[***sustainability***](http://www.mdpi.com/journal/sustainability)



**Mô hình hóa và quản lý Dữ liệu lớn trong cơ sở dữ liệu—Một đánh giá tài liệu có hệ thống**

Diana Martinez-Mosquera 1,\* , Rosa Navarrete 1 và Sergio Lujan-Mora 2

1

Khoa Tin học và Khoa học Máy tính, Escuela Politécnica Nacional, 170525 Quito, Ecuador; rosa.navarrete@epn.edu.ec

Khoa Phần mềm và Hệ thống Máy tính, Đại học Alicante, 03690 Alicante, Tây Ban Nha; sergio.lujan@ua.es

2

\* Thư từ: diana.martinez@epn.edu.ec; Điện thoại: +593-02-2976300

Nhận: 25 tháng 11 năm 2019; Chấp nhận: ngày 9 tháng 1 năm 2020; Đã xuất bản: 15 tháng 1 năm 2020

Tóm tắt: Công việc được trình bày trong bài báo này được thúc đẩy bởi sự thừa nhận rằng một đánh giá tài liệu có hệ thống (SLR) đầy đủ và cập nhật nhằm hợp nhất tất cả các nỗ lực nghiên cứu về mô hình hóa và quản lý Dữ liệu lớn đang bị thiếu. Nghiên cứu này trả lời ba câu hỏi nghiên cứu. Câu hỏi đầu tiên là số lượng bài báo được xuất bản về quản lý và mô hình hóa Dữ liệu lớn đã phát triển như thế nào theo thời gian. Câu hỏi thứ hai là liệu nghiên cứu có tập trung vào dữ liệu bán cấu trúc và/hoặc phi cấu trúc hay không và kỹ thuật nào được áp dụng. Cuối cùng, câu hỏi thứ ba xác định xu hướng và lỗ hổng nào tồn tại theo ba khái niệm chính: nguồn dữ liệu, mô hình hóa và cơ sở dữ liệu. Kết quả là, 36 nghiên cứu, được thu thập từ các thư viện kỹ thuật số khoa học quan trọng nhất và trong khoảng thời gian từ 2010 đến 2019, đã được coi là có liên quan. Hơn nữa, chúng tôi trình bày một phân tích thư mục hoàn chỉnh để cung cấp thông tin chi tiết về các tác giả và dữ liệu xuất bản trong một tài liệu. SLR này tiết lộ những sự thật rất thú vị. Chẳng hạn, Mối quan hệ thực thể và hướng tài liệu là những mô hình được nghiên cứu nhiều nhất ở cấp độ trừu tượng hóa khái niệm và logic tương ứng và MongoDB là triển khai thường xuyên nhất tại vật lý. Hơn nữa, 2,78% nghiên cứu đã đề xuất các phương pháp hướng đến cơ sở dữ liệu lai với trường hợp thực tế cho dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc.

Từ khóa: dữ liệu lớn; ban quản lý; người mẫu; phê bình văn học

1. Giới thiệu

Thuật ngữ lập mô hình Dữ liệu lớn đã trở nên phổ biến vào năm 2011, như thể hiện trong Hình 1. Hình này cho thấy các tìm kiếm trong Google Xu hướng liên quan đến lập mô hình Dữ liệu lớn, được tăng cường từ năm 2011 trở đi. Các tìm kiếm trước năm 2004 không được hiển thị vì Google Xu hướng không lưu trữ dữ liệu trước đó. Trong những năm gần đây, các nhà nghiên cứu đã củng cố nỗ lực nghiên cứu các mô hình mới để xử lý Dữ liệu lớn. Do đó, mô hình hóa và quản lý Dữ liệu lớn mới trong các phương pháp tiếp cận cơ sở dữ liệu đã xuất hiện, phù hợp với các yêu cầu mới. Do đó, các kỹ thuật mới trong bối cảnh cơ sở dữ liệu đã phát triển theo hướng Không Chỉ SQL (NoSQL).

Công việc được trình bày trong bài báo này được thúc đẩy bởi sự thừa nhận rằng một đánh giá tài liệu có hệ thống hoàn chỉnh (SLR) hợp nhất tất cả các nỗ lực nghiên cứu để quản lý và mô hình hóa Dữ liệu lớn trong cơ sở dữ liệu. SLR là cách tốt nhất để thu thập, tóm tắt và đánh giá tất cả các bằng chứng khoa học về một chủ đề [1]. Nó cho phép mô tả các lĩnh vực nghiên cứu được các nhà nghiên cứu quan tâm nhiều nhất và ít nhất. Xem xét các vấn đề đã được phơi bày, SLR được thực hiện trong công việc này có thể góp phần giải quyết vấn đề thiếu sót này bằng cách thu thập và phân tích chi tiết về nghiên cứu được công bố từ năm 2010 đến 2019. Để làm cơ sở cho SLR của chúng tôi, chúng tôi đã tuân thủ các nguyên tắc do Kitchenham đề xuất [1]. Hơn nữa, bài viết này trình bày một phân tích thư mục đầy đủ và tóm tắt bằng chứng hiện có về nghiên cứu về mô hình hóa và quản lý Dữ liệu lớn trong cơ sở dữ liệu. Với thông tin thu được từ quá trình phân tích được thực hiện, chúng tôi sẽ xác định các xu hướng và lỗ hổng trong nghiên cứu đã công bố để cung cấp nền tảng cho các nghiên cứu mới. Từ phân tích được thực hiện, chúng tôi sẽ xác định các xu hướng và lỗ hổng trong nghiên cứu được công bố để cung cấp nghiên cứu. Kết quả là, 1376 bài báo được lấy từ các thư viện khoa học và 36 nghiên cứu đã được chọn là nền tảng cho nghiên cứu mới. Kết quả là, 1376 bài báo được lấy từ các thư viện khoa học và 36 liên quan, thích hợp. Tất cả các nỗ lực nghiên cứu đã được vạch ra để trả lời ba câu hỏi nghiên cứu đã xác định các nghiên cứu đã được chọn là có liên quan. Trong nghiên cứu này. Mục tiêu chính của chúng tôi là hợp nhất các công việc chính để cung cấp nhận thức về các xu hướng câu hỏi nghiên cứu được xác định trong nghiên cứu này. Mục tiêu chính của chúng tôi là hợp nhất các công việc chính để cung cấp và những khoảng trống liên quan đến mô hình Dữ liệu lớn. nhận thức về các xu hướng và lỗ hổng liên quan đến mô hình Dữ liệu lớn.

Phần còn lại của bài viết này được tổ chức như sau. Đầu tiên là phần Giới thiệu, bao gồm ý nghĩa của các thuật ngữ khác nhau được thảo luận trong nghiên cứu này. Thứ hai, phần Phương pháp trình bày quy trình, ý nghĩa của các thuật ngữ khác nhau được thảo luận trong nghiên cứu này. Thứ hai, phần Phương pháp trình bày được sử dụng để thực hiện việc lập kế hoạch, tiến hành và báo cáo SLR. Giai đoạn lập kế hoạch mô tả các quy trình được sử dụng để thực hiện việc lập kế hoạch, tiến hành và báo cáo SLR. Giai đoạn tiến hành trình bày việc đưa vào và lựa chọn mục tiêu và biện minh, câu hỏi nghiên cứu và chiến lược. Giai đoạn tiến hành trình bày các tiêu chí cho khối văn bản cuối cùng của các nghiên cứu được lựa chọn. Tiêu chí đưa vào và lựa chọn cho tập cuối cùng của các nghiên cứu được chọn. Tiểu mục đầu tiên, Phân tích thư mục, bao gồm giai đoạn báo cáo, bao gồm thông tin của tác giả, chẳng hạn như tiểu mục, Phân tích thư mục, bao gồm giai đoạn báo cáo, bao gồm thông tin của tác giả, liên kết và quốc gia và dữ liệu liên quan về công việc của họ; ví dụ: thông tin xuất bản, chẳng hạn như liên kết và quốc gia và dữ liệu liên quan về tác phẩm của họ; ví dụ, xuất bản số lần trích dẫn, nguồn kinh phí, năm, thư viện số, hệ số tác động, thứ hạng. Tiểu mục thứ hai, thông tin, số lần trích dẫn, nguồn tài trợ, năm, thư viện số, hệ số tác động, thứ hạng. Các Tổng quan Văn học có Hệ thống, trình bày sơ đồ các nghiên cứu được lựa chọn theo ba tiêu chí chính. Tiểu mục thứ hai, Tổng quan tài liệu có hệ thống, trình bày sơ đồ các nghiên cứu được chọn các khái niệm trong một ma trận khái niệm liên quan đến nguồn dữ liệu, mô hình hóa và cơ sở dữ liệu. Thứ ba, phần Kết quả trả lời các câu hỏi nghiên cứu trong ba tiểu mục. Tổng quan tài liệu có hệ thống, trình bày sơ đồ các nghiên cứu được chọn các khái niệm trong một ma trận khái niệm liên quan đến nguồn dữ liệu, mô hình hóa và cơ sở dữ liệu. thứ ba theo ba khái niệm chính trong một ma trận khái niệm liên quan đến nguồn dữ liệu, mô hình hóa và tiểu mục, Thảo luận, nêu bật những phát hiện có liên quan trong nghiên cứu SLR để xác định hiện tại cơ sở dữ liệu. Tiểu mục thứ ba, Thảo luận, nêu bật những phát hiện liên quan trong nghiên cứu SLR theo thứ tự xu hướng và lỗ hổng. Cuối cùng, kết luận và công việc trong tương lai được Graphical user interface, chart, line chart

Description automatically generatedtrình bày. Để xác định các xu hướng và lỗ hổng hiện có. Cuối cùng, kết luận và công việc trong tương lai được trình bày.

1.1. Khái niệm dữ liệu lớn

Trong phần này, chúng tôi mô tả các khái niệm chính liên quan đến Dữ liệu lớn, nhằm cung cấp một cái nhìn tổng quát, tổng quan cho người đọc và nền tảng của các thuật ngữ được thảo luận sau.

1.1.1. Sơ lược về lịch sử dữ liệu lớn

Việc sản xuất và xử lý khối lượng lớn dữ liệu bắt đầu được nhiều nhà nghiên cứu quan tâm.nhiều năm trước. Đến năm 1944, các ước tính về quy mô của các thư viện, vốn tăng nhanh hàng năm, đã được thực hiệntrong các trường đại học Mỹ [3]. Năm 1997, tại Viện Kỹ sư Điện và Điện tử (IEEE) đã được thực hiện tại các trường đại học Mỹ [3]. Hội nghị về Trực quan hóa, thuật ngữ “Dữ liệu lớn” lần đầu tiên được sử dụng trong một nghiên cứu về trực quan hóa tập dữ liệu lớn[4], trình bày một nghiên cứu về trực quan hóa bộ dữ liệu lớn [4].

Dữ liệu lớn là từ thông dụng trong những năm gần đây, tức là một cách diễn đạt thời thượng trong các hệ thống thông tin. Dân số nói chung liên quan đến thuật ngữ. Dữ liệu lớn với nghĩa đen của khối lượng dữ liệu lớn. Tuy nhiên, Dữ liệu lớn là một thuật ngữ chung được sử dụng để chỉ các bộ dữ liệu lớn và phức tạp phát sinh từ dữ liệu. Sự kết hợp của Big Data V nổi tiếng đặc trưng cho nó [5].

1.1.2. Đặc tính dữ liệu lớn

Như đã đề cập trước đây, Dữ liệu lớn không chỉ đề cập đến khối lượng dữ liệu lớn sẽ được xử lý. Tại khi bắt đầu nghiên cứu Dữ liệu lớn, khối lượng, tốc độ và sự đa dạng của chúng được coi là cơ bản các đặc điểm, được gọi là ba Vs của Dữ liệu lớn. Sau những tiến bộ trong nghiên cứu, các V mới, chẳng hạn như giá trị và tính xác thực, đã được thiết lập. Hiện nay, có tác giả đề xuất tới 42 đặc trưng cần thiết để coi dữ liệu là Big Data nên họ định nghĩa 42 Vs cho Big Data [6]. Đối với mục đích nghiên cứu của chúng tôi, chúng tôi sẽ chỉ đề cập đến mười V của Dữ liệu lớn, được trình bày trong một nghiên cứu khoa học [7]. Bảng 1 tóm tắt từng đặc điểm, cùng với một mô tả ngắn gọn.

**Bảng 1. 10 Vs Bigdata.**

|  |  |
| --- | --- |
| **Đặc tính** | **Mô tả ngắn gọn** |
| Volume | Bộ dữ liệu lớn |
| Velocity | Tỷ lệ tạo dữ liệu cao |
| Variety | Loại định dạng dữ liệu khác nhau |
| Variability | Dữ liệu nhất quán, có thể sử dụng cho nhiều mục đích cụ thể |
| Viscosity | Biến thể vận tốc dữ liệu |
| Virality | Tốc độ truyền dữ liệu |
| Veracity | Độ chính xác của dữ liệu |
| Validity | Đánh giá dữ liệu |
| Visualization | Biểu tượng dữ liệu |
| Value | Dữ liệu hữu ích để truy xuất thông tin |

1.1.3. Khối lượng và Vận tốc

Để đối phó với các đặc điểm Khối lượng và Tốc độ của Dữ liệu lớn, hệ sinh thái và kiến ​​trúc giải pháp, chẳng hạn như lambda và kappa, đã được tạo ra. Cả hai kiến ​​trúc đề xuất một cấu trúc của các lớp để xử lý Dữ liệu lớn; sự khác biệt chính giữa chúng là lambda đề xuất một lớp cho hàng loạt xử lý dữ liệu và một lớp khác để truyền dữ liệu, trong khi kappa đề xuất một lớp duy nhất cho cả hai đợt và xử lý luồng [8]. Máy ảnh SLR này tập trung vào mô hình hóa dữ liệu, một khái niệm liên quan đến Đa dạng đặc trưng, ​​được giải thích tiếp theo.

1.1.4. Đa dạng

Sự đa dạng là một đặc điểm đề cập đến các loại dữ liệu khác nhau và các danh mục và quản lý của một kho dữ liệu lớn. Theo đặc điểm này, Dữ liệu lớn đã được phân loại thành cấu trúc, dữ liệu bán cấu trúc và phi cấu trúc [9,10]. Các tiểu mục tiếp theo giải thích chi tiết từng loại dữ liệu. Trong Dữ liệu lớn, dữ liệu có cấu trúc được biểu diễn dưới dạng bảng, trong bảng tính hoặc quan hệ cơ sở dữ liệu [10]. Để đối phó với loại dữ liệu này, các công nghệ và kỹ thuật được sử dụng. Tuy nhiên, theo báo cáo được trình bày bởi công ty CISCO, loại dữ liệu chỉ chiếm 10% tổng số dữ liệu hiện có trong năm 2014 [11]. Vì vậy, điều rất quan trọng là phải phân tích các 90% dữ liệu còn lại, tương ứng với dữ liệu bán cấu trúc và dữ liệu phi cấu trúc sẽ được mô tả dưới đây. Dữ liệu bán cấu trúc được coi là dữ liệu không tuân theo một cấu trúc chính thức, chẳng hạn như một mô hình cơ sở dữ liệu quan hệ. Tuy nhiên, họ trình bày một tổ chức nội bộ tạo điều kiện thuận lợi cho quá trình xử lý của nó; chẳng hạn, nhật ký của máy chủ ở định dạng giá trị được phân tách bằng dấu phẩy (csv), tài liệu ở định dạng Đánh dấu mở rộng định dạng ngôn ngữ (XML), Ký hiệu đối tượng JavaScript (JSON) và JSON nhị phân (BSON), v.v. Một số tác giả có thể coi XML và JSON là có cấu trúc [10].

**Dữ liệu phi cấu trúc**

Dữ liệu phi cấu trúc được coi là những dữ liệu không có lược đồ được xác định trước hoặc không có tổ chức trong cấu trúc của chúng [12]. Trong loại dữ liệu này là tài liệu văn bản, email, dữ liệu cảm biến, tệp âm thanh,tệp hình ảnh, tệp video, dữ liệu từ trang web, trò chuyện, hồ sơ sức khỏe điện tử, dữ liệu truyền thông xã hội và dữ liệu không gian-thời gian, trong số những dữ liệu khác [9]. Theo CISCO, khối lượng dữ liệu phi cấu trúc giữa năm 2017 và 2022 dự kiến ​​tăng gấp 12 lần [13]. Để hỗ trợ Sự đa dạng, Khối lượng và Tốc độ của Dữ liệu lớn, không liên quan, phân tán và mã nguồn mở hệ thống lưu trữ dữ liệu đã được tạo ra. Các hệ thống này bao gồm khả năng mở rộng theo chiều ngang, tuyến tính hóa, tính sẵn sàng cao và khả năng chịu lỗi. Thông thường, những cơ sở dữ liệu này được gọi là NoSQL.

1.2. NoSQL

Các mô hình hiện có để xử lý dữ liệu thông thường không đủ và không phù hợp để xử lý Yêu cầu Dữ liệu lớn. Vì lý do đó, ở cấp độ lưu trữ dữ liệu, việc giới thiệu các phương pháp mới, chẳng hạn như cơ sở dữ liệu NoSQL, là bắt buộc. NoSQL đề cập đến Không chỉ SQL, thuật ngữ được sử dụng cho tất cả cơ sở dữ liệu phi quan hệ [14]. Cơ sở dữ liệu NoSQL được coi là không có lược đồ, vì chúng được thiết kế để làm việc không có cấu trúc [14]; tuy nhiên, trên thực tế, cần có một mô hình tự cung tự cấp để xác định dữ liệu sẽ được tổ chức và truy xuất từ ​​cơ sở dữ liệu như thế nào. Để giải quyết yêu cầu này, một số mô hình dữ liệu NoSQL được đề xuất.

Mô hình dữ liệu

Một mô hình dữ liệu là một đại diện cho cấu trúc của dữ liệu để xử lý và tổ chức [15].

Một mô hình dữ liệu được coi là một yếu tố chính để lưu trữ, phân tích và xử lý trong các hệ thống lưu trữ. Hiện tại, các hệ thống lưu trữ được phân thành hai nhóm lớn, quan hệ và không quan hệ. Trong quan hệ, các mô hình nổi tiếng là Mối quan hệ thực thể (ER), Thực thể mở rộng mối quan hệ (EER), Key-Cube và Đa chiều, trong số những thứ khác. Mục tiêu của bài viết này không phải là để trình bày một nghiên cứu sâu về các mô hình được coi là cổ điển: chúng nổi tiếng và không cần phải giải thích. Chúng tôi chỉ phát triển một nghiên cứu về các mô hình mới đối với Dữ liệu lớn. Đối với các hệ thống không quan hệ, có cơ sở dữ liệu NoSQL; đối với họ, các mô hình dữ liệu được phân loại thành bốn loại chính [9]:

1. Định hướng theo cột

2. Định hướng tài liệu

3. Đồ thị

4. Khóa-giá trị

Định hướng theo cột

Trong mô hình này, dữ liệu được biểu diễn dưới dạng bảng theo cột và hàng. Các cột làcó thể nhận dạng bằng một khóa phân vùng là duy nhất và bắt buộc và các hàng bằng một nhóm tùy chọn chìa khóa. Khóa chính là sự kết hợp của khóa phân vùng và cụm. Về cơ bản, lược đồ của các bảng bao gồm một tập hợp các cột, khóa chính và kiểu dữ liệu [16]. Đối với quản lý cơ sở dữ liệu các hệ thống (DBMS) sử dụng mô hình dữ liệu hướng cột, có thể kể đến Accumulo, AmazonSimpleDB, Cassandra, Cloudata, Druid, Elassandra, Flink, HBase, Hortonworks, HPCC, Hypertable, IBM Informix, Kudu, MonetDB, Scylla và Splice Machine, trong số những công ty khác [17].

