ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC BÁCH KHOA

THUYẾT TRÌNH NHÓM
MÔN HỌC: DỮ LIỆU LỚN

PHÂN TÍCH DỮ LIỆU LỚN IOT DỰA VÀO LƯỢNG SỬ DỤNG ĐIỆN VÀ NƯỚC

Giảng viên : PGS.TS Thoại Nam

Thành viên. : Huỳnh Thanh Tâm – 2370193
Trần Cao Bảo Ân – 2370297

Thành phố Hồ Chí Minh, tháng 5 năm 2024

Nội dung

1. Giới thiệu
2. Công nghệ sử dụng
3. Hiện thực hệ thống
4. Kết luận

1. GIỚI THIỆU

Giới thiệu

Hiện thực dự án theo dõi lượng sử dụng điện và nước trong thành phố thông minh (smart city):

+ Cung cấp giải pháp theo dõi máy đo lượng nước và máy đo điện theo thời gian thực, để phát hiện lỗi phát sinh nếu có của từng máy đo,
+ Theo dõi lượng tiêu thụ điện và nước của từng hộ dân.

2. CÔNG NGHỆ SỬ DỤNG

Công nghệ sử dụng

1. Apache Spark
2. Apache Kafka
3. MongoDB
4. Airflow
5. Metabase
6. Faker

Công nghệ sử dụng

1. Apache Spark

Spark là một công cụ xử lý dữ liệu lớn hiệu quả, Spark có thể phân tích dữ liệu theo bó (Batch) hoặc theo dòng (Stream) với thư viện Spark Streaming.

Công nghệ sử dụng

2. Apache Kafka

- Message queue trong luồng xử lý dữ liệu,
- Giúp luồng xử lý dữ liệu và luồng thu thập dữ liệu được độc lập với nhau

Công nghệ sử dụng

3. MongoDB

MongoDB là một cơ sở dữ liệu NoSQL được sử dụng để lưu trữ dữ liệu đã được xử lý bởi Spark. MongoDB là một cơ sở dữ liệu linh hoạt và có thể mở rộng, lý tưởng để lưu trữ lượng dữ liệu lớn được tạo ra bởi hệ thống giám sát thông minh.

Công nghệ sử dụng

4. Airflow

- Mã nguồn mở
- Airflow sẽ được sử dụng để lên lịch cho Spark phân tích dữ liệu theo bó.

Công nghệ sử dụng

5. Metabase

- Mã nguồn mở
- Metabase cho phép sử dụng các câu lệnh truy vấn native của MongoDB, giúp tận dụng được sự linh hoạt của các câu truy vấn Mongo.

Công nghệ sử dụng

6. Faker

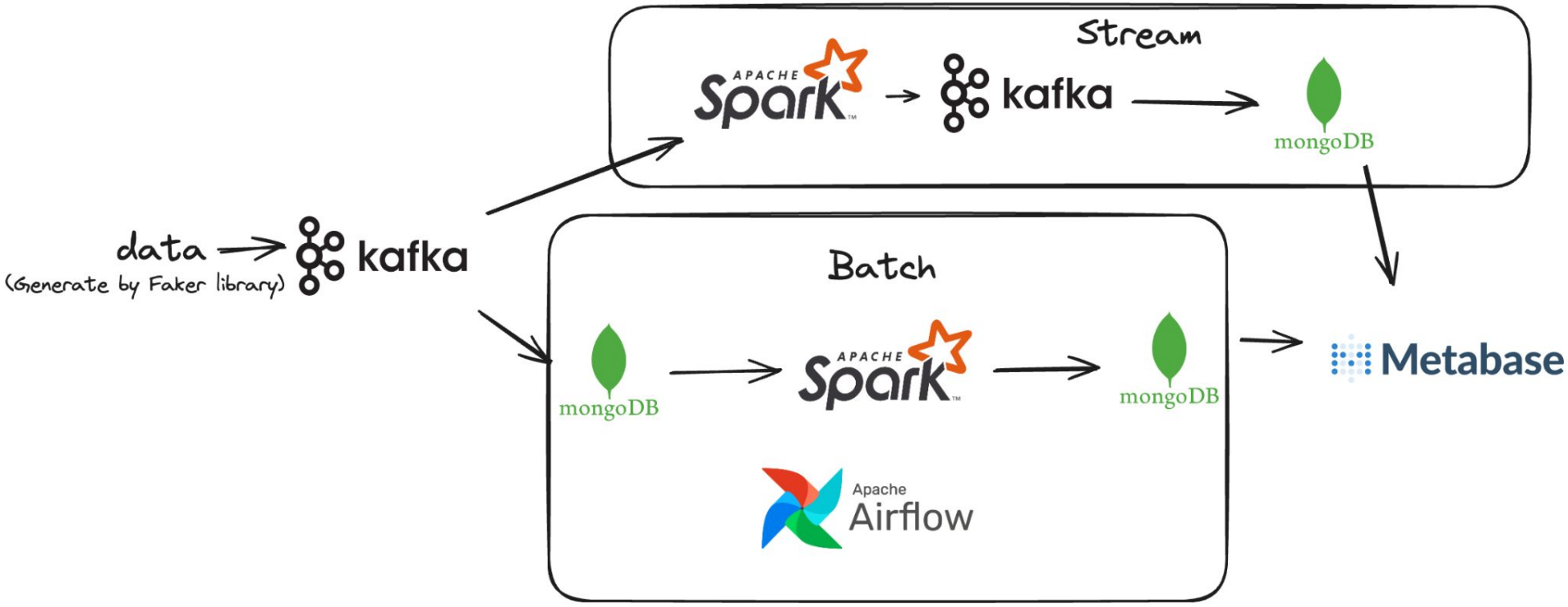
- Thư viện python dùng để tạo dữ liệu giả mô phỏng.

Công nghệ sử dụng

- Ngoài ra còn có những công nghệ khác như Docker để đóng gói hệ thống, ZooKeeper để quản lý Kafka,.

3. Hiện thực hệ thống

Hiện thực hệ thống



4. KẾT LUẬN

Kết luận

- Hiện thực được dự án theo dõi lượng điện và nước sử dụng theo thời gian thực, phân tích dữ liệu và trực quan hóa dữ liệu một cách đầy đủ và liên tục.

Hạn chế

- Các công nghệ khi khởi tạo chưa thống nhất (Airflow và Spark vẫn phải khởi tạo riêng) làm cho việc setup mất nhiều thời gian, nên đóng gói Airflow và Spark vô chung môi trường Docker với các công nghệ khác.
- Hiện chỉ sử dụng một cluster Kafka, Mongo và Spark, thực tế các công nghệ này có thể khởi tạo nhiều cluster theo hướng phân tán, giúp cho việc xử lý dữ liệu lớn được hiệu quả hơn.