ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

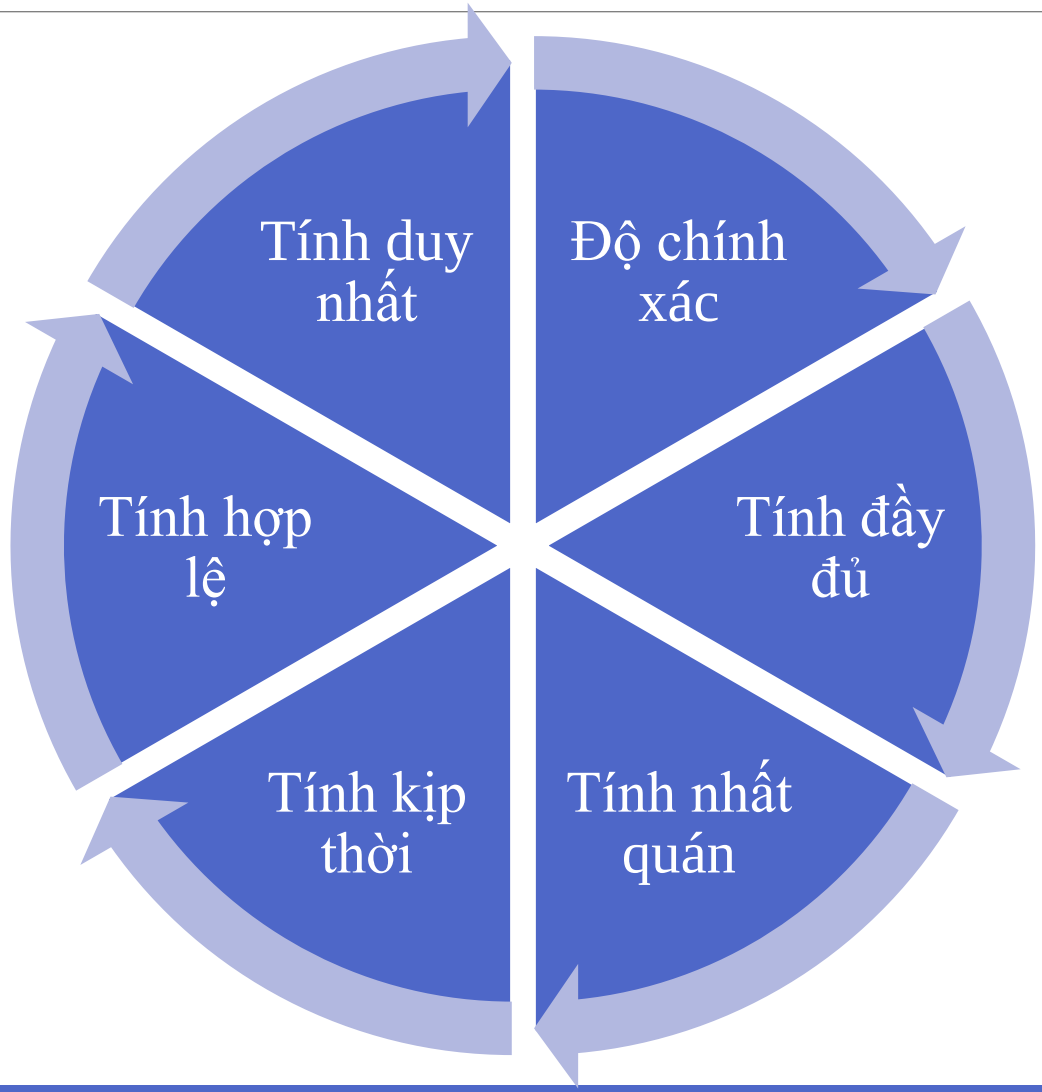
# CHƯƠNG 4
# Chất lượng dữ liệu

Biên soạn: ThS. Nguyễn Thị Anh Thư

# Nội dung

1. Tổng quan
2. Khuôn khổ và độ đo
3. Phương pháp quản lý chất lượng
4. Quản lý vòng đời dữ liệu
5. Vấn đề và xu hướng
6. Bài tập
7. Tổng kết

# 1. Tổng quan




**Chất lượng dữ liệu (Data quality)** là sự phù hợp cho việc sử dụng dữ liệu. Đề cập đến mức độ dữ liệu đáp ứng nhu cầu dự kiến.

**Chất lượng dữ liệu** có thể bao gồm nhiều *khía cạnh (dimension)* khác nhau, tùy thuộc vào trường hợp sử dụng cụ thể. Một số khía cạnh phổ biến như hình bên:

# 1. Tổng quan

**Khía cạnh dữ liệu (Data Dimensions)** là các thuộc tính của chất lượng dữ liệu, có thể đo lường chính xác để chỉ ra mức chất lượng tổng thể của dữ liệu.