Định hướng tài liệu

Trong mô hình này, dữ liệu được lưu trữ theo cặp khóa-giá trị, tài liệu giá trị ở định dạng XML, JSON hoặc BSON. Mỗi tài liệu có thể có các tài liệu con, chỉ mục, trường và thuộc tính lồng nhau [15]. Như ví dụ DBMS sử dụng mô hình dữ liệu hướng tài liệu có thể kể đến ArangoDB, Azure, BagriDB, Đám mây dữ liệu, CouchDB, DocumentDB, đàn hồi, IBM Cloudant, MongoDB, NOSDB, RAVENDB, RETHINKDB, SEQUOIADB, TORODB và UNQULITE, trong số những người khác [17].

Đồ thị

Mô hình này bao gồm một biểu đồ chứa các nút và cạnh. Một nút đại diện cho một thực thể và một cạnh đại diện cho mối quan hệ giữa các thực thể. Có một số cấu trúc đồ thị: không mong muốn/được định hướng, đồ thị được dán nhãn, đồ thị được quy kết, bigdata, đa hình, hypergraphs và đồ thị lồng nhau, trong số các biểu đồ khác [10]. Một số ví dụ về DBM sử dụng mô hình dữ liệu đồ thị là Allegraph, Arangodb, Infinite Graph, Graphbase, HyperGraphDB, Infogrid, Meronymousy, Neo4j, Onyx DataBase, Titan, Trinity, Virtuoso OpenLink, SparkSee và WhitedB [17].

Giá trị khóa

Trong mô hình này, dữ liệu được biểu thị bằng một bộ giá trị khóa. Khóa biểu thị một định danh duy nhất được lập chỉ mục với một giá trị đại diện cho dữ liệu thuộc loại, cấu trúc và kích thước tùy ý [10]. Các khóa và chỉ mục thứ cấp không được hỗ trợ [15]. Aerospike, Azure Table Storage, Bangdb, Berkeley DB, DynamoDB, GeniedB, KeyDB, Redis, Riak, Scalaris, Voldemort, trong số những người khác [17] là những ví dụ về DBM sử dụng mô hình dữ liệu giá trị khóa. Bảng 2 tóm tắt các đặc điểm chính của các mô hình dữ liệu NoQuery, chẳng hạn như khái niệm chính, cấu trúc, kỹ thuật của nó để tạo ra mô hình dữ liệu, ưu điểm và nhược điểm.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Đặc điểm / Mô hình dữ liệu | Định hướng theo cột | Định hướng tài liệu | Hướng đồ thị | Giá trị khóa |
| Ý tưởng | Một mô hình cho phép biểu diễn dữ liệu trong các cột | Một mô hình cho phép biểu diễn dữ liệu thông qua  văn bản có cấu trúc | Một mô hình cho phép biểu diễn dữ liệu và các kết nối của chúng | Mô hình cho phép biểu diễn dữ liệu ở định dạng đơn giản (khóa và giá trị) |
| Kết cấu | Các giá trị trong một cột được lưu trữ liên tiếp | Lồng các cặp khóa-giá trị | Tập hợp các liên kết giữa các đối tượng (các cạnh) | Một khóa đại diện cho bất kỳ thực thể nào  thuộc tính |
| Mỗi tài liệu được xác định bởi một số nhận dạng duy nhất |
|  | Bất kỳ giá trị nào cũng có thể là một tài liệu có cấu trúc |
|  | Khóa và giá trị được phân tách bằng dấu hai chấm “:” | Giá trị có thể thuộc bất kỳ kiểu dữ liệu nào |
| Giá trị có thể thuộc bất kỳ dữ liệu kiểu nào | Các cặp khóa-giá trị được phân tách bằng dấu phẩy “,” |
|  | Dữ liệu được đặt trong dấu ngoặc nhọn biểu thị |
|  | Dữ liệu được đặt trong dấu ngoặc vuông biểu thị tập hợp mảng |
| Với nén:  Mã hóa nhẹ  Mã hóa vectơ bit  Mã hóa từ điển  Hệ quy chiếu mã hóa  Mã hóa vi sai | Mô hình không chuẩn hóa với nhiều cấu trúc hơn (siêu dữ liệu)  Mô hình phẳng không chuẩn hóa |  |
| Kỹ thuật | Biểu đồ có hướng đơn giản, nhiều đồ thị vô hướng, nhiều đồ thị có hướng | NA |
| Với thuật toán tham gia | Đã tan vỡ, tương đương với chuẩn hóa  (https://pdfs.semanticscholar.org/ea15/  945ce9ec0c12b92794b8ace69ce44ebe40cc.pdf) | Đồ thị có trọng số |
| Với sự vật chất hóa muộn | Siêu đồ thị |
| Tuple tại một thời điểm | Biểu đồ lồng nhau |

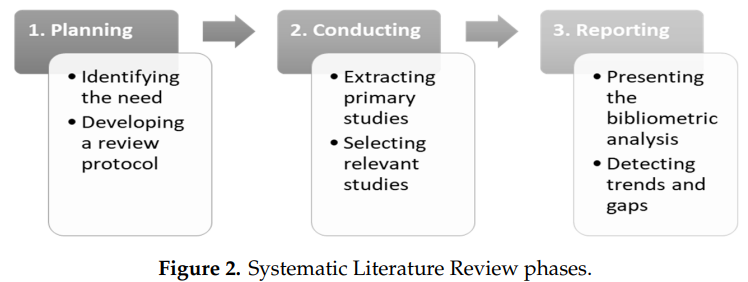
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Các ứng dụng | Dữ liệu người tiêu dùng  Dữ liệu hàng tồn kho | Tài liệu JSON  Tài liệu XML | Mạng xã hội  Chuỗi cung ứng  Hồ sơ bệnh án  Hoạt động CNTT  Phương tiện giao thông | Hồ sơ người dùng và thuộc tính của họ |
| Thuận lợi | Hiệu suất cao trong các hoạt động tải và truy vấn  Nén và phân vùng dữ liệu hiệu quả (theo cả chiều ngang và chiều dọc)  Khả năng mở rộng  Hỗ trợ xử lý song song lớn  Rất thích hợp cho khối lượng công việc Xử lý Phân tích Trực tuyến và Xử lý Giao dịch Trực tuyến | Hỗ trợ nhiều loại tài liệu  Hỗ trợ các giao dịch nguyên tử, nhất quán, cách ly và độ bền  Khả năng mở rộng  Thích hợp cho dữ liệu phức tạp, tài liệu và mảng được lồng vào nhau | Mô hình hóa dễ dàng  Truy vấn nhanh chóng và đơn giản  Khả năng mở rộng | Thiết kế và thực hiện dễ dàng  Khả năng chịu lỗi  Dư  Khả năng mở rộng  Tốc độ cao |
| Nhược điểm | Khó sử dụng cột rộng  Sự chậm trễ trong việc truy vấn dữ liệu cụ thể | thông tin trùng lặp trên nhiều tài liệu, mâu thuẫn trong các thiết kế phức tạp | Thiếu ngôn ngữ khai báo chuẩn  Hỗ trợ đồng thời hạn chế và song song | Ngôn ngữ truy vấn rất cơ bản  Một số truy vấn chỉ có thể phụ thuộc vào khóa chính |

1.3. Mức độ trừu tượng hóa dữ liệu

Nói chung, trong thiết kế của cả cơ sở dữ liệu quan hệ và NoSQL, ba cấp độ trừu tượng được sử dụng: khái niệm, logic và vật lý. Mô hình hóa dữ liệu được hiểu là kỹ thuật ghi lại các đặc trưng của các phần tử dữ liệu trong một bản đồ mô tả dữ liệu được sử dụng trong một quá trình. Mô hình dữ liệu minh họa cách các phần tử dữ liệu được tổ chức và liên quan [18]. Các phương pháp luận mô hình hóa quan hệ đã thiết lập tốt các thủ tục, là kết quả của nhiều thập kỷ nghiên cứu [16]; tuy nhiên, đối với cơ sở dữ liệu NoSQL, các phương pháp lập mô hình, đặc biệt cho Dữ liệu lớn, là một chủ đề mới tiếp tục được nghiên cứu.

Mô hình hóa dữ liệu ở cấp độ khái niệm có liên quan chặt chẽ đến phạm vi của quy trình kinh doanh. Do đó, mô hình khái niệm là bất khả tri về công nghệ và độc lập với cơ sở dữ liệu sẽ được sử dụng. Do đó, các mô hình đã biết cho cơ sở dữ liệu quan hệ có thể được sử dụng trong cơ sở dữ liệu không quan hệ. Ở cấp độ logic, mô hình hóa tập trung vào mô hình dữ liệu sẽ được sử dụng. Đối với cơ sở dữ liệu NoSQL, việc mô hình hóa nhằm thể hiện cấu trúc dữ liệu của các mô hình hướng cột, hướng tài liệu, đồ thị hoặc khóa-giá trị, như đã mô tả trước đây. Ở cấp độ vật lý, mô hình sẽ đại diện cho lược đồ riêng của cơ sở dữ liệu đã chọn; nghĩa là việc triển khai cụ thể của cơ sở dữ liệu NoSQL, chẳng hạn như Cassandra hoặc MongoDB.

**2. Phương pháp**

Nghiên cứu SLR này được thực hiện dựa trên các hướng dẫn do Kitchenham đề xuất [1], dẫn đến sự phân chia ba giai đoạn: (1) lập kế hoạch cho nghiên cứu SLR, (2) tiến hành nghiên cứu SLR và (3) báo cáo nghiên cứu SLR. Hình 2 tóm tắt các giai đoạn trong nghiên cứu SLR của chúng tôi.

 Hình 2. Các giai đoạn ôn tập tài liệu có hệ thống.

Các hoạt động bao gồm trong giai đoạn đầu tiên là kiểm tra sự tồn tại của các nghiên cứu SLR khác về chủ đề quan tâm; xác định quy trình đánh giá sẽ được áp dụng để giảm thiểu các thành kiến có thể có từ các nhà nghiên cứu liên quan; khái niệm mục tiêu, luận cứ và câu hỏi nghiên cứu; và cuối cùng, xác định chiến lược sẽ theo đuổi. Trong giai đoạn thứ hai, các hoạt động là thu thập các nghiên cứu sơ cấp từ các nguồn được xác định bởi bước lập kế hoạch và sau đó áp dụng các tiêu chí để đưa vào để chỉ chọn những nghiên cứu liên quan đến mô hình Dữ liệu lớn. Cuối cùng, trong giai đoạn thứ ba, các hoạt động tương ứng là thực hiện phân tích sinh trắc học đầy đủ để cung cấp thông tin về tác giả và dữ liệu xuất bản trong một tài liệu duy nhất. Hơn nữa, mỗi nghiên cứu có liên quan được lựa chọn đã được ánh xạ đến một số khái niệm chính. Kết quả phân tích cho phép chúng tôi trả lời các câu hỏi nghiên cứu. Mỗi giai đoạn sẽ được trình bày kỹ hơn trong các phần phụ tiếp theo.

2.1. Lập kế hoạch Nghiên cứu SLR

Các mục tiêu chính của giai đoạn này là xác định sự cần thiết của một nghiên cứu SLR và phát triển một quy trình đánh giá.

2.1.1. Xác định nhu cầu cho một nghiên cứu SLR

Theo các đề xuất được cung cấp bởi một số công trình [1,19], chúng tôi tìm kiếm SLR và các ấn phẩm tương tự liên quan đến quản lý và lập mô hình Dữ liệu lớn, để xác minh xem có lỗ hổng nào có thể được che đậy với SLR được đề xuất trong công việc này hay không. Chúng tôi tìm thấy bốn tác phẩm có niên đại 2015 [5], 2016 [20], 2017 [15] và 2018 [10], được mô tả bên dưới

Ribeiro, Silva và Rodrigues da Silva [5] đã hoàn thành một cuộc khảo sát vào năm 2015 tập trung vào mô hình dữ liệu và phân tích dữ liệu. Mặc dù không phải là một nghiên cứu SLR, nhưng công trình mô tả một số khái niệm có liên quan đến các mô hình Dữ liệu lớn. Các tác giả đã xác định bốn mô hình dữ liệu chính cho Dữ liệu lớn — khóa-giá trị, tài liệu, cột rộng và biểu đồ — cũng được mô tả trong công trình của chúng tôi. Họ cũng trình bày một bản tóm tắt ngắn gọn về các mức trừu tượng, khái niệm, ngôn ngữ, công cụ mô hình hóa và hỗ trợ công cụ cơ sở dữ liệu. Tuy nhiên, nghiên cứu của họ không chi tiết như của chúng tôi, và họ cũng không trình bày phân tích sinh trắc học. Ví dụ: thiếu thông tin về các mô hình dữ liệu được sử dụng ở cấp độ khái niệm, logic và vật lý, các kỹ thuật được sử dụng để chuyển đổi theo các cấp độ trừu tượng khác nhau, các xu hướng nghiên cứu, nguồn, loại và mô hình tập dữ liệu nào cho Dữ liệu lớn là được nghiên cứu nhiều nhất và v.v. Hơn nữa, SLR của chúng tôi được cập nhật vào tháng 8 năm 2019. Tuy nhiên và tương tự như chúng tôi, nghiên cứu đặc biệt chú ý đến thực tế là mô hình Dữ liệu lớn và quản lý trong cơ sở dữ liệu phải được xem xét để nghiên cứu, lập tài liệu và phát triển, vì chúng chứng minh mô hình dữ liệu cần thiết như một phương tiện để cải thiện quá trình phát triển trong Dữ liệu lớn. Tuy nhiên, chúng không bao gồm các tiêu chí mà chúng tôi đã đề cập trước đây.

Sousa và Val Cura [20] bao gồm khung thời gian từ 2012 đến 2016. Họ trình bày một nghiên cứu SLR về mô hình logic cho cơ sở dữ liệu NoSQL. Các tác giả đề cử 12 bài báo và phân loại chúng theo các cấp độ khái niệm, logic và vật lý. Họ cũng xác định các đề xuất mô hình hóa cho cơ sở dữ liệu NoSQL, cho các đề xuất di chuyển và lớp của cơ sở dữ liệu NoSQL. Chúng tôi không coi đó là một tác phẩm hoàn chỉnh, vì trong nghiên cứu của mình, chúng tôi đã kiểm tra 1376 bài báo về quản lý và lập mô hình Dữ liệu lớn. Hơn nữa, họ kết luận rằng không có nghiên cứu nào về chuyển đổi mô hình dữ liệu từ mức khái niệm sang mức logic tồn tại vào thời điểm này, mặc dù phát hiện của chúng tôi cho thấy sự tồn tại của một số nghiên cứu liên quan đến nó.

Davoudian, Chen và Liu [10] trình bày một nghiên cứu kỹ lưỡng về tất cả các khái niệm và kỹ thuật được sử dụng trong cơ sở dữ liệu NoSQL; mô hình dữ liệu được sử dụng trong Dữ liệu lớn được mô tả nhưng trong công việc của chúng tôi, chúng tôi cũng trình bày một nghiên cứu sâu về các phương pháp mô hình Dữ liệu lớn. Đây được coi là một công trình có liên quan nhưng nó không hiển thị phân tích thư mục của tất cả các tác giả, hội nghị và tạp chí, trong số các thông tin liên quan khác để biết được xu hướng và lỗ hổng trong chủ đề nghiên cứu này. Ngoài ra, công việc của chúng tôi tập trung vào việc xem xét từng nghiên cứu được thực hiện trong nghiên cứu để cung cấp cho các nhà nghiên cứu hướng dẫn về các phương pháp tiếp cận hiện tại và hướng đi trong tương lai.

Wu, Sark và Zhu [15] xác định một số cơ sở dữ liệu NoSQL và tập trung để so sánh chúng theo mô hình dữ liệu và định lý của chúng, điều này chỉ ra rằng một hệ thống phân tán chỉ có thể đảm bảo đồng thời hai trong ba đặc tính sau: Tính nhất quán, Tính khả dụng và Dung sai phân vùng ( CHỮ HOA) [21]. Công việc này cũng không bao gồm một SLR như được trình bày trong tác phẩm này.

Hơn nữa, các cuộc khảo sát khác gần đây liên quan đến Dữ liệu lớn đã được công bố; ví dụ, một mô tả hiện đại về phương pháp luận được phát triển cho phân tích Dữ liệu lớn đa phương tiện và những thách thức, kỹ thuật, ứng dụng, chiến lược và triển vọng tương lai [22]. Một nghiên cứu khác trình bày và phân tích chi tiết giai đoạn hiện tại của môi trường và nền tảng Dữ liệu lớn và các thuật toán thu gom rác có sẵn [23]. Những công trình này không bao gồm phạm vi các câu hỏi nghiên cứu của chúng tôi về mô hình hóa và quản lý Dữ liệu lớn, cũng như không đạt được cùng mức độ chi tiết và chính xác.

Trong tiểu mục tiếp theo, chúng tôi trình bày chi tiết sự phát triển của đề cương đánh giá của chúng tôi, khẳng định các mục tiêu và sự biện minh của chúng tôi cũng như các câu hỏi nghiên cứu.

2.1.2. Phát triển một quy trình đánh giá

Một quy trình đánh giá là cần thiết để giảm thiểu bất kỳ sự thiên vị nào có thể xảy ra từ các nhà nghiên cứu và nó phải được xác định trước khi tiến hành SLR [1]. Vì vậy, trong giai đoạn này, chúng tôi đã hình thành phương pháp áp dụng. Đầu tiên, chúng tôi đề xuất các mục tiêu phát triển cụ thể và lý do tương ứng cho công việc của chúng tôi. Sau đó, chúng tôi đã xây dựng ba câu hỏi nghiên cứu với mục đích tóm tắt các bằng chứng hiện có về mô hình hóa và quản lý Dữ liệu lớn. Cuối cùng, chúng tôi đã xây dựng một chiến lược để thực hiện nghiên cứu SLR này một cách hiệu quả.

Mục tiêu và chứng minh

Mục tiêu đầu tiên là trình bày thông tin về nghiên cứu phù hợp nhất về mô hình hóa và quản lý Dữ liệu lớn trong một phân tích thư mục toàn diện. Nghiên cứu này bao gồm một số nghiên cứu từ các thư viện kỹ thuật số và thông tin chi tiết về tác giả, chi nhánh tổ chức của họ, quốc gia và chi tiết xuất bản, chẳng hạn như năm xuất bản và các yếu tố tác động của chúng trong Báo cáo trích dẫn của Tạp chí (JCR) và Xếp hạng Tạp chí Scimago (SJR) cho các tạp chí và trong Xếp hạng CORE cho các Hội nghị.

Dựa trên những phát hiện của chúng tôi, mục tiêu thứ hai là phát hiện các cách tiếp cận khác nhau cho mô hình Dữ liệu lớn được sử dụng trong các nghiên cứu khác nhau để xác định xu hướng và khoảng cách trong ba khái niệm chính, nguồn, mô hình và cơ sở dữ liệu. Nghiên cứu SLR được thực hiện trong nghiên cứu này có thể tập trung tất cả các nghiên cứu liên quan đến mô hình Dữ liệu lớn vào một tài liệu duy nhất, nhằm mang lại lợi ích cho ngành, học viện và cộng đồng.

Câu hỏi nghiên cứu

Giai đoạn này bao gồm giai đoạn quan trọng nhất của quá trình phát triển giao thức [1]. Do đó, chúng tôi đặc biệt lưu ý khi làm theo các đề xuất của Kitchenham. Thứ nhất, chúng tôi xác định ba tác nhân trong cộng đồng: (1) nhà nghiên cứu, (2) nhà phân tích thông tin và (3) nhà phát triển phần mềm, những người nghiên cứu, lập tài liệu và triển khai các giải pháp cho việc quản lý và lập mô hình Dữ liệu lớn trong cơ sở dữ liệu. Thứ hai, chúng tôi đã xem xét việc thu thập tất cả các phương pháp tiếp cận liên quan đến mô hình dữ liệu theo định hướng Dữ liệu lớn. Thứ ba, về kết quả, chúng tôi dự định tóm tắt các phát hiện và xác định các xu hướng và khoảng trống trong chủ đề được nghiên cứu. Nghiên cứu này đưa ra các câu hỏi nghiên cứu sau:

Câu hỏi nghiên cứu (RQ1): Số lượng bài báo được xuất bản về mô hình hóa và quản lý Dữ liệu lớn đã thay đổi như thế nào theo thời gian?

