BÀI DỊCH NHÓM 10

Tên thành viên: Nguyễn Quang Thọ

Nguyễn Hoài Sơn

Đặng Lê Quỳnh Mai

Đề tài 29: **Khảo sát công nghệ lưu trữ dữ liệu lớn**

Tóm tắt: Lưu trữ dữ liệu lớn là nền tảng của xử lý và phân tích dữ liệu lớn. Thông qua việc nghiên cứu và tổng kết công nghệ xử lý chính của lưu trữ dữ liệu, bài báo này lần lượt tìm hiểu và phân tích bốn khía cạnh sau: hệ thống tệp phân tán, cơ sở dữ liệu NoSQL, thiết bị cơ sở dữ liệu và công nghệ lưu trữ dữ liệu kiểu mới của kiến ​​trúc MPP. Ngoài ra, bài viết này đưa ra một số khuyến nghị áp dụng cho các môi trường khác nhau nhằm nắm bắt các trạng thái phát triển của công nghệ lưu trữ dữ liệu từ các góc độ khác nhau. Bài báo này tóm tắt về phân đoạn tệp, các kịch bản phù hợp và ưu nhược điểm của hệ thống tệp phân tán, đồng thời chủ yếu phân tích và tóm tắt các lý thuyết và kịch bản phù hợp của bốn mô hình lưu trữ dữ liệucủa cơ sở dữ liệu NoSql. Hơn nữa, bài viết này điều tra và kết luận những phát triển và tính năng của thiết bị cơ sở dữ liệu một cách chi tiết . Đồng thời, phác thảo kiến ​​trúc MPP (Xử lý song song hàng loạt), một công nghệ lưu trữ dữ liệu mới. Cuối cùng, các xu hướng nghiên cứu về công nghệ lưu trữ được đưa ra, cung cấp các tài liệu tham khảo cho nghiên cứu về công nghệ lưu trữ dữ liệu lớn.

Từ khoá: Big Data Storage, No Sql, Phát hành tập tin, Database All-in-one Machine, MPP Kiến trúc

Giới thiệu

Trong vài thập kỷ qua, với việc mở rộng quy mô ứng dụng, các dịch vụ web đã phát triển từ dạng đơn lẻ sang dạng đa phương tiện, dẫn đến cấu trúc và biểu mẫu dữ liệu đa dạng, đồng thời tăng trưởng dữ liệu theo cấp số nhân. International Data Corporation (IDC) dự đoán rằng trong tương lai, kích thước dữ liệu sẽ tăng gấp đôi sau mỗi hai năm [1]. Người tiên phong trong nghiên cứu dữ liệu lớn—McKinsey & Company, một công ty tư vấn ở Hoa Kỳ, định nghĩa dữ liệu lớn là: tập hợp dữ liệu có quy mô vượt quá khả năng thu nhận, lưu trữ, quản lý và phân tích của các công cụ cơ sở dữ liệu thông thường [2]. Hệ thống lưu trữ dữ liệu truyền thống đã đạt đến điểm tắc nghẽn và không thể hoàn thành xử lý dữ liệu kịp thời. Dữ liệu lớn có như vậycác tính năng như dung lượng cao, nhiều kiểu dữ liệu, mật độ giá trị thấp, tốc độ xử lý cao, quan hệ động phức tạp giữa các dữ liệu. Và yêu cầu về tính sẵn sàng, khả năng mở rộng và độ tin cậy cao [3], đặt ra những thách thức đối với công nghệ lưu trữ dữ liệu truyền thống.

Bài báo này nghiên cứu và tổng kết các loại công nghệ lưu trữ dữ liệu mới, phục vụ cho các bài toán quy mô ứng dụng Dự án Quỹ: Quỹ Khoa học Tự nhiên Quốc gia Trung Quốc (61462070); Dự án quy hoạch khoa học và công nghệ vùng (20130364). mở rộng, tăng trưởng dữ liệu nhanh chóng, nhiều kiểu dữ liệu. Đầu tiên, điều tra và phân tích tình trạng nghiên cứu của hệ thống tệp phân tán. Bên cạnh đó, phân tích và so sánh các đặc điểm của các hệ thống tệp phân tán chính, chẳng hạn như GFS [4], HDFS [5], GlusterFS [6], GridFS [7], TFS [8], Luster [9], FastDFS [10]. Ngoài ra, bằng cách nghiên cứu bốn mô hình dữ liệu của cơ sở dữ liệu NoSql, tương phản công nghệ lưu trữ của mô hình giá trị khóa, mô hình kiểu cột, mô hình tài liệu và mô hình đồ họa. Và nghiên cứu khả năng ứng dụng và tính năng kiến ​​trúc của máy tất cả trong một dữ liệu lớn hiện tại. Cuối cùng, giới thiệu định nghĩa và kịch bản của các loại cụm cơ sở dữ liệu mới sử dụng khung MPP (Xử lý song song lớn) [11] và dự báo các xu hướng nghiên cứu có thể xảy ra của công nghệ lưu trữ trong tương lai.

