**CÔNG TY CỔ PHẦN MISA JSC**

**---🙠**🕮**🙢---**



**NGÀY HỘI SẢN XUẤT**

**NGHIÊN CỨU GIẢI PHÁP ĐỒNG BỘ DỮ LIỆU PHỤC VỤ TÌM KIẾM DỮ LIỆU LỚN**

**CHO CÁC SẢN PHẨM CỦA MISA**

BÁO CÁO NGHIÊN CỨU

Nhóm: Dream Chaser From CRM

**HÀ NỘI, 08-2022**

**CÔNG TY CỔ PHẦN MISA JSC**

**---🙠**🕮**🙢---**



**NGÀY HỘI SẢN XUẤT**

**NGHIÊN CỨU GIẢI PHÁP ĐỒNG BỘ DỮ LIỆU PHỤC VỤ TÌM KIẾM DỮ LIỆU LỚN**

**CHO CÁC SẢN PHẨM CỦA MISA**

BÁO CÁO NGHIÊN CỨU

Nhóm: Dream Chaser From CRM

**Cán bộ hướng dẫn: Nguyễn Quang Hoàng**

**HÀ NỘI, 08-2022**

# **NHẬN XÉT CỦA BAN GIÁM KHẢO**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….………………………………………………………….**

**……………………………………….……………………………………………………………………………………………………………………………………………………………**

|  |
| --- |
| Hà Nội, ngày …. Tháng …. Năm 2022 |
| Ký tên |
|  |

# **LỜI NÓI ĐẦU**

Với sự phát triển mạnh mẽ của công ty cổ phần MISA, sản phầm của công ty ngày càng phát triển cả về số lượng và chất lượng, kéo theo đó là tập dữ liệu khách hàng khổng lồ từ các phần mềm. MISA luôn mong muốn đem lại cho người dung trải nghiệm dịch vụ tốt nhất nhằm đem lại hiệu quả về năng suất cũng như chất lượng công việc lên hàng đầu, luôn đề cao tinh thần phụ sự xã hội. Tuy nhiên, cùng với việc phát triển lớn mạnh của các phần mềm nhà MISA thì dữ liệu từ khách hàng là vô cùng lớn, việc dữ liệu lớn được lưu trữ trên cơ sở dữ liệu quan hệ cho ra một hiệu suất truy vấn không được tốt, MISA cũng đã có những sự cố liên quan đến việc tìm kiếm dữ liệu lớn. Các công cụ tìm kiếm ra đời nhằm giải quyết tình trạng này như: Apache Solr, MeiliSearch, Sphinx, Manticore và Elasticsearch là một trong những công cụ tìm kiếm giúp tôi ưu việc tìm kiếm nhanh dữ liệu lớn. Tuy nhiên, việc phát sinh vấn đề đồng bộ dữ liệu giữa cơ sở dữ liệu quan hệ sang cơ sở dữ liệu tìm kiếm là vấn đề lớn nhất, Apache Kafka là một nền tảng theo kiến trúc phân tán cho phép lưu trữ sự kiện và xử lý dữ liệu luồng mã nguồn mở được phát triển bởi Apache Software, với khả năng chịu lỗi cao và là hệ thống nhắn tin nhanh, kafka đang dần được thay thế cho hệ thống nhắn tin truyền thống. Nó được sử dụng cho các hệ thống nhắn tin thông thường trong các ngữ cảnh khác nhau.

