**A black and grey logo

Description automatically generated**

**CÔNG TY CỔ PHẦN KATALYST**

147 NGUYỄN CỬU VÂN, P.17, BÌNH THẠNH | WWW.KATALYST.VN

**DBT DOCUMENT AND GUIDELINE**

dÀNH CHO ….

[1. Giới thiệu về DBT 1](#_Toc169873767)

[1.1 Ưu điểm của việc sử dụng dbt\_project 1](#_Toc169873768)

[1.2 Cấu trúc của dbt\_project 1](#_Toc169873769)

[1.2.1 Profile.yml 2](#_Toc169873770)

[1.2.2 Dbt\_project.yml 2](#_Toc169873771)

[1.2.3 Macros 2](#_Toc169873772)

[1.2.4 Models 2](#_Toc169873773)

[1.2.4.1 Source 2](#_Toc169873774)

[1.2.4.2 Integration 3](#_Toc169873775)

[1.2.5 Seeds 3](#_Toc169873776)

[1.2.6 Target 3](#_Toc169873777)

[1.2.7 Tests 3](#_Toc169873778)

[2. Materializations 3](#_Toc169873779)

[2.1 Table 3](#_Toc169873780)

[2.2 Incremental 3](#_Toc169873781)

[2.3 Streaming 4](#_Toc169873782)

[3. Set up dbt project 4](#_Toc169873783)

[3.1 Activate dbt environment 4](#_Toc169873784)

[3.2 Setup variales 4](#_Toc169873785)

[3.2.1 Variable configs 4](#_Toc169873786)

[3.2.2 Models configs 5](#_Toc169873787)

[3.3 DBT Syntaxs specific 6](#_Toc169873788)

[3.4 Install dbt\_packages 7](#_Toc169873789)

[3.4.1 Setup dependencies.yml file 7](#_Toc169873790)

[4. Phát triển vùng integration 7](#_Toc169873791)

[4.1 Raw vault 7](#_Toc169873792)

[4.1.1 Structure 7](#_Toc169873793)

[4.1.1.1 Batch 8](#_Toc169873794)

[4.1.1.2 Stream 8](#_Toc169873795)

[4.1.2 Phát triển vùng RawVault 8](#_Toc169873796)

[4.1.2.1 YAML config + macros 8](#_Toc169873797)

[a. Hard rule của các cột hệ thống DataVault được thêm vào dbt project.yaml: 8](#_Toc169873798)

[b. Templates Hub, Lnk, Sat, Lsat: 9](#_Toc169873799)

[c. Define model yaml trong các template 10](#_Toc169873800)

[d. Template đầy đủ 11](#_Toc169873801)

[4.1.2.2 SQL Syntax 17](#_Toc169873802)

[4.2 Biz vault 18](#_Toc169873803)

[4.2.1 Structure 18](#_Toc169873804)

[4.3 Dbt docs 18](#_Toc169873805)

[4.4 Schema.yml 19](#_Toc169873806)

[4.5 File Source.yaml file 20](#_Toc169873807)

[5. Sample end to end 20](#_Toc169873808)

[5.1 Activate dbt environment 20](#_Toc169873809)

[5.2 Setup các variales cho file project.yml 20](#_Toc169873810)

[5.2.1 Setup variables config 20](#_Toc169873811)

[5.2.2 Setup models config 21](#_Toc169873812)

[5.2.3 Install dbt\_packages 22](#_Toc169873813)

[5.3 Phát triển vùng integration 22](#_Toc169873814)

[5.3.1 Define source 22](#_Toc169873815)

[5.3.2 Phát triển vùng raw vault 23](#_Toc169873816)

[5.3.2.1 Phát triển cấu trúc vùng raw vault 23](#_Toc169873817)

[5.3.2.2 Phát triển các models 23](#_Toc169873818)

[a. Stream 23](#_Toc169873819)

[b. Batch 30](#_Toc169873820)

[5.3.3 Phát triển vùng bizvault 39](#_Toc169873821)

[5.3.4 Document và Testing 41](#_Toc169873822)

[5.3.4.1 Document 41](#_Toc169873823)

[5.3.4.2 Testing 42](#_Toc169873824)

[5.4 Vận hành dbt project 43](#_Toc169873825)

***DANH SÁCH SAMPLE CODE***

[Exp\_Code 1: Config streaming in project.yml file 4](#_Toc169880305)

[Exp\_Code 2: Other variable configurations in project.yml 4](#_Toc169880306)

[Exp\_Code 3: Dv\_system columns in project.yml: 4](#_Toc169880307)

[Exp\_Code 4: Models configs in project.yml: 5](#_Toc169880308)

[Exp\_Code 5: Vars parameters 6](#_Toc169880309)

[Exp\_Code 6: Checkpoint\_path 7](#_Toc169880310)

[Exp\_Code 7: Intitial\_date and checkpoint\_path 7](#_Toc169880311)

[Exp\_Code 8: Setup dependencies.yml file 8](#_Toc169880312)

[Exp\_Code 9: Dv\_system columns hard rule in project.yml file 9](#_Toc169880313)

[Exp\_Code 10: Hub\_template 10](#_Toc169880314)

[Exp\_Code 11: Lnk\_template 10](#_Toc169880315)

[Exp\_Code 12: Sat\_der\_template 10](#_Toc169880316)

[Exp\_Code 13: Sat\_template 11](#_Toc169880317)

[Exp\_Code 14: Sat\_snp\_template 11](#_Toc169880318)

[Exp\_Code 15: Full hub\_template 12](#_Toc169880319)

[Exp\_Code 16: Full lnk\_template 13](#_Toc169880320)

[Exp\_Code 17: Full sat\_template 13](#_Toc169880321)

[Exp\_Code 18: Full sat\_template 14](#_Toc169880322)

[Exp\_Code 19: Full sat\_snp\_template 15](#_Toc169880323)

[Exp\_Code 20: Full lsat\_der\_template 16](#_Toc169880324)

[Exp\_Code 21: Full lsat\_template 16](#_Toc169880325)

[Exp\_Code 22: Full lsat\_snp\_template 17](#_Toc169880326)

[Exp\_Code 23: Sql syntax with source(‘db’,’table’) example 18](#_Toc169880327)

[Exp\_Code 24: Dbt docs example 19](#_Toc169880328)

[Exp\_Code 25: Hub schema.yml 20](#_Toc169880329)

[Exp\_Code 26: Source.yml example 21](#_Toc169880330)

[Exp\_Code 27: Setup other variables config 21](#_Toc169880331)

[Exp\_Code 28: Setup dv\_system columns 21](#_Toc169880332)

[Exp\_Code 29: Setup models config 22](#_Toc169880333)

[Exp\_Code 30: Setup packages in dependencies.yml 23](#_Toc169880334)

[Exp\_Code 31: Hub\_customer.sql 24](#_Toc169880335)

[Exp\_Code 32: Hub\_card.sql 25](#_Toc169880336)

[Exp\_Code 33: Lnk\_customer\_card.sql 25](#_Toc169880337)

[Exp\_Code 34: Sat\_der\_customers.sql 26](#_Toc169880338)

[Exp\_Code 35: Sat\_der\_card.sql 28](#_Toc169880339)

[Exp\_Code 36: Lsat\_der\_customer\_card.sql 30](#_Toc169880340)

[Exp\_Code 37: Sat\_card.sql 31](#_Toc169880341)

[Exp\_Code 38: Sat\_customers.sql 32](#_Toc169880342)

[Exp\_Code 39: Sat\_snp\_card.sql 35](#_Toc169880343)

[Exp\_Code 40: Sat\_snp\_customers.sql 36](#_Toc169880344)

[Exp\_Code 41: Lsat\_customer\_card.sql 38](#_Toc169880345)

[Exp\_Code 42: Lsat\_snp\_customer\_card.sql 39](#_Toc169880346)

[Exp\_Code 43: Dim\_card.sql 41](#_Toc169880347)

[Exp\_Code 44: Hub\_doc.md 43](#_Toc169880348)

[Exp\_Code 45: Schema.yml 43](#_Toc169880349)

1. Giới thiệu về DBT
   1. Ưu điểm của việc sử dụng dbt\_project

* Version control và collaborate trong team

Version control trong dbt project cho phép các teammates có thể đồng thời cùng làm việc và phối hợp với nhau trên một project duy nhất một cách an toàn, đồng thời track toàn bộ các thay đổi code.

