International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

Big Data and NoSQL Databases Architecture: A Review

**Moko, Anasuodei**

Khoa Khoa học Máy tính và Tin học,

Đại học Liên bang Otuoke,

Bang Bayelsa, Nigeria.

**Asagba, Prince Oghenekaro ,**

Khoa Khoa học Máy tính,

Đại học Port Harcourt,

Bang Rivers, Nigeria.

***Tóm tắt***

*Dữ liệu trong thời gian gần đây ngày càng trở nên lớn hơn đặc biệt là tăng về khối lượng, khiến dữ liệu phi cấu trúc trở nên khó xử lý và quản lý, vì cơ sở dữ liệu có cấu trúc sẽ gặp khó khăn khi xử lý dữ liệu phi cấu trúc do kích thước của nó quá lớn. Do đó, để hiểu được Dữ liệu lớn, cần có các cấu trúc và phương pháp mới, bài báo này xem xét Dữ liệu lớn và các tính năng của nó, bốn kiến trúc NoSQL có sẵn để xử lý Dữ liệu lớn, điểm mạnh, điểm yếu và các loại có sẵn.*

***Từ khóa:*** *Big Data, Unstructured, NoSQL, Database, Structured*

1. **Giới thiệu**

Một thách thức công nghệ lớn mà thế giới đang phải đối mặt là quản lý và lưu trữ dữ liệu, trong đó hàng triệu dữ liệu trong thời gian gần đây đang được tạo ra với tốc độ dưới nano giây. Do đó, việc xử lý số lượng lớn dữ liệu là một thách thức đáng kể và do đó với sự gia tăng dân số thì cần phải có công nghệ quản lý và thu thập dữ liệu hiện đại nhất.

(Sultana và cộng sự 2017) nhấn mạnh rằng nhu cầu xử lý và tạo dữ liệu nhanh chóng trong thời gian gần đây đã dẫn đến thực tế là hơn 2,6 nghìn tỷ dữ liệu đang được tạo ra hàng ngày.

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 1 |

Họ tiếp tục dự đoán rằng trong tương lai sẽ có sự gia tăng theo cấp số nhân của việc sử dụng và tạo ra dữ liệu. Nêu rõ rằng việc quản lý và lưu trữ dữ liệu trong thời đại dữ liệu lớn, nơi phần mềm thông thường đang bị đẩy đến giới hạn bởi lượng dữ liệu khổng lồ, tạo ra nhu cầu thay đổi đáng kể trong việc đầu tư, thu thập và lưu trữ dữ liệu để sử dụng trong tương lai, do đó khiến các tổ chức khác nhau phải nỗ lực có ý thức để bảo mật và lưu giữ mọi phần dữ liệu tiềm năng. Mặc dù dữ liệu thường không có cấu trúc, nó có thể được tạo ra từ nhiều nguồn khác nhau, chẳng hạn như bài đăng trên phương tiện truyền thông xã hội, nội dung đa phương tiện với kho lưu trữ tự động. Email, truy vấn công cụ tìm kiếm, kho tài liệu quản lý nội dung, dữ liệu cảm biến thuộc nhiều loại khác nhau, trao đổi chứng khoán, hình ảnh vệ tinh, hệ thống giám sát và ứng dụng v.v…

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

Nó đôi khi được hiểu là 'Not Only SQL', do đó nó có thể diễn đạt thực tế là các công nghệ khác được sử dụng trong các ứng dụng web được phân phối hàng loạt, bên cạnh các công nghệ dữ liệu quan hệ.

Quan trọng nhất, công nghệ NoSQL là bắt buộc nếu dịch vụ web yêu cầu tính khả dụng cao. Nó là một cấu trúc lưu trữ phân tán rộng rãi, có cấu trúc hệ thống quản lý cơ sở dữ liệu hữu ích cơ bản. Dữ liệu thực tế của các giá trị chính được lưu trữ trong các cặp, cột hoặc nhóm cột, tài liệu và đồ thị. Do đó, các khái niệm dự phòng khác nhau được hỗ trợ để đảm bảo tránh thất bại và tăng tính khả dụng của các hệ thống cơ sở dữ liệu NoSQL. (Francis, 2019)

Đối với tất cả các phương pháp quản lý dữ liệu phi quan hệ, thuật ngữ NoSQL được sử dụng để đáp ứng các tiêu chí được liệt kê dưới đây:

1. Thông tin không được lưu trong bảng
2. Ngôn ngữ SQL không phải là ngôn ngữ cơ sở dữ liệu.

Tuy nhiên, công nghệ cơ sở dữ liệu có tính tương đối với công nghệ NoSQL cần được mở rộng để tạo ra khả năng truy cập toàn cầu không đổi đối với các dịch vụ mà nó cung cấp cho các ứng dụng Web sâu rộng hoặc các ứng dụng xử lý cơ sở dữ liệu lớn. (Ali, 2019)

"Dữ liệu lớn" trong khái niệm tạo dữ liệu rất hữu ích, vì nó giúp lưu trữ và quản lý vô số thông tin được tạo ra mỗi giây, mỗi ngày. Mặc dù đã có các hệ thống quản lý cơ sở dữ liệu quan hệ (RDMS) để lưu trữ dữ liệu cố định trong một số năm, khả năng mở rộng, tính nhất quán, hiệu quả hệ thống, thu thập dữ liệu và tích hợp trích xuất dữ liệu vẫn chưa được giải quyết.

Dữ liệu lớn là công cụ cho phép chúng ta hiểu được nhiều ý nghĩa hơn từ việc học bằng máy móc dữ liệu và những công cụ cho phép chúng ta duy trì khối lượng dữ liệu cao hơn với độ chi tiết cao hơn bao giờ hết. Những công nghệ Dữ liệu lớn này đã được Google đi tiên phong, và họ đã tìm thấy con đường của mình dưới hình thức Hadoop vào cộng đồng CNTT rộng lớn hơn.

NoSQL hoạt động để giúp xử lý các yêu cầu về *Dung lượng, Tính đa dạng và Tốc độ dữ liệu lớn* (Andreas & Michael, 2019) giải thích rằng Dữ liệu lớn vẫn chưa có định nghĩa ràng buộc. Tuy nhiên, nhiều chuyên gia dữ liệu sẽ đồng ý về ba chữ V (3V): *Dung lượng* cho khối lượng dữ liệu mở rộng; *Có nhiều định dạng*, dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc; *Tốc độ* để xử lý dữ liệu tốc độ cao và thời gian thực.

(Nzar & Dashne 2019) đã phân loại các tính năng của Dữ liệu lớn thành ― 5V: dung lượng, tốc độ, sự đa dạng, tính xác thực và giá trị như được mô tả dưới đây:

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 2 |

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

A picture containing text, electronics

Description automatically generated

**Hình 1**: Big Data 5 V 's (Nguồn: Nzar & Dashne, 2019)

1. *Dung lượng*: Hiển thị số lượng lớn dữ liệu, chẳng hạn như dữ liệu cho thiết bị di động, được sử dụng cho các chức năng khác nhau.
2. *Tốc độ*: chỉ định tốc độ hoặc tần suất tạo, cập nhật, xử lý và truy cập dữ liệu.
3. *Sự đa dạng:* dữ liệu được truy cập thông qua các loại thiết bị khác nhau, chẳng hạn như video, ảnh,...
4. *Giá trị*: giải thích cách rút ra kiến thức hữu ích từ các bộ dữ liệu khổng lồ. Khía cạnh quan trọng nhất của bất kỳ công cụ dữ liệu lớn nào là giá trị, vì nó cho phép tạo ra kiến thức có giá trị.
5. *Tính xác thực*: Đề cập đến độ chính xác và giá trị lớn của thông tin.

(Khan, et al, 2017) phân loại từ 1-6 V của dữ liệu lớn, phát triển thành dữ liệu có giá trị, biến nó thành 7 V. Bảy chữ V là – *Dung lượng, Tốc độ, Đa dạng, Biến đổi, Độ chính xác, Hình ảnh hóa và Giá trị.*

Chart, pie chart

Description automatically generated  
 **Hình 2:** Bảy chữ V của dữ liệu lớn (Nguồn: Syed et al, 2019)

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 3 |

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

1. **Rà soát các công trình liên quan**

Kalid và cộng sự (2017) Bài báo nhấn mạnh tài liệu giải quyết các vấn đề và tìm giải pháp thông qua việc sử dụng các phương pháp dữ liệu lớn tương phản để quản lý cơ sở dữ liệu No SQL; BigTable, DybanoDB và Cassandra. Công trình cũng chỉ ra rằng các công ty lớn không có hệ thống lưu trữ chính xác không cần phải lưu trữ và quản lý dữ liệu lớn. Nghiên cứu cho thấy rằng BigTable từ Google và DynamoDB từ Amazon là rất quan trọng và mang lại lợi nhuận cho chính họ và Casandra là sự kết hợp của cả hai hệ thống. Md.Razu và cộng sự (2018) đã nghiên cứu cơ sở dữ liệu NoSQL để xử lý Dữ liệu lớn, bao gồm các vấn đề về cấu trúc và giao dịch. Công trình cũng nhấn mạnh các hướng nghiên cứu và thách thức liên quan đến việc xử lý Dữ liệu lớn mà nghiên cứu tin rằng thông tin là đáng kinh ngạc để xem xét tài liệu trên Cơ sở dữ liệu dữ liệu lớn NoSQL, bao gồm dữ liệu cấu trúc, các vấn đề và phương pháp thu thập thông tin hữu ích được đo trong thời gian thực.

