**Đại học Bách Khoa Hà Nội**

**Đề tài**

**Thu thập và phân tích hành vi và xu hướng nghe nhạc của người dung**

**GVHD:**

Trần Việt Trung

Thân Quang Khoát

Nguyễn Thị Oanh

Bùi Thị Mai Anh

**Khoa:** Khoa học máy tính

**Trường:** Công nghệ thông tin và truyền thông

**Thành viên nhóm**

Trần Nhật Huy 20204988

Bùi Đức Dũng 20204954

Hoàng Hà Trung Đức 20205066

Vũ Đức Lương 20205099

Nguyễn Doãn Tuấn 20205040

# Phần 1: Giới thiệu vấn đề - Giới thiệu bài toán

Bài toán thu thập hành vi người dùng từ trang web Nghe nhạc và phân tích dữ liệu là một yếu tố quan trọng trong việc hiểu và tối ưu hóa trải nghiệm người dùng. Trong thế giới trực tuyến ngày nay, thông tin về hành vi và sở thích của người dùng đóng vai trò quyết định đối với thành công của một doanh nghiệp.

Bằng cách thu thập và phân tích dữ liệu hành vi người dùng, ta có thể khám phá những thông tin quan trọng như số lượng người nghe, bài hát được nghe nhiều nhất, bài hát được tìm kiếm hay được yêu thích nhiều nhất. Những dữ liệu này cung cấp cái nhìn sâu sắc về xu hướng và ưu thích của người dùng, giúp tìm ra cách tối ưu hóa các chiến lược phát triển, từ việc quảng cáo đến phát hành sản phẩm âm nhạc.

Với thông tin này, có thể hiểu rõ hơn về nhu cầu của người dùng, tăng cường khả năng dự đoán và tương tác với họ. Ngoài ra, việc phân tích dữ liệu hành vi người dùng còn giúp xác định hiệu quả của các chiến dịch tiếp thị và quảng cáo, định hình lại các chiến lược marketing và cải thiện trải nghiệm người dùng trên trang web.

Với những lợi ích mà bài toán này mang lại, việc thu thập và phân tích dữ liệu hành vi người dùng từ trang web Nghe nhạc trở thành một công cụ quan trọng giúp nắm bắt thị trường, tăng cường độ cạnh tranh và mang lại lợi nhuận trong môi trường trực tuyến ngày nay.

# Phần 2: Công nghệ nền tảng

**\* Snowplow**

- Snowplow là một nền tảng mạnh mẽ cho việc thu thập, xử lý và lưu trữ dữ liệu hành vi người dùng từ các ứng dụng web, di động và các nguồn dữ liệu khác. Snowplow cung cấp cho người dùng khả năng tự tùy chỉnh việc thu thập dữ liệu.

## Các thành phần được ứng dụng trong project

* + Tracker tạo dữ liệu sự kiện và gửi dữ liệu này đến Collector của người dùng. Snowplow có các tracker bao gồm web, mobile, server, desktop và IoT. Ngoài ra webhook cho phép phần mềm của bên thứ 3 gửi các luồng nội bộ của riêng họ đến Collector của người dùng để xử lý thêm.
  + Collector nhận dữ liệu sự kiện từ tracker hoặc webhook, lưu dữ liệu thô vào bộ lưu trữ, sau đó gửi chúng theo pipeline đến Enrich.
  + Iglu server: Dữ liệu muốn tracking được định nghĩa trong iglu server, có thể tùy biến dữ liệu thu thập theo mong muốn của tổ chức
  + Sau khi được xác thực, mỗi sự kiện sẽ được làm giàu bằng Enrichments mà người dùng đã cấu hình cho pipeline.
  + Sự kiện đã được làm giàu được gửi tới luồng thời gian thực. Tại thời điểm này, dữ liệu có thể được chuyển tiếp đễn các ứng dụng khác để cung cấp cho các use cases thời gian thực.

## Lợi ích của Snowplow

* Kiến trúc vững chắc có thể xử lý hàng tỷ sự kiện mỗi ngày.
* Có hơn 20 SDKs để thu thập dữ liệu từ web, mobile, server-side và các nguồn khác.
* Cách tiếp cận độc đáo dựa trên schemas và xác thực giúp đảm bảo dữ liệu người dùng được làm sạch nhất có thể.
* Hơn 15 Enrichments để tận dụng tối đa dữ liệu của người dùng.
* Gửi dữ liệu đến các kho dữ liệu và luồng phổ biến. Snowplow rất phù hợp với ngăn xếp dữ liệu hiện đại.

