TCVN TIÊU CHUẨN QUỐC GIA

TCVN 13239-3:2023
ISO/IEC 20547-3:2020

Xuất bản lần 1

CÔNG NGHỆ THÔNG TIN – KIẾN TRÚC THAM CHIẾU DỮ LIỆU LỚN – PHẦN 3: KIẾN TRÚC THAM CHIẾU

Information technology – Big data reference architecture – Part 3: Reference architecture

HÀ NỘI – 2023

Mục lục

Lời nói đầu

TCVN 13239-3:2023 hoàn toàn tương đương với ISO/IEC 20547-3:2020

TCVN 13239-3:2023 do Viện Công nghiệp phần mềm và Nội dung số Việt Nam biên soạn, Bộ Thông tin và Truyền thông đề nghị, Tổng cục Tiêu chuẩn Đo lường Chất lượng thẩm định, Bộ Khoa học và Công nghệ công bố.

TCVN 13239 (ISO/IEC 20547) về Công nghệ thông tin – Kiến trúc tham chiếu dữ liệu lớn gồm:

- TCVN 13239-1:2023 (ISO/IEC 20547-1:2020), Phần 1: Khung và quy trình ứng dụng;
- TCVN 13239-2:2020 (ISO/IEC 20547-2:2018), Phần 2: Các trường hợp sử dụng và yêu cầu dẫn suất;
- TCVN 13239-3:2023 (ISO/IEC 20547-3:2020), Phần 3: Kiến trúc tham chiếu;
- TCVN 13239-4:2023 (ISO/IEC 20547-4:2020), Phần 4: Bảo mật và quyền riêng tư.
- TCVN 13239-5:2020 (ISO/IEC 20547-5:2018), Phần 5: Lộ trình tiêu chuẩn.

Công nghệ thông tin – Kiến trúc tham chiếu dữ liệu lớn - Phần 3: Kiến trúc tham chiếu

Information technology – Big data reference architecture – Part 3: Reference architecture

1 Phạm vi áp dụng

Tiêu chuẩn này quy định kiến trúc tham chiếu dữ liệu lớn. Kiến trúc tham chiếu bao gồm các khái niệm và góc nhìn kiến trúc.

Kiến trúc tham chiếu trong tiêu chuẩn này xác định hai góc nhìn kiến trúc:

- Góc nhìn người dùng: định nghĩa các vai trò/vai trò phụ, mối quan hệ và các hoạt động của chúng trong hệ sinh thái dữ liệu lớn;
- Góc nhìn chức năng: định nghĩa các lớp kiến trúc và các lớp của các thành phần chức năng bên trong, thực thi các hoạt động của các vai trò/vai trò phụ trong góc nhìn người dùng.

Mục đích của kiến trúc tham chiếu dữ liệu lớn:

- Cung cấp một ngôn ngữ chung cho các bên liên quan;
- Khuyến khích việc tuân thủ các tiêu chuẩn, thông số kỹ thuật và các khuôn mẫu chung;
- Cung cấp một phương thức công nghệ nhất quán để giải quyết các vấn đề tương tự;
- Tạo điều kiện cho việc tìm hiểu sự phức tạp trong vận hành dữ liệu lớn;
- Mô tả và hiểu rõ các thành phần, quy trình và các hệ thống dữ liệu lớn khác nhau, trong bối cảnh một mô hình khái niệm dữ liệu lớn tổng thể;
- Cung cấp một tài liệu kỹ thuật tham khảo cho các chính phủ, các cơ quan, đơn vị và những người dùng khác để nắm bắt, thảo luận, phân loại và so sánh các giải pháp dữ liệu lớn;
- Tạo điều kiện cho việc phân tích khả năng tương thích, tính linh hoạt, khả năng tái sử dụng và mở rộng của các tiêu chuẩn đề xuất.

2 Tài liệu viện dẫn

Các tài liệu viện dẫn sau rất cần thiết cho việc áp dụng tiêu chuẩn này. Đối với các tài liệu viện dẫn ghi năm công bố thì áp dụng phiên bản được nêu. Đối với các tài liệu viện dẫn không ghi năm công bố thì áp dụng phiên bản mới nhất, bao gồm cả các sửa đổi, bổ sung (nếu có).

TCVN 10249-2:2013 (ISO 8000-2), Chất lượng dữ liệu – Phần 2: Từ vựng;

TCVN 13238:2020 (ISO/IEC 20546:2019), Công nghệ thông tin — Dữ liệu lớn — Tổng quan và từ vựng;

TCVN 12481:2019 (ISO/IEC 17789), Công nghệ thông tin – Tính toán đám mây – Kiến trúc tham chiếu;

ISO/TS 8000-60, Data quality – Part 60: Data quality management: Overview (Phần 60: Quản lý chất lượng dữ liệu);

ISO 8000-61, Data quality – Part 61: Data quality management: process reference model (Phần 61: Quản lý chất lượng dữ liệu: Mô hình tham chiếu quy trình);

ISO/IEC 38500, Information technology – Governance of IT for the organization (Quản trị công nghệ thông tin cho tổ chức);

ISO/IEC 38505-1, Information technology – Governance of IT – Governance of data – Part 1: Application of ISO/IEC 38500 to the governance of data (Phần 1: Áp dụng ISO/IEC 38500 trong quản trị dữ liệu);

ISO/IEC TR 38505-2, Information technology — Governance of IT — Governance of data — Part 2: Implications of ISO/IEC 38505-1 for data management (Phần 2: Ý nghĩa của ISO/IEC 38505 cho quản trị dữ liệu);

ISO 55000, Asset management — Overview, principles and terminology (Tổng quan, các nguyên tắc và thuật ngữ);

ISO 55001, Asset management — Management systems — Requirements (Các yêu cầu);

ISO 55002, Asset management — Management systems — Guidelines for the application of ISO 55001 (Hướng dẫn áp dụng ISO 55001);

ISO/IEC/IEEE 42010, Systems and software engineering — Architecture description (Mô tả kiến trúc).

3 Thuật ngữ và định nghĩa

Tiêu chuẩn này sử dụng các thuật ngữ và định nghĩa được nêu trong các tiêu chuẩn: TCVN 10249-2:2013 (ISO 8000-2), ISO/TS 8000-60, ISO 8000-61, ISO/IEC 38500, ISO/IEC 38505-1, ISO/IEC TR 38505-2, ISO 55000, ISO 55001, ISO 55002, ISO/IEC/IEEE 42010, TCVN 13238:2020 (ISO/IEC 20546), TCVN 12481:2019 (ISO/IEC 17789) và các thuật ngữ, định nghĩa sau:

3.1

Dữ liệu (Data)

Sự thể hiện thông tin chính thức phù hợp cho truyền thông, diễn giải hoặc xử lý.

CHÚ THÍCH 1: Dữ liệu có thể được xử lý bởi con người hoặc các phương tiện tự động.

[Nguồn: TCVN 13238:2020 (ISO/IEC 20546:2019), 3.1.5]

3.2

Kiến trúc tham chiếu (Reference architecture)

Nguồn thông tin có cơ sở về một chủ đề cụ thể để hướng dẫn và ràng buộc việc thuyết minh các kiến trúc và giải pháp khác nhau.

CHÚ THÍCH 1: Tiêu chuẩn này sử dụng định nghĩa về kiến trúc tham chiếu từ DoD "Reference architecture description" [7].

CHÚ THÍCH 2: Kiến trúc tham chiếu thường đóng vai trò là nền tảng cho các kiến trúc giải pháp và cũng có thể được sử dụng để so sánh và liên kết các bản thuyết minh các kiến trúc và giải pháp.

3.3

Thông tin (Information)

Dữ liệu (3.1) được xử lý, tổ chức và liên kết để tạo ra ý nghĩa.

CHÚ THÍCH 1: Thông tin được nhắc đến bao gồm các sự việc, khái niệm, đối tượng, sự kiện, ý tưởng, quy trình...

3.4

Hoạt động (Activity)

Một hoạt động cụ thể hoặc một tập các tác vụ.

[Nguồn: TCVN 12481:2019 (ISO/IEC 17789:2014), 3.2.1]

3.5

Tri thức (Knowledge)

Thông tin (3.3) được duy trì, xử lý và diễn giải.

[Nguồn: ISO 5127:2017, 3.1.1.17]

3.6

Thành phần chức năng (Functional component)

Khối xây dựng chức năng cần thiết cho một hoạt động (3.4) nào đó, được định hướng bởi một quá trình thực thi.

[Nguồn: TCVN 12481:2019 (ISO/IEC 17789:2014), 3.2.3]

3.7

Quản trị dữ liệu (Data governance)

Thuộc tính hoặc tính năng cần được phối hợp và thực hiện bởi một tập các hoạt động (3.4) nhằm mục đích thiết kế, thực thi và giám sát kế hoạch chiến lược để quản lý tài sản dữ liệu.

CHÚ THÍCH 1: Việc quản trị dữ liệu được mô tả trong ISO/IEC 38505-1.

CHÚ THÍCH 2: Tài sản dữ liệu được hiểu là một tập hợp các mục dữ liệu hoặc thực thể dữ liệu có lợi ích thực hoặc tiềm năng đối với một tổ chức. Tài sản dữ liệu là một tập con của tài sản được định nghĩa trong ISO 55000. Lợi ích là ưu thế đối với một tổ chức về các tri thức có thể áp dụng khi vận hành một một hệ thống phân tích. Lợi ích thường được gắn liền với dữ liệu lớn do người ta cho rằng dữ liệu lớn có nhiều tiềm năng mà trước đây thường không được xem xét tới.

CHÚ THÍCH 3: Kế hoạch chiến lược để quản lý tài sản dữ liệu là tài liệu quy định rõ cách thức quản lý dữ liệu (3.15) phù hợp với chiến lược tổ chức. Thuật ngữ này đồng nghĩa với kế hoạch quản lý tài sản chiến lược (SAMP) được định nghĩa trong ISO 55000 trên phương diện dữ liệu.

3.8

Chất lượng dữ liệu (Data quality)

Mức độ mà các đặc tính của dữ liệu đáp ứng các nhu cầu đã được đưa ra và áp dụng khi được sử dụng trong các điều kiện cụ thể.

[Nguồn: ISO/IEC 25024:2015, 4.11]

3.9

Quản lý chất lượng dữ liệu (Data quality management)

Các hoạt động phối hợp để chỉ đạo và kiểm soát một tổ chức khi nói đến chất lượng dữ liệu.

[Nguồn: TCVN 10249-2:2013 (ISO 8000-2:2018), 3.4.9]

3.10

Bên tham gia (Party)

Thể nhân hoặc pháp nhân, hoặc nhóm của thể nhân hoặc pháp nhân, có thể dưới hình thức doanh nghiệp hoặc không.

[Nguồn: TCVN 12481:2019 (ISO/IEC 17789:2014), 7.2.3]

3.11

Chính sách (Policy)

Mục đích và định hướng của một tổ chức do ban lãnh đạo của tổ chức xây dựng.

[Nguồn: ISO 55000:2014, 3.1.18, đã sửa đổi – Thuật ngữ này đã được sửa đổi thành dạng số ít và dấu chấm cuối cùng đã bị xóa.]

3.12

Vai trò (Role)

Tập hợp các hoạt động (3.4) nhằm phục vụ một mục đích chung.

[Nguồn: TCVN 12481:2019 (ISO/IEC 17789:2014), 3.2.7]

3.13

Luồng (Stream)

Danh mục các đối tượng dòng được gắn vào một cổng của một đối tượng dòng.

[Nguồn: ISO/IEC 10179:1996, 4.33, đã sửa đổi –xóa mục phía trước và bổ sung dấu chấm cuối cùng.]

3.14

Vai trò phụ (Sub-role)

Tập hợp con của các hoạt động (3.4) của một vai trò (3.12) cụ thể.

[Nguồn: TCVN 12481:2019 (ISO/IEC 17789:2014), 3.2.9]

3.15

Quản lý dữ liệu (Data management)

Tập hợp các hoạt động (3.4) nhằm triển khai kiến trúc dữ liệu lớn nhằm đáp ứng tốt nhất các mục đích kinh doanh bằng cách tuân thủ kế hoạch chiến lược trong việc đánh giá quản lý dữ liệu.

3.16

Vòng đời dữ liệu (Data lifecycle)

Các giai đoạn trong quản lý dữ liệu.

CHÚ THÍCH 1: Trong tiêu chuẩn này, mục tiêu của vòng đời (được định nghĩa trong ISO 55000) là dữ liệu.

3.17

Giao diện lập trình ứng dụng (Application programming interface - API)

Nơi phần mềm ứng dụng sử dụng các chức năng của ngôn ngữ lập trình để gọi các dịch vụ.

[Nguồn: ISO/IEC 18012-2:2012, 3.1.4, đã sửa đổi – CHÚ THÍCH 1 đã bị gỡ bỏ và dấu chấm cuối đã bị xóa]

4 Chữ viết tắt

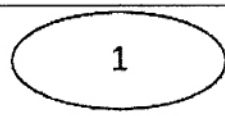
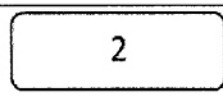
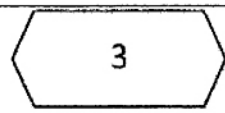
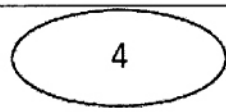
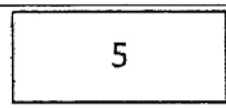
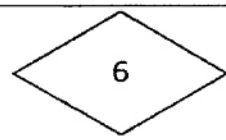
ACID	atomicity, consistency, isolation, and durability	Tính nguyên tử, tính nhất quán, sự cô lập và tính bền bỉ
API	application programming interface	Giao diện lập trình ứng dụng
CEP	complex event processing	Xử lý sự kiện phức tạp

CPU	central processing unit	Đơn vị xử lý trung tâm
BDA	big data auditor	Nhà kiểm toán dữ liệu lớn
BDAP	big data application provider	Nhà cung cấp ứng dụng dữ liệu lớn
BDAcP	big data access provider	Nhà cung cấp dịch vụ truy cập dữ liệu lớn
BDAnP	big data analytics provider	Nhà cung cấp dịch vụ phân tích dữ liệu lớn
BDC	big data consumer	Người dùng dữ liệu lớn
BDCP	big data collection provider	Nhà cung cấp dịch vụ thu thập dữ liệu lớn
BDFP	big data framework provider	Nhà cung cấp khung chức năng dữ liệu lớn
BDIP	big data infrastructure provider	Nhà cung cấp cơ sở hạ tầng dữ liệu lớn
BDP	big data provider	Nhà cung cấp dữ liệu lớn
BDPlaP	big data platform provider	Nhà cung cấp nền tảng dữ liệu lớn
BDPreP	big data preparation provider	Nhà cung cấp dịch vụ chuẩn bị dữ liệu lớn
BDProP	big data processing provider	Nhà cung cấp dịch vụ xử lý dữ liệu lớn
BDRA	big data reference architecture	Kiến trúc tham chiếu dữ liệu lớn
BDSD	big data service developer	Nhà phát triển dịch vụ dữ liệu lớn
BDSO	big data system orchestrator	Người điều hành hệ thống dữ liệu lớn
BDSP	big data service partner	Đối tác dịch vụ dữ liệu lớn
BDVP	big data visualization provider	Nhà cung cấp dịch vụ trực quan hóa dữ liệu lớn
DG	data governance	Quản trị dữ liệu
DM	data manager	Người quản lý dữ liệu
DQM	data quality manager	Người quản lý chất lượng dữ liệu
PII	personally identifiable information	Thông tin định danh cá nhân
RA	reference architecture	Kiến trúc tham chiếu

5 Quy ước

Các sơ đồ trong tiêu chuẩn này được trình bày theo các quy ước trong Bảng 1 dưới đây. Các ký hiệu này được sử dụng như mô tả trong ISO/IEC 17789.

Bảng 1 – Chú giải các sơ đồ được sử dụng trong tài liệu

Đối tượng	Ý nghĩa
	Bên tham gia
	Vai trò
	Vai trò phụ
	Hoạt động
	Thành phần chức năng
	Các khía cạnh xuyên suốt

6 Khái niệm kiến trúc tham chiếu dữ liệu lớn

6.1 Khái quát chung

Tiêu chuẩn này xác định vai trò của BDRA là điểm tham chiếu cơ bản cho việc tiêu chuẩn hóa dữ liệu lớn và cung cấp một khung kiến trúc tổng thể cho các khái niệm, nguyên tắc cơ bản của hệ thống dữ liệu lớn.

Tiêu chuẩn này mô tả các mối quan hệ logic giữa các vai trò/vai trò phụ, các hoạt động, các thành phần chức năng và các khía cạnh xuyên suốt tạo nên kiến trúc hệ thống dữ liệu lớn.

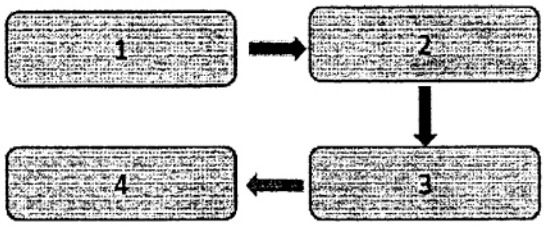
Các tiêu chuẩn có thể liên quan đến một số mối quan hệ. Các tiêu chuẩn kết hợp với một mối quan hệ có thể được sử dụng để:

- Xác định mức độ của luồng thông tin hoặc bất kỳ khả năng tương tác nào khác;
- Đảm bảo các mức chất lượng theo quy định (ví dụ: mức độ an toàn, bảo mật hoặc mức độ chất lượng dịch vụ).

Các mối quan hệ logic được định nghĩa trong kiến trúc này là một phần có ý nghĩa quan trọng trong việc xác định BDRA và các hành vi đi kèm. Mối quan hệ này mô tả các vấn đề như: các loại luồng thông tin giữa các thành phần chức năng trong BDRA.

6.2 Các góc nhìn

Dữ liệu lớn có thể được mô tả bằng cách sử dụng quan điểm góc nhìn. Bốn góc nhìn khác nhau được sử dụng trong BDRA (xem Hình 1 và Bảng 2):



Chỉ dẫn

1 Góc nhìn người dùng

2 Góc nhìn chức năng

3 Góc nhìn thực thi

4 Góc nhìn triển khai

Hình 1 – Sự chuyển đổi giữa các góc nhìn kiến trúc

Bảng 2 – Các góc nhìn BDRA

Góc nhìn BDRA	Mô tả góc nhìn BDRA	Phạm vi
Góc nhìn người dùng	Hệ sinh thái của dữ liệu lớn với các bên liên quan (được sử dụng trong ISO/IEC/IEEE 42010), các vai trò, vai trò phụ và các hoạt động dữ liệu lớn	Trong phạm vi tiêu chuẩn
Góc nhìn chức năng	Các chức năng cần thiết để hỗ trợ các hoạt động dữ liệu lớn	Trong phạm vi tiêu chuẩn
Góc nhìn thực thi	Các chức năng cần thiết cho việc áp dụng dữ liệu lớn trong cấu phần dịch vụ và (hoặc) cấu phần cơ sở hạ tầng	Ngoài phạm vi tiêu chuẩn
Góc nhìn triển khai	Cách mà các chức năng của dữ liệu lớn được triển khai về mặt kỹ thuật khi sử dụng các thành phần cơ sở hạ tầng sẵn có hoặc khi bổ sung các thành phần mới cho hệ thống cơ sở hạ tầng sẵn có này	Ngoài phạm vi tiêu chuẩn

CHÚ THÍCH: Tiêu chuẩn này đề cập đến chi tiết về góc nhìn người dùng và góc nhìn chức năng, mà không mô tả góc nhìn thực thi và góc nhìn triển khai vì có liên quan đến công nghệ và việc triển khai dữ liệu lớn cụ thể của các nhà cung cấp trong thực tế. Do vậy, những góc nhìn đó không thuộc phạm vi của tiêu chuẩn này.

