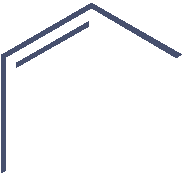
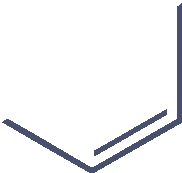
[Sensors](https://www.mdpi.com/journal/sensors)



*Bài viết*

**Kho dữ liệu lớn cho các ứng dụng dữ liệu nhạy cảm với chăm sóc sức khỏe**

# Arsalan Shahid *∗*, Thiên An Ngọc Nguyễn &; M-Tahar Kechadi

Trường Khoa học Máy tính, Đại học Dublin, Belfield, Dublin 4, Ireland; [an-thien-ngoc.nguyen@ucdconnect.ie](mailto:an-thien-ngoc.nguyen@ucdconnect.ie) (T.-A.N.N.); [tahar.kechadi@ucd.ie](mailto:tahar.kechadi@ucd.ie) (M.-T.K.)

**\*** Thư từ: [arsalan.shahid@ucd.ie](mailto:arsalan.shahid@ucd.ie)

[Kiểm tra ROR](https://www.mdpi.com/1424-8220/21/7/2353?type=check_update&version=3)

[**Cập nhật**](https://www.mdpi.com/1424-8220/21/7/2353?type=check_update&version=3)

**Trích dẫn:** Shahid, A.; Nguyễn, T.-A.N.; Kechadi, M.-T. Dữ liệu lớn

Kho cho các ứng dụng dữ liệu nhạy cảm với chăm sóc sức khỏe. *Cảm biến* **2021**, *21*,

2353. [https://doi.org/10.3390/](https://doi.org/10.3390/s21072353)  [S21072353](https://doi.org/10.3390/s21072353)

Biên tập viên học thuật: Biswanath Samanta Ngày nhận: 8 Tháng Ba 2021

**Tóm tắt:** Béo phì là một vấn đề sức khỏe cộng đồng lớn trên toàn thế giới và tỷ lệ béo phì ở trẻ em là mối quan tâm đặc biệt. Các can thiệp hiệu quả để ngăn ngừa và điều trị béo phì ở trẻ em nhằm mục đích thay đổi hành vi và phơi nhiễm ở cấp độ cá nhân, cộng đồng và xã hội. Tuy nhiên, việc theo dõi và đánh giá những thay đổi như vậy là rất khó khăn. Dự án EU Horizon 2020 "Dữ liệu lớn chống béo phì ở trẻ em (BigO)" nhằm thu thập dữ liệu quy mô lớn từ một số lượng lớn trẻ em sử dụng các công nghệ cảm biến khác nhau để tạo ra các mô hình tỷ lệ béo phì toàn diện cho các dự đoán dựa trên dữ liệu về các chính sách cụ thể trên cộng đồng. Nó tiếp tục cung cấp giám sát thời gian thực về các phản ứng của dân số, được hỗ trợ bởi phân tích và trực quan hóa dữ liệu thời gian thực có ý nghĩa. Vì BigO liên quan đến việc giám sát và lưu trữ dữ liệu cá nhân liên quan đến hành vi của nhóm dân số dễ bị tổn thương, nên việc thể hiện dữ liệu, bảo mật và kiểm soát truy cập là rất quan trọng. Trong bài viết này, chúng tôi trình bày ngắn gọn kiến trúc hệ thống BigO và tập trung vào các thành phần cần thiết của hệ thống liên quan đến kiểm soát truy cập dữ liệu, lưu trữ, ẩn danh và các giao diện tương ứng với phần còn lại của hệ thống. Chúng tôi đề xuất kiến trúc kho dữ liệu ba lớp: Lớp back-end bao gồm một hệ thống quản lý cơ sở dữ liệu để thu thập dữ liệu, khử nhận dạng và ẩn danh các tập dữ liệu gốc. Các quyền dựa trên vai trò và chế độ xem bảo mật được triển khai trong lớp kiểm soát truy cập. Cuối cùng, lớp điều khiển điều chỉnh các giao thức truy cập dữ liệu cho bất kỳ truy cập dữ liệu và phân tích dữ liệu nào. Chúng tôi trình bày thêm các phương pháp biểu diễn dữ liệu và các mô hình lưu trữ xem xét các cơ chế bảo mật và quyền riêng tư. Các kế hoạch bảo mật và quyền riêng tư dữ liệu được đưa ra dựa trên các loại cá nhân được thu thập, loại người dùng, lưu trữ dữ liệu, truyền dữ liệu và phân tích dữ liệu. Chúng tôi thảo luận chi tiết về những thách thức của việc bảo vệ quyền riêng tư trong ứng dụng theo hướng dữ liệu phân tán lớn này và triển khai các giao thức phân tích dữ liệu nhận thức được quyền riêng tư mới để đảm bảo rằng các mô hình được đề xuất đảm bảo quyền riêng tư và bảo mật của bộ dữ liệu. Cuối cùng, chúng tôi trình bày kiến trúc hệ thống BigO và việc triển khai nó tích hợp các giao thức nhận thức về quyền riêng tư.

**Từ khóa:** đại diện dữ liệu lớn; dữ liệu chăm sóc sức khỏe; bảo mật dữ liệu lớn; mô hình nhận thức về quyền riêng tư

Đã chấp nhận: 25 Tháng Ba 2021

Xuất bản: 28 Tháng Ba 2021

**Lưu ý của nhà xuất bản:** MDPI vẫn trung lập đối với các khiếu nại về quyền tài phán trong các bản đồ đã xuất bản và các liên kết thể chế.

[](https://creativecommons.org/)

**Bản quyền:** *◯c* 2021 của các tác giả. Người được cấp phép MDPI, Basel, Thụy Sĩ. Bài viết này là một bài viết truy cập mở được phân phối theo các điều khoản và điều kiện của giấy phép Creative Commons Ghi công (CC BY) [(https://](https://creativecommons.org/licenses/by/4.0/)  [creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/) 4.0/).

# Giới thiệu

Với sự gia tăng tỷ lệ phổ biến thu nhập trên toàn thế giới, gần 7,8% bé trai và 5,6% bé gái bị béo phì ở trẻ em vào năm 2016 [[1](#_bookmark27)]. Vì béo phì là một vấn đề sức khỏe cộng đồng toàn cầu lớn [[2](#_bookmark28)], với ý nghĩa chi phí đáng kể cho cả cá nhân và xã hội nói chung, các biện pháp mạnh mẽ phải được thực hiện để can thiệp vào bệnh béo phì ở trẻ em. Trẻ em béo phì dễ gặp phải một loạt các vấn đề sức khỏe cả về thể chất và tâm lý xã hội [[3](#_bookmark29)]. Béo phì cũng góp phần vào các vấn đề sức khỏe quan trọng như bệnh tiểu đường loại 2 và bệnh tim mạch vành phát triển ở thời thơ ấu [[4](#_bookmark30)].

Trong hai thập kỷ qua, căn nguyên béo phì đã được nghiên cứu kỹ lưỡng trong các ngành khoa học y sinh, tức là các thuật ngữ về biến thể di truyền. Tuy nhiên, biến thể di truyền giữa mọi người đã được phát hiện để làm sáng tỏ gần 1,5% sự thay đổi giữa các cá nhân trong chỉ số khối cơ thể (BMI) [[5](#_bookmark31)]. Các phương pháp dịch tễ học hiện tại phần lớn đã khám phá vấn đề béo phì như một yếu tố nguy cơ của bệnh không lây nhiễm nhưng chưa nghiên cứu nó như là kết quả của hành vi trong môi trường béo phì [[6](#_bookmark32)]. Với sự thiếu tin cậy của

*Sensors* **2021**, *21*, 2353. <https://doi.org/10.3390/s21072353> <https://www.mdpi.com/journal/sensors>

Can thiệp dược lý và sự xâm lấn của các thủ tục phẫu thuật, sự phụ thuộc ngày càng tăng đang được đưa vào các biện pháp khắc phục hành vi.

Các can thiệp nhắm vào các khía cạnh khác nhau của mô hình hành vi của trẻ em, chẳng hạn như cách chúng ăn, cách chúng di chuyển, cách chúng tương tác trong môi trường và cách chúng ngủ [[7](#_bookmark33)] có thể có kết quả tích cực đối với bệnh béo phì ở trẻ em [8,9]. Thật không may, hầu hết các hành động y tế công cộng toàn cầu bị giới hạn trong các 'chính sách bao trùm' bừa bãi và việc thiếu sâu sắc khung giám sát và đánh giá chung cũng không giúp ích gì [[10](#_bookmark36)].

Trong thập kỷ qua, những tiến bộ trong lĩnh vực thiết bị di động và thiết bị đeo được đã tạo ra một thế hệ nghiên cứu chăm sóc sức khỏe sáng tạo mới [[11](#_bookmark37)]. Những nỗ lực sáng tạo như dự án SPLENDID EU [[12](#_bookmark38)] đẩy ranh giới trong việc sử dụng công nghệ trong cuộc sống hàng ngày để sàng lọc, theo dõi và bình thường hóa các hành vi liên quan đến sự tiến triển của bệnh béo phì. Song song, những thành tựu mới trong lĩnh vực thu thập, xử lý và phân tích Dữ liệu lớn cho phép mở rộng những nỗ lực đó trên quy mô lớn hơn, cho phép người dân tích cực tham gia vào việc định hình lại các chính sách y tế ở cấp khu vực. Thu thập dữ liệu lớn nhằm mục đích khám phá mối quan hệ nhân quả giữa hành vi, môi trường xây dựng và béo phì.

Những phát triển gần đây trong các lĩnh vực khoa học thay đổi hành vi, y tế công cộng, nhi khoa lâm sàng, ICT, khoa học công dân và phân tích dữ liệu lớn có thể được khai thác để thực hiện các dự án nghiên cứu đa ngành nhằm giải quyết việc phòng ngừa và điều trị béo phì ở trẻ em và thanh thiếu niên ở cấp độ dân số. Dự án H2020 "BigO: Dữ liệu lớn chống béo phì ở trẻ em" ([http://bigoprogram.eu](http://bigoprogram.eu/), truy cập ngày 21 tháng 3 năm 2021) là một ví dụ về những nỗ lực nghiên cứu như vậy. Dự án BigO nhằm xác định lại cách các chiến lược nhắm vào tỷ lệ béo phì ở trẻ em được triển khai trong các xã hội châu Âu. Những chiến lược mới này sử dụng các phương pháp thu thập dữ liệu công dân-nhà khoa học thu thập dữ liệu quy mô lớn bằng cách sử dụng các công nghệ di động khác nhau (điện thoại thông minh, dây đeo cổ tay, mandometer) [[13](#_bookmark39)]. Mục tiêu chính là tạo ra các mô hình dự đoán toàn diện về ma trận phụ thuộc tỷ lệ béo phì để nghiên cứu hiệu quả của các chính sách cụ thể đối với cộng đồng và giám sát thời gian thực về phản ứng dân số. Dữ liệu được các nhà khoa học công dân tải lên cơ sở hạ tầng đám mây BigO để tổng hợp, phân tích và trực quan hóa. Việc thu thập dữ liệu quy mô lớn trong BigO cho phép các nhà nghiên cứu tạo ra các mô hình để phân tích các yếu tố nguy cơ hành vi và dự đoán tỷ lệ béo phì, thông qua các hiệp hội với các mô hình hành vi cộng đồng và tỷ lệ béo phì tại địa phương.

Vì BigO liên quan đến việc giám sát và lưu trữ dữ liệu cá nhân (ảnh về bữa ăn, đồ uống, quảng cáo thực phẩm và đồ uống và dữ liệu GPS) liên quan đến hành vi của nhóm dân số dễ bị tổn thương (tức là trẻ em và thanh thiếu niên), việc thể hiện dữ liệu, bảo mật và kiểm soát truy cập là một thách thức. Trong nghiên cứu này, trước tiên chúng tôi đã phát triển các thành phần cần thiết của hệ thống BigO liên quan đến truy cập và lưu trữ dữ liệu, bao gồm định nghĩa và triển khai giao diện của chúng với các thành phần hệ thống khác. Điều này đã tạo điều kiện thuận lợi đáng kể cho việc tổng hợp dữ liệu, phân tích dữ liệu và trực quan hóa, đồng thời tuân thủ quyền riêng tư dữ liệu của các cá nhân và bảo mật của toàn bộ hệ thống. Chính xác hơn, chúng tôi đã phát triển kiến trúc kho dữ liệu ba lớp cho BigO. bao gồm lớp back-end, lớp kiểm soát truy cập và lớp điều khiển. Lớp back-end bao gồm một hệ thống quản lý cơ sở dữ liệu để thu thập dữ liệu, khử nhận dạng và ẩn danh. Các quyền dựa trên vai trò và chế độ xem bảo mật được triển khai trong lớp kiểm soát truy cập. Cuối cùng, lớp điều khiển điều chỉnh các giao thức truy cập dữ liệu.

Việc chia sẻ dữ liệu cá nhân nhạy cảm trong lĩnh vực chăm sóc sức khỏe được quy định bởi các quy tắc và luật pháp như Chỉ thị 95/46 / EC của Liên minh Châu Âu và HIPAA của Hoa Kỳ [[14,15](#_bookmark40)]. Hai cách tiếp cận thường được chấp nhận để bảo vệ quyền riêng tư: (a) lấy sự đồng ý của bệnh nhân và (b) ẩn danh dữ liệu cá nhân. Cách tiếp cận đầu tiên duy trì tính chính xác của kết quả phân tích nhưng tốn thời gian, không linh hoạt và dễ bị vi phạm dữ liệu hơn khi được sử dụng để thu thập dữ liệu từ một nhóm lớn bệnh nhân [[16](#_bookmark42)]. Tuy nhiên, cách tiếp cận ẩn danh linh hoạt hơn nhưng tác động đến chất lượng và phân tích dữ liệu; đặc biệt là trong trường hợp dữ liệu chiều cao [[17](#_bookmark43)].

Xem xét độ phức tạp cao của dữ liệu BigO, trước tiên chúng tôi giới thiệu lược đồ đại diện dữ liệu và các mô hình lưu trữ có tính đến quyền riêng tư và bảo mật.

Các kế hoạch bảo mật và quyền riêng tư dữ liệu được đưa ra dựa trên các loại dữ liệu cá nhân được thu thập trong các giao thức lưu trữ, truyền dữ liệu và truy cập dữ liệu. Chúng tôi thảo luận chi tiết về những thách thức của việc bảo vệ quyền riêng tư và triển khai các giao thức phân tích dữ liệu nhận thức về quyền riêng tư hiệu quả đối với quyền riêng tư cho trẻ em. Lưu ý rằng hệ thống được thiết kế để nhận ra và bỏ qua thông tin cá nhân dư thừa. Cuối cùng, chúng tôi đã triển khai kiến trúc hệ thống BigO tích hợp các giao thức nhận thức về quyền riêng tư.

Phần còn lại của bài viết được sắp xếp như sau: Phần [2](#_bookmark0) trình bày các thách thức ẩn danh dữ liệu và các yêu cầu hệ thống để phân tích dữ liệu chăm sóc sức khỏe cùng với các kiến trúc nhận thức về quyền riêng tư hiện có. Tiếp theo là tổng quan về hệ thống BigO và các phương pháp thu thập dữ liệu với luồng dữ liệu dựa trên loại người dùng trong Phần [3](#_bookmark4) và [4](#_bookmark7), tương ứng. Trong Phần [5](#_bookmark10), chúng tôi trình bày kiến trúc kho dữ liệu BigO và biểu diễn dữ liệu lớn trong hệ thống lưu trữ. Phần [6](#_bookmark13) thảo luận về các cân nhắc về bảo mật và quyền riêng tư cho BigO cùng với giao thức phân tích dữ liệu nhận thức được quyền riêng tư. Phần [7](#_bookmark18) dành cho việc triển khai hệ thống và cách giao thức nhận biết quyền riêng tư được tích hợp trong kiến trúc hệ thống. Trong Phần [8](#_bookmark20), chúng tôi cung cấp các cuộc thảo luận về ưu và nhược điểm của công việc của chúng tôi và nêu bật các hướng nghiên cứu trong tương lai. Cuối cùng, Phần [9](#_bookmark21) kết thúc bài báo.

# Bảo mật dữ liệu chăm sóc sức khỏe

Trước tiên, chúng tôi trình bày tổng quan ngắn gọn về các thách thức ẩn danh dữ liệu trong các hệ thống chăm sóc sức khỏe dựa trên dữ liệu và sau đó chúng tôi liệt kê các yêu cầu về quyền riêng tư và bảo mật cần thiết để giám sát, chia sẻ và phân tích dữ liệu cá nhân trong các hệ thống chăm sóc sức khỏe. Chúng tôi trình bày thêm một đánh giá ngắn gọn về các hệ thống hiện có và đánh giá chúng dựa trên các yêu cầu này.

* 1. *Ẩn danh và chia sẻ dữ liệu*

Các thuộc tính dữ liệu (hoặc kích thước) trong một tập dữ liệu chăm sóc sức khỏe điển hình có thể được chia thành bốn loại dựa trên độ nhạy cảm và mối quan hệ của chúng với chủ đề; (1) nhạy cảm,

(2) không nhạy cảm, (3) có thể nhận dạng và (4) gần như có thể nhận dạng. Bảng [1](#_bookmark1) cho thấy, bên cạnh các thuộc tính nhạy cảm và có thể nhận dạng rõ ràng, có một số tính năng được gọi là định danh gần như có thể được kết hợp để xác định các cá nhân cụ thể. Bảo vệ quyền riêng tư của một phần dữ liệu nhạy cảm, có thể nhận dạng hoặc gần như có thể nhận dạng được là một thách thức quan trọng. Ẩn danh dữ liệu duy trì ngưỡng tin cậy thấp để liên kết thông tin nhạy cảm với một cá nhân [[18](#_bookmark44)].

