Xây dựng nền tảng kiến trúc dữ liệu

Đoàn Thanh Tám

Giới thiệu

Đoàn Thanh Tám
tamdt9@viettel.com.vn

- 15+ năm kinh nghiệm trong lĩnh vực CNTT, đặc biệt là các hệ thống Big Data, DWH, Data Lake, ...
- Chứng chỉ chuyên môn quốc tế về CNTT: Cloudera, Oracle, Google, TDWI, ...
- Giải thưởng trong nước/ quốc tế: Sao Khuê, IT World Award, CEM Excellence Awards
- Vai trò: Senior SE, Tech Lead, PM, PGĐ TT CNTT, GĐ TT PTDL, PB QTDL TĐ, PB CNTT TĐ
- Admin nhóm Fb Cộng đồng Big Data Việt Nam
- Tác giả/ đồng tác giả của 10+ bài báo quốc tế về Machine Learning, Deep Learning

Nội dung

1. Quyết định dựa trên dữ liệu
2. Giới thiệu kiến trúc Lakehouse
3. Ứng dụng Big Data tại Viettel

Quyết định dựa trên dữ liệu

Ra quyết định dựa trên dữ liệu

Ra quyết định nhanh và chính xác hơn

Dữ liệu chính xác và kịp thời sẽ giúp các doanh nghiệp, tổ chức tạo ra các quyết định nhanh hơn và tốt hơn

Giảm chi phí

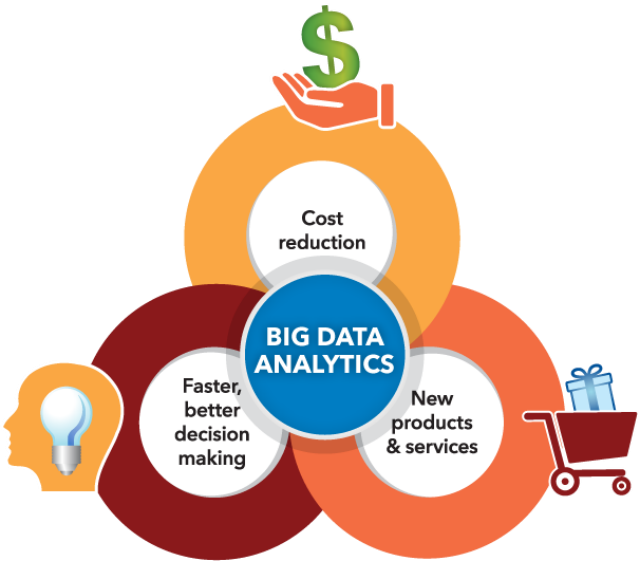
Bằng việc ứng dụng các thuật toán học máy, học sâu vào trong hoạt động vận hành, tối ưu sản xuất kinh doanh, BDA có thể giúp giảm chi phí vận hành

Launching sản phẩm và dịch vụ mới

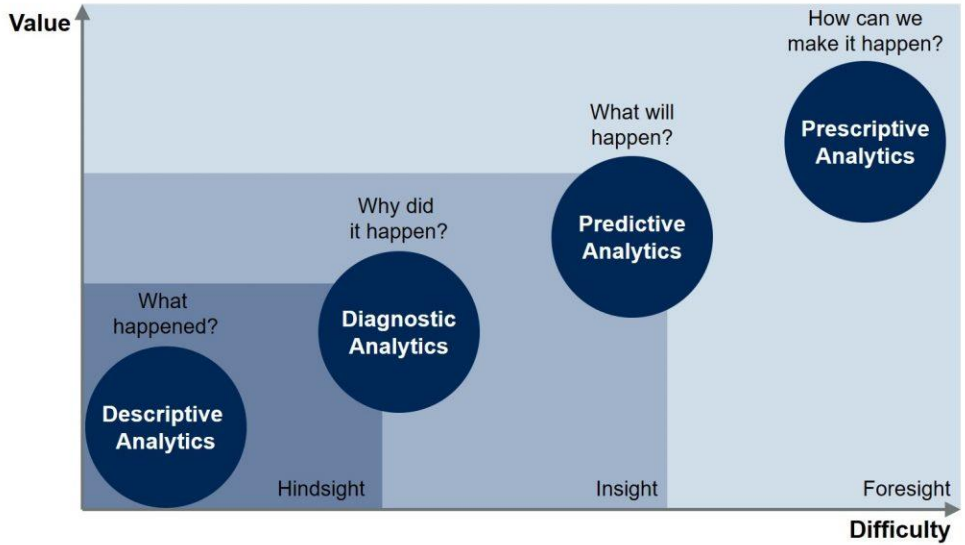
Bằng việc sử dụng dữ liệu về hành vi, sở thích, mong muốn của khách hàng, BDA giúp các doanh nghiệp, tổ chức đưa ra các sản phẩm dịch vụ phù hợp với khách hàng hơn

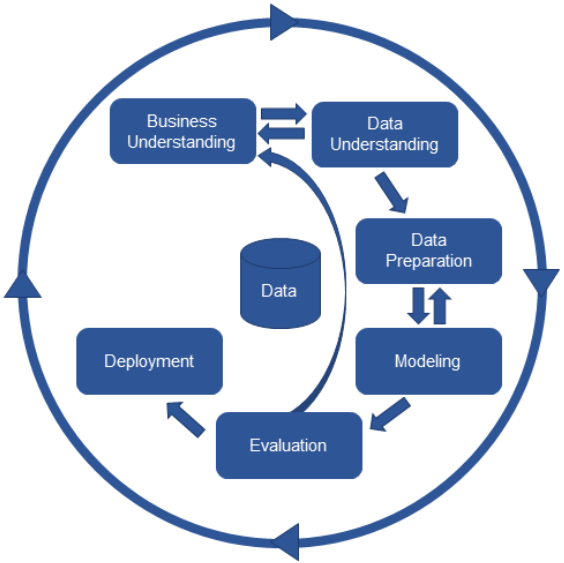
Tăng trải nghiệm khách hàng

Cải thiện chất lượng dịch vụ, tiết kiệm chi phí dựa trên dữ liệu khách hàng trong quá khứ giúp tăng trải nghiệm và kết nối hơn với khách hàng



Các cấp độ PTDL & Quy trình PTDL





(1) https://blogs.gartner.com/jason-mcnellis/2019/11/05/youre-likely-investing-lot-marketing-analytics-getting-right-insights/

(2) https://www.ibm.com/docs/sr/spss-modeler/saas?topic=dm-crisp-help-overview

Ứng dụng Big Data trong Viễn thông

Viễn thông (Telecomunications): một trong số những ngành có lượng dữ liệu khổng lồ từ người dùng dịch vụ, với sự bùng nổ của smartphone trong những năm gần đây

Tối ưu mạng lưới

Xác định các khu vực, vị trí có dung lượng vượt mức và định tuyến lại băng thông khi cần thiết nhằm đảm bảo chất lượng dịch vụ, tăng mức độ hài lòng của khách hàng

Giữ chân khách hàng

Phân tích dữ liệu về chất lượng dịch vụ (cuộc gọi nghe có rõ ràng không, người dùng xem video trên mạng di động có bị giật không, ...), sự tiện lợi của dịch vụ, mức độ tiêu dùng của khách hàng dành cho viễn thông, ...