Tham khảo thêm về dbt version control tại [đây](https://docs.getdbt.com/docs/collaborate/git-version-control).

* Dbt speaks the language of analyst – SQL

Việc tran'sform data trong dbt project được thực hiện bằng việc run các models. Models trong dbt được viết bằng SQL, là một câu lệnh select và được lưu với định dạng “.sql”. Ngoài ra, dbt models cũng có thể là Python model.

Tham khảo thêm về sql models và python models tại đây: [sql models](https://docs.getdbt.com/docs/build/sql-models) & [python models](https://docs.getdbt.com/docs/build/python-models)

* Testing

Có thể define tests với dạng file yaml. Những hình thức test phổ biến nhất thường là:

* + - Uniqueness: test các columns không được phép có giá trị trùng lặp
    - Not null: test các columns không cho phép có giá trị null
    - Referential integrity: test các giá trị column có tồn tại relationship với parent table
    - Source freshness: test dữ liệu SLA mới
    - Accepted values: check giá trị column thuộc list đã được defined.

Tham khảo thêm về dbt testing tại [đây](https://docs.getdbt.com/docs/build/data-tests).

* Documentation

Dbt cho phép generate document cho một dbt project và render dưới hình thức như một website. Document dbt project bao gồm thông tin của:

* + - Thông tin về project (information about your project): bao gồm model code, DAG, và bất cứ tests được thêm cho columns,….
    - Thông tin về data warehouse: bao gồm datatypes, table sizes.
    - Hơn nữa, dbt cũng cho phép thêm các mô tả (add description) cho models, columns, sources, và nhiều thông tin khác để hoàn thiện mặt document.

Tham khảo thêm về dbt doc tại [đây](https://docs.getdbt.com/docs/collaborate/documentation).

* 1. Cấu trúc của dbt\_project

A screenshot of a computer

Description automatically generatedMột dbt\_project được tổ chức theo cây sơ đồ:

* + 1. Profile.yml
* Default path là “./profiles.yml” hoặc “~./dbt/profile.yml”, có thể lưu ở một directory khác:
  + - Thêm enviroment variable DBT\_PROFILES\_DIR=path/to/dir
    - Thêm option --profile s-dir path/to/dir vào các execute command của dbt.
* Chứa các config để dbt core có thể kết nối tới các service khác trong data platform
  + 1. Dbt\_project.yml
* Dbt-project.yml file chứa các config để dbt core biết cách vận hành project.
* Default path là directory đang đứng, có thể lưu ở một directory khác.
  + - Thêm eviroment variable DBT\_PROJECT\_DIR=path/to/dir
    - Thêm option --project-dir path/to/dir vào các execute command của dbt.
* Chủ yếu để define các config chung và các biến global vars cho các models.
* Xem thêm về dbt\_project.yml file tại [đây](https://docs.getdbt.com/reference/dbt_project.yml).
  + 1. Macros

Dbt macros về cơ bản là các “control structures”, tương tự như các functions trong ngôn ngữ lập trình. Các macros cho phép viết SQL codes có thể sử dụng lặp đi lặp lại cho quá trình transform data cũng như combine các câu lệnh và vòng lặp SQL, giúp nâng cao hiệu quả sử dụng dbt project.

Tìm hiểu thêm về dbt macros tại [đây](https://docs.getdbt.com/docs/build/jinja-macros).

* + 1. Models

A black text on a white background

Description automatically generatedMột dbt\_project model được tổ chức theo cây sơ đồ:

* + - 1. Source

Source được định nghĩa bởi một file Source.yaml dùng để lưu trữ metadata tại vùng Peristent Staging Area, chuẩn bị cho việc transform dữ liệu.

* + - 1. Integration

Vùng Integration được tổ chức thành 2 sub-folders:

* **Raw Vault:** Lưu trữ các logic transform và metadata vào Raw Vault từ vùng Persistent staging area theo mô hình data vault.
* **Biz Vault:** Lưu trữ các logic transform và metadata từ vùng raw vault, phục vụ cho các mục đích phát triển Data Mart theo mô hình star-schema.
  + 1. Seeds

Seeds là các file csv trong dbt project (thường được lưu tại folder “seeds”), mà dbt có thể load vào data warehouse bằng cách sử dụng “dbt seed” command.

Tìm hiểu thêm về dbt seeds tại [đây](https://docs.getdbt.com/docs/build/seeds).

* + 1. Target

Target là một folder chứa tất cả các file đầu ra của quá trình chạy dbt, bao gồm các scripts được compile và run, các actifacts.

* + 1. Tests

Chứa các singular tests - tests được định nghĩa bằng các SQL files với mỗi SQL file chứa một câu lệnh select.

1. Materializations

Materializations sẽ ảnh hưởng trực tiếp đến thời gian run model. Một vài materializations sẽ làm tăng performance của model nhưng lại làm giảm performance của downstream models và queries, một số khác lại cải thiện performance nhưng lại làm tăng độ phức tạp của model.

* 1. Table

Table có cơ chế đơn giản nhất là create or replace vào target. Table nên được sử dụng với các queries có logic phức tạp mà mỗi khi query sẽ làm giảm performance của downstream queries.

* 1. Incremental

Cơ chế insert append thay vì overwrite như table.

Macro “is\_incremental()” được sử dụng để phân biệt giữa initial load (chạy lần đầu + tạo bảng) và incremental load (bảng đã tồn tại) trong model jinja2.

Incremental nên được sử dụng với bảng size M hoặc lớn hơn, model run time size L hoặc lớn hơn, và với source data thay đổi ít mỗi ngày

* Incremental nên được set “never full refesh” nếu bảng có size XL hoặc lớn hơn, source data không update old records, và có ít rủi ro shema từ source thay đổi.
* Incremental có thể sử dụng một bảng tạm (temporary table) hoặc một view cho việc incremental load và có thể có performance không giống nhau. Nên cân nhắc đến các phương án khác nhau để tạo ra một incremental model. Chi tiết hơn, có thể tham khảo tại [đây](https://docs.getdbt.com/reference/resource-configs/snowflake-configs#temporary-tables).
  1. Streaming
* Là 1 custom materialization được thêm vào hỗ trợ spark structured streaming thay vì spark sql đơn giản.
* Chỉ hỗ trợ trigger once do bản chất dbt chỉ nên chạy batch, chủ yếu tạo checkpoint và từ đó lấy được phần delta của source table
* Các config nên được thêm vào dbt project.yaml như sau:

Exp\_Code 1: Config streaming in project.yml file

|  |  |
| --- | --- |
| 1  2  3 | stream:  +materialized: streaming  +streaming\_checkpoint\_path: s3a://data/tmp/checkpoint/ktl\_dbt\_test" |

1. Set up dbt project
   1. Activate dbt environment

Sau khi access vào dbt project, sử dụng command “source ./.env/bin/” để activate dbt enviroment.