Công nghệ lưu trữ dữ liệu lớn:

Bài báo này nghiên cứu và phân tích công nghệ lưu trữ dữ liệu lớn từ bốn khía cạnh sau: hệ thống tệp phân tán, cơ sở dữ liệu NoSQL, công nghệ lưu trữ dữ liệu kiểu mới của kiến ​​trúc MPP và cơ sở dữ liệu máy tất cả trong một. Ngoài ra, bài báo này còn đưa ra một số khuyến nghị áp dụng cho các môi trường khác nhau nhằm nắm bắt các trạng thái phát triển của công nghệ lưu trữ dữ liệu từ các góc độ khác nhau.

2.1. Hệ thống tệp phân tán:

Hệ thống tệp là cơ sở của chương trình ứng dụng. Tuy nhiên, với sự phát triển của ứng dụng mạng, dữ liệu tăng lên nhanh chóng. Vì vậy, công nghệ lưu trữ dữ liệu lớn đã trở thành nhiệm vụ chính của các doanh nghiệp và tổ chức nghiên cứu. Do khả năng lưu trữ bị hạn chế, các hệ thống lưu trữ truyền thống khó có thể giải quyết vấn đề lưu trữ dữ liệu lớn. Vì vậy, chúng tôi sử dụng hệ thống tệp phân tán để chuyển tải hệ thống tới nhiều nút. Hệ thống tệp phân tán cung cấp dung lượng lưu trữ polyme và băng thông I/O, do đó có thể dễ dàng mở rộng quy mô của hệ thống [12]. Nói chung, việc xác định xem một hệ thống tệp phân tán có thành công hay không phụ thuộc vào ba yếu tố sau: dữ liệuchế độ lưu trữ, tốc độ đọc, cơ chế bảo mật. Tuy nhiên, vẫn có những cải tiến được thực hiện trong hệ thống tệp phân tán. Ví dụ: GFS và HDFS được thiết kế cho các tệp lớn không thể đáp ứng yêu cầu lưu trữ của nhiều tệp nhỏ. Tần suất truy cập của tệp nhỏ cao, dẫn đến tần suất truy cập vào đĩa cứng cao. Vì vậy, hiệu suất của I/O đã bị giảm xuống [13]. Các tệp nhỏ cũng có thể dẫn đến việc tạo ra một số lượng lớn siêu dữ liệu, điều này sẽ ảnh hưởng đến khả năng khôi phục và quản lý máy chủ siêu dữ liệu, sau đó dẫn đến giảm hiệu suất tổng thể. Hơn nữa, vì tệp tương đối nhỏ nên dễ tạo ra sự phân mảnh tệp, gây lãng phí dung lượng ổ đĩa. Tạo liên kết cho từngtập tin sẽ gây ra sự chậm trễ mạng [14]. Bảng 1 tóm tắt và so sánh rộng rãi một số hệ thống tệp phân tán phổ biến, chẳng hạn như GFS, HDFS, TFS, Lustre, v.v.