Khuyến nghị sản phẩm

Các nhà mạng có thể dựa trên sở thích, hành vi của khách hàng để xây dựng thiết kế các sản phẩm có các đặc tính phù hợp với mối quan tâm của từng lớp khách hàng

https://www.oracle.com/a/ocom/docs/top-22-use-cases-for-big-data.pdf

Ứng dụng Big Data trong Tài chính

Khối lượng dữ liệu của các tổ chức tài chính, ngân hàng không thực sự lớn như dữ liệu viễn thông tuy nhiên dữ liệu của các tổ chức tài chính lại có chất lượng tốt và có mức độ tin tưởng cao hơn (chữ V thứ 5 trong đặc tính 6Vs của Big Data – Veracity)

Phát hiện gian lận

Bằng việc sử dụng Big Data, các công ty/ tổ chức tài chính có thể xác định được các mẫu (pattern) biểu thị hành vi gian lận trong giao dịch tài chính từ đó áp dụng các biện pháp cần thiết để chống lại các hành vi gian lận này

Quản lý rủi ro

Hệ thống quản lý rủi ro dựa trên Big Data để phát hiện rủi ro, gian lận tiềm ẩn trong thời gian thực sẽ giúp các tổ chức tài chính, ngân hàng không phải chịu tổn thất về doanh thu

Cá nhân hóa trải nghiệm khách hàng

Dựa trên dữ liệu lớn và các mô hình học máy để dự đoán mức độ hài lòng của khách hàng, chủ động thực hiện các chiến dịch chăm sóc khách hàng một cách chủ động để tăng trải nghiệm khách hàng

Ứng dụng Big Data trong Y tế

Y tế và chăm sóc sức khỏe là lĩnh vực có nguồn dữ liệu khổng lồ. Sử dụng Big Data, các tổ chức y tế và chăm sóc sức khỏe đã có thể dự đoán được các xu hướng bệnh, phát hiện sớm các biểu hiện các bệnh hiểm nghèo cũng như cung cấp cho bệnh nhân các dịch vụ y tế và chăm sóc sức khỏe tốt hơn

Nghiên cứu gen

Thông qua sử dụng lượng dữ liệu khổng lồ từ các nguồn khác nhau, các nhà nghiên cứu có thể xác định được các gen bệnh cũng như các dấu hiệu lâm sàng để giúp bệnh nhân xác định được chính xác các vấn đề sức khỏe mà họ có thể gặp phải trong tương lai

Dự đoán bệnh nhân

Bằng cách thu thập dữ liệu đầy đủ dữ liệu quá khứ thông qua các công nghệ Big Data, các tổ chức y tế và chăm sóc sức khỏe có thể dự báo được số lượng bệnh nhân có thể có trong tương lai thông qua các kỹ thuật "time series analysis"

Hỗ trợ (Personal Health Assistants)

Các tổ chức y tế và chăm sóc sức khỏe cũng có thể sử dụng Big Data để cung cấp các cách điều trị tốt hơn và cải thiện chất lượng chăm sóc bệnh nhân mà không làm tăng chi phí thông qua những hệ thống như Patient 360 để cung cấp một cái nhìn đa chiều của từng bệnh nhân

Ứng dụng Big Data trong Giáo dục

Big Data cho phép các nhà trường, tổ chức giáo dục cơ hội tích hợp các hệ thống, ứng dụng và nền tảng quan trọng giúp tăng hiệu quả đào tạo.

Cá nhân hóa chương trình đào tạo

Big Data giúp thiết kế và khuyến nghị chương trình đào tạo, lộ trình học tập phù hợp với từng học viên

Tăng hiệu quả quản lý học tập

Dữ liệu thời gian người học, kết quả đánh giá học tập giúp tùy chỉnh chương trình giảng dạy một cách hiệu quả. Big Data cũng giúp phân loại học viên theo sở thích, phân nhóm học viên hiệu quả vào các hoạt động nhóm.

Dự đoán hiệu quả học tập

BDA giúp theo dõi những học viên cần được chú ý nhiều hơn trong một chương trình cụ thể, giúp có thể tác động đến kết quả học tập của học viên bằng cách phát triển những kỹ năng mà họ còn yếu kém.

Ứng dụng Big Data trong Bán lẻ

Các công nghệ Big Data được sử dụng để giúp các đơn vị, tổ chức tối ưu chi phí, tăng trải nghiệm khách hàng, cũng như phát triển sản phẩm

Phát triển sản phẩm

Big Data có thể giúp các doanh nghiệp bán lẻ dự đoán nhu cầu của khách hàng thông qua việc phân loại các thuộc tính chính của các sản phẩm mà khách hàng quan tâm dựa trên doanh số bán được, hành vi mua của khách hàng trong quá khứ

Tối ưu giá

Big Data có thể giúp các nhà bán lẻ tìm ra điểm cân bằng về giá với từng phân khúc khách hàng nhằm tối đa lợi nhuận của các nhà bán lẻ, đồng thời cũng không làm khách hàng rời bỏ sản phẩm vì giá cao

Tăng trải nghiệm khách hàng

Các doanh nghiệp bán lẻ cũng có thể sử dụng Big Data để cung cấp cho khách hàng các trải nghiệm tốt hơn trong hành trình khách hàng (Customer Jouney) thông qua Customer 360

Các ứng dụng Big Data phổ biến

Tối ưu vận hành

Giảm chi phí vận hành thông qua các use case tối ưu chi phí nhằm tối đa hóa lợi nhuận

Phát hiện gian lận

Xác định được các mẫu (pattern) biểu thị hành vi gian lận từ đó áp dụng các biện pháp cần thiết để chống lại các hành vi gian lận này

Mở rộng/ giữ chân khách hàng

Ứng dụng Big Data và PTDL có thể giúp tổ chức/ doanh nghiệp duy trì tập khách hàng cũng như phát triển khách hàng mới

Quản lý rủi ro

Phát hiện rủi ro, gian lận tiềm ẩn trong thời gian thực sẽ giúp các tổ chức tài chính, ngân hàng không phải chịu tổn thất về doanh thu

Tối ưu doanh thu/ chi phí

Tìm đúng khách hàng, offer đúng sản phẩm và đúng thời điểm nhằm tăng doanh thu cho doanh nghiệp. Ngoài ra, doanh nghiệp có thể sử dụng tài sản dữ liệu để tạo ra các nguồn doanh thu mới

Tăng trải nghiệm khách hàng

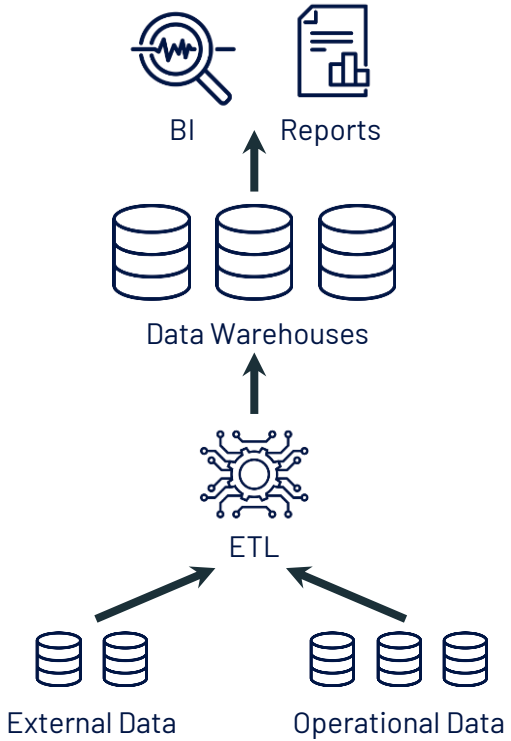
Các doanh nghiệp có thể sử dụng Big Data và PTDL để cung cấp cho khách hàng các trải nghiệm tốt hơn trong hành trình khách hàng (Customer Jouney) thông qua Customer 360

