Trâm Anh

**METADATA**

1. **Data definition and mapping metadata**
   1. **Data definition metadata**

Định nghĩa: là danh sách tất cả các cột từ mọi bảng trong DDS và NDS.

Vd:

Project:

1. **NDS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table\_key | Column\_key | Description | Sample\_value | Source\_column\_key |
| 1 | 1 | id Bang. Mỗi Bang có 1 id duy nhất dùng để định danh Bang | 10 |  |
| 1 | 2 | Mã code Bang. Mỗi Bang có 1 mã code duy nhất | 12 |  |
| 1 | 3 | Tên Bang | Florida |  |
| 1 | 4 | Tên viết tắt của Bang | FL |  |
| 2 | 1 | id Hạt. Mỗi Hạt có 1 id duy nhất dùng để định danh Hạt | 5 |  |
| 2 | 2 | Mã fips của Hạt. Mỗi Hạt có 1 mã fips duy nhất | 01009 |  |
| 2 | 3 | Tên Hạt | Butler |  |
| 2 | 4 | Tên ASCII của Hạt | Butler |  |
| 2 | 5 | Tên đầy đủ của Hạt | Butler County |  |
| 2 | 6 | Mã code Hạt | 013 |  |
| 2 | 7 | Vĩ độ | 3,17524E + 16 |  |
| 2 | 8 | Kinh độ | -8,66803E + 16 |  |
| 2 | 9 | Dân số | 18981 |  |
| 2 | 10 | *Mã code Bang. Khóa ngoại, tham chiếu đến state\_code của bảng States* | 01 |  |
| 2 | 11 | Ngày tạo dữ liệu | 1/1/2021 4:30:00 PM |  |
| 2 | 12 | Ngày cập nhật dữ liệu | 12/31/2021 11:12:00 PM |  |
| 3 | 1 | id mỗi lần đo lường AQI. Mỗi lần đo sẽ có 1 id duy nhất định danh lần đo đó | 1 |  |
| 3 | 2 | Ngày ghi nhận AQI | 1/1/2021 |  |
| 3 | 3 | Giá trị AQI đo được | 7 |  |
| 3 | 4 | Chỉ số xác định AQI | PM2.5 |  |
| 3 | 5 | Địa điểm xác định | 02-013-0002 |  |
| 3 | 6 | Số lượng địa điểm báo cáo |  |  |
| 3 | 7 | *Mã loại chất lượng không khí. Khóa ngoại tham chiếu đến category\_id của bảng Category* | 1 |  |
| 3 | 8 | *Mã fips Hạt. Khóa ngoại tham chiếu đến county\_fips của bảng Counties* | 2013 |  |
| 3 | 9 | Ngày tạo dữ liệu | 1/1/2021 4:30:00 PM |  |
| 3 | 10 | Ngày cập nhật dữ liệu | 12/31/2021 11:12:00 PM |  |
| 4 | 1 | Khóa chính, mỗi loại chất lượng không khí có 1 category\_id duy nhất | 1 |  |
| 4 | 2 | Tên loại chất lượng không khí | Good |  |
| 4 | 3 | Mức AQI tối thiểu cho loại chất lượng không khí | 0 |  |
| 4 | 4 | Mức AQI tối đa cho loại chất lượng không khí | 50 |  |
| 4 | 5 | Màu sắc AQI | Green |  |
| 4 | 6 | Mô tả AQI | Air quality is satisfactory, and air pollution poses little or no risk |  |

