TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO THỰC TẬP TỐT NGHIỆP

ĐỀ TÀI

ỨNG DỤNG CHANGE DATA CAPTURE (CDC) ĐỂ CHUYỂN ĐỔI DỮ LIỆU GIỮA CÁC HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

Giảng viên hướng dẫn : Cô PHẠM THỊ MIÊN
Sinh viên thực hiện : HOÀNG GIA KIỆT
Lớp : CÔNG NGHỆ THÔNG TIN
Khoá : 62

Tp. Hồ Chí Minh, năm 2025

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO THỰC TẬP TỐT NGHIỆP

ĐỀ TÀI

ỨNG DỤNG CHANGE DATA CAPTURE (CDC) ĐỂ CHUYỂN ĐỔI DỮ LIỆU GIỮA CÁC HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

Giảng viên hướng dẫn : Cô PHẠM THỊ MIÊN
Sinh viên thực hiện : HOÀNG GIA KIỆT
Mã sinh viên : 6251071049
Lớp : CÔNG NGHỆ THÔNG TIN
Khoá : 62

Tp. Hồ Chí Minh, năm 2025

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

NHIỆM VỤ THIẾT KẾ THỰC TẬP TỐT NGHIỆP

Mã sinh viên: 6251071049 **Họ tên SV**: Hoàng Gia Kiệt

Khoá: 62 **Lớp**: Công nghệ thông tin

1. Tên đề tài

"Ứng dụng Change Data Capture (CDC) để chuyển đổi dữ liệu giữa các hệ quản trị cơ sở dữ liệu"

2. Nhiệm vụ thực tập tốt nghiệp

Về kiến thức

Ứng dụng được Change Data Capture (CDC) vào hệ thống backend.

Tìm hiểu và áp dụng Kafka và Debezium để xử lý dữ liệu thay đổi trong MySQL và MongoDB.

Sử dụng Django và Django Rest Framework để xây dựng ứng dụng mẫu.

Về kỹ năng

Nghiên cứu và phát triển các tính năng thực tế của hệ thống phục vụ yêu cầu doanh nghiệp.

Nâng cao kỹ năng làm việc nhóm, phối hợp với phòng ban backend để triển khai giải pháp.

Ứng dụng các công nghệ phù hợp vào việc xây dựng và phát triển hệ thống backend.

Thời gian thực tập

Ngày bắt đầu thực tập tốt nghiệp: Ngày 06 tháng 01 năm 2025.

Ngày hoàn thành báo cáo thực tập tốt nghiệp: Ngày 15 tháng 3 năm 2025.

Thông tin công ty

Tên công ty: Chi nhánh Công ty Cổ phần Viễn thông FPT.

Địa chỉ: Tòa nhà FPT, Lô L.29B-31B-33B đường Tân Thuận, KCX Tân Thuận, Phường Tân Thuận Đông, Quận 7, TP Hồ Chí Minh.

3. Giảng viên hướng dẫn

Họ và tên: ThS. Phạm Thị Miên.

Đơn vị công tác: Bộ môn Công nghệ thông tin Phân hiệu Trường Đại học Giao thông Vận tải Phân hiệu tại TP. Hồ Chí Minh.

TP. Hồ Chí Minh, ngày ... tháng ... năm 2025

Sinh viên thực hiện

Hoàng Gia Kiệt

LỜI CẢM ƠN

Lời đầu tiên em xin gửi lời cảm ơn sâu sắc đến cô Phạm Thị Miên đã truyền đạt kiến thức, hỗ trợ và giúp đỡ em hoàn thành học phần Thực tập tốt nghiệp.

Em cũng xin gửi lời cảm ơn đến Quý thầy cô Bộ môn Công nghệ thông tin Trường Đại học Giao thông Vận tải Phân hiệu tại TP. Hồ Chí Minh đã truyền đạt những kiến thức nền tảng cho chúng em và hỗ trợ em khi có những khó khăn hoàn thành đề tài của học phần.

Ngoài ra, em xin gửi lời cảm ơn anh Nguyễn Đông Đông đã hỗ trợ em trong quá trình thực hiện dự án.

Em xin chân thành cảm ơn cô./.

TP. Hồ Chí Minh, ngày ... tháng ... năm 2025

Sinh viên thực hiện

Hoàng Gia Kiệt

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày ... tháng ... năm 2025

Giảng viên hướng dẫn

ThS. Phạm Thị Miên

MỤC LỤC

DANH MỤC VIẾT TẮT

STT	Viết tắt	Ý nghĩa	Diễn giải
1	CDC	Change Data Capture	
2	AI	Artifical Intelligence	Trí tuệ nhân tạo
3	IoT	Internet of Things	Internet vạn vật
4	CNTT	Công nghệ thông tin	
5	DRF	Django REST Framework	
6	NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
7	ML	Machine Learning	Học máy
8	OOP	Object Oriented Programming	Lập trình hướng đối tượng
9	Web	Website	Trang web
10	CSDL	Cơ sở dữ liệu	
11	SQL	Structured Query Language	Ngôn ngữ truy vấn mang có cấu trúc
12	URL	Uniform Resource Locator	Hệ thống định vị tài nguyên thống nhất
13	CSRF	Cross Site Request Forgery	Giả mạo yêu cầu liên trang
14	API	Application Programming Interface	Giao diện lập trình ứng dụng
15	JSON	Javascript Object Notation	
16	XML	Xml Extensible Markup Language	
17	HTTP	Hypertext Transfer Protocol	Giao thức truyền tải siêu văn bản

STT	Viết tắt	Ý nghĩa	Diễn giải
18	RDBMS	Relational Database Management System	hệ thống quản lý cơ sở dữ liệu quan hệ
19	SSPL	Server Side Public License	Giấy phép công cộng phía máy chủ
20	BSON	Binary JSON	
21	RDBMS	Relational Database Management System	Hệ quản trị cơ sỡ dữ liệu quan hệ
22	FTEL	FPT Telecom	

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

CHƯƠNG 1: TỔNG QUAN VỀ CÔNG TY

1.1. Giới thiệu về tập đoàn FPT

Hình 1.1. Logo tập đoàn FPT

Năm 1988, 13 nhà khoa học trẻ thành lập Công ty FPT với mong muốn xây dựng "một tổ chức kiểu mới, giàu mạnh bằng nỗ lực lao động sáng tạo trong khoa học kỹ thuật và công nghệ, làm Khách hàng hài lòng, góp phần hưng thịnh Quốc gia, đem lại cho mỗi thành viên của mình điều kiện phát triển đầy đủ nhất về tài năng và một cuộc sống đầy đủ về vật chất, phong phú về tinh thần."

Không ngừng đổi mới, liên tục sáng tạo và luôn tiên phong mang lại cho Khách hàng các sản phẩm/ giải pháp/ dịch vụ công nghệ tối ưu nhất, FPT trở thành Công ty CNTT-VT lớn nhất trong khu vực kinh tế tư nhân của Việt Nam với gần 37.180 Cán bộ Nhân viên, trong đó có 24.068 kỹ sư CNTT, lập trình viên, chuyên gia công nghệ; hệ thống 46 chi nhánh, văn phòng tại 27 quốc gia và vùng lãnh thổ bên ngoài Việt Nam. FPT cũng là doanh nghiệp dẫn đầu trong các lĩnh vực: Xuất khẩu phần mềm, Tích hợp hệ thống; Phát triển phần mềm; Dịch vụ CNTT.

Trong 34 năm qua, FPT không chỉ tiên phong xây dựng, phát triển các phần mềm thương hiệu Việt; đưa công nghệ vào cuộc sống; hiện đại hóa các ngành kinh tế xương sống của Quốc gia; đẩy mạnh giáo dục & đào tạo thế hệ trẻ theo hướng thực học, thực nghiệp, mà còn tiên phong trong lĩnh vực xuất khẩu phần mềm, góp phần đưa trí tuệ Việt Nam ra thế giới. Trong nước, hầu hết các hệ thống thông tin lớn trong các cơ quan nhà nước và các ngành kinh tế trọng điểm của Việt Nam đều do FPT xây dựng và phát triển.

Trong cuộc cách mạng 4.0, FPT là Công ty Việt Nam tiên phong trong việc nghiên cứu và phát triển các công nghệ mới về trí tuệ nhân tạo, dữ liệu lớn, điện toán đám mây, di động,… FPT cũng là doanh nghiệp tiên phong đồng hành cùng với các tập đoàn công nghệ hàng đầu thế giới để tạo nên các nền tảng công nghệ số tiên tiến nhất như GE (Predix), Siemens (MindSphere), Airbus (Skywise), Amazon AWS…

Vị thế của FPT trên toàn cầu đã được công nhận và khẳng định thông qua danh sách Khách hàng gồm hơn 700 doanh nghiệp lớn trên thế giới, đặc biệt trong đó có gần 100 Khách hàng nằm trong danh sách Fortune 500. Một số tên tuổi khách hàng lớn có thể kể đến Toshiba, Hitachi, Airbus, Deutsche Bank, Unilever, Panasonic…

Với định hướng tiên phong nghiên cứu và ứng dụng các xu hướng công nghệ mới nhất, FPT sẽ tiếp tục là đơn vị đi đầu về chuyển đổi số cho Khách hàng, đưa các công nghệ mới như AI, Big Data, IoT… vào các giải pháp trong mọi lĩnh vực như giao thông thông minh, y tế thông minh, chính phủ số… [1]s

1.1.1. Tầm nhìn chiến lược

FPT tiếp tục theo đuổi mục tiêu lớn dài hạn là trở thành doanh nghiệp số và đứng trong Top 50 công ty hàng đầu thế giới về cung cấp dịch vụ, giải pháp chuyển đổi số toàn diện vào năm 2030.

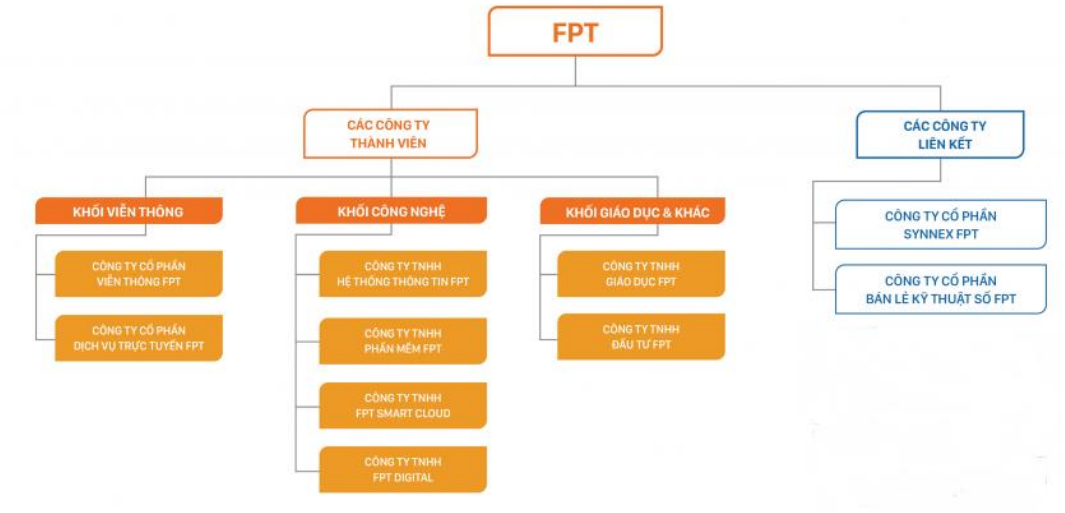
Trong giai đoạn 2021 – 2023, FPT mong muốn trở thành đồng minh tin cậy của các doanh nghiệp, tổ chức đem lại những trải nghiệm số xuất sắc thông qua quản trị, vận hành dựa trên dữ liệu gần thời gian thực. Để đạt được mục tiêu trên, FPT xây dựng các chương trình hành động cân bằng/toàn diện ở cả ba khía cạnh kinh doanh, công nghệ và con người.

1.1.2. Mạng lưới toàn cầu

Với hệ thống 46 văn phòng tại nước ngoài, chúng tôi có thể cùng lúc sử dụng nguồn lực trên toàn cầu và tại Việt Nam để cung cấp dịch vụ/giải pháp cho Khách hàng một cách hiệu quả nhất.

1.1.3. Công ty con và công ty liên kết

FPT cung cấp giải pháp CNTT tổng thể trong 3 lĩnh vực kinh doanh cốt lõi gồm Công nghệ, Viễn thông, Giáo dục với 08 Công ty thành viên trực thuộc và 02 công ty liên kết chính.



Hình 1.2. Công ty con và công ty liên kết

1.2. Giới thiệu về công ty FPT Telecom

Hình 1.3. Logo công ty FPT Telecom

Là công ty thành viên thuộc Tập đoàn FPT, Công ty Cổ phần Viễn thông FPT (tên gọi tắt là FPT Telecom) hiện là một trong những nhà cung cấp dịch vụ Viễn thông và Internet hàng đầu khu vực.

