XÂY DỰNG DATAWAREHOUSE CHO BỘ DỮ LIỆU PHÂN TÍCH CẢM XÚC TRONG TIẾNG VIỆT

VÕ VĂN NAM - TRẦN CÔNG HÙNG

GVHD: THS. LAM MAI

1. GIỚI THIỆU

Giới thiệu về Data Sentiment Analysis

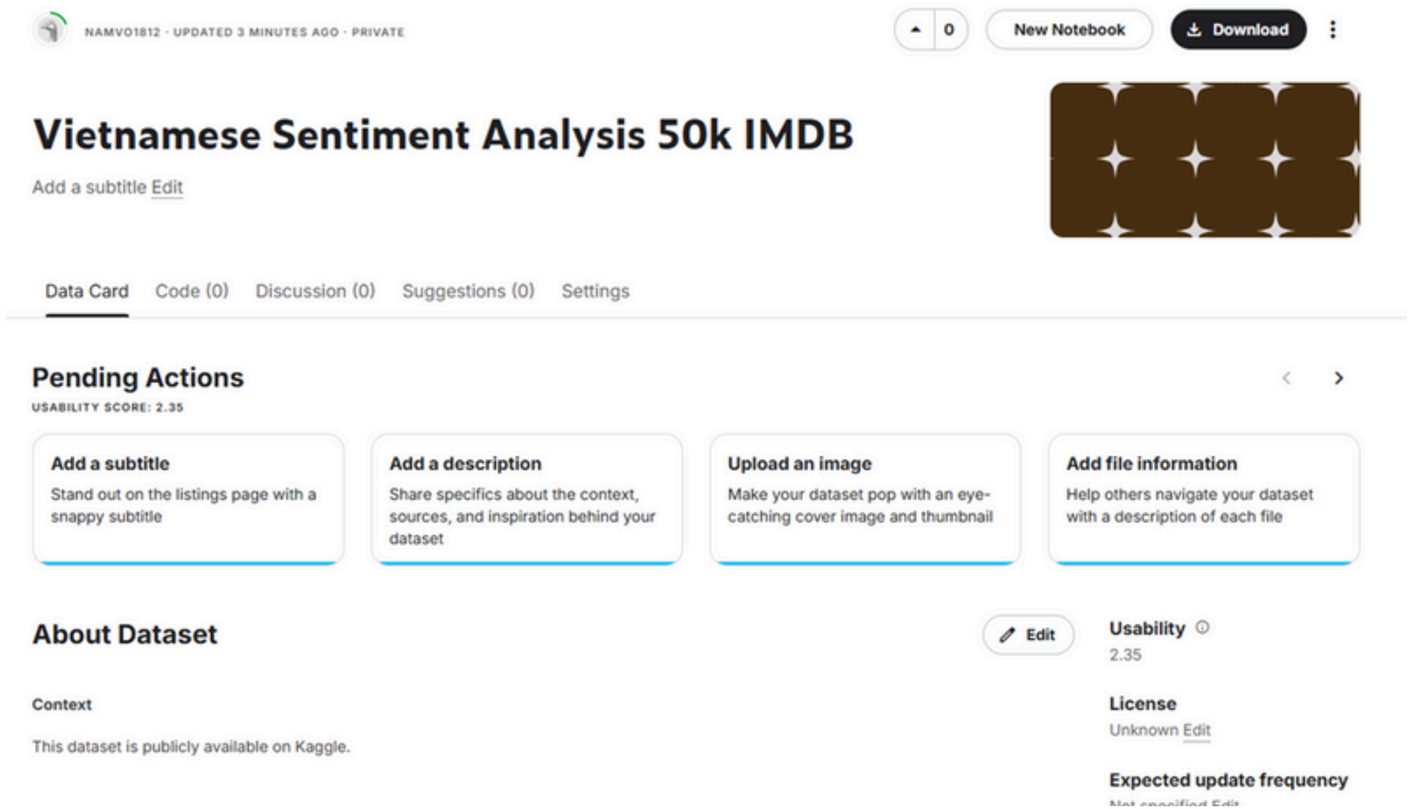
- **Sentiment Analysis**, hay còn gọi là **Opinion Mining**, là **quá trình phân tích** và **nhận diện cảm xúc**(positve, negative, neutral) từ văn bản.
- **Cung cấp thông tin quan trọng** về **ý kiến** và tình hình **cảm xúc** của người dùng đối với một **sản phẩm, dịch vụ** hoặc **sự kiện**.