- **Độ chính xác (Accuracy)**: Dữ liệu có bị lỗi không? Liệu có đại diện chính xác cho thế giới thực không?
- **Tính đầy đủ (Completeness)**: Dữ liệu có bao gồm tất cả thông tin cần thiết cho mục đích sử dụng không? Có bất kỳ giá trị bị thiếu?
- **Tính nhất quán (Consistency)**: Dữ liệu có được định dạng nhất quán trong toàn bộ tập dữ liệu không? Có bất kỳ định nghĩa hoặc đơn vị xung đột nào được sử dụng không?
- **Tính kịp thời (Timeliness)**: Dữ liệu có được cập nhật và phản ánh tình trạng hiện tại không?
- **Tính hợp lệ (Validity)**: Dữ liệu có tuân thủ các quy tắc và ràng buộc đã xác định không? Ví dụ: có giá trị nào nằm ngoài phạm vi dự kiến không?
- **Tính duy nhất (Uniqueness)**: Có bất kỳ mục nhập trùng lặp nào trong dữ liệu không?

# 1. Tổng quan

**Khung chất lượng dữ liệu (data quality framework)** là một cách tiếp cận có cấu trúc để quản lý và đảm bảo chất lượng dữ liệu.

**Khung chất lượng dữ liệu** giống như một lộ trình xác định cách tổ chức xử lý dữ liệu của người quản lý hoặc tổ chức để đảm bảo độ tin cậy và tính hữu ích của dữ liệu.

Lộ trình này liên quan đến:

- **Bối cảnh dữ liệu (context of data)**
- **Hệ thống thông tin (information system)**
- **Loại hình kinh doanh (type of business)**

# 1. Tổng quan

Thành phần chính của khung chất lượng dữ liệu:

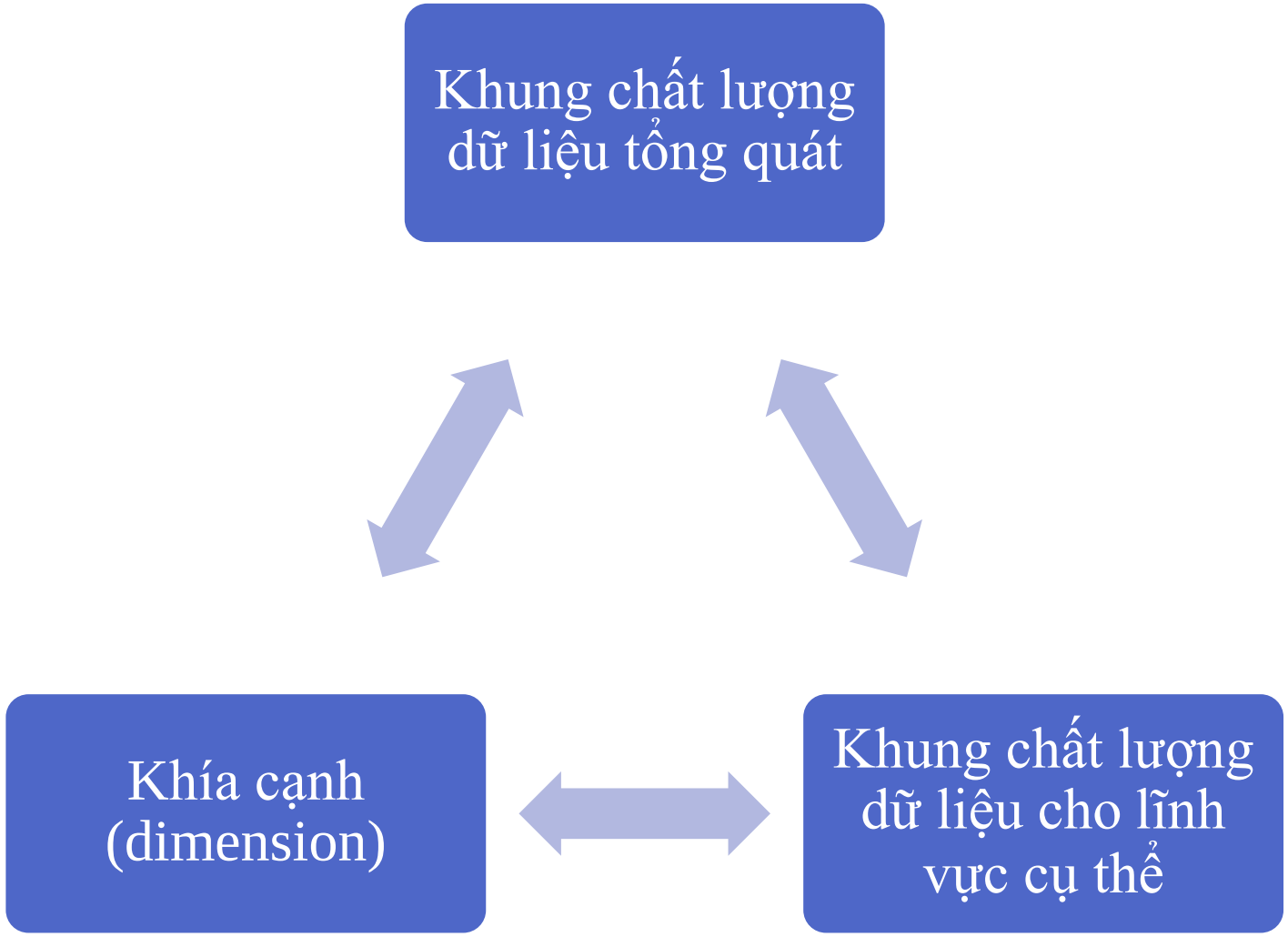

*Bằng cách triển khai khung chất lượng dữ liệu, các tổ chức có thể tự tin đưa ra quyết định dựa trên dữ liệu. Biết rằng dữ liệu là chính xác, nhất quán và đáng tin cậy.*