6.3 Tổng quan về góc nhìn người dùng

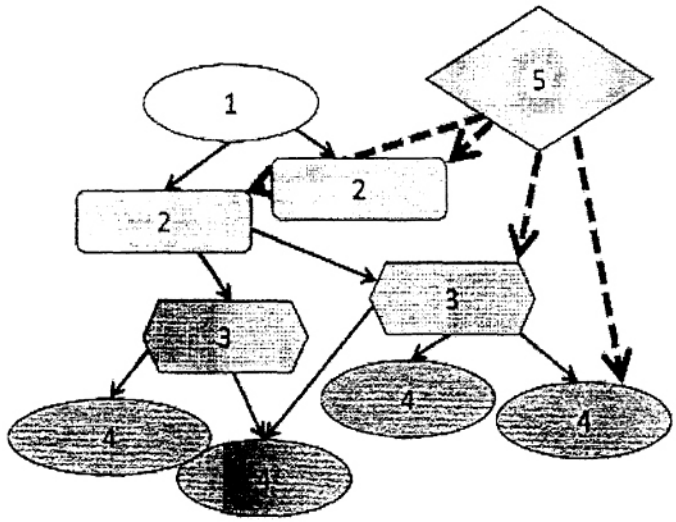
Góc nhìn người dùng tập trung vào hệ sinh thái của dữ liệu lớn với các khái niệm sau:

- Các bên tham gia: thể nhân hoặc pháp nhân, hoặc nhóm thể nhân/pháp nhân, dưới hình thức doanh nghiệp hay không;

- Các vai trò và vai trò phụ: vai trò là một tập hợp các hoạt động dữ liệu lớn phục vụ một mục đích chung. Vai trò phụ là một tập hợp con của các hoạt động dữ liệu lớn để thực hiện một vai trò nhất định. Các vai trò phụ khác nhau có thể chia sẻ các hoạt động dữ liệu lớn trong một vai trò nhất định;
- Các hoạt động: hoạt động là việc thực hiện một hoặc một tập hợp các nhiệm vụ cụ thể. Các hoạt động dữ liệu lớn cần có mục đích và mang lại một hoặc nhiều kết quả và những kết quả này có được bằng cách sử dụng các thành phần chức năng.
- Các khía cạnh xuyên suốt: các khía cạnh xuyên suốt có thể được chia sẻ và ảnh hưởng đến nhiều vai trò, các hoạt động dữ liệu lớn. Các khía cạnh xuyên suốt có thể ánh xạ đến các chức năng nhiều lớp cùng các thành phần chức năng liên quan của chúng để thực hiện các hoạt động trong khía cạnh xuyên suốt.

CHÚ THÍCH: Một Bên tham gia có thể đảm nhận nhiều hơn một vai trò tại bất kỳ thời điểm nào và có thể tham gia vào một tập con các hoạt động cụ thể của vai trò đó. Ví dụ, các Bên tham gia có thể là: các tập đoàn lớn, các doanh nghiệp vừa và nhỏ, các cơ quan chính phủ, các tổ chức học thuật và các cá nhân.

Hình 2 minh họa các thực thể được xác định cho góc nhìn người dùng.



Chỉ dẫn

1 Bên tham gia
2 Vai trò
3 Vai trò phụ
4 Hoạt động
5 Khía cạnh xuyên suốt

Hình 2 – Các thực thể của góc nhìn người dùng

6.4 Tổng quan về góc nhìn chức năng

Góc nhìn chức năng là một góc nhìn mang tính công nghệ trung lập về các chức năng cần thiết để tạo thành một hệ thống dữ liệu lớn. Góc nhìn chức năng mô tả sự phân bổ các chức năng cần thiết để hỗ trợ các hoạt động dữ liệu lớn.

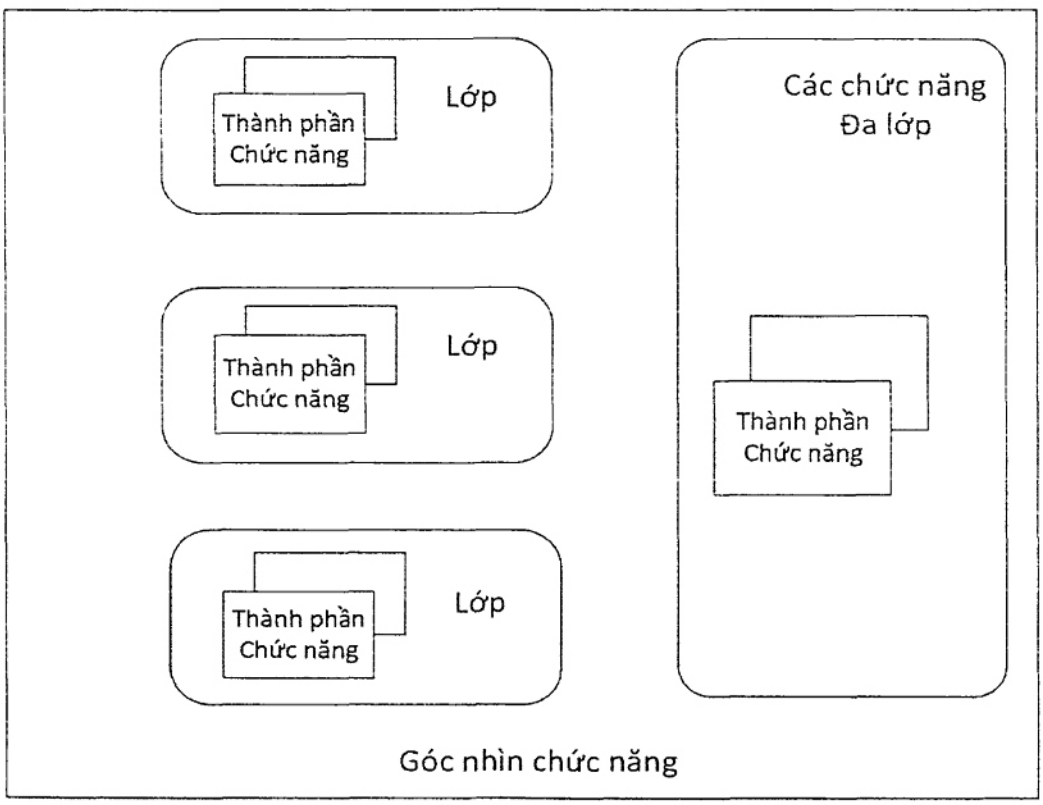
Kiến trúc chức năng cũng xác định sự phụ thuộc giữa các chức năng.

Góc nhìn chức năng đề cập đến các khái niệm dữ liệu lớn sau:

- **Các thành phần chức năng:** một thành phần chức năng là một khối chức năng thành phần cần thiết để tham gia vào một hoạt động, do một hoạt động thực thi thực hiện;
- **Các lớp chức năng:** một lớp là một tập các thành phần chức năng cung cấp các khả năng tương tự hoặc phục vụ một mục đích chung;
- **Các chức năng nhiều lớp:** các chức năng nhiều lớp bao gồm các thành phần chức năng có khả năng có thể được sử dụng trên nhiều lớp chức năng và những lớp chức năng này được nhóm lại thành các tập con.

CHÚ THÍCH: Không phải tất cả các lớp hoặc các thành phần chức năng đều phải được khởi tạo trong một hệ thống dữ liệu lớn cụ thể.

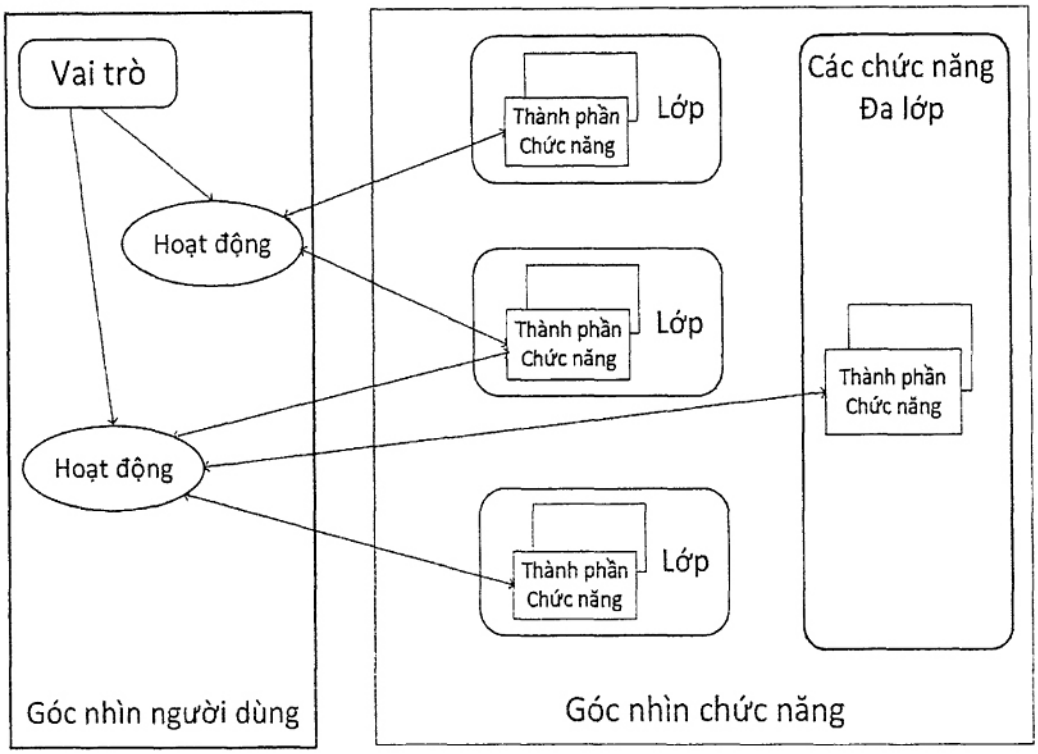
Hình 3 mô tả các khái niệm về các thành phần chức năng, các lớp và các chức năng nhiều lớp.



Hình 3 – Phân lớp chức năng

6.5 Mối quan hệ giữa góc nhìn người dùng và góc nhìn chức năng

Hình 4 minh họa cách mà góc nhìn người dùng cung cấp tập các hoạt động dữ liệu lớn được thể hiện trong góc nhìn chức năng.



Hình 4 – Từ góc nhìn người dùng đến góc nhìn chức năng

6.6 Mối quan hệ của góc nhìn người dùng và góc nhìn chức năng với các khía cạnh xuyên suốt

Các khía cạnh xuyên suốt là cấu phần của cả góc nhìn người dùng và góc nhìn chức năng của dữ liệu lớn.

Trong góc nhìn người dùng, các khía cạnh xuyên suốt ảnh hưởng đến các vai trò và vai trò phụ và tác động trực tiếp hoặc gián tiếp đến các hoạt động mà các vai trò đó thực hiện.

Trong góc nhìn chức năng, các khía cạnh xuyên suốt tác động đến các thành phần chức năng, và được sử dụng khi thực hiện các hoạt động được mô tả trong góc nhìn người dùng (Hình 4).

Các khía cạnh xuyên suốt của dữ liệu lớn được mô tả trong mục 9, gồm có:

- Tính bảo mật và tính riêng tư;
- Quản lý;
- Quản trị dữ liệu.

7 Góc nhìn người dùng

7.1 Vai trò, vai trò phụ và hoạt động của dữ liệu lớn

Do các dịch vụ phân tán và việc triển khai các dịch vụ này diễn ra ở cốt lõi của dữ liệu lớn, tất cả các hoạt động liên quan đến dữ liệu lớn có thể được phân loại thành ba nhóm chính: sử dụng dữ liệu lớn, cung cấp dịch vụ phân tích dữ liệu lớn và cung cấp dữ liệu.

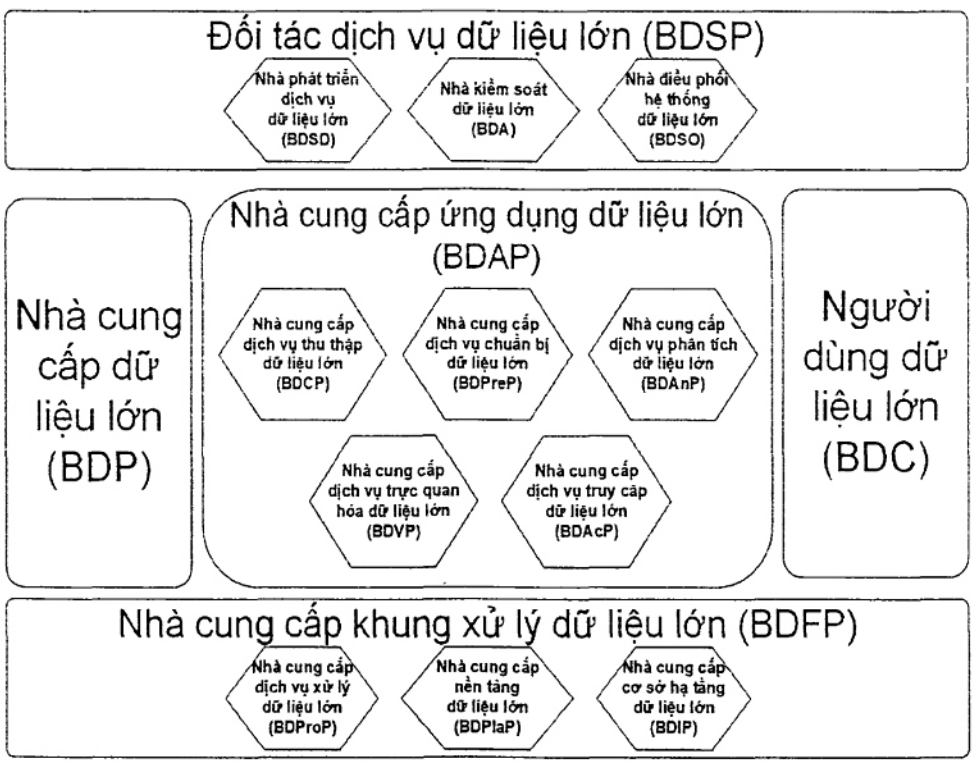
Mục này mô tả về một số vai trò phổ biến và vai trò phụ liên quan đến dữ liệu lớn.

Điều quan trọng cần lưu ý là tại một thời điểm bất kỳ, một chủ thể có thể đóng nhiều hơn một vai trò . Khi đóng một vai trò nhất định, chủ thể đó có thể hạn chế đóng một hoặc nhiều vai trò phụ. Vai trò phụ là một tập con của các hoạt động dữ liệu lớn của một vai trò nhất định.

Như trình bày trong Hình 5, các vai trò của dữ liệu lớn là:

- Nhà cung cấp ứng dụng dữ liệu lớn (BDAP) (xem 7.2);
- Nhà cung cấp khung chức năng dữ liệu lớn (BDFP) (xem 7.3);
- Đối tác dịch vụ dữ liệu lớn (BDSP) (xem 7.4);
- Nhà cung cấp dữ liệu lớn (BDP) (xem 7.5);
- Người dùng dữ liệu lớn (BDC) (xem 7.6).

CHÚ THÍCH: Nhà cung cấp dữ liệu lớn là bất kỳ nhà cung cấp dữ liệu nào cho BDRA.



Hình 5 – Các vai trò của dữ liệu lớn

Phụ lục B cung cấp các ví dụ minh họa về mối quan hệ của các vai trò trong hệ sinh thái dữ liệu lớn.

Mỗi vai trò phụ thể hiện trong Hình 5 được mô tả chi tiết hơn trong mục 7.2 đến 7.6.

7.2 Vai trò: Đơn vị cung cấp ứng dụng dữ liệu lớn (BDAP)

7.2.1 Khái quát chung

BDAP thao túng toàn bộ vòng đời của dữ liệu lớn. Đây là nơi kết hợp các tính năng chung trong góc nhìn người dùng về kiến trúc tham chiếu dữ liệu lớn như trong Hình 5 để tạo ra hệ thống dữ liệu cụ thể.

CHÚ THÍCH 1: Mặc dù các hoạt động của các nhà cung cấp ứng dụng là như nhau dù giải pháp có liên quan đến dữ liệu lớn hay không, thì các phương pháp và kỹ thuật cũng thay đổi bởi vì dữ liệu và xử lý dữ liệu được thực hiện song song giữa các tài nguyên.

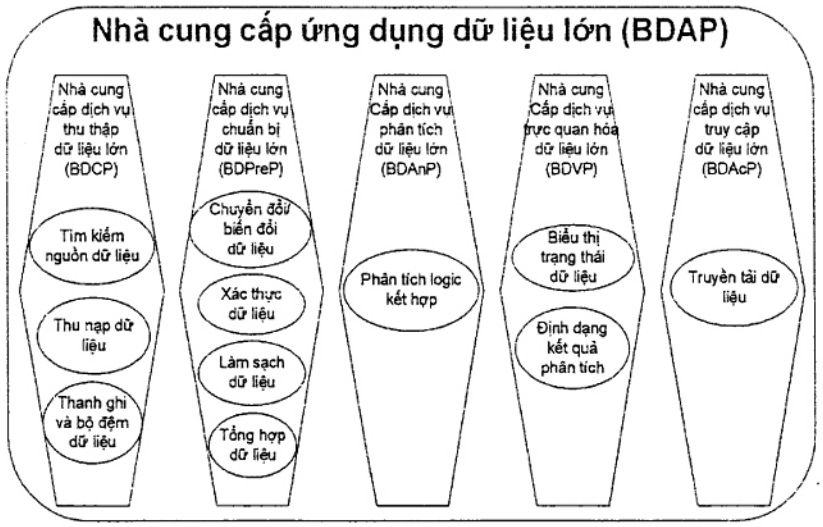
CHÚ THÍCH 2: Khi dữ liệu truyền qua hệ sinh thái, chúng được xử lý và biến đổi theo những cách khác nhau để trích xuất giá trị từ thông tin. Mỗi hoạt động của nhà cung cấp ứng dụng dữ liệu lớn có thể được thực hiện bởi các bên liên quan độc lập và được triển khai như các dịch vụ độc lập.

CHÚ THÍCH 3: BDAP có thể là một thực thể đơn lẻ hoặc một tập các nhà cung cấp ứng dụng dữ liệu lớn cụ thể, mỗi nhà cung cấp thực hiện các bước khác nhau trong một vòng đời của dữ liệu lớn. Mỗi hoạt động của nhà cung cấp ứng dụng dữ liệu lớn có thể là một dịch vụ chung do bởi nhà cung cấp dữ liệu hoặc người sử dụng dữ liệu lớn yêu cầu, như máy chủ web, máy chủ tệp, tập hợp của một hoặc nhiều chương trình ứng dụng hoặc kết hợp.

CHÚ THÍCH 4: BDAP chịu trách nhiệm thực hiện, kiểm tra và xác nhận các quy tắc, yêu cầu nghiệp vụ về chất lượng dữ liệu và các chỉ số đảm bảo việc quản lý dữ liệu chính xác trong hệ thống dữ liệu lớn. Bất kỳ nhà cung cấp ứng dụng dữ liệu lớn nào cũng có thể áp dụng các yêu cầu về chất lượng dữ liệu trong suốt vòng đời của dữ liệu lớn.