Vahid (2016) trong Dữ liệu lớn: Bây giờ và sau đó, mở ra rằng sự tăng trưởng của dữ liệu và kiến thức phụ thuộc vào sự sẵn có của nó trong tay của người tiêu dùng. Điều này rất đơn giản, vì điện thoại di động, máy tính xách tay, PC và nhiều thiết bị khác có thể dễ dàng truy cập. Trong trường hợp này, tăng trưởng kiến thức đang nhanh chóng tràn ngập - vì nội dung mới được tạo ra bởi chính người dùng để đáp ứng sự cần thiết của việc quản lý, sử dụng, phân loại và bảo mật dữ liệu đó. Bài báo cố gắng sắp xếp và phân tích các quy trình, các vấn đề tiềm ẩn, nghiên cứu và sáng kiến Dữ liệu lớn và hiển thị hiện tượng gia tăng mãnh liệt này từ một chân trời tiềm năng.

Ali et al (2019) cho rằng NoSQL được thay đổi bởi hầu hết các cơ sở dữ liệu quan hệ được sử dụng để lưu trữ dữ liệu, nhưng nó không thay thế hoàn toàn SQL, như tên chỉ ra. Bài báo thảo luận về cơ sở dữ liệu SQL và NoSQL, sự tương phản của SQL thông thường với cơ sở dữ liệu NoSQL của Big Data Analytics, các mẫu dữ liệu NoSQL, các biểu mẫu lưu trữ dữ liệu NoSQL, các tính năng và tính năng của từng lưu trữ dữ liệu, các ưu và nhược điểm của NoSQL và RDBMS.

1. **Ngôn ngữ truy vấn có cấu trúc (SQL)**

SQL (ngôn ngữ truy vấn có cấu trúc) mặt khác theo truyền thống là các cơ sở dữ liệu nổi tiếng nhất từ đầu những năm 1970. Một ví dụ là Cơ sở dữ liệu quan hệ (được truy cập bằng cách sử dụng SQL – ví dụ MySQL) nơi dữ liệu được lưu trữ trong một bảng có các hàng và cột. Các nhà phát triển hồi đó chủ yếu thực hiện các thiết kế của họ theo mô hình phát triển phần mềm thác nước. Điều này có nghĩa là mỗi giai đoạn phát triển phần mềm đều được lên kế hoạch tốt trước khi bắt đầu phát triển bằng cách sử dụng một mối quan hệ thực thể phức tạp kỹ lưỡng để đảm bảo rằng tất cả những gì cần thiết trong cơ sở dữ liệu đã được suy nghĩ cẩn thận và cung cấp (Schaefer, 2015). Mặc dù cơ sở dữ liệu quan hệ hữu ích hơn nhiều, có những thách thức đối với các nhà phát triển phần mềm, chẳng hạn như nếu có một chút cải thiện trong chu kỳ phát triển của phần mềm, các nhà phát triển sẽ phải đấu tranh để giữ cho dự án dưới ngân sách quy định và phần mềm có thể không đáp ứng được nhu cầu của người dùng. Một lý do khác là đã có sự gia tăng đột biến về số lượng dữ liệu được tạo ra, đó là do một loại cơ sở dữ liệu mới NoSQL.

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 4 |

Một cách tiếp cận quản lý mới được coi là cần thiết để hỗ trợ các ứng dụng như xem xét các tệp nhật ký theo thời gian thực, giao dịch thương mại điện tử và dữ liệu được đăng lên

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

phương tiện truyền thông xã hội có dung lượng khổng lồ. Một cách tiếp cận khác phải được đưa vào thực tế để xử lý sự gia tăng dữ liệu bất thường này được tạo ra và để vượt qua tất cả những thách thức nêu trên. Các cơ sở dữ liệu được giới thiệu, là các cơ sở dữ liệu NoSQL, mặc dù có một số mức độ không hiệu quả do khối lượng sản xuất dữ liệu lớn và thiếu hỗ trợ cho các thuộc tính ACID.