Thành lập ngày 31/01/1997, khởi nguồn từ Trung tâm Dịch vụ Trực tuyến do 4 thành viên sáng lập cùng sản phẩm mạng Intranet đầu tiên của Việt Nam mang tên "Trí tuệ Việt Nam – TTVN", sản phẩm được coi là đặt nền móng cho sự phát triển của Internet tại Việt Nam.

Với sứ mệnh tiên phong mang Internet, mang kết nối đến với người dân Việt Nam cùng mong muốn lớn lao mỗi gia đình Việt Nam đều sử dụng ít nhất một dịch vụ của Công ty, FPT Telecom đang nỗ lực thực thi Chiến lược "Mang đến trải nghiệm tuyệt vời cho khách hàng" trên cơ sở phát huy giá trị văn hóa cốt lõi "Lấy khách hàng làm trọng tâm" và nền tảng sức mạnh công nghệ FPT, từ đó tiên phong trở thành Nhà cung cấp dịch vụ số có trải nghiệm khách hàng vượt trội, tốt nhất tại Việt Nam. [2]

1.2.1. Lĩnh vực kinh doanh

- Cung cấp hạ tầng mạng viễn thông cho dịch vụ Internet băng rộng
- Dịch vụ giá trị gia tăng trên mạng Internet, điện thoại di động
- Dịch vụ Truyền hình trả tiền
- Dịch vụ tin nhắn, dữ liệu, thông tin giải trí trên mạng điện thoại di động
- Thiết lập hạ tầng mạng và cung cấp các dịch vụ Viễn thông, Internet
- Xuất nhập khẩu thiết bị Viễn thông và Internet.
- Dịch vụ Viễn thông cố định nội hạt
- Dịch vụ Viễn thông giá trị gia tăng
- Dịch vụ Viễn thông cố định đường dài trong nước
- Cung ứng dịch vụ trung gian thanh toán

1.2.2. Các dịch vụ cho khách hàng đại chúng

- Dịch vụ truy nhập Internet băng rộng cố định mặt đất công nghệ FTTH/xPON
- Truyền hình FPT Play sở hữu gần 200 kênh truyền hình trong nước và quốc tế, cung cấp các ứng dụng học tập online, rèn luyện thể thao tại nhà, ứng dụng Sự kiện trực tuyến...Cho phép người dùng trải nghiệm nhiều nhu cầu: học tập, giải trí, chơi game...trên đa nền tảng thiết bị.
- Dịch vụ nội dung, ứng dụng trên Internet: FPT Play Box (Voice Remote), FPT Play, Foxy, Hi FPT, Fshare, Fsend...
- Ví điện tử, Cổng thanh toán điện tử Foxpay
- Dịch vụ, sản phẩm IoT/Smart Home: FPT Camera, iHome.

1.2.3. Các dịch vụ cho khách hàng tổ chức, doanh nghiệp

- Truyền dẫn số liệu: Trong nước (kết nối nội hạt, kết nối liên tỉnh) và quốc tế (IPLC, MPLS, IEPL).
- Kênh thuê riêng Internet: NIX, GIA, Asia Transit.
- Dịch vụ thoại: Trong nước (Điện thoại cố định, VoIP, đầu số 1800/1900) và quốc tế.
- Dữ liệu trực tuyến: Tên miền, lưu trữ dữ liệu và email, thuê máy chủ và chỗ đặt máy chủ, thuê tủ Rack.
- Dịch vụ quản lý: Hội nghị truyền hình, điện toán đám mây, tích hợp hệ thống, dịch vụ bảo mật.
- Ví Doanh nghiệp, Cổng thanh toán điện tử Foxpay; Dịch vụ Hỗ trợ thu hộ chi hộ; Và các dịch vụ hỗ trợ thanh toán khác
- Dịch vụ Điện toán đám mây – FPT HI GIO CLOUD: là dịch vụ nền tảng điện toán đám mây (Cloud Infrastructure Service) được phát triển bởi FPT Telecom và Internet Initiative Japan (IIJ).

1.2.4. Các giải thưởng tiêu biểu

- Giải Vàng tại Giải thưởng Kinh doanh quốc tế IBA Stevie 2021 - Hạng mục Sản phẩm mới
- Top 10 Doanh nghiệp Công nghệ thông tin Việt Nam 2021
- Giải thưởng "Nhà cung cấp dịch vụ Internet Cố định được khách hàng hài lòng nhất về Chất lượng Dịch vụ và Chăm sóc Khách hàng năm 2021" của IDG.
- Top 10 Doanh nghiệp Hạ tầng số xuất sắc 2020 – Chương trình Top 10 Doanh nghiệp CNTT Việt Nam 2020.
- Top 10 Doanh nghiệp tiêu biểu ASIA/ ASIA Typical Enterprise.
- 02 Giải thưởng hạng mục "Giải pháp ứng dụng cho công dân/cộng đồng thông minh" thuộc Giải thưởng Thành phố thông minh Việt Nam 2020 cho 2 sản phẩm: Truyền hình FPT và FPT Play - FPT Play Box.
- Giải thưởng hạng mục "Giải pháp an ninh, an toàn, cấp cứu, cứu nạn" thuộc Giải thưởng Thành phố thông minh Việt Nam 2020 cho sản phẩm FPT Camera.

- Top Doanh nghiệp niêm yết có Năng lực cạnh tranh tốt nhất năm 2019 - Vietnam the Best Company.
- Giải thưởng "Nhà cung cấp dịch vụ Internet Cố định được khách hàng hài lòng nhất về Chất lượng Dịch vụ và Chăm sóc Khách hàng năm 2019" của IDG.

Và các giải thưởng khác…

1.2.5. Các chứng chỉ quốc tế

Hình 1.4. Các chứng chỉ quốc tế (1)

Hình 1.5. Các chứng chỉ quốc tế (2)

1.3. Quá trình thực tập

Thời gian thực tập tại công ty theo hợp đồng là 12 tháng, từ ngày 06/01/2025 đến ngày 31/12/2025. Quá trình thực tập tại công ty được trình bày qua bảng sau:

Tuần 1 (06 – 10/01)	- Công ty chào đón sinh viên - Đào tạo về công ty FPT Telecom - Đào tạo quy tắc ứng xử, các kỹ năng làm việc - Làm quen với văn hoá của phòng ban - Tìm hiểu về Django, Django REST Framework, Streaming data
Tuần 2 (13 – 17/01)	- Tìm hiểu về MySQL, MongoDB, OpenSearch, CDC… - Thực hiện task demo nhỏ về CDC

Tuần 3 (20 – 24/01)	- Thực hiện task chỉnh sửa API, cách thức xử lý của chức năng cho dự án của công ty (1)
Tuần 4 (Nghỉ tết)	
Tuần 5 (03 – 07/02)	Tiếp tục làm task (1) được giao
Tuần 6 (10 – 14/02)	- Hoàn thành task (1) - Tìm hiểu về unit test - Chọn đề tài
Tuần 7 (17 – 21/02)	- Thực hiện task viết unit test cho dự án công ty (2) - Lên kế hoạch thực hiện đề tài, xây dựng báo cáo
Tuần 8 (24 – 28/02)	- Hoàn thành task (2) - Dựng source code dự án - Xây dựng Chương 1, 2, 3
Tuần 9 (03 – 07/3)	- Xây dựng Chương 4, 5 - Xây dựng chương trình demo
Tuần 10 (10 – 14/3)	- Hoàn thành báo cáo

Bảng 1.1. Quá trình thực tập tại công ty

CHƯƠNG 2: TỔNG QUAN VỀ ĐỀ TÀI

2.1. Tình hình nghiên cứu

Trong quá trình thực tập tại công ty FPT Telecom (FTEL), em có cơ hội làm việc cùng phòng phát triển hệ thống backend và tìm hiểu về cách công ty áp dụng Change Data Capture (CDC) vào hệ thống. Hiện tại, FTEL sử dụng CDC để chuyển đổi dữ liệu giữa các hệ quản trị cơ sở dữ liệu nhằm đảm bảo dữ liệu được cập nhật liên tục và kịp thời giữa các thành phần hệ thống.

Change Data Capture (CDC) là một kỹ thuật quan trọng trong quản trị dữ liệu, cho phép theo dõi và xử lý thay đổi trong cơ sở dữ liệu theo thời gian thực. Các nghiên cứu trước đây đã chỉ ra rằng CDC giúp cải thiện hiệu suất trong việc đồng bộ dữ liệu giữa các hệ thống khác nhau, giảm thiểu độ trễ và đảm bảo tính nhất quán.

Bên cạnh đó, việc áp dụng CDC trong môi trường microservice, hệ thống phân tán và xử lý dữ liệu thời gian thực đang là xu hướng phổ biến. Đặc biệt, các tổ chức có nhu cầu đồng bộ dữ liệu từ cơ sở dữ liệu quan hệ (RDBMS) sang NoSQL (như MongoDB) để phục vụ mục đích phân tích, báo cáo, đồng bộ dữ liệu… ngày càng quan tâm đến CDC.

2.2. Lý do chọn đề tài

Trong thời gian thực tập tại FTEL, em được tìm hiểu về cách công ty đang áp dụng CDC để chuyển đổi dữ liệu giữa MySQL và MongoDB trong hệ thống.

Việc đồng bộ dữ liệu giữa các hệ quản trị cơ sở dữ liệu khác nhau là một thách thức lớn trong nhiều hệ thống phần mềm. Các phương pháp truyền thống như ETL (Extract, Transform, Load) thường có độ trễ cao và yêu cầu tài nguyên lớn. Do đó, CDC như một giải pháp hiệu quả giúp truyền tải dữ liệu theo thời gian thực mà không ảnh hưởng đến hiệu suất hệ thống, giúp hệ thống luôn được vận hành mà không ảnh hưởng đến trải nghiệm của khách hàng.

Hơn nữa, với sự phát triển của các hệ thống phân tán và microservice, nhu cầu cập nhật dữ liệu liên tục giữa các thành phần hệ thống ngày càng trở nên quan trọng. Đề

tài này tập trung vào việc ứng dụng CDC để chuyển đổi dữ liệu giữa MySQL và MongoDB, giúp tối ưu hóa quy trình đồng bộ dữ liệu trong các hệ thống hiện đại.

2.3. Mục tiêu đề tài

- Tìm hiểu và áp dụng kỹ thuật CDC để đồng bộ hoá dữ liệu giữa MySQL và MongoDB.
- Tìm hiểu và áp dụng Kafka làm message broker để truyền dữ liệu thay đổi từ MySQL đến MongoDB.
- Xây dựng được hệ thống mẫu để minh hoạ việc được ứng dụng trong thực tế.

2.4. Nội dung đề tài

- Nghiên cứu tổng quan về Change Data Capture (CDC).
- Tìm hiểu cách hoạt động của Debezium và Kafka trong việc xử lý thay đổi dữ liệu.
- Xây dựng hệ thống CDC giữa MySQL và MongoDB sử dụng Django, Django Rest Framework, Kafka và Debezium.

2.5. Nội dung cuốn báo cáo

Chương 1: Tổng quan về công ty

Chương 2: Tổng quan về đề tài

Chương 3: Cơ sở lý thuyết

Chương 4: Thiết kế cơ sở dữ liệu

Chương 5: Xây dựng hệ thống

CHƯƠNG 3: CƠ SỞ LÝ THUYẾT

3.1. Ngôn ngữ lập trình Python

3.1.1. Giới thiệu về Python

Python là một ngôn ngữ lập trình bậc cao, thông dịch, và đa mục đích, được tạo ra bởi Guido van Rossum và ra mắt lần đầu vào năm 1991. Python nổi tiếng với cú pháp đơn giản, dễ học, và dễ đọc, giúp lập trình viên phát triển phần mềm một cách nhanh chóng và hiệu quả. Python được thiết kế với triết lý "Code rõ ràng hơn là code phức tạp", khiến nó trở thành lựa chọn lý tưởng cho người mới bắt đầu cũng như các lập trình viên giàu kinh nghiệm. [3]

Hình 3.1. Logo ngôn ngữ lập trình Python

Python hỗ trợ nhiều paradigm lập trình, bao gồm lập trình hướng đối tượng (OOP), lập trình hàm, và lập trình thủ tục. Ngôn ngữ này đi kèm với một thư viện chuẩn phong phú (Standard Library) cung cấp nhiều module và hàm có sẵn, giúp giải quyết các tác vụ phổ biến như xử lý chuỗi, làm việc với file, kết nối mạng và quản lý cơ sở dữ liệu. Python cũng có một hệ sinh thái rộng lớn với các thư viện và framework mạnh mẽ như NumPy, Pandas, TensorFlow, Django, và Flask, hỗ trợ phát triển từ khoa học dữ liệu, trí tuệ nhân tạo đến phát triển web và tự động hóa.

