

# FETEK

Delivering Excellence Data  
& Software Services





About Our

# DATA SERVICE

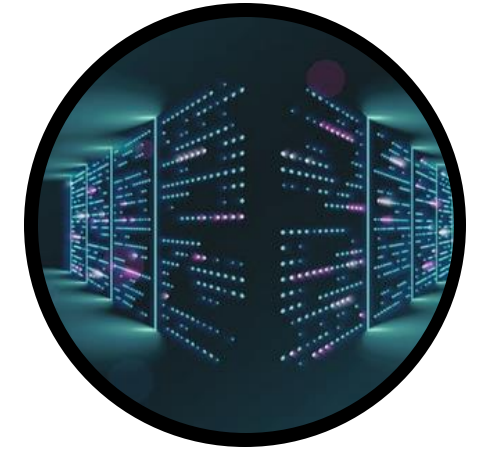
We are a dynamic, innovative company dedicated to providing top-notch software development services tailored to meet the unique needs of our clients. With a team of highly skilled professionals, we specialize in delivering cutting-edge solutions that drive business growth and efficiency.



## Data Transformation



## Report Development



## Data Lake & Lake House



## Data Product

Customer 360  
CRM System

Real-time Monitoring  
NOC PRO System

Campaign Marketing  
Campaign System

Reconciliation system  
RAFM

# MỤC LỤC

I

Tổng quan yêu cầu

II

Tổng quan giải pháp đề xuất

III

Quy trình triển khai

IV

Timeline triển khai setup hạ tầng

V

Danh sách tính năng chi tiết

VI

Demo trực quan hóa dữ liệu



## I. TỔNG QUAN YÊU CẦU

**559** CHỈ SỐ

### MỤC TIÊU

Xây dựng hệ thống báo cáo end-to-end toàn diện, đáp ứng nhu cầu báo cáo của tổ chức, hỗ trợ quá trình ra quyết định dựa trên dữ liệu chính xác và kịp thời.

**117** BÁO CÁO

### PHẠM VI

Hệ thống bao gồm các bước từ thu thập dữ, xử lý dữ liệu, lưu trữ, cho đến trực quan hóa thông tin qua các công cụ báo cáo hiện đại như Superset.

# LỢI ÍCH



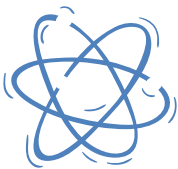
## **Nâng cao khả năng ra quyết định:**

Triển khai xây dựng hệ thống báo cáo không chỉ giảm tải cho đội ngũ kỹ thuật mà còn tăng tốc độ hiệu quả trong việc ra quyết định dựa trên dữ liệu. Điều này giúp tối ưu hóa quy trình kinh doanh và nâng cao khả năng cạnh tranh thông qua việc tận dụng dữ liệu hiệu quả



## **Khả năng giám sát và quản lý truy cập :**

Hệ thống hỗ trợ Role-Based Access Control (RBAC) để kiểm soát quyền truy cập vào dữ liệu, bảng báo cáo và các tài nguyên khác. Quản trị viên có thể dễ dàng chỉ định quyền xem hoặc chỉnh sửa báo cáo



## **Tự động cập nhật dữ liệu :**

Hệ thống có khả năng lập lịch, xử lý dữ liệu tự động và tạo báo cáo theo định kỳ hoặc sự kiện cụ thể ( VD: Hàng tuần, hàng tháng, hoặc khi có dữ liệu mới )



## **Khả năng tác động và chia sẻ :**

Hệ thống hỗ trợ việc chia báo cáo trực tiếp với người dùng hoặc nhóm thông qua giao diện web của superset người dùng có thể nhận phản hồi hoặc chỉnh sửa báo cáo nhanh chóng

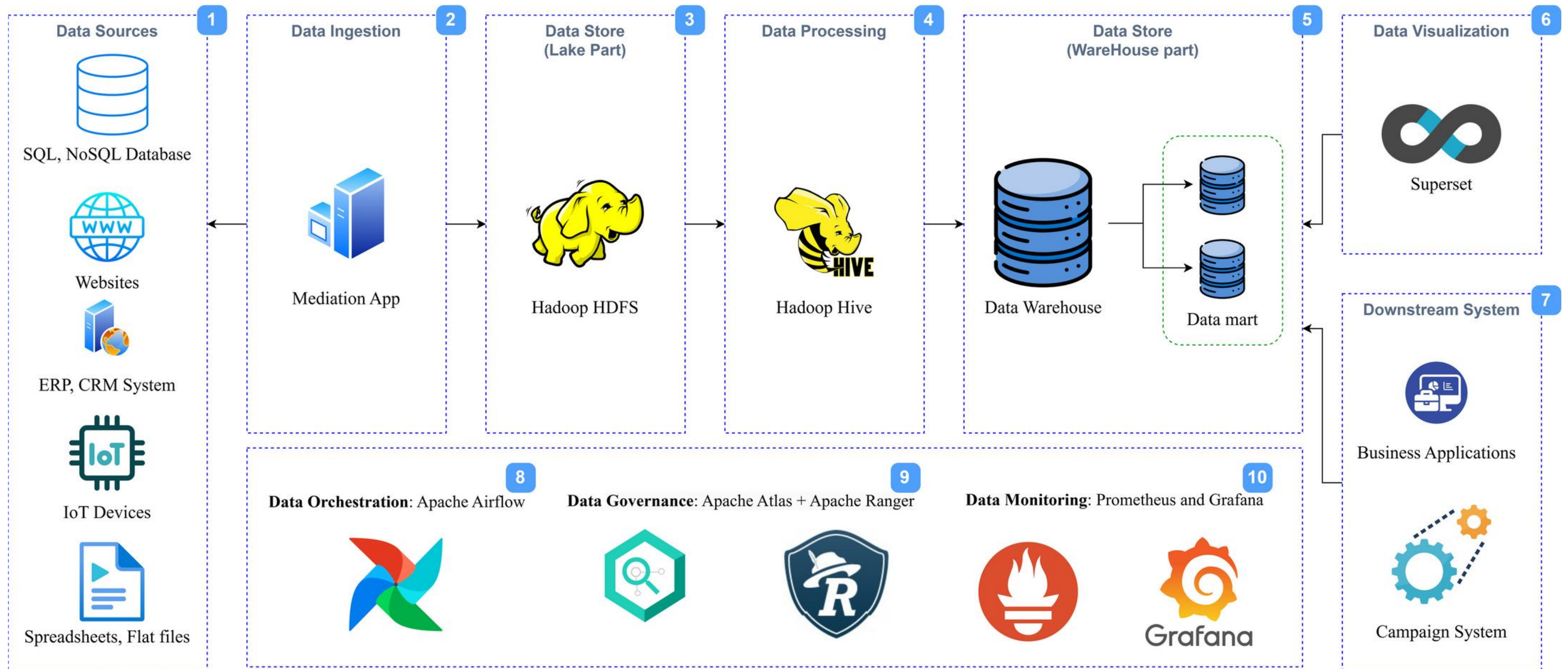


## **Khả năng tích hợp với nhiều nguồn dữ liệu :**

Sử dụng mediation giúp tăng khả năng tích hợp với nhiều nguồn dữ liệu khác nhau như cơ sở dữ liệu, API, Files để thu thập dữ liệu đầu vào giúp dễ dàng kết nối, khai thác và phân tích dữ liệu từ nhiều nguồn



## II. Tổng quan giải pháp đề xuất



## II. Tổng quan giải pháp đề xuất

01

### Data Sources:

Dữ liệu được lưu trữ từ nhiều nguồn khác nhau như hệ thống ERP, CRM, logs hệ thống, và các cảm biến IoT. Dữ liệu có thể có cấu trúc, phi cấu trúc, hoặc bán cấu trúc

02

### Data Ingestion:

Sử dụng Mediation App để trích xuất dữ liệu theo nhiều phương thức khác nhau (Batch Ingestion, Stream Ingestion, API ingestion, File Ingestion)