Data Sentiment Analysis trong tiếng Việt

- **Sự thiếu sót** về tài nguyên **dữ liệu** và **công cụ phân tích**, đặc biệt là so với các ngôn ngữ phổ biến khác như **tiếng Anh**.
- **Sự thiếu sót** các **kho dữ liệu lưu trữ** những phản hồi của người dùng một cách **tối ưu** và **tự động.**

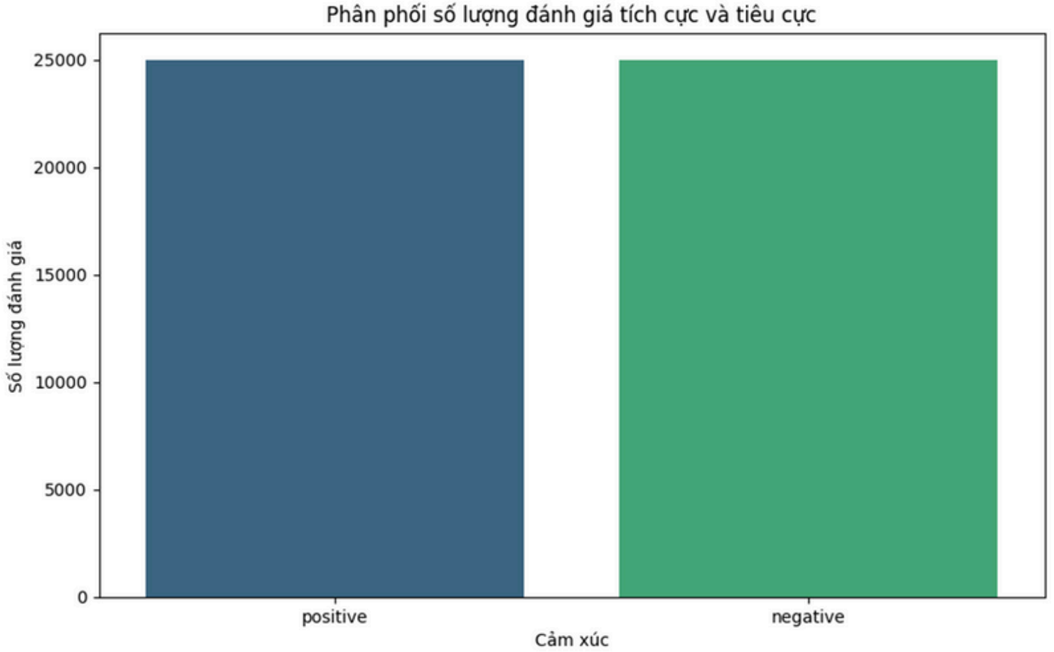
2. NỘI DUNG

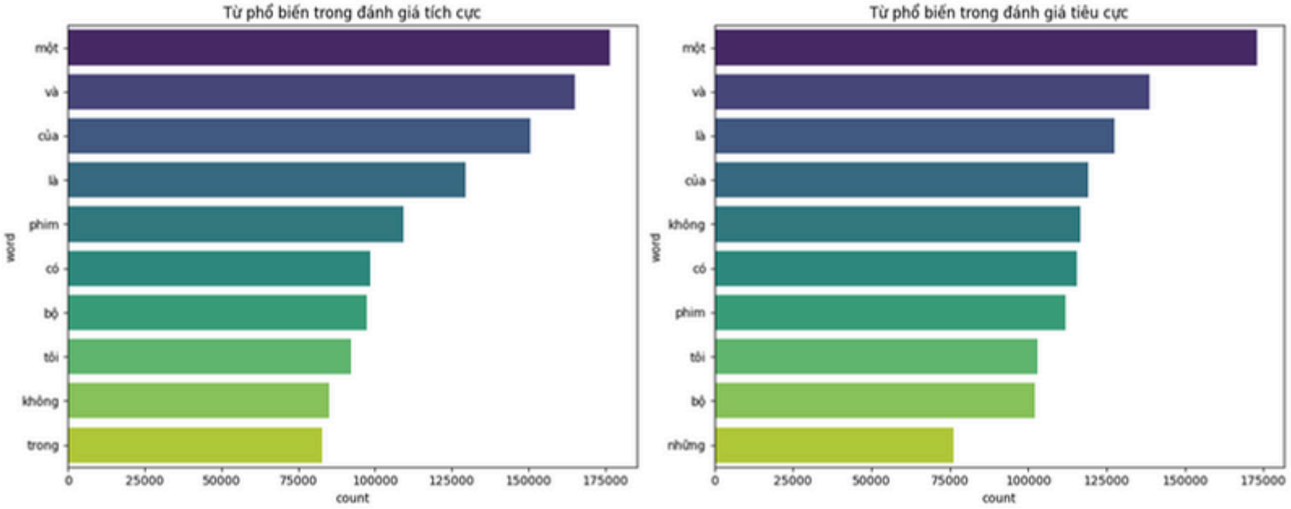
Dữ liệu thô với **50.000 câu** đánh giá phim từ **nền tảng IMDB** được **translate** và **tiền xử lý** để import vào DB



