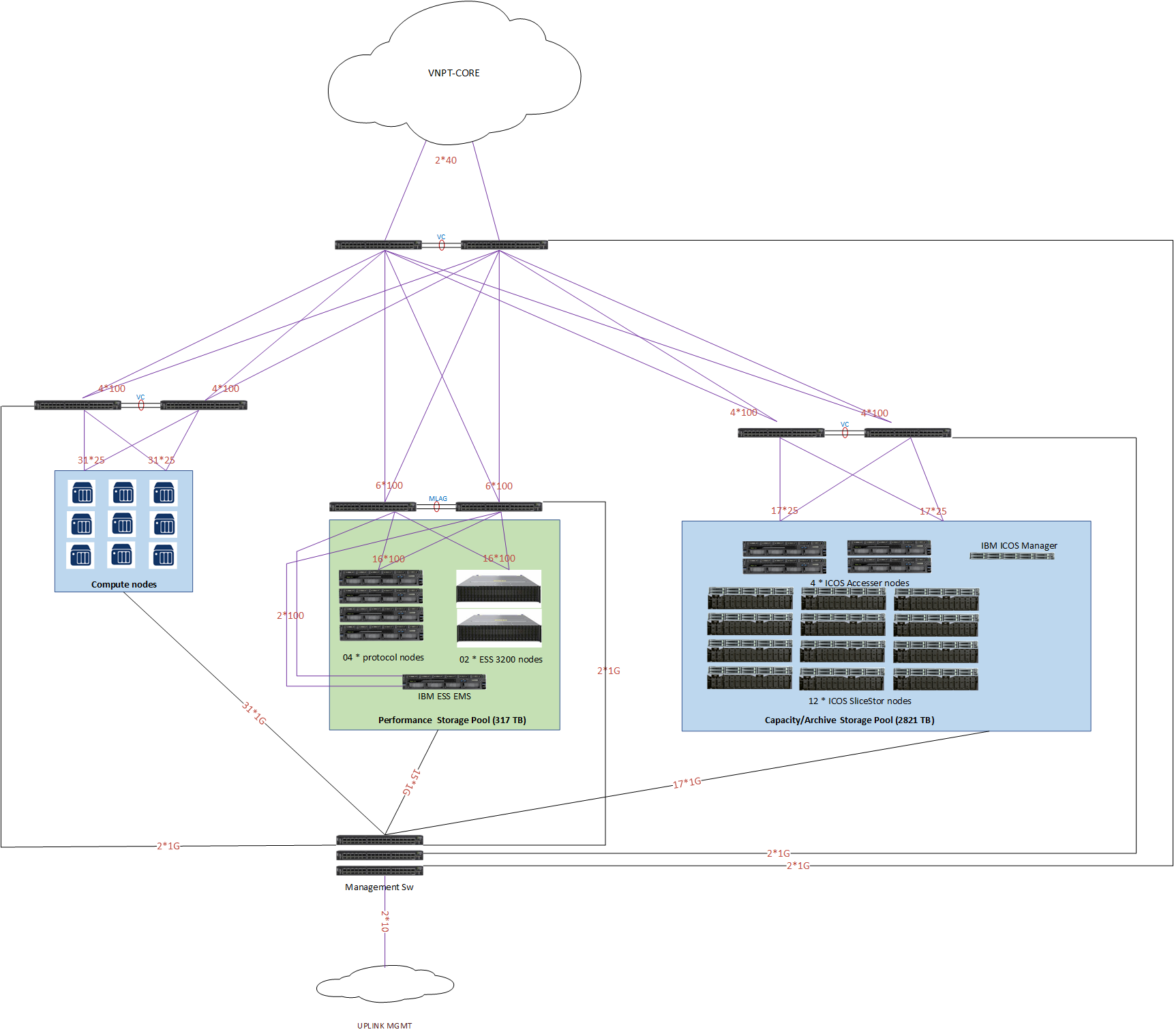
1. Lên kế hoạch về CSHT
2. Thiết kế vật lý

Phần vật lý sẽ bao gồm các thiết bị lưu trữ và máy chủ như sau:

Bảng 3: Chi tiết số lượng thiết bị lắp đặt theo Rack[[8]](#footnote-8)

|  |  |  |  |
| --- | --- | --- | --- |
| **Rack** | **Name** | **Quantity** | **Unit** |
| **ESS** | IBM Switch 8831-00M | 2 | 1 |
| IBM Protocols Node - 5105-22E | 5 | 2 |
| IBM ESS3200 - 5141-FN1 | 2 | 2 |
| **Total** | **9** | **16** |
| **ICOS-01** | IBM COS Manager Node - 4957-M10 | 1 | 1 |
| IBM COS Accessor Node - 4957-A10 | 2 | 1 |
| IBM COS Controller Node - 4957-C10 | 6 | 1 |
| IBM COS JBOD J11 - 4957-J11 | 6 | 4 |
| **Total** | **15** | **33** |
| **ICOS-02** | IBM COS Accessor Node - 4957-A10 | 2 | 1 |
| IBM COS Controller Node - 4957-C10 | 6 | 1 |
| IBM COS JBOD J11 - 4957-J11 | 6 | 4 |
| Juniper QFX5120-48Y | 2 | 1 |
| **Total** | **16** | **34** |
| **Dell-01** | Juniper QFX5120-32C | 2 | 1 |
| Dell R6525 Management Node | 1 | 1 |
| Dell R6525 Compute Node | 7 | 1 |
| **Total** | **10** | **10** |
| **Dell-02** | Juniper QFX5120-48Y | 2 | 1 |
| Dell R6525 Management Node | 1 | 1 |
| Dell R6525 Compute Node | 7 | 1 |
| **Total** | **9** | **9** |
| **Dell-03** | Juniper EX3400-48T | 2 | 1 |
| Dell R6525 Management Node | 1 | 1 |
| Dell R6525 Compute Node | 7 | 1 |
| **Total** | **11** | **11** |
| **Dell-04** | Juniper QFX5120-48Y | 2 | 1 |
| Dell R6525 Compute Node | 7 | 1 |
| **Total** | **9** | **9** |

1. Thiết kế mạng, hệ thống máy chủ

****

Hình 19: Sơ đồ mạng của hệ thống Data Lake

Hệ thống mạng cho dự án Data Lake được thiết kế đảm bảo tính dự phòng cũng như traffic cho toàn bộ hệ thống. Về mặt traffic được chia làm 2 phần:

*Traffic dữ liệu của các server*: Các server sẽ được kết nối 2 chân lên các cụm switch tương ứng để phòng trường hợp xảy ra sự cố khi một thiết bị switch lỗi.

* Traffic trong một cụng server: Sẽ được chuyển mạnh qua các cụm switch của cụm server đó nếu chung vlan, subnet mạng.
* Traffic giữa các cụm server: Sẽ được chuyển mạch qua cụm spine nếu chung vlakees n, subnet mạng.
* Traffic giữa các cụm server với bên ngoài hoặc khác vlan, subnet mạng: sẽ được chuyển mạnh qua hệ thống mạng core của VNPT.

*Traffic quản trị thiết bị*: Toàn bộ thiết bị trong hệ thống sẽ được quản trị OOB[[9]](#footnote-9), các cổng manager trên thiết bị sẽ được kết nối đến switch quản trị. Lớp mạng quản trị sẽ được phía VNPT cung cấp. Traffic quản trị sẽ không liên quan gì đến traffic dữ liệu của hệ thống.

Với hệ thống máy chủ, IBM® và Red Hat đã hợp tác để cung cấp một giải pháp chung sử dụng IBM Cloud Private và OpenShift. Red Hat OpenShift Container Platform (RHOCP) là một tập hợp các thành phần mô-đun và các dịch vụ được xây dựng trên cơ sở hạ tầng container Kubernetes. RHOCP bổ sung các khả năng cung cấp nền tảng PaaS cho môi trường Production như remote management, multitenancy, tăng cường bảo mật, giám sát và kiểm tra, quản lý vòng đời ứng dụng và giao diện tự phục vụ cho các nhà phát triển, tất cả đều sử dụng Red Hat Enterprise Linux CoreOS làm hệ điều hành cơ bản. Mô hình giải pháp như sau:

1. Mô hình OpenShift cho môi trường Bare metal [[10]](#footnote-10)bao gồm 2 thành phần:

- Thành phần Red Hat OpenShift Container Platform: 3 node management và 27 node worker.

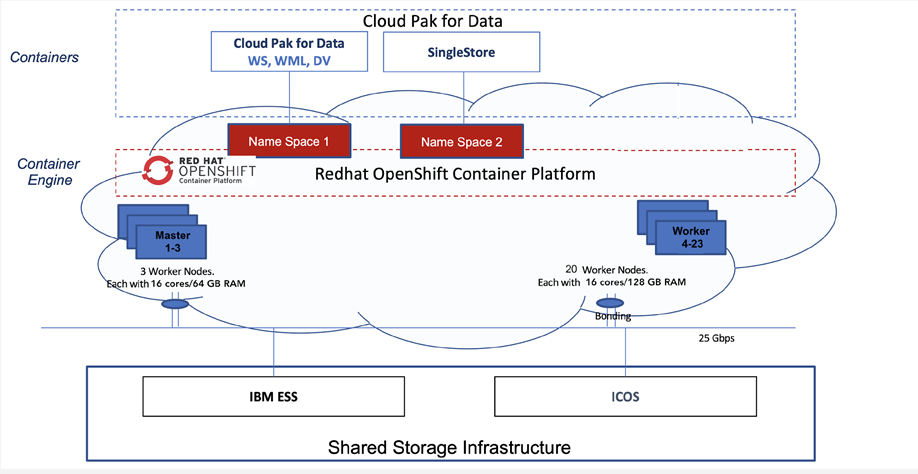
- Thành phần môi trường: 1 node Bastion chạy các dịch vụ môi trường bao gồm: NAT gateway, DNS, DHCP, Web/File Server, Ldap, Proxy và Load Balancer.

1. Mô hình Openshift cho triển khai cụm Cloud Pak for Data

Hệ thống được triển khai trên 1 cụm Red Hat OpenShift Container Platform cho phép chia thành 2 vùng namespace.

Namespace 1: chứa các dịch vụ IBM Private Cloud(Cloud Pak for Data) bao gồm WS, WDL, DV.

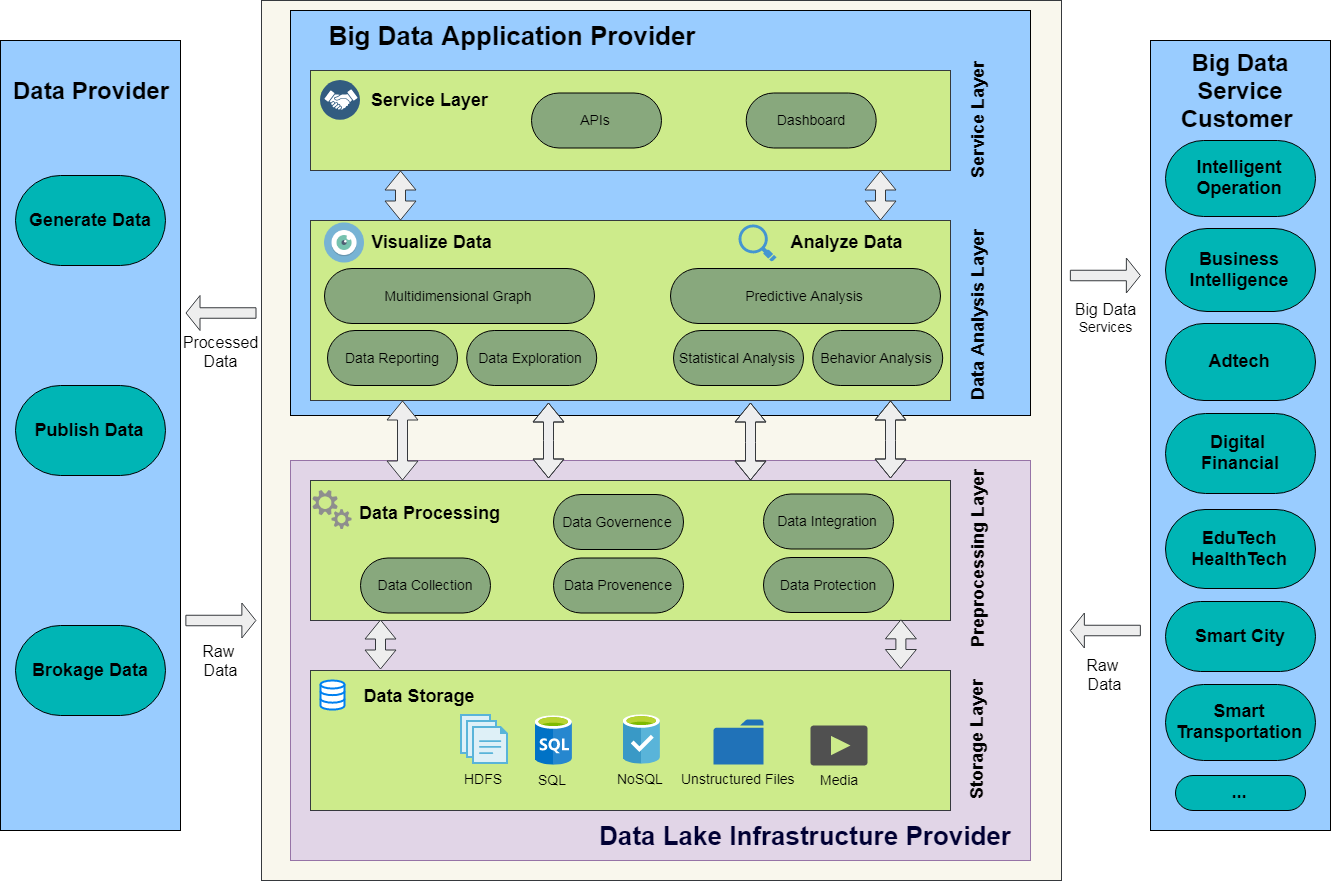
Namespace 2: chứa dịch vụ cơ sở dữ liệu Singstore.



Hình 20: Mô hình Openshift cho triển khai cụm Cloud Pak for Data

1. Thiết kế phần mềm

Dựa theo khuyến cáo của tổ chức ITU theo tiêu chuẩn ITU-T Y.3600 (11/2015), kiến trúc thành phần chức năng của nền tảng VNPT Big Data được tổ chức theo mô hình dưới đây:



Hình 21: Kiến trúc thành phần chức năng của nền tảng Big Data

*Nguồn: ITU (2015)*

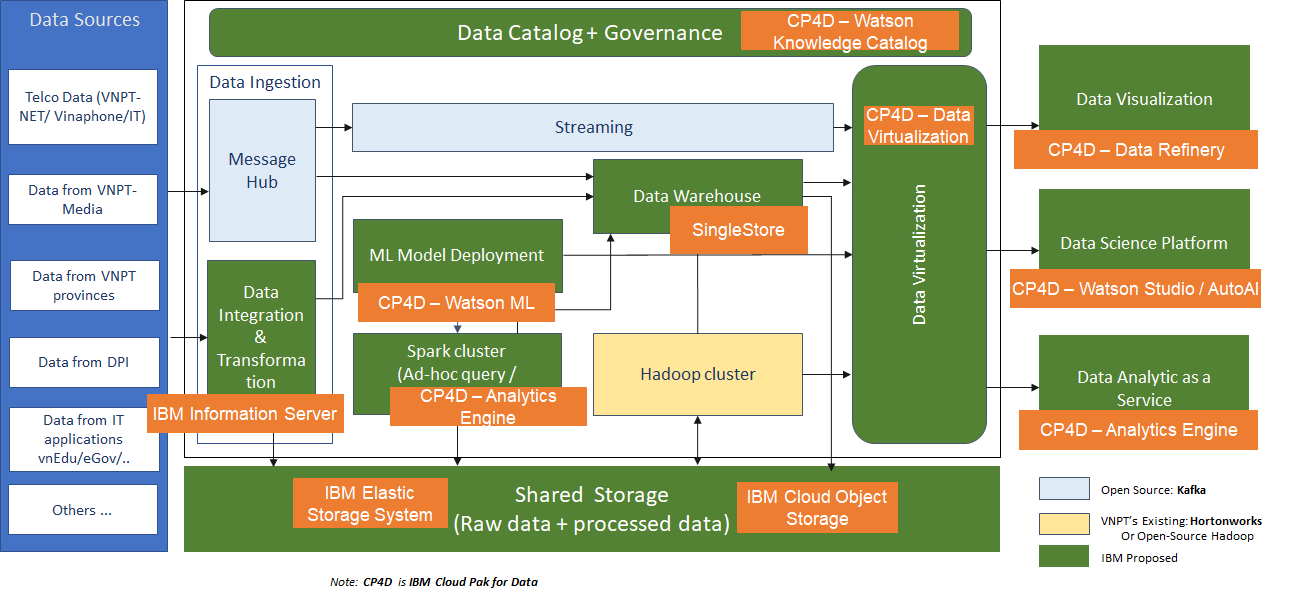
Đơn vị cung cấp dữ liệu (Data Provider): Thực hiện chức năng sản sinh dữ liệu để cung cấp cho Hồ Dữ liệu Big Data một cách trực tiếp hoặc thông qua đơn vị môi giới dữ liệu.

Đơn vị cung cấp dịch vụ dữ liệu lớn (Big Data Application Provider): Có vai trò cung cấp nền tảng cho việc lưu trữ, xử lý cũng như phân tích, cung cấp dịch vụ trên cơ sở dữ liệu lớn.

Đơn vị sử dụng dịch vụ dữ liệu lớn (Big Data Service Customer): Là đơn vị sử dụng kết quả sản phẩm của đơn vị cung cấp dịch vụ dữ liệu lớn (Big Data Service Provider), thông qua các giao diện kết nối

Trong đó, nền tảng Hồ dữ liệu lớn Data Lake đóng vai trò của đơn vị cung cấp hạ tầng dữ liệu lớn và bao gồm hai lớp chức năng: Lớp lưu trữ dữ liệu (Storage Layer) và lớp tiền xử lý dữ liệu (Preprocessing Layer).

Với kiến trúc trên, tương ứng với các thành phần là các giải pháp của IBM như sau:

****

Hình 22: Kiến trúc hệ thống hồ dữ liệu tương ứng với giải pháp IBM

Theo từng pha của dự án, tài nguyên cần thiết để triển khai phần mềm như sau:

**Pha 1**

Khối compute cluster: 1732 core

Khối management: 03 máy chủ

* OpenShift Master

3 master nodes: 32 cores, 512 GB RAM

* Cloud Pak for Data & SingleStore

5 compute nodes: 64 cores, 512GB GB RAM

* Analytics Engine

23 compute nodes: 64 cores, 512GB GB RAM

**Pha 2**

Khối compute cluster: 1280 core

* Cloud Pak for Data & SingleStore

4 compute nodes: 64 cores, 512GB GB RAM

* Analytics Engine

16 compute nodes: 64 cores, 512GB GB RAM

1. Xác định chiến lược hoạt động

Chiến lược hoạt động liên quan đến các quyết định hàng ngày về việc sử dụng tài nguyên, tổ chức công việc và quản lý rủi ro để ứng dụng hồ dữ liệu một cách hiệu quả. Có năm chiến lược được đưa ra như sau:

Một, tối ưu hóa hiệu suất dữ liệu. Xác định và giải quyết các vấn đề về hiệu suất dữ liệu. Điều này bao gồm việc tối ưu hóa thời gian truy xuất dữ liệu, tối ưu hóa lưu trữ và xử lý dữ liệu, cải thiện khả năng mở rộng của hệ thống và tối ưu hóa các quy trình xử lý dữ liệu. Về tối ưu hóa hiệu xuất truy xuất dữ liệu, các dữ liệu được truy xuất từ kho dữ liệu sẽ được phân vùng, chia theo column hoặc row để tối ưu truy vấn. Việc lưu trữ dữ liệu sẽ phân chia ra khu vực hot data và cold data, riêng với dữ liệu được lưu ở kho dữ liệu SingleStore sẽ được nén lại để giảm thiểu dung lượng lưu trữ.