Bảng 1. Danh sách phân tích các hệ thống tệp phân tán phổ biến [15-20]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tên | Phân đoạn tệp | Sao lưu hệ thống | Ưu điểm | Nhược điểm | Ứng dụng |
| GFS | Các tệp được lưu trữ trong GFS được chia thành các khối có kích thước cố định | Mỗi khối được sao chép sang nhiều máy chủ chunk , lưu 3 bản sao theo mặc định | GFS sẽ không coi lỗi phần cứng là bất thường. Thông thường, một bản cập nhật được xử lý bằng cách thêm dữ liệu mới thay vì thay đổi dữ liệu hiện có. | Không áp dụng cho lưu trữ tệp nhỏ. Các tệp nhỏ bổ sung sẽ làm giảm hiệu suất. | Tập dữ liệu lớn phân tán lớn . Kích thước dữ liệu thường nằm trong khoảng 4G~40G. |
| HDFS | Các tệp lớn được chia thành một số khối có kích thước mặc định là 64MB. Mỗi khối sẽ lưu trữ một số bản sao trên nhiều nút dữ liệu. | Hỗ trợ sao chép dữ liệu. Lưu trữ nhiều bản sao qua các nút khác nhau. | Khả năng mở rộng rất mạnh. Một phiên bản HDFS duy nhất có thể hỗ trợ hàng chục triệu tài liệu. Và có khả năng thời gian thực cao. | Không thể được sử dụng cho các tình huống yêu cầu độ trễ thấp trong truy cập dữ liệu. Không thể lưu trữ các tệp nhỏ có quy mô lớn. | Tập dữ liệu rất lớn có kích thước ở mức GB đến TB . |
| GlusterFS | Không hỗ trợ phân đoạn tập tin | Hỗ trợ sao chép dữ liệu và cung cấp không gian tên toàn cầu. Nhiều bản sao của nhiều tệp có thể được lưu trữ trong các máy chủ khác nhau. Trong khi đọc một bản sao, hệ thống sẽ chọn bản sao gần nhất một cách thuận lợi. | Hỗ trợ CIFS, NFS và máy gốc sử dụng ứng dụng khách GlusterFS. Hệ thống nhiều tệp có thể được triển khai trên hệ thống tệp phân tán ảo. Sử dụng bảng điều khiển dành cho quản trị viên, quản lý cập nhật trung tâm một cách dễ dàng. Tất cả các nút có thể được sử dụng để lấy dữ liệu | Chỉ có thể quản lý hệ thống trên một máy chủ, không có dự phòng. Không thể thêm một nút mới nửa chừng. Không biết cách thêm nhiều đĩa vào mỗi nút; Không có chính sách bảo mật trong GlusterFS | Hỗ trợ lưu trữ tệp lớn ở cấp độ PB. |
| TFS | Không hỗ trợ phân đoạn tệp. Một số lượng lớn các tệp nhỏ sẽ được hợp nhất thành một tệp lớn. | TFS lưu trữ các tệp dữ liệu theo khối và lưu trữ nhiều bản sao trong trường hợp bảo mật dữ liệu | Hoạt động đơn giản, với khả năng mở rộng và cân bằng tải mượt mà . Hỗ trợ chia tỷ lệ tuyến tính. Có thể dễ dàng mở rộng đến mức PB. | Khi đồng thời cao và kích thước tệp lớn hơn 5 MB, các lỗi nghiêm trọng sẽ phát sinh trong TFS. Trong những trường hợp hiếm hoi, hỗ trợ lưu trữ tệp lớn. Không hỗ trợ danh mục và sự cho phép của người dùng. | Việc lưu trữ và xử lý một lượng lớn dữ liệu phi cấu trúc và hình ảnh khổng lồ trên trang web taobao . |
| GridFS | Hỗ trợ chia một file lớn thành nhiều file tài liệu nhỏ. | GridFS lưu trữ dữ liệu tệp và siêu dữ liệu tệp trong MongoDB. Sao chép tệp để đối phó với chuyển đổi dự phòng và tích hợp dữ liệu. Và cũng có thể được sử dụng để đọc tiện ích mở rộng, sao lưu nóng hoặc được sử dụng làm nguồn dữ liệu xử lý hàng loạt ngoại tuyến | Nó dựa trên mô hình cấu trúc của bộ lưu trữ đối tượng, giảm thêm chi phí truy cập trong thời gian chạy. GridFS được thiết kế để điều chỉnh cơ chế truy cập, nhằm