**Bảng 1.** Các loại thông tin trong Hồ sơ sức khỏe điện tử (I: có thể nhận dạng, Q-I: gần như có thể nhận dạng, S: nhạy cảm).

|  |  |  |  |
| --- | --- | --- | --- |
| **Các loại thuộc tính** | **Sự miêu tả** | **Ví dụ** | **Sự riêng tư** |
| Số nhận dạng | Nhận dạng người | Tên, email, địa chỉ | số điện thoại *I* |
| Dân số | Phân loại người cho một nhóm dân số cụ thể | chủng tộc, tuổi, giới tính, khu vực, mã bưu chính, giáo dục, nghề nghiệp, tình trạng hôn nhân | Hỏi I |
| Sinh trắc học cá nhân | Thông tin y tế liên quan đến sức khỏe thể chất | X-Quang, MRI, siêu âm, huyết áp, cholesterol, nhịp tim, dị ứng, sự cố ICU, báo cáo xét nghiệm | *S* |
| Thông tin lâm sàng | Tiền sử bệnh | chẩn đoán, liều lượng, dịch vụ điều trị, medi- cation, gặp phải, vấn đề, liệu pháp | Q-I, S |
| Thông tin tâm thần | Liên quan đến các vấn đề tâm lý, tâm lý và tâm lý xã hội | vấn đề về giấc ngủ, tâm lý, ăn kiêng quá mức, rối loạn tâm lý tình dục | *S* |
| Thông tin về phong cách sống và hoạt động | Liên quan đến các hoạt động thể chất, lối sống | hoạt động thể chất, chế độ tập luyện, dinh dưỡng, tiêu hao năng lượng thông qua các bài tập | *S* |
| Bảo hiểm và các vấn đề tài chính | Liên quan đến thanh toán, bồi hoàn, bảo hiểm | DRG, lớp tài chính, nhà cung cấp chính và chuyên gia | Q-I, S |

Kết quả của việc chia sẻ dữ liệu Liên kết bản ghi [[18](#_bookmark44)] và liên kết thuộc tính [[19](#_bookmark45)] là hai loại tấn công chính, thường được sử dụng. Cả hai điều này đều có thể được ngăn chặn bằng cách sử dụng các mô hình bảo mật, chẳng hạn như k-ẩn danh [[20](#_bookmark46)] và l-diversity [[19](#_bookmark45)]. Một số phiên bản mở rộng của các mô hình bảo mật này, bao gồm (*X*, *Y*)-*ẩn danh*, *LKC-privacy* và

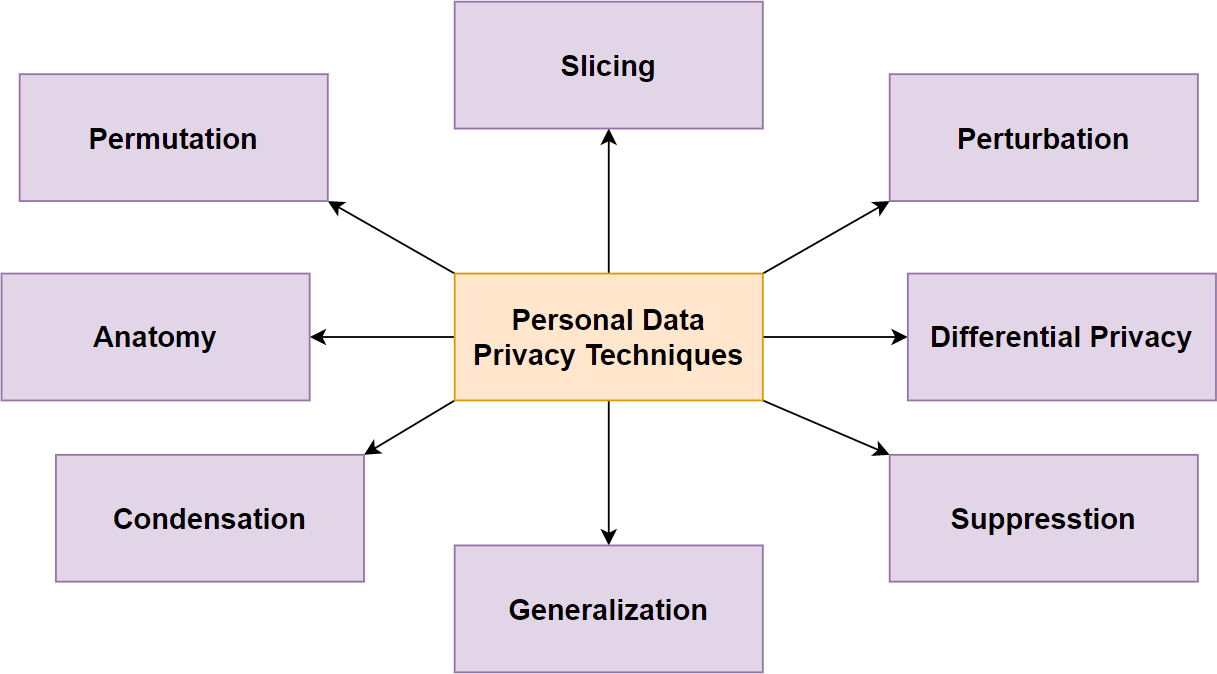
T-closeness  [[18](#_bookmark44)]; có thể được thực hiện bằng cách sử dụng các hoạt động ẩn danh khác nhau, chẳng hạn như thêm tiếng ồn, khái quát hóa, xáo trộn và nhiễu loạn [[21](#_bookmark47)]. Tuy nhiên, một số hoạt động này, bao gồm nhiễu loạn và khái quát hóa tế bào, ảnh hưởng đến chất lượng của bộ dữ liệu và khả năng sử dụng của chúng để khám phá kiến thức (bằng thuật toán khai thác) hoặc truy xuất thông tin.

* 1. *Yêu cầu về dữ liệu chăm sóc sức khỏe*

Sau đây là các yêu cầu cần thiết của bộ dữ liệu chăm sóc sức khỏe, một mặt, cần được bảo vệ và giữ kín và mặt khác có thể được phân tích để trích xuất kiến thức hữu ích có thể thúc đẩy nghiên cứu và thực hành chăm sóc sức khỏe. Ý tưởng chính là tạo ra một môi trường nơi dữ liệu riêng tư và nhạy cảm được phân tích mà không tiết lộ danh tính của các cá nhân.

* + - Quyền riêng tư: Điều này rất cần thiết chủ yếu trong lĩnh vực chăm sóc sức khỏe. Hồ sơ của bệnh nhân và các thuộc tính dữ liệu của họ rất dễ bị tấn công. Vì vậy, bắt buộc phải đưa ra các cơ chế bảo vệ để bảo vệ quyền riêng tư của bệnh nhân và cá nhân khi chia sẻ bộ dữ liệu chăm sóc sức khỏe.
    - Chất lượng dữ liệu: Dữ liệu chất lượng cao là điều cần thiết cho việc khai thác và phân tích dữ liệu (vì không có dữ liệu chất lượng dẫn đến không có kết quả chất lượng). Do đó, dữ liệu được chia sẻ nên duy trì các giá trị thuộc tính tốt đủ chi tiết để phục vụ mục đích khai thác và phân tích. Người ta cũng phải cẩn thận tính đến lời nguyền của tính chiều cao [[17](#_bookmark43)] và duy trì tính trung thực ở cấp độ hồ sơ cá nhân.
    - Tính linh hoạt: Bảo vệ quyền riêng tư phải đủ linh hoạt cho các nhiệm vụ phân tích và kỹ thuật khai thác khác nhau. Cách tiếp cận lý tưởng là thực hiện các giải pháp bảo vệ quyền riêng tư độc lập với các thuật toán khai thác và mục đích nghiên cứu.
    - Khả năng tương thích: Các mô hình bảo vệ quyền riêng tư phải tuân thủ và hỗ trợ kiến trúc tham chiếu hệ thống.
    - Tiện ích: Cung cấp một mức độ hỗ trợ để cho phép các nhà nghiên cứu truy cập lại dữ liệu bệnh nhân theo các cơ chế đạo đức và kiểm soát truy cập thích hợp.
  1. *Các kỹ thuật tăng cường quyền riêng tư dữ liệu hiện có*

Quyền riêng tư dữ liệu là một khái niệm phức tạp và đa chiều đã trở thành mối đe dọa nghiêm trọng trong thời đại hiện đại. Nó được định nghĩa trong một bối cảnh pháp lý, triết học cũng như trong bối cảnh kỹ thuật. Các kỹ thuật bảo mật thông tin cá nhân nhằm giải quyết các vấn đề liên quan đến thông tin cá nhân của một người và việc tiếp xúc với nó. Hình [1](#_bookmark2) cung cấp một cái nhìn phân loại về các kỹ thuật bảo mật dữ liệu hiện có liên quan đến ẩn danh và ngẫu nhiên hóa dữ liệu để đảm bảo quyền riêng tư của thông tin nhạy cảm.



**Hình 1.** Phân loại các kỹ thuật bảo mật dữ liệu cá nhân. Được thông qua từ [[22](#_bookmark48)].

Các hệ thống chăm sóc sức khỏe hiện tại (Phần [2.4](#_bookmark3)) thiếu sự tích hợp nhận thức về quyền riêng tư

Các giao thức theo ba chiều chính bao gồm lưu trữ, truyền tải và xử lý. Một số khung lý thuyết đã giới thiệu các phương pháp ẩn danh dữ liệu để duy trì quyền riêng tư của cá nhân, nhưng chúng làm giảm chất lượng của các nhiệm vụ phân tích dữ liệu.

Sau đây, chúng tôi xem xét và đánh giá các khuôn khổ hiện có dựa trên các yêu cầu nêu trên.

* 1. *Phân tích so sánh các hệ thống chăm sóc sức khỏe hiện đại hiện có*

Trước tiên, chúng tôi xem xét các dự án nghiên cứu gần đây của châu Âu giải quyết phòng chống béo phì. Dự án ToyBox [[23](#_bookmark49)] được trình bày và các chương trình toàn diện, tiết kiệm chi phí và liên quan đến gia đình để hỗ trợ phòng chống béo phì trong thời thơ ấu. Dự án MOODFOOD [[24](#_bookmark50)] đã khám phá việc ngăn ngừa trầm cảm bằng cách xem xét chế độ ăn uống, hành vi ăn uống của bệnh nhân và béo phì. Dự án đã nghiên cứu hơn 990 người tham gia từ Hà Lan, Vương quốc Anh, Đức và Tây Ban Nha. Dự án đã thực hiện các nghiên cứu tương quan với tiền sử trầm cảm và ăn uống của bệnh nhân. Kết quả cho thấy các hoạt động ăn uống không lành mạnh là yếu tố hàng đầu gây trầm cảm. Dự án SPOTLIGHT của EU [[25](#_bookmark51)] đã tiến hành một nghiên cứu có hệ thống để xác định các yếu tố can thiệp béo phì ở cấp độ cá nhân. Dự án đã xem xét vai trò của các yếu tố xã hội và môi trường liên quan đến béo phì và tiến hành phân tích định tính các can thiệp đa cấp xem xét phạm vi, hiệu quả, thực hiện và áp dụng. Dự án Daphne [[26](#_bookmark52)] đã phát triển một nền tảng chăm sóc sức khỏe dữ liệu dưới dạng dịch vụ để giải quyết vấn đề sức khỏe, cân nặng, hoạt động thể chất và lối sống bằng cách liên kết các nền tảng công nghệ với các hỗ trợ lâm sàng. Cần lưu ý rằng không có dự án nào nói trên, liên quan đến phòng chống béo phì, cung cấp các giao thức nhận thức về quyền riêng tư và bảo mật để xử lý thông tin cá nhân nhạy cảm của bệnh nhân.

Việc thu thập dữ liệu cá nhân bằng công nghệ cảm biến rất có giá trị để hiểu và chữa các vấn đề chăm sóc sức khỏe khác nhau như béo phì. Tuy nhiên, với sự gia tăng nhanh chóng của các công nghệ cảm biến, dữ liệu cá nhân nhạy cảm như hoạt động và đặc điểm của người dùng phải chịu các mối đe dọa về quyền riêng tư. Một bộ quy tắc như Quy định bảo vệ dữ liệu chung (GDPR) [[27](#_bookmark53)] hoạt động như một hướng dẫn về cách các công ty nên thu thập và chia sẻ dữ liệu cá nhân của công dân EU bằng cách đưa ra các giải pháp nhận thức về quyền riêng tư. Để tuân thủ các giao thức GDPR, điều cần thiết là phải có được sự đồng ý của người dùng, đặc biệt là trong lĩnh vực chăm sóc sức khỏe. Rantos et al. [[28](#_bookmark54)] đề xuất ADvoCATE để giúp người dùng kiểm soát sự đồng ý liên quan đến việc truy cập dữ liệu cá nhân của họ được thu thập từ các cảm biến và công nghệ đeo được. Cái gọi là lution được đề xuất cũng hướng dẫn các bộ điều khiển và xử lý dữ liệu để đáp ứng các yêu cầu GDPR. Larrucea et al. [[29](#_bookmark55)] đã trình bày các mô hình tham chiếu kiến trúc tuân thủ GDPR và các công cụ quản lý được gửi cho ngành chăm sóc sức khỏe. Các tác giả đã xác định các mối đe dọa bảo mật và quyền riêng tư tiềm ẩn trong các mô hình kiến trúc và sử dụng các công cụ ẩn dữ liệu để đảm bảo quyền riêng tư trong khi chia sẻ hồ sơ sức khỏe. Mustafa et al. [[30](#_bookmark56)] đã trình bày đánh giá toàn diện về các yêu cầu về quyền riêng tư đối với các ứng dụng y tế di động trong bối cảnh GDPR và đánh giá các yêu cầu về quyền riêng tư của hệ thống hỗ trợ theo dõi, chẩn đoán sớm và phát hiện các tình huống bất thường ở những bệnh nhân mắc Bệnh phổi tắc nghẽn mãn tính (COPD).

Có một vài nghiên cứu lý thuyết cũng đã được tiến hành, trong đó đề xuất phát triển kiến trúc kho dữ liệu trong các cơ sở chăm sóc sức khỏe. Sahama et al. [[31](#_bookmark57)] đã đề xuất một kiến trúc kho dữ liệu để giải quyết các vấn đề tích hợp dữ liệu. Các tác giả nhấn mạnh sự cần thiết phải khám phá quyền truy cập an toàn vào các mô hình kho dữ liệu trong khi tôn trọng các hệ thống hỗ trợ quyết định chăm sóc sức khỏe bằng cách sử dụng các cấu trúc dữ liệu dựa trên bằng chứng, dựa trên trường hợp và dựa trên vai trò. Ali Fahem Neamah [[32](#_bookmark58)] đã đề xuất một kho dữ liệu linh hoạt và có thể mở rộng để xây dựng kiến trúc hồ sơ sức khỏe điện tử. Các tác giả nhấn mạnh một số vấn đề của hệ thống để hỗ trợ phát triển ứng dụng di động, bao gồm khả năng tương thích với các nền tảng và thiết bị rất lớn. Poenaru et al. [[33](#_bookmark59)] đã đề xuất các giải pháp tiên tiến dưới dạng kho dữ liệu để lưu trữ thông tin y tế để giải quyết một số vấn đề, chẳng hạn như các tính năng mô hình hóa dữ liệu phức tạp, cấu trúc phân loại và tích hợp dữ liệu. Tất cả các giải pháp kho dữ liệu nói trên không quan tâm đến bất kỳ vấn đề riêng tư và bảo mật nào để lưu trữ, truy cập và phân tích dữ liệu.

Cùng với việc cải thiện các giao thức thu thập và tích hợp dữ liệu trong các hệ thống chăm sóc sức khỏe bằng cách lấy sự đồng ý của bệnh nhân khiếu nại GDPR, điều quan trọng không kém là lưu trữ dữ liệu cá nhân theo cách an toàn và nhận thức về quyền riêng tư mà không làm ảnh hưởng đến chất lượng dữ liệu để thực hiện các nhiệm vụ phân tích định kỳ. Do đó, các giao thức ẩn danh dữ liệu đã được triển khai trong một số khuôn khổ chăm sóc sức khỏe. Tuy nhiên, trong lĩnh vực dữ liệu lớn; Có một số thuộc tính, cũng có thể được coi là số nhận dạng gần như, ẩn danh dữ liệu trở nên không tầm thường và dẫn đến mất một lượng lớn thông tin [[17](#_bookmark43)]. Datafly, được đề xuất bởi Sweeney et al. [[34](#_bookmark60)], sử dụng hồ sơ của người nhận dữ liệu và các yêu cầu dữ liệu toàn cầu để thực hiện ẩn danh dữ liệu. Tuy nhiên, việc tích hợp Datafly trong các hệ thống chăm sóc sức khỏe rất khó khăn do việc triển khai nó như một chương trình độc lập. Hơn nữa, nó không tính đến các cuộc tấn công liên kết thuộc tính và lời nguyền của tính chiều cao. Các phần mở rộng cho Datafly, như được đề xuất trong [[35](#_bookmark61)] giải quyết các vấn đề về quyền riêng tư nhưng khả năng tương thích và chất lượng dữ liệu vẫn giữ nguyên. Agarwal et al. [[36](#_bookmark62)] đề xuất Hippocrates bằng cách cung cấp các giao thức và dịch vụ quản lý tiết lộ dữ liệu. Khung hỗ trợ kiểm soát truy cập, ẩn danh và kiểm tra nhưng thiếu tính linh hoạt, nhận thức về quyền riêng tư, chất lượng dữ liệu và khả năng tương thích. Khi được đánh giá cho các nhiệm vụ phân tích khác nhau, Hippocrates mất tính trung thực ở mức kỷ lục và gây ra kết quả khai thác sai [[18](#_bookmark44)] vì các hoạt động ẩn danh, chẳng hạn như nhiễu loạn và tổng quát hóa tế bào. Prasser et al. [[37](#_bookmark63)] đã đề xuất một khung ARX giải quyết các vấn đề về quyền riêng tư bằng cách triển khai các mô hình nhận thức về quyền riêng tư nhưng thiếu chất lượng dữ liệu do không xem xét tác động của tính chiều cao.

Nguyen et al. [[38](#_bookmark64)] đã trình bày các giao thức nhận thức về quyền riêng tư cho các hệ thống Hồ sơ sức khỏe điện tử (EHR) sử dụng các chế độ xem được bảo mật cùng với kiến trúc phần mềm trung gian cấp cao. Tran et al. [[39](#_bookmark65)] đã đề xuất một kiến trúc phân tán dựa trên mô hình để bảo mật trong lưu trữ và phân tích dữ liệu chăm sóc sức khỏe kiểm soát quyền truy cập vào dữ liệu nhạy cảm và kiểm soát nhiệm vụ truyền giữa các nút phân tán. Hai chiến lược này có hiệu quả và cần được nghiên cứu thêm.

* 1. *Tóm tắt*

Các nghiên cứu gần đây trong lĩnh vực chăm sóc sức khỏe đã tập trung vào việc phát triển các giao thức tuân thủ GDPR dưới dạng hệ thống quản lý sự đồng ý để thu thập dữ liệu. Các hệ thống kho dữ liệu chăm sóc sức khỏe hiện đại hiện có giải quyết các vấn đề khác nhau, chẳng hạn như tính linh hoạt, khả năng mở rộng, tích hợp dữ liệu và khả năng tương thích hệ thống phần mềm. Tuy nhiên, các hệ thống này thiếu sự tích hợp của các giao thức nhận thức về quyền riêng tư và bảo mật để lưu trữ, truy cập và phân tích dữ liệu. Một số khuôn khổ đã giới thiệu các phương pháp ẩn danh dữ liệu để duy trì quyền riêng tư của các cá nhân, nhưng khi được đánh giá cho các nhiệm vụ phân tích mang lại kết quả khai thác không chính xác. Chất lượng dữ liệu bị suy giảm hơn nữa khi tác động của tính chiều cao không được tính đến. Một số công trình đã trình bày các giao thức nhận thức về quyền riêng tư bằng cách triển khai các chế độ xem bảo mật và kiểm soát truy cập để lưu trữ EHR, nhưng chúng yêu cầu điều tra thêm trong các cơ sở chăm sóc sức khỏe trong thế giới thực.

Tóm lại, kiến trúc kho dữ liệu chăm sóc sức khỏe nên thực hiện các giao thức nhận thức về quyền riêng tư để giám sát và lưu trữ dữ liệu cá nhân. Các mục tiêu chính trong khi xây dựng một bộ dữ liệu chăm sóc sức khỏe lành mạnh bao gồm bảo vệ dữ liệu và quyền riêng tư, và sự phù hợp để phân tích và khai thác kiến thức sâu sắc để cải thiện nghiên cứu và thực hành chăm sóc sức khỏe. Nói cách khác, các nền tảng chăm sóc sức khỏe dựa trên dữ liệu nên cố gắng phát triển một môi trường nơi dữ liệu cá nhân và nhạy cảm riêng tư có thể được phân tích mà không tiết lộ danh tính của các cá nhân.

Dự án BigO nhằm thu thập dữ liệu quy mô lớn của hơn 25.000 cá nhân trong các trường học và phòng khám để phát triển các chính sách phòng chống béo phì hiệu quả. Việc thu thập, lưu trữ và xử lý dữ liệu quy mô lớn như vậy đòi hỏi phải nghiên cứu và phát triển theo hướng thực hiện các giao thức và khuôn khổ bảo mật và quyền riêng tư mạnh mẽ trong kiến trúc kho dữ liệu. Trong công việc này, chúng tôi đề xuất kiến trúc kho dữ liệu ba lớp bao gồm:

1. Một lớp back-end với hệ thống quản lý cơ sở dữ liệu để thu thập dữ liệu, khử nhận dạng và ẩn danh các bộ dữ liệu gốc.
2. Các quyền dựa trên vai trò và chế độ xem bảo mật được triển khai trong lớp con-trol truy cập.
3. Lớp điều khiển điều chỉnh các giao thức truy cập dữ liệu cho bất kỳ truy cập dữ liệu và phân tích dữ liệu nào.

Chúng tôi trình bày thêm các phương pháp biểu diễn dữ liệu và các mô hình lưu trữ có tính đến các cơ chế bảo mật và quyền riêng tư. Các kế hoạch bảo mật và quyền riêng tư dữ liệu được đưa ra dựa trên các loại cá nhân được thu thập, loại người dùng, lưu trữ dữ liệu, truyền dữ liệu và phân tích dữ liệu. Cuối cùng, chúng tôi trình bày kiến trúc hệ thống BigO và việc triển khai nó tích hợp các giao thức nhận thức về quyền riêng tư.