BDAP bao gồm năm vai trò phụ sau, được thể hiện như trong Hình 6:

- Nhà cung cấp dịch vụ thu thập dữ liệu lớn (BDCP) (xem 7.2.2);
- Nhà cung cấp dịch vụ chuẩn bị dữ liệu lớn (BDPreP) (xem 7.2.3);
- Nhà cung cấp dịch vụ phân tích dữ liệu lớn (BDAnP) (xem 7.2.4);
- Nhà cung cấp dịch vụ trực quan hóa dữ liệu lớn (BDVP) (xem 7.2.5);
- Nhà cung cấp dịch vụ truy cập dữ liệu lớn (BDAcP) (xem 7.2.6);



Hình 6 – Các hoạt động dữ liệu lớn liên quan đến vai trò phụ của nhà cung cấp ứng dụng dữ liệu lớn

7.2.2 Vai trò phụ: Đơn vị cung cấp ứng dụng thu thập dữ liệu lớn (BDCP)

BDCP là một vai trò phụ của BDAP, chịu trách nhiệm thu thập dữ liệu lớn từ nhà cung cấp dữ liệu. Đây có thể là một dịch vụ chung, như máy chủ tệp, máy chủ web để chấp nhận hoặc thực hiện việc thu thập các dữ liệu cụ thể hoặc có thể là một dịch vụ ứng dụng cụ thể được thiết kế để lấy dữ liệu hoặc nhận dữ liệu từ nhà cung cấp dữ liệu.

Các hoạt động của BDCP gồm:

- Hoạt động tìm nguồn dữ liệu: tập trung vào việc tìm kiếm và lưu trữ thông tin nguồn dữ liệu như một dạng siêu dữ liệu, mà có thể được sử dụng để giữ lại hoặc lưu trữ dữ liệu;
- Việc thu thập dữ liệu tập trung vào việc chuyển đổi dữ liệu có sẵn (ví dụ: tài liệu web, dữ liệu blog, vv...) thành một biểu mẫu có thể được xử lý bởi hệ thống;
- Hoạt động của thanh ghi và bộ đệm dữ liệu tập trung vào việc lưu trữ dữ liệu vào thanh ghi dữ liệu hoặc lưu trữ dữ liệu trước khi chuyển nó sang các tác vụ hoặc quy trình khác.

7.2.3 Vai trò phụ: Đơn vị cung cấp ứng dụng chuẩn bị dữ liệu lớn (BDPreP)

BDPreP là một vai trò phụ của BDAP, có nhiệm vụ chuẩn bị dữ liệu từ dữ liệu thô sang dữ liệu sẵn sàng để phân tích.

Các hoạt động của BDPreP gồm:

- Hoạt động chuyển đổi dữ liệu: tập trung vào việc chuyển đổi dữ liệu hoặc thông tin từ định dạng này sang định dạng khác;
- Hoạt động xác thực dữ liệu: tập trung vào việc đảm bảo tính chính xác của dữ liệu dựa trên các ràng buộc xác thực như tính đúng đắn, ý nghĩa, bảo mật và quyền riêng tư...;
- Hoạt động làm sạch dữ liệu: tập trung vào việc phát hiện phần dữ liệu không chính xác và sửa chúng bằng cách thay thế, sửa đổi hoặc xóa;
- Hoạt động tổng hợp dữ liệu: tập trung vào việc kết hợp hai hoặc nhiều dữ liệu thành một tập dữ liệu dạng mẫu tổng hợp.

Việc xác thực dữ liệu và làm sạch dữ liệu phải được hướng dẫn bằng việc áp dụng quản lý chất lượng dữ liệu.

7.2.4 Vai trò phụ: Đơn vị cung cấp ứng dụng phân tích dữ liệu lớn (BDAnP)

BDAnP là một vai trò phụ của BDAP, có nhiệm vụ phân tích dữ liệu lớn nhằm đáp ứng các yêu cầu của thuật toán để xử lý dữ liệu nhằm tạo ra thông tin chi tiết đáp ứng mục tiêu kỹ thuật.

Hoạt động của BDAnP gồm có một hoạt động logic phân tích đi kèm liên quan đến việc mô hình hóa các quy trình dữ liệu với logic đã cho để trích xuất thông tin từ dữ liệu dựa trên các yêu cầu của ứng dụng.

7.2.5 Vai trò phụ: Đơn vị cung cấp ứng dụng trực quan (BDVP)

BDVP là một vai trò phụ của BDAP, có nhiệm vụ biểu diễn thông tin nguồn dữ liệu hoặc kết quả phân tích dữ liệu cho người dùng dữ liệu lớn. Mục tiêu của hoạt động này là định dạng và biểu diễn dữ liệu theo hướng tối ưu hóa thông tin và kiến thức giao tiếp.

Các hoạt động của BDVP như sau:

- Hoạt động biểu thị trạng thái dữ liệu liên quan đến việc mô tả trạng thái dữ liệu trong bộ lưu trữ dữ liệu, bao gồm: trực quan hóa, phân loại...;
- Hoạt động định dạng kết quả phân tích liên quan đến việc định dạng dữ liệu đã xử lý để việc truyền tải thông tin rõ ràng và hiệu quả. Hoạt động này có thể bao gồm biểu diễn trực quan, che phủ...

7.2.6 Vai trò phụ: Đơn vị cung cấp ứng dụng truy cập dữ liệu lớn (BDAcP)

BDAcP là một vai trò phụ của BDAP, có nhiệm vụ trao đổi dữ liệu lớn giữa ứng dụng dữ liệu lớn và nhà cung cấp dữ liệu hoặc người dùng dữ liệu lớn.

Hoạt động BDAcP bao gồm hoạt động truyền dữ liệu tập trung vào việc truyền hoặc di chuyển dữ liệu lớn từ hệ thống này sang hệ thống khác mà vẫn đảm bảo tính toàn vẹn, liên tục, bảo mật và quyền riêng tư trong quá trình truyền dữ liệu.

7.3 Vai trò: Đơn vị cung cấp khung xử lý dữ liệu lớn (BDFP)

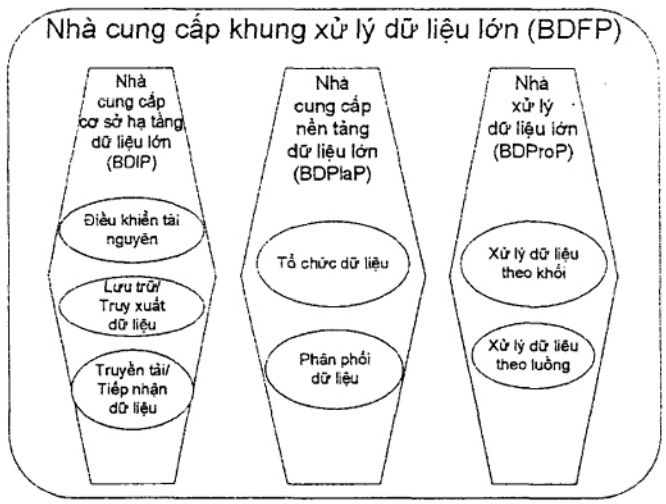
7.3.1 Khái quát chung

BDFP gồm một hoặc nhiều phân cấp được tổ chức theo các đối tượng của các thành phần. Không có yêu cầu nào về việc các đối tượng ở cùng một cấp nhất định trong hệ thống phân cấp phải có cùng một công nghệ.

CHÚ THÍCH: Trong thực tế, hầu hết việc triển khai dữ liệu lớn sử dụng phương pháp lai, kết hợp nhiều phương pháp tiếp cận công nghệ để đảm bảo sự linh hoạt hoặc để đáp ứng toàn bộ các yêu cầu do các nhà cung cấp ứng dụng dữ liệu lớn đưa ra.

BDFP bao gồm ba vai trò phụ sau, được trình bày như trong Hình 7:

- Nhà cung cấp cơ sở hạ tầng dữ liệu lớn (BDIP) (xem 7.3.2);
- Nhà cung cấp nền tảng dữ liệu lớn (BDPlaP) (xem 7.3.3);
- Nhà cung cấp xử lý dữ liệu lớn (BDProP) (xem 7.3.4);



Hình 7 – Các hoạt động dữ liệu lớn liên quan đến vai trò phụ của nhà cung cấp Khung xử lý dữ liệu lớn

7.3.2 Vai trò phụ: Đơn vị cung cấp cơ sở hạ tầng dữ liệu lớn (BDIP)

BDIP là một vai trò phụ của BDFP, có nhiệm vụ cung cấp tài nguyên hệ thống bao gồm các hệ thống cơ sở (như hệ thống mạng, tính toán, lưu trữ...) và môi trường vật lý (như phòng máy, nguồn điện, điều hòa không khí...).

Các hoạt động của BDIP gồm:

- Hoạt động điều khiển tài nguyên: tập trung vào việc xử lý hoặc kiểm soát các tài nguyên vật lý hoặc tài nguyên ảo;
- Hoạt động lưu trữ/truy xuất dữ liệu: liên quan đến việc duy trì và truy xuất dữ liệu từ kho lưu trữ (thao tác dữ liệu ở trạng thái nghỉ);
- Hoạt động truyền tải/tiếp nhận dữ liệu tập trung vào việc truyền dữ liệu qua mạng (đưa dữ liệu vào chuyển động).

7.3.3 Vai trò phụ: Đơn vị cung cấp nền tảng dữ liệu lớn (BDPlaP)

BDPlaP là một vai trò phụ của BDFP, có nhiệm vụ cung cấp các nền tảng để tổ chức và phân phối dữ liệu lớn trên hạ tầng dữ liệu lớn.

Các hoạt động của BDPlaP gồm:

- Hoạt động tổ chức dữ liệu: liên quan đến việc sắp xếp, lập chỉ mục và liên kết dữ liệu theo những cách phù hợp với các ứng dụng và phân tích cụ thể;
- Hoạt động phân phối dữ liệu: liên quan đến việc phân bổ dữ liệu trên các tài nguyên hạ tầng cơ sở để tối đa hóa vị trí dữ liệu cho hiệu suất tính toán phân tán.

7.3.4 Vai trò phụ: Đơn vị xử lý dữ liệu lớn (BDProP)

BDProP là một vai trò phụ của BDFP, có nhiệm vụ hỗ trợ quá trình tính toán và phân tích cho các hoạt động của BDAP.

Các hoạt động của BDProP gồm:

- Xử lý dữ liệu theo khối: xử lý dữ liệu theo khối lượng lớn và trên cơ sở không liên tục. Quá trình xử lý theo khối được áp dụng khi thời gian phản hồi không phải là quan trọng. Quá trình xử lý khối thường liên quan đến khối lượng dữ liệu hoặc độ phức tạp của phân tích;
- Xử lý dữ liệu theo luồng: xử lý dữ liệu liên tục với số lượng nhỏ (thường là các bản ghi riêng lẻ hoặc các phần tử dữ liệu). Xử lý dữ liệu theo luồng được sử dụng khi thời gian phản hồi là quan trọng và thường liên quan đến tốc độ của dữ liệu.

7.4 Vai trò: Đối tác dịch vụ dữ liệu lớn (BDSP)

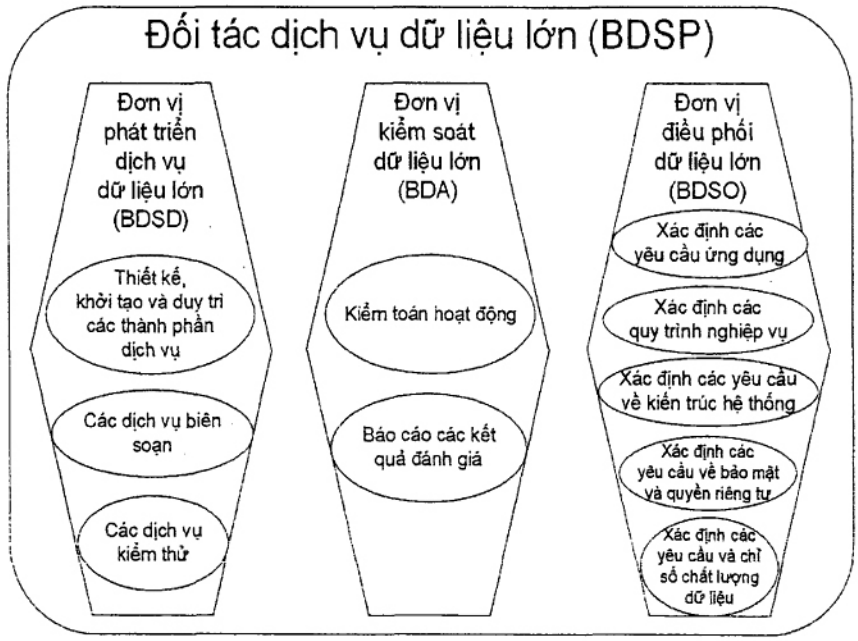
7.4.1 Khái quát chung

BDSP là một vai trò tham gia vào việc hỗ trợ hoặc bổ trợ cho các hoạt động giữa nhà cung cấp ứng dụng dữ liệu lớn, nhà cung cấp khung dữ liệu lớn, nhà cung cấp dữ liệu lớn hoặc người sử dụng dữ liệu lớn, hoặc tất cả.

Hoạt động dữ liệu lớn của BDSP thay đổi tùy thuộc vào loại đối tác mối quan hệ của họ với các vai trò khác trong hệ sinh thái dữ liệu lớn.

BDSP gồm có ba vai trò phụ, như thể hiện trong Hình 8:

- Đơn vị phát triển dịch vụ dữ liệu lớn (BDSD) (xem 7.4.2);
- Đơn vị kiểm soát dữ liệu lớn (BDA) (xem 7.4.3);
- Đơn điều phối hệ thống dữ liệu lớn (BDSO) (xem 7.4.4).



Hình 8 – Các hoạt động dữ liệu lớn liên quan đến vai trò phụ của đối tác dịch vụ dữ liệu lớn

7.4.2 Vai trò phụ: Đơn vị phát triển dịch vụ dữ liệu lớn (BDSD)

BDSD là một vai trò phụ của BDSP, có nhiệm vụ thiết kế, phát triển, thử nghiệm và duy trì việc thực hiện một dịch vụ dữ liệu lớn. Điều này có thể bao gồm việc soạn thảo quy trình triển khai dịch vụ từ các dịch vụ đã triển khai.

Các hoạt động của BDSD gồm có:

- Hoạt động thiết kế, khởi tạo và duy trì các thành phần dịch vụ liên quan đến việc thiết kế và khởi tạo các thành phần, phần mềm nằm trong quá trình triển khai dịch vụ dữ liệu lớn và cung cấp các bản sửa lỗi hoặc cải tiến cho việc triển khai dịch vụ;
- Hoạt động dịch vụ biên soạn tập trung vào việc soạn thảo các quy trình dịch vụ sử dụng các dịch vụ có sẵn bằng phương thức trung gian, tổng hợp;
- Hoạt động dịch vụ kiểm thử tập trung vào việc kiểm thử các thành phần và dịch vụ do nhà phát triển dịch vụ dữ liệu lớn cung cấp.

7.4.3 Vai trò phụ: Đơn vị kiểm toán dữ liệu lớn (BDA)

BDA là một vai trò phụ của BDSP, có nhiệm vụ thực hiện kiểm toán việc cung cấp và sử dụng các dịch vụ dữ liệu lớn. Kiểm toán dữ liệu lớn bao gồm: tính xác thực của các nguồn dữ liệu, quá trình vận hành, hiệu năng, bảo mật và quyền riêng tư; đồng thời điểm tra xem các tiêu chí kiểm toán có được thỏa mãn hay không.

CHÚ THÍCH 1: Có nhiều thông số kỹ thuật khác nhau cho các tiêu chí kiểm toán, ví dụ: giải quyết các vấn đề về bảo mật [8].

Các hoạt động của BDA gồm có:

- Hoạt động kiểm toán hiệu quả yêu cầu hoặc thu thập bằng chứng kiểm toán, tiến hành bất kỳ thử nghiệm bắt buộc nào trên hệ thống hoặc dữ liệu được đánh giá và thu thập bằng chứng theo chương trình;
- Hoạt động báo cáo kết quả kiểm toán liên quan đến việc cung cấp một bản báo cáo dạng văn bản về kết quả của cuộc kiểm toán.

CHÚ THÍCH 2: BDA chịu trách nhiệm đánh giá chất lượng dữ liệu, định nghĩa và đánh giá các mức dịch vụ chất lượng dữ liệu, đo lường và giám sát chất lượng dữ liệu liên tục.

7.4.4 Vai trò phụ: Đơn vị điều phối hệ thống dữ liệu lớn (BDSO)

BDSO là một vai trò phụ của BDSP, cung cấp các yêu cầu tổng thể mà hệ thống phải đáp ứng, bao gồm các yêu cầu về chính sách, quản trị, kiến trúc, tài nguyên và nghiệp vụ, cũng như các hoạt động giám sát để đảm bảo hệ thống tuân thủ các yêu cầu đó.

Các hoạt động của BDSO gồm có:

- Hoạt động xác định các yêu cầu ứng dụng đề cập đến các yêu cầu tổng thể mà ứng dụng dữ liệu lớn cần phải đáp ứng;
- Hoạt động xác định quy trình nghiệp vụ đề cập đến một tập các hoạt động nghiệp vụ được sắp xếp thành từng phần để thực hiện một mục đích nhất định của đơn vị hoặc một bộ phận của đơn vị nhằm đạt được một số kết quả cuối cùng như mong đợi;
- Hoạt động xác định các yêu cầu về kiến trúc hệ thống đề cập đến các yêu cầu về khái niệm để xác định cấu trúc, hành vi và góc nhìn của một hệ thống dữ liệu lớn;
- Hoạt động xác định các yêu cầu về bảo mật và quyền riêng tư tập trung vào việc xác định yêu cầu về bảo mật và quyền riêng tư theo góc nhìn quản trị;
- Hoạt động xác định các yêu cầu và chỉ số chất lượng dữ liệu tập trung vào việc phát triển và nâng cao nhận thức về chất lượng dữ liệu và định nghĩa các quy tắc, yêu cầu, chỉ số về nghiệp vụ chất lượng dữ liệu.

7.5 Vai trò: Đơn vị cung cấp dữ liệu lớn (BDP)

Nhà cung cấp dữ liệu lớn (BDP) cung cấp dữ liệu cho chính họ hoặc cho những đối tượng khác. Để thực hiện vai trò của mình, BDP tạo ra một bản tóm tắt nhiều loại nguồn dữ liệu khác nhau như dữ liệu thô hoặc dữ liệu đã được hệ thống khác chuyển đổi trước đó và cung cấp chúng qua các giao diện chức năng khác nhau.

CHÚ THÍCH: Khái niệm về nhà cung cấp dữ liệu không phải là mới, khả năng thu thập và phân tích dữ liệu lớn hơn đã mở ra những tiềm năng mới về cung cấp dữ liệu có giá trị.

Hình 9 – Các hoạt động dữ liệu lớn liên quan đến nhà cung cấp dữ liệu lớn

Các hoạt động của BDP như sau (xem Hình 9):

- Hoạt động cung cấp dữ liệu có sẵn được tập trung vào việc mở ra hoặc phân phối nguồn dữ liệu ra bên ngoài hệ thống theo mục tiêu ban đầu;
- Hoạt động của kiểu nguồn dữ liệu tóm tắt liên quan đến việc xuất bản siêu dữ liệu hoặc danh mục dữ liệu nhằm mục đích phân phối dữ liệu qua thanh ghi.

CHÚ THÍCH: Khi cung cấp dữ liệu cho đối tượng khác, nhà cung cấp dữ liệu lớn có thể giám sát dữ liệu và quản lý các vấn đề về chất lượng dữ liệu được quy định bởi việc quản lý chất lượng dữ liệu.

7.6 Vai trò: Người dùng dữ liệu lớn (BDC)

Người dùng dữ liệu lớn (BDC) nhận kết quả từ đầu ra của hệ thống dữ liệu lớn. Theo nhiều khía cạnh, BDC nhận được giao diện chức năng cùng loại với giao diện mà nhà cung cấp dữ liệu lớn (BDP) đưa ra cho nhà cung cấp ứng dụng dữ liệu lớn (BDAP). Sau khi hệ thống gia tăng giá trị vào các nguồn dữ liệu ban đầu, BDAP cung cấp loại giao diện chức năng tương tự cho người dùng dữ liệu lớn (BDC).

