**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 TECHNICAL DOCUMENT**

dÀNH CHO ….

[1. Ktl\_autovaul 1](#_Toc171941508)

[1.1 Overview 1](#_Toc171941509)

[1.2 Installation 1](#_Toc171941510)

[1.3 Macros 2](#_Toc171941511)

[1.3.1 Hub\_transform 2](#_Toc171941512)

[1.3.2 Lnk\_transform 3](#_Toc171941513)

[1.3.3 Sat\_transform 4](#_Toc171941514)

[1.3.4 Lsat\_transform 5](#_Toc171941515)

[1.4 Materializations 7](#_Toc171941516)

[1.4.1 Table, view, incremental 7](#_Toc171941517)

[1.4.2 Streaming 7](#_Toc171941518)

[2. Ktl-dbt 8](#_Toc171941519)

[2.1 Overview 8](#_Toc171941520)

[2.2 Installation 8](#_Toc171941521)

[2.3 Autovault 8](#_Toc171941522)

[2.4 ERD documents 9](#_Toc171941523)

[3. K8s deployment 11](#_Toc171941524)

[3.1 Docker image 11](#_Toc171941525)

[3.2 Deployment 11](#_Toc171941526)

[3.2.1 Configurations 11](#_Toc171941527)

[3.2.2 Execution 12](#_Toc171941528)

***DANH SÁCH SAMPLE CODE***

[Exp\_Code 1: Add package autovault vào file package.yml 1](#_Toc171945117)

[Exp\_Code 2: Install package với dbt deps 1](#_Toc171945118)

[Exp\_Code 3: Add project variables 1](#_Toc171945119)

[Exp\_Code 4: yaml model config hub 3](#_Toc171945120)

[Exp\_Code 5: Calling hub template 3](#_Toc171945121)

[Exp\_Code 6: yaml model config lnk 4](#_Toc171945122)

[Exp\_Code 7: calling lnk template 4](#_Toc171945123)

[Exp\_Code 8: yaml model config sat 5](#_Toc171945124)

[Exp\_Code 9: calling sat\_der template 5](#_Toc171945125)

[Exp\_Code 10: calling sat template 5](#_Toc171945126)

[Exp\_Code 11: calling sat\_snp template 6](#_Toc171945127)

[Exp\_Code 12: yaml model config lsat 6](#_Toc171945128)

[Exp\_Code 13: calling lsat\_der template 7](#_Toc171945129)

[Exp\_Code 14: calling lsat template 7](#_Toc171945130)

[Exp\_Code 15: calling lsat\_snp template 7](#_Toc171945131)

[Exp\_Code 16: adding streaming configurations 8](#_Toc171945132)

[Exp\_Code 17: ktl installation 8](#_Toc171945133)

[Exp\_Code 18: load YAML configurations 9](#_Toc171945134)

[Exp\_Code 19: load sat\_der và sat\_snp YAML configurations 9](#_Toc171945135)

[Exp\_Code 20: ERD documents 10](#_Toc171945136)

[Exp\_Code 21: tải các files jar cho Spark 11](#_Toc171945137)

[Exp\_Code 22: Build và push vào một một registry 11](#_Toc171945138)

[Exp\_Code 23: configmaps.yaml configurations 11](#_Toc171945139)

[Exp\_Code 24: K8S\_CONTAINER\_IMAGE 12](#_Toc171945140)

[Exp\_Code 25: base64 value from original value 13](#_Toc171945141)

[Exp\_Code 26: example 13](#_Toc171945142)

[Exp\_Code 27: Deploy DBT trên k8s cluster 13](#_Toc171945143)

[Exp\_Code 28: Execute a DBT job on k8s 13](#_Toc171945144)

1. Ktl\_autovaul
   1. Overview

* Detail of project:

gitlab: [🔗14.241.249.100](http://14.241.249.100:7979/de-team/ktl_autovault)​

* Dbt package dùng để generate Datavault models
* Support Tầng RawVault, bao gồm:
  + - Hub
    - Lnk
    - Sat/Lsat
    - Lsat-effectivity (testing)
* Support adapters:
  + - Spark (session)
    - Oracle
  1. Installation

