BỘ THÔNG TIN VÀ TRUYỀN THÔNG

**VIỆN CÔNG NGHIỆP PHẦN MỀM VÀ NỘI DUNG SỐ VIỆT NAM**

**THUYẾT MINH**

**Dự thảo Tiêu chuẩn quốc gia**

**Kiến trúc tham chiếu Dữ liệu lớn**

**Hà Nội – 8/2022**

**MỤC LỤC**

[1. Tình hình xây dựng tiêu chuẩn về kiến trúc tham chiếu dữ liệu lớn 2](#_Toc114490251)

[2. Tổng quan về chuẩn kiến trúc tham chiếu dữ liệu lớn 3](#_Toc114490252)

[2.1. Khái quát chung 3](#_Toc114490253)

[2.2. Mô tả về kiến trúc tham chiếu dữ liệu lớn 4](#_Toc114490254)

[3. Các vấn đề đối với tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn 6](#_Toc114490255)

[3.1 Các yêu cầu chung 6](#_Toc114490256)

[3.2. Mô hình kiến trúc và các thành phần, vai trò, chức năng 7](#_Toc114490257)

[4. Lựa chọn tài liệu tham chiếu để xây dựng tiêu chuẩn về kiến trúc tham chiếu dữ liệu lớn 12](#_Toc114490258)

[4.1 Các tài liệu mô tả về kiến trúc tham chiếu dữ liệu lớn trên thế giới 12](#_Toc114490259)

[4.2 Lựa chọn tài liệu tham chiếu chính để xây dựng tiêu chuẩn 18](#_Toc114490260)

[5. Xây dựng dự thảo tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn 21](#_Toc114490261)

[5.1 Phương pháp xây dựng 21](#_Toc114490262)

[5.2 Nội dung dự thảo tiêu chuẩn 21](#_Toc114490263)

[Tài liệu tham khảo 27](#_Toc114490264)

# 1. Tình hình xây dựng tiêu chuẩn về kiến trúc tham chiếu dữ liệu lớn

Ngày nay, dữ liệu lớn là một khái niệm không còn xa lạ trong xu thế chuyển đổi số đang diễn ra trên toàn thế giới. Để có thể vận hành, khai thác hiệu quả những lợi ích mà dữ liệu lớn mang lại, rất nhiều các tập đoàn, các công ty lớn hoạt động trong lĩnh vực công nghệ đã xây dựng những mô hình kiến trúc dữ liệu lớn để ứng dụng trong việc khai thác, vận hành kho dữ liệu của riêng mình như: Google, Microsoft, Amazon, Facebook, Netflix, Oracle…

Bên cạnh đó, cũng có nhiều các tổ chức tiêu chuẩn, các viện nghiên cứu, các hiệp hội công nghệ quốc tế cũng đưa ra các mô hình kiến trúc tham chiếu của mình để khuyến nghị các tổ chức, đơn vị đã, đang và sẽ triển khai dữ liệu lớn tham khảo như một giải pháp tuân thủ các tiêu chuẩn chung trong lĩnh vực dữ liệu lớn, tiêu biểu như:

* Bộ tài liệu ISO/IEC 20547 của Tổ chức tiêu chuẩn quốc tế (ISO), được biên soạn bởi Ủy ban kỹ thuật chung ISO/IEC JTC1, tiểu ban SC42 (chuyên về chuẩn hóa trong lĩnh vực trí tuệ nhân tạo và dữ liệu lớn);
* Tài liệu ITU-T Y.3605 của Liên minh viễn thông quốc tế (ITU), được biên soạn bởi Bộ phận Tiêu chuẩn hóa viễn thông (ITU-T) (đơn vị chịu trách nhiệm nghiên cứu các vấn đề liên quan đến kỹ thuật, vận hành và đưa ra khuyến nghị chuẩn hóa);
* Bộ tài liệu NIST.SP.1500 của Viện tiêu chuẩn và công nghệ quốc gia Mỹ (NIST), được biên soạn bởi Nhóm công tác về dữ liệu lớn của NIST (NBD-PWG) (nhóm nghiên cứu được hình thành với mục tiêu xây dựng, phát triển một cộng đồng có nhận thức chung về các khái niệm, phân loại, kiến trúc tham chiếu và lộ trình công nghệ trong lĩnh vực dữ liệu lớn).

# 2. Tổng quan về chuẩn kiến trúc tham chiếu dữ liệu lớn

## 2.1. Khái quát chung

Kiến trúc tham chiếu là một mô hình kiến trúc tổng thể, cung cấp nguồn thông tin có căn cứ về một lĩnh vực, chủ đề cụ thể nhằm mục đích hướng dẫn và hạn chế việc tạo ra nhiều kiến trúc và giải pháp. Kiến trúc tham chiếu thường đóng vai trò là nền tảng tham chiếu cho các giải pháp kiến trúc và cũng có thể được sử dụng cho mục đích so sánh và định hướng.

Kiến trúc tham chiếu dữ liệu lớn cần đảm bảo các mục tiêu sau:

- Cung cấp một ngôn ngữ chung cho các bên liên quan;

- Khuyến khích việc tuân thủ các tiêu chuẩn, thông số kỹ thuật và khuôn mẫu chung;

- Cung cấp các phương pháp nhất quán để triển khai hệ thống, giải quyết các vấn đề tương tự;

- Minh họa và nâng cao nhận thức về các thành phần, quy trình và hệ thống dữ liệu lớn khác nhau;

- Cung cấp tài liệu tham khảo kỹ thuật cho các đơn vị, tổ chức và người dùng hiểu, thảo luận, phân loại và so sánh các giải pháp dữ liệu lớn khác nhau;

- Tạo điều kiện thuận lợi cho việc phân tích các tiêu chuẩn về khả năng tương tác, tính di động, khả năng tái sử dụng và mở rộng hệ thống.

Kiến trúc tham chiếu dữ liệu lớn không đại diện cho kiến trúc hệ thống của một hệ thống dữ liệu lớn cụ thể nào, mà là công cụ để mô tả, thảo luận và phát triển kiến trúc hệ thống cụ thể bằng cách sử dụng các khung tham chiếu. Mô hình không bị ràng buộc với bất kỳ sản phẩm, dịch vụ hoặc thực thi tham chiếu của một nhà cung cấp cụ thể nào, cũng như không xác định các giải pháp mang tính quy định, ngăn cản sự đổi mới, sáng tạo.

## 2.2. Mô tả về kiến trúc tham chiếu dữ liệu lớn

Kiến trúc tham chiếu dữ liệu lớn là một mô hình khái niệm tổng thể của một hệ thống dữ liệu lớn. Nó cung cấp giao diện tương tác giữa các thành phần chức năng logic và các thành phần khác trong một hệ thống dữ liệu lớn tổng thể. Kiến trúc tham chiếu dữ liệu lớn là một khái niệm nằm ngoài phạm vi công nghệ và được sử dụng như một khuôn mẫu tham chiếu kỹ thuật chung để phát triển các kiến trúc hệ thống ứng dụng dữ liệu lớn cụ thể khác nhau. Mục tiêu là thiết lập một kiến trúc tham chiếu dữ liệu lớn mở, cho phép các kỹ sư hệ thống, nhà khoa học dữ liệu, nhà phát triển phần mềm và nhà phát triển giải pháp, chính sách của các đơn vị, tổ chức có thể tương tác đề giải quyết các vấn đề liên quan đến các hệ thống dữ liệu lớn.

