**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🖎🕮✍**

Logo

Description automatically generated

**Specification**

**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

20130283 - Nguyễn Phạm Quốc Huy

20130250 - Lê Thị Mỹ Hạnh

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** |
| La Thi Thuy Hao | [20130253@st.hcmuaf.edu.vn](mailto:20130253@st.hcmuaf.edu.vn) | 0364677119 | Leader |
| Nguyen Pham Quoc Huy | [20130283@st.hcmuaf.edu.vn](about:blank) | 0374162824 | Member |
| Le Thi My Hanh | [20130250@st.hcmuaf.edu.vn](mailto:20130250@st.hcmuaf.edu.vn) | 0384887437 | Member |
|  |  |  |  |

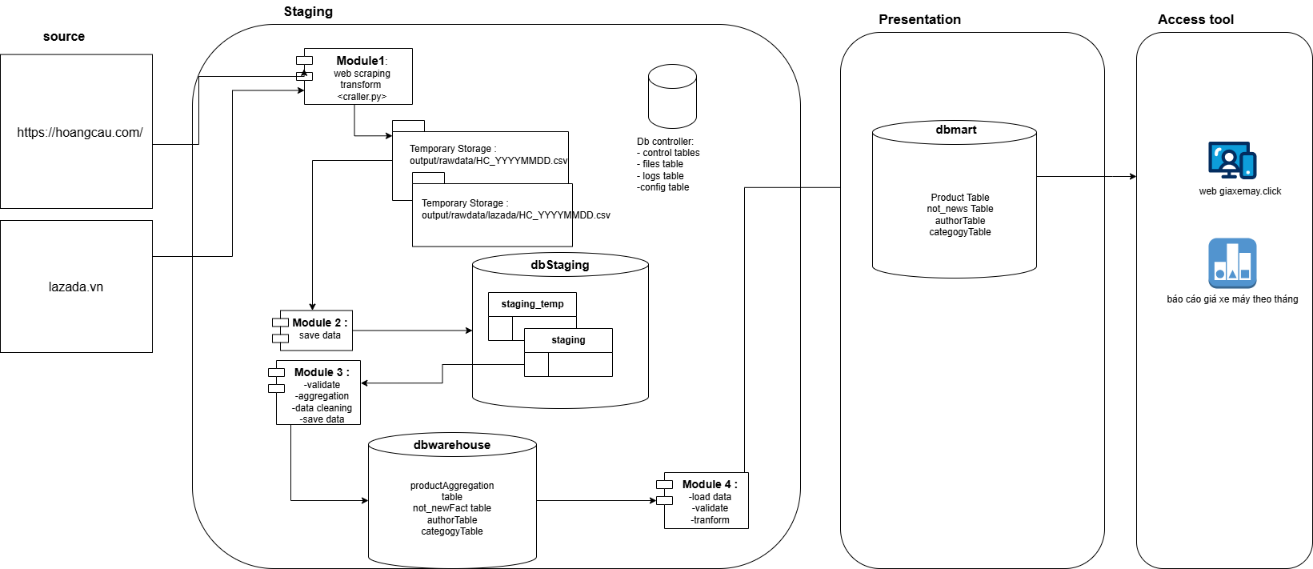
## 1.3 References

|  |  |  |  |
| --- | --- | --- | --- |
| **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