03

### Data Store (Lake Part):

Dữ liệu từ các nguồn sau khi được tích hợp sẽ được lưu trữ trong Data Lake (Hadoop HDFS). Đây là nơi chứa dữ liệu thô từ tất cả các nguồn mà không qua xử lý lớn, giữ lại toàn bộ thông tin ban đầu

04

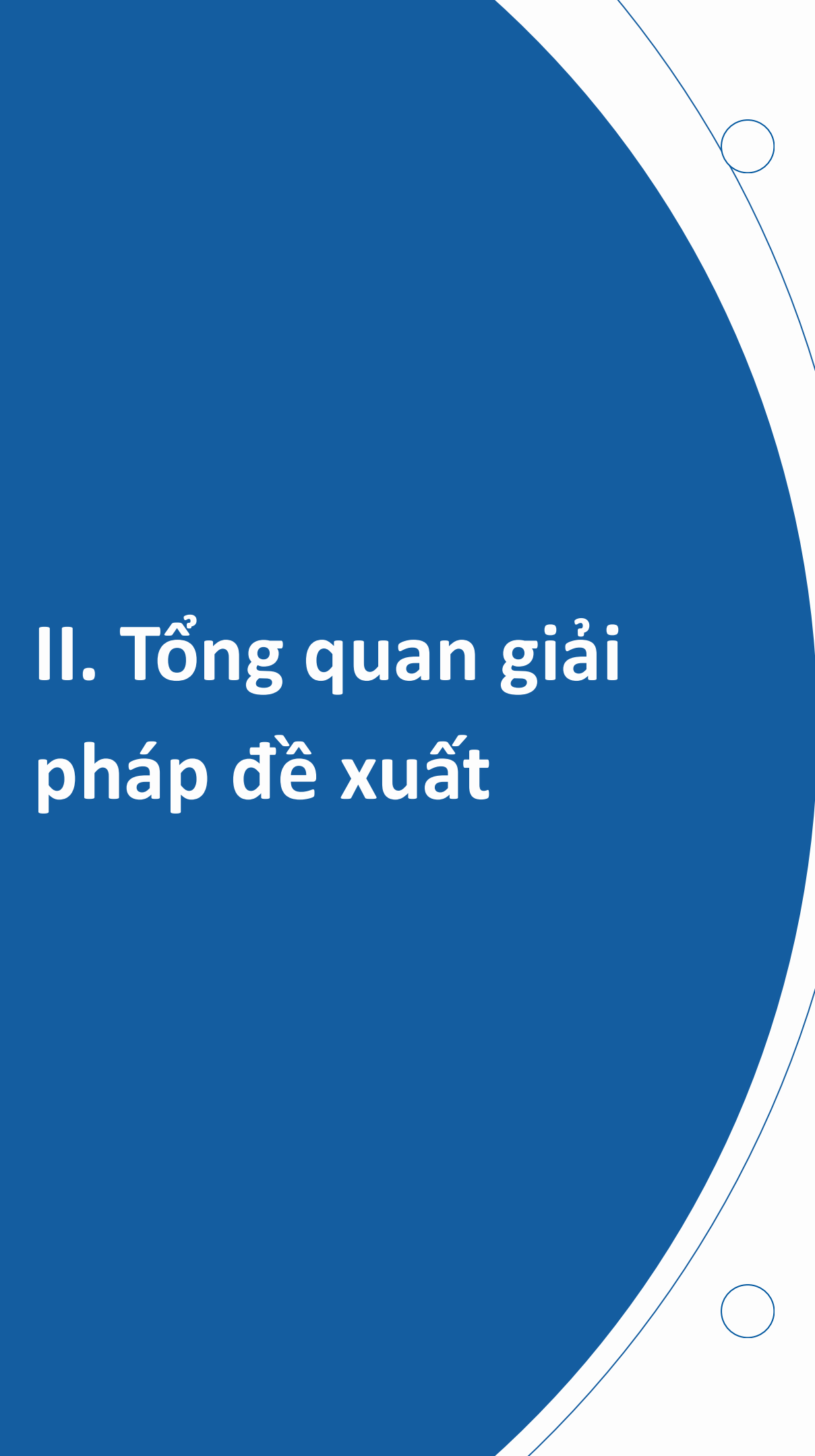
### Data Processing:

Hadoop Hive sẽ được sử dụng để xử lý, biến đổi và chuẩn hóa dữ liệu trước khi đưa vào Data Warehouse

05

### Data Store (Warehouse Part):

Dữ liệu đã qua xử lý sẽ được lưu trữ trong Data Warehouse (Oracle Database). Tùy theo nhu cầu sử dụng, dữ liệu có thể được phân tách thành các Data Mart để phục vụ các mục đích báo cáo và phân tích cụ thể



## II. Tổng quan giải pháp đề xuất

06

**Data Visualization:**

Dữ liệu được lưu trữ trong Data Warehouse sẽ được sử dụng để tạo ra các báo cáo, dashboard nhằm phục vụ cho việc phân tích và hỗ trợ ra quyết định

07

**Downstream System:**

Thay vì tập trung vào báo cáo và dashboard, dữ liệu từ Data Warehouse được sử dụng để phục vụ các hệ thống nghiệp vụ khác nhau của doanh nghiệp như Campaign System, BCCS, ...

08

**Data Orchestration:**

Airflow chịu trách nhiệm điều phối các task và quy trình trong toàn bộ kiến trúc, đảm bảo rằng các tiến trình từ trích xuất, xử lý, lưu trữ đến sử dụng dữ liệu diễn ra trơn tru và theo lịch trình

09

**Data Governance:**

Đảm bảo mọi hoạt động liên quan đến dữ liệu được quản lý theo các tiêu chuẩn về chất lượng, bảo mật và tuân thủ các quy định pháp lý

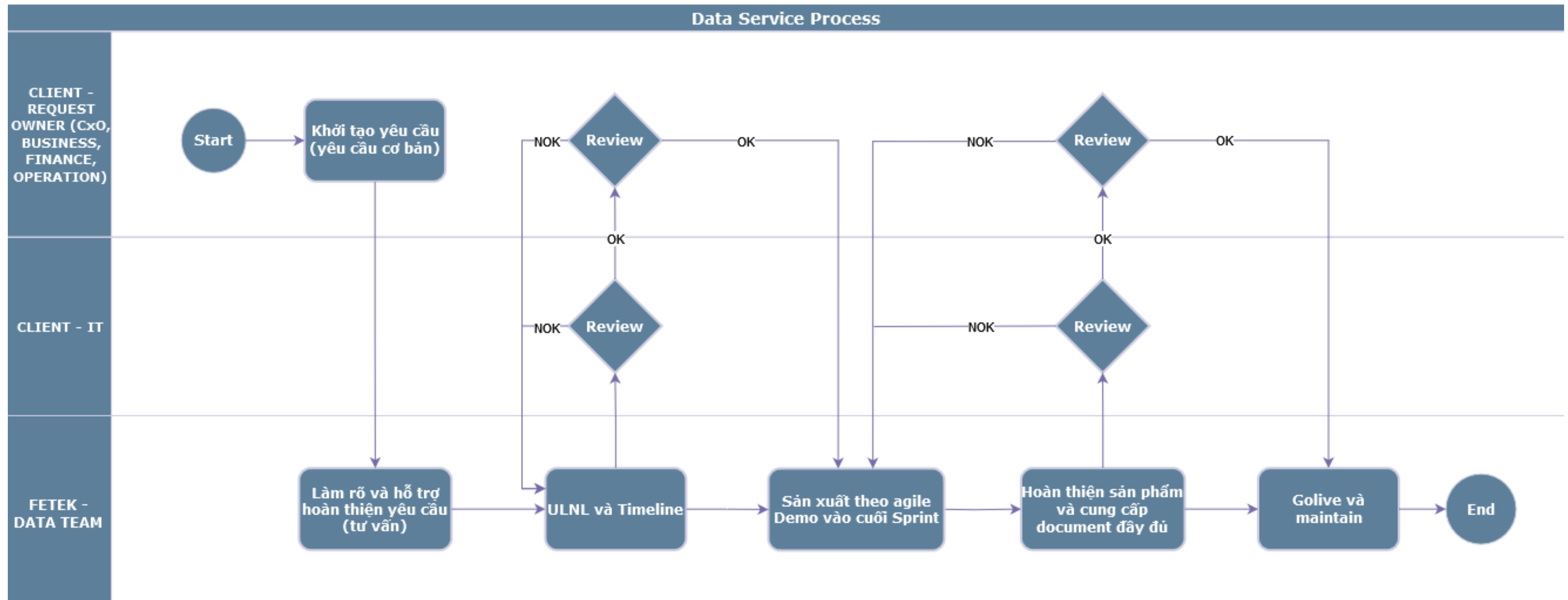
10

**Data Monitoring:**

Prometheus và Grafana chịu trách nhiệm theo dõi và ghi nhận toàn bộ hoạt động của hệ thống, từ hiệu suất xử lý đến việc giám sát các vấn đề kỹ thuật