thích ứng với hiệu suất ứng dụng nhanh hơn do chế độ I/O đưa ra. Có thể cung cấp hiệu suất I/O nhanh nhất mà cụm ứng dụng cần. GridFS đảm bảo cân bằng tải trong từng thiết bị lưu trữ, thay vì lấp đầy một nút. | Tốc độ đọc tệp từ GridFS chậm hơn so với đọc trực tiếp từ hệ thống tệp. Nếu tệp lớn và được lưu trữ dưới dạng nhiều tệp, không thể khóa tất cả các khối tệp khi sửa đổi tệp lớn này. Khi thay đổi tài liệu được lưu trữ trong GridFS, nó chỉ có thể xóa tài liệu cũ trước rồi lưu lại tài liệu | Thích hợp cho các tệp lớn hiếm khi cần thay đổi |
| Lustre | Chia dữ liệu thành một số đối tượng cố định. Mỗi đối tượng chứa một số khối dữ liệu. Khi một khối dữ liệu được ghi vào đối tượng vượt quá dung lượng của nó, lần ghi tiếp theo sẽ được lưu trữ trong đối tượng tiếp theo. Lustre có thể phân phối tệp tối đa thành 160 đối tượng để lưu trữ. | Lustre cung cấp hai công cụ sao lưu. Một cái được sử dụng để quét hệ thống tệp và một cái khác được sử dụng để sao lưu gói và khôi phục áp suất . | Có thể cung cấp khả năng chia sẻ dữ liệu và xử lý song song. Khả năng mở rộng là rất mạnh mẽ. Có thể cung cấp công nghệ chuyển đổi dự phòng cho siêu dữ liệu và dữ liệu mục tiêu dưới sự quản lý của Lustre, đạt được quyền truy cập với độ tin cậy cao. Cơ chế quản lý phân tán có thể đạt được kiểm soát đồng thời. Cung cấp quyền truy cập trong nhiều giao thức mạng | Rất khó để thực hiện phản chiếu dữ liệu. Chuyển đổi dự phòng giữa các nút dựa trên công nghệ nhịp tim của bên thứ ba . Chỉ có hai nút quản lý siêu dữ liệu. Nếu quy mô hệ thống đã đạt đến quy mô nhất định, nút quản lý sẽ đạt đến mức quá tải. Lustre kernel chỉ có thể được triển khai trên Linux với một số hạn chế. | Hỗ trợ lưu trữ tệp lớn ở cấp độ PB và phù hợp với cụm máy tính lớn hoặc siêu máy tính |
| Ceph | Áp dụng mẫu RAIDO trên nhiều đĩa cứng. Phân tán dữ liệu liên tục trên nhiều ổ đĩa để truy cập, nhằm thích nghi với việc cân bằng tải. | Hỗ trợ sao chép dữ liệu. Có nhiều máy chủ siêu dữ liệu. | Lưu trữ dữ liệu và siêu dữ liệu riêng biệt. Quản lý siêu dữ liệu bằng cách sử dụng phân phối động. Có bộ lưu trữ đối tượng phân tán tự động đáng tin cậy. | Công nghệ chưa trưởng thành, có thể chưa áp dụng được vào môi trường sản xuất. | Hỗ trợ lưu trữ tệp lớn ở cấp độ PB. |
| FastDFS | FastDFS không lưu trữ tệp theo khối. Các tệp do máy khách tải lên tương ứng với các tệp được lưu trữ trên máy chủ. | FastDFS áp dụng chế độ lưu trữ theo nhóm. Các máy chủ lưu trữ trong cùng một nhóm sao lưu lẫn nhau. | Máy chủ FastDFS chỉ có hai ký tự, nút theo dõi và lưu trữ . Vì vậy, nó có tính năng nhẹ. FastDFS áp dụng chế độ lưu trữ theo nhóm, linh hoạt và được kiểm soát chặt chẽ. Trong FastDFS, mỗi nút là nút chính, với cấu trúc ngang hàng. Có thể thay đổi số lượng trình theo dõi bất cứ lúc nào, theo áp lực của máy chủ. | FastDFS không lưu trữ tệp theo khối. Vì vậy, nó không phù hợp với kịch bản tính toán phân tán . Dung lượng lưu trữ bị giới hạn bởi một máy chủ lưu trữ duy nhất. | Phù hợp với dịch vụ có lưu lượng truy cập cao , với tệp là nhà cung cấp dịch vụ, chẳng hạn như trang web album ảnh , trang web video , v.v. Và số lượng lớn tệp nhỏ, 4K ~ 500M. |

