

# HỆ THỐNG THÔNG TIN ĐA PHƯƠNG TIỆN TRÊN Đám Mây

## A MULTIMEDIA INFORMATION SYSTEM IN THE CLOUD

NGUYỄN QUỐC THOẠI ANH<sup>1,a</sup>, NGUYỄN QUỐC ANH THU<sup>2,b</sup>

<sup>1</sup> Lớp Cao học CNTT K1. Khoa Công nghệ thông tin, Trường Đại học SPKT Vĩnh Long

<sup>2</sup> Lớp Cao học CNTT K1. Khoa Công nghệ thông tin, Trường Đại học SPKT Vĩnh Long

Email: <sup>a</sup>thoaianh.thvl@gmail.com, <sup>b</sup>athu.c3pq@gmail.com

*Nhận bài (Received): 25/3/2022; Phản biện (Reviewed): 26/4/2022; Chấp nhận (Accepted): 15/7/2022*

### TÓM TẮT:

Việt Nam là một trong những quốc gia đầu tiên trên thế giới ban hành chương trình chuyển đổi số quốc gia, đưa Việt Nam trở thành quốc gia có nhận thức về chuyển đổi số song hành cùng các quốc gia tiên tiến trên thế giới. Đặc biệt, trong năm 2020-2021 dù ảnh hưởng đại dịch COVID-19 nhưng quá trình chuyển đổi số quốc gia diễn ra mạnh mẽ và quyết liệt trên tất cả các ngành, các lĩnh vực, dễ thấy nhất là Chính phủ đã xây dựng hệ thống cơ sở dữ liệu quốc gia về dân cư và hệ thống sản xuất, cấp và quản lý căn cước công dân. Không nằm ngoài xu thế đó, Đài Phát thanh và Truyền hình tỉnh Vĩnh Long từng bước số hóa các nội dung chương trình. Trước đây quản lý nội dung bằng băng từ, hiện nay việc quản lý đã được số hóa thành các tập tin. Vì vậy, việc quản lý khối lượng dữ liệu khổng lồ là việc làm cấp thiết hiện nay. Để quản lý tốt dữ liệu hiện có và cập nhật dữ liệu mới hàng ngày, cần xây dựng hệ thống thông tin đa phương tiện trên đám mây.

**Từ khóa:** Điện toán đám mây, dữ liệu lớn, CouchDB, Kafka

### ABSTRACT:

Vietnam is one of the first countries in the world to issue a national digital transformation program, making Vietnam become a country being aware of digital transformation in parallel with advanced countries in the world. In particular, in the years 2020 and 2021, despite the impact of the COVID-19 epidemic, the national digital transformation took place vigorously and drastically in all sectors and fields. Most notably, the Government has built the national database system on the population and the system of producing, issuing and managing citizens' identification. As a result, Vinh Long Radio and Television Station has gradually digitized their program contents. Previously, the content management was through magnetic tapes, but now the management has been digitized into files. Therefore, the management of huge volumes of data is an urgent task today. In order to well manage the existing data and update new data on a daily basis, it is necessary to build a multimedia information system in the cloud.

**Keywords:** Cloud computing, Big data, CouchDB, Kafka

### 1. MỞ ĐẦU

Việc kiểm soát lượng thông tin cực lớn, quản lý các chương trình tại một Đài Truyền hình đòi hỏi những giải pháp tốt,

do đó thách thức quản lý thông tin lưu trữ, việc truy xuất sau này và tương tác thông tin. Những thách thức trong việc quản lý thông tin như sau: *Lưu trữ thông tin:* nhiều