2. NỘI DUNG

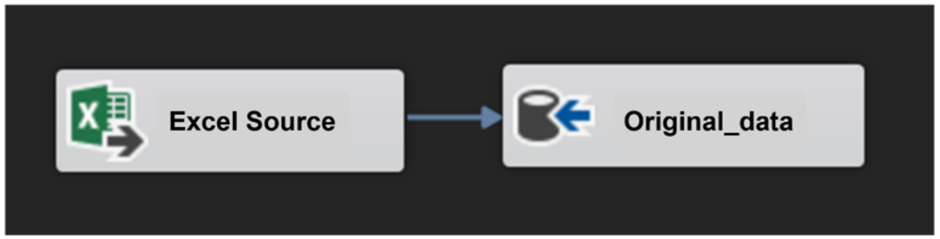
Dữ liệu được **trực quan hóa** bằng các công cụ như **Matplotlib**, **Seaborn**,... giúp dễ dàng đánh giá bộ dữ liệu

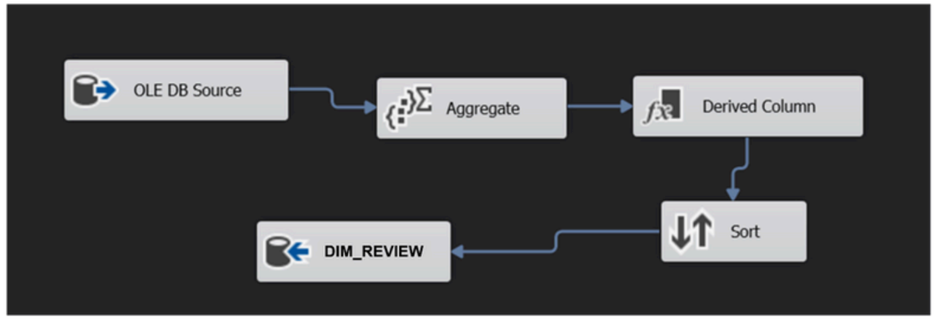


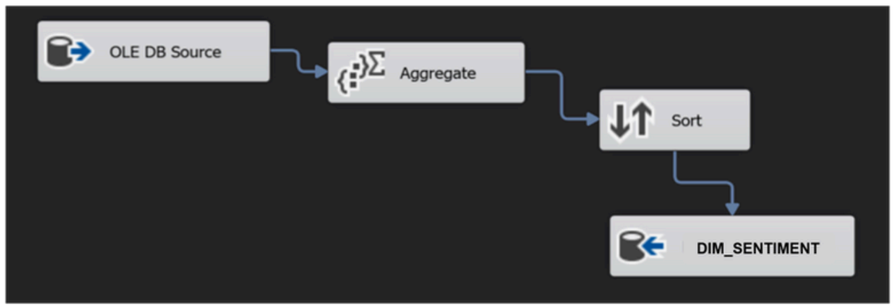