2.2. Cơ sở dữ liệu NOSQL:

Với sự tăng trưởng nhanh chóng về kích thước dữ liệu của người dùng doanh nghiệp và nâng cao nhu cầu của người dùng về mức độ dịch vụ, cơ sở dữ liệu quan hệ truyền thống có một số hạn chế. Cơ sở dữ liệu truyền thống sử dụng tệp phẳng dựa trên bản ghi có cấu trúc để lưu trữ tất cả dữ liệu ứng dụng, dẫn đến sự không khớp giữa ứng dụng và cơ sở dữ liệu. Điều này xảy ra khi ứng dụng được mã hóa bằng ngôn ngữ khai báo. Cấu trúc của nó hoàn toàn khác với các cơ sở dữ liệu này [21]. Trước đây, để cải thiện hiệu suất của hệ thống, các thành phần và tài nguyên được mở rộng theo chiều dọc. Tuy nhiên, do bộ lưu trữ và ứng dụng không còn tách biệt nên mỗi lần mở rộng tài nguyên sẽ làm gián đoạn dịch vụ và đặt lại ứng dụng.

Hầu hết dữ liệu chúng tôi tạo ra là dữ liệu không đồng nhất. Sự tồn tại của một số lượng lớn dữ liệu có cấu trúc và phi cấu trúc gây khó khăn cho việc xác định trước mô hình dữ liệu quan hệ thống nhất và hoàn hảo, đồng thời khả năng mở rộng theo chiều ngang của cơ sở dữ liệu quan hệ là kém [22]. Hầu hết các cơ sở dữ liệu quan hệ không hỗ trợ lưu trữ phân tán quy mô lớn. Đồng thời, khó có thể đáp ứng yêu cầu thời gian thực về tính đồng thời cao và lượng dữ liệu lớn. Vì vậy, công nghệ lưu trữ cơ bản không chỉ linh hoạt để cho phép dữ liệu được lưu trữ ở dạng tự nhiên mà còn đáp ứng nhu cầu của biên giới.

So với cơ sở dữ liệu quan hệ, hệ thống lưu trữ cơ sở dữ liệu NoSQL hỗ trợ lưu trữ và quản lý động dữ liệu khối. Nó tránh được sự phức tạp không cần thiết, có thông lượng cao và có thể xử lý tốt tỷ lệ ngang. Và khả năng chịu lỗi cao có thể lưu trữ dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc để tránh ánh xạ quan hệ đối tượng. Ý tưởng thiết kế của cơ sở dữ liệu NoSQL là trích xuất cơ chế lập chỉ mục của cơ sở dữ liệu quan hệ, kết hợp chiến lược lưu trữ phân tán và xóa những thứ không cần thiết đối với một số vấn đề trong hệ thống SQL. Do đó, nó đạt được hiệu quả tương đối tốt, khả năng mở rộng và tính linh hoạt [23].

Cơ sở dữ liệu Nosql chủ yếu được chia thành: lưu trữ khóa-giá trị, lưu trữ dựa trên cột, lưu trữ tài liệu, lưu trữ đồ họa.

Cơ sở dữ liệu khóa-giá trị được thiết kế để hỗ trợ các hoạt động truy vấn đơn giản, để lại các hoạt động phức tạp cho lớp ứng dụng. Tập dữ liệu sẽ ánh xạ khóa tới một hoặc một tập hợp các giá trị. Tức là key là từ khóa duy nhất để tìm từng địa chỉ dữ liệu, cũng có nghĩa là không thể thiếu. Giá trị là nội dung mà dữ liệu thực sự lưu trữ. Lưu trữ khóa-giá trị cung cấp một bảng băm với các cặp khóa-giá trị trên các máy chủ từ xa của một cụm phân tán, để triển khai ánh xạ từ khóa đến giá trị. Giá trị băm dựa trên khóa định vị trực tiếp địa chỉ của dữ liệu, đạt được truy vấn đồng thời cao nhanh chóng và cũng hỗ trợ hoạt động của dữ liệu hàng loạt. Lưu trữ khóa-giá trị được chia thành loại khóa-giá trị, loại tài liệu khóa và loại cột khóa[24]. Loại khóa-cột là sự mở rộng điển hình của các cặp khóa-giá trị của loại khóa-giá trị. Do tính đơn giản và khả năng mở rộng linh hoạt, nó cũng là xu hướng chính của mô hình dữ liệu.

Lưu trữ dựa trên cột tổng quát thay thế các cột bằng họ cột. Ý tưởng của cơ sở dữ liệu quan hệ là lưu trữ tất cả các bảng bằng một dòng trên đĩa. Nghĩa là, một danh sách các mục được liên kết với cùng một id hàng cụ thể sẽ được lưu trữ cùng nhau [25]. Vì các ngân hàng hoặc tổ chức tài chính cần duy trì một số lượng lớn các bản ghi liên quan, không đảm bảo rằng tất cả các giá trị luôn được lưu trữ một cách liên tục. Trong cơ sở dữ liệu của dữ liệu cột, toàn bộ cột của bảng được lưu trữ cùng nhau, được ánh xạ tới một khóa. Vì tất cả các mục được liệt kê đều có chỉ mục nên chúng tôi chỉ có thể tìm kiếm một phần của bảng. Một cột cũng có thể có các cột lồng nhau theo cấu trúc phân cấp và một trong số đó là siêu cột [26]. Điều này cung cấp truy vấn đơn giản vàtruy cập nhanh, đồng thời tránh được chi phí không cần thiết khi tìm khóa duy nhất của bản ghi.