1. **Dữ liệu lớn và cơ sở dữ liệu**

Dữ liệu lớn có thể được lưu trữ bằng cách sử dụng cả Cơ sở dữ liệu có cấu trúc (MySQL là cơ sở dữ liệu quan hệ) và Cơ sở dữ liệu không cấu trúc (MongoDB là cơ sở dữ liệu không quan hệ). Xem xét sự thay đổi về thời gian phản hồi của từng loại cơ sở dữ liệu, các thuật toán khác nhau cần được phân tích để tăng cường hiệu suất trong giám sát hệ thống theo thời gian thực đối với cả SQL và NoSQL cập nhật và chèn dữ liệu lớn.

Hệ thống cơ sở dữ liệu NoSQL được gọi là hệ thống lưu trữ dựa trên web miễn là chúng đáp ứng các yêu cầu sau:

* Mô hình: Mô hình cơ sở dữ liệu cơ bản không phải là quan hệ.
* Ít nhất 3 V: Một lượng lớn dữ liệu (dung lượng), cấu trúc dữ liệu linh hoạt (đa dạng) và xử lý theo thời gian thực được bao gồm trong hệ thống cơ sở dữ liệu (tốc độ).
* Lược đồ: Lược đồ cơ sở dữ liệu tập hợp không bị ràng buộc bởi hệ thống quản lý cơ sở dữ liệu.
* Cấu trúc: Cấu trúc cơ sở dữ liệu hỗ trợ mở rộng theo chiều ngang và các ứng dụng web phân tán đầy đủ.
* Sao chép: Hệ thống quản lý cơ sở dữ liệu hỗ trợ việc sao chép dữ liệu.
* Đảm bảo tính nhất quán: tính nhất quán có thể được đảm bảo với độ trễ để ưu tiên tính sẵn có và khả năng chịu đựng cao của các phân vùng. (2019 Andreas & Michael)

1. **Cấu trúc Big Data NoSQL**

So với 'Not SQL', NoSQL có nghĩa là 'Not Only SQL' vì nhiều người coi nó là một loại cơ sở dữ liệu giúp thực hiện các hoạt động dữ liệu lớn và lưu trữ chúng ở định dạng hợp lệ. Nó thường được sử dụng vì sự đơn giản và phạm vi dịch vụ rộng. Mô hình kiến trúc của nó cung cấp một cách hợp lý để dữ liệu được xử lý trên cơ sở dữ liệu. (Nzar&Dashne,2019)

Các tính năng NoSQL:

* Giản đồ tự do
* Ultimately consistent (as in the BASE property)
* Sao chép các kho dữ liệu để loại bỏ một điểm hỏng duy nhất.
* Có khả năng xử lý nhiều loại dữ liệu và khối lượng lớn dữ liệu.

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 5 |

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

1. **Kiểu cấu trúc của NOSQL:**

Dữ liệu được lưu trữ trong NoSQL tuân theo bất kỳ bốn mẫu cấu trúc dữ liệu.

1. Cơ sở dữ liệu lưu trữ khóa giá trị
2. Cơ sở dữ liệu hàng cột
3. Cơ sở dữ liệu tài liệu
4. Cơ sở dữ liệu đồ thị
5. **Cơ sở dữ liệu lưu trữ khóa giá trị**

Một trong những mô hình cơ sở dữ liệu NoSQL cơ bản nhất là mô hình này. Dữ liệu được thu thập theo mô hình Key-Value Pairs, như tên gọi của nó. Một loạt các chuỗi, số nguyên hoặc ký tự thường là khóa, nhưng nó cũng có thể là một dạng dữ liệu nâng cao hơn. Thông thường, giá trị được kết nối hoặc đồng liên quan đến khóa. Các cơ sở dữ liệu cho lưu trữ cặp khóa-giá trị thường lưu trữ thông tin dưới dạng bảng băm trong đó mỗi khóa là duy nhất. Giá trị có thể ở bất kỳ dạng nào (Ký hiệu đối tượng JavaScript (JSON), Đối tượng lớn nhị phân (BLOB), chuỗi, v.v.). Phong cách cấu trúc này thường được sử dụng trong các trang web mua sắm hoặc các ứng dụng thương mại điện tử và tài sản quan trọng của nó là khả năng quản lý rộng rãi khối lượng dữ liệu, tải nặng và dễ dàng sử dụng các khóa để truy xuất dữ liệu.