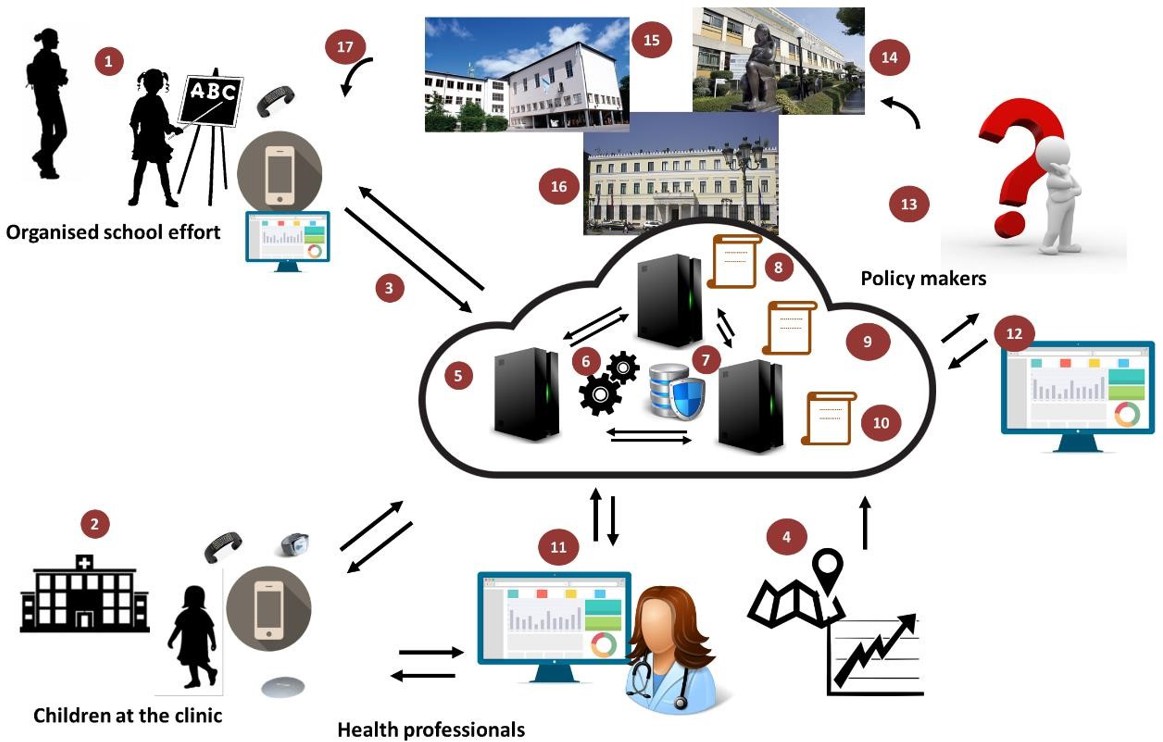
# Hệ thống BigO—Tổng quan

Phần này trình bày các mục tiêu Bigo, các nguồn dữ liệu và các phép đo, và các bên liên quan để hiểu nguyên nhân gây béo phì ở trẻ em.

BigO cố gắng cung cấp một hệ thống sáng tạo, cho phép Cơ quan Y tế Công cộng (PHA) đánh giá cộng đồng của họ dựa trên nguy cơ mắc bệnh béo phì và thực hiện các hành động tại địa phương, dựa trên bằng chứng khách quan. Hình [2](#_bookmark5) cung cấp một cái nhìn tổng quan về nền tảng BigO. Trẻ em hoặc thanh thiếu niên trong các trường học và phòng khám tham gia với tư cách là nhà khoa học công dân, để cung cấp dữ liệu thông qua các cảm biến của thiết bị di động của họ (điện thoại hoặc đồng hồ thông minh). Họ tương tác với hệ thống thông qua các ứng dụng điện thoại di động và / hoặc đồng hồ thông minh, cổng thông tin trường học (thông qua giáo viên của họ), cổng thông tin lâm sàng (thông qua bác sĩ lâm sàng của họ) và cổng cộng đồng trực tuyến. Các cổng web trực tuyến cung cấp thông tin chi tiết từ dữ liệu được thu thập bởi nền tảng BiGo. Chúng cũng cho phép sinh viên hình dung các bản tóm tắt đơn giản của dữ liệu và định lượng đóng góp của họ cho sáng kiến BigO và cách hành động chống béo phì. Các ứng dụng di động và cổng web hoạt động như người thu thập dữ liệu cũng như cơ chế tương tác, để giúp người dùng đóng góp với dữ liệu của họ và hiểu tại sao và làm thế nào dữ liệu của họ hữu ích trong trường hợp này. Dữ liệu của họ được sử dụng để đo lường các chỉ số hành vi và Điều kiện bên ngoài cục bộ (LEC) của môi trường, có liên quan đến béo phì ở trẻ em. Để đo LEC, dữ liệu cũng được thu thập từ các nguồn bên ngoài có sẵn công khai, chẳng hạn như bản đồ, Hệ thống thông tin địa lý (GIS) và các dịch vụ của cơ quan thống kê. Thông tin thu thập được xử lý bằng cách sử dụng các công cụ phân tích, trực quan hóa và mô phỏng BigO, trích xuất các chỉ số có ý nghĩa mô tả hành vi, môi trường, cũng như các mô hình về mối quan hệ của chúng. Các phép đo kết quả hỗ trợ hoạt động của Cố vấn chính sách, Người hoạch định chính sách và các dịch vụ cố vấn lâm sàng và trường học của hệ thống.

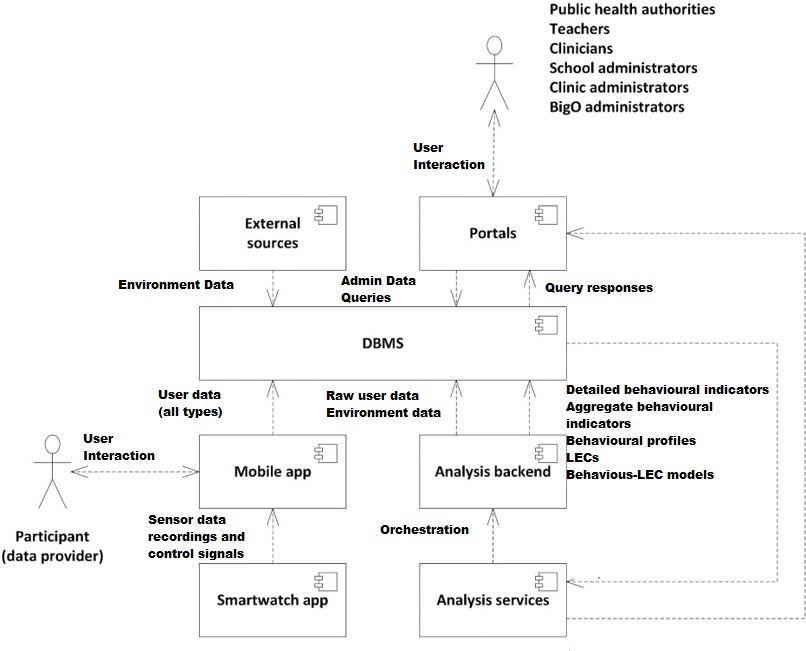
Tóm lại, từ góc độ kỹ thuật, BigO đo lường và nghiên cứu hai yếu tố chính: (1) điều kiện bên ngoài cục bộ và (2) mô hình hành vi cá nhân. Dữ liệu thu thập được từ các yếu tố này được diễn giải bởi người dùng BigO được liệt kê bên dưới:

1. Trẻ em và thanh thiếu niên trong độ tuổi (9–18 tuổi) là nhà cung cấp dữ liệu như:
   * Học sinh của trường, thông qua các nỗ lực có tổ chức của trường về các dự án xung quanh hoạt động thể chất, ăn và ngủ.
   * Bệnh nhân tham gia các phòng khám béo phì.
   * Cá nhân tình nguyện viên.
2. Giáo viên điều hành các nỗ lực tổ chức của trường với học sinh.
3. Bác sĩ lâm sàng điều trị bệnh nhân tại phòng khám.
4. Cán bộ Y tế Công cộng (nhà nghiên cứu hoặc nhà hoạch định chính sách) đánh giá các chỉ số hành vi của trẻ em / thanh thiếu niên trong một khu vực địa lý kết hợp các Điều kiện bên ngoài địa phương (LEC) có liên quan đến béo phì.
5. Quản trị viên cho trường học, phòng khám và toàn bộ nền tảng Bigo.



**Hình 2.** Tổng quan về hệ thống Bigo. (1) Nhà khoa học công dân. (2) Trẻ em được theo dõi tại các phòng khám, thông qua đồng hồ thông minh và ứng dụng di động. (3) Truyền dữ liệu ẩn danh và mã hóa. (4) Nguồn dữ liệu ngoài (bản đồ, POI, thống kê khu vực). (5) Tổng hợp và xử lý dữ liệu đám mây BigO. (6) Thư viện và công cụ phân tích và trực quan hóa dữ liệu. (7) Lưu trữ cơ sở dữ liệu phân tán an toàn. (8) Dịch vụ tư vấn chính sách. (9) Dịch vụ hoạch định chính sách. (10) Dịch vụ cố vấn lâm sàng và trường học. (11) Một bác sĩ lâm sàng sử dụng các công cụ web để theo dõi và hướng dẫn trẻ em. (12) Công cụ web để hỗ trợ quyết định của nhà hoạch định chính sách. (13) Các nhà hoạch định chính sách xác định tình trạng béo phì ở trẻ em. (14, 15, 16) Các chính sách được áp dụng cho bệnh viện, trường học và cấp cộng đồng hoặc khu vực. (17) Các chính sách được áp dụng ảnh hưởng đến các nhà khoa học công dân, đóng vòng lặp và bắt đầu một vòng thu thập và phân tích dữ liệu khác.

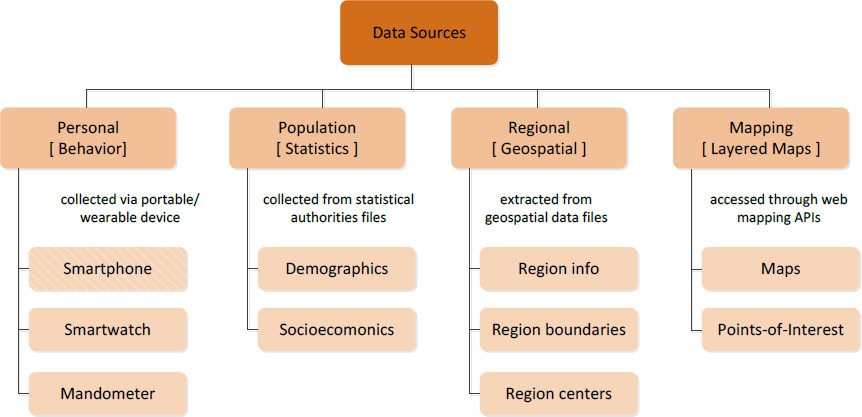
Hình [3](#_bookmark6) tóm tắt luồng dữ liệu dựa trên các nhóm người dùng nói trên. Ban đầu, giáo viên và bác sĩ lâm sàng chèn dữ liệu vào hệ thống (đăng ký người dùng ban đầu). Trẻ em, cả chủ động (bằng cách ghi lại quảng cáo thực phẩm hoặc bữa ăn) và thụ động (bằng cách phát hiện chuyển động tự động) chia sẻ dữ liệu của chúng bằng cách gửi nó vào hệ thống. Tất cả dữ liệu này được tổ chức và thể hiện trong Hệ thống quản lý cơ sở dữ liệu (DBMS) và được tải lên các máy chủ, nơi nó có thể được xem, xử lý và phân tích ngay lập tức trong các cổng thông tin.



**Hình 3.** Luồng thông tin BiGo.

# Thu thập dữ liệu BigO

Hình [4](#_bookmark8) cho thấy bốn loại dữ liệu chính được thu thập bởi BigO; (1) nguồn dữ liệu mỗi âm thanh hoặc hành vi, (2) Nguồn dữ liệu dân số, (3) Nguồn dữ liệu khu vực và (4) Nguồn dữ liệu bản đồ. Những điều này được thảo luận ngắn gọn trong phần sau.



**Hình 4.** Phân loại phân cấp các nguồn dữ liệu thô cần thiết của BigO. Điện thoại thông minh được phân biệt trực quan với anh chị em của nó, vì nó là nguồn dữ liệu thô lai (do ảnh quảng cáo thực phẩm).

1. Nguồn dữ liệu cá nhân (Hành vi): Dữ liệu thô này được thu thập ở cấp độ cá nhân từ các nhà khoa học công dân liên quan đến các mẫu hành vi có liên quan đến nghiên cứu BigO (ví dụ: cách một người di chuyển, ăn, ngủ). Dữ liệu thô trong danh mục này được thu thập từ các thiết bị di động và / hoặc thiết bị đeo được cá nhân. Các nguồn này được phân loại thêm theo thiết bị thu thập dữ liệu cảm giác di động; (tức là,

(a) Điện thoại thông minh, (b) Đồng hồ thông minh và (c) Mandometer).

Chúng tôi kết hợp các thiết bị trong ba cài đặt dựa trên các yêu cầu của hệ thống BigO đòi hỏi phải thu thập dữ liệu và tính khả dụng của các cảm biến ngoại vi (Bảng [2](#_bookmark9)).

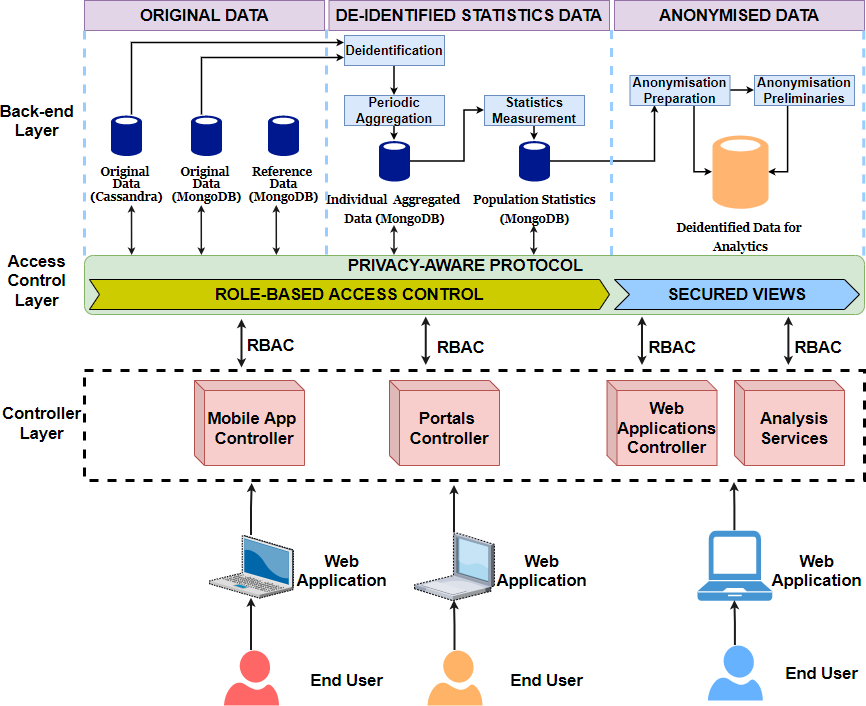
1. Nguồn dữ liệu dân cư (Thống kê): Các nguồn dữ liệu thô chứa thông tin về dân cư cư trú tại một khu vực nhất định (nông thôn, thành phố, v.v.). Các nhà cung cấp dữ liệu bao gồm các cơ quan thống kê quốc gia của các quốc gia liên quan đến BigO. Tùy thuộc vào loại thống kê dân số, các nguồn dữ liệu thô được phân loại thêm là (a) Nguồn dữ liệu nhân khẩu học và (b) Các nguồn dữ liệu thô kinh tế xã hội liên quan đến dân số của một thành phố quan tâm và các khu vực hành chính của họ.
2. Nguồn dữ liệu khu vực (Không gian địa lý): Các nguồn này kết hợp dữ liệu không gian địa lý được liên kết với các khu vực quan tâm của BigO (quốc gia, thành phố hoặc khu vực thành phố hành chính).
3. Ánh xạ nguồn dữ liệu (Layered Maps): Dữ liệu ánh xạ web được cung cấp bởi API của bên thứ 3. Tùy thuộc vào loại dữ liệu, các nguồn này được phân loại thêm thành a) Bản đồ (tức là bản đồ địa hình tương tác) và (b) Điểm ưa thích (PoIs).

**Bảng 2.** Các chế độ thu thập dữ liệu được triển khai trong hệ thống Bigo. Tất cả các chế độ bao gồm việc sử dụng điện thoại thông minh. Tên viết tắt: Điện thoại thông minh (SP), Vòng đeo tay (WB) và Mandometer (MM).

|  |  |  |  |
| --- | --- | --- | --- |
| **Kiểu dữ liệu được thu thập** | **Ánh sáng** | **Chuẩn** | **Nâng cao** |
| Gia tốc đo | SP | SP / WB | SP / WB |
| GPS | SP | WB (có GPS) | WB |
| Tự báo cáo bữa ăn | SP | SP | SP |
| Quét mã vạch thực phẩm | SP | SP | SP |
| Hình ảnh thực phẩm | SP | SP | SP |
| Hành vi ăn uống trong bữa ăn | - | Sử dụng MM hạn chế | Sử dụng MM mở rộng |

# Kiến trúc kho dữ liệu lớn

Phần này trình bày kiến trúc kho dữ liệu lớn và các phương pháp biểu diễn dữ liệu được triển khai trong BigO có tính đến bảo mật và quyền riêng tư của dữ liệu. Kiến trúc kho dữ liệu BigO (DW) được thể hiện trong Hình [5](#_bookmark11). DW bao gồm các hệ thống lưu trữ khác nhau tùy thuộc vào độ nhạy và loại truy cập theo yêu cầu của dữ liệu ở bất kỳ giai đoạn nào trong vòng đời của nó trong hệ thống Bigo. Nó được thiết kế theo kiến trúc ba tầng. Ba tầng (hoặc các lớp trong trường hợp này) là lớp back-end, lớp kiểm soát truy cập và lớp điều khiển.



**Hình 5.** Kiến trúc kho dữ liệu Bigo.

* Dữ liệu gốc: Trong hệ thống BigO, dữ liệu thô được lưu trữ theo hai lược đồ chính. Các lược đồ này được thực hiện bằng cách sử dụng hai cơ sở dữ liệu hiện đại; MongoDB và Cas- sandra. Chúng tôi đã tạo ra hai lược đồ vì hai lý do khác nhau: thứ nhất là vì lý do bảo mật và quyền riêng tư. Tách làm giảm nguy cơ vi phạm truy cập dữ liệu và thao túng có chủ ý. Thứ hai là hai lược đồ không được sử dụng theo cùng một cách trong quá trình phân tích và từ các vai trò khác nhau. Ví dụ, Cassandra được sử dụng để lưu trữ dữ liệu chuỗi thời gian, trong khi MongoDB được sử dụng cho phần còn lại. Hơn nữa, các nguồn dữ liệu bên ngoài cũng được sử dụng trong quá trình phân tích, bao gồm cơ sở dữ liệu thẩm quyền và cơ sở dữ liệu thống kê quốc gia.
* Dữ liệu thống kê đã hủy nhận dạng: Tất cả thông tin về danh tính của người dùng sẽ bị xóa. Điều này bao gồm dữ liệu tổng hợp riêng lẻ và thống kê dân số, tức là dữ liệu thống kê có nguồn gốc và được tính toán từ dữ liệu gốc và dữ liệu tham chiếu.
* Dữ liệu ẩn danh: Ngoài việc hủy nhận dạng, chúng tôi còn ẩn danh dữ liệu để dữ liệu gốc không thể được khôi phục từ kiến thức được trích xuất trong quá trình phân tích dữ liệu, sử dụng các thuật toán khai thác dữ liệu.
  1. *Lớp kiểm soát truy cập*

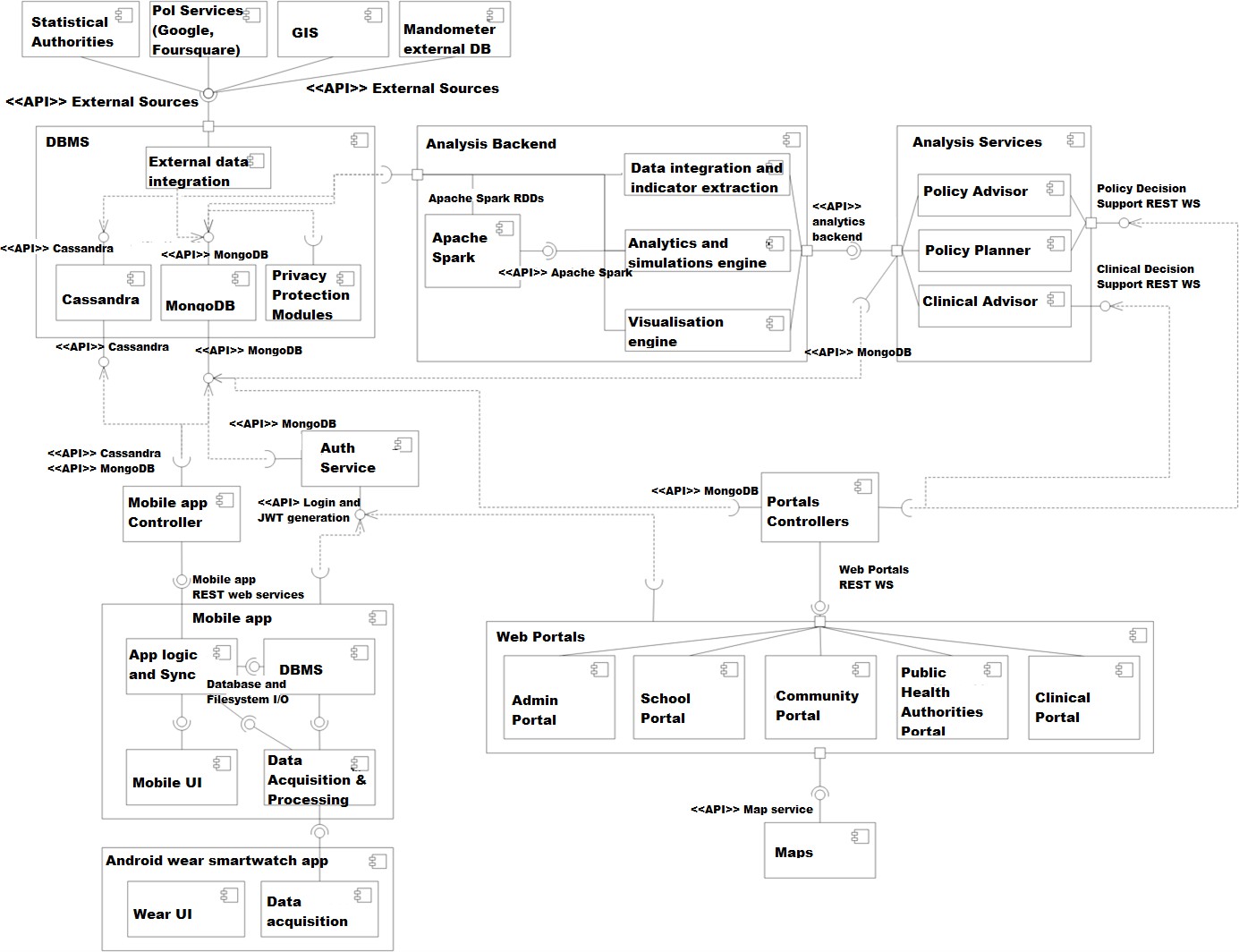
Lớp này bao gồm kiểm soát truy cập dựa trên vai trò và chế độ xem được bảo mật. Kiểm soát truy cập dựa trên vai trò kiểm tra quyền của bộ điều khiển khi họ yêu cầu truy cập dữ liệu gốc và dữ liệu thống kê không xác định dựa trên vai trò của họ. Kiểm soát truy cập dựa trên vai trò chỉ nhận yêu cầu dữ liệu từ bộ điều khiển ứng dụng dành cho thiết bị di động hoặc bộ điều khiển cổng thông tin. Trong khi đó, các chế độ xem bảo mật kiểm soát kiến thức được trích xuất từ dữ liệu ẩn danh. Các chế độ xem bảo mật chỉ nhận các yêu cầu dữ liệu từ các ứng dụng web, bộ điều khiển và dịch vụ phân tích.