### III. Quy trình triển khai



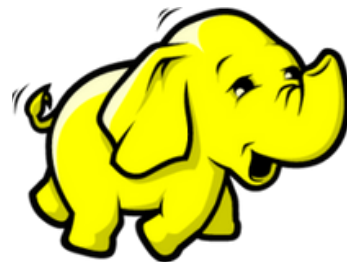
# IV. Timeline triển khai setup hạ tầng

Giai đoạn chuẩn bị		11/2024	12/2024	01/2025	02/2025
A.1	Hoàn thành thủ tục hợp đồng	A.1, A.2, A.3			
A.2	Cài đặt hạ tầng, ứng dụng				
A.3	Khảo sát chi tiết yêu cầu báo cáo				
Phát triển báo cáo					
B.1	Nhóm báo cáo về giao dịch và dịch vụ		B1		
B.2	Nhóm báo cáo về cân đối tài khoản, số dư và tài chính			B2	
B.3	Nhóm báo cáo về khuyến mãi, chương trình viên thông				B3, B4
B.4	Đào tạo self-serve, chuyển giao công nghệ				



We are here

## V. Danh sách tính năng chi tiết



Apache Atlas





# DATA INGESTION



## Mediation

### Quản lý kết nối:

Giao diện quản lý kết nối tập trung, cho phép kết nối đến nhiều loại data source:

- Batching Ingestion: hỗ trợ kết nối dữ liệu dạng database với nhiều loại database khác nhau (Oracle, MySQL, MSSQL...)
- File Ingestion: hỗ trợ kết nối dữ liệu dạng file (SFTP, SSH...)
- Streaming Ingestion: hỗ trợ kết nối dữ liệu dạng streaming near real time (Kafka, RabbitMQ...)
- API Ingestion: hỗ trợ kết nối dữ liệu dạng API (Rest, Soap...)

### Tích hợp dữ liệu:

Cho phép tích hợp dữ liệu đầu ra của hệ thống Mediation Gateway với nhiều database khác nhau, từ RDBMS (Oracle, MySQL...) đến No-SQL (Hadoop) đến API service (đối tác, vendor...)

### Xử lý dữ liệu:

- Cho phép cấu hình tham số cho từng bước xử lý dữ liệu (get, unzip, convert, filter, combine, push, load/write)
- Cấu hình DataFlow để đảm bảo thiết lập 01 luồng dữ liệu hoàn chỉnh (ví dụ một luồng mới cần get => convert => push => load hoặc get => unzip => filter => write)
- Cấu hình module đảm bảo chất lượng dữ liệu, ví dụ so sánh số lượng bản ghi get và convert...
- Cho phép xử lý dữ liệu near real time với các luồng cần mức độ SLA cao

# DATA INGESTION



## Mediation

### Vận hành dữ liệu:

Đảm bảo đầy đủ các tính năng vận hành dữ liệu:

- Quản lý toàn trình trên giao diện, cho phép cấu hình các tham số
- Quản lý và phân quyền người dùng theo các mức khác nhau (phát triển, vận hành, sử dụng dữ liệu), cho phép khai báo và phê duyệt luồng mới trên giao diện.
- Quản lý nguồn dữ liệu, cho phép tùy biến các tham số tần suất truy vấn (để đảm bảo hiệu năng cho nguồn), lượng dữ liệu transfer...
- Cảnh báo qua email/sms. Cấu phần xử lý lỗi
- Cho phép thực hiện với môi trường/dữ liệu test khi triển khai một luồng dữ liệu mới

### Khả năng mở rộng:

Cho phép mở rộng cả theo chiều ngang và chiều rộng

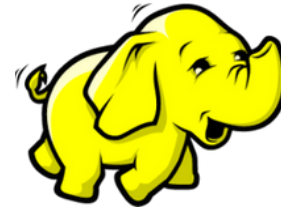
### Xử lý dữ liệu:

- Tích hợp các tính năng đảm bảo an toàn thông tin
- Mã hóa chuỗi kết nối
- Lọc các dữ liệu nhạy cảm trước khi load data vào các hệ thống phân tích như (cell, address...)
- Đảm bảo kết nối trên đường truyền (TLS, SSL...)

### Khả năng chịu lỗi:

- Đảm bảo tính phân tán (distributed, master/slave)
- Các cụm master cần có cơ chế active/active
- Tự động retry các luồng dữ liệu lỗi và đảm bảo chất lượng dữ liệu không trùng lặp (idempotent)

# DATA STORE



## Hadoop HDFS

### LƯU TRỮ DỮ LIỆU PHÂN TÁN

HDFS chia nhỏ dữ liệu thành các khối (blocks) và lưu trữ chúng trên nhiều máy chủ khác nhau, giúp đảm bảo độ tin cậy và khả năng phục hồi dữ liệu khi có sự cố phần cứng

### KHẢ NĂNG CHỊU LỖI

HDFS sao lưu mỗi khối dữ liệu trên nhiều nodes khác nhau (mặc định là 3 bản sao), đảm bảo dữ liệu vẫn truy cập được khi có sự cố ở một hoặc nhiều nút lưu trữ

### KHẢ NĂNG MỞ RỘNG

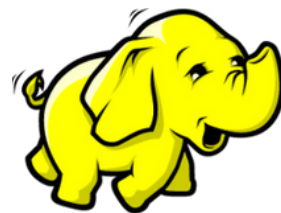
HDFS hỗ trợ mở rộng dễ dàng bằng cách thêm các Node mới vào hệ thống mà không cần dừng hoạt động, cho phép hệ thống quản lý khối lượng dữ liệu ngày càng lớn mà không ảnh hưởng đến hiệu năng

### TÍNH TOÀN VỆ DỮ LIỆU

HDFS thực hiện kiểm tra tính toàn vẹn của dữ liệu bằng cách sử dụng checksum, giúp phát hiện và khôi phục các khối dữ liệu bị lỗi hoặc bị hỏng trong quá trình lưu trữ và truy xuất.



# DATA STORE



## Hadoop HDFS

TRUY CẬP TỐC ĐỘ CAO	CƠ CHẾ DATA LOCALITY	HADOOP ARCHIVE (HAR)	BẢO MẬT VÀ QUYỀN TRUY CẬP
HDFS được tối ưu hóa để xử lý các tác vụ truy cập và ghi dữ liệu tốc độ cao, giúp tăng hiệu quả xử lý với các tập dữ liệu lớn, phù hợp với các ứng dụng như phân tích dữ liệu và xử lý dữ liệu lớn	HDFS tận dụng cơ chế Data Locality, tức là di chuyển các tiến trình xử lý dữ liệu đến nơi lưu trữ dữ liệu, thay vì di chuyển dữ liệu, giúp giảm tải băng thông mạng và tối ưu hóa hiệu suất xử lý	HAR là một tính năng giúp nén và lưu trữ các tập tin dữ liệu trong HDFS, giảm kích thước lưu trữ và chi phí, đồng thời vẫn đảm bảo khả năng truy xuất dữ liệu một cách dễ dàng khi cần	HDFS cung cấp các tính năng bảo mật cơ bản như quyền truy cập dựa trên dùng và nhóm, và hỗ trợ tích hợp với Kerberos để tăng cường bảo mật, quản lý quyền truy cập dữ liệu trong Data Lake một cách an toàn và có kiểm soát