Cơ sở lý luận: Mối quan tâm của chúng tôi là hợp nhất, thông qua phân tích thư mục, tất cả các nỗ lực nghiên cứu cho chủ đề, cung cấp cho các nhà nghiên cứu khả năng biết tất cả thông tin về tác giả và dữ liệu xuất bản trong một tài liệu duy nhất. Do đó, người đọc có thể biết các nghiên cứu, trong chủ đề quan tâm của chúng tôi, đã phát triển như thế nào theo thời gian, ai là tác giả đã đóng góp đáng kể cho chủ đề, đâu là nghiên cứu được trích dẫn nhiều nhất và quốc gia nào quan tâm nhất đến chủ đề nghiên cứu này, cũng như các tạp chí và hội nghị nào liên quan đến chủ đề này và thư viện khoa học nào có phần lớn các nghiên cứu về mô hình hóa và quản lý Dữ liệu lớn. Ngoài ra, chúng tôi muốn biết liệu các nghiên cứu này chủ yếu được tài trợ bởi ngành công nghiệp hay học viện.

Câu hỏi nghiên cứu (RQ2): Có nghiên cứu nào tập trung vào các cách tiếp cận đối với dữ liệu bán cấu trúc và phi cấu trúc không và những kỹ thuật nào cần áp dụng?

Cơ sở lý luận: Mục tiêu của chúng tôi là tìm hiểu xem liệu các nghiên cứu có tập trung vào dữ liệu bán cấu trúc và dữ liệu phi cấu trúc, theo dữ liệu được chỉ định trong tiểu mục Khái niệm dữ liệu lớn, bao gồm hầu hết dữ liệu có sẵn. Ngoài ra, chúng tôi dự định trình bày những mô hình mà các nhà nghiên cứu đề xuất ở mỗi cấp độ trừu tượng hóa mô hình và xác định ba khái niệm chính: nguồn, mô hình hóa và cơ sở dữ liệu:

* For source: The dataset sources and data types;

• For modeling: The data abstraction levels, the data model proposed at conceptual, logical and physical levels, the techniques used for transformations between abstraction levels, the applied modeling language, the modeling methodology and the proposed tools for automatic model transformation;

• For database: The database type and the evaluation and performance comparison between models

Câu hỏi nghiên cứu (RQ3): Các xu hướng và khoảng trống trong mô hình hóa và quản lý Dữ liệu lớn là gì?

Cơ sở lý luận: Dựa trên dữ liệu thu được trong RQ2, mối quan tâm chính của chúng tôi là trình bày các giải pháp do các nhà nghiên cứu đề xuất trong chủ đề này trong một công trình tổng hợp. Mục tiêu là cho phép các nhà nghiên cứu tập trung nỗ lực của họ vào các lỗ hổng và giải pháp cho phép tiêu chuẩn hóa các phương pháp hiện đang tồn tại hoặc phương pháp mới.

Chiến lược

Chiến lược để tiến hành tổng hợp đầy đủ các nghiên cứu về chủ đề quan tâm bao gồm bốn hành động:

1.Tìm kiếm các nghiên cứu sơ cấp từ các thư viện số khoa học, chủ yếu xem xét liệu: (1) chúng chứa các tài liệu nghiên cứu đã được lập chỉ mục, (2) có tần suất cập nhật cơ sở dữ liệu cao hay không và (3) chúng xuất bản các nghiên cứu liên quan về chủ đề chúng ta quan tâm. Các nguồn được liệt kê dưới đây tuân thủ các yêu cầu mong muốn, để tập trung vào việc xem xét có hệ thống của chúng tôi về các nghiên cứu có liên quan:

• IEEE Xplore

• ScienceDirect

• Scopus

• Web of Science (WoS)

Hơn nữa, theo một nghiên cứu toàn diện, đánh giá chất lượng của 28 hệ thống tìm kiếm khoa học, Google Scholar không phù hợp làm hệ thống tìm kiếm chính, trong khi ScienceDirect, Scopus và WoS phù hợp để tổng hợp bằng chứng trong SLR [24]

2. Áp dụng các tiêu chí đưa vào các nghiên cứu chính, để chọn các nghiên cứu đó liên quan đến quản lý và lập mô hình Dữ liệu lớn, chúng tôi đã tiến hành tìm kiếm quy trình đối sánh cụm từ cụ thể trong tiêu đề, tóm tắt và từ khóa của bài báo. Dựa trên các câu hỏi nghiên cứu của chúng tôi, hai thuật ngữ tìm kiếm chính đã được rút ra: dữ liệu lớn và mô hình. Các thuật ngữ được chọn sau khi kết hợp một số tùy chọn, để có được một số lượng đáng kể các bài báo và các thuật ngữ này bao gồm phần lớn các nghiên cứu đề cập đến chủ đề nghiên cứu của chúng tôi.

Do thực tế là các thư viện kỹ thuật số được chọn không có chung một cú pháp tìm kiếm, chúng tôi đã liệt kê tất cả các chuỗi tìm kiếm được áp dụng trong mỗi một. Từ “mô hình” đã được sử dụng vì một số nghiên cứu sử dụng thuật ngữ này khi đề cập đến mô hình:

• IEEE Xplore - (((“Tiêu đề tài liệu”: “dữ liệu lớn” và “mô hình dữ liệu”) HOẶC “Tóm tắt”: “dữ liệu lớn” và “mô hình dữ liệu”) HOẶC “Điều khoản chỉ mục”: “dữ liệu lớn” và “dữ liệu người mẫu")

• ScienceDirect — Tiêu đề, tóm tắt, từ khóa: “dữ liệu lớn” và “mô hình dữ liệu”

• Scopus — TITLE-ABS-KEY (“dữ liệu lớn” VÀ “mô hình dữ liệu”)

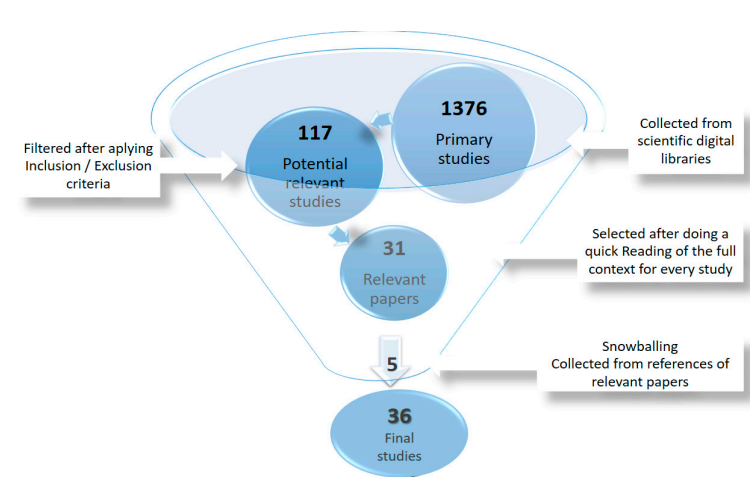
• WoS — TS = (“dữ liệu lớn” VÀ “mô hình dữ liệu”). TS liên quan đến các trường Chủ đề bao gồm tiêu đề, tóm tắt và từ khóa.

Đối với các tiêu chí bao gồm, chỉ các nghiên cứu được viết bằng tiếng Anh và được xuất bản trong các hội nghị hoặc tạp chí mới được xem xét. Mặc dù không có yếu tố giới hạn ngày tháng nào được xác định trong tiêu chí tìm kiếm của chúng tôi, nhưng có thể thấy rằng không có kết quả nào trước năm 2010 được trả về bởi bất kỳ thư viện khoa học đã chọn nào. Những kết quả này phù hợp với Hình 1, trong đó một báo cáo từ Google Xu hướng chứng minh rằng thuật ngữ Dữ liệu lớn bắt đầu trở nên phổ biến vào năm 2011;

3. Xem lại các nghiên cứu lần thứ hai thông qua việc đọc nội dung của bài báo cho phép chúng tôi loại bỏ những nội dung không liên quan đến ngữ cảnh của chủ đề chúng ta quan tâm;

4. Kỹ thuật ném bóng tuyết đã được áp dụng để xác định vị trí các bài báo có liên quan bổ sung theo các tài liệu tham khảo hiện có từ trong các nghiên cứu đã được xem xét.

2.2. Tiến hành Nghiên cứu SLR

Việc tìm kiếm nghiên cứu SLR được thực hiện vào tháng 8 năm 2019. Hình 3 cung cấp đại diện của quá trình lựa chọn được áp dụng cho các nghiên cứu. Các nghiên cứu trùng lặp đã bị loại bỏ khỏi giai đoạn nghiên cứu có khả năng liên quan

Hình 3. Quy trình lựa chọn.

2.2.1. Giai đoạn hòa nhập

Trong giai đoạn này, chúng tôi đã chọn các nghiên cứu liên quan đến mô hình Dữ liệu lớn và kiểm tra tiêu đề, tóm tắt và từ khóa của chúng theo chiến lược đã lên kế hoạch trước đó. Chúng tôi chỉ xem xét các nghiên cứu bằng tiếng Anh giải quyết các câu hỏi nghiên cứu của chúng tôi và được xuất bản trong các hội nghị hoặc tạp chí. Mục tiêu chính của chúng tôi là xác định các cách tiếp cận khác nhau để lập mô hình và quản lý dữ liệu trong kho dữ liệu một cách tổng quát cho các loại dữ liệu khác nhau ở ba cấp độ trừu tượng. Kết quả là, 1 bài báo tiếng Trung và 27 bài báo tương ứng với sách, chương sách, thư từ, ghi chú hoặc bài xã luận, đã bị loại bỏ. Ngoài ra, chúng tôi cũng loại bỏ 1259 bài báo, mặc dù đề cập đến mô hình dữ liệu, đề cập đến các ứng dụng cụ thể hoặc không liên quan đến tính ổn định của dữ liệu nhưng liên quan đến nhập dữ liệu, hồ dữ liệu hoặc phân tích dữ liệu. Từ giai đoạn này, 117 nghiên cứu đã được chấp nhận.

2.2.2. Giai đoạn lựa chọn

Ở giai đoạn này, việc xem xét nhanh toàn bộ nội dung của mọi nghiên cứu cho phép chúng tôi chỉ chọn những nghiên cứu liên quan đến mô hình Dữ liệu lớn. Điều này dẫn đến việc chấp nhận 31 nghiên cứu, loại bỏ 70 bài báo và lọc 16 công trình trùng lặp. Ở giai đoạn này, chúng tôi đã loại bỏ các giấy tờ trùng lặp. Sau khi quét toàn bộ nội dung của 31 nghiên cứu được chọn này, chúng tôi cũng đưa vào năm bài báo mới sau khi đánh giá quả cầu tuyết. Cuối cùng, 36 nghiên cứu đã tạo thành kho tài liệu cuối cùng của chúng tôi để báo cáo nghiên cứu SLR.

2.3. Báo cáo Nghiên cứu SLR

Mục tiêu của bước này là trả lời các câu hỏi nghiên cứu được nêu ra trong đề cương xem xét. Với mục đích này, nghiên cứu này được chia thành ba phần. Trong phần đầu tiên, chúng tôi thực hiện phân tích thư mục để trả lời RQ1; trong phần thứ hai, chúng tôi trình bày tổng quan tài liệu với dữ liệu phù hợp nhất của các phương pháp tiếp cận trong một ma trận khái niệm để trả lời RQ2; và trong phần thứ ba, chúng tôi thảo luận về các xu hướng và khoảng trống để trả lời RQ3. Phần 3 trình bày chi tiết các kết quả thu thập được từ các hoạt động được mô tả trong giai đoạn này.

**3. Kết quả**

Trong phần này, chúng tôi trả lời các câu hỏi nghiên cứu thông qua các hoạt động dưới đây:

1. Phân tích thư mục, để thu thập thông tin về các tác giả và dữ liệu xuất bản, các tác giả và quốc gia có nhiều đóng góp hơn trong chủ đề, tác động của các nghiên cứu được chọn và cách nghiên cứu đã phát triển trong suốt thời gian, cũng như các tạp chí và hội nghị thủ tục nơi các nghiên cứu được xuất bản

2. Tổng quan tài liệu để lập bản đồ các nghiên cứu theo ba khái niệm chính - nguồn, mô hình hóa và cơ sở dữ liệu - trong một ma trận khái niệm. Trong khái niệm nguồn, chúng tôi phân tích các nguồn và kiểu dữ liệu của tập dữ liệu. Sự bền vững. Trong khái niệm mô hình hóa, chúng tôi phân tích các mức trừu tượng hóa dữ liệu, các mô hình dữ liệu được đề xuất ở các mức khái niệm, logic và vật lý, các kỹ thuật được sử dụng để thực hiện chuyển đổi giữa các mức trừu tượng, ngôn ngữ mô hình hóa được áp dụng, phương pháp mô hình hóa và các công cụ mô hình hóa được đề xuất. Ở khái niệm cơ sở dữ liệu, chúng tôi phân tích kiểu và tiến hành đánh giá và so sánh hiệu suất giữa các mô hình;

3. Một cuộc thảo luận để xác định các xu hướng và khoảng trống trong việc quản lý và lập mô hình Dữ liệu lớn.

3.1. Phân tích Bibliometric

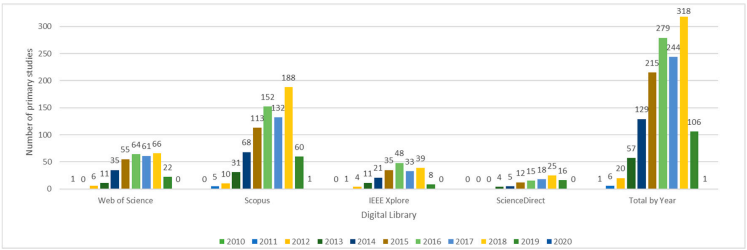
Mục tiêu của phân tích này là để trả lời RQ1. Để trả lời phần đầu về cơ sở lý luận của câu hỏi này, chúng tôi phân tích kết quả của các giai đoạn thu nhận và lựa chọn. Trong Hình 4, chúng tôi tóm tắt kết quả của giai đoạn bao gồm và nêu bật một số phát hiện:

• Tốc độ tăng trưởng trung bình hàng năm của các bài báo đã xuất bản tuân theo Công thức (1)

y = 30,309x - 29,2

• Trước năm 2010, không có nghiên cứu liên quan nào về mô hình Dữ liệu lớn được công bố

• Kể từ năm 2015, số lượng các nghiên cứu đã tăng lên đáng kể và trong năm 2018, đã có 318 bài báo được xuất bản. Năm 2019, đã có 106 xuất bản trước tháng 8

• Scopus xếp hạng cao nhất trong số các nguồn được xem xét, với 760 tác phẩm được thu thập, tiếp theo là WoS với 321 tác phẩm, IEEE Xplore với 200 và ScienceDirect với 95

Hình 4. Số lượng nghiên cứu chính theo năm và nguồn.

Kết quả của giai đoạn lựa chọn được trình bày trong Hình 5, được sắp xếp theo nguồn và năm. Chúng ta có thể làm nổi bật findi sau

1. Trước năm 2013, không có nghiên cứu liên quan nào được tìm thấy;

2. Năm mà chúng tôi tìm thấy số lượng nghiên cứu nhiều nhất về mô hình Dữ liệu lớn là năm 2018. Tuy nhiên, điều quan trọng cần làm là năm 2019 đang diễn ra và cuối cùng có thể có nhiều nghiên cứu hơn nữa.

3. Với 27 bài báo, Scopus là nguồn có số lượng nghiên cứu liên quan cao nhất, tiếp theo là WoS và IEEE Xplore với hai bài mỗi loại. ScienceDirect không báo cáo bất kỳ bài báo liên quan nào về chủ đề này.

Bảng 3 tóm tắt dữ liệu chính của từng bài báo được chọn, trong số đó, tài liệu tham khảo, tên tác giả đầu tiên và công ty liên kết, quốc gia nơi nghiên cứu được thực hiện, nhận dạng của tạp chí hoặc hội nghị, thư viện kỹ thuật số, năm xuất bản, số lượng trích dẫn trong Scopus, ứng dụng kiến thức và sự tồn tại của kinh phí.

Bảng 3. Phân tích Bibliometric.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tên liệu tham khảo | Tên tác giả đầu tiên | Liên kết đầu tiên của tác giả | Quốc Gia | ID tạp chí/ Hội nghị | Thư Viện Số | Năm Xuất Bản | Trích dẫn từ Scopus | Ứng dụng kiến thức | Kinh phí |
| 25 | Jie Song | Cao đẳng phần mềm Đông Bắc | Trung Quốc | J1 | Scopus | 2019 | 0 | Học Viện | Có |
| 26 | Laurent Thiry | Đại học Haute Alsace | Pháp | J2 | Scopus | 2018 | 0 | Học Viện | N/A |
| 27 | Victor martins de Sousa | Đại học UNIFACCAMP | Brazil | C1 | Scopus | 2018 | 1 | Học Viện | Có |
| 28 | Igor Zeˇcevi´c | Đại học Novi Sad | Serbia | J3 | Scopus | 2018 | 2 | Học Viện | Có |
| 29 | Antonio M. Rinaldi | Đại học Naples Federico II | Ý | C2 | Scopus | 2018 | 1 | Học Viện | N/A |
| 30 | Shady Hamouda | Cao đẳng Công nghệ Emirates | Các Tiểu Vương Quốc Ả Rập Thống Nhất | C3 | Scopus | 2018 | 1 | Học Viện | N/A |
| 31 | Dippy Aggarwal | Đại học Cincinnati | Mỹ | J4 | Scopus | 2018 | 0 | Học Viện | N/A |
| 32 | Alfonso de la Vega | Đại học Cantabria | Tây Ban Nha | C4 | Scopus | 2018 | 0 | Học Viện | Có |
| 33 | Xu Chen | Đại học Bắc Minzu | Trung Quốc | J5 | Scopus | 2018 | 0 | Học Viện | N/A |
| 18 | Maribel Yasmina Santos | Đại học Minho | Bồ Đào Nha | J6 | Scopus | 2017 | 10 | Học Viện | Có |
| 34 | KwangchuShin | Đại học Kook Min | Hàn Quốc | J7 | Scopus | 2017 | 7 | Học Viện | Có |
| 35 | Fatma Abdelhedi | Đại học Toulouse Capitole | Pháp | C5 | Scopus | 2017 | 1 | Học Viện | N/A |
| 36 | Aravind Mohan | Wayne State University | Mỹ | C6 | Scopus | 2016 | 7 | Học Viện | Có |
|  |  |  |  |  |  |  |  |  |  |