Kiến trúc tham chiếu dữ liệu lớn sử dụng cấu trúc phân lớp để thể hiện các khái niệm ở mức cao của một hệ thống dữ liệu lớn và phân loại các thành phần chung. Về thành phần, kiến trúc tham chiếu dữ liệu lớn bao gồm một loạt các thành phần logic ở các cấp độ khái niệm khác nhau. Các thành phần logic này được chia thành các cấp độ, từ cao đến thấp là các vai trò, hoạt động và thành phần chức năng. Các thành phần logic ở mức cao nhất là các vai trò, bao gồm: hệ thống điều phối, cung cấp dữ liệu, cung cấp ứng dụng, cung cấp khung chức năng, người dùng, quản lý, bảo mật và quyền riêng tư. Các thành phần logic mức hai là các hoạt động của mỗi vai trò. Các thành phần logic mức ba là các thành phần chức năng cần thiết để thực hiện các hoạt động.

Bố cục tổng thể của kiến trúc tham chiếu dữ liệu lớn được tổ chức theo hai chiều đại diện cho chuỗi giá trị dữ liệu lớn, đó là: chuỗi giá trị thông tin (trục ngang) và chuỗi giá trị CNTT (trục dọc). Trong chuỗi giá trị thông tin, giá trị của dữ liệu lớn được thực hiện thông qua các hoạt động như thu thập, tiền xử lý, phân tích, trực quan hóa và truy cập dữ liệu. Trong chuỗi giá trị CNTT, giá trị của dữ liệu lớn đạt được bằng cách cung cấp cho các hệ thống ứng dụng dữ liệu lớn: mạng, cơ sở hạ tầng, nền tảng, công cụ ứng dụng và các dịch vụ CNTT khác để lưu trữ và vận hành dữ liệu lớn. Các nhà cung cấp ứng dụng dữ liệu lớn nằm ở giao điểm của hai chiều, cho thấy phân tích dữ liệu lớn và thực hiện nó mang lại giá trị cho các bên liên quan trên cả hai chuỗi giá trị.

Các thành phần mô hình chính đại diện cho các vai trò kỹ thuật khác nhau tồn tại trong mỗi hệ thống dữ liệu lớn: hệ thống điều phối, cung cấp dữ liệu, cung cấp ứng dụng, cung cấp khung chức năng và người dùng dữ liệu lớn. Hai thành phần mô hình rất quan trọng khác là quyền riêng tư và quản lý bảo mật, đại diện cho các thành phần cung cấp dịch vụ và chức năng cho các thành phần chính khác của hệ thống dữ liệu lớn. Chức năng của hai thành phần mô hình này là cực kỳ quan trọng và được tích hợp trong mọi giải pháp kiến trúc dữ liệu lớn.

Kiến trúc tham chiếu dữ liệu lớn có thể được sử dụng cho các hệ thống phức tạp bao gồm nhiều hệ thống dữ liệu lớn, như hệ thống xếp chồng hoặc chuỗi, để người dùng dữ liệu lớn của một trong các hệ thống có thể đóng vai trò cung cấp dữ liệu lớn cho một hệ thống dữ liệu lớn khác.

#

# 3. Các vấn đề đối với tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn

## 3.1 Các yêu cầu chung

Việc xây dựng, phát triển kiến trúc tham chiếu dữ liệu lớn đòi hỏi sự hiểu biết sâu sắc về kiến trúc hệ thống, các vấn đề kỹ thuật và những đặc trưng tiêu biểu của dữ liệu lớn. Do vậy, phần lớn các tổ chức nghiên cứu, xây dựng tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn trên thế giới sử dụng các phương pháp khảo sát, thu thập và tổng hợp, đánh giá thông tin từ các mô hình, kiến trúc của các tổ chức, tập đoàn công nghệ tiêu biểu đã và đang cung cấp các dịch vụ dữ liệu lớn trên thế giới.

Tuy nhiên, vì dữ liệu lớn là một khái niệm rất rộng, mang tính tổng quát cũng như các đặc thù riêng của từng yêu cầu nghiệp vụ nên các kết quả nghiên cứu vẫn có sự khác biệt nhất định. Báo cáo của các tổ chức, đơn vị nghiên cứu đã chỉ ra rằng, một kiến trúc tham chiếu dữ liệu lớn cần phải đảm bảo các yêu cầu sau:

* Các mô hình khái niệm, góc nhìn kiến trúc;
* Các thành phần của kiến trúc;
* Các vai trò, hoạt động của các thành phần;
* Các thành phần khác.

Trong đó:

* Các mô hình khái niệm, góc nhìn kiến trúc thể hiện, mô tả được các khía cạnh kiến trúc khác nhau, mang đến một cái nhìn toàn diện về kiến trúc tổng thể của một hệ thống dữ liệu lớn.
* Đối với kiến trúc dữ liệu lớn của mỗi tổ chức, đơn vị xây dựng đều có những khác biệt nhất định. Tuy nhiên, về cơ bản các kiến trúc tham chiếu cần phải có ít nhất các thành phần sau:
* Nhà cung cấp dữ liệu/dữ liệu lớn;
* Nhà cung cấp ứng dụng dữ liệu lớn;
* Nhà cung cấp khung chức năng xử lý dữ liệu lớn;
* Người dùng dữ liệu/dữ liệu lớn;
* Các đối tác dịch vụ.
* Các vai trò, hoạt động của các thành phần trong kiến trúc dữ liệu lớn cần được xác định, mô tả cụ thể;
* Các thành phần khác là các thành phần xuyên suốt quy trình, vòng đời của một hệ thống dữ liệu lớn. Có vai trò tác động, hỗ trợ trực tiếp đến các vai trò, các hoạt động của các thành phần trong kiến trúc dữ liệu lớn.

## 3.2. Mô hình kiến trúc và các thành phần, vai trò, chức năng

Mô hình kiến trúc tham chiếu hệ thống dữ liệu lớn được xây dựng bởi các thành phần khái niệm, với mục đích giúp người đọc hiểu và nắm được ý nghĩa ở mức tổng quát. Trong đó, sự tương tác và mối quan hệ giữa các thành phần, vai trò được biểu diễn bằng các mũi tên, liên kết chúng với nhau, hoặc cũng có thể là những thực thể bao phủ các thành phần, vai trò bên trong.