# DATA PROCESSING



**Hadoop Hive**

## TRUY VẤN DỮ LIỆU BẰNG SQL

Hive cung cấp giao diện truy vấn với cú pháp giống SQL, giúp người dùng dễ dàng tương tác với dữ liệu lớn mà không cần viết code phức tạp

## XỬ LÝ DỮ LIỆU QUY MÔ LỚN

Hive được thiết kế như một giải pháp kho dữ liệu lớn, hỗ trợ xử lý hàng petabyte dữ liệu thông qua hệ thống phân tán của Hadoop

## PHÂN CHIA DỮ LIỆU (Partition)

Hive cho phép người dùng phân chia dữ liệu thành các phần nhỏ hơn dựa trên một hoặc nhiều khóa, giúp tối ưu hóa các truy vấn khi xử lý dữ liệu lớn

## NHÓM DỮ LIỆU (Bucket)

Bucketing giúp phân loại dữ liệu vào các nhóm nhỏ hơn dựa trên hàm băm (Hash Function), giúp tăng tốc truy vấn đối với các tập dữ liệu có phân phối không đều

# DATA PROCESSING



**Hadoop Hive**

## EXTERNAL TABLES

Hive hỗ trợ tạo bảng tham chiếu dữ liệu từ bên ngoài mà không di chuyển dữ liệu vào HDFS, thuận tiện cho việc quản lý dữ liệu đến từ nhiều nguồn.

## XỬ LÝ DỮ LIỆU ĐẶC BIỆT

Hive hỗ trợ cơ chế SerDe (Serializer/Deserializer) để xử lý các định dạng dữ liệu đặc biệt (như JSON, XML), giúp hệ thống dễ dàng tích hợp với các nguồn dữ liệu khác nhau.

## HIVE METASTORE

Hive Metastore lưu trữ thông tin metadata về cấu trúc bảng và các cột dữ liệu, giúp tăng cường khả năng truy cập nhanh và quản lý dữ liệu hiệu quả.

## HỖ TRỢ GIAO DỊCH ACID

Hive hỗ trợ các giao dịch ACID (Atomicity, Consistency, Isolation, Durability), đảm bảo tính toàn vẹn của dữ liệu khi có nhiều quá trình đọc/ghi đồng thời



# DATA STORE (WAREHOUSE PART)

## ORACLE DATABASE

### HIỆU SUẤT CAO

Oracle tối ưu hóa cho việc xử lý các truy vấn phức tạp trong Data Warehouse với công nghệ như Parallel Processing và Partitioning, giúp cải thiện hiệu suất truy xuất dữ liệu lớn, đảm bảo khả năng đáp ứng nhanh chóng các truy vấn lớn

### PHÂN CHIA DỮ LIỆU (Partition)

Oracle hỗ trợ phân vùng dữ liệu giúp chia nhỏ bảng và Index theo nhiều phương pháp (range, list, hash, composite) để tăng tốc độ truy vấn và quản lý dữ liệu hiệu quả hơn, giảm lượng dữ liệu cần quét khi thực hiện các truy vấn lớn trong Warehouse

### IN-MEMORY STORE

Oracle hỗ trợ công nghệ In-Memory giúp lưu trữ dữ liệu ở dạng cột thay vì dạng dòng, tối ưu cho các tác vụ analytics bằng cách tăng tốc độ truy xuất các cột dữ liệu lớn và giảm thời gian thực hiện các phép tính trên dữ liệu

### REAL APPLICATION CLUSTERS (RAC)

Oracle RAC cho phép Oracle chạy trên nhiều server (cluster) cùng một lúc, đảm bảo tính sẵn sàng cao HA và khả năng mở rộng ngang (horizontal scalability) cho các hệ thống Data Warehouse, đảm bảo không bị gián đoạn khi có sự cố xảy ra



# DATA STORE (WAREHOUSE PART)

## ORACLE DATABASE

### BẢO VỆ DỮ LIỆU

Oracle Database cung cấp các công nghệ bảo mật mạnh mẽ như TDE, Data Redaction, và VPD để bảo vệ dữ liệu nhạy cảm, đảm bảo tính bảo mật khi lưu trữ và truy cập dữ liệu trong môi trường Data Warehouse

### KHẢ NĂNG NÉN DỮ LIỆU TIÊN TIẾN

Oracle hỗ trợ nhiều cơ chế nén dữ liệu như Advanced Compression và Hybrid Columnar Compression giúp giảm kích thước lưu trữ của dữ liệu, tối ưu hóa không gian lưu trữ, đồng thời tăng tốc độ truy vấn nhờ vào việc giảm thiểu lượng dữ liệu cần xử lý.

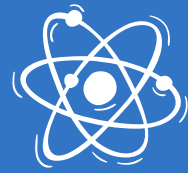
### KHẢ NĂNG MỞ RỘNG

Oracle có khả năng mở rộng linh hoạt theo nhu cầu về quy mô dữ liệu và khối lượng công việc, từ việc xử lý trên một máy chủ đơn lẻ đến hệ thống phân tán trên nhiều server

### PHÂN TÍCH DỮ LIỆU ĐA CHIỀU (OLAP)

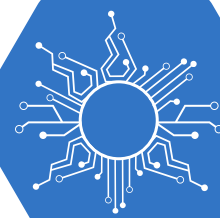
Oracle hỗ trợ tích hợp các công cụ phân tích dữ liệu đa chiều Oracle OLAP và các chức năng phân tích nâng cao Oracle Advanced Analytics, giúp phân tích sâu rộng trên dữ liệu Data Warehouse mà không cần di chuyển dữ liệu ra ngoài hệ thống





### KHẢ NĂNG TỰ TẠO VÀ QUẢN LÝ DASHBOARD

Người dùng có khả năng truy cập vào nguồn dữ liệu, tự chọn các chỉ số và biểu đồ phù hợp, cũng như sắp xếp các biểu đồ vào 1 Dashboard mà không cần có kiến thức lập trình



### TRUY CẬP VÀ XỬ LÝ DỮ LIỆU TỪ NHIỀU NGUỒN

Superset hỗ trợ kết nối với nhiều nguồn dữ liệu từ cơ sở dữ liệu quan hệ, Cloud đến các hệ thống dữ liệu phi cấu trúc. Người dùng có khả năng kết nối với dữ liệu mà họ có quyền truy cập, sau đó dễ dàng xử lý và tạo các bảng hoặc biểu đồ từ dữ liệu đó



### QUẢN LÝ TRUY CẬP VÀ PHÂN QUYỀN

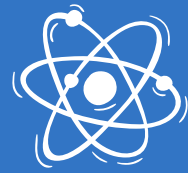
Superset tích hợp với hệ thống phân quyền Role-Based Access Control (RBAC), cho phép quản trị viên định nghĩa các quyền cụ thể cho từng người dùng hoặc nhóm người dùng. Điều này đảm bảo rằng chỉ những người có quyền mới có thể truy cập và tạo dashboard trên những nguồn dữ liệu cụ thể



### KHẢ NĂNG CHIA SẺ VÀ NHÚNG DASHBOARD

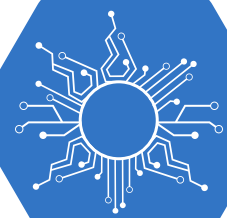
Superset cho phép người dùng chia sẻ dashboard với người khác thông qua liên kết hoặc mời đồng nghiệp trực tiếp trên hệ thống. Superset cũng hỗ trợ nhúng các biểu đồ hoặc dashboard vào các hệ thống bên ngoài thông qua iframe hoặc API





### KHẢ NĂNG MỞ RỘNG

Superset có kiến trúc modular, cho phép mở rộng và thêm mới các loại biểu đồ hoặc tích hợp các công cụ phân tích khác thông qua plugin. Nó cũng có thể dễ dàng mở rộng về mặt hiệu suất để hỗ trợ hàng trăm người dùng đồng thời