Hình 10 – Các hoạt động dữ liệu lớn liên quan đến người dùng dữ liệu lớn

Các hoạt động của BDC như sau (xem Hình 10):

- Hoạt động sử dụng dữ liệu lớn tập trung vào việc sử dụng kết quả phân tích dữ liệu lớn hoặc sử dụng các giao diện ứng dụng do nhà cung cấp ứng dụng dữ liệu lớn cung cấp cho mục đích nghiệp vụ của người dùng dữ liệu lớn;

- Hoạt động đánh giá dữ liệu lớn liên quan đến việc đánh giá chất lượng của dữ liệu lớn hoặc ứng dụng dữ liệu lớn dưới dạng ý kiến phản hồi.

8 Các khía cạnh xuyên suốt

8.1 Khái quát chung

Các khía cạnh xuyên suốt bao gồm:

- Bảo mật và quyền riêng tư: khía cạnh này liên quan đến cách các hệ thống và dữ liệu được bảo đảm bằng cách duy trì tính bảo mật, toàn vẹn và khả dụng của chúng khỏi các rủi ro và cách các thông tin nhận dạng cá nhân được bảo vệ để tránh bị sử dụng trái phép;
- Quản lý: khía cạnh này liên quan đến cách mà các thành phần của hệ thống và tài nguyên được phân bổ, cấu hình, sử dụng và giám sát;
- Quản trị dữ liệu: khía cạnh này liên quan đến cách dữ liệu được kiểm soát và quản lý trong hệ thống trong suốt vòng đời của nó.

8.2 Bảo mật và quyền riêng tư

Các vấn đề về bảo mật và quyền riêng tư ảnh hưởng đến tất cả các vai trò và vai trò phụ khác trong hệ sinh thái dữ liệu lớn và các thành phần chức năng của BDRA. Bảo mật và quyền riêng tư tương tác với nhà điều hành hệ thống dữ liệu lớn về chính sách, các yêu cầu và việc kiểm tra quản lý; cũng như với cả nhà cung cấp ứng dụng dữ liệu lớn và nhà cung cấp khung dữ liệu lớn để phát triển, triển khai và vận hành.

Các vấn đề liên quan đến bảo mật trong dữ liệu lớn bao gồm:

- Tính bảo mật: đảm bảo rằng các hệ thống và dữ liệu không được cung cấp hoặc tiết lộ cho các cá nhân, thực thể hoặc các quy trình trái phép;
- Tính toàn vẹn: đảm bảo rằng hệ thống và dữ liệu là chính xác và đầy đủ;
- Tính khả dụng: đảm bảo rằng các hệ thống và dữ liệu có thể truy cập và sử dụng được theo yêu cầu bởi một thực thể có thẩm quyền.

Các vấn đề liên quan đến quyền riêng tư trong dữ liệu lớn bao gồm:

- Tính không liên kết: đảm bảo rằng một PII chính có thể sử dụng nhiều tài nguyên hoặc dịch vụ mà không ai khác có thể liên kết những giá trị này lại với nhau;
- Tính minh bạch: đảm bảo rằng việc đạt được một mức độ rõ ràng, phù hợp của các quy trình trong quá trình xử lý dữ liệu liên quan đến quyền riêng tư để việc thu thập, xử lý và sử dụng thông tin có thể được nắm bắt và xây dựng lại bất kỳ lúc nào;
- Khả năng can thiệp: đảm bảo rằng các PII chính, người kiểm soát PII, bộ xử lý PII và các cơ quan giám sát có thể can thiệp vào tất cả các quá trình xử lý dữ liệu liên quan đến quyền riêng tư. (xem ISO/IEC 20547-4[28], ISO/IEC 27000[29])

8.3 Quản lý

Các đặc điểm của dữ liệu lớn về khối lượng, vận tốc, sự đa dạng và biến đổi đòi hỏi một nền tảng quản lý hệ thống và phần mềm linh hoạt để cung cấp, cấu hình gói và phần mềm và quản lý chúng; cùng với việc giám sát, quản lý tài nguyên và hiệu suất. Quản lý dữ liệu lớn đòi hỏi sự xem xét về hệ thống, dữ liệu, bảo mật và quyền riêng tư ở quy mô lớn, đồng thời duy trì chất lượng dữ liệu ở mức cao và khả năng truy cập an toàn.

Các vấn đề liên quan đến quản lý trong dữ liệu lớn gồm những điều sau:

- Cung ứng (phân bổ): là hành động cấu hình tài nguyên hệ thống để hỗ trợ một tác vụ cụ thể. Việc phân bổ có thể diễn ra ở nhiều cấp trên toàn bộ kiến trúc hệ thống, từ phân bổ tài nguyên cho máy ảo đến phân bổ tài nguyên cho một công việc cụ thể trên một hoặc nhiều nút. Những vấn đề này liên quan đến hiệu quả của việc sử dụng và cấu hình các tài nguyên để hỗ trợ một hoặc nhiều nhiệm vụ.
- Cấu hình: liên quan đến việc thiết lập các tham số thích hợp trong các phần tử hệ thống để thực thi và sử dụng tài nguyên hệ thống một cách tối ưu.
- Quản lý gói: liên quan đến việc quản lý các gói cơ sở cho các thành phần hệ thống để duy trì tính bảo mật và độ tin cậy trong hoạt động của hệ thống.
- Quản lý tài nguyên: liên quan đến việc sử dụng tài nguyên trong hệ thống như thế nào để hỗ trợ khối lượng công việc khác nhau, do hệ thống hỗ trợ theo mức độ ưu tiên.

8.4 Quản trị dữ liệu

Quản trị dữ liệu là một thuộc tính hoặc một tính năng cần được phối hợp và thực hiện bởi tập các hoạt động của các vai trò và vai trò phụ trong góc nhìn người dùng để đảm bảo dữ liệu được sử dụng trong các quy trình nghiệp vụ tạo ra giá trị và đáp ứng hiệu quả các yêu cầu của nghiệp vụ.

Quản trị dữ liệu đưa ra và xác định:

- Chiến lược tổ chức liên quan đến việc quản lý dữ liệu để đảm bảo rằng dữ liệu phù hợp với hoạt động kinh doanh;
- Chiến lược quản lý chất lượng dữ liệu. Chiến lược này là một tập các ràng buộc và hành động nhằm đảm bảo dữ liệu đáp ứng được các yêu cầu chất lượng được xác định bởi nghiệp vụ (xem Phụ lục C để biết thêm chi tiết).

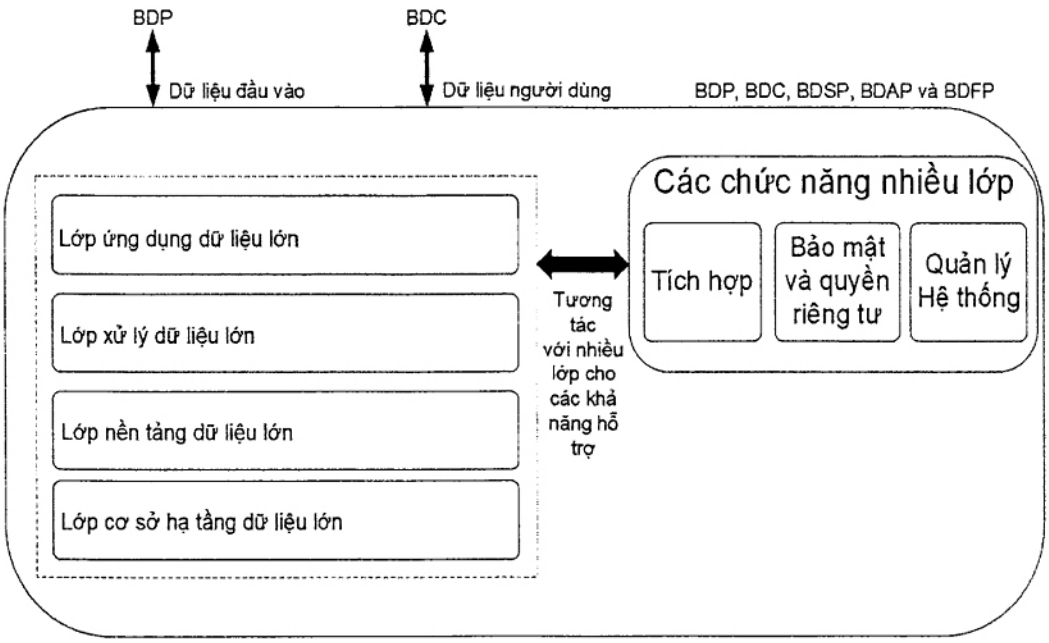
9 Góc nhìn chức năng

9.1 Kiến trúc chức năng

9.1.1 Khái quát chung

Kiến trúc chức năng cho dữ liệu lớn mô tả dữ liệu lớn dưới dạng tập hợp ở mức cao của các lớp thành phần chức năng. Các lớp chức năng đại diện cho tập các thành phần chức năng có tính năng tương tự để thực hiện các hoạt động dữ liệu lớn được mô tả trong mục 8 đối với các vai trò và vai trò phụ khác nhau liên quan đến dữ liệu lớn, thông số kỹ thuật và thực thi kiến trúc dữ liệu lớn.

Kiến trúc chức năng mô tả các thành phần chức năng dưới dạng kiến trúc phân lớp, trong đó các loại chức năng cụ thể được nhóm thành từng lớp như minh họa trong Hình 12.



Hình 11 – Kiến trúc dựa trên lớp BDRA

BDP và BDC được biểu diễn như trên có thể nằm ngoài hệ thống dữ liệu lớn đang được phát triển kiến trúc hoặc các thành phần bên trong (vì một nhà cung cấp ứng dụng trong kiến trúc dữ liệu lớn có thể cung cấp đầu vào hoặc sử dụng đầu vào từ nhà cung cấp ứng dụng khác trong kiến trúc). Phụ lục A cung cấp thông tin bổ sung về cách ánh xạ góc nhìn chức năng của kiến trúc tham chiếu dữ liệu lớn sang kiến trúc tham chiếu tích hợp hệ thống khác.

Các vai trò và hoạt động của góc nhìn người dùng trong dữ liệu lớn bao gồm: BDP, BDC, BDSP, BDAP và BDFP được thực hiện bởi bốn lớp chức năng và/hoặc các chức năng nhiều lớp như biểu diễn trong Hình 11. Với mục đích xác định một kiến trúc cụ thể, phương pháp tốt nhất được khuyến nghị là lập tài liệu kiến trúc về các thành phần chức năng cụ thể cung cấp giao diện từ các lớp đó tới kiến trúc dữ liệu lớn.

9.1.2 Kiến trúc phân lớp

9.1.2.1 Khái quát chung

Kiến trúc phân lớp được sử dụng trong BDRA có bốn lớp, cộng với một tập các chức năng trải dài trên các lớp. Bốn lớp là:

- Lớp ứng dụng dữ liệu lớn (xem 9.1.2.2);
- Lớp xử lý dữ liệu lớn (xem 9.1.2.3);
- Lớp nền tảng dữ liệu lớn (xem 9.1.2.4);
- Lớp hạ tầng dữ liệu lớn (xem 9.1.2.5).

Các chức năng trải dài trên các lớp được gọi là các chức năng nhiều lớp.

Kiến trúc phân lớp được biểu diễn trong Hình 11 và mỗi lớp bên trong của kiến trúc phân lớp được mô tả trong các mục 9.1.2.2 đến 9.1.2.5.

9.1.2.2 Lớp ứng dụng dữ liệu lớn

Lớp ứng dụng dữ liệu lớn cung cấp các chức năng hỗ trợ ứng dụng, bao gồm các chức năng thu thập, chuẩn bị, phân tích, hiển thị và truy cập dữ liệu lớn. Các chức năng này đạt được thông qua các giao diện với BDP, lớp xử lý dữ liệu lớn, lớp nền tảng dữ liệu lớn và BDC.

9.1.2.3 Lớp xử lý dữ liệu lớn

Lớp xử lý dữ liệu lớn cung cấp các thành phần khung và thư viện để thực hiện các phép phân tích được chỉ định bởi lớp nhà cung cấp ứng dụng. Trong lớp này, các thành phần quản lý thực hiện các tác vụ phân tích trên toàn hệ thống. Các thành phần thường tương tác với lớp nền tảng để xác định nơi lưu trữ dữ liệu trên hệ thống và hướng các phân tích cho dữ liệu đó đến nút tương ứng để cung cấp vị trí dữ liệu cho các tác vụ tính toán. Chúng cũng tương tác với các thành phần quản lý tài nguyên trong các chức năng nhiều lớp để cân bằng các phép tính toán trên toàn hệ thống.

9.1.2.4 Lớp nền tảng dữ liệu lớn

Lớp nền tảng dữ liệu lớn cung cấp các thành phần lưu trữ và tổ chức cho các dữ liệu do hệ thống xử lý. Các thành phần này thu thập tài nguyên từ lớp tài nguyên và trong trường hợp lưu trữ trên bộ nhớ, chúng phối hợp với các thành phần quản lý tài nguyên trong các chức năng nhiều lớp để thu thập tài nguyên cần thiết. Các thành phần của lớp nền tảng tập trung chủ yếu vào việc cung cấp phương thức tổ chức dữ liệu hiệu quả để nhà cung cấp ứng dụng có thể truy cập và xử lý các lớp trong hệ thống.

9.1.2.5 Lớp hạ tầng dữ liệu lớn

Lớp hạ tầng dữ liệu lớn là nơi tập trung các tài nguyên, bao gồm các thiết bị thường được sử dụng trong trung tâm dữ liệu như máy chủ, thiết bị chuyển mạch mạng và thiết bị định tuyến, thiết bị lưu trữ và cả phần mềm tương ứng dành cho dữ liệu lớn chạy trên máy chủ và các thiết bị khác như hệ điều hành máy chủ, phần mềm giám sát, trình điều khiển thiết bị và phần mềm quản lý hệ thống chung.

Lớp hạ tầng dữ liệu lớn cũng đại diện và chứa các chức năng của mạng truyền tải dữ liệu lớn được yêu cầu để cung cấp kết nối mạng cơ bản giữa nhà cung cấp ứng dụng dữ liệu lớn và BDP/BDC, cũng như nội bộ nhà cung cấp ứng dụng dữ liệu lớn và giữa các nhà cung cấp ứng dụng dữ liệu lớn với nhau.

9.1.3 Chức năng nhiều lớp

Các chức năng nhiều lớp bao gồm một loạt các thành phần chức năng tương tác với các thành phần chức năng của bốn lớp khác ở trên để cung cấp các khả năng hỗ trợ, bao gồm và không giới hạn:

- Khả năng bảo mật hệ thống (xác thực, ủy quyền, kiểm tra, xác nhận, mã hóa);
- Khả năng tích hợp (liên kết các thành phần khác nhau để đạt được chức năng cần thiết);
- Khả năng quản lý (triển khai, cấu hình, giám sát, tài nguyên đa khách hàng, tính khả dụng cao và vòng đời dữ liệu lớn).

Các chức năng nhiều lớp được mô tả ở trên có thể hỗ trợ các khía cạnh xuyên suốt hoặc các hoạt động từ các vai trò có khả năng ứng dụng rộng rãi của kiến trúc hệ thống.

9.2 Các thành phần chức năng

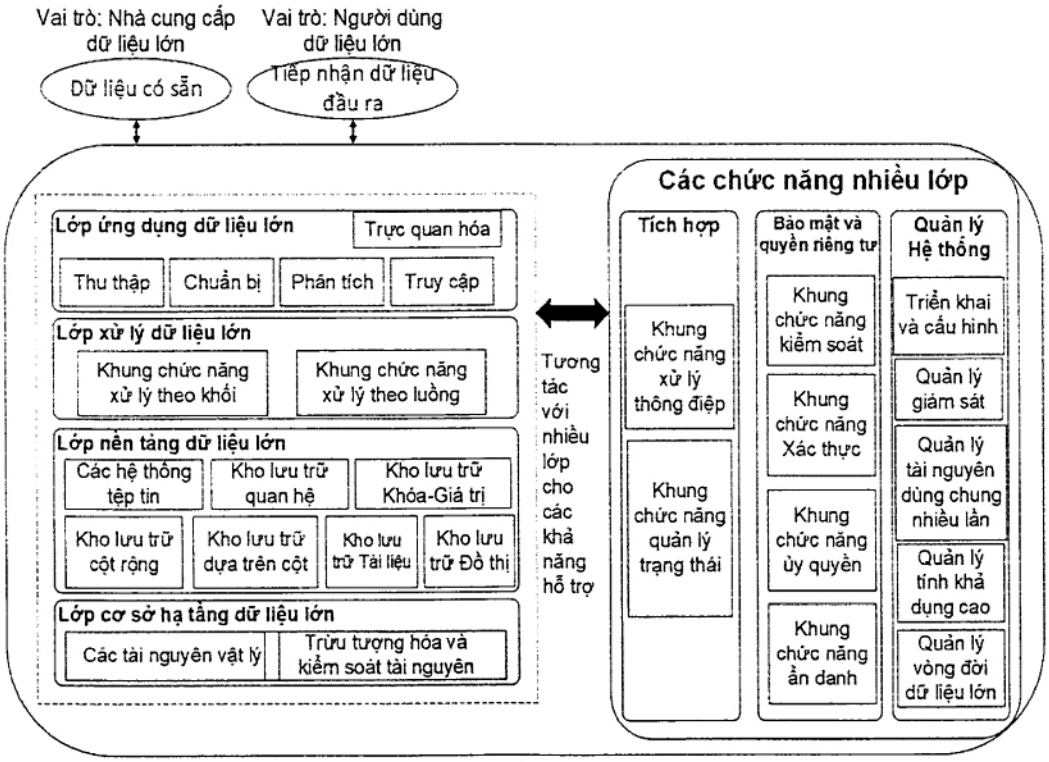
9.2.1 Khái quát chung

Khoản mục phụ này mô tả kiến trúc dữ liệu lớn dưới dạng tập hợp chung của các thành phần chức năng dữ liệu lớn. Thành phần chức năng là một phần tử chức năng của BDRA được sử dụng để thực hiện một hoạt động hoặc một số phần của hoạt động và có một cấu phần thực thi trong việc thực hiện một kiến trúc cụ thể, ví dụ: một thành phần phần mềm, một hệ thống con hoặc một ứng dụng.

Hình 12 mô tả tổng quan chung về các thành phần chức năng BDRA được tổ chức theo kiến trúc phân lớp.

Thuật ngữ Khung được sử dụng cho tên của các thành phần chức năng trong Hình 12 và các khoản mục văn bản liên quan được định nghĩa trong ISO/IEEE 11073-10201 như sau: "một cấu trúc của các quy trình và thông số kỹ thuật được thiết kế để hỗ trợ việc hoàn thành một nhiệm vụ cụ thể".

CHÚ THÍCH: Với phạm vi của các ứng dụng/các lĩnh vực liên quan đến dữ liệu lớn và sự phát triển nhanh chóng của công nghệ dữ liệu lớn, việc mô tả một danh sách đầy đủ các thành phần chức năng có thể có bên trong các lớp này là một khối lượng khổng lồ và có thể không bao giờ đủ. Do đó, nội dung này chỉ trình bày danh mục khái quát chung của các thành phần.



Hình 12 – Các thành phần chức năng của BDRA

9.2.2 Thành phần chức năng của lớp ứng dụng dữ liệu lớn

9.2.2.1 Khái quát chung

Lớp ứng dụng dữ liệu lớn với các thành phần chức năng hỗ trợ các hoạt động của nhà cung cấp ứng dụng dữ liệu lớn. Nó cung cấp giao diện chính cho các thành phần bên ngoài bao gồm các nhà cung cấp dữ liệu lớn và người dùng dữ liệu lớn. Các thành phần ở đây gọi các thành phần trong lớp xử lý dữ liệu lớn và lớp nền tảng dữ liệu lớn để thực thi các hoạt động của lớp ứng dụng dữ liệu lớn. Dưới đây là các thành phần chức năng chính trong lớp này.