**Các thành phần, vai trò chính:**

* Nhà điều phối/điều hành hệ thống: 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, kiểm soát để đảm bảo việc tuân thủ của hệ thống. Những đối tượng trong thành phần này bao gồm: lãnh đạo các tổ chức, đơn vị; nhà khoa học dữ liệu; kiến trúc sư hệ thống; kiến trúc sư phần mềm…Vai trò của họ là xác định và tích hợp các hoạt động ứng dụng dữ liệu cần thiết vào hệ thống theo chiều dọc. Nhà điều phối hệ thống có thể là con người hoặc phần mềm, hoặc kết hợp cả hai. Chức năng chính của vai trò này là thiết lập cấu hình và quản lý các thành phần khác trong kiến trúc hệ thống dữ liệu lớn để thực thi các công việc trong hệ thống.
* Nhà cung cấp dữ liệu: đóng vai trò cung cấp dữ liệu cho hệ thống dữ liệu lớn, đối tượng trong vai trò này có thể là: các doanh nghiệp, tổ chức chính phủ, nhà khoa học, các phần mềm ứng dụng, nhà cung cấp ứng dụng…Trong hệ thống dữ liệu lớn, các hoạt động của nhà cung cấp dữ liệu thường bao gồm: thu thập, lưu trữ, biến đổi/chuyển đổi, làm sạch, tạo siêu dữ liệu và chính sách kiểm soát truy cập hoặc triển khai thông qua các giao diện lập trình ứng dụng…
* Nhà cung cấp ứng dụng dữ liệu lớn: có vai trò thực hiện một loạt các hoạt động trong vòng đời dữ liệu để đáp ứng các yêu cầu hệ thống và các yêu cầu về bảo mật và quyền riêng tư do nhà điều hành/điều phối hệ thống thiết lập. Nhà cung cấp ứng dụng dữ liệu lớn xây dựng một hệ thống ứng dụng dữ liệu lớn cụ thể bằng cách kết hợp các tài nguyên chung và khả năng dịch vụ trong khung chức năng xử lý dữ liệu lớn, đồng thời đóng gói logic nghiệp vụ và các chức năng thành các thành phần kiến trúc.
* Nhà cung cấp khung chức năng, xử lý dữ liệu lớn: có vai trò cung cấp các nguồn lực và khả năng dịch vụ chung cho các nhà cung cấp ứng dụng dữ liệu lớn để tạo ra các hệ thống ứng dụng dữ liệu lớn cụ thể. Bao gồm: các trung tâm dữ liệu, các nhà cung cấp dịch vụ đám mây, các cụm máy chủ riêng lẻ…Các hoạt động và chức năng của nhà cung cấp khung chức năng xử lý dữ liệu lớn có thể kể đến như: cung cấp cơ sở hạ tầng, nền tảng dữ liệu, khung chức năng, quản lý tài nguyên…
* Người dùng dữ liệu lớn: là đối tượng nhận kết quả đầu ra của hệ thống dữ liệu lớn. Tương tự như nhà cung cấp dữ liệu, người dùng dữ liệu lớn có thể là người dùng cuối hoặc các hệ thống ứng dụng khác sử dụng đầu ra từ nhà cung cấp dữ liệu. Các hoạt động của người dùng dữ liệu lớn bao gồm: tìm kiếm/truy xuất, tải về, phân tích cục bộ, tạo báo cáo, trực quan hóa…Người dùng dữ liệu lớn sử dụng các giao diện hoặc dịch vụ do nhà cung cấp ứng dụng dữ liệu lớn cung cấp để truy cập thông tin và họ quan tâm như: báo cáo, truy cập và kiết xuất dữ liệu.
* Bảo mật và quyền riêng tư: trong kiến trúc tham chiếu dữ liệu lớn, vai trò bảo mật và quyền riêng tư tác động, ảnh hưởng trực tiếp tới các thành phần, chức năng khác. Đóng vai trò như một khía cạnh xuyên suốt trong toàn bộ kiến trúc, quy trình hoạt động của một hệ thống dữ liệu lớn.
* Quản lý: vai trò này gồm hai nhóm hoạt động chính là quản lý hệ thống và quản lý vòng đời dữ liệu lớn. Nhóm quản lý hệ thống bao gồm các hoạt động như: cung cấp, cấu hình, quản lý gói, quản lý phần mềm, quản lý lưu trữ, quản lý dung lượng, quản lý tài nguyên, quản lý hiệu suất…của cơ sở hạ tầng dữ liệu lớn. Quản lý vòng đời dữ liệu lớn bao gồm các quy trình xử lý trong một vòng đời dữ liệu lớn, các hoạt động và chức năng của nó có vai trò xác minh xem dữ liệu có thể được xử lý chính xác bởi hệ thống dữ liệu lớn trong mỗi quy trình của vòng đời hay không.
* Quản trị: là một thuộc tính hoặc tính năng cần được phối hợp thực hiện với các hoạt động của các thành phần, vai trò khác để đảm bảo dữ liệu được sử dụng trong quy trình nghiệp vụ tạo ra giá trị và đáp ứng hiệu quả các yêu cầu của nghiệp vụ.

**Các chức năng công nghệ chính:**

* Thu thập dữ liệu: cần đảm bảo dữ liệu được thu thập theo thời gian thực hoặc kịp thời từ các nguồn dữ liệu khác nhau và gửi đến hệ thống lưu trữ hoặc hệ thống phần mềm trung gian để xử lý. Các hình thức thu thập dữ liệu chủ yếu có thể từ các thiết bị, các trang web, các ứng dụng…
* Tiền xử lý dữ liệu: các hệ thống dữ liệu lớn thường có một hoặc nhiều nguồn dữ liệu, có thể bao gồm các cơ sở dữ liệu đồng nhất hoặc không đồng nhất…Dữ liệu đến từ các nguồn thu thập theo thời gian thực dễ bị các tác động dẫn đến việc thừa, thiếu hoặc xung đột thông tin. Ngoài ra, các thuật toán và công nghệ triển khai trong quá trình xử lý, phân tích và trực quan hóa rất phúc tạp, đa dạng, do vậy chúng cũng cần thực hiện một số thao tác xử lý dữ liệu trước khi thực hiện.
* Lưu trữ dữ liệu: lưu trữ và truy cập phân tán là công nghệ chính trong lưu trữ dữ liệu lớn, chúng cần phải đảm bảo tính tiết kiệm, hiệu quả và có khả năng chịu lỗi. Công nghệ lưu trữ phân tán liên quan trực tiếp đến loại phương tiện lưu trữ dữ liệu và hình thức tổ chức, quản lý dữ liệu. Các phương tiện lưu trữ dữ liệu chính hiện tại gồm có bộ nhớ, đĩa, băng…Các hình thức tổ chức và quản lý dữ liệu chính bao gồm việc sắp xếp, tổ chức hàng, cột, giá trị và các mối quan hệ.
* Xử lý dữ liệu: hiện nay, các mô hình tính toán, xử lý dữ liệu lớn phổ biến bao gồm: mô hình tính toán Map/Reduce, mô hình tính toán DAG, mô hình tính toán BSP và một số hình thức khác.
* Phân tích dữ liệu: bao gồm các công nghệ thống kê, phân tích dữ liệu phân tán, công nghệ khai thác phân tán và học sâu (deep learning). Kỹ thuật thống kê, phân tích phân tán về cơ bản có thể được thực hiện trực tiếp bằng kỹ thuật xử lý dữ liệu, trong khi đó kỹ thuật khai thác phân tán và học sâu có thể được chia thành: phân cụm, phân loại, phân tích liên kết…
* Trực quan hóa dữ liệu: sử dụng đồ họa máy tính và công nghệ xử lý hình ảnh để chuyển đổi dữ liệu thành đồ họa hoặc hình ảnh, hiển thị lên màn hình và thực hiện xử lý tương tác. Nó liên quan đến nhiều lĩnh vực kỹ thuật như đồ họa máy tính, xử lý hình ảnh, thị giác máy tính và tương tác giữa người/máy. Tương tác dữ liệu, thông tin giữa dữ liệu lớn và người dùng là mục tiêu quan trọng của trực quan hóa dữ liệu. Công nghệ trực quan hóa dữ liệu biểu thị mỗi mục dữ liệu trong cơ sở dữ liệu như một phần tử duy nhất và số lượng lớn các tập dữ liệu tạo thành hình ảnh dữ liệu. Mỗi giá trị thuộc tính của dữ liệu được biểu diễn dưới dạng dữ liệu đa chiều và dữ liệu có thể được quan sát từ các chiều khác nhau giúp việc quan sát, phân tích dữ liệu đạt hiệu quả cao hơn.
* Bảo mật và quyền riêng tư: là lĩnh vực ảnh hưởng trực tiếp (hoặc gián tiếp) tới tất cả các thành phần, vai trò và hoạt động trong kiến trúc dữ liệu lớn. Bảo mật và quyền riêng tư tương tác với Nhà điều hành/điều phối hệ thống về chính sách, yêu cầu và công tác kiểm soát, quản lý, cũng như với cả nhà cung cấp ứng dụng dữ liệu lớn, nhà cung cấp khung chức năng xử lý dữ liệu lớn để phát triển, thực thi và vận hành hệ thống. Các vấn đề bảo mật gồm có: tính bảo mật, tính toàn vẹn, tính khả dụng… Các vấn đề quyền riêng tư gồm có: tính không liên kết, tính minh bạch và khả năng can thiệp…
* Quản lý và quản trị: các đặc điểm của dữ liệu lớn về khối lượng, vận tốc, sự đa dạng và biến hóa đòi hỏi một nền tảng quản lý và phần mềm điều hành linh hoạt để cung cấp, cấu hình và quản lý chúng. Cùng với việc giám sát, quản lý tài nguyên và hiệu suất, việc quản lý và quản trị dữ liệu lớn đòi hỏi sự xem xét kỹ lượng về hệ thống, dữ liệu, bảo mật và quyền riêng tư ở quy mô tổng quát, đồng thời duy trì chất lượng dữ liệu và khả năng an toàn, bảo mật ở mức cao. Các vấn đề liên quan đến quản lý dữ liệu lớn gồm: cung ứng (phân bổ), cấu hình, quản lý gói, quản lý tài nguyên…Các vấn đề liên quan đến quản trị được xác định gồm: chiến lược tổ chức trong quản lý dữ liệu và chiến lược quản lý chất lượng dữ liệu.