### HỖ TRỢ NHIỀU PLUGIN

Superset hỗ trợ các plugin tùy chỉnh, cho phép người dùng phát triển và tích hợp thêm các loại biểu đồ hoặc trực quan hóa khác ngoài những loại mặc định có sẵn



### SQL-LAB

SQL Lab cho phép người dùng trực tiếp viết và thực thi các truy vấn SQL trên dữ liệu, với khả năng xem kết quả tức thời, phù hợp cho các nhà phân tích dữ liệu muốn tùy chỉnh truy vấn chi tiết



### HIỆU NĂNG CAO VỚI DỮ LIỆU LỚN

Được tối ưu hóa để xử lý các tập dữ liệu lớn với khả năng chạy truy vấn trên hàng triệu bản ghi mà vẫn đảm bảo hiệu suất cao và thời gian phản hồi nhanh chóng

# DATA ORCHESTRATION



## Quản lý luồng công việc theo DAG

Airflow sử dụng mô hình Directed Acyclic Graphs (DAGs) để định nghĩa workflow. Mỗi DAG biểu diễn một chuỗi các tác vụ được thực hiện theo thứ tự đã định, giúp tổ chức các luồng công việc phức tạp một cách rõ ràng và trực quan



## Khả năng quản lý và lập lịch thực thi Task

Airflow hỗ trợ lập lịch tự động cho các luồng công việc theo thời gian định trước, cho phép chạy các Task định kỳ theo lịch trình mong muốn. Người dùng có thể quản lý lịch chạy mà không cần sự phê duyệt từ quản trị viên



## Quản lý phụ thuộc giữa các Task

Airflow hỗ trợ xác định rõ ràng sự phụ thuộc giữa các tác vụ, đảm bảo rằng một tác vụ chỉ chạy khi các tác vụ phụ thuộc trước đó đã hoàn thành, giúp quản lý luồng công việc một cách hiệu quả



## Khả năng mở rộng

Airflow có kiến trúc phân tán, dễ dàng mở rộng quy mô. Nó cũng hỗ trợ việc sử dụng nhiều worker và có khả năng phân phối task một cách hiệu quả. Đồng thời, Airflow cho phép tổ chức quản lý workflow dưới dạng mã nguồn, giúp dễ dàng tích hợp với hệ thống CI/CD

# DATA ORCHESTRATION



## Phân quyền vào bảo mật

Airflow hỗ trợ tích hợp với hệ thống quản lý người dùng (như LDAP, OAuth) và phân quyền truy cập DAG dựa trên vai trò. Điều này cho phép tổ chức quy định người nào có thể xem, chỉnh sửa, hay thực thi các DAG khác nhau



## Khả năng giám sát và khắc phục sự cố

Giao diện web của Airflow cung cấp log chi tiết cho từng task, trạng thái chạy, và khả năng khởi động lại workflow khi task thất bại. Người dùng cũng có thể thiết lập số lần retry và xử lý lỗi tự động



## Khả năng xử lý ngược (Backfill)

Tính năng backfill giúp thực thi lại các DAG trong quá khứ mà chưa được chạy, đảm bảo rằng không có dữ liệu hoặc tác vụ nào bị bỏ sót khi có các điều kiện đặc biệt như server down hoặc thay đổi lịch trình



## Cảnh báo vào thông báo

Airflow có thể được thiết lập để gửi thông báo qua email, Slack, hoặc các hệ thống cảnh báo khác khi một tác vụ hoàn thành hoặc thất bại, giúp người quản lý hệ thống luôn cập nhật được trạng thái của các luồng công việc

# DATA GOVERNANCE



## Apache Atlas

QUẢN LÝ METADATA	THEO DÕI LUỒNG DỮ LIỆU	PHÂN LOẠI VÀ GÁN NHÃN DỮ LIỆU	POLICY ENFORCEMENT
Apache Atlas cho phép quản lý và tổ chức metadata của các tập dữ liệu trong hệ thống, cung cấp một bức tranh tổng quan về các nguồn dữ liệu, định dạng và mối quan hệ giữa chúng	Tính năng này giúp theo dõi và hiển thị đường đi của dữ liệu từ nguồn gốc đến nơi tiêu thụ, giúp kiểm soát và hiểu rõ cách dữ liệu di chuyển và biến đổi trong toàn bộ hệ sinh thái dữ liệu.	Cho phép gán các thẻ (tag) và phân loại cho các tập dữ liệu để quản lý tốt hơn, đồng thời giúp dễ dàng tìm kiếm và áp dụng các quy tắc quản trị cụ thể cho từng loại dữ liệu khác nhau	Atlas tích hợp với Apache Ranger để đảm bảo các chính sách quản lý dữ liệu được thực thi chính xác, giúp bảo vệ dữ liệu nhạy cảm và đảm bảo tuân thủ quy định.



# DATA GOVERNANCE



## Apache Atlas

BẢO MẬT VÀ QUẢN LÝ QUYỀN TRUY CẬP	TÍCH HỢP VỚI HỆ SINH THÁI HADOOP	DATA GOVERNANCE FRAMEWORK	THEO DÕI VÀ ĐẢM BẢO TUÂN THỦ
Atlas hỗ trợ quản lý quyền truy cập vào các đối tượng dữ liệu dựa trên vai trò và chính sách, giúp bảo vệ quyền riêng tư của dữ liệu và ngăn ngừa truy cập trái phép.	Apache Atlas tích hợp với các công cụ khác trong hệ sinh thái Hadoop như Hive, HBase, Kafka để theo dõi và quản lý metadata toàn diện trong môi trường dữ liệu lớn.	Apache Atlas cung cấp một khung quản trị dữ liệu đầy đủ với khả năng tùy chỉnh và mở rộng, giúp các tổ chức xây dựng các chính sách và quy trình quản trị dữ liệu một cách linh hoạt và dễ dàng.	Tính năng này ghi lại các hành động liên quan đến dữ liệu và cung cấp khả năng báo cáo để đảm bảo rằng các quy định pháp lý và yêu cầu tuân thủ dữ liệu được thực hiện chính xác.

# DATA GOVERNANCE



**Apache Ranger**

## CẬP NHẬT CHÍNH SÁCH ĐỘNG

Ranger cho phép cập nhật và áp dụng các chính sách truy cập dữ liệu mà không cần khởi động lại hệ thống, giúp đảm bảo tính liên tục và bảo mật của hệ thống dữ liệu.

## CHÍNH SÁCH DỰA TRÊN NHÃN DỮ LIỆU

Tính năng này cho phép quản lý và thực thi các chính sách truy cập dựa trên các tags được gán cho dữ liệu, giúp bảo mật dữ liệu dễ dàng hơn với các quy tắc phân loại và kiểm soát quyền truy cập dựa trên nhãn

## QUẢN LÝ USER VÀ GROUP

Hive Metastore lưu trữ thông tin metadata về cấu trúc bảng và các cột dữ liệu, giúp tăng cường khả năng truy cập nhanh và quản lý dữ liệu hiệu quả.

## ẢN DỮ LIỆU NHẠY CẢM

Hive hỗ trợ các giao dịch ACID (Atomicity, Consistency, Isolation, Durability), đảm bảo tính toàn vẹn của dữ liệu khi có nhiều quá trình đọc/ghi đồng thời

# DATA GOVERNANCE



**Apache Ranger**

## ACCESS CONTROL POLICIES

Apache Ranger cho phép quản lý chi tiết quyền truy cập dữ liệu cho từng User hoặc Group dựa trên các chính sách truy cập cụ thể, giúp bảo vệ dữ liệu nhạy cảm

## QUẢN LÝ BẢO MẬT TẬP TRUNG

Ranger cung cấp bảng điều khiển tập trung để quản lý các chính sách bảo mật cho toàn bộ hệ sinh thái Hadoop, giúp đơn giản hóa việc theo dõi và kiểm soát bảo mật trong các môi trường phân tán.

## AUDIT LOGGING

Apache Ranger ghi lại toàn bộ các hành động truy cập dữ liệu của người dùng, bao gồm việc thực thi chính sách, đảm bảo rằng tất cả các hoạt động truy cập đều có thể theo dõi và kiểm tra khi cần thiết.

