

Lesson 14:

THIẾT KẾ MỘT BIG DATA PLATFORM

- Thiết kế kiến trúc nền tảng xử lý dữ liệu lớn
- Định cỡ tài nguyên và ước lượng đầu tư
- Thực hành xây dựng một Big Data Platform

Giảng Viên: Trần Đăng Hòa

Trợ Giảng: Nguyễn Chí Thanh





Trần Đăng Hòa

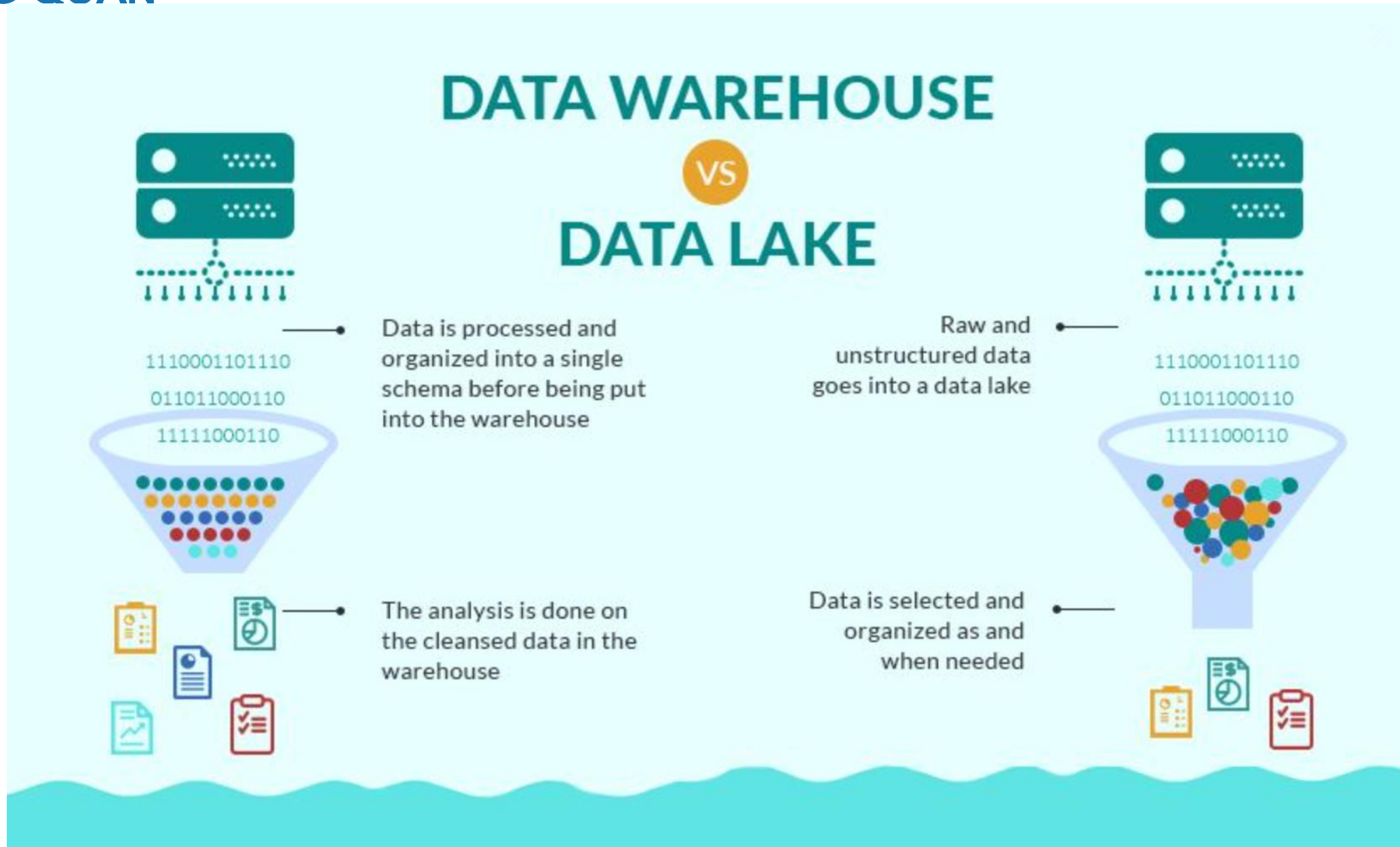
Big Data Architecture & Data Governance

Chuyên gia. **Trần Đăng Hòa**

- Xây dựng, thiết kế Data warehouse
- Thiết kế kiến trúc Data Lake
- Phân tích và khoa học dữ liệu (DA - DS)
- Kiến trúc hạ tầng và mạng lưới hệ thống Big Data
- Quản trị dữ liệu
- Admin 'Cộng đồng Big Data Việt Nam'

MỤC LỤC

- 1 TỔNG QUAN
- 2 CÁC KHÁI NIỆM CƠ BẢN
- 3 ỨNG DỤNG & GIÁ TRỊ MANG LẠI DEMO
- 4 KIẾN TRÚC TỔNG QUAN
- 5 THIẾT KẾ CHI TIẾT
- 6 CÁC ỨNG DỤNG KHAI THÁC

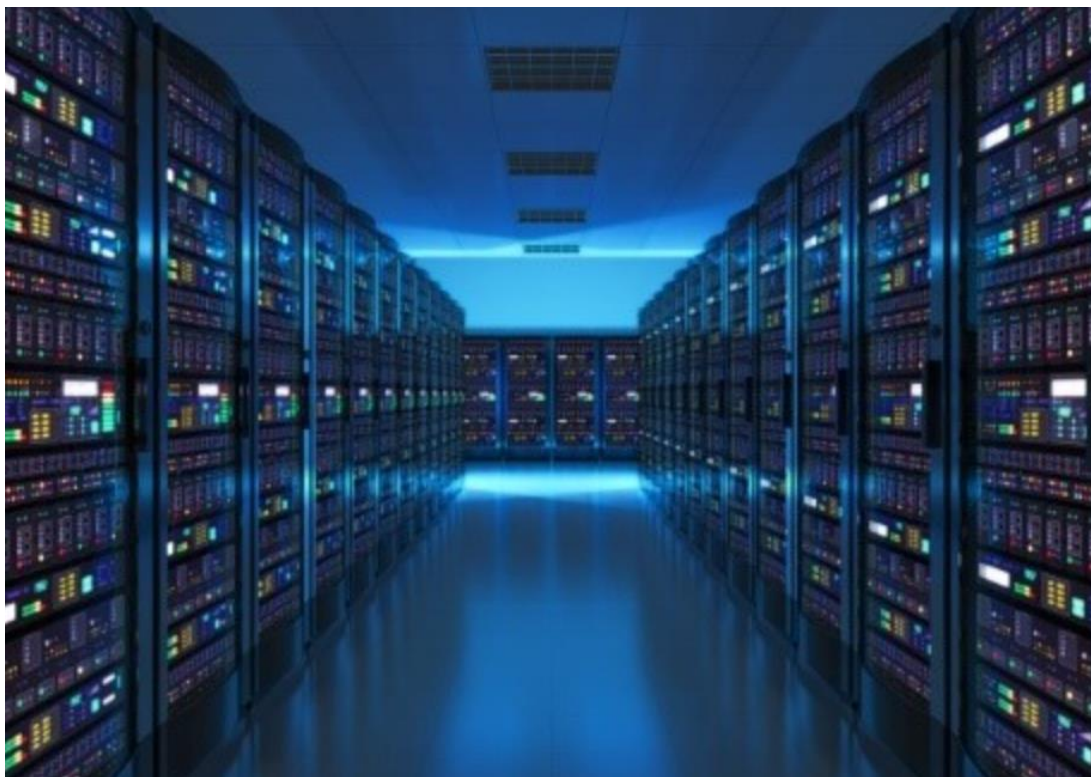


TỔNG QUAN

Kho dữ liệu đã làm tốt vai trò của nó, nhưng theo thời gian, những mặt trái của công nghệ này đã trở nên rõ rệt hơn:

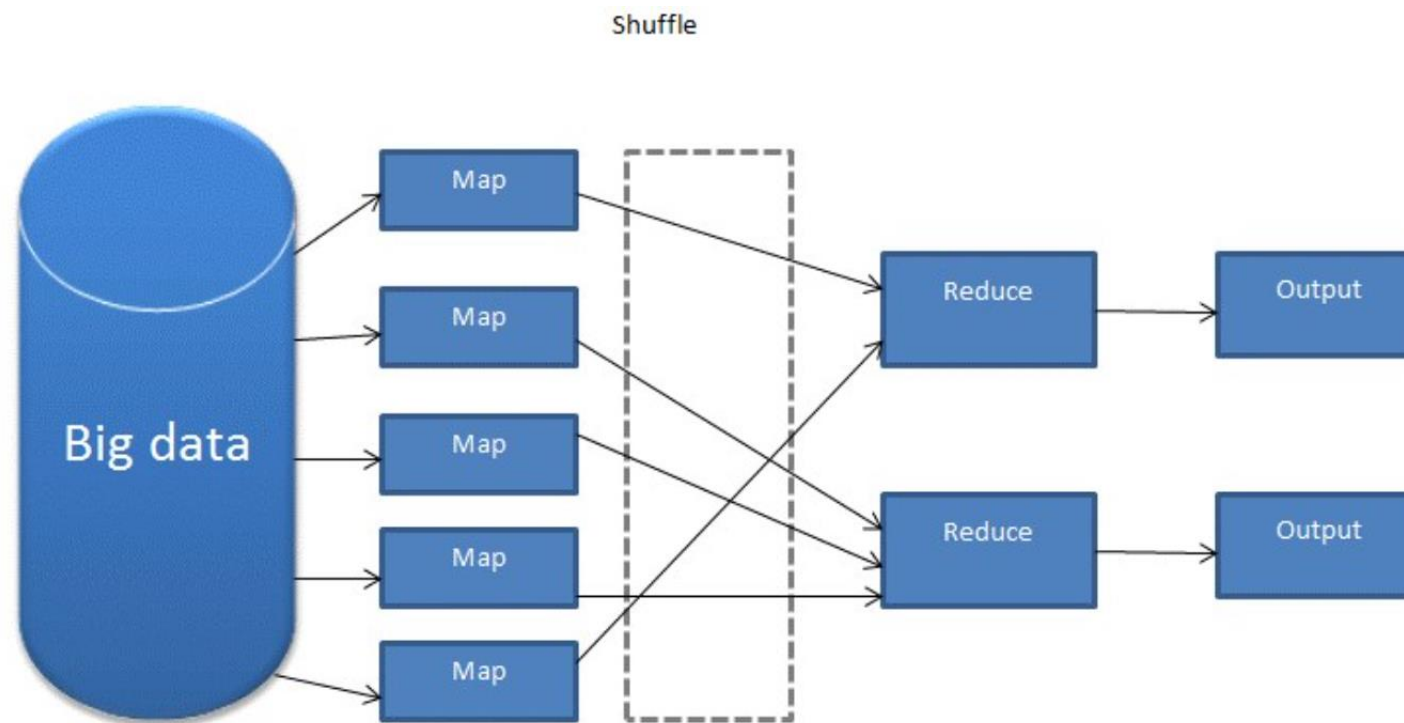
- Không có khả năng lưu trữ dữ liệu chưa được xử lý, thô
- Đắt tiền, phần cứng và phần mềm độc quyền
- Khó mở rộng quy mô do sự kết hợp chặt chẽ giữa bộ nhớ và sức mạnh tính toán

Sự vươn lên của Internet và Big Data tiền đề cho Data Lake:



TỔNG QUAN

Với sự gia tăng của “Big Data” vào đầu những năm 2000, các công ty nhận thấy rằng họ cần phải thực hiện phân tích trên các tập dữ liệu mà không thể phù hợp khi hình dung trên một máy tính. Hơn nữa, loại dữ liệu họ cần để phân tích không phải lúc nào cũng có cấu trúc gọn gàng - các công ty cũng cần những cách để sử dụng dữ liệu phi cấu trúc (unstructured data)



TỔNG QUAN

Data Lake là gì ?

Data Lake là một siêu Kho lưu trữ mà bạn có thể lưu trữ khối lượng rất lớn dữ liệu không cấu trúc, bán cấu trúc và cấu trúc. Nó được tạo ra để lưu mọi loại dữ liệu với định dạng gốc mà không giới hạn dung lượng, bản ghi hay số file. Nó lưu trữ một lượng dữ liệu lớn để tăng khả năng phân tích và tích hợp đa nền tảng.