định dạng và kích thước lớn (text, cơ sở dữ liệu, video HD, audio); *Chia sẻ thông tin*: Truy xuất và tìm kiếm: bị trùng lặp, mỗi bộ phận khác nhau, nhu cầu sử dụng nội dung khác nhau và tìm theo ngữ cảnh của nội dung; *An toàn dữ liệu và bảo mật thông tin*: Cực kỳ quan trọng, nếu ồ ồng bị hư thì hậu quả rất lớn; *Tính mở rộng của hệ thống*: Lượng thông tin ngày càng nhiều thì khả năng mở rộng cao; *Tính sẵn dùng*: Lúc nào cần là có, mọi lúc, mọi nơi; *Tốc độ*: cũng là điều hết sức quan trọng; *Băng thông*: phải được thông suốt, không bị nghẽn nút thắt cổ chai. Từ những yêu cầu dữ liệu ngày càng lớn, mức độ phức tạp của dữ liệu và người dùng ngày càng cao. Dữ liệu cũng ngày càng phức tạp về định dạng dữ liệu, kiểu dữ liệu, đòi hỏi xử lý video có dung lượng lớn. Một hệ thống quản lý tri thức để khám phá và khai thác dữ liệu lớn trở thành nhu cầu cấp thiết hiện nay. Vì lý do này, chúng tôi đề xuất “xây dựng hệ thống thông tin đa phương tiện trên đám mây” vừa đáp ứng nhu cầu nghiệp vụ, vừa phù hợp với xu thế công nghệ 4.0 hiện nay.

Điện toán đám mây (cloud computing), còn gọi là điện toán máy chủ ảo, là mô hình điện toán sử dụng các công nghệ máy tính và phát triển dựa vào mạng Internet. Thuật ngữ “đám mây” ở đây là lối nói ẩn dụ chỉ mạng Internet (dựa vào cách được bố trí của nó trong sơ đồ mạng máy tính) và như sự liên tưởng về độ phức tạp của các cơ sở hạ tầng chứa trong nó. Ở mô hình điện toán này, mọi khả năng liên quan đến công nghệ thông tin đều được cung cấp dưới dạng các “dịch vụ”, cho phép người sử dụng truy cập các dịch vụ công nghệ từ một nhà cung cấp nào đó “trong đám mây” mà không cần phải có các kiến thức, kinh nghiệm về công nghệ đó, cũng như không cần quan tâm đến các cơ sở hạ tầng phục vụ công nghệ đó.

Từ nhu cầu thực tế công việc, chúng

tôi đề xuất xây dựng hệ thống thông tin đa phương tiện quy mô lớn trên đám mây. Với các nội dung nghiên cứu như sau: Thiết kế hệ thống thông tin đa phương tiện trên đám mây, giải pháp lưu trữ hiệu quả thông tin đa phương tiện, giải pháp truy cập/tìm kiếm thông tin đa phương tiện trên metadata, giải pháp xây dựng thao tác cập nhật dữ liệu đa phương tiện.

## 2. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

### 2.1. Nghiên cứu lý thuyết

*Điện toán đám mây*: Điện toán đám mây là việc cung cấp các dịch vụ khác nhau thông qua Internet. Các tài nguyên này bao gồm các công cụ và ứng dụng như lưu trữ dữ liệu, máy chủ, cơ sở dữ liệu, mạng và phần mềm. Điện toán đám mây là một lựa chọn phổ biến cho mọi người và doanh nghiệp vì tiết kiệm chi phí, tăng năng suất, tốc độ và hiệu quả, hiệu suất và bảo mật.