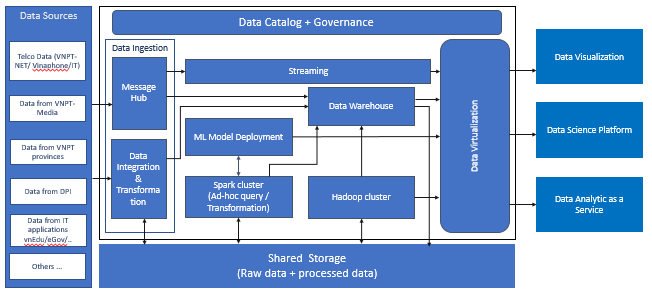
Hai, tối ưu hóa việc sử dụng tài nguyên. Hồ dữ liệu yêu cầu nhiều tài nguyên để triển khai và hoạt động, do đó cần phải tối ưu hóa việc sử dụng tài nguyên để giảm thiểu chi phí. Các công việc liên quan đến tối ưu hóa việc sử dụng tài nguyên sẽ bao gồm việc tận dụng các công nghệ mới, tối ưu hóa dữ liệu, cấu hình hệ thống, theo dõi và đánh giá hiệu suất. Việc này cần trao đổi và đưa ra chiến lược cụ thể giữa nội bộ VNPT và VNPT với đơn vị triển khai.

Ba, tăng cường bảo mật. Ứng dụng Multi-tenant để tổ chức sử dụng hệ thống hồ dữ liệu. Mỗi nhóm người dùng được cung cấp một môi trường riêng biệt và được phân tách hoàn toàn với nhóm khác. Mỗi nhóm người dùng chỉ có quyền truy cập vào môi trường của mình. Môi trường bao gồm CPU, memory và storage riêng biệt.

Bốn, xây dựng bộ danh mục dữ liệu. Khai báo các danh mục dữ liệu (bao gồm: trong phần quản trị dữ liệu của hồ dữ liệu (CP4D – Watson Knowledge Catalog) để giúp tìm kiếm, truy xuất, quản lý dữ liệu nhanh chóng và chính xác hơn.

Năm, đảm bảo tính sẵn sàng. Triển khai và kiểm tra các giải pháp backup và recovery để đảm bảo tính sẵn sàng của hệ thống.

1. Triển khai xây dựng Data Lake cho tập đoàn VNPT
2. Kiến trúc tổng quan



Hình 23: Kiến trúc tổng quan về hồ dữ liệu cho tập đoàn VNPT

*Message Hub*: Dữ liệu truyền về dưới dạng các message/event và được thu nhận về theo thời gian thực. Các dữ liệu này có thể thể được sử dụng để phân tích ngay trong quá trình truyền, nhận dữ liệu và sau đó đưa vào khai thác & truy xuất.

*Data Integration & Transformation*: Quá trình Tích hợp & Chuyển đổi dữ liệu sẽ xử lý việc thu thập, định tính, xử lý và di chuyển dữ liệu để chuẩn bị lưu trữ trong Kho lưu trữ dữ liệu, sau đó được sử dụng cho các ứng dụng và hệ thống Phân tích / Truy cập sau này. Chuyển đổi dữ liệu: Áp dụng các luật nghiệp vụ (business rules) để chuẩn hóa dữ liệu phù hợp với mô hình dữ liệu đích. Chất lượng dữ liệu: Bao gồm các quy tắc chuẩn hóa và chất lượng dữ liệu trên dữ liệu nguồn để đảm bảo rằng nó tuân thủ các tiêu chuẩn và được mọi người cần sử dụng hiểu rõ. Tải dữ liệu: Quá trình tải hoặc chèn dữ liệu vào kho lưu trữ đích (nguồn phân tích) và làm cho dữ liệu luôn sẵn sàng để sử dụng. Việc này có thể được lập kế hoạch để xử lý theo lịch định kỳ, hoặc xử lý theo thời gian thực.

*Data warehouse*: Kho dữ liệu - nơi tập trung lưu trữ dữ liệu tinh là các dữ liệu đã qua xử lý và sẵn sàng để sử dụng cho các bài toán nghiệp vụ khác nhau, chủ yếu là các bài toán khai thác dữ liệu qua các công cụ BI và tạo lập báo cáo. Mô hình dữ liệu (Data Mart) – các engine của mô hình dữ liệu được thiết lập trước phù hợp với từng loại báo cáo, hay API dữ liệu với các mục đích nghiệp vụ cụ thể.

*Spark cluster*: Các nền tảng cho phép xử lý phân tán các tập dữ liệu lớn trên các cụm máy tính. Tuỳ theo mục đích sử dụng khác nhau mà có thể sử dụng Spark hay Hadoop. Spark: có lợi thế xử lý ngay trong bộ nhớ một cách nhanh chóng, nhanh gấp nhiều lần trên ổ đĩa, và phù hợp cho streaming workload, các truy vấn tương tác, và học máy. Hadoop: cho phép mở rộng quy mô xử lý từ các hệ thống máy tính đơn lẻ thành hàng ngàn máy quy mô lớn để có thể đem đến năng lực tính toán và lưu trữ cục bộ giúp đẩy nhanh hiệu quả tính toán.

*ML Model Deployment*: Nền tảng giúp xây dựng các mô hình học máy, các thuật toán và kỹ thuật khai phá dữ liệu để phục vụ cho các bài toán phân tích nâng cao.

*Data Virtualization*: Lớp ảo hoá dữ liệu cung cấp một nền tảng trung gian để đảm bảo quản lý việc truy cập dữ liệu từ bên ngoài vào các hệ thống lõi của hệ thống một cách an toàn và hiệu quả thông qua các dịch vụ & API dữ liệu.

*Shared storage*: Hạ tầng lưu trữ chia sẻ cho phép lưu trữ cả dữ liệu thô chưa được xử lý/chưa cần khai thác lẫn cả dữ liệu đã được xử lý. Do lượng dữ liệu lưu trữ rất lớn nên cần có các kỹ thuật về phần mềm cũng như công nghệ phần cứng phù hợp để giúp truy cập dữ liệu một cách nhanh chóng và tối ưu nhất.

*Data Visualization*: Các công cụ trực quan hoá dữ liệu như Báo cáo, Dashboard hay các công cụ tìm kiếm sẽ giúp cho người dùng nghiệp vụ (business user) dễ dàng tiếp cận với dữ liệu và thông tin khai thác được.

*Data Science Platform*: Cung cấp các công cụ, các mô hình, thuật toán AI/ML để các nhà khoa học dữ liệu, kỹ sư dữ liệu tạo ra các mô hình khai thác dữ liệu một cách hiệu quả theo từng bài toán nghiệp vụ.

*Data Analytic as a service*: Cung cấp khả năng phân tích dữ liệu theo yêu cầu, dưới dạng các dịch vụ, là nơi mà người dùng có thể tự thao tác để sử dụng khai thác dữ liệu mà ít phụ thuộc vào sự trợ giúp của các bộ phận kỹ thuật.

Tương ứng với kiến trúc tổng quan là các sản phẩm cụ thể đã được liệt kê trong phần kế hoạch CSHT về phần mềm (kiến trúc hệ thống hồ dữ liệu tương ứng với giải pháp IBM) như sau:

Bảng 4: Chi tiết các sản phẩm được sử dụng trong hồ dữ liệu

|  |  |  |
| --- | --- | --- |
| **Thành phần** | **Chức năng** | **Sản phẩm** |
| Storage | Hot | IBM ESS using Spectrum Scale CSI |
| Cold | IBM Cloud Object Store (ICOS) |
| Software | Spark | IBM Analytics Engine |
| SQL Engine | Big Data RDBMS – SingleStore  Data Virtualization – CP4D Data Virtualization |
| Analytics Platform | CP4D - Watson Studio |
| ETL & Data Quality | Information Server |
| Data Governance with Linage & Catalog | CP4D - Watson Knowledge Catalog (WKC) |

1. Xác định nguồn dữ liệu và người dùng hệ thống

Nguồn dữ liệu của hệ thống Data Lake được đến từ các sản phẩm, dịch vụ mà tập đoàn VNPT cung cấp. Có thể liệt kê 5 nguồn dữ liệu chính của tập đoàn cho hồ dữ liệu như sau:

Bảng 5: Thông tin nguồn dữ liệu

|  |  |
| --- | --- |
| **Nguồn dữ liệu** | **Mô tả** |
| Telco Data (VNPT-NET/Vinaphone/IT) | Những dữ liệu về dịch vụ viễn thông , bao gồm các thông tin về điện thoại di động, băng thông internet, truyền hình và các dịch vụ khác. Ví dụ:   * Dữ liệu lưu lượng mạng: Đây là các dữ liệu về lưu lượng dữ liệu được truyền qua mạng, bao gồm dữ liệu tải và tải xuống, thời gian truy cập, vị trí của người dùng, nội dung truy cập, v.v. * Dữ liệu về điện thoại di động: Bao gồm thông tin về cuộc gọi, tin nhắn, dữ liệu di động, vị trí của người dùng, v.v. * Dữ liệu truyền hình: Bao gồm các dữ liệu về lượt xem, thời gian xem, thói quen xem và sở thích của người dùng, v.v. |
| Data from VNPT-Media | Dữ liệu của dịch vụ truyền hình và truyền thông trên nền tảng kỹ thuật số bao gồm các thông tin về nội dung phim ảnh, truyền hình, âm nhạc, sách báo và các hình thức truyền thông khác. |
| Data from VNPT provinces | Dữ liệu từ các chi nhánh VNPT trên các tỉnh thành |
| Data from DPI | Dữ liệu phản ánh toàn bộ luồng khách hàng sử dụng 3G/4G truy cập web và app |
| Data from IT applications vnEdu/eGov/… | Những dữ liệu từ các ứng dụng mà các công ty con của tập đoàn cung cấp như giáo dục, chính phủ |

Người dùng hệ thống Data Lake của VNPT bao gồm các nhân viên và đội ngũ chuyên môn trong công ty hoặc tổ chức sử dụng hệ thống này để truy cập, phân tích và sử dụng dữ liệu, có thể chia làm bốn nhóm chính sau:

*Nhóm xử lý dữ liệu*, là những người làm nhiệm vụ thu gom, tổng hợp, biến đổi, dữ liệu, xem xét các số liệu thống kê, đi sâu vào dữ liệu, tìm kiếm những thông tin chi tiết mới sẽ được sử dụng để nhấn mạnh chiến lược kinh doanh, đưa ra các quyết định chung cấp chiến lược, các quyết định chiến thuật, mục tiêu phương hướng phát triển. Đó có thể là những người làm Data Analyst, Data Engineer, Data Scientist, Data Quality Analyst, BI, …

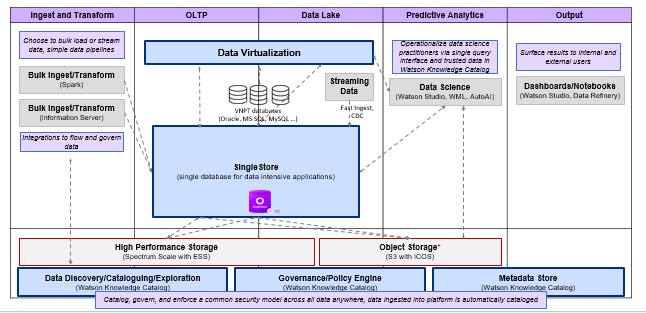
*Nhóm điều hành*, là những người điều hành doanh nghiệp, thúc đẩy sự thành công của tổ chức bằng cách cải thiện hiệu quả phân tích hoạt động và không ngừng tìm cách giảm chi phí. Hệ thống hồ dữ liệu cung cấp một cái nhìn tổng quan về tổ chức cho phép Giám đốc điều hành phát hiện các xu hướng trong toàn bộ cấu trúc của doanh nghiệp - thông tin chi tiết có thể hỗ trợ: Tăng trưởng kinh doanh, sự đổi mới, hiệu quả hoạt động để đưa ra các quyết định cấp chiến lược.

*Nhóm người dùng doanh nghiệp,* các nhà quản lý, người đang tìm cách để giúp một bộ phận hoạt động hiệu quả hơn và hiệu quả hơn. Người dùng doanh nghiệp thường chưa được đào tạo trong việc sử dụng các tiện ích của hồ dữ liệu nhưng có khả năng nắm bắt các kiến thức cơ bản về báo cáo và có thể sử dụng thành công thông tin kinh doanh để báo cáo về các hoạt động kinh doanh, đưa ra các quyết định cấp chiến thuật và chiến lược. Họ cần sử dụng thông tin từ hồ dữ liệu để đưa ra những quyết định liên quan đến việc xây dựng, điều khiển, kiểm soát mục tiêu của bộ phận mình, phương thức điều hành ra sao và liên kết phối hợp giữa các nhóm như thế nào để đạt mục tiêu chung.

*Nhóm CNTT,* là người sử dụng hồ dữ liệu để đưa ra quyết định cấp chiến thuật: đảm bảo rằng người dùng doanh nghiệp đang tận dụng tối đa từ hồ dữ liệu như phân tích, quản trị dữ liệu và bảo mật dữ liệu ra sao. Họ chịu trách nhiệm về cơ sở hạ tầng, các công cụ sẵn có trong công ty và họ quản lý các quyền và vai trò của nhân viên nên cần đưa ra các quyết định liên quan đến vấn đề này.

1. Chiến lược thu thập dữ liệu

Để đưa ra chiến lược thu thập, trước hết xác định luồng dữ liệu trong hồ dữ liệu. Dữ liệu ban đầu từ các nguồn sẽ được nhập (ingest) và biến đổi (transform) bằng các công cụ như Spark, Information Server (DataStage) hay pipeline của SingleStore vào trong khu vực lưu trữ là Spectrum Scale with ESS (với hot data), S3 with ICOS (với cold data) và vùng lưu trữ của SingleStore (với dữ liệu dùng để phân tích). Dữ liệu được lưu trữ sẽ được sử dụng cho việc triển khai các phân tích, ứng dụng ML, DL, AI, v.v. Sau đó trả về kết quả là các báo cáo, dashboard, dự báo, v.v. Ban đầu, dữ liệu từ các hệ thống nguồn được thu thập theo hai dạng là dạng streaming (truyền dữ liệu liên tục) và dạng batch (truyền dữ liệu theo lô). Đối với dạng streaming sử dụng chức năng pipeline của SingleStore để nhập dữ liệu. Còn đối với truyền dữ liệu dạng batch sử dụng cả SingleStore và IBM InfoSphere DataStage.

****

Hình 24: Tổng quan luồng dữ liệu

**Pipeline trong SingleStore** sẽ được sử dụng để nhập dữ liệu dạng streaming và batch bằng cách thiết lập tham số BATCH\_INTERVAL. Pipeline trong SingleStore hỗ trợ truyền dữ liệu từ nhiều nguồn như: Apache Kafka, Amazon S3, Azure Blob, HDFS, Filesystem Extractor, GCS (Google Cloud Storage) và định dạng file JSON, Avro, Parquet và CSV. Trước khi truyền dữ liệu vào cần phải tạo sẵn database và table để lưu dữ liệu. Với pipeline của SingleStore người dùng có thể biến đổi dữ liệu theo định dạng mong muốn trước khi truyền vào, việc nhập dữ liệu có thể diễn ra song song, nhiều pipeline có thể cùng truyền vào một table, loại bỏ các bản ghi trùng lặp tại thời điểm nhập để làm sạch dữ liệu theo thời gian thực và hơn hết là tự động nhập khi có dữ liệu mới đến để phục vụ cho kiểu truyền streaming. Những tính năng trên được cung cấp với câu lệnh CREATE PIPELINE, CREATE PIPELINE ... INTO PROCEDURE và CREATE PIPELINE ... WITH TRANSFORM.

Cấu trúc một câu lệnh tạo pipeline như sau:

*CREATE [OR REPLACE] [AGGREGATOR] PIPELINE [IF NOT EXISTS] <pipeline\_name> AS*

*LOAD DATA { kafka\_configuration | s3\_configuration | filesystem\_configuration*

*| azure\_blob\_configuration | hdfs\_configuration | gcs\_configuration | link\_configuration}*

*[BATCH\_INTERVAL <milliseconds>]*

*[(ENABLE|DISABLE) OUT\_OF\_ORDER OPTIMIZATION]*

*[WITH TRANSFORM ('uri', 'executable', 'arguments [...]') ]*

*[REPLACE | IGNORE | SKIP { ALL | CONSTRAINT | DUPLICATE KEY | PARSER } ERRORS]*

*{ INTO TABLE <table\_name> | INTO PROCEDURE <procedure\_name> }*

*{ <json\_format\_options> | <avro\_format\_options> | <parquet\_format\_options> | <csv\_format\_options> }*

*[ (<column\_name>, ... ) ]*

*[SET <column\_name> = <expression>,... | <pipeline\_source\_file>]*

*[WHERE <expression>,...]*

*[ON DUPLICATE KEY UPDATE <column\_name> = <expression>, [...]]*

*[NULL DEFINED BY <string> [OPTIONALLY ENCLOSED]]*

AS LOAD DATA: người dùng có thể tải dữ liệu bằng cách chỉ định nguồn dữ liệu từ Kafka, S3, filesystem, Azure, HDFS, GCS, Link. Với một số nguồn dữ liệu cần chỉ định thông tin xác thực cần khai báo thêm CONFIG và CREDENTIALS.

BATCH\_INTERVAL: tham số chỉ định khoảng thời gian (tính bằng milliseconds) mà pipeline đợi trước khi kiểm tra nguồn dữ liệu để tìm dữ liệu mới, sau khi tất cả dữ liệu hiện có đã được tải từ nguồn dữ liệu, giá trị mặc định là 2500.

OUT\_OF\_ORDER OPTIMIZATION: Theo mặc định, các bản ghi có thể được chèn không theo thứ tự, khi tham số này được đặt là DISABLE, các bản ghi khi được nhập vào database sẽ theo thứ tự xuất hiện trong phân vùng.

WITH TRANSFORM: với tính năng này, người dùng có thể tự định nghĩa việc biến đổi dữ liệu, nó có thể là tập lệnh với một ngôn ngữ nào đó. Một transform nhận data từ nguồn, định hình, biến đổi theo lệnh được tạo và trả kết quả về cho pipeline. Pipeline sẽ nhập dữ liệu được trả về vào trong database. Các tham số với transform có thể là:

* uri: URI của biến đổi là vị trí mà từ đó có thể tải xuống chương trình do người dùng cung cấp có thể là dạng http:// hoặc file://. Nếu URI chứa một tarball có phần mở rộng .tar.gz hoặc .tgz, thì nội dung của nó sẽ được trích xuất tự động. Nếu uri chứa một tarball, tham số program phải được chỉ định. Ngoài ra, nếu URI chỉ định chính tên tệp chương trình do người dùng cung cấp (chẳng hạn như file://localhost/root/path/to/my-transform.py), thì tham số program và arguments có thể để trống.
* program: Tên tệp của chương trình do người dùng cung cấp để áp dụng cho pipeline. Tham số này là bắt buộc nếu tarball được chỉ định làm điểm cuối cho biến đổi url. Nếu url chỉ định chính tệp chương trình do người dùng cung cấp, tham số này có thể trống.
* arguments: Các đối số được truyền cho biến đổi khi chạy. Mỗi đối số phải được phân cách bằng khoảng trắng.