NHỮNG THÁCH THỨC VỚI DATA LAKES ?

Khó khăn đầu tiên: Độ tin cậy của data

Nếu không có các công cụ thích hợp, Data Lakes có thể gặp phải các vấn đề về độ tin cậy, khiến các nhà khoa học và phân tích data gặp khó khăn trong việc lập luận về data. Trong kì này, chúng ta sẽ khám phá một số nguyên nhân gốc rễ trong các vấn đề về độ tin cậy của dữ liệu trên Data lakes.

Xử lý lại data do luồng dữ liệu (Data Pipeline) phát sinh lỗi ?

Xác thực data và thực thi chất lượng

Kết hợp hàng loạt và truyền trực tuyến data

Với số lượng data được thu thập theo thời gian thực ngày càng tăng, Data lakes cần khả năng dễ dàng nắm bắt và kết hợp data truyền trực tuyến với lịch sử và data dây chuyền để chúng luôn được cập nhật.

Cập nhật hàng loạt, hợp nhất và xóa bỏ

Data Lakes có thể chứa một lượng lớn data và các công ty cần có cách để thực hiện các thao tác cập nhật, hợp nhất và xóa một cách đáng tin cậy trên data đó để data đó luôn được cập nhật.

Hiệu suất truy vấn

Hiệu suất truy vấn là yếu tố chính thúc đẩy sự hài lòng của người dùng đối với các công cụ phân tích data lakes. Đối với người dùng thực hiện phân tích data tương tác, khám phá bằng SQL, phản hồi nhanh chóng cho các truy vấn phổ biến là điều cần thiết.

Quản lý metadata:

Data lakes phát triển để trở thành nhiều petabyte hoặc nhiều hơn có thể bị tắc nghẽn không phải do data mà do metadata đi kèm với nó.

CÁC KHÁI NIỆM CƠ BẢN TRONG DATA LAKE



CÁC KHÁI NIỆM CƠ BẢN TRONG DATA LAKE

Data Ingestion: Cung cấp và triển khai các công cụ đồng bộ và tiền xử lý dữ liệu để đưa vào Data Lake (Hiện có nhiều công cụ hỗ trợ đa dạng các loại dữ liệu cũng như cách thức đồng bộ, tham khảo [Nifi](#))

Data Storgare: Việc lưu trữ dữ liệu trên Data Lake đòi hỏi phải có tính mở rộng, chi phí thấp và cho phép truy cập nhanh tới dữ liệu cần khai phá và đặc biệt hỗ trợ đa định dạng.

Data Governance: Quản trị dữ liệu là một quá trình quản lý tính khả dụng, khả năng tương tác, bảo mật, tri thức nghiệp vụ và tính toán vẹn của dữ liệu trong tổ chức.

Security: Bảo mật và An toàn thông tin cần được thực hiện trong mọi lớp của hồ dữ liệu. Nó bắt đầu với việc lưu trữ, xử lý và khai thác. Đơn giản là việc cấm truy cập các tầng với những người không được cho phép. Nó nên hỗ trợ nhiều công cụ truy cập dữ liệu thông qua giao diện hoặc các màn hình quản lý.

Data Quality: Chất lượng dữ liệu là một thành phần thiết yếu của hệ thống Dữ liệu, đặc biệt với Data Lake. Dữ liệu khi sử dụng phải đảm bảo tính chính xác, toàn vẹn và kịp thời như vậy sẽ đem lại giá trị cho kinh doanh. Rủi ro về dữ liệu không chính xác, thiếu sẽ dẫn tới các quyết định sai lầm.

CÁC KHÁI NIỆM CƠ BẢN TRONG DATA LAKE

Data Discovery: Khai phá dữ liệu là một bước quan trọng trước khi bạn có thể bắt đầu phân tích chúng. Trong giai đoạn này, các kỹ thuật làm sạch và gán cấu trúc được sử dụng để gia tăng giá trị cho dữ liệu, giúp chúng được tổ chức và diễn giải dễ hiểu.

Data Auditing: Kiểm soát tác động với dữ liệu giúp theo dõi các tác động, thay đổi, rủi ro và tính tuân thủ đối với những người sử dụng.

Data Lineage: Dòng đời dữ liệu cho phép người dùng nắm được nguồn gốc của thông tin mình đang sử dụng, nơi nó đến, di chuyển qua và những gì được tác động lên nó. Nó giúp giảm các lỗi trong bước phân tích dữ liệu từ gốc tới đích.

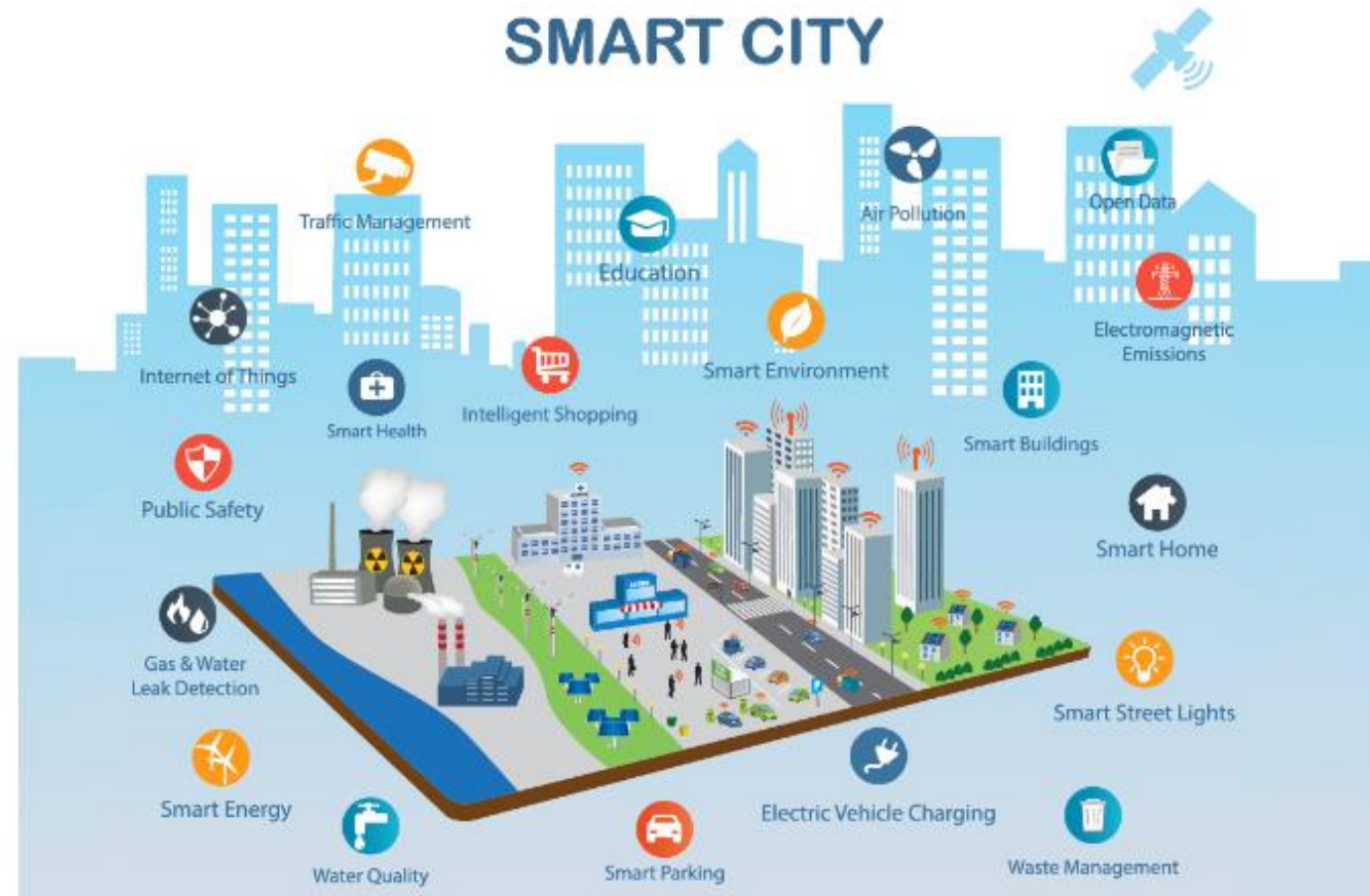
Data Exploration: Không giống như Data Discovery, giai đoạn này người sử dụng cần phân tích dữ liệu, họ cần lấy mẫu và thử nghiệm với các thông tin thu thập. Các thông tin cần được phối hợp với nhau để dễ dàng phát triển và xây dựng các bài toán phù hợp.

ỨNG DỤNG & GIÁ TRỊ MANG LẠI

Các công ty lớn đang sử dụng Data Lake như thế nào

Hiện có 6 nhóm ứng dụng phổ biến của Data Lake cho các doanh nghiệp:

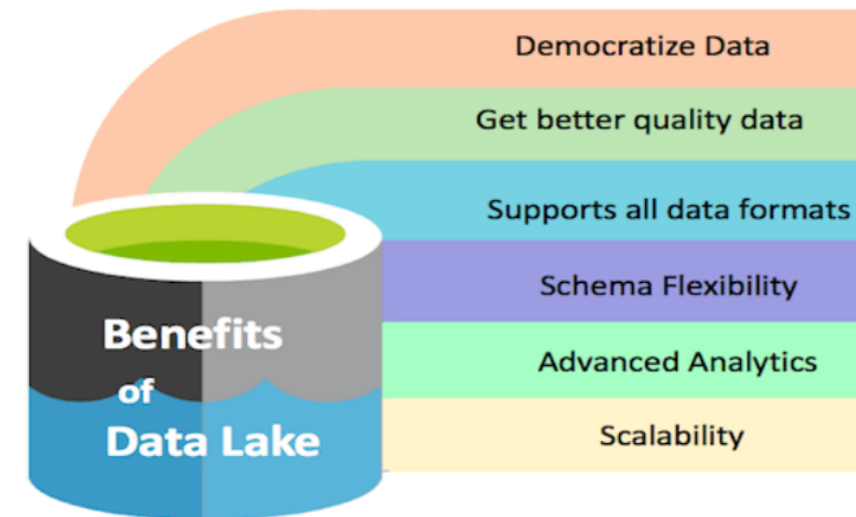
1. Thành phố thông minh (Smart City)
2. Internet vạn vật (IOT)
3. Khoa học và đời sống
4. An ninh mạng và Bảo mật
5. Khách hàng và Marketing
6. Tư vấn và hỗ trợ



Vậy Data Lake có những lợi ích và bất cập gì ?

Lợi ích

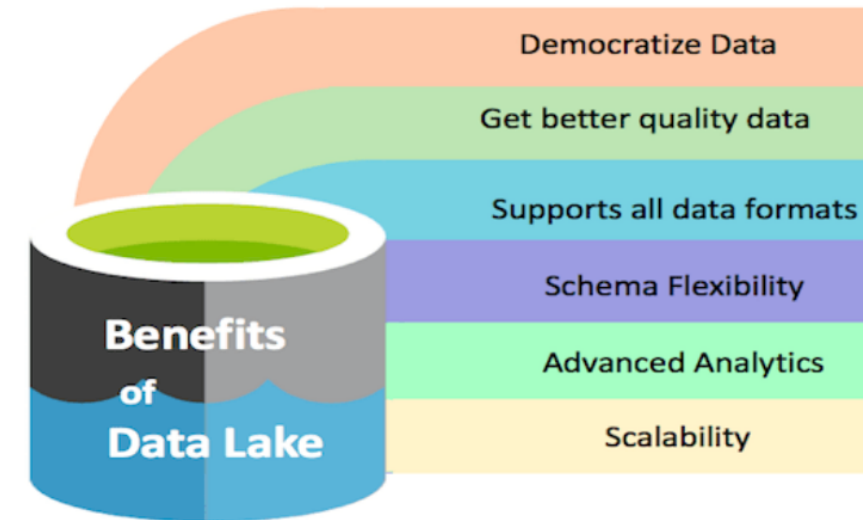
- Hệ thống hóa toàn bộ quy trình xử lý, khai thác và phân tích dữ liệu
- Cung cấp khả năng mở rộng linh hoạt và tối ưu chi phí
- Cung cấp khả năng không giới hạn về định dạng và lưu trữ dữ liệu
- Giảm chi phí sở hữu, lưu trữ dữ liệu lâu dài
- Cho phép lưu trữ hợp lý về chi phí
- Thích nghi nhanh với việc thay đổi
- Tập trung được toàn bộ các nguồn dữ liệu từ nhiều nguồn khác nhau
- Người dùng từ mọi nơi trên toàn cầu có thể dễ dàng truy cập và khai thác



Vậy Data Lake có những lợi ích và bất cập gì ?