Python hoạt động trên nhiều hệ điều hành như Windows, macOS, Linux, và là một ngôn ngữ mã nguồn mở, được cộng đồng phát triển rộng khắp hỗ trợ. Python được sử dụng trong nhiều lĩnh vực như phát triển phần mềm, phân tích dữ liệu, học máy (ML), trí tuệ nhân tạo (AI), tự động hóa, và xử lý ngôn ngữ tự nhiên (NLP). Nhờ tính linh hoạt, mạnh mẽ và khả năng mở rộng, Python đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất hiện nay.

3.1.2. Hệ sinh thái và thư viện, công nghệ phổ biến

Phát triển web: Django, Flask.

Khoa học và phân tích dữ liệu: Pandas, NumPy, Matplotlib.

Học máy và trí tuệ nhân tạo: TensorFlow, PyTorch, Scikit-learn.

Phát triển game: Pygame.

Cào dữ liệu: BeautifulSoup, Scrapy.

Ứng dụng desktop: Tkinter, PyQt.

IoT: MicroPython, Raspberry Pi. [4]

3.1.3. Ưu điểm

Cú pháp đơn giản, dễ học: Cú pháp gần gũi với ngôn ngữ tự nhiên, dễ đọc và viết với người mới bắt đầu.

Thư viện phong phú: Hệ sinh thái với nhiều thư viện và framework hỗ trợ nhiều lĩnh vực (AI, ML, Web, dữ liệu).

Đa nền tảng: Chạy trên hầu hết các hệ điều hành như Windows, macOS, và Linux.

Lập trình đa mục đích: Hỗ trợ nhiều paradigms (hướng đối tượng, hàm, thủ tục).

Cộng đồng lớn: Cộng đồng mạnh mẽ và tài liệu hỗ trợ dồi dào.

Khả năng mở rộng: Có thể tích hợp với các ngôn ngữ khác như C/C++ để tăng hiệu suất.

3.1.4. Nhược điểm

Hiệu suất thấp: Chạy chậm hơn các ngôn ngữ biên dịch như C++ hay Java do là ngôn ngữ thông dịch.

Quản lý bộ nhớ kém: Tiêu tốn nhiều bộ nhớ, không thích hợp cho các ứng dụng cần tối ưu cao.

GIL (Global Interpreter Lock): Giới hạn khả năng xử lý đa luồng trong một số tác vụ.

Hạn chế trong lập trình di động: Không phổ biến trong phát triển ứng dụng di động.

Không phù hợp cho ứng dụng thời gian thực: Thiếu tốc độ xử lý cần thiết cho các hệ thống thời gian thực.

3.2. Công nghệ Django

3.2.1. Giới thiệu về Django

Django là một web framework Python bậc cao, giúp phát triển nhanh chóng và thiết kế thực tế, sạch sẽ. Được xây dựng bởi các nhà phát triển có kinh nghiệm, Django giải quyết các rắc rối trong khi phát triển web, vì vậy lập trình viên chỉ cần tập trung vào việc viết ứng dụng của mình mà không cần phải phát minh lại "bánh xe". Django là 1 phần mềm mã nguồn mở và miễn phí. [5]

3.2.2. Kiến trúc của Django

Django tuân thủ theo thiết kế kiến trúc được gọi là MVT (Model – View – Template).

- Model:
    - Đại diện cho dữ liệu, có thể xem các class là 1 bảng trong CSDL.
    - Thao tác với CSDL.
- View:
    - Xử lý logic của ứng dụng và thao tác với model.
    - Đóng vai trò như là bộ não của ứng dụng.
- Template:
    - Xác định giao diện người dùng hiển thị như thế nào.
    - Là mẫu cho các trang web.

3.2.3. Các tính năng chính

Giao diện admin: Django cung cấp giao diện admin được tích hợp sẵn để quản lý dữ liệu ứng dụng. Tính năng này sẽ được tự động tạo ra dựa trên model và có thể tuỳ chỉnh.

ORM (Object-Relational Mapping): ORM của Django cho phép các nhà phát triển sử dụng Python để tương tác với CSDL thay vì SQL. Hỗ trợ nhiều hệ quản trị CDSL như: MySQL, PostgreSQL, MariaDB, Oracle, SQLite.

Định tuyến URL: Bộ điều phối URL của Django ánh xạ các mẫu URL đến views, giúp dễ dàng thiết kế các URL sạch và có thể đọc được.

Xử lý form: Django cung cấp tính năng xử lý form mạnh mẽ, bao gồm các tính năng xác thực dữ liệu và bảo mật để bảo vệ chống lại các cuộc tấn công CSRF.

Hệ thống xác thực: Django bao gồm một hệ thống xác thực toàn diện với đăng nhập người dùng, đăng xuất, quản lý mật khẩu và quyền. [6]

3.3. Công nghệ Django REST Framework

3.3.1. Giới thiệu về Django REST Framework

Django REST Framework (DRF) là một thư viện mạnh mẽ dành cho việc xây dựng các API (Application Programming Interface) trên nền tảng Django, một trong những framework phổ biến nhất trong việc phát triển web bằng Python. [7]

3.3.2. Các thành phần chính của DRF

Serializers: Là công cụ giúp chuyển đổi dữ liệu giữa các định dạng phức tạp như mô hình Django và định dạng JSON, XML hoặc các định dạng khác. Đây là thành phần cốt lõi giúp DRF xử lý dữ liệu đầu vào và đầu ra một cách hiệu quả.

Views: Trong DRF cung cấp các lớp để xử lý các yêu cầu HTTP và trả về các phản hồi thích hợp. DRF hỗ trợ hai loại views chính: Function-Based Views (FBV) và Class-Based Views (CBV). Các views này có thể được tùy chỉnh hoặc sử dụng các lớp có sẵn như *APIView*, *GenericAPIView* hoặc các lớp *viewset* như *ModelViewSet*.

Authentication & Permissions: DRF hỗ trợ nhiều cơ chế xác thực khác nhau như Token Authentication, Session Authentication, OAuth, và các cơ chế tùy chỉnh khác. Permissions cho phép kiểm soát truy cập vào các view hoặc các phương thức cụ thể, giúp bảo vệ API khỏi các truy cập trái phép.

Routers: Tự động ánh xạ các URL tới các viewset tương ứng, giúp đơn giản hóa việc định tuyến các yêu cầu trong ứng dụng.

Pagination: DRF cung cấp các lớp phân trang để kiểm soát số lượng dữ liệu trả về trong mỗi yêu cầu, giúp tối ưu hóa hiệu suất của API khi làm việc với các tập dữ liệu lớn.

3.4. Hệ quản trị cơ sở dữ liệu MySQL

3.4.1. Giới thiệu về MySQL

MySQL là một mã nguồn mở, hệ thống quản lý cơ sở dữ liệu quan hệ - Relational Database Management System (RDBMS) có hiệu quả trong việc lưu trữ, sắp xếp và truy xuất thông tin. Được phát triển bởi Oracle và dựa trên Ngôn ngữ truy vấn có cấu trúc - Structured Query Language (SQL), được sử dụng phổ biến trong các ứng dụng Web, cổng mua sắm và ứng dụng dữ liệu lớn (Big Data). [8]

3.4.2. Những tính năng chính của MySQL

Mã nguồn mở (Open source)

MySQL là phần mềm nguồn mở cho phép người dùng tải xuống MySQL và bắt đầu sử dụng nó mà không mất bất kỳ khoản phí nào. Nó có Giấy phép công cộng ((GPL) GNU. Nó mang lại sự tự do cho các dự án nguồn mở và làm cho MySQL trở nên phù hợp đối với các doanh nghiệp quan tâm đến ngân sách phát triển.

Hỗ trợ nhiều nền tảng (Supported Flatforms)

MySQL hỗ trợ tất cả các nền tảng hệ điều hành chính bao gồm Oracle Linux, Solaris, Ubuntu, SUSE, Debian, Windows và macOS. Nó cũng hỗ trợ các ngôn ngữ và trình điều khiển phổ biến như Perl, Ruby, Go, Rust, C, C++, C# và ODBC.

Cơ sở dữ liệu quan hệ (Relational database)

Dữ liệu trong MySQL được sắp xếp theo mô hình quan hệ và được lưu trữ trong các bảng, mỗi hàng được liên kết với một bảng ghi liên quan. Các phương pháp tiếp cận có cấu trúc đảm bảo rằng dữ liệu này là thống nhất, do đó giúp dễ dàng truy xuất và xử lý thông tin.

Ngôn ngữ truy vấn có cấu trúc (SQL)

Ngôn ngữ chính được MySQL sử dụng để giao tiếp với CSDL là SQL. Người dùng có thể tạo, đọc, cập nhật hoặc xóa dữ liệu bằng các lệnh SQL để thực hiện công việc phát triển và quản CSDL.

Khả năng mở rộng (Scalability)

MySQL phù hợp cho các ứng dụng cơ sở dữ liệu quy mô nhỏ. Chúng ta có thể mở rộng quy mô theo chiều dọc (thêm nhiều tài nguyên hơn như CPU, Bộ nhớ, Đĩa và I/O) và định cấu hình các cụm để chia tỷ lệ theo chiều ngang cho các yêu cầu ứng dụng của bạn. Các công ty như Facebook và X (Twitter) sử dụng hệ thống cơ sở dữ liệu MySQL để xử lý khối lượng công việc của người dùng ở quy mô lớn.

Tính linh hoạt (Flexibility)

MySQL cho phép các nhà phát triển sử dụng SQL truyền thống hoặc NoSQL. Nó cũng hỗ trợ dữ liệu quan hệ và tài liệu JSON trong cùng một CSDL cũng như trong cùng một ứng dụng.

Hiệu suất (Performance)

Hiệu suất CSDL là điều cần thiết khi chọn một hệ thống quản lý cơ sở dữ liệu. MySQL bao gồm tối ưu hóa lưu trữ dữ liệu, tăng khả năng mở rộng, truy xuất dữ liệu nâng cao... Nó cung cấp nhiều cơ chế khác nhau cho các chuyên gia CSDL để đảm bảo hiệu suất được tối ưu hóa nhằm tối đa hóa hiệu suất hệ thống.

Cộng đồng và hỗ trợ (Community and support)

MySQL có cộng đồng người dùng và nhà phát triển cơ sở dữ liệu hoạt động tích cực lớn. Chúng ta có thể sử dụng diễn đàn MySQL để thảo luận về các chủ đề và vấn đề nhằm giúp chúng ta khắc phục sự cố.

3.5. Hệ quản trị cơ sở dữ liệu MongoDB

3.5.1. Giới thiệu về MonoDB

MongoDB là một CSDL hướng tài liệu mã nguồn mở được thiết kế để lưu trữ dữ liệu quy mô lớn và cho phép chúng ta làm việc với dữ liệu đó hiệu quả. Nó được phân

loại theo CDSL NoSQL vì việc lưu trữ và truy xuất dữ liệu trong MongoDB không ở dạng bảng. CSDL MongoDB được phát triển và quản lý bởi MongoDB. Inc. theo SSPL (Giấy phép công cộng phía máy chủ - Server Side Public License) và ban đầu được phát hành vào tháng 2 năm 2009.

Nó cũng cung cấp hỗ trợ trình điều khiển chính thức cho tất cả các ngôn ngữ phổ biến như C, C ++, C#và .NET, GO, Java, Node.js, Perl, PHP, Python, Motor, Ruby, Scala, Swift, Mongoid. Vì vậy, chúng ta có thể tạo một ứng dụng bằng cách sử dụng bất kỳ ngôn ngữ nào trong số này. Ngày nay, có rất nhiều công ty sử dụng MongoDB như Facebook, Nokia, eBay, Adobe, Google… để lưu trữ lượng dữ liệu. [9]

3.5.2. Ưu điểm

Tính linh hoạt: MongoDB là một hệ thống cơ sở dữ liệu phi quan hệ, nó cung cấp khả năng lưu trữ dữ liệu bất cứ khi nào, bất cứ nơi đâu, không cần phải tuân thủ một mô hình quan hệ cụ thể.

Khả năng mở rộng: MongoDB có khả năng mở rộng dễ dàng, nhờ tính năng sharding cho phép phân chia dữ liệu thành nhiều phần và lưu trữ trên nhiều máy chủ.

Tốc độ truy xuất nhanh: MongoDB có thể đáp ứng các yêu cầu truy vấn dữ liệu trong thời gian ngắn hơn so với các hệ thống cơ sở dữ liệu quan hệ truyền thống.

Tính khả dụng cao: MongoDB cung cấp tính năng sao lưu và phục hồi dữ liệu, giúp người dùng bảo vệ dữ liệu của mình khỏi những rủi ro.

Dễ sử dụng: MongoDB cung cấp các công cụ quản lý dữ liệu trực quan và dễ sử dụng, giúp người dùng tối ưu hóa hiệu suất và quản lý cơ sở dữ liệu một cách dễ dàng.

Dễ dàng tích hợp với Big Data Hadoop.

3.5.3. Nhược điểm

Cần sử dụng bộ nhớ cao để lưu trữ dữ liệu (data storage).