**\* Kafka :**

- Snowplow hỗ trợ đẩy dữ liệu đến nhiều nguồn khác nhau như: Kinesis (AWS), Pubsub (GCP), Kafka, NSQ, RabbitMQ. Trong đó, Kafka là một mã nguồn mở và có hiệu năng cao, được rất nhiều công ty trên thế giới sử dụng và tin dùng. Kafka được chọn làm nơi trung gian cho việc truyền thông tin giữa các thành phần trong hệ thống. Việc sử dụng Kafka trong hệ thống cũng mang lại nhiều lợi ích. Nó tạo ra một cầu nối tin cậy và linh hoạt giữa các phần của hệ thống, cho phép dữ liệu được truyền tải một cách liên tục và đồng bộ. Ngoài ra, Kafka còn hỗ trợ tính năng chia sẻ dữ liệu và công cụ xử lý luồng dữ liệu (stream processing), giúp chúng ta thực hiện các công việc phân tích và xử lý dữ liệu trực tiếp từ Kafka.

**\* Metabase**

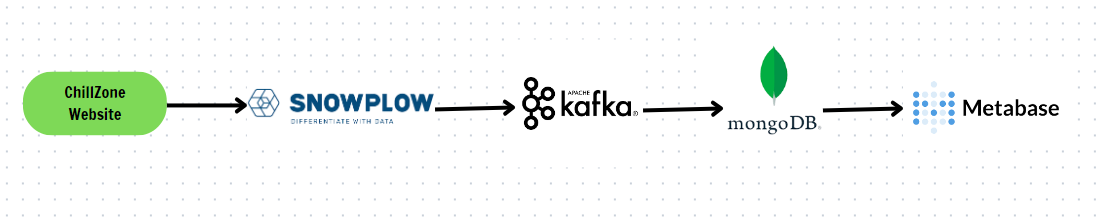
- Metabase là một công cụ mã nguồn mở và miễn phí được sử dụng để trực quan hóa và khám phá dữ liệu. Nó cung cấp một giao diện đơn giản và trực quan cho người dùng không phải là nhà phân tích dữ liệu để tương tác với các nguồn dữ liệu và tạo ra các báo cáo, biểu đồ và truy vấn dữ liệu một cách dễ dàng

# Phần 3: Triền khai hệ thống

## Mục tiêu

## Thu thập hành vi người dùng từ trang Web Nghe nhạc theo thời gian thực. Từ đó phân tích dữ liệu và đưa ra những dữ liệu hữu ích ( có bao nhiêu người nghe, tìm kiếm bài hát, bài hát nào được nghe, tìm kiếm, yêu thích nhiều nhất, 1 người đã nghe 1 bài hát bao nhiêu lần )