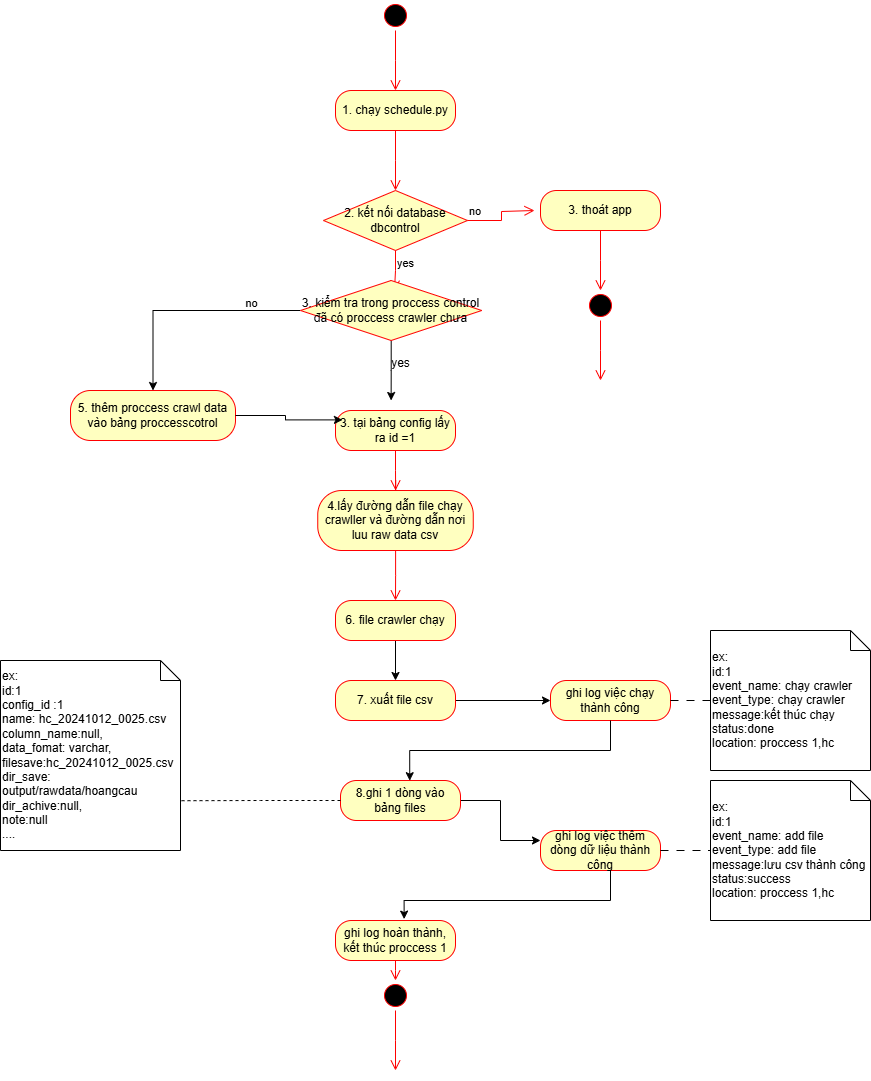
Structure diagram: https://drive.google.com/file/d/1c-Ohf3aeL9Y7pB42WFEEtypjsyU80cYe/view?usp=sharing