Diagram

Description automatically generated

**Hình 3**. Một ví dụ về khóa giá trị

Các hạn chế liên quan đến cơ sở dữ liệu lưu trữ giá trị chính là sự phức tạp của nó trong việc xử lý các truy vấn sẽ cố gắng bao gồm nhiều cặp khóa giá trị có thể trì hoãn đầu ra và có thể khiến dữ liệu xung đột với nhiều mối quan hệ với nhiều người.

Một vài ví dụ :

* DynamoDB (được phát triển bởi Amazon)
* Berkeley DB (được phát triển bởi Oracle)

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 6 |

* REDIS: Một cửa hàng khóa giá trị nguồn mở nâng cao, còn được gọi là máy chủ cấu trúc dữ liệu vì các khóa có thể bao gồm chuỗi, hàm băm, danh sách, bộ và các bộ được sắp International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

xếp. Sản phẩm này, được viết bằng C/C++, cực kỳ nhanh chóng, làm cho nó hoàn hảo để thu thập dữ liệu trong thời gian thực.

* Riak: Một nguồn mở là một cơ sở dữ liệu phân tán mạnh mẽ, có thể dự đoán khả năng và đơn giản hóa việc tạo ra bằng cách tạo mẫu, phát triển và triển khai các ứng dụng một cách nhanh chóng. Được viết bằng Erlang và C, công nghệ này cung cấp chức năng chống lỗi/lỗi vượt qua, một API toàn diện và linh hoạt hoàn hảo cho các hệ thống kiểm soát điểm bán hàng và nhà máy.
* VoltDB: Cơ sở dữ liệu có thể mở rộng trong bộ nhớ cung cấp tính nhất quán của AXIT giao dịch hoàn chỉnh và thông lượng cực cao, tự giới thiệu là NewSQL. Công nghệ này dựa vào phân đoạn và sao chép để đạt được ảnh chụp nhanh dữ liệu có sẵn cao và ghi nhật ký lệnh bền vững bằng cách sử dụng các quy trình được lưu trữ Java (để phục hồi sự cố), làm cho nó lý tưởng cho thị trường vốn, mạng kỹ thuật số, dịch vụ mạng và cho cờ bạc trực tuyến.

1. **Cơ sở dữ liệu hàng cột**

Mô hình này sử dụng lưu trữ dữ liệu trong các ô riêng lẻ được chia thành các cột, thay vì lưu trữ dữ liệu trong các cấu trúc quan hệ. Các cơ sở dữ liệu định hướng cột chỉ hoạt động trên các cột. Chúng cùng nhau lưu trữ một lượng lớn dữ liệu trong các cột. Định dạng cột và tiêu đề sẽ phân kỳ từ hàng này sang hàng khác. Mỗi cột được xử lý khác nhau, nhưng vẫn, giống như các cơ sở dữ liệu thông thường, mỗi cột riêng lẻ sẽ chứa một số cột khác. (Niharika) 2020

Về cơ bản, các cột ở dạng chế độ lưu trữ này. Dữ liệu có sẵn và có thể thực hiện các truy vấn như Số, TRUNG BÌNH, ĐẾM trên các cột một cách dễ dàng.

Table

Description automatically generated

**Hình 4.** Ví dụ về lưu trữ hàng cột

Những trở ngại đối với hệ thống này bao gồm: các giao dịch nên tránh hoặc không được hỗ trợ, các truy vấn có thể làm giảm hiệu suất cao khi nối bảng, cập nhật và xóa bản ghi làm giảm hiệu quả lưu trữ và có thể khó thiết kế các lược đồ phân vùng/lập chỉ mục hiệu quả.

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 7 |

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

Một vài ví dụ:

* HBase: HBase là một Cửa hàng dữ liệu lớn phân tán, di động được mô hình hóa theo công nghệ BigTable của Google, cơ sở dữ liệu Hadoop.
* Google 's BigTable
* Cassandra: Một hệ thống quản lý cơ sở dữ liệu phân tán mã nguồn mở được xây dựng để quản lý khối lượng dữ liệu rất lớn nằm rải rác trên nhiều máy chủ mà không có một điểm hỏng hóc nào trong khi cung cấp một dịch vụ có khả năng truy cập cao. Được viết bằng
* Java, sản phẩm này phù hợp nhất cho phân tích dữ liệu thời gian thực không giao dịch với khả năng mở rộng tuyến tính và khả năng chịu lỗi đã được chứng minh kết hợp với các chỉ mục cột.

