**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🖎🕮✍**

Logo

Description automatically generated

**Specification cá nhân**

**Đề tài: báo cáo giá xe máy**

Giảng viên hướng dẫn: Nguyễn Đức Công Song

Môn học: Data warehouse

Nhóm: 3

Thành viên:

211330419 – Nguyễn Hoàng Lập

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Author | Changes |
| 8/12/2024 | V1.0 | Nguyễn Hoàng Lập | Data Feed Specification |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[1.0 Overview 1](#_1t3h5sf)

[1.1 Data Feed Specification 1](#_4d34og8)

[1.2 Contact Information 1](#_2s8eyo1)

[1.3 References 1](#_17dp8vu)

[2.0 Transaction Definition 2](#_3rdcrjn)

[2.1 Overview 2](#_26in1rg)

[2.2 Data Feed Process Flow 3](#_lnxbz9)

[2.1.1 module 1 – Get data from source web to csv file: 3](#_35nkun2)

[2.2.2 module 2 – Save data from csv file to DBStaging 4](#_1ksv4uv)

[2.2.3 module 3 – Get data from DBStaging and save in DBNew 5](#_44sinio)

[2.2.4 module 4 – Save data from DBNew to DBMart 6](#_2jxsxqh)

[2.3 Validation Guidelines 7](#_z337ya)

[2.3.1 module 1- Get data from source web to csv file: 7](#_3j2qqm3)

[2.3.2. module 2 – Save data from csv file to DBStaging 7](#_1y810tw)

[2.3.3 module 3 – Get data from DBStaging and save in DBNew 7](#_4i7ojhp)

[2.3.4 module 4 – Save data from DBNew to DBMart 8](#_2xcytpi)

[2.4 Attribute Specification 8](#_1ci93xb)

[2.4.1 Scraping data (GetDataWeb\_yyyyMMdd\_HHmmss.csv): 8](#_3whwml4)

[Attribute Specification 8](#_2bn6wsx)

[2.5 Code Values 9](#_qsh70q)

[2.6 Data Source Extraction and Exception Handling Procedures 9](#_3as4poj)

[2.6.1 Special Extract Procedure: Source](#_1pxezwc) 🡪 Import 9

[2.6.2 Error Code 9](#_49x2ik5)

[2.6.3 Exception Handling Procedure (Not Applicable) 9](#_2p2csry)

[2.6.4 Special Design Consideration (Not Applicable) 9](#_147n2zr)

[3.0 Feed Architecture 9](#_3o7alnk)

[3.1 Overview 9](#_23ckvvd)

[3.2 Transport Mechanism 10](#_ihv636)

[3.3 Feed Characteristics 10](#_32hioqz)

[3.3.1 Staging schema: DBController 10](#_1hmsyys)

[3.3.2 Staging schema: DBStaging 12](#_41mghml)

[3.3.3 Staging schema: DBNew 13](#_2grqrue)

[3.3.4 Warehouse schema: DBMart 16](#_vx1227)

[3.4 Data/Process Controls 18](#_3fwokq0)

[3.5 Checksum Definition (Not Applicable) 18](#_1v1yuxt)

[3.6 Control File Format (Not Applicable) 19](#_4f1mdlm)

[3.7 FTP Setup (Not Applicable) 19](#_2u6wntf)

[3.8 Operations Interface (Not Applicable) 19](#_19c6y18)

[4.0 SLA Negotiation (Not Applicable) 19](#_3tbugp1)

[Appendix B – Attribute specification for each relational table 20](#_28h4qwu)

[Appendix C – Error Codes & Exception Handling 20](#_nmf14n)

[Appendix D – Raw log file example 20](#_37m2jsg)

[5. 0 Database Diagram 20](#_1mrcu09)

[5.1. Overview 20](#_46r0co2)

[5.2. Database dbcontroller 20](#_2lwamvv)

[5.3. Database dbstaging 22](#_111kx3o)

[5.4. Database dbnew: 22](#_3l18frh)

[5.5. Database dbmart: 24](#_206ipza)

# 1.0 Overview

## 1.1 Data Feed Specification

This document provides a detailed description for scraping data from the web to the application system.