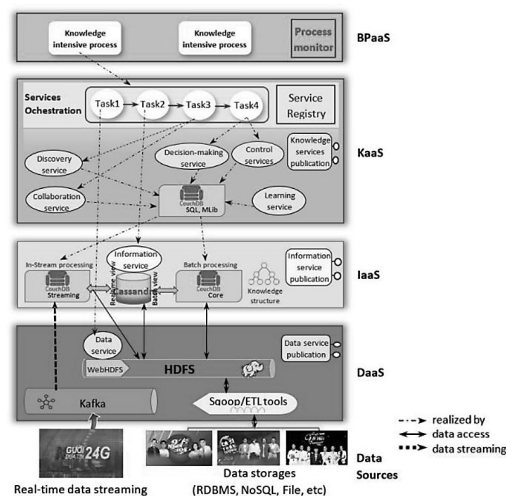
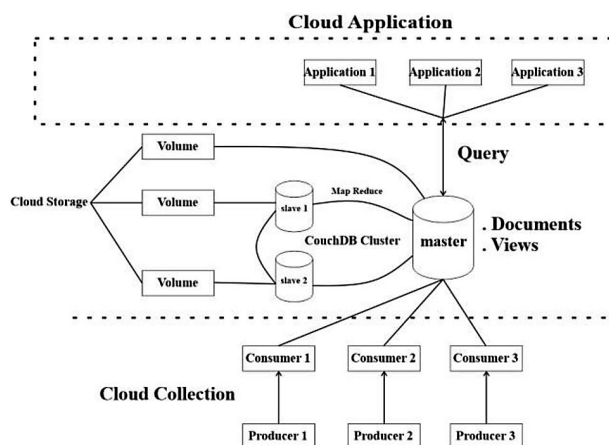
*BigData*: Định nghĩa dữ liệu lớn của Gartner: Dữ liệu lớn là khối lượng thông tin cực lớn, tốc độ và tính đa dạng cao, đòi hỏi các hình thức xử lý thông tin sáng tạo, hiệu quả (Beyer and Laney 2012). *Volume* thể hiện lượng thông tin có thể đến từ các nguồn dữ liệu khác nhau. *Velocity* là tất cả về tốc độ truyền và xử lý dữ liệu. *Variety* là sự đa dạng đề cập đến nhiều nguồn dữ liệu và các loại dữ liệu khác nhau cả có cấu trúc, không cấu trúc và bán cấu trúc.

*Cơ sở dữ liệu CouchDB*: CouchDB là một cơ sở dữ liệu mã nguồn mở được phát triển bởi nền tảng phần mềm Apache. Trọng tâm là tính dễ sử dụng, giao diện web. CouchDB là một cơ sở dữ liệu lưu trữ tài liệu NoSQL.

*Apache Kafka*: là một nền tảng xử lý dòng dữ liệu phân tán, thời gian thực, khả năng chịu được lỗi cao và là hệ thống nhắn tin nhanh. Có khả năng mở rộng ngang và chuyển giao dữ liệu đáng tin cậy.

## 2.2. Kết quả ứng dụng thực tiễn

### 2.2.1 Thiết kế tổng quát hệ thống



Hình 1. Thiết kế tổng quát hệ thống và mô tả quá trình thực hiện số hóa nghiệp vụ

**Producer:** Đóng vai trò như người tạo ra các bài viết, kịch bản và video ban đầu từ các phóng viên tác nghiệp hoặc nhà sản xuất. Liên tục gửi các text, video đến hàng đợi Kafka.

**Consumer:** đóng vai trò như một đơn vị tổ chức sản xuất hoặc biên tập viên. Sản xuất các chương trình, nhận các text, video từ hàng đợi Kafka.

CouchDB Server lưu trữ cơ sở dữ liệu có tên làm ultimedia, nơi lưu trữ Document. Mỗi Document được đặt tên duy nhất trong cơ sở dữ liệu và CouchDB cung cấp API RESTful HTTP để đọc và cập nhật (thêm, sửa, xóa) tài liệu cơ sở dữ liệu.

Thành phần còn lại là các Application (Web truy cập/tìm kiếm và cập nhật dữ liệu đa phương tiện) đã được đăng ký thông tin với hệ thống trên đám mây.

#### 2.2.2 Thiết kế thành phần

**Apache Kafka:** Công cụ kafka\_2.13-2.7.0 để thu thập, truyền và thu nhận dữ liệu từ khắp nơi, được chuyển về từ các phóng viên, nhà sản xuất. Sử dụng Python 2.7.17 để lập trình việc thu thập, truyền và thu nhận dữ liệu.