Giới thiệu kiến trúc Lakehouse

Data Warehouse

Được xây dựng với mục đích BI và báo cáo. Hạn chế:

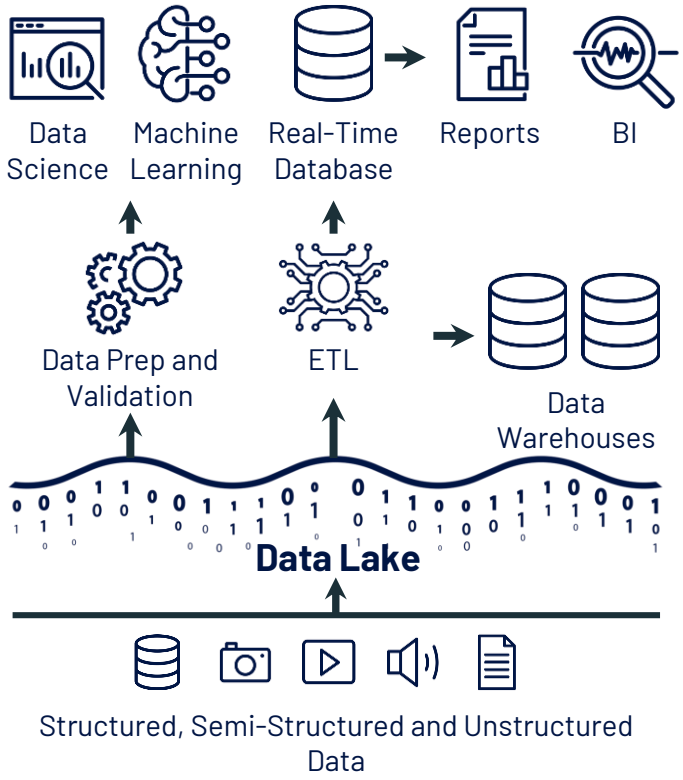
- Không hỗ trợ video, audio, text
- Không hỗ trợ data science, ML
- Không hỗ trợ realtime streaming
- Định dạng đóng & xác định trước



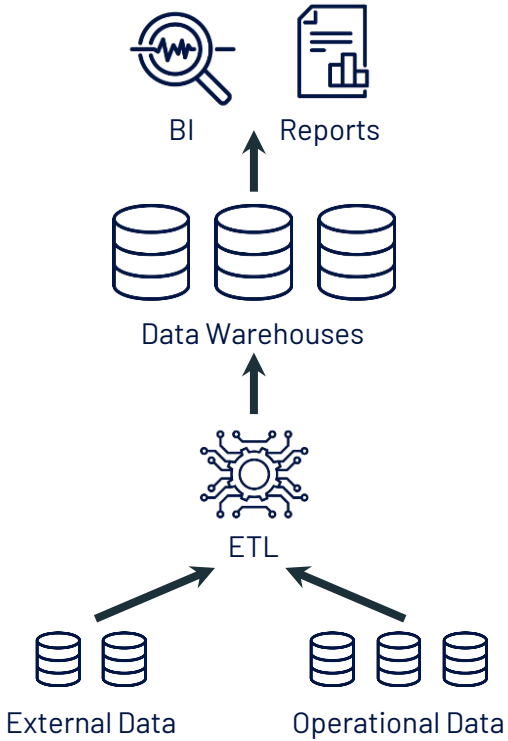
Data Lake

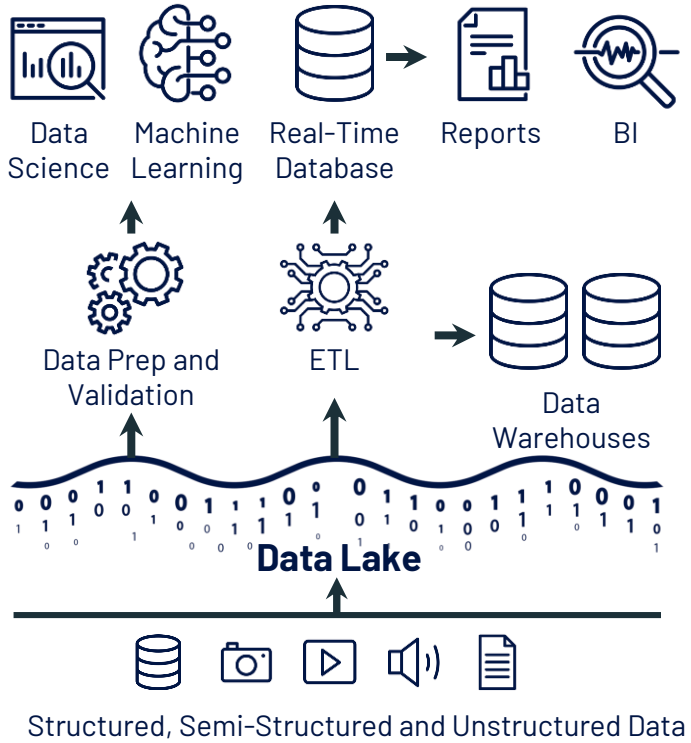
Có thể lữu trữ toàn bộ dữ liệu của tổ chức/ doanh nghiệp phục vụ các yêu cầu trong tương lai. Tuy nhiên:

- Hạn chế hỗ trợ BI
- Cài đặt phức tạp
- Hiệu năng không cao
- Dữ liệu chưa đáng tin cậy