# 1. Tổng quan

Thành phần chính của khung chất lượng dữ liệu:

- **Định nghĩa**: Thiết lập các định nghĩa rõ ràng cho từng *khía cạnh (dimension)* của chất lượng dữ liệu phù hợp với nhu cầu của tổ chức.
- **Tiêu chuẩn**: Đặt điểm chuẩn hoặc ngưỡng cụ thể cho từng *khía cạnh (dimension)* để đo lường chất lượng dữ liệu.
- **Quy trình**: Triển khai các quy trình xác thực, làm sạch, chuyển đổi và giám sát dữ liệu để đảm bảo dữ liệu đáp ứng các tiêu chuẩn.
- **Công cụ**: Sử dụng các công cụ để tự động kiểm tra chất lượng dữ liệu, xác định và khắc phục sự cố cũng như tạo báo cáo.
- **Quản trị chất lượng dữ liệu**: Thiết lập vai trò và trách nhiệm quản lý chất lượng dữ liệu trong tổ chức (*đảm bảo dữ liệu chính xác, đầy đủ và tuân thủ các quy định*). Cung cấp các bước cải thiện chất lượng dữ liệu.

# 2. Khuôn khổ và độ đo



# Khung chất lượng dữ liệu tổng quát

| Tên viết tắt | Tên đầy đủ | Năm |
|---|---|---|
| **TDQM** [1] | Total Data Quality Management | 1998 |
| TIQM [2] | Total Information Quality Management | 1999 |
| **AIMQ** [3] | A Methodology for Information Quality assessment | 2002 |
| **DQA** [4] | Data Quality Assessment | 2002 |
| HIQM [5] | Hybrid Information Quality Management | 2006 |
| CDQ [6] | Comprehensive Data Quality | 2008 |
| DQPA [7] | A Data Quality Practical Approach | 2009 |
| HDQM [8] | Heterogeneous Data Quality Methodology | 2011 |
| DQAF [9] | Data Quality Assessment Framework | 2013 |
| **TBDQ** [10] | Task-Based Data Quality | 2016 |
| OODA DQ [11] | Data Quality Improvement Through OODA Methodology | 2017 |

# Khung chất lượng dữ liệu cho lĩnh vực cụ thể

| Tên viết tắt | Tên đầy đủ | Năm |
|---|---|---|
| **DQF4CT** [12] | Data Quality Issues Specifically For Classification Tasks | 2018 |
| VIoTF [13] | Valid.IoT Framework | 2018 |
| PPF [14] | Pre-Processing Framework | 2019 |
| **HDQF-EF** [15] | Hybrid Data Quality Framework in EF (Environmental Footprint) Tools | 2021 |
| IDQ-MDM [16] | A Framework for Improving Data Quality Throughout the MDM Implementation Process | 2022 |
| **DC-AI** [17] | A Data-centric AI Framework | 2023 |
| **ISO/IEC 25012** [18] | ISO/IEC 25012 Framework for Software Vulnerability Datasets | 2023 |
| **RWDQF** [19] | Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation | 2024 |

# Khía cạnh (dimension)

Mỗi tiêu chí có thể được đánh giá theo cả hai phương pháp chủ quan và khách quan:

| | Đánh giá chủ quan | Đánh giá khách quan |
|---|---|---|
| **Phương pháp** | • Thực hiện khảo sát lấy ý kiến đánh giá của những người tham gia vào quy trình sử dụng dữ liệu.<br>• Bảng khảo sát (Google Forms, …) về các tiêu chí (dimension) như tính chính xác, tính đầy đủ, … theo thang điểm. | • Chuyển đổi các tiêu chí (dimension) như tính chính xác, tính đầy đủ, … thành số liệu và có thể đo lường được.<br>• Tính chính xác: Có thể sử dụng các phương pháp thống kê như độ lệch chuẩn, khoảng tin cậy để đánh giá độ chính xác của dữ liệu. |

# Khía cạnh (dimension)

Mỗi tiêu chí có thể được đánh giá theo cả hai phương pháp chủ quan và khách quan:

| | Đánh giá chủ quan | Đánh giá khách quan |
|---|---|---|
| **Ưu điểm** | • Nhanh chóng và dễ thực hiện.<br>• Không yêu cầu kiến thức chuyên môn cao.<br>• Có thể đánh giá các yếu tố khó định lượng. | • Chính xác và nhất quán.<br>• Dựa trên các tiêu chí cụ thể.<br>• Có thể so sánh và đo lường. |
| **Nhược điểm** | • Phụ thuộc vào kinh nghiệm và kiến thức của người đánh giá.<br>• Có thể không chính xác và không nhất quán.<br>• Khó so sánh và đo lường. | • Tốn thời gian và công sức để thực hiện.<br>• Yêu cầu kiến thức chuyên môn cao.<br>• Khó đánh giá các yếu tố khó định lượng. |