1. **Cơ sở dữ liệu tài liệu**

Ở dạng các cặp khóa-giá trị, cơ sở dữ liệu bản ghi tìm nạp và tích lũy thông tin, nhưng ở đây các giá trị được gọi là tài liệu. Một cấu trúc dữ liệu phức tạp có thể được biểu diễn dưới dạng văn bản. Tài liệu có thể ở dạng văn bản, mảng, chuỗi, JSON (JavaScript Object Notation), XML (Extensible Markup Language) hoặc bất kỳ định dạng nào khác. Việc sử dụng các tài liệu lồng nhau là vô cùng phổ biến. Nó có hiệu quả cao vì hầu hết các thông tin được tạo ra thường ở dạng JSON và không có cấu trúc.

Shape, rectangle

Description automatically generated

**Hình 5**. Ví dụ về tài liệu

Định dạng này cực kỳ hữu ích và thích hợp cho dữ liệu bán cấu trúc, và nó rất đơn giản để lấy và xử lý tài liệu từ lưu trữ. Các nhược điểm liên quan đến hệ thống này bao gồm yếu tố thách thức của việc xử lý nhiều tài liệu và hoạt động không chính xác của các hoạt động tổng hợp.

Vài ví dụ về các cơ sở dữ liệu:

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 8 |

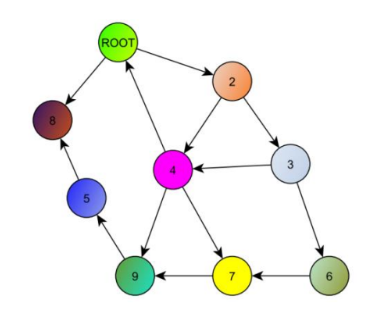
* MongoDB: Cơ sở dữ liệu NoSQL có khả năng mở rộng, hiệu suất cao, nguồn mở này có tính năng lưu trữ định hướng tài liệu (giống JSON), hỗ trợ đầy đủ chỉ mục, sao chép và cập nhật nhanh chóng tại chỗ từ "humongous".Sản phẩm này phù hợp cho các truy vấn động, cấu trúc dữ liệu động, được viết bằng C/C++ và nếu bạn yêu thích các chỉ mục của bạn trên Bản đồ/Giảm.

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

* CouchDB: Cũng là một cơ sở dữ liệu nguồn mở tập trung vào sự dễ dàng lưu trữ dữ liệu trong một loạt các tài liệu JSON, mỗi tài liệu có định nghĩa riêng về lược đồ. Cuối cùng tính nhất quán này được thực thi bởi ngữ nghĩa ACID ngăn chặn việc khóa các tệp cơ sở dữ liệu trong quá trình viết. Sản phẩm này, được viết bằng Java, phù hợp cho các ứng dụng dựa trên web quản lý số lượng lớn dữ liệu được tổ chức lỏng lẻo.

1. **Cơ sở dữ liệu đồ thị**

Mô hình cấu trúc này rõ ràng đề cập đến việc lưu trữ và quản lý thông tin trong đồ thị. Đồ thị về cơ bản là các cấu trúc đại diện cho mối quan hệ giữa hai hoặc nhiều đối tượng trong một số dữ liệu. Các đối tượng hoặc thực thể được gọi là các nút và được kết nối với các mối quan hệ được gọi là cạnh. Có một mã định danh duy nhất trên mỗi cạnh. Đối với đồ thị, mỗi nút đóng vai trò là một điểm tiếp xúc. Trong các mạng xã hội có nhiều và số lượng lớn các thực thể, mô hình này được sử dụng rất rộng rãi và mỗi thực thể có một hoặc nhiều đặc điểm được liên kết bởi các cạnh.



**Hình 6.** Ví dụ về đồ thị

Có các bảng được kết nối lỏng lẻo trong mô hình cơ sở dữ liệu quan hệ, trong khi đồ thị thường mạnh mẽ và cứng nhắc về bản chất, có tốc độ di chuyển nhanh hơn do kết nối và cho phép dữ liệu không gian được xử lý dễ dàng, nhưng kết nối không chính xác có thể dẫn đến vòng lặp vô hạn. (Ian, 2016)

Một vài ví dụ về cơ sở dữ liệu:

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 9 |

* Neo4J: Nền tảng đồ thị và cơ sở dữ liệu đồ thị gốc hàng đầu là Neo4J: Neo4j. Đối với mức độ bảo mật của doanh nghiệp và hiệu suất và độ tin cậy cao bằng cách phân cụm, nó có sẵn cả dưới dạng nguồn mở và thông qua giấy phép thương mại. Cypher, ngôn ngữ truy vấn đồ thị của Neo4j, rất đơn giản để tìm hiểu và có thể sử dụng các bộ công cụ mã nguồn mở mới được phát hành, "Cypher trên Apache Spark (CApS) và Cypher cho Gremlin để hoạt động trên các sản phẩm dựa trên Neo4j, Apache Spark và Gremlin."
* FlockDB (Được sử dụng bởi Twitter): FlockDB dễ dàng hơn so với các cơ sở dữ liệu đồ thị khác vì nó cố gắng giải quyết ít vấn đề hơn. Nó phù hợp theo chiều ngang và được tối International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

ưu hóa cho các môi trường trên đường truyền, độ trễ thấp, thông lượng cao như các trang web.

* ArangoDB: loại cơ sở dữ liệu đồ thị yêu cầu một cơ sở dữ liệu, ngôn ngữ Một Truy vấn, Ba mô hình cho dữ liệu. Khả năng vô hạn. ArangoDB là một cơ sở dữ liệu NoSQL đa mô hình gốc phát triển nhanh chóng, với hơn một triệu lượt tải xuống.
* OrientDB: OrientDB là mô hình đa DBMS phân tán đầu tiên với Trình điều khiển đồ thị thực. Multi-Model có nghĩa là NoSQL thế hệ thứ hai có khả năng quản lý các tên miền phức tạp với hiệu quả đáng kinh ngạc.
* Titan: Titan là một cơ sở dữ liệu đồ thị có thể mở rộng được thiết kế để lưu trữ và truy vấn các đồ thị trải rộng trên một cụm nhiều máy bao gồm hàng trăm tỷ đỉnh và cạnh. Titan là một cơ sở dữ liệu giao dịch có thể tạo điều kiện cho hàng ngàn người dùng đồng thời thực hiện các giao dịch đồ thị phức tạp theo thời gian thực.
* DataStax: Trong một môi trường thay đổi nhanh chóng, nơi khát vọng lớn mạnh, DataStax giúp các doanh nghiệp phát triển mạnh mẽ và các công nghệ mới xuất hiện hàng ngày.
* Amazon Neptune: Amazon Neptune với bộ dữ liệu được kết nối cao để tạo và chạy các ứng dụng an toàn, nhanh chóng và có dịch vụ cơ sở dữ liệu đồ thị được quản lý hoàn toàn dễ dàng. (Niharika, 2020).

|  |  |  |
| --- | --- | --- |
| * **S/N** | **Điểm mạnh** | **Nhược điểm** |
| **i.** | * Phù hợp, có thế mạnh lưu trữ và tra cứu Big Data. | Yêu cầu một hình dung  cơ sở hạ tầng đắt đỏ. |
| **ii.** | Là ứng dụng tập trung. | Quá phức tạp |
| **iii.** | Hỗ trợ dữ liệu HUGE  khả năng | Người sử dụng có thể sẽ thiếu các kỹ năng chuyên môn ở mức tương đối vì hệ thống này còn khá mới và không phải ai cũng biết sử dụng nó một cách thành thạo. |
| **iv.** | Thu nạp dữ liệu nhanh  (tải) | Không có lược đồ |
| **v.** | Tốc độ tra cứu nhanh  (trên các cụm) | Trọng tâm hẹp, chủ yếu được lưu trữ nhưng nó cung cấp rất ít chức năng |
| **vi.** | Cho phép truyền dữ liệu  và dữ liệu hiệu suất cao  không cần máy chủ | Không đủ cho phân tích  Truy vấn (tổng hợp,  số liệu, BI) |

1. **Điểm mạnh và điểm yếu của NOSQL:**
2. **Kết luận:**

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 10 |

Công trình này đã xem xét và nghiên cứu dữ liệu lớn trong thời gian gần đây và tìm ra cách để xử lý khối lượng dữ liệu ngày càng tăng một cách đặc biệt. Dữ liệu trong thời gian gần đây đã trở nên lớn hơn và lớn hơn để xử lý đặc biệt tăng về khối lượng, gây ra dữ liệu phi cấu trúc cụ thể trở nên khó xử lý và quản lý, bởi vì cơ sở dữ liệu có cấu trúc gặp khó khăn khi xử lý dữ liệu phi cấu trúc do kích International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

thước của nó, được gọi là Dữ liệu lớn. Để hiểu được Dữ liệu lớn, cấu trúc và phương pháp mới là cần thiết, công việc này cũng đã kiểm tra các kiến trúc cơ sở dữ liệu NoSQL dữ liệu lớn khác nhau, các loại liên quan đến chúng, tầm quan trọng và cách sử dụng.