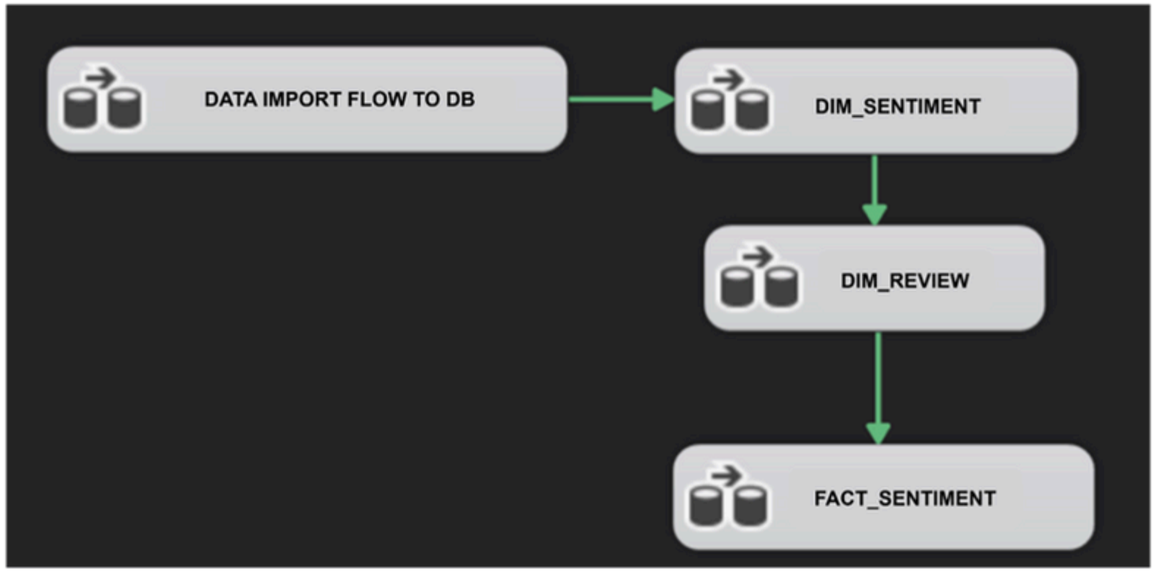
2. NỘI DUNG

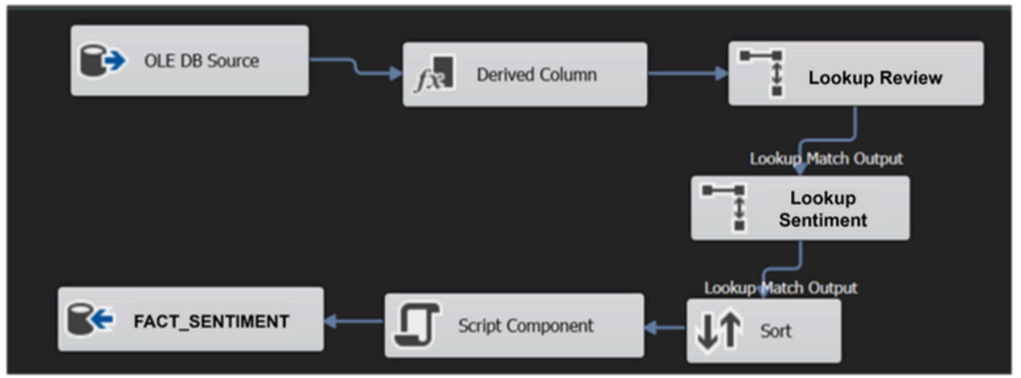
Triển khai **ELT** bằng **SSIS**, giúp **đơn giản hóa**, **tự động hóa** và **tối ưu hóa** quy trình tích hợp dữ liệu từ bộ dữ liệu thô