# 3. Phương pháp quản lý chất lượng

Dựa vào Bối cảnh dữ liệu (context of data), Hệ thống thông tin (information system) và Loại hình kinh doanh (type of business) để lựa chọn khung chất lượng dữ liệu áp dụng trên các tiêu chí sau:

- Chức năng
- Có linh hoạt áp dụng cho nhiều trường hợp
- Khía cạnh (dimension) đề xuất sử dụng
- Dimension có thể chuyển thành số liệu và có thể đo lường được
- Các loại dữ liệu có thể đo lường
- Hướng dẫn các bước cụ thể để thực hiện triển khai thực tế

# Khung chất lượng dữ liệu tổng quát

| Tên viết tắt | Chức năng | Linh hoạt |
|---|---|---|
| **TDQM** | Liên tục đánh giá, cải thiện và duy trì chất lượng dữ liệu. | - |
| TIQM | Quản lý chất lượng thông tin tổng thể trong một tổ chức. | - |
| **AIMQ** | Đánh giá và cải thiện chất lượng tổng thể của thông tin trong một tổ chức. | - |
| **DQA** | Cung cấp một cách tiếp cận có hệ thống để đánh giá khách quan chất lượng dữ liệu. | Có |
| HIQM | Hỗ trợ giám sát và phục hồi chất lượng dữ liệu | Có |
| CDQ | Đánh giá và cải thiện chất lượng dữ liệu cho nguồn dữ liệu có cấu trúc và web. | Có |
| DQPA | Xác định dữ liệu liên quan và đánh giá chất lượng dữ liệu. | - |
| HDQM | Đánh giá và cải thiện chất lượng dữ liệu trong tổ chức cần quản lý nhiều loại dữ liệu. | - |
| DQAF | Đánh giá chất lượng dữ liệu thống kê. | - |
| **TBDQ** | Cải thiện chất lượng dữ liệu phù hợp với các tổ chức có cơ sở hạ tầng CNTT yếu. | Có |
| OODA DQ | Liên tục cải thiện chất lượng dữ liệu thông qua vòng lặp OODA. | - |

# Khung chất lượng dữ liệu tổng quát

| Tên viết tắt | Khía cạnh | Định lượng cụ thể | Structured | Semi-structured | Unstructured | Triển khai |
|---|---|---|---|---|---|---|
| **TDQM** | Tùy chỉnh | Có | Có | Có | Có | Có |
| TIQM | Tùy chỉnh | Có | Có | Có | Có | - |
| **AIMQ** | Tùy chỉnh | - | Có | Có | Có | Có |
| **DQA** | Xác định | Có | Có | - | - | Có |
| HIQM | Xác định | - | Có | Có | - | - |
| CDQ | Xác định | Có | Có | Có | Một phần | Có |
| DQPA | Tùy chỉnh | Có | Có | - | - | - |
| HDQM | Tùy chỉnh | Có | Có | Có | Có | Có |
| DQAF | Tùy chỉnh | Có | Có | - | - | Có |
| **TBDQ** | Xác định | Có | Có | - | - | Có |
| OODA DQ | Tùy chỉnh | - | Có | Có | Có | Có |

# Khung chất lượng dữ liệu cho lĩnh vực cụ thể

| Tên viết tắt | Chức năng | Linh hoạt |
|---|---|---|
| **DQF4CT** | Cải thiện độ chính xác của các tác vụ phân loại trong quá trình tiền xử lý. | - |
| VIoTF | Đảm bảo chất lượng và độ tin cậy của dữ liệu cảm biến trong hệ thống IoT. | Có |
| PPF | Cải thiện chất lượng dữ liệu thời tiết lớn trong giai đoạn tiền xử lý. | - |
| **HDQF-EF** | Đánh giá chất lượng dữ liệu trong các công cụ dấu chân sinh thái (EF). | - |
| IDQ-MDM | Quy trình 5 bước đảm bảo chất lượng dữ liệu trong Quản lý dữ liệu chủ (MDM). | Có |
| **DC-AI** | Tự động hóa các nhiệm vụ phân tích dữ liệu thăm dò (EDA) và chuẩn bị dữ liệu. | Có |
| **ISO/IEC 25012** | Cung cấp cấu trúc được tiêu chuẩn hóa để đánh giá chất lượng dữ liệu khi xử lý các bộ dữ liệu về lỗ hổng phần mềm. | Có |
| **RWDQF** | Đánh giá chất lượng của dữ liệu trong thế giới thực (RWD) về nghiên cứu ung thư. | - |

# Khung chất lượng dữ liệu cho lĩnh vực cụ thể

| Tên viết tắt | Khía cạnh | Định lượng cụ thể | Structured | Semi-structured | Unstructured | Triển khai |
|---|---|---|---|---|---|---|
| **DQF4CT** | - | - | Có | - | - | Có |
| VIoTF | Xác định | Có | Có | Có | - | Có |
| PPF | - | - | Có | Có | Có | Có |
| **HDQF-EF** | Tùy chỉnh | Có | Có | Có | - | Có |
| IDQ-MDM | - | - | Có | - | - | Có |
| **DC-AI** | - | - | Có | Một phần | - | - |
| **ISO/IEC 25012** | Tùy chỉnh | Có | Có | Một phần | Một phần | Có |
| **RWDQF** | Xác định | Có | Có | Có | - | Có |

