LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN

NHÓM 25

Đề tài: Phân tích dữ liệu từ Twitter
để dự đoán kết quả bầu cử tổng thống Mỹ 2020

CÁC THÀNH VIÊN

1.	PHẠM VĂN HÙNG	20173151
2.	TRẦN VĂN HOÀNG	20173129
3.	TRẦN VĂN ĐIỆP	20173014
4.	TRẦN PHƯƠNG THẢO	20173083
5.	LÊ TRỌNG NHÂN	20173292

ĐẶT VẤN ĐỀ

Theo TS Lê Huy Khôi- Viện Nghiên cứu chiến lược, chính sách công thương (Bộ Công Thương) cho rằng, chính sách kinh tế của Việt Nam với Mỹ sẽ đi theo chiến lược quốc gia trong quan hệ giữa hai nước, dù ai làm Tổng thống Mỹ. Tuy vậy, nếu Biden làm Tổng thống, chính sách kinh tế của Mỹ có thể linh hoạt, nhẹ nhàng hơn so với Tổng thống Donald Trump. Xu hướng chuyển dịch đầu tư sang khu vực Đông Nam Á trong đó có Việt Nam sẽ chậm hơn.

Nguồn: Báo An ninh thủ đô

ĐẶT VẤN ĐỀ

17:00 05/11/2020 |

Trang info.cz mới đây đăng bài viết của tác giả Vojtech Kristen trích dẫn nhận định của 4 nhà kinh tế học Štěpán Hájek, Vít Hradil, Lukáš Kovanda và Anna Píchová về sự tác động của cuộc bầu cử Mỹ đến thị trường tài chính, trong đó nhấn mạnh quan điểm rằng, thị trường chứng khoán sẽ chịu ảnh hưởng nhiều nhất từ kết quả của cuộc bầu cử này.

Nguồn: Tạp chí tài chính

ĐẶT VẤN ĐỀ

- Do đó, dự đoán được kết quả bầu cử tổng thống sẽ giúp:
  - Tìm ra được đường lối đối ngoại phù hợp trong điều kiện mới.
  - Chuẩn bị trước với những thách thức trong tương lai (kinh tế, năng lượng, an ninh quốc phòng,…)
  - Tìm ra được các cơ hội mới để phát triển (đầu tư vào các xu hướng mới)

CÁC BƯỚC TIẾN HÀNH

- Phân tích bài toán.
- Thiết kế kiến trúc chương trình.
- Triển khai

PHÂN TÍCH BÀI TOÁN

1. Dữ liệu:

- Lựa chọn nguồn thông điệp để có thể cập nhật các thông tin nhanh chóng, hiệu quả.
- Lựa chọn các từ khoá từ nguồn thông điệp (từ các bài báo, từ các top hot search).
- Lấy thông tin từ nguồn thông điệp để xử lý (thông qua các API, các