## **Kiến trúc hệ thống**

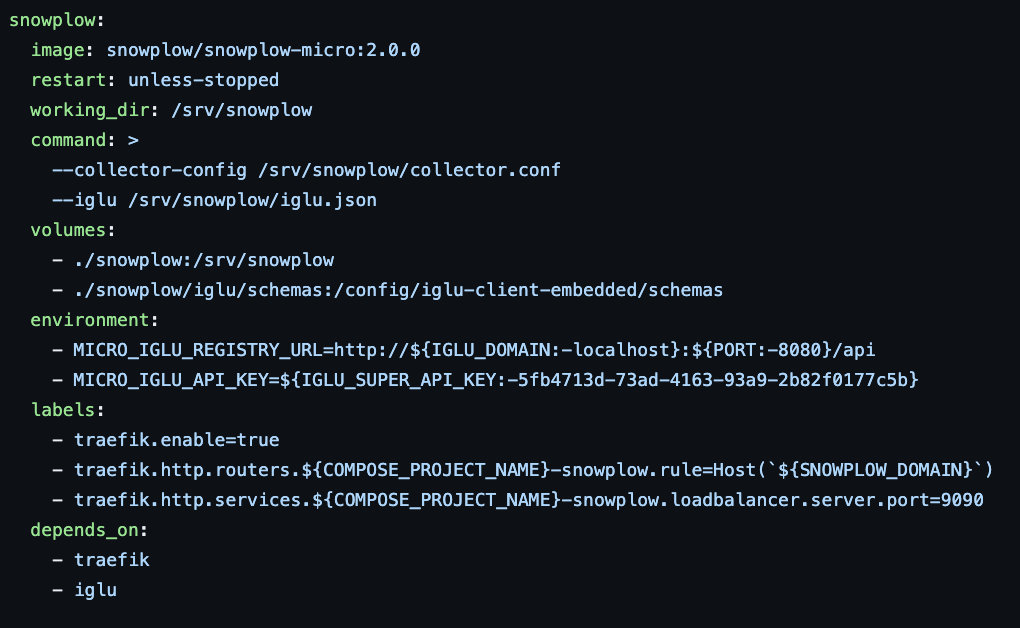


Mô tả:

* Khi người dùng thao tác với trang web, hành động của người dùng được tracker thu thập và gửi đến snowplow server theo thời gian thực với đầy đủ thông tin ( địa chỉ IP, thời gian, thông tin người dùng, thông tin bài hát, loại hành động,....) dưới dạng json object
* Dữ liệu sau khi thu thập được được đưa vào kafka để xử lý streaming. Tại đây, tiến hành làm sạch dữ liệu, lấy ra những dữ liệu hữu ích để phân tích ( thông tin người dùng, thông tin bài hát, thể loại bài hát,... ) . Sau đó, gửi lên broker server qua thành phần producer để lưu trữ và chia sẻ dữ liệu giữa các ứng dụng trong hệ thống .Do có khả năng chịu tải cao và mở rộng lớn bằng cách thêm một hoặc nhiều broker mới vào cluster, dữ liệu tại đây được đảm bảo sự mất mát đến mức tối đa và luôn tuân theo thứ tự gửi sự kiện, đảm bảo luôn nhận được dữ kiện mới nhất khi được gửi lên.
* Sau đó thành phần consumer sẽ thực hiện lấy dữ liệu trong kafka server và lập tức đẩy vào nơi lưu trữ trên mongodb để tiến hành trực quan hóa dữ liệu.
* Để trực quan hóa dữ liệu một cách nhanh chóng và hiệu quả, sử dụng metabase.

## **Các bước thực hiện**

1. Xây dựng website nghe nhạc có các chức năng:
   * Đăng nhập, đăng kí
   * Nghe nhạc
   * Filter theo nghệ sĩ
   * Filter theo thể loại bài hát
2. Xác định context: song ,user
   * Song: id, name, type, artist
   * User: id , name
3. Xác định các events:
   * Tìm kiếm nhạc
   * Nghe nhạc
   * Thêm vào danh sách yêu thích
   * Filter theo thể loại của bài hát
   * Filter theo nghệ sĩ
4. Triển khai Snowplow micro bằng Docker



1. Cài đặt Iglu server để định nghĩa và validate dữ liệu, khởi tạo schema và config làm giàu sự kiện được gửi bởi tracking code