* 1. Setup variales
     1. Variable configs

Cần set up các variable configs trong file project.yml để sử dụng dbt project, bao gồm các columns hệ thống (dv\_system columns) của hệ thống datavault:

* Các variable khác:

Exp\_Code 2: Other variable configurations in project.yml

|  |  |
| --- | --- |
| 1  2  3  4 | vars:  initial\_date: "1970-01-01"  dbt\_date:time\_zone: Asia/Ho\_Chi\_Minh  disable\_dbt\_artifacts\_autoupload: true |

* Các dv\_system columns:

Exp\_Code 3: Dv\_system columns in project.yml:

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43 | dv\_system:  columns:  - target: dv\_kaf\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ kafka'  source:  name: current\_ts  dtype: timestamp  - target: dv\_kaf\_ofs  dtype: bigint  description: 'số offset của kafka'  source:  name: offset  dtype: bigint  - target: dv\_cdc\_ops  dtype: string  description: 'loại CDC (change data capture) bao gồm :"R"="read", "I"="insert", "U"="update" và "D"="delete"'  source:  name: op\_type  dtype: string  - target: dv\_src\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được insert từ source'  source:  name: cast(op\_ts/1000 as timestamp)  dtype: timestamp  - target: dv\_src\_rec  dtype: string  description: 'Tên bảng source của dữ liệu'  source:  name: table  dtype: string  - target: dv\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ vùng raw vault'  source:  name: current\_timestamp()  dtype: timestamp |

* + 1. Model configs

Model configs trong file project.yml định nghĩa các thuộc tính của models như materialized, file format, schema, tblproperties.

* Exp:

Exp\_Code 4: Models configs in project.yml:

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25 | models:  dbt\_project\_bk:  *# Config indicated by + and applies to all files under models/example/*  +materialized: table  +file\_format: iceberg  +schema: test\_dbt\_bk  +tblproperties:  "hive.engine.enabled": "true"  "read.parquet.vectorization.enabled": "true"  "read.parquet.vectorization.batch-size": "10000"  integration:  raw\_vault:  stream:  +materialized: streaming  +streaming\_checkpoint\_path: "s3a://data/tmp/checkpoint/test\_dbt\_bk"    elementary:  +file\_format: iceberg  +schema: test\_dbt\_bk  +tblproperties:  "hive.engine.enabled": "true"  "read.parquet.vectorization.enabled": "true"  "read.parquet.vectorization.batch-size": "10000" |

* 1. DBT Syntaxs specific
* Dbt compile
  + - Syntax: dbt compile –select <path.to.model>
    - dbt compile render sql scripts trong model và lưu nó trong target/compiled/profile\_name/models/…
    - dbt compile được sử dụng khi chỉ muốn lấy sql script hoàn chỉnh chứ chưa cần query ngay
    - Ví dụ: dbt compile —select models.integration.raw\_vault.stream.hub sẽ lấy kết quả trong target/compiled/profile\_name/models/integration/raw\_vault/stream/hub và sử dụng nó làm input cho 1 spark job bên ngoài.
* Dbt run
  + - Syntax: dbt run --select <path.to.model> --vars "<dictionary>" [--full-refresh]
    - Dbt run render sql scripts và execute
    - Các model incremental của Raw Vault nếu được viết bằng dbt macros sẽ cần thêm parameters bằng –vars. VD:

Exp\_Code 5: Vars parameters

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10 | {  *# start date of initial load*  'initial\_date': 'yyyy-MM-dd',  *# start date for incremental load*  'incre\_start\_date': 'yyyy-MM-dd',  *# end date for incremental load*  'incre\_end\_date': 'yyyy-MM-dd',  } |

* + - Các model streaming thì cần thêm

Exp\_Code 6: Checkpoint\_path

|  |  |  |
| --- | --- | --- |
| 1 |  | { 'checkpoint\_path': "path/to/checkpoints" } |

* + - Các parameter initial\_date và checkpoint\_path có thể thêm vào dbt\_project.yml nếu không cần thay đổi liên tục

Exp\_Code 7: Intitial\_date and checkpoint\_path

|  |  |
| --- | --- |
| 1  2  3 | vars:  initial\_date: "2024-05-16"  checkpoint\_path: "s3a://data/checkpoints/demo\_dp" |

* + - Thêm option --full-refresh nếu muốn drop tất cả target trong model và ingest lại từ đầu
* Dbt test

Dbt test run custom data tests và schema tests. Các lỗi (nếu có) vi phạm rule test đã đặt ra sẽ được hiển thị.

* Dbt seed

Dbt seed sẽ load file csv thuộc data-path vào data warehouse.

* Dbt\_run\_changed

Dbt\_run\_changed chỉ run models có sự thay đổi

* Cycle\_logs

Cycle\_logs sẽ xóa tất cả dbt logs

* Make dbt-docs

Tạo một local container chứa dbt docs trên web-browser, tại localhost:8081

**Có thể tham khảo thêm các document dbt syntaxs & commands tại:**

* [***dbt Command reference***](https://docs.getdbt.com/reference/dbt-commands)
* [***dbt seed***](https://handbook.gitlab.com/handbook/business-technology/data-team/platform/dbt-guide/#seeds)
* [***Jinja\_reference***](https://docs.getdbt.com/reference/dbt-jinja-functions)
  1. Install dbt\_packages

Các dbt packages có thể hiểu là các dbt projects độc lập, sử dụng các models và macros để xử lý một vấn đề nhất định, ví dụ như testing chất lượng data.

* + 1. Setup dependencies.yml file
* Thêm file với tên dependencies.yml vào dbt project, với level cùng cấp với file dbt\_project.yml.
* Định nghĩa các packages cần dùng vào file.
* Sử dụng command “dbt deps” để install các packages.
* Dependencies.yml Exp:

Exp\_Code 8: Setup dependencies.yml file

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | packages:  - package: dbt-labs/dbt\_utils  version: 1.2.0  - package: calogica/dbt\_expectations  version: 0.10.3  - package: elementary-data/elementary  version: 0.15.2 |

1. Phát triển vùng integration

Integration là vùng nhận dữ liệu từ source để thực hiện quá trình transform. Integration được phân thành 2 sub – folders: **raw vault** và **biz vault**.

* 1. Raw vault

Raw Vault: Lưu trữ các logic transform và metadata từ vùng Persistent staging area với mô hình data vault, bao gồm các entities chính: hub, link, sat và sat\_link (lsat).

* + 1. Structure

Có 2 sub-folders là batch và stream, đây là phương thức xử lý dữ liệu tương ứng cho các loại entities.

* + - 1. Batch

Batch sử dụng để xử lý và lưu trữ dữ liệu cuối ngày (end of date - EOD)

Batch\_processing:

* Last (sat snapshot ngày T-1)
* Sat (sat chính)
  + - 1. Stream

Stream sử dụng để tiếp nhận và xử lý dữ liệu near real-time

Streaming\_processing:

* Hub
* Lnk
* sat\_der (sat phát sinh),
* lsat\_der (sat\_link phát sinh).
  + 1. Phát triển vùng RawVault

Có 2 cách phát triển vùng RawVault:

* YAML config + macros
* SQL syntax
  + - 1. YAML config + macros

Để phát triển vùng RawVaul theo phương pháp yaml config + macros, cần:

* thêm vào dbt project.yaml các hard rule của các cột hệ thống DataVault (dv\_system columns)
* sử dụng các template: Hub, Lnk, Sat, Lsat
* define model\_yml trong các template
  + - * 1. Hard rule của các cột hệ thống DataVault được thêm vào dbt project.yaml:

Exp\_Code 9: Dv\_system columns hard rule in project.yml file

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43 | dv\_system:  columns:  - target: dv\_kaf\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ kafka'  source:  name: current\_ts  dtype: timestamp  - target: dv\_kaf\_ofs  dtype: bigint  description: 'số offset của kafka'  source:  name: offset  dtype: bigint  - target: dv\_cdc\_ops  dtype: string  description: 'loại CDC (change data capture) bao gồm :"R"="read", "I"="insert", "U"="update" và "D"="delete"'  source:  name: op\_type  dtype: string  - target: dv\_src\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được insert từ source'  source:  name: cast(op\_ts/1000 as timestamp)  dtype: timestamp  - target: dv\_src\_rec  dtype: string  description: 'Tên bảng source của dữ liệu'  source:  name: table  dtype: string  - target: dv\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ vùng raw vault'  source:  name: current\_timestamp()  dtype: timestamp |

* + - * 1. Templates Hub, Lnk, Sat, Lsat:
* Hub\_template

Exp\_Code 10: Hub\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10 | {%- set model\_yml -%}  -- model yaml from autovault service  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ hub\_transform(model=model, dv\_system=dv\_system, materialized="streaming") }}  -- support: streaming, table, view  -- default: streaming |

* Lnk\_template

Exp\_Code 11: Lnk\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10 | {%- set model\_yml -%}  -- model yaml from autovault service  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lnk\_transform(model=model, dv\_system=dv\_system, materialized="streaming") }}  -- support: streaming, table, view  -- default: streaming |

* Sat\_template
  + - Sat\_der\_template

Exp\_Code 12: Sat\_der\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9 | {%- set model\_yml -%}  -- model yaml from autovault service  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_der\_transform(model=model, dv\_system=dv\_system) }}  -- support: streaming |

* + - Sat\_template

Exp\_Code 13: Sat\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9 | {%- set model\_yml -%}  -- model yaml from autovault service  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_transform(model=model, dv\_system=dv\_system) }}  -- support: incremental |

* + - Sat\_snp\_template

Exp\_Code 14: Sat\_snp\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9 | {%- set model\_yml -%}  -- model yaml from autovault service  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_snp\_transform(model=model, dv\_system=dv\_system) }}  -- support: incremental + unique\_key |

* Lsat\_template

Tương tự Sat, nhưng thay prefix các macro bằng lsat\_

* + - * 1. Define model yaml trong các template

Các model\_yml attribute cần được khai báo vào các template trên để execute các sql models, bao gồm:

* Collision\_code

Code được sử dụng để phân biệt nguồn hệ thống (source systeam) data.

* Target\_schema

Schema chứa bảng target (target\_table), ở đây cụ thể là integration

* Target\_entity\_type

Loại bảng theo mô hình datavault của bảng target (target\_table): hub, lnk, sat, lsat

* Target\_table

Tên bảng target

* Description

Mô tả bảng target (target\_table),

* Source\_schema

Schema của bảng source (source\_table) từ vùng Peristent Staging Area.

* Source\_table

Tên bảng source (source\_table) từ vùng Peristent Staging Area.

* Columns
  + - Target: tên cột target đến
    - Dtype: data type của cột target
    - Key\_type: column type theo mô hình datavault:

+ hash\_key\_lnk

+ hash\_key\_hub

+ hash\_key\_sat

+ hash\_diff

+ Nếu không khai báo key\_type thì column type không thuộc column type của mô hình datavault

* + - Source: tên source column tại vùng Peristent Staging Area mà target column được tạo ra trực tiếp. Nếu target column được generate không chỉ từ target source, khai báo source tại columns sẽ được khai báo null, cụ thể với trường hợp của hash\_key\_sat.
* Parent: table cấp cha theo thiết kế của hệ thống datavault.
  + - * 1. Template đầy đủ
* Hub\_template

Exp\_Code 15: Full hub\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_hub  source:  -  - target:  dtype:  key\_type: biz\_key  source:  name:  dtype:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ hub\_transform(model=model, dv\_system=dv\_system) }} |

* Lnk\_template

Exp\_Code 16: Full lnk\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description:  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_lnk  source:  -  -  - target:  dtype: string  key\_type: hash\_key\_hub  source:  -  parent:  - target:  dtype: string  key\_type: hash\_key\_hub  source:  -  parent:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lnk\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_template
  + - Sat\_der

Exp\_Code 17: Full sat\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description: ""  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target:  dtype: string  key\_type: hash\_key\_hub  source:  -  - target:  dtype: string  key\_type: hash\_diff  - target:  dtype:  source:  name:  dtype:  …………………  …………………  …………………  parent\_table:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* + - Sat

Exp\_Code 18: Full sat\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description:  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target:  dtype: string  key\_type: hash\_key\_hub  source:  -  - target:  dtype: string  key\_type: hash\_diff  - target:  dtype:  source:  name:  dtype:  …………………  …………………  …………………  parent\_table:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_snp

Exp\_Code 19: Full sat\_snp\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description:  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target:  dtype: string  key\_type: hash\_key\_hub  source:  -  - target:  dtype: string  key\_type: hash\_diff  - target:  dtype:  source:  name:  dtype:  …………………  …………………  …………………  parent\_table:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* Lsat\_template
  + - Lsat\_der

Exp\_Code 20: Full lsat\_der\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description: ''  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_sat  source: 'null'  - target:  dtype: string  key\_type: hash\_key\_lnk  source:  -  -  - target:  dtype: string  key\_type: hash\_diff  - target  dtype:  source:  name:  dtype:  …………………  …………………  …………………  parent\_table:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lsat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* + - Lsat

Exp\_Code 21: Full lsat\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description: ''  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_sat  source: 'null'  - target:  dtype: string  key\_type: hash\_key\_lnk  source:  -  -  - target:  dtype: string  key\_type: hash\_diff  - target:  dtype:  source:  name:  dtype  …………………  …………………  …………………  parent\_table:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lsat\_transform(model=model, dv\_system=dv\_system) }} |

* + - Lsat\_snp

Exp\_Code 22: Full lsat\_snp\_template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34 | {%- set model\_yml -%}  collision\_code:  target\_schema:  target\_entity\_type:  target\_table:  description: ''  source\_schema:  source\_table:  columns:  - target:  dtype: string  key\_type: hash\_key\_sat  source: 'null'  - target:  key\_type: hash\_key\_lnk  source:  -  -  - target:  dtype: string  key\_type: hash\_diff  - target:  dtype:  source:  name:  dtype:  …………………  …………………  …………………  parent\_table:  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lsat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* + - 1. SQL Syntax
* User có thể viết script manually hoặc sử dụng Autovault để generate ra sql scripts và thêm vào dbt models.
* default materialized là table.
* cần phải chỉnh thủ công tên source table để tạo relation trong linage, từ `db.table` thành:
  + - `source('db', 'table')` nếu ingest từ source
    - `ref('table')` nếu transform từ 1 model khác
    - `source\_view('db', 'table')` nếu stream từ source (materialized=stream)
* Exp:

Exp\_Code 23: Sql syntax with source(‘db’,’table’) example

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | select  sha2(  coalesce(  nullif(rtrim(upper(cast(ln\_ac\_nbr as string))), ''),  '-1'  ) || '#~!' || 'dp',  256  ) as dv\_hkey\_hub\_account,  current\_timestamp() as dv\_kaf\_ldt,  1 as dv\_kaf\_ofs,  'I' as dv\_cdc\_ops,  current\_timestamp() as dv\_src\_ldt,  'test' as dv\_src\_rec,  current\_timestamp() as dv\_ldt,  'dp' as dv\_ccd  from  {{source('source', 'loan\_info')}}  where  and ln\_ac\_nbr is not null |

* 1. Biz vault

Biz Vault là nơi lưu trữ các logic transform và metadata từ vùng raw vault, phục vụ cho các mục đích phát triển Data Mart theo mô hình star-schema đáp ứng nhu cầu nghiệp vụ

* + 1. Structure
* Markdown file for document

Markdown file lưu trữ document có ý nghĩa về mặt bussiness và logic transform cho các SQL scripts.