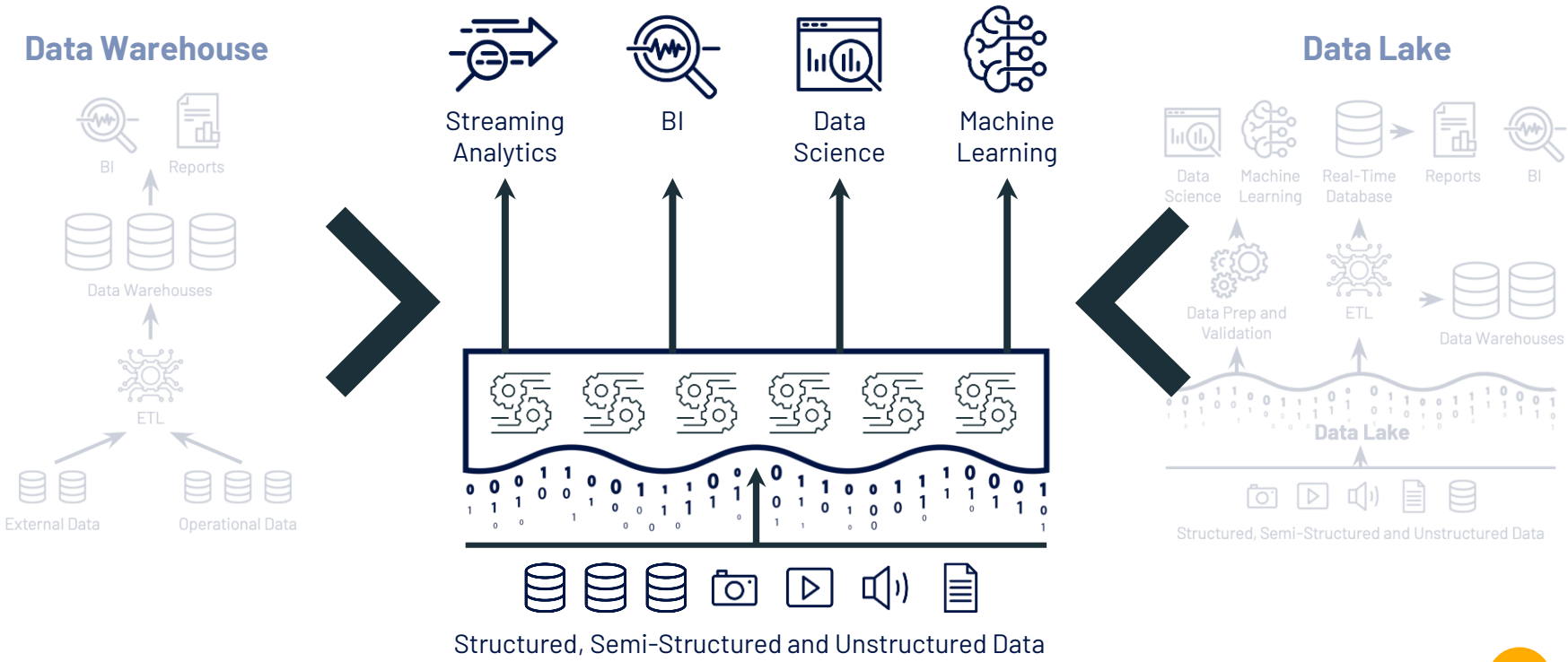


Data Platform = DWH + DL





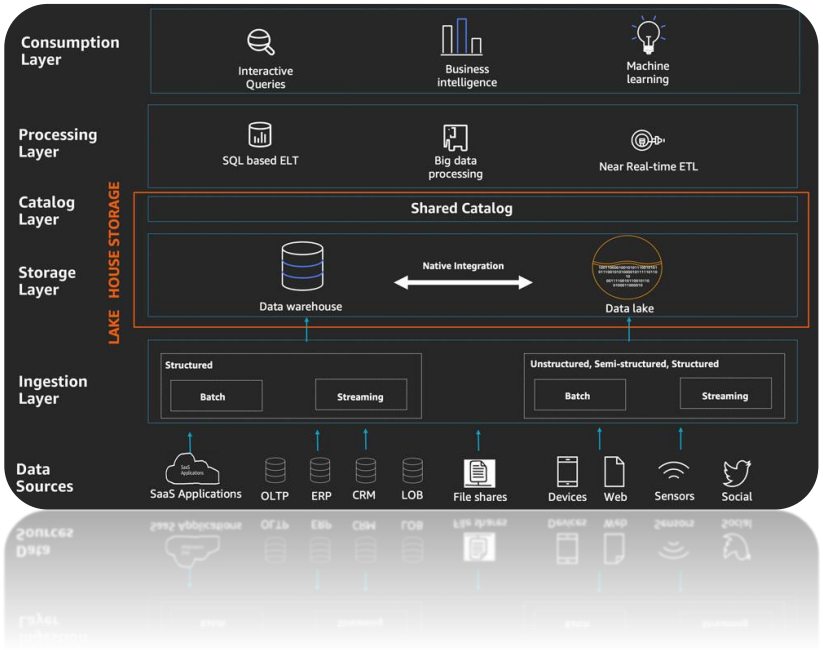
Kiến trúc Lakehouse



Kiến trúc Lakehouse

Các đặc điểm chính của Lakehouse:

- Hỗ trợ các định dạng và loại dữ liệu khác nhau
- Đảm bảo tính tin cậy và nhất quán của dữ liệu
- Hỗ trợ tính chất công việc dữ liệu khác nhau (BI, DS, ML, DA)
- Khả năng sử dụng các công cụ BI trực tiếp trên dữ liệu nguồn



Xây dựng nền tảng dữ liệu Lakehouse

Các thành phần chính để xây dựng một nền tảng dữ liệu theo kiến trúc Lakehouse

1. Kho dữ liệu (Data Lake) (lưu trữ trên cloud, định dạng mở, blob)
2. Engine để truy vấn dữ liệu lưu trữ trong Kho dữ liệu
3. Lớp transaction để đảm bảo tính nhất quán của dữ liệu
4. Workflow để xử lý dữ liệu (ETL, Data Pipeline)
5. Lớp bảo mật (Security), quản trị dữ liệu (Data Governance)
6. Công cụ hỗ trợ các tác vụ khác nhau:
    a. Truy vấn SQL
    b. Các công cụ BI và dashboards
    c. Các chương trình ML, DL, AI
    d. Dữ liệu thời gian thực realtime, streaming

Ưu điểm của Data Lake

Chi phí lưu trữ rẻ
Chi phí lưu trữ rẻ và có thể mở rộng dễ dàng.

Lưu trữ các định dạng dữ liệu khác nhau
Video, audio, text, structured, unstructured

Định dạng lưu trữ mở
Sử dụng định dạng Parquet, rất nhiều công cụ hỗ trợ

Các thách thức đối với Data Lake

1. Khó bổ sung dữ liệu
Dữ liệu được bổ sung sẽ không được cập nhật đồng nhất

2. Khó cập nhật DL đang có
Việc cập nhật thông tin, xóa dữ liệu đang có sẽ tốn chi phí và tài nguyên

3. Job bị lỗi giữa chừng
Cơ chế kiểm soát dữ liệu khi job bị lỗi chưa tốt

4. Vận hành Real-time
Sử dụng cả streaming và batch dẫn tới sự bất đồng nhất về dữ liệu

5. Chi phí quản lý phiên bản data lớn
Chi phí để quản lý việc thay đổi trong dữ liệu là rất lớn

6. Khó quản lý metadata lớn
Với các kho dữ liệu lớn, rất khó để quản lý siêu dữ liệu trong toàn bộ kho dữ liệu

7. Quá nhiều file nhỏ
Data Lake xử lý không tốt trong trường hợp có quá nhiều file nhỏ

8. Chất lượng dữ liệu
Chất lượng của dữ liệu không được đảm bảo trong Data Lake

1. Khó bổ sung dữ liệu
2. Khó cập nhật DL đang có
3. Job bị lỗi giữa chừng
4. Vận hành Real-time
5. Chi phí quản lý phiên bản data lớn
6. Khó quản lý metadata lớn
7. Quá nhiều file nhỏ
8. Chất lượng dữ liệu

ACID Transactions

Áp dụng cơ chế giao dịch ACID

- Cập nhật khi đảm bảo đầy đủ các điều kiện ràng buộc

```
/path/to/table/_delta_log
 - 0000.json   { Add file1.parquet
                 Add file2.parquet
                 ...
 - 0001.json
 - 0002.json
 - ...
 - 0010.parquet
```

1. Khó bổ sung dữ liệu
2. Khó cập nhật DL đang có
3. Job bị lỗi giữa chừng
4. Vận hành Real-time
5. Chi phí quản lý phiên bản data lớn
6. Khó quản lý metadata lớn
7. Quá nhiều file nhỏ
8. Chất lượng dữ liệu

ACID Transactions

Áp dụng cơ chế giao dịch ACID

- Cập nhật khi đảm bảo đầy đủ các điều kiện ràng buộc

Xem lịch sử giao dịch

- Tất cả các transactions được lưu lại, cho phép xem lịch sử cập nhật dữ liệu

```
SELECT * FROM events
TIMESTAMP AS OF ...

SELECT * FROM events
VERSION AS OF ...
```