Rủi ro:

- Sau một thời gian Data Lake có thể mất đi sự liên kết và ràng buộc giữa các thành phần
- Rủi ro lớn trong thiết kế do kiến trúc gồm rất nhiều thành phần và phân hệ phức tạp
- Việc lưu trữ lớn dữ liệu không cấu trúc có thể dẫn tới rối loạn dữ liệu, phức tạp và không thể khai thác được
- Do lưu trữ lâu và nhiều hơn lên cũng làm tăng chi phí
- Khó kiểm soát bảo mật và truy cập



THÀNH PHẦN CHÍNH CỦA DATA LAKE



BUSINESS ANALYST

PHÂN TÍCH KINH DOANH

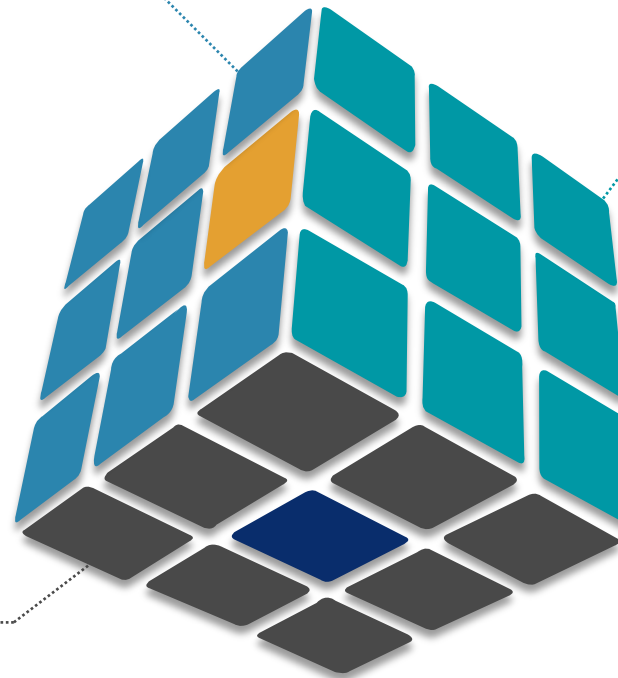
Các nhà lãnh đạo Doanh nghiệp đưa ra những quyết định chiến lược đối với hoạt động kinh doanh của doanh nghiệp.



DATA WAREHOUSE

KHO DỮ LIỆU

Chứa dữ liệu tổng hợp của doanh nghiệp



DATA MINING

KHAİ PHÁ DỮ LIỆU

Các kỹ thuật dùng để khai phá dữ liệu và phát hiện tri thức như phân loại (*Classification*), phân nhóm (*Clustering*), phát hiện luật kết hợp (*Association Rule*), Dự đoán (*Predcition*)

Thiết kế kiến trúc dữ liệu trong Data Lake

Data Lake là một cách tiếp cận hoàn toàn mới giữa sự kết hợp sức mạnh của Big Data và khả năng Self-service. Nhiều doanh nghiệp hiện nay đã phát triển hoặc triển khai hệ thống này trong hoạt động điều hành, sản xuất kinh doanh.

Vậy cách thiết kế hệ thống và tổ chức dữ liệu trong Data Lake như thế nào?

Khái niệm cơ bản

Thiết kế kiến trúc hạ tầng

Thiết kế lưu trữ

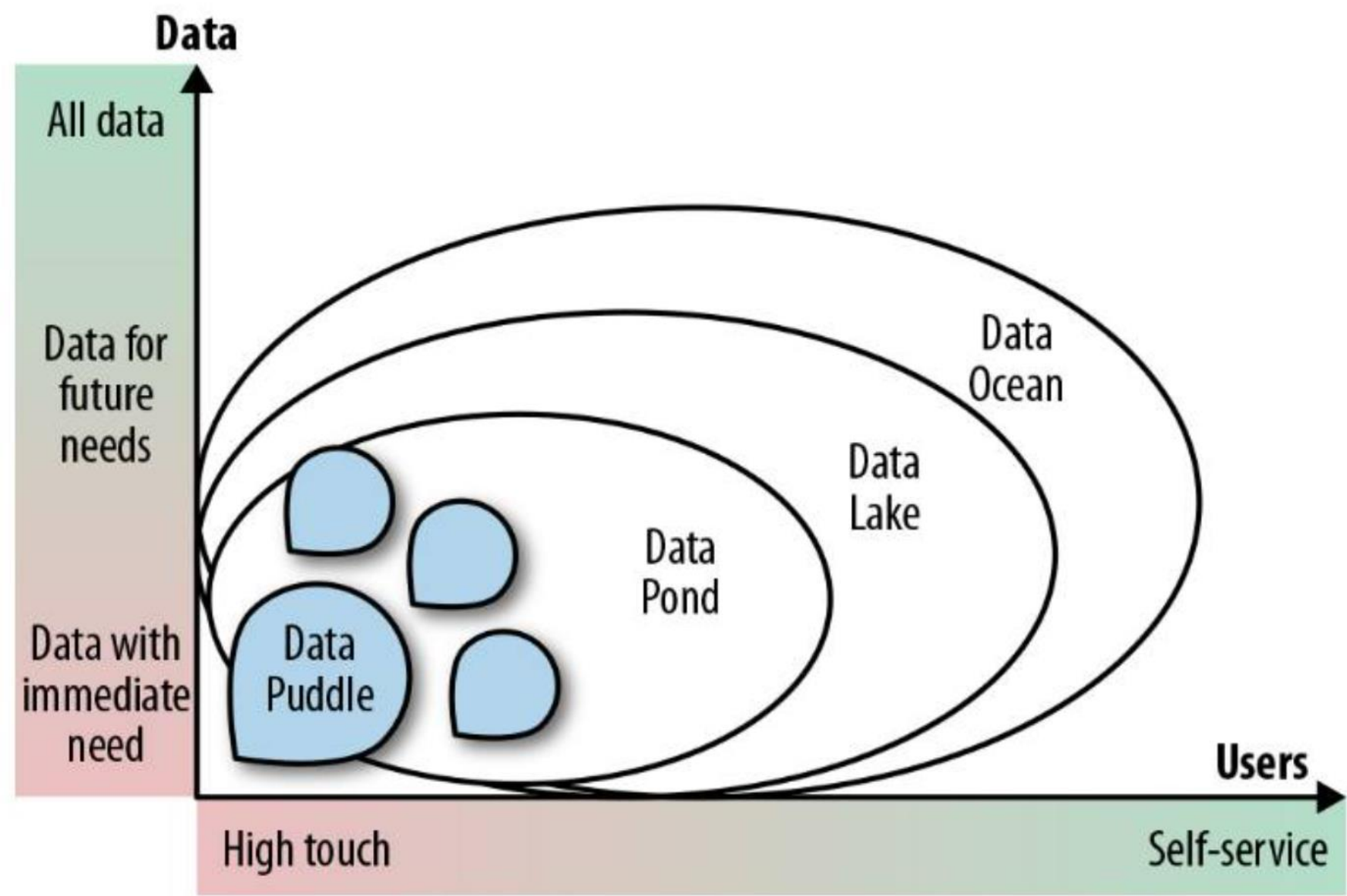
Thiết kế kiến trúc dữ liệu trong Data Lake

Khái niệm

Data Lake là một khái niệm tương đối mới, vì vậy để xác định được các kiến trúc xây dựng bạn có thể tham khảo một số định nghĩa sau đây:

- *Data Puddle* là một vùng dữ liệu, cơ bản như một Data Mart với một mục đích hoặc phục vụ cho một dự án. Áp dụng trong những bước đầu tiên áp dụng công nghệ Big Data.
- *Data Pond* là một tập hợp các vùng dữ liệu, có thể coi như một kho dữ liệu nhưng được thiết kế chưa tối ưu, giúp giảm tải kho dữ liệu truyền thống hiện có. Tuy có chi phí công nghệ thấp hơn, khả năng mở rộng tốt nhưng lại đòi hỏi chi phí CNTT lớn, kèm theo việc hạn chế trong tính khả dụng của dữ liệu nên nó không thực sự giúp tối ưu khả năng Self-service và Data-driven hỗ trợ ra quyết định cho người dùng doanh nghiệp.
- *Data Lake* khác với Data Pond ở 2 đặc điểm quan trọng: đầu tiên, nó hỗ trợ khả năng Self-Service, nơi mà người dùng có thể tìm và sử dụng các dữ liệu mà họ muốn mà không cần nhờ tới sự trợ giúp của bộ phận CNTT. Hai là nó nhằm mục đích chứa các dữ liệu mà ngay hiện tại doanh nghiệp hay các cá nhân cũng chưa có nhu cầu sử dụng.
- *Data Ocean* mở rộng khả năng self-service dữ liệu và data-driven hỗ trợ ra quyết định dữ liệu, bất cứ nơi nào có thể, bất kể nó có có tải vào hệ thống Data Lake hay không.

Thiết kế kiến trúc dữ liệu trong Data Lake



Thiết kế kiến trúc dữ liệu trong Data Lake

Vậy cần làm gì để xây dựng thành công một Data Lake? Tương tự như bất kỳ dự án nào, bắt buộc phải có việc liên kết nó với chiến lược của công ty kèm theo việc đầu tư và điều hành xuyên suốt. Ngoài ra, cần xác định 3 điều kiện chính trước khi bắt đầu:

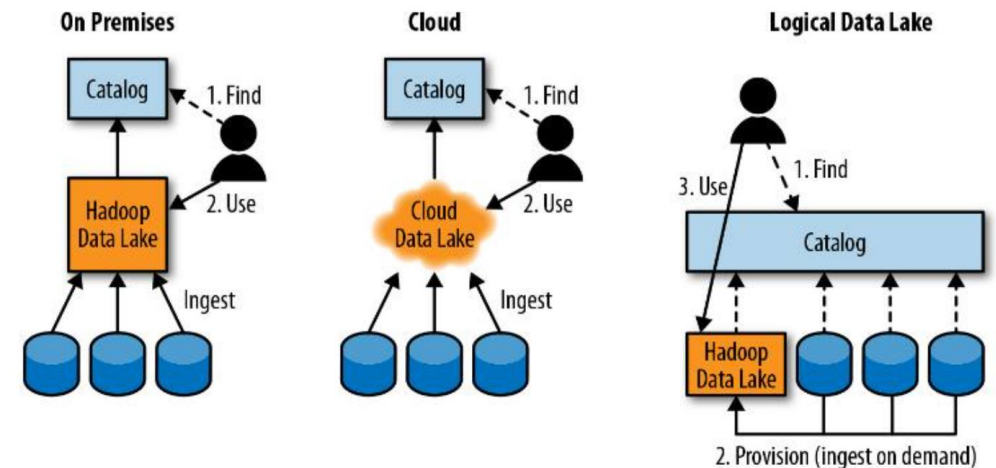
- Nền tảng phù hợp: đánh giá và lựa chọn giữa Hadoop, Amazon Web Service, Microsoft Azure,...
- Dữ liệu : mục đích lưu càng nhiều dữ liệu càng tốt với định dạng gốc
- Giao diện tương tác: Khả năng cung cấp Self-service ở mức độ đơn giản cho người dùng, đảm bảo người dùng có thể tự tìm kiếm và khai thác

Thiết kế kiến trúc dữ liệu trong Data Lake

Thiết kế kiến trúc hạ tầng

Roap map: Chúng ta đã có các điều kiện cần, vậy các bước chính cần thực hiện là gì?