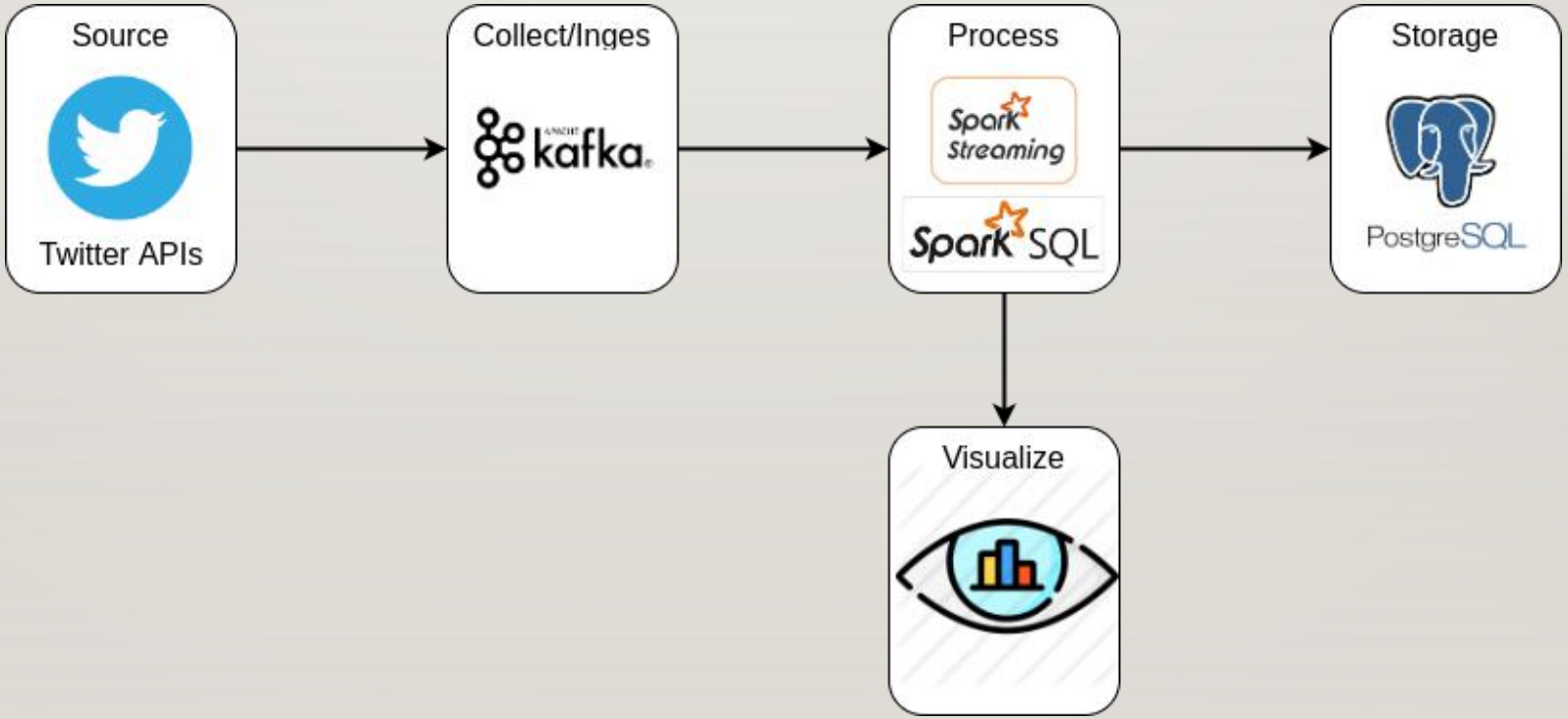
2. Xử lý dữ liệu:

- Định dạng dữ liệu thu được (là các luồng có cấu trúc)
- Xử lý dữ liệu (đánh giá các thông tin thu được từ dữ liệu bằng mô hình Stanford NLP Core).

3. Trực quan hoá dữ liệu đã xử lý

THIẾT KẾ KIẾN TRÚC CHƯƠNG TRÌNH.

Mô hình luồng dữ liệu:



THIẾT KẾ KIẾN TRÚC CHƯƠNG TRÌNH.

Twitter APIs:

- Twitter là một mạng xã hội thông dụng với số lượng người dùng lớn, với lượt tương tác nhiều.
- Được cung cấp miễn phí cho phép các nhà phát triển có thể sử dụng nền tảng Twitter cho nhiều mục đích khác nhau.

=> Phù hợp cho bài toán đã đề ra với một khối lượng thông tin ở mức thử nghiệm chương trình.

THIẾT KẾ KIẾN TRÚC CHƯƠNG TRÌNH.

Kafka

- Là một hệ thống message pub/sub phân tán.
- Có khả năng truyền một lượng lớn dữ liệu thieo thời gian thực, trong trường hợp bên nhận chưa nhận thì message vẫn được lưu trong hàng đợi và cả trên ổ đĩa để đảm bảo an toàn.
- Dữ liệu cũng được replicate trong cluster giúp phòng tránh mất dữ liệu.

THIẾT KẾ KIẾN TRÚC CHƯƠNG TRÌNH.

Spark Streaming/ Spark SQL

- Apache Spark là một open source cluster computing framework.
- Spark Streaming là một phần bổ sung cho Spark để xử lý lượng dữ liệu lớn theo thời gian thực và đảm bảo chống chịu lỗi.
- Spark SQL là một mô-đun để xử lý dữ liệu có cấu trúc trong Spark.