Tập lệnh python thu thập và truyền dữ liệu “send\_kafka.py”.

Tập lệnh python thu nhận dữ liệu “receive\_kafka.py”.

Hệ thống thu thập dữ liệu Kafka là các producer liên tục gửi dữ liệu về từ các phóng viên ở khắp nơi gồm video và kịch bản.

Producer tách hình ảnh và âm thanh của Video, sau đó tách từng frame của hình ảnh để truyền dữ liệu qua hệ thống Kafka cùng kịch bản.

Hệ thống nhận dữ liệu Kafka là các consumer liên tục nhận dữ liệu từ hệ thống truyền dữ liệu Kafka.

Consumer nhận từng frame hình để nhận diện khuôn mặt sau đó ghép lại thành file video xử lý có xác nhận tên của khuôn mặt đã nhận diện và một file đã ghép với file âm thanh tái tạo lại file gốc.

Sau khi trích xuất tên của các nhân vật trong video, cùng với nội dung của kịch bản, thực hiện đưa các nội dung nhận được vào cơ sở dữ liệu.

#### Các thuật toán xử lý

Xử lý ở bên gửi thông tin

**Bước 1:** send với topic “start” với nội dung:

```
d = {'name': name, 'w': w, 'h': h,
     'fps': fps, 'rate': rate, 'samp': samp}
```

Trong đó: name: là tên video; w, h, fps: chiều rộng, chiều cao, fps của video rate: fps của audio; samp: nchannels của video

**Bước 2:** Tạo 1 luồng khác với Thread dùng để gửi audio (dùng thư viện moviepy để lấy audio từ file video, duyệt qua từng frame của audio rồi chuyển sang kiểu bytes => gửi thông qua kafka với topic “audio”)

**Bước 3:** Kiểm tra nếu có truyền tham số file mô tả thì tạo 1 luồng gửi file mô tả (dùng thư viện docx để đọc file doc/docx chuyển thành text rồi gửi thông qua kafka với topic “text”)

**Bước 4:** dùng thư viện moviepy để lấy video từ file, duyệt qua từng frame hình và chuyển sang kiểu byte => gửi thông qua kafka trên luồng chính với topic “frame”

*Xử lý ở bên nhận thông tin*

Mô hình nhận dạng khuôn mặt:

Dùng Multi-task Cascaded Convolutional Networks (MTCNN) để xác định vị trí khuôn mặt.

Dùng FaceNet để trích xuất đặc trưng khuôn mặt.

Dùng Cosine Similarity để so khớp khuôn mặt với CSDL khuôn mặt.

Tạo CSDL khuôn mặt: Thu thập dữ liệu khuôn mặt của các diễn viên, người nổi tiếng lưu vào thư mục TapHuanLuyen, ảnh của mỗi người sẽ lưu trong 1 thư mục riêng, mỗi người thu thập ít nhất 5 ảnh để huấn luyện nhận diện khuôn mặt.

Tạo CSDL khuôn mặt bằng cách: dùng duyệt qua từng thư mục => dùng MTCNN để lấy ảnh khuôn mặt => resize ảnh về 160x160 => dùng faceNet lấy đặc trưng =>

lưu đặc trưng thu được vào mảng “emb” và đồng thời lưu tên thư mục vào mảng “label\_array” => sau khi duyệt qua hết tất cả thư mục thì dùng thư viện pickle để lưu 2 mảng “emb” và “label\_array” ra thành file “data\_face.pkl” (đây là cơ sở dữ liệu khuôn mặt)

*Quá trình nhận và xử lý dữ liệu:*

Tạo ra 4 Kafka Consumer với các Topic: start, frame, audio, text.