SKIP ... ERRORS: cho phép định nghĩa kịch bản lỗi, khi gặp trường hợp này sẽ loại bỏ hàng gặp lỗi đó.

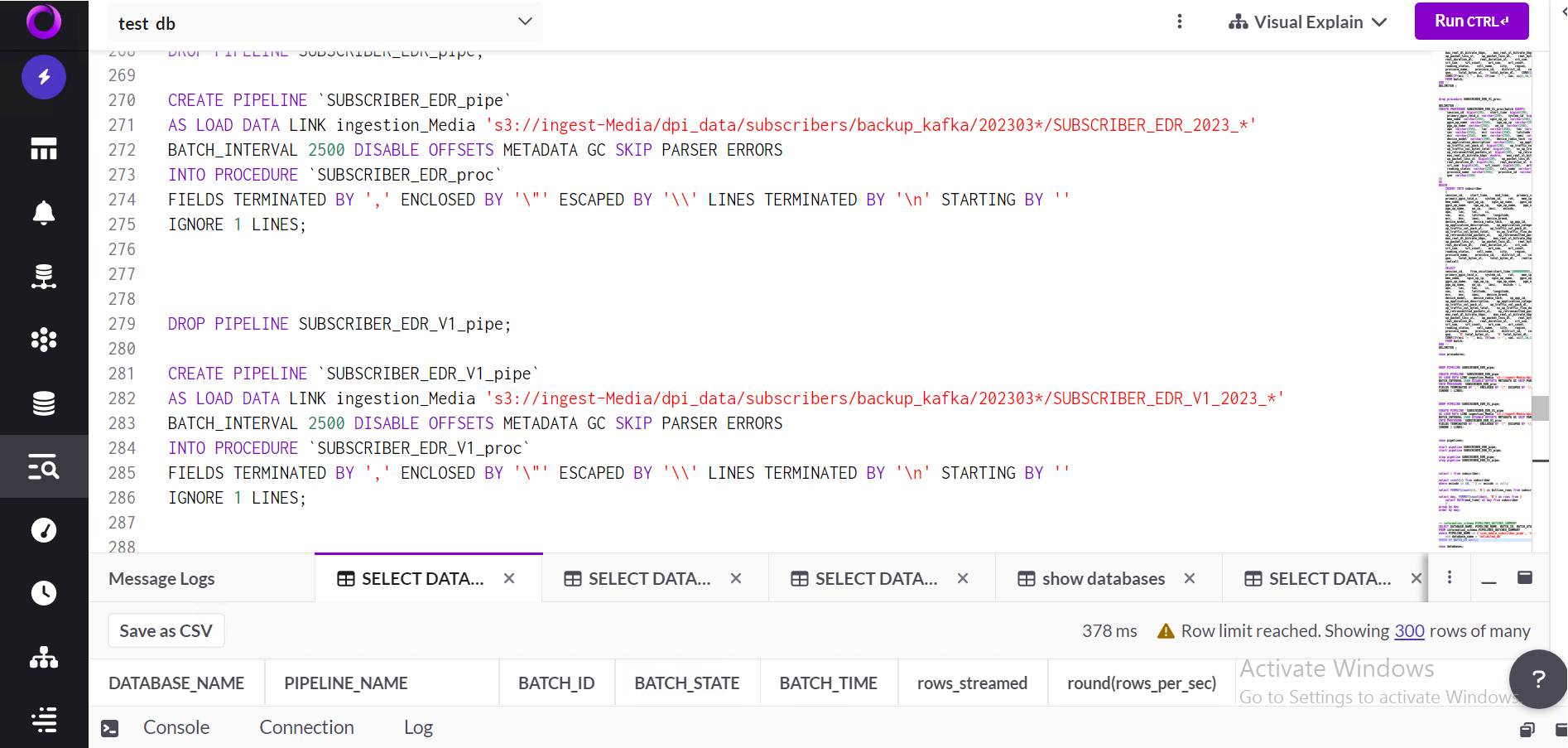
IGNORE ... ERRORS: không giống như SKIP, IGNORE khi gặp lỗi sẽ biến đổi hàng được insert thành đúng định dạng và insert vào table, mọi giá trị không hợp lệ sẽ được cập nhật về giá trị default. Vì vậy, mệnh đề này sẽ không đảm bảo tính toàn vẹn dữ liệu.

OR REPLACE: khi mệnh đề này được cung cấp thì pipeline được tạo sẽ thay thế pipeline đã tồn tại, pipeline mới sẽ được định nghĩa lại. REPLACE cũng có thể được dùng với IGNORE và SKIP, tuy nhiên nó sẽ đứng sau mệnh đề SKIP hoặc IGNORE.

INTO TABLE: table được truyền dữ liệu vào.

INTO PROCEDURE: Tạo một quy trình sử dụng stored procedure để định hình dữ liệu được trích xuất từ ​​nguồn dữ liệu. (SingleStore)

Thu thập dữ liệu streaming với SingleStore bằng cách tạo pipeline như sau:



Hình 25: Tạo pipeline nhập dữ liệu streaming

Hình trên là mẫu hai pipeline lấy dữ liệu từ dịch vụ lưu trữ S3 của VNPT và đẩy vào procedure để định hình lại dữ liệu. FIELDS TERMINATED BY ',' có nghĩa các trường trong tệp dữ liệu được phân tách bằng dấu phẩy. ENCLOSED BY '\"' có nghĩa là giá trị của mỗi trường dữ liệu được bao quanh bởi dấu \". ESCAPED BY '\\' được sử dụng để chỉ định ký tự thoát (bỏ qua các ký tự đặc biệt, ở đây là ký tự \ trong dữ liệu đầu vào). LINES TERMINATED BY '\n' trong câu lệnh được sử dụng để chỉ định ký tự được sử dụng để phân tách các dòng trong tệp đầu vào, ở đây là ký tự xuống dòng. STARTING BY '' được sử dụng để chỉ định ký tự bắt đầu của mỗi dòng trong tệp đầu vào. IGNORE 1 LINES dùng để chỉ định cho SingleStore bỏ qua dòng đầu tiên trong tệp đầu vào trước khi bắt đầu chèn dữ liệu.



Hình 26: Pipeline running

Các tính năng của pipeline trong SingleStore làm cho nó trở thành một giải pháp thay thế mạnh mẽ cho các phần mềm ETL trung gian của bên thứ ba trong nhiều tình huống bởi 6 lý do sau đây:

* Tải liên tục một cách dễ dàng: các pipeline giám sát các thư mục hoặc Kafka khi các tệp hoặc tin nhắn mới đến và tự động tải chúng.
* Khả năng mở rộng: Các pipeline đã mở rộng quy mô với SingleStoreDB Cloud workspaces cũng như các hệ thống nguồn dữ liệu phân tác như Kafka và lưu trữ đám mây như Amazon S3.
* Hiệu suất cao: Dữ liệu trong pipeline được tải song song từ nguồn dữ liệu đến các leaves SingleStore. Trong hầu hết các trường hợp, điều này cải thiện thông lượng bằng cách bỏ qua bộ tổng hợp. Ngoài ra, pipeline đã được tối ưu hóa cho khả năng tranh chấp khóa và đồng thời thấp.
* Exactly-once Semantics: Kiến trúc của pipeline đảm bảo rằng các giao dịch được xử lý chính xác một lần, ngay cả trong trường hợp chuyển đổi dự phòng.
* Gỡ lỗi: Pipeline giúp gỡ lỗi từng bước trong quy trình ETL dễ dàng hơn bằng cách lưu trữ metadata đầy đủ về các transaction.
* Đồng thời: Nhiều pipeline có thể chèn dữ liệu vào một bảng. (SingleStore)

**IBM InfoSphere DataStage** hỗ trợ nhập dữ liệu từ các nguồn rất đa dạng với dự án xây dựng hồ dữ liệu cho VNPT, DataStage được sử dụng để ingest dữ liệu theo dạng batch. Với DataStage, người dùng sẽ định nghĩa flow (luồng) dữ liệu từ hệ thống nguồn đến hệ thống đích.

Các thành phần cơ bản của một luồng dữ liệu bao gồm:

* Data sources nguồn dữ liệu để đọc
* Stages là các giai đoạn chuyển đổi dữ liệu
* Data targets là đích của dữ liệu để ghi
* Links là các kết nối giữa sources, stages và targets.

Để khởi chạy flow, cần tạo job với flow và khởi chạy, có thể đặt lịch chạy các job theo nhu cầu sử dụng. Các bước để tạo một DataStage flow như sau:

Bước 1: Mở một project đã tồn tại hoặc tạo một project mới.

Bước 2: Trên tab Assets, chọn New asset → Graphical builders → DataStage

Bước 3: Trong trag Create a DataStage flow, chọn một trong hai phương thức để tạo luồng DataStage:

* Chọn New tab và điền các thông tin cần thiết cho DataStage flow và nhấn Create. Một luồng DataStage trống sẽ được khởi tạo.
* Chọn tab Local file và đẩy lên file ISX hoặc ZIP từ local và nhấn Create. Khi nhập thành công, đóng trang nhập lại và mở flow đã nhập ở Assets tab.

Bước 4: Kéo các connectors (nguồn dữ liệu) hoặc stages (các giai đoạn biến đổi) từ giao diện để tạo luồng dữ liệu theo nhu cầu sử dụng.

Bước 5: Nhấn đúp chuột vào một node để cài đặt cấu hình.

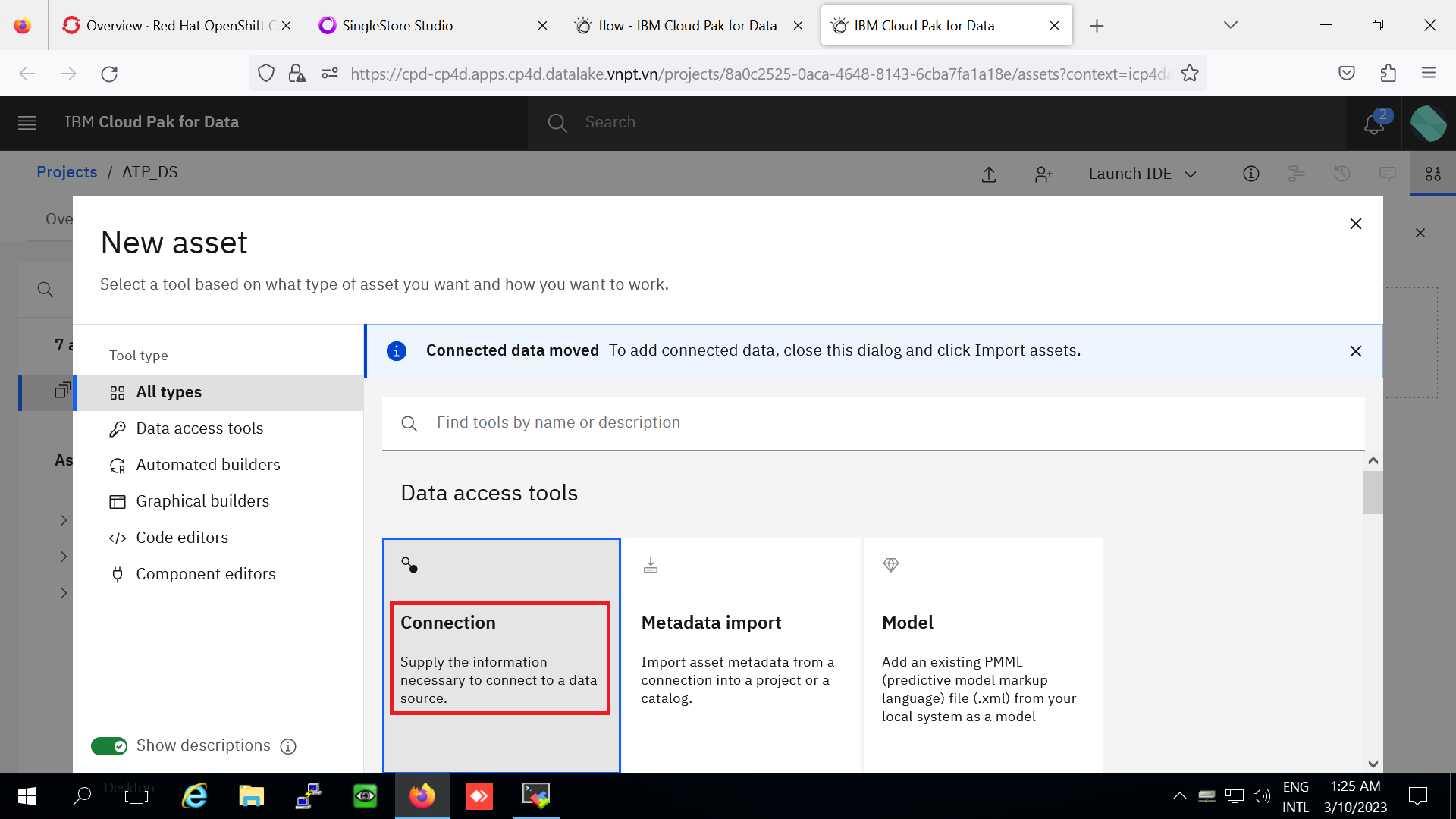
Bước 6: Click Run khi hoàn thành việc tạo flow. Flow sẽ được tự động lưu, biên dịch và chạy. Người dùng có thể xem logs của biên dịch và job run.

Sau khi quy trình được biên dịch thành công việc, bạn có thể chạy lại công việc, đặt lịch, giám sát công việc và cập nhật môi trường mà bạn muốn chạy công việc đó.

Dưới đây là mô tả chi tiết các thành phần cơ bản trong một luồng DataStage. Trước hết là nguồn dữ liệu, với DataStage, cung cấp kết nối đến các nguồn dữ liệu như:

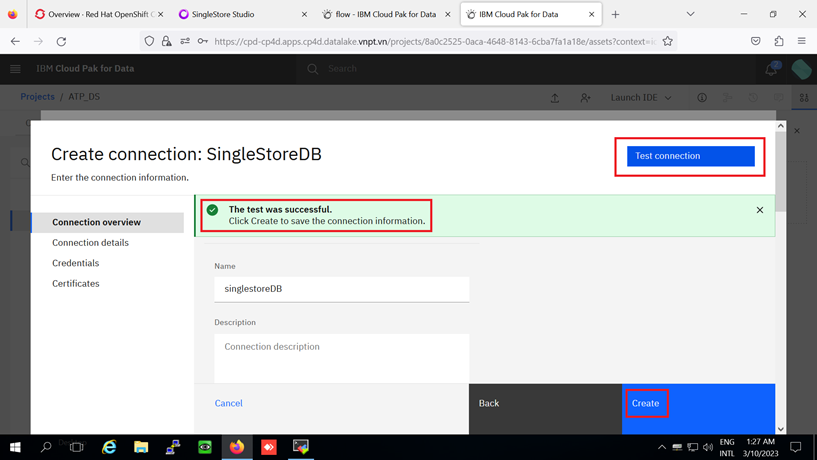
Amazon S3, Azure, BDFS (Linux only), BigQuery, Cassandra, Classic federation, Cloud object storage, Complex flat file, Data set, Db2, Distributed transaction, DRS, External source, External target, File, File set, FTP enterprise, Google cloud storage, Greenplum, Hierarchical (JSON composer step, JSON parser step, REST step, Test assembly, Details inspector, Create / view contract libraries, Administration), Hbase, Hierarchical, Hive, Informix enterprise, Informix load, ISD input, ISD output, Java Integration, JDBC, Kafka, Lookup file set, Netezza, ODBC, Oracle, Salesforce, SAP Odata, Sequential file, Snowflake, Sybase enterprise, Sybase IQ12 load, Sybase OC, Teradata, WebSphere MQ. (IBM, 2021)

Dưới đây là ví dụ về việc kết nối đến nguồn dữ liệu cơ chế DB Connection tới SingleStoreDB. Bước đầu tiên, cần tạo một asset mới và chọn tools là Connection. Sau đó chọn connection là SingleStore và điền các thông tin truy cập đến SingleStore.



Hình 27: Tạo connection mới

Sau khi điền đầy đủ thông tin kết nối đến SingleStore chọn Test connection để kiểm tra kết nối có thành công hay không. Nếu thành công chọn Create để tạo connection.

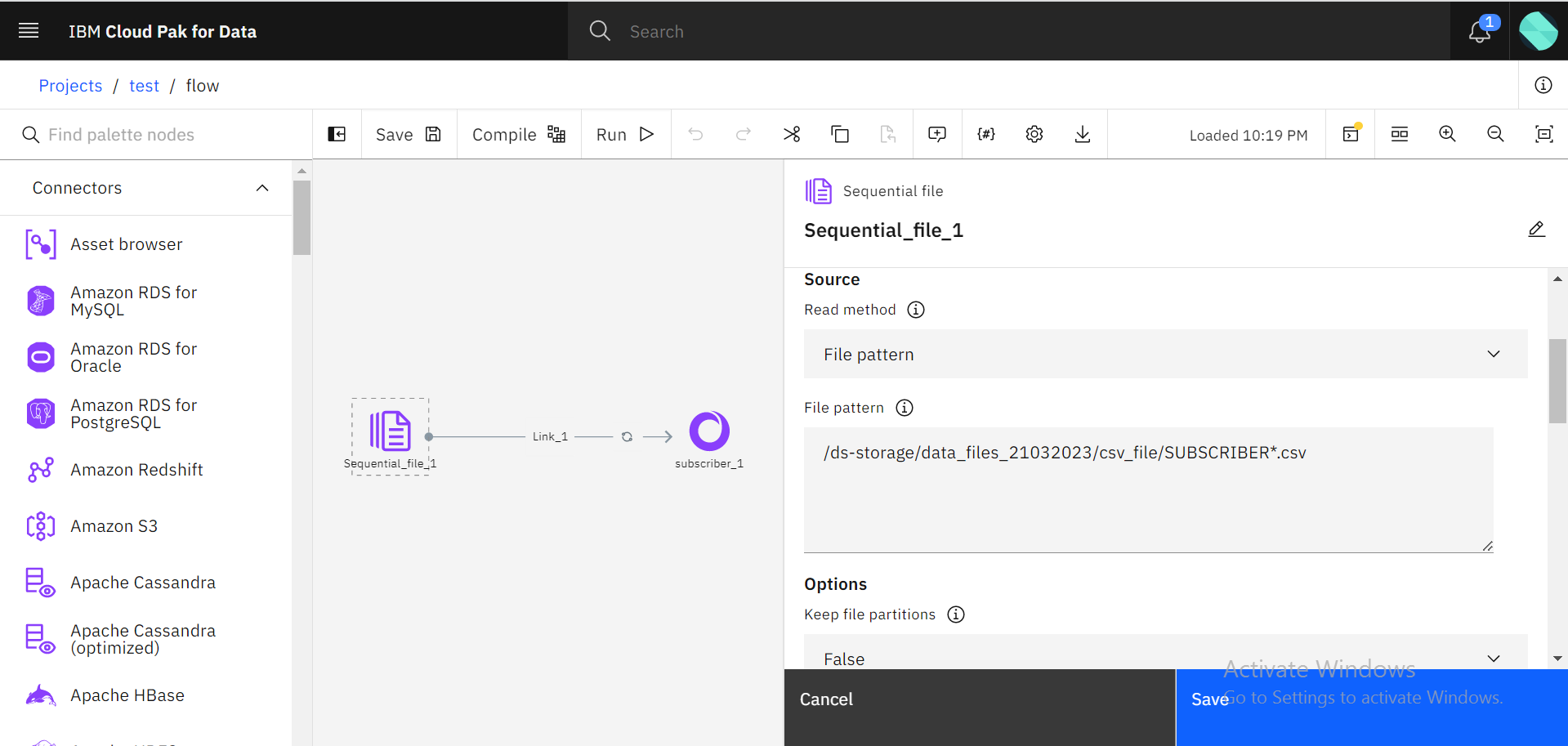


Hình 28: Tạo kết nối tới SingleStore

Tiếp đến là biến đổi dữ liệu trước khi đầy vào đích. DataStage hỗ trợ các giai đoạn biến đổi dữ liệu như:

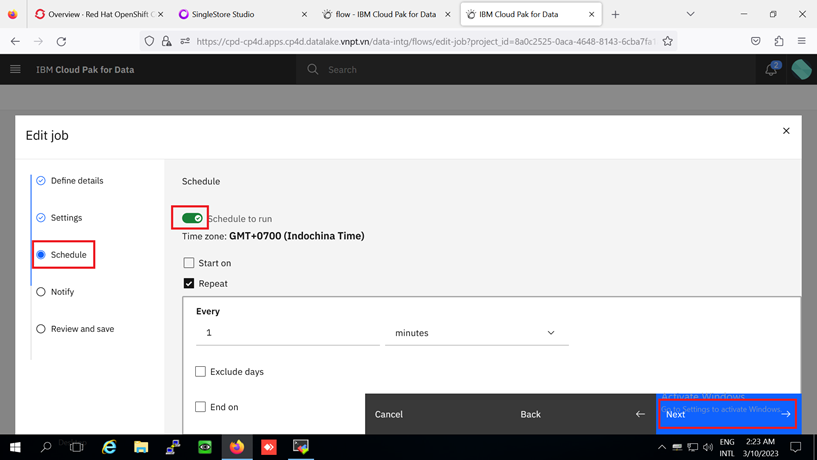
Bảng 6: Các giai đoạn biến đổi được DataStage hỗ trợ

|  |  |
| --- | --- |
| **Stage** | **Mô tả** |
| Aggregator | Phân loại các hàng dữ liệu từ một liên kết vào duy nhất thành các nhóm và tính tổng hoặc các hàm tổng hợp khác cho từng nhóm. |
| Change Apply | Là một giai đoạn xử lý. Lấy tập dữ liệu thay đổi có chứa các thay đổi trong tập dữ liệu trước và sau từ Change Capture và áp dụng các thao tác thay đổi được mã hóa cho tập dữ liệu trước để tính toán tập dữ liệu sau. |
| Change Capture | Giai đoạn xử lý so sánh hai bộ dữ liệu và tạo bản ghi về sự khác biệt. |
| Checksum | Tạo giá trị checksum từ các cột được chỉ định trong một hàng và thêm checksum vào hàng để kiểm tra tính hợp lệ của mỗi hàng khi được ghi vào target. |
| Column Export | Xuất dữ liệu từ một số cột có kiểu dữ liệu khác nhau thành một cột duy nhất có kiểu dữ liệu ustring, string hoặc nhị phân. |
| Column Generator | Thêm các cột vào dữ liệu đến và tạo dữ liệu giả cho các cột này với mỗi hàng dữ liệu được xử lý. Sau đó xuất ra tập dữ liệu mới. |
| Column Import | Import dữ liệu từ một cột và xuất sang một hoặc nhiều cột. |
| Compare | Thực hiện so sánh các bản ghi theo từng cột trong hai bộ dữ liệu đầu vào được sắp xếp trước. |
| Compress | Sử dụng tiện ích nén UNIX hoặc GZIP để nén tập dữ liệu. Chuyển đổi một tập dữ liệu từ một chuỗi các bản ghi thành một luồng dữ liệu nhị phân thô. |
| Copy | Sao chép một bộ dữ liệu đầu vào thành một số bộ dữ liệu đầu ra. |
| Decode | Giải mã tập dữ liệu bằng cách sử dụng lệnh giải mã UNIX, chẳng hạn như gzip mà bạn cung cấp. |
| Encode | Mã hóa mã hóa tập dữ liệu bằng cách sử dụng lệnh mã hóa UNIX, chẳng hạn như gzip, mà bạn cung cấp. |
| Expand | Giai đoạn xử lý sử dụng tiện ích giải nén UNIX hoặc GZIP để mở rộng tập dữ liệu. |
| External Filter | Lọc dữ liệu đầu vào từ các nguồn khác nhau theo các điều kiện được xác định trước. |
| Filter | Chuyển các bản ghi tập dữ liệu đầu vào thỏa mãn các yêu cầu đã chỉ định và lọc ra tất cả các bản ghi khác. |
| Funnel | Sao chép nhiều bộ dữ liệu đầu vào sang một bộ dữ liệu đầu ra duy nhất. Bạn có thể sử dụng thao tác này để kết hợp các tập dữ liệu riêng biệt thành một tập dữ liệu chung lớn hơn và duy nhất. |
| Head | Chọn N hàng đầu tiên từ mỗi phân vùng của tập dữ liệu đầu vào và sao chép các hàng đã chọn sang tập dữ liệu đầu ra. |
| Hierarchical data transformation | Tạo các phép biến đổi phân cấp, phân tích cú pháp, soạn dữ liệu JSON và XML, đồng thời gọi các dịch vụ web REST với hiệu suất cao và khả năng mở rộng. |
| Join | Thực hiện các thao tác nối trên hai hoặc nhiều tập dữ liệu đầu vào cho giai đoạn và sau đó xuất tập dữ liệu kết quả. |
| Lookup | Sử dụng để thực hiện các thao tác tra cứu trên tập dữ liệu được đọc vào bộ nhớ từ bất kỳ Parallel job stage khác có thể xuất dữ liệu. Cách sử dụng phổ biến nhất để tra cứu là ánh xạ các mã ngắn trong tập dữ liệu đầu vào vào thông tin mở rộng từ bảng tra cứu, sau đó được nối với dữ liệu đầu vào và đầu ra. |
| Merge | Cho phép ghép nối các bản ghi từ các nguồn dữ liệu khác nhau thông qua một trường chung hoặc một tập hợp các trường chung. |
| Modify | Biến đổi schema của tập dữ liệu đầu vào. |
| Remove Duplicates | Lấy một tập dữ liệu được sắp xếp duy nhất làm đầu vào, loại bỏ tất cả các hàng trùng lặp và ghi kết quả vào tập dữ liệu đầu ra. |
| Row Generator | Tạo ra một tập hợp dữ liệu mô phỏng phù hợp với siêu dữ liệu đã chỉ định. Sử dụng để kiểm tra job nhưng không có sẵn dữ liệu thực để xử lý. |
| Sample | Trích xuất ra mẫu từ tập dữ liệu đầu vào. |
| Slowly Changing Dimension stage | Cho phép xác định các cột cần theo dõi sự thay đổi của chúng trong một khoảng thời gian và cập nhật các giá trị mới của chúng trong bảng dữ liệu. Hỗ trợ SCD loại 1 (ghi đè) và loại 2 (thêm). |
| Sort | Thực hiện sắp xếp. |
| Web Service | Truy cập các hoạt động dịch vụ web trong flow hoặc job. |

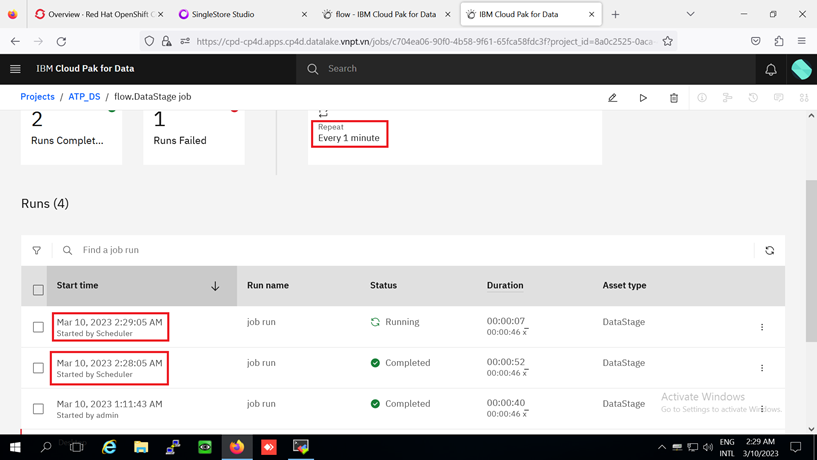


Hình 29: Tạo luồng nhập dữ liệu từ nguồn tới đích với DataStage

Đối với DataStage, có tính năng đặt lịch tự động chạy các tiến trình ETL. Người dùng có thể đặt lịch job chạy ETL theo đúng yêu cầu. Trong giao diện Project, chọn job muốn đặt lịch và vào Edit job để thiết lập các thông số.



Hình 30: Thiết lập lịch chạy cho job

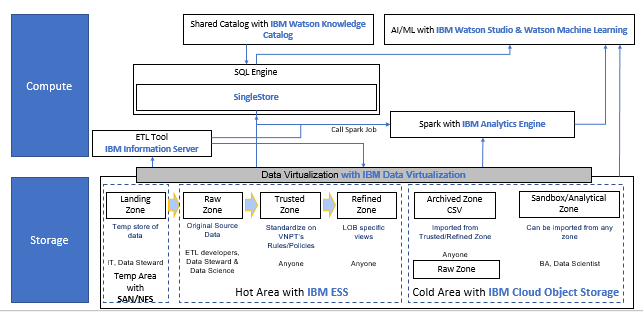


Hình 31: Lịch chạy của job được tạo thành công

1. Lưu trữ dữ liệu

Trong hệ thống lưu trữ dữ liệu của hồ dữ liệu cho VNPT bao gồm IBM ESS, ICOS và phần lưu trữ của database SingleStore. Thành phần lưu trữ tốc độ cao (Hot) sử dụng IBM Elastic Storage System 3200, dung lượng lưu trữ khả dụng 340 TiB, được sử dụng cho việc lưu trữ các dữ liệu yêu cầu tốc độ cao hoặc thường xuyên được truy cập, thay đổi. Thành phần lưu trữ dung lượng cao (Cold) sử dụng giải pháp IBM Cloud Object Storage (ICOS), dung lượng lưu trữ khả dụng 3054 TiB, phục vụ cho việc lưu trữ các dữ liệu yêu cầu dung lượng cao, không yêu cầu đọc/ghi tốc độ cao và thường xuyên thay đổi. Giữa hai thành phần lưu trữ tốc độ cao và dung lượng cao có khả năng tự động tiering. Hệ thống lưu trữ được mở rộng theo chiều ngang (Scale-Out) bằng cách thêm node lưu trữ dung lượng và hiệu năng đáp ứng nhu cầu lưu trữ dữ liệu.

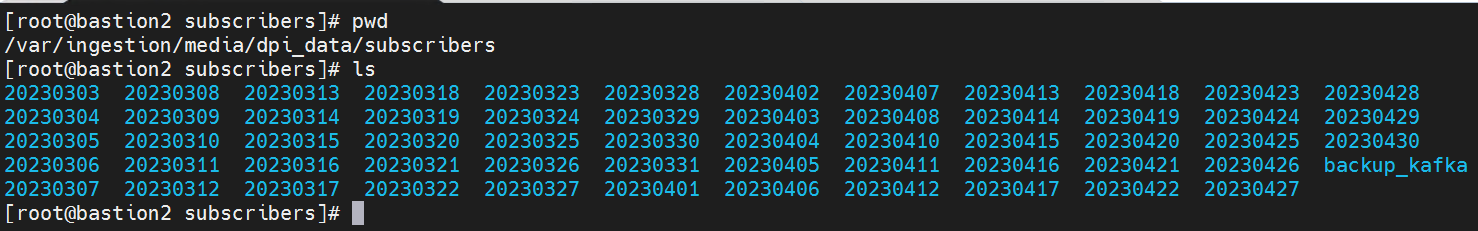
Để đảm bảo hiệu suất truy vấn cao, đáp ứng được cả OLTP và OLAP hệ thống tách biệt khu vực lưu trữ và tính toán/xử lý thành hai tầng riêng biệt.

****

Hình 32: Tổng quan các chức năng ở tầng tính toán/xử lý và lưu trữ

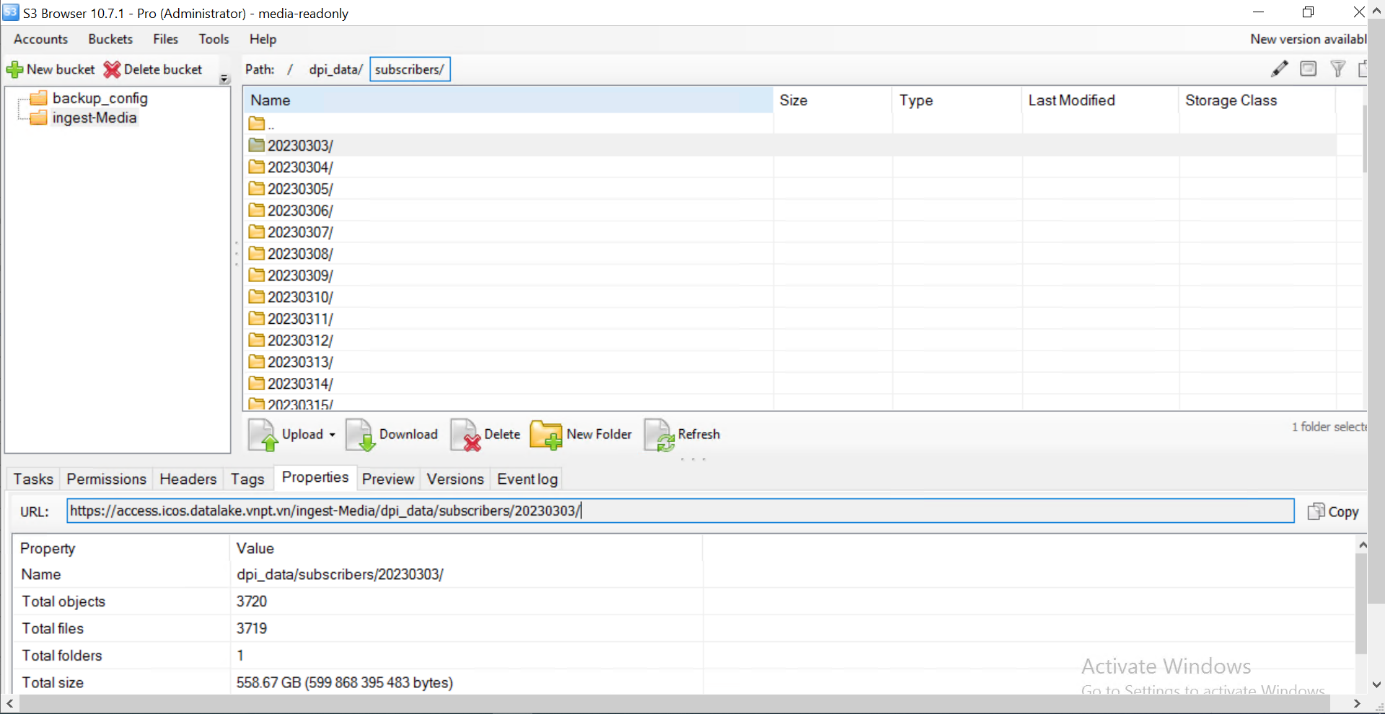
Trong tầng lưu trữ, phân vùng dữ liệu thô (Raw Zone) là khu vực chứa mọi dữ liệu thô được thu thập bởi khối Data Ingestion. Phân vùng dữ liệu được tin cậy (Trusted Zone) nhằm mục đích trở thành nơi lưu trữ, sở hữu và tập trung số liệu chuẩn xác của toàn Tập đoàn, trở thành nguồn số liệu hệ quy chiếu cho các đơn vị khác so sánh và hiệu chỉnh. Phân vùng dữ liệu được làm sạch có cấu trúc (Structured and Refined Data Zone) là phân vùng chuẩn hóa lại, làm sạch lại, làm giàu dữ liệu ở phân vùng dữ liệu thô để trích xuất ra một phân vùng dữ liệu đáng tin cậy hơn, chính xác hơn và dễ dàng sử dụng hơn. Phân vùng dữ liệu phân tích (Analytics Zone) lưu trữ các dữ liệu phục vụ quá trình phân tích, tính toán, xử lý cũng như kết quả phân tích, thống kê, tổng hợp dữ liệu trên nền tảng VNPT Data Lake.

Thành phần lưu trữ hot data IBM Elastic Storage System 3200. IBM ESS là giải pháp software defined storage bao gồm một tập hợp phần cứng và phần mềm đã được tích hợp thành một cụm lưu trữ. Các thành phần chính của giải pháp IBM ES bao gồm: Phần cứng IBM ESS 3200 bao gồm máy chủ quản trị (IBM Elastic Storage System Management Server), máy chủ dữ liệu (Data Server hardware), các khay đĩa lưu trữ (Storage enclosures), các protocol node (Protocol nodes) và các switch kết nối (Network switches). Tại đây, lưu trữ các dữ liệu thường xuyên thay đổi, với dữ liệu sau một tháng không truy cập sẽ được lưu trên Object Storage.



Hình 33: Dữ liệu lưu trữ tại ESS

Thành phần lưu trữ cold data IBM Cloud Object Storage (ICOS) như sau:



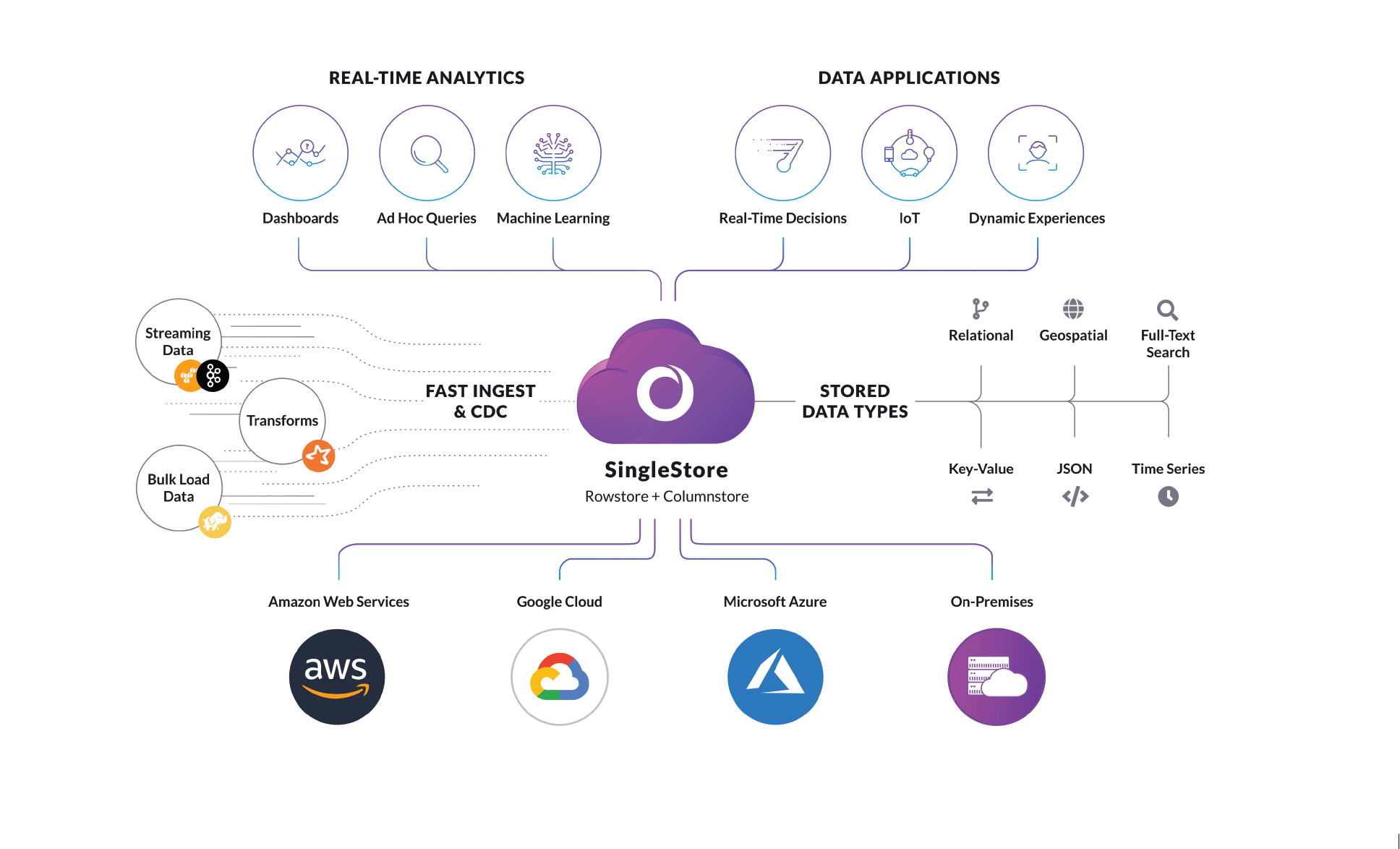
Hình 34: Dữ liệu lưu trữ tại ICOS

Hệ thống lưu trữ ICOS là hệ thống lưu trữ dữ liệu phi cấu trúc, các dữ liệu được ghi xuống hệ thống là các Object, khác với lưu trữ theo kiểu block hoặc theo file mỗi Object sẽ bao gồm các metadata và data được lưu trữ xuống hệ thống ICOS. Hệ thống lưu trữ ICOS được thiết kế để lưu trữ dữ liệu lớn lên tới hàng exabyte. IBM COS là một giải pháp lưu trữ dạng software-defined storage, IBM COS là giải pháp lưu trữ có khả năng sẵn sàng và độ tin cậy cao sử dụng các máy chủ x86. IBM COS tận dụng được tất cả những ưu thế của các máy chủ x86 về khả năng giám sát, quản trị và hiệu năng.