9.2.2.2 Thành phần chức năng thu thập

Thành phần chức năng thu thập được sử dụng để thiết lập cơ chế nhập dữ liệu từ nhà cung cấp dữ liệu lớn và lưu trữ dữ liệu cho các quy trình tiếp theo:

- Thiết lập kết nối;
- Nhập dữ liệu;
- Lưu trữ dữ liệu.

Thành phần này liên quan đến việc đưa dữ liệu vào hệ thống. Các thành phần này có thể thực hiện hiệu quả các chức năng của chúng dựa trên khối lượng và tốc độ của dữ liệu đầu vào.

9.2.2.3 Thành phần chức năng chuẩn bị

Thành phần chức năng chuẩn bị được sử dụng để chuẩn bị dữ liệu phù hợp cho một quá trình phân tích cụ thể. Các chức năng chi tiết bao gồm: tổng hợp dữ liệu, làm sạch dữ liệu, chuyển đổi/biến đổi dữ liệu, tạo trường tính toán dữ liệu, tối ưu hóa dữ liệu, phân vùng dữ liệu, tóm tắt dữ liệu, căn chỉnh dữ liệu, xác thực dữ liệu, ảo hóa dữ liệu và lưu trữ dữ liệu đã chuẩn bị. Ảo hóa dữ liệu là một cách tiếp cận để quản lý dữ liệu mà ứng dụng có thể truy cập và thay đổi dữ liệu mà không cần biết đến định dạng vật lý và lưu trữ của dữ liệu. Biến đổi dữ liệu là sự thay đổi dữ liệu từ định dạng này sang định dạng khác, bao gồm: mã hóa/giải mã, nén/giải nén, phân rã, hoán đổi và chuẩn hóa dữ liệu.

9.2.2.4 Thành phần chức năng phân tích

Thành phần chức năng phân tích được sử dụng để đóng gói các phép toán chuyên dụng cần thực hiện trên dữ liệu để tìm kiếm thông tin và/hoặc khai thác kiến thức để đáp ứng các yêu cầu ứng dụng bằng sách sử dụng các thuật toán được chỉ định.

CHÚ THÍCH 1: Các lớp thuật toán của học máy bao gồm (không giới hạn): mối tương quan, phân loại, tổng hợp dữ liệu, tích hợp dữ liệu, khai thác dữ liệu, trí tuệ nhân tạo, nhận dạng mẫu, mô hình dự toán, hồi quy, phân tích cụm, phân tích không gian, phân tích âm thanh, phân tích hình ảnh, phân tích văn bản...Các thuật toán phân tích văn bản bao gồm phân tích cảm tính, nhận dạng thực thể được đặt tên và phát hiện chủ đề. Các thuật toán học máy bao gồm: tương quan, phân loại, nhận dạng mẫu, mô hình dự đoán, hồi quy, phân tích cụm và phân tích không gian. Trong nhiều trường hợp, các hệ thống dữ liệu lớn kết hợp một số loại thuật toán này thành một luồng công việc trên dữ liệu. Ví dụ: một hệ thống có thể sử dụng tính năng nhận dạng thực thể được đặt tên để trích xuất các thực thể cụ thể (người, địa điểm, tổ chức...) từ các đoạn văn bản rời rạc, sau đó cung cấp thông tin đó dưới dạng các tính năng vào một thuật toán K láng giềng gần nhất hoặc thuật toán phân cụm K-mean để phân loại các khối văn bản.

CHÚ THÍCH 2: Một lớp của chức năng phân tích là phân tích dữ liệu hoạt động, tức là phân tích các tệp nhật ký, dữ liệu trạng thái hệ thống, thông tin cảnh báo...để vận hành và bảo trì hệ thống. Đặc trưng truy vấn và phân tích điển hình bao gồm tìm kiếm tệp văn bản nhật ký, phân tích tổng hợp đa chiều...Các thuật toán phân tích số bao gồm: biến đổi fourier nhanh, đại số tuyến

tính và phương pháp N-Body. Thuật toán đồ thị bao gồm: phát hiện cộng đồng, tìm kiếm đồ thị con, tìm đường kính, hệ số phân cụm, xếp hạng trang, tập lớn nhất, thành phần được kết nối, độ trung tâm trung gian, đường dẫn ngắn nhất.

CHÚ THÍCH 3: Các đặc điểm quan trọng của các thuật toán này đối với dữ liệu lớn là chúng cần có khả năng hoạt động song song trong lớp xử lý và giải quyết đặc tính phân tán của dữ liệu trong lớp nền tảng.

9.2.2.5 Thành phần chức năng trực quan hóa

Thành phần chức năng trực quan hóa được sử dụng để trình bày dữ liệu lớn đã được phân tích cho người dùng dữ liệu lớn một cách có nghĩa. Các chức năng chi tiết bao gồm:

- Trực quan hóa dữ liệu khai phá (đa chiều, đa phân giải, tương tác, hoạt họa, mô phỏng, đồ họa thống kê, kết xuất bề mặt, kết xuất khối lượng);
- Trực quan hóa kiến thức/giải thích (trình bày tóm tắt các báo cáo và khách hàng).

CHÚ THÍCH: Các khía cạnh quan trọng của việc trực quan hóa dữ liệu lớn là trình bày các bộ dữ liệu lớn theo cách có thể dễ dàng vận hành và có thể hiểu được. Ngoài ra, nó có thể cần hoạt động trên dữ liệu theo hình thức song song phân tán.

9.2.2.6 Thành phần chức năng truy cập

Thành phần chức năng truy cập được sử dụng để cung cấp cho người dùng dữ liệu lớn quyền truy cập vào các kết quả của lớp ứng dụng dữ liệu lớn. Các chức năng chi tiết bao gồm:

- Quản lý quyền truy cập;
- Xuất dữ liệu (Ví dụ: thông qua giao diện lập trình ứng dụng, giao thức hoặc ngôn ngữ truy vấn);
- Truy cập dữ liệu an toàn.

CHÚ THÍCH: Người dùng dữ liệu lớn kết nối qua thành phần chức năng này bằng các dịch vụ web, giao diện người dùng và/hoặc các API, giao thức...được sử dụng để truy cập/trích xuất dữ liệu. Vấn đề duy nhất đối với dữ liệu lớn ở đây liên quan đến cách trình bày dữ liệu cho người dùng dữ liệu lớn khi tính tới thách thức lớn về các khía cạnh khối lượng và tốc độ.

9.2.3 Thành phần chức năng của lớp xử lý dữ liệu lớn

9.2.3.1 Khái quát chung

Các thành phần của lớp xử lý dữ liệu lớn chủ yếu tập trung vào hiệu suất (ví dụ như tạo ra kết quả tính toán trong một khoảng thời gian nhất định). Lớp xử lý dữ liệu lớn cung cấp các thành phần chức năng chủ yếu để hỗ trợ các đặc điểm của dữ liệu lớn về khối lượng, vận tốc và sự đa dạng. Lớp xử lý dữ liệu lớn thông qua các công cụ xử lý khác nhau trên các bộ lưu trữ dữ liệu khác nhau và tính toán theo lịch trình trên bộ lưu trữ gần hoặc cục bộ. Lớp này cung cấp các chức năng tóm tắt cho các hoạt động của lớp ứng dụng dữ liệu lớn. Hoạt động của người dùng được tóm lược dưới dạng nguồn dữ liệu, bộ lọc, bản đồ, cửa sổ, tổng hợp...Lớp xử lý dữ liệu lớn hoàn thành quá trình thực thi với dữ liệu truyền từ toán tử này sang toán tử khác và từ đầu vào đến đầu ra. Xử lý dữ liệu song song được thực hiện trong lớp này.

CHÚ THÍCH 1: Trong các hệ thống cơ sở dữ liệu truyền thống, các thành phần của lớp xử lý dữ liệu lớn được gọi là công cụ thực thi. Lớp xử lý dữ liệu lớn liên quan chủ yếu đến về thời gian thực thi. Từ khóa "lớn" không chỉ có ý nghĩa là dữ liệu lớn từ nguồn; trên thực tế, dữ liệu trung gian có thể lớn hơn dữ liệu thô.

CHÚ THÍCH 2: Trong các hoạt động song song, các thành phần của lớp xử lý thường phân bổ công việc cho các nút trong cụm đầu tiên dựa trên vị trí dữ liệu (ví dụ như dữ liệu trong lớp nền tảng cần thiết cho tính toán nằm trên nút) và sau đó là dựa trên tài nguyên bộ nhớ và CPU.

CHÚ THÍCH 3: Một ví dụ về điều này là mô hình lập trình ánh xạ/thu nhỏ, nơi việc tính toán trên các bản ghi riêng lẻ được phân phối đến các nút dựa trên vị trí dữ liệu trong giai đoạn ánh xạ và sau đó kết quả từ mỗi nút được hợp nhất và sắp xếp trong giai đoạn thu nhỏ.

Lớp xử lý dữ liệu lớn sử dụng các công cụ xử lý khác nhau trên các bộ lưu trữ dữ liệu khác nhau và lập lịch tính toán trên bộ lưu trữ gần hoặc cục bộ.

Thông thường, các khung trong lớp xử lý dữ liệu lớn được phân loại dựa trên số lượng thành phần và tốc độ xử lý của chúng. Các hình thức phân loại phổ biến là theo khối (lô) hoặc theo phần tử (luồng).

9.2.3.2 Thành phần chức năng của khung chức năng xử lý theo khối

Thành phần chức năng của khung chức năng xử lý theo khối chủ yếu nhằm giải quyết vấn đề về khối lượng. Chức năng này cần một khối các phần tử làm đơn vị cơ bản để xử lý. Các phần tử này bị khóa lại để tạo thành một khối dựa trên sự phân phối của chúng trong lớp nền tảng để xử lý nhằm tối đa hóa vị trí dữ liệu. Sau khi mỗi nút đã xử lý khối phần tử của nó, kết quả được chuyển tiếp đồng bộ hoặc không đồng bộ sang bước tiếp theo, có thể là một vòng xử lý khác (giống như được thực hiện theo mô hình song song đồng bộ số lượng lớn) hoặc tổng hợp các kết quả (giống như được thực hiện trong mô hình ánh xạ/thu nhỏ). Thời gian cần thiết để hoàn thành việc phân tích khối có thể thay đổi từ vài giờ đến vài giây tùy thuộc vào việc phân tích và đặc tính dữ liệu. Các truy vấn đặc biệt và các ứng dụng báo cáo phân tích hoạt động hàng ngày có thể cần thời gian phản hồi khác nhau. Nếu thời gian phản hồi trong vòng vài phút, vài giờ hoặc lâu hơn, quy trình này được gọi là xử lý ngoại tuyến. Nếu thời gian phản hồi nằm trong khoảng vài giây hoặc dưới một giây, nó được gọi là xử lý tương tác. Tuy nhiên, khi một hệ thống được thiết kế để tương tác, điều đó không có nghĩa là tất cả thời gian phản hồi đều nằm trong phạm vi vài giây hoặc dưới một giây. Một phân tích/truy vấn được viết không tốt, một yêu cầu có các kết nối phức tạp giữa các dữ liệu hoặc một truy vấn phải xử lý một khối lượng bản ghi lớn có thể mất vài phút hoặc vài giờ để hoàn thành.

9.2.3.3 Thành phần chức năng của khung chức năng xử lý theo luồng (Streaming)

9.2.3.3.1 Khái quát chung

Thành phần chức năng của khung chức năng xử lý theo luồng chủ yếu nhằm giải quyết vấn đề tốc độ. Mô hình quy trình là dạng ống và mọi phần tử được chuyển tiếp đến vị trí xử lý tiếp theo với độ trễ tối thiểu. Phản hồi tức thì là mối quan tâm chính và mọi phần tử đều có giá trị trong thời điểm này. Trong khi đó, một số hoạt động yêu cầu các phần tử bị khóa lại hoặc được lưu vào bộ đệm. Tuy nhiên, trong một tình huống lý tưởng, dữ liệu truyền liên tục qua ống. Thành phần chức năng khung xử lý thông điệp (xem 9.2.6.2.2) được sử dụng để giao tiếp giữa các toán tử qua các nút. Khi dữ liệu quá lớn và/hoặc quá nhanh mà hệ thống không theo kịp, hệ thống có thể sử dụng bộ nhớ tạm thời, chọn cách loại bỏ dữ liệu dư thừa hoặc buộc phải sử dụng cơ chế giới hạn tốc độ với nhà sản xuất để tránh sự cố xảy ra với hệ thống.

Đặc trưng cơ bản của khung chức năng xử lý theo luồng là luồng dữ liệu. Luồng dữ liệu bên trong là một đồ thị xoay chiều có hướng chứa toán tử là đỉnh và luồng sự kiện là cạnh. Toán tử có thể được song song hóa và luồng sự kiện có thể được phân vùng. Xử lý sự kiện phức tạp (CEP) tiên tiến hơn so với

luồng thuần túy và có thể truy vấn được, điều này bổ sung thêm các đặc điểm thực tế hơn: sắp xếp thứ tự sự kiện, đảm bảo xử lý sự kiện, trạng thái lưu trữ và phân vùng luồng/toán tử song song.

Bốn đặc điểm được mô tả trong các mục từ 9.2.3.3.2 đến 9.2.3.3.5.

9.2.3.3.2 Sắp xếp thứ tự sự kiện

Sắp xếp thứ tự sự kiện được đảm bảo bởi mốc thời gian toàn cục tùy chỉnh hoặc chuỗi ID, cả hai đều được đánh dấu bởi bộ cấp dữ liệu. Thứ tự sự kiện có thể được xử lý theo thời gian hoặc số lượng. Thứ tự sự kiện có liên quan đến cửa sổ luồng. Khi thời gian sự kiện được sử dụng, sắp xếp sự kiện có nghĩa là sự kiện phải được xem xét, đánh giá trong toán tử cửa sổ theo thứ tự của dấu thời gian. Các sự kiện không theo thứ tự và bị trì hoãn nên được sắp xếp lại, loại bỏ hoặc đánh giá ngay lập tức. Khi bộ đếm sự kiện được sử dụng, sắp xếp sự kiện có nghĩa là sự kiện phải được xem xét, đánh giá trong toán tử cửa sổ trong dãy ID. Thời gian sự kiện hoặc chuỗi ID cần phải tăng dần đều.

9.2.3.3.3 Đảm bảo xử lý sự kiện

Các sự kiện phải được xử lý với cơ chế chịu lỗi khi xảy ra sự cố. Đặc biệt, khi luồng được phân vùng, toán tử được song song và dữ liệu được phân phối. Dữ liệu được lưu trữ trong bộ nhớ và dữ liệu được lưu trữ liên tục trong hệ thống tệp phải được đảm bảo trong khoảng thời gian cửa sổ. Hai giai đoạn quan trọng cần được chú ý đặc biệt là tiếp nhận trước khi xử lý (Receiver) và cam kết sau khi xử lý (Processor). Đảm bảo xử lý sự kiện thường được chia thành ba lớp sau:

- Tối đa một lần: ý nghĩa của lớp này là giai đoạn tiếp nhận sẽ nhận một lần từ nguồn dữ liệu và không cần duy trì phần bù đã nhận, và giai đoạn xử lý không được bảo đảm. Sự kiện đã nhận có thể nhận được nhưng không có kết quả trả về. Việc này đơn giản và có độ trễ thấp, nhưng tính chính xác không được đảm bảo.
- Ít nhất một lần: ý nghĩa của lớp này là giai đoạn tiếp nhận có thể lặp lại, tiếp nhận một sự kiện nhiều lần và giai đoạn xử lý có thể xử lý các sự kiện lặp đi lặp lại. Tất cả các sự kiện có thể được tiếp nhận và xử lý, nhưng kết quả có thể không chính xác. Việc bổ sung cơ chế bảo trì phần bù thủ công cần được hỗ trợ để đáp ứng việc thực hiện lại sự kiện và cơ chế nhân bản có thể được hỗ trợ để làm giảm bớt việc xử lý lặp lại. Việc này làm tăng thêm chi phí nhưng có thể đạt được độ trễ thấp và mức độ đảm bảo nhất định.
- Chính xác một lần: sự kiện được nhận và xử lý một lần, không bị mất và không thực hiện lại. Cả hai giai đoạn tiếp nhận và xử lý đều được đảm bảo. Cả hai giai đoạn đều cần có khả năng chịu lỗi độc lập và cơ chế khôi phục hư hỏng để tạo thành kho lưu trữ nguyên tử và bền vững.

9.2.3.3.4 Trạng thái lưu trữ

Các khung chức năng xử lý theo luồng đặc trưng có mô hình quy trình dạng ống, khi CEP dựa trên các khung chức năng xử lý luồng cần thêm các điều kiện bổ sung để hỗ trợ hoạt động xử lý cửa sổ cho các truy vấn liên tục. Trong kỹ thuật này, các sự kiện được lưu trữ trong một khoảng thời gian để tạo ra cửa sổ. Trong CEP truyền thống, kích thước cửa sổ thường nhỏ và các sự kiện được lưu trữ trong một bộ đệm. Trong CEP hiện đại áp dụng cho dữ liệu lớn, số lượng các sự kiện cần xử lý bằng phương pháp cửa sổ có thể rất lớn, do đó việc lưu trữ trạng thái có thể hỗ trợ cho các luồng có lưu lượng lớn. Bộ nhớ

bổ sung là cần thiết để đáp ứng khả năng chịu lỗi và khôi phục hư hỏng; sao chép, ghi nhật ký (WAL) và điểm kiểm tra là các phương pháp truyền thống để giải quyết vấn đề này, do vậy lưu trữ trạng thái có thể hỗ trợ vấn đề phân tán và ACID ở một mức độ nào đó nhằm đảm bảo sự cân bằng giữa hiệu suất và độ chính xác.

9.2.3.3.5 Phân vùng luồng/toán tử song song

Đặc tính này liên quan đến khả năng mở rộng. Luồng và toán tử được thực hiện trong đồ thị xoay chiều có hướng. Mục tiêu của các khung chức năng xử lý luồng là song song với việc thực thi ở mức độ lớn nhất. Phân vùng luồng phục vụ phân phối sự kiện và song song toán tử phục vụ tính toán song song. Bộ lập lịch biểu thực hiện tính toán song song với các sự kiện cục bộ. Phân vùng luồng theo khóa (như ID cảm biến, ID người dùng, ID tài khoản) và chức năng tổng hợp được đánh giá biệt lập trên luồng đã được phân vùng. Siêu dữ liệu luồng, điều phối giao tiếp, phân bổ tài nguyên động và chiến lược tìm nạp kéo/đẩy là cần thiết để giúp phân vùng luồng trong môi trường phân tán. Tính song song của toán tử là cần thiết để đáp ứng yêu cầu tính toán nhanh, nhưng cần có thêm cơ chế để giúp điều phối trạng thái toàn cục (rào chắn, thứ tự sự kiện). Các toán tử thường là lần lượt, một số toán tử có thể áp dụng cho chuỗi để giảm chi phí giao tiếp mạng.

9.2.4 Thành phần chức năng của lớp nền tảng dữ liệu lớn

9.2.4.1 Khái quát chung

Các thành phần của lớp nền tảng dữ liệu lớn cung cấp dịch vụ lưu trữ, tổ chức và truy xuất dữ liệu để hỗ trợ các lớp cao hơn. Theo đó, lớp này cung cấp tổ chức và phân phối dữ liệu logic kết hợp với các phương thức hoặc giao diện lập trình ứng dụng truy cập liên kết (APIs). Điều này cũng có thể bao gồm đăng ký dữ liệu và các dịch vụ siêu dữ liệu cùng với các mô tả dữ liệu ngữ nghĩa ví dụ như bản thể chính thức hoặc phân loại.