Tạo 1 luồng với Thread để nhận dữ liệu audio, khi nhận được tất cả dữ liệu thì dùng thư viện wavio để save audio lại với đuôi mở rộng là .wav

Tạo thêm 1 luồng với Thread để nhận dữ liệu từ file mô tả, dùng thư viện docx để lưu kết quả lại thành file docx.

Tại luồng chính tiến hành nhận các frame ảnh và xử lý:

Nhận tất cả các frame hình để lưu lại với dạng no\_audio (dùng thư viện cv2 lưu video).

Ngoài ra, cứ mỗi 5 frame hình thì tiến hành nhận dạng khuôn mặt, dùng MTCNN để lấy các khuôn mặt => resize về 160x160 => FaceNet lấy đặc trưng => dùng Cosine Similarity để so khớp khuôn mặt với CSDL khuôn mặt => lớn hơn 0.8 thì trả về tên tương ứng => chỉ trả về khuôn mặt dự đoán được có độ chính xác cao nhất và những khuôn mặt có độ chính xác thấp hơn không quá 0.04 (tránh nhận những khuôn mặt sai).

Lưu kèm 1 video xử lý khuôn mặt (dùng thư viện cv2 lưu video).

Sau khi tất cả dữ liệu đã nhận được hết thì dùng thư viện moviepy để ghép video no\_audio và file audio => video hoàn chỉnh, sau đó xóa các file không cần thiết.

Dùng thư viện CouchDB để kết nối đến CSDL CouchDB và update kết quả lên CSDL với dạng Json.



Hình 2. Đánh dấu nhận diện nhân vật

### Apache CouchDB

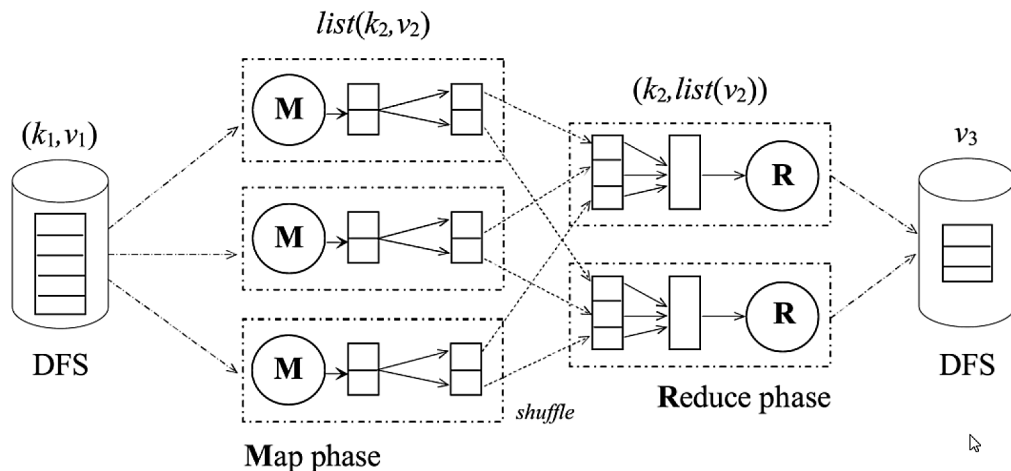
Cài đặt Apache CouchDB v.3.1.1 để lưu trữ cơ sở dữ liệu.

Truy cập địa chỉ local [http://127.0.0.1:5984/\\_utils](http://127.0.0.1:5984/_utils) để chạy cơ sở dữ liệu CouchDB trên trình duyệt web hoặc địa chỉ IP của các

node đã cấu hình trong cluster.

### MapReduce

MapReduce là một mô hình lập trình song song và phân tán thích hợp chạy trên các cụm máy tính có quy mô đến hàng nghìn nút theo cách có thể chịu được lỗi.