**MỤC LỤC**

[NHẬN XÉT CỦA BAN GIÁM KHẢO 3](#_Toc112272886)

[MỤC LỤC 5](#_Toc112272887)

[DANH MỤC HÌNH ẢNH 6](#_Toc112272888)

[DANH MỤC BẢNG BIỂU 7](#_Toc112272889)

[CHƯƠNG 1: GIỚI THIỆU CHUNG 8](#_Toc112272890)

[1.1 Lý do chọn đề tài 8](#_Toc112272891)

[1.2 Giới thiệu 8](#_Toc112272892)

[1.1.1 Bài toán tìm kiếm nhanh với dữ liệu lớn 8](#_Toc112272893)

[1.1.2 Tìm kiếm nhanh trên các công cụ tìm kiếm 8](#_Toc112272894)

[1.1.3 Cơ chế đồng bộ dữ liệu 8](#_Toc112272895)

[1.3 Tiềm năng ứng dụng của bài toán tìm kiếm dữ liệu lớn và đồng bộ dữ liệu real-time tại công ty cổ phần MISA JSC 8](#_Toc112272896)

[1.4 Mục tiêu của nghiên cứu 8](#_Toc112272897)

[1.5 Cấu trúc bài nghiên cứu 8](#_Toc112272898)

[CHƯƠNG 2. ĐỒNG BỘ DỮ LIỆU REAL-TIME GIỮA CSDL GỐC VÀ CSDL TÌM KIẾM 9](#_Toc112272899)

[2.1 Cơ chế đồng bộ CDC 9](#_Toc112272900)

[2.2 Đồng bộ theo cơ chế Binlog Mysql 9](#_Toc112272901)

[2.3 Message Queue 10](#_Toc112272902)

[2.4. Các nền tảng và website hiện nay 13](#_Toc112272903)

[2.5 Kết luận 13](#_Toc112272904)

[CHƯƠNG 3. KẾT QUẢ THỰC NGHIỆM 14](#_Toc112272905)

[CHƯƠNG 4. KẾT LUẬN 15](#_Toc112272906)

[TÀI LIỆU THAM KHẢO 16](#_Toc112272907)

# **DANH MỤC HÌNH ẢNH**

**DANH MỤC BẢNG BIỂU**

# **CHƯƠNG 1: GIỚI THIỆU CHUNG**

## Vấn đề đưa ra

Hiện tại, với 40 sản phẩm, và tập dữ liệu khách hàng lớn của MISA, dữ liệu được lưu trữ trên cơ sở dữ liệu quan hệ đến một thời điểm sẽ rất lớn, dẫn đến hiệu suất truy vấn không được tốt. Qua khảo sát, cũng cho thấy một số sản phẩm của MISA đã gặp những sự cố liên quan đến việc tìm kiếm dữ liệu bị chậm, treo chương trình.

Tìm kiếm nhanh với dữ liệu lớn là một trong những vấn đề thách thức hầu hết các công ty trong và ngoài nước. Theo khảo sát tại các dự án tại Khối sản xuất, có tới trên 50% dự án đang gặp vấn đề về bài toán tìm kiếm dữ liệu lớn. Trong đó, 3 dự án đã có cách giải quyết, tuy nhiên những giải pháp đồng bộ dữ liệu giữa cơ sở dữ liệu quan hệ với cơ sở dữ liệu tìm kiếm đều gặp nhiều vấn đề về hiệu năng đồng bộ và ảnh hưởng đến tính đúng đắn của dữ liệu, đây là thách thức rõ nhất, vấn đề này sẽ được tìm hiểu và giải quyết trong bài nghiên cứu này.

## 1.2 Giới thiệu

### **1.1.1 Bài toán tìm kiếm nhanh với dữ liệu lớn**

### **1.1.2 Tìm kiếm nhanh trên các công cụ tìm kiếm**

### **1.1.3 Cơ chế đồng bộ dữ liệu**

## 1.3 Tiềm năng ứng dụng của bài toán tìm kiếm dữ liệu lớn và đồng bộ dữ liệu real-time tại công ty cổ phần MISA

## 1.4 Mục tiêu của nghiên cứu

## 1.5 Cấu trúc bài nghiên cứu

Dựa trên mục tiêu cụ thể đã trình bày, bài nghiên cứu được chia ra làm bốn chương với nội dung cụ thể như sau:

Chương 1: Giới thiệu chung

Chương 2: Đồng bộ dữ liệu real-time giữa CSDL gốc và CSDL tìm kiếm

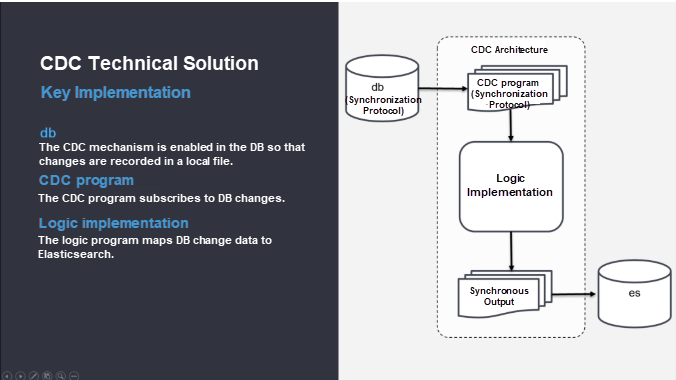
Chương 3: Kết quả thực nghiệm

Chương 4: Kết luận

**CHƯƠNG 2. ĐỒNG BỘ DỮ LIỆU REAL-TIME GIỮA CSDL GỐC VÀ CSDL TÌM KIẾM**

**2.1 Cơ chế đồng bộ CDC**

Dữ liệu thay đổi được ghi lại từ DB, được đẩy sang một chương trình trung gian và được đẩy đồng bộ đến Elaticsearch bằng cách sử dụng logic của chương trình trung gian. Dựa trên cơ chế CDC, dữ liệu chính xác được trả về ở tốc độ cực nhanh để đáp ứng với các truy vấn. Giải pháp này ít ảnh hưởng đến luồng chạy chính của các ứng dụng do đó, nó có thể được trừu tượng hóa và tách biệt với các hệ thống, làm cho nó phù hợp cho việc sử dụng quy mô lớn. Điều này được minh họa trong hình sau.

**

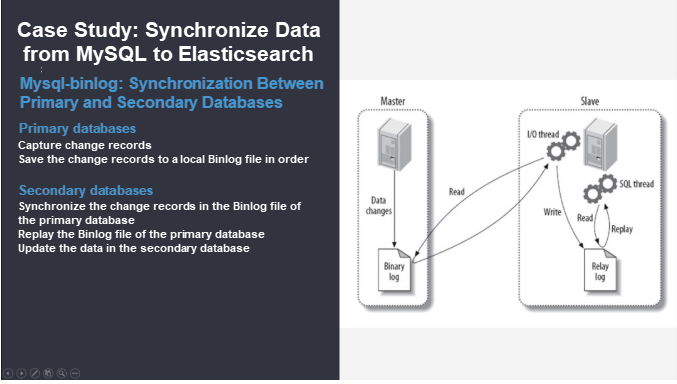
## 2.2 Đồng bộ theo cơ chế Binlog Mysql

Đồng bộ dữ liệu sử dụng cơ chế Binlog Mysql Binlog là một cơ chế chức năng được cung cấp bởi MySQL. BinLog là một tệp nhị phân trên đĩa, chứa tất cả các sự kiện thay đổi nội dung hoặc cấu trúc của cơ sở dữ liệu MySQL, ví dụ: chèn, cập nhật, xóa. Tất cả các thay đổi sẽ được lưu thêm vào (append only) vào một file log được đánh số thứ tự. Mỗi thay đổi được đặc trưng bởi hai tham số: file log được ghi vào vị trí offset trên file log. MySql Binlog có ba loại:

* Statement Based: lưu trữ các câu lệnh làm thay đổi dữ liệu
* Row based: lưu trữ các dữ liệu thay đổi
* Mixed: lưu trữ cả statement và data thay đổi của từng row