* SQL script

Các SQL scripts để phát triển Data Mart theo mô hính Star-schema

* Schema.yaml

File schema.yml được lưu trực tiếp trong các “biz\_vault” directory, nhằm định nghĩa model properties và models tests.

* 1. Dbt docs

Dbt docs là một chức năng của dbt cho phép tạo và xem các document cho data models, sources, … Trong dbt\_project folder, Docs được lưu dưới dạng markdown files trong các directory tương ứng hoặc sử dụng description trực tiếp trong các file SQL.

* Exp:

Exp\_Code 24: Dbt docs example

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11 | *## Tables*  {% docs hub\_customer %}  Thông tin mã khách hàng được lưu trữ tập trung tại bảng hub\_customer  Bảng hub\_account được load bởi apache spark structured stream  {% enddocs %}  *## Columns*  {% docs dv\_hkey\_hub\_customer %}  hash key customer id được sinh ra bởi customer id bằng hàm SHA256  {% enddocs %} |

Tìm hiểu thêm về dbt docs tại [đây](https://docs.getdbt.com/docs/build/documentation).

* 1. Schema.yml

File schema.yml được lưu trực tiếp trong các “models” directory, nhằm định nghĩa các models properties và models tests.

* Exp:

A screenshot of a computer

Description automatically generated

File schema.yml tại directory hub định nghĩa cho các hub models như sau

Exp\_Code 25: Hub schema.yml

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33 | models:  - name: hub\_customer  description: '{{ doc("hub\_customer") }}'  columns:  - name: dv\_hkey\_hub\_customer  description: '{{ doc("dv\_hkey\_hub\_customer") }}'  data\_type: string  tests:  - not\_null  - unique  - dbt\_expectations.expect\_column\_values\_to\_be\_of\_type:  column\_type: string  - dbt\_expectations.expect\_column\_values\_to\_match\_regex:  regex: "^[A-Z0-9]"  - relationships:  to: ref('sat\_der\_customers')  field: dv\_hkey\_hub\_customer    - name: dv\_src\_ldt  data\_type: timestamp  tests:  - dbt\_expectations.expect\_row\_values\_to\_have\_recent\_data:  datepart: day  interval: 50  - name: dv\_kaf\_ofs  data\_type: long  tests:  - dbt\_expectations.expect\_column\_values\_to\_be\_between:  min\_value: 0  max\_value: 999999  tests:  - elementary.schema\_changes\_from\_baseline:  tags: ["elementary"] |

* Hub model: hub\_customer
* Description: trỏ đến dbt document có dạng file .md trong trong directory “/hub”
* Column dv\_hkey\_hub\_customer có description trỏ đến dbt document có dạng file .md trong trong directory “/hub”. Column dv\_hkey\_hub\_customer có data\_type string.
* Tương tự với các columns còn lại
* Tests: khai báo các hình thức test cho các columns và model hub\_custmer.

Tham khảo thêm về file schema.yml (properties.yml) tại [đây](https://docs.getdbt.com/reference/model-configs).

* 1. File Source.yaml file

Source.yaml dùng để lưu trữ metadata tại vùng Peristent Staging Area, chuẩn bị cho việc transform dữ liệu.

* Exp:

Exp\_Code 26: Source.yml example

|  |  |
| --- | --- |
| 1  2  3  4  5 | - name: source  schema: source  tables:  - name: loan\_info  - name: customer |

1. Sample end to end

Dưới đây là ví dụ step by step cách sử dụng 1 folder dbt-project, với giả định như sau:

* Dbt directory: “dbt/dbt\_project\_bk”
* Folder với môi trường đã được set up hoàn chỉnh và sẵn sàng để sử dụng
* Sử dụng dbt power user trên Visual Studio Code (VS Code)
  1. Activate dbt environment
* B1: Access vào folder dbt:

[dev@spark ~]$ cd dbt

* B2: Activate dbt eviroment

[dev@spark dbt]$ source ./.env/bin/activate

* B3: Access vào folder dbt project cần hoạt động

(.env) [dev@spark dbt]$ cd dbt\_project\_bk

* 1. Setup các variales cho file project.yml

Tại file dbt\_project.yml, set up các variable configs và models config để sử dụng dbt project, bao gồm các date variables, các columns hệ thống (dv\_system columns) của hệ thống datavault và config cho các models:

* + 1. Setup variables config
* Các variable khác:

Exp\_Code 27: Setup other variables config

|  |  |
| --- | --- |
| 1  2  3  4 | vars:  initial\_date: "1970-01-01"  dbt\_date:time\_zone: Asia/Ho\_Chi\_Minh  disable\_dbt\_artifacts\_autoupload: true |

* Các dv\_system columns:

Exp\_Code 28: Setup dv\_system columns

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38 | dv\_system:  columns:  - target: dv\_kaf\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ kafka'  source:  name: current\_ts  dtype: timestamp  - target: dv\_kaf\_ofs  dtype: bigint  description: 'số offset của kafka'  source:  name: offset  dtype: bigint  - target: dv\_cdc\_ops  dtype: string  description: 'loại CDC (change data capture) bao gồm :"R"="read", "I"="insert", "U"="update" và "D"="delete"'  source:  name: op\_type  dtype: string  - target: dv\_src\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được insert từ source'  source:  name: cast(op\_ts/1000 as timestamp)  dtype: timestamp  - target: dv\_src\_rec  dtype: string  description: 'Tên bảng source của dữ liệu'  source:  name: table  dtype: string  - target: dv\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ vùng raw vault'  source:  name: current\_timestamp()  dtype: timestamp |

* + 1. Setup models config

Exp\_Code 29: Setup models config

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24 | models:  dbt\_project\_bk:  *# Config indicated by + and applies to all files under models/example/*  +materialized: table  +file\_format: iceberg  +schema: test\_dbt\_bk  +tblproperties:  "hive.engine.enabled": "true"  "read.parquet.vectorization.enabled": "true"  "read.parquet.vectorization.batch-size": "10000"  integration:  raw\_vault:  stream:  +materialized: streaming  +streaming\_checkpoint\_path: "s3a://data/tmp/checkpoint/test\_dbt\_bk"  elementary:  +file\_format: iceberg  +schema: test\_dbt\_bk  +tblproperties:  "hive.engine.enabled": "true"  "read.parquet.vectorization.enabled": "true"  "read.parquet.vectorization.batch-size": "10000" |

* + 1. Install dbt\_packages

Install các packages phục vụ cho việc testing và kiểm soát chất lượng data.

* B1: Tạo file dependencies.yml:

Thêm file với tên dependencies.yml vào dbt project, với level cùng cấp với file dbt\_project.yml.

* B2: Set up file

Định nghĩa các packages cần dùng vào file:

Exp\_Code 30: Setup packages in dependencies.yml

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | packages:  - package: dbt-labs/dbt\_utils  version: 1.2.0  - package: calogica/dbt\_expectations  version: 0.10.3  - package: elementary-data/elementary  version: 0.15.2 |

* B3: Install packages

Sử dụng command “dbt deps” để install các packages.