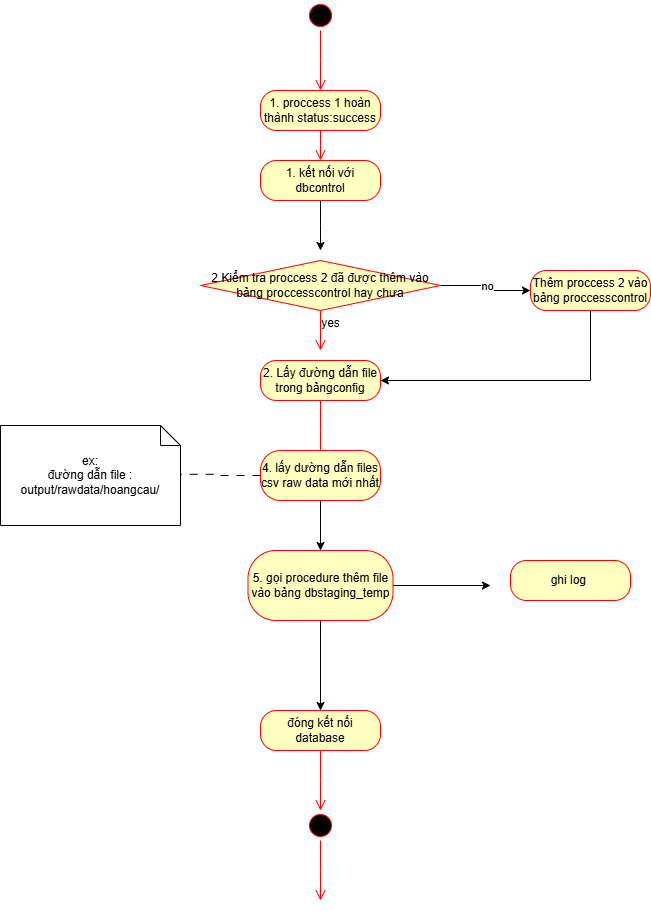
## 2.2 Data Feed Process Flow

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

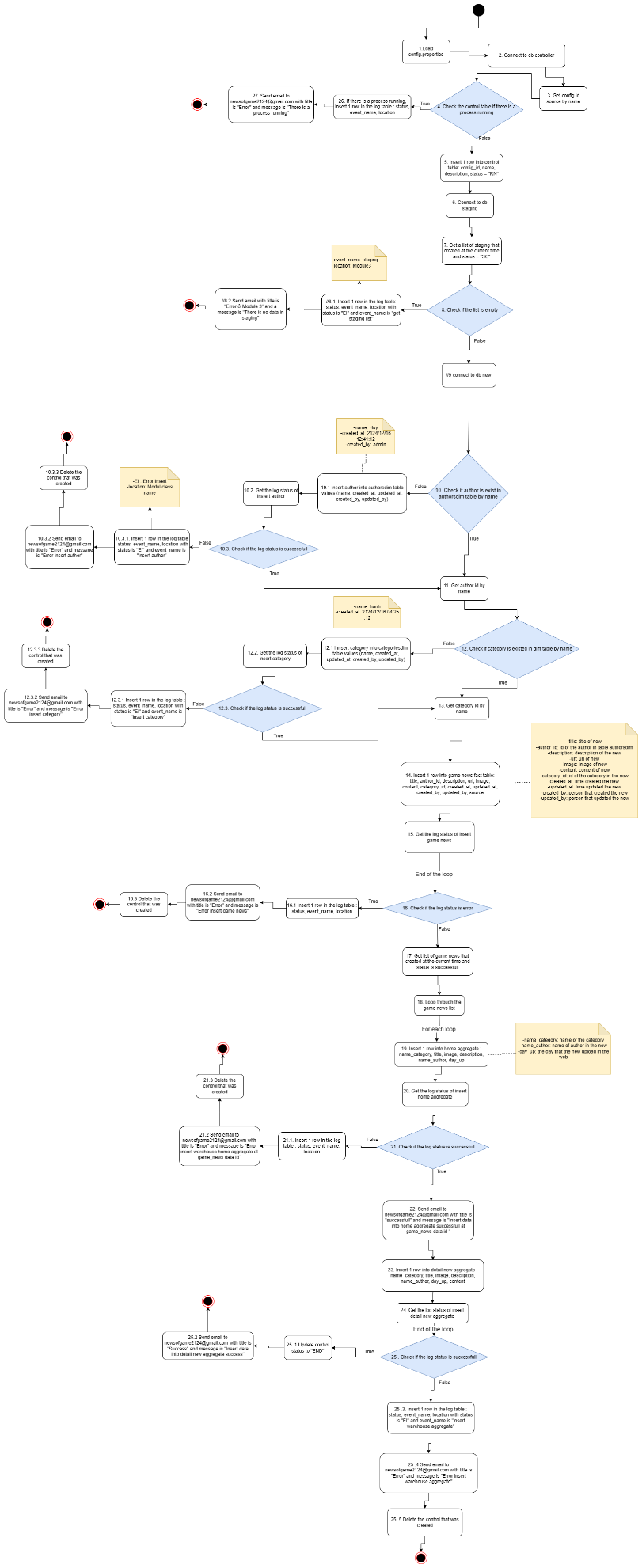
### module 1 – Get data from source web to CVS file:



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



### module 3 – Get data from dbstaging and save in dbwarehouse



### module 4 – Save data from dbwarehouse to dbmart

A diagram of a diagram

Description automatically generated with medium confidence

## 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 'dbcontroller' 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 GetDataWeb\_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 GetDataWeb\_yyyyMMdd\_HHmmss.csv to the 'staging' 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 newsofgame2023@gmail.com 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.

### 2.3.3 module 3 – Get data from dbstaging and save in dbnew

When retrieving data from the 'dbstaging' database diagram and saving it to the 'dbnew' database diagram:

* Check the control stable if the status = “RN”. If this condition is false proceed to the next step. Otherwise, insert 1 row in the log table : status, event\_name, location, and send email to newsofgame2023@gmail.com with title is "Error" and message is "There is a process running" and then, stop.
* Check if the staging list is empty. If this condition is false proceed to the next step. Otherwise, insert 1 row in the log table : status, event\_name, location with status is "staging table is empty", and send email to newsofgame2023@gmail.com with title is "Error" and message is "There is no data in staging" and then, stop.
* Check if author is exist in authorsdim table by name. If this condition is true proceed to the next step. Otherwise, Insert author into authorsdim table values (name, created\_at, updated\_at, created\_by, updated\_by).
* Every time a insert being used to insert data to table, Check if the log status is successful. If this condition is false, Insert 1 row in the log table