**Bảng 3 : Cont**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| Tên liệu tham khảo | Tên tác giả đầu tiên | Liên kết của tác giả | Quốc Gia | ID tạp chí/ Hội nghị | Thư Viện Số | Năm Xuất Bản | Trích dẫn từ Scopus | Ứng dụng kiến thức | Kinh phí |
| 37 | Massimo Villari | Đại học Messina | Ý | C7 | Scopus | 2016 | 2 | Học Viện | N/A |
| 38 | Maribel Yasmina Santo | Đại học Minho | Bồ Đào Nha | C8 | Scopus | 2016 | 10 | Học Viện | Có |
| 39 | Maribel Yasmina Santo | Đại học Minho | Bồ Đào Nha | J8 | Scopus | 2016 | 8 | Học Viện | Có |
| 40 | Ganesh B. Solanke | PCCoE, Nigd | Ấn Đụ | C9 | Scopus | 2018 | 0 | Học Viện | N/A |
| 41 | Vincent Reniers | Đại học  KU Leuven | Bỉ | C10 | Scopus | 2018 | 0 | Học  Viện | Có |
| 42 | Fatma Abelhedi | Đại học  Toulouse Capitole University | Pháp | C11 | Scopus | 2016 | 4 | Học Viện | N/A |
| 43 | Max Chevalier | Đại Học | Pháp | C12 | Scopus | 2016 | 6 | Học Viện | ANRT |
| 44 | Shreya Banerjee | Đại Học National Institute of Technology | Ấn Độ | C13 | Scopus | 2015 | 6 | Học Viện | N/A |
| 16 | Artem Chebotko | DataStax INC. | Mỹ | C14 | Scopus | 2015 | 43 | Công nghiệp | Có |
| 45 | Wendou Feng | Đại học Guangxi | Trung Quốc | C15 | Scopus | 2015 | 2 | Học Viện | Có |
| 46 | Ling Chen | Đại học  Zhejiang | Trung Quốc | C16 | Scopus | 2015 | 1 | Học Viện | Có |
| 47 | Max chevalier | Dại học Toulouse | Pháp | C17 | Scopus | 2015 | 14 | Học Viện | N/A |
| 48 | Dewi W. Wardani | Đại học  Sebelas Maret | Indonesia | C18 | Scopus | 2014 | 7 | Học Viện | N/A |
| 49 | Ming Zhe | Đại học công nghệ Hubei | Trung Quốc | C19 | Scopus | 2013 | 0 | Học Viện | N/A |
| 50 | Mohamed Nadjib Mami | Đại học Bonn | Đức | C20 | Scopus | 2016 | 5 | Học Viện | Có |
| 51 | Dan Han | Đại học Alberta | Canada | C21 | WoS | 2013 | 0 | Học Viện | Có |
| 52 | Zhiyun Zheng | Đại học Zhengzhou | Trung Quốc | C22 | IEEE | 2014 | 1 | Học Viện | Có |
| 53 | Dongqi We | Đại học  Geosciences | Trung Quốc | C23 | IEEE | 2014 | 3 | Học Viện | N/A |
| 14 | Karamjit Kaur | Đại học Thapar | Ấn Độ | C24 | IEEE | 2013 | 59 | Học Viện | N/A |
| 54 | Michael J. Mior | Đại học Waterloo | Canada | J1 | IEEE | 2017 | 12 | Học Viện | Có |
| 55 | Max Chevalier | Đại học Toulouse | Pháp | C25 | Scopus | 2016 | 4 | Học Viện | Có |
| 56 | Harley Vera | Đại học của Brasília | Brazil | C26 | Scopus | 2015 | 8 | Học Viện | N/A |
| 57 | Robert T. Mason | Đại học Regis | Mỹ | C27 | N/A | 2015 | 0 | Học Viện | N/A |

**3.1.1 Tác giả**

Trong hình số 6, có thể xác minh tên của các tác giả đầu tiên có đóng góp lớn cho chủ đề này. Như  vậy, Maribel Yasmina Santos đến từ Bồ Đào Nha và Max Chevalier dến từ Pháp chiếm vị trí đầu tiên với ba bài viết mỗi người và Fatma abdelhedi đế từ Pháp đứng vị trí thứ 2 với 2 bài báo. Theo dữ liệu quan sát cho thấy, hai nghiên cứu của Santos đã được xuất bản vào năm 2016 và một nghiên cứu khác vào năm 2017. Nghiên cứu của họ được thực hiện với sự cộng tác của Đại học Minho và đây là nghiên cứu duy nhất từ Bồ Đào Nha được trình bày trong tập hợp các nghiên cứu cuối cùng.

Chart

Description automatically generated

**Hình 6 : Đóng góp của tác giả.**

Về Pháp, các tác giả Chevalier, Abdelhedi và Laurent Thiry đã nghiên cứu chủ đề này và bổ sung tổng cộng sáu đóng góp, một xuất bản năm 2015, 3 xuất bản năm 2016 và một xuất bản năm 2017 và 2018. Nghiên cứu của họ được liên kết với Đại học Haute Alsace và Đại học Toulouse. Các quốc gia như Mỹ, Trung Quốc và Ấn Độ đã có một số đóng góp từ các tác giả khác nhau. Đối với Mỹ, có bốn bài báo, năm 2015 Robert Mason của Đại học Regis và Artem Chebotko từ DataStax Inc. đã trình bày mỗi người một bài báo, vào năm 2016 Aravind Mohan từ Wayne Đại học Bang và năm 2018, Dippy Aggarwal của Đại học Cincinnati, cũng đã trình bày các phương pháp tiếp cận của họ. Ở Trung Quốc, sáu tác giả và tổ chức đã đóng góp, vào năm 2013 Ming Zhe Hubei của Đại học Công nghệ, năm 2014 Donggi Wei từ Đại học Khoa học Địa chất, ở 2015 Ling Chen từ Đại học Chiết Giang và Wenduo Feng của Đại học Quảng Tây, trên 2018 Xu Đại học Chen của North Minzu và vào năm 2019, Jie Song từ Đại học Phần mềm, Đông Bắc. Có ba nghiên cứu từ Ấn Độ được xuất bản bởi Karamjit Kaur từ Đại học Thapar vào năm 2013, Shreya Banerjee của Viện Công nghệ Quốc gia năm 2015 và Ganesh Soanke từ PCCoE, Nigdi năm 2018. Tổng cộng, 15 tổ chức khác nhau, một từ ngành công nghiệp và 14 từ học viện đã trình bày các công trình liên quan và có thể thấy rằng ngay cả trong năm 2018 và 2019, chủ đề này vẫn đang được điều tra tích cực.

**3.1.2. Quốc gia và Năm**

Trong phần này, sau khi loại bỏ 16 kết quả trùng lặp, 101 trong số 117 bài báo được thu thập sau giai đoạn đưa vào được lấy làm mẫu. Những bài báo này chứa nghiên cứu liên quan đến lĩnh vực chúng tôi quan tâm và cho phép chúng tôi phân tích nhiều bài báo hơn. Hình 7 cho thấy các quốc gia dẫn đầu về chủ đề được quan tâm là Mỹ và Trung Quốc, với lần lượt là 17 và 16 bài báo. Quốc gia nơi Chevalier thực hiện nghiên cứu của mình, Pháp, đứng ở vị trí thứ ba với tám bài báo và mỗi nước có bảy ấn phẩm, Ý và Tây Ban Nha chiếm vị trí thứ tư. Cuối cùng, Đức chiếm vị trí thứ năm với sáu nghiên cứu.

A screenshot of a computer

Description automatically generated with low confidence**Hình 7: Đóng góp theo năm và quốc gia.**

Đối với Mỹ, bốn bài báo đã được công bố vào năm 2014, một bài báo năm 2015, sáu bài báo năm 2016, hai bài báo năm 2017, ba bài báo năm 2018 và một bài báo năm 2019. 2014, có nhiều đóng góp nhất vào năm 2016 và tiếp tục đến năm 2019. Về Trung Quốc, bài báo đầu tiên của họ được xuất bản vào năm 2013, tiếp theo là ba bài báo vào năm 2014, hai bài báo vào năm 2015, trong các năm 2016, 2017 và 2018, ba bài báo mỗi năm và một bài báo vào năm 2019. Có thể thấy rằng Trung Quốc đã bắt đầu nghiên cứu vào năm 2013 và vẫn đang tiếp tục điều tra chủ đề này. Điều đáng nói là các ấn phẩm bài báo liên tục được quan sát từ năm 2014 đến năm 2018. Pháp bắt đầu vào năm 2015 với một bài báo, bốn bài báo vào năm 2016, một bài báo vào năm 2017 và hai bài báo vào năm 2018. Quốc gia này bắt đầu nghiên cứu vào năm 2015 và năm 2016 là năm với nhiều đóng góp hơn. Ý và Tây Ban Nha cũng bắt đầu nghiên cứu vào năm 2015. Ý đã xuất bản nhiều bài báo hơn trong năm 2016 và Tây Ban Nha vào năm 2018. Về năm 2019, chỉ có Tây Ban Nha đã xuất bản một bài báo. Năm 2013, Đức bắt đầu nghiên cứu với một bài báo và bài báo xuất bản cuối cùng của nước này được tìm thấy vào năm 2018.

Tóm lại, từ năm 2015 trở đi, nhiều quốc gia bắt đầu đóng góp vào sản xuất khoa học về chủ đề này, tăng gấp đôi số lượng bài báo được xuất bản vào năm 2016. Trong năm 2018 và 2019, xu hướng này vẫn tiếp tục. Tuy nhiên, năm 2019 vẫn đang tiếp diễn; do đó, có khả năng nhiều nghiên cứu sẽ được công bố trước cuối năm nay.

**3.1.3. trích dẫn**

Bảng 3 trình bày số lần trích dẫn của các nghiên cứu trong Scopus. Hình 8 trình bày bài báo có tác động lớn nhất, có 59 trích dẫn và được xuất bản bởi Karamjit Kaur từ Ấn Độ, tiếp theo là bài báo của Artem Chebotko từ Mỹ, với 43 trích dẫn. Điều quan trọng cần nhấn mạnh là cả hai tác giả cũng thuộc về các quốc gia có nhiều đóng góp hơn.

Chart, histogram

Description automatically generated

**Hình 8: Số lượng trích dẫn trong Scopus.**

Bài báo được trích dẫn nhiều nhất có tiêu đề "Mô hình hóa và truy vấn dữ liệu trong cơ sở dữ liệu NoSQL" và được xuất bản vào năm 2013. Bài báo được trích dẫn nhiều thứ hai có tiêu đề "Phương pháp lập mô hình dữ liệu lớn cho Apache Cassandra" và được xuất bản vào năm 2015. Thông tin chi tiết về các ấn phẩm này được trình bày trong phần SLR. Cũng có thể lưu ý, theo Bảng 3, 97,22% bài báo thuộc về học viện và 52,78% bài báo được tài trợ. Theo tiêu chí của chúng tôi, chủ đề này được coi là có mức độ liên quan cao vì kinh phí được phân bổ trong các dự án dành cho nghiên cứu. Bảng 4 và 5 cung cấp thông tin cho người đọc về các tạp chí và hội nghị nơi các nghiên cứu được công bố; yếu tố tác động của chúng cũng được trình bày trong JCR và SJR, và đối với các hội nghị,

thứ hạng của họ. Điều quan trọng cần nhấn mạnh là 75% các nghiên cứu đã được trình bày trong các hội nghị, do đó chúng ta có thể dự đoán rằng trong năm hiện tại có những nghiên cứu vẫn đang được tiến hành, chưa đạt đến giai đoạn cuối cùng.

**Bảng 4 : Thông tin của các tạp chí nơi các nghiên cứu liên quan đã được trình bày.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID Tạp Chí | Tên tạp chí | Quốc Gia | JCR IF | SJR | ID nghiên cứu |
| J1 | Giao dịch của IEEE về Kiến thức và Kỹ thuật dữ liệu | Mỹ | 3.86 | 1.1 | 1.33 |
| J2 | Tạp chí dữ liệu lớn | Vương Quốc Anh | N/A | 1.1 | 2 |
| J3 | Hệ thống thông tin doanh nghiệp | Vương Quốc Anh | 2.12 | 0.7 | 4 |
| J4 | Những tiến bộ trong hệ thống thông tin và máy tính | Đức | N/A | 0.2 | 7 |
| J5 | Màng lọc | Serbia | 0.79 | 0.4 | 9 |
| J6 | Tạp chí phân tích quản lý | Vương Quốc Anh | N/A | N/A | 10 |
| J7 | Tạp chí quốc tế về ứng dụng nghiên cứu kỹ thuật | Ấn Độ | N/A | 0.1 | 11 |
| J8 | Ghi chú bài giảng trong Khoa học Máy tính | Đức | 0.4 | 0.3 | 16 |

**Bảng 5: Cont**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| ID Hội nghị | Tên hội nghị | Xếp hạng Core 2018 | ID Nghiên cứu |
| C1 | Hội nghị quốc tế lần thứ 20 về thông tin  Tích hợp và các ứng dụng dựa trên web và  Dịch vụ | C | 3 |
| C2 | Hội nghị quốc tế lần thứ 10 về quản lý  Hệ sinh thái kỹ thuật số | Không có xếp hạng | 5 |
| C3 | Hội nghị quốc tế 2017 về đổi mới và ứng dụng dữ liệu lớn | Không có xếp hạng | 6 |
| C4 | Hội nghị quốc tế lần thứ 8 về mô hình và dữ liệu kỹ thuật | Không có xếp hạng | 8 |
| C5 | Hội nghị quốc tế lần thứ 19 về doanh nghiệp  Hệ thống thông tin | C | 12 |
| C6 | Đại hội quốc tế IEEE thứ 5 về Dữ liệu lớn | Không có xếp hạng | 13 |
| C7 | Hội nghị chuyên đề IEEE 2016 về máy tính và giao tiếp | B | 14 |
| C8 | Hội nghị C\* quốc tế lần thứ 9 về Khoa học máy tính và Kỹ thuật phần mềm | Không có xếp hạng | 15 |
| C9 | Hội nghị quốc tế 2017 về máy tính, giao tiếp, kiểm soát và tự động hóa | Không có xếp hạng | 17 |
| C10 | Hội nghị quốc tế IEEE 2017 về dữ liệu lớn | Không có xếp hạng | 18 |
| C11 | Hội nghị chung quốc tế lần thứ 8 về khám phá kiến ​​thức, kỹ thuật kiến ​​thức và quản lý kiến ​​thức | C | 19 |
| C12 | Hội nghị quốc tế lần thứ 18 về hệ thống thông tin doanh nghiệp | C | 20 |
| C13 | Hội nghị quốc tế IEEE 2015 về tin học công nghiệp | Không có xếp hạng | 21 |
| C14 | Đại hội quốc tế IEEE lần thứ 4 về dữ liệu lớn | Không có xếp hạng | 22 |
| C15 | Hội nghị quốc tế IEEE 2015 về Thành phố thông minh/Socialcom/Sustaincom cùng với Datacom | Không có xếp hạng | 23 |
| C16 | Hội nghị quốc tế IEEE 2015 về Dữ liệu lớn đa phương tiện | Không có xếp hạng | 24 |
| C17 | Hội nghị quốc tế lần thứ 17 về phân tích dữ liệu lớn và khám phá kiến ​​thức | Không có xếp hạng | 25 |
| C18 | Hội nghị quốc tế 2014 về máy tính, kiểm soát, tin học và các ứng dụng của nó | Không có xếp hạng | 26 |
| C19 | Hội nghị quốc tế 2013 về Khoa học và ứng dụng máy tính | Không có xếp hạng | 27 |
| C20 | Hội nghị quốc tế lần thứ 18 về phân tích dữ liệu lớn và khám phá kiến ​​thức | Không có xếp hạng | 28 |
| C21 | Hội nghị quốc tế thứ sáu của IEEE 2013 về điện toán đám mây | B | 29 |
| C22 | Đại hội quốc tế IEEE lần thứ 3 về Dữ liệu lớn | Không có xếp hạng | 30 |
| C23 | Hội nghị quốc tế năm 2014 về máy tính cho nghiên cứu và ứng dụng không gian địa lý | Không có xếp hạng | 31 |
| C24 | Hội nghị quốc tế IEEE 2013 về dữ liệu lớn | Không có xếp hạng | 32 |
| C25 | Hội nghị quốc tế thứ 10 của IEEE về các thách thức nghiên cứu trong khoa học thông tin | B | 34 |
| C26 | Hội thảo quốc tế hàng năm lần thứ 2 về quản lý thông tin và dữ liệu lớn | B | 35 |
| C27 | Thông báo Hội nghị Giáo dục Khoa học & CNTT | C | 36 |
|  |  |  |  |

**3.1.4. Tạp chí**

Chúng tôi trình bày trong Bảng 4 Danh sách các tạp chí nơi các nghiên cứu có liên quan được lựa chọn được công bố.Bảng chứa số nhận dạng tạp chí được chỉ định, tên tạp chí, quốc gia tạp chí, tác độngYếu tố (IF) trong JCR và SJR và ID nghiên cứu liên quan. Chúng tôi đã xem xét điều quan trọng là hiển thị JCRNếu và SJR, vì chúng là các chỉ số liên quan đến chất lượng nghiên cứu theo số lượngvề các trích dẫn của các nghiên cứu được công bố và tầm quan trọng của chúng trong nghiên cứu khoa học

**3.1.5. Hội nghị**

Chúng tôi trình bày trong Bảng 5 Các chi tiết của các hội nghị nơi một số nghiên cứu có liên quan được trình bày.Định danh hội nghị được chỉ định, tên hội nghị, xếp hạng cốt lõi và các nghiên cứu tương ứngĐịnh danh được liệt kê. Chúng tôi đã sử dụng Hiệp hội Giáo dục và Nghiên cứu Điện toán Xếp hạng Hội nghị củaAustralasia (Core), 2018. Xếp hạng này được tạo ra bởi một hiệp hội của các khoa khoa học máy tínhtừ các trường đại học ở Úc và New Zealand. Hiệp hội này cung cấp bảng xếp hạng hội nghị trong các ngành tính toán dựa trên sự kết hợp của các chỉ số, bao gồm tỷ lệ trích dẫn, gửi giấy vàTỷ lệ chấp nhận. Phạm vi xếp hạng được thể hiện bằng các chữ cái A\*, A, B và C - A\* là tốt nhất và c tồi tệ nhất.

Thông qua phân tích được thực hiện, câu hỏi nghiên cứu RQ1 được trả lời chi tiết quan trọng. theo thứ tự để trả lời hai câu hỏi nghiên cứu tiếp theo, mỗi bài báo được chọn là có liên quan phân tích, sau khi đọc đầy đủ của mỗi người trong số họ.

**3.2. Đánh giá văn học có hệ thống**

Mục tiêu của phần này là trả lời câu hỏi nghiên cứu thứ hai, RQ2. Để tuân thủ điều nàyMục tiêu, chúng tôi dựa vào ma trận khái niệm [58] được biên soạn trong Phụ lục A. Ở đó, chúng tôi tổng hợp tài liệu về mỗi một trong số 36 bài báo bao gồm tập đoàn nghiên cứu cuối cùng. Tiếp theo, mỗi khái niệm chínhmà chúng tôi đã đề cập trong DSLR này sẽ được mô tả. Chủ yếu, ba miền được phân tích:

1. Nguồn

2. Mô hình hóa

3. Cơ sở dữ liệu

3.2.1. Nguồn

Tại phần này, chúng tôi phân tích các nguồn dữ liệu và loại dữ liệu. Các nguồn dữ liệu cho phép chúng tôibiết liệu nghiên cứu đã được thực hiện trong môi trường thực tế hay trong môi trường thử nghiệm vớidữ liệu mô phỏng. Việc sử dụng bộ dữ liệu trong thế giới thực rất quan trọng để xác minh việc tuân thủ khối lượng,Vận tốc, tính xác thực và giá trị đặc trưng cho dữ liệu lớn. Như đã đề cập trong phần 1.1.4, theo mộtNghiên cứu [11], 90% dữ liệu tồn tại trên thế giới tương ứng với dữ liệu bán cấu trúc và không cấu trúc.

Vì lý do này, khái niệm này cho phép chúng tôi xác nhận nếu nghiên cứu được định hướng theo các loại dữ liệu này.Nguồn đặt dữ liệu sau khi phân tích 36 bài viết được chọn, người ta đã xác định rằng 22 bài viết đã sử dụng bộ dữ liệu mẫuĐối với các đề xuất của họ, 10 bài viết đã sử dụng bộ dữ liệu trong thế giới thực và bốn bài không trình bày bất kỳ ví dụ nào vềCác giải pháp vì lý do này, họ không đề cập đến bất kỳ loại dữ liệu nào. Do đó, nó đã kết luận rằng hơn 50% các nghiên cứu liên quan đã không trình bày các đề xuất được xác minh của họ với các bộ dữ liệu trong thế giới thực.Bằng cách không sử dụng

bộ dữ liệu trong thế giới thực, hành vi của các giải pháp trong môi trường sản xuấtkhông thể được xác minh.