Không được phép lưu trữ hơn 16MB data trong tài liệu.

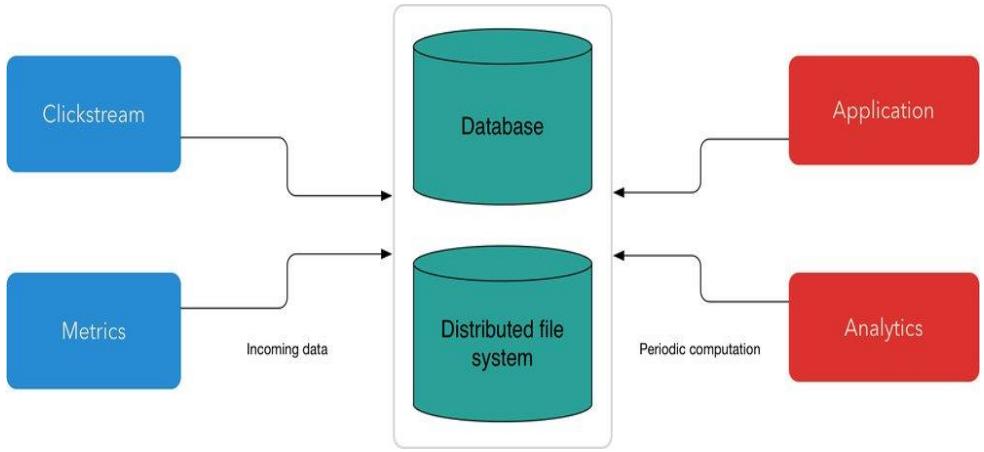
Data nesting trong BSON cũng bị hạn chế, bạn không được phép nest data quá 100 cấp độ.

3.6. Xử lý dữ liệu streaming (Streaming Data)

3.6.1. Giới thiệu về Streaming Data

Streaming data là loại dữ liệu được truyền đi theo khối lượng lớn và liên tục, giúp các tổ chức xử lý nhanh chóng và hiệu quả. Với Streaming data, các doanh nghiệp có thể theo dõi và phân tích các thông tin quan trọng ngay lập tức, giúp họ đưa ra các quyết định đúng đắn và kịp thời. [10]

3.6.2. Các thành phần của streaming data



Hình 3.2. Thành phần của streaming data

Dữ liệu nguồn: Là nguồn dữ liệu từ các thiết bị, cảm biến hoặc các nguồn dữ liệu khác mà chúng ta muốn theo dõi và phân tích. Ví dụ: Cảm biến IoT, logs hệ thống, giao dịch ngân hàng…

Quy trình xử lý: Là giai đoạn mà dữ liệu được phân tích và xử lý ngay lập tức. Các công nghệ thường được sử dụng bao gồm Apache Kafka, Apache Flink và Spark Streaming.

Dữ liệu xuất: Sau khi xử lý, dữ liệu sẽ được đưa ra dưới dạng báo cáo, dashboard, hoặc được lưu trữ trong cơ sở dữ liệu để phục vụ cho các mục đích phân tích sau này.

3.6.3. Ứng dụng thực tế của Streaming Data

Phát hiện gian lận trong giao dịch tài chính:

- Hệ thống ngân hàng sử dụng Streaming Data để phát hiện các giao dịch bất thường theo thời gian thực, giúp ngăn chặn gian lận tài chính.

- Ví dụ: Các tổ chức tài chính như Visa, Mastercard sử dụng Apache Kafka và Spark Streaming để phân tích dữ liệu giao dịch ngay khi phát sinh.

Giám sát và cảnh báo hệ thống IT:

- Dữ liệu log từ server và ứng dụng được thu thập và phân tích theo thời gian thực để phát hiện sự cố nhanh chóng.
- Ví dụ: Netflix sử dụng Apache Kafka và Elasticsearch để theo dõi hiệu suất hệ thống và xử lý lỗi ngay lập tức.

Xử lý dữ liệu trong lĩnh vực truyền thông và giải trí:

- Các nền tảng như YouTube, Spotify, TikTok sử dụng Streaming Data để đề xuất nội dung cá nhân hóa dựa trên hành vi người dùng.
- Ví dụ: Netflix phân tích dữ liệu xem phim theo thời gian thực để tối ưu hóa chất lượng video và đề xuất nội dung phù hợp.

3.7. Hệ thống Apache Kafka

3.7.1. Giới thiệu về Apache Kafka

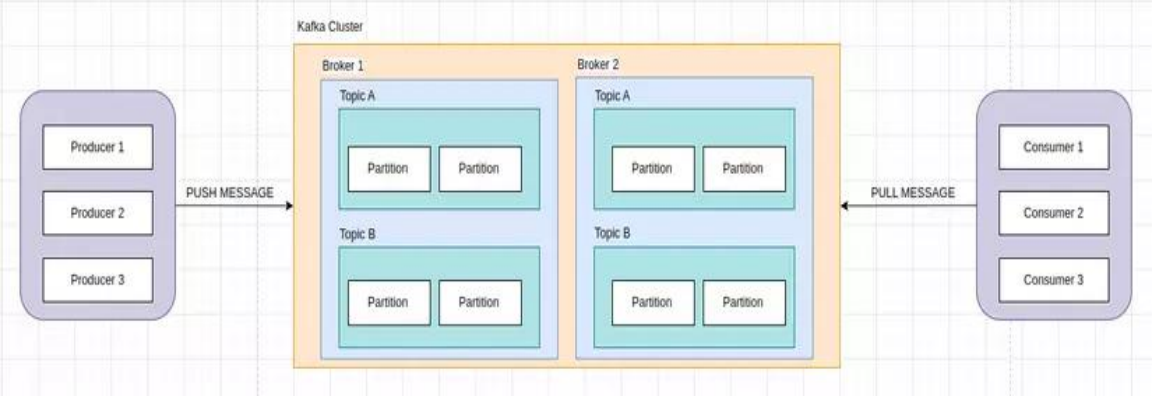
Apache Kafka là một kho dữ liệu phân tán được tối ưu hóa để thu nạp và xử lý dữ liệu truyền phát theo thời gian thực. Dữ liệu truyền phát là dữ liệu được tạo ra liên tục từ hàng nghìn nguồn dữ liệu khác nhau, các nguồn này thường gửi các bản ghi dữ liệu đồng thời. Nền tảng truyền phát cần phải xử lý luồng dữ liệu liên tục này và xử lý dữ liệu theo trình tự và tăng dần.

Kafka cung cấp ba chức năng chính cho người dùng:

- Xuất bản và đăng ký các luồng bản ghi.
- Lưu trữ hiệu quả các luồng bản ghi theo thứ tự tạo bản ghi.
- Xử lý các luồng bản ghi trong thời gian thực.

Kafka chủ yếu được dùng để xây dựng các quy trình dữ liệu truyền phát trong thời gian thực và các ứng dụng thích ứng với luồng dữ liệu đó. Kafka kết hợp nhắn tin, lưu trữ và xử lý luồng nhằm hỗ trợ hoạt động lưu trữ, phân tích cả dữ liệu lịch sử lẫn dữ liệu trong thời gian thực. [11]

3.7.2. Kiến trúc của Apache Kafka



Hình 3.3. Kiến trúc của Apache Kafka

Cluster: Một tập hợp các máy chủ Broker bao gồm ít nhất 1 Broker nhưng thường là nhiều Broker hoạt động cùng nhau. Cluster Kafka có vai trò quan trọng trong việc cung cấp tính mở rộng, tính nhất quán và độ tin cậy cho việc xử lý dữ liệu thời gian thực.

Broker: Là thành phần cốt lõi của Apache Kafka, đại diện cho máy chủ xử lý và lưu trữ dữ liệu Kafka. Một cụm Kafka thường bao gồm nhiều Broker và mỗi Broker có thể xử lý Producer (sản xuất) và Consumer (tiêu thụ) dữ liệu. Broker chịu trách nhiệm quản lý và lưu trữ các Partition của các chủ đề Topic.

Topic: Dữ liệu trong Apache Kafka được phân loại thành các Topic (chủ đề). Mỗi Topic là một luồng dữ liệu độc lập và có thể được coi là một danh sách các tin nhắn liên quan. Producer gửi dữ liệu tới các Topic và Consumer đọc dữ liệu từ các Topic. Topic cho phép các ứng dụng chỉ quan tâm đến các loại dữ liệu cụ thể mà nó muốn tiêu thụ.

Partition: Mỗi Topic có thể được chia thành nhiều phân vùng. Partition là một phần nhỏ của chủ đề và đóng vai trò quan trọng trong việc phân tán dữ liệu và tăng hiệu suất. Mỗi Partition được lưu trữ trên một Broker và dữ liệu được đọc và ghi vào từng Partition một.

Producer: Producer là thành phần của Apache Kafka cho phép ứng dụng gửi dữ liệu tới các Topic. Producer sử dụng giao thức Kafka để kết nối với Broker và đưa dữ liệu vào các Partition của các chủ đề.

Consumer: Là thành phần cho phép ứng dụng lấy dữ liệu từ các Topic Kafka. Consumer cũng sử dụng giao thức Kafka để kết nối với Broker và đọc dữ liệu từ các Partition. Có thể có nhiều Consumer đọc từ cùng 1 Topic và Kafka đảm bảo rằng mỗi tin nhắn chỉ được đọc bởi 1 Consumer.

ZooKeeper: Là một hệ thống quản lý tập trung được sử dụng để quản lý và duy trì trạng thái của các Broker trong một cụm Kafka. Nó chịu trách nhiệm trong việc theo dõi và quản lý các Broker giúp Kafka hoạt động ổn định và đảm bảo tính nhất quán. [12]

3.8. Change Data Capture

3.8.1. Giới thiệu về Change Data Capture

Change Data Capture (CDC) là một kỹ thuật cũng như một kiểu kiến trúc để thiết kế hệ thống mà qua đó ta có thể theo dõi được những thay đổi phát sinh ở phía source database.

Với việc sử dụng CDC, ta có thể dễ dàng sao chép, nhân bản (replicate) hoặc chuyển đổi (migrate) dữ liệu giữa nhiều database khác nhau trong thời gian thực (real-time).

Nhờ CDC, hệ thống sẽ có thể từ từ load được những thay đổi ở phía source database (incremental load) thay vì phải load một lượng lớn dữ liệu theo từng đợt (bulk load). [13]

3.8.2. Tại sao Change Data Capture lại rất quan trọng đối với doanh nghiệp

Ngày nay, dữ liệu là trung tâm trong cách các doanh nghiệp hiện đại vận hành và là yếu tố chính thúc đẩy chuyển đổi số và ra quyết định kinh doanh. Các kiến trúc dữ liệu hiện đại đang ngày càng gia tăng. Các công ty đang chuyển dữ liệu của họ từ cơ sở hạ tầng tại chỗ lên đám mây bao gồm kho dữ liệu đám mây và hồ dữ liệu. Doanh nghiệp đang chuyển từ quản lý dữ liệu theo lô (Batch) sang quản lý dữ liệu theo thời gian thực (streaming). Nhưng họ vẫn gặp khó khăn trong việc theo kịp với khối lượng, sự đa dạng và tốc độ dữ liệu đang gia tăng. Các kiến trúc đám mây mới đang giải quyết những thách thức này. Chúng bao gồm kho dữ liệu đám mây, hồ dữ liệu đám mây và phát trực tuyến dữ liệu.

Nhưng tuổi thọ của dữ liệu đang giảm dần. Khi dữ liệu nhạy cảm với thời gian, giá trị của nó đối với doanh nghiệp sẽ nhanh chóng hết hạn. Những hiểu biết về dữ liệu theo thời gian thực là thước đo mới cho sự thành công trong kỷ nguyên số. Khi một công ty không thể hành động ngay lập tức, họ sẽ bỏ lỡ cơ hội kinh doanh. Những hiểu biết về dữ liệu cung cấp giá trị lớn hơn gấp bội so với phân tích truyền thống, nhưng giá trị đó sẽ nhanh chóng lỗi thời và không mang lại giá trị cao. [14]

3.8.3. Ứng dụng của CDC

Đồng bộ hóa/nhân bản cơ sở dữ liệu truyền thống

Thường thì việc quản lý thay đổi dữ liệu liên quan đến việc nhân bản dữ liệu theo lô (Batch). Với nhu cầu ngày càng tăng về việc ghi lại và phân tích dữ liệu theo luồng thời gian thực (Real-time), các công ty không thể ngừng hoạt động và sao chép toàn bộ cơ sở dữ liệu để quản lý thay đổi dữ liệu. CDC cho phép nhân bản liên tục trên các tập dữ liệu nhỏ hơn. Nó cũng chỉ giải quyết các thay đổi gia tăng.