# Độ chính xác – Đánh giá khách quan

**1. Phân tích thống kê:**

- **Tỷ lệ lỗi:** Tính toán tỷ lệ phần trăm dữ liệu sai so với tổng số dữ liệu.
- **Độ lệch chuẩn:** Đo lường mức độ phân tán của dữ liệu so với giá trị trung bình.
- **Hệ số tương quan:** Đánh giá mối liên hệ giữa các biến trong tập dữ liệu.

**2. So sánh với nguồn dữ liệu đáng tin cậy:**

- So sánh dữ liệu với một nguồn dữ liệu được công nhận là chính xác.
- Sử dụng phương pháp "ground truth" để xác định giá trị thực tế của dữ liệu.

**3. Xác minh dữ liệu:**

- Thực hiện kiểm tra thủ công hoặc tự động để xác định lỗi trong dữ liệu.
- Sử dụng các công cụ xác minh dữ liệu để xác định các giá trị bất thường.

**4. Phân tích độ nhạy:**

- Đánh giá mức độ ảnh hưởng của lỗi dữ liệu đến kết quả phân tích.

# Tính đầy đủ – Đánh giá khách quan

**1. Phân tích thống kê:**

- **Tỷ lệ dữ liệu thiếu:** Tính toán tỷ lệ phần trăm dữ liệu thiếu so với tổng số dữ liệu.
- **Phân tích phân bố:** Đánh giá sự phân bố của dữ liệu để xác định các giá trị bị thiếu một cách có hệ thống.

**2. So sánh với nguồn dữ liệu đáng tin cậy:**

- So sánh dữ liệu với một nguồn dữ liệu được công nhận là đầy đủ.
- Sử dụng phương pháp "ground truth" để xác định các giá trị cần thiết cho dữ liệu.

**3. Xác minh dữ liệu:**

- Thực hiện kiểm tra thủ công hoặc tự động để xác định các giá trị bị thiếu trong dữ liệu.
- Sử dụng các công cụ xác minh dữ liệu để xác định các trường dữ liệu bị thiếu.

**4. Phân tích độ nhạy:**

- Đánh giá mức độ ảnh hưởng của dữ liệu thiếu đến kết quả phân tích.

# Tính nhất quán – Đánh giá khách quan

## 1. Phân tích thống kê:

- **Kiểm tra thống kê:** Sử dụng các kiểm tra thống kê như kiểm tra chi-squared, kiểm tra t-test, v.v. để xác định xem có sự khác biệt đáng kể nào giữa các giá trị dữ liệu hay không.
- **Phân tích độ tương quan:** Đánh giá mối liên hệ giữa các thuộc tính dữ liệu để xác định xem chúng có nhất quán với nhau hay không.

## 2. So sánh với nguồn dữ liệu đáng tin cậy:

- So sánh dữ liệu với một nguồn dữ liệu được công nhận là nhất quán.
- Sử dụng phương pháp "ground truth" để xác định giá trị chính xác của dữ liệu.

## 3. Xác minh dữ liệu:

- Thực hiện kiểm tra thủ công hoặc tự động để xác định các giá trị không nhất quán trong dữ liệu.
- Sử dụng các công cụ xác minh dữ liệu để xác định các giá trị bất thường.

## 4. Phân tích độ nhạy:

- Đánh giá mức độ ảnh hưởng của các giá trị không nhất quán đến kết quả phân tích.

# Tính kịp thời – Đánh giá khách quan

## 1. Phân tích độ trễ:

- **Tính toán độ trễ trung bình:** Thời gian trung bình để dữ liệu mới được cập nhật vào hệ thống.
- **Tính toán độ trễ tối đa:** Thời gian tối đa để dữ liệu mới được cập nhật vào hệ thống.
- **Phân tích phân bố độ trễ:** Phân tích sự phân bố thời gian để dữ liệu mới được cập nhật vào hệ thống.

## 2. So sánh với thời gian thực:

- **So sánh thời gian cập nhật dữ liệu với thời gian thực tế:** Xác định xem dữ liệu có được cập nhật kịp thời hay không.
- **Sử dụng các nguồn dữ liệu thời gian thực:** So sánh dữ liệu với các nguồn dữ liệu thời gian thực để xác định độ trễ.

## 3. Phân tích ảnh hưởng của độ trễ:

- **Đánh giá mức độ ảnh hưởng của độ trễ đến việc ra quyết định.**
- **Sử dụng các mô hình mô phỏng:** Sử dụng các mô hình mô phỏng để đánh giá tác động của độ trễ đến hiệu quả hoạt động.