1. **DDS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table\_key | Column\_key | Description | Sample\_value | Source\_column\_key |
| 5 | 1 | Khóa Bang. Mỗi Bang có 1 khóa duy nhất dùng để định danh Bang | 10 |  |
| 5 | 2 | Tên Bang | Florida |  |
| 5 | 3 | Mã code Bang. Mỗi Bang có 1 mã code duy nhất | 12 |  |
| 5 | 4 | Tên viết tắt của Bang |  |  |
| 6 | 1 | Khóa Hạt. Mỗi Hạt có 1 id duy nhất dùng để định danh Hạt | 5 |  |
| 6 | 2 | Tên Hạt | Butler |  |
| 6 | 3 | Mã fips của Hạt. Mỗi Hạt có 1 mã fips duy nhất | 01009 |  |
| 6 | 4 | Khóa Bang. Khóa ngoại, tham chiếu đến StateKey của bảng Dim\_State | 01 |  |
| 7 | 1 | Khóa ngày tháng năm. Mỗi ngày tháng năm có 1 khóa ngày định danh ngày tháng năm đó | 1 |  |
| 7 | 2 | Ngày tháng năm | 12/31/2021 |  |
| 7 | 3 | Năm | 2021 |  |
| 7 | 4 | Quý | 4 |  |
| 7 | 5 | Tháng | 12 |  |
| 7 | 6 | Ngày | 31 |  |
| 8 | 1 | Id loại AQI |  |  |
| 8 | 2 | Tên loại |  |  |
| 8 | 3 | Giá trị nhỏ nhất |  |  |
| 8 | 4 | Giá trị lớn nhất |  |  |
| 8 | 5 | Màu AQI |  |  |
| 8 | 6 | Mô tả |  |  |
| 9 | 1 | Id tham số |  |  |
| 9 | 2 | Tên tham số |  |  |
| 10 | 1 | Ìd bảng fact |  |  |
| 10 | 2 | Ì Hạt |  |  |
| 10 | 3 | Id ngày |  |  |
| 10 | 4 | Id loại AQI |  |  |
| 10 | 5 | Id tham số |  |  |
| 10 | 6 | Ý nghĩa AQI |  |  |
| 10 | 7 | Độ lệch chuẩn |  |  |
| 10 | 8 | Giá trị AQI nhỏ nhất |  |  |
| 10 | 9 | Giá trị AQI lớn nhất |  |  |
| 10 | 10 | Số ngày có loại AQI là “Very Unhealthy” |  |  |
| 10 | 11 | Tổng AQI |  |  |
| 10 | 12 | Căn bậc 2 tổng AQI |  |  |
| 10 | 13 | Số ngày trong từng loại AQI |  |  |

* 1. **Mapping metadata**

Định nghĩa: Mapping Metadata cho biết dữ liệu đến từ đâu trong hệ thống nguồn, và cách mà nó được lưu trữ qua từng data store

Vd:

Project:

1. **Meata struture meta data:**

Định nghĩa: Data Structure Metadata là loại metadata chứa thông tin về cấu trúc và kiểu dữ liệu của các bảng và cột trong datawarehouse. Có thể bao gồm các thông tin như định nghĩa bảng, chỉ mục, khóa chính, khóa ngoại, các ràng buộc về giá trị null hoặc duy nhất

* 1. **Ds\_data\_store table:**

Định nghĩa: liệt kê tất cả các data store trong data warehouse

Vd:

Project:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Key | Data store | Description | Collation | Current\_size | Growth |
| 1 | Stage | Staging area | SQL\_latin1\_general\_CP1\_CI\_AS |  |  |
| 2 | NDS | Normalized data store | SQL\_latin1\_general\_CP1\_CI\_AS |  |  |
| 3 | DDS | Dimensional data store | SQL\_latin1\_general\_CP1\_CI\_AS |  |  |
| 4 | Meta | Metadata | SQL\_latin1\_general\_CP1\_CI\_AS |  |  |

* 1. **Ds\_table table:**

Định nghĩa: liệt kê tất cả các bảng trong data warehouse

Vd:

Project:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Key | Name | Entity\_type | Data\_store | Description |
| 1 | StageDB.dbo.Counties |  |  | Bảng lưu trữ thông tin Hạt ở StageDB |
| 2 | StageDB.dbo.AQI\_Measurements |  |  | Bảng lưu trữ thông tin đo AQI ở StageDB |
| 3 | NDSDB.dbo.State |  |  | Bảng lưu trữ thông tin Bang ở NDSDB |
| 4 | NDSDB.dbo.County |  |  | Bảng lưu trữ thông tin Hạt ở NDSDB |
| 5 | NDSDB.dbo.Category |  |  | Bảng lưu trữ thông tin loại AQI ở NDSDB |
| 6 | NDSDB.dbo.AQI\_Measurement |  |  | Bảng lưu trữ thông tin đo AQI ở NDSDB |
| 7 | Dim\_State |  |  | Bảng chiều Bang, lưu trữ thông tin Bang ở NDSDB |
| 8 | Dim\_County |  |  | Bảng chiều Hạt, lưu trữ thông tin Hạt ở NDSDB |
| 9 | Dim\_Date |  |  | Bảng chiều Ngày, lưu trữ thông tin Hạt ở NDSDB |
| 10 | Dim\_Category |  |  | Bảng chiều loại AQI, lưu trữ thông tin loại AQI ở NDSDB |
| 11 | Dim\_DefiningParameter |  |  | Bảng chiều tham số, lưu trữ thông tin tham số ở NDSDB |
| 12 | AQI\_Fact |  |  | Bảng fact, lưu trữ thông tin AQI |

* 1. **Ds\_table\_type:**

Định nghĩa: liệt kê tất cả các loại bảng trong data warehouse

Vd:

Project:

|  |  |  |
| --- | --- | --- |
| Key | Table\_type | Description |
| 1 | Dimension | Bảng chiều |
| 2 | Fact | Bảng fact |
| 3 | Stage | Bảng stage |
| 4 | Metadata | Bảng metadata |

* 1. **Ds\_column\_type table:**

Liệt kê các loại thành phần cột có trong DDS

Project:

|  |  |  |
| --- | --- | --- |
| Column type | Location | Description |
| Surrogate key | DDS dimension tables | A single not null column that uniquely identifies a row in a dimension table |
| Natural key | DDS dimension tables | Uniquely identifies a dimension row in the source system |
| Dimensional attribute | DDS dimension tables | Decribes a particular property of dimension |
| Measure | DDS Fact tables | Columns in the fact table that contain business measurements or transaction values |
| Fact key | DDS Fact tables | A single not null column that uniquely identifies a row on a fact table |

* 1. **Ds\_table\_column table:**

Liệt kê các tất cả các cột trong mỗi bảng

1. **Source system metadata:**

Source table data profile:

Trong đó:

* + **table\_key:** khóa bảng
  + **rows:** số lượng dòng
  + **row\_size:** kích thước của một dòng
  + **columns:** số lượng cột
  + **has\_timestamp:** có nhãn thời gian hay không

Source column data profile:

Trong đó:

* + **column\_key:** khóa cột
  + **table\_key:** khóa bảng
  + **unique\_values:** số giá trị duy nhất
  + **minimum:** giá trị nhỏ nhất
  + **maximum:** giá trị lớn nhất
  + **average:** giá trị trung bình
  + **max\_length:** độ dài tối đa
  + **null:** số giá trị null

1. **ETL process metadata:**

Định nghĩa: Để theo dõi quy trình ETL (Extract, Transform, Load) metadata, người ta thương dùng 3 bảng bao gồm một bảng luồng dữ liệu (data flow table), một bảng gói (package table), và một bảng trạng thái (status table).

Bảng package:

Bảng status:

Bảng ETL data flow:

Thuộc tính package và status sẽ là khóa ngoại tham chiếu tới 2 bảng kia

1. **Data quality metadata:**

Định nghĩa: Data Quality metadata chứa các quy tắc chất lượng dữ liệu, bao gồm tên và mô tả của quy tắc, loại quy tắc, danh mục quy tắc, kiểm tra tham chiếu chéo, hoặc kiểm tra dữ liệu nội bộ trong kho dữ liệu, mức độ rủi ro, trạng thái, hành động

* 1. **Data quality rule table:**

Trong đó:

Rule\_type: giá trị E là lỗi, W là cảnh báo (này hỏi chat gpt nên cũng không chắc hoàn toàn)

Rule\_category: này hỏi chat gpt nó ra I là tính toàn vẹn, C là tính nhất quán nhưng thấy có vẻ không hợp lý lắm

Status: A (active), I (inactive)

Action: R (reject), A (accept)

* 1. **Data warehouse user table:**
  2. **Data Quality Notification Table:**

Trong đó recipent là khóa ngoại sẽ tham chiếu đến user\_key của bảng warehouse user

1. **Data quality metadata:**

Định nghĩa: Được sử dụng để hỗ trợ các quản trị viên của data warehouse giám sát và bảo trì hiệu quả các quy trình dữ liệu.

Ví dụ: nếu chúng ta cần phải xác nhận các ETL đã chạy thành công hay chưa thì phải xem trong audit metadata để kiểm tra số dòng đã tải vô trong mỗi bảng.

Bảng event category:

Bảng event type:

Bảng log:

Trong đó

event\_type và event\_category sẽ là khóa ngoại tham chiếu đến event\_type\_key và event\_category\_key của 2 bảng trên.

object là khóa ngoại tham chiếu đến key của bảng ds\_table (phần struture metadata)

data\_flow tham chiếu đến key của bảng data\_flow (phần ETL progress metadata)

1. **Usage metadata:**

Định nghĩa: Usage metadata ghi nhận thông tin mỗi khi một người dùng truy cập báo cáo, cube, hoặc mô hình khai phá dữ liệu thông qua cổng web. Trước khi liên kết trên cổng web chuyển hướng người dùng đến báo cáo hoặc cubes, hệ thống sẽ ghi lại một bản ghi trong usage metadata. Điều này giúp cho quản trị viên có thể theo dõi đối tượng nào đã được truy cập và thời điểm truy cập diễn ra

Vd:

Trong đó:

user là khóa ngoại tham chiếu đến user\_key trong bảng Data warehouse user

object là khóa ngoại tham chiếu đến key của bảng ds\_table (phần struture metadata)

parameter thì vẫn không rõ nó là tham số gì