* Check các packages đã được install vào dbt project

A black screen with white text

Description automatically generated

* 1. Phát triển vùng integration
     1. Define source

Với source data thuộc DBA psa, có các tables: customers, customers\_corecard, customers\_los, file source.yml được định nghĩa như sau:

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10 | version: 2  sources:  - name: psa  database: psa  schema: psa  tables:  - name: customers  - name: customers\_corecard  - name: customers\_los |

* + 1. Phát triển vùng raw vault
       1. Phát triển cấu trúc vùng raw vault
* B1: Tạo sub-folders: batch và stream tại directory: “dbt\_project\_bk/models/integration/raw\_vault”
* B2: Tạo các sub-folders theo thiết kế datavault tại stream:

“dbt\_project\_bk/models/integration/raw\_vault/stream/hub”

“dbt\_project\_bk/models/integration/raw\_vault/stream/lnk”

“dbt\_project\_bk/models/integration/raw\_vault/stream/sat”

“dbt\_project\_bk/models/integration/raw\_vault/stream/lsat”

* B3: Tạo các sub-folders tại batch:

“dbt\_project\_bk/models/integration/raw\_vault/batch/sat”

“dbt\_project\_bk/models/integration/raw\_vault/stream/lsat”

* + - 1. Phát triển các models

Sử dụng các template tương ứng, lần lượt tạo các sql models. Cụ thể, với data như đã định nghĩa tại source.yml, có thể phát triển model với data của customers và card:

* Hub:
  + - Hub customers
    - Hub card
* Lnk:
  + - Lnk customer\_card
* Sat:
  + - Sat customers
    - Sat card
* Lsat:
  + - Lsat\_customer\_card
      * 1. Stream
* Hub\_customer.sql

Exp\_Code 31: Hub\_customer.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: hub  target\_table: hub\_customer  source\_schema: psa  source\_table: customers  columns:  - target: dv\_hkey\_hub\_customer  dtype: string  key\_type: hash\_key\_hub  source:  - cst\_no  - target: customer\_no  dtype: decimal(38,0)  key\_type: biz\_key  source:  name: cst\_no  dtype: decimal(38,0)  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ hub\_transform(model=model, dv\_system=dv\_system, materialized = 'table') }} |

* Hub\_card.sql

Exp\_Code 32: Hub\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: hub  target\_table: hub\_card  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_hkey\_hub\_card  dtype: string  key\_type: hash\_key\_hub  source:  - card\_no  - target: card\_no  dtype: decimal(38,0)  key\_type: biz\_key  source:  name: card\_no  dtype: decimal(38,0)  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ hub\_transform(model=model, dv\_system=dv\_system, materialized = 'table') }} |

* Lnk\_customer\_card.sql

Exp\_Code 33: Lnk\_customer\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: lnk  target\_table: lnk\_customer\_card  description:  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_hkey\_lnk\_customer\_card  dtype: string  key\_type: hash\_key\_lnk  source:  - customer\_no  - card\_no  - target: dv\_hkey\_hub\_customer  dtype: string  key\_type: hash\_key\_hub  source:  - customer\_no  parent: hub\_customer  - target: dv\_hkey\_hub\_card  dtype: string  key\_type: hash\_key\_hub  source:  - card\_no  parent: hub\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lnk\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_der\_customers.sql

Exp\_Code 34: Sat\_der\_customers.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: sat  target\_table: sat\_customers  description: ""  source\_schema: psa  source\_table: customers  columns:  - target: dv\_hkey\_sat\_customer  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target: dv\_hkey\_hub\_customer  dtype: string  key\_type: hash\_key\_hub  source:  - cst\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CST\_NM  dtype: string  source:  name: CST\_NM  dtype: string  - target: CST\_FULL\_NM  dtype: string  source:  name: CST\_FULL\_NM  dtype: string  - target: CST\_TYPE  dtype: string  source:  name: CST\_TYPE  dtype: string  - target: BR\_CD  dtype: decimal(38,0)  source:  name: BR\_CD  dtype: decimal(38,0)  - target: ADDR\_PRI  dtype: string  source:  name: ADDR\_PRI  dtype: string  - target: ADDRESS\_LINE4  dtype: string  source:  name: ADDRESS\_LINE4  dtype: string  - target: ADDRESS\_LINE3  dtype: string  source:  name: ADDRESS\_LINE3  dtype: string  - target: CREATE\_DT  dtype: string  source:  name: CREATE\_DT  dtype: string  - target: RECORD\_STAT  dtype: string  source:  name: RECORD\_STAT  dtype: string  - target: SEX  dtype: string  source:  name: SEX  dtype: string  - target: ID\_NUMBER  dtype: string  source:  name: ID\_NUMBER  dtype: string  - target: TYPE\_OF\_ID  dtype: string  source:  name: TYPE\_OF\_ID  dtype: string  - target: DATE\_OF\_ISSUE  dtype: string  source:  name: DATE\_OF\_ISSUE  dtype: string  - target: PLACE\_OF\_ISSUE  dtype: string  source:  name: PLACE\_OF\_ISSUE  dtype: string  - target: PHONE  dtype: string  source:  name: PHONE  dtype: string  - target: BIRTH\_DAY  dtype: string  source:  name: BIRTH\_DAY  dtype: string  - target: INCOME\_LEVCD  dtype: string  source:  name: INCOME\_LEVCD  dtype: string  - target: EDUC\_LEVCD  dtype: string  source:  name: EDUC\_LEVCD  dtype: string  - target: MAIL\_TYPCD  dtype: string  source:  name: MAIL\_TYPCD  dtype: string  - target: OCCPTN\_CD  dtype: string  source:  name: OCCPTN\_CD  dtype: string  parent\_table: test\_dbt\_bk.hub\_customer  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_der\_card.sql

Exp\_Code 35: Sat\_der\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: sat  target\_table: sat\_card  description: ""  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_hkey\_sat\_card  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target: dv\_hkey\_hub\_card  dtype: string  key\_type: hash\_key\_hub  source:  - card\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CUSTOMER\_NO  dtype: string  source:  name: CUSTOMER\_NO  dtype: string  - target: CARD\_NAME  dtype: string  source:  name: CARD\_NAME  dtype: string  - target: CARD\_FULL\_NAME  dtype: string  source:  name: CARD\_FULL\_NAME  dtype: string  - target: CARD\_TYPE  dtype: string  source:  name: CARD\_TYPE  dtype: string  - target: ACCT\_NO  dtype: string  source:  name: ACCT\_NO  dtype: string  - target: ADDR\_LINE1  dtype: string  source:  name: ADDR\_LINE1  dtype: string  - target: ADDR\_LINE2  dtype: string  source:  name: ADDR\_LINE2  dtype: string  parent\_table: test\_dbt\_bk.hub\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* Lsat\_der\_customer\_card.sql

Exp\_Code 36: Lsat\_der\_customer\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: lsat  target\_table: lsat\_customer\_card  description: ''  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_lsat\_customer\_card  dtype: string  key\_type: hash\_key\_sat  source: 'null'  - target: dv\_hkey\_lnk\_customer\_card  dtype: string  key\_type: hash\_key\_lnk  source:  - customer\_no  - card\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CUSTOMER\_NO  dtype: string  source:  name: CUSTOMER\_NO  dtype: string  - target: CARD\_NAME  dtype: string  source:  name: CARD\_NAME  dtype: string  - target: CARD\_FULL\_NAME  dtype: string  source:  name: CARD\_FULL\_NAME  dtype: string  - target: CARD\_TYPE  dtype: string  source:  name: CARD\_TYPE  dtype: string  - target: ACCT\_NO  dtype: string  source:  name: ACCT\_NO  dtype: string  - target: ADDR\_LINE1  dtype: string  source:  name: ADDR\_LINE1  dtype: string  - target: ADDR\_LINE2  dtype: string  source:  name: ADDR\_LINE2  dtype: string  parent\_table: test\_dbt\_bk.lnk\_customer\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lsat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* + - * 1. Batch
* Sat\_card.sql