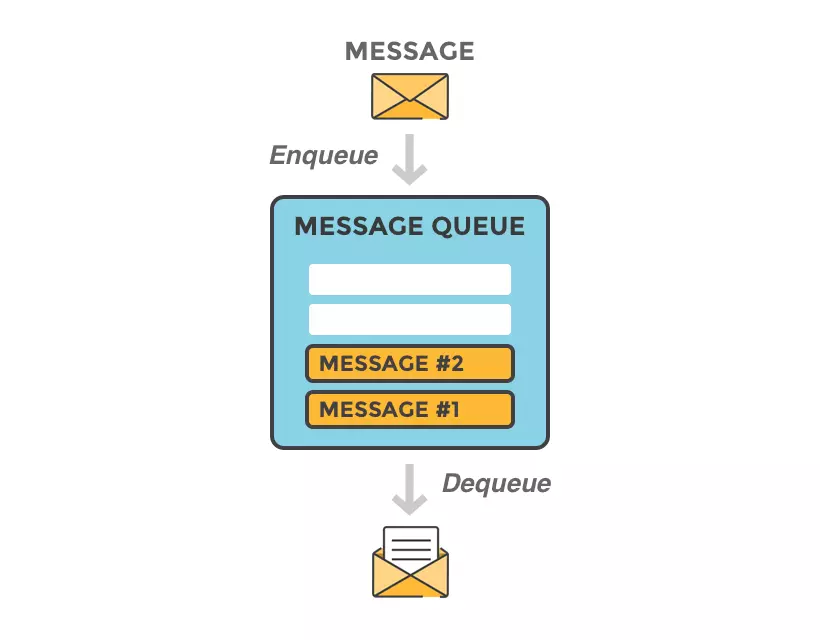
Mysql hỗ trợ cơ chế replicate binary log. MySql sẽ streamming tất cả các thay đổi của database từ master tới slave. Toàn bộ binary log sẽ được replicate realtime tới slave. Nó ban đầu được thiết kế để đồng bộ hóa giữa cơ sở dữ liệu chính và phụ:

* Cơ sở dữ liệu chính ghi dữ liệu vào tệp binlog khi cơ chế BinLog được bật.
* Cơ sở dữ liệu phụ lấy dữ liệu binlog từ cơ sở dữ liệu chính, phát lại dữ liệu binlog và cập nhật dữ liệu trong cơ sở dữ liệu phụ.
* Khi bạn bật Binlog, lưu ý rằng: Một cơ sở dữ liệu chính thường được liên kết bởi nhiều cơ sở dữ liệu phụ và phải cập nhật các thay đổi thời gian thực cho hệ thống, dẫn đến tắc nghẽn tài nguyên máy chủ. Các bảng được đồng bộ hóa phải có các khóa chính.



## 2.3 Message Queue

Định nghĩa: Message queue là một kiến trúc cung cấp giao tiếp không đồng bộ. Ý nghĩa của queue ở đây chính là 1 hàng đợi chứa message chờ để được xử lý tuần tự theo cơ chế vào trước thì ra trước (FIFO - First In First Out).

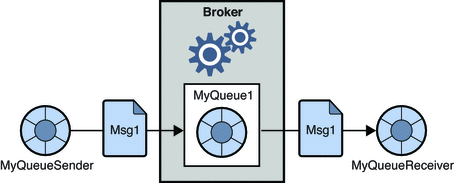
Một message là các dữ liệu cần vận chuyển giữa người gửi và người nhận. Vậy có thể hiểu đơn giản, message queue giống như một hòm thư email của chúng ta. Email có lẽ là ví dụ tốt nhất về giao tiếp không đồng bộ. Khi một email được gửi đi, người gửi tiếp tục xử lý những thứ khác mà không cần phản hồi ngay lập tức từ người nhận. Cách xử lý tin nhắn này tách người gửi khỏi người nhận để họ không cần phải tương tác với hàng đợi tin nhắn cùng một lúc.