1. Triển khai cơ sở hạ tầng cho lưu trữ (Hadoop là một lựa chọn không tồi)
2. Tổ chức Data Lake (tạo các Zone để phân vùng cho các người dùng, dữ liệu khác nhau).
3. Thiết lập Self-service (tạo các danh mục quản lý dữ liệu, thiết lập queyefn và cung cấp các công cụ khai thác, phân tích dữ liệu).
4. Vận hành và cung cấp Data Lake cho người dùng

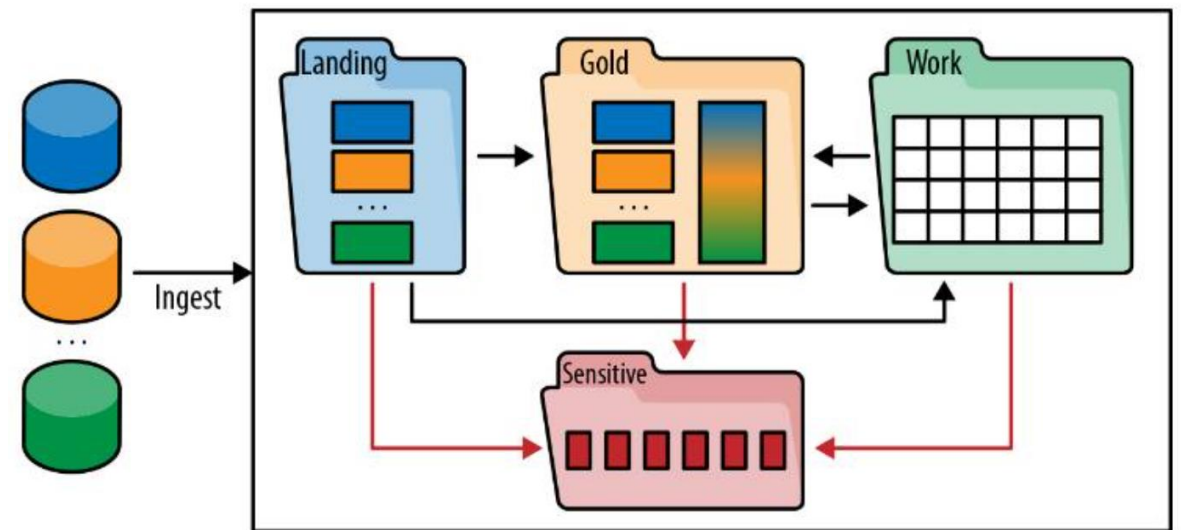


Thiết kế kiến trúc dữ liệu trong Data Lake

Thiết kế lưu trữ

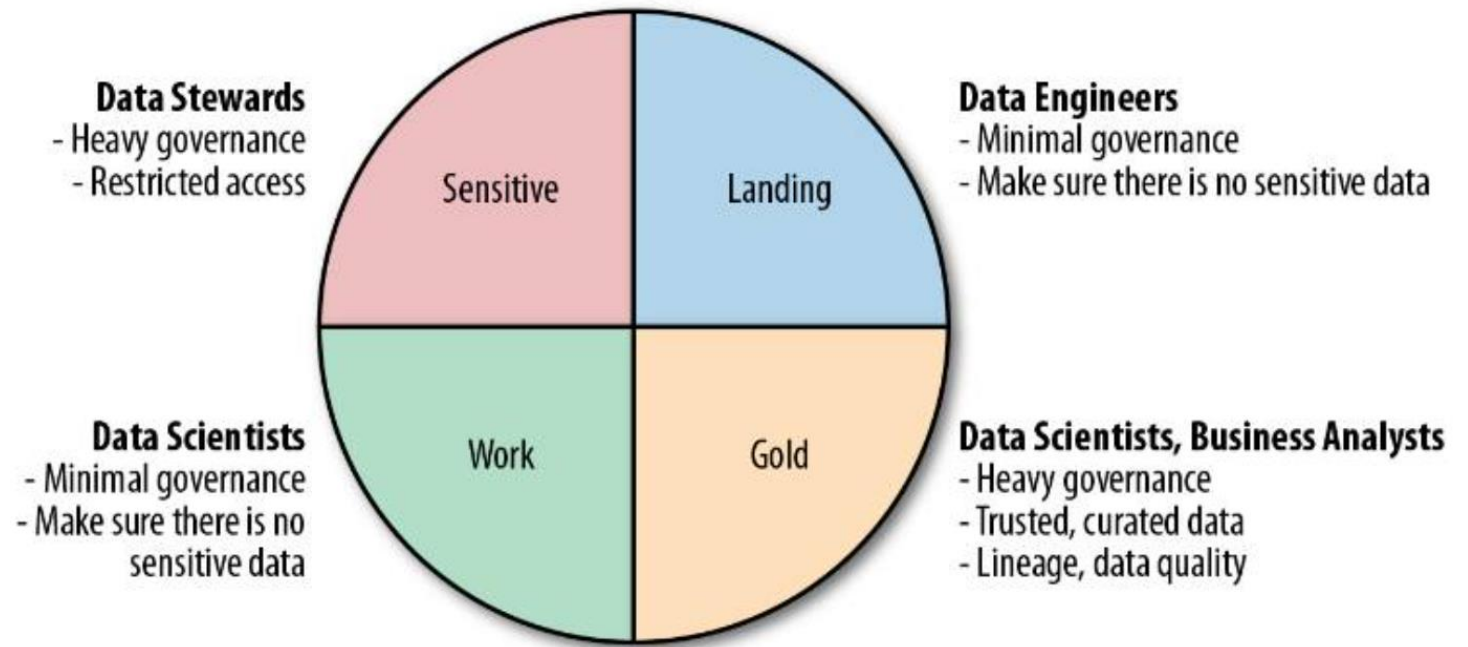
Hầu hết các Data Lake được tổ chức theo cùng một cách, với các Zone khác nhau:

- **Raw hay Landing Zone** là nơi dữ liệu được đưa vào và xử lý, làm chuẩn với mục tiêu giống với hiện trạng ban đầu tối đa nhất.
- **Gold hay Production Zone** là nơi lưu trữ dữ liệu đã được xử lý, tổng hợp sạch sẽ.
- **Dev hay Work Zone** là nơi có nhiều nhân sự phát triển, phân tích, khai phá làm việc và được tổ chức theo nhu cầu của người dùng, theo dự án hoặc theo chủ đề, khi hoàn thành sản phẩm triển khai, dữ liệu sẽ được chuyển lên Gold Zone.
- **Sensitive Zone** là nơi chưa dữ liệu nhạy cảm, dữ liệu mã hóa phục vụ để trao đổi với các hệ thống ngoài Data Lake.

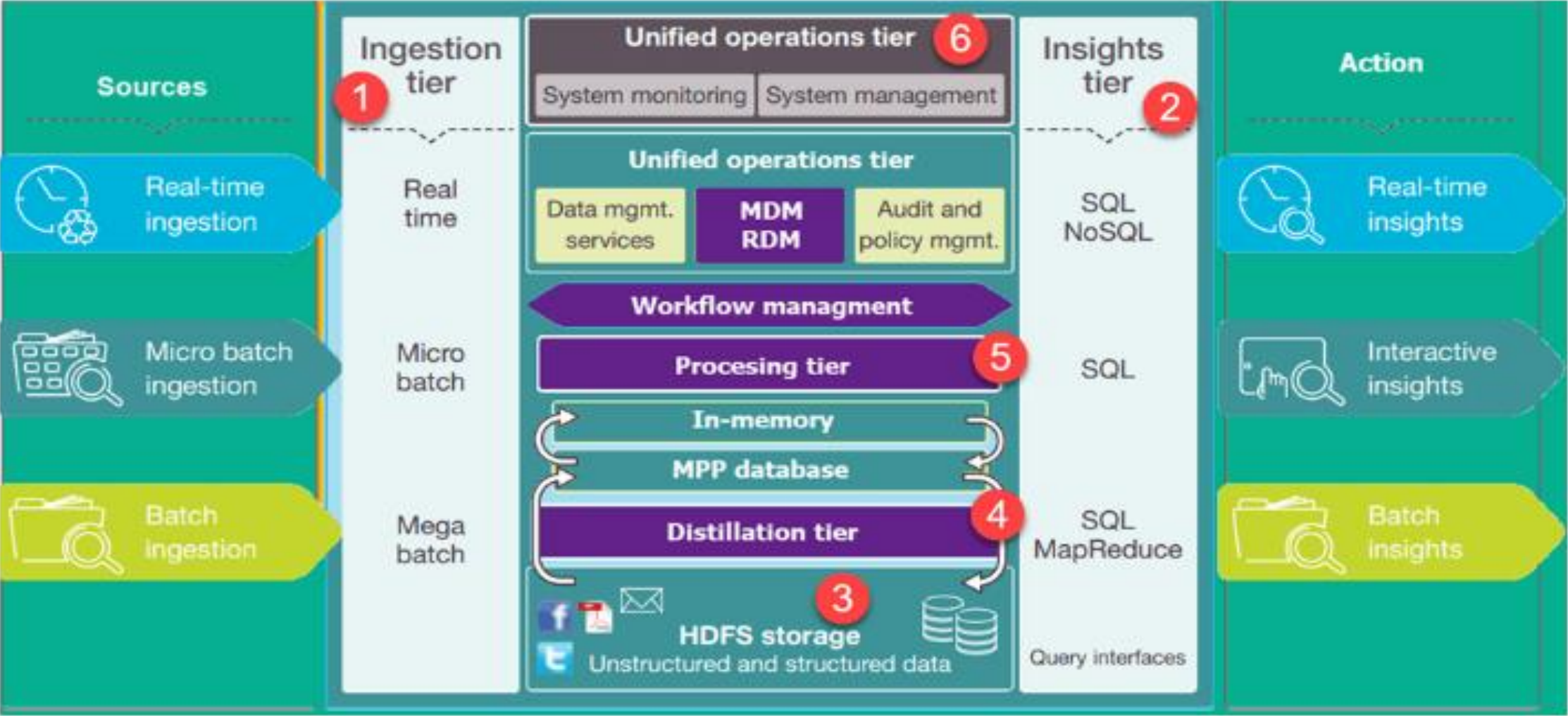


Thiết kế kiến trúc dữ liệu trong Data Lake

Với các Zone khác nhau, sẽ có cách chính sách quản lý và phân quyền phù hợp. Ví dụ, dữ liệu trong Gold Zone thường được tổ chức chặt chẽ, đảm bảo chất lượng và độ chính xác. Những người dùng khác nhau sẽ có nhu cầu với từng vùng, những người phân tích kinh doanh sẽ sử dụng chủ yếu Gold Zone để khai thác, các nhân viên phát triển, vận hành hệ thống sẽ đưa và xử lý dữ liệu vào Raw Zone sau đó chuyển sang Gold Zone, những người thử nghiệm hoặc xây dựng bài toán học máy sẽ sử dụng Work Zone để làm việc trước khi triển khai sản phẩm



KIẾN TRÚC CƠ BẢN CỦA MỘT DATA LAKE



KIẾN TRÚC CƠ BẢN CỦA MỘT DATA LAKE

Data Lake có 5 tầng quan trọng:

Tầng nạp dữ liệu (Ingestion Tier) : Tầng này nằm bên trái của kiến trúc. Dữ liệu có thể được tải vào Data Lake từ nhiều Nguồn (Data Source) thông qua thời gian thực (Real Time) hoặc theo lô (Batches)

Tầng khai phá (Insights Tier): Tầng này nằm ở phía trên bên phải của hình nơi sẽ sử dụng dữ liệu từ hệ thống. Các truy vấn SQL, NoSQL, SQL MapReduce sẽ được sử dụng để khai thác và phân tích dữ liệu.