The first section of the document provides a definition of the data to be included in the feed, and the attribute specifications. The attribute specification includes definitions, formats and validation rules.

The next section describes the architecture for the feed processes. This includes the transport mechanism (FTP), feed characteristics (format, naming convention, frequency, etc), data/process controls (control counts, checksum, etc.), and operations interface.

With the completion of testing, an implementation schedule is negotiated, and the data feed is put into production status.

## 1.2 Contact Information

The following personnel are designated as subject area experts for data feed development.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Contact Email** | **Contact Phone** | **Role** |
| Nguyễn Hoàng Lập | [21130419@st.hcmuaf.edu.vn](mailto:21130419@st.hcmuaf.edu.vn) | 0343820047 | Leader |

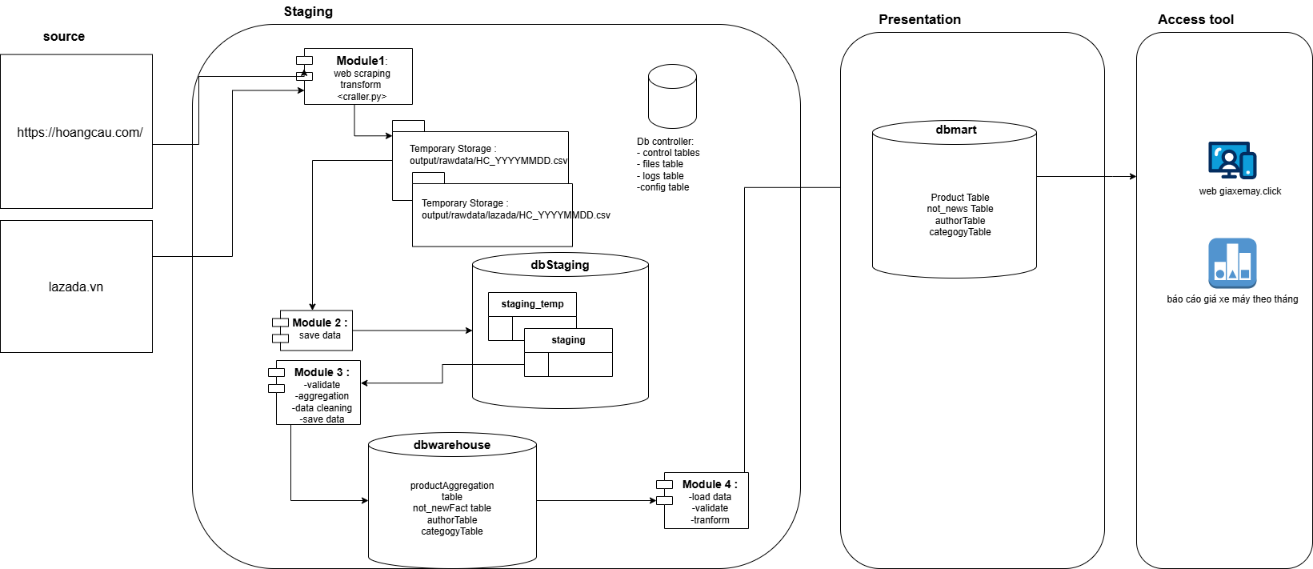
## 1.3 References

|  |  |  |  |
| --- | --- | --- | --- |
| **Document** | **Version** | **Date** | **Author** |
| Tomcat\_Access\_Logs\_Datafeed\_Specification\_Example | V1.0 | 4/10/2023 | Nguyen Duc Cong Song |
|  |  |  |  |
|  |  |  |  |