## 4. Sử dụng các công cụ giám sát:

- **Sử dụng các công cụ giám sát để theo dõi độ trễ của dữ liệu.**

# Tính hợp lệ – Đánh giá khách quan

**1. Phân tích thống kê:**

- **Kiểm tra tính chính quy:** Sử dụng các kiểm tra thống kê như kiểm tra Shapiro-Wilk, v.v. để xác định xem dữ liệu có phân phối chính quy hay không.
- **Kiểm tra ngoại lệ:** Sử dụng các phương pháp như phân tích IQR (Interquartile Range) hoặc z-score để xác định các giá trị ngoại lệ trong dữ liệu.
- **Phân tích tương quan:** Đánh giá mối liên hệ giữa các thuộc tính dữ liệu để xác định xem dữ liệu có hợp lý hay không.

**2. So sánh với nguồn dữ liệu đáng tin cậy:**

- So sánh dữ liệu với một nguồn dữ liệu được công nhận là hợp lệ.
- Sử dụng phương pháp "ground truth" để xác định giá trị thực của dữ liệu.

**3. Xác minh dữ liệu:**

- Thực hiện kiểm tra thủ công hoặc tự động để xác định các giá trị không hợp lệ trong dữ liệu.
- Sử dụng các công cụ xác minh dữ liệu để xác định các giá trị bất thường.

**4. Phân tích độ nhạy:**

- Đánh giá mức độ ảnh hưởng của các giá trị không hợp lệ đến kết quả phân tích.

# Tính duy nhất – Đánh giá khách quan

**1. Phân tích thống kê:**

- **Kiểm tra thống kê:** Sử dụng các kiểm tra thống kê như kiểm tra chi-squared, kiểm tra t-test, v.v. để xác định xem có sự khác biệt đáng kể nào giữa các giá trị dữ liệu hay không.
- **Phân tích độ tương quan:** Đánh giá mối liên hệ giữa các thuộc tính dữ liệu để xác định xem chúng có liên quan đến nhau hay không.

**2. So sánh với nguồn dữ liệu đáng tin cậy:**

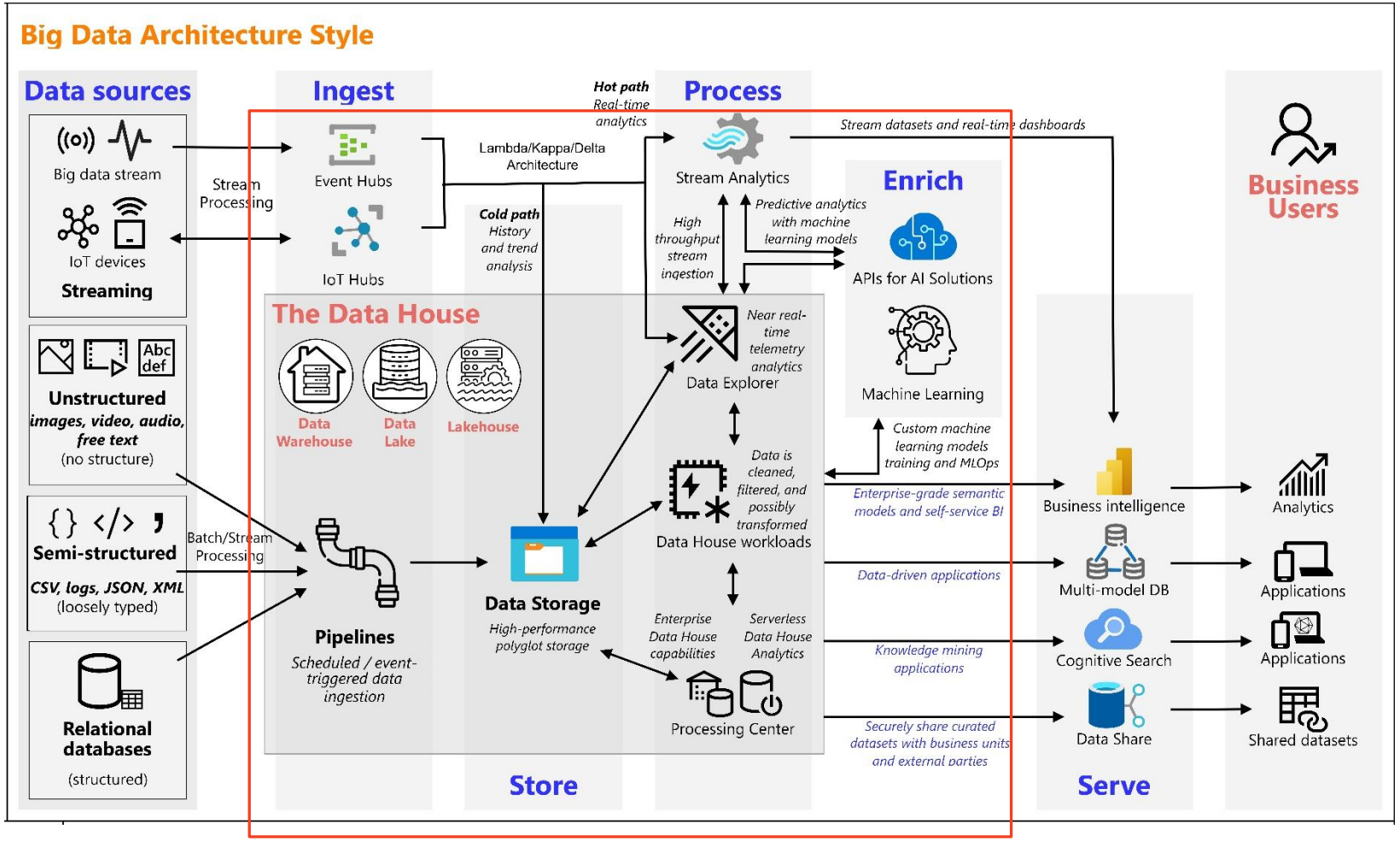
- So sánh dữ liệu với một nguồn dữ liệu được công nhận là duy nhất.
- Sử dụng phương pháp "ground truth" để xác định giá trị thực của dữ liệu.