Phân tích dữ liệu theo luồng thời gian thực và tiếp nhận dữ liệu Cloud Data Lake

Khía cạnh khó khăn nhất của việc quản lý hồ dữ liệu đám mây là giữ cho dữ liệu luôn cập nhật. Với kiến trúc dữ liệu hiện đại, các công ty có thể liên tục tiếp nhận dữ liệu CDC vào Data Lake thông qua Data Pipeline tự động. Điều này tránh việc di chuyển hàng terabyte dữ liệu một cách không cần thiết qua mạng. Chúng ta có thể tập trung vào sự thay đổi trong dữ liệu, tiết kiệm chi phí tính toán và mạng. Với sự hỗ trợ cho các công nghệ như Apache Spark để xử lý theo thời gian thực, CDC là công nghệ nền tảng để thúc đẩy phân tích thời gian thực tiên tiến. Chúng ta cũng có thể hỗ trợ các trường hợp sử dụng AI và ML.

Phát hiện gian lận trong tài chính

Đối với các tổ chức dựa trên dữ liệu, trải nghiệm khách hàng rất quan trọng để giữ chân và phát triển cơ sở khách hàng của họ. Một ví dụ điển hình là trong lĩnh vực tài chính. Nếu một ngân hàng lớn gặp phải sự gia tăng đột ngột về các hoạt động gian lận, họ cần phân tích theo thời gian thực để chủ động cảnh báo khách hàng về các gian lận tiềm ẩn. Dữ liệu giao dịch cần được thu thập từ cơ sở dữ liệu theo thời gian thực.

Sau đó, nó có thể chuyển đổi và làm phong phú dữ liệu để công cụ giám sát gian lận có thể chủ động gửi tin nhắn và email cảnh báo cho khách hàng. Khi đó, khách hàng có thể thực hiện hành động khắc phục ngay lập tức. [14]

3.9. Công cụ Debezium

3.9.1. Giới thiệu về Debezium

Debezium là một tập hợp các dịch vụ phân tán để nắm bắt các thay đổi trong cơ sở dữ liệu để các ứng dụng có thể thấy những thay đổi đó và phản hồi với chúng. Debezium ghi lại tất cả các thay đổi ở cấp hàng trong mỗi bảng cơ sở dữ liệu trong luồng sự kiện thay đổi và các ứng dụng chỉ cần đọc các luồng này để xem các sự kiện thay đổi theo cùng thứ tự mà chúng xảy ra. [15]

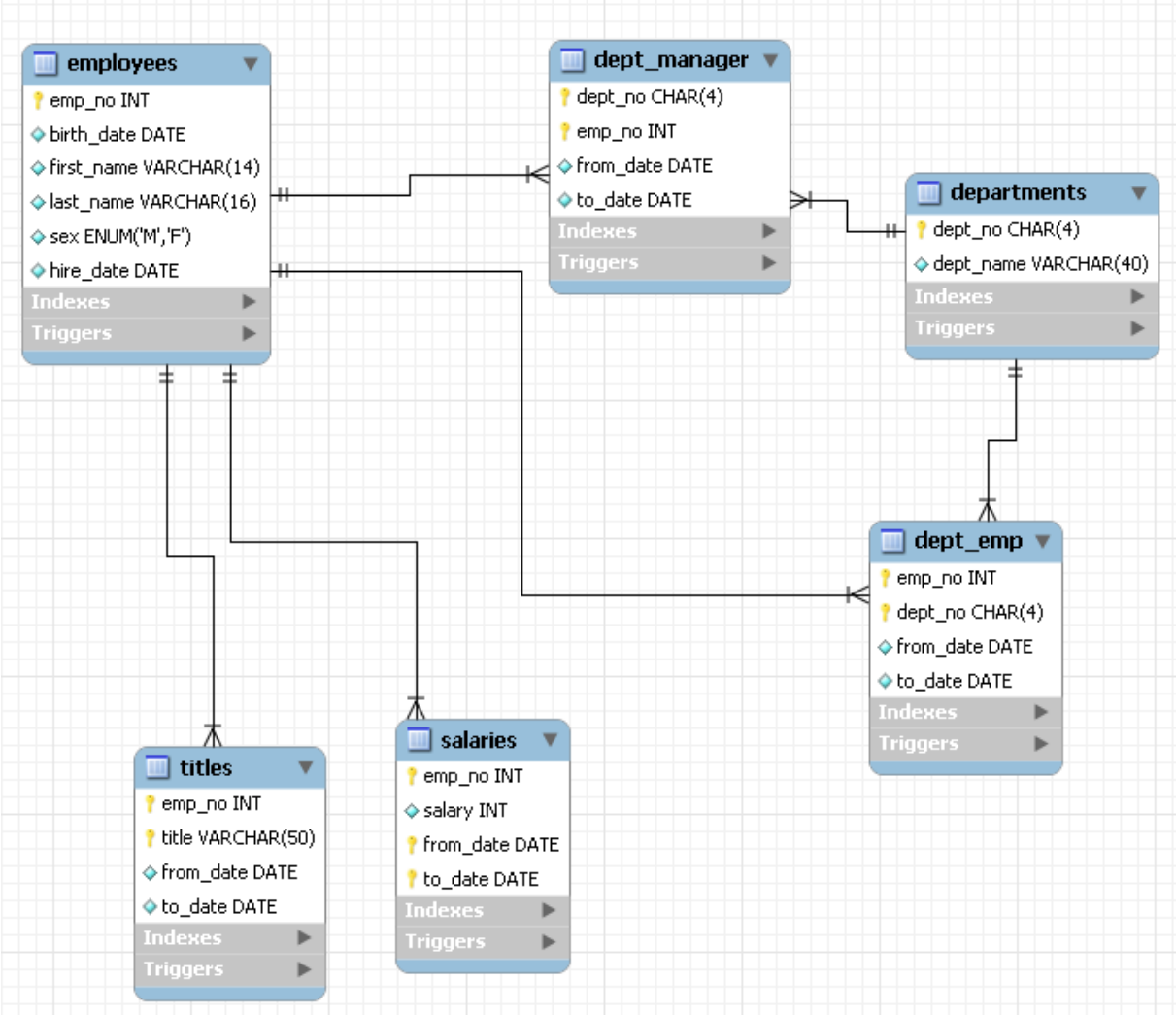
3.9.2. Ứng dụng của Debezium

Vì có khả năng phát và nhận dữ liệu realtime nên debezium được sử dụng phát triển trong rất nhiều dự án khác nhau có thể kể tới như:

- CDC, Datalake, ETL: các ứng dụng nhận và xử lý dữ liệu có thể sử dung debezium để nhận và phân tích dữ liệu.
- Xem số liệu thời gian thực (có vụ tôi muốn xem các đơn hàng 1 cách realtime, hoặc tôi muốn xử lý hậu đơn hàng một cách realtime)
- Phát hiện giao dịch gian lận: debezium là 1 cách đơn giản để nhận và kiểm tra các giao dịch liên tục. [16]

CHƯƠNG 4: THIẾT KẾ CƠ SỞ DỮ LIỆU

4.1. Tổng quan



Hình 4.1. Lược đồ quan hệ

Trong dự án này, cơ sở dữ liệu được sử dụng là một bộ dữ liệu mẫu "**Employees Sample Database**" [17] được cung cấp bởi MySQL. Bộ dữ liệu mẫu này nhằm hỗ trợ việc nghiên cứu và triển khai giải pháp chuyển đổi dữ liệu từ MySQL qua MongoDB. Bộ dữ liệu bao gồm các bảng trong MySQL và được sử dụng để mô phỏng môi trường làm việc thực tế, phục vụ cho việc thực nghiệm và kiểm chứng các phương pháp đồng bộ dữ liệu.

4.2. Đặc tả các bảng dữ liệu

4.2.1. Bảng EMPLOYEES

STT	Tên trường	Kiểu dữ liệu	Khoá	Diễn giải
1	Emp_No	Integer	Chính	Mã nhân viên
2	Birth_Date	Date		Ngày sinh
3	First_Name	Varchar(14)		Tên nhân viên
4	Last_Name	Varchar(16)		Họ nhân viên
5	Gender	Enum		Giới tính
6	Hire_Date	Date		Ngày tuyển dụng

Bảng 4.1. Đặc tả bảng EMPLOYEES

4.2.2. Bảng DEPARTMENTS

STT	Tên trường	Kiểu dữ liệu	Khoá	Diễn giải
1	Dept_No	Char(4)	Chính	Mã phòng ban
2	Dept_Name	Varchar(40)		Tên phòng ban

Bảng 4.2. Đặc tả bảng DEPARTMENTS

4.2.3. Bảng DEPT_EMP

STT	Tên trường	Kiểu dữ liệu	Khoá	Diễn giải
1	Emp_No	Integer	Ngoại	Mã nhân viên
2	Dept_No	Char(10)	Ngoại	Mã phòng ban
3	From_Date	Date		Ngày bắt đầu
4	To_Date	Date		Ngày kết thúc

Bảng 4.3. Đặc tả bảng DEPT_EMP

4.2.4. Bảng TITLES

STT	Tên trường	Kiểu dữ liệu	Khoá	Diễn giải
1	Emp_No	Integer	Ngoại	Mã nhân viên
2	Title	Varchar(50)	Ngoại	Chức danh
3	From_Date	Date		Ngày bắt đầu
4	To_Date	Date		Ngày kết thúc

Bảng 4.4. Đặc tả bảng TITLES

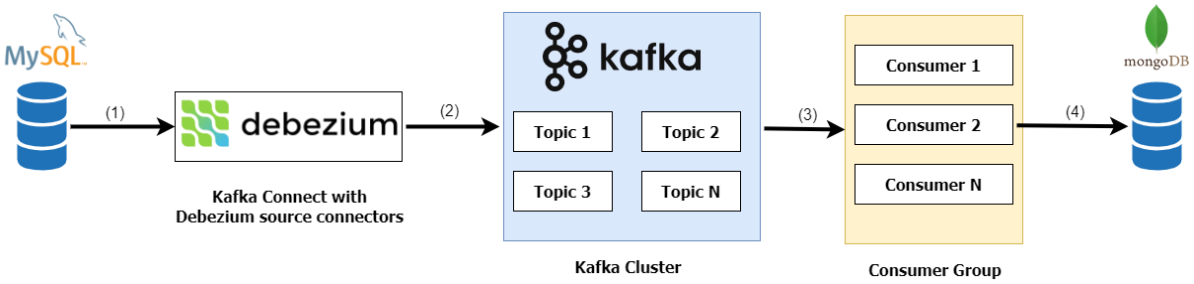
4.2.5. Bảng SALARIES

STT	Tên trường	Kiểu dữ liệu	Khoá	Diễn giải
1	Emp_No	Integer	Ngoại	Mã nhân viên
2	Salary	Integer		Mức lương
3	From_Date	Date		Ngày bắt đầu
4	To_Date	Date		Ngày kết thúc

Bảng 4.5. Đặc tả bảng SALARIES

CHƯƠNG 5: XÂY DỰNG HỆ THỐNG

5.1. Tổng quan kiến trúc hệ thống



Hình 5.1. Sơ đồ CDC

(1) Lắng nghe thay đổi trong MySQL (Debezium Source Connector)

- **Debezium** được kết nối với **MySQL** thông qua **binlog** để theo dõi các thay đổi dữ liệu (INSERT, UPDATE, DELETE).
- Khi có thay đổi trong MySQL, Debezium trích xuất dữ liệu và đóng gói thành sự kiện CDC.
- Sự kiện CDC này sẽ được gửi vào một Kafka.

(2) Gửi sự kiện thay đổi vào Kafka

- **Kafka** đóng vai trò làm message broker, giúp đảm bảo dữ liệu từ MySQL có thể được gửi đi một cách ổn định và theo thứ tự.
- Debezium Source Connector gửi sự kiện CDC từ MySQL vào các Kafka topic tương ứng.
- Từng bảng trong MySQL có một Kafka topic riêng biệt.