Các bộ dữ liệu trong thế giới thực chính được sử dụng trong các nghiên cứu là dữ liệu cảm biến, siêu dữ liệu hình ảnh,Các trang web xuất bản và tài liệu điện tử, như Hình 9 trình bày. Như chúng ta có thể thấy, những bộ dữ liệu đó làđược phân loại là dữ liệu phi cấu trúc mà chúng tôi phân tích trong khái niệm tiếp theo. Ngoài ra, xử lý hàng loạt làđược sử dụng bởi hầu hết các phương pháp này, trong khi xử lý thời gian thực được đề xuất bởi một nghiên cứu về dữ liệu mô hình hóa cho các chuyến bay thương mại ở Hoa Kỳ [18]

Chart, bar chart, waterfall chart

Description automatically generated

**Hình 9: Các loại bộ dữ liệu trong thế giới thực được sử dụng.**

Trong Bảng 6, chúng tôi tóm tắt mười nghiên cứu với các bộ dữ liệu trong thế giới thực được trình bày trên Hình 9, theo thứ tựđể biết theo ứng dụng nào, các nghiên cứu đã được xây dựng và liệu chúng có tuân thủKhối lượng, vận tốc, tính xác thực và đặc điểm giá trị. Đặc điểm đa dạng được phân tích trongKiểu dữ liệu tiểu mục tiếp theo. Từ Bảng 6, chúng ta có thể thấy rằng 90% các nghiên cứu không biện minh cho đặc tính vận tốc

**Bảng 6 :Phân tích các bộ dữ liệu trong thế giới thực được sử dụng trong các nghiên cứu có liên quan.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Bộ dữ liệu trong thế giới thực | ID | Miền | Dung lượng | Tốc độ | Tính xác thực | Giá trị |
| Tài liệu điện tử | 27 | Tài liệu điện tử của chính phủ điện tử | Một số lượng lớn các tài liệu điện tử của chính phủ | Không có sẵn | Tài liệu là luật và quy định | Quản lý quá trình chuyển nhà ở |
| Hình ảnh siêu dữ liệu | 5 | Hình ảnh từ một máy chủ web | Mạng với 8000 mối quan hệ và 5990 nút | Không có sẵn | Cơ sở kiến ​​thức nổi tiếnghọa sĩ | Có được hình ảnh liên quan đến các họa sĩ nổi tiếng |
| Cảm biến | 9 | Dữ liệu không gian mờ | Không có sẵn | Không có sẵn | Data from the real movement of the tropical cyclones Saomei and Bopha under the influence of subtropical high | Analyzing meteorological phenomena |
| 13 | Phương tiện vào OpenXC | 14 exabyte mỗi năm | Lên đến 5GB/giờ | Dữ liệu từ các thiết bị được cài đặt trong phương tiện | Cung cấp cái nhìn sâu sắc về mức độ rủi ro dựa trên hành vi lái xe của trình điều khiển |
| Trang web | 4 | Ứng dụng nông nghiệp dựa trên web | Không có sẵn | Không có sẵn | Dữ liệu từ ứng dụng Dịch vụ Thông tin dựa trên Nông nghiệp chính xác (PAIS) | Cung cấp dịch vụ trực tuyến cho nông dân, với 24/7 quyền truy cập vào hình ảnh của cây trồng |
| 10 | Các chuyến bay thương mại ở Mỹ | Hơn 123 triệu hồ sơ | Không có sẵn | Các chuyến bay nội địa ở Hoa Kỳ có được từ RITA-BTS | Trình bày hành vi của các công ty liên quan đến các chuyến bay hoàn thành và bị hủy bỏ |
| 18 | Xem lại trang web epinions | Không có sẵn | Không có sẵn | Dữ liệu từ người tiêu dùng trực tuyến | Xác định sở thích của người dùng |
| 30 | Microblog Sina | 1.75GB | Không có sẵn | Dữ liệu từ 1500 hồ sơ người dùng và microblog của họ | Không có sẵn |
| 32 | Slashdot | Không có sẵn | Không có sẵn | Bài viết của người dùng | Tìm thông tin hữu ích về bài đăng của người dùng |
| 33 | EasyAntiCheat | Khối lượng lớn từ dữ liệu thời gian thực của hành vi của người chơi | Không có sẵn | Khối lượng công việc một phần được trích xuất từ ​​các trò chơi nhiều người chơi | Xác định các mẫu để phát hiện gian lận |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |

**Loại dữ liệu**

Về mặt này, 18 nghiên cứu từ các bài viết có liên quan trình bày các giải pháp cho dữ liệu phi cấu trúc và 12 bài viết cho dữ liệu bán cấu trúc. Theo những dữ liệu đó, có thể xác minh rằng nghiên cứu vềMô hình hóa dữ liệu phi cấu trúc và bán cấu trúc theo xu hướng hiện tại 83,33% thời gian.Một thực tế thú vị khác là cũng có những nghiên cứu đề xuất các phương pháp mô hình hóa cho dữ liệu có cấu trúc. Điều này là do, để dữ liệu được coi là dữ liệu lớn, chúng cũng phải tuân thủ đặc điểm đa dạng. Những dữ liệu có cấu trúc này được phân tích trong tám nghiên cứu và đến từCơ sở dữ liệu quan hệ.

**3.2.2. Mô hình**

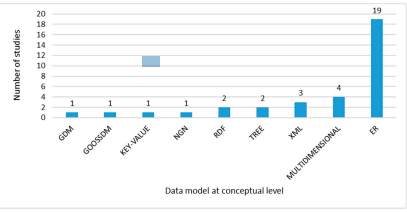
Chart, waterfall chart

Description automatically generatedTrong phần này, từ tập hợp cuối cùng được chọn, chúng tôi đã phân tích các mức độ trừu tượng dữ liệu được đề xuất, các mô hình được trình bày ở các cấp độ khái niệm, logic và vật lý, các phương pháp được đề xuất để chuyển đổi giữa các cấp độ trừu tượng, ngôn ngữ mô hình hóa, phương pháp và công cụ.Mức độ trừu tượng hóa dữ liệuĐối với khái niệm này, chúng tôi dự định xác định mức độ trừu tượng dữ liệu nào đã được đề cập bởiNghiên cứu cho các giải pháp mô hình dữ liệu. Như đã đề cập trong tiểu mục Dữ liệu lớn, có baCác mức được sử dụng cho cơ sở dữ liệu quan hệ cũng được sử dụng trong các cửa hàng NoQuery: khái niệm, logic và vật lý.Theo Hình 10, trong đó tóm tắt dữ liệu thu được trong Phụ lục A từ 36 nghiên cứu, 25Các phương pháp hiện tại để mô hình hóa dữ liệu ở ba cấp độ trừu tượng, do đó, những nghiên cứu đó có thể làđược coi là các tác phẩm hoàn chỉnh đạt được việc thực hiện vật lý các đề xuất của họ trong bộ lưu trữ NoQuery. Các nghiên cứu khác chỉ bao gồm một hoặc hai cấp độ mặc dù có thể, trong tương lai, các tác phẩm của họ sẽ chứng minh các phương pháp của họ ở cả ba cấp độ.

**Hình 10: Mức độ trừu tượng hóa dữ liệu để mô hình hóa dữ liệu lớn.**

Khái niệm tiếp theo trình bày mô hình dữ liệu được đề xuất bởi các tác giả cho từng mức trừu tượng dữ liệu.

Mô hình dữ liệu ở cấp độ khái niệm

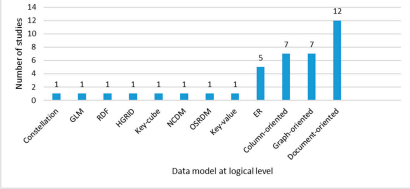
Khái niệm này bao gồm các mô hình được trình bày trong mỗi nghiên cứu từ tập đoàn cuối cùng ở cấp độ trừu tượng khái niệm. Như chúng tôi đã đề cập trước đây, cấp độ này là công nghệ không chính đáng và không có hạn chế nào liên quan đến việc sử dụng các mô hình nổi tiếng được áp dụng cho cơ sở dữ liệu quan hệ.Hình 11 trình bày 19 tác phẩm sử dụng mô hình ER ở cấp độ trừu tượng khái niệm. Trong những điều này19 Tác phẩm, người ta đề xuất việc sử dụng mối quan hệ thực thể nhị phân mở rộng (EBER) [27], một mô hình dựa trên ER áp dụng các loại thuộc tính khác nhau và vai trò thống trị. Một nghiên cứu khác từ 19 tác phẩm, đề xuất mối quan hệ thực thể được làm phong phú (EER) [37] với ký hiệu đồ họa cho việc thể hiện dữ liệu lớn

**Hình 11: Mô hình dữ liệu ở mức độ trừu tượng khái niệm.**

Hơn nữa, việc sử dụng mô hình dữ liệu đa chiều được quan sát thấy trong bốn nghiên cứu. Giả định rằng điều này có nguồn gốc từ sự quan tâm ngày càng tăng đối với các nhà dữ liệu và datamarts để xử lý phân tích trực tuyến (OLAP), trong đó việc sử dụng các truy vấn đặc biệt là phổ biến. Ngoài ra, ba bài báo đề xuất sử dụng mô hình XML, tương ứng với một biểu diễn trừu tượng của các đoạn XML. Tám công trình còn lại khác đề xuất các mô hình độc lập, chẳng hạn như mô hình dữ liệu chung (GDM),Mô hình dữ liệu bán cấu trúc theo định hướng đối tượng đồ thị (GoossDM), giá trị khóa, ký hiệu đồ họa mới (NGN), khung mô tả tài nguyên (RDF), cây và có hai tác phẩm không đề xuất bất kỳ mô hình nào.

**Mô hình dữ liệu ở cấp độ logic**

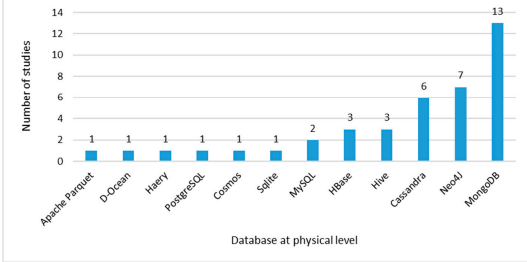
Ở mức độ trừu tượng hóa logic, theo dữ liệu thu được trong Hình 12, mô hình xu hướng được định hướng bằng tài liệu với 12 nghiên cứu, tiếp theo là định hướng biểu đồ và định hướng cột, với bảy nghiên cứu. Như được trình bày chi tiết trong tiểu mục của các khái niệm dữ liệu lớn, có bốn mô hình được sử dụng rộng rãi trong giá trị khóa NoQuery: định hướng cột, định hướng tài liệu và đồ thị; Tuy nhiên, giá trị chính đã được nghiên cứu ở mức độ trừu tượng này chỉ trong một đề xuất



**Hình 12: Mô hình dữ liệu ở mức độ trừu tượng logic.**

ER cũng đã được đề xuất như một mô hình cấp độ logic trong ba nghiên cứu và tám nghiên cứu còn lạiCác nghiên cứu đã đề xuất các giải pháp độc lập như chòm sao, mô hình logic chung (GLM),Mô hình dữ liệu bộ sưu tập RDF, HGRID, NoQuery (NCDM), Chế độ dữ liệu quan hệ có thể mở rộng mở (OSRDM)và Key-Cube. Ngoài ra, năm nghiên cứu không đề xuất bất kỳ mô hình ở cấp độ này.Dữ liệu thu được trong phần này sẽ được so sánh với dữ liệu từ dữ liệu sau,trong đó xác định các lưu trữ dữ liệu được nghiên cứu nhiều nhất

**Mô hình dữ liệu ở cấp độ vật lý**

Ở cấp độ này, việc triển khai vật lý của các mô hình trong một DBM cụ thể được xác định. Theo kết quả thu được trong Hình 13, xu hướng là hướng tới việc thực hiện trong MongoDB với 13 đề xuất, tiếp theo là NEO4J với bảy nghiên cứu và cuối cùng, Cassandra với sáu nghiên cứu. Những dữ liệu này khớp với dữ liệu được trình bày trong Hình 12, trong đó xu hướng ở cấp độ logic hướng tới các mô hình hướng đến tài liệu, hướng đến cột và định hướng đồ thị.

**Hình 13: Data Model at physical abstraction level.**

MongoDB là một NOSQL DBM định hướng tài liệu lưu trữ dữ liệu trong JSON. Mỗi tài liệu có mã định danh duy nhất, được sử dụng làm khóa chính [40]. DBM này được sử dụng bởi Foursquare, Sourceforge, CERN và Tổ chức nghiên cứu hạt nhân châu Âu, trong số các công ty khác [59].

Neo4J là một DBM NoSQL định hướng đồ thị, tổ chức dữ liệu của nó thông qua các nhãn để nhóm các nút và cạnh, còn được gọi là mối quan hệ và cả hai nút và cạnh có thể có các thuộc tính dưới dạng các cặp giá trị khóa [31]. DBM này đặc biệt được sử dụng bởi Infojobs, một công ty tư nhân cho các tìm kiếm việc làm [59].

Cassandra là một DBM NOQL định hướng cột, đại diện cho dữ liệu ở dạng bảng theo các cột và hàng [16]. Các công ty lớn, chẳng hạn như Facebook và Twitter, sử dụng DBMS này [59].

Chúng tôi nhận thấy rằng MongoDB là DBM được nghiên cứu nhiều nhất vì các công ty lớn sử dụng nó, có lẽ là do các đặc điểm hỗ trợ của nó cho tập hợp và các chỉ số thứ cấp hoạt động truy vấn và khả năng chịu đựng nhất quán và phân vùng được đề cập trong tiểu mục khái niệm dữ liệu lớn. Hơn nữa, đây là các cơ sở dữ liệu nguồn mở với hiệu suất có khả năng mở rộng, linh hoạt và tốt nhất so với cơ sở dữ liệu quan hệ. Những kết quả này cho chúng ta ý tưởng về một xu hướng trong mỗi mô hình dữ liệu đã biết, theo định hướng tài liệu của Google với MongoDB, định hướng cột với Cassandra và đồ thị với NEO4J.

Ngoài ra còn có việc triển khai NoSQL HBase và Hive DBM trên quy mô nhỏ hơn và cơ sở dữ liệu quan hệ, trong đó PostgreSQL và MySQL là một trong những cơ sở dữ liệu nổi tiếng nhất. Điều đáng nói là có những nghiên cứu đề xuất các giải pháp lai với các triển khai bao gồm các cơ sở dữ liệu khác nhau. Thông tin chi tiết về các nghiên cứu này sẽ được trình bày trong phần cơ sở dữ liệu.

Tóm lại, hầu hết các nghiên cứu đề xuất việc sử dụng mô hình ER ở cấp độ khái niệm, một mô hình định hướng tài liệu ở cấp độ logic và việc thực hiện MongoDB ở cấp độ vật lý.

**Chuyển đổi giữa các mức độ trừu tượng**

Theo Hình 10, có 25 nghiên cứu được lựa chọn trình bày phương pháp tiếp cận của họ ở ba cấp độ trừu tượng và tám nghiên cứu ở hai cấp độ khác biệt với logic đến vật lý, khái niệm đến logic và khái niệm đến vật lý. Theo khái niệm này, các phương pháp được đề xuất cho sự chuyển đổi giữa các mức độ trừu tượng hóa dữ liệu được trình bày.

Như được trình bày trong Hình 14, có 19 nghiên cứu trong đó các tác giả đề xuất các quy tắc lập bản đồ mới lạ của riêng họ, cho thấy nghiên cứu riêng biệt tồn tại về chủ đề này. Vì vậy, rất khó để quyết định cái nào là phù hợp nhất khi chọn bất kỳ trong số chúng. Một khía cạnh thú vị khác là 12 nghiên cứu không xác định các quy tắc biến đổi và có sáu nghiên cứu đề xuất các biến đổi dựa trên các kỹ thuật khác, chẳng hạn như thuật toán tuyến tính hóa (LA), ngôn ngữ chuyển đổi ATL (ATL), Hoberman Heuristic (HH), thuật toán Cardinality ( AC), Lý thuyết danh mục (CT) và ràng buộc không gian khối lượng công việc (WSC).

Chart

Description automatically generated

**Hình 14: Chuyển đổi giữa các mức độ trừu tượng dữ liệu**

Nói chung, các tác giả đề xuất thuật toán dưới đây lấy một mô hình làm đầu vào, áp dụng các quy tắc chuyển đổi của riêng họ và tạo ra một mô hình khác làm đầu ra:

Input1: Cấp độ khái niệm: ***C*n =} {*Ci*},** trong đó ***Ci*** thuộc mỗi yếu tố ***i*** từ mô hình khái niệm

Quy tắc chuyển đổi : *R***= *rj RJ****,* trong đó ***RJ***  thuộc về từng *J* quy tắc hoặc ràng buộc từ các quy tắc ánh xạ

được xác định bởi các tác giả.

OUTPUT1: Mức logic: ***L* = *C* ∪ *R* = {*lk*}** trong đó ***lk***  thuộc mỗi phần tử ***k*** từ mô hình logic. Input2 = input 1

Quy tắc chuyển đổi’**,: *R*j = *r*j *j*,** trong đó ***r*j** *j***,** thuộc về từng quy tắc***j*,** hoặc ràng buộc từ các quy tắc ánhxạ được xác định bởi các tác giả.

OUTPUT2: Cấp độ vật lý: ***P* = *L* ∪ *R*j = *pk***,, trong đó ***pk*** thuộc mỗi phần tử ***k*** từ mô hình vật lý.

**Mô hình ngôn ngữ**

Chart

Description automatically generatedTrong khái niệm này, điều quan trọng là phải làm rõ rằng một mô hình dữ liệu mô tả các đặc điểm của dữ liệu, cấu trúc, hoạt động của nó và các ràng buộc của nó; Trong khi đó, mô hình dữ liệu là quá trình tạo ra các mô hình này. Mục đích của mô hình dữ liệu là cho các mô hình được sử dụng và hiểu bởi tất cả các nhà lập mô hình, nhà phát triển và những người khác làm việc trong lĩnh vực phát triển/kỹ thuật phần mềm theo cách tiêu chuẩn hóa. Do đó, Hình 15 trình bày các kết quả thu được từ ánh xạ được thực hiện trong Phụ lục A từ 36 bài báo có liên quan đã chọn.

**Hình 15: Ngôn ngữ mô hình dữ liệu.**

Theo Hình 15, không có xu hướng của ngôn ngữ mô hình hóa dữ liệu. Có 27 nghiên cứu không xác định một ngôn ngữ được tiêu chuẩn hóa được sử dụng để trực quan hóa các mô hình. Chỉ có sáu nghiên cứu được điều chỉnh theo một tiêu chuẩn như ngôn ngữ mô hình hóa thống nhất (UML) và bốn nghiên cứu khác đề xuất việc sử dụng ngôn ngữ mô hình hóa của riêng họ, như sơ đồ Chebokto, mô hình dữ liệu bán cấu trúc đối tượng đồ thị (GOOSDM), mở rộng metamodel nhẹ và phần mở rộng metamodel và XML. Trong số sáu nghiên cứu trình bày các mô hình của họ với việc sử dụng UML, hai trong số chúng sử dụng nó trong mô hình cấp độ khái niệm [35,45], người ta sử dụng nó trong các mô hình khái niệm và logic [14] và ba người sử dụng nó trong tất cả , mức độ logic và vật lý [32,34,42].

Theo một số tác giả [16,41] và một số kinh nghiệm thực hiện, một sự khác biệt quan trọng giữa cơ sở dữ liệu quan hệ và cơ sở dữ liệu NoQuery là cái sau không yêu cầu chuẩn hóa; Đó là, họ hỗ trợ dữ liệu trùng lặp. Trong tình huống này, mô hình hóa dữ liệu trong cơ sở dữ liệu NoQuery thường bắt đầu với việc xây dựng các câu hỏi về cách tham khảo dữ liệu cơ sở dữ liệu. Những câu hỏi này sẽ xác định các thực thể và mối quan hệ giữa các thực thể đó. Mô hình mới này chuyển từ quy trình mô hình hóa dựa trên dữ liệu sang quy trình mô hình hóa dựa trên truy vấn. Do đó, trong khái niệm sau đây, các phương pháp mô hình hóa dữ liệu được đề xuất trong tập hợp cuối cùng của các bài viết được chọn sẽ được phân tích.