Cơ sở dữ liệu tài liệu là một loại cơ sở dữ liệu quan hệ khác, được sử dụng để lưu trữ dữ liệu bán cấu trúc. Có XML ( ngôn ngữ đánh dấu mở rộng), JSON (ký hiệu đối tượng JavaScript) hoặc các định dạng tương tự khác [27]. Một tài liệu có thể được xem như một dòng trong cơ sở dữ liệu quan hệ, chứa tất cả các thông tin liên quan của tài liệu. Một nhóm bao gồm nhiều tài liệu và mỗi tài liệu có thể có các mẫu khác nhau cũng như số lượng và loại lưu trữ dữ liệu khác nhau [28]. Lưu trữ tin nhắn văn bản là tối ưu hóa đặc biệt. Do các tập dữ liệu liên quan được lưu trữ nhiều, chi phí cho thao tác THAM GIA SQL được tiết kiệm. Mặc dù cơ sở dữ liệu là thiết kế không lược đồ, nhưng nó lưu trữ các bản ghi bán cấu trúc và có cấu trúc phân cấp.

Cơ sở dữ liệu đồ thị phù hợp nhất với tìm kiếm ứng dụng và truyền tải, chẳng hạn như tìm các liên kết có liên quan trên LinkedIn, tìm bạn bè trên Facebook [29], v.v. Nó chú ý nhiều hơn đến mối quan hệ giữa các mục dữ liệu hơn là bản thân dữ liệu. Họ tối ưu hóa tốc độ truyền tải nhanh và sử dụng thuật toán đồ thị một cách hiệu quả. Ví dụ: đường đi ngắn nhất trước để tìm sự liên quan giữa các thông tin, v.v.

Bảng 2 phân tích và kết luận các loại lưu trữ chính của cơ sở dữ liệu NoSQL:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loại cơ sở dữ liệu | Ưu điểm | Nhược điểm | Mô hình dữ liệu | Kịch bản ứng dụng | Các trường hợp |
| Lưu trữ khóa-giá trị | Có hiệu suất đọc và ghi đồng thời rất cao. Dữ liệu được lập chỉ mục và phân đoạn theo giá trị chính. Tìm kiếm nhanh chóng và mô hình dữ liệu đơn giản. | Dữ liệu không có cấu trúc và không hỗ trợ hoạt động dữ liệu logic phức tạp . | ánh xạ khóa-giá trị giữa khóa và giá trị | Bộ đệm nội dung. Chủ yếu được sử dụng cho hệ thống nhật ký. | Dynamo, Redis, Voldemort |
| lưu trữ dựa trên cột | Tìm kiếm nhanh, khả năng mở rộng tốt và tiết kiệm nhiều thao tác I/O. Nó dễ dàng hơn cho phần mở rộng phân tán. | Chức năng tương đối hạn chế. | Lưu trữ theo cột , trong đó dữ liệu trong cùng một cột được lưu trữ trên cùng một trang | Hệ thống tập tin phân tán | Bigtable, Cassandra, HBase HyperTable |
| lưu trữ tài liệu | Không cần xác định trước cấu trúc dữ liệu . Sử dụng tài liệu có định dạng cụ thể thay vì bộ làm đơn vị lưu trữ dữ liệu. | Hiệu quả truy vấn chưa cao, thiếu cú ​​pháp truy vấn thống nhất. | Giá trị trỏ đến dữ liệu có cấu trúc. | ứng dụng wed | CouchDB, MongoDB XML Database ThruDB |
| Đồ họa lưu trữ | Sử dụng lý thuyết đồ thị và thuật toán liên quan để cải thiện hiệu suất lưu trữ, dữ liệu quản lý và vận hành. | Chức năng tương đối hạn chế. | Cấu trúc đồ thị | Mạng xã hội, biểu đồ mối quan hệ | Neo4j, GraphDB InfoGrdi |