1. Khó bổ sung dữ liệu
2. Khó cập nhật DL đang có
3. Job bị lỗi giữa chừng
4. Vận hành Real-time
5. Chi phí quản lý phiên bản data lớn
6. Khó quản lý metadata lớn
7. Quá nhiều file nhỏ
8. Chất lượng dữ liệu

Sử dụng Spark

- Spark được xây dựng để xử lý khối lượng rất lớn dữ liệu
- Meta-Data được lưu dưới định dạng Parquet
- Dữ liệu được phân nhỏ, lưu bộ đệm, tối ưu cho việc truy xuất nhanh
- Dữ liệu và siêu dữ liệu gắn với dữ liệu luôn tồn tại song2, không cần việc đồng bộ

1. Khó bổ sung dữ liệu
2. Khó cập nhật DL đang có
3. Job bị lỗi giữa chừng
4. Vận hành Real-time
5. Chi phí quản lý phiên bản data lớn
6. Khó quản lý metadata lớn
7. Quá nhiều file nhỏ
8. Chất lượng dữ liệu

Hợp nhất tệp

Tự động tối ưu tổ chức tệp, cho phép truy cập nhanh

- Phân vùng: đối với các truy vấn đặc thù
- Bỏ bớt các tệp không cần thiết
- Tổ chức tối ưu truy vấn nhiều cột

```
OPTIMIZE events
ZORDER BY (eventType)
```

1. Khó bổ sung dữ liệu
2. Khó cập nhật DL đang có
3. Job bị lỗi giữa chừng
4. Vận hành Real-time
5. Chi phí quản lý phiên bản data lớn
6. Khó quản lý metadata lớn
7. Quá nhiều file nhỏ
8. Chất lượng dữ liệu

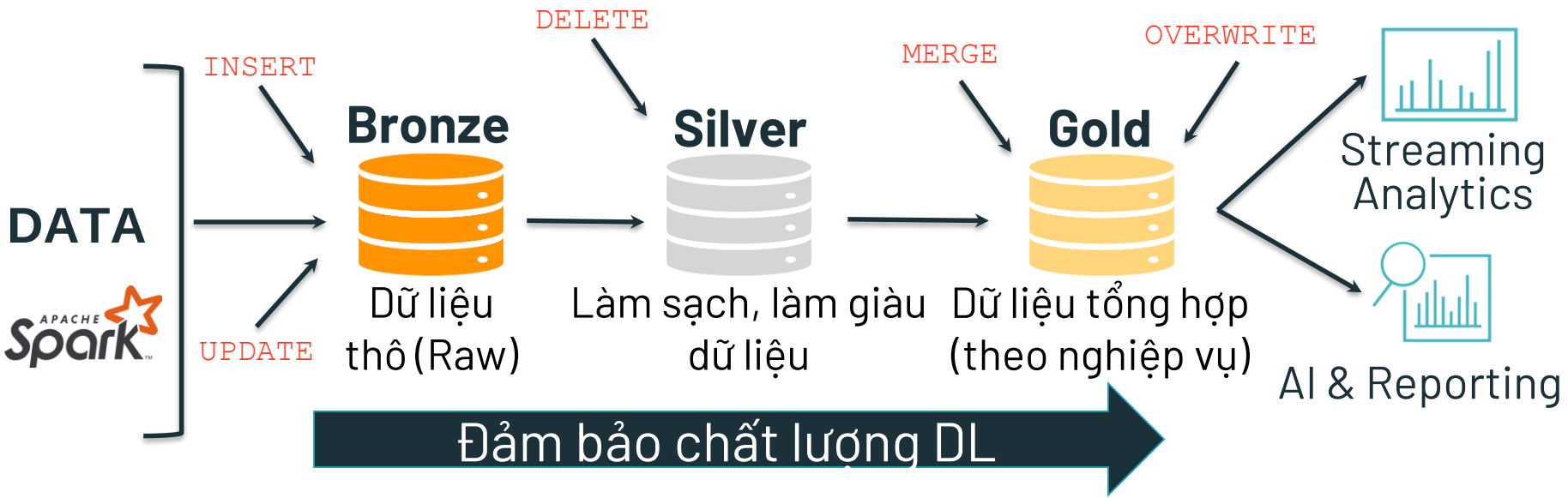
Sử dụng schema

Sử dụng schema để kiểm tra và đảm bảo chất lượng dữ liệu

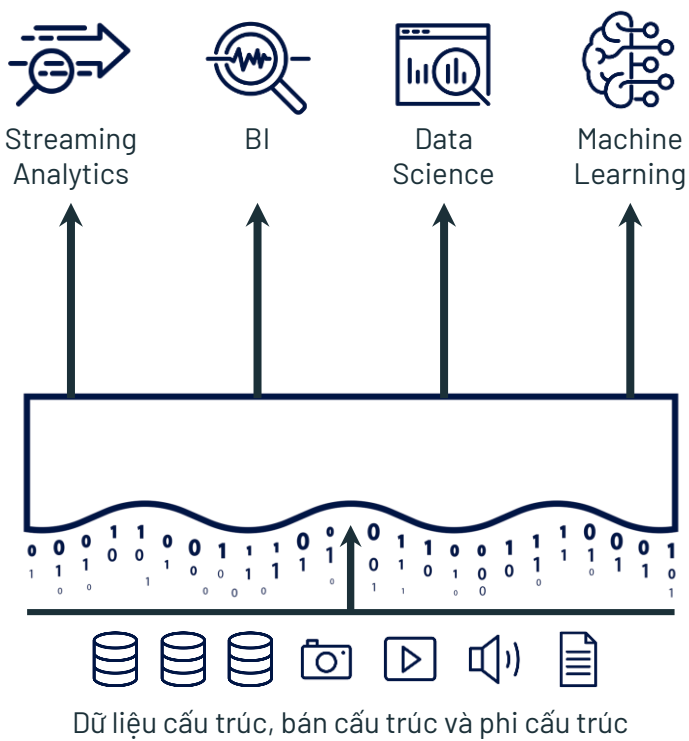
- Tất cả dữ liệu trong các bảng phải tuân thủ nghiêm các luật nghiệp vụ
- Bao gồm cả các tác vụ thay đổi schema trong quá trình merge, tác động dữ liệu

```
MERGE INTO events
USING changes
ON events.id = changes.id
WHEN MATCHED THEN
  UPDATE SET *
WHEN NOT MATCHED THEN
  INSERT *
```

Luồng dữ liệu



Lớp lưu trữ



Sử dụng 1 nền tảng dữ liệu cho tất cả các use-case

Lớp giao dịch theo cấu trúc

Data Lake lưu trữ tất cả dữu liệu

Ứng dụng Big Data tại Viettel

Ứng dụng Big Data

Up/ Cross-sell

Mua gói tốt hơn **1.4- 5.4 lần**
Doanh thu & ARPU tốt hơn **3.6 lần**

Churn Prediction

Churn rate giảm **54%**

Recharge

Tỷ lệ nạp thẻ tốt hơn **2.1- 3 lần**

Pre to Postpaid

Tỷ lệ chuyển đổi tăng **~2.6 lần**

Nâng cao lòng trung thành KH

Nâng tỷ lệ duy trì từ **70% lên 85%**

Reactivate KH

35% KH quay lại dịch vụ

TELCOS

- Recommend ưu đãi
- Tìm kiếm KH tiềm năng
- Dự đoán nguy cơ lỗi, hỏng TB
- Dự đoán lưu lượng data