**Phương pháp mô hình hóa**

Như đã đề cập trong khái niệm trước, phần này nhằm mục đích tiết lộ phương pháp mô hình hóa dữ liệu được đề xuất bởi các nghiên cứu. Theo kết quả đạt được, xu hướng của các đề xuất là sử dụng phương pháp dựa trên dữ liệu được trình bày trong 33 nghiên cứu. Mô hình hóa dựa trên dữ liệu là một kỹ thuật, dựa trên cách tổ chức dữ liệu trong bộ dữ liệu và cách chúng được lấy từ các hệ thống bên ngoài, tạo ra tất cả các thành phần để biểu diễn mô hình [60]. Chỉ có năm nghiên cứu đề xuất mô hình điều khiển truy vấn; Cần phải đề cập rằng các nghiên cứu đề xuất mô hình hướng khối lượng công việc, cũng dựa trên điều khiển truy vấn, đã được xem xét trong năm nghiên cứu này.

**Công cụ mô hình**

A picture containing application

Description automatically generatedChúng tôi xem xét điều quan trọng là phải biết liệu các nghiên cứu được lựa chọn cũng đề xuất một công cụ máy tính để hỗ trợ xây dựng mô hình từ đầu, xác nhận các mô hình được xây dựng và hỗ trợ chuyển đổi tự động giữa các mức độ trừu tượng. Trong số các kết quả thu được trong ma trận khái niệm, như trong Hình 16, 29 nghiên cứu không đề xuất bất kỳ công cụ nào. Chỉ có hai nghiên cứu đề xuất việc sử dụng khung mô hình Eclipse (EMF). Tương tự, có năm nghiên cứu đề xuất các công cụ riêng biệt như Kashlev Data Modeler (KDM) [16], tập lệnh trong Haskell [26], Mortadelo [32], Neoclipse [14], Trình đánh giá lược đồ NoQSQL (NoSE) [54]. Điều đáng nói là tất cả các công cụ này cho phép mô hình hóa ở ba cấp độ trừu tượng hóa dữ liệu

**Hình 16: Công cụ mô hình hóa dữ liệu (Khung mô hình hóa Eclipse (EMF), Trình mô hình dữ liệu Kashlev (KDM), Trình đánh giá lược đồ NoSQL (NoSE)).**

**3.2.2 Cơ sở dữ liệu**

Trong phần này, chúng tôi đã xác định các loại cơ sở dữ liệu được đề xuất và so sánh đánh giá và hiệu suất được thực hiện bởi các nghiên cứu.

**Loại cơ sở dữ liệu**

Ở khía cạnh này, chúng tôi trình bày các loại cơ sở dữ liệu mà các nghiên cứu có liên quan được lựa chọn đã đề xuất. Chúng đã được phân loại thành hai nhóm chính, đồng nhất và giống lai. Bằng cách đồng nhất, chúng tôi có nghĩa là những cơ sở dữ liệu nơi dữ liệu được triển khai trong một cơ sở dữ liệu. Bằng các giống lai, chúng tôi có nghĩa là các hệ thống nơi có một số cơ sở dữ liệu được triển khai có thể là quan hệ và/hoặc NoQuery. Theo một số nghiên cứu [14,61,62], do các đặc điểm đa dạng của dữ liệu lớn, việc thiết kế và quản lý cơ sở dữ liệu đã trở nên phức tạp, do đó, các hệ thống được định hướng theo hệ thống tồn tại polyglot (PPS). Điều này có nghĩa là, khi dữ liệu lớn được lưu trữ, tốt hơn là sử dụng các công nghệ lưu trữ khác nhau, đó là cơ sở dữ liệu lai, để các ứng dụng có thể chọn một công nghệ phù hợp nhất tùy thuộc vào dữ liệu họ cần [61]. Sự tồn tại của polyglot là thuật ngữ được sử dụng khi một ứng dụng có thể truy vấn dữ liệu từ các cơ sở dữ liệu NoQuery khác nhau [14].

Theo kết quả của DSLR của chúng tôi, 20 nghiên cứu đề xuất các giải pháp đồng nhất, đó là, chúng tập trung vào một loại cơ sở dữ liệu duy nhất và chỉ có tám giải pháp lai. Điều đáng nói là, trong tám nghiên cứu này, không có giải pháp nào thực hiện các mô hình dữ liệu sau: E-R, định hướng tài liệu, định hướng cột và đồ thị. Tương tự như vậy, tám nghiên cứu không xác định bất kỳ loại thực hiện vật lý nào.

Trong số các nghiên cứu trình bày các giải pháp cho ít nhất ba loại DBM khác nhau là một loại đề xuất triển khai trong SQLite, MongoDB, MySQL và Neo4J [26]; Một cái khác đề xuất các triển khai trong MySQL, MongoDB và Cosmos [28]; và một người khác đề xuất các triển khai ở Cassandra, MongoDB và Neo4j [35].

**Đánh giá và so sánh hiệu suất**

Tương tự, trong tiểu mục mô hình hóa, 27 nghiên cứu không trình bày mô hình tiêu chuẩn và, theo tiểu mục công cụ mô hình, 29 nghiên cứu không xác định một công cụ mô hình hóa tự động. Chỉ dựa trên thông tin này, rất khó để chọn bất kỳ đề xuất nào. Vì lý do này, một số tác phẩm đã thực hiện đánh giá các phương pháp được đề xuất của họ, dựa trên thời gian tải dữ liệu và thời gian thực hiện truy vấn. Từ các kết quả được yêu cầu trong Phụ lục A, tám nghiên cứu đã gửi đánh giá các đề xuất của họ về thời gian thực hiện truy vấn [25,26,40,48,51,52,54,55], một nghiên cứu đánh giá thời gian chuyển đổi mô hình [47] và một nghiên cứu khác So sánh thời gian tải dữ liệu [43]. Cuối cùng, một số bài viết đã đề cập đến tính hữu ích của kỹ thuật đảo ngược để xác minh tính hợp lệ của mô hình đề xuất [63]. Chỉ có một nghiên cứu tập trung vào khía cạnh này nhưng nó không kiểm tra nó [31].

Theo dữ liệu của các khái niệm chính được phân tích trước đây, các xu hướng và khoảng trống được tìm thấy trong nghiên cứu DSLR của chúng tôi được trình bày trong cuộc thảo luận.

**3.3 Thảo luận**

Trong phần này, chúng tôi trả lời RQ3. Đầu tiên, chúng tôi đề cập đến các thông tin liên quan đạt được từ phân tích thư mục. Từ phân tích này, có thể xác minh sự phát triển của chủ đề nghiên cứu này. Kể từ năm 2015, số lượng nghiên cứu đã tăng khoảng 100 lần và năm 2018 là năm với nhiều ấn phẩm nhất. Scopus là nguồn giữ số lượng nghiên cứu liên quan cao nhất. Các quốc gia có đóng góp lớn nhất trong chủ đề mô hình dữ liệu lớn là Hoa Kỳ và Trung Quốc và hơn 50% các nghiên cứu được tài trợ.

Những dữ liệu cho thấy sự quan tâm tồn tại trong chủ đề trong cộng đồng nghiên cứu. Mặc dù nhiều nghiên cứu đề cập đến phía phân tích dữ liệu lớn, kỹ thuật chăm sóc các nền tảng, kỹ thuật trích xuất và tải và chuyển đổi dữ liệu cho đến khi đạt được lưu trữ cần thiết, không có tầm quan trọng hơn, vì chúng bao gồm điểm bắt đầu theo yêu cầu của các quy trình phân tích trong đặt hàng để tạo ra giá trị từ dữ liệu.

Tiếp theo, chúng tôi sẽ lấy thông tin đạt được từ SLR làm tài liệu tham khảo để khám phá các xu hướng và khoảng trống trong chủ đề quan tâm.

**3.3.1 Xu hướng**

Để tóm tắt các xu hướng được quan sát, Hình 17 giúp làm tài liệu tham khảo để trình bày ba khái niệm chính mà phân tích này tập trung vào: Nguồn, mô hình và cơ sở dữ liệu.

**Nguồn**

Đối với khía cạnh nguồn, các nguồn dữ liệu được nghiên cứu nhiều nhất là dữ liệu phi cấu trúc, cụ thể là thông tin trang web.

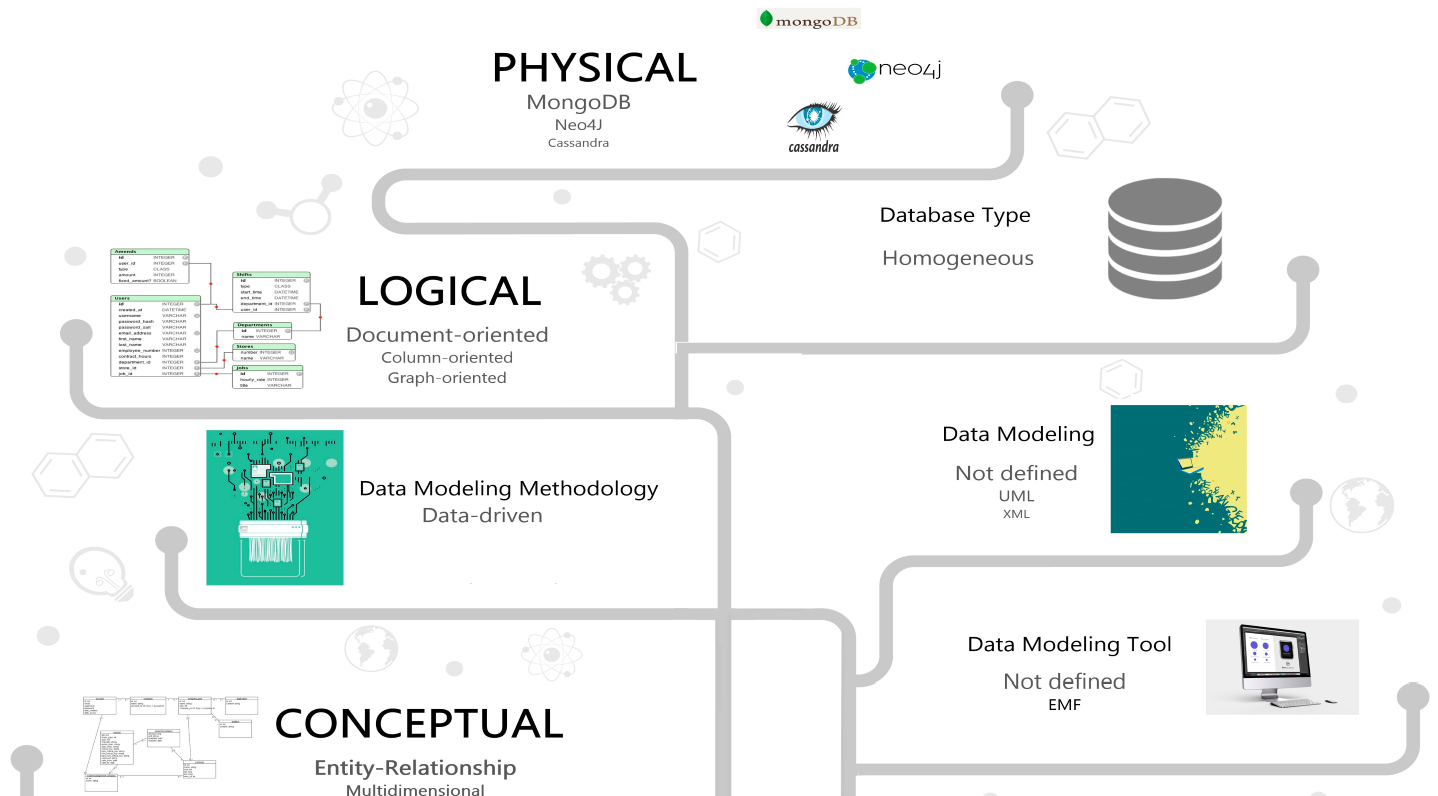
**Mô hình**

Đối với khía cạnh mô hình hóa, bảy xu hướng chính đã được xác định trong kho văn bản được phân tích:

1. Hầu hết các nghiên cứu trình bày các đề xuất của họ ở mức độ trừu tượng khái niệm, logic và vật lý;
2. Mô hình ER được sử dụng nhiều nhất trong các phương pháp ở cấp độ trừu tượng khái niệm, tiếp theo là mô hình đa chiều và, thứ ba, XML;
3. Ở cấp độ trừu tượng hóa logic, mô hình được nghiên cứu nhiều nhất được định hướng bằng tài liệu, tiếp theo là định hướng cột và đồ thị;
4. Ở cấp độ trừu tượng vật lý, việc triển khai tập trung nhiều hơn vào DBM MongoDB, tiếp theo là Neo4j và Cassandra. Do đó, các tiêu chuẩn thực tế sau đây đã xuất hiện, MongoDB cho mô hình dữ liệu định hướng tài liệu, Cassandra cho định hướng cột và NEO4J cho mô hình dữ liệu đồ thị. Những dữ liệu này được hỗ trợ bởi thông tin thống kê từ xếp hạng DB -Engines - mức độ phổ biến xu hướng của Công ty CNTT rắn kể từ tháng 12 năm 2019 [64];
5. Phương pháp mô hình được đề xuất nhiều nhất là dựa trên dữ liệu;
6. Không có xu hướng rõ ràng đối với phương pháp mô hình hóa dữ liệu nhưng có các đề xuất với UML và XML;
7. Không có công cụ mô hình dữ liệu nào được định nghĩa là một xu hướng nhưng một số nghiên cứu đã sử dụng EMF;
8. Về các lĩnh vực ứng dụng khác nhau, ER thường được sử dụng ở cấp độ trừu tượng khái niệm trong các nghiên cứu trường hợp khác nhau. Ở cấp độ logic, các phương pháp di chuyển từ các mô hình di chuyển từ quan hệ sang định hướng tài liệu, đồ thị và hướng cột được đề xuất. Cụ thể, đối với dữ liệu không gian và dữ liệu truyền tải, các giải pháp cho mô hình đồ thị được đề xuất. Và, đối với các định dạng XML và JSON từ các máy chủ web, các giải pháp cho tất cả các mô hình dữ liệu NoQuery được đề xuất

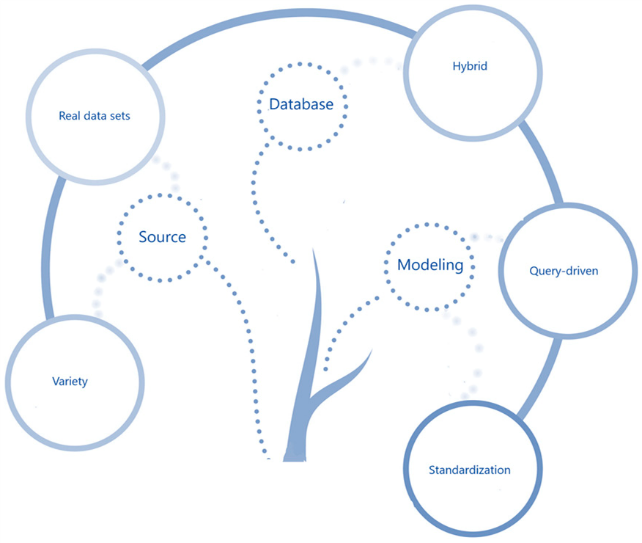
**Cơ sở dữ liệu**

Cho khía cạnh cơ sở dữ liệu, 55,55% các nghiên cứu tập trung vào các loại cơ sở dữ liệu đồng nhất, vì vậy xu hướng đối với một hệ thống polyglot liên tục không phải là người quan sát



Hình 17. Xu hướng mô hình và quản lý dữ liệu lớn.

**3.3.2. Khoảng cách**

Hình 18 cho phép chúng tôi xác định cũng có ba khái niệm chính để phân tích các khoảng trống: Nguồn, Mô hình hóa và Cơ sở dữ liệu

Hình 18. Khoảng cách cho mô hình và quản lý dữ liệu lớn

Nguồn

cho khía cạnh nguồn, hai khoảng trống chính đã được xác định trong tập đoàn được phân tích:

1. Một trong những mục tiêu chính của dữ liệu lớn là sử dụng dữ liệu để tạo giá trị; Giá trị này sẽ phụ thuộc vào nhu cầu của doanh nghiệp. Vì lý do này, điều rất quan trọng là các nghiên cứu được xác nhận với các trường hợp sử dụng thực tế, vì chỉ với việc sử dụng các bộ dữ liệu thực mới có thể được tạo ra giá trị có được tạo ra hay không. Kết quả của DSLR của chúng tôi, chỉ có mười nghiên cứu, tương ứng với 27,78%, sử dụng bộ dữ liệu thực; 16,67% trong số các tác phẩm này trình bày các nghiên cứu trường hợp của họ với dữ liệu từ các trang web, 5,56% sử dụng dữ liệu từ các cảm biến và 2,78% liên quan đến tài liệu điện tử Dữ liệu và hình ảnh siêu dữ liệu;
2. Ngoài ra, các tính năng chính dữ liệu lớn khác phải được đảm bảo, chẳng hạn như khối lượng, vận tốc và tính xác thực. Theo kết quả của Bảng 6, không phải tất cả các tính năng này đều được chứng minh trong các phương pháp;
3. Để tuân thủ sự đa dạng, các nghiên cứu nên xem xét rằng dữ liệu có thể có ở bất kỳ định dạng nào: có cấu trúc, bán cấu trúc hoặc không cấu trúc. Do đó, các phương pháp đề xuất nên giải quyết bất kỳ loại nào trong số này. Chỉ 2,78% đã đề xuất một giải pháp cho cả ba loại dữ liệu. 47,22% các nghiên cứu còn lại chỉ đưa ra các phương pháp tiếp cận cho dữ liệu phi cấu trúc, chỉ 30,56% đối với dữ liệu có cấu trúc, 13,89% đối với dữ liệu bán cấu trúc, 2,78% kết hợp với dữ liệu có cấu trúc và bán cấu trúc và 2,78%

**Mô hình hóa**

cho khía cạnh mô hình hóa, năm khoảng trống chính được tìm thấy trong nghiên cứu của chúng tôi tương ứng với phương pháp mô hình hóa và tiêu chuẩn hóa. Các tiêu chuẩn được coi là quan trọng, vì chúng đảm bảo cách giải thích phổ biến, thống nhất, khả năng đọc của mô hình dữ liệu, tính di động giữa các công cụ, nền tảng và ứng dụng cơ sở dữ liệu, trong số các cơ sở khác cho các nhà quản lý dự án, nhà phân tích, nhà thiết kế, nhà phát triển và người dùng cuối của cơ sở dữ liệu. Tóm lại, kết quả như sau:

1. Các mô hình có nguy cơ không thực tế vì tính chất đa dạng của các luồng dữ liệu [65]. 6. Một nghiên cứu [66] chứng minh rằng việc xử lý dữ liệu, về mặt thời gian, với việc sử dụng điều khiển truy vấn đòi hỏi thời gian xử lý ít hơn so với dựa trên dữ liệu. Điều này, đến lượt nó, dẫn đến một lượng nhỏ ENE tiêu thụ
2. Chưa có hệ thống NoQuery nào nổi lên như một tiêu chuẩn hoặc là một tiêu chuẩn thực tế;
3. Không có việc sử dụng rõ ràng một số ngôn ngữ hoặc phương pháp tiêu chuẩn hóa để mô hình hóa dữ liệu ở cấp độ logic và vật lý. Theo kết quả của DSLR của chúng tôi, chỉ ở cấp độ khái niệm mới có thể được coi là một xu hướng, có thể bởi vì mức độ khái niệm là công nghệ-bất khả tri;
4. Như đã đề cập trong tiểu mục ngôn ngữ mô hình hóa, cho cơ sở dữ liệu NoQuery, mô hình mới cho quy trình mô hình hóa được điều khiển theo truy vấn. Tuy nhiên, chỉ có năm nghiên cứu tập trung vào phương pháp này
5. Để phân tích dữ liệu lớn, việc thực hiện các hệ thống hiệu quả là bắt buộc. Trong phương pháp dựa trên dữ liệu, các mô hình được thiết kế trước khi đánh giá những truy vấn nào sẽ được thực hiện. Hạn chế của giải pháp này là tất cả dữ liệu sẽ được lưu trữ, khi chỉ cần một phần giới hạn để trả lời các truy vấn cần thiết. Ngược lại, trong các phương pháp điều khiển truy vấn, một tập hợp các truy vấn phải được thể hiện, đánh giá và tích hợp trước khi mô hình hóa, để lập kế hoạch có thể tập trung vào các câu trả lời cho các truy vấn cần thiết. Đối với các giải pháp có dữ liệu thời gian thực được sử dụng, các mô hình dựa trên dữ liệu có nguy cơ không thực tế vì tính chất đa dạng của các luồng dữ liệu [65].
6. Một nghiên cứu [66] chứng minh rằng việc xử lý dữ liệu, về mặt thời gian, với việc sử dụng điều khiển truy vấn đòi hỏi thời gian xử lý ít hơn so với dựa trên dữ liệu. Điều này, đến lượt nó, dẫn đến lượng năng lượng tiêu thụ nhỏ hơn và do đó, tuổi thọ dài hơn của thiết bị
7. Theo một nghiên cứu khác [67], việc sử dụng phương pháp điều khiển truy vấn cho phép người dùng thực hiện các truy vấn bằng ngôn ngữ người dùng tự nhiên tập trung vào các khu vực quan tâm có liên quan;

Các kỹ thuật như kỹ thuật đảo ngược không được tính đến; Vì vậy, chi tiết này có thể bổ sung cho các nghiên cứu. Phương pháp này được sử dụng rộng rãi trong cơ sở dữ liệu quan hệ trên các dự án hoạt động với cơ sở dữ liệu tồn tại và khi không có mô hình tài liệu nào tồn tại (hoặc có nguy cơ các mô hình này đã lỗi thời).