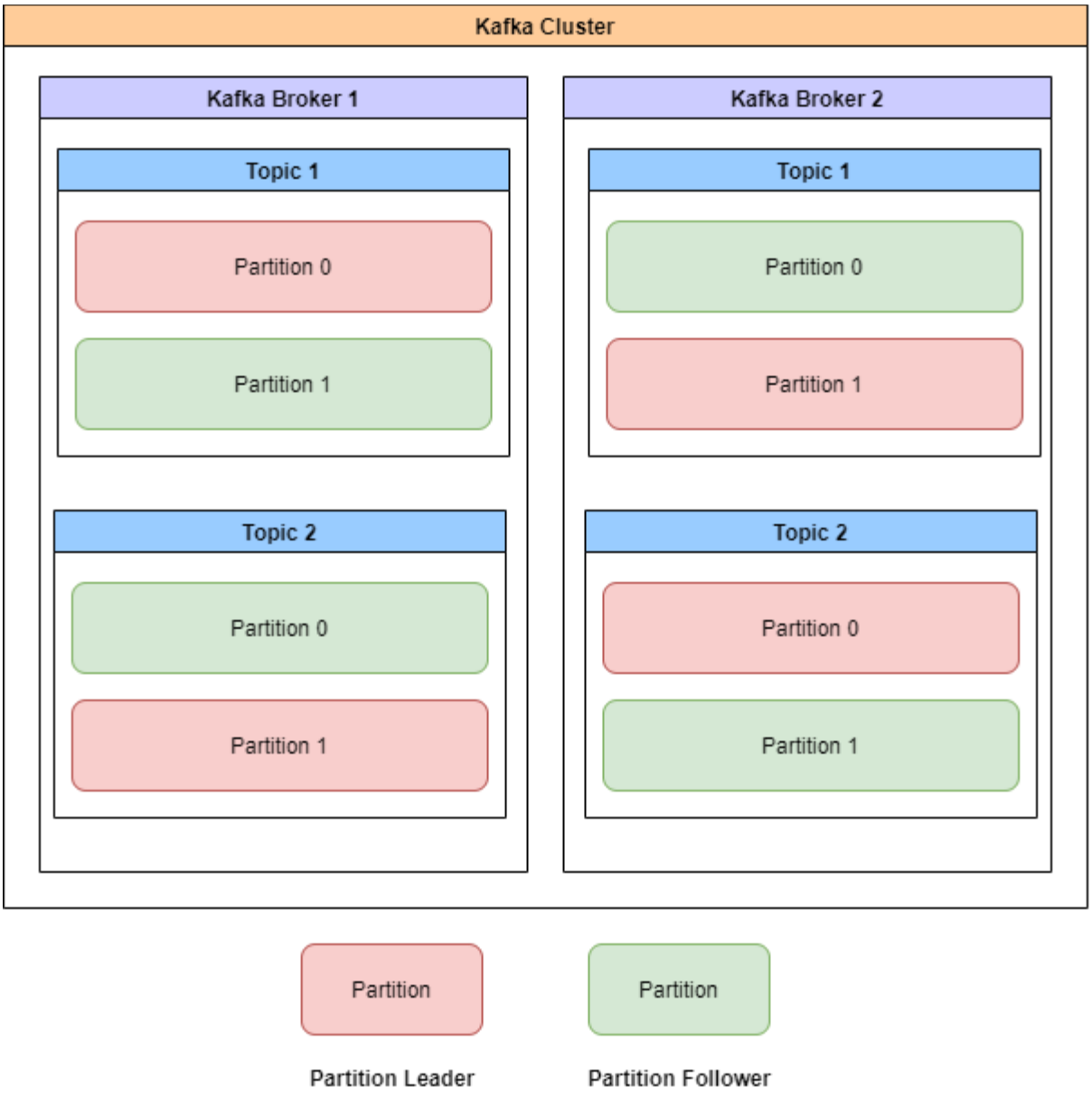
(3) Tiêu thụ sự kiện và ghi vào MongoDB (Kafka Consumer)

- Một ứng dụng Kafka Consumer (Django hoặc một service khác) sẽ đọc dữ liệu từ Kafka topic.
- Consumer này sẽ xử lý và chuyển đổi dữ liệu để phù hợp với MongoDB.

(4) Ghi dữ liệu vào MongoDB

- Sau khi Consumer xử lý dữ liệu, nó sẽ ghi dữ liệu vào **MongoDB**, giúp MongoDB luôn được cập nhật theo MySQL.

5.2. Kiến trúc cụm Kafka của hệ thống



Hình 5.2. Kiến trúc minh hoạ cụm Kafka

Tổng quan về Kafka Cluster

- Sơ đồ mô tả một Kafka Cluster gồm 2 broker:

  - Kafka Broker 1 (9092)
  - Kafka Broker 2 (9093)

- Zookeeper không được hiển thị, nhưng nó đóng vai trò quản lý metadata và leader election (bầu cử).

Các thành phần chính trong sơ đồ

- Các Topic: Topic 1, Topic 2 (Thực tế sẽ đặt tên khác).

- Các Partition: Mỗi topic có 2 partition (Partition 0 và Partition 1).

- Partition Leader (ô màu đỏ): Là nơi producer ghi dữ liệu.

- Partition Follower (Replica) (ô màu xanh): Sao chép dữ liệu từ leader.

Cách thức hoạt động

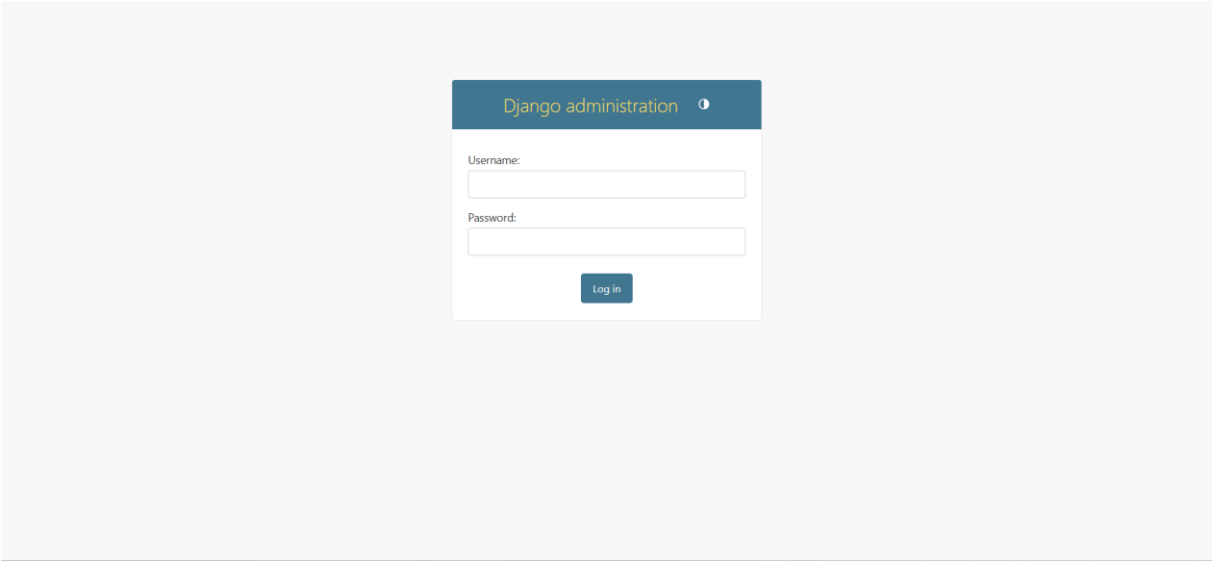
1. Producer gửi dữ liệu đến **Partition Leader** của topic tương ứng.

2. Broker chứa **Partition Leader** sẽ ghi dữ liệu và sao chép dữ liệu cho các **Follower** (Replica).

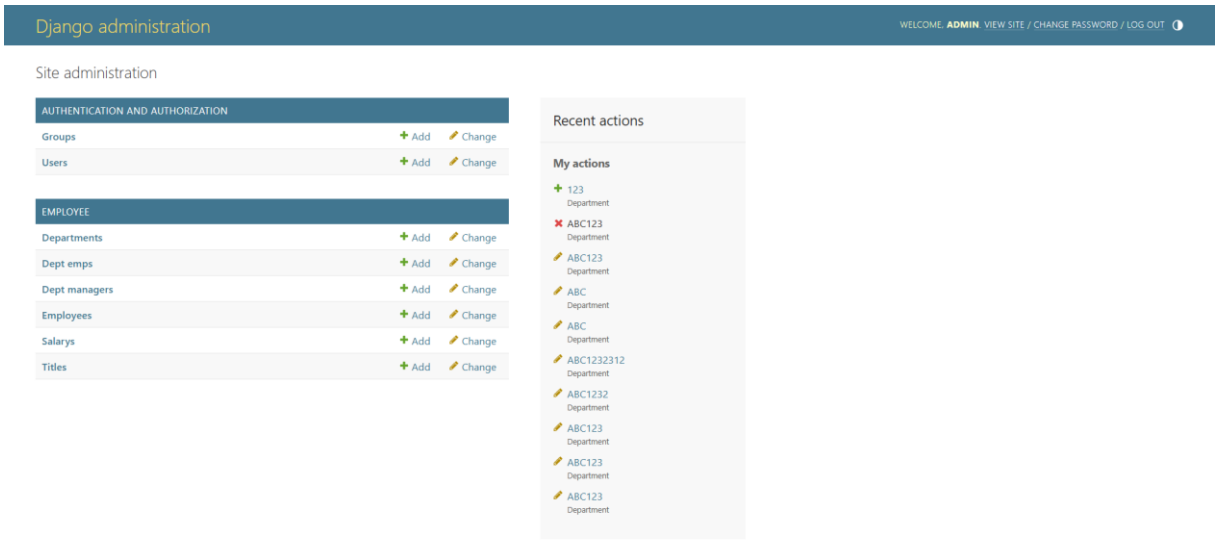
3. Consumer có thể đọc từ **bất kỳ Partition** nào nhưng thường sẽ ưu tiên đọc từ **Leader**.