## PHÂN QUYỀN THEO TÁC VỤ

Ranger cung cấp cơ chế phân quyền chi tiết cho từng tác vụ (như đọc, ghi, xóa) đối với từng đối tượng dữ liệu, cho phép kiểm soát chính xác và linh hoạt đối với dữ liệu.



**Time-Series Database:** Prometheus lưu trữ các số liệu (metrics) dưới dạng chuỗi thời gian, mỗi chuỗi thời gian bao gồm các giá trị được ghi nhận theo thời gian thực, giúp dễ dàng theo dõi các biến động trong hệ thống.



**Thu thập số liệu theo cơ chế Pull:** Prometheus chủ động truy xuất số liệu từ các endpoint theo một lịch trình định trước, giảm tải cho các hệ thống theo dõi và đảm bảo rằng chỉ các số liệu mới nhất được thu thập.



**PromQL:** PromQL cho phép truy vấn và phân tích số liệu phức tạp từ cơ sở dữ liệu chuỗi thời gian của Prometheus, hỗ trợ các phép tính như trung bình, tổng, tỉ lệ và các hàm xử lý dữ liệu khác.



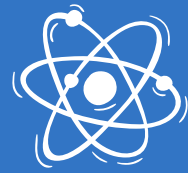
**Tích hợp hệ thống cảnh báo:** Prometheus tích hợp với Alertmanager để thiết lập và gửi cảnh báo khi một số điều kiện nhất định được thỏa mãn, giúp người quản lý hệ thống nhận thông báo ngay khi có sự cố.



**Service Discovery:** Prometheus hỗ trợ khám phá tự động các dịch vụ và endpoint cần giám sát, giúp giảm thiểu việc phải cấu hình thủ công, đặc biệt hữu ích trong các môi trường động như Kubernetes.

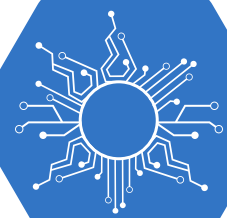






## HỖ TRỢ CÁC EXPORTERS

Các exporter của Prometheus giúp thu thập số liệu từ các hệ thống khác như hệ điều hành, cơ sở dữ liệu, ứng dụng web, cung cấp khả năng giám sát rộng rãi cho nhiều loại môi trường.



## HIGH AVAILABILITY

Prometheus có thể được triển khai trong môi trường phân tán để đảm bảo tính sẵn sàng cao, với các instance backup có khả năng thu thập và lưu trữ số liệu nếu instance chính gặp sự cố.



## CHÍNH SÁCH LƯU TRỮ DỮ LIỆU

Prometheus cung cấp các tùy chọn để cấu hình thời gian lưu trữ dữ liệu, giúp tối ưu hóa việc quản lý dung lượng lưu trữ trong hệ thống mà vẫn duy trì đủ dữ liệu để phân tích.



## RELABELING

Tính năng relabeling cho phép thay đổi nhãn (label) của số liệu trước khi lưu trữ, giúp tổ chức và xử lý các số liệu theo nhu cầu giám sát cụ thể của từng môi trường hoặc ứng dụng.



# DATA MONITORING



## CUSTOM DASHBOARD

HDFS chia nhỏ dữ liệu thành các khối (blocks) và lưu trữ chúng trên nhiều máy chủ khác nhau, giúp đảm bảo độ tin cậy và khả năng phục hồi dữ liệu khi có sự cố phần cứng

## MULTI-SOURCE DATA SUPPORT

Grafana không chỉ hỗ trợ Prometheus mà còn tích hợp với nhiều hệ thống khác như MySQL, Elasticsearch, InfluxDB, giúp giám sát tập trung dữ liệu từ nhiều nguồn trong một giao diện duy nhất

## INTERACTIVE VISUALIZATION

Người dùng có thể tương tác với các biểu đồ, ví dụ như zoom, lọc, thay đổi khoảng thời gian trực tiếp trên dashboard, giúp việc khám phá và phân tích dữ liệu trở nên dễ dàng hơn

## ALERTING INTEGRATION

Grafana cho phép thiết lập cảnh báo dựa trên các ngưỡng hoặc điều kiện nhất định, giúp theo dõi các chỉ số quan trọng và gửi thông báo qua email, Slack, PagerDuty hoặc các kênh khác

# DATA MONITORING



## CUSTOM DASHBOARD

Grafana cho phép tạo bảng điều khiển trực quan với các biểu đồ và widget, giúp người dùng có thể dễ dàng theo dõi và hiển thị các số liệu quan trọng từ Prometheus hoặc nhiều nguồn dữ liệu khác

## HỖ TRỢ NHIỀU NGUỒN DỮ LIỆU

Grafana không chỉ hỗ trợ Prometheus mà còn tích hợp với nhiều hệ thống khác như MySQL, Elasticsearch, InfluxDB, giúp giám sát tập trung dữ liệu từ nhiều nguồn trong một giao diện duy nhất

## TRỰC QUAN HÓA DỮ LIỆU TƯƠNG TÁC

Người dùng có thể tương tác với các biểu đồ, ví dụ như zoom, lọc, thay đổi khoảng thời gian trực tiếp trên dashboard, giúp việc khám phá và phân tích dữ liệu trở nên dễ dàng hơn

## TÍCH HỢP HỆ THỐNG CẢNH BÁO

Grafana cho phép thiết lập cảnh báo dựa trên các ngưỡng hoặc điều kiện nhất định, giúp theo dõi các chỉ số quan trọng và gửi thông báo qua email, Slack, PagerDuty hoặc các kênh khác

# DATA MONITORING



## TEMPLATING

Tính năng templating của Grafana giúp tạo các dashboard động, cho phép thay đổi các tham số đầu vào (như tên server, dịch vụ, hoặc vùng dữ liệu) để hiển thị các số liệu khác nhau mà không cần tạo mới hoàn toàn.

## ANNOTATIONS

Người dùng có thể thêm chú thích trực tiếp lên các biểu đồ để đánh dấu các sự kiện quan trọng, giúp dễ dàng theo dõi và phân tích nguyên nhân của các biến động trong dữ liệu.