1. Phân tích dữ liệu

Việc phân tích dữ liệu được sử dụng với CSDL SingleStore và Analytics Engine Powered by Apache Spark. Với **SingleStore**, cung cấp các giải pháp để nâng cao hiệu suất với việc phân tích dữ liệu, đáp ứng yêu cầu của VNPT là hỗ trợ cả OLAP và OLTP.

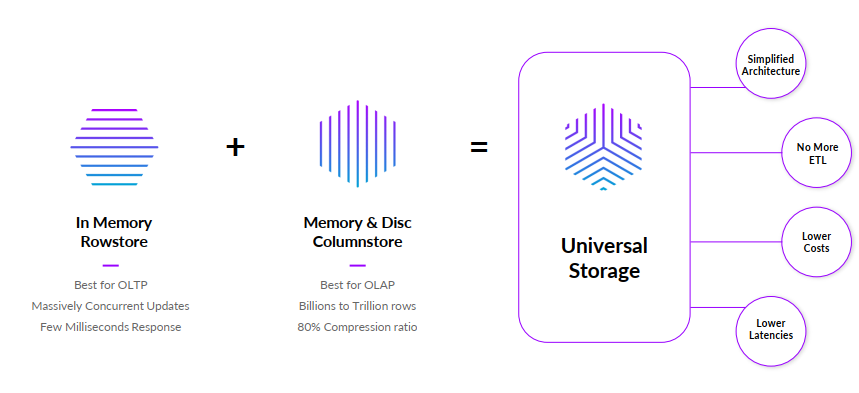
Được xây dựng với kiến ​​trúc lưu trữ ba tầng độc đáo và được thiết kế cho thời gian phản hồi tính bằng mili giây, SingleStoreDB là cơ sở dữ liệu SQL phân tán nhanh nhất thế giới dành cho các ứng dụng phân tích thời gian thực. Bằng cách kết hợp OLTP và OLAP với một cấu trúc đa mô hình trong một công cụ duy nhất, SingleStoreDB loại bỏ các cản trở đến hiệu suất và việc di chuyển dữ liệu không cần thiết.



Hình 35: SingleStore kết hợp OLTP và OLAP

*Nguồn: SingleStore, 2022*

Định dạng Universal Storage của SingleStore tập hợp hiệu suất quét bảng nhanh của columnstore và hiệu suất tìm kiếm cao của rowstore, cùng với lưu trữ đối tượng đám mây.



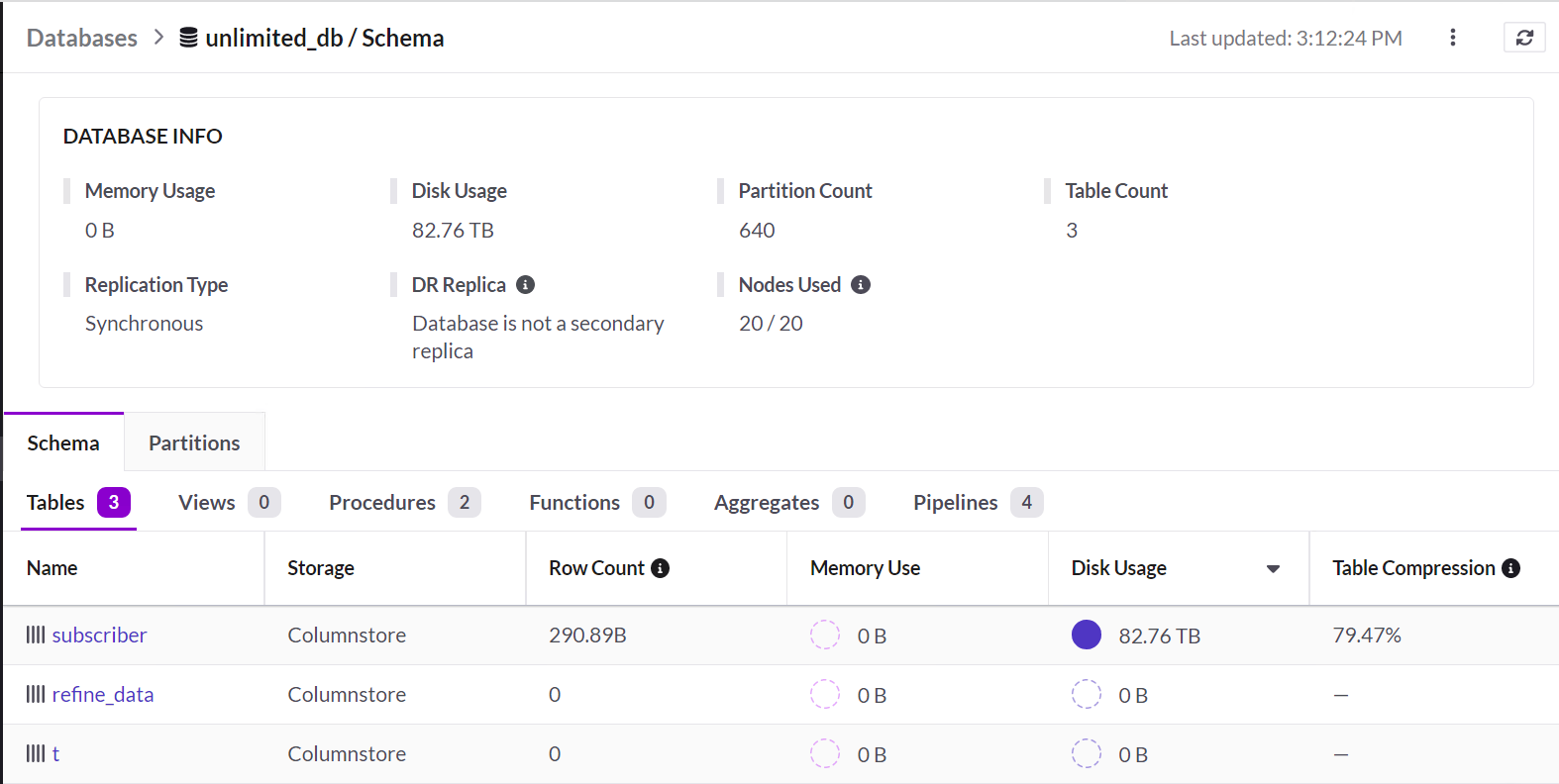
Hình 36: Mô tả Universal Storage

*Nguồn: SingleStore, 2022*

Rowstore được lưu trữ trong bộ nhớ nên có thể update đồng thời, phản hồi nhanh chỉ với mili giây, ứng dụng với OLTP. Columnstore được lưu trữ trên đĩa, nên dung lượng lưu trữ lớn hơn và phù hợp với OLAP. Tuy nhiên, với rowstore dung lượng lưu trữ không lớn và chi phí cao. Vì vậy, SingleStore đưa ra giải pháp cho phép rowstore lưu trữ nhiều dữ liệu hơn trong cùng một dung lượng RAM, để giảm thiểu chi phí mà vẫn đảm bảo hiệu suất cao, phương sai thấp với mỗi lần tìm kiếm. Việc này được thực hiện nhờ việc nén các sparse trong bộ nhớ. Và một cách khác là hỗ trợ columnstore trong việc tìm kiếm, cho phép truy cập, đọc và ghi đồng thời. Việc này đạt được nhờ sử dụng hash indexes và lược đồ row-level locking cho kho lưu trữ cột, cộng với quyền truy cập phân vùng con, một phương pháp để đọc các phần nhỏ của cột kho lưu trữ cột một cách độc lập và hiệu quả. (SingleStore, 2022)

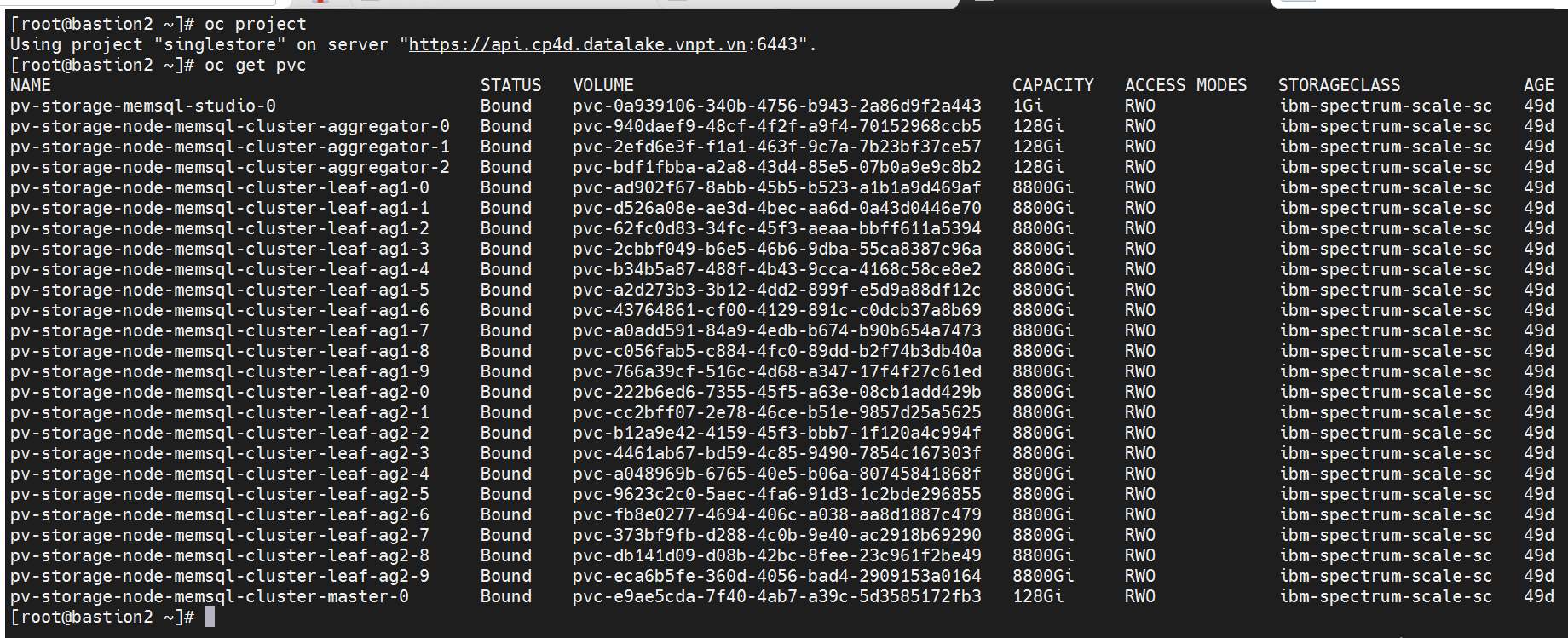
Trước hết, nén các sparse trong bộ nhớ. SingleStore phát triển một phương pháp nén rowstore data trong bộ nhớ. Bản ghi sẽ được chia thành hai phần: phần có chiều rộng cố định chứa các trường non-sparse, index key và phần có chiều rộng thay đổi chứa các trường sparse. Các trường sparse được nén lại để giảm dung lượng.

Columnstore có thể tìm kiếm với tính đồng thời cao. Với phương pháp nén các sparse trong bộ nhớ có thể mất nhiều thời gian và đồng thời dữ liệu vẫn phải lưu trữ trong RAM. Vì vậy phương pháp nâng cao Columstore được sử dụng. Các bảng columnstore được phân chia thành các segments. Trong mỗi segment, các cột được lưu trữ độc lập – trong các phần liền kề của tệp hoặc trong chính tệp. Các đoạn cột được lưu trữ này được gọi là các column segments. Để tăng tốc độ truy vấn vào một hàng trong columnstore, hệ thống sẽ tính toán vị trí của dữ liệu cho hàng đó, sau đó chỉ đọc phần của column segments cần thiết để cụ thể hóa hàng đó. Để có thể biết được vị trí của hàng và làm việc đó hiệu quả, sử dụng hash indexes và lược đồ row-level locking (chi tiết tham khảo https://www.singlestore.com/blog/memsql-singlestore-then-there-was-one/) cho kho lưu trữ cột.



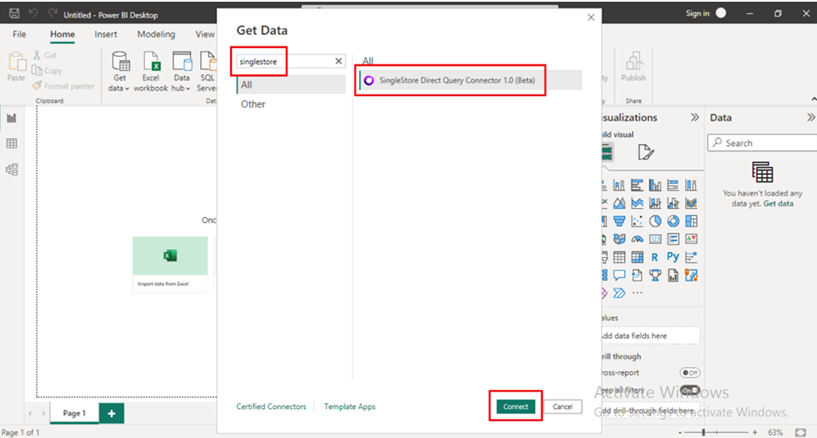
Hình 37: Tỷ lệ nén dữ liệu thực tế của SingleStore với các bảng dữ liệu của VNPT

Thành phần lưu trữ trên database SingleStore được cấu tạo với 4 aggregator và 20 leaf, dữ liệu sẽ được tự động chia sẻ giữa các node trong cùng một cluster. Đây cũng là một yếu tố giúp tăng tốc độ truy vấn dữ liệu.



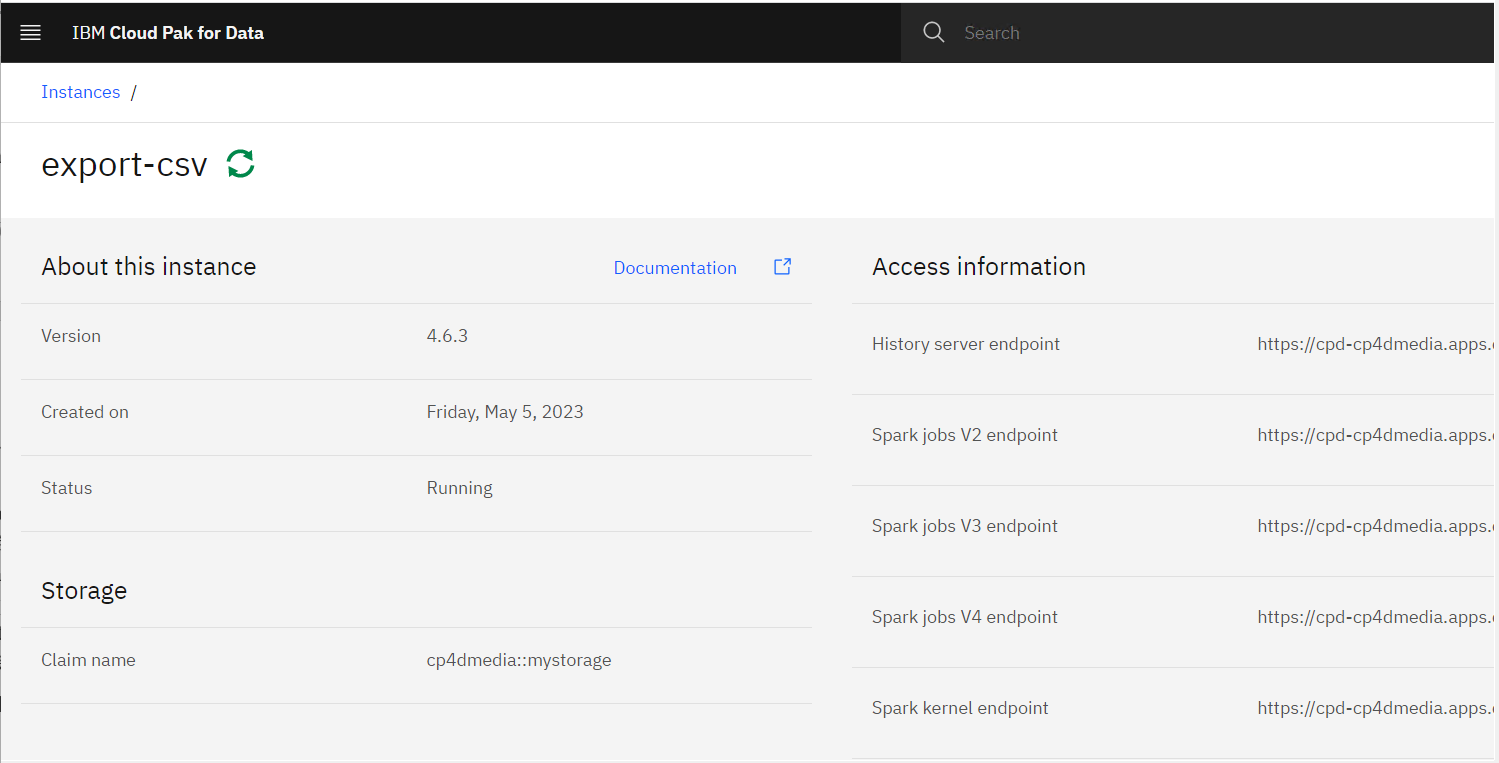
Hình 38: Thông tin các node trong SingleStore

Không chỉ cung cấp cơ chế phục vụ cho OLTP và OLAP, SingleStoreDB cũng cung cấp các kết nối từ các BI tool để phục vụ cho việc xây dựng các báo cáo, dashboard cho tập đoàn VNPT. Các BI tool được hỗ trợ bao gồm: Cognos Analytics, Power BI, Tableau, v.v.