4. Nếu một **Partition Leader** bị lỗi, Kafka sẽ tự động bầu chọn một **Replica** lên làm Leader để duy trì hoạt động.

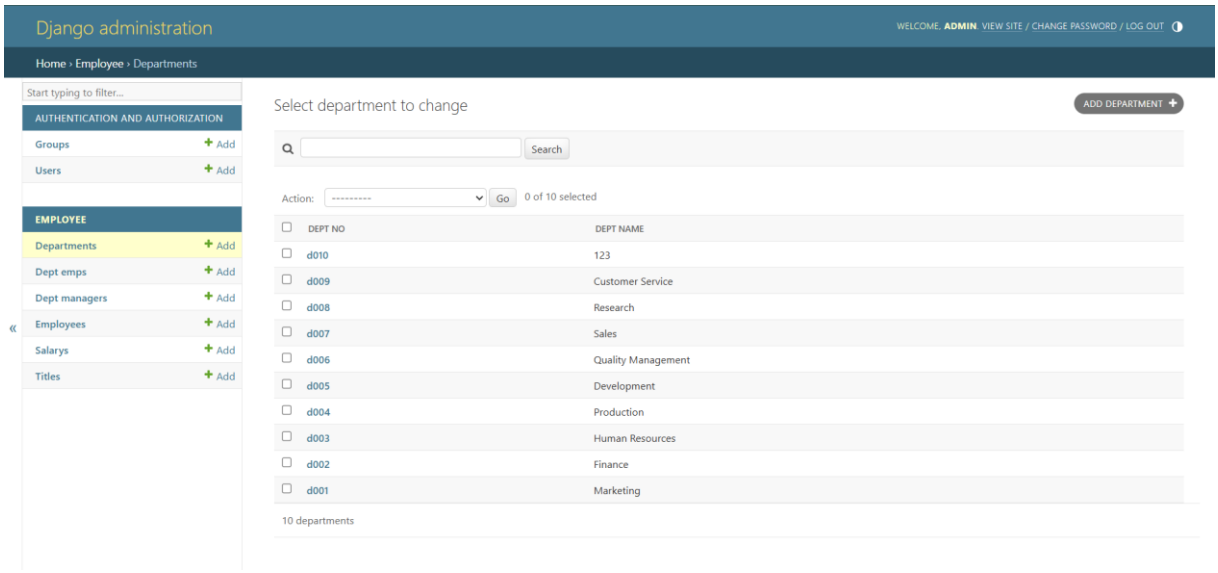
5.3. Giao diện hệ thống



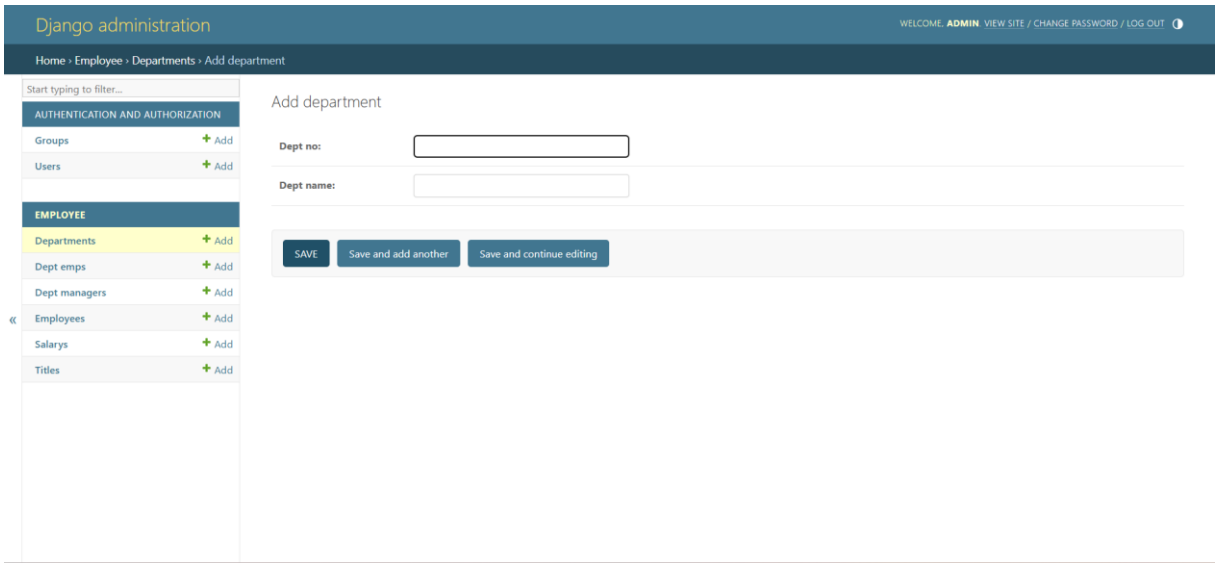
Hình 5.3. Trang đăng nhập admin



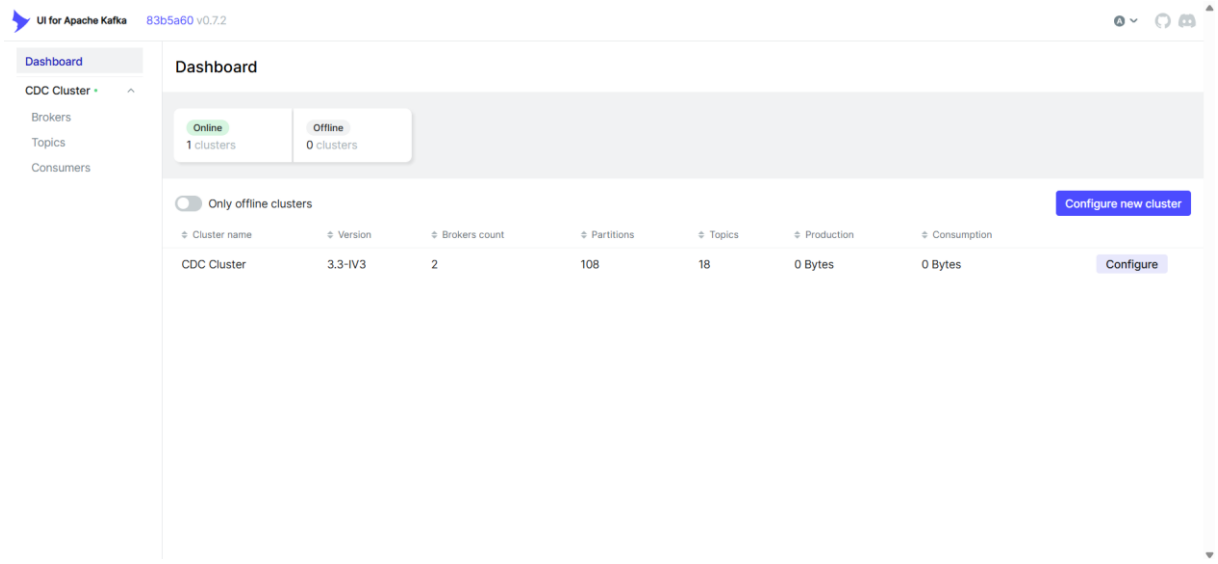
Hình 5.4. Màn hình chính admin



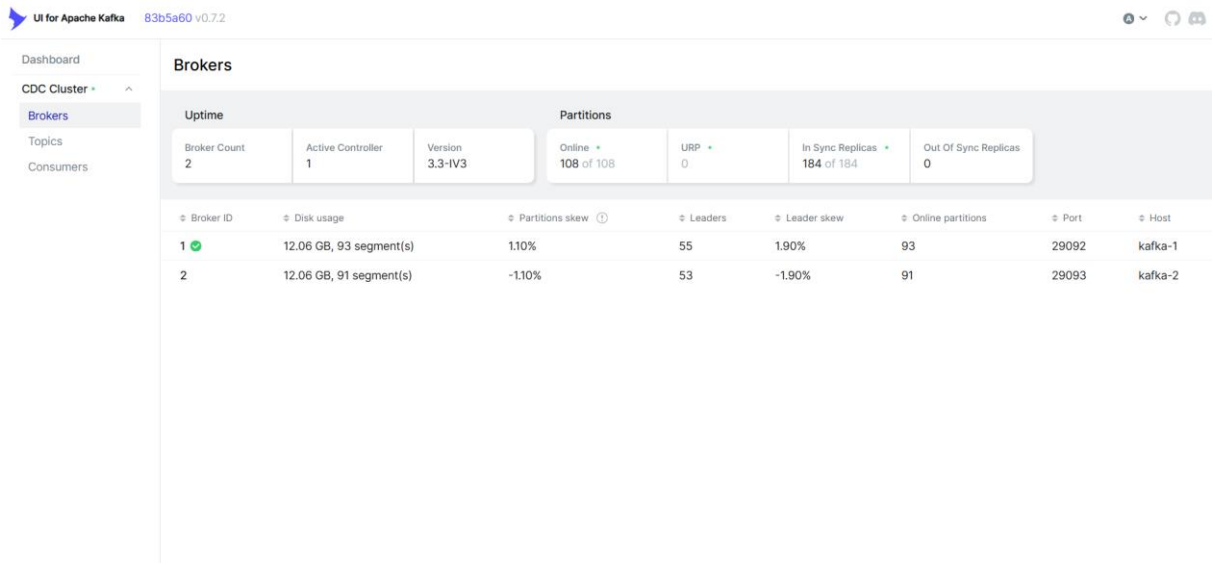
Hình 5.5. Trang quản lý Department



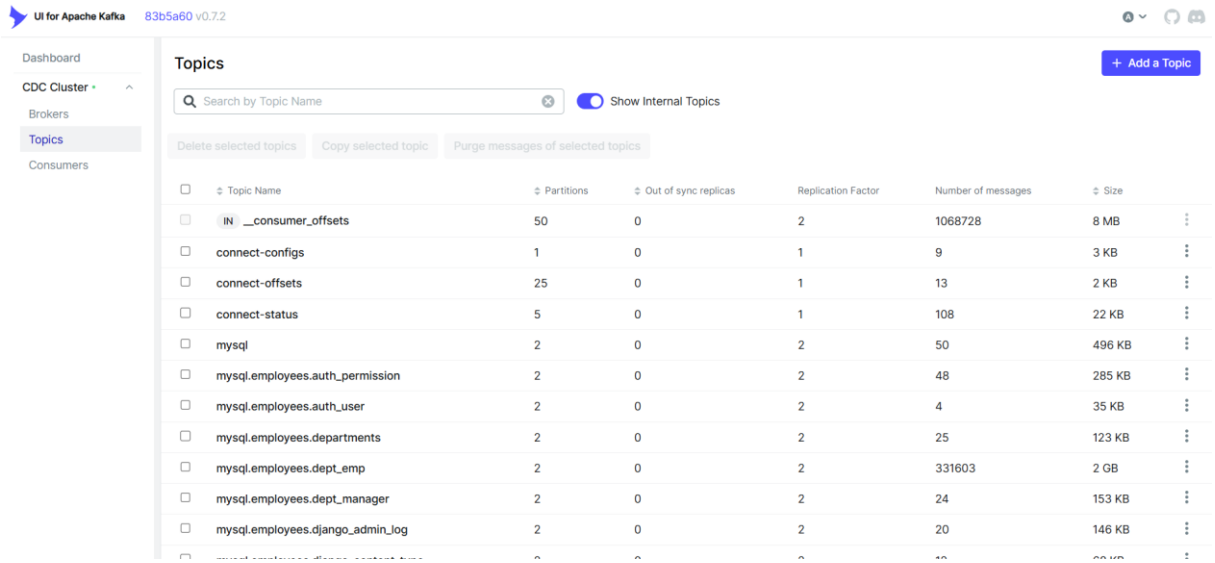
Hình 5.6. Trang thêm Department



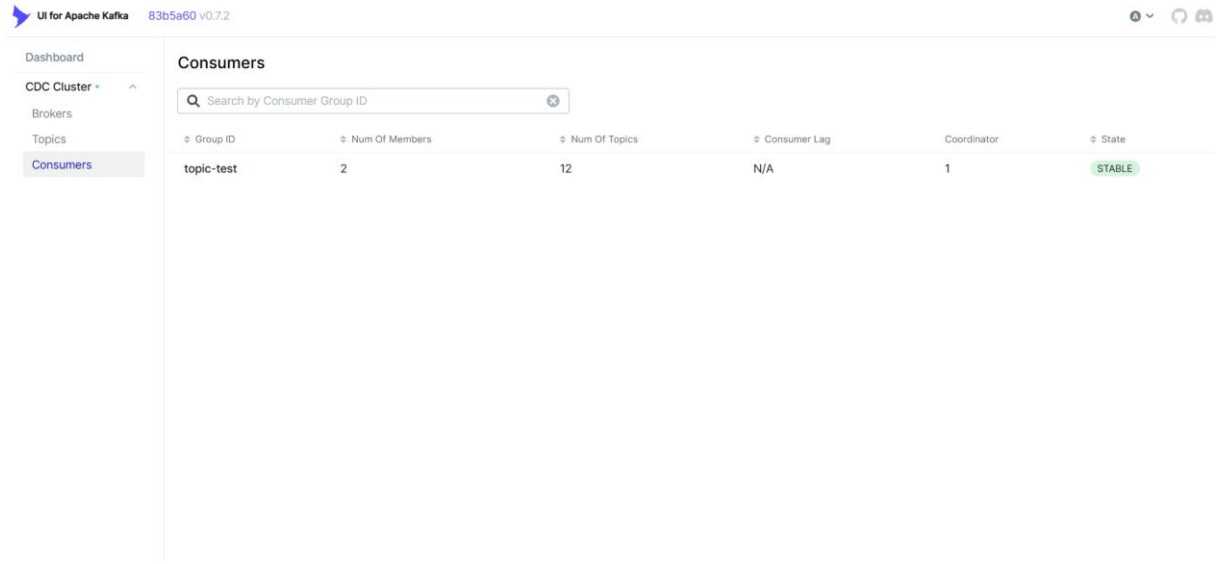
Hình 5.7. Trang chủ quản lý cụm Kafka



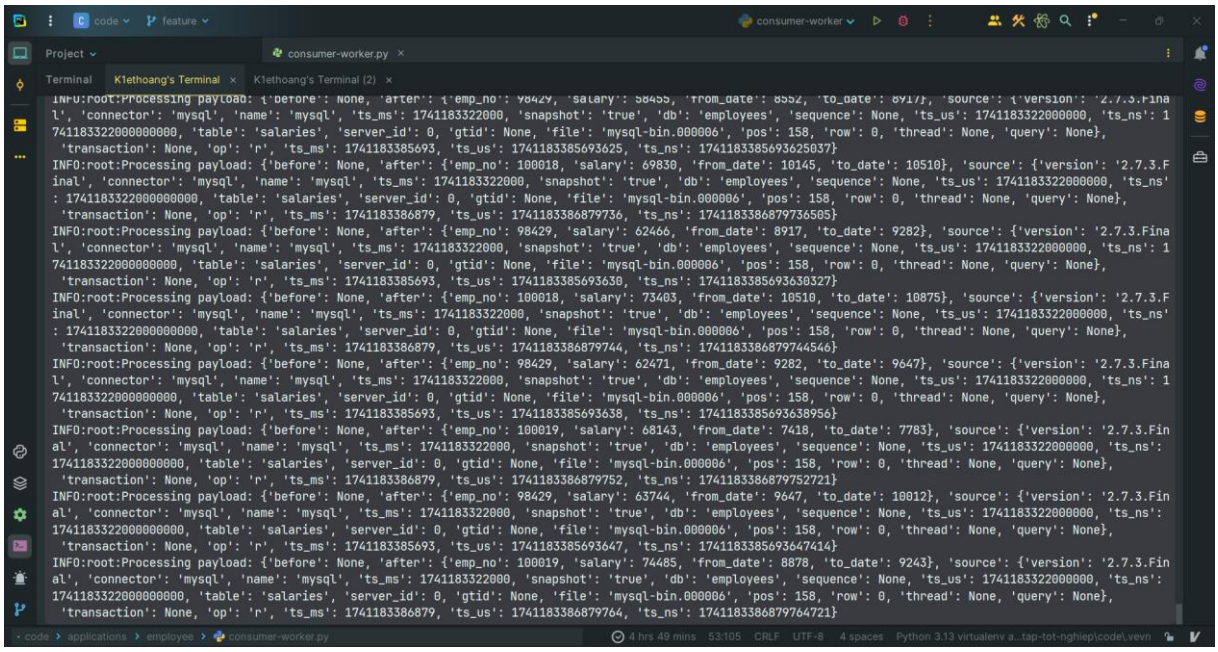
Hình 5.8. Trang Broker



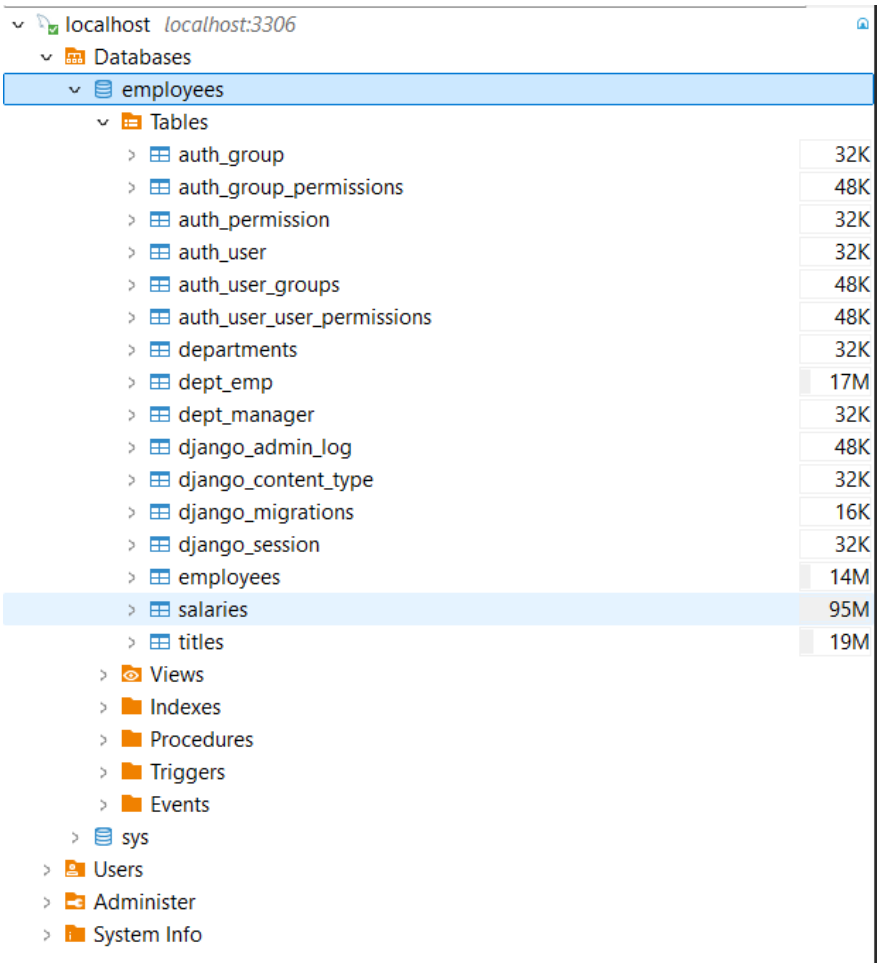
Hình 5.9. Trang Topic



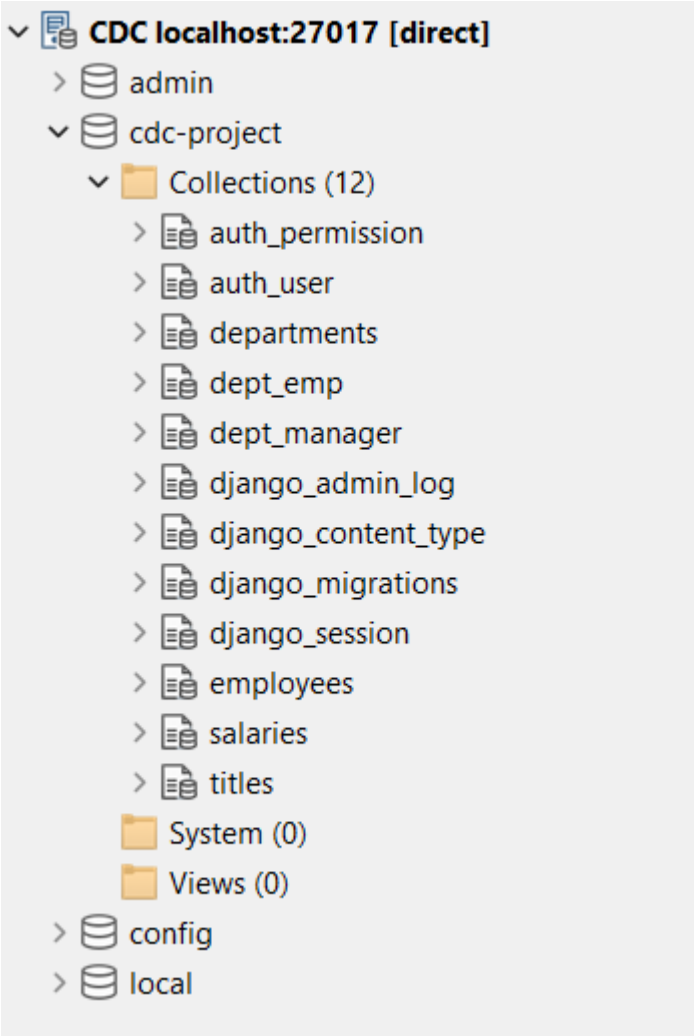
Hình 5.10. Trang Consumer



Hình 5.11. Log chuyển đổi dữ liệu



Hình 5.12. Cơ sở dữ liệu MySQL



Hình 5.13. Cơ sở dữ liệu MongoDB

KẾT LUẬN

1. Kết quả đạt được

Nghiên cứu và triển khai thành công hệ thống Change Data Capture (CDC) để đồng bộ dữ liệu giữa MySQL và MongoDB theo thời gian thực, sử dụng Debezium và Apache Kafka. Hệ thống đạt được các kết quả sau:

- Theo dõi và ghi nhận thay đổi từ MySQL, sử dụng Debezium để trích xuất dữ liệu và Kafka để truyền tải.
- Sử dụng Kafka với 2 partition, giúp phân tán tải và tăng hiệu suất xử lý dữ liệu.
- Tiêu thụ dữ liệu từ Kafka topic và cập nhật MongoDB theo thời gian thực, đảm bảo dữ liệu luôn đồng bộ giữa các hệ thống.
- Áp dụng Docker để triển khai Kafka, Debezium và MongoDB, giúp hệ thống dễ dàng mở rộng và quản lý.
- Tích hợp CDC vào ứng dụng Django, giúp backend có thể xử lý dữ liệu hiệu quả hơn.

Bên cạnh đó, quá trình thực hiện đề tài giúp em nâng cao kỹ năng làm việc với Kafka, Django, Docker cũng như hiểu rõ hơn về hệ thống xử lý dữ liệu thời gian thực.

2. Hạn chế

Khả năng mở rộng của Kafka: Hiện tại, topic chỉ có 2 partition, có thể gây hạn chế về throughput khi lượng dữ liệu tăng cao.

Quản lý consumer: Hiện tại, hệ thống có một consumer group, cần tối ưu thêm để cân bằng tải giữa các partition.

Chưa có cơ chế retry khi xảy ra lỗi: Nếu một message thất bại khi ghi vào MongoDB, hiện chưa có cơ chế xử lý tự động.

Chưa thử nghiệm với dữ liệu lớn: Hệ thống chưa được kiểm tra trong môi trường production với lượng dữ liệu lớn và tải cao.S

3. Hướng phát triển

Tăng số lượng partition trong Kafka, giúp cải thiện khả năng mở rộng và tăng tốc độ xử lý.

Triển khai thêm consumer trong cùng consumer group, để tận dụng tối đa khả năng song song của Kafka.

Tích hợp cơ chế retry và error handling khi ghi dữ liệu vào MongoDB, tránh mất dữ liệu khi xảy ra lỗi.

Thử nghiệm với môi trường production, kiểm tra hiệu suất và tối ưu hệ thống khi xử lý dữ liệu lớn.

Nâng cấp hệ thống với Apache Flink hoặc Spark Streaming, giúp xử lý dữ liệu mạnh mẽ hơn trong thời gian thực.

TÀI LIỆU THAM KHẢO

[1] “fpt telecom,” [Trực tuyến]. Available: https://fpt.vn/vi/ve-fpt-telecom/tap-doan-fpt.html. [Đã truy cập 21 02 2025].

[2] “fpt telecom,” [Trực tuyến]. Available: https://fpt.vn/vi/ve-fpt-telecom/gioi-thieu-chung.html. [Đã truy cập 21 02 2025].

[3] “wikipedia,” [Trực tuyến]. Available: https://en.wikipedia.org/wiki/Python_(programming_language). [Đã truy cập 25 02 2025].

[4] “geeksforgeeks,” [Trực tuyến]. Available: https://www.geeksforgeeks.org/introduction-to-python/. [Đã truy cập 25 02 2025].