5.1 Cài đặt Iglu server



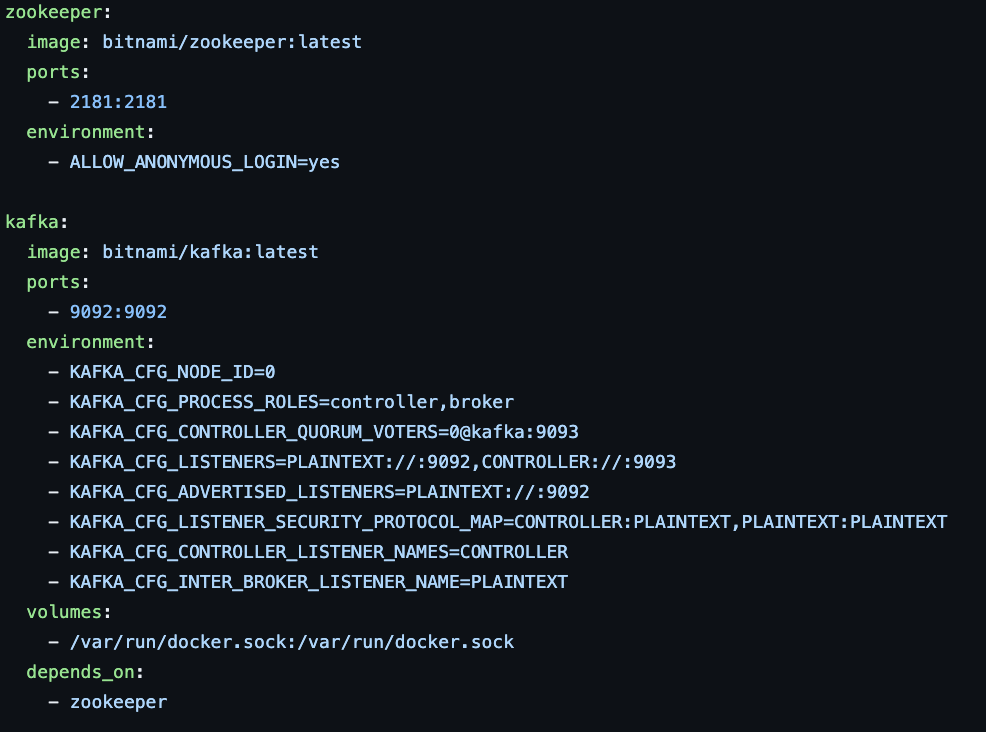
5.2 Khởi tạo schema định nghĩa các dữ kiện







1. Triển khai kafka trên môi trường Nodejs bằng Docker và xử lý dữ liệu



Khởi tạo producer để gửi dữ liệu tới các topic đã định nghĩa trong broker server



1. Lưu trữ dữ liệu thu thập được vào mongodb.

Khởi tạo 1 consumer để nhận dữ liệu liên tục theo thời gian từ 1 topic trên broker server. Mỗi khi có 1 dữ liệu được gửi lên, dữ liệu đó sẽ được consumer nhận được và lưu vào trong database