Hình 39: PowerBI kết nối tới SingleStore

**Analytics Engine Powered by Apache Spark** là công cụ do Apache Spark cung cấp, người dùng có thể chạy các analytical và machine learning job trên cụm Spark. Instance này cũng hỗ trợ tương thích với Spark Streaming. Để tạo một instance Analytic Engine by Apache Spark sử dụng giao diện web của CP4D, truy cập tab Instance chọn New Instance. Sau đó chọn loại instance là Analytics Engine Powered by Apache Spark và điền các thông số cần thiết. Mỗi instance Analytics Engine Powered by Apache Spark được gán với một volume service instance, tại đây sẽ chứa các Driver log và Spark event log của mỗi job Spark.



Hình 40: Thông tin instance Analytics Engine Powered by Apache Spark

Để submit Spark job có hai cách.

Cách thứ nhất: Submit với file processing trong Storage Volume.

Bước 1: Tạo Storage Volume

Tại bước này cần tối thiểu 2 Storage Volume được dùng, một Storage Volume để tạo instance Spark ở bước chuẩn bị và một hoặc nhiều Storage Volume khác để lưu các processing file và data file.

Bước 2: Tạo Access Token

token = $(curl -k -X POST <V4\_JOBS\_API\_ENDPOINT> -H "Authorization: Bearer <ACCESS\_TOKEN>" -d @input.json)

Bước 3: Start History để lưu lịch sử submit job

curl -ik -X POST <V4\_JOBS\_API\_ENDPOINT>/<instance\_id>/spark\_history\_server -H "Authorization: Bearer $token"

Bước 4: Submit Spark job

Việc submit Spark sẽ được thực hiện với instance `submit-spark` với Payload bên dưới. Mouth path là path để định nghĩa nơi application được mount.

curl -k -X POST <V4\_JOBS\_API\_ENDPOINT>/<instance\_id>/spark\_applications -H "Authorization: Bearer $token" -d '{

"application\_details": {"application": "/td/ex\_csv.py"},

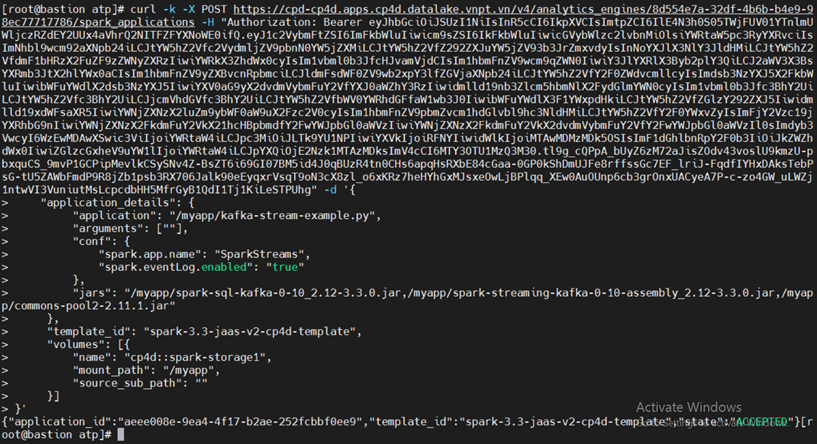
"volumes": [{"name": "cp4dmedia::mystorage", "mount\_path": "/td"}]

}'

Sau khi được submit, một responese được trả về kết quả việc submit và application\_id. Để check trạng thái job sử dụng payload:

curl -k -X GET <V4\_JOBS\_API\_ENDPOINT>/<instance\_id>/spark\_applications/<app\_id> -H "Authorization: Bearer $token"

Sau khi tạo được instance Spark, sẽ submit Spark job, một job được submit thành công sẽ có response như dưới đây:



Hình 41: Submit job thành công

Cách thứ hai: Submit với file processing trong Deployment space. Deployment space là một nơi chứ các model, dataset, source code và quản lý các quá trình deploy các model. Để sử dụng tính năng này cần thực hiện các bước sau:

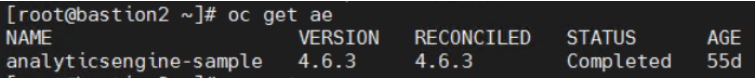
Bước 1: Enabling advanced features

oc patch AnalyticsEngine <analyticsengine-cr-name> --namespace ${PROJECT\_CPD\_INSTANCE} --type merge --patch '{"spec": {"serviceConfig":{"sparkAdvEnabled":true}}}'

Kiểm tra tính năng nâng cao này đã được bật chưa sử dụng cmd:

oc get ae –w

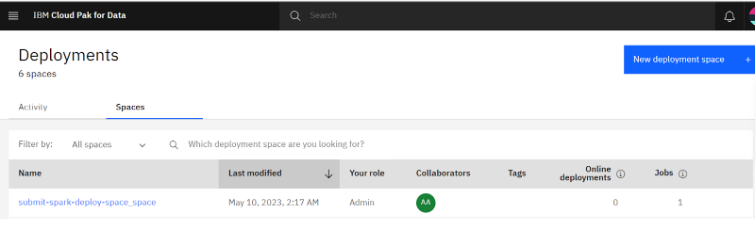
Đến khi có kết quả sau:



Hình 42: Advanced feature được enable

Bước 2: Tạo Spark instance

Instance `submit-spark-deploy-space` đã được tạo với volume `cp4dmedia::submit-spark-deployspace2`. Sau khi tạo, Deployment space đã được tạo cùng với Spark insntace với thông tin `Deployment space name`.



Hình 43: Deployment space được tạo gán với instance

Bước 3: Tạo Acceess Token tương tự như cách 1

Bước 4: Submit Spark job với payload

Cách model, dataset, source code lưu trữ tại assets của Deployment space được truy cập thông qua directory /home/spark/space/assets/data\_asset.

curl -k -X POST https://cpd-cp4dmedia.apps.cp4d.datalake.vnpt.vn/v4/analytics\_engines/<instance\_id>/spark\_applications -H "Authorization: Bearer $token" -d '{

"application\_details": {

"application": "/home/spark/space/assets/data\_asset/export\_csv\_python.py"

} ,

"volumes": [{

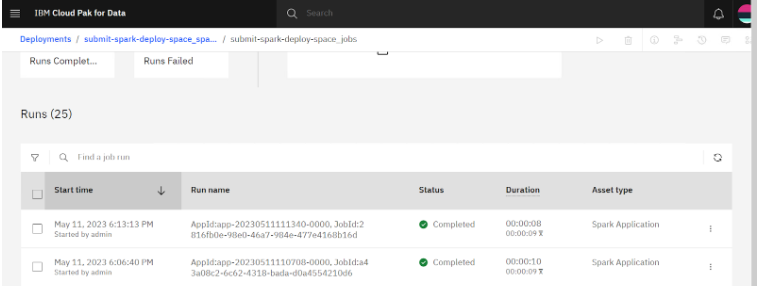
"name": "cp4dmedia::submit-spark-data",

"mount\_path": "/myapp"

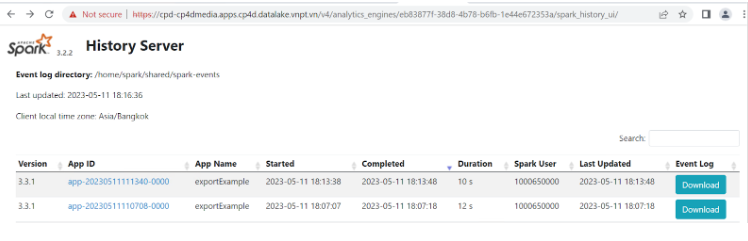
}]

}'

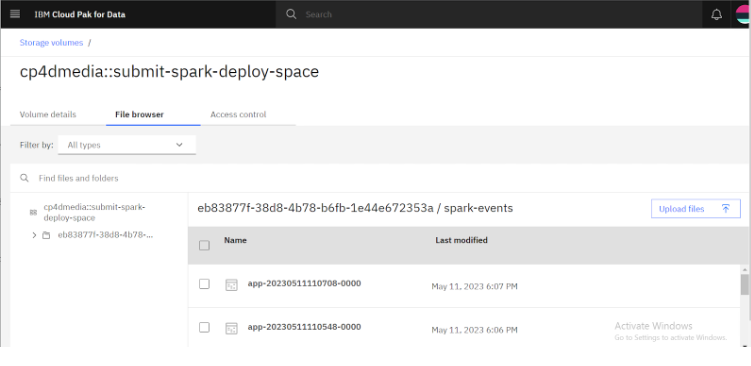
Trong giao diện Deployment space, Jobs, người dùng có thể xem driver logs của App.



Hình 44: Thông tin các App khi job được submit thành công



Hình 45: Spark History Server lưu log của mỗi App



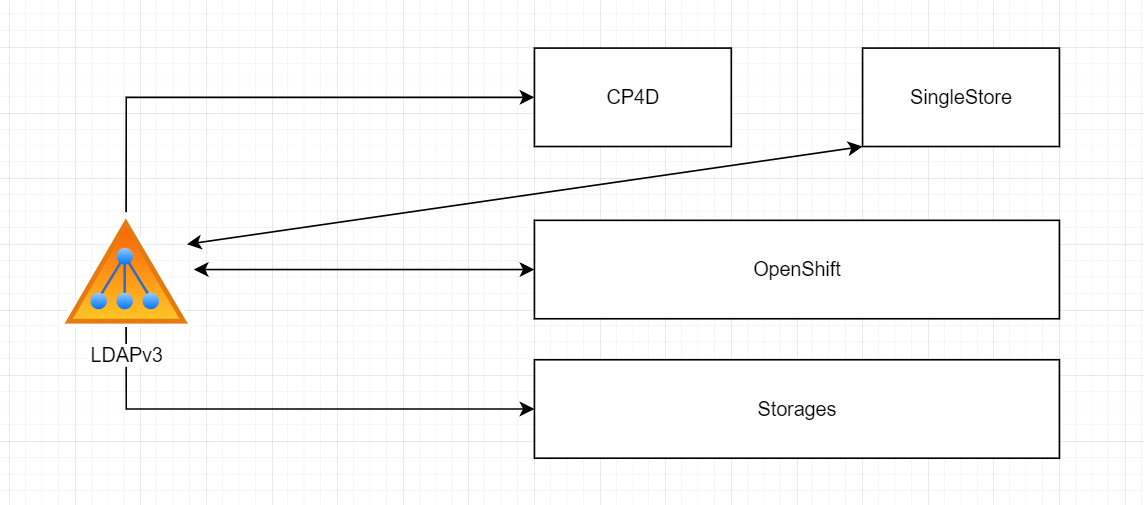
Hình 46: Storage Volume lưu event log của App

1. Quản lý và quản trị dữ liệu

Việc quản lý và quản trị dữ liệu trong Data Lake tại VNPT được triển khai bằng việc phân chia tài nguyên lưu trữ, tài nguyên tính toán, phân quyền, xử lý dữ liệu, theo dõi và quản lý chất lượng dữ liệu, sao lưu, phục hồi dữ liệu, tối ưu hóa hiệu suất. Do giới hạn về thời gian, nội dung trình bày trong khóa luận cũng như bảo mật thông tin, trong phần này sẽ chỉ trình bày về việc phân quyền và tài nguyên với user cùng với việc quản trị dữ liệu (giới hạn trong việc tạo quy tắc bảo vệ dữ liệu và áp dụng quy tắc đó để bảo vệ thông tin nhạy cảm) với Watson Knowledge Catalog.

* 1. Phân chia quyền và tài nguyên với user

Xuất phát từ yêu cầu của VNPT-Media về việc khởi tạo và thu hồi user từ một nơi duy nhất cho toàn bộ hệ thống hồ dữ liệu nên sử dụng mô hình quản lý user tập trung với LDAP server. Các thành phần khác của hệ thống như phần Storage (ESS, ICOS), Openshift, SingleStore, CP4D sẽ tích hợp với LDAP server này để xác thực. Việc phân quyền nếu đồng bộ được Groups/Roles từ LDAP server về từng hệ thống thì sẽ thực hiện mapping Groups/Roles trên hệ thống đó, nếu không đồng bộ được sẽ phải thực hiện cấu hình phân quyền thủ công tại các hệ thống. Cùng với đó là sử dụng kiến trúc multi-tenant[[11]](#footnote-11) để phân chia tài nguyên cho từng user.



Hình 47: Mô hình tích hợp LDAP với hệ thống hồ dữ liệu

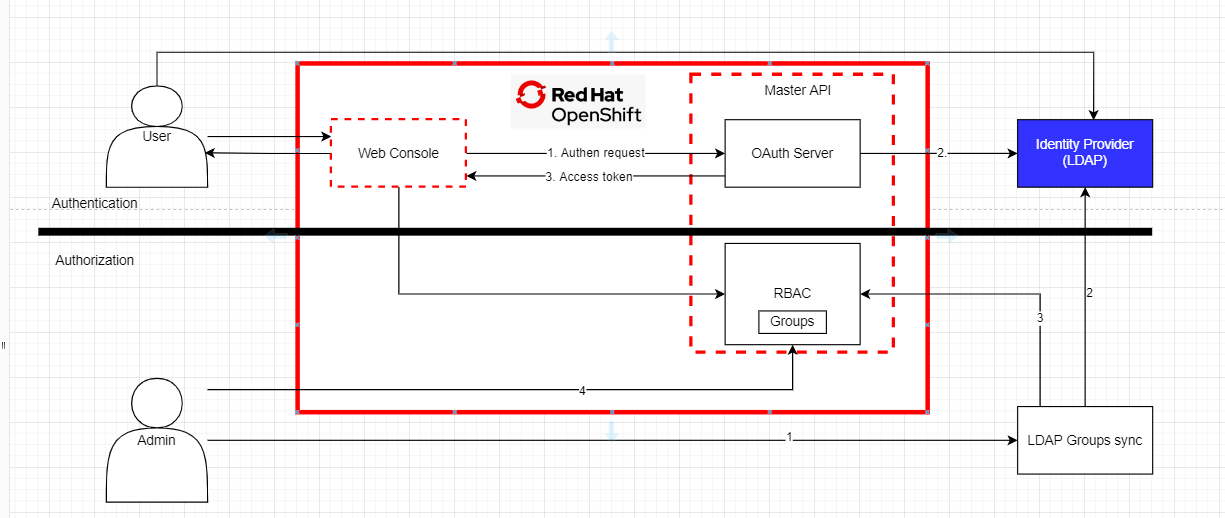
Ban đầu, thực hiện cài đặt một Active Directory/ bật LDAP server. Tiếp theo sẽ thực hiện tạo các user, groups. Theo yêu cầu của VNPT, các groups sẽ bao gồm:

Bảng 7: Các groups trong hệ thống LDAP

|  |  |
| --- | --- |
| VNPT\_Media\_Data\_Ingestion | VNPT\_NET\_Data\_Stewards |
| VNPT\_Media\_Data\_Engineer | VNPT\_NET\_Data\_Export |
| VNPT\_Media\_Data\_Analytic | VNPT\_NET\_Tenant\_Admin |
| VNPT\_Media\_Stewards | VNPT\_Vinaphone\_Data\_Ingestion |
| VNPT\_Media\_Data\_Export | VNPT\_Vinaphone\_Data\_Engineer |
| VNPT\_Media\_Tenant\_Admin | VNPT\_Vinaphone\_Data\_Analytic |
| VNPT\_IT\_Data\_Ingestion | VNPT\_Vinaphone\_Stewards |
| VNPT\_IT\_Data\_Engineer | VNPT\_Vinaphone\_Data\_Export |
| VNPT\_IT\_Data\_Analytic | VNPT\_Vinaphone\_Tenant\_Admin |
| VNPT\_IT\_Stewards | Master\_Tenant\_Data\_Ingestion |
| VNPT\_IT\_Data\_Export | Master\_Tenant\_Data\_Engineer |
| VNPT\_IT\_Tenant\_Admin | Master\_Tenant\_Data\_Analytic |
| VNPT\_NET\_Data\_Ingestion | Master\_Tenant\_Stewards |
| VNPT\_NET\_Data\_Engineer | Master\_Tenant\_Data\_Export |
| VNPT\_NET\_Data\_Analytic | Master\_Tenant\_Tenant\_Admin |

Tiếp theo sẽ cấu hình tích hợp các hệ thống (Storage, Openshift, CP4D, Singlestore) với LDAP server để thực hiện xác thực tập trung qua LDAP server. Nếu phần mềm hỗ trợ đồng bộ với LDAP thì thực hiện đồng bộ các groups từ LDAP server về các hệ thống và cấu hình xác thực, phân quyền trên phần mềm sử dụng thông tin từ hệ thống LDAP.

**Tích hợp LDAP với hệ thống OpenShift**



Hình 48: Tích hợp hệ thống LDAP với hệ thống OpenShift

Bước 1: Trên Openshift thực hiện cấu hình Identity Provider trỏ đến LDAP server

Bước 2: Trên Openshift thực hiện cấu hình đồng bộ groups với LDAP

2.1 Quản trị viên định cấu hình LDAP Group Sync để đồng bộ hóa các nhóm được chỉ định và người dùng của họ với OpenShift và định cấu hình đồng bộ hóa để chạy theo lịch trình.

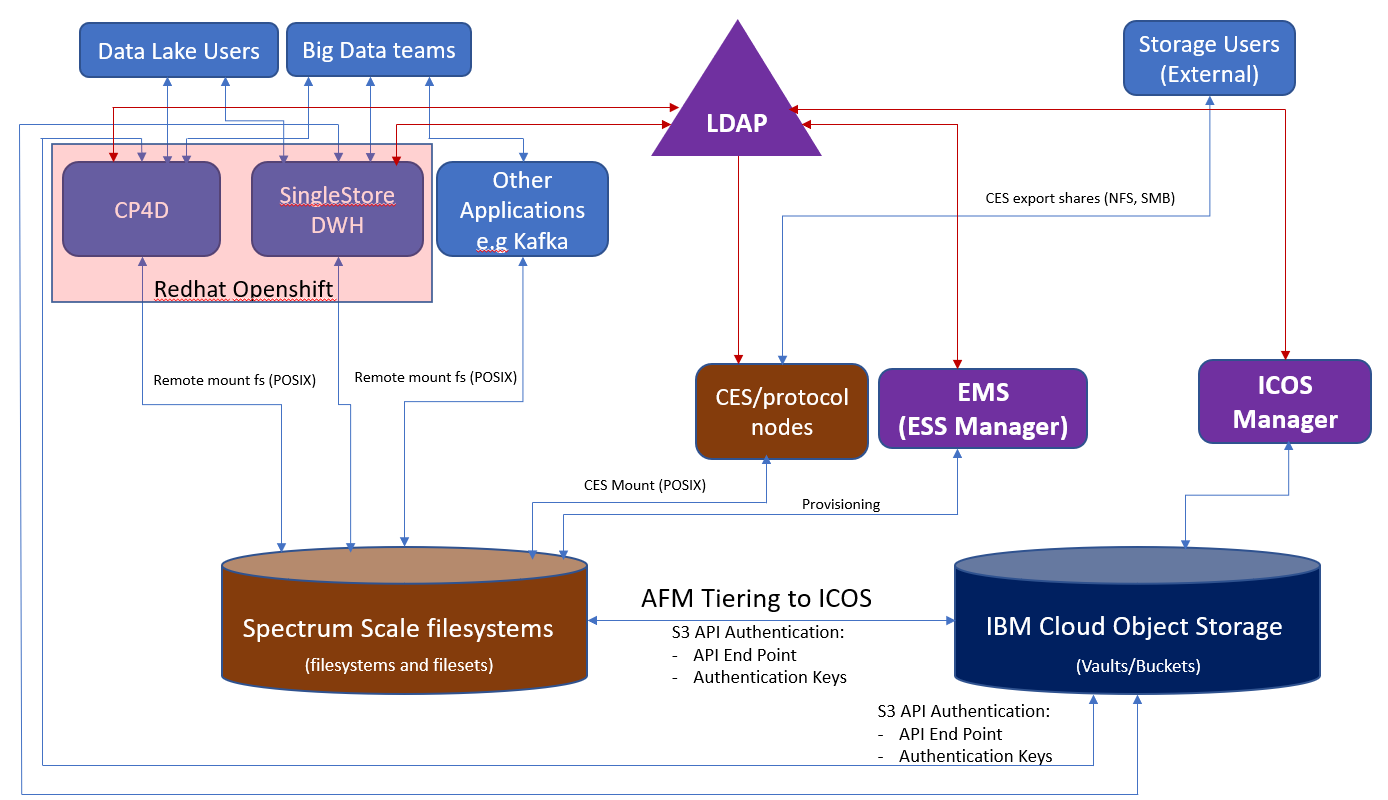
2.2 Trên mỗi LDAP Group Sync các nhóm được chỉ định và người dùng được liên kết của họ được lấy từ LDAP.