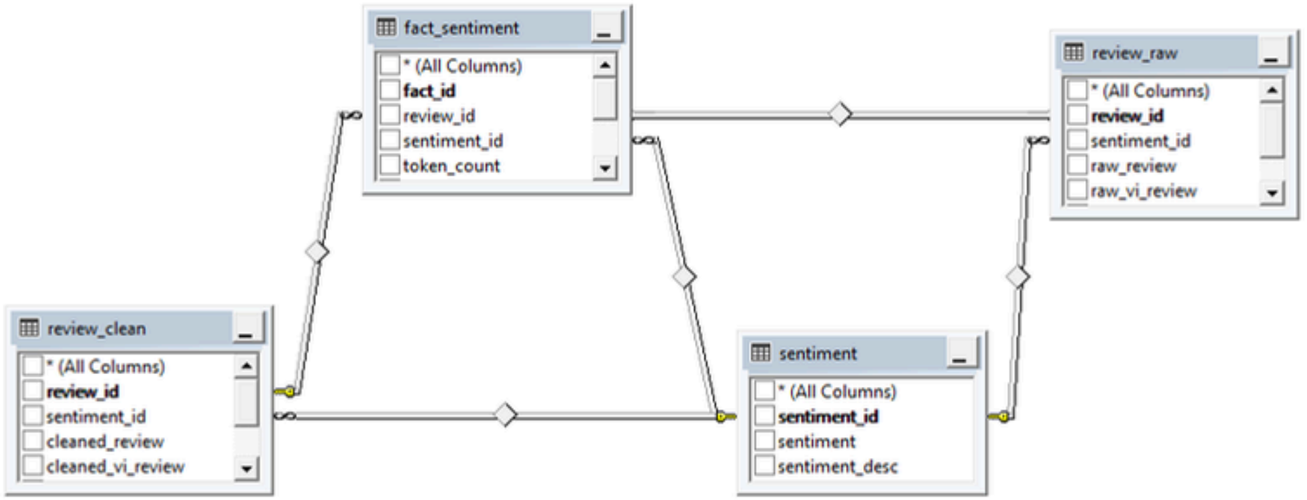






3. KẾT QUẢ

- Nắm rõ các **khái niệm cơ bản** về kho dữ liệu, các tính chất của một kho dữ liệu cần có.
- Trang bị kiến thức về các công cụ **SSIS**.
- Xây dựng được một **kho dữ liệu** hoàn chỉnh với chủ đề **50.000 câu** đánh giá phim từ người dùng của các nhà phát triển trên **IMBD**.
- Trình bày **tối ưu hóa câu truy vấn**.




```
SELECT sentiment, COUNT(*) AS total_reviews
FROM fact_sentiment_analysis
GROUP BY sentiment;
```

	sentiment	total_review
1	negative	25000
2	positive	25000

```
Select vi_review, COUNT(*) AS total_review
from dbo.VI_IMDB
where vi_review like '%hay%' and sentiment = 'positive'
group by vi_review
```

	vi_review	total_review
1	Ai cũng biết và đã từng nói: Không ai có thể đóng S...	1
2	- Một nhóm cướp cướp chuyến hàng vàng mà đoàn t...	1
3	- SPOILER NHỎ TẠI ĐÂY! -Khi tôi xem các phiếu bầu...	1
4	Tôi cho rằng đây không phải là bộ phim hay nhất từn...	1
5	Tôi có thể nói gì đây? Đây là một trong những bộ phi...	1
6	" Before Sunrise " là một câu chuyện tình yêu tuyệt v...	1
7	" Tôi đã vật lộn với cái chết. Đó là cuộc thi tẻ nhạt nh...	1
8	"... nhịp điệu quá mạnh... giờ chúng tôi là những dị nh...	1
9	"2001: A Space Odyssey" là một chuyến du hành vũ tr...	1
10	"A Bug's Life" giống như một thanh kẹo được yêu thí...	1
11	"A Family Affair" đưa chúng ta trở lại thời kỳ ít phức tạ...	1
12	"A Fare to Remember" là một bộ phim hoàn toàn phải ...	1
13	"A Guy Thing" có thể không phải là một tác phẩm kin...	1
14	"Ác quỷ trong bóng tối" là tập phim yêu thích của Will...	1
15	"Admissions" là một bộ phim truyền hình hay mặc dù c...	1
16	"Ah Ritchie đã làm một bộ phim xã hội đen khác với ...	1
17	"Ahh...I don't Order No Amazing Hit Show"....."We'll y...	1
18	"Ai Sẽ Yêu Con Tôi" Bộ phim buồn nhất mà tôi từng x...	1
19	"Americans Next Top Model" là chương trình thực tế ...	1
20	"Anchors Aweigh" là sản phẩm của đơn vị sản xuất n...	1

4. HẠN CHẾ/GIẢI PHÁP

Những hạn chế hiện tại

- **Chưa áp dụng** được **so sánh kết quả** của nhiều phương pháp **Data Mining**.
- Quá trình **SSIS** còn **rườm rà** chưa được **tối ưu**.
- Các câu **truy vấn MDX** chưa nâng cao

Giải pháp trong tương lai

- **Tiếp tục nghiên cứu** và **tích hợp** các **công cụ phân tích dữ liệu nâng cao** để tạo **báo cáo**, **thống kê**, và **bảng điều khiển** nhằm tối ưu hóa.
- Nghiên cứu việc **tích hợp** các công nghệ mới như **Blockchain** để tăng cường tính bảo mật và minh bạch của hệ thống kho dữ liệu.
- **Khám phá** các phương pháp mới trong **Phân tích Dữ liệu Lớn (Big Data Analytics)** và **Trực quan hóa Dữ liệu Tương tác (Interactive Data Visualization),...**

Thanks For Listening
Q&A