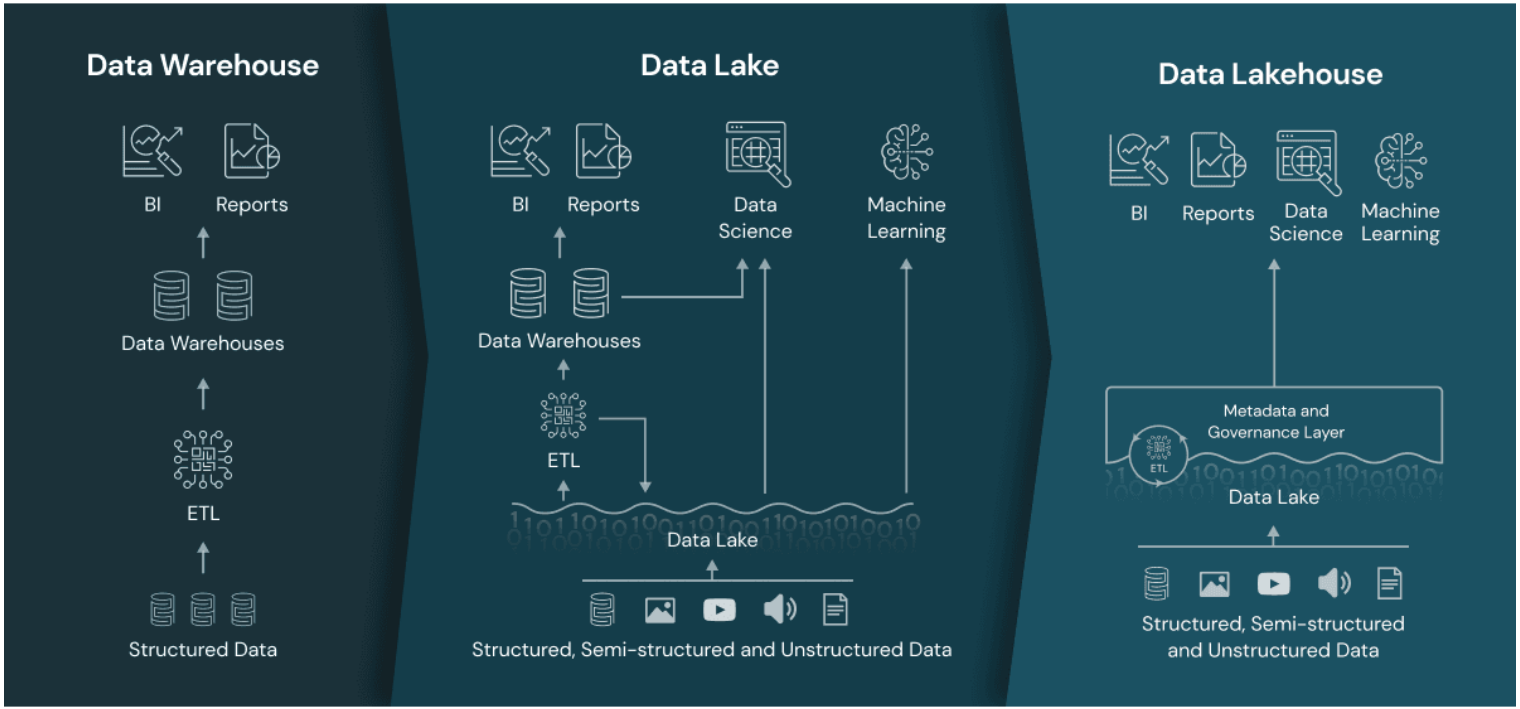
FINTECH

- Kích thích tiêu dùng
- Giữ chân khách hàng
- Tái kích hoạt

LOGISTIC

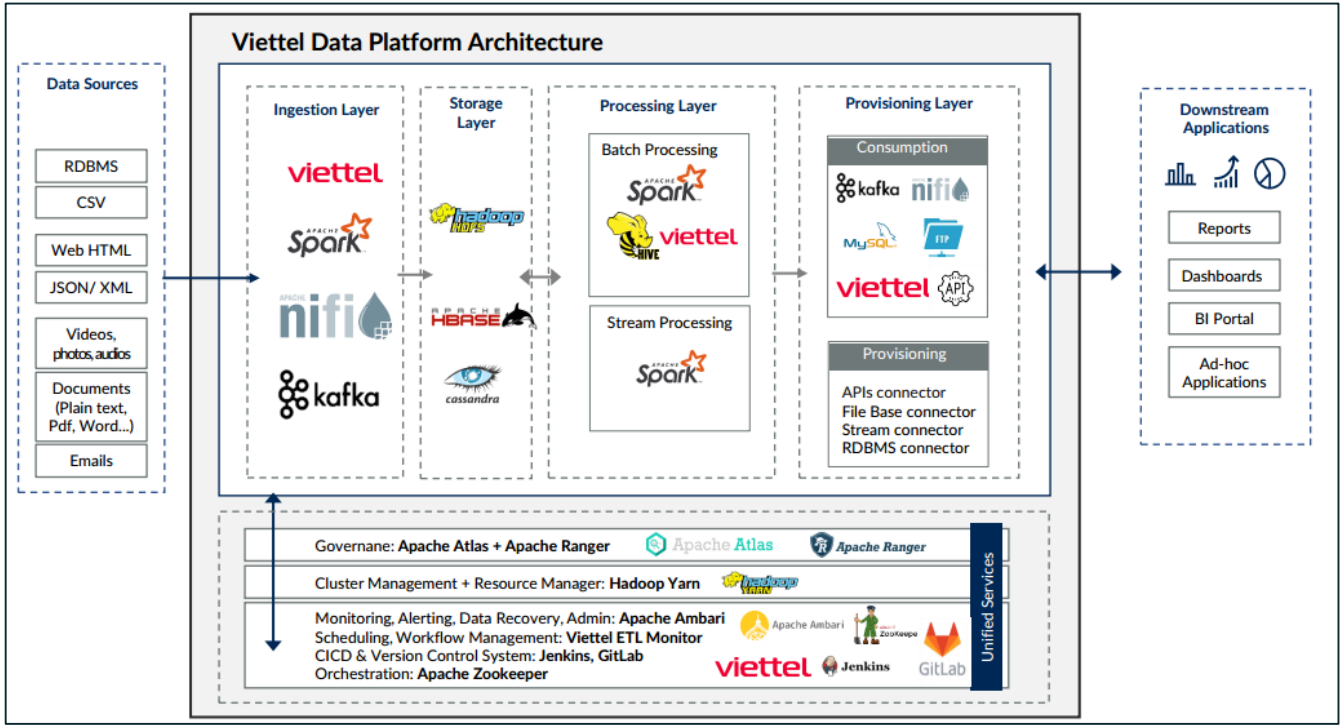
- Tối ưu vị trí đặt Hub
- Phân tích KH tiềm năng
- Tối ưu đường vận chuyển

Phát triển nền tảng dữ liệu



https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

Kiến trúc Viettel Data Lake



Ingestion Layer
Đồng bộ dữ liệu

Storage Layer
Lưu trữ dữ liệu

Procesing Layer
Xử lý, làm giàu dữ liệu

Unified Services
Quản trị nền tảng

Downstream Applications
Các công cụ khai thác dữ liệu

Phân lớp Viettel Data Lake

LỚP THU THẬP

- Lấy dữ liệu từ các hệ thống bên ngoài
- Nhận dữ liệu đẩy vào
- Xử lý real-time, mini-batch, batch