CHÚ THÍCH: Một khía cạnh quan trọng khi xây dựng lớp này là lựa chọn hoặc cải tiến tổ chức dữ liệu và các phương thức lưu trữ để nâng cao độ khả dụng dữ liệu và hiệu suất truy vấn hoặc truy xuất dữ liệu. Đặc biệt là với sự gia tăng nhanh chóng về khối lượng của dữ liệu lớn (như: tài chính, ngân hàng, truyền thông, công nghiệp sản xuất) và các kịch bản dịch vụ, người dùng yêu cầu nâng cao hiệu suất truy vấn và phân tích khác nhau bằng cách giảm sự trùng lặp và dư thừa trong lưu trữ dữ liệu.

Các mục phụ từ 9.2.4.2 đến 9.2.4.8 mô tả các danh mục chung của các thành phần này.

9.2.4.2 Thành phần chức năng các hệ thống tệp

Hệ thống tệp tổ chức các khối dữ liệu (thường được định nghĩa là bản ghi) được truy cập như một thực thể được đặt tên trong một không gian tên xác định. Trong khi hệ thống tệp cục bộ thường được sử dụng trong các hệ thống dữ liệu lớn để lưu trữ dữ liệu trung gian cục bộ cho một nút xử lý, thì các hệ thống tệp phân tán lại phổ biến hơn nhiều để lưu trữ dữ liệu liên tục. Sự khác biệt là các hệ thống tệp phân tán quản lý việc phân phối và nhân bản các khối dữ liệu thông qua các nút và không gian tên thay vì được lưu trữ cùng với dữ liệu được quản lý thông qua một dịch vụ tên trung tâm thường chạy theo cách thức chủ/tớ hoặc cách thức đa chủ để cung cấp khả năng chịu lỗi.

Các hệ thống tệp phân tán (còn được gọi là các hệ thống tệp cụm) nhằm giải quyết các vấn đề về lưu lượng do đặc điểm về khối lượng và tốc độ của dữ liệu lớn, kết hợp lưu lượng vào/ra trên nhiều thiết bị

(trục chính) tại mỗi nút, với khả năng điều phối phần dư thừa và chuyển đổi dự phòng hoặc nhân bản dữ liệu ở mức khối trên nhiều nút. Việc nhân bản dữ liệu của một hệ thống tệp phân tán được thiết kế đặc biệt cho phép sử dụng phần cứng bán sẵn không đồng nhất trên cụm dữ liệu lớn. Do đó, nếu một ổ đĩa đơn hoặc toàn bộ nút bị lỗi, dữ liệu sẽ không bị mất vì nó đã được sao chép trên các nút khác và lưu lượng ít bị ảnh hưởng nhất vì quá trình xử lý đó có thể được chuyển đến các nút khác. Ngoài ra, tính năng nhân bản có khả năng đọc dữ liệu và ghi lần đầu đồng thời ở mức cao.

Kho đối tượng phân tán (DOS) (còn được gọi là kho đối tượng toàn cục) là ví dụ tiêu biểu về tổ chức hệ thống tệp phân tán. Không giống như các phương pháp tiếp cận được mô tả ở trên, nơi sử dụng phương pháp tiếp cận không gian tên phân cấp trên hệ thống tệp truyền thống, DOS cung cấp một không gian tên phẳng với mã định danh duy nhất trên toàn cục (GUID) cho bất kỳ đoạn dữ liệu nào. Nói chung, dữ liệu lưu trữ được định vị thông qua một truy vấn dựa trên danh mục siêu dữ liệu trả về các GUID được liên kết. GUID thường triển khai phần mềm cơ bản cùng với vị trí lưu trữ của dữ liệu cần quan tâm. Các kho lưu trữ đối tượng này được phát triển và giới thiệu để lưu trữ các đối tượng dữ liệu rất lớn, từ các tập dữ liệu hoàn chỉnh đến các đối tượng riêng lẻ lớn (như hình ảnh có độ phân giải cao trong phạm vi kích thước hàng chục gigabyte [GB]).

9.2.4.3 Thành phần chức năng lưu trữ quan hệ

Trong mô hình lưu trữ quan hệ, dữ liệu được lưu trữ dưới dạng các hàng với mỗi trường đại diện cho một cột được tổ chức thành một bảng dựa trên tổ chức dữ liệu logic.

CHÚ THÍCH: Việc triển khai các mô hình lưu trữ quan hệ dữ liệu lớn đã tương đối hoàn thiện và được một số tổ chức áp dụng. Các công cụ này đang phát triển rất nhanh chóng trong việc tập trung cải thiện thời gian phản hồi. Nhiều phương thức triển khai dữ liệu lớn cải tiến mạnh mẽ để mở rộng các truy vấn quan hệ. Về cơ bản, các truy vấn được chia thành các giai đoạn nhưng quan trọng hơn là việc xử lý các bảng đầu vào được phân phối trên nhiều nút (thường dưới dạng tác vụ ánh xạ/thu nhỏ).

Nơi lưu trữ dữ liệu thực tế có thể là các tệp phẳng (được phân cách hoặc có độ dài cố định) trong đó mỗi bản ghi/dòng trong tập tin đại diện cho một hàng trong bảng. Phương pháp này ngày càng áp dụng nhiều định dạng lưu trữ nhị phân được tối ưu hóa cho các hệ thống tập tin phân tán. Những định dạng này thường sử dụng chỉ mục ở mức khối và tổ chức hướng theo cột của dữ liệu để cho phép truy cập vào các trường riêng lẻ trong bản ghi mà không cần đọc toàn bộ bản ghi. Mặc dù vậy, hầu hết các mô hình lưu trữ quan hệ dữ liệu lớn vẫn là các hệ thống theo dạng khối được thiết kế cho các truy vấn rất phức tạp tạo ra ma trận tích hữu hướng trung gian rất lớn từ các phép nối, vì vậy ngay cả truy vấn đơn giản nhất cũng có thể mất hàng chục giây để hoàn thành.

9.2.4.4 Thành phần chức năng lưu trữ Khóa-Giá trị

Các nguyên tắc lưu trữ Khóa-Giá trị là nền tảng cho tất cả các mô hình lưu trữ và lập chỉ mục khác. Từ góc độ dữ liệu lớn, các kho lưu trữ này thể hiện một cách hiệu quả các mô hình bộ nhớ truy cập ngẫu nhiên. Mặc dù dữ liệu được lưu trữ trong các giá trị có thể phức tạp tùy ý về cấu trúc, nhưng việc xử lý độ phức tạp đó phải được cung cấp bởi các ứng dụng thực hiện lưu trữ mà các công cụ này thường chỉ cung cấp một điểm chỉ dẫn tới một khối dữ liệu. Kho lưu trữ Khóa-Giá trị cũng có xu hướng hoạt động tốt nhất cho mối quan hệ 1-1 (ví dụ: mỗi khóa liên quan đến một giá trị duy nhất) nhưng cũng có thể tác động để ánh xạ khóa tới danh sách các giá trị đồng nhất. Khi các khóa ánh xạ nhiều giá trị của các

kiểu/cấu trúc không đồng nhất hoặc khi các giá trị từ một khóa cần được ghép nối với các giá trị cho một khóa khác hoặc cùng một khóa thì cần phải có logic ứng dụng tùy chỉnh. Yêu cầu đối với logic tùy chỉnh này thường là ngăn các kho lưu trữ Khóa-Giá trị mở rộng quy mô hiệu quả đối với một số vấn đề nhất định.

Kho lưu trữ Khóa-Giá trị thường đáp ứng tốt với các bản cập nhật khi ánh xạ là 1-1 và giá trị về kích thước/độ dài của dữ liệu không thay đổi. Khả năng xử lý các bản chèn vào của kho lưu trữ Khóa-Giá trị thường phụ thuộc vào thao tác thực hiện cơ bản. Kho lưu trữ Khóa-Giá trị nói chung cũng đòi hỏi những nỗ lực đáng kể (cả về phương pháp thủ công hoặc tính toán) để đáp ứng những thay đổi đối với cấu trúc dữ liệu cơ bản của các giá trị. Kho lưu trữ Khóa-Giá trị phân tán là cách thực thi thường xuyên nhất được sử dụng trong các ứng dụng dữ liệu lớn. Một vấn đề sẽ luôn cần được giải quyết (nhưng không phải duy nhất để triển khai Khóa-Giá trị) là việc phân phối khóa trên không gian của các Khóa-Giá trị.

Đặc biệt, các khóa phải được lựa chọn một cách cẩn thận để tránh sai lệch trong phân phối dữ liệu trên toàn cụm. Khi dữ liệu bị sai lệch trong một phạm vi nhỏ, nó có thể dẫn đến các điểm nóng về tính toán trên toàn cụm nếu quá trình thực thi đang cố gắng tối ưu hóa vị trí dữ liệu. Nếu dữ liệu là động (các khóa mới đang được thêm vào) cho quá trình này, thì tại một thời điểm nào đó, dữ liệu có thể yêu cầu tái cân bằng trên toàn cụm. Việc thực hiện tối ưu hóa phi cục bộ sử dụng nhiều phương pháp tiếp cận khác nhau như băm, ngẫu nhiên, hoặc vòng lặp để phân phối dữ liệu và không có chiều hướng bị sai lệch và xuất hiện điểm nóng. Tuy nhiên, các công cụ này vận hành không ổn định khi xử lý các vấn đề liên quan đến yêu cầu tổng hợp trên tập dữ liệu.

9.2.4.5 Thành phần chức năng lưu trữ cột rộng

Trong khi dữ liệu quan hệ truyền thống lưu trữ dữ liệu theo các hàng giá trị liên quan, lưu trữ dạng cột tổ chức dữ liệu theo các nhóm giá trị tương đồng. Hai hình thức này chỉ khác biệt đôi chút, ở chỗ, trong cơ sở dữ liệu quan hệ thì toàn bộ nhóm các cột được gắn với một vài khóa chính (thường là một hoặc nhiều cột) để tạo bản ghi. Trong lưu trữ dạng cột, giá trị của mỗi cột là một khóa và các giá trị cột tương tự trỏ đến các hàng được liên kết. Trường hợp đơn giản nhất của lưu trữ dạng cột là lưu trữ nhiều hơn một Khóa-Giá trị với các vai trò của khóa và giá trị được đảo ngược. Theo nhiều cách, lưu trữ dữ liệu dạng cột rất giống với các chỉ mục trong cơ sở dữ liệu quan hệ. Ngoài ra, việc thực hiện các lưu trữ dạng cột rộng theo mô hình bản đồ được sắp xếp đa chiều rải rác, phân tán (nơi các mảng byte ngẫu nhiên được định danh/tiếp cận dựa trên các khóa dòng và cột) đưa ra một mức phân đoạn bổ sung bên ngoài bảng, hàng và cột của mô hình quan hệ, và được gọi là họ cột. Các kho cột rộng bổ sung thêm một nhân tố cũng được gọi là họ cột.

9.2.4.6 Thành phần chức năng lưu trữ dựa trên cột

Bằng cách tổ chức và lưu trữ dữ liệu theo cột (thay vì theo hàng trong lưu trữ dựa trên hàng), cơ sở dữ liệu cột rất phù hợp cho các ứng dụng dữ liệu lớn đòi hỏi nhiều phân tích phổ rộng, chẳng hạn như truy vấn OLAP (Online Analytic processing) đa chiều, truy vấn quét lớn và nhỏ. Các kỹ thuật sắp xếp, lập chỉ mục và nén dựa trên cột khác nhau (như lập chỉ mục đa chiều, mã hóa từ điển...) có thể được áp dụng để tăng hiệu suất truy vấn.

9.2.4.7 Thành phần chức năng lưu trữ tài liệu

Các kho lưu trữ tài liệu ngày nay đã phát triển để đưa vào khả năng tìm kiếm và lập chỉ mục mở rộng cho dữ liệu có cấu trúc và siêu dữ liệu, đó là lý do tại sao chúng thường được gọi là kho lưu trữ dữ liệu bán cấu trúc. Trong kho dữ liệu hướng tài liệu, mỗi tài liệu sẽ đóng gói và mã hóa siêu dữ liệu, các trường và bất kỳ bản trình bày nào khác của bản ghi đó. Mặc dù khá tương đồng với một hàng trong bảng quan hệ, nhưng một lý do khiến các kho lưu trữ tài liệu đã phát triển và trở nên phổ biến là hầu hết các cách thực hiện không bắt buộc một lược đồ cố định hoặc không đổi. Mặc dù các phương pháp tốt nhất cho thấy các nhóm tài liệu phải liên quan với nhau về mặt logic và chứa dữ liệu tương tự nhau, nhưng không có yêu cầu nào rằng chúng phải giống nhau hoặc thậm chí hai tài liệu bất kỳ phải chứa các trường giống nhau. Đó là một lý do mà các kho lưu trữ tài liệu thường phổ biến với các tập dữ liệu có các trường dữ liệu thưa thớt, vì thông thường sẽ tốn ít chi phí hơn so với các hệ thống RDBMS truyền thống, nơi mà các cột giá trị rỗng thực được lưu trữ. Các nhóm tài liệu trong kho lưu trữ dạng này thường được gọi là các bộ sưu tập và giống như Khóa-Giá trị lưu trữ một số loại khóa tham chiếu duy nhất cho mỗi tài liệu.

9.2.4.8 Thành phần chức năng lưu trữ đồ thị

Mặc dù các trang mạng truyền thông xã hội đã thúc đẩy khả năng hiển thị và phát triển của các kho lưu trữ đồ thị (quá trình xử lý được thảo luận bên dưới), các kho lưu trữ đã trở thành một phần quan trọng trong nhiều lĩnh vực từ tình báo quân sự và chống khủng bố cho đến lập kế hoạch/điều hướng tuyến đường và web ngữ nghĩa trong thời gian dài. Các kho lưu trữ đồ thị biểu diễn dữ liệu dưới dạng một loạt các nút, các cạnh và các thuộc tính trên đó. Việc phân tích dựa trên kho lưu trữ đồ thị bao gồm đường dẫn ngắn nhất và phân trang để định hướng thực thể và đối sánh đồ thị.

Các cách tiếp cận lưu trữ đồ thị có thể được xem như một cách thực thi chuyên biệt của một sơ đồ lưu trữ tài liệu với hai dạng tài liệu (các nút và các mối quan hệ). Ngoài ra, một trong những yếu tố quan trọng trong việc phân tích dữ liệu đồ thị là xác định vị trí của nút hoặc cạnh trong đồ thị khi bắt đầu phân tích. Để làm được điều này, hầu hết các cơ sở dữ liệu đồ thị thực hiện lập các chỉ mục trên các thuộc tính của nút hoặc cạnh. Không giống như các cách tiếp cận lưu trữ dữ liệu quan hệ và lưu trữ dữ liệu khác, hầu hết các cơ sở dữ liệu đồ thị có xu hướng sử dụng các khóa nhân tạo/khóa giả hoặc các hướng dẫn để xác định tính duy nhất của các nút và cạnh. Điều này làm cho các đặc tính/thuộc tính có thể dễ dàng bị thay đổi bởi cả những thay đổi thực tế trong dữ liệu (người nào đó thay đổi tên của họ) hoặc khi có thêm thông tin (như vị trí tốt hơn cho một số mục dữ liệu hoặc sự kiện) mà không cần thay đổi các chỉ dẫn đến/từ các mối quan hệ.

Thông thường, các kiến trúc phân tán để xử lý đồ thị gán các phần của đồ thị cho các nút trong hệ thống, sau đó các nút này sử dụng các phương pháp truyền tin để truyền đạt các thay đổi trong đồ thị hoặc giá trị của các phép tính toán theo một đường dẫn. Ngay cả những đồ thị nhỏ cũng nhanh chóng được nâng lên thành lĩnh vực của dữ liệu lớn khi một người đang tìm kiếm các mẫu hoặc khoảng cách nhiều hơn một hoặc hai mức phân cách giữa các nút của đồ thị.

Tùy thuộc vào mật độ của đồ thị, điều này có thể nhanh chóng gây ra sự bùng nổ tổ hợp về số lượng các điều kiện/mẫu cần được kiểm tra. Việc triển khai chuyên biệt một kho lưu trữ đồ thị được gọi là khung

mô tả tài nguyên (RDF), là một phần của họ các thông số kỹ thuật từ World Wide Web Consortium (W3C) thường được liên kết trực tiếp với web ngữ nghĩa và các khái niệm liên quan. Bộ ba RDF, bao gồm một chủ thể (Mr X), một vị ngữ (Live at) và một đối tượng (Mockingbird Lane). Do đó, tập hợp bộ ba RDF biểu diễn một đồ thị gắn nhãn có hướng. Nội dung các kho lưu trữ RDF thường được mô tả bằng cách sử dụng các ngôn ngữ bản thể học chính thức như OWL hoặc ngôn ngữ lược đồ RDF (RDFS), ngôn ngữ này thiết lập các ý nghĩa và mô hình ngữ nghĩa của dữ liệu cơ bản. Để hỗ trợ tích hợp theo chiều ngang tốt hơn (Smith, et al., 2012) [16] các phần mở rộng của tập dữ liệu không đồng nhất đối với khái niệm RDF như khung mô tả dữ liệu (RDF) (Yoakum-Stover & Malyuta, 2008) [17] đã được đề xuất bổ sung thêm để hỗ trợ tốt hơn khả năng tương thích và phân tích ngữ nghĩa. Các kho lưu trữ dữ liệu đồ thị hiện tại đang thiếu nhiều APIs hoặc ngôn ngữ truy vấn được chuẩn hóa. Tuy nhiên, W3C đã phát triển ngôn ngữ truy vấn SPARQL cho RDF, hiện tại đang ở trạng thái khuyến nghị và một số hệ thống như Sesame đang trở nên phổ biến để làm việc với RDF và các kho lưu trữ dữ liệu hướng đồ thị khác.

9.2.5 Thành phần chức năng của lớp tài nguyên

9.2.5.1 Khái quát chung

Các thành phần chức năng của lớp tài nguyên bao gồm:

- Trừu tượng hóa và kiểm soát tài nguyên;
- Các tài nguyên vật lý.

9.2.5.2 Thành phần chức năng kiểm soát và trừu tượng hóa tài nguyên

Thành phần chức năng kiểm soát và trừu tượng hóa tài nguyên được các nhà cung cấp ứng dụng dữ liệu lớn (BDAP) sử dụng để cung cấp quyền truy cập vào các tài nguyên tính toán vật lý thông qua việc trừu tượng hóa phần mềm. Việc trừu tượng hóa tài nguyên cần đảm bảo sử dụng hiệu quả, an toàn và đáng tin cậy đối với cơ sở hạ tầng bên dưới. Tính năng kiểm soát của thành phần chức năng cho phép quản lý các tính năng trừu tượng hóa tài nguyên.

CHÚ THÍCH 1: Khi hệ thống dữ liệu lớn được triển khai trong môi trường điện toán đám mây, các chức năng trừu tượng hóa tài nguyên được cung cấp bởi môi trường điện toán đám mây như được định nghĩa trong ISO/IEC 17789 [6].

Thành phần chức năng kiểm soát và trừu tượng hóa tài nguyên cho phép các nhà cung cấp ứng dụng dữ liệu lớn (BDAP) cung cấp các đặc tính như khả năng co giãn, tổng hợp tài nguyên và tự phục vụ theo yêu cầu. Thành phần chức năng kiểm soát và trừu tượng hóa tài nguyên có thể bao gồm các yếu tố phần mềm như giám sát, các máy ảo, lưu trữ dữ liệu ảo và chia sẻ thời gian.

Đối với mạng, đây là các tài nguyên truyền dữ liệu từ thành phần này sang thành phần khác trong lớp cơ sở hạ tầng. Bên cạnh đó, cơ sở hạ tầng mạng cũng có thể bao gồm việc triển khai tự động, khả năng cung cấp hoặc các tác vụ và tác vụ giám sát diện rộng trên cơ sở hạ tầng được tận dụng bởi các yếu tố quản lý/giao tiếp để triển khai một mô hình cụ thể.