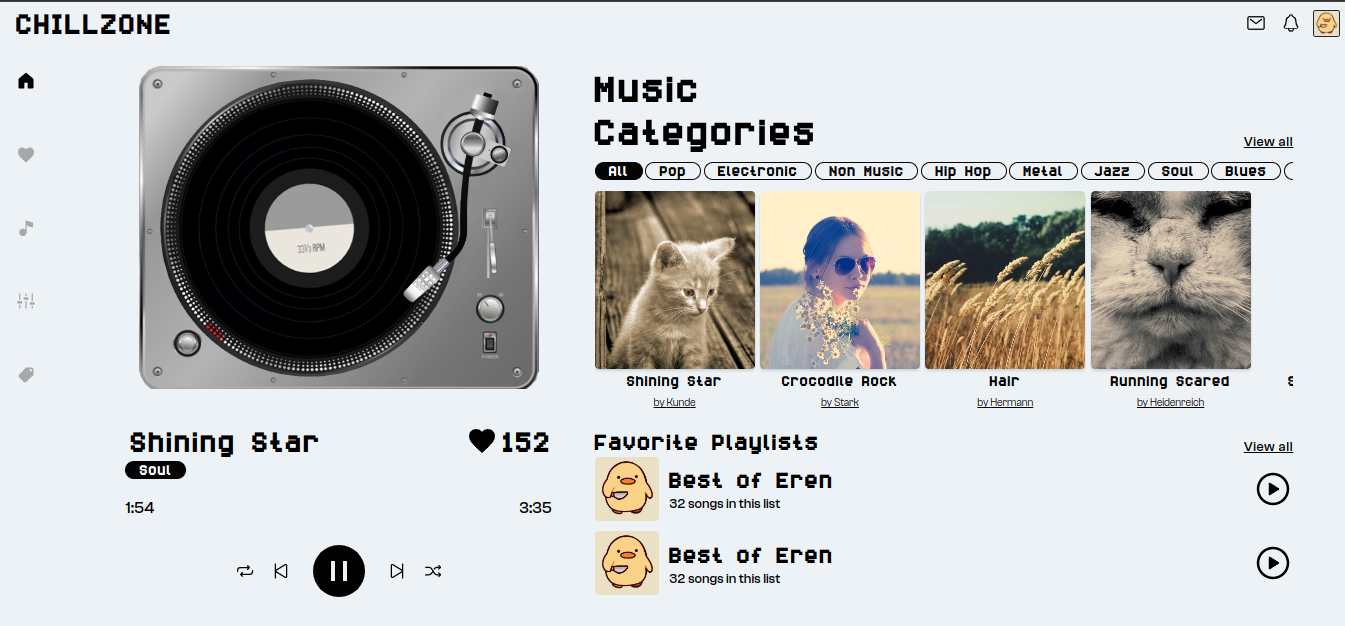


1. Kết nối metabase với mongodb và tiến hành trực quan hóa dữ liệu bằng metabase

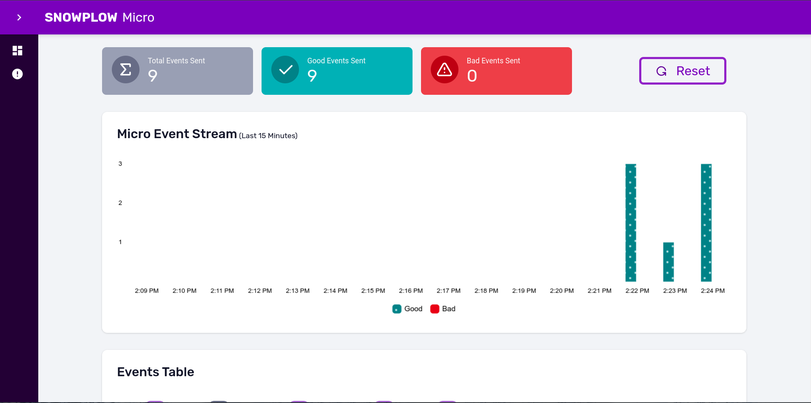
# Phần 4: Kết quả đánh giá

## Kết quả đánh giá

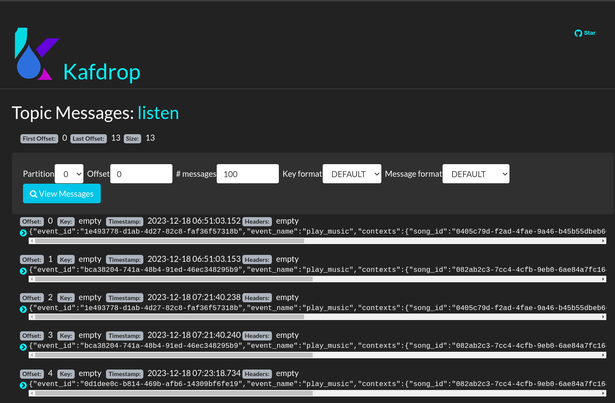
Xây dựng được trang web nghe nhạc cơ bản



Xây dựng được hệ thống thu thập dữ liệu theo thời gian thực



Trực quan hóa dữ liệu thu thập được



## **Trực quan hóa dữ liệu thu thập được**

## Mô tả các biểu đồ phân tích dữ liệu

|  |  |  |
| --- | --- | --- |
| STT | Tên biểu đồ | Mô tả |
| 1 | Biểu đồ các hành động được người dùng thực hiện | Biểu đồ thể hiện:Lượng quan tâm tới các nghệ sĩ**Lượng quan tâm tới các thể loại âm nhạc**Lượng nghe nhạc trên trang web |
| 2 | Biểu đồ thống kê só lượng user đã sử dụng trang web ( đã đăng kí tài khoản và không đăng kí tài khoản) | Biều đồ thể hiện :Tổng số lượng user đã đăng kí tài khoán sử dụng trang web**Tổng số lượng user chưa đăng kí tài khoản mà có sử dụng trang web** |
| 3 | Biểu đồ thống kê số lượt quan tâm/ nghe tới các thể loại âm nhạc trên trang web | Biểu đồ thể hiện :10 thể loại nhạc được nghe nhiều nhất |
| 4 | Biểu đồ Thống kê bài hát được nghe nhiều nhất đến thời điểm hiện tại | Biểu đồ thể hiện:10 bài hát được nghe nhiều nhất. |
|  |  |  |

## 

## 

## Mỗi khi dữ liệu thu thập được thêm vào , các biểu đồ sẽ cập nhập lại dữ liệu mới nhất mỗi 5 phút 1 lần.