[5] “djangoproject,” [Trực tuyến]. Available: https://www.djangoproject.com/. [Đã truy cập 25 02 2025].

[6] kanithkar_baskaran, “medium,” [Trực tuyến]. Available: https://medium.com/@kanithkar_baskaran/essential-features-of-django-that-make-it-a-developers-favorite-6c916692c890. [Đã truy cập 25 02 2025].

[7] leanhnam. [Trực tuyến]. Available: https://lethanhnamwork.com/tim-hieu-chi-tiet-django-rest-framework-la-gi/#:~:text=Django%20REST%20Framework%20(DRF)%20l%C3%A0,ph%C3%A1t%20tri%E1%BB%83n%20web%20b%E1%BA%B1ng%20Python..

[8] “Quest,” [Trực tuyến]. Available: https://www.quest.com/learn/what-is-my-sql.aspx. [Đã truy cập 25 02 2025].

[9] “geeksforgeeks,” [Trực tuyến]. Available: https://www.geeksforgeeks.org/what-is-mongodb-working-and-features/. [Đã truy cập 25 02 2025].

[10] [Trực tuyến]. Available: https://hazelcast.com/foundations/event-driven-architecture/streaming-data/. [Đã truy cập 25 02 2025].

[11] “amazon,” [Trực tuyến]. Available: https://aws.amazon.com/vi/what-is/apache-kafka/. [Đã truy cập 26 02 2025].

[12] “viblo,” [Trực tuyến]. Available: https://viblo.asia/p/tong-quan-ve-apache-kafka-he-thong-xu-ly-du-lieu-thoi-gian-thuc-phan-tan-5OXLA5XkLGr. [Đã truy cập 26 02 2025].

[13] T-Rex, “Atekco,” [Trực tuyến]. Available: https://atekco.io/1665116042835-tat-tan-tat-ve-change-data-capture-va-debezium-phan-1/. [Đã truy cập 26 02 2025].

[14] “fptcloud,” [Trực tuyến]. Available: https://fptcloud.com/change-data-capture-dong-bo-du-lieu-tu-dong-cho-doanh-nghiep/. [Đã truy cập 26 02 2025].

[15] “debezium,” [Trực tuyến]. Available: https://debezium.io/documentation/faq/#what_is_debezium. [Đã truy cập 26 02 2025].

[16] “viblo,” [Trực tuyến]. Available: https://viblo.asia/p/debezium-series-debezium-la-gi-ung-dung-thuc-te-5OXLAA2YLGr#_ung-dung-cua-debezium-3. [Đã truy cập 26 02 2025].

[17] “mysql,” [Trực tuyến]. Available: https://dev.mysql.com/doc/employee/en/. [Đã truy cập 28 02 2025].

PHỤ LỤC

1. Mã nguồn

github.com/K1ethoang/Uni_ThucTapTotNghiep-2025

2. Hướng dẫn cài đặt, cấu hình và sử dụng

2.1. Cơ sở dữ liệu

github.com/datacharmer/test_db

2.2. Docker và Docker Compose

- viblo.asia/p/cai-dat-docker-tren-windows-10-3Q75w6gelWb

- viblo.asia/p/cach-tao-docker-compose-3P0lPeaP5ox

2.3. Cài đặt Python

kungfutech.edu.vn/bai-viet/python/huong-dan-cai-dat-python

2.4. Chạy dự án

- Điều kiện tiên quyết:

+ Import dữ liệu vào MySQL

+ Tạo file .env

+ Cài đặt docker

+ Tạo môi trường ảo cho dự án python

+ Tải các thư viện trong file requirements.txt

- Chạy lệnh docker compose để build container cần thiết cho dự án:

+ docker compose --env-file .env -f .\docker\docker-compose.yml up -d --build

+ docker exec -it kafka-connect bash

+ Kiểm tra Kafka Connect đã cấu hình xong chưa: curl -s http://localhost:8083.

- Nếu kết quả trả về: {"version":"3.7.0","commit":"2ae524ed625438c5","kafka_cluster_id":"g

FGPx47qQcitvVKwm3ECQQ"} tiếp tục chạy lệnh sau để đăng ký connector: sh /register-connector.sh

- Migrate dự án:

python manage.py makemigrations

python manay.py migrate

- Tạo tài khoản admin:

python manage.py createsuperuser --username admin

- Chạy dự án:

python manage.py runserver (tab terminal 1)

python applications/employee/consumer-worker.py (tab terminal 2)

- Mở 2 đường link sau trong trình duyệt:

- Trang admin: http://127.0.0.1:8000/admin
- Trang kafka: http://127.0.0.1:5000