LỚP TỔNG HỢP

- Tổng hợp dữ liệu từ nhiều nguồn
- Tạo bảng dữ liệu tổng hợp từ các bảng thô (raw)

LỚP LƯU TRỮ

- Lưuu trữ nhiều định dạng khác nhau: text, video, voice,
- Lưu trữ nhiều mức độ khác nhau: block, file, object,

LỚP PHÂN TÍCH

- Tạo môi trường phân tích dữ liệu
- Cho phép thực hiện các các công nghệ AI, ML, DL

LỚP QUẢN TRỊ

- Quản lý siêu dữ liệu
- Quản lý master-data, reference-data
- Bảo mật dữ liệu
- Giám sát, vận hành, cảnh báo

Tính năng nổi bật Viettel Lakehouse

Nền tảng dữ liêu hiện đại, hỗ trợ **nhanh chóng** việc **xây dựng kho dữ liệu** tập trung trên môi trường cloud-native.

Tích hợp sẵn công cụ cho **Data Engineer/ Data Scientist/ Data Analyst** sử dụng và triển khai nghiệp vụ.

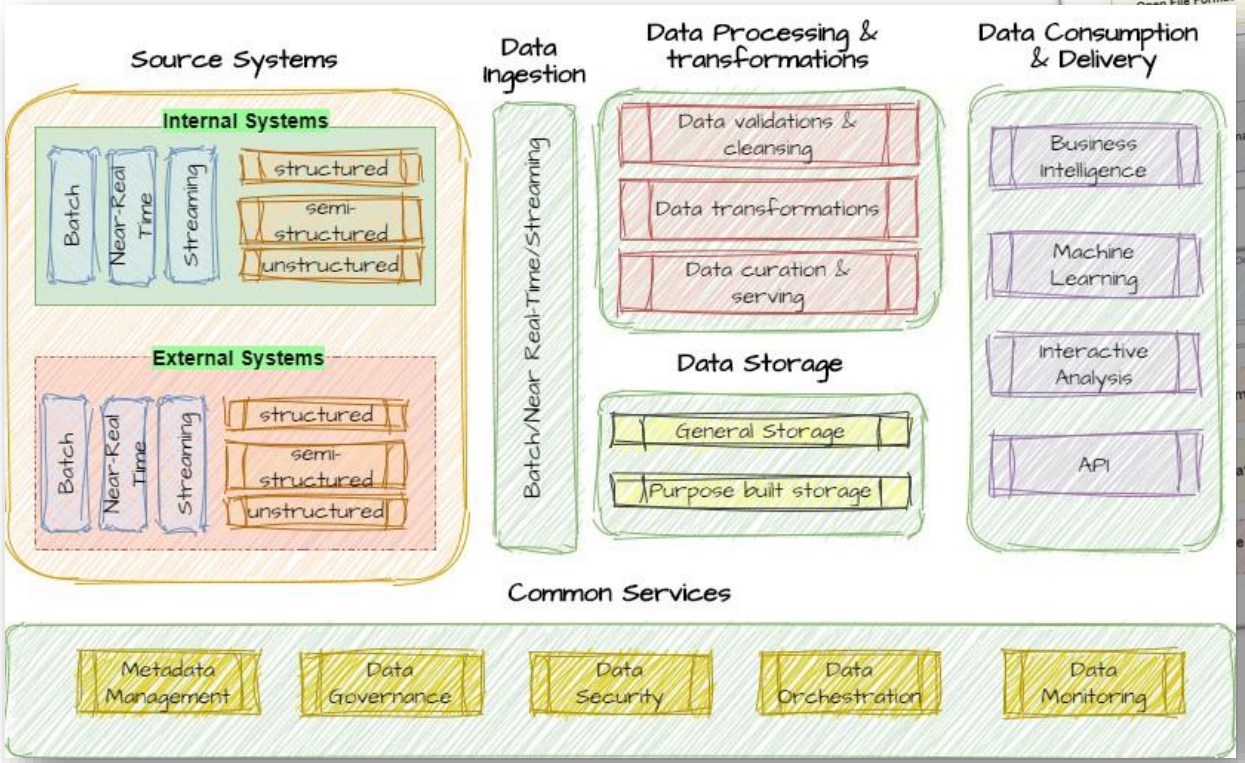
Quản trị **metadata** của toàn bộ dữ liệu với Data Catalog

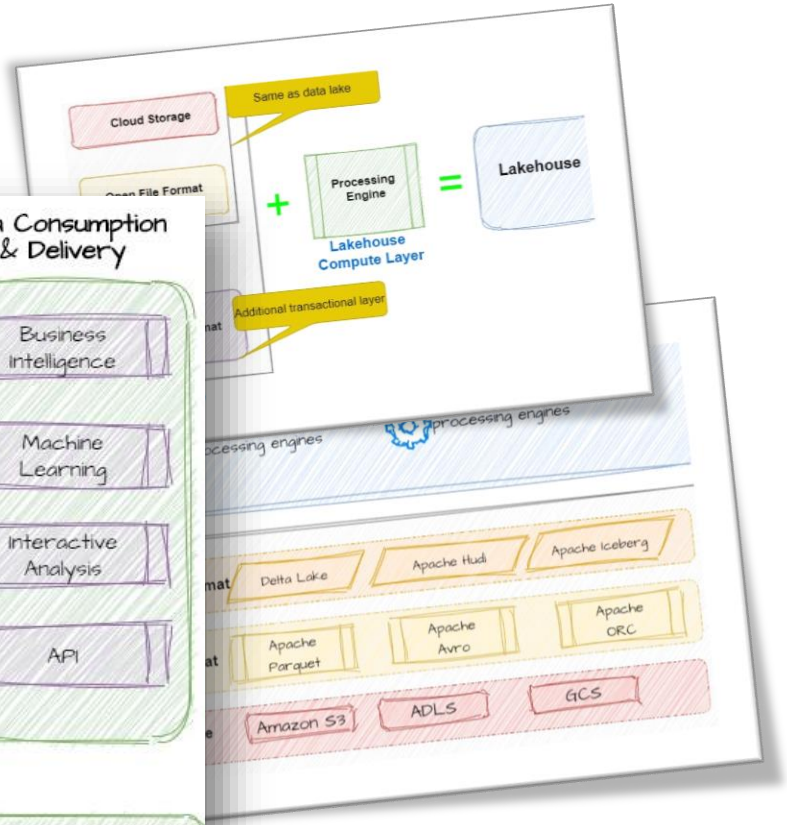
Tính năng hiện đại của định dạng **Delta File** trên kiến trúc **Lakehouse** như hỗ trợ ACID, metadata scalable tới hàng TB, Time travel dữ liệu, Audit history...

Xác thực tập trung và **phân quyền** sử dụng dữ liệu.

Hỗ trợ quản lý và triển khai trên **nhiều cụm Cloud** sử dụng một nền tảng duy nhất.