# 2.0 Transaction Definition

## 2.1 Overview

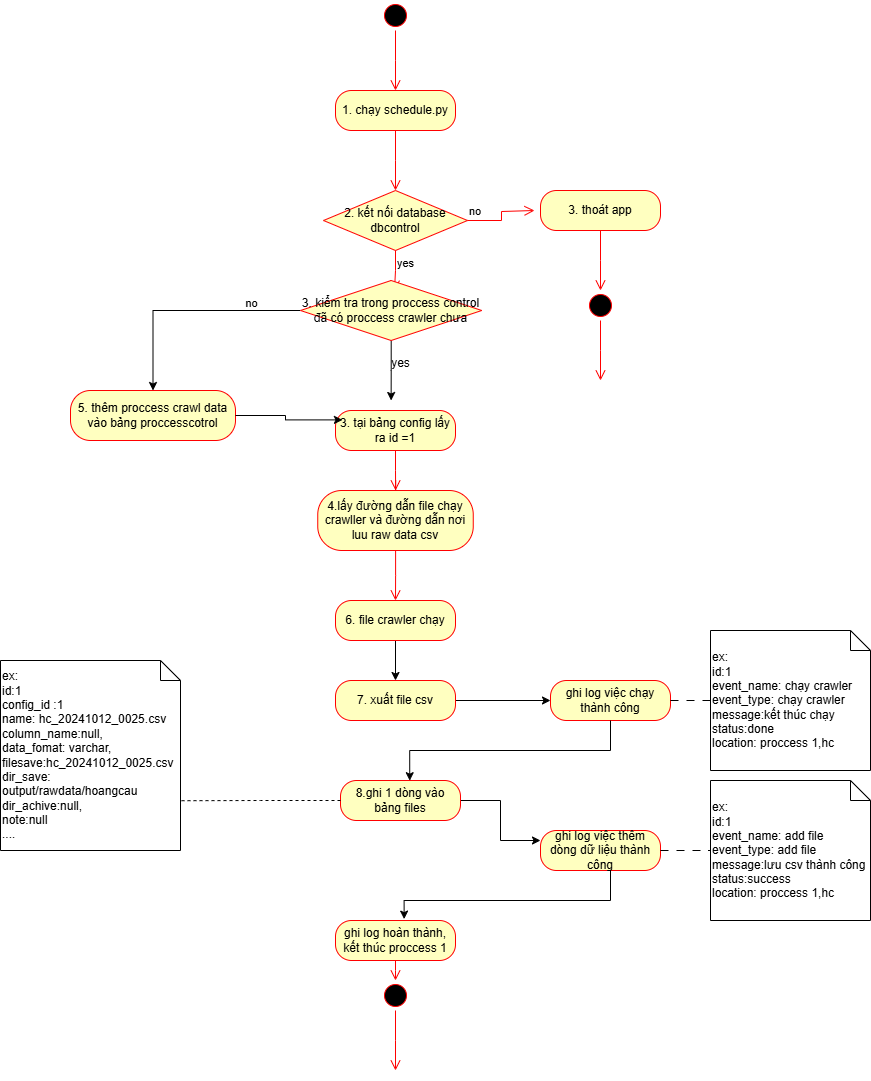
**Structure diagram**: https://drive.google.com/file/d/1c-Ohf3aeL9Y7pB42WFEEtypjsyU80cYe/view?usp=sharing