# 4. Lựa chọn tài liệu tham chiếu để xây dựng tiêu chuẩn về kiến trúc tham chiếu dữ liệu lớn

## 4.1 Các tài liệu mô tả về kiến trúc tham chiếu dữ liệu lớn trên thế giới

Hiện nay, trên thế giới đã có nhiều tổ chức, đơn vị nghiên cứu và tập đoàn công nghệ lớn nghiên cứu và đưa ra các tài liệu về kiến trúc tham chiếu. Trong đó, các tài liệu kiến trúc mang tính tiêu chuẩn, có giá trị tham chiếu là: ISO/IEC 20547 của Tổ chức Tiêu chuẩn hóa Quốc tế (ISO), ITU-T Y.3605 của Liên minh viễn thông quốc tế (ITU) và NIST.SP.1500 của Viện Tiêu chuẩn và Công nghệ Quốc gia Mỹ (NIST).

**Bộ tài liệu tiêu chuẩn ISO/IEC 20547 về Kiến trúc tham chiếu dữ liệu lớn** là tài liệu được viết bởi Ủy ban kỹ thuật chung ISO/IEC JTC1, Công nghệ thông tin, tiểu ban 42 (chuyên về chuẩn hóa trong lĩnh vực trí tuệ nhân tạo và dữ liệu lớn). Bộ tài liệu tiêu chuẩn được xây dựng gồm 5 phần, bao gồm:

* Tài liệu ISO/IEC TR 20547-1:2020: cung cấp cho người dùng cái nhìn tổng quan về khung kiến trúc tham chiếu và quy trình áp dụng trong việc phát triển kiến trúc.
* Tài liệu ISO/IEC TR 20547-2:2018: cung cấp một tập hợp các trường hợp sử dụng dữ liệu lớn và phân loại các trường hợp sử dụng đó thành các gợi ý cân nhắc về kỹ thuật mà kiến trúc sư dữ liệu lớn và người triển khai hệ thống có thể xem xét.
* Tài liệu ISO/IEC 20547-3:2020: mô tả kiến trúc tham chiếu theo góc nhìn người dùng và góc nhìn chức năng. Kiến trúc sư hệ thống có thể sử dụng các góc nhìn đó để mô tả, xây dựng các hệ thống dữ liệu lớn cụ thể.
* Tài liệu ISO/IEC 20547-4:2020: mô tả các khía cạnh bảo mật và quyền riêng tư của dữ liệu lớn.
* Tài liệu ISO/IEC TR 20547-5: 2020: mô tả các tiêu chuẩn liên quan đến dữ liệu lớn, cả đang tồn tại và đang xây dựng, cùng với các ưu tiên xây dựng tiêu chuẩn dữ liệu lớn trong tương lai dựa trên phân tích sự thiếu hụt tiêu chuẩn.

Trong đó, tài liệu ISO/IEC 20547-3:2020 là tài liệu chính, mô tả về kiến trúc tham chiếu dữ liệu lớn. Kiến trúc tham chiếu được trình bày trong tài liệu này cung cấp các khung kiến trúc tổng quát để mô tả các thành phần, quy trình và hệ thống dữ liệu lớn nhằm thiết lập một ngôn ngữ chung cho các bên liên quan khác nhau, tham gia trong lĩnh vực dữ liệu lớn. Kiến trúc tham chiếu trong tài liệu này không đại diện cho kiến trúc dữ liệu lớn của một hệ thống cụ thể nào. Thay vào đó, nó là một công cụ để mô tả, thảo luận và phát triển các kiến trúc hệ thống cụ thể bằng cách sử dụng một khung tham chiếu kiến trúc. Tài liệu được xây dựng gồm các nội dung chính sau:

* Các quy ước, ký hiệu mô tả trong các sơ đồ, hình vẽ;
* Các khái niệm và nguyên tắc cơ bản; các góc nhìn kiến trúc; các thành phần vai trò/vai trò phụ, hoạt động, thành phần chức năng, các khía cạnh xuyên suốt và mối quan hệ giữa chúng trong kiến trúc hệ thống dữ liệu lớn;
* Góc nhìn người dùng phân loại các hoạt động liên quan đến dữ liệu lớn thành 03 nhóm chính: cung cấp dữ liệu, cung cấp dịch vụ phân tích dữ liệu lớn và sử dụng dữ liệu lớn. Các vai trò chính của dữ liệu lớn cũng được xác định, gồm: nhà cung cấp ứng dụng dữ liệu lớn, nhà cung cấp khung chức năng xử lý dữ liệu lớn, đối tác dịch vụ dữ liệu lớn, nhà cung cấp dữ liệu lớn, người dùng dữ liệu lớn.
* Góc nhìn chức năng mô tả dữ liệu lớn dưới dạng tập hợp ở mức cao của các lớp thành phần chức năng. Kiến trúc chức năng mô tả các thành phần chức năng dưới dạng kiến trúc phân lớp, trong dó các chức năng cụ thể được nhóm thành từng lớp, gồm có: lớp ứng dụng dữ liệu lớn, lớp xử lý dữ liệu lớn, lớp nền tảng dữ liệu lớn, lớp cơ sở hạ tầng dữ liệu lớn; các chức năng đa lớp cũng được xác định gồm: lớp tích hợp, lớp bảo mật và quyền riêng tư và lớp quản lý hệ thống.
* Các khía cạnh xuyên suốt được mô tả là cấu phần của cả góc nhìn người dùng và góc nhìn chức năng trong hệ thống dữ liệu lớn. Trong góc nhìn người dùng, các khía cạnh xuyên suốt ảnh hưởng đến các vai trò/vai trò phụ và tác động trực tiếp hoặc gián tiếp tới các hoạt động mà chúng thực hiện. Trong góc nhìn chức năng, các khía cạnh xuyên suốt tác động đến các thành phần chức năng. Các khía cạnh xuyên suốt được xác định, gồm: tính bảo mật và quyền riêng tư, quản lý và quản trị dữ liệu.