Follow các bước sau để install ktl\_autovault project:

* Add package autovault vào file package.yml trong project directory

Exp\_Code 1: Add package autovault vào file package.yml

|  |  |
| --- | --- |
| 1  2  3 | packages:  - git: "http://reader:7PmXk-DsZo8bX\_Y6zKZH@14.241.249.100:7979/de-team/ktl\_autovault.git"  revision: main |

* Install package với dbt deps:

Exp\_Code 2: Install package với dbt deps

|  |  |
| --- | --- |
| 1 | dbt deps --upgrade |

* Add project variables:

Exp\_Code 3: Add project variables

|  |  |
| --- | --- |
| 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 | vars:  *# start date of raw vault batch initial load*  initial\_date: "yyyy-mm-dd"    *# start date value for EOD (default: no start date or 1900-01-01)*  *# incre\_start\_date:*  *# end date value for EOD (default: current date)*  *# incre\_end\_date:*  *# default data type of every hash-key in raw vault*  *# accept: binary/string (or raw/varchar2(...) for oracle)*  dv\_hash\_key\_dtype: binary    *# default hash method of every hash-key in raw vault*  *# accept: md5/sha1/sha256/sha384/sha512*  dv\_hash\_method: sha256    *# all data vault system columns*  dv\_system:  columns:  *# require*  - 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:  dtype:  - target: dv\_src\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được insert từ source'  source:  name:  dtype:  - target: dv\_src\_rec  dtype: string  description: 'Tên bảng source của dữ liệu'  source:  name:  dtype:  - target: dv\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ vùng raw vault'  source:  name: current\_timestamp()  dtype: timestamp    *# optional*  - target: dv\_kaf\_ldt  dtype: timestamp  description: 'Thời điểm dữ liệu được load từ kafka'  source:  name:  dtype:  - target: dv\_kaf\_ofs  dtype: bigint  description: 'số offset của kafka'  source:  name:  dtype: |

* 1. Macros
     1. Hub\_transform
* Macro hub\_transform sẽ render SQL queries từ table source/PSA tạo thành Raw\_vault hub table
* Support materializations:
  + - Table
    - Incremental
    - Streamting
* Cần định nghĩa model config vào yaml file

Exp\_Code 4: yaml model config hub

|  |  |
| --- | --- |
| 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 | target\_entity\_type: hub  *# table configurations*  collision\_code:  target\_schema:  target\_table:  source\_schema:  source\_table:  *# columns configurations*  columns:  *# hash-key-hub*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_hub  source: [] *# list of source columns name*    *# biz-key*  - target: *# column name*  dtype: *# data type*  key\_type: biz\_key  source:  *# source column*  name: *# column name*  dtype: *# data type* |

* Calling template

Exp\_Code 5: Calling hub template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.hub\_transform(model=model, dv\_system=dv\_system) }} |

* + 1. Lnk\_transform
* Macro lnk\_transform sẽ render SQL queries từ table source/PSA tạo thành Raw\_vault lnk table
* Support materializations:
  + - Table
    - Incremental
    - streaming
* Cần định nghĩa model config vào yaml file

Exp\_Code 6: yaml model config lnk

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23 | target\_entity\_type: lnk  *# table configurations*  collision\_code:  target\_schema:  target\_table:  source\_schema:  source\_table:  *# columns configurations*  columns:  *# hash-key-lnk*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_lnk  source: [] *# list of source columns name*    *# hash-key-hubs*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_hub  source: [] *# list of source columns name*  parent: *# foreign hub table name* |

* Calling template

Exp\_Code 7: calling lnk template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.lnk\_transform(model=model, dv\_system=dv\_system) }} |

* + 1. Sat\_transform
* Render SQL query từ một bảng source/PSA vào Raw Vault Sat-derivate table -> sat/sat-snapshot tables.
* Support materializations:
  + - Incremental
    - Table (sat-derivative only)
    - Streaming (sat-derivative only)
* Cần định nghĩa model config vào yaml file