## TEAMS & PERMISSIONS

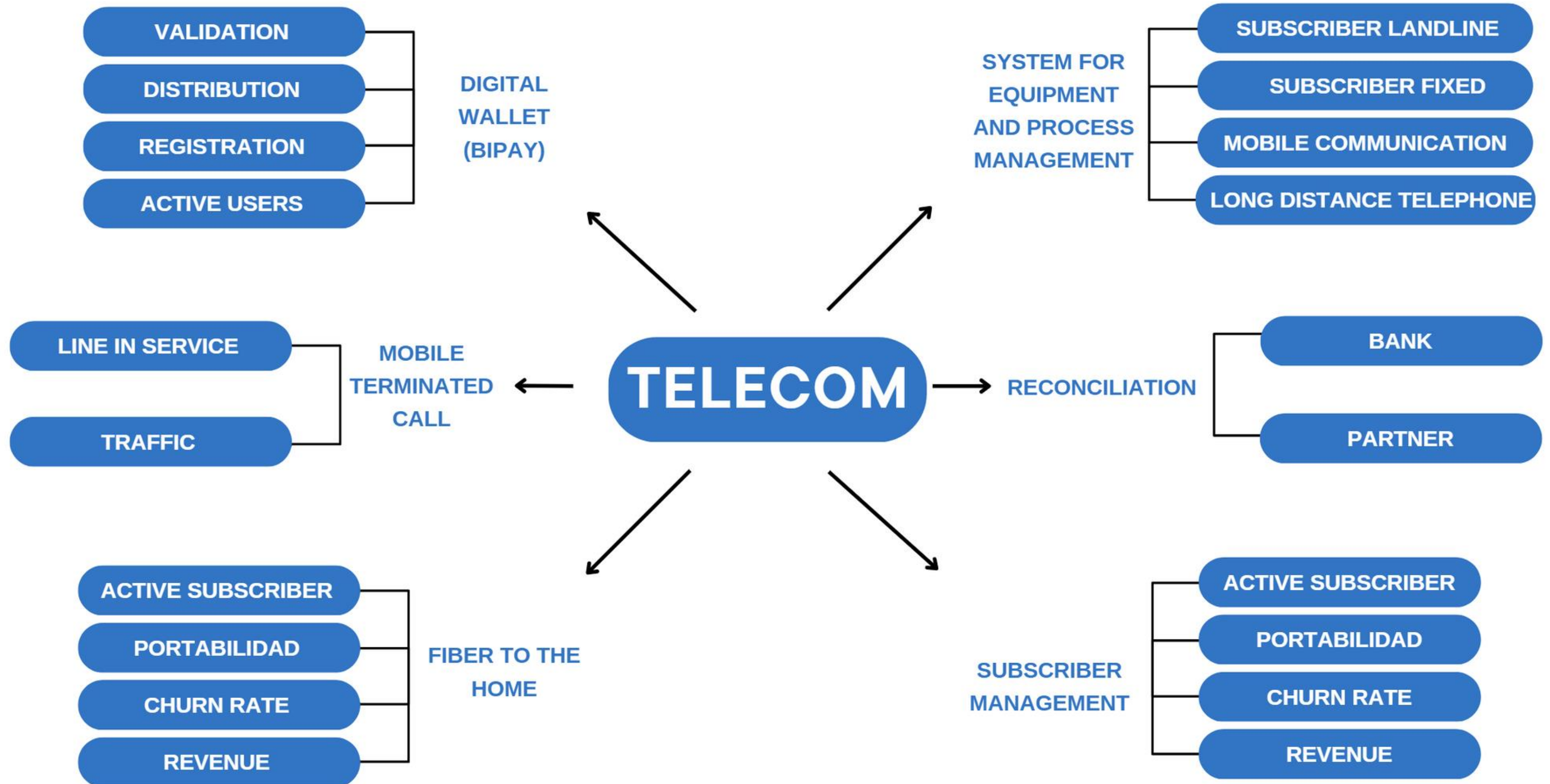
Grafana hỗ trợ quản lý người dùng và quyền truy cập, cho phép phân quyền cho từng nhóm hoặc cá nhân để truy cập hoặc chỉnh sửa các bảng điều khiển, đảm bảo tính bảo mật và kiểm soát trong tổ chức.

## SNAPSHOT SHARING

Grafana cho phép người dùng tạo và chia sẻ snapshot của các dashboard dưới dạng ảnh tĩnh hoặc liên kết có thể chia sẻ, giúp dễ dàng trao đổi thông tin mà không yêu cầu người nhận có quyền truy cập hệ thống.