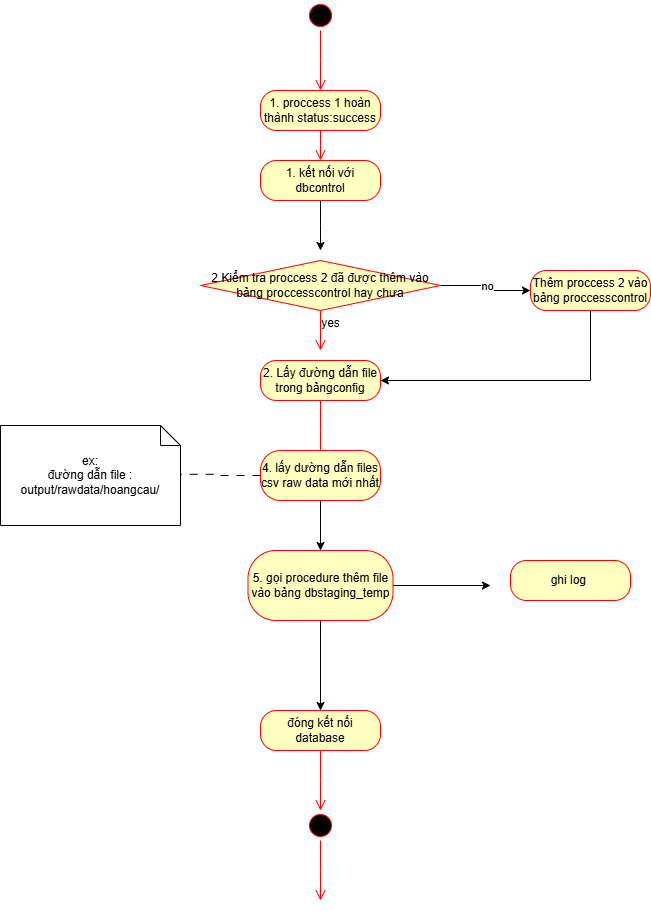
## 2.2 Data Feed Process Flow

Link: <https://app.diagrams.net/#G1rg0yEjkTXLaBvQXlotavHTqYEgqHloeB>

### module 1 – Get data from source web to CVS file:



### module 2 – Save data from CVS file to dbstaging



## Validation Guidelines

### 2.3.1 module 1- Get data from source web to CVS file:

When retrieving data from the source website and saving it in both a CSV file and the 'dbcontroll' database diagram:

* Check the controls table if the status = “RN”. If this condition is false, proceed to the next step. Otherwise, stop.
* Check the files table if the status = “done”. If this condition is false, load data from the configs table. Otherwise, stop.
* Check if scraping is successful. If this condition is false, it will delete 1 row in the controls table and in the files table, then insert 1 row in the logs table to write the bug and then send an email to 21130419@gmail.com with content that scraping web is error at + error time. Otherwise, it will move the previous file in the output folder to data\_archive and write logs.

### 2.3.2. module 2 – Save data from CVS file to dbstaging

When retrieving data from the CSV file (GetDataWeb\_yyyyMMdd\_HHmmss.csv) and saving it to the 'dbstaging' database diagram:

* Check the 'controls' table to if the status is "RN." If this condition is false, proceed to the next step. Otherwise, stop.
* Check the 'files' table to determine if the status is "done." If this condition is false, retrieve the paths of the CSV. Otherwise, stop.
* Successfully retrieve the file hoangcau\_yyyyMMdd\_HHmmss.csv. If this condition is true, insert one row into the 'controls' table and the 'logs' table. Then, establish a connection with the 'dbstaging' database, truncate data in the ‘staging’ table, retrieve data from the CSV file, and insert it into the ‘staging’ table. Otherwise, stop.
* Check if the insertion of data from the file hoangcau\_yyyyMMdd\_HHmmss.csv to the 'staging\_temp' table was successful. If this condition is true, insert one row into the 'logs' table and update the status of one row in the 'files' table. Otherwise, delete one row in the 'controls' table, insert one row into the 'logs' table, and then send an email to 21130419@st.hcmuaf.edu.vn with the content "Save data from files to DBStaging error at" + error time.
* Check if the successful update in the 'files' table has a status of 'done.' If this condition is true, insert one row into the 'logs' table and update the status in the 'controls' table. Otherwise, insert one row into the 'logs' table.
* Check if the successful update in the 'controls' table has a status of 'SC.' If this condition is true, insert one row into the 'logs' table and stop. Otherwise, insert one row into the 'logs' table.

## 2.4 Attribute Specification

This section provides concise definitions of each attribute, including field format and edit criteria. For more information about the terms used for attribute definition