Exp\_Code 37: Sat\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: sat  target\_table: sat\_card  description: ""  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_hkey\_sat\_card  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target: dv\_hkey\_hub\_card  dtype: string  key\_type: hash\_key\_hub  source:  - card\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CUSTOMER\_NO  dtype: string  source:  name: CUSTOMER\_NO  dtype: string  - target: CARD\_NAME  dtype: string  source:  name: CARD\_NAME  dtype: string  - target: CARD\_FULL\_NAME  dtype: string  source:  name: CARD\_FULL\_NAME  dtype: string  - target: CARD\_TYPE  dtype: string  source:  name: CARD\_TYPE  dtype: string  - target: ACCT\_NO  dtype: string  source:  name: ACCT\_NO  dtype: string  - target: ADDR\_LINE1  dtype: string  source:  name: ADDR\_LINE1  dtype: string  - target: ADDR\_LINE2  dtype: string  source:  name: ADDR\_LINE2  dtype: string  parent\_table: test\_dbt\_bk.hub\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_customers.sql

Exp\_Code 38: Sat\_customers.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: sat  target\_table: sat\_customers  description: ""  source\_schema: psa  source\_table: customers  columns:  - target: dv\_hkey\_sat\_customer  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target: dv\_hkey\_hub\_customer  dtype: string  key\_type: hash\_key\_hub  source:  - cst\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CST\_NM  dtype: string  source:  name: CST\_NM  dtype: string  - target: CST\_FULL\_NM  dtype: string  source:  name: CST\_FULL\_NM  dtype: string  - target: CST\_TYPE  dtype: string  source:  name: CST\_TYPE  dtype: string  - target: BR\_CD  dtype: decimal(38,0)  source:  name: BR\_CD  dtype: decimal(38,0)  - target: ADDR\_PRI  dtype: string  source:  name: ADDR\_PRI  dtype: string  - target: ADDRESS\_LINE4  dtype: string  source:  name: ADDRESS\_LINE4  dtype: string  - target: ADDRESS\_LINE3  dtype: string  source:  name: ADDRESS\_LINE3  dtype: string  - target: CREATE\_DT  dtype: string  source:  name: CREATE\_DT  dtype: string  - target: RECORD\_STAT  dtype: string  source:  name: RECORD\_STAT  dtype: string  - target: SEX  dtype: string  source:  name: SEX  dtype: string  - target: ID\_NUMBER  dtype: string  source:  name: ID\_NUMBER  dtype: string  - target: TYPE\_OF\_ID  dtype: string  source:  name: TYPE\_OF\_ID  dtype: string  - target: DATE\_OF\_ISSUE  dtype: string  source:  name: DATE\_OF\_ISSUE  dtype: string  - target: PLACE\_OF\_ISSUE  dtype: string  source:  name: PLACE\_OF\_ISSUE  dtype: string  - target: PHONE  dtype: string  source:  name: PHONE  dtype: string  - target: BIRTH\_DAY  dtype: string  source:  name: BIRTH\_DAY  dtype: string  - target: INCOME\_LEVCD  dtype: string  source:  name: INCOME\_LEVCD  dtype: string  - target: EDUC\_LEVCD  dtype: string  source:  name: EDUC\_LEVCD  dtype: string  - target: MAIL\_TYPCD  dtype: string  source:  name: MAIL\_TYPCD  dtype: string  - target: OCCPTN\_CD  dtype: string  source:  name: OCCPTN\_CD  dtype: string  parent\_table: test\_dbt\_bk.hub\_customer  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_snp\_card.sql

Exp\_Code 39: Sat\_snp\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: sat  target\_table: sat\_card  description: ""  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_hkey\_sat\_card  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target: dv\_hkey\_hub\_card  dtype: string  key\_type: hash\_key\_hub  source:  - card\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CUSTOMER\_NO  dtype: string  source:  name: CUSTOMER\_NO  dtype: string  - target: CARD\_NAME  dtype: string  source:  name: CARD\_NAME  dtype: string  - target: CARD\_FULL\_NAME  dtype: string  source:  name: CARD\_FULL\_NAME  dtype: string  - target: CARD\_TYPE  dtype: string  source:  name: CARD\_TYPE  dtype: string  - target: ACCT\_NO  dtype: string  source:  name: ACCT\_NO  dtype: string  - target: ADDR\_LINE1  dtype: string  source:  name: ADDR\_LINE1  dtype: string  - target: ADDR\_LINE2  dtype: string  source:  name: ADDR\_LINE2  dtype: string  parent\_table: test\_dbt\_bk.hub\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* Sat\_snp\_customers.sql

Exp\_Code 40: Sat\_snp\_customers.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: sat  target\_table: sat\_customers  description: ""  source\_schema: psa  source\_table: customers  columns:  - target: dv\_hkey\_sat\_customer  dtype: string  key\_type: hash\_key\_sat  source: "null"  - target: dv\_hkey\_hub\_customer  dtype: string  key\_type: hash\_key\_hub  source:  - cst\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CST\_NM  dtype: string  source:  name: CST\_NM  dtype: string  - target: CST\_FULL\_NM  dtype: string  source:  name: CST\_FULL\_NM  dtype: string  - target: CST\_TYPE  dtype: string  source:  name: CST\_TYPE  dtype: string  - target: BR\_CD  dtype: decimal(38,0)  source:  name: BR\_CD  dtype: decimal(38,0)  - target: ADDR\_PRI  dtype: string  source:  name: ADDR\_PRI  dtype: string  - target: ADDRESS\_LINE4  dtype: string  source:  name: ADDRESS\_LINE4  dtype: string  - target: ADDRESS\_LINE3  dtype: string  source:  name: ADDRESS\_LINE3  dtype: string  - target: CREATE\_DT  dtype: string  source:  name: CREATE\_DT  dtype: string  - target: RECORD\_STAT  dtype: string  source:  name: RECORD\_STAT  dtype: string  - target: SEX  dtype: string  source:  name: SEX  dtype: string  - target: ID\_NUMBER  dtype: string  source:  name: ID\_NUMBER  dtype: string  - target: TYPE\_OF\_ID  dtype: string  source:  name: TYPE\_OF\_ID  dtype: string  - target: DATE\_OF\_ISSUE  dtype: string  source:  name: DATE\_OF\_ISSUE  dtype: string  - target: PLACE\_OF\_ISSUE  dtype: string  source:  name: PLACE\_OF\_ISSUE  dtype: string  - target: PHONE  dtype: string  source:  name: PHONE  dtype: string  - target: BIRTH\_DAY  dtype: string  source:  name: BIRTH\_DAY  dtype: string  - target: INCOME\_LEVCD  dtype: string  source:  name: INCOME\_LEVCD  dtype: string  - target: EDUC\_LEVCD  dtype: string  source:  name: EDUC\_LEVCD  dtype: string  - target: MAIL\_TYPCD  dtype: string  source:  name: MAIL\_TYPCD  dtype: string  - target: OCCPTN\_CD  dtype: string  source:  name: OCCPTN\_CD  dtype: string  parent\_table: test\_dbt\_bk.hub\_customer  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ sat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* Lsat\_customer\_card.sql

Exp\_Code 41: Lsat\_customer\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: lsat  target\_table: lsat\_customer\_card  description: ''  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_lsat\_customer\_card  dtype: string  key\_type: hash\_key\_sat  source: 'null'  - target: dv\_hkey\_lnk\_customer\_card  dtype: string  key\_type: hash\_key\_lnk  source:  - customer\_no  - card\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CARD\_NAME  dtype: string  source:  name: CARD\_NAME  dtype: string  - target: CARD\_FULL\_NAME  dtype: string  source:  name: CARD\_FULL\_NAME  dtype: string  - target: CARD\_TYPE  dtype: string  source:  name: CARD\_TYPE  dtype: string  - target: ACCT\_NO  dtype: string  source:  name: ACCT\_NO  dtype: string  - target: ADDR\_LINE1  dtype: string  source:  name: ADDR\_LINE1  dtype: string  - target: ADDR\_LINE2  dtype: string  source:  name: ADDR\_LINE2  dtype: string  parent\_table: test\_dbt\_bk.lnk\_customer\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lsat\_transform(model=model, dv\_system=dv\_system) }} |