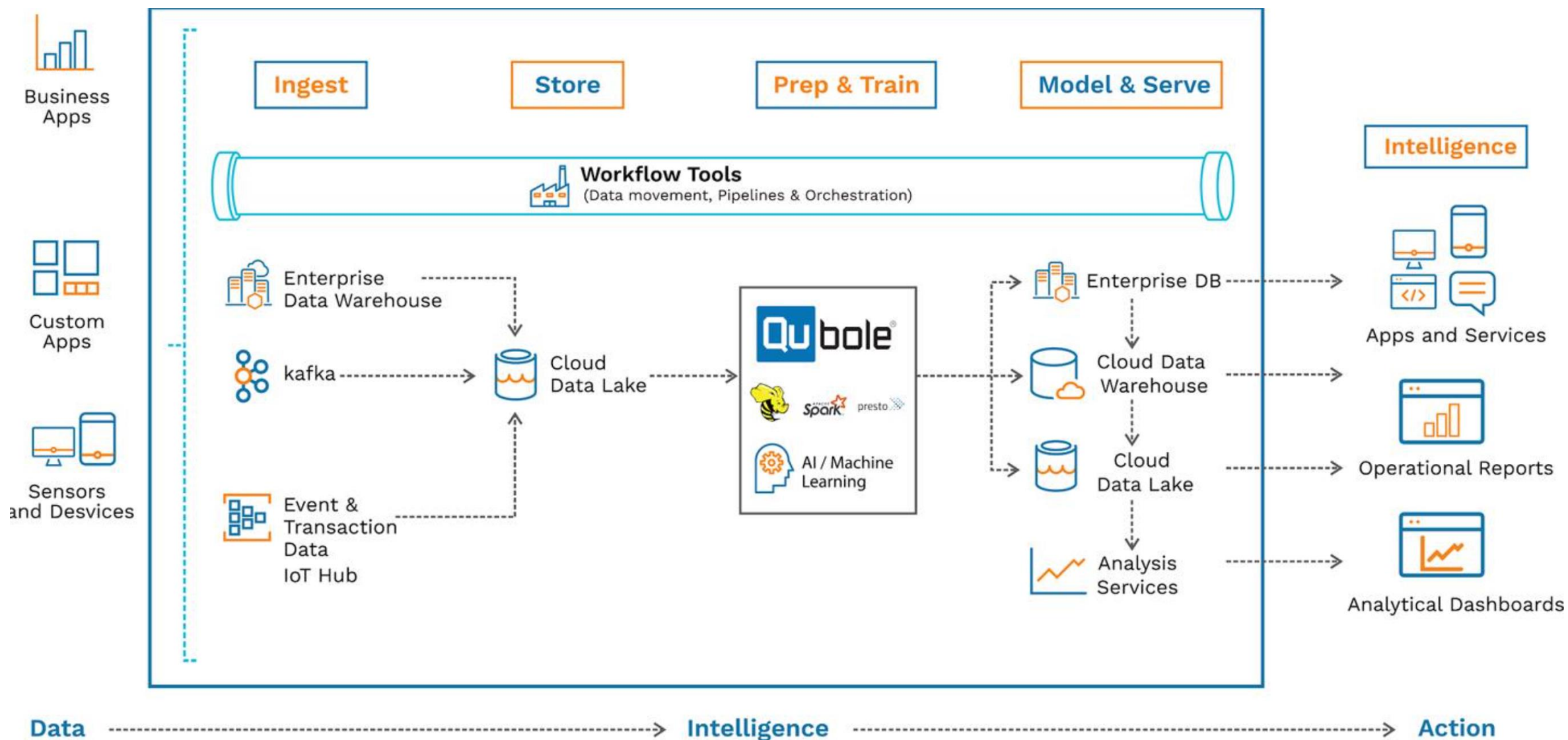
Tầng lưu trữ (Storage): Tầng này hiện hầu hết các hệ thống sử dụng HDFS với ưu điểm về chi phí, tính linh hoạt, chịu lỗi và khả năng mở rộng dễ dàng đặc biệt hiệu quả với các dữ liệu cấu trúc và phi cấu trúc. Đây là tầng sẽ lưu trữ toàn bộ dữ liệu của hệ thống.

Tầng tiền xử lý (Distillation tier): Vai trò lấy dữ liệu trực tiếp từ tầng lưu trữ sau đó làm sạch và chuyển sang dữ liệu có cấu trúc, giúp dễ dàng hơn cho việc phân tích.

Tầng xử lý (Processing tier): Xử lý và chạy các thuật toán phân tích, hỗ trợ người dùng truy vấn thời gian thực, tương tác theo lô với mục đích sinh ra các dữ liệu có cấu trúc để dễ dàng phân tích.

Tầng giám sát, vận hành (Operations tier): Chi phối quản lý và giám sát hệ thống, bao gồm cả việc quản lý chất lượng dữ liệu, danh mục dữ liệu, bảo mật và quy trình khai thác, sử dụng hệ thống

DATA LAKE ARCHITECTURE



THIẾT KẾ CHI TIẾT

1. Ingest Data (Đưa dữ liệu vào hệ thống)

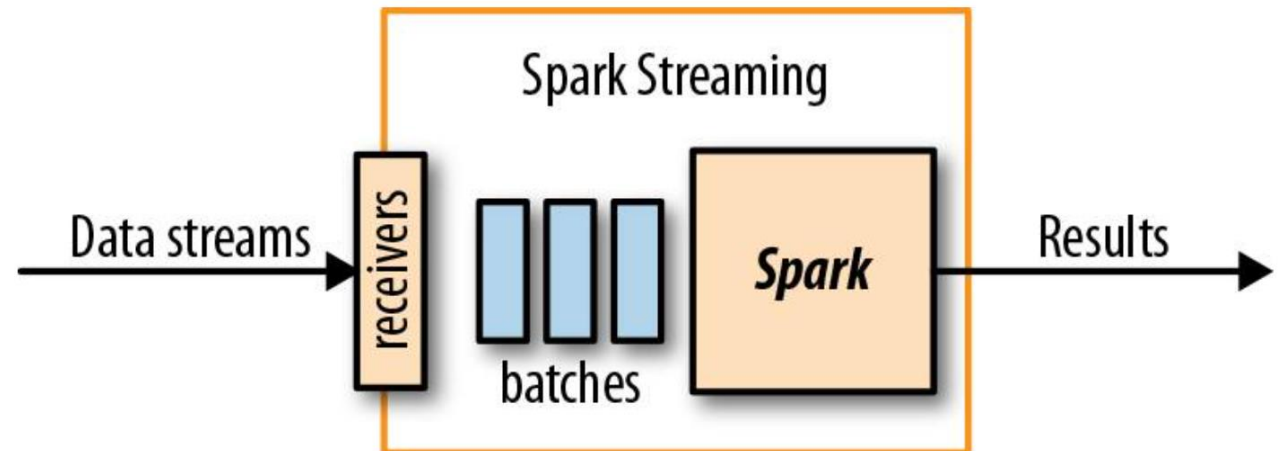
- Việc đưa dữ liệu vào hệ thống cần lưu ý:

+ Dữ liệu được đưa vào sẽ phục vụ lưu trữ và đảm bảo cho các dịch vụ sử dụng phía sau nên Data phải đảm bảo thời gian, độ trễ dữ liệu phù hợp.

+ Trong các trường hợp việc xử lý theo lô mất nhiều thời gian, đặc biệt các hệ thống Recommendation đòi hỏi độ trễ mức Near RealTime thì Ingest Data chính là bước bạn thực hiện nó.

Ồ Vì vậy, bạn cần một công cụ đảm bảo tự động gom dữ liệu thành nhóm. Công cụ này giúp bạn tổ chức các luồng Streaming tự động, ở đây mọi người có thể sử dụng stack Nifi + Kafka + Spark Streaming và Kiniesis Stream để chuyển dữ liệu lên Amazon S3 hoặc Redshift trong trường hợp bạn sử dụng Cloud.

- Với Spark Streaming đã phát triển đáng kể về khả năng và tính đơn giản, cho phép bạn phát triển các ứng dụng Streaming. Nó sử dụng các API để kết nối với các nguồn dữ liệu như HDFS, Flume, Kafka, Python. Có thể đọc dữ liệu từ HDFS, Flume, Kafka



THIẾT KẾ CHI TIẾT

2. Cảnh báo, giám sát và quản lý dữ liệu

- Khi bạn bắt đầu phát triển Data Lake của mình thì cấu trúc trong nó phải được xác định rõ.
- Bạn cần quản trị các Data Flow để đảm bảo các luồng ETL luôn hoạt động hoặc trong trường hợp xảy ra lỗi phải được cảnh báo và giám sát.

a. Giám sát, cảnh báo Data Lake

- Giám sát là việc cực kỳ quan trọng trong việc thành công của Data Lake.
- Ngày ngày các hệ thống giám sát cung cấp một bộ dịch vụ phong phú như bảng điều khiển (dashboard), tin nhắn, email, gọi, ..nhằm thông báo sớm các bất thường
- LogicMonitor, Datadog và VictorOps là những công cụ mà bạn có thể tham khảo
- Hiện Data Lake bên mình đang tự phát triển công cụ riêng do nhu cầu, ví dụ hình minh họa dưới



THIẾT KẾ CHI TIẾT

b. Công cụ quản lý dữ liệu

- Việc quản lý dữ liệu bao gồm bảo mật, an toàn thông tin và quản lý danh mục dữ liệu trong hệ thống
- Apache Ranger đóng vai trò trong việc đảm bảo phân quyền truy xuất, nó là một framework cho phép thiết lập các chính sách quản lý quyền của những người sử dụng hệ thống
- Apache Atlas là một Low-Level Service trong Hadoop Stack và là công cụ hỗ trợ việc quản lý danh mục dữ liệu, hỗ trợ Data Lineage bằng việc giúp mọi người hiểu Meta Data, dòng đời, công thức và các định nghĩa liên quan của dữ liệu

THIẾT KẾ CHI TIẾT

3. Chuẩn bị dữ liệu

- Bước xử lý và tổng hợp sẽ tạo ra những giá trị cho dữ liệu, bước này sẽ bổ sung các thông tin danh mục, id, hoặc tính toán lên các giá trị so sánh cùng kỳ, ..
- Công cụ thường sử dụng ở đây là Hive SQL do khả năng triển khai nhanh, phù hợp và dễ sử dụng với đa số người dùng
- Apache Hive cũng hỗ trợ tương thích với Atlast để đồng bộ danh mục dữ liệu tự động
- Hive được tổ chức để trở thành một Data Warehouse trên nền Hadoop/MapReduce, nó biên dịch SQL -> Spark Job -> MapReduce.
- Hive Metastore lưu trữ cấu trúc của bảng, giúp chuyển các thư mục HDFS thành dữ liệu có cấu trúc để Query
- Bạn có sử dụng Spark SQL, Presto, Tez, Impala , Drill để tương tác với Hive

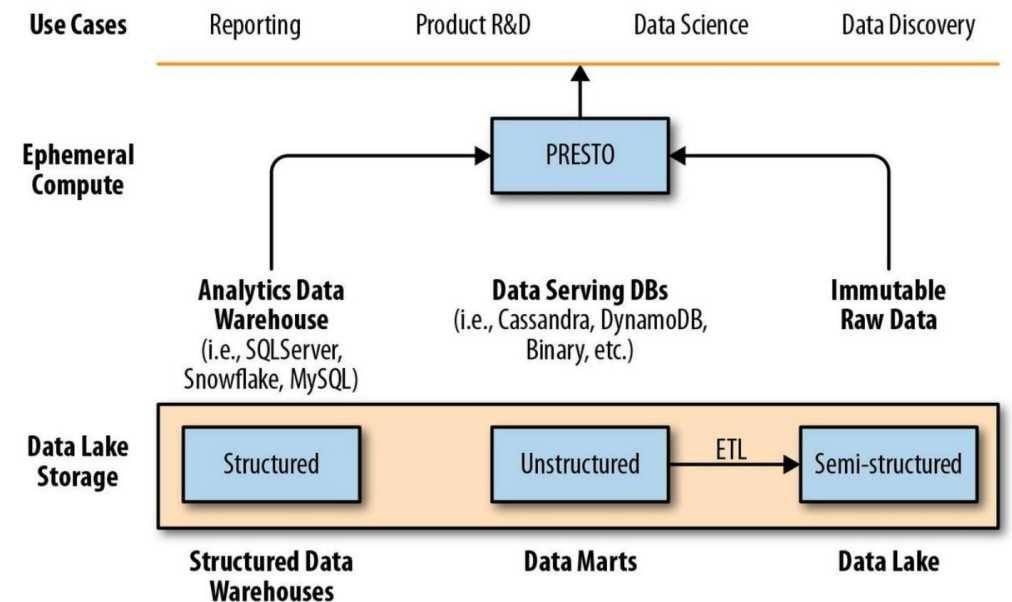
THIẾT KẾ CHI TIẾT

4. Model và Serve Data

- Phần học máy ở đây có thể dùng Spark ML, nó là một mảng riêng nên mình sẽ không đi sâu, các bạn có thể join group Cộng đồng Big Data Analytics.
- Trong phần này, bạn có thể triển khai các usecase phổ biến về quản lý Khách hàng như Churn, Upsell, Cross Sell,...