* 1. *Lớp điều khiển*

Có bốn loại bộ điều khiển, đó là bộ điều khiển ứng dụng dành cho thiết bị di động, bộ điều khiển cổng thông tin, bộ điều khiển ứng dụng web và dịch vụ phân tích. Lớp này nhận các yêu cầu dữ liệu từ người dùng hoặc các thành phần khác của hệ thống. Họ kiểm tra quyền của người dùng, sử dụng mã thông báo và vai trò của các thành phần, bất cứ khi nào họ nhận được yêu cầu dữ liệu từ người dùng.

* 1. *Lưu trữ và tích hợp dữ liệu*

Hệ thống lưu trữ được xác định để triển khai kho dữ liệu BigO sử dụng MongoDB và Cassandra. MongoDB là cơ sở dữ liệu ứng dụng chính, trong khi Cassandra được sử dụng để lưu trữ khối lượng lớn dữ liệu chuỗi thời gian. Bộ điều khiển di động, cổng Web và ứng dụng Web truy cập trực tiếp vào hệ thống lưu trữ bằng văn bản chỉ thông qua API REST. Hơn nữa, các API này thực hiện các yêu cầu tích hợp và hợp nhất dữ liệu để dữ liệu, được thu thập từ các nguồn khác nhau, được lưu trữ trong một đại diện thống nhất trong kho dữ liệu. Hình [6](#_bookmark12) cho thấy luồng dữ liệu giữa các mô-đun hệ thống khác nhau.



**Hình 6.** Sơ đồ thành phần của hệ thống BigO.

Ngoài các cơ sở dữ liệu cốt lõi, BigO cung cấp quyền truy cập vào hệ thống tệp phân tán là HDFS của Hadoop. Mục đích của nó là giữ các tệp tập dữ liệu lớn có thể được trích xuất trực tiếp từ cơ sở dữ liệu, ở định dạng thuận tiện, có thể được sử dụng bởi các thuật toán phân tích dữ liệu (mô-đun). Một quy trình định kỳ cập nhật hàng ngày nội dung tệp từ cơ sở dữ liệu theo một khoảng thời gian đều đặn. Phụ lục [A.1](#_bookmark22) và [A.2](#_bookmark25) trình bày lược đồ chi tiết

từ cơ sở dữ liệu cốt lõi. Sau đây, chúng tôi thảo luận về dữ liệu được lưu trữ trong mỗi cơ sở dữ liệu cốt lõi.

* + 1. Dữ liệu MongoDB

Dữ liệu được lưu trữ trong MongoDB được thu thập từ bốn nguồn khác nhau, được mô tả dưới đây:

* + - * Bộ điều khiển cổng thông tin: Có năm cổng thông tin trong hệ thống BigO; Cổng thông tin quản trị, cổng thông tin trường học, cổng thông tin cộng đồng, cổng thông tin cơ quan y tế công cộng và cổng thông tin lâm sàng. Dữ liệu được thu thập từ tất cả các cổng này và được tích hợp trước khi lưu trữ cơ sở dữ liệu MongoDB.
      * Bộ điều khiển di động: Bộ điều khiển này xử lý dữ liệu được thu thập từ điện thoại di động hoặc đồng hồ thông minh. Dữ liệu được xử lý trước, tích hợp và sau đó được chuyển đến MongoDB. Dữ liệu này có tầm quan trọng đặc biệt, vì nó nhạy cảm và cần được xử lý cẩn thận.
      * Dịch vụ phân tích: Thành phần phân tích dữ liệu xử lý toàn bộ Bộ dữ liệu được lưu trữ trong hai DB. Thành phần phân tích sử dụng môi trường điện toán Spark. Kết quả phân tích được lưu trữ lại tại cơ sở dữ liệu MongoDB.
      * Dịch vụ phân tích back-end: Dịch vụ này truy cập dữ liệu được lưu trữ trong cả hai DB để trích xuất các chỉ số hành vi. Các thuộc tính dữ liệu mới này sau đó được lưu trữ trong MongoDB. Dịch vụ phân tích back-end được thực thi trong môi trường Spark.
    1. Dữ liệu Cassandra

DB này chủ yếu lưu trữ dữ liệu thô được thu thập từ các nguồn bên ngoài và ứng dụng dành cho thiết bị di động.

* + - * Nguồn bên ngoài: Các nguồn dữ liệu bên ngoài bao gồm dữ liệu hành vi cá nhân và dữ liệu dân số bên ngoài. Các bộ dữ liệu này được truyền trực tiếp đến cơ sở dữ liệu thông qua các mô-đun tích hợp dữ liệu ngoài (Hình [6](#_bookmark12)). Ví dụ: dữ liệu về thiết bị của từng cá nhân.
      * Bộ điều khiển ứng dụng dành cho thiết bị di động: Dữ liệu từ các ứng dụng di động được thu thập thông qua điện thoại thông minh hoặc đồng hồ thông minh được lưu trữ bằng nhà cung cấp nội dung ứng dụng dành cho thiết bị di động trong DB của ứng dụng dành cho thiết bị di động. Dữ liệu được lưu trữ sau đó được đồng bộ hóa với Cassandra DB. (Hình [6](#_bookmark12)).

# Bảo mật dữ liệu và quyền riêng tư

Quyền riêng tư và bảo mật là những yêu cầu cơ bản đối với bất kỳ hệ thống thông tin nào. Mặc dù bảo mật dữ liệu và quyền riêng tư dữ liệu có liên quan với nhau, nhưng chúng giải quyết các vấn đề khác nhau và cần các biện pháp đối phó khác nhau. Bảo mật dữ liệu nhằm mục đích bảo vệ tính bảo mật, tính toàn vẹn, tính khả dụng và không từ chối. Trong khi quyền riêng tư dữ liệu ngăn chặn dữ liệu được chia sẻ tiết lộ thông tin nhạy cảm về chủ sở hữu tương ứng của chúng. Các phương pháp bảo mật dữ liệu sau đây đã được triển khai trong BigO.

* Lưu trữ an toàn: Kiến trúc hệ thống BigO được triển khai bằng các nền tảng chính thống, chẳng hạn như Cassandra, MongoDB, hệ thống quản lý cơ sở dữ liệu SQLite, hệ thống tệp HDFS của Hadoop, Android và iOS. Các hệ thống tiêu chuẩn này sử dụng các công nghệ mã hóa tích hợp để bảo mật cả các thành phần hệ thống và dữ liệu mà chúng chứa. Phần [6.1](#_bookmark14) Cung cấp thêm chi tiết về từng loại lưu trữ.
* Truyền thông an toàn: Dữ liệu phải được bảo mật khi được chuyển giữa các mô-đun khác nhau của hệ thống. Giao tiếp giữa các mô-đun BigO sử dụng các giao thức bảo mật, chẳng hạn như SSL, TLS hoặc HTTPS (Xem phần [6.1](#_bookmark14) để biết thêm chi tiết).
* Kiểm soát truy cập dữ liệu: Kiểm soát truy cập là một vấn đề phức tạp trong các hệ thống lớn, chẳng hạn như BigO. Do đó, chúng tôi đã triển khai một giải pháp đầy đủ để kiểm soát truy cập bao gồm các chính sách nhất quán, cơ chế rõ ràng và mạnh mẽ để đăng ký, xác thực và ủy quyền. Chúng được tóm tắt như sau:
  + Lưu trữ ứng dụng dành cho thiết bị di động: Trong BigO, bộ lưu trữ dữ liệu ứng dụng dành cho thiết bị di động chỉ có thể được truy cập bằng mặt sau của điện thoại di động đó và chỉ chứa dữ liệu cá nhân của chủ sở hữu. Do đó, không cần kiểm soát truy cập chi tiết trong bối cảnh này. Nó sử dụng một điều khiển truy cập đơn giản dựa trên tên người dùng và mật khẩu.
  + Lưu trữ tệp phụ trợ: Đây là bộ lưu trữ tạm thời để xử lý dữ liệu thô và được sử dụng bởi back-end di động thông qua bộ điều khiển của nó. Giống như lưu trữ ứng dụng dành cho thiết bị di động, nó chỉ cần kiểm soát truy cập cơ bản với tên người dùng và mật khẩu.
  + Máy chủ cơ sở dữ liệu: Đây là bộ lưu trữ chính và chứa tất cả dữ liệu BigO để phân tích. Dữ liệu có sẵn cho nhiều người dùng cuối tùy thuộc vào vai trò của họ. Cơ chế Kiểm soát truy cập dựa trên vai trò (RAC) được sử dụng với các chính sách cụ thể để cấp vai trò và quyền. Mỗi cơ sở dữ liệu được truy cập thông qua một API RESTful. (Mục [6.2](#_bookmark15) để biết chi tiết).

Các biện pháp bảo vệ quyền riêng tư dữ liệu của BigO dựa trên loại dữ liệu cá nhân (Bảng [2](#_bookmark9)).

Các kiểu dữ liệu này được đưa ra dưới đây:

* Dữ liệu cảm biến quán tính, chuyển động và Mandometer: Các loại dữ liệu này được thu thập và lưu trữ trên các thiết bị cá nhân (đồng hồ thông minh, điện thoại di động, Mandometers). Chỉ dữ liệu thống kê và tổng quát mới được trích xuất và gửi đến máy chủ BiGo. Những loại dữ liệu này không làm tăng rủi ro về quyền riêng tư.
* Hình ảnh: Tất cả các bức ảnh được xem xét bởi quản trị viên BigO. Bất kỳ ảnh nào được coi là không liên quan, không đứng đắn hoặc tiết lộ danh tính của người dùng sẽ bị xóa khỏi hệ thống.
* Dữ liệu có thể nhận dạng: Tất cả các thuộc tính, chẳng hạn như tên người dùng, deviceID, v.v., sẽ bị xóa trước khi lưu trữ dữ liệu trong kho dữ liệu. Hoạt động xóa các thuộc tính đó được gọi là hủy nhận dạng (Phần [6.3](#_bookmark16)).
* Dữ liệu gần như có thể nhận dạng: Tất cả các thuộc tính dữ liệu, chẳng hạn như quốc gia, khu vực, trường học, phòng khám, chiều cao, cân nặng, giới tính, năm sinh, câu trả lời tự đánh giá, vị trí của ảnh, v.v., đều được xử lý bằng cách ẩn danh. Tuy nhiên, do những tác động xấu đến chất lượng dữ liệu, một giao thức nhận thức về quyền riêng tư được áp dụng để tính đến loại dữ liệu này (Phần [6.3](#_bookmark16)).
  1. *Bảo vệ dữ liệu*

Ứng dụng dành cho thiết bị di động: Ứng dụng dành cho thiết bị di động được xây dựng cho cả Android và iOS. Cả hai hệ điều hành đều có các tính năng bảo mật tích hợp. Chúng tôi khai thác các tính năng này để bảo vệ dữ liệu do ứng dụng thu thập, ngay cả trong trường hợp sử dụng quyền hệ thống và tệp mặc định [[17](#_bookmark43)]. Ví dụ, các tính năng bảo mật cốt lõi của hệ điều hành Android bao gồm Android Application Sandbox cách ly dữ liệu ứng dụng và thực thi mã với các ứng dụng khác. Các tính năng tương tự cũng được cung cấp trong iOS.

Ứng dụng BigO lưu trữ tất cả dữ liệu (bao gồm dữ liệu CSDL SQLite, ảnh thu được trong ứng dụng và tùy chọn được chia sẻ) trong bộ nhớ trong của thiết bị. Tất cả dữ liệu đó sẽ bị xóa ngay khi nó được đồng bộ hóa / tải lên máy chủ BigO.

Máy chủ cơ sở dữ liệu: Hệ thống MongoDB và Cassandra kết hợp các mô hình bảo mật rất thú vị. Ví dụ, mô hình bảo mật MongoDB được chia thành bốn phần chính: xác thực, ủy quyền, kiểm tra và mã hóa. Chúng tôi giải thích ngắn gọn cách các tính năng bảo vệ bảo mật dữ liệu được MongoDB và Cassandra hỗ trợ như sau.

* Xác thực: MongoDB tích hợp các cơ chế bảo mật bên ngoài bao gồm Giao thức truy cập thư mục nhẹ (LDAP) [[40](#_bookmark66)], Windows Active Directory, Kerberos [[41](#_bookmark67)] và

chứng chỉ x.509 PKI [[42](#_bookmark68)] để thực thi lại kiểm soát truy cập vào cơ sở dữ liệu.

* Ủy quyền: Vai trò do người dùng xác định có thể được xác định trong MongoDB để định cấu hình các quyền chi tiết cho người dùng hoặc ứng dụng dựa trên các đặc quyền họ cần. Hơn nữa, người ta có thể xác định các chế độ xem chỉ hiển thị một tập hợp con dữ liệu từ một bộ sưu tập nhất định.
* Kiểm toán: Để tuân thủ quy định, mô hình bảo mật MongoDB ghi lại nhật ký kiểm toán gốc để theo dõi quyền truy cập và các hoạt động được thực hiện đối với cơ sở dữ liệu.
* Mã hóa: Hệ thống bảo mật MongoDB cung cấp dữ liệu mã hóa dữ liệu trên mạng, đĩa và sao lưu. Bằng cách mã hóa các tệp cơ sở dữ liệu trên đĩa, người ta loại bỏ cả chi phí quản lý và hiệu suất của các cơ chế mã hóa bên ngoài.
* Giám sát và sao lưu: MongoDB cung cấp nhiều công cụ khác nhau, bao gồm Mongostat, Mongotop và MongoDB Management Service (MMS) để giám sát cơ sở dữ liệu.

Các đỉnh đột ngột trong tải CPU và bộ nhớ của hệ thống máy chủ và các bộ đếm hoạt động cao trong cơ sở dữ liệu có thể chỉ ra một cuộc tấn công từ chối dịch vụ.

Apache Cassandra là một hệ thống cơ sở dữ liệu NoSQL và nó không dựa trên một kiến trúc được chia sẻ, chẳng hạn như MongoDB. Nó dựa vào DataStax Enterprise (DSE) [[43](#_bookmark69)] để cung cấp các tính năng bảo mật như Mã hóa dữ liệu trên chuyến bay và trên thiết bị lưu trữ, Xác thực, Ủy quyền và Kiểm tra dữ liệu. DataStax Enterprise (DSE) [[43](#_bookmark69)] tích hợp gắn kết với bất động sản công nghệ hiện có, bao gồm hỗ trợ cho Active Directory (AD), Giao thức truy cập thư mục nhẹ (LDAP), Kerberos, Cơ sở hạ tầng khóa công khai (PKI) và Giao thức tương tác quản lý khóa (KMIP). Chúng được giải thích dưới đây.

* Mã hóa: Nó duy trì tính bảo mật dữ liệu. Thông thường, mã hóa dữ liệu DB rơi vào hai loại: Mã hóa trên thiết bị lưu trữ và Mã hóa trên chuyến bay. Đầu tiên đề cập đến việc bảo vệ dữ liệu được lưu trữ trên bộ nhớ liên tục. Thứ hai đề cập đến mã hóa dữ liệu khi nó di chuyển qua mạng giữa các nút hoặc máy khách và các nút trong cụm DSE.
* Mã hóa dữ liệu minh bạch DSE (TDE): là tính năng chịu trách nhiệm mã hóa dữ liệu lưu trữ trong hệ thống DSE. DSE TDE bảo vệ dữ liệu nhạy cảm ở trạng thái lưu trữ bằng cách sử dụng tệp khóa mã hóa cục bộ hoặc khóa mã hóa Giao thức tương tác quản lý khóa (KMIP) được lưu trữ và quản lý từ xa.
* Xác thực: đề cập đến quá trình thiết lập danh tính của người hoặc hệ thống thực hiện một hoạt động đối với cơ sở dữ liệu. Xác thực hợp nhất DSE tạo điều kiện kết nối với bốn cơ chế chính để xác thực, như được mô tả bên dưới. Nó mở rộng các lược đồ xác thực tương tự cho cơ sở dữ liệu, DSE Search và DSE Analytics.
* Ủy quyền: Trong DSE, ủy quyền xác định tài nguyên nào (ví dụ: bảng, không gian khóa, v.v.) có thể được đọc, viết hoặc sửa đổi bởi một thực thể được kết nối, cũng như các cơ chế kết nối của chúng. Nó sử dụng mô hình GRANT / REVOKE để ủy quyền để ngăn chặn mọi truy cập không đúng vào dữ liệu và sử dụng ba cơ chế để ủy quyền cho người dùng: Kiểm soát truy cập dựa trên vai trò (RBAC), Kiểm soát truy cập cấp hàng (RLAC), Xác thực proxy.
* Kiểm toán: Kiểm toán dữ liệu cho phép theo dõi và ghi lại tất cả các hoạt động của người dùng được thực hiện trên cơ sở dữ liệu để ngăn chặn truy cập trái phép vào thông tin và đáp ứng các yêu cầu tuân thủ. Với DSE, tất cả hoặc một tập hợp con của một hoạt động diễn ra trên cụm DataStax được ghi lại cùng với danh tính của người dùng và thời gian hoạt động được thực hiện. Kiểm toán hiệu quả trong DSE được thực hiện thông qua cơ chế log4J được tích hợp vào nền tảng.
* Trình điều khiển: DataStax cung cấp trình điều khiển cho C / C ++, C #, Java, Nodejs, ODBC, Python, PHP và Ruby hoạt động với mọi kích thước cụm cho dù được triển khai tại chỗ hay trung tâm dữ liệu đám mây. Các trình điều khiển này được cấu hình với một số tính năng, chẳng hạn như SSL để đảm bảo người dùng tương tác với các cụm DSE một cách an toàn và bảo mật.
  + 1. Lưu trữ tệp phụ trợ

Lưu trữ tệp phụ trợ được sử dụng trong BigO để lưu trữ dữ liệu gia tốc kế thô và con quay hồi chuyển cho mục đích phát triển. Dữ liệu đo lường quán tính như vậy không thể được sử dụng để xác định người dùng và không gây ra bất kỳ rủi ro bảo mật nào. Dữ liệu được lưu trữ ở định dạng nhị phân tùy chỉnh trong một máy chủ Unix an toàn.