Exp\_Code 8: yaml model config sat

|  |  |
| --- | --- |
| 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 | target\_entity\_type: lnk  *# table configurations*  collision\_code:  target\_schema:  target\_table:  source\_schema:  source\_table:  *# columns configurations*  columns:  *# hash-key-lnk*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_lnk  source: [] *# list of source columns name*    *# hash-key-hubs*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_hub  source: [] *# list of source columns name*  parent: *# foreign hub table name*    - ... |

* Calling template
  + - Sat\_der table

Exp\_Code 9: calling sat\_der template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.sat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* + - Sat\_table

Exp\_Code 10: calling sat template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.sat\_transform(model=model, dv\_system=dv\_system) }} |

* + - Sat\_snp\_table

Exp\_Code 11: calling sat\_snp template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.sat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* + 1. Lsat\_transform
* Render SQL query từ bảng source/PSA vào bảng Raw Vault Lsat - derivative table -> Lsat/lsat-snapshot tables.
* Support materializations:
  + - Incremental
    - Table (lsat-derivative only)
    - Streaming (lsat-derivative only)
* Cần định nghĩa model config vào yaml file

Exp\_Code 12: yaml model config lsat

|  |  |
| --- | --- |
| 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 | target\_entity\_type: lsat  *# table configurations*  collision\_code:  target\_schema:  target\_table: *# lsat table name (not derivative or snapshot)*  source\_schema:  source\_table:  parent\_table: *# foreign link table*  *# columns configurations*  columns:  *# hash-key-sat*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_sat    *# hash-key-lnk*  - target: *# column name*  dtype: *# data type, accept: string/binary*  key\_type: hash\_key\_lnk  source: [] *# list of source columns name*    *# hash-diff*  - target: *# column name*  dtype: *# data type*  key\_type: hash\_diff    *# dependent-keys*  - target: *# column name*  dtype: *# data type*  key\_type: dependent\_key  source:  *# source column*  name:  dtype:  - ...    *# other attributes*  - target: *# column name*  dtype: *# data type*  source:  *# source column*  name:  dtype:  - ... |

* Calling template
  + - Lsat\_der\_table

Exp\_Code 13: calling lsat\_der template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.lsat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* + - Lsat\_table

Exp\_Code 14: calling lsat template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.lsat\_transform(model=model, dv\_system=dv\_system) }} |

* + - Lsat\_snp\_table

Exp\_Code 15: calling lsat\_snp template

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | {%- set model\_yml -%}  *# YAML model config*  {%- endset -%}  {%- set model = fromyaml(model\_yml) -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.lsat\_snp\_transform(model=model, dv\_system=dv\_system) }} |

* 1. Materializations
     1. Table, view, incremental

Tham khảo về table, view, incremental materializations tại : [[](https://docs.getdbt.com/docs/build/materializations)Materializations | dbt Developer Hub](https://docs.getdbt.com/docs/build/materializations)​.

* + 1. Streaming
* Chỉ support adapter Spark
* Chỉ support table format Inceberg
* Core là Spark Structured Streaming, nhưng chỉ support table trigger Once
* Cần có thêm các configurations, trong đó:

Exp\_Code 16: adding streaming configurations

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | models:  profile\_name:  +materialized: streaming  *# method of execute python model*  +submission\_method: spark\_session  *# path to checkpoint of Spark Structured Streaming*  +streaming\_checkpoint\_path: path/to/checkpoint |

* + - Submission\_method:
      * Cách thức execute python model, mặc định là submit đến một DataBrick host
      * Original dbt không support local Spark session
      * Spark\_session: phương thức mới được định nghĩa không python package ktl-dbt, support local Spark session
    - Streaming\_checkpoint\_path:
      * Streaming\_checkpoint\_path là đường dẫn đến checkpoint của Spark Structured Streaminng
      * Giá trị có thể là là tệp local hoặc aws s3
      * Giá trị này chỉ là tiền tố của final path - đường dẫn mà sẽ trở thành *path/to/checkpoint/source\_schema\_source\_table\_to\_target\_schema\_target\_table*

1. Ktl-dbt
   1. Overview

* Detail of project:

gitlab: [🔗14.241.249.100](http://14.241.249.100:7979/de-team/ktl_autovault)​

* Dbt package dùng để extending features của DBT
* Support dbt package ktl\_autovault
  1. Installation
* Install:

Exp\_Code 17: ktl installation

|  |  |
| --- | --- |
| 1  2 | *# version 1.1.1*  pip install ktl-dbt@git+http://pip:gJFSz2wPYjw-JyFgGUJx@14.241.249.100:7979/de-team/ktl-dbt.git@v1.1.1 |

* Sau khi install, tất cả dbt CLI có thể được run bởi ktl-dbt alias.
  1. Autovault
* Trước khi run dbt CLI thực sự, trước tiên ktl-dbt chuẩn bị các macros hỗ trợ cho ktl\_autovault
* Tất cả các YAML entities configurations ktl\_autovault có thể được định nghĩa trong một central directory thay vì định nghĩa rải rác trong jinja2 model
* A screenshot of a computer

  Description automatically generatedCác thực hiện là thêm một đường dẫn ktl\_autovault\_configs mới đến dbt project, ví dụ:
* Các files config này sử dụng template giống với ktl\_autovault
* Trong jinja2 models, sử dụng dv\_config macro để load YAML configurations

Exp\_Code 18: load YAML configurations

|  |  |
| --- | --- |
| 1  2  3  4 | {%- set model = dv\_config('hub\_table\_name') -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.hub\_transform(model=model, dv\_system=dv\_system) }} |

* Lưu ý là (L)sat\_derivativity (sat/lsat-der) và (L)sat\_snapshot sử dụng cùng config với tên (L)sat (không có “der” hoặc “snp” prefix)

Exp\_Code 19: load sat\_der và sat\_snp YAML configurations

|  |  |
| --- | --- |
| 1  2  3  4 | {%- set model = dv\_config('sat\_table\_name') -%}  {%- set dv\_system = var("dv\_system") -%}  {{ ktl\_autovault.sat\_der\_transform(model=model, dv\_system=dv\_system) }} |

* 1. ERD documents
* Một tính năng mới của ktl-dbt là visualizing mối quan hệ giữa các thực thể Datavault với DBT dóc
* Sau khi run ktl-dtbt docs generate, ktl-dbt sẽ vẽ ERD dựa trên các ktl\_autovault\_configs và lưu lại dưới dạng PNG images vào đường dẫn target
* Trong document của model, sử dụng markdown

|  |  |
| --- | --- |
|  | ```render\_dv\_erd\_docs(table\_1,table\_2,...)``` |

với một list các Datavault entities để render ERD giữa chúng.

**Ví dụ:**

Exp\_Code 20: ERD documents

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | {% docs hub\_account %}  Thông tin id của tài khoản được lưu trữ tập trung tại bảng hub\_account  Bảng hub\_account được load bởi apache spark structured stream  ```render\_dv\_erd\_docs(hub\_account,sat\_account)```  {% enddocs %} |

**A screenshot of a computer

Description automatically generatedKết quả:**

* A screenshot of a computer

  Description automatically generatedDùng ```render\_dv\_erd\_docs(all)``` để include tất cả các entities, ví dụ:

1. K8s deployment
   1. Docker image

* Copy files từ [🔗14.241.249.100](http://14.241.249.100:7979/de-team/dbt/-/tree/dev/docker-build)​
* Nếu sử dụng Spark adapter và thêm external jar packages, append links vào jars.txt file, sau đó run

Exp\_Code 21: tải các files jar cho Spark

|  |  |
| --- | --- |
| 1 | wget -i jars.txt -P ./jars |

* Build và push vào một một registry

Exp\_Code 22: Build và push vào một một registry

|  |  |
| --- | --- |
| 1  2 | docker build . -t registry\_host/ktl-dbt-spark3.5:0.1.0 --no-cache  docker push registry\_host/ktl-dbt-spark3.5:0.1.0 |

* 1. Deployment
     1. Configurations
* Copy files từ [🔗14.241.249.100](http://14.241.249.100:7979/de-team/dbt/-/tree/dev/k8s-deployment)​
* Tạo file configmaps.yaml từ configmaps.yaml.template, điền configurations, ví dụ:

Exp\_Code 23: configmaps.yaml configurations

|  |  |
| --- | --- |
| 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 | apiVersion: v1  kind: ConfigMap  metadata:  namespace: spark-jobs  name: ktl-dbt-configmaps  data:  *# s3/minio storage address and bucket*  AWS\_ENDPOINT: http://192.168.1.151  AWS\_BUCKET: data    *# hive metastore address*  HIVE\_METASTORE\_URIS: thrift://192.168.1.156:9083  HIVE\_WAREHOUSE\_DIR: s3a://data/warehouse    *# kubernetes configurations*  K8S\_NAMESPACE: spark-jobs  K8S\_SERVICE\_ACCOUNT: spark  K8S\_SECRET\_REF: ktl-dbt-secrets  K8S\_CONTAINER\_IMAGE: registry\_host/ktl-dbt-spark3.5:0.1.0    *# dbt project git repo and git sync options*  GIT\_SYNC\_REPO: http://14.241.249.100:7979/de-team/dbt.git  GIT\_SYNC\_BRANCH: dev  GIT\_SYNC\_ROOT: /git  GIT\_SYNC\_DEST: dbt-project-repo  GIT\_SYNC\_WAIT: "10"  GIT\_SYNC\_ONE\_TIME: "false"    *# mount directory S3FS\_MOUNTPOINT to AWS\_BUCKET in order to save DBT artifacts*  S3FS\_MOUNTPOINT: /mnt/minio  DBT\_PROFILES\_DIR: /git/dbt-project-repo/.dbt  DBT\_PROJECT\_DIR: /git/dbt-project-repo/ktl\_dbt  DBT\_TARGET\_PATH: /mnt/minio/tmp/dbt/ktl\_dbt/target |

* Hãy chắc rằng K8S\_CONTAINER\_IMAGE giống với image của dbt-project container trong deployment.yaml

Exp\_Code 24: K8S\_CONTAINER\_IMAGE

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23 | apiVersion: apps/v1  kind: Deployment  metadata:  namespace: spark-jobs  name: ktl-dbt  labels:  app: ktl-dbt  spec:  replicas: 1  selector:  matchLabels:  app: ktl-dbt  template:  metadata:  labels:  app: ktl-dbt  spec:  serviceAccountName: spark  containers:  - name: dbt-project  image: registry\_host/ktl-dbt-spark3.5:0.1.0  imagePullPolicy: Always  ... |

* Tạo file secrets.yaml từ *secrets.yaml.template,* điền configurations thiết yếu với base64 value, bao gồm s3 storage keys và git access token (hoặc password).
* Để có được base64 value from original value, run:

Exp\_Code 25: base64 value from original value

|  |  |
| --- | --- |
| 1 | echo -n '<original value>' | base64 |

* Example:

Exp\_Code 26: example

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12 | apiVersion: v1  kind: Secret  metadata:  namespace: spark-jobs  name: ktl-dbt-secrets  data:  AWS\_ACCESS\_KEY\_ID: QVdTX0FDQ0VTU19LRVlfSUQ= *# base64 of 'AWS\_ACCESS\_KEY\_ID'*  AWS\_SECRET\_ACCESS\_KEY: QVdTX1NFQ1JFVF9BQ0NFU1NfS0VZ *# base64 of 'AWS\_SECRET\_ACCESS\_KEY'*  AWS\_REGION: YXV0bw== *# base64 of 'auto'*  GIT\_SYNC\_USERNAME: Z2l0LXVzZXI= *# base64 of 'git-user'*  GIT\_SYNC\_PASSWORD: Z2l0LXBhc3N3b3Jk *# base64 of 'git-user'* |

* + 1. Execution
* Deploy DBT trên k8s cluster:

Exp\_Code 27: Deploy DBT trên k8s cluster

|  |  |
| --- | --- |
| 1 | kubectl apply -f /path/to/config/directory |

* Execute a DBT job:

Exp\_Code 28: Execute a DBT job on k8s

|  |  |
| --- | --- |
| 1 | kubectl -n spark-jobs exec -it deploy/ktl-dbt -- ktl-dbt <options> |