Nhóm 16

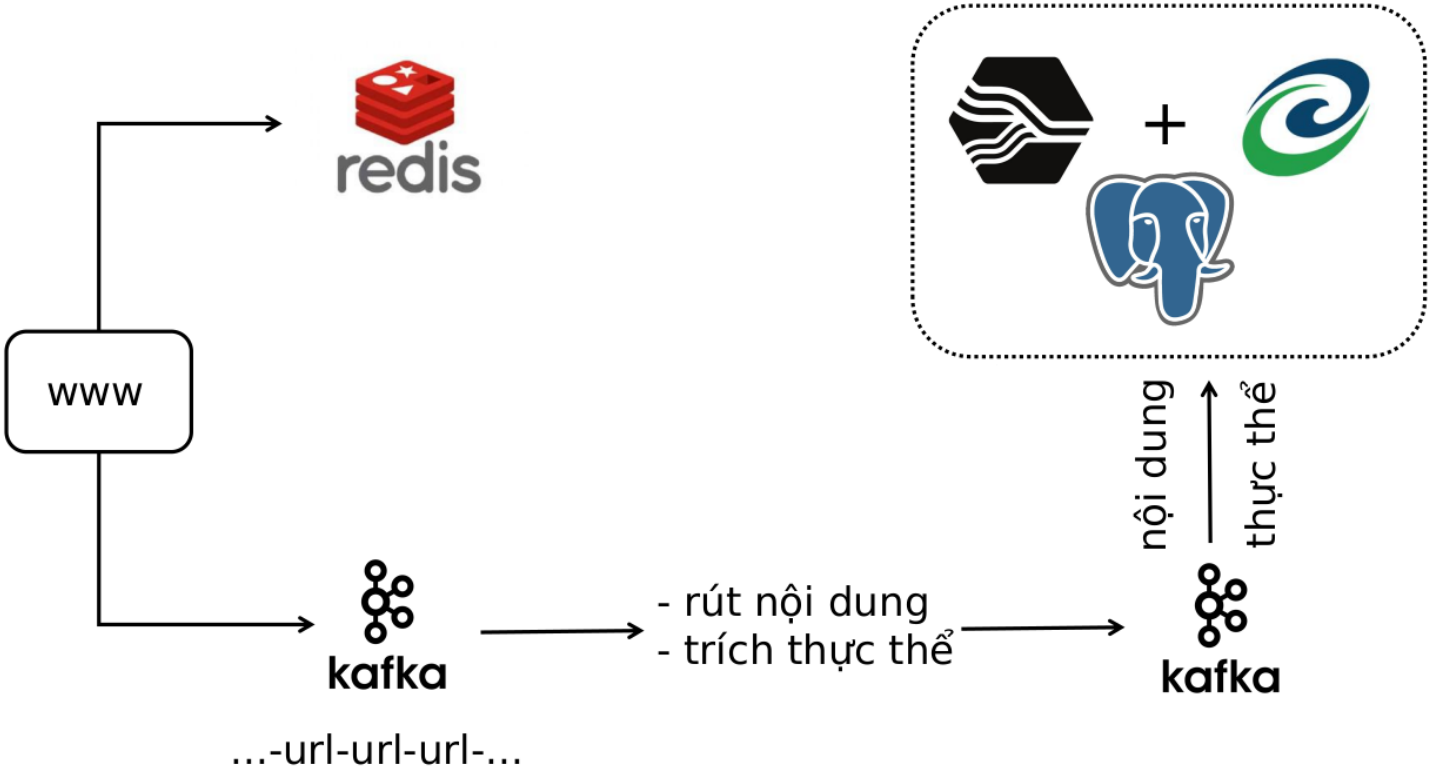
Xây dựng hệ thống Rút trích tin tức từ Internet và thống kê các từ khóa chính xuất hiện

Nguyễn Châu Long – CH1802052
Nguyễn Tân Kim – CH1801030
Nguyễn Chí Thương – CH1801015
Phạm Đức Duy – CH1801024

Nội dung

- Kiến trúc hệ thống
- Xử lý luồng dữ liệu
- Lưu trữ & Thống kê
- Hiện thực đề tài
- Hướng phát triển đề tài