**Cơ sở dữ liệu**

Cho khía cạnh cơ sở dữ liệu, khoảng cách chính được tìm thấy có liên quan đến các đặc điểm đa dạng của dữ liệu lớn. Để quản lý tính năng này một cách chính xác, cơ sở dữ liệu NoQuery phải được nhắm mục tiêu vào PPS. Một PPS có thể được lấy thông qua lưu trữ trong cơ sở dữ liệu lai; Đó là, nó hỗ trợ các công nghệ lưu trữ khác nhau. Chỉ có tám nghiên cứu đã đề xuất các giải pháp được định hướng cho cơ sở dữ liệu lai và, trong số này, chỉ có một nghiên cứu đã chứng minh điều này với một trường hợp thực tế và quản lý cả dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc và lưu trữ chúng trong các cơ sở dữ liệu khác nhau [28].

1. **Kết luận**

Là một hạn chế, chúng tôi biết rằng chúng tôi không thể thu thập tất cả các nghiên cứu tồn tại về chủ đề này, do số lượng lớn các từ đồng nghĩa và ràng buộc trong các thuật ngữ tìm kiếm chính. Tuy nhiên, số lượng lớn các nghiên cứu chính, một tập hợp 1376 bài báo đã được lấy, cung cấp một bộ dữ liệu hoàn chỉnh và đáng tin cậy.

Nghiên cứu SLR về mô hình và quản lý dữ liệu lớn được thực hiện trong bài viết này đã xác định 36 nghiên cứu có liên quan. Những công trình này đã được lựa chọn dựa trên quá trình bao gồm và lựa chọn các tài liệu nghiên cứu từ các thư viện khoa học kỹ thuật số cho đến tháng 8 năm 2019; Không có bài báo được xuất bản trước năm 2010.

Chúng tôi đã đặt ra ba câu hỏi nghiên cứu đã được trả lời thông qua phân tích thư mục, máy ảnh DSLR và một cuộc thảo luận. Kết quả của nghiên cứu thư mục cung cấp rất nhiều thông tin liên quan. Ví dụ, từ năm 2015 trở đi, số lượng nghiên cứu đã tăng đáng kể vì đây là năm mà nhiều quốc gia bắt đầu đóng góp cho chủ đề quan tâm. Các quốc gia hàng đầu trong chủ đề này là Hoa Kỳ và Trung Quốc. Hoa Kỳ có một trong những nghiên cứu có ảnh hưởng nhất.

Các kết quả từ máy ảnh DSLR và từ cuộc thảo luận cũng tiết lộ một số sự thật rất thú vị. Ví dụ, hơn 50% các nghiên cứu không xác minh các đề xuất của họ với các bộ dữ liệu trong thế giới thực. Đặc tính vận tốc dữ liệu lớn không được chứng minh bởi 90% các nghiên cứu. ER là mô hình được sử dụng nhiều nhất ở cấp độ trừu tượng khái niệm, theo định hướng tài liệu là mô hình được nghiên cứu nhiều nhất ở cấp độ trừu tượng logic và DBM MongoDB là triển khai thường xuyên nhất ở cấp độ vật lý. Hơn nữa, là những khoảng trống chính, chúng tôi đã xác định việc thiếu các đánh giá đề xuất và một vài nghiên cứu tập trung vào phương pháp điều khiển truy vấn và các giải pháp cơ sở dữ liệu lai

Sự đóng góp của nghiên cứu DSLR này, bằng cách làm rõ kiến ​​thức về chủ đề cụ thể của mô hình dữ liệu lớn, có thể hỗ trợ các nhà nghiên cứu và học viên trong việc cải thiện các dự án dữ liệu lớn của họ. Các tác phẩm có liên quan được thu thập có thể là điểm khởi đầu hữu ích cho các nghiên cứu mới về mô hình dữ liệu lớn.

Cuối cùng, chúng tôi biết rằng, do nhu cầu phân tích khối lượng dữ liệu lớn với nhiều cấu trúc khác nhau, đến tần số cao, nghiên cứu cơ sở dữ liệu trở nên tập trung hơn vào NOCL. Tuy nhiên, NoQuery DBMS không thể có được một số điểm mạnh đã có trong cơ sở dữ liệu quan hệ, chẳng hạn như khả năng hỗ trợ tính nhất quán và tính khả dụng cùng một lúc; Vì lý do này, NewsQL đã xuất hiện và có thể đặt ra một giải pháp cho các vấn đề mà cả hai DBMS phải đối mặt. Là một công việc trong tương lai, chúng tôi cũng hy vọng sẽ nghiên cứu hệ thống cơ sở dữ liệu mới này.

Chúng tôi hy vọng sẽ giữ cho máy ảnh DSLR cập nhật và trình bày kết quả cho các khái niệm khác. Hơn nữa, dựa trên các lỗ hổng, chúng tôi sẽ thúc đẩy nghiên cứu của chúng tôi về các khía cạnh này.

**Đóng góp của tác giả:** D.M.-M. đóng góp vào việc thụ thai và thiết kế công việc, thu thập dữ liệu, phân tích và giải thích dữ liệu; S.L.-M. và R.N., Đánh giá quan trọng về tính chính xác của nghiên cứu, nội dung trí tuệ và sự chấp thuận cuối cùng của phiên bản sẽ được gửi. Tất cả các tác giả đã đọc và đồng ý với phiên bản được xuất bản của bản thảo.

**Tài trợ:** Nghiên cứu này không nhận được tài trợ bên ngoài.

**Lời cảm ơn:** Chúng tôi cảm ơn sự đóng góp của ngành để biết nhu cầu trong các dự án dữ liệu lớn của họ ở phía cửa hàng

**Xung đột lợi ích:** Các tác giả tuyên bố không có xung đột lợi ích.

**Phụ lục A. Ma trận khái niệm**

**Bộ dữ liệu**

**Mẫu**

**Mô hình**

**Đánh giá**

**Chuyển đổi và**

**ID Tác giả Tiêu đề Loại cơ sở dữ liệu**

**giữa Mô hình Công cụ Phương Pháp so sánh**

**Nguồn Cấu trúc Bán cấu trúc Phi cấu trúc Khái niệm Hợp lý Vật lý**

**các mức độ ngôn ngữ mô hình Mô hình hóa hiệu suất**

**trừu tượng**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Jie SONG,  Hongyan HE,  Richard 1  THOMAS,  Yubin BAO, Ge  YU | Haery: a Hadoop based Query System on  Accumulative and  High-dimensional  Data Model for Big  Data | Dữ liệu mẫu | - | - | Tệp văn bản | giá trị cốt lõi | Giá trị đa chiều | Haery | Thuật toán hóa tuyến tính hóa thuật toán hóa | Không xác định | NA | Điều khiển theo mô hình  điều khiển truy vấn | Đồng nhất | truy vấn |
| Laurent Thiry,  Heng Zhao  2 and Michel  Hassenforder | Categories for  (Big) Data models and optimization | Dữ liệu mẫu | - | csv | - | Thực thể - mối quan hệ (ER) | ER đồ thị tài liệu | Sqlite  Mongo  MySql  Neo4J | Lý thuyết thể loại (CT) | Not defined | Kịch bản Haskell | mô hình điều khiển | Hỗn hợp | truy vấn |
| Victor Martins de Sousa, Luis  3  Mariano del  Val Cura | Logical Design of  Graph Databases from an Entity Relationship  Conceptual Model | Dữ liệu mẫu | - | - | trang mạng | Nhị phân mở rộng ER (EB-ER) | đồ thị | Neo4j | Thuật toán ánh xạ các ràng buộc về thuật toán các ràng buộc thuật toán thuật toán thuật toán | Not defined | NA | mô hình điều khiển | Đồng nhất | NA |
| Igor Zecˇevic´,  Petar Bjeljac,  Branko Perišic´,  Stevan 4  Stankovski,  Danijel Venus & Gordana  Ostojic´ | Model-driven development of  hybrid databases  using lightweight metamodel extensions | Dữ liệu thực | dữ liệu quản lý | Tệp JSON | trang mạng | ER | ER đồ thị tài liệu giá trị cốt lõi | MySQL  MongoDB  Cosmos | Lập bản đồ quy tắc | lightweight metamodel extensions | NA | mô hình điều khiển | Hỗn hợp | NA |
| Antonio M.  Rinaldi, 5  Cristiano  Russo | A Semantic-based  Model to represent  Multimedia Big  Data | Dữ liệu thực | - | - | Siêu dữ liệu từ hình ảnh | - | đồ thị | Neo4j | NA | Not defined | NA | mô hình điều khiển  mô hình truy vấn | Đồng nhất | NA |
| Shady  Hamoud, 6  Zurinahni  Zainol | Document-Oriented  Data Schema for  Relational  Database  Migration to  NoSQL | Dữ liệu mẫu | Dữ liệu cơ sở dữ liệu quan hệ | - | - | ER | tài liệu | MongoDB | Quy tắc lập bản  đồ quy trình  chuẩn hóa &  phân ly quá trình | Not defined | NA | mô hình điều khiển | Đồng nhất | NA |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dippy  7 Aggarwal,  Karen C. Davis | Employing Graph  Databasesas a  Standardization  Model for  Addressing  Heterogeneity and  Integration | Dữ liệu mẫu | Dữ liệu cơ sở dữ liệu quan hệ RDF | csv | - | ER Khung mô tả tài nguyên (RDF) | đồ thị | Neo4j | Lập bản đồ quy tắc | Not defined | NA | mô hình điều khiển | Đồng nhất | NA |
| Alfonso de la  Vega, Diego  García-Saiz,  8 Carlos Blanco,  Marta Zorrilla, and Pablo  Sánchez | Mortadelo: A  Model-Driven  Framework for  NoSQL Database  Design | Dữ liệu mẫu | Dữ liệu cơ sở dữ liệu quan hệ | - | - | Mô hình dữ liệu chung (GDM) | Tài liệu cột | Cassandra  MongoDB | Lập bản đồ quy tắc | Unified  Modeling  Language  (UML) | Mortadelo | mô hình điều khiển | Hỗn hợp | NA |
| Xu Chena, Li  9 Yanb, Weijun  Lia, Fu Zhangc | Fuzzy  Spatio-temporal  Data Modeling  Based on XML  Schema | Dữ liệu thực | - | - | cảm biến | XML | - | - | NA | Tree | NA | mô hình điều khiển | NA | NA |
| Maribel  Yasmina  10 Santos, Bruno  Martinho &  Carlos Costa | Modeling and implementing big  data warehouses for decision  support | Dữ liệu thực | đa chiều | - | trang mạng | ER | ER | Hive | Lập bản đồ quy tắc | Not defined | NA | mô hình điều khiển | Đồng nhất | NA |
| Kwangchul  Shin,  Chulhyun 11  Hwang,  Hoekyung  Jung | NoSQL Database Design Using  UML Conceptual  Data Model Based on Peter Chen’s  Framework | Dữ liệu mẫu | - | csv | - | ER | ER | - | Lập bản đồ quy tắc | UML | NA | mô hình điều khiển | NA | NA |
| Fatma  Abdelhedi,  Amal Ait 12  Brahim, Faten  Atigui, Gilles  Zurfluh | Logical Unified  Modeling For  NoSQL DataBases | NA | sơ đồ lớp từ cơ sở dữ liệu quan hệ | - | - | ER | GLM | Cassandra  MongoDB  Neo4J | Lập bản đồ quy tắc | UML at  conceptual | EMF | mô hình điều khiển | Hỗn hợp | NA |
| Victor Martins de Sousa, Luis  13  Mariano del  Val Cura | A NoSQL Data  Model For Scalable  Big Data Workflow  Execution | Dữ liệu thực | - | - | cảm biến | NA | NCDM | NA | NA | NA | NA | NA | NA | NA |
| Massimo  Villari, Antonio  Celesti, 14  Maurizio  Giacobbe and  Maria Fazio | Enriched E-R  Model to Design  Hybrid Database for Big Data  Solutions | NA | - | - | Hồ sơ y tế sức khỏe điện tử | EE-R | - | - | NA | NA | NA | mô hình điều khiển | NA | NA |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Maribel  Yasmina 15  Santos, Carlos  Costa | Data Warehousing in Big Data From  Multidimensional to Tabular Data  Models | Dữ liệu mẫu | đa chiều - | - | đa chiều | chòm sao | Hive | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển | Đồng nhất | NA |
| Maribel  Yasmina 16  Santos, Carlos  Costa | Data Models in  NoSQL Databases for Big Data  Contexts | Dữ liệu mẫu | Dữ liệu cơ sở dữ  liệu quan hệ | - | ER | Cột ER | Hbase Hive | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển | Hỗn hợp | NA |
| Ganesh B.  17 Solanke, K.  Rajeswari | Migration of  Relational  Database to  MongoDB and  Data Analytics using Naïve Bayes  Classifier based on  Mapreduce  Approach | Dữ liệu mẫu | Dữ liệu cơ sở dữ  liệu quan hệ | - | ER | tài liệu | MongoDB | Lập bản đồ quy tắc | NA | Các kịch bản SQL | mô hình điều khiển | Đồng nhất | truy vấn |
| Vincent  Reniers,  Dimitri Van 18  Landuyt,  Ansar Rafique, Wouter Joosen | Schema Design Support for  Semi-Structured  Data: Finding the  Sweet Spot between NF and  De-NF | Dữ liệu thực | - - | trang mạng | ER | tài liệu | - | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển  mô hình công việc | Đồng nhất | NA |
| Fatma  Abdelhedi,  Amal Ait 19  Brahim, Faten Atigui and  Gilles Zurfluh | Big Data and Knowledge Management:  How to Implement Conceptual  Models in NoSQL Systems? | Dữ liệu mẫu | - - | cảm biến | ER | Cột | Hbase Cassandra | Lập bản đồ quy tắc | UML XML | EMF | mô hình điều khiển | Hỗn hợp | NA |
| Max Chevalier,  Mohammed El  Malki, Arlind  20 Kopliku,  Olivier Teste and Ronan  Tournier | Document-oriented  Models for Data  Warehouses | Dữ liệu mẫu | - Tệp JSON | - | đa chiều | tài liệu | MongoDB | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển | Đồng nhất | tải dữ liệu |
| Shreya  Banerjee,  Renuka Shaw,  21  Anirban Sarkar,  Narayan C  Debnath | Towards Logical  Level Design of  Big Data | Dữ liệu mẫu | Dữ liệu cơ sở dữ  liệu quan hệ | - | GOOSSDM | tài liệu | MongoDB | Lập bản đồ quy tắc | GOOSSDM | NA | mô hình điều khiển | Đồng nhất | NA |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Artem  Chebotko,  22 Andrey  Kashlev,  Shiyong Lu | A Big Data  Modeling  Methodology for  Apache Cassandra | Dữ liệu mẫu | - | - | trang mạng | ER | Cột | Cassandra | Quy tắc ánh xạ các mẫu ánh xạ quy trình ứng dụng | Chebotko  Diagrams at logical level | KDM | mô hình truy vấn | Đồng nhất | NA |
| Wenduo Feng∗,  Ping Gu, Chao 23  Zhang, Kai  Zhou | Transforming UML Class  Diagram into  Cassandra Data  Model with  Annotations | Dữ liệu mẫu | - | Tệp JSON | - | ER | Cột | Cassandra | ATL  Ngôn ngữ chuyển đổi (ATL)  Lập bản đồ quy tắc | UML at  conceptual | NA | mô hình điều khiển | Đồng nhất | NA |
| Ling Chen, Jian  Shao, Zhou Yu,  Jianling Sun, 24  Fei Wu,  Yueting  Zhuang | RAISE: A Whole  Process Modeling Method for  Unstructured Data  Management | NA | - | - | trang mạng | XML | - | D-Ocean | NA | NA | NA | mô hình điều khiển | Đồng nhất | NA |
| M. Chevalier, M. El Malki, A. 25 Kopliku, O.  Teste and R.  Tournier | Implementation of  Multidimensional Databases with  Document-Oriented  NoSQL | Dữ liệu mẫu | - | Tệp JSON | - | đa chiều | tài liệu | MongoDB | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển | Đồng nhất | Biến đổi |
| Dewi W.  26 Wardani, Josef  Küng | Semantic Mapping  Relational to  Graph Model | Dữ liệu mẫu | Dữ liệu cơ sở dữ liệu quan hệ | - | - | ER | đồ thị dữ liệu liên kết RDF | Neo4j | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển | Đồng nhất | truy vấn |
| Ming Zhe, 27  Kang Ruihua | A Data Modeling  Approach for  Electronic  Document based on Metamodel | Dữ liệu thực | - | Tệp JSON | tài liệu điện tử | Cây | - | - | NA | NA | NA | mô hình điều khiển | NA | NA |
| Mohamed  Nadjib Mami,  Simon Scerri,  28 S¨oren Auer and  Maria-Esther  Vidal | Towards  Semantification of  Big Data  Technology | Dữ liệu mẫu | - | - | trang mạng | RDF RDF-XML | Cột | Apache Parquet | NA | NA | NA | NA | NA | NA |
| Dan Han, Eleni 29  Stroulia | HGrid: A Data  Model for Large  Geospatial Data  Sets in Hbase | Dữ liệu mẫu | - | - | cảm biến | - | Hgrid | Hbase | NA | NA | NA | mô hình điều khiển | Đồng nhất | truy vấn |
| Zhiyun Zheng,  Zhimeng Du,  30  Lun Li, Yike  Guo | Big Data-Oriented  Open Scalable  Relational Data  Model | Dữ liệu thực | - | - | trang mạng | - | OSRDM | - | NA | NA | NA | mô hình điều khiển | NA | truy vấn |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dongqi Wei,  Chaoling Li,  Wumuti 31  Naheman,  Jianxin Wei,  Junlu Yang | Organizing and  Storing Method for Large-scale  Unstructured Data  Set with Complex  Content | NA | - - | Dữ liệu khoa học địa chất | Cây | - | - | NA | NA | NA | mô hình điều khiển | NA | NA |
| Karamjit Kaur, 32  Rinkle Rani | Modeling and  Querying Data in  NoSQL Databases | Dữ liệu thực | - - | trang mạng | ER | đồ thị tài liệu | MongoDB  Neo4J | NA | UML at  conceptual and document | Neoclipse | mô hình điều khiển | Hỗn hợp | NA |
| Michael J.  Mior, Kenneth  33 Salem, Ashraf  Aboulnaga and  Rui Liu | NoSE: Schema  Design for NoSQL  Applications | Dữ liệu thực | - - | trang mạng | ER | Cột | Cassandra | Bị ràng buộc không gian khối lượng công việc | NA | NoSE | mô hình truy vấn | Đồng nhất | truy vấn |
| Max Chevalier,  Mohammed El  Malki, Arlind  34 Kopliku,  Olivier Teste,  Ronan  Tournier | Document-Oriented Data Warehouses: Models and  Extended Cuboids | Dữ liệu mẫu | Dữ liệu cơ sở dữ  liệu quan hệ | - | đa chiều | tài liệu | MongoDB  PostgrSQL | Lập bản đồ quy tắc | NA | NA | mô hình điều khiển | Hỗn hợp | truy vấn |
| Harley Vera, Wagner  Boaventura,  Maristela  35 Holanda,  Valeria  Guimarães,  Fernanda Hondo | Data Modeling for NoSQL  Document-Oriented  Databases | Dữ liệu mẫu | - - | cảm biến | NGN | tài liệu | MongoDB | NA | NA | NA | mô hình điều khiển | Đồng nhất | NA |
| Robert T.  36  Mason | NoSQL Databases and Data  Modeling  Techniques for a  Document-oriented  NoSQL Database | Dữ liệu mẫu | - - | - | ER | tài liệu | MongoDB | Hoberman heuristic | NA | NA | mô hình điều khiển | Đồng nhất | NA |