**Tài liệu ITU-T Y.3605 của Liên minh Viễn thông Quốc tế (ITU)** được xây dựng bởi bộ phận Tiêu chuẩn hóa viễn thông ITU (ITU-T) và có vai trò khuyến nghị trong công tác xây dựng, phát triển kiến trúc hệ thống dữ liệu lớn. Tài liệu này cung cấp mô tả về các khái niệm kiến trúc tham chiếu, góc nhìn người dùng, góc nhìn chức năng và các khía cạnh xuyên suốt của hệ thống dữ liệu lớn. Các nội dung chính của tài liệu gồm có:

* Các quy ước được sử dụng trong các sơ đồ, mô hình minh họa kiến trúc tham chiếu dữ liệu lớn trong tài liệu;
* Các khái niệm, góc nhìn về kiến trúc tham chiếu dữ liệu lớn trong tài liệu được xác định gồm: các góc nhìn kiến trúc, mối quan hệ giữa các góc nhìn;
* Góc nhìn người dùng mô tả hệ sinh thái dữ liệu lớn, các vai trò/vai trò phụ, xác định các hoạt động cần thiết của các vai trò/vai trò phụ để cung cấp, sử dụng các dịch vụ dữ liệu lớn, cũng như các mối quan hệ giữa các vai trò. Hệ sinh thái dữ liệu lớn được xác định trong góc nhìn người dùng gồm có: nhà cung cấp dữ liệu, nhà cung cấp dịch vụ dữ liệu lớn, người dùng dịch vụ dữ liệu lớn.
* Góc nhìn chức năng mô tả các chức năng cần thiết để hỗ trợ các hoạt động dữ liệu lớn. Góc nhìn này mô tả khung phân lớp cho dữ liệu lớn và các thành phần chức năng của chúng. Các lớp chức năng được xác định gồm có: lớp truy cập, lớp ứng dụng, lớp xử lý, lớp nguồn dữ liệu; chức năng nhiều lớp được xác định gồm có: các hệ thống hỗ trợ phân tích, quản lý vòng đời dữ liệu và bảo mật. Các thành phần chức năng của các lớp chức năng được xác định gồm có:
* Lớp truy cập: chức năng kiểm soát truy cập;
* Lớp ứng dụng: chức năng phân tích và trực quan hóa;
* Lớp xử lý: chức năng tổng hợp, làm sạch và trích xuất, chuyển đổi, nạp/tải dữ liệu;
* Lớp nguồn dữ liệu: chức năng thu thập dữ liệu, lưu trữ dữ liệu;
* Chức năng nhiều lớp:
* Hệ thống hỗ trợ phân tích: chức năng danh mục dữ liệu, quản lý quy trình làm việc, giám sát và báo cáo;
* Quản lý vòng đời dữ liệu: chức năng quản lý xuất xứ dữ liệu, quản lý cấp phép dữ liệu, quản lý bảo quản dữ liệu;
* Bảo mật: chức năng quản lý quyền riêng tư, quản lý xác thực, quản lý ủy quyền;
* Các khía cạnh xuyên suốt có thể được chia sẻ và tác động đến nhiều vai trò, hoạt động và các thành phần chức năng, gồm có: dữ liệu dự phòng, hiệu năng và chất lượng dữ liệu;
* Các yêu cầu về bảo mật liên quan được tham chiếu đến các tài liệu đã ban hành của ITU-T bao gồm: [b-ITU-T Y.2201], [b-ITU-T Y.2701], [b-ITU-T X.1751], [b-ITU-T X.1750].

**Bộ tài liệu NIST.SP.1500 của Viện Tiêu chuẩn và Công nghệ Quốc gia Mỹ (NIST)** có tên là Khung khả năng tương tác dữ liệu lớn được xây dựng bởi nhóm công tác về dữ liệu lớn của NIST (NBD-PWG). NBD-PWG đã xây dựng bộ tài liệu gồm 07 tập:

* NIST.SP.1500-1: Định nghĩa;
* NIST.SP.1500-2: Phân loại;
* NIST.SP.1500-3: Các trường hợp sử dụng và yêu cầu chung;
* NIST.SP.1500-4: Bảo mật và quyền riêng tư;
* NIST.SP.1500-5: Khảo sát kiến trúc sách trắng;
* NIST.SP.1500-6: Kiến trúc tham chiếu;
* NIST.SP.1500-7: Lộ trình tiêu chuẩn.

Trong đó, tài liệu NIST.SP.1500-6 là tài liệu chính về kiến trúc tham chiếu dữ liệu lớn. Tài liệu này cung cấp mô hình khái niệm về công nghệ, cơ sở hạ tầng và các vấn đề liên quan trong kiến trúc tham chiếu dữ liệu lớn. Mô hình khái niệm này được xây dựng bằng phương pháp khảo sát kiến trúc hệ thống dữ liệu lớn đã được triển khai của các tổ chức, đơn vị hoạt động trong lĩnh vực công nghệ để đảm bảo một cách tiếp cận sản phẩm thực tiễn và đa dạng. Kiến trúc tham chiếu dữ liệu lớn của NIST (NBDRA) nắm bắt hai chiều chuỗi giá trị kinh tế dữ liệu lớn là: thông tin (nơi giá trị được tạo ra bằng cách thu thập, tích hợp, phân tích và áp dụng kết quả vào các dịch vụ dữ liệu) và công nghệ thông tin ( nơi giá trị được tạo ra bằng cách cung cấp mạng kết nối, cơ sở hạ tầng, nền tảng và công cụ hỗ trợ các ứng dụng dựa trên dữ liệu dọc). Các nội dung chính của tài liệu như sau:

* Yêu cầu của kiến trúc tham chiếu ở mức cao: đòi hỏi kiến thức sâu về các vấn đề kỹ thuật và các mối quan tâm hiện tại đối với kiến trúc hệ thống dữ liệu lớn. Bao gồm:
* Các trường hợp và yêu cầu sử dụng;
* Khảo sát kiến trúc hệ thống dữ liệu lớn;
* Phân loại.
* Mô hình khái niệm NBDRA đại diện cho một hệ thống dữ liệu lớn, bao gồm:
* Năm thành phần chức năng logic được kết nối bằng các giao diện tương tác là: hệ thống điều phối, nhà cung cấp ứng dụng dữ liệu lớn, nhà cung cấp khung chức năng xử lý dữ liệu lớn, nhà cung cấp dữ liệu và người dùng dữ liệu;
* Hai cấu trúc bao bọc các thành phần, đại diện cho bản chất đan xen giữa khả năng quản lý, bảo mật và quyền riêng tư;
* Thành phần chức năng của NBDRA được chi tiết hóa thành các chức năng cụ thể:
* Hệ thống điều phối: Xác định và tích hợp các hoạt động ứng dụng dữ liệu cần thiết vào một hệ thống hoạt động theo chiều dọc;
* Nhà cung cấp dữ liệu: Đưa các nguồn cấp dữ liệu hoặc thông tin mới vào hệ thống dữ liệu lớn;
* Nhà cung cấp ứng dụng dữ liệu lớn: thực hiện vòng đời dữ liệu để đáp ứng các yêu cầu về tính bảo mật và quyền riêng tư, cũng như các yêu cầu do hệ thống điều phối xác định;
* Nhà cung cấp khung chức năng xử lý dữ liệu lớn: thiết lập khung chức năng tính toán để thực thi các ứng dụng chuyển đổi nhất định, đồng thời bảo vệ tính riêng tư và tính toàn vẹn của dữ liệu;
* Người dùng dữ liệu: gồm người dùng cuối hoặc các hệ thống khác sử dụng kết quả của dữ liệu lớn.
* Cấu trúc quản lý của NBDRA: các đặc điểm của dữ liệu lớn về khối lượng, tốc độ, sự đa dạng và khả năng biến đổi đòi hỏi một nền tảng quản lý linh hoạt để lưu trữ, xử lý và quản lý dữ liệu. Cấu trúc quản lý của NBDRA bao gồm hai nhóm hoặt động chung: Quản lý hệ thống và quản lý vòng đời dữ liệu lớn. Cụ thể:
* Quản lý hệ thống: bao gồm các hoạt động như cung cấp, cấu hình, quản lý gói, quản lý phần mềm, quản lý sao lưu, quản lý tính năng, quản lý tài nguyên và quản lý hiệu suất;
* Quản lý vòng đời dữ liệu lớn: liên quan đến các hoạt động xoay quanh vòng đời dữ liệu như thu thập, chuẩn bị/quản lý, phân tích, trực quan hóa và truy cập.
* Cấu trúc bảo mật và quyền riêng tư của NBDRA: bảo mật và quyền riêng tư là một khía cạnh cơ bản, bao hàm và ảnh hưởng tới năm thành phần chính của NBDRA. Các chi tiết bổ sung về cấu trúc bảo mật và quyền riêng tư được trình bày, diễn giải cụ thể trong NIST.SP.1500-4: Bảo mật và quyền riêng tư.