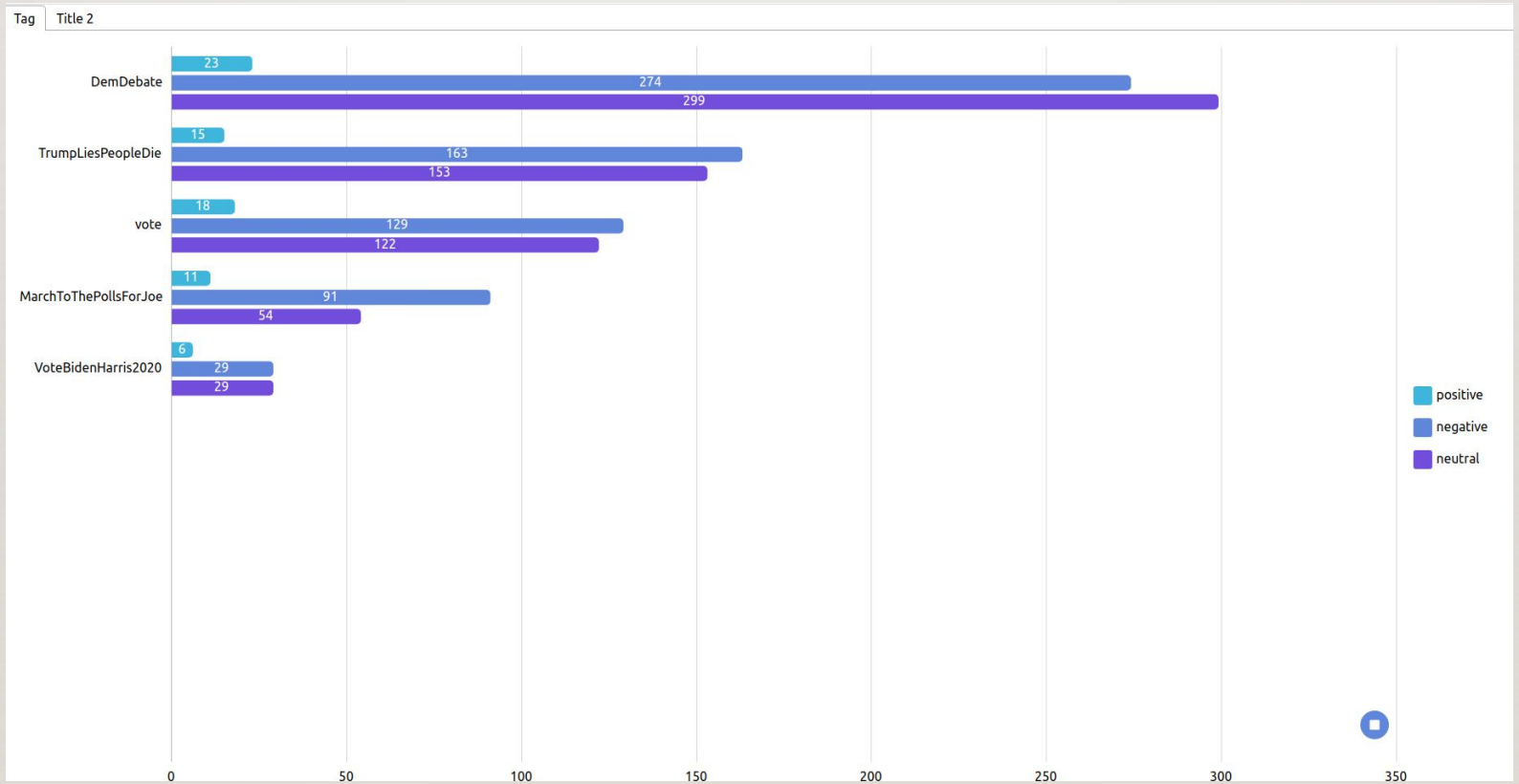
THIẾT KẾ KIẾN TRÚC CHƯƠNG TRÌNH.

Postgres SQL

- Là một cơ sở dữ liệu có quan hệ.
- Là một phần mềm mã nguồn mở miễn phí.

DEMO

KẾT QUẢ ĐẠT ĐƯỢC



KẾT QUẢ ĐẠT ĐƯỢC