# Tham khảo

1. Kitchenham, B. *Procedures for Performing Systematic Reviews*; Keele University: Keele, UK, 2004; Volume 33, pp. 1–26.
2. Google. Google Trends. Available online: [https://trends.google.es/trends/explore?date=all&q=%22big% 20data%22](https://trends.google.es/trends/explore?date=all&q=%22big%20data%22) (accessed on 23 August 2019).
3. Rider, F. *The Scholar and the Future of the Research Library: A Problem and Its Solution*; Hadham Press: New York, NY, USA, 1944; pp. 98–100.
4. Cox, M.; Ellsworth, D. Application-controlled demand paging for out-of-core visualization. In Proceedings of the 8th IEEE Conference on Visualization, Phoenix, AZ, USA, 24 October 1997; pp. 235–244.
5. Ribeiro, A.; Rodrigues da Silva, A. Data Modeling and Data Analytics: A Survey from a Big Data Perspective.

*J. Softw. Eng. Appl.* **2015**, *8*, 617–634. [[CrossRef]](http://dx.doi.org/10.4236/jsea.2015.812058)

1. Shafer, T. The 42 V’s of Big Data and Data Science. Available online: [https://www.elderresearch.com/blog/42v-of-big-data](https://www.elderresearch.com/blog/42-v-of-big-data) (accessed on 23 August 2019).
2. Manogaran, G.; Thota, C.; Lopez, D.; Vijayakumar, V.; Abbas, K.M.; Sundarsekar, R. Big Data Knowledge System in Healthcare. In *Internet of Things and Big Data Technologies for Next Generation Healthcare*; Springer International Publishing: Cham, Switzerland, 2017; Volume 23, pp. 133–157.
3. Persico, V.; Pescapé, A.; Picariello, A.; Sperlí, G. Benchmarking big data architectures for social networks data processing using public cloud platforms. *Future Gener. Comput. Syst.* **2018**, *89*, 98–109. [[CrossRef]](http://dx.doi.org/10.1016/j.future.2018.05.068)
4. Costa, C.; Santos, Y. Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges. *Int. J. Comput. Sci.* **2017**, *44*, 1–17.
5. Davoudian, A.; Chen, L.; Liu, M. A Survey on NoSQL Stores. *ACM Comput. Surv.* **2018**, *51*, 1–43. [[CrossRef]](http://dx.doi.org/10.1145/3158661)
6. CISCO. Big Data: Not Just Big, but Different—Part 2. Available online: [https://www.cisco.com/c/dam/en\_ us/about/ciscoitatwork/enterprise-networks/docs/i-bd-04212014-not-just-big-different.pdf](https://www.cisco.com/c/dam/en_us/about/ciscoitatwork/enterprise-networks/docs/i-bd-04212014-not-just-big-different.pdf) (accessed on 10 September 2019).
7. O’Sullivan, P.; Thompson, G.; Clifford, A. Applying data models to big data architectures. *IBM J. Res. Dev.* **2014**, *58*, 18:1–18:11. [[CrossRef]](http://dx.doi.org/10.1147/JRD.2014.2352474)
8. CISCO VNI. Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper. Available online: [https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/](https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.htm)

[white-paper-c11-741490.htm](https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.htm) (accessed on 10 September 2019).

1. Karamjit, K.; Rinkle, R. Modeling and querying data in NoSQL databases. In Proceedings of the 1st IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 1–7.
2. Wu, D.; Sakr, S.; Zhu, L. Big Data Storage and Data Models. In *Handbook of Big Data Technologies*; Springer International Publishing: Cham, Switzerland, 2017; pp. 3–29. [[CrossRef]](http://dx.doi.org/10.1007/978-3-319-49340-4_1)
3. Chebotko, A.; Kashlev, A.; Lu, S. A Big Data Modeling Methodology for Apache Cassandra. In Proceedings of the 4th IEEE International Congress on Big Data, New York, NY, USA, 27 June–2 July 2015; pp. 238–245.
4. Edlich, S. List of NoSQL Database Management Systems. Available online: <http://nosql-database.org/>(accessed on 15 September 2019).
5. Santos, M.Y.; Martinho, B.; Costa, C. Modelling and implementing big data warehouses for decision support. *J. Manag. Anal.* **2017**, *4*, 111–129. [[CrossRef]](http://dx.doi.org/10.1080/23270012.2017.1304292)
6. Centre for Reviews and Dissemination; (University of York: Centre for Reviews and Dissemination, York, UK). Undertaking systematic reviews of research on effectiveness: CRD’s guidance for carrying out or commissioning reviews. Personal communication, 2001.
7. Martins de Sousa, V.; del Val Cura, L.M. Modelagem Lógica para Bancos de Dados NoSQL: Uma revisão sistemática. *Anais WCF* **2016**, *3*, 32–39.
8. Brewer, E.A. Towards robust distributed systems. In Proceedings of the ACM Symposium on Principles of Distributed Computing, Portland, Oregon, 16–19 July 2000; Volume 7.
9. Pouyanfar, S.; Yang, Y.; Chen, S.; Shyu, M.L.; Iyengar, S. Multimedia big data analytics: A survey. *ACM Comput. Surv.* **2018**, *51*, 10. [[CrossRef]](http://dx.doi.org/10.1145/3150226)
10. Bruno, R.; Ferreira, P. A Study on Garbage Collection Algorithms for Big Data Environments. *ACM Comput.*

*Surv.* **2018**, *51*, 20. [[CrossRef]](http://dx.doi.org/10.1145/3156818)

1. Gusenbauer, M.; Haddaway, N. Which Academic Search Systems are Suitable for Systematic Reviews or

Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed and 26 other Resources. *Res. Synth. Methods* **2019**. [[CrossRef]](http://dx.doi.org/10.1002/jrsm.1378) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/31614060)

1. Song, J.; He, H.; Thomas, R.; Bao, Y.; Yu, G. Haery: A Hadoop based Query System on Accumulative and High-dimensional Data Model for Big Data. *IEEE Trans. Knowl. Data Eng.* **2019**. [[CrossRef]](http://dx.doi.org/10.1109/TKDE.2019.2904056)
2. Thiry, L.; Zhao, H.; Hassenforder, M. Categories for (Big) Data models and optimization. *J. Big Data* **2018**, *5*, 1–20. [[CrossRef]](http://dx.doi.org/10.1186/s40537-018-0132-9)
3. Martins de Sousa, V.; del Val Cura, L.M. Logical Design of Graph Databases from an Entity-Relationship Conceptual Model. In Proceedings of the 20th International Conference on Information Integration and Web-Based Applications and Services, Yogyakarta, Indonesia, 19–21 November 2018; pp. 183–189.
4. Zecˇevic´, I.; Bjeljac, P.; Perišic´, B.; Stankovski, S.; Venus, D.; Ostojic´, G. Model driven development of hybrid databases using lightweight metamodel extensions. *Enterp. Inf. Syst.* **2018**, *12*, 1221–1238. [[CrossRef]](http://dx.doi.org/10.1080/17517575.2018.1445295)
5. Rinaldi, A.; Russo, C. A Semantic-based Model to represent Multimedia Big Data. In Proceedings of the 10th International Conference on Management of Digital EcoSystems, Tokyo, Japan, 25–28 September 2018; pp. 31–38.
6. Hamouda, S.; Zainol, Z. Document-Oriented Data Schema for Relational Database Migration to NoSQL. In Proceedings of the 2017 International Conference on Big Data Innovations and Applications, Prague, Czech Republic, 21–23 August 2018; pp. 43–50.
7. Aggarwal, D.; Davis, K. Employing Graph Databases as a Standardization Model for Addressing Heterogeneity and Integration. *Adv. Intell. Syst. Comput.* **2018**, *561*, 109–138. [[CrossRef]](http://dx.doi.org/10.1007/978-3-319-56157-8_6)
8. De la Vega, A.; García-Saiz, D.; Blanco, C.; Zorrilla, M.; Sánchez, P. Mortadelo: A Model-Driven Framework for NoSQL Database Design. In Proceedings of the 8th International Conference on Model and Data Engineering, Marrakesh, Morocco, 24–26 October 2018; pp. 41–57.
9. Chen, X.; Yan, L.; Li, W.; Zhang, F. Fuzzy spatio-temporal data modeling based on XML schema. *Filomat* **2018**, *32*, 1663–1677. [[CrossRef]](http://dx.doi.org/10.2298/FIL1805663C)
10. Shin, K.; Hwang, C.; Jung, H. NoSQL Database Design Using UML Conceptual Data Model Based on Peter Chen’s Framework. *Int. J. Appl. Eng. Res.* **2017**, *12*, 632–636.
11. Abdelhedi, F.; Brahim, A.A.; Atigui, F. Logical unified modeling for NoSQL databases. In Proceedings of the 19th International Conference on Enterprise Information Systems, Porto, Portugal, 26–29 April 2017; pp. 249–256.
12. Mohan, A.; Ebrahimi, M.; Lu, S.; Kotov, A. A NoSQL Data Model for Scalable Big Data Workflow Execution. In Proceedings of the 2016 IEEE International Congress on Big Data, San Francisco, CA, USA, 27 June–2 July 2016; pp. 52–59.
13. Villari, M.; Celesti, A.; Giacobbe, M.; Fazio, M. Enriched E-R Model to Design Hybrid Database for Big Data Solutions. In Proceedings of the 2016 IEEE Symposium on Computers and Communication, Messina, Italy, 27–30 June 2016; pp. 163–166.
14. Santos, M.Y.; Costa, C. Data Warehousing in Big Data: From Multidimensional to Tabular Data Models. In Proceedings of the 9th International C\* Conference on Computer Science and Software Engineering, Porto, Portugal, 20–22 July 2016; Volume 20, pp. 51–60.
15. Santos, M.Y.; Costa, C. Data Models in NoSQL Databases for Big Data Contexts. In Proceedings of the International Conference on Data Mining and Big Data, Bali, Indonesia, 25–30 June 2016; Volume 9714, pp. 475–485.
16. Solanke, G.B.; Rajeswari, K. Migration of Relational Database to MongoDB and Data Analytics using Naïve Bayes Classifier based on Mapreduce Approach. In Proceedings of the 2017 International Conference on Computing, Communication, Control and Automation, Maharashtra, India, 17–18 July 2017; pp. 1–6.
17. Reniers, V.; Van Landuyt, D.; Rafique, A.; Joosen, W. Schema Design Support for Semi-Structured Data: Finding the Sweet Spot between NF and De-NF. In Proceedings of the 2017 IEEE International Conference on Big Data, Boston, MA, USA, 11–14 December 2017; pp. 2921–2930.
18. Abdelhedi, F.; Ait Brahim, A.; Atigui, F.; Zurfluh, G. Big Data and Knowledge Management: How to Implement Conceptual Models in NoSQL Systems. In Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), Porto, Portugal, 9–11 November 2016; pp. 235–240.
19. Chevalier, M.; El Malki, M.; Kopliku, A.; Teste, O.; Tournier, R. Document-Oriented Models for Data

Warehouses: NoSQL Document-Oriented for Data Warehouses. In Proceedings of the 18th International

Conference on Enterprise Information Systems (ICEIS 2016), Rome, Italy, 25–28 April 2016; pp. 142–149. [[CrossRef]](http://dx.doi.org/10.5220/0005830801420149)

1. Banerjee, S.; Shaw, R.; Sarkar, A.; Debnath, N.C. Towards Logical Level Design of Big Data. In Proceedings of the 13th IEEE International Conference on Industrial Informatics, Cambridge, UK, 22–24 July 2015; pp. 1665–1671.
2. Feng, W.; Gu, P.; Zhang, C.; Zhou, K. Transforming UML Class Diagram into Cassandra Data Model with Annotations. In Proceedings of the IEEE International Conference on Smart City/SocialCom/SustainCom, Chengdu, China, 19–21 December 2015; pp. 798–805.
3. Chen, L.; Shao, J.; Yu, Z.; Sun, J.; Wu, F.; Zhuang, Y. RAISE: A Whole Process Modeling Method for Unstructured Data Management. In Proceedings of the 2015 IEEE International Conference on Multimedia Big Data, Beijing, China, 20–22 April 2015; pp. 9–12.
4. Chevalier, M.; Malki, M.; Kopliku, A.; Teste, O.; Tournier, R. Implementation of Multidimensional Databases with Document-Oriented NoSQL. *Lect. Notes Comput. Sci.* **2015**, *9263*, 379–390. [[CrossRef]](http://dx.doi.org/10.1007/978-3-319-22729-0_29)
5. Wardani, D.; Küng, J. Semantic Mapping Relational to Graph Model. In Proceedings of the 2014 International Conference on Computer, Control, Informatics and Its Applications, Bandung, Indonesia, 21–23 October 2014; pp. 160–165.
6. Zhe, M.; Ruihua, K. A Data Modeling Approach for Electronic Document Based on Metamodel. In Proceedings of the 2013 International Conference on Computer Sciences and Applications, Wuhan, China, 14–15 December 2013; pp. 829–832.
7. Mami, M.N.; Scerri, S.; Auer, S.; Vidal, M.E. Towards Semantification of Big Data Technology. In Proceedings of the 18th International Conference on Big Data Analytics and Knowledge Discovery, Porto, Portugal, 6–8 September 2016; pp. 376–390.
8. Han, D.; Stroulia, E. HGrid: A Data Model for Large Geospatial Data Sets in HBase. In Proceedings of the 6th International Conference on Cloud Computing, Shanghai, China, 9–11 November 2013; pp. 910–917.
9. Zheng, Z.; Du, L.; Guo, Y. BigData oriented open scalable relational data model. In Proceedings of the 3rd IEEE International Congress on Big Data, Washington, DC, USA, 27–30 October 2014; pp. 398–405.
10. Wei, D.; Li, C.; Naheman, W.; Wei, J.; Yang, J. Organizing and Storing Method for Large-scale Unstructured Data Set with Complex Content. In Proceedings of the 5th International Conference on Computing for Geospatial Research and Application, Washington, DC, USA, 4–6 August 2014; pp. 70–76.
11. Mior, M.; Salem, K.; Aboulnaga, A.; Liu, R. NoSE: Schema design for NoSQL applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2275–2289. [[CrossRef]](http://dx.doi.org/10.1109/TKDE.2017.2722412)
12. Chavalier, M.; El Malki, M.; Kopliku, A.; Teste, O.; Tournier, R. Document-Oriented Data Warehouses: Models and Extended Cuboids. In Proceedings of the 10th IEEE International Conference on Research Challenges in Information Science, Grenoble, France, 1–3 June 2016; pp. 1–11.
13. Vera,H.; Boaventura,W.; Holanda,M.; Guimaraes,V.; Hondo,F.DatamodelingforNoSQLdocument-oriented databases. In Proceedings of the CEUR Workshop, Turin, Italy, 28–29 September 2015; pp. 129–135.
14. Mason, R.T. NoSQL Databases and Data Modeling Techniques for a Document-oriented NoSQL Database. In Proceedings of the Informing Science & IT Education Conference, Tampa, FL, USA, 2–5 July 2015; pp. 259–268.
15. Webster, J.; Watson, R. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *Manag. Inf.*

*Syst.* **2002**, *26*, 13–23.

1. ACENS. Bases de Datos NoSQL. Qué son y Tipos que nos Podemos Encontrar. Available online: [https: //www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf](https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf) (accessed on 20 September 2019).
2. IBM and IBM Knowledge Center. Data Driven Modeling. Available online: [https://www.ibm.com/support/ knowledgecenter/en/SSGTJF/com.ibm.help.omcloud.omniconfig.doc/productconcepts/c\_OC\_DDMIntro. html](https://www.ibm.com/support/knowledgecenter/en/SSGTJF/com.ibm.help.omcloud.omniconfig.doc/productconcepts/c_OC_DDMIntro.html) (accessed on 20 September 2019).
3. Abelló, A. Big Data Design. In Proceedings of the 18th International Workshop on Data Warehousing and OLAP, Melbourne, Australia, 23 October 2015; pp. 35–38.
4. Schaarschmidt, M.; Gessert, F.; Ritter, N. Towards automated polyglot persistence. In Proceedings of the Datenbanksysteme für Business, Technologie und Web, Stuttgart, Germany, 6–7 March 2015.
5. Sevilla Ruiz, D.; Morales, S.F.; Garcia Molina, J. Inferring Versioned Schemas from NoSQL Databases and Its

Applications. *Lect. Notes Comput. Sci.* **2015**, *9381*, 467–480. [[CrossRef]](http://dx.doi.org/10.1007/978-3-319-25264-3_35)

1. Solid, I.T. DB-Engines Ranking—Trend Popularity. Available online: [https://db-engines.com/en/ranking\_ trend](https://db-engines.com/en/ranking_trend) (accessed on 1 January 2020).
2. Dell’Aglio, D.; Balduini, M.; Della Valle, E. Applying semantic interoperability principles to data stream management. In *Data Management in Pervasive Systems*; Springer: Cham, Switzerland, 2015; pp. 135–166. [[CrossRef]](http://dx.doi.org/10.1007/978-3-319-20062-0_7)
3. Haghighi, M. Market-based resource allocation for energy-efficient execution of multiple concurrent applications in wireless sensor networks. In *Mobile, Ubiquitous, and Intelligent Computing*; Springer: Berlin, Germany, 2014; pp. 173–178. [[CrossRef]](http://dx.doi.org/10.1007/978-3-642-40675-1_27)
4. Bajcsy, R.; Joshi, A.; Krotkov, E.; Zwarico, A. Landscan: A natural language and computer vision system for analyzing aerial images. In Proceedings of the 9th International Joint Conference on Artificial Intelligence, Los Angeles, CA, USA, 18–23 August 1985; Volume 2, pp. 919–921.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license [(http://creativecommons.org/licenses/by/4.0/)](http://creativecommons.org/licenses/by/4.0/./).