2.3 Trên mỗi LDAP Group Sync, các nhóm được chỉ định và người dùng được liên kết của họ được cập nhật trong nhóm OpenShift RBAC.

2.4 Quản trị viên áp dụng (các) Vai trò OpenShift cho các nhóm OpenShift để xác định ma trận ủy quyền.

Trên Openshift bằng việc sử dụng thành phần RBAC như rules, roles và bindings thì người quản trị có thể chỉ định và gán quyền cho user.

**Tích hợp LDAP với hệ thống Storage (ESS và ICOS)**



Hình 49: Tích hợp hệ thống Storage

Để truy cập tài nguyên lưu trữ chúng ta chia ra 2 đối tượng:

* External user (ví dụ là các đơn vị nghiệp vụ ở VNPT đẩy dữ liệu trực tiếp vào hệ thống lưu trữ thông qua các giao thức như NFS, SMB) sẽ được cấu hình riêng phân quyền thông qua CES/protocol nodes để thực hiển đẩy dữ liệu vào các phân vùng trên ESS.
* Internal user là các user trên hệ thống data lake ở đây chúng ta thấy hệ thống CP4D và Singlestore cũng như Openshift sẽ được tích hợp với LDAP Server, toàn bộ việc truy cập vào vùng lưu trữ sẽ được thực hiện ở mức ứng dụng ở đây là CP4D. Ngoài ra chúng ta cũng sẽ cấu hình tích hợp CES, EMS, ICOS với LDAP thực hiện xác thực tập trung, với yêu cầu multi tenancy Spectrum Scale sử dụng RBAC và filesets, còn ICOS sử dụng organization. Việc cấu hình auto-tiering sử dụng ILM và được cấu hình ở mức hệ thống.

**Tích hợp LDAP với hệ thống CP4D**

Nhìn từ phía ứng dụng thì người dùng sẽ truy cập và sử dụng các dịch vụ của CP4D, các dịch vụ này hoàn toàn tích hợp được với LDAP để thực hiện xác thực tập trung, việc cấu hình truy cập có thể thực hiện qua GUI cho các dịch vụ WKC, Spark, DataStage (Information server).

Đối với việc phân chia tài nguyên sử dụng cho từng user, sử dụng kiến trúc multi-tenancy. Với khả năng tạo ra nhiều instance dịch vụ khác nhau trên nền tảng CP4D như IBM DataStage, IBM Analytic Engine (Spark instance) cũng như kiến trúc lưu trữ của hệ CSDL SingleStore có thể support được mô hình multi-tenancy cho VNPT như sau:

Diagram

Description automatically generated

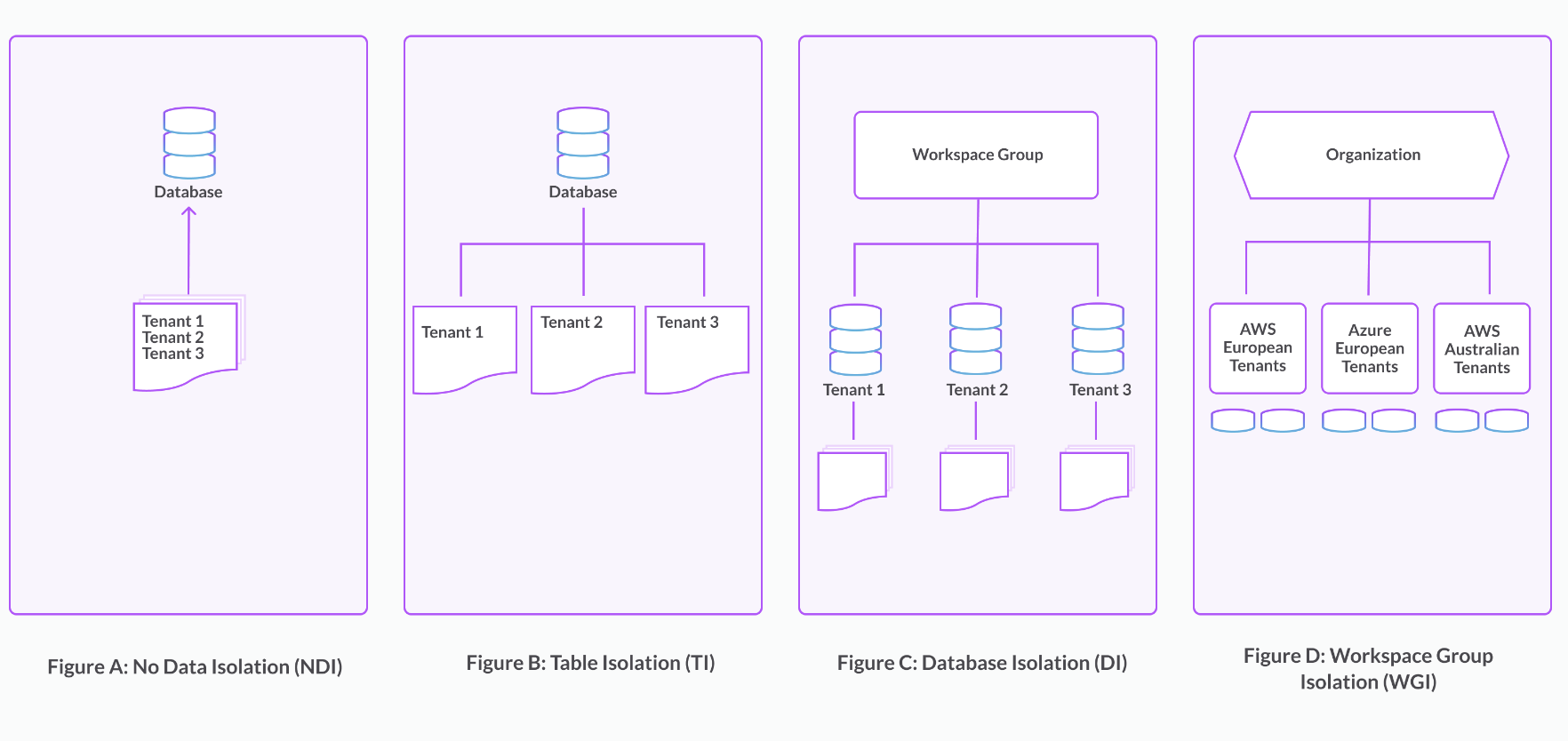
Hình 50: Kiến trúc multi-tenancy cho VNPT

Trên hệ thống CP4D, người dùng của từng đơn vị tác nghiệp trong VNPT (ví dụ: VNPT IT, VNPT Net, VNPT Media) đều có thể tự tạo các project để lưu trữ các Asset trên hệ thống CP4D cho riêng mình. Các Asset có thể tạo và lưu trữ trong một project của CP4D có thể là các kết nối đến database, các bảng dữ liệu hoặc dataset (text file, csv, v.v..), các job DataStage, các job chạy trên môi trường Spark cluster, v.v. Đối với người quản trị hệ thống có thể tạo sẵn các instance của DataStage, Analytic Engine và gán quyền truy cập cho từng instance đối với từng đối tượng người dùng cụ thể (username hoặc user group).

**Tích hợp LDAP với hệ thống SingleStore**

SingleStore có thể đồng bộ với LDAP bằng việc sử dụng công cụ sdb-admin sync-ldap để import users và groups từ LDAP server. Công cụ cũng hỗ trợ tự động đồng bộ user và groups do đó đáp ứng khả năng quản lý tập trung user ở một vị trí trung tâm. LDAP tool cung cấp đồng bộ 1 chiều từ LDAP server, phân quyền sẽ được thực hiện cấu hình trên hệ thống SingleStore.

Đối với hệ CSDL SingleStore, người quản trị có thể xây dựng hệ thống theo định hướng multi-tenancy trên mức database hoặc schema. SingleStore cho phép thiết lập nhiều mức data isolation[[12]](#footnote-12) trên hệ thống như hình bên dưới:



Hình 51: Các mức thiết lập theo định hướng multi-tenancy của SingleStore

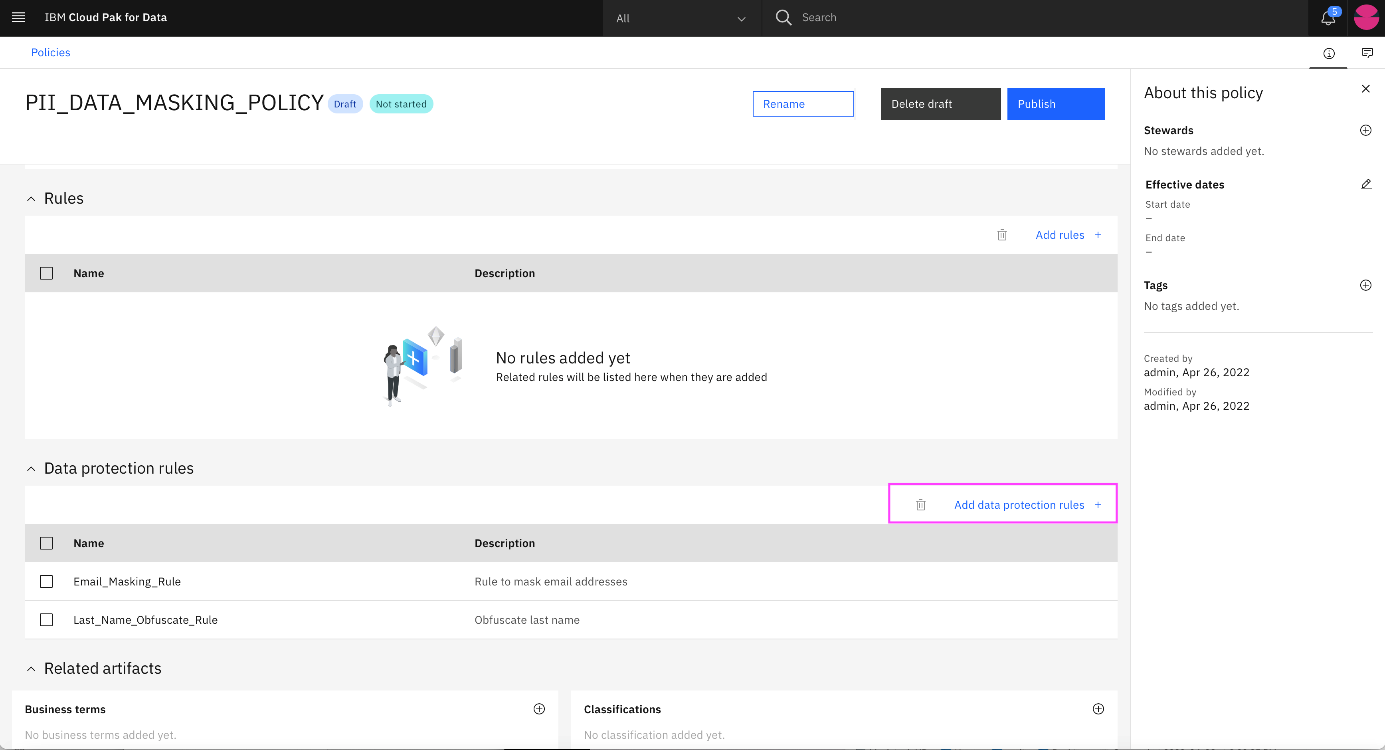
Đối với khả năng lưu trữ trong hệ CSDL, SingleStore cung cấp khả năng tiering dữ liệu giữa các lớp storage khác nhau thông qua tính năng Local Storage và Unlimited Storage. Trong đó lớp Local Storage có thể định nghĩa trên hệ thống ESS phục vụ cho việc caching và lưu trữ dữ liệu online với tốc độ truy cập cao và lớp Unlimited Storage có thể định nghĩa trên hệ thống ICOS cho việc lưu trữ dữ liệu lịch sử với tốc độ truy cập thấp hơn.

* 1. Quản trị dữ liệu với WKC

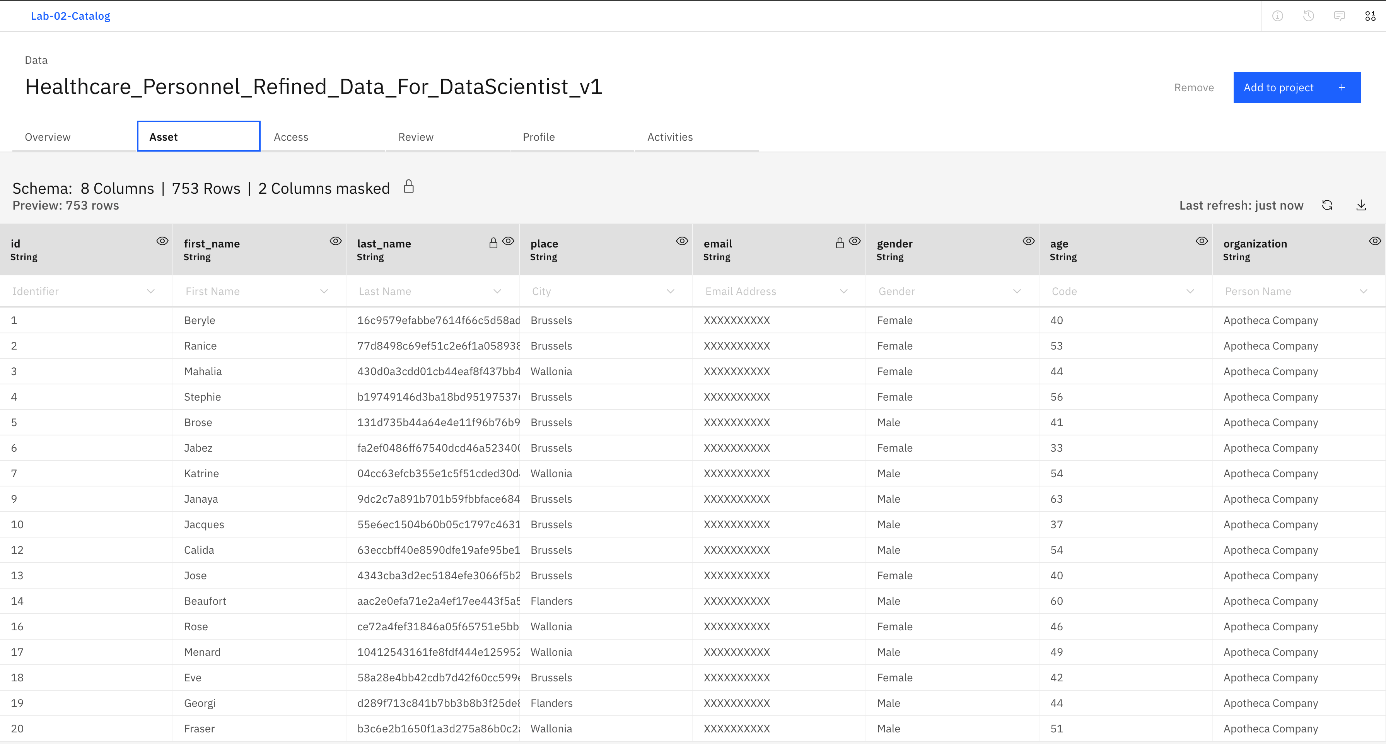
Quản trị và bảo mật dữ liệu là những thách thức lớn khi dữ liệu mở rộng trên cơ sở hạ tầng tại chỗ và đám mây. Dự án Data Lake của VNPT sử dụng WKC trong bộ CP4D để quản trị dữ liệu. WKC giúp các tổ chức cung cấp dữ liệu đáng tin cậy và có ý nghĩa bằng cách cung cấp nền tảng quản lý danh mục doanh nghiệp, an toàn được hỗ trợ bởi khung quản trị dữ liệu. Các tính năng chính của WKC là:

* Ảo hóa dữ liệu thời gian thực
* Tự động khám phá dữ liệu và tạo siêu dữ liệu
* Sử dụng học máy trích xuất bảng thuật ngữ kinh doanh từ hầu hết các thuật ngữ quy định
* Mặt nạ dữ liệu động để bảo vệ dữ liệu nhạy cảm
* Tự động quét và đánh giá rủi ro đối với dữ liệu phi cấu trúc bằng cách sử dụng Watson Catalog InstaScan

Trong dự án xây dựng hồ dữ liệu cho tập đoàn VNPT, WKC đang được triển khai ứng dụng bảo vệ dữ liệu nhạy cảm bằng cách tạo các chính sách và quy tắc bảo vệ dữ liệu. Các chính sách được sử dụng để mô tả và ghi lại các hướng dẫn, quy định, tiêu chuẩn hoặc thủ tục của VNPT nhằm đảm bảo rằng tài sản dữ liệu và thông tin được quản lý và sử dụng đúng cách. Một số ví dụ về chính sách là Thỏa thuận chia sẻ dữ liệu và xử lý dữ liệu nhạy cảm. Chính sách cần phải được xuất bản trước khi thêm các quy tắc bảo vệ dữ liệu. Quy tắc bảo vệ dữ liệu kiểm soát quyền truy cập vào nội dung. Quy tắc bảo vệ dữ liệu dựa trên tiêu chí, điều kiện và hành động mà người dùng xác định. Chúng sử dụng các thuật ngữ được xác định trước, chẳng hạn như thuật ngữ kinh doanh (business terms), lớp dữ liệu (data classes) hoặc phân loại (classifications), trong các biểu thức để xác định điều kiện. (Simon Cambridge, Lakshmana Ekambaram, Stephen D. Gawtry, Vasfi Gucer, Audrey Holloman, Frank Ketelaars, Darren King, Karen Medhat, Mark Moloney, Payal Patel, Neil Patterson, Deepak Rangarao, Mark Simmonds, Malcolm Singh, Tamara Tatian, Henry L.Quach, 2022)



Hình 52: Chính sách và các quy tắc bảo vệ dữ liệu



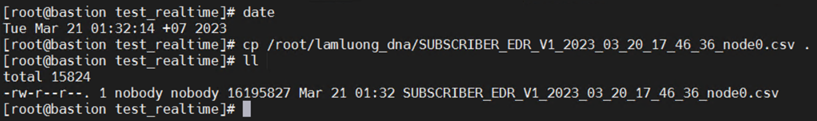
Hình 53: Dữ liệu sau khi áp dụng chính sách và quy tắc bảo vệ

1. Kết quả đạt được sau khi triển khai

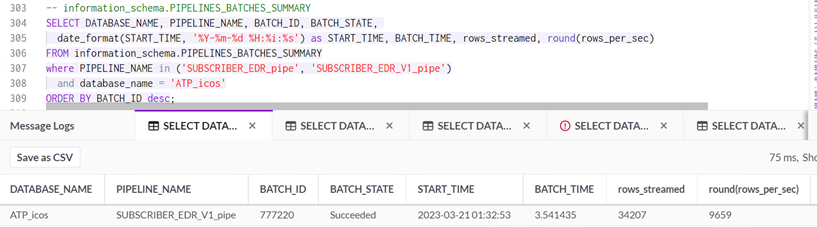
Sau khi triển khai dự án hồ dữ liệu cho Tập đoàn VNPT, đã giải quyết được hầu hết các yêu cầu của Tập đoàn, đối với Phase đầu tiên đã đáp ứng được toàn bộ các yêu cầu của phía khách hàng. Các yêu cầu chưa được đáp ứng vẫn tiếp tục được triển khai trong những Phase tiếp theo. Trong Phase thứ nhất, đã đạt được những kết quả tích cực trong cả mặt vận hành, kỹ thuật cũng như kinh doanh của VNPT.