Đối với điện toán, sự phân bổ hợp lý của cơ sở hạ tầng cụm/điện toán có thể thay đổi từ một kết cấu dầy đặc các máy vật lý trong tủ rack thành một tập các máy ảo chạy trên một nhà cung cấp dịch vụ đám mây hoặc một tập các máy được kết nối thiếu chặt chẽ được phân bố trên toàn cầu cung cấp truyền truy cập vào các tài chuyên máy tính chưa sử dụng.

CHÚ THÍCH 2: Hypervisor là một bộ phận của phần mềm, phần đệm hoặc phần cứng máy tính để tạo và chạy các máy ảo. với hình thức này, một hypervisor chạy trực tiếp trên phần cứng máy tính và quản lý nhiều máy ảo bao gồm các hệ điều hành (OS) và các ứng dụng.

9.2.5.3 Thành phần chức năng các tài nguyên vật lý

Thành phần chức năng tài nguyên vật lý biểu thị cho các yếu tố mà nhà cung cấp ứng dụng dữ liệu lớn cần để chạy và quản lý hệ thống dữ liệu lớn mà họ cung cấp.

Tài nguyên vật lý bao gồm tài nguyên phần cứng như các máy tính (CPU và bộ nhớ), các thiết bị mạng (bộ định tuyến, tường lửa, thiết bị chuyển mạch, thiết bị kết nối mạng, thiết bị đấu nối mạng), các thành phần lưu trữ (đĩa cứng) và các yếu tố cơ sở hạ tầng tính toán vật lý khác. Những tài nguyên này có thể bao gồm những thành phần bên trong trung tâm dữ liệu đám mây (như máy chủ điện toán, máy chủ lưu trữ và các mạng trung tâm dữ liệu nội bộ), và bên ngoài các trung tâm dữ liệu, điển hình là các tài nguyên mạng như các mạng liên trung tâm dữ liệu và các mạng truyền tải lõi.

Đối với mạng, các đặc điểm về khối lượng và tốc độ của dữ liệu lớn thường là những yếu tố thúc đẩy việc thực hiện các kết nối bên trong và bên ngoài của cơ sở hạ tầng mạng.

Đối với máy tính, đây là các máy chủ vật lý thực thi và lưu giữ phần mềm của các thành phần hệ thống dữ liệu lớn khác. Cơ sở hạ tầng máy tính cũng thường bao gồm các hệ điều hành cơ bản và các dịch vụ liên quan được dùng để kết nối các tài nguyên cụm với nhau thông qua các phần tử mạng.

Đối với lưu trữ, đây là những tài nguyên cung cấp sự ổn định của dữ liệu trong một hệ thống dữ liệu lớn. Cơ sở hạ tầng lưu trữ có thể bao gồm bất kỳ tài nguyên nào, từ các đĩa cục bộ biệt lập đến các hệ thống mạng lưu trữ vùng (SAN) hoặc hệ thống mạng lưu trữ gắn liền (NAS).

Đây là những tài nguyên vật lý của nhà máy/xưởng (nguồn điện, làm mát) cần được tính đến khi thiết lập một bản thể của hệ thống dữ liệu lớn. Trong khi các thành phần tài nguyên có thể được triển khai trực tiếp trên các tài nguyên vật lý hoặc các tài nguyên ảo thì ở một mức độ nào đó, tất cả các tài nguyên đều có sự hiện diện vật lý. Các tài nguyên vật lý thường được sử dụng để triển khai nhiều thành phần được lặp lại trên một số lượng lớn các nút vật lý để cung cấp khả năng mở rộng theo chiều ngang. Ảo hóa thường được sử dụng để đạt được tính co giãn và linh hoạt trong việc phân bổ tài nguyên vật lý và thường được gọi là dịch vụ cơ sở hạ tầng (IaaS) trong cộng đồng điện toán đám mây.

Ở dạng này, các đơn vị tăng tốc là tài nguyên cải thiện hiệu quả cho tốc độ tính toán, lưu trữ hoặc truyền tải của hệ thống dữ liệu lớn. Khối lượng, sự đa dạng và tốc độ của dữ liệu lớn yêu cầu tốc độ xử lý cao hơn và linh hoạt hơn so với dạng truyền thống.

CHÚ THÍCH: Các đơn vị tăng tốc cho tính toán bao gồm nhưng không giới hạn ở đơn vị xử lý đồ họa, mảng cổng tùy chỉnh để tăng tốc bằng mạch tích hợp field-programmable gate array (FPGA).

9.2.6 Thành phần chức năng nhiều lớp

9.2.6.1 Khái quát chung

Các chức năng nhiều lớp bao gồm một loạt các thành phần chức năng cung cấp dịch vụ cho các thành phần chức năng trong các lớp khác.

9.2.6.2 Thành phần chức năng của lớp tích hợp

9.2.6.2.1 Khái quát chung

Các thành phần chức năng của lớp tích hợp cung cấp các dịch vụ để kết nối chức năng của các thành phần trong cùng một lớp hoặc trên các lớp khác nhau.

Các thành phần chức năng tích hợp có thể bao gồm nhưng không giới hạn:

- Khung chức năng xử lý thông điệp (xem 9.2.6.2.2);
- Khung chức năng quản lý trạng thái (xem 9.2.6.2.3).

9.2.6.2.2 Thành phần chức năng của khung xử lý thông điệp

Thành phần chức năng của khung xử lý thông điệp là cung cấp các dịch vụ, ví dụ: trong hình thức API, để định tuyến và trao đổi thông điệp, bao gồm nhưng không giới hạn việc xếp hàng, truyền tải và nhận dữ liệu đáng tin cậy giữa các nút trong một cụm được chia tỉ lệ theo chiều ngang, hoặc các thành phần trong cùng một hoặc trên các lớp theo chiều dọc khác nhau được định nghĩa trong Hình 12. Ví dụ: một tài nguyên mạng trong lớp tài nguyên có thể gửi một thông điệp về tình trạng sức khỏe của nó tới các thành phần quản lý hệ thống thông qua các API được cung cấp bởi các khung chức năng xử lý thông điệp.

9.2.6.2.3 Thành phần chức năng khung quản lý trạng thái

Thành phần chức năng khung quản lý trạng thái được sử dụng bởi các thành phần chức năng để duy trì hoặc bảo toàn trạng thái qua các nút trong môi trường phân tán, để đảm bảo trạng thái nhất quán và ổn định, tránh xảy ra lỗi tài nguyên hoặc hệ thống. Thông tin trạng thái được duy trì có thể được nhập vào các thành phần quản lý hệ thống để theo dõi hoặc quản lý tài nguyên.

9.2.6.3 Thành phần chức năng của lớp bảo mật và quyền riêng tư

9.2.6.3.1 Khái quát chung

Các thành phần bảo mật và quyền riêng tư được sử dụng để tạo điều kiện thuận lợi cho khả năng tương tác trong BDRA mà không ảnh hưởng đến quyền riêng tư, tính bảo mật hoặc tính toàn vẹn. Các thành phần bảo mật và quyền riêng tư được kết hợp chặt chẽ với tất cả các thành phần chức năng thông qua các API.

CHÚ THÍCH: Các thành phần bảo mật và quyền riêng tư tạo thành một khía cạnh cơ bản của kiến trúc tham chiếu. Đây là các thành phần chính bao trùm hoặc xuyên suốt, cho thấy rằng tất cả các thành phần đều bị ảnh hưởng bởi các vấn đề về bảo mật và quyền riêng tư. Do đó, vai trò của bảo mật và quyền riêng tư được mô tả chính xác trong mối quan hệ với các thành phần nhưng không mở rộng thành các chi tiết nhỏ hơn, có thể chính xác hơn nhưng phải được chuyển sang một kiến trúc tham chiếu bảo mật và quyền riêng tư chi tiết hơn. Dưới đây là các danh mục chung của các thành phần được triển khai để hỗ trợ các khía cạnh bảo mật và quyền riêng tư.

Các thành phần bảo mật và quyền riêng tư giao tiếp và tận dụng một số thành phần quản lý hệ thống để thực hiện thu thập và theo dõi dữ liệu.

9.2.6.3.2 Thành phần chức năng của khung kiểm toán

Thành phần chức năng của khung kiểm toán được sử dụng bởi các thành phần khác để ghi lại các sự kiện trong hệ thống. Sự kiện có thể liên quan đến những người dùng, những thành phần, những công việc và hành động như chạy, dừng, truy cập dữ liệu, cập nhật dữ liệu...Các thành phần này thường tận dụng các thành phần của lớp nền tảng để ghi và duy trì dữ liệu của chúng nhưng có thể vì mục đích bảo

mật mà duy trì dữ liệu bên ngoài kiến trúc dữ liệu lớn. Các dấu vết hoặc nhật ký kiểm toán được duy trì bởi các thành phần này có thể được sử dụng để giúp cho việc truy vết nguồn gốc của dữ liệu, để khôi phục dữ liệu/trạng thái trong trường hợp thành phần hệ thống bị lỗi, hoặc để phân tích chính xác sự cố hoặc sự xâm nhập vào hệ thống.

9.2.6.3.3 Thành phần chức năng của khung xác thực

Thành phần chức năng của khung xác thực cung cấp quyền kiểm soát truy cập vào dữ liệu và các dịch vụ cơ bản trong khác thành phần khác và việc truy cập vào hệ thống từ tất cả các yếu tố bên ngoài. Việc xác thực liên quan đến việc cung cấp một số yếu tố định danh (như tên người dùng) và một khóa truy cập hoặc các khóa (như mật khẩu hoặc chứng nhận) được xác minh dựa trên một kho lưu trữ tham chiếu. Thông thường, thành phần được xác thực giao tiếp với thành phần mà chúng muốn truy cập bằng cách cung cấp mã định danh và khóa. Sau đó, thành phần được truy cập sẽ gọi các dịch vụ xác thực và nhận được câu trả lời về việc cho phép hay từ chối quyền truy cập. Trong điều kiện lý tưởng nhất, các dịch vụ xác thực nên được tập trung trong một thành phần duy nhất, sự xuất hiện của nhiều thành phần trên tất cả các lớp có thể liên quan đến các lớp hoặc các thành phần khác nhau yêu cầu các thành phần xác thực khác nhau.

9.2.6.3.4 Thành phần chức năng của khung ủy quyền

Thành phần chức năng khung ủy quyền hỗ trợ ánh xạ người dùng hoặc mã định danh thành phần với các quyền ưu tiên mà họ có trong việc truy cập tài nguyên (cả dữ liệu và xử lý) trong cụm.

CHÚ THÍCH: Quyền ưu tiên có thể áp dụng cho tài nguyên hoặc phần tử bất kỳ nào đó trong cụm là quyền đọc hoặc truy cập, ghi, xóa, thực thi, di chuyển và kết thúc.

Các quyền ưu tiên có thể áp dụng ở các mức độ chi tiết khác nhau trong tài nguyên. Ví dụ: nhiều nền tảng dữ liệu lớn hiện đang triển khai kiểm soát quyền truy cập ở mức trường/phần tử thay vì kiểm soát ở mức bản ghi hoặc tệp/tập dữ liệu.

9.2.6.3.5 Thành phần chức năng của khung ẩn danh

Thành phần chức năng của khung ẩn danh hỗ trợ duy trì quyền riêng tư hoặc bảo mật cho dữ liệu bằng cách xáo trộn một hoặc nhiều phần tử dữ liệu để chúng không thể dễ dàng liên kết với các phần tử dữ liệu khác.

CHÚ THÍCH: Một ví dụ điển hình là ẩn danh thông tin định dạng cá nhân (PII) của mọi người để bảo vệ quyền riêng tư của họ. Các thành phần này thường thực hiện các hàm băm một chiều để tạo ra các giá trị duy nhất mà không dễ để đảo ngược về giá trị ban đầu của chúng.

Các dịch vụ ủy quyền được sử dụng để xác định xem một người dùng hoặc một dịch vụ nhất định có thể truy cập vào dữ liệu gốc hoặc dữ liệu riêng hay chỉ có quyền truy cập vào dữ liệu đã bị xáo trộn.

9.2.6.4 Thành phần chức năng của lớp quản lý hệ thống

9.2.6.4.1 Khái quát chung

Các thành phần chức năng của lớp quản lý hệ thống cung cấp một loạt các dịch vụ cài đặt, triển khai, cấu hình và giám sát cho các thành phần chức năng trong các lớp dọc, bao gồm nhưng không giới hạn ở:

- Triển khai và cấu hình (xem 9.2.6.4.2);

- Giám sát và cảnh báo (xem 9.2.6.4.3);
- Quản lý tài nguyên dùng chung nhiều lần (xem 9.2.6.4.4);
- Quản lý tính khả dụng cao (xem 9.2.6.4.5);
- Thành phần chức năng quản lý vòng đời dữ liệu lớn (xem 9.2.6.4.6).

9.2.6.4.2 Thành phần chức năng triển khai và cấu hình

Các thành phần triển khai và cấu hình cung cấp các chức năng để cài đặt, triển khai và cấu hình (lại) các gói và các dịch vụ trên các lớp khác nhau.

9.2.6.4.3 Thành phần chức năng giám sát và cảnh báo

Các thành phần giám sát và cảnh báo cung cấp các chức năng giám sát trạng thái và hiệu suất của các tài nguyên và dịch vụ được triển khai trên các lớp khác nhau và gửi các cảnh báo đến các thành phần quản lý phù hợp khi xảy ra các sự kiện nghiêm trọng hoặc đáng báo động. Để phục vụ điện toán cụm và độ co giãn, việc quản lý tài nguyên theo yêu cầu là cần thiết để tận dụng một số lượng lớn các loại tài nguyên và dịch vụ khác nhau trên bất kỳ lớp. Ví dụ: các cảnh báo có thể được gửi đến thành phần quản lý tài nguyên đa khách hàng hoặc thành phần quản lý khả dụng cao để kích hoạt việc cấu hình lại tài nguyên hoặc dịch vụ.

9.2.6.4.4 Thành phần chức năng quản lý tài nguyên đa khách hàng

Thành phần quản lý tài nguyên đa khách hàng cung cấp các chức năng để phân bổ tài nguyên chuyên biệt cho các dịch vụ dữ liệu lớn có nhu cầu sử dụng khác nhau. Đa khách hàng là một kỹ thuật phổ biến được sử dụng nhiều trong điện toán đám mây, cho phép chia sẻ tài nguyên và cung cấp QoS giữa những người sử dụng khác nhau. Các tài nguyên được phân lập và cung cấp cho khách hàng có thể là lớp tài nguyên (như CPU và kho lưu trữ), lớp nền tảng (các hệ thống tệp tin hoặc cơ sở dữ liệu), lớp xử lý (như khung chức năng xử lý đơn hoặc kết hợp), đến lớp ứng dụng dữ liệu lớn (các dịch vụ cụ thể được cung cấp cho người thuê dịch vụ). Vì mục tiêu của điện toán cụm và độ co giãn, các giao diện tiêu chuẩn cần được cho phép để quản lý tài nguyên theo yêu cầu và có thể sử dụng một số lượng khác nhau và các loại tài nguyên và dịch vụ khác nhau trên bất kỳ lớp nào.

9.2.6.4.5 Thành phần chức năng quản lý tính khả dụng cao

Các thành phần quản lý lý tính khả dụng cao cung cấp các chức năng để thiết lập chính sách, triển khai và cấu hình các dịch vụ tĩnh hoặc động liên quan đến việc cung cấp dự phòng, sao lưu dữ liệu hoặc tài nguyên, thay thế dự phòng và di chuyển dữ liệu, để đối mặt và phục hồi khi xảy ra lỗi. Một hệ thống dữ liệu lớn, từ lớp tài nguyên đến lớp trung gian, có thể gặp phải nhiều loại lỗi khác nhau, như các lỗi CPU hoặc bộ lưu trữ, các lỗi tại nút đơn lẻ hoặc cụm, lỗi nguồn hoặc các thiết bị mất điện bất chợt. Các thành phần quản lý tính khả dụng cao có thể nhận đầu vào từ các thành phần giám sát và cảnh báo, và cấu hình các tài nguyên hoặc dịch vụ trực tiếp hoặc thông qua thành phần quản lý tài nguyên nhiều bên thuê.

9.2.6.4.6 Thành phần chức năng quản lý vòng đời dữ liệu lớn

9.2.6.4.6.1 Khái quát chung

Các thành phần chức năng quản lý vòng đời dữ liệu lớn cung cấp các chức năng để quản lý vòng đời dữ liệu lớn từ thời điểm dữ liệu được nhập vào hệ thống thông qua các chức năng nhập dữ liệu cho đến

khi chúng được xử lý hoặc xóa khỏi hệ thống. Các thành phần này có thể bao gồm nhưng không chỉ là quá trình quản lý siêu dữ liệu hoặc quản lý chất lượng dữ liệu.

9.2.6.4.6.2 Thành phần chức năng quản lý siêu dữ liệu

Quản lý siêu dữ liệu đề cập đến các chức năng và khả năng quản lý siêu dữ liệu được tạo ra trong mỗi giai đoạn của vòng đời dữ liệu lớn, từ bước nhập, tiền xử lý, xử lý, phân tích, lưu trữ, tiêu hủy hoặc loại bỏ.

CHÚ THÍCH: Việc quản lý siêu dữ liệu là hết sức cần thiết với các hệ thống dữ liệu lớn, vì:

- Khối lượng siêu dữ liệu trong kỷ nguyên dữ liệu lớn lớn hơn đáng kể so với trước đây và không ngừng tăng lên;
- Hệ thống quản lý siêu dữ liệu phù hợp là công cụ cho quá trình khai thác và phân tích dữ liệu, vì siêu dữ liệu cung cấp thông tin về cách dữ liệu có thể được xử lý hoặc sử dụng.

9.2.6.4.6.3 Thành phần chức năng quản lý chất lượng dữ liệu

Quản lý chất lượng dữ liệu đề cập đến việc thiết lập và triển khai các vai trò, chính sách, hoạt động và quy trình liên quan đến tính chính xác, tính toàn vẹn và đầy đủ của dữ liệu trong suốt vòng đời dữ liệu lớn.

CHÚ THÍCH: Quản lý chất lượng dữ liệu là điều cần thiết đối với các hệ thống dữ liệu lớn, vì chất lượng dữ liệu thấp như dữ liệu không đầy đủ, không đúng hoặc quá lỗi thời có thể ảnh hưởng đến hiệu quả của quá trình khai thác dữ liệu, cản trở các kết quả có ích hoặc dẫn tới sai sót ở đầu ra.