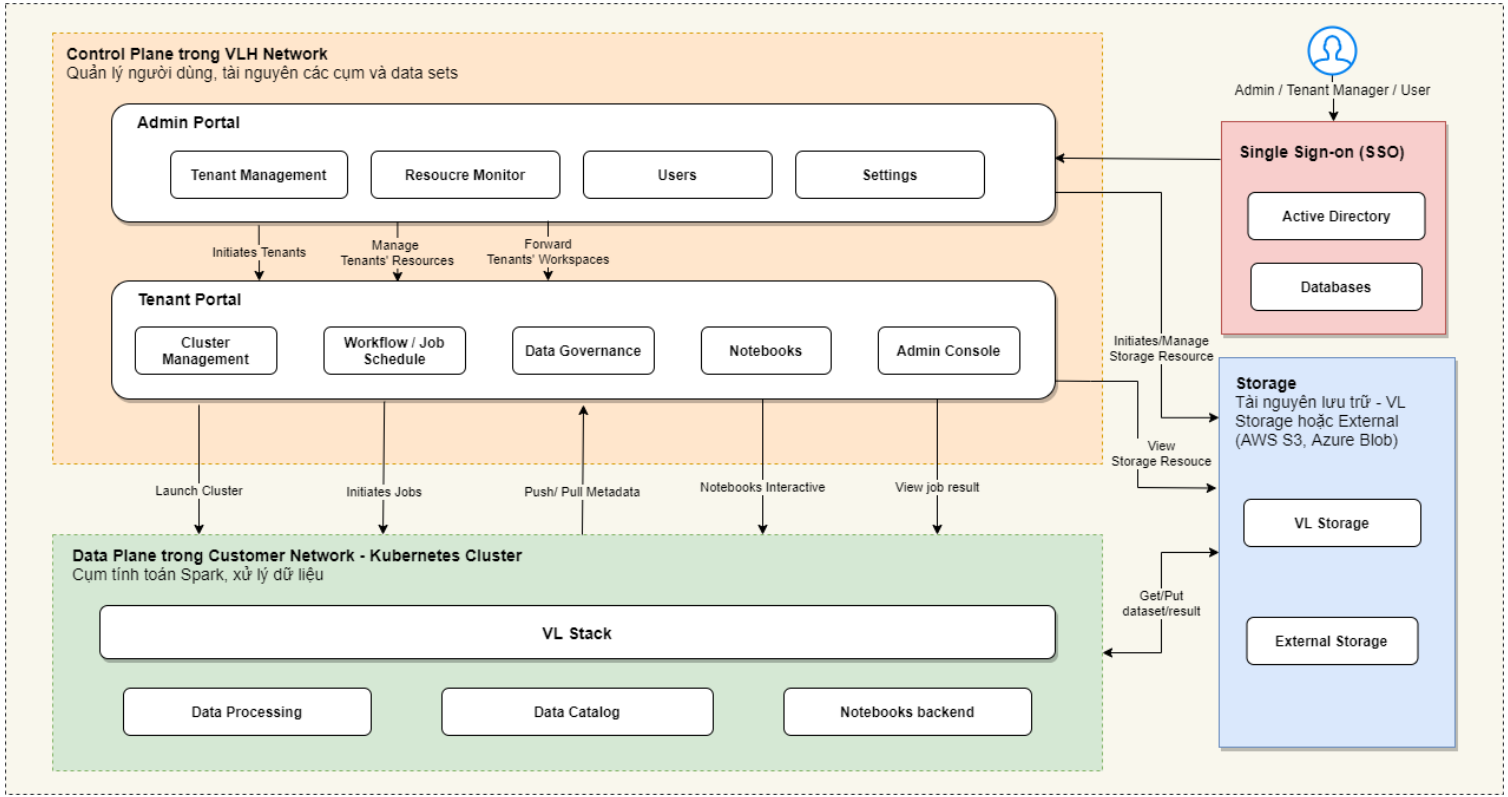
Kiến trúc Viettel Lakehouse



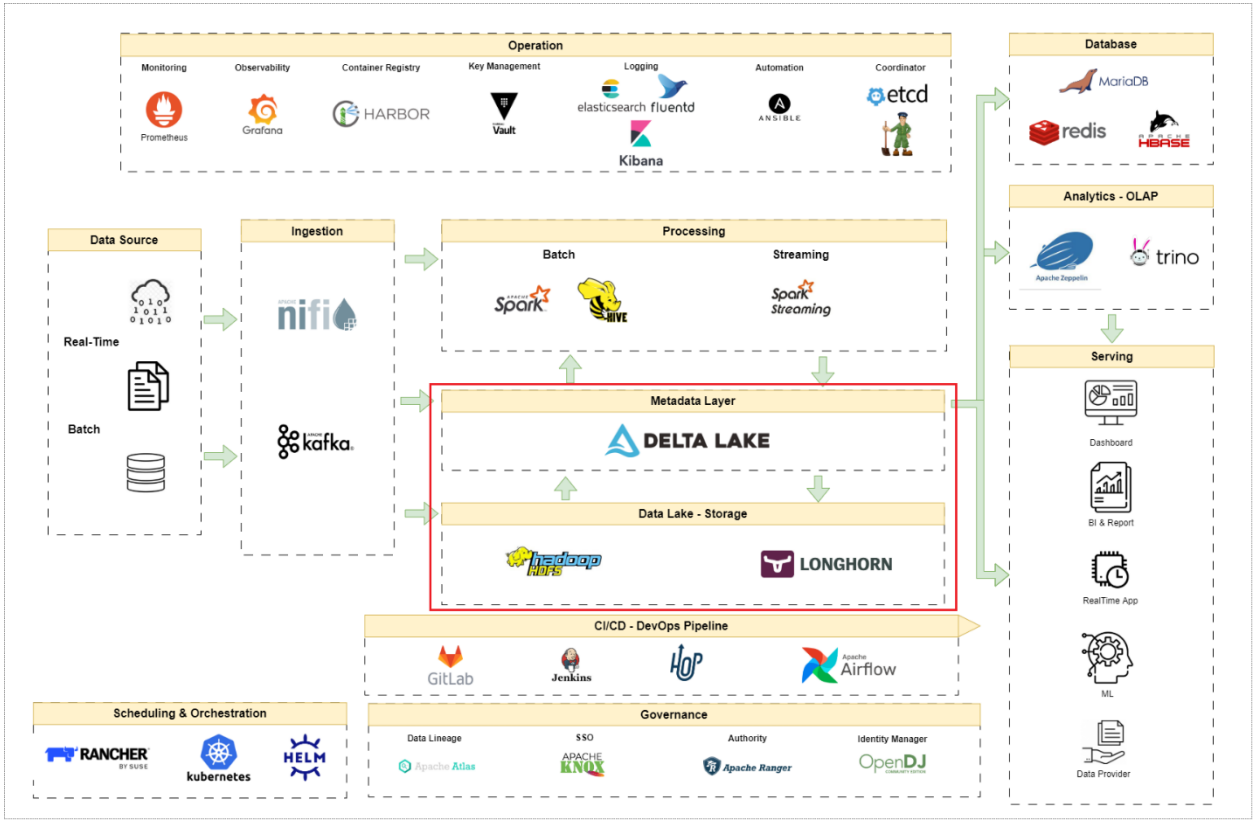


Practical Lakehouse Architecture (Gaurav Ashok Thalpati)

Kiến trúc Viettel Lakehouse



Công nghệ Viettel Lakehouse



Trân trọng cảm ơn !

Đoàn Thanh Tám
Deputy Chief Information Technology | Senior Big Data Architect

(+84) 983 585 283

tamdt9@viettel.com.vn | doanthanhtam283@gmail.com

https://www.linkedin.com/in/doanthanhtam283/

https://www.facebook.com/doanthanhtam283