## 4.2 Lựa chọn tài liệu tham chiếu chính để xây dựng tiêu chuẩn

Tại Việt Nam, chiến lược chuyển đổi số đã được ban hành và xu thế chuyển đổi số đang diễn ra rất nhanh chóng theo xu hướng chung trên toàn thế giới. Do vậy, việc sớm ban hành các tiêu chuẩn trong lĩnh vực dữ liệu lớn và đặc biệt là kiến trúc tham chiếu dữ liệu lớn là vô cùng cần thiết, nó sẽ cung cấp cho các tổ chức, đơn vị hoạt động trong lĩnh vực công nghệ, các kiến trúc sư hệ thống một khuôn mẫu chung, thống nhất và đồng bộ trong việc xây dựng giải pháp cho các hệ thống dữ liệu lớn tại Việt Nam.

Để thực hiện việc xây dựng và ban hành tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn tại Việt Nam thì việc lựa chọn tài liệu tiêu chuẩn, làm cơ sở tham chiếu là hết sức quan trọng. Tài liệu được chọn không những cần đáp ứng được các tiêu chuẩn chung về kiến trúc tham chiếu dữ liệu lớn, mà còn phải phù hợp với các đặc thù của công nghệ, kỹ thuật và lộ trình chuẩn hóa tại Việt Nam.

Căn cứ vào tình hình xây dựng tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn và các tài liệu tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn của các tổ chức uy tín trên thế giới đã được trình bày ở nội dung 4.1. Nhóm nghiên cứu đã tổng hợp, phân tích và đánh giá như sau:

**Bảng tổng hợp, so sánh các tài liệu tiêu chuẩn Kiến trúc tham chiếu dữ liệu lớn trên thế giới**

|  |  |  |  |
| --- | --- | --- | --- |
| **Các tiêu chí** | **Tài liệu ISO 20547-3:2020** | **Tài liệu ITU-T Y.3605** | **Tài liệu NIST NIST.SP.1500-6** |
| Phương pháp xây dựng | được xây dựng trên cơ sở kế thừa, vận dụng một loạt các tiêu chuẩn cùng lĩnh vực đã được ban hành trước đó của ISO. | được xây dựng trên cơ sở kế thừa, vận dụng các tiêu chuẩn đã được ban hành trước đó của ITU. | Khảo sát, thu thập sách trắng về kiến trúc hệ thống dữ liệu lớn của các doanh nghiệp, tổ chức hoạt động công nghệ, các đơn vị nghiên cứu công nghệ cao tại Mỹ. |
| Mô hình, góc nhìn kiến trúc | 02 góc nhìn kiến trúc | 02 góc nhìn kiến trúc | 01 Mô hình khái niệm |
| **- Góc nhìn người dùng**: các thành phần tham gia, các vai trò/vai trò phụ, các hoạt động của chúng.**- Góc nhìn chức năng:** mô tả kiến trúc của các chức năng cần thiết trong hoạt động của dữ liệu lớn. | **- Góc nhìn người dùng**: các thành phần tham gia, các vai trò, hoạt động của chúng.**- Góc nhìn chức năng:** mô tả kiến trúc của các chức năng cần thiết trong hoạt động của dữ liệu lớn. | **Mô hình khái niệm** với các thành phần chức năng logic và mối quan hệ giữa chúng được thể hiện bằng các giao diện tương tác. |
| Thành phần, vai trò, chức năng | **Góc nhìn người dùng:** 05 thành phần vai trò, trong đó:- Đối tác dịch vụ dữ liệu lớn: 03 vai trò phụ, 10 hoạt động chính;- Nhà cung cấp ứng dụng dữ liệu lớn: 05 vai trò phụ, 11 hoạt động chính;- Nhà cung cấp khung dữ liệu lớn: 03 vai trò phụ, 07 hoạt động chính;- Nhà cung cấp dữ liệu lớn: 02 hoạt động chính;- Người dùng dữ liệu lớn: 02 hoạt động chính. | **Góc nhìn người dùng:** 03 thành phần vai trò, trong đó:- Nhà cung cấp dữ liệu: 02 vai trò phụ, 09 hoạt động chính;- Nhà cung cấp dịch vụ dữ liệu lớn: 08 hoạt động chính;- Người dùng dịch vụ dữ liệu lớn: 03 hoạt động chính. | **05 thành phần chức năng chính:**- Nhà điều hành hệ thống: 08 hoạt động chính;- Nhà cung cấp dữ liệu: 07 hoạt động chính;- Nhà cung cấp ứng dụng dữ liệu lớn: 03 hoạt động chính;- Nhà cung cấp khung dữ liệu lớnL 03 hoạt động chính- Người dùng dữ liệu lớn: 04 hoạt động chính. |
|  | **Góc nhìn chức năng:** 04 lớp chức năng và 01 chức năng đa lớp.- Lớp ứng dụng dữ liệu lớn: 04 chức năng- Lớp xử lý dữ liệu lớn: 02 chức năng- Lớp nền tảng dữ liệu lớn: 07 chức năng- Lớp cơ sở hạ tầng dữ liệu lớn: 02 chức năng- Chức năng đa lớp:* Lớp Tích hợp: 02 chức năng
* Lớp Bảo mật và quyền riêng tư: 04 chức năng

- Lớp Quản lý hệ thống: 05 chức năng | **Góc nhìn chức năng:** 04 lớp chức năng và 01 chức năng đa lớp**.**- Lớp truy cập: 01 chức năng- Lớp ứng dụng: 02 chức năng- Lớp xử lý: 03 chức năng- Lớp nguồn dữ liệu: 02 chức năng- Chức năng đa lớp:* Lớp Hệ thống hỗ trợ phân tích: 03 chức năng
* Lớp Quản lý vòng đời dữ liệu: 03 chức năng

- Lớp Bảo mật: 03 chức năng | **02 thành phần cấu trúc NBDRA:**- Cấu trúc quản lý NBDRA: 02 hoạt động chính- Cấu trúc bảo mật và quyền riêng tư: được trình bày cụ thể trong NIST.SP.1500-4 |

***Tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn của NIST:*** được tiến hành xây dựng từ những năm 2015, đến nay NBDRA đã xuất bản thêm 02 phiên bản cập nhật vào các năm 2018, 2019. NBDRA đưa ra 01 mô hình khái niệm và 05 thành phần chức năng chính. Về cơ bản NBDRA đã đảm bảo được các yếu tố cần thiết của một kiến trúc tham chiếu dữ liệu lớn. Tuy nhiên, NBDRA được xây dựng theo phương pháp thu thập các trường hợp sử dụng và tiến hành khảo sát sách trắng về kiến trúc hệ thống dữ liệu lớn của các tổ chức, đơn vị lớn hoạt động trong lĩnh vực nghiên cứu, phát triển công nghệ cao tại Mỹ. Phương pháp này tuy giúp cho kiến trúc tham chiếu dữ liệu lớn của NIST có được tính thực tiễn và ứng dụng cao, nhưng lại mang những đặc thù riêng, phù hợp với hạ tầng công nghệ vô cùng phát triển của nước Mỹ. Bên cạnh đó, việc chỉ đưa ra mô hình khái niệm khiến cho kiến trúc tham chiếu của NIST thiếu đi góc nhìn toàn diện, đa chiều, điều vô cùng quan trọng đối với một tiêu chuẩn kiến trúc tham chiếu. Do vậy, nhóm nghiên cứu đánh giá tài liệu này không phù hợp với việc tham chiếu xây dựng tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn tại Việt Nam.

***Tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn của ISO:*** được xây dựng bởi tiểu ban SC42 của Ủy ban kỹ thuật chung ISO/IEC JTC1 chuyên về việc xây dựng tiêu chuẩn trong lĩnh vực trí tuệ nhân tạo và dữ liệu lớn của Tổ chức tiêu chuẩn hóa quốc tế (ISO), ban hành năm 2020. Tài liệu này được xây dựng trên cơ sở kế thừa, vận dụng một loạt các tiêu chuẩn cùng lĩnh vực đã được ban hành trước đó của ISO như: ISO 8000-2, Data quality – Part 2: Vocabulary; ISO/TS 8000-60, Data quality – Part 60: Data quality management: Overview; ISO 8000-61, Data quality – Part 61: Data quality management: process reference model; ISO/IEC/IEEE 42010, Systems and software engineering — Architecture description; ISO/IEC 20546, Information technology — Big data — Overview and vocabulary…Do vậy, tài liệu đã đảm bảo được tính đồng bộ và nhất quán trong lộ trình xây dựng và phát triển tiêu chuẩn. Về nội dung, tài liệu ISO/IEC 20547-3 mô tả rất đầy đủ và toàn diện về các góc nhìn kiến trúc, các vai trò, thành phần chức năng cấu thành nên một kiến trúc tham chiếu dữ liệu lớn đầy đủ. Đặc biệt, tài liệu này xác định phạm vi nằm ngoài các yếu tố kỹ thuật đối với các hệ thống dữ liệu lớn cụ thể nên không bị chi phối bởi nền tảng cơ sở hạ tầng công nghệ của từng quốc gia. Do vậy, nhóm nghiên cứu đánh giá đây là tài liệu phù hợp nhất để làm cơ sở tham chiếu, xây dựng tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn tại Việt Nam.

***Tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn của ITU:*** được xây dựng bởi bộ phận Tiêu chuẩn hóa viễn thông (ITU-T), ban hành năm 2020. Cũng giống như tài liệu ISO/IEC 20547-3:2020, tài liệu ITU-T Y.3605 cũng được xây dựng trên cơ sở kế thừa, vận dụng các tiêu chuẩn đã được ban hành trước đó của ITU như: Recommendation ITU-T Y.3600 (2015), Big data – cloud computing based requirements and capabilities; Recommendation ITU-T Y.3601 (2018), Big data – framework and requirements for data exchange; Recommendation ITU-T Y.3603 (2019), Big data – requirements and conceptual model of metadata for data catalogue; Recommendation ITU-T Y.3604 (2020), Big data – overview and requirements for data preservation. Về nội dung, tài liệu ITU-T Y.3605 cũng xác định khuyến nghị nằm ngoài các yếu tố, đặc điểm kỹ thuật của một hệ thống cụ thể và trình bày các góc nhìn kiến trúc, các vai trò/vai trò phụ, hoạt động và các thành phần chức năng như tài liệu ISO/IEC 20547-3:2020. Tuy nhiên, mức độ chi tiết của các nội dung không được cụ thể, mạch lạc và đầy đủ như tài liệu của ISO. Do vậy, nhóm nghiên cứu đánh giá tài liệu này cũng ở mức tương đối, mức độ phù hợp kém hơn so với tài liệu tiêu chuẩn ISO/IEC 20547-3:2020 của ISO.

Bên cạnh đó, Bộ Khoa học và Công nghệ Việt nam hiện cũng đã ban hành một số tiêu chuẩn trong lĩnh vực dữ liệu lớn dựa trên nguyên tắc chấp thuận nguyên vẹn các tài liệu tiêu chuẩn của Tổ chức Tiêu chuẩn hóa quốc tế (ISO). Cụ thể gồm:

* Tiêu chuẩn Công nghệ thông tin – Dữ liệu lớn – Tổng quan và từ vựng (TCVN 13238:2020): tiêu chuẩn này được xây dựng hoàn toàn tương đương với tài liệu tiêu chuẩn ISO/IEC 20546:2019.
* Tiêu chuẩn Công nghệ thông tin – Kiến trúc tham chiếu dữ liệu lớn – Phần 2: Các trường hợp sử dụng và yêu cầu dẫn suất (TCVN 13239-2:2020): tiêu chuẩn này được xây dựng hoàn toàn tương đương với tài liệu tiêu chuẩn ISO/IEC TR 20547-2:2018.
* Tiêu chuẩn Công nghệ thông tin – Kiến trúc tham chiếu dữ liệu lớn – Phần 5: Lộ trình tiêu chuẩn (TCVN 13239-5:2020): tiêu chuẩn này được xây dựng hoàn toàn tương đương với tài liệu tiêu chuẩn ISO/IEC TR 20547-5:2018.

Do vậy, xét về mức độ phù hợp, chi tiết của nội dung cũng như để đảm bảo tính nhất quán, đồng bộ trong lộ trình xây dựng và phát triển tiêu chuẩn dữ liệu lớn tại Việt Nam, nhóm nghiên cứu đề xuất, lựa chọn tài liệu tiêu chuẩn ISO/IEC 20547-3:2020 làm tài liệu tham chiếu để xây dựng Dự thảo tiêu chuẩn Kiến trúc tham chiếu dữ liệu lớn tại Việt Nam.

# 5. Xây dựng dự thảo tiêu chuẩn kiến trúc tham chiếu dữ liệu lớn

## 5.1 Phương pháp xây dựng

Dự thảo tiêu chuẩn được xây dựng bằng cách sử dụng phương pháp chấp thuận nguyên vẹn các nội dung của tiêu chuẩn quốc tế ISO/IEC 20547-3:2020 và được bố cục và trình bày theo đúng các quy định hiện hành của Việt Nam tại Thông tư số 03/2011/TT-BTTTT ngày 4/01/2011 quy định hoạt động xây dựng quy chuẩn kỹ thuật quốc gia và tiêu chuẩn quốc gia thuộc Bộ Thông tin và Truyền thông cũng như các văn bản quy phạm pháp luật có liên quan về xây dựng tiêu chuẩn, quy chuẩn kỹ thuật quốc gia và phù hợp Tiêu chuẩn Việt Nam TCVN 1-2:2008 quy định về trình bày và thể hiện nội dung tiêu chuẩn quốc gia.