Kiến trúc cơ bản của message queue rất đơn giản, bao gồm các thành phần như sau:

* Message: Thông tin được gửi (có thể là text, binary hoặc JSON)
* Producer: Service tạo ra thông tin, đưa thông tin vào message queue.
* Message Queue: Nơi chứa những message này, cho phép producer và consumer có thể trao đổi với nhau
* Consumer: Service nhận message từ message queue và xử lý
* Một service có thể vừa làm producer, vừa làm consumer

Một số message queue phổ biển hiện nay: Kafka, RamitMQ, Pulsar, ZeroMQ, IronMQ, RocketMQ, ..

Tại sao nên sử dụng Message queue.

* Ưu điểm:
  + Dễ scaling hệ thống: Vào giờ cao điểm, nhiều truy vấn, ta có thể tăng số lượng consumer lên để xử lý được nhiều messege hơn. Khi không cần ta có thể giảm lại.
  + Phân tán hệ thống: Giúp phân tách hệ thống thành nhiều service nhỏ hơn, mỗi service chỉ xử lý 1 chức năng nhất định theo cấu trúc [microservice](https://lcdung.top/phat-trien-phan-mem-theo-kien-truc-microservice/).
  + Đảm bảo duration/recovery: Do message đã được lưu trong queue, khi 1 service đang xử lý nhưng bị crash, ta không lo bị mất data vì có thể lấy message từ trong queue ra và retry. Trong 1 hệ thống có nhiều consumer, nếu vài consume crash cũng không làm crash cả hệ thống
  + Hỗ trợ rate limit, batch process: Trong trường hợp khả năng xử lý của hệ thống có hạn (chỉ có thể xử lý 100 lượt release/s) mà phải xử lý 10000 lượt release. Với message queue, ta có thể lấy từng lượt release chưa xử lý trong queue ra xử lý từ từ, không sợ bị mất.
* Nhược điểm:
  + Làm phức tạp hệ thống
  + Phải có message format: Từ 2 phía producer và consumer cần phải thống nhất format message để có thể gửi và nhận message.
  + Monitor Queue là cần thiết: Bạn cần phải theo dõi queue của mình để đảm bảo queue không quá nhiều hay đầy queue.
  + Khó xử lý đồng bộ: Không phải lúc nào message queue cũng là lựa chọn hàng đầu khi chúng ta xây dựng hệ thống. Sẽ có nhiều trường hợp hệ thống bắt buộc cần phải xử lý đồng bộ giữa các service, khi đó ta sẽ cần lựa chọn những cơ chế phù hợp hơn như Remote Procedure Invocation (RPI).
* Các loại message queue.
  + Point-to-point: Message queue có thể là kiểu point-to-point, tức là khi đó ta chỉ có một hàng đợi và một consumer duy nhất dể xử lý các tin nhắn trong hàng đợi:
  + Diagram

    Description automatically generatedPublisher-Subscriber. message queue có thể sử dụng định dạng Publisher-Subscriber, trong đó publisher (nhà sản xuất) gửi tin nhắn đến hàng đợi (trong trường hợp này được gọi là Topic) và tất cả subscriber (người đăng ký) vào cùng 1 Topic đều sẽ nhận được tin nhắn trong Topic đó:

## 2.4. Các nền tảng và website hiện nay

Magento là một nền tảng thương mại điện tử mã nguồn mở được viết bằng PHP. Nó sử dụng các framework PHP khác như Laminas và Symfony. Mã nguồn Magento được phân phối theo Bản quyền Phần mềm Mở v3.0. Magento được mua bởi Adobe Inc vào tháng 5/2018. Công cụ tìm kiếm được tíc hợp sẵn trong mangento cloud: Elasticsearch là chức năng mới được thêm vào ở các phiên bản Magento 2.4. Magento Cloud giải quyết bài toán đồng bộ bằng cách sử dụng zookeeper và kafka.

## 2.4 Tool đồng bộ bên thứ ba

**2.5 Kết luận**

**CHƯƠNG 3. KẾT QUẢ THỰC NGHIỆM**

# **CHƯƠNG 4. KẾT LUẬN**

# **TÀI LIỆU THAM KHẢO**