**Tài liệu tham khảo**

Ali, W., Shafique, M. U., Majeed, M. A., & Raza, A. (2019). Comparison between SQL and NoSQL Databases and Their Relationship with Big Data Analytics. Asian Journal of Research in Computer Science, 4(2), 1–10.

<https://doi.org/10.9734/ajrcos/2019/v4i230108>

Andreas Meier, & Michael Kaufmann. (2019). SQL & NoSQL Databases Models, Languages, Consistency Options and Architectures for Big Data Management. Springer Vieweg. https://doi.org/10.1007/978-3-658-24549-8

Best Graph Databases in 2020. (2020, May 27). G2. https://www.g2.com/categories/graphdatabases

Big Data and NoSQL Technologies | DB Best Chronicles. (2012, June 15). DB BEST. https://www.dbbest.com/blog/big-data-nosql-technologies/ Column-Oriented Database Technologies | DB Best Chronicles. (2012, July 24). DB BEST. <https://www.dbbest.com/blog/column-oriented-database-technologies/>

Francis, K. K. (2019). NoSQL Databases for Big Data Management: Review and Application in Mobile Commerce. ResearchGate, 9.

<https://doi.org/10.13140/RG.2.2.10239.87204>

Garrett, Alley. (2019, March 4). What Is Big Data Architecture? - DZone Big Data.

Dzone.Com. <https://dzone.com/articles/what-is-big-data-architecture>

Guy Harrison. (2015). Next Generation Databases NoSQL, NewSQL, and Big Data. APRESS.

Ian. (2016, June 16). What is a Graph Database? | Database.Guide.

<https://database.guide/what-is-a-graph-database/>

Kalid, S., Syed, A., Mohammad, A., & Halgamuge, M. N. (2017). Big-data NoSQL databases: A comparison and analysis of ―Big-Table‖, ―DynamoDB‖, and ―Cassandra‖. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(, 89–93. <https://doi.org/10.1109/ICBDA.2017.8078782>

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 11 |

Khan, M. A., Uddin, M. F., & Gupta, N. (2014). Seven V’s of Big Data understanding Big Data to extract value. Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, 1–5.

International Journal of Applied Science and Mathematical Theory E- ISSN 2489-009X P-ISSN 2695-1908, Vol 7. No.1 2020 [www.iiardpub.org](http://www.iiardpub.org)

<https://doi.org/10.1109/ASEEZone1.2014.6820689>

Le, J. (2019, November 20). An Introduction to Big Data: NoSQL. Medium. https://medium.com/cracking-the-data-science-interview/an-introduction-to-big-datanosql-96b882f35e50

Liyakathunisa Syed, Saima Jabeen, S. Manimala, & Hoda Ahmed Galal Elsayed. (2019, January). (5) Data Science Algorithms and Techniques for Smart Healthcare Using IoT and Big Data Analytics: Towards Smarter Algorithms. ResearchGate. https://www.researchgate.net/publication/330723399\_Data\_Science\_Algorithms\_and\_Te chniques\_for\_Smart\_Healthcare\_Using\_IoT\_and\_Big\_Data\_Analytics\_Towards\_Smarte r\_Algorithms#fullTextFileContent

Md. Razu, A., Mst. Arifa, K., Md. Asraf, A., & Kenneth, S. (2018). A literature review on NoSQL database for big data processing. International Journal of Engineering & Technology, 7(2), 902–906.

<https://doi.org/10.14419/ijet.v7i2.12113>

Niharika, P. (2020, January 2). NoSQL Data Architecture Patterns. GeeksforGeeks. <https://www.geeksforgeeks.org/nosql-data-architecture-patterns/>

Raouf, D & Ali, N. (2019). Improving the performance of big data databases. Kurdistan Journal of Applied Research. 4. 206-220. 10.24017/science.2019.2.20.

Schaefer, L. (2015). What is NoSQL? Available at

https://www.mongodb.com/nosqlexplained

The basics of NoSQL databases—And why we need them. (2019, January 31). FreeCodeCamp.Org.

https://www.freecodecamp.org/news/nosql-databases5f6639ed9574

Vahid Rahmati. (2016). Big Data: Now and Then. International Journal of Emerging Computing Methods in Engineering, 1(2), 1–6.

https://www.researchgate.net/publication/309458088\_Big\_Data\_Now\_and\_Then

|  |  |
| --- | --- |
| IIARD – Viện nghiên cứu và phát triển học thuật quốc tế | Trang 12 |