* Lsat\_snp\_customer\_card.sql

Exp\_Code 42: Lsat\_snp\_customer\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57 | {%- set model\_yml -%}  collision\_code: mdm  target\_schema: test\_dbt\_bk  target\_entity\_type: lsat  target\_table: lsat\_customer\_card  description: ''  source\_schema: psa  source\_table: customers\_corecard  columns:  - target: dv\_lsat\_customer\_card  dtype: string  key\_type: hash\_key\_sat  source: 'null'  - target: dv\_hkey\_lnk\_customer\_card  dtype: string  key\_type: hash\_key\_lnk  source:  - customer\_no  - card\_no  - target: dv\_hsh\_dif  dtype: string  key\_type: hash\_diff  - target: CARD\_NAME  dtype: string  source:  name: CARD\_NAME  dtype: string  - target: CARD\_FULL\_NAME  dtype: string  source:  name: CARD\_FULL\_NAME  dtype: string  - target: CARD\_TYPE  dtype: string  source:  name: CARD\_TYPE  dtype: string  - target: ACCT\_NO  dtype: string  source:  name: ACCT\_NO  dtype: string  - target: ADDR\_LINE1  dtype: string  source:  name: ADDR\_LINE1  dtype: string  - target: ADDR\_LINE2  dtype: string  source:  name: ADDR\_LINE2  dtype: string  parent\_table: test\_dbt\_bk.lnk\_customer\_card  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ lsat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* + 1. Phát triển vùng bizvault

Tại biz\_vault, các sql models được phát triển theo mô hình star-schema. Vd:

* Dim\_card.sql
* Dim\_customers.sql

**Dim\_card.sql**

Exp\_Code 43: Dim\_card.sql

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101  102 | {{-  config(  materialized='table'  )  -}}  with  cte\_rnk as (  select  \*,  row\_number() over (  partition by dv\_hkey\_hub\_card, dv\_src\_ldt order by dv\_ldt  ) rnk  from {{ ref('sat\_der\_card') }}  where 1 = 1  ),  cte\_latest\_rnk1 as (select \* from cte\_rnk where rnk = 1),  cte\_new\_hsh\_dif as (  select  \*,  lag(dv\_hsh\_dif) over (  partition by dv\_hkey\_hub\_card order by dv\_src\_ldt  ) prev\_dv\_hsh\_dif  from cte\_latest\_rnk1  ),  cte\_dim\_card\_origin as (  select  hub.dv\_hkey\_hub\_card as sk\_card,  hub.card\_no,  CUSTOMER\_NO,  CARD\_NAME,  CARD\_FULL\_NAME,  CARD\_TYPE,  ACCT\_NO,  ADDR\_LINE1,  ADDR\_LINE2,  sat.dv\_src\_ldt as dv\_start\_date,  lead(sat.dv\_src\_ldt, 1, date '9999-12-31') over (  partition by sat.dv\_hkey\_hub\_card order by sat.dv\_src\_ldt  ) dv\_end\_date,  sat.dv\_ldt  from cte\_new\_hsh\_dif sat  left join  {{ ref('hub\_card') }} hub  on hub.dv\_hkey\_hub\_card = sat.dv\_hkey\_hub\_card  where  1 = 1  and (prev\_dv\_hsh\_dif <> dv\_hsh\_dif or prev\_dv\_hsh\_dif is null)  and sat.dv\_cdc\_ops <> 'D'  ),  cte\_dim\_card as (  select  dv\_ldt as dv\_ldt,  current\_timestamp() as cob\_date,  sk\_card,  card\_no,  CUSTOMER\_NO,  CARD\_NAME,  CARD\_FULL\_NAME,  CARD\_TYPE,  ACCT\_NO,  ADDR\_LINE1,  ADDR\_LINE2  from cte\_dim\_card\_origin  )  ,cte\_union as (  SELECT  dv\_ldt  ,cob\_date,  sk\_card,  card\_no,  CUSTOMER\_NO,  CARD\_NAME,  CARD\_FULL\_NAME,  CARD\_TYPE,  ACCT\_NO,  ADDR\_LINE1,  ADDR\_LINE2  FROM cte\_dim\_card  )  ,cte\_union\_rnk as (  select \*  ,row\_number() OVER ( PARTITION BY card\_no order by dv\_ldt desc) rnk  from cte\_union  ), cte\_final as (  select  dv\_ldt  ,cob\_date,  sk\_card,  card\_no,  CUSTOMER\_NO,  CARD\_NAME,  CARD\_FULL\_NAME,  CARD\_TYPE,  ACCT\_NO,  ADDR\_LINE1,  ADDR\_LINE2  from cte\_union\_rnk  where 1=1  and rnk = 1  )  select \* from cte\_final |

* + 1. Document và Testing
       1. Document
* B1: Tạo file document với định dạng .md vào các directory tương ứng
* B2: Tạo docs vào file

**Ex: Hub\_doc.md**

Exp\_Code 44: Hub\_doc.md

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24 | *## Tables*  {% docs hub\_card %}  Thông tin số card được lưu trữ tập trung tại bảng hub\_card  Bảng hub\_card được load bởi apache spark structured stream  {% enddocs %}  *## Columns*  {% docs dv\_hkey\_hub\_card %}  hash key card id được sinh ra bởi card no bằng hàm SHA256  {% enddocs %}  ----  *## Tables*  {% docs hub\_customer %}  Thông tin mã khách hàng được lưu trữ tập trung tại bảng hub\_customer  Bảng hub\_account được load bởi apache spark structured stream  {% enddocs %}  *## Columns*  {% docs dv\_hkey\_hub\_customer %}  hash key customer id được sinh ra bởi customer id bằng hàm SHA256  {% enddocs %} |

* + - 1. Testing

Dbt testing được định nghĩa trong các schema.yml tại các directory tương ứng.

* Exp: file schema.yml tại hub

Exp\_Code 45: Schema.yml

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35 | version: 2  models:  - name: hub\_customer  description: '{{ doc("hub\_customer") }}'  columns:  - name: dv\_hkey\_hub\_customer  description: '{{ doc("dv\_hkey\_hub\_customer") }}'  data\_type: string  tests:  - not\_null  - unique  - dbt\_expectations.expect\_column\_values\_to\_be\_of\_type:  column\_type: string  - dbt\_expectations.expect\_column\_values\_to\_match\_regex:  regex: "^[A-Z0-9]"  - relationships:  to: ref('sat\_der\_customers')  field: dv\_hkey\_hub\_customer    - name: dv\_src\_ldt  data\_type: timestamp  tests:  - dbt\_expectations.expect\_row\_values\_to\_have\_recent\_data:  datepart: day  interval: 50  - name: dv\_kaf\_ofs  data\_type: long  tests:  - dbt\_expectations.expect\_column\_values\_to\_be\_between:  min\_value: 0  max\_value: 999999  tests:  - elementary.schema\_changes\_from\_baseline:  tags: ["elementary"] |

* 1. Vận hành dbt project

Sử dụng các dbt syntaxs, dbt commands để vận hành các models trong dbt project cũng như kiểm soát chất lượng data.

Ví dụ, sau khi đã xây dựng các models, sử dụng command “dbt run --select hub\_customer” để excute hub\_customer.sql. Sau đó, sử dụng command “dbt test --select hub\_customer”. Với các tests đã define tại schema.yml tại hub directoy, kết quả test nhận được:

A screenshot of a computer program

Description automatically generated