5. Khai thác dữ liệu

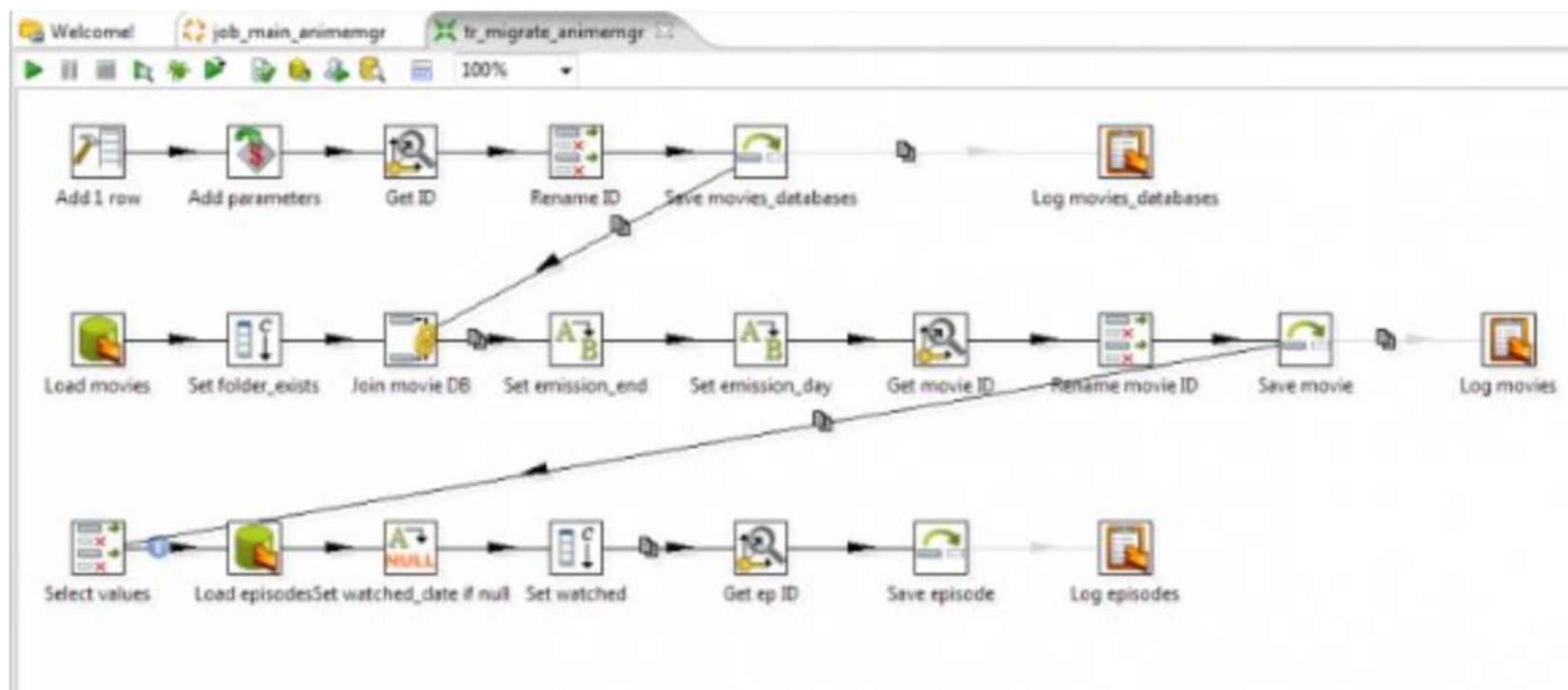
- Các công cụ BI sẽ cung cấp cho bạn khả năng trình diễn và khai thác dữ liệu hiệu quả
- Ví dụ như Power BI, Tableau, Qlik đều là những công cụ cho phép bạn cấu hình, tùy chỉnh và tương tác với dữ liệu hiệu quả. Nó đều hỗ trợ kết nối với Impala, Presto, Tez,...



THIẾT KẾ CHI TIẾT

6. Triển khai và tự động hóa

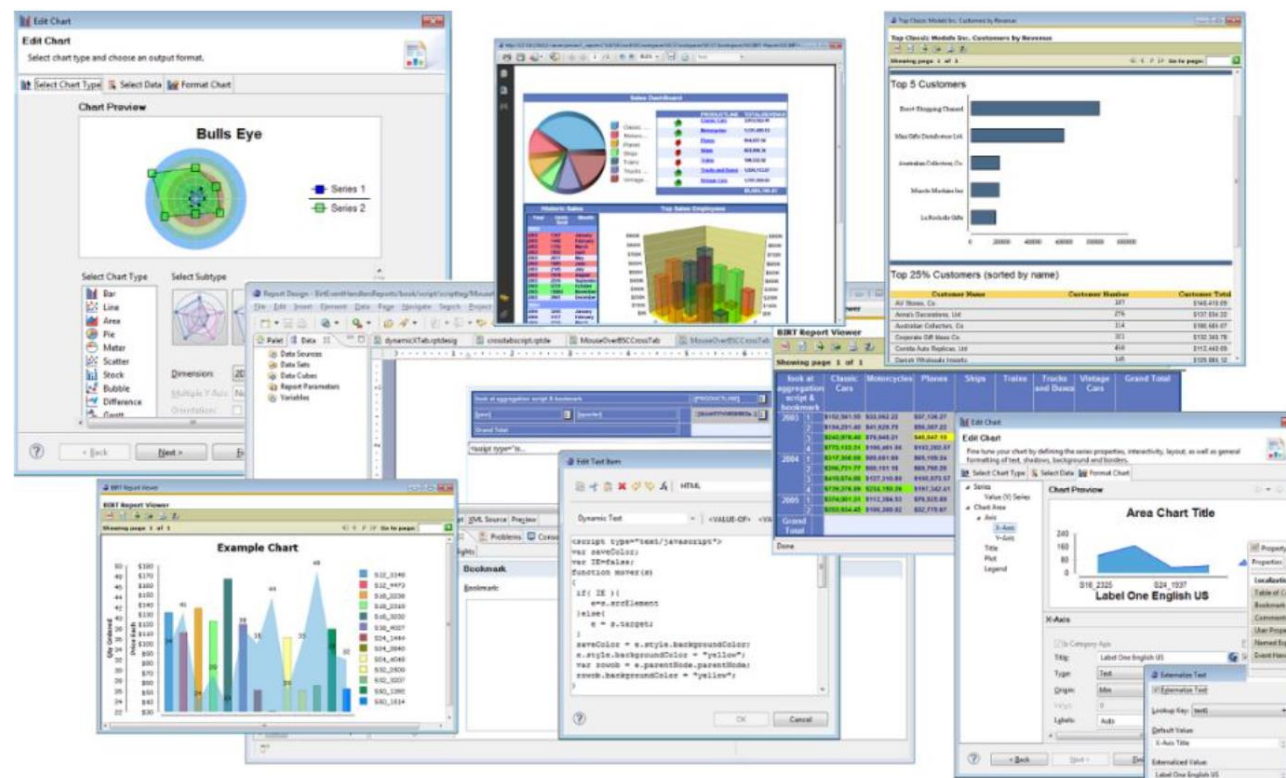
- Sau khi bạn đã có Job xử lý, tổng hợp dữ liệu điều cần làm tiếp theo là triển khai để chúng chạy định kỳ theo lịch, đó chính là Work Flow
- Hiện có một số Work Flow Open source phổ biến như Airflow, Ozie, ..
- Ngoài ra, trong các Flow phức tạp, bạn có thể sử dụng các công cụ ETL để tích hợp được nhiều tính năng hơn, ví dụ như Pentaho



CÁC ỨNG DỤNG KHAI THÁC

1. Hệ thống báo cáo Regular

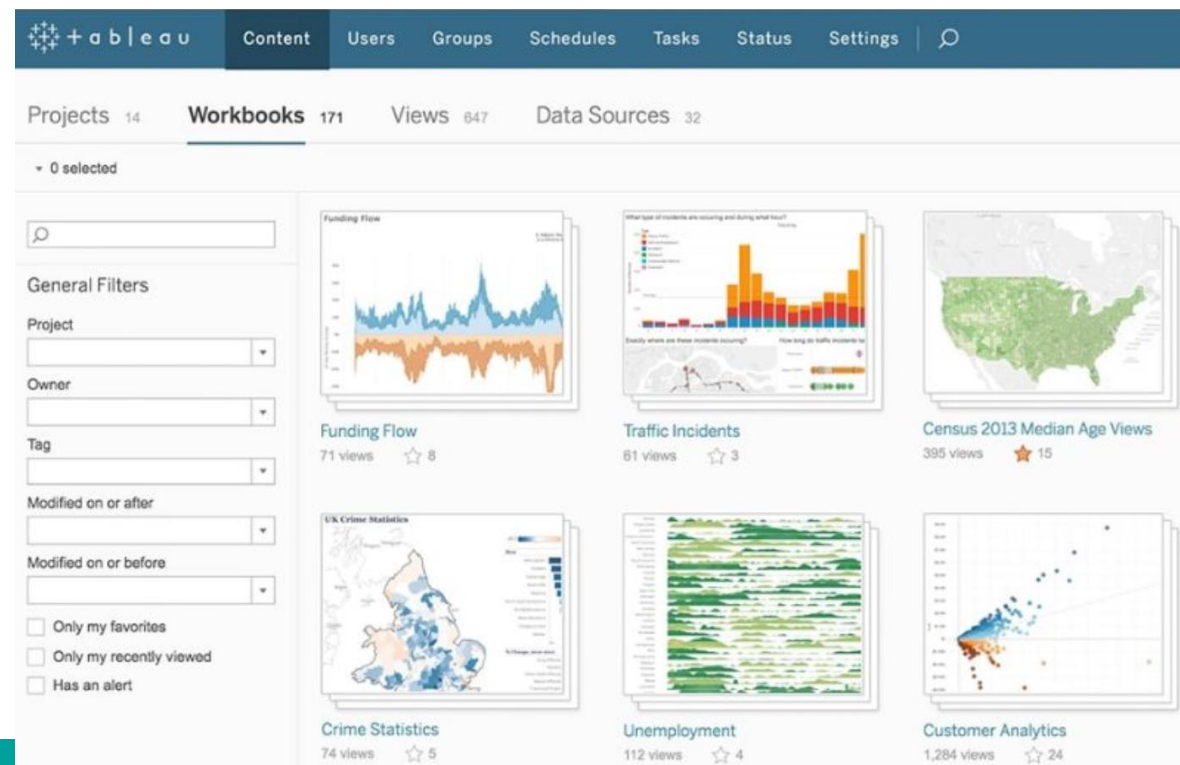
- Số liệu và báo cáo luôn gắn liền với nhau, phục vụ các nhu cầu thiết yếu trong điều hành của doanh nghiệp
- Các kho dữ liệu xây dựng, mục đích đầu tiên đều hướng tới việc đảm bảo cung cấp các số liệu cần thiết, định kỳ cho các đơn vị kinh doanh, nghiệp vụ.
- Bạn có thể sử dụng Birt Report, một Opensource tương đối dễ dàng trong việc triển khai công cụ liên quan tới báo cáo tĩnh:



CÁC ỨNG DỤNG KHAI THÁC

2. Hệ thống báo cáo động

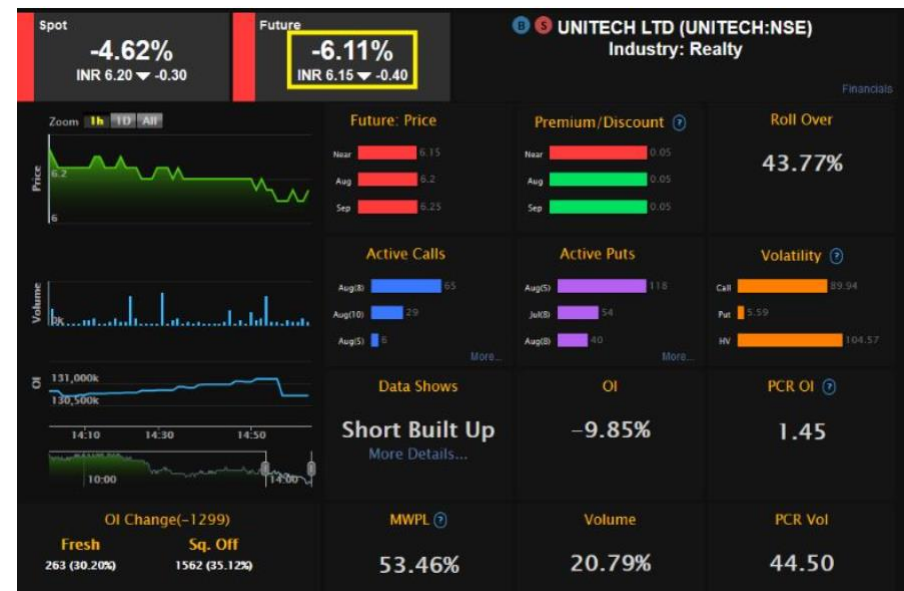
- Những báo cáo tĩnh cần một khoảng thời gian để xây dựng, chỉnh sửa, lại không cho người dùng được tương tác mà phụ thuộc hoàn toàn vào đội ngũ phát triển, lập trình viên
- Những hệ thống báo cáo động đã ra đời giúp kinh doanh có thể chủ động thực hiện các phân tích, đánh giá và tạo những view cho riêng mình
- Hiện với nhóm sản phẩm này, thì Tableau vẫn là một lựa chọn Top đầu nhưng lại không phải là một sản phẩm Open Source



CÁC ỨNG DỤNG KHAI THÁC

3. Hệ thống điều hành kinh doanh

- Đa số các hệ thống báo cung cấp số liệu dạng Offline, mang tính lịch sử và không hỗ trợ ra quyết định tức thời
- Khó khăn trong giao việc, kiểm soát các bất thường, chiến trình hay sản lượng thực hiện của từng nhân viên, đơn vị hay các bộ phận
- Hệ thống điều hành kinh doanh sẽ giải quyết các vấn đề này, giao việc tới từng nhân viên, cửa hàng hay các đơn vị, kiểm soát số liệu chỉ tiêu theo kế hoạch từng giờ, ngày
- Cảnh báo các bất thường doanh thu, sản lượng, cung cấp công cụ hỗ trợ chỉ đạo từ xa, ...
- Hiện nhóm sản phẩm này chưa có sản phẩm thương mại hay opensource, các công ty tự chủ động phát sinh theo nhu cầu của mình.

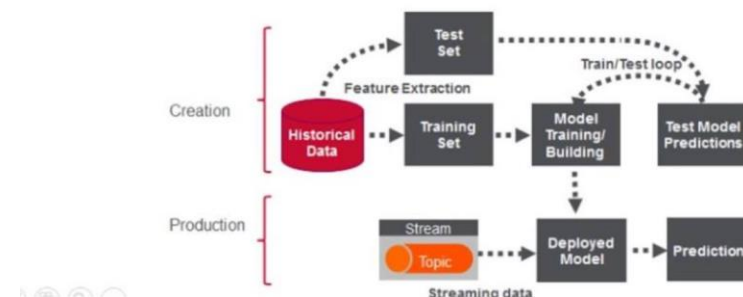


CÁC ỨNG DỤNG KHAI THÁC

4. Hệ thống kiểm soát gian lận

- Dữ liệu về các giao dịch được lưu đầy đủ trên Data Lake
- Các chương trình khuyến mại, tặng thưởng của khách hàng được lưu trữ đầy đủ bao gồm cả chi phí thanh toán với đối tác hay các giao dịch cưỡng,,
- Dựa vào các thông tin trên bạn có thể xây dựng một hệ thống kiểm soát gian lận
- Các hệ thống kiểm soát gian lận nhằm mục đích phát hiện sớm các Case rủi ro ảnh hưởng tới doanh thu, chi phí hay sản lượng
- Một số khách hàng thường tìm cách để được hưởng nhiều khuyến mãi, sử dụng một số công cụ hỗ trợ để gian lận thông tin,..việc xây dựng nhóm sản phẩm này là hết sức cần thiết đối với những công ty lớn có hàng triệu khách hàng
- Hiện sản phẩm này cũng chủ yếu do các công ty xây dựng và may đo theo nhu cầu của riêng mình

Credit Card Fraud Detection



CÁC ỨNG DỤNG KHAI THÁC

5. Hệ thống truyền thông

- Các chương trình khuyến mãi hay quảng cáo sẽ cần một hệ thống để quản lý vì:
 - + Mỗi khách sẽ có những hành vi, sở thích khác nhau bạn sẽ cần may đo một gói sản phẩm phù hợp
 - + Chiến dịch truyền thông cần được quản lý kết quả, đánh giá và điều hành liên tục để kịp có các điều chỉnh với thị trường
 - + Lựa chọn các khách hàng được tặng thưởng hay cần gìn giữ, chăm sóc
- Hiện sản phẩm này cũng chủ yếu do các công ty xây dựng và may đo theo nhu cầu của riêng mình



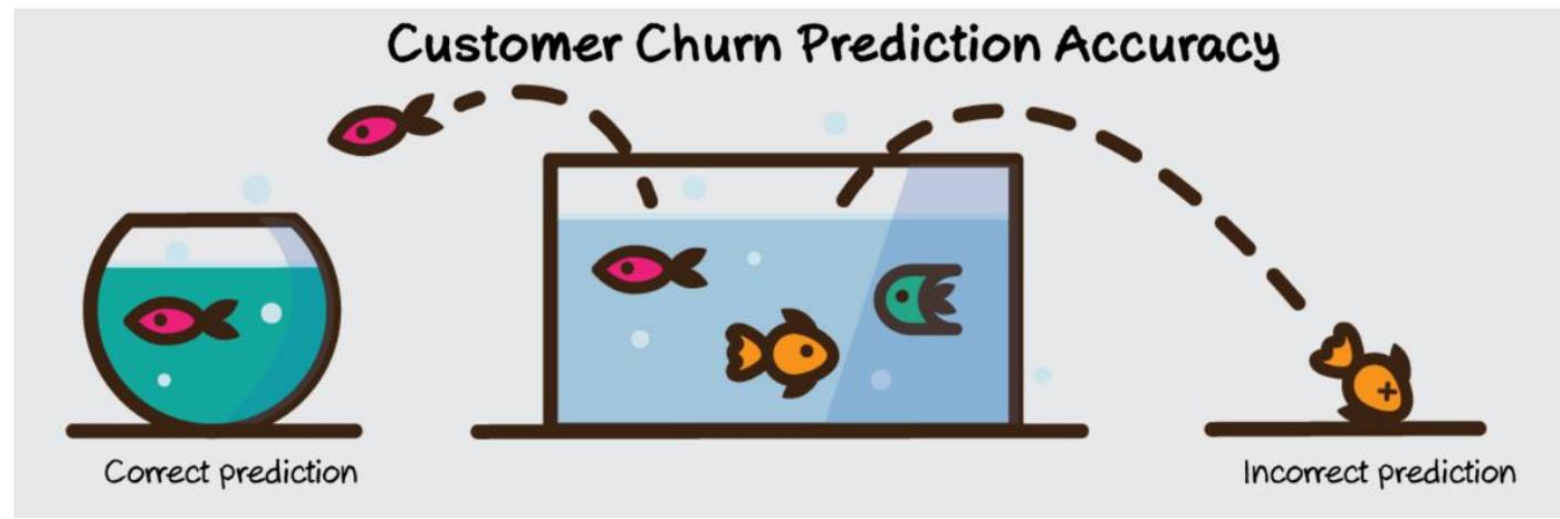
**What
is your
Strategy?**



CÁC ỨNG DỤNG KHAI THÁC

6. Hệ thống Phân tích dữ liệu

- Các hệ thống trên hầu hết điều hành ở hiện tại, bạn cần một hệ thống có khả năng dự đoán tương lai, hay sở thích của một khách trong tương lai ?
- Không thể dùng tới các usecase về phân tích dữ liệu
- Một số bài toán phổ biến hay ứng dụng như Churn Prediction (dự đoán khách hàng hủy), Upsell (kích thích tăng tiêu dùng) , Cross Sell (Bán chéo sản phẩm, mua sp A có thể thích sp



Hiệu năng một hệ thống Data Lake tại Việt Nam

1.Streaming Processing

Với luồng Streaming Processing, hệ thống xData Lake chia thành 3 nhóm:

- Realtime (xử lý < 1ph)
- Near Realtime (xử lý < 10ph)
- Micro Batch

a. Real time

- Phục vụ xử lý các sự kiện yêu cầu tức thời, ví dụ cảnh báo, sự cố, chiến dịch, độ phức tạp tương đối thấp khi xử lý
- Hệ thống sử dụng Spark Streaming, phần storage vật lý được thay thế toàn bộ là SSD và phần App chuyển từ HDFS sang Redis. Tốc độ xử lý đạt mức < 1s.

Hiệu năng một hệ thống Data Lake tại Việt Nam

b. Near Realtime

- Phục vụ xử lý các sự kiện yêu cầu có độ trễ, ví dụ tích lũy hành vi, đếm lượt sử dụng thỏa mãn một điều kiện nào đó, ..
- Hệ thống sử dụng Spark Streaming, phần storage vật lý được thay thế toàn bộ là SSD và phần App là HDFS, message queue sử dụng Kafka. Tốc độ xử lý đạt mức < 1ph.

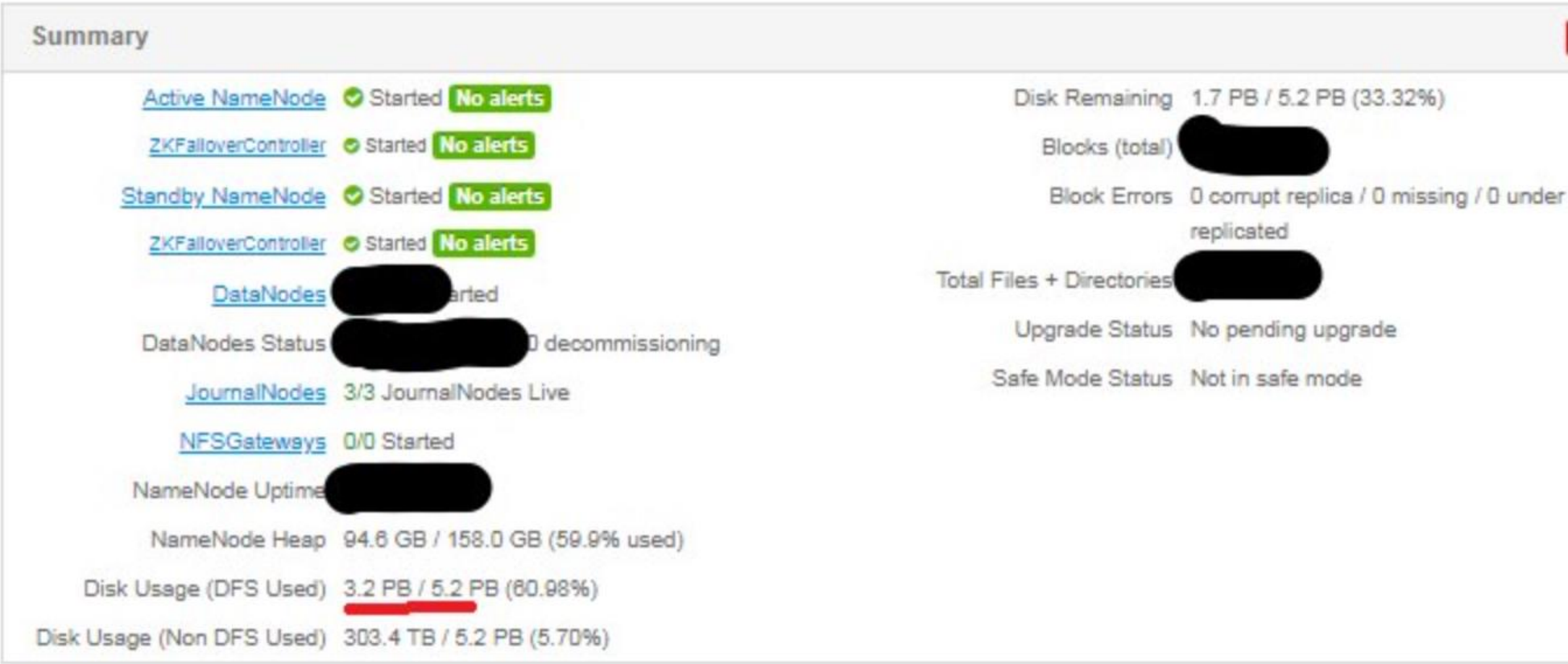
c. Micro Batch

- Phục vụ xử lý các sự kiện yêu cầu có độ trễ, ví dụ tích lũy hành vi so với ngày n-1, cộng dồn lũy kế tháng, số lượng bản lớn mức tỉ trở lên, ..
- Hệ thống sử dụng Spark Streaming, phần storage vật lý được thay thế toàn bộ là SSD và phần App là HDFS, message queue sử dụng Kafka. Tốc độ xử lý đạt mức < 15ph cho khối lượng bản ghi khoảng 50 – 75 tỷ / ngày.