* + 1. Truyền dữ liệu

Dữ liệu được truyền giữa (a) đồng hồ thông minh và điện thoại di động, (b) điện thoại di động của người dùng và máy chủ BigO và (c) giữa các máy chủ khác nhau của BigO. Giao tiếp giữa đồng hồ thông minh và điện thoại di động diễn ra qua kênh Bluetooth được mã hóa và kẻ tấn công tiềm năng phải ở gần người dùng. Hơn nữa, dữ liệu chỉ có thể truy cập được bằng ứng dụng BigO trên điện thoại di động; Tất cả các ứng dụng khác đang chạy trên điện thoại không thể nhận được dữ liệu được truyền. Giao tiếp giữa điện thoại di động và

Máy chủ BigO được mã hóa bằng SSL 2048-bit. Cuối cùng, do tất cả các máy chủ BigO là một phần của cùng một trung tâm dữ liệu, việc truyền tải giữa các máy chủ là không quan trọng.

* 1. *Kiểm soát truy cập dữ liệu*
     1. Đăng ký

Thủ tục đăng ký phụ thuộc vào loại người dùng. Các thủ tục khác nhau là:

* + - * Quản trị viên Bigo: Điều này được tạo ra bởi các nhà phát triển BigO và nó đã được sửa. Một người quản lý quảng cáo BigO có thể đăng ký quản trị viên trường học và phòng khám. Tương tự cũng có thể xem xét các hình ảnh đã gửi để xóa hình ảnh không phù hợp hoặc hình ảnh xâm phạm quyền riêng tư của cá nhân.
      * Quản trị viên trường học có thể thêm/chỉnh sửa chi tiết trường học và đăng ký giáo viên.
      * Quản trị viên phòng khám có thể thêm/chỉnh sửa chi tiết phòng khám và đăng ký bác sĩ lâm sàng.
      * Giáo viên có thể tạo nhóm, chỉnh sửa nhóm học sinh và chi tiết cá nhân học sinh, chẳng hạn như BMI, lịch tập thể dục ở trường, v.v. và có thể tạo mã đăng ký cho học sinh.
      * Các bác sĩ lâm sàng có thể tạo mã đăng ký cho bệnh nhân và chỉnh sửa chi tiết từng bệnh nhân, chẳng hạn như BMI.
      * Học sinh có thể đăng ký thông qua ứng dụng di động BigO bằng mã đăng ký do giáo viên cung cấp. Khi giáo viên tạo tài khoản học sinh, mã đăng ký được tạo và lưu trữ trong cơ sở dữ liệu. Học sinh nhập mã đăng ký lần đầu tiên sử dụng ứng dụng. Sau khi mã đăng ký được "đổi", học sinh được đăng ký trong hệ thống và mã đăng ký không còn hiệu lực.
      * Bệnh nhân đăng ký với mã đăng ký do bác sĩ lâm sàng của họ cung cấp; giống như học sinh.
    1. Xác thực:

Xác thực người dùng diễn ra trên một máy chủ xác thực chuyên dụng thông qua JSON Web Tokens. Quá trình này như sau:

* + - * Mỗi người dùng có một tên người dùng và mật khẩu (đối với sinh viên và bệnh nhân, chúng được tạo tự động và lưu trữ trên điện thoại di động mà không có sự tham gia của người dùng). Mật khẩu được muối và băm, và mật khẩu được mã hóa được lưu trữ trong cơ sở dữ liệu.
      * Khi điện thoại di động cần truy cập điểm cuối REST bị hạn chế, trước tiên nó sẽ yêu cầu JSON Web Token (JWT) từ máy chủ xác thực, bằng cách trình bày thông tin đăng nhập của người dùng. Các thông tin đăng nhập cũng được muối (với cùng một muối) và băm, và máy chủ xác thực so sánh các mật khẩu được mã hóa. Nếu chúng khớp, nó cung cấp cho người dùng một JWT hợp lệ.
      * Sử dụng JWT, ứng dụng dành cho thiết bị di động và các cổng / ứng dụng web có thể truy cập các điểm cuối REST bị hạn chế, cho đến khi hết hạn. Sau khi hết hạn, ứng dụng dành cho thiết bị di động yêu cầu JWT mới từ máy chủ xác thực và quá trình này được lặp lại.
    1. Phép:

Các vai trò của BigO bao gồm quản trị viên Bigo, quản trị viên trường học, quản trị viên phòng khám, giáo viên, bác sĩ lâm sàng, học sinh và bệnh nhân, vai trò của cơ quan y tế công cộng và sinh viên tình nguyện. Kiểm soát truy cập được thực hiện ở lớp ứng dụng bởi các con- troller BigO (Bộ điều khiển ứng dụng di động, bộ điều khiển cổng thông tin, bộ điều khiển ứng dụng web). Chúng tôi cũng bao gồm kiểm soát truy cập cấp bộ sưu tập dựa trên vai trò ở lớp cơ sở dữ liệu (được hỗ trợ nguyên bản trong MongoDB), như cơ chế bảo mật và bảo vệ dữ liệu bổ sung.

* 1. *Bảo vệ quyền riêng tư dữ liệu*

Trong BigO, dữ liệu được lưu trữ trong bộ lưu trữ ứng dụng dành cho thiết bị di động (bao gồm hệ thống tệp di động và cơ sở dữ liệu SQLite), lưu trữ tệp phụ trợ, cơ sở dữ liệu Cassandra và cơ sở dữ liệu MongoDB. Ngoại trừ dữ liệu trong cơ sở dữ liệu MongoDB, dữ liệu ở những nơi khác được truy cập và sử dụng bởi chủ sở hữu dữ liệu và các mô-đun nội bộ. Cơ sở dữ liệu MongoDB hiếm khi được chia sẻ trực tiếp để phân tích. Điều này nhằm ngăn chặn dữ liệu được chia sẻ tiết lộ thông tin nhạy cảm của chủ sở hữu dữ liệu. Do đó, phần này tập trung vào việc bảo vệ quyền riêng tư của cơ sở dữ liệu MongoDB khi dữ liệu này được chia sẻ cho mục đích phân tích.

Trước tiên, chúng tôi thảo luận về những thách thức của việc bảo vệ quyền riêng tư trong các phần tiếp theo. Sau đó, chúng tôi thảo luận thêm về cách giảm thiểu tác động của những thách thức này với giao thức nhận thức về quyền riêng tư.

* + 1. Hủy nhận dạng và bút danh

Việc hủy nhận dạng che giấu danh tính thực của chủ sở hữu dữ liệu bằng cách xóa tất cả các trường có thể trực tiếp nhận dạng một cá nhân, chẳng hạn như tên, số điện thoại và email. Trong trường hợp của chúng tôi, việc hủy nhận dạng không ảnh hưởng đến chất lượng kết quả phân tích dữ liệu.

Khi dữ liệu được hủy nhận dạng và chia sẻ, một mã định danh ngẫu nhiên mới được sử dụng để đặt tên cho các cá nhân. Trong một số trường hợp, cần phải giữ mối quan hệ giữa số nhận dạng cũ và mới để cập nhật dữ liệu đã hủy nhận dạng. Do đó, bút danh được sử dụng để tạo ID giả cho các cá nhân và mã hóa mối liên hệ giữa ID thật và giả. Tuy nhiên, trong hệ thống BiGo, "tên người dùng" và "display\_id" được sử dụng trong các giao diện với người dùng cuối trong khi ID con chỉ được sử dụng bởi các mô-đun bên trong. Do đó, chúng tôi không cần tạo ID giả cho trẻ em khi dữ liệu của chúng được chia sẻ. ID của trẻ em có thể được lưu giữ để phân biệt các cá nhân trong tập dữ liệu được chia sẻ mà không tiết lộ danh tính của chúng.

Ngoài ra, các trường không cần thiết để phân tích có thể được loại bỏ cùng một lúc. Điều quan trọng là phải làm rõ rằng bộ sưu tập cơ sở dữ liệu "Ảnh" không được coi là nguồn thông tin có thể nhận dạng. Điều này là do ảnh đã tải lên được kiểm tra bởi các thuật toán nhận dạng khuôn mặt và quản trị viên BigO để đảm bảo rằng chúng chỉ chứa quảng cáo thực phẩm và bữa ăn.

* + 1. Ẩn danh

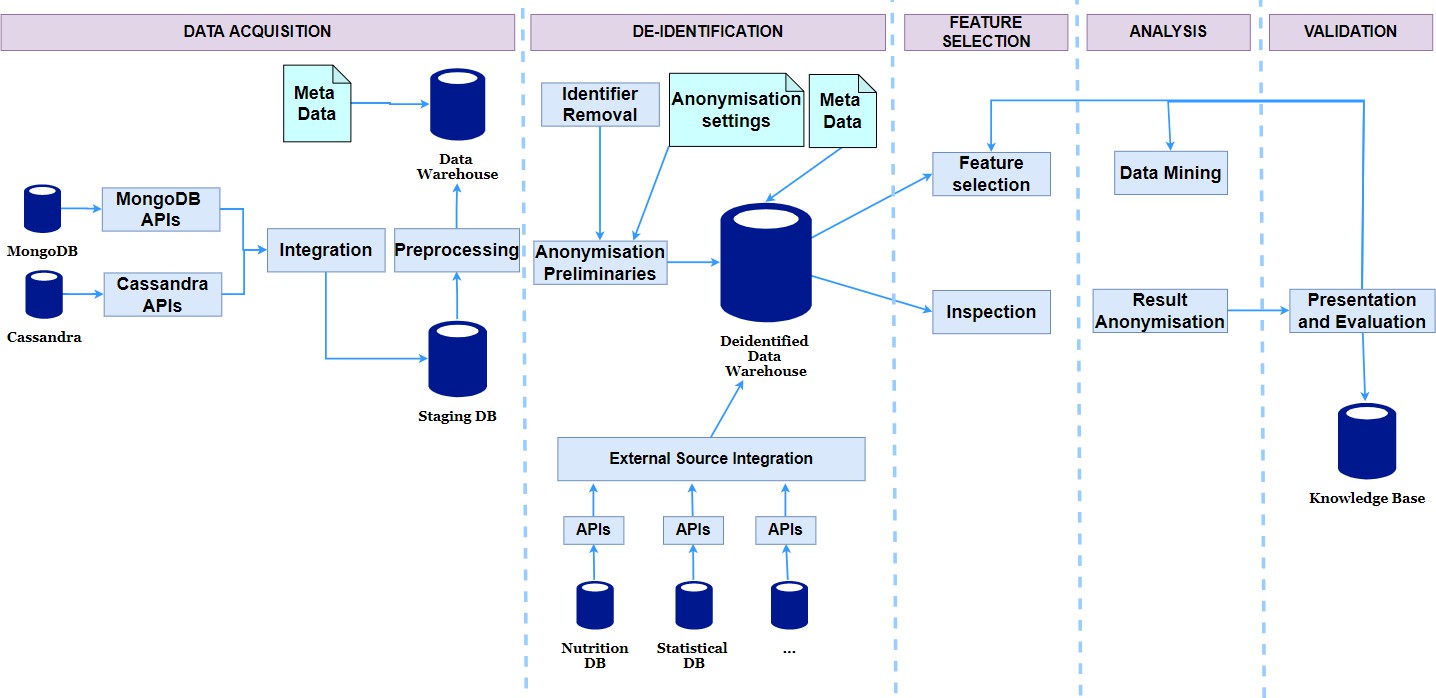
Thực tế việc hủy nhận dạng và bút danh là không đủ để giữ an toàn cho tập dữ liệu vì các định danh gần như. Không giống như các trường có thể nhận dạng, việc loại bỏ các định danh gần như có thể ảnh hưởng đến chất lượng phân tích dữ liệu. Các mô hình bảo mật, chẳng hạn như k-ẩn danh, l- đa dạng và LKC-privacy để ngăn chặn các cuộc tấn công liên kết bản ghi và thuộc tính, sử dụng các hoạt động ẩn danh khác nhau, chẳng hạn như khái quát hóa, ngăn chặn, xáo trộn, nhiễu loạn và thêm tiếng ồn. Tuy nhiên, một số hoạt động (ví dụ: xáo trộn, khái quát hóa tế bào và nhiễu loạn) chỉ phù hợp để phân tích cụ thể và có thể làm cho các bộ dữ liệu không có sẵn cho các thuật toán khai thác.

BigO là một hệ thống dữ liệu lớn với các bộ dữ liệu chiều cao, đặc biệt là khi nhiều bộ sưu tập cơ sở dữ liệu được kết hợp. Do đó, việc ẩn danh, sử dụng k-ẩn danh và các phiên bản mở rộng của nó, trên các bộ dữ liệu chiều cao gây ra biến dạng dữ liệu nghiêm trọng dẫn đến chất lượng phân tích dữ liệu kém.

* + 1. Giao thức phân tích dữ liệu nhận thức được quyền riêng tư

Giao thức nhận biết quyền riêng tư được đề xuất được thể hiện trong Hình [7](#_bookmark17). Giao thức được thiết kế để giải quyết các vấn đề về các thuộc tính có thể nhận dạng và gần như có thể nhận dạng khi chia sẻ dữ liệu để phân tích và nó xem xét vấn đề tính chiều cao.

Trong các phương pháp truyền thống, các nhiệm vụ phân tích được áp dụng cho dữ liệu ẩn danh có chất lượng bị suy giảm. Ý tưởng tổng thể của giao thức là hủy nhận dạng nhưng không ẩn danh dữ liệu. Thay vì ẩn danh dữ liệu và chia sẻ nó để phân tích, chúng tôi cung cấp các chế độ xem được bảo mật đặc biệt cho các nhà khoa học dữ liệu để kiểm tra các bộ dữ liệu và sau đó chạy phân tích trên các bộ dữ liệu không xác định nhưng không ẩn danh. Các quan điểm đặc biệt cho phép những người dùng cuối này kiểm tra dữ liệu từ nhiều góc độ khác nhau nhưng không tiết lộ mối liên hệ giữa thông tin nhạy cảm và bệnh nhân. Các nhà khoa học dữ liệu có thể chọn các phương pháp lựa chọn tính năng và thuật toán phân tích được chạy trên dữ liệu không xác định nhưng không ẩn danh. Vì dữ liệu không chứa số nhận dạng, kết quả được phát hiện có chất lượng cao. Hơn nữa, quá trình chạy này được quản lý bởi hệ thống và các nhà khoa học không thể truy cập dữ liệu không ẩn danh. Trước khi được công bố, kết quả được kiểm tra xem không tiết lộ thông tin bất ngờ. Một số chi tiết của các mô hình kết quả được lọc để đảm bảo quyền riêng tư cho trẻ em. Các bước của giao thức như sau:



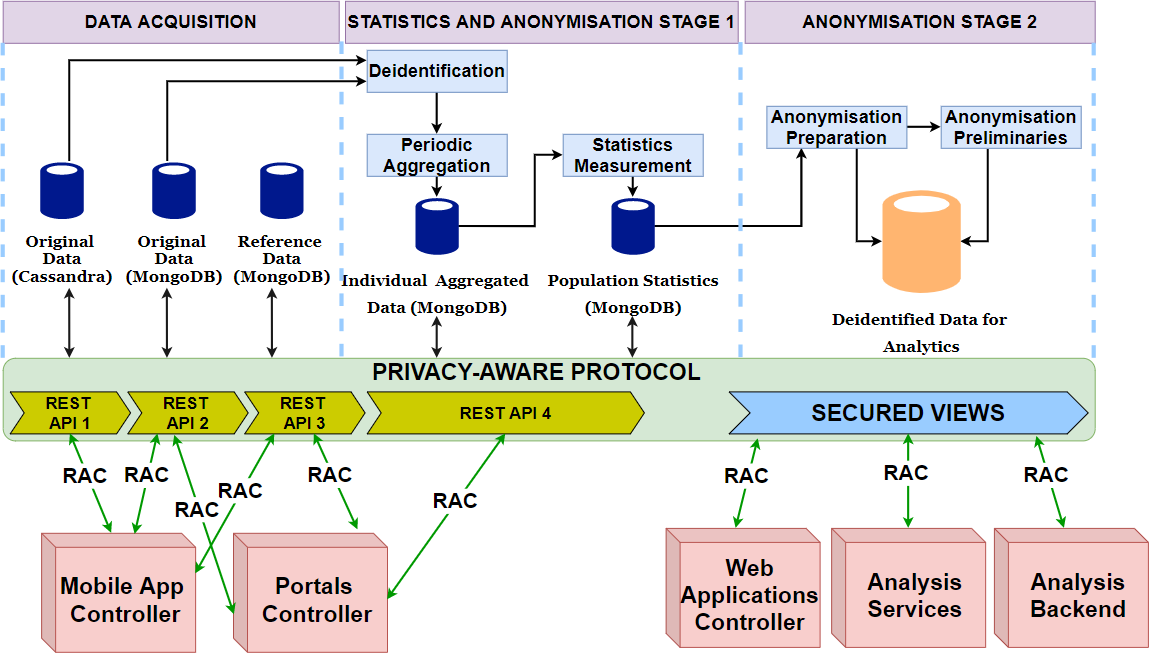
**Hình 7.** Giao thức phân tích dữ liệu nhận biết quyền riêng tư.

* + - * Hủy nhận dạng: Các thuộc tính có thể nhận dạng và các thuộc tính không quan trọng để phân tích sẽ bị xóa trong bước này.
      * Chuẩn bị ẩn danh: Một số thuộc tính yêu cầu các phương pháp xử lý nhỏ hỗ trợ tạo chế độ xem an toàn. Trường hợp này thường xảy ra cho đến nay và số nhưng không phải là thuộc tính mèo. Ví dụ, thường không cần thiết phải giữ các giá trị chiều cao chi tiết. Vì vậy, chúng tôi làm tròn chúng thành phạm vi. Việc chuyển đổi này không phải là một hoạt động ẩn danh và nó chỉ thay đổi một chút nội dung dữ liệu và thông tin trong dữ liệu để phân tích gần như được bảo toàn. Do đó, chúng tôi coi bước này là chuẩn bị ẩn danh. Một công việc quan trọng khác của nhiệm vụ này là tạo ra các sơ bộ ẩn danh (bao gồm cả cây phân loại cho các thuộc tính gần như có thể nhận dạng) được sử dụng để tạo các chế độ xem được bảo mật.
      * Tạo chế độ xem bảo mật: Chế độ xem bảo mật được tạo ra để giúp các nhà khoa học dữ liệu kiểm tra và hiểu tập dữ liệu từ nhiều góc độ khác nhau nhưng không tiết lộ mối liên hệ giữa bệnh nhân và thông tin nhạy cảm của họ. Có ba loại chế độ xem bảo mật:
        + Chế độ xem thống kê: Điều này cung cấp các biện pháp, chẳng hạn như độ lệch chuẩn, phạm vi miền và thống kê giá trị cho các thuộc tính được tính toán tự động.
        + Chế độ xem ẩn danh: Điều này cung cấp toàn bộ chế độ xem của các bộ dữ liệu được chia sẻ. Để bảo vệ quyền riêng tư, chúng tôi đã áp dụng thuật toán Quyền riêng tư và Ẩn danh trong Bảo mật Thông tin (PAIS) [[44](#_bookmark70)] để đạt được mô hình bảo mật LKC [[45](#_bookmark71)]. LKC-privacy ngăn chặn các cuộc tấn công liên kết bản ghi và thuộc tính cho các bộ dữ liệu chiều cao. PAIS sử dụng chiến lược tìm kiếm từ trên xuống trên cây phân loại để tìm khái quát hóa dưới mức tối ưu cho các bản ghi. Đối với các nhiệm vụ phân tích chung, chi phí phân biệt được sử dụng làm thước đo để chọn chuyên môn tốt nhất.
        + Chế độ xem giải phẫu: Vì k-ẩn danh là một điều kiện của quyền riêng tư LKC, kết quả của PAIS gặp phải vấn đề về tính chiều cao. Do đó, các chế độ xem ẩn danh có thể cung cấp quan điểm quá chung chung về các định danh gần như. Quan điểm chi tiết hoặc giải phẫu cũng được cung cấp bằng cách sử dụng kỹ thuật giải phẫu.
      * Lựa chọn tính năng: Sau khi kiểm tra các bộ dữ liệu với các chế độ xem khác nhau, các nhà khoa học dữ liệu có thể chọn các phương pháp chuyển đổi, lựa chọn tính năng và trích xuất thích hợp để tạo dữ liệu đầu vào thích hợp cho các nhiệm vụ phân tích ứng dụng cụ thể của họ. Việc xử lý được thực hiện trên tập dữ liệu không xác định và không ẩn danh.
      * Khai thác dữ liệu và ẩn danh kết quả: Các nhà khoa học dữ liệu có thể chọn các phương pháp phân tích khác nhau. Kết quả trả về có thể quá chi tiết trong một số trường hợp. Ví dụ, một cây quyết định (đầu ra của thuật toán phân loại ở trên) có các nút lá chi tiết liên kết với một số cá nhân đặc biệt. Do đó, các kết quả khai thác phải được kiểm tra và lọc trước khi công bố để các nhà nghiên cứu đảm bảo quyền riêng tư của trẻ em.
      * Trình bày và đánh giá: Các mô hình kết quả được đánh giá và các nhà phân tích dữ liệu có thể được bắt đầu lại phân tích của họ từ bước kiểm tra nếu cần thiết.