2.3. Cơ sở dữ liệu Máy tất cả trong một

Trong những năm gần đây, đối mặt với việc xử lý và lưu trữ dữ liệu hàng loạt, nhiều nhà sản xuất phần cứng truyền thống đề xuất giải pháp tích hợp---máy tất cả trong một cơ sở dữ liệu, giải pháp này đã trở thành một điểm nóng. Bằng dạng sản phẩm của máy tất cả trong một, nó đơn giản hóa sự phức tạp của việc triển khai và quản lý cơ sở hạ tầng của trung tâm dữ liệu, giải quyết vấn đề liên tục mở rộng tài nguyên phần cứng cơ bản ở thời đại dữ liệu lớn, yêu cầu của tất cả trong một một máy và chi phí lưu trữ dữ liệu lớn. Các nhà sản xuất quốc tế như IBM, Oracle, EMC tung ra các sản phẩm và giải pháp tích hợp cho dữ liệu lớn [35]. Theo sau họ, các nhà sản xuất Trung Quốc cũng phát triểncơ sở dữ liệu riêng tất cả trong một máy. Ví dụ, máy cơ sở dữ liệu tất cả trong một của Huawei tận dụng lợi thế kiến ​​trúc phần cứng của nó về điện toán, lưu trữ và hội tụ mạng, cũng như tính năng của thông lượng cao và IOPS cao, tích hợp các đặc điểm tuyệt vời của thẻ mạng thông minh , SSD và phần cứng khác, giải quyết tắc nghẽn hiệu suất giữa tính toán và lưu trữ. Bộ xử lý dữ liệu lớn XData của Shuguang tách biệt đơn vị lưu trữ dữ liệu và đơn vị xử lý. Bằng cách xây dựng phần mềm trung gian dịch vụ hiệu quả, trùng hợp nút lưu trữ dữ liệu cơ bản áp dụng cấu trúc không chia sẻ thành một hình ảnh hệ thống xử lý dữ liệu duy nhất. Vỏ máy tất cả trong một đám mây dữ liệu lớn Langchaocác phiên kỹ thuật như lưu trữ dữ liệu, xử lý dữ liệu, trình bày dữ liệu, v.v. Ngoài ra còn có máy đa năng điện toán đám mây khối dữ liệu Yunchuang Storage, máy đa năng dữ liệu lớn Zhongzhiheda, dữ liệu lớn dựa trên Zhiyitu Hadoop- máy trong một [36].

Máy tất cả trong một cơ sở dữ liệu thường phù hợp với mô hình dữ liệu có quan hệ lưu trữ phức tạp. Đồng thời, điện toán cần tính giao dịch cao và nhất quán. Nói chung, cấu hình máy chủ công cụ cơ sở dữ liệu phụ thuộc vào nhu cầu đồng thời và cấu hình máy chủ nút lưu trữ cơ sở dữ liệu phụ thuộc vào yêu cầu kích thước dữ liệu [37]. Máy tất cả trong một cơ sở dữ liệu áp dụng kiến ​​trúc xử lý dữ liệu lớn phân tán đầy đủ, tích hợp phần cứng và phần mềm trong một hệ thống. Với sự tăng trưởng của dữ liệu người dùng và mở rộng kinh doanh, nó có thể được cải thiện bằng cách mở rộng chiều dài phần cứng và cũng có thể đạt được tuyến tínhmở rộng quy mô bằng cách thêm các nút theo chiều rộng, đảm bảo hiệu suất của độ trễ thấp, thông lượng cao và tính liên tục của doanh nghiệp [38]. Máy tất cả trong một là sự kết hợp giữa phần mềm và phần cứng, được thiết kế hoàn toàn để xử lý lưu trữ dữ liệu hàng loạt. Và nó được tạo thành từ một tập hợp các máy chủ tích hợp, thiết bị lưu trữ, hệ điều hành, hệ quản trị cơ sở dữ liệu và phần mềm được cài đặt sẵn để quản lý dữ liệu. Nó cung cấp giải pháp lưu trữ dữ liệu lớn, chủ yếu cho thị trường kho dữ liệu lớn. Và khả năng thông lượng cao của nó tạo điều kiện giải quyết vấn đề nút cổ chai I / O. Người dùng có thể chọn các loạt sản phẩm khác nhau theo yêu cầu, tùy chỉnh theo yêu cầu. Tuy nhiên, cơ sở dữ liệu tất cả trong một máy cũng phải đối mặt vớinhững thách thức. Trong thời đại dữ liệu lớn, lượng dữ liệu ngày càng tăng lên một cách đáng kinh ngạc. Do đó, nếu người dùng cần mở rộng máy tất cả trong một, họ chỉ có thể thêm một tủ thiết bị, dẫn đến việc mở rộng không linh hoạt. Và bởi vì phần mềm tất cả trong một có tính tích hợp cao, nên nó khó có thể được triển khai trong các môi trường khác. Trong một số ngành, nhu cầu thay đổi nhanh chóng hơn. Vì vậy, mô hình kinh doanh sẽ thay đổi rất nhanh với nó. Việc sử dụng máy tất cả trong một sẽ hạn chế các thao tác của doanh nghiệp ngược lại. Nhưng trong một số ứng dụng tương đối thuần thục và ổn định, máy tất cả trong một thể hiện giá trị của việc đơn giản hóa CNTT.