## 5.2 Nội dung dự thảo tiêu chuẩn

Nội dung của Dự thảo tiêu chuẩn **“Kiến trúc tham chiếu dữ liệu lớn”**, được biên soạn hoàn toàn tương đương với tài liệu ISO/IEC 20547-3:2020, bao gồm các nội dung chính:

**Tên dự thảo TCVN: Kiến trúc tham chiếu dữ liệu lớn.**

**Nội dung gồm:**

**1 Phạm vi áp dụng**

**2 Tài liệu tham khảo**

**3 Thuật ngữ và định nghĩa**

**4 Thuật ngữ viết tắt**

**5 Quy ước**

**6 Các khái niệm kiến trúc tham chiếu dữ liệu lớn**

*6.1 Khái quát chung*

*6.2 Các góc nhìn*

*6.3 Tổng quan về góc nhìn người dùng*

*6.4 Tổng quan về góc nhìn chức năng*

*6.5 Mối quan hệ giữa góc nhìn người dùng và 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*

**7 Góc nhìn người dùng**

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

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

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

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

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

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

**8 Các khía cạnh xuyên suốt**

 *8.1 Khái quát chung*

 *8.2 Bảo mật và quyền riêng tư*

 *8.3 Quản lý*

 *8.4 Quản trị dữ liệu*

**9 Góc nhìn chức năng**

 *9.1 Kiến trúc chức năng*

 *9.2 Các thành phần chức năng*

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

**Phụ lục B:** Các ví dụ về mối quan hệ của các vai trò trong hệ sinh thái dữ liệu lớn

**Phụ lục C:**

**Bảng đối chiếu nội dung TCVN và tài liệu tham khảo**

|  |  |  |
| --- | --- | --- |
| **Dự thảo TCVN** | **Tài liệu gốc ISO/IEC 20547-3:2020** | **Hình thức xây dựng** |
| **1. Phạm vi áp dụng** | ISO/IEC 20547-3:2020, Mục 1 | Chấp thuận nguyên vẹn |
| **2. Tài liệu tham khảo** | ISO/IEC 20547-3:2020, Mục 2 | Chấp thuận có sửa đổi, bổ sung một số tài liệu TCVN: TCVN 10249-2:2013 (ISO 8000-2), TCVN 13238:2020 (ISO/IEC 20546), TCVN 12481:2019 (ISO/IEC 17789) |
| **3. Thuật ngữ và định nghĩa** | ISO/IEC 20547-3:2020, Mục 3 | Chấp thuận có sửa đổi, bổ sung một số thuật ngữ, định nghĩa theo các TCVN đã ban hành: 3.1; 3.4; 3.5; 3.6; 3.9; 3.10; 3.12 và 3.14. |
| **4. Thuật ngữ viết tắt** | ISO/IEC 20547-3:2020, Mục 4 | Chấp thuận nguyên vẹn |
| **5. Quy ước** | ISO/IEC 20547-3:2020, Mục 5 | Chấp thuận nguyên vẹn |
| **6. Các khái niệm kiến trúc tham chiếu dữ liệu lớn** | ISO/IEC 20547-3:2020, Mục 6 | Chấp thuận nguyên vẹn |
| 6.1. Khái quát chung | ISO/IEC 20547-3:2020, Mục 6.1 | Chấp thuận nguyên vẹn |
| 6.2. Các góc nhìn | ISO/IEC 20547-3:2020, Mục 6.2 | Chấp thuận nguyên vẹn |
| 6.3. Tổng quan về góc nhìn người dùng | ISO/IEC 20547-3:2020, Mục 6.3 | Chấp thuận nguyên vẹn |
| 6.4. Tổng quan về góc nhìn chức năng | ISO/IEC 20547-3:2020, Mục 6.4 | Chấp thuận nguyên vẹn |
| 6.5. Mối quan hệ giữa góc nhìn người dùng và góc nhìn chức năng | ISO/IEC 20547-3:2020, Mục 6.5 | Chấp thuận nguyên vẹn |
| 6.6. Mối quan hệ của góc nhìn người dùng và góc nhìn chức năng với các khía cạnh xuyên suốt | ISO/IEC 20547-3:2020, Mục 6.6 | Chấp thuận nguyên vẹn |
| **7. Góc nhìn người dùng** | ISO/IEC 20547-3:2020, Mục 7 | Chấp thuận nguyên vẹn |
| 7.1. Các vai trò, vai trò phụ và hoạt động của dữ liệu lớn | ISO/IEC 20547-3:2020, Mục 7.1 | Chấp thuận nguyên vẹn |
| 7.2. Vai trò: Đơn vị cung cấp ứng dụng dữ liệu lớn | ISO/IEC 20547-3:2020, Mục 7.2 | Chấp thuận nguyên vẹn |
| 7.3. Vai trò: Đơn vị cung cấp khung xử lý dữ liệu lớn | ISO/IEC 20547-3:2020, Mục 7.3 | Chấp thuận nguyên vẹn |
| 7.4. Vai trò: Đối tác dịch vụ dữ liệu lớn | ISO/IEC 20547-3:2020, Mục 7.4 | Chấp thuận nguyên vẹn |
| 7.5. Vai trò: Đơn vị cung cấp dữ liệu lớn | ISO/IEC 20547-3:2020, Mục 7.5 | Chấp thuận nguyên vẹn |
| 7.6. Vai trò: Người dùng dữ liệu lớn | ISO/IEC 20547-3:2020, Mục 7.6 | Chấp thuận nguyên vẹn |
| **8. Các khía cạnh xuyên suốt** | ISO/IEC 20547-3:2020, Mục 8 | Chấp thuận nguyên vẹn |
| 8.1. Khái quát chung | ISO/IEC 20547-3:2020, Mục 8.1 | Chấp thuận nguyên vẹn |
| 8.2. Bảo mật và quyền riêng tư | ISO/IEC 20547-3:2020, Mục 8.2 | Chấp thuận nguyên vẹn |
| 8.3. Quản lý | ISO/IEC 20547-3:2020, Mục 8.3 | Chấp thuận nguyên vẹn |
| 8.4. Quản trị dữ liệu | ISO/IEC 20547-3:2020, Mục 8.4 | Chấp thuận nguyên vẹn |
| **9. Góc nhìn chức năng** | ISO/IEC 20547-3:2020, Mục 9 | Chấp thuận nguyên vẹn |
| 9.1. Khái quát chung | ISO/IEC 20547-3:2020, Mục 9.1 | Chấp thuận nguyên vẹn |
| 9.2. Các thành phần chức năng | ISO/IEC 20547-3:2020, Mục 9.2 | Chấp thuận nguyên vẹn |
| **Phụ lục A:** Ánh xạ góc nhìn chức năng của kiến trúc tham chiếu dữ liệu lớn sang kiến trúc tham chiếu tích hợp hệ thống khác | ISO/IEC 20547-3:2020, Annex A (informative) | Chấp thuận nguyên vẹn |
| **Phụ lục B:** Các ví dụ về mối quan hệ của các vai trò trong hệ sinh thái dữ liệu lớn | ISO/IEC 20547-3:2020, Annex B (informative) | Chấp thuận nguyên vẹn |
| **Phụ lục C:** | ISO/IEC 20547-3:2020, Annex C (informative) | Chấp thuận nguyên vẹn |
| **Thư mục viện dẫn** | ISO/IEC 20547-3:2020, Bibliography | Chấp thuận nguyên vẹn |

# Tài liệu tham khảo

1. ISO/IEC 20547-3:2020, Information technology – Big data reference architecture – Part 3: Reference architecture
2. NIST.SP.1500-6, Nist big data interoperability framework: Volume 6, Reference Architecture
3. ITU-T Y.3605, Big data – Reference architecture