**3. Xác minh dữ liệu:**

- Thực hiện kiểm tra thủ công hoặc tự động để xác định các giá trị không duy nhất trong dữ liệu.
- Sử dụng các công cụ xác minh dữ liệu để xác định các giá trị bất thường.

**4. Phân tích độ nhạy:**

- Đánh giá mức độ ảnh hưởng của các giá trị không duy nhất đến kết quả phân tích.




# 4. Quản lý vòng đời dữ liệu

- Đánh giá và đảm bảo chất lượng phải diễn ra ở từng giai đoạn của vòng đời. Các biện pháp được sử dụng sẽ thay đổi ở từng giai đoạn.
- Chất lượng dữ liệu ở mỗi giai đoạn được ghi lại và truyền đạt rõ ràng.
- Có thể cần quay lại các giai đoạn trước đó trong vòng đời để khắc phục các vấn đề về chất lượng dữ liệu.

# 5. Vấn đề và xu hướng

## Vấn đề

- Khối lượng dữ liệu ngày càng tăng
- Thiếu sự đồng nhất
- Dữ liệu không chính xác
- Dữ liệu không đầy đủ
- Dữ liệu lỗi thời
- Thiếu quy trình quản lý chất lượng dữ liệu
- Bảo mật dữ liệu

# 5. Vấn đề và xu hướng

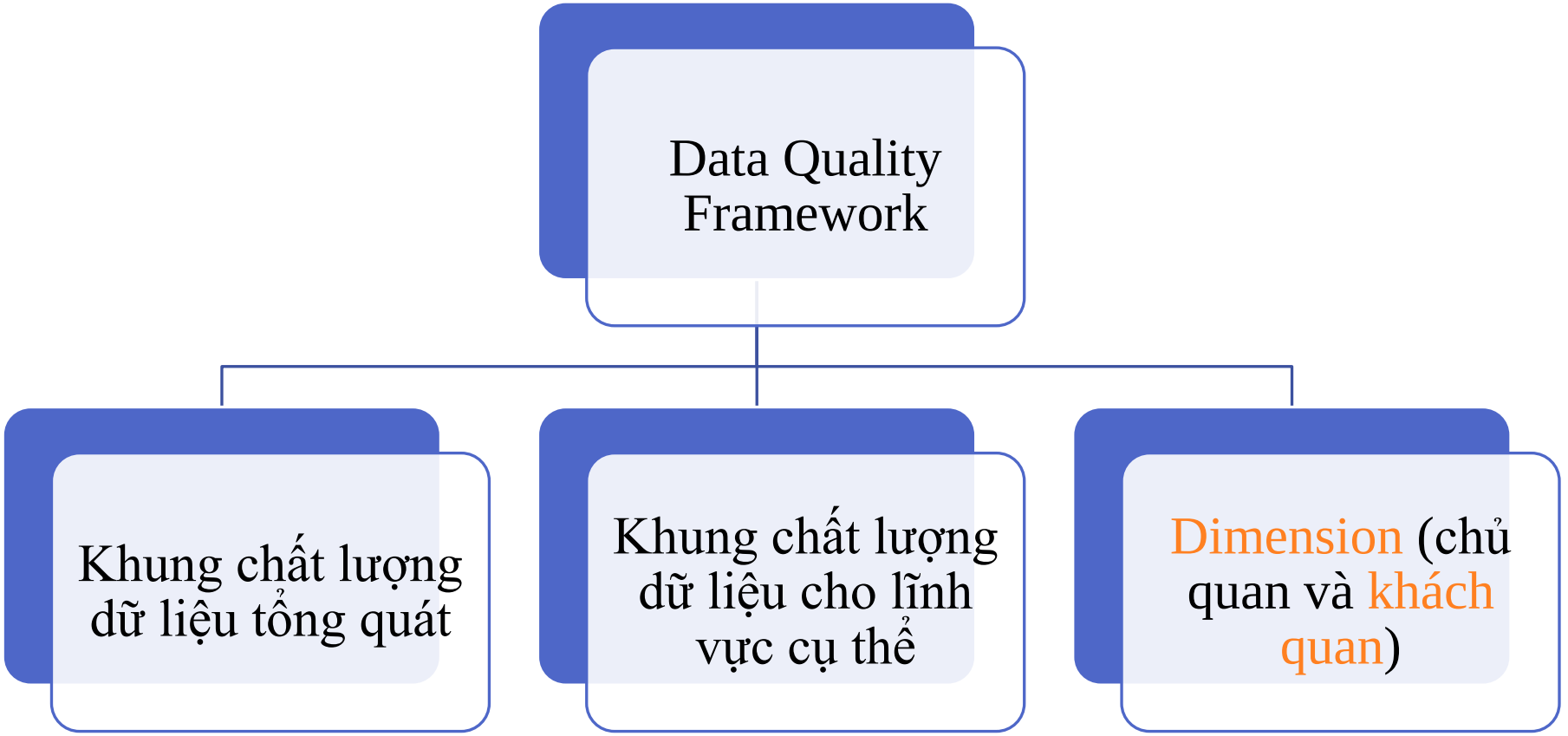
## Xu hướng

- AI & ML
- Quản lý dữ liệu chủ động
- Nền tảng dữ liệu tích hợp
- Blockchain

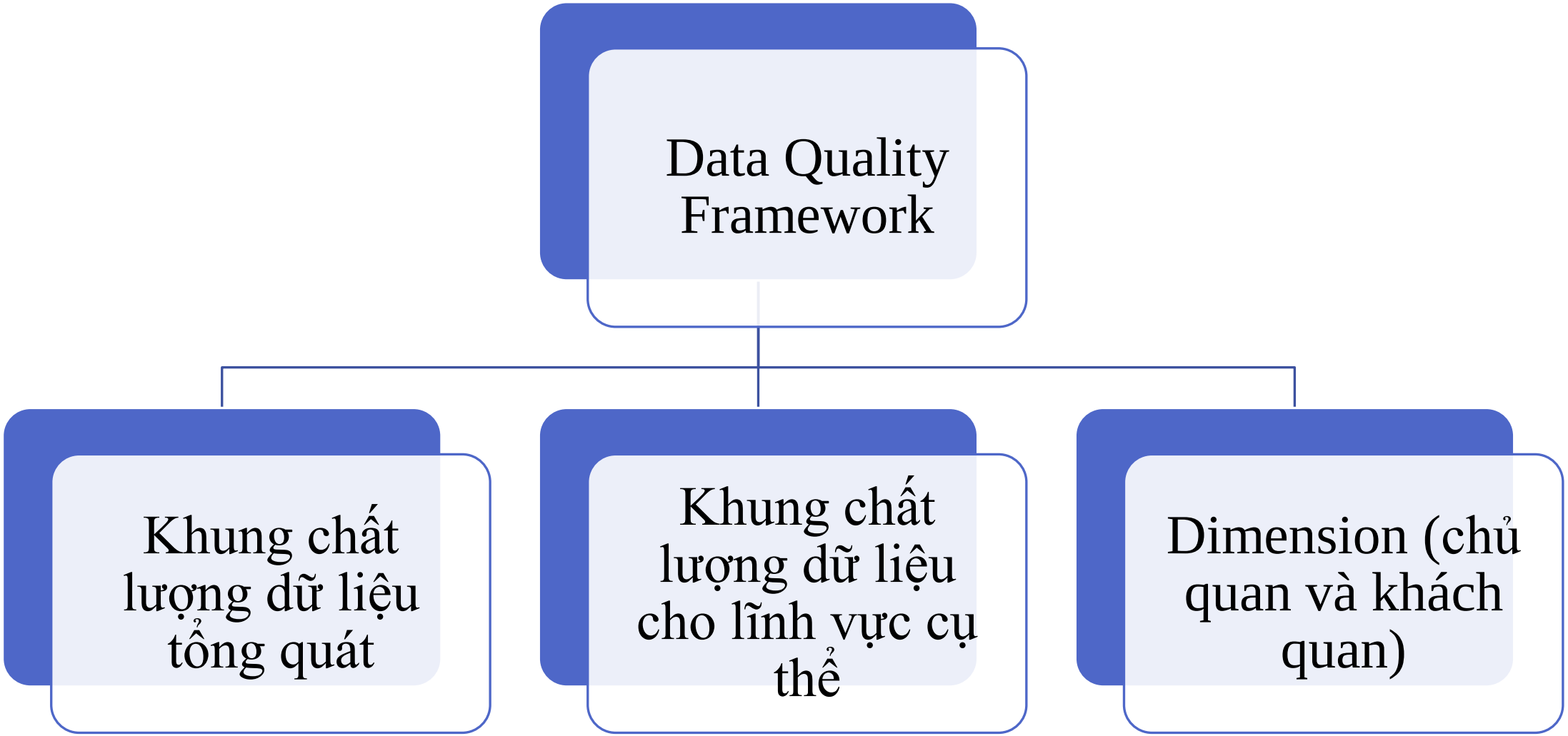
# 6. Bài tập

Tìm hiểu công cụ và thư viện hỗ trợ - Python.

- Công dụng, cú pháp sử dụng, ví dụ demo cách áp dụng.
- Vận dụng vào tập dữ liệu khai thác của nhóm.



# 7. Tổng kết


Data Quality Framework
Khung chất lượng dữ liệu tổng quát
Khung chất lượng dữ liệu cho lĩnh vực cụ thể
Dimension (chủ quan và khách quan)

# Question & Answer