Kiến trúc hệ thống



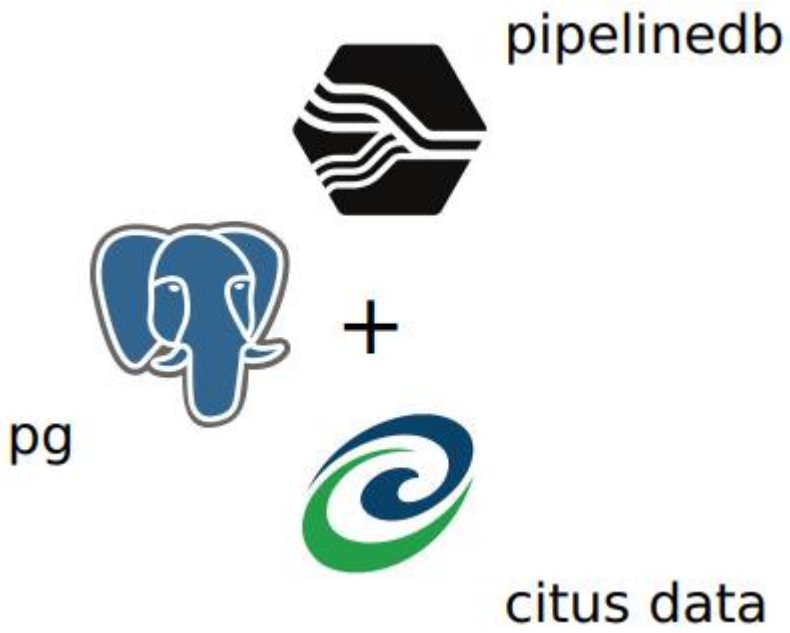
Xử lý luồng dữ liệu (Kafka + Redis)

- ✓ **Tiếp nhận luồng dữ liệu đến**
- ✓ **Làm trung gian xử lý luồng dữ liệu**
- ✓ **Lưu các trang đã duyệt**
- ✓ **Kiểm tra trang đã duyệt hay chưa**

Kafka: https://kafka.apache.org/
Redis: https://redis.io/

Lưu trữ & Thống kê

Ý tưởng: postgresql + tiện ích mở rộng



- ✓ pg: postgresql
- ✓ citus data: phân tán dữ liệu
- ✓ pipelinedb: đếm các thực thể theo thời gian thực

* *pipeline_kafka + pipelinedb*

Pipelinedb: https://www.pipelinedb.com/
Citus: https://www.citusdata.com/
Pipeline kafka: https://github.com/pipelinedb/pipeline_kafka

Citusdata: Bảng phân tán

Node A

Accounts table (shard 1)

account_id	name	created_at
1	CNN	2016-07-12
5	Comcast	2016-07-19
...	...	...
1252	Walmart	2016-08-02

Campaigns table (shard 3)

campaign_id	name	account_id
1202	tv series	1
1204	superbowl	1
...	...	...
352042	chocolate	1252

Node B

Accounts table (shard 2)

account_id	name	created_at
2	AT&T	2016-07-13
3	Exxon	2016-07-14
...	...	...
1253	UPS	2016-08-03

Campaigns table (shard 4)

campaign_id	name	account_id
2742	gas state	3
2743	my phone	2
...	...	...
352423	new phone	2

PipelineDB: Khung nhìn liên tục

```
INSERT INTO stream (x) VALUES (4)
INSERT INTO stream (x) VALUES (8)
INSERT INTO stream (x) VALUES (7)
INSERT INTO stream (x) VALUES (0)
INSERT INTO stream (x) VALUES (7)
```

```
SELECT count(*) FROM stream
```

count
0

```
SELECT sum(count) FROM view, worker
```

count
76

- Đếm thực thể theo thời gian thực
- Bảng stream
- Khung nhìn cập nhật liên tục
- Tích hợp kafka

Postgresql + tiện ích

- ✓ Dễ sử dụng
- ✓ Postgresql = SQL + NoSQL
- ✓ Nhiều tiện ích mở rộng
- ✓ PipelineDB: thực hiện được 25 tỷ (*) giao dịch mỗi ngày (trên 1 node 32 vCPUs cloud)
- ✓ Citusdata: dữ liệu phân tán + truy vấn phân tán
- ✓ Tích hợp trên cloud (AWS, Azure, Alibaba cloud,...)
- ✓ Mã nguồn mở + giấy phép GNU và AGPL v3.0

Hiện thực đề tài

- Docker compose: slave 4
- Master + slave 1 - 3: Postgresql 11 + Citusdata
- Master: Postgres master (coordinator node)
- Slave 1 - 3: Postgres slave (worker node)
- Master: PipelineDB
- Slave 4: docker zookeeper, kafka, redis, python,...

Hướng phát triển đề tài

- Mở Redis, Kafka theo chiều rộng + phân tán
- Cân bằng tải (HA Proxy, Istio, Doctorkafka,...)
- Mở rộng dịch vụ rút tin (replicate hoặc kubernetes)

  .....

NHÓM 16
Xin cảm ơn
Thầy và các anh chị