### 2.4.1 Scraping data (hoangcau\_yyyyMMdd\_HHmmss.csv):

### Attribute Specification

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Field Name** | **Format** | **Mask** | **Edit Rules** | **Description** | **Example** |
|  | id | text |  |  | Id of row | 1 |
|  | model\_name | Text |  |  | Model of 1 motobike |  |
|  | type | Text |  |  | Type of motobike | XE tay ga |
|  | color | Text |  |  | color | N/A |
|  | price | Text |  |  | price | 16,000,000â‚« |
|  | price\_range | Text |  |  | Price range, if under 20.000.000 is cheap  From 20.000.000 to 40.000.000 is average  From 40 to 80 is high | N/A |
|  | brand | Text |  |  | The brand of motobike | DAELIM |
|  | version | Text |  |  | Version of motobike | N/A |
|  | name | Text |  |  | Name of motobike | DAELIM SIRIUS |
|  | engine\_capacity |  |  |  |  |  |
|  | engine\_type |  |  |  |  |  |
|  | transmission\_type |  |  |  |  |  |
|  | features |  |  |  |  |  |
|  | image\_url |  |  |  |  | https://product.hstatic.net/200000281285/product/daelim-sirius-trang\_50c38e0b00da4f53a7feb57daa622b90\_master.jpg |
|  | source\_url |  |  |  |  | https://hoangcau.com/products/daelim-sirius |
|  | source\_pid |  |  |  | Id of product in web source |  |
|  | source\_SkuId |  |  |  |  | N/A |
|  | source\_name |  |  |  |  | hoangcau.com |
|  | status |  |  |  |  | active |
|  | create\_at |  |  |  |  | 46:00.4 |
|  | source\_id |  |  |  |  | N/A |

## 2.5 Code Values

## 2.6 Data Source Extraction and Exception Handling Procedures

### 2.6.1 Special Extract Procedure:

### 2.6.2 Error Code

### 2.6.3 Exception Handling Procedure (Not Applicable)

#### Can not crawl data in web source:

* Operator can try when this error occurred:
  + Kill process, restart again
  + Set schedule to crawl per 1hour
  + Do not try to edit code because it’s not the most effiecent way to solve
  + Change the source url
* Why this error occurred:
  + The web using service (cloudflare) to tracking ip and the request, then prevent “spam” and crawl request like this tool
  + The web using service to prevent “robot”
* Delete the control that was created

Whenever encountered orther exception:

* Insert row into log table with status, event\_name, location
* Send email to 21130419@st.hcmuaf.edu.vn with title and message
* Delete the control that was created

### 2.6.4 Special Design Consideration (Not Applicable)

# 3.0 Feed Architecture

## 3.1 Overview

The key components of the feed architecture are:

**Transport Mechanism**.

Data will flow directly from source web to GetDataWeb\_yyyyMMdd\_HHmmss.csv.

**Feed Characteristics**.

Log file : hoangcau\_yyyyMMdd\_HHmmss.csv

Frequency : Daily

Content : Contains data from **previous** day

Estimated Table size : ~ 16kB/day

Estimated monthly size seq : 16\*30 = 480kB/month

Ram using: 1085 mb

**Data/Process Controls**.

**Operations Interface.**

## Transport Mechanism

This process is flat file to database data transfer.

## 3.3 Feed Characteristics

**File Format:**

The file formats accepted currently are ASCII CSV (comma separated values) and ASCII pipe-delimited. Note that files in either “DOS” (lines terminated by a carriage-return & linefeed) or “Unix” (terminated by a linefeed) are acceptable as long as the feed is consistently one or the other. Also note that the file format should not change during the transfer process to avoid checksum errors.

Though not recommended, if the data files are very large they may be compressed. Acceptable compression formats are z (compress), gzip, zip and jar. When the file is compressed, the standard extension for the compression software is used, and a compression entry is required in the control file