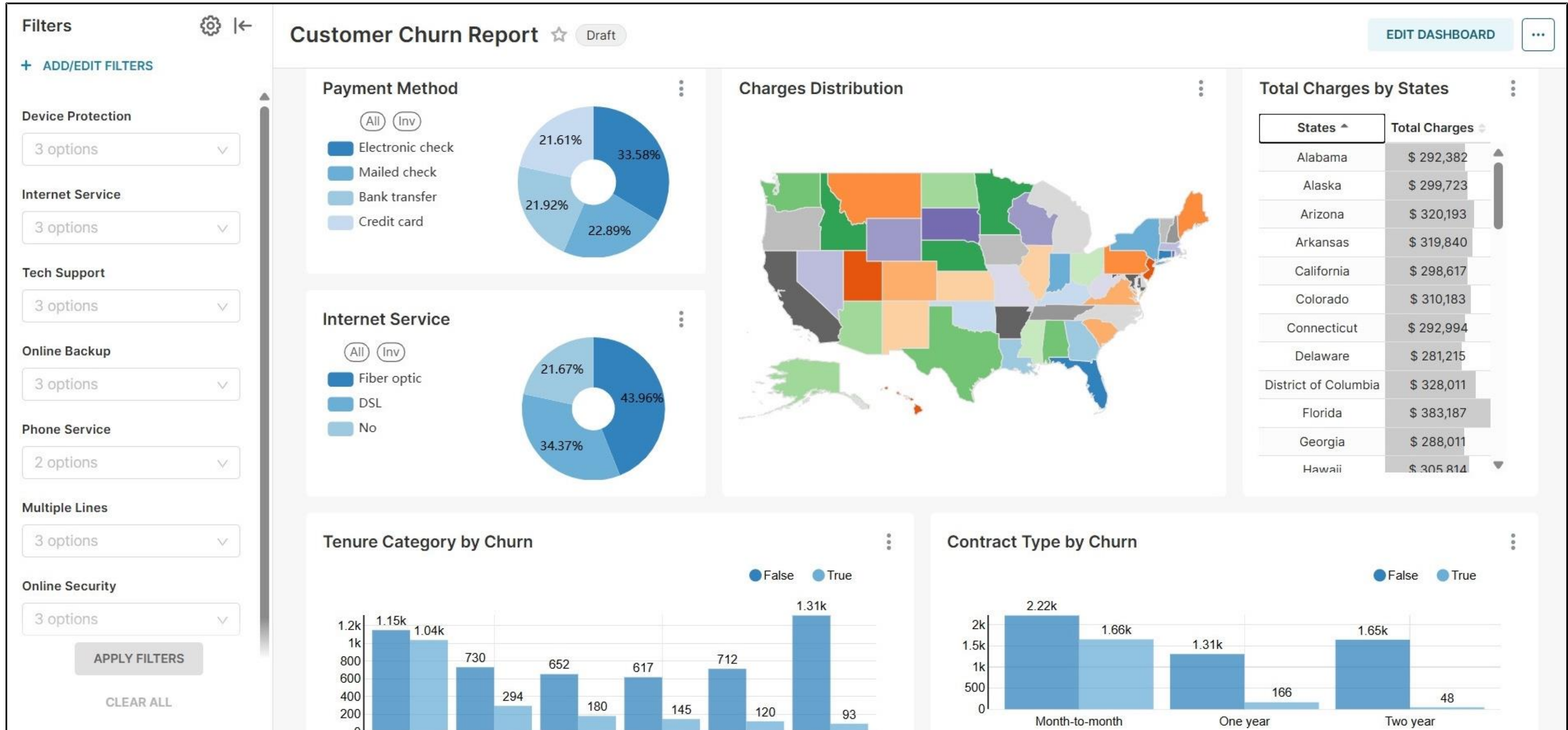


# REPORT LIST FOR TELECOMMUNICATION



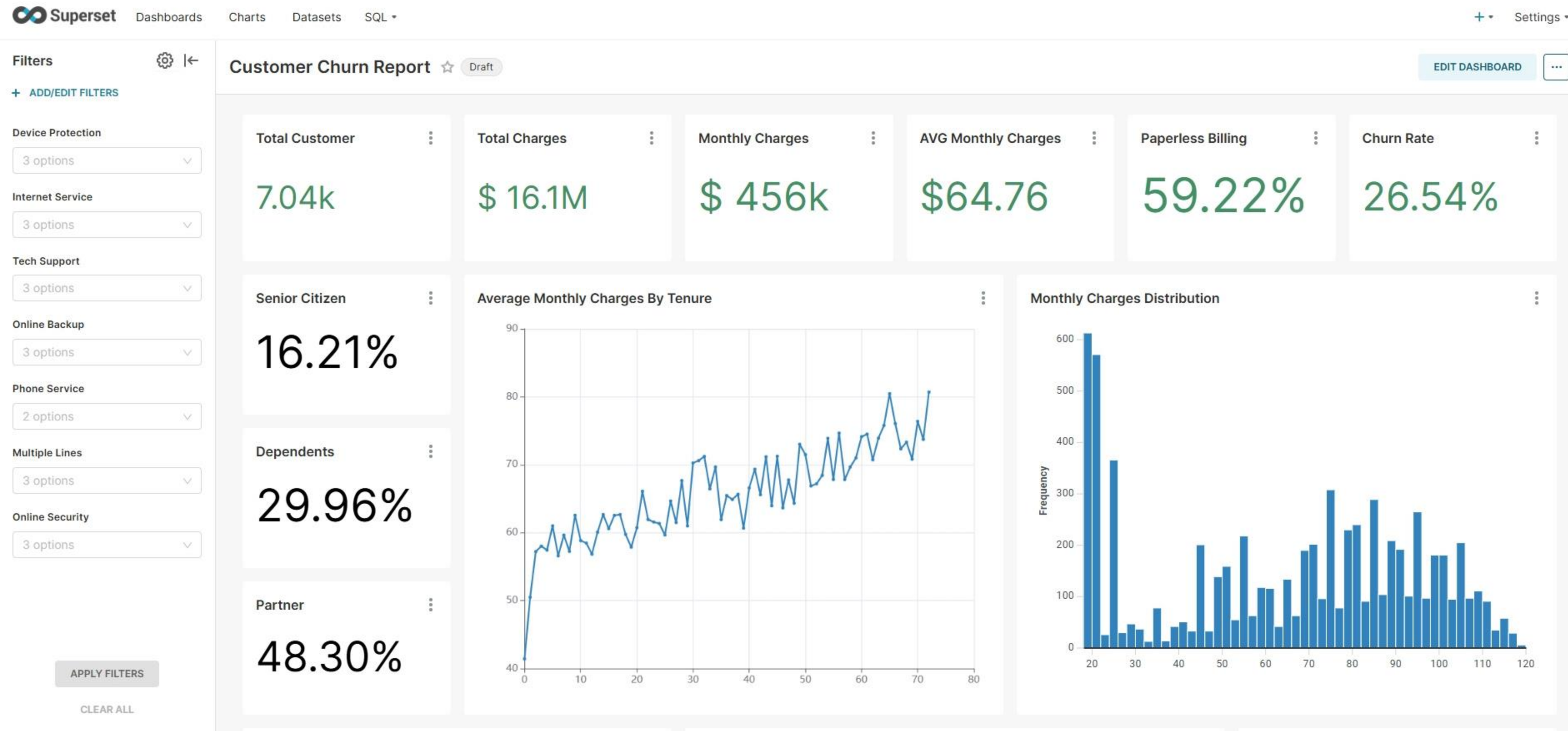
## VI. Demo trực quan hóa dữ liệu với Superset

**Link:** <http://62.84.186.190:8088/>  
**User:** demo\_dashboard  
**Password:** FETEK123@123





# VI. Demo trực quan hóa dữ liệu với Superset



# VI. Demo trực quan hóa dữ liệu với Superset - Charts

## Create a new chart

### 1 Choose a dataset

Choose a dataset ▾

[Add a dataset or view instructions](#)

### 2 Choose chart type

 All charts


Recommended tags ▾

# Popular


# ECharts

# Advanced-Analytics

Category ▾

 Correlation


 Distribution

 Evolution

 Flow


 KPI

 Map

 Part of a Whole

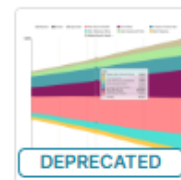
 Ranking

 Table

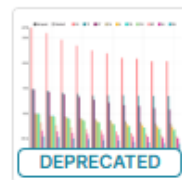
 Other

Tags ▾

Search all charts



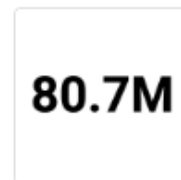
Time-series  
Area Chart  
(legacy)



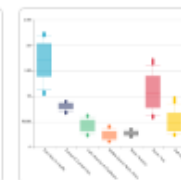
Time-series  
Bar Chart  
(legacy)



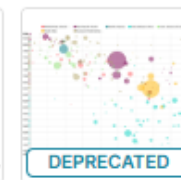
Big Number  
with Trendline



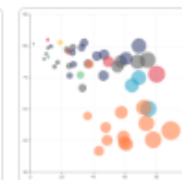
Big Number



Box Plot



Bubble Chart  
(legacy)



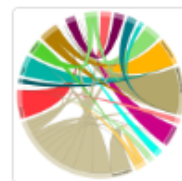
Bubble Chart



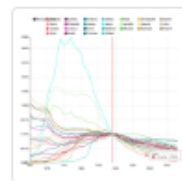
Bullet Chart



Calendar  
Heatmap



Chord  
Diagram



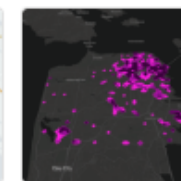
Time-series  
Percent  
Change



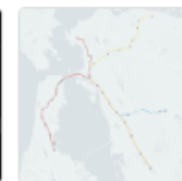
Country Map



deck.gl Arc



deck.gl  
Contour



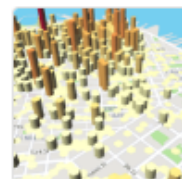
deck.gl  
Geojson



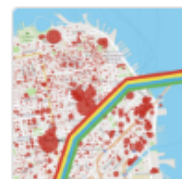
deck.gl Grid



deck.gl  
Heatmap



deck.gl 3D  
Hexagon



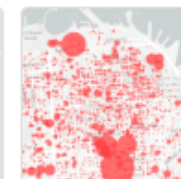
deck.gl  
Multiple  
Layers



deck.gl Path



deck.gl  
Polygon



deck.gl  
Scatterplot



deck.gl  
Screen Grid



Bar Chart  
(legacy)

Please select both a Dataset and a Chart type to proceed

CREATE NEW CHART



# VI. Demo trực quan hóa dữ liệu với Superset - SQL Lab



Superset

Dashboards

Charts

Datasets

SQL

+

Settings

Query superset.Churn\_Retension

Query superset.Churn\_Retension

+

DATABASE

bigquery Google BigQuery

SCHEMA

superset

SEE TABLE SCHEMA

Churn\_Retension

Churn\_Retension

customerID VARCHAR

gender VARCHAR

SeniorCitizen INTEGER

Partner BOOLEAN

Dependents BOOLEAN

tenure INTEGER

PhoneService BOOLEAN

MultipleLines VARCHAR

InternetService VARCHAR

OnlineSecurity VARCHAR

OnlineBackup VARCHAR

DeviceProtection VARCHAR

TechSupport VARCHAR

StreamingTV VARCHAR

StreamingMovies VARCHAR

Contract VARCHAR

PaperlessBilling BOOLEAN

PaymentMethod VARCHAR

MonthlyCharges FLOAT

TotalCharges FLOAT

numAdminTickets INTEGER

numTechTickets INTEGER

Churn BOOLEAN

States VARCHAR

Country VARCHAR

1 SELECT \* FROM superset.Churn\_Retension

RUN

LIMIT: 1 000

SAVE

COPY LINK

RESULTS

QUERY HISTORY

PREVIEW: 'CHURN\_RETENSION'

COPY TO CLIPBOARD

Filter results

100 rows

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
0661-KQHNK	Female	0	true	true	6	true	No	No	No internet service	No internet service
9732-OUYRN	Female	0	true	false	49	true	No	No	No internet service	No internet service
0967-BMLBD	Female	0	true	true	11	true	No	No	No internet service	No internet service
1401-FTHFQ	Male	0	true	true	23	true	No	No	No internet service	No internet service
9824-QCJPK	Male	0	true	false	36	true	No	No	No internet service	No internet service
4709-LKHYG	Female	0	true	true	29	true	No	No	No internet service	No internet service
4323-ELYYP	Male	0	true	true	13	true	No	No	No internet service	No internet service
1269-FOYWN	Male	0	true	true	44	true	No	No	No internet service	No internet service
4955-VCWBI	Female	0	true	true	43	true	No	No	No internet service	No internet service

## VI. Demo trực quan hóa dữ liệu với Superset - Databases



amazon REDSHIFT	Google BigQuery	snowflake	
presto	databricks	druid	<b>FIREBOLT</b>
Timescale	<b>[ROCKSET]</b>	PostgreSQL	MySQL
SQL Server	IBM DB2	SQLite	SYBASE
MariaDB	<b>VERTICA</b>	<b>ORACLE</b>	
Greenplum	ClickHouse	<b>Exasol</b>	monetdb
			pinot
<b>teradata.</b>	yugabyteDB	Databend	StarRocks





THANK YOU