Một, dữ liệu từ các nguồn tại VNPT Media đã được lưu trữ tập trung trong hồ dữ liệu và được phân loại với hai vùng hot và cold riêng biệt với dữ liệu trong vòng 3 năm thay vì 30 ngày như trước đây.

Hai, dữ liệu được ingest near real time trực tiếp từ nguồn với dạng streaming, ngay sau khi dữ liệu từ nguồn được đẩy vào phân vùng ingestion sau khoảng 15s tính năng afm của storage sẽ đẩy dữ liệu sang object storage, tại đây pipeline của SingleStoreDB sẽ load dữ liệu near real-time, thực hiện transform và lưu vào tables. Từ lúc xuất hiện dữ liệu trên vùng ingestion đến lúc dữ liệu xuất hiện trên table của SingleStoreDB nhỏ hơn 1 phút.

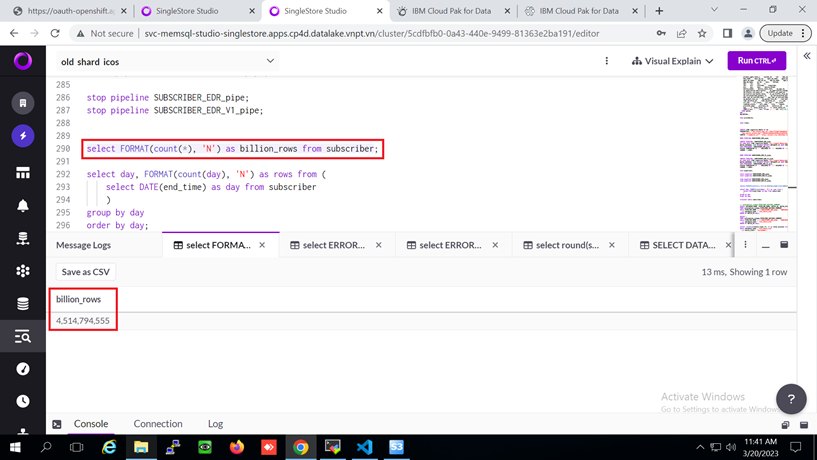


Hình 54: Dữ liệu được đưa vào vùng ingestion

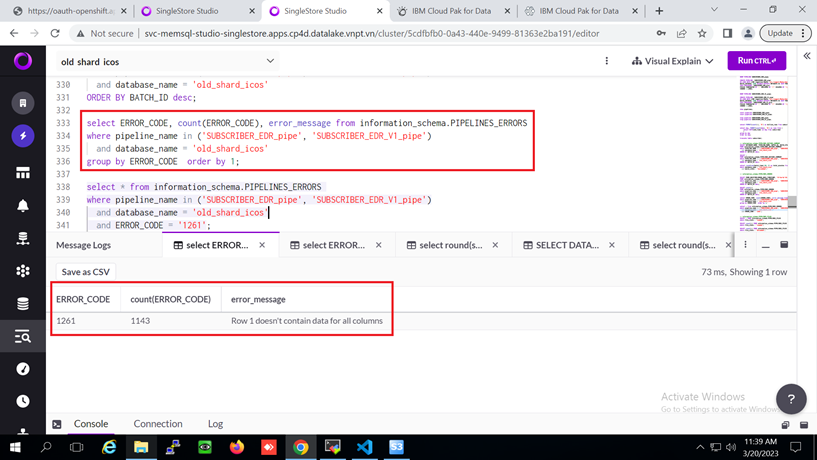


Hình 55: Dữ liệu được ingest near real time

Ba, ingest dữ liệu theo batch với các bài toán phân tích với hiệu suất cao và tỷ lệ lỗi thấp. Kiểm tra khi ingest dữ liệu từ ICOS với 17554 file và 4.514.813.252 dòng chỉ với 16.79 phút với số lỗi là 1143.

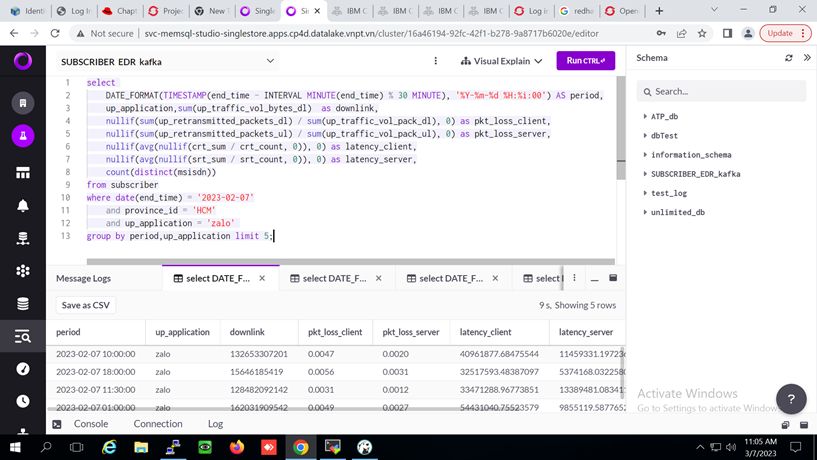


Hình 56: Ingest dữ liệu dạng batch



Hình 57: Tổng số lỗi khi ingest dữ liệu với 4.514.794.555 bản ghi

Bốn, thời gian truy vấn dữ liệu nhanh hơn. Kiểm tra với tất cả các câu truy vấn bên VNPT cung cấp, SingleStoreDB trả về kết quả nhanh hơn rất nhiều so với mong đợi của VNPT. Với câu truy vấn tính KPI cho tất cả các APP trên địa bàn Hồ Chí Minh, VNPT mong đợi kết quả trả về trong 50s, thực tế chỉ mất 9s để đạt kết quả trên.



Hình 58: Kết quả truy vấn

Năm, phân chia quyền và tài nguyên cho từng người dùng thành công. Sau khi triển khai, đã tạo được users/groups trên LDAP Server và tích hợp được với CP4D, SingleStore và Storage. Các users/groups đã được phân chia quyền cũng như tài nguyên đúng như yêu cầu của VNPT.

Sáu, giảm thiểu thời gian tạo các report và dashboard. Trước đây, mất tới 5 giờ để sinh một report hay dashboard thì sau khi triển khai thời gian đã được rút ngắn còn 30 phút. Giúp các nhà quản lý đưa ra quyết định nhanh hơn, mang lại một số tín hiệu tích cực trong kinh doanh như: Tảng tỷ lệ gia hạn các gói cước trả sau, tăng từ 5-10% ARPU[[13]](#footnote-13) (Average Revenue Per User).

Bảy, được lên kế hoạch ứng dụng data science cho các công việc như: đề xuất các gói cước tốt nhất qua tiếp thị đa kênh, trang web, app, SMS, Email, v.v.; dự đoán điểm rời bỏ của khách hàng; phân tích khách hàng 360 độ và nhiều các ứng dụng khác.

1. Kết luận chương 3

Trong chương này tác giả đã mô tả kế hoạch xây dựng hồ dữ liệu từ việc xác định thách thức, phát triển và triển khai giải pháp chiến lược, dự đoán sự tăng trưởng của dữ liệu đến việc lên kế hoạch về CSHT cũng như xác định các chiến lược hoạt động. Tiếp đó là việc triển khai thực tế tại Tập đoàn VNPT bao gồm: kiến trúc tổng quan, nguồn dữ liệu và người dùng hệ thống, các bước để thực hiện chiến lược thu thập, lưu trữ, phân tích và quản trị dữ liệu ra sao cũng đã được trình bày. Đặc biệt là cùng với lý thuyết, tác giả cũng đã trình bày cả việc thực tế cài đặt, sử dụng các thành phần trong Data Lake ra sao. Kết quả mang lại sau khi triển khai thực tế rất khả quan và đã đáp ứng được tất cả các yêu cầu của khách hàng trong Phase thứ nhất.

# **KẾT LUẬN**

Trong bài viết này, đã trình bày về tổng quan đơn vị triển khai dự án là Tập đoàn VNPT, lý thuyết liên quan đến hồ dữ liệu và hơn hết là quá trình xây dựng một hệ thống Data Lake cho Tập đoàn VNPT. Để đáp ứng nhu cầu quản lý và sử dụng dữ liệu lớn của Tập đoàn, hệ thống Data Lake được thiết kế để tích hợp và lưu trữ các nguồn dữ liệu khác nhau từ các nguồn dữ liệu đa dạng của Tập đoàn. Ngoài ra, bài khóa luận đã đưa ra các bộ công cụ được áp dụng trong khi triển khai thực tế việc lưu trữ, xử lý và quản lý dữ liệu trong hệ thống Data Lake. Về cơ bản, dự án đã thành công trong phase đầu tiên và đưa vào triển khai thực tế tại VNPT mang lại nhiều lợi ích cho Tập đoàn. Có thể thấy được các **điểm đạt được** trong bài viết như sau:

Thứ nhất, đã mô tả được những thực trạng, thách thức và yêu cầu thực tế tại tạp VNPT đối với việc xây dựng Data Lake.

Thứ hai, trình bày được những kiến thức, lý thuyết cần thiết, nền tảng cho việc triển khai xây dựng Data Lake tại Tập đoàn.

Thứ ba, đưa ra được những thông tin tổng quát nhất về kiến trúc, kết cấu, cách hoạt động cũng như lợi ích của các thành phần được sử dụng khi xây dựng Data Lake theo khung kiến trúc tham chiếu của IBM cho Data Lake.

Thứ tư, kế hoạch xây dựng được mô tả đầy đủ từ việc xác định thách thức, phát triển giải pháp, dự đoán sự tăng trưởng đến CSHT và chiến lược hoạt động. Tất cả đều được thực thi thực tế tại doanh nghiệp và mang lại hiệu quả cao.

Thứ năm, các phương án triển khai từ kiến trúc, nguồn dữ liệu, người dùng hệ thống, chiến lược thu thập, phân tích, quản trị và quản lý dữ liệu đã được trình bày chi tiết, kết hợp giải thích cách hoạt động cũng như cài đặt và sử dụng thực tế. Tất cả đã được ứng thực và kiểm tra đầy đủ, đáp ứng tất cả các bài kiểm tra của VNPT trong phase thứ nhất.

Thứ sáu, mang đến kết quả tích cực ngay sau khi triển khai. Thực tế đã cho thấy việc ứng dụng Data Lake tại VNPT đã mang lại nhiều kết quả rất tốt không chỉ về mặt cải thiện hiệu suất công việc mà còn đã có điểm sáng trong việc kinh doanh tại VNPT.

Tuy nhiên, bên cạnh những kết quả đạt được như trên, bài viết cũng còn một số điểm **hạn chế** và cũng là hướng phát triển trong thời gian tới như sau:

Chưa trình bày hết tất cả các thành phần trong dự án thực tế. Cũng do phạm vi trình bày trong một bài khóa luận, vấn đề thời gian cũng như bảo mật dữ liệu mà bài khóa luận chưa thể trình bày hết những thành phần được triển khai tại dự án trong thực tế.

Một số thành phần trình bày chưa được chi tiết và đầy đủ. Với kiến thức hạn chế cũng như chưa được tham gia từ đầu và vào tất cả các phần trong dự án nên phần trình bày của một số phần trong bài luận chưa được chi tiết và đầy đủ. Trong tương tai, đây sẽ là các phần mà em nghiên cứu và triển khai tiếp trong các phase tiếp theo.

Việc quản trị dữ liệu chưa hoàn thiện và còn nhiều thiếu sót. Trong phần quản trị dữ liệu, mới trình bày được một số thông tin cơ bản về chính sách và các quy tắc trong việc quản trị. Trong thời gian sắp tới sẽ cùng với VNPT xây dựng và hoàn thiện các thông tin để triển khai quản trị dữ liệu như: Data mapping and classification (ánh xạ và phân loại dữ liệu), Business glossary (thuật ngữ kinh doanh) và Data catalog (danh mục dữ liệu). Từ đó, đưa vào ứng dụng thực tế với WKC trong bộ CP4D.

Chưa đưa ra được các phương án cải thiện hiệu suất phân tích dữ liệu. Hiệu suất hiện tại của việc phân tích dữ liệu mới chỉ đạt được nhờ các tính năng của các công cụ sử dụng. Tuy nhiên, về phía người dùng cũng cần đưa ra các phương án sử dụng để đạt được hiệu quả tốt nhất. Đây là điểm bài viết chưa trình bày được, trong giai đoạn tiếp theo đây sẽ là một hướng phát triển và hoàn thiện hơn nữa.

# **TÀI LIỆU THAM KHẢO**

*Alibaba Clouds.* (2020). Truy cập từ https://www.alibabacloud.com/blog/data-lake-concepts-characteristics-architecture-and-case-studies\_596910

Coates, B. (2022). *Introduction to SingleStoreDB.*

*Cổng thông tin điện tử Ủy ban quản lý vốn nhà nước tại doanh nghiệp.* (n.d.). Truy cập từ http://cmsc.gov.vn

Dilmegani, C. (n.d.). *Prescriptive Analytics: Optimize Business Decisions in 2023.*

Dixons, J. (2010, 10). *Hadoop and Data Lakes.* Truy cập từ https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes

*Google Cloud.* (n.d.). Truy cập từ https://cloud.google.com/learn/what-is-object-storage

Hopkins, K. (n.d.). *Data Bricks.* Truy cập từ https://www.databricks.com/customers/att

*IBM.* (2021, February 26). Truy cập từ https://www.ibm.com/docs

*IBM.* (2021). Truy cập từ https://www.ibm.com/docs/en/iis/11.7?topic=designer-supported-connectors-stages

Oreilly. (n.d.). Truy cập từ https://learning.oreilly.com

Saurabh Gupta, V. G. (2018). *Practical Enterprise Data Lake Insights.*

Simon Cambridge, Lakshmana Ekambaram, Stephen D. Gawtry, Vasfi Gucer, Audrey Holloman, Frank Ketelaars, Darren King, Karen Medhat, Mark Moloney, Payal Patel, Neil Patterson, Deepak Rangarao, Mark Simmonds, Malcolm Singh, Tamara Tatian, Henry L.Quach. (2022). *IBM Cloud Pak for Data Version 4.5 A practical, hands-on guide with best practices, examples, use cases, and walk-throughs.*

*SingleStore.* (n.d.). Truy cập từ https://docs.singlestore.com/managed-service/en/reference/sql-reference/pipelines-commands/create-pipeline.html

*SingleStore.* (n.d.). Truy cập từ https://docs.singlestore.com/managed-service/en/load-data/load-data-with-pipelines.html

*SingleStore.* (2022). Truy cập từ https://www.singlestore.com/blog/why-singlestoredb-real-time-analytics-applications

*Talend.* (n.d.). Truy cập từ https://www.talend.com/resources/data-lake-vs-data-warehouse

*VNPT.* (n.d.). Truy cập từ https://vnpt.com.vn/gioi-thieu/su-menh-tam-nhin

*VNPT.* (n.d.). Truy cập từ https://vnpt.com.vn/gioi-thieu/cac-moc-phat-trien

VNPT. (2021). *Báo cáo tài chính Tập đoàn Bưu chính Viễn thông Việt Nam.*

*Vodafone.* (2022). Truy cập từ https://www.vodafone.com/news/technology/vodafone-cardinality-io-google-cloud-smarter-pan-european-network-performance-platform

1. công ty viễn thông đa quốc gia là một trong những công ty viễn thông lớn nhất tại Hoa Kỳ, có trụ sở tại Tháp Whitacre ở vùng trung tâm Dallas, Texas. [↑](#footnote-ref-1)
2. một công ty thông tin di động đa quốc gia Anh quốc, thuộc top những công ty viên thông viễn thông lớn nhất trên thế giới. [↑](#footnote-ref-2)
3. một công ty nghiên cứu thị trường và cung cấp số liệu thống kê từ các lĩnh vực như kinh doanh, công nghệ, xã hội và nhiều lĩnh vực khác. [↑](#footnote-ref-3)
4. Các thiết bị có khả năng xử lý, lưu trữ và truyền tải dữ liệu và thông tin tại nguồn thu thập thông tin. Thường được đặt tại các vị trí gần với người dùng hoặc các thiết bị cảm biến, các thiết bị này thường được trang bị tính năng xử lý dữ liệu và truyền tải thông tin với tốc độ cao và độ trễ thấp. [↑](#footnote-ref-4)
5. HTAP là một kiến trúc hỗ hợp xử lý dữ liệu trực tuyến (OLTP) và phân tích dữ liệu trực tuyến (OLAP) đồng thời. [↑](#footnote-ref-5)
6. Blob store là một dịch vụ lưu trữ đối tượng trên đám mây, cho phép lưu trữ và quản lý các tệp nhị phân. Blob store được tối ưu hóa để lưu trữ và phân phối các tệp lớn và cho phép truy cập trực tiếp từ các ứng dụng khác nhau trên đám mây. [↑](#footnote-ref-6)
7. những thông tin giá trị và sâu sắc mà ta rút ra từ quá trình phân tích dữ liệu. [↑](#footnote-ref-7)
8. Rack là thiết bị dùng để lưu trữ, quản lý và bảo vệ các thiết bị và hệ thống kết nối mạng, như máy chủ, thiết bị lưu trữ dữ liệu, switch mạng, bộ lưu điện (UPS), và các thiết bị khác. [↑](#footnote-ref-8)
9. Out-of-band tách biệt quản lý và kiểm soát hệ thống với lưu lượng mạng chính. [↑](#footnote-ref-9)
10. thuật ngữ chỉ tới việc sử dụng một máy chủ (server) vật lý để chạy một hệ thống phần mềm hoặc ứng dụng, thay vì sử dụng một hệ thống ảo hóa hoặc một môi trường điều hành được cài đặt sẵn trên máy chủ. [↑](#footnote-ref-10)
11. Kiến trúc phần mềm cho phép user sử dụng chung một ứng dụng hoặc hệ thống, mà vẫn được phân biệt và bảo mật dữ liệu. Mỗi user sẽ có một phạm vi dữ liệu, quyền truy cập, quyền sử dụng và cấu hình riêng biệt. [↑](#footnote-ref-11)
12. là một khái niệm trong cơ sở dữ liệu để đảm bảo rằng các giao dịch và truy cập dữ liệu diễn ra một cách độc lập và an toàn. [↑](#footnote-ref-12)
13. Doanh thu trung bình trên một khách hàng là giá trị dùng để đánh giá sự thành công của một mạng viễn thông di động, hay còn gọi là chỉ số doanh thu bình quân của một thuê bao/tháng; là chỉ số dùng để đánh giá tính hiệu quả của các nhà khai thác, cung cấp dịch vụ viễn thông. [↑](#footnote-ref-13)