Hình 3. Thực thi phân tán MapReduce

Một công việc MapReduce bao gồm hai giai đoạn riêng biệt, đó là giai đoạn map và giai đoạn reduce. Mỗi giai đoạn thực thi một chức năng do người dùng xác định trên một cặp giá trị khóa. Hàm map do người dùng xác định (M) nhận một cặp đầu vào  $(k_1, v_1)$  và xuất ra danh sách các cặp khóa/ giá trị trung gian  $h \langle k_2, v_2 \rangle$ .

$$(k_1, v_1) \xrightarrow{\text{map}} \text{list}(k_2, v_2) \quad (1)$$

Các giá trị trung gian được liên kết với cùng một khóa  $k_2$  được nhóm theo khung và sau đó được chuyển đến hàm reduce để tổng hợp các giá trị.

$$(k_2, \text{list}(v_2)) \xrightarrow{\text{reduce}} v_3 \quad (2)$$

MapReduce được thiết kế để xử lý một tập dữ liệu duy nhất. Việc kết hợp một số đầu vào với một khung MapReduce rất phức tạp. Mặc dù các phép tham gia trong MapReduce có thể được thực hiện theo nhiều cách, nhưng hiệu suất tương đối của các thuật toán khác nhau phụ thuộc vào các giả định nhất định như kích thước của đầu vào, ràng buộc dữ liệu và tỷ lệ tham gia.

Truy vấn/tìm kiếm nội dung dữ liệu trên đám mây theo mô hình MapReduce và cập nhật dữ liệu đa phương tiện trên đám mây

Mục đích của hệ thống truy vấn video trên đám mây nhằm tìm kiếm những nội dung trong video gần giống với từ khóa đầu vào, truy vấn video dựa theo nội dung trong mô hình Mapreduce trên đám mây nhằm đạt được hiệu suất cao.

Kết quả được hiển thị trên giao diện web cho người dùng trên đám mây là: Thông tin về các trường bao gồm Tên, thể loại, đạo diễn, kịch bản, biên tập, File; Các file tương ứng với nội dung được tìm kiếm.

*Cập nhật dữ liệu trên đám mây:*

*Chức năng thêm dữ liệu:* Người dùng nhập thông tin tên, thể loại, đạo diễn, kịch bản, biên tập, file trong giao diện web thêm trên đám mây.

*Chức năng chỉnh sửa dữ liệu:* Người dùng chọn thông tin cần chỉnh sửa, nhập

thông tin chỉnh sửa, nếu hợp lý thông tin mới sẽ được thay thế trong CSDL trên đám mây.

*Chức năng xóa dữ liệu:* Người dùng chọn thông tin cần xóa, xác nhận xóa, thông tin sẽ được xóa khỏi CSDL trong đám mây.

### 2.3. Đóng góp của nghiên cứu

Nghiên cứu liên quan đến thành phần thiết kế, kiến trúc được đề xuất bao gồm một tập hợp các cấu trúc, mô hình và phương thức (Le Dinh Thang, et al 2016. “Towards a Service-Oriented Architecture for Knowledge Management in Big Data Era.” IJIT vol.14, no.4 2018). Các thiết kế này dựa trên các nguyên tắc nghiên cứu khoa học thiết kế, bao gồm các thành phần như cấu trúc, mô hình, phương pháp và thực thi.

**Bảng 1. Mô tả các thành phần thiết kế**

| Thành phần  | Mô tả                                                                                                                           |
|-------------|---------------------------------------------------------------------------------------------------------------------------------|
| Cấu trúc    | Các loại khái niệm khác nhau liên quan đến các đối tượng tri thức trong một kiến trúc dựa trên dữ liệu và môi trường nghiệp vụ. |
| Mô hình     | Một tập hợp các mối quan hệ giữa các đối tượng thông tin ở các cấp độ khác nhau trong kiến trúc.                                |
| Phương pháp | Một tập hợp các hoạt động hỗ trợ quá trình phát triển kiến thức                                                                 |
| Thực thi    | Thực hiện tại các phòng chuyên môn                                                                                              |