# Triển khai kiến trúc hệ thống BigO nhận biết quyền riêng tư

* 1. *Mô tả các thay đổi kiến trúc*

Để sử dụng giao thức nhận biết quyền riêng tư, được mô tả trong Hình [7](#_bookmark17), vào kiến trúc BigO (Hình [6](#_bookmark12)), chúng tôi đã cập nhật kiến trúc như thể hiện trong Hình [8](#_bookmark19). Có thể thấy rằng một số mô-đun trung gian được thêm vào giữa các bước chuẩn bị khử nhận dạng và ẩn danh. Các mô-đun này hỗ trợ tổng hợp định kỳ và tính toán trước thống kê dân số. Các bản cập nhật chính của kiến trúc được liệt kê dưới đây:



**Hình 8.** Kiến trúc BigO nhận biết quyền riêng tư.

* Tách cơ sở dữ liệu MongoDB: Không giống như sơ đồ thành phần BigO (Hình [6](#_bookmark12)) với một cơ sở dữ liệu MongoDB, các bộ sưu tập của cơ sở dữ liệu này được tách thành ba cơ sở dữ liệu:
  + Cơ sở dữ liệu MongoDB đầu tiên (Original data): Cơ sở dữ liệu này chứa dữ liệu quản trị và dữ liệu được thu thập/đo lường, bao gồm các bộ sưu tập USERS, CHIL- DREN, MEALS, TIMELINES, FOOD\_ADVERTISEMENTS, DAILY\_ANSWERS và PHOTOS.
  + Second MongoDB database (Dữ liệu tham khảo): Cơ sở dữ liệu này lưu trữ dữ liệu không liên quan đến cá nhân và được sử dụng để tham khảo. Danh sách bao gồm các bộ sưu tập TRƯỜNG HỌC, PHÒNG KHÁM, NHÓM, KHU VỰC và PUBLIC\_POIS.
  + Cơ sở dữ liệu MongoDB thứ ba (bao gồm Dữ liệu tổng hợp cá nhân và Thống kê dân số): Dữ liệu tổng hợp cá nhân bao gồm các bộ sưu tập tóm tắt dữ liệu hành vi peri-odically của từng đứa trẻ như HÀNG NGÀY, HÀNG TUẦN và THỐNG KÊ. Dữ liệu Thống kê Dân số bao gồm các bộ sưu tập COUNTERS, PUB- LIC\_POIS\_VOTES, GEOHASH\_VOTES, GEOHASH\_ATTRIBUTES và HISTOGRAMS.
* Tách API: Kiến trúc BigO được cập nhật hỗ trợ bốn API khác nhau để truy cập cơ sở dữ liệu Cassandra và ba cơ sở dữ liệu MongoDB.
* Mô-đun khử nhận dạng: Mô-đun này loại bỏ các trường có thể nhận dạng cũng như các trường không cần thiết để phân tích và không yêu cầu bút danh.
* Phân hệ tổng hợp định kỳ: Mô-đun này tổng hợp định kỳ dữ liệu hành vi của trẻ em.
* Mô-đun đo lường thống kê: Mô-đun này tính toán trước số liệu thống kê của một số dân cư được sử dụng cho các tính năng trực quan hóa và tạo chế độ xem thống kê.
* Mô-đun chuẩn bị ẩn danh: Mô-đun này thực hiện các tác vụ được mô tả trong bước Chuẩn bị ẩn danh của giao thức nhận thức về quyền riêng tư nói trên. Đầu ra của mô-đun này là Sơ bộ ẩn danh và dữ liệu không xác định để phân tích.
* Sơ bộ ẩn danh: Chúng được lưu ở định dạng JSON hoặc XML.
* Dữ liệu không xác định để phân tích: Việc lưu trữ dữ liệu này không nên lưu trữ các bộ sưu tập rời rạc như trong cơ sở dữ liệu của Dữ liệu tổng hợp cá nhân và Thống kê dân số. Dữ liệu để phân tích phải là các bộ dữ liệu kết hợp ở các định dạng thuận tiện cho việc tạo các chế độ xem và thuật toán khai thác được bảo mật. Một lựa chọn tốt là các tệp CSV được lưu trữ trong bộ lưu trữ hệ thống tệp của Hadoop.
* Chế độ xem bảo mật: Dữ liệu để phân tích được truy cập thông qua chế độ xem bảo mật. Có các mô-đun chịu trách nhiệm tạo các chế độ xem bảo mật và chạy các lựa chọn tính năng / thuật toán khai thác trên dữ liệu đã hủy nhận dạng để phân tích.
  1. *Quy trình cập nhật thay đổi dữ liệu*

Điều quan trọng là phải kiểm tra khi có thay đổi trong lưu trữ dữ liệu gốc. Trong thực tế, có hai loại cập nhật dữ liệu có thể:

* Thay đổi dữ liệu quản trị: Dữ liệu quản trị (ví dụ: email, tên, địa chỉ) được nhập thủ công để đôi khi có lỗi cần cập nhật. Vì kiểu dữ liệu này không được trích xuất để lưu trữ trong các cơ sở dữ liệu khác, việc đồng bộ hóa không phải là vấn đề.
* Chèn các biện pháp mới: Các biện pháp hành vi được tải lên từ ứng dụng dành cho thiết bị di động đến cơ sở dữ liệu gốc thường xuyên. Sau một khoảng thời gian nhất định, để phản ánh những thay đổi trong dữ liệu gốc, dữ liệu tóm tắt mới được thêm vào kho lưu trữ dữ liệu tổng hợp riêng lẻ và số liệu thống kê hiện có được cập nhật trong kho lưu trữ Thống kê dân số. Các sơ bộ ẩn danh và Dữ liệu không xác định để phân tích cũng được tính toán lại.

# Hình ảnh hiện tại, khuyến nghị và định hướng tương lai

Trong phần này, chúng tôi xem xét các bài học kinh nghiệm và các khuyến nghị của chúng tôi để đề cập đến kiến trúc kho dữ liệu lớn nhận thức được quyền riêng tư và các bước hướng tới các hướng nghiên cứu tiếp theo.

Trước tiên, chúng tôi nhấn mạnh một số vấn đề chính trong kho dữ liệu hiện có, bao gồm cả việc thiếu khả năng đối phó với lời nguyền của tính chiều cao. Trong các khuôn khổ truyền thống, các nhiệm vụ phân tích được áp dụng cho dữ liệu ẩn danh trong đó chất lượng dữ liệu bị suy giảm. Công việc của chúng tôi trình bày kiến trúc kho dữ liệu BigO và tích hợp hơn nữa một giao thức nhận thức về quyền riêng tư mới để giải quyết các vấn đề về các thuộc tính dữ liệu có thể nhận dạng và gần như có thể nhận dạng trong khi chia sẻ dữ liệu cho các tác vụ phân tích. Chúng tôi trình bày các giao thức bảo mật để thu thập, tích hợp và đánh giá dữ liệu ở hai cấp độ khác nhau, đó là cấp ứng dụng và cấp cơ sở dữ liệu. Chúng tôi cũng trình bày các giao thức kiểm soát truy cập dữ liệu cho các giai đoạn, chẳng hạn như đăng ký, xác thực và ủy quyền.

Ý tưởng chính là hủy nhận dạng dữ liệu cá nhân nhưng không nên ẩn danh. Kiến trúc kho dữ liệu BigO sử dụng các chế độ xem bảo mật cho các nhà khoa học dữ liệu để khai thác các bộ dữ liệu. Các chế độ xem đặc biệt cho phép người dùng cuối thực hiện các nhiệm vụ khám phá của chúng tôi mà không tiết lộ mối liên hệ giữa thông tin nhạy cảm và bệnh nhân tương ứng.

Các giao thức bảo mật và quyền riêng tư được triển khai trong hệ thống BigO có thể được áp dụng cho bất kỳ ứng dụng dựa trên dữ liệu nào ngoài lĩnh vực chăm sóc sức khỏe, nơi có nguy cơ mất thông tin cá nhân nhạy cảm trong quá trình thu thập, lưu trữ, truyền và truy cập dữ liệu trong khi thực hiện các tác vụ phân tích. Bản chất của các giao thức này là không làm giảm chất lượng dữ liệu, người ta sẽ có thể thực hiện các nhiệm vụ khai thác và phân tích dữ liệu mà không có nguy cơ mất thông tin nhạy cảm.

Trong tương lai, chúng tôi có kế hoạch xác định các chính sách và quy trình quản trị dữ liệu phù hợp và chính xác để tránh mọi vi phạm dữ liệu, vi phạm quyền riêng tư của bất kỳ cá nhân nào và tạo ra kết quả chất lượng cao vì lợi ích của xã hội và công dân. Quản trị dữ liệu rất quan trọng vì dự án BigO xử lý dữ liệu cá nhân và trẻ em nhạy cảm cần được quản lý và thao tác với mức độ chăm sóc và tin cậy cao. Do đó, phương pháp vòng đời quản trị dữ liệu phải đảm bảo tính bảo mật, chất lượng và tính toàn vẹn của dữ liệu.

# Kết luận

Vì béo phì đã trở thành một vấn đề sức khỏe cộng đồng toàn cầu nghiêm trọng với những tác động đối với cả cá nhân và xã hội nói chung, các can thiệp hành vi và các yếu tố cộng đồng môi trường phải được tính đến để can thiệp vào thời thơ ấu. BigO sử dụng các phương pháp thu thập dữ liệu công dân-nhà khoa học và các công nghệ khác nhau (điện thoại thông minh, dây đeo cổ tay, Mandometer) để tạo ra các mô hình toàn diện về tỷ lệ béo phì. Thu thập dữ liệu trong BigO cho phép các nhà nghiên cứu tạo ra các mô hình phân tích các yếu tố nguy cơ hành vi và dự đoán tỷ lệ béo phì, thông qua các hiệp hội với các mô hình hành vi cộng đồng và tỷ lệ béo phì tại địa phương.

Giám sát và lưu trữ dữ liệu cá nhân làm cho việc thể hiện dữ liệu, bảo mật và kiểm soát truy cập là nhiệm vụ đầy thách thức. Bài viết này lần đầu tiên triển khai các thành phần truy cập và lưu trữ dữ liệu của BigO bao gồm các giao diện với các thành phần hệ thống khác để cho phép tổng hợp dữ liệu, phân tích dữ liệu và trực quan hóa hoạt động trơn tru. Chúng tôi đã trình bày kiến trúc kho dữ liệu linh hoạt ba lớp cho BigO bao gồm lớp back-end, lớp kiểm soát truy cập và lớp điều khiển.

Chúng tôi tiếp tục triển khai các giao thức biểu diễn và chia sẻ dữ liệu trong cơ sở dữ liệu BigO và các mô hình lưu trữ xem xét các khía cạnh quyền riêng tư và bảo mật. Các kế hoạch bảo mật và quyền riêng tư dữ liệu được đưa ra dựa trên các loại dữ liệu cá nhân được thu thập trong các khía cạnh lưu trữ dữ liệu, truyền dữ liệu và truy cập dữ liệu. Chúng tôi đã trình bày những thách thức trong việc bảo vệ quyền riêng tư và triển khai các giao thức phân tích dữ liệu mới nhận thức được quyền riêng tư để đảm bảo rằng các mô hình kết quả đảm bảo quyền riêng tư cho trẻ em. Cuối cùng, chúng tôi đã triển khai kiến trúc hệ thống BigO tích hợp các giao thức nhận thức về quyền riêng tư nói trên.

**Đóng góp của tác giả:** Khái niệm hóa, A.S., T.-A.N.N. và M.-T.K.; quản lý dữ liệu, T.-A.N.N. và M.-T.K.; phân tích chính thức, A.S., T.-A.N.N. và M.-T.K.; mua lại nguồn vốn, M.-T.K.; Điều tra, A.S., T.-A.N.N. và M.-T.K.; PHƯƠNG PHÁP LUẬN, A.S., T.-A.N.N. và M.-T.K.; quản trị dự án, A.S.,

và M.-T.K.; TÀI NGUYÊN, A.S., T.-A.N.N. và M.-T.K.; phần mềm, A.S., M.-T.K.; giám sát, M.-T.K.;

xác nhận, A.S., T.-A.N.N. và M.-T.K.; hình dung, A.S., T.-A.N.N. và M.-T.K.; viết — bản thảo gốc, A.S.; viết — xem xét và chỉnh sửa, A.S., và M.-T.K. Tất cả các tác giả đã đọc và đồng ý với phiên bản đã xuất bản của bản thảo.

**Tài trợ:** Công việc dẫn đến những kết quả này là một phần của dự án EU H2020: BigO: Dữ liệu lớn chống béo phì ở trẻ em (Tài trợ số 727688, [https://BigOprogram.eu](https://BigOprogram.eu/), truy cập vào ngày 21 tháng 3 năm 2021). Dự án này là một phần của Chương trình Sức khỏe, Thay đổi Nhân khẩu học và Hạnh phúc của Cộng đồng Châu Âu của EU H2020. Công trình này cũng được hỗ trợ bởi Quỹ Khoa học Ireland theo số tài trợ SFI / 12 / RC / 2289\_P2.

**Tuyên bố của Hội đồng Đánh giá Tổ chức:** Nghiên cứu được thực hiện theo các hướng dẫn của Tuyên bố Helsinki và được Ủy ban Đạo đức Nghiên cứu Con người Thể chế của Đại học Dubin phê duyệt (2 tháng 9 năm 2019).

**Tuyên bố đồng ý được thông báo:** Sự đồng ý được thông báo đã được lấy từ tất cả các đối tượng tham gia vào nghiên cứu.

**Tuyên bố về tính khả dụng của dữ liệu:** Chia sẻ dữ liệu không áp dụng cho bài viết này.

**Lời cảm ơn:** Nhóm nghiên cứu mong muốn ghi nhận các cộng tác viên nghiên cứu Anastasios Delopoulos, Eirini Lekka, Ioannis Ioakeimidis, Christos Diou, Isabel Perez, Daniel Ferri, và tất cả các thành viên Hiệp hội Dự án Bigo; nhóm lâm sàng và hành chính W82GO, và khoa vật lý trị liệu tại Sức khỏe trẻ em Ireland, Temple Street, Dublin; và các bệnh nhân tham gia và cha mẹ của họ.

**Xung đột lợi ích:** Các tác giả tuyên bố không có xung đột lợi ích. Các nhà tài trợ không có vai trò trong việc thiết kế nghiên cứu; trong việc thu thập, phân tích hoặc giải thích dữ liệu; bằng văn bản của bản thảo, hoặc trong quyết định công bố kết quả.

# Viết tắt

Các chữ viết tắt sau đây được sử dụng trong bản thảo này:

BigO Dữ liệu lớn cho bệnh béo phì

GPS Hệ thống định vị toàn cầu

HIPAA Đạo luật về trách nhiệm giải trình và cung cấp thông tin bảo hiểm y tế EHR Hồ sơ sức khỏe điện tử

PHA Cơ quan Y tế Công cộng LEC Điều kiện bên ngoài địa phương

GIS Hệ thống thông tin địa lý DBMS Hệ quản trị cơ sở dữ liệu

API Giao diện lập trình ứng dụng JWT Mã thông báo web JSON

# Phụ lục A. Thiết kế lược đồ

Phần này trình bày thiết kế lược đồ cho cơ sở dữ liệu cốt lõi trong BigO.

*Phụ lục A.1. Lược đồ MongoDB*

Hình [A1](#_bookmark23)–[A5](#_bookmark24) cho thấy lược đồ của MongoDB. Lược đồ được tổ chức trong các bộ sưu tập như sau:

* Khu vực: Bộ sưu tập này lưu trữ các tính năng thú vị của các khu vực hành chính để đại diện cho tất cả các loại khu vực địa lý hoặc hành chính. Do đó, một khu vực có thể là một thị trấn, một thành phố, một tỉnh hoặc thậm chí là một quốc gia. Mỗi khu vực có một loạt các tọa độ tạo thành ranh giới của nó. Thông tin ranh giới sẽ được cung cấp bởi các cơ quan công quyền. Tùy thuộc vào loại khu vực, các trường đặc điểm khu vực có thể được thêm vào sau.
* Geohash: Do các đặc điểm đặc biệt như cấu trúc phân cấp và dễ dàng lưu trữ trong cơ sở dữ liệu, geohash được sử dụng trong hệ thống BigO để thể hiện thông tin địa lý. Dữ liệu về các hoạt động của trẻ em tại một số khu vực được ghi lại. Tuy nhiên, chỉ có dữ liệu ẩn danh và thống kê được lưu trữ trong Geohash\_attributes thu thập và Geohash\_votes.
* POI công cộng: Hai bộ sưu tập Public\_pois và Public\_pois\_votes lưu trữ thông tin hoạt động tổng hợp của trẻ em tại các điểm quan tâm công cộng. Do đó, cấu trúc của bộ sưu tập Geohash\_votes và Public\_pois\_votes gần như giống nhau.
* Trường học và Phòng khám: Những bộ sưu tập này lưu trữ thông tin của các trường học và phòng khám có tổ chức.
* Nhóm: Học sinh tham gia vào hệ thống BigO thông qua các trường có tổ chức được chia thành các nhóm. Mỗi nhóm được quản lý bởi một giáo viên.
* Người dùng và trẻ em: Hệ thống BigO có nhiều loại người dùng nhưng chỉ có dữ liệu trẻ em được thu thập để nghiên cứu. Do đó, bộ sưu tập Người dùng được sử dụng để quản lý các lĩnh vực quản trị trong khi bộ sưu tập Trẻ em bao gồm các trường được ghi lại để nghiên cứu. Các trường có sẵn của người dùng phụ thuộc vào vai trò của họ. Một số trường đặc biệt được chia sẻ vì chúng cần thiết cho các truy vấn phổ biến trên hai bộ sưu tập. Sự trùng lặp này đòi hỏi một chi phí bổ sung nhỏ khi chèn một đứa trẻ mới nhưng có thể cải thiện hiệu suất cho nhiều hoạt động.
* Ảnh: Các trường "ảnh" trong bộ sưu tập Các bữa ăn và Food\_Advertisements trong phiên bản trước được tách ra để lưu trữ trong Bộ sưu tập Ảnh. Nhóm tất cả các ảnh trong một bộ sưu tập cụ thể sẽ làm trơn tru việc quản lý và xác minh ảnh.
* Timelines and Mobility: Collection Mobility trong phiên bản cũ được thay thế và mở rộng sang Timelines bộ sưu tập. Bộ sưu tập mới không chỉ chứa địa điểm ghé thăm mà còn cả các hoạt động du lịch của trẻ em.
* Daily\_anwsers: Bộ sưu tập mới này chứa câu trả lời cho các câu hỏi hàng ngày.
* Bộ đếm: Cần tạo ID hiển thị cho trẻ em (trong Bộ sưu tập Người dùng). Các ID này được tạo từ hàm gọi trình tự tự động tăng get- NextChildDisplayId ("trẻ em"). Sau khi được gọi, hàm đơn giản này thêm 1 vào trường "child\_seq" của tài liệu có "\_id" = "con" trong Bộ đếm bộ sưu tập và trả về giá trị mới làm ID hiển thị mới. Kỹ thuật tương tự và Bộ đếm bộ sưu tập có thể được sử dụng để tạo trình tự tăng tự động khác.
* Quảng cáo về giấc ngủ, khả năng di chuyển, bữa ăn và thực phẩm: Các bộ sưu tập này lưu trữ dữ liệu về các hoạt động hàng ngày của trẻ em. Đặc biệt, quảng cáo bữa ăn và thực phẩm bao gồm hình ảnh do trẻ em chụp.
* Thói quen ăn uống: Bộ sưu tập này chứa thông tin về thói quen ăn uống có thể được trích xuất và tổng hợp từ Bữa ăn.
* Thống kê: Bộ sưu tập được sử dụng để lưu trữ thông tin thống kê, chẳng hạn như số lượng ảnh.
* Hàng ngày và hàng tuần: Dữ liệu của các hoạt động có thể được tổng hợp hàng ngày và hàng tuần và sau đó được lưu trữ trong các bộ sưu tập này.