**Data Size:** Approximately: ~16kB/Day

**Data Frequency:** Daily

**Delivery Location:** This process is flat file to database data transfer.

### 3.3.2 Staging schema: dbstaging

**Data Location:**

~database/dbstaging.db – in gcpdw0

STAGING schema: dbstaging.staging\_temp table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | **id** | bigint AI PK |  |  |  | 1 |
|  | model\_name | text |  |  |  |  |
|  | type | text |  |  |  | Xe gắn máy |
|  | color | text |  |  |  | red |
|  | price | text |  |  |  | 19.650.000 ₫ |
|  | price\_range | text |  |  |  |  |
|  | brand | text |  |  |  |  |
|  | version | text |  |  |  |  |
|  | name | text |  |  |  | [HCM][Xe máy trả góp 0%] - Xe Máy Honda WAVE ALPHA 2024 - Phiên Bản Đặc Biệt |
|  | engine\_capacity | text |  |  |  |  |
|  | engine\_type | text |  |  |  |  |
|  | transmission\_type | text |  |  |  |  |
|  | features | text |  |  |  |  |
|  | image\_url | text |  |  |  |  |
|  | source\_url | text |  |  |  | https://www.lazada.vn/products/hcmxe-may-tra-gop-0-xe-may-honda-wave-alpha-2024-phien-ban-dac-biet-i2547623022.html |
|  | source\_pid | text |  |  |  |  |
|  | source\_SkuId | text |  |  |  |  |
|  | source\_name | text |  |  |  | lazada |
|  | status | text |  |  |  | active |
|  | create\_at | text |  |  |  | 2024-12-04 19:20:45.776565 |

## 3.4 Data/Process Controls

Data and process controls are put in place to ensure that the data feed is accurate, complete and timely.

The key controls are:

| **Control** | **Description** | **Implementation** |
| --- | --- | --- |
| SC | Status Success | Report scrap success, save data success, insert success, update status success. |
| Running | Running | Report scrap is running, save data is running. |
| ES | Error succes | Report scrap error |
| EI | Error insert | Report insert error |
| ERRor | Error | Report update status error |
| END | End | Report control is fisnish |

## 

## Checksum Definition (Not Applicable)

Since data is acquired via direct access to the system, checksum definition are *not applicable* for this data feed.

## 3.6 Control File Format (Not Applicable)

Since data is acquired via direct access to the system, control files are *not applicable* for this data feed.

## 3.7 FTP Setup (Not Applicable)

Since data is acquired via direct access to the system, FTP Setup is *not applicable* for this data feed.

## 3.8 Operations Interface (Not Applicable)

The definition of operational procedures for feed management, including exception handling and problem escalation.

This process definition will explicitly define the actions to be taken by each set of operations staff for manual components of the process. It will also identify the monitoring requirements for both sets of staff.

The problem resolution section will identify all possible error conditions and provide an appropriate course of action.

The problem escalation section identifies the situations in which the problems should be escalated, and the actions to be taken for escalation.

A contact list will be developed for each organization. This will provide names, phone numbers and areas of responsibility.

# 4.0 SLA Negotiation (Not Applicable)

Scraping data from 5p.m to 12p.m everyday and after 30 minutes, it scrapes one time.

**Appendix A – Attribute Specification**

The following values are used for definition of an attribute format.

| **Attribute Format** | **Description** | **Example** |
| --- | --- | --- |
|  |  |  |

For the definition of masks for date attributes, the following values are used:

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| yyyyMMdd\_HHmmss | Year month day – hour minute second | 20231213\_151036 |
|  |  |  |

For the definition of masks for string attributes, the following values are used:

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| NONE | No preference | get data from source |
| UPPER | All upper case | GET DATA FROM SOURCE |
| LOWER | All lower case | get data from source |
| INITCAP | All words are capitalized | Get data from source |

For the definition of masks (for attributes such as part no's), the following values are used:

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
|  |  |  |

The edit rules defined for data feed attributes are:

| **Edit Rule** | **Description** | **Example** |
| --- | --- | --- |
|  |  |  |
|  |  |  |