2.4. Cụm cơ sở dữ liệu mới của kiến ​​trúc MPP

MPP là hệ thống xử lý song song quy mô lớn, là một loại phương pháp để mở rộng tài nguyên hệ thống, chủ yếu là xử lý song song. Điều này có nghĩa là một máy tính duy nhất có nhiều bộ xử lý mạng [39]. Mở rộng theo chiều ngang là mục tiêu thiết kế chính của cơ sở dữ liệu kiến ​​trúc MPP. Nó được liên kết bởi nhiều máy chủ SMP thông qua internet của các nút cố định, cộng tác cho các nhiệm vụ chung. Từ cấp độ người dùng, nó là một hệ thống máy chủ, hỗ trợ mô hình quan hệ dữ liệu chặt chẽ. Đặc điểm lớn nhất là mỗi nút chỉ có thể truy cập tài nguyên cục bộ của riêng họ, không có chia sẻ. Cụm cơ sở dữ liệu sử dụng kiến ​​trúc MPP có thể hỗ trợ hiệu quả việc lưu trữ dữ liệu có cấu trúc khối lượng lớn ở mức PB. Nó là nền tảngtrên kiến ​​trúc Không có gì được chia sẻ. Bằng công nghệ xử lý dữ liệu lớn của lưu trữ cột, chỉ mục thô, v.v. và kết hợp mô hình tính toán phân tán với hiệu suất cao, nó hoàn thành hỗ trợ kỹ thuật cho ứng dụng lưu trữ của lớp phân tích. Môi trường hoạt động chủ yếu là PC giá rẻ và nó có ưu điểm là hiệu suất cao và khả năng mở rộng cao [40]. Nó có thể cải thiện hiệu suất xử lý dữ liệu, cải thiện tải quá trình dữ liệu, nâng cao hiệu quả xử lý dữ liệu hàng loạt và giảm chi phí tổng thể của việc xử lý từng TB. Do đó nó đã được sử dụng rộng rãi trong kho dữ liệu doanh nghiệp thuộc lĩnh vực phân tích dữ liệu có cấu trúc và thế hệ mới.

3.Triển vọng của công nghệ lưu trữ Bigdata

Do số lượng lớn dữ liệu không cấu trúc và bán cấu trúc, cơ sở dữ liệu quan hệ truyền thống đã bất lực. Tuy nhiên, công nghệ lưu trữ mới, chẳng hạn như cơ sở dữ liệu NoSQL và hệ thống tệp phân tán, vượt trội hơn so với lưu trữ truyền thống, bất kể khả năng chịu lỗi, khả năng mở rộng và tính di động của dữ liệu. Và nó thích hợp để lưu trữ dữ liệu liên tục và quản lý lưu trữ dữ liệu hàng loạt [41]. Nhưng đối với hiệu suất xử lý dữ liệu theo thời gian thực, có một khoảng cách nhất định giữa công nghệ lưu trữ mới và cơ sở dữ liệu quan hệ. Vì vậy, mỗi cái đều có mặt tốt của nó. Hiện tại, sự kết hợp của cơ sở dữ liệu quan hệ và hệ thống xử lý song song phân tán có thể nâng cao hiệu quả lưu trữ, xử lý tốc độ và tốc độ phân tích [42]. Cách làm này cũng là xu hướng hot trong thời gian tới. Vấn đề cốt lõi của công nghệ lưu trữ dữ liệu lớn là hiệu suất. Một kỹ thuật và nền tảng duy nhất không còn có thể đáp ứng nhu cầu tăng trưởng bùng nổ dữ liệu và yêu cầu phân tích và lưu trữ dữ liệu từ các nhà khai thác. Trong quá trình phát triển tiếp theo, cơ sở dữ liệu kiểu mới sẽ dần được trộn lẫn với hệ sinh thái Hadoop hoặc hệ sinh thái Spark [43], cung cấp hỗ trợ SQL và giao dịch cho ứng dụng. Sử dụng Hadoop hoặc Spark để xử lý dữ liệu bán cấu trúc, phi cấu trúc. Vì vậy trong tương lai, lưu trữ cũng sẽ được phát triển theo hướng kết hợp giữa cụm cơ sở dữ liệu song song MPP và cụm Hadoop / Spark. Ngoài ra, với sự phát triển bùng nổ của dữ liệu doanh nghiệp, máy tất cả trong một dữ liệu lớn chắc chắn sẽ trở thành một công nghệ hot và được sử dụng rộng rãi. Bằng cách nghiên cứu công nghệ lưu trữ dữ liệu mới, bài báo này tóm tắt và đối chiếu hệ thống tệp phân tán, cơ sở dữ liệu NoSQL, máy cơ sở dữ liệu tất cả trong một và cụm cơ sở dữ liệu kiểu mới của kiến ​​trúc MPP từ các góc độ khác nhau. Và xu hướng nghiên cứu trong tương lai đã được đưa ra. Lưu trữ dữ liệu lớn vẫn đang trong giai đoạn phát triển nhanh chóng. Không gian phát triển còn rất lớn, cần các nhà nghiên cứu không ngừng khám phá.