Phân tích tĩnh được sử dụng để kiểm tra cấu trúc của kiến trúc đề xuất và phân tích kiến trúc được sử dụng để nghiên cứu sự phù hợp của thành phần thiết kế với kiến trúc hệ thống thông tin hiện tại. Truy vấn/tìm kiếm và cập nhật dữ liệu video trên đám mây sử dụng mô hình MapReduce là một kỹ thuật tạo thuận lợi cho việc tìm kiếm và cập nhật dữ liệu của các video trong bộ cơ sở dữ liệu lớn. Trong kỹ thuật này, phân tích dữ liệu video được thực hiện trên đặc trưng trích xuất từ các video. Từ những thông tin đã được trích xuất, nghiên cứu đề xuất lập

chỉ mục dữ liệu các video theo các tiêu chí khác nhau, dữ liệu được phân tích để hỗ trợ truy cập/tìm kiếm và cập nhật dữ liệu trên Metadata trên đám mây. Tất cả quá trình lập chỉ mục, tìm kiếm và cập nhật dữ liệu video sẽ được thực hiện bằng mô hình MapReduce trong CouchDB và công cụ tìm kiếm Elasticsearch.

### 3. KẾT LUẬN

Hệ thống đã thực hiện quy trình số hóa nghiệp vụ sản xuất chương trình cho Đài Phát thanh và Truyền hình Vĩnh Long. Hệ thống sử dụng kiến trúc hệ thống quản lý thông tin (KMS) hiện đại trên nền tảng

đám mây. Điều này có ý nghĩa hết sức quan trọng trong thời đại 4.0 và quá trình số hóa ngày càng mạnh mẽ như hiện nay.

Những thành phần của hệ thống được xây dựng bao gồm: Cơ sở dữ liệu đa phương tiện cho Đài Phát thanh và Truyền hình Vĩnh Long trên công nghệ đám mây. Thành phần Producer: đóng vai trò như các nguồn video và text từ các phóng viên và quay phim. Thu thập dữ liệu text và video, tách video thành từng frame hình và audio. Sau đó chuyển (text, từng frame hình, audio) qua hệ thống Kafka streaming để truyền dữ liệu. Thành phần Consumer: đóng vai trò như những biên tập viên sản

xuất chương trình dựa trên các nguồn từ Producer. Nhận từng frame hình, audio, text từ kafka streaming. Sau đó nhận diện đối tượng, tái tạo file video.

Hệ thống quản lý cơ sở dữ liệu CouchDB: sau khi nhận dữ liệu, phân tích dữ liệu và đưa vào hệ thống quản lý cơ sở dữ liệu CouchDB.

Sau cùng, từ việc triển khai hệ thống trên đám mây cho thấy tính khả thi của giải pháp đề xuất xây dựng “hệ thống thông tin đa phương tiện trên đám mây” vừa đáp ứng nhu cầu nghiệp vụ, vừa phù hợp với xu thế công nghiệp 4.0 hiện nay.

### TÀI LIỆU THAM KHẢO:

- [1]. Thang Le Dinh, Phan Thuong Cang, Trung Bui, Manh Chien Vu. 2018. TOWARDS A SERVICE-ORIENTED ARCHITECTURE FOR KNOWLEDGE MANAGEMENT IN BIG DATA ERA. International Journal of Intelligent Information Technologies. 14. 24-38. <http://doi.org/10.4018/IJIT.2018100102>.
- [2]. Thang Le Dinh, Thuong-Cang Phan, Trung Bui. Twenty-second Americas Conference on Information Systems, San Diego, 2016. “Towards an Architecture for Big Data-Driven Knowledge Management Systems”.
- [3]. <https://docs.couchdb.org/en/stable> (20,05,2021).
- [4]. <https://tecadmin.net/how-to-install-apache-kafka-on-ubuntu-20-04> (20,05,2021).