Hiệu năng một hệ thống Data Lake tại Việt Nam

Cấu hình Cluster

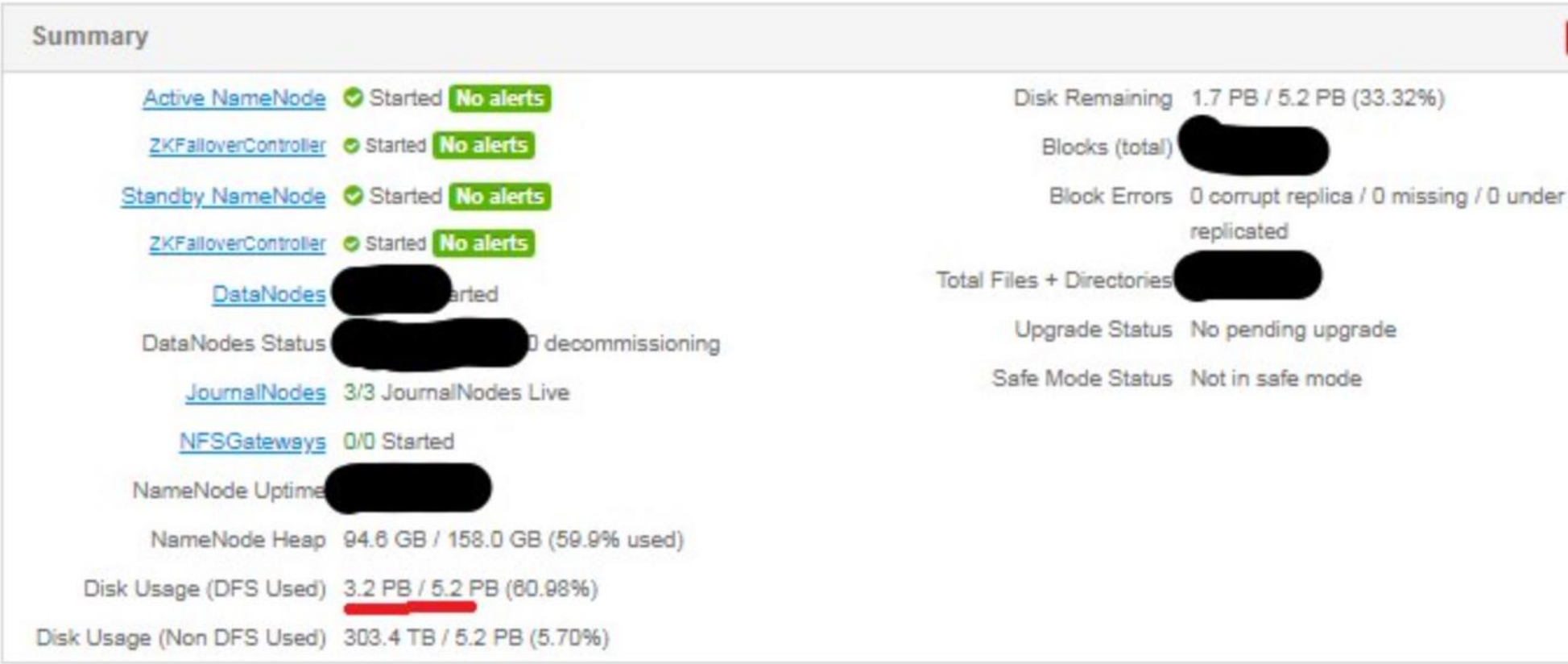
- x Data Lake có 5 Cluster, với những vai trò khác nhau, ví dụ Production, Lab, Dev, AI&ML, Streaming
- Trong đó có 2 Cluster lớn với cấu hình lần lượt:
 - + Cluster1: HDFS 5.5PB, RAM 13.34 TB, 4360 vCore, HDP 2.7



Hiệu năng một hệ thống Data Lake tại Việt Nam

Cấu hình Cluster

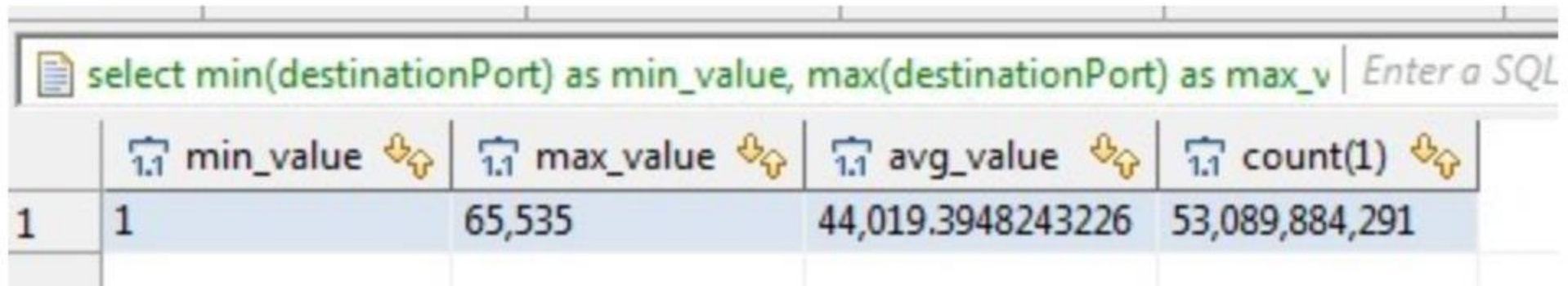
- x Data Lake có 5 Cluster, với những vai trò khác nhau, ví dụ Production, Lab, Dev, AI&ML, Streaming
- Trong đó có 2 Cluster lớn với cấu hình lần lượt:
 - + Cluster1: HDFS 5.5PB, RAM 13.34 TB, 4360 vCore, HDP 2.7



Hiệu năng một hệ thống Data Lake tại Việt Nam

3. Batch Processing

- Phần Batch Processing sử dụng Spark hoặc Hive SQL cho các luồng xử lý theo lô
- Tùy từng nghiệp vụ sẽ lựa chọn công cụ, tài nguyên phù hợp để đảm bảo yêu cầu cho bài toán đặt ra
- Hiện mình kiểm tra Aggregate (min, max, avg) một bảng có khối lượng 53 tỷ bản ghi và dung lượng khoảng vài TB thì mất 27s là ra kết quả



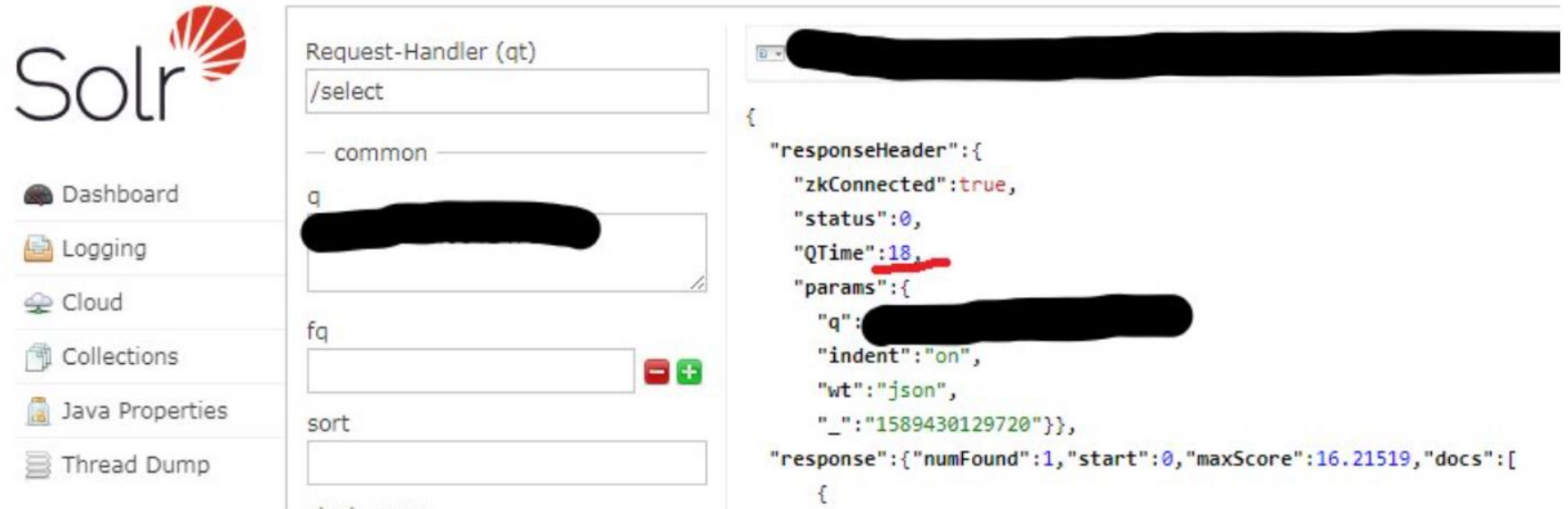
The screenshot shows a SQL query execution interface. At the top, there is a text input field containing the SQL query: `select min(destinationPort) as min_value, max(destinationPort) as max_v`. To the right of the query is a placeholder text "Enter a SQL". Below the query input, there is a table with 5 columns: "min_value", "max_value", "avg_value", and "count(1)". Each column header has a small icon of a document with a checkmark and a yellow arrow pointing up and down. The table has one data row with the following values: "1", "65,535", "44,019.3948243226", and "53,089,884,291".

	min_value	max_value	avg_value	count(1)
1	1	65,535	44,019.3948243226	53,089,884,291

Hiệu năng một hệ thống Data Lake tại Việt Nam

4. Search Engine

- X Data Lake sử dụng Solr cho phần Search Engine
- Phần storage sử dụng SSD và tối đa hóa việc Cache RAM để đảm bảo tốc độ
- Mình kiểm tra hiệu năng hiện một bảng hơn 80 triệu bản ghi, tìm một phần tử mất 18ms.



The screenshot displays the Solr Admin interface. On the left is a sidebar with navigation links: Dashboard, Logging, Cloud, Collections, Java Properties, and Thread Dump. The main area is titled 'Request-Handler (qt)' and contains several input fields: '/select' for the request handler, a 'common' section, a 'q' (query) field with a redacted value, an 'fq' (filter query) field, and a 'sort' field. To the right of the 'q' field is a redacted area. Below the 'fq' field are minus and plus icons. The 'sort' field is empty. On the far right, a JSON response is shown, with a redacted header. The visible JSON content includes:

```
{  
  "responseHeader": {  
    "zkConnected": true,  
    "status": 0,  
    "QTime": 18,  
    "params": {  
      "q": [REDACTED],  
      "indent": "on",  
      "wt": "json",  
      "_": "1589430129720" },  
    "response": { "numFound": 1, "start": 0, "maxScore": 16.21519, "docs": [
```

THỰC HÀNH

Thiết kế hệ thống Data Lake cho phép:

- Đồng bộ dữ liệu realtime từ hệ thống ghi log giao dịch
- Đồng bộ dữ liệu offline thông tin khách hàng
- Tổng hợp báo cáo ngày, lũy kế và tháng

Định cỡ:

- Hàng ngày có khoảng 100 triệu giao dịch ~ 50GB/ngày
- Có 10 triệu khách hàng ~ 7GB
- Thời gian lưu trữ dữ liệu chi tiết 1 năm
- Thời gian lưu trữ dữ liệu khách hàng 5 năm
- Thời gian lưu trữ dữ liệu tổng hợp 3 năm

Trân trọng cảm ơn!