**Vùng**

\_id: objectId

last\_update\_datetime: chuỗi // ISO8601 tên: chuỗi

kiểu:xâu // phân loại (thành phố, quốc gia, v.v.) geohashes: [chuỗi] hoặc điều này hoặc ranh giới ranh giới: {

loại: "Đa giác",

tọa độ: [

[ điểm 11, điểm 12, ..., điểm 1N, điểm 11], [ điểm 21, điểm 22, ..., điểm 2N, điểm 2N],

...

]} // trình tự điểm của ranh giới khu vực

được cung cấp bởi các cơ quan công quyền

Điểm: [Lon, lat]

center: {type: "Điểm", tọa độ: [lon, lat]}

zoom\_level: số

Kinh tế xã hội

average\_income: số // EUR/người/năm

education\_stats: [{

Cấp độ: chuỗi // phân loại

Tỷ lệ phần trăm: Số // %

}]

unemployment\_rate: số // %

Các lĩnh vực bổ sung (môi trường đô thị, tiếp xúc với thực phẩm

tiếp thị, v.v.) khi cần thiết

INDEX:

({loại: 1, tên: 1}, {duy nhất: đúng})

{tên: 1}

{ranh giới: "2dsphere"}

{average\_income: 1}

{unemployment\_rate: -1}

**Hình A1.** MongoDB Schema—Phần 1.

\_id: objectId

**geohash\_attributes**

last\_update\_datetime: chuỗi // ISO8601 geohash: chuỗi

has\_highest\_resolution: bool

hash\_center: {

loại: "Điểm", tọa độ: [lon, lat]

}

hash\_box: {

kiểu: "Đa giác", tọa độ: [

[ point\_bl, point\_ul, point\_ur, point\_br, point\_bl]

]} // b: dưới, u: trên, l: trái, r: phải hash\_area: số

Các trường cấp độ hoạt động

steps\_per\_min: số // activity\_counts\_per\_min trung bình : số // trung bình

Các trường bổ sung nếu cần

INDEX:

{geohash: 1}

{has\_highest\_resolution: -1}

{hash\_box: "2dsphere"}

{hash\_center: "2dsphere"}

**geohash\_votes**

\_id: objectId geohash: chuỗi

datetime\_start: chuỗi // ISO8601 datetime\_end: chuỗi // ISO8601 near\_home: int // (1: rất gần, 2: gần, 3: bình thường

4: xa, 5: rất xa) activity\_counts\_per\_min: số // trung bình

steps\_per\_min: số lượng // bữa ăn trung bình : int // số bữa ăn

Các trường "hành vi" bổ sung khi cần thiết

INDEX:

{geohash: 1, datetime\_start: -1}

**Hình A2.** MongoDB Schema—Phần 2.

**public\_pois\_votes**

\_id: objectId

public\_poi\_id: chuỗi // Google (hoặc nhà cung cấp khác)

chìa khóa duy nhất của PoI datetime\_start:xâu ISO8601 datetime\_end:xâu ISO8601

near\_home: int // (1: rất gần, 2: gần, 3: bình thường

4: xa, 5: rất xa) activity\_counts\_per\_min: số // trung bình steps\_per\_min: số trung bình Bữa ăn: Int Số lượng bữa ăn

Các trường "hành vi" bổ sung khi cần thiết

INDEX:

{public\_poi\_id: 1, datetime\_start: -1}

**public\_pois**

\_id: objectId

public\_poi\_id: chuỗi // Google (hoặc nhà cung cấp khác)

chìa khóa duy nhất của PoI

Tên: chuỗi

location: {type: "Point", tọa độ: [lon, lat]} type: string // categorical (from list of BigO types) description: string

INDEX:

{loại: 1, tên: 1}

{tên: 1}

{public\_poi\_id: 1}

{vị trí: "2dsphere"}

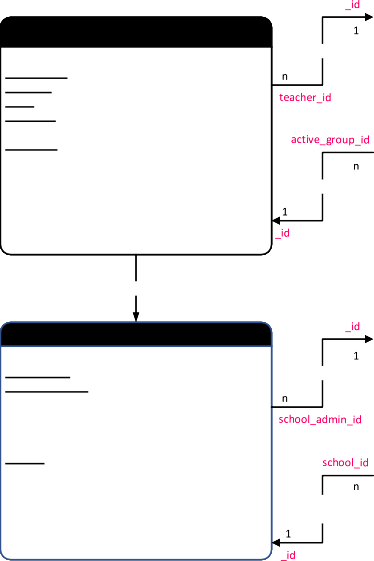
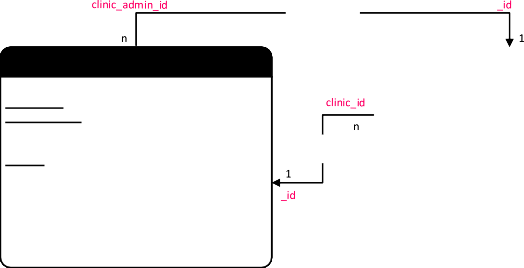
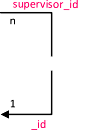
**Hình A3.** MongoDB Schema—Phần 3.

có học sinh

/Giáo viên

có bệnh nhân

/Bác sĩ



Đại diện bởi

**Bệnh viện**

**Người dùng**

**trẻ em**

\_id: objectId

clinic\_name: string clinic\_admin\_id: USERS.\_ID phone: string

Địa chỉ: Vị trí chuỗi: {

loại: "Điểm",

Tọa độ: [lon, lat]

}

có bệnh nhân

/Bác sĩ

được giám sát bởi

INDEX:

{clinic\_name: 1}

{vị trí: "2dsphere"}

**Nhóm**

\_id: objectId group\_name: chuỗi school\_id: SCHOOLS.\_ID active: bool

start\_date: chuỗi // ISO8601 end\_date: chuỗi // ISO8601 teacher\_id: USERS.\_ID mô tả: chuỗi

được giám sát bởi

\_Id

Quốc gia, Geohash, school\_id, clinic\_id

INDEX:

({school\_id: 1, start\_day: -1, group\_name: 1},

{duy nhất: đúng})

{teacher\_id: 1}

{start\_day: -1}

n school\_id thuộc về

1 \_id

có học sinh

Là

user\_id,

Quốc gia, Geohash, school\_id, clinic\_id

**Trường**

\_id: objectId

school\_name: xâu Đại diện bằng cách

school\_admin\_id: USERS.\_ID

Kiểu: chuỗi // (mầm non, tiểu học,

THCS, THPT) Điện thoại: String

Địa chỉ: String

location: {type: "Điểm", tọa độ: [lon, lat]}

INDEX:

{school\_name: 1}

{loại: 1}

{vị trí: "2dsphere"}

có học sinh

/Giáo viên

current\_height\_cm: số // cm

current\_weight\_kg: số // kg current\_bmi: số current\_bmi\_class: int

(1: thiếu cân rất nghiêm trọng,

2: thiếu cân nghiêm trọng, 3: thiếu cân,

4: bình thường, 5: tiền béo phì, 6: béo phì loại I,

7: béo phì loại II, 8: béo phì loại III) Chiều cao: [{

update\_datetime: chuỗi // ISO8601 Giá trị: Số Cm

}]

Trọng lượng: [{

update\_datetime: chuỗi // ISO8601 Giá trị: Số Kg

}]

BMI: [{

update\_datetime: chuỗi // ISO8601

Giá trị: Số

lớp: int // (1: thiếu cân rất nghiêm trọng,

2: thiếu cân nghiêm trọng,

3: thiếu cân, 4: bình thường, 5: tiền béo phì,

6: béo phì loại I, 7: béo phì loại II,

8: béo phì loại III)

}]

Bảng câu hỏi: [{

update\_datetime: chuỗi // ISO8601 weight\_assessment: int

(5: nhiều hơn, 4: nhiều hơn, 3: giống nhau,

2: ít hơn, 1: ít hơn nhiều) eating\_amount: int

(5: nhiều hơn, 4: nhiều hơn, 3: giống nhau,

2: ít hơn, 1: ít hơn nhiều) on\_diet: bool eating\_speed: int

(5: nhanh hơn nhiều, 4: nhanh hơn, 3: giống nhau,

2: chậm hơn, 1: chậm hơn nhiều) active\_degree: int

(5: nhiều hơn, 4: nhiều hơn, 3: giống nhau,

2: ít hơn, 1: ít hơn nhiều) sleeping\_quality: int

(5: rất tốt, 4: tốt, 3: trung bình,

2: xấu, 1: rất tệ) sleeping\_weekday\_time: chuỗi // ISO8601 wakeup\_weekday\_time: chuỗi // ISO8601 sleeping\_weekend\_time: chuỗi // ISO8601 wakeup\_weekend\_time: chuỗi // ISO8601

}]

Chương trình học hàng ngày school\_start\_time: String ISO8601 school\_end\_time: chuỗi ISO8601 exercise\_num: int

exercise\_periods: [int] // phút của mỗi lần breaks\_num: int

breaks\_periods: [int] // phút mỗi lần

meals\_num: int

meals\_periods: [int] // phút mỗi lần

INDEX:

({user\_id: 1}, {unique: true})

({quốc gia: 1, current\_bmi\_class: -1}, {thưa thớt: đúng})

({school\_id: 1, current\_bmi\_class: -1}, {thưa thớt: đúng})

({clinic\_id: 1, current\_bmi\_class: -1}, {thưa thớt: đúng}) ({current\_bmi\_class: -1}, {thưa thớt: đúng})

(F, M, NA)

Giới tính: Chuỗi

birth\_year: int

\_id: objectId

user\_id: Quốc gia USERS.\_ID: NGƯỜI DÙNG. GEOHASH QUỐC GIA : NGƯỜI DÙNG. BĂM ĐỊA LÝ

school\_id: NGƯỜI DÙNG. SCHOOL\_ID clinic\_id: NGƯỜI DÙNG. CLINIC\_ID

\_id: objectId username: chuỗi

Mật khẩu: chuỗi // được mã hóa hoặc băm

Vai trò: [chuỗi] // (trẻ em, giáo viên, bác sĩ lâm sàng

school\_admin, clinic\_admin,

global\_admin, nhà hoạch định chính sách)

E-mail: chuỗi

Biệt danh: chuỗi // Tên hiển thị registration\_code: chuỗi

Số điện thoại di động

(bị xóa sau khi đăng ký) hoặc TTL

Quốc gia: chuỗi // Phân loại

geohash: chuỗi // geohash với tối đa 5 chữ cái

==================Nhóc con==================

display\_id: getNextChildDisplayId('children')

bắt buộc đối với thiết bị conID: chuỗi // bắt buộc đối với trẻ em

Vai trò phụ: chuỗi // (sinh viên, bệnh nhân, tình nguyện viên) supervisor\_id: USERS.\_ID // giáo viên hoặc bác sĩ lâm sàng school\_id: SCHOOLS.\_ID

clinic\_id: CLINICS.\_ID active\_group\_id: GROUPS.\_ID

inactive\_group\_ids: [GROUPS.\_ID] // để lưu trữ

=================Sư huynh================

Tên: chuỗi

school\_id: SCHOOLS.\_ID

Điện thoại: String

=================Bác sĩ lâm sàng================

Tên: Chuỗi clinic\_id: CLINICS.\_ID Điện thoại: Chuỗi

================Các vai trò khác===============

Tên: Chuỗi Điện thoại: Chuỗi Tổ chức: Chuỗi

INDEX:

({tên người dùng: 1}, {unique: true})

({biệt danh: 1}, {duy nhất: đúng})

({roles: 1, subrole: 1}) ({supervisor\_id: 1}, {sparse: true}) ({school\_id: 1}, {sparse: true})

({clinic\_id: 1}, {thưa thớt: true}) ({active\_group\_id: 1}, {thưa thớt: true})

**Hình A4.** MongoDB Schema—Phần 4.

**trẻ em**

**Quầy**

\_id: giá trị chuỗi là "con" cho child\_seq

child\_seq: int

**Hình ảnh**

\_id \_id: objectId

ảnh: binData

**food\_advertisements**

\_Id

có

photo\_id

\_id: objectId

user\_id: TRẺ EM. USER\_ID datetime: chuỗi // ISO8601 photo\_id: PHOTOS.\_ID

có

**Bữa ăn**

location: {type: "Điểm", tọa độ: [lon, lat]}

location\_bearing: số location\_accuracy: số

ghi

user\_id

user\_id orientation\_a: số // phương vị

orientation\_p: số // pitch

orientation\_r: số // cuộn

photo\_id

annotation\_type: chuỗi // phân loại

annotation\_place: chuỗi

INDEX:

({user\_id: 1, datetime: -1}, {unique: true})

{user\_id: 1, vị trí: "2dsphere"}

{vị trí: "2dsphere", user\_id: 1}

**daily\_answers**

\_id: objectId

user\_id: TRẺ EM. USER\_ID

datetime: chuỗi // ISO8601

cho

user\_id user\_id

Tâm trạng: int // (5: rất hạnh phúc, 4: hạnh phúc, 3: trung tính,

2: không vui, 1: rất bất hạnh)

Các trường bổ sung (câu trả lời) nếu cần

INDEX:

({user\_id: 1, datetime: -1}, {unique: true})

({tâm trạng: 1}, {thưa thớt: đúng})

ăn

user\_id

user\_id

**Các mốc thời gian**

\_id: objectId

user\_id: TRẺ EM. USER\_ID start\_datetime: chuỗi // ISO8601 duration\_mins: int

Mỗi tài liệu đại diện cho

HOẶC là một địa điểm đã ghé thăm HOẶC một phong trào

1

có

Một địa điểm đã ghé thăm

location\_type: chuỗi // phân loại

n near\_home: int // (1: rất gần, 2: gần, 3: bình thường

4: xa, 5: rất xa)

user\_id user\_id

Một phong trào

movement\_method: chuỗi // (đi bộ, chạy, đạp xe,

xe máy, ô tô, public\_transport, khác) Khoảng cách: [số] // km

Các đặc điểm môi trường bổ sung khi cần thiết INDEXES:

({user\_id: 1, start\_datetime: -1}, {unique: true}) ({location\_type: 1}, {thưa thớt: true}) ({movement\_method: 1}, {thưa thớt: true})

\_id: objectId

user\_id: TRẺ EM. USER\_ID

start\_datetime: chuỗi // ISO8601 photo\_id: PHOTOS.\_ID location\_type: chuỗi // phân loại

(nhà, nhà hàng, thức ăn nhanh, v.v.) meal\_type: chuỗi // (bữa sáng, bữa trưa,

bữa tối, bữa ăn nhẹ, những người khác)

food\_attributes: [{

food\_type: chuỗi // (ví dụ: cá, thịt lợn, v.v.) Kích thước: chuỗi // (nhỏ, vừa, lớn) Chuẩn bị: chuỗi // Phân loại

(nấu tại nhà, nhà hàng,

mang đi, thực phẩm đóng gói, v.v.)

}]

drink\_attributes: [{

drink\_type: chuỗi // (ví dụ: chanh, soda, v.v.) Kích thước: Số lượng // mililit

include\_milk: bool include\_sugar: bool

}]

Mandometer đo m\_duration: số // giây m\_raw\_plate\_weights: [{

m\_raw\_time: số m\_raw\_weight: số // gram

}]

m\_plate\_weights: [{ m\_time: số

m\_weight: số // gram

}]

m\_intake\_curve\_a: Số // Đường cong đầu vào AX2 + BX m\_intake\_curve\_b: Số // Trục đường cong đầu vào2 + bx m\_intake\_amount: số // g m\_intake\_rate\_init: số // g / giây m\_intake\_rate\_avg: số // g / giây m\_intake\_rate\_deceleration: số // g / giây 2 m\_food\_addition\_num: int m\_food\_addition\_weights: [số] // g m\_bite\_times: [số]

m\_bite\_size\_mean: số// g m\_bite\_freq: số// cắn / giây

satiety\_before\_level: int // (1: rất đói, 2: đói,

3: bình thường, 4: đầy đủ, 5: rất đầy đủ) satiety\_before\_time: số

vài giây trước khi bữa ăn bắt đầu

satiety\_after\_level: int // (1: rất đói, 2: đói,

3: bình thường, 4: đầy đủ, 5: rất đầy đủ)

satiety\_after\_time: số

giây sau bữa ăn bắt đầu mood\_after\_meal: int // (5: rất hạnh phúc, 4: hạnh phúc,

3: trung lập, 2: không hạnh phúc, 1: rất bất hạnh)

INDEX:

({user\_id: 1, start\_datetime: -1}, {unique: true})

{user\_id: 1, location\_type: 1}

{user\_id: 1, meal\_type: 1}

{meal\_type: 1, location\_type: 1}

{user\_id: 1, "food\_attributes.food\_type": 1}

{user\_id: 1, "drink\_attributes.drink\_type": 1}

{mood\_after\_meal: -1}

\_id: objectId

user\_id: USERS.\_ID gia: NGƯỜI DÙNG. GEOHASH QUỐC GIA : NGƯỜI DÙNG. BĂM ĐỊA LÝ

school\_id: NGƯỜI DÙNG. SCHOOL\_ID

clinic\_id: NGƯỜI DÙNG. CLINIC\_ID

giống: xâu (F, M, NA)

birth\_year: int

current\_height\_cm: số // cm current\_weight\_kg: số // kg current\_bmi: số current\_bmi\_class: int

(1: thiếu cân rất nghiêm trọng,

2: thiếu cân nghiêm trọng, 3: thiếu cân,

4: bình thường, 5: tiền béo phì, 6: béo phì loại I,

7: béo phì loại II, 8: béo phì loại III)

Chiều cao: [{

update\_datetime: chuỗi // ISO8601 Giá trị: Số Cm

}]

Trọng lượng: [{

update\_datetime: chuỗi // ISO8601 Giá trị: Số Kg

}]

BMI: [{

update\_datetime: chuỗi // ISO8601

Giá trị: Số

lớp: int // (1: thiếu cân rất nghiêm trọng,

2: thiếu cân nghiêm trọng,

3: thiếu cân, 4: bình thường, 5: tiền béo phì,

6: béo phì loại I, 7: béo phì loại II,

8: béo phì loại III)

}]

Bảng câu hỏi: [{

update\_datetime: chuỗi // ISO8601

weight\_assessment: int

(5: nhiều hơn, 4: nhiều hơn, 3: giống nhau,

2: ít hơn, 1: ít hơn nhiều) eating\_amount: int

(5: nhiều hơn, 4: nhiều hơn, 3: giống nhau,

2: ít hơn, 1: ít hơn nhiều) on\_diet: bool eating\_speed: int

(5: nhanh hơn nhiều, 4: nhanh hơn, 3: giống nhau,

2: chậm hơn, 1: chậm hơn nhiều) active\_degree: int

(5: nhiều hơn, 4: nhiều hơn, 3: giống nhau,

2: ít hơn, 1: ít hơn nhiều) sleeping\_quality: int

(5: rất tốt, 4: tốt, 3: trung bình,

2: xấu, 1: rất tệ) sleeping\_weekday\_time: chuỗi // ISO8601 wakeup\_weekday\_time: chuỗi // ISO8601 sleeping\_weekend\_time: chuỗi // ISO8601 wakeup\_weekend\_time: chuỗi // ISO8601

}]

Chương trình học hàng ngày school\_start\_time: String ISO8601 school\_end\_time: chuỗi ISO8601 exercise\_num: int

exercise\_periods: [int] // phút mỗi lần

breaks\_num: int

breaks\_periods: [int] // phút của mỗi lần meals\_num: int

meals\_periods: [int] // phút mỗi lần

INDEX:

({user\_id: 1}, {unique: true})

({quốc gia: 1, current\_bmi\_class: -1}, {thưa thớt: đúng})

({school\_id: 1, current\_bmi\_class: -1}, {thưa thớt: đúng})

({clinic\_id: 1, current\_bmi\_class: -1}, {thưa thớt: true})

({current\_bmi\_class: -1}, {thưa thớt: đúng})

**Hình A5.** MongoDB Schema—Phần 5.

có

*Phụ lục A.2. Lược đồ Cassandra*

Hình [A6](#_bookmark26) cho thấy các bộ sưu tập trong Cassandra. Bây giờ chúng tôi liệt kê chúng dưới đây:

* Physical\_activity\_by\_user: lưu trữ dữ liệu của các hoạt động (ví dụ: đi bộ, đứng, ngồi, chạy, v.v.) cho mỗi người dùng.
* Physical\_activity\_by\_date: lưu trữ dữ liệu các hoạt động (ví dụ: đi bộ, đứng, ngồi, chạy, v.v.) của người dùng cho từng ngày giờ cụ thể.