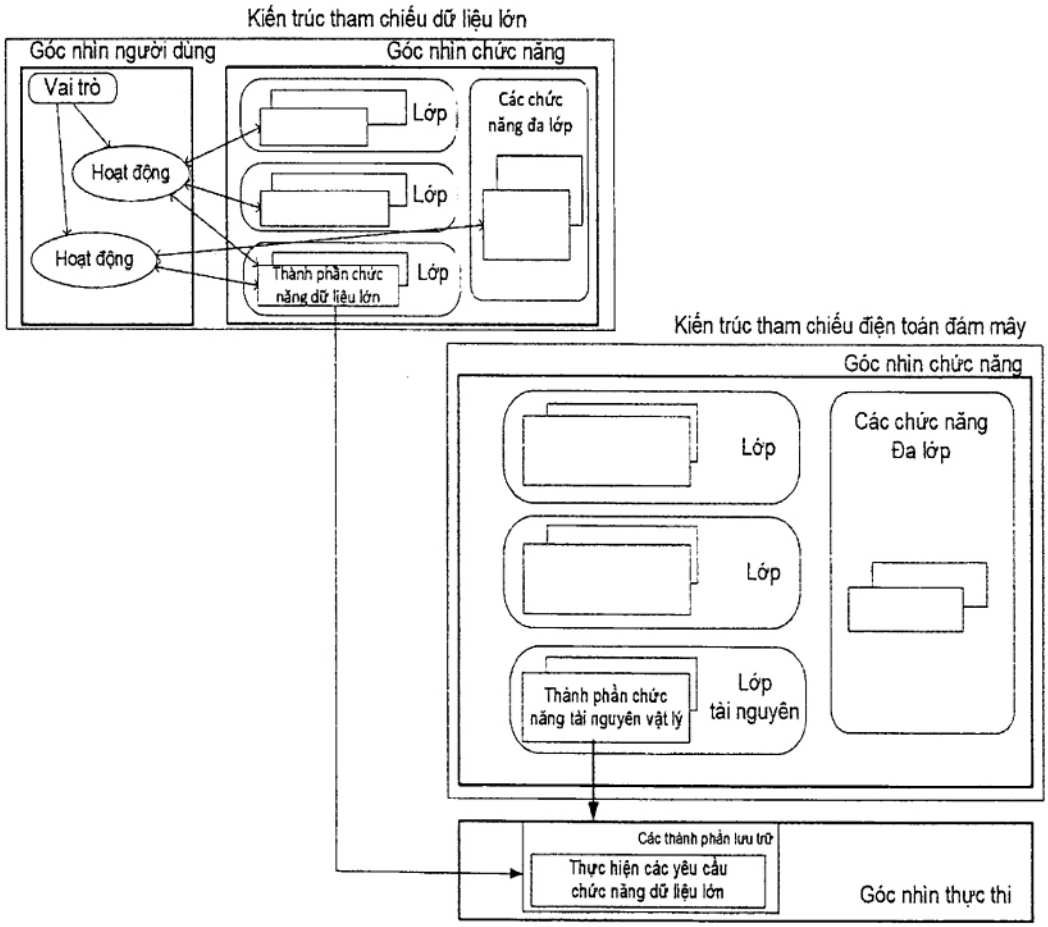
Các chức năng quản lý chất lượng dữ liệu tác động qua lại với mỗi lớp chức năng theo hàng dọc vì chất lượng dữ liệu bị ảnh hưởng bởi việc nhập, kết hợp, phân tích, lưu trữ, hình ảnh hóa dữ liệu và quá trình sử dụng dữ liệu.

Phụ lục A

(Tham khảo)

Ánh xạ góc nhìn chức năng của kiến trúc tham chiếu dữ liệu lớn sang kiến trúc tham chiếu tích hợp hệ thống khác

Đối với dữ liệu lớn, Góc nhìn người dùng là duy nhất. Góc nhìn chức năng có thể được áp dụng trên hệ thống hoặc dịch vụ đích. Ví dụ: kiến trúc tham chiếu điện toán đám mây (ISO/IEC 17789) định nghĩa góc nhìn chức năng của riêng nó cho điện toán đám mây. Nếu giải pháp dữ liệu lớn được triển khai trên môi trường điện toán đám mây, góc nhìn chức năng dữ liệu lớn có thể được ánh xạ sang góc nhìn chức năng của điện toán đám mây (xem Hình A.1)



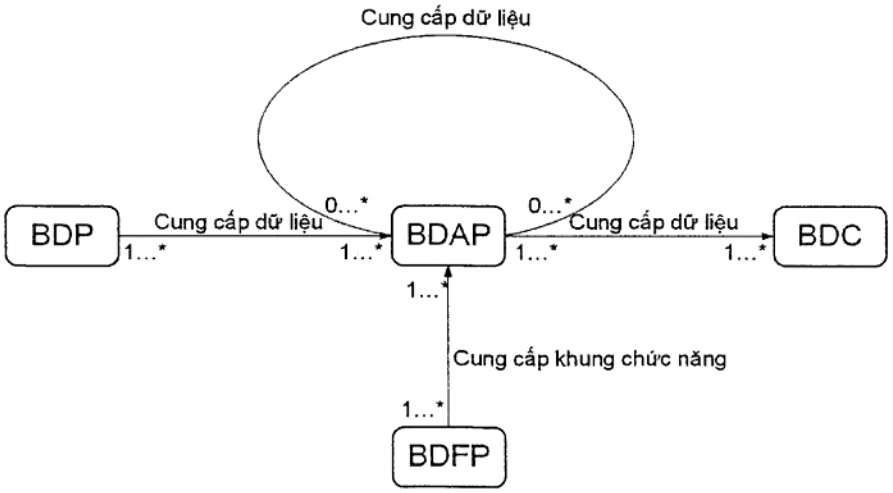
Hình A.1 – Ánh xạ góc nhìn chức năng của kiến trúc tham chiếu dữ liệu lớn sang góc nhìn chức năng của kiến trúc tham chiếu điện toán đám mây

Phụ lục B

(Tham khảo)

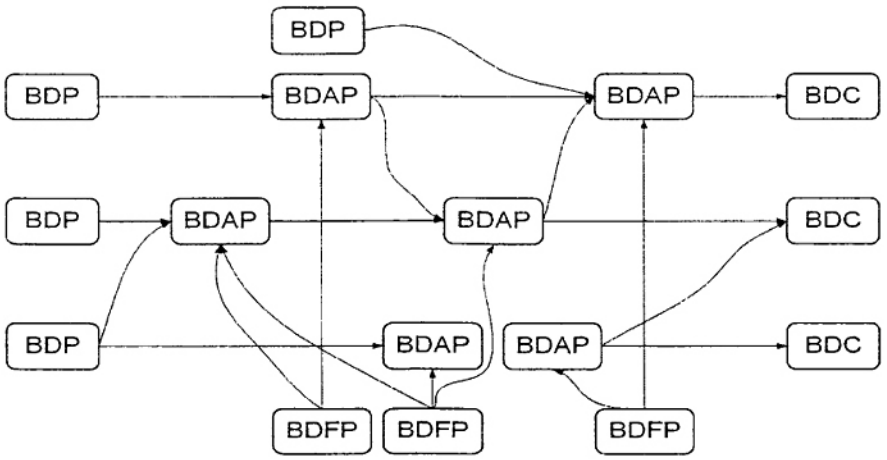
Các ví dụ về mối quan hệ của các vai trò trong hệ sinh thái dữ liệu lớn

Hình B.1 mô tả việc xây dựng kiến trúc tham chiếu dưới dạng biểu đồ lớp UML. Lưu ý rằng, lớp BDAP có một liên kết "cung cấp dữ liệu" lặp lại, để tính đến khả năng một BDAP cung cấp dữ liệu cho một BDAP khác, làm cho việc cung cấp dữ liệu có thể trở thành chuỗi thông qua nhiều BDAP.



Hình B.1 – Biểu đồ lớp UML của kiến trúc tham chiếu

Hình B.2 mô tả mạng lưới các mối quan hệ giữa các thực thể vai trò dựa trên lược đồ UML ở trên. Lưu ý rằng việc cung cấp dữ liệu phân tầng qua nhiều BDAP tại đây.



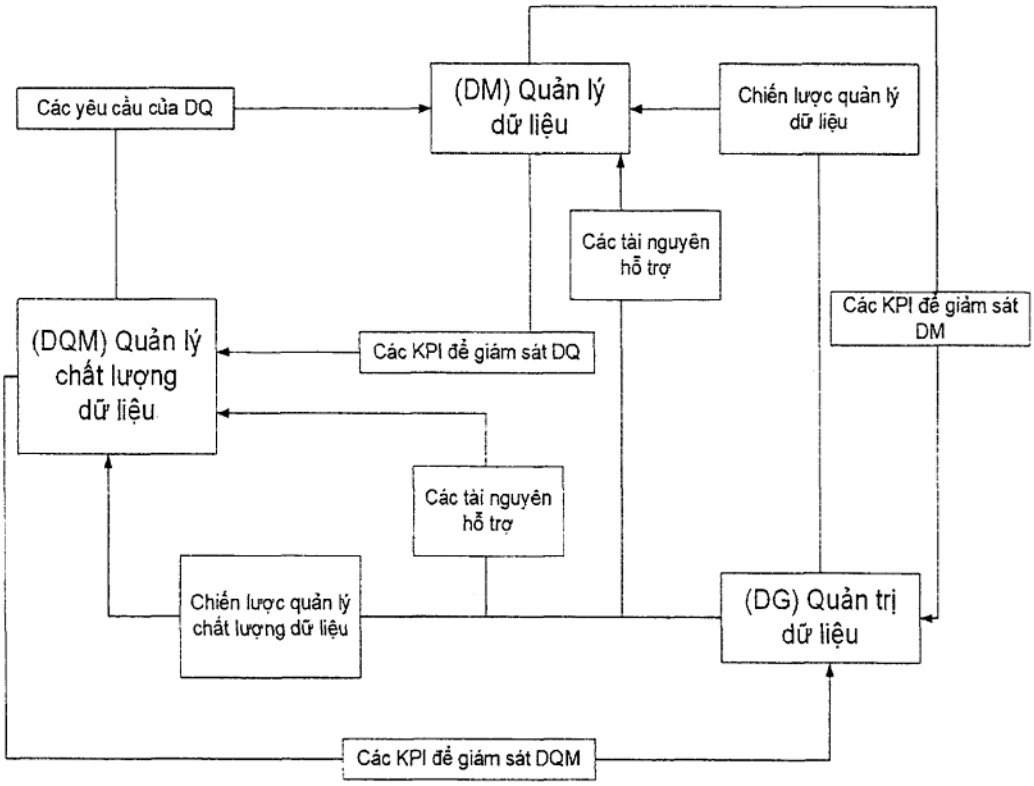
Hình B.2 – Ví dụ về mạng lưới các mối quan hệ giữa các thực thể vai trò dữ liệu lớn

Phụ lục C

(Tham khảo)

Mục đích chính của phụ lục này là xác định các khái niệm về quản trị dữ liệu, quản lý chất lượng dữ liệu và quản lý dữ liệu trong khuôn khổ dữ liệu lớn. Các khái niệm giới thiệu trong phụ lục này chủ yếu được lấy từ các tài liệu ISO liên quan, cũng như các tài liệu đã được công nhận và sử dụng rộng rãi từ các hiệp hội nghề nghiệp cụ thể như: DAMA, Viện Quản trị dữ liệu (DGI, Data Governance Institute), Viện Dữ liệu tổng thể (MDI, Master Data Institute), Hiệp hội chất lượng Dữ liệu và Thông tin Tây Ban Nha (AECDI, Asociación Española para la Calidad de Datos y Información), ISA và Hiệp hội quốc tế về chất lượng thông tin và dữ liệu (IAIDQ).

Cách tiếp cận được trình bày dựa trên ý tưởng kết hợp quan điểm "Dữ liệu là tài sản của tổ chức" và "Dữ liệu là sản phẩm", do đó dữ liệu cần được quản trị như một tài sản và được quản lý như một sản phẩm. Hình C.1 mô tả mối quan hệ giữa ba khái niệm trong phạm vi của khoản mục này.



Hình C.1 – Mối quan hệ giữa các khái niệm DM, DQM và DG

Quản trị dữ liệu (DG) là một chức năng của tổ chức (một tập hợp các hoạt động của tổ chức) chịu trách nhiệm đảm bảo rằng dữ liệu được sử dụng trong các quy trình nghiệp vụ sẽ tạo ra giá trị và đáp ứng hiệu quả các nhu cầu nghiệp vụ.

Quản lý dữ liệu (DM) là một tập hợp các hoạt động nhằm hỗ trợ vòng đời dữ liệu (thu thập, mô tả, lưu trữ, xử lý và tiêu hủy) theo quan điểm kỹ thuật (DAMA, 2009).

DG lập kế hoạch và xác định chiến lược của tổ chức liên quan đến quản lý dữ liệu để đảm bảo rằng dữ liệu được điều chỉnh phù hợp với nghiệp vụ. DM nhận được yêu cầu chiến lược này và tập hợp các nguồn lực để hỗ trợ việc thực hiện chiến lược đó. DM thực thi và triển khai chiến lược quản lý dữ liệu và cung cấp cho DG một bộ chỉ số để theo dõi trạng thái của các hành động đã tiến hành trong chiến lược đã đề ra.

Hình C.1 mô tả mối quan hệ này

Quản lý chất lượng dữ liệu (DQM) là chức năng của tổ chức nhằm xác nhận xem dữ liệu có đạt mức chất lượng phù hợp với yêu cầu nghiệp vụ hay không. Mức chất lượng dữ liệu phù hợp chỉ ra hiệu lực của các kết quả của quy trình nghiệp vụ sử dụng dữ liệu có sẵn.

DG xây dựng chiến lược quản lý chất lượng dữ liệu. Chiến lược này là một tập hợp các ràng buộc và hành động nhằm đảm bảo rằng dữ liệu đáp ứng được các yêu cầu chất lượng đã được đề ra. DG cung cấp chiến lược quản lý chất lượng dữ liệu và các nguồn lực hỗ trợ cho DQM. DQM đưa ra các yêu cầu, chỉ số và tiêu chí quyết định chất lượng dữ liệu dựa trên những ràng buộc và hành động để kiểm soát và cải thiện các mức độ chất lượng dữ liệu nếu cần thiết một cách hiệu quả.

Các yêu cầu và cách thức đo lường mức chất lượng dữ liệu được cung cấp cho DM để triển khai và thực thi các chỉ số chất lượng dữ liệu do DQM xác định. Các kết quả về chỉ số chất lượng dữ liệu được chuyển lại cho DQM, nơi chịu trách nhiệm xác định xem yêu cầu về chất lượng dữ liệu của tổ chức có được đáp ứng hay không.

DQM cung cấp cho DG một bộ chỉ số về hiệu quả của các hoạt động chất lượng dữ liệu.

DG yêu cầu bộ phận IT, nhân lực hoặc tài chính cung cấp các nguồn lực cần thiết để đảm bảo tính khả thi của các chức năng DQM và DM.

Các khái niệm đã giới thiệu được điều chỉnh (nếu có thể) theo các tiêu chuẩn ISO đã công bố hoặc đang xây dựng. Theo nghĩa này, chúng tôi xem xét các giả thuyết sau đây:

- Trên thực tế, dữ liệu rất có giá trị đối với các tổ chức, vì vậy dữ liệu có thể được coi là tài sản và được quản trị phù hợp để đáp ứng các mục tiêu của tổ chức. Theo giả thuyết này, có thể xem xét rằng:

Xử lý dữ liệu là tài sản có thể được hiểu theo các nguyên tắc của ISO 55000[24];

Quản trị dữ liệu có thể được hiểu theo các nguyên tắc của ISO/IEC 38500[21];

- Trên thực tế, dữ liệu có thể vừa được coi là nguyên liệu thô, vừa là kết quả của quá trình "xử lý dữ liệu". Lúc này, dữ liệu là một sản phẩm. Theo giả thuyết này, có thể xem xét rằng:

Quản lý chất lượng dữ liệu thường được hiểu theo các nguyên tắc của bộ tiêu chuẩn ISO 8000;

Các đặc tính của chất lượng dữ liệu như một sản phẩm cũng như định nghĩa của các chỉ số đi kèm, có thể được xác định trong ISO/IEC 25012[9], ISO/IEC 25024[10] và ISO 8000-8[30].

Các khái niệm được trích xuất và tùy chỉnh từ bộ tiêu chuẩn này có thể được bổ sung một cách thuận tiện bởi bất kỳ tiêu chuẩn hiện có nào khác liên quan đến các chủ đề cụ thể về quản lý dữ liệu (như ISO 22745[11]), hoặc bất kỳ tiêu chuẩn nào khác phù hợp để quản lý dữ liệu hoặc quản lý chất lượng dữ liệu

trong các lĩnh vực cụ thể (như ISO 19157[12], ISO 13119[13], ISO/TR 21707[14] hoặc ISO/HL7 10781[15]).

Ngay cả khi đó là các vấn đề về ngôn ngữ, việc sử dụng các thuật ngữ "Quản trị dữ liệu lớn", "Quản lý dữ liệu lớn" và "Quản lý chất lượng dữ liệu lớn" không tương đương với "Quản trị dữ liệu trong dữ liệu lớn [Dự án|Hệ sinh thái]", "Quản lý dữ liệu trong dữ liệu lớn [Dự án|Hệ sinh thái]" và "Quản lý chất lượng dữ liệu trong dữ liệu lớn [Dự án|Hệ sinh thái]", vì các khái niệm DG, DQM và DM đã vượt qua giới hạn của việc sử dụng dữ liệu đơn thuần.

Thư mục tài liệu tham khảo

[1] Colella P., Defining software requirements for scientific computing. Slide of 2004 presentation included in David Patterson's 2005 talk. http://www.lanl.gov/orgs/hpc/salishan/salishan2005/davidpatterson.pdf.

[2] Patterson D., Yelick K., Dwarf Mind. A View From Berkeley. http://view.eecs.berkeley.edu/wiki/Dwarf_Mine.

[3] United States Census Bureau, The "72-Year Rule." https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html. Accessed March 3, 2015.

[4] Apache Hadoop., Web HDFS REST API. https://hadoop.apache.org/docs/r1.0.4/webhdfs.html. #FsURIvsHTTP_URL. Accessed Feb 24, 2017.

[5] ISO/IEC 20546, Information technology — Big data — Overview and vocabulary.

[6] ISO/IEC 17789, Information technology — Cloud computing — Reference architecture.

[7] DoD Reference Architecture Description https://dodcio.defense.gov/Portals/0/Documents/DIEA/Ref_Archi_Description_Final_v1_18Jun10.pdf.

[8] ISO/IEC 27002, Information technology — Security techniques — Code of practice for information security controls.

[9] ISO/IEC 25012, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model.

[10] ISO/IEC 25024, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality.

[11] ISO 22745, Industrial automation systems and integration — Open technical dictionaries and their application to master data.

[12] ISO 19157, Geographic information — Data quality.

[13] ISO 13119, Health informatics — Clinical knowledge resources — Metadata.

[14] ISO/TR 21707, Intelligent transport systems — Integrated transport information, management and control — Data quality in ITS systems.

[15] ISO/HL7 10781, Health Informatics — HL7 Electronic Health Records-System Functional Model, Release 2 (EHR FM).

[16] Smith B., Malyuta T., Mandrick W.S., Fu C., Parent K., Patel M., (2012). Horizontal Integration of Warfighter Intelligence Data: A Shared Semantic Resource for the Intelligence Community. In Proceedings of the Conference on Semantic Technology in Intelligence, Defense and Security (STIDS), CEUR_. pp. 1–8.

[17] Yoakum-Stover S., Malyuta T., Unified Integration Architecture for Intelligence Data." Proceedings of DAMA International Europe Conference, London, UK. 2008.

[18] ISO 8000-2, Data quality — Part 2: Vocabulary.

[19] ISO/TS 8000-60, Data quality — Part 60: Data quality management: Overview.

[20] ISO 8000-61, Data quality — Part 61: Data quality management: Process reference model.

[21] ISO/IEC 38500, Information technology — Governance of IT for the organization.

[22] ISO/IEC 38505-1, Information technology — Governance of IT — Governance of data — Part 1: Application of ISO/IEC 38500 to the governance of data.

[23] ISO/IEC TR 38505-2, Information technology — Governance of IT — Governance of data — Part 2: Implications of ISO/IEC 38505-1 for data management.

[24] ISO 55000, Asset management — Overview, principles and terminology.

[25] ISO 55001, Asset management — Management systems — Requirements.

[26] ISO 55002, Asset management — Management systems — Guidelines for the application of ISO 55001.

[27] ISO/IEC/IEEE 42010, Systems and software engineering — Architecture description.

[28] ISO/IEC 20547-4, Information technology — Big data reference architecture — Part 4: Security and Privacy.

[29] ISO/IEC 27000, Information technology — Security techniques — Information security management systems — Overview and vocabulary.

[30] ISO 8000-8:2015, Data quality — Part 8: Information and data quality: Concepts and measuring.