# Appendix B – Attribute specification for each relational table

# Appendix C – Error Codes & Exception Handling

# Appendix D – Raw log file example

# 5. Database Diagram

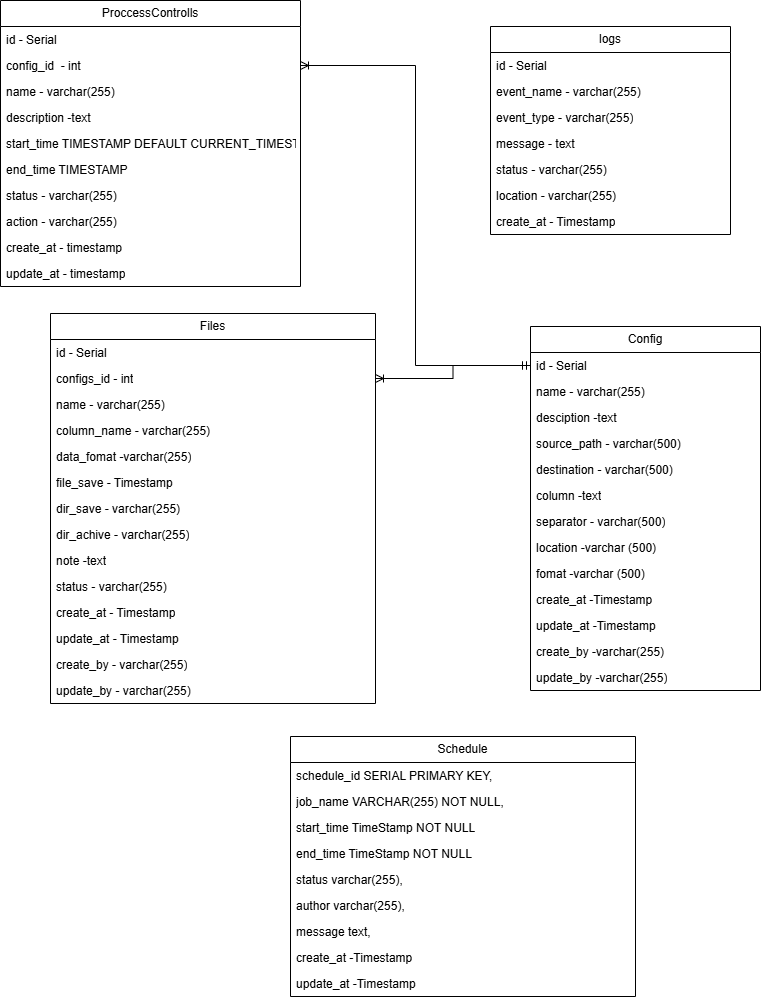
## 5.1. Overview

Data is decided into many parts, including: control, staging, new, mart

|  |  |
| --- | --- |
| **Database Name** | **Description** |
| dbcontrol | Store process and control proccess |
| Dbstaging | Store the raw data when import csv obtained from the source and data transformed to right type in datawarehouse |
| dbwarehouse | Store the cleaned and processed data from staging. |
| dbmart | Store the cleaned data and update it with the latest information from the "dbwarehouse." |

## 

## 5.2. Database dbcontroll



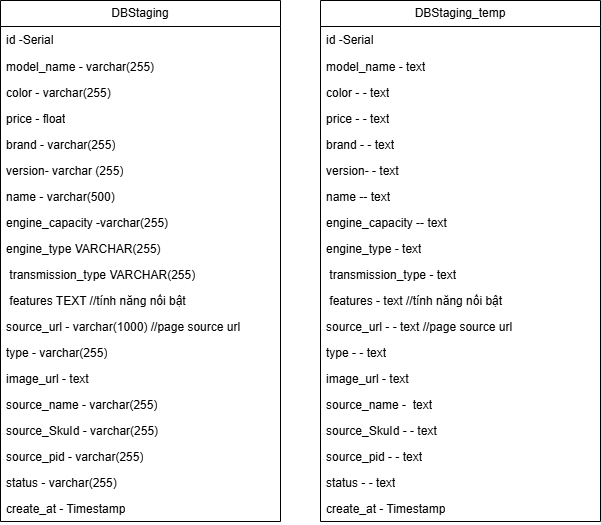
## In there:

## - Dbcontrol schema: logs table will be the place to store records of all processes from scraping data to saving and processing them

## - Dbcontrol schema: proccesscontrol table will be the place to store information about processes that will run in the system

## - Dbcontrol schema: files table will be the place to store scraped csv file information

## - Dbcontrol schema: configs table is where the system's configuration information is stored, from where the scratch file is stored to the mail information that will receive notifications.5.3. Database dbstaging



In there:

- Dbstaging schema: staging\_temp table saved data from hoangcau\_yyyyMMdd\_HHm

mss.csv

- Dbstaging schema: staging table saved data “dbstaging.stagingtemp” table which transform to right data type same to warehouse

## 5.4 Link github: https://github.com/lapnguyen331/pj\_datawarehouse.git

~END~