**physical\_activity\_by\_user**

user\_id: văn bản TRẺ EM. USER\_ID Ngày: Văn bản ISO8601

Giờ: Phút INT : INT

Các bước: activity\_counts Int: Double

năng lượng: đôi // MET/phút

cấp độ: int // (ít vận động: 1, ánh sáng: 2, trung bình: 3,

mạnh mẽ: 4, rất mạnh mẽ: 5)

loại: văn bản // phân loại (ngồi, đứng, đi bộ,

chạy bộ, chạy, đi du lịch, v.v.) PK((user\_id, ngày), giờ, phút)

**physical\_activity\_by\_date**

Ngày: Văn bản ISO8601 giờ: int

phút: int

user\_id: Nhắn tin // TRẺ EM. USER\_ID

Các bước: activity\_counts Int: Double

năng lượng: đôi // MET/phút

mức: trong ít vận động: 1, ánh sáng: 2, vừa phải: 3,

Mạnh mẽ: 4, rất mạnh mẽ: 5

loại: văn bản // phân loại (ngồi, đứng, đi bộ,

chạy bộ, chạy, đi du lịch, v.v.) PK((ngày), giờ, phút, user\_id)

**Hình A6.** Lược đồ cơ sở dữ liệu Cassandra.

# Tham khảo

1. Abarca-Gómez, L.; Abdeen, Z.A.; Hamid, Z.A.; Abu-Rmeileh, N.M.; Acosta-Cazares, B.; Acuin, C.; Adams, RJ; Aekplakorn, W.; Afsana, K.; Aguilar-Salinas, C.A.; et al. Xu hướng trên toàn thế giới về chỉ số khối cơ thể, thiếu cân, thừa cân và béo phì từ năm 1975 đến năm 2016: Một phân tích tổng hợp của 2416 nghiên cứu đo lường dựa trên dân số trong 128 *·* 9 triệu trẻ em, thanh thiếu niên và người lớn. *Lancet* **2017**, *390*, 2627–2642. [[CrossRef](http://doi.org/10.1016/S0140-6736(17)32129-3)]
2. Dobbs, R.; Manyika, J. Cuộc khủng hoảng béo phì. *Cairo Linh mục Glob. Aff.* **Năm 2015**, *5*, 44–57.
3. Macaulay, E.; Donovan, E.; Leask, M.; Bloomfield, F.; Vickers, M.; Dearden, P.; Thợ làm bánh, P. Tầm quan trọng của cuộc sống sớm trong béo phì ở trẻ em và các bệnh liên quan: Một báo cáo từ Hội nghị thượng đỉnh chiến lược Gravida 2014. *J. Dev. Orig. Sức khỏe Dis.* **2014**, *5*, 398–407. [[CrossRef](http://dx.doi.org/10.1017/S2040174414000488)] [[PubMed](http://www.ncbi.nlm.nih.gov/pubmed/25308169)]
4. Bhave, S.; Bavdekar, A.; Otiv, M. Lực lượng đặc nhiệm quốc gia IAP về phòng chống bệnh người lớn ở trẻ em: Béo phì ở trẻ em. *Nhi khoa Ấn Độ.* **2004**, *41*, 559–576. [[PubMed](http://www.ncbi.nlm.nih.gov/pubmed/15235162)]
5. Cộng tác viên, G.O. Ảnh hưởng sức khỏe của thừa cân và béo phì ở 195 quốc gia trong hơn 25 năm. *N. Engl. J. Med.* **Năm 2017**, *377*, 13–27. [[Tham chiếu chéo](http://dx.doi.org/10.1056/NEJMoa1614362)]
6. Di Cesare, M.; Soric', M.; Bovet, P.; Miranda, JJ; Bhutta, Z.; Stevens, G.A.; Laxmaiah, A.; Kengne, AP; Bentham, J. Gánh nặng dịch tễ học của béo phì ở trẻ em: Một dịch bệnh trên toàn thế giới đòi hỏi hành động khẩn cấp. *BMC Med.* **Năm 2019**, *17*, 1–20. [[Tham chiếu chéo](http://dx.doi.org/10.1186/s12916-019-1449-8)]
7. Daumit, G.L.; Dickerson, FB; Wang, N.Y.; Dalcin, A.; Jerome, G.J.; Anderson, C.A.; Trẻ, D.R.; Frick, K.D.; Yu, A.; Gennusa, J.V., III; et al. Một can thiệp giảm cân hành vi ở những người mắc bệnh tâm thần nghiêm trọng. *N. Engl. J. Med.* **2013**, *368*, 1594–1602. [[Tham chiếu chéo](http://dx.doi.org/10.1056/NEJMoa1214530)]
8. Katzmarzyk, P.T.; Barreira, TV; Broyles, S.T.; Rượu sâm banh, C.M.; Chaput, JP; Fogelholm, M.; Hu, G.; Johnson, W.D.; Kuriyan, R.; Kurpad, A.; et al. Nghiên cứu quốc tế về béo phì ở trẻ em, lối sống và môi trường (ISCOLE): Thiết kế và phương pháp. *BMC Y tế công cộng* **2013**, *13*, 900. [[CrossRef](http://dx.doi.org/10.1186/1471-2458-13-900)]
9. Blake-Lamb, T.L.; Ổ khóa, L.M.; Perkins, M.E.; Baidal, JAW; Cheng, E.R.; Taveras, E.M. Can thiệp cho bệnh béo phì ở trẻ em trong 1000 ngày đầu tiên một đánh giá có hệ thống. *Am. J. Prev. Med.* **2016**, *50*, 780–789. [[Tham chiếu chéo](http://dx.doi.org/10.1016/j.amepre.2015.11.010)]
10. Briggs, AD; Mytton, O.T.; Kehlbacher, A.; Tiffin, R.; Rayner, M.; Scarborough, P. Ảnh hưởng tổng thể và thu nhập cụ thể đối với tỷ lệ thừa cân và béo phì của thuế đồ uống có đường 20% ở Anh: Nghiên cứu mô hình đánh giá rủi ro kinh tế lượng và so sánh. *BMJ* **2013**, *347*, f6189. [[Tham chiếu chéo](http://dx.doi.org/10.1136/bmj.f6189)]
11. Dương, H.J.; Kang, J.H.; Kim, O.H.; Choi, M.; Ồ, M.; Nam, J.; Sung, E. Can thiệp để ngăn ngừa béo phì ở trẻ em bằng điện thoại thông minh và thiết bị đeo được: Một giao thức cho một thử nghiệm có đối chứng không ngẫu nhiên. *Quốc tế J. Môi trường. Res. Y tế công cộng* **2017**, *14*, 184. [[CrossRef](http://dx.doi.org/10.3390/ijerph14020184)] [[PubMed](http://www.ncbi.nlm.nih.gov/pubmed/28208839)]
12. Maramis, C.; Diou, C.; Ioakeimidis, I.; Lekka, I.; Dudnik, G.; Sao Hỏa, M.; Maglaveras, N.; Bergh, C.; Delopoulos, A. Ngăn ngừa béo phì và rối loạn ăn uống thông qua sửa đổi hành vi: Tầm nhìn tuyệt vời. Trong Kỷ yếu của Hội nghị quốc tế lần thứ 4 năm 2014 về truyền thông di động không dây và chăm sóc sức khỏe chuyển đổi chăm sóc sức khỏe thông qua đổi mới trong công nghệ di động và không dây (MOBIHEALTH), Athens, Hy Lạp, 3–5 tháng 11 năm 2014; 7–10.
13. Delopoulos, A. Dữ liệu lớn chống béo phì ở trẻ em, Dự án Bigo. Trong Kỷ yếu của Hội nghị chuyên đề quốc tế lần thứ 32 của IEEE năm 2019 về các hệ thống y tế dựa trên máy tính (CBMS), Cordoba, Tây Ban Nha, 5–7 tháng 6 năm 2019; 64–66.
14. Berman, JJ Các vấn đề bảo mật cho các công cụ khai thác dữ liệu y tế. *Tạo hình. Trí tuệ. Med.* **Năm 2002**, *26*, 25–36. [[Tham chiếu chéo](http://dx.doi.org/10.1016/S0933-3657(02)00050-7)]
15. Elger, BS; Iavindrasana, J.; Iacono, L.L.; Müller, H.; Roduit, N.; Mùa hè, P.; Wright, J. Chiến lược trao đổi dữ liệu y tế cho nghiên cứu lâm sàng thứ cấp, liên tổ chức. *Máy tính. Phương pháp Chương trình Biomed.* **2010**, *99*, 230–251. [[Tham chiếu chéo](http://dx.doi.org/10.1016/j.cmpb.2009.12.001)]
16. Ponemon, tôi. *Nghiên cứu điểm chuẩn hàng năm lần thứ sáu về quyền riêng tư và bảo mật dữ liệu chăm sóc sức khỏe*; Báo cáo kỹ thuật; Viện Ponemon LLC: Thành phố Traverse, MI, Hoa Kỳ, 2016.
17. Aggarwal, C.C. Về k-ẩn danh và lời nguyền của tính chiều. Trong Kỷ yếu của VLDB, Trondheim, Na Uy, 30 tháng 8–2 tháng 9 năm 2005; Tập 5, các trang 901–909.
18. Fung, B.C.; Vương, K.; Fu, A.W.C.; Philip, S.Y. *Giới thiệu về xuất bản dữ liệu bảo vệ quyền riêng tư: Các khái niệm và kỹ thuật*; Báo chí CRC: Boca Raton, FL, Hoa Kỳ, 2010.
19. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-đa dạng: Quyền riêng tư ngoài K-ẩn danh. *ACM Trans. Biết. Vũ trường. Dữ liệu* **2007**, *1*, 3-es. [[CrossRef](http://dx.doi.org/10.1145/1217299.1217302)]
20. Sweeney, L. K-ẩn danh: Một mô hình để bảo vệ quyền riêng tư. *Int. J. Không chắc chắn. Fuzziness-knowl.-dựa trên syst.* **2002**, *10*, 557–570. [[Tham chiếu chéo](http://dx.doi.org/10.1142/S0218488502001648)]
21. Nelson, GS. *Ý nghĩa thực tế của việc chia sẻ dữ liệu: Sơ lược về quyền riêng tư dữ liệu, ẩn danh và hủy nhận dạng*; Báo cáo kỹ thuật; Công nghệ ThotWave: Chapel Hill, NC, Hoa Kỳ, 2015.
22. Kanwal, T.; Anjum, A.; Khan, A. Bảo vệ quyền riêng tư trong đám mây sức khỏe điện tử: Phân loại, yêu cầu về quyền riêng tư, phân tích tính khả thi và cơ hội. *Rầm rầm rầm. Máy tính.* **2021**, *24*, 293–317. [[Tham chiếu chéo](http://dx.doi.org/10.1007/s10586-020-03106-1)]
23. Manios, Y.; Grammatikaki, E.; Androutsos, O.; Chinapaw, M.; Gibson, E.; Buijs, G.; Iotova, V.; Socha, P.; Annemans, L.; Wildgruber, A.; et al. Một cách tiếp cận có hệ thống để phát triển can thiệp dựa trên mẫu giáo để ngăn ngừa béo phì ở trẻ em ở độ tuổi mẫu giáo: Nghiên cứu ToyBox. *Béo phì. Rev.* **2012**, *13*, 3–12. [[CrossRef](http://dx.doi.org/10.1111/j.1467-789X.2011.00974.x)]
24. Paans, N.P.; Bot, M.; Brouwer, I.A.; Visser, M.; Roca, M.; Kohls, E.; Watkins, E.; Penninx, B.W. Mối liên quan giữa trầm cảm và phong cách ăn uống ở bốn quốc gia châu Âu: Nghiên cứu phòng chống MooDFOOD. *J. Tâm lý. Res.* **Năm 2018**, *108*, 85–92. [[Tham chiếu chéo](http://dx.doi.org/10.1016/j.jpsychores.2018.03.003)]
25. Lakerveld, J.; Glonti, K.; Rutter, H. Mối tương quan cá nhân và bối cảnh của các hành vi liên quan đến béo phì và béo phì: Dự án SPOTLIGHT. *Béo phì. Khải huyền* **2016**, *17*, 5–8. [[Tham chiếu chéo](http://dx.doi.org/10.1111/obr.12384)]
26. Vượn, C.; Del Pozo, G.B.; Andrés, J.; Lobstein, T.; Manco, M.; Lewy, H.; Bergman, E.; O'Callaghan, D.; Doherty, G.; Kudraut- seva, O.; et al. Nền tảng dữ liệu như một dịch vụ để cung cấp lối sống lành mạnh và y tế dự phòng: Khái niệm và cấu trúc của dự án DAPHNE. *JMIR Res. Nguyên sinh.* **2016**, *5*, E222. [[Tham chiếu chéo](http://dx.doi.org/10.2196/resprot.6589)]
27. Voigt, P.; Von dem Bussche, A. Quy định bảo vệ dữ liệu chung của EU (GDPR). In *A Practical Guide*, ấn bản 1; Nhà xuất bản Quốc tế Springer: Cham, Thụy Sĩ, 2017; Tập 10, tr. 3152676.
28. Rantos, K.; Drosatos, G.; Demertzis, K.; Ilioudis, C.; Papanikolaou, A.; Kritsas, A. ADvoCATE: Một nền tảng quản lý sự đồng ý để xử lý dữ liệu cá nhân trong IoT bằng công nghệ blockchain. Trong Kỷ yếu của Hội nghị Quốc tế về An ninh Công nghệ Thông tin và Truyền thông, Bucharest, Romania, 14–15 tháng 11 năm 2018; 300–313.
29. Ấu trùng, X.; Moffie, M.; Asaf, S.; Santamaria, I. Hướng tới một cách tuân thủ GDPR để đảm bảo ngành chăm sóc sức khỏe xuyên biên giới châu Âu 4.0. *Máy tính. Đứng. Giao diện* **2020**, *69*, 103408. [[Tham chiếu chéo](http://dx.doi.org/10.1016/j.csi.2019.103408)]
30. Mustafa, Hoa Kỳ; Pflugel, E.; Philip, N. Một khung bảo mật mới cho các ứng dụng m-health an toàn: Trường hợp của GDPR. Trong Kỷ yếu của Hội nghị Quốc tế IEEE lần thứ 12 năm 2019 về An ninh, An toàn và Bền vững Toàn cầu (ICGS3), London, Vương quốc Anh, 16–18 tháng 1 năm 2019; 1–9.
31. Sahama, T.; Croll, P. Kiến trúc kho dữ liệu cho kho dữ liệu lâm sàng. In Proceedings of the ACSW Frontiers 2007: Proceedings of 5th Australasian Symposium on Grid Computing and e-Research, 5th Australasian Information Security Workshop (Privacy Enhancement Technologies), and Australasian Workshop on Health Knowledge Management and Discovery, Victoria, Australia, 30 January – 2 February 2007; 227–232.
32. Neamah, A.F. Kho dữ liệu linh hoạt: Hướng tới xây dựng kiến trúc hồ sơ sức khỏe điện tử tích hợp. Trong Kỷ yếu của Hội nghị Quốc tế về Điện tử và Truyền thông Thông minh (ICOSEC) năm 2020, Tamilnadu, Ấn Độ, 10–12 tháng 9 năm 2020; 1038–1042.
33. Poenaru, C.E.; Merezeanu, D.; Dobrescu, R.; Posdarascu, E. Các giải pháp tiên tiến để lưu trữ thông tin y tế: Kho dữ liệu lâm sàng. Trong Kỷ yếu của Hội nghị Kỹ thuật Sinh học và Sức khỏe Điện tử 2017 (EHB), Sinaia, Romania, 22–24 tháng 6 năm 2017;

37–40.

1. Sweeney, L. Datafly: Một hệ thống cung cấp ẩn danh trong dữ liệu y tế. Trong *Bảo mật cơ sở dữ liệu XI*; Springer: Berlin/Heidelberg, Đức, 1998; 356–381.
2. Tưởng, Y.C.; Hsu, T.s.; Kuo, S.; Liau, C.J.; Wang, D.W. Giữ bí mật khi chia sẻ cơ sở dữ liệu y tế với hệ thống Cellsecu. *Int. J. Med. Thông báo.* **Năm 2003**, *71*, 17–23. [[Tham chiếu chéo](http://dx.doi.org/10.1016/S1386-5056(03)00030-3)]
3. Agrawal, R.; Johnson, C. Bảo mật hồ sơ sức khỏe điện tử mà không cản trở luồng thông tin. *Int. J. Med. Thông báo.* **Năm 2007**,

*76*, 471–479. [[CrossRef](http://dx.doi.org/10.1016/j.ijmedinf.2006.09.015)] [[PubMed](http://www.ncbi.nlm.nih.gov/pubmed/17204451)]

1. Prasser, F.; Kohlmayer, F.; Lautenschläger, R.; Kuhn, KA ARX — Một công cụ toàn diện để ẩn danh dữ liệu y sinh. Trong Kỷ yếu của Kỷ yếu Hội nghị chuyên đề hàng năm AMIA. Hiệp hội Tin học Y tế Hoa Kỳ, Washington, DC, Hoa Kỳ, 19–21 tháng 5 năm 2014; Tập 2014, tr. 984.
2. Nguyễn, T.A.; Lê-Khac, N.A.; Kechadi, M.T. Phần mềm trung gian phân tích dữ liệu nhận biết quyền riêng tư cho các hệ thống dựa trên dữ liệu của cô ấy. Trong Kỷ yếu Hội nghị Quốc tế về Kỹ thuật Dữ liệu và Bảo mật Tương lai, Thành phố Hồ Chí Minh, Việt Nam, 29 Tháng Mười Một – 1 Tháng Mười Hai 2017; 335–350.
3. Trần, N.H.; Nguyễn Ngọc, T.A.; Lê-Khac, N.A.; Kechadi, M. Một mô hình truy cập nhận thức bảo mật cho hệ thống EHR theo hướng dữ liệu. *arXiv*

**2019**, arXiv:1908.10229.

1. Zeilenga, K. *Giao thức truy cập thư mục nhẹ (LDAP): Bản đồ lộ trình đặc tả kỹ thuật*; Báo cáo kỹ thuật, RFC 4510, tháng Sáu; Quỹ OpenLDAP: Minden, NV, Hoa Kỳ, 2006.
2. Mặt trời, J.; Gao, Z. Cải thiện cơ chế bảo mật ứng dụng di động dựa trên Kerberos. Trong Kỷ yếu của Hội thảo Quốc tế lần thứ 4 về Kỹ thuật Vật liệu và Khoa học Máy tính năm 2019, Bangkok, Thái Lan, 17–19 tháng 5 năm 2019; 108–112.
3. Tewari, H.; Hughes, A.; Weber, S.; Barry, T. X509Cloud — Khung cho PKI phổ biến. Trong Kỷ yếu của Hội nghị Truyền thông Quân sự IEEE MILCOM 2017—2017 (MILCOM), Baltimore, MD, Hoa Kỳ, 23–25 tháng 10 năm 2017; 225–230.
4. Hoa Kỳ, IC Bảo mật và bảo vệ cơ sở dữ liệu Cassandra với IBM Security Guardium. Có sẵn trực tuyến: [https://www.ibm.com/](https://www.ibm.com/developerworks/library/se-secure-protect-cassandra-databases-ibm-security-guardium-trs/index.html) [developerworks/library/se-secure-protect-cassandra-databases-ibm-security-guardium-trs/index.html](https://www.ibm.com/developerworks/library/se-secure-protect-cassandra-databases-ibm-security-guardium-trs/index.html) (truy cập ngày 5 tháng Mười năm 2020).
5. Xiong, L.; Truta, T.M.; Fotouhi, F. Báo cáo về hội thảo quốc tế về quyền riêng tư và ẩn danh trong xã hội thông tin (PAIS 2008). *ACM SIGMOD Rec.* **Năm 2009**, *37*, 108–111. [[Tham chiếu chéo](http://dx.doi.org/10.1145/1519103.1519125)]
6. Rafiei, M.; Wagner, M.; van der Aalst, WM TLKC-mô hình bảo mật cho quá trình khai thác. Trong Kỷ yếu của Hội nghị Quốc tế về Thách thức Nghiên cứu trong Khoa học Thông tin, Limassol, Síp, 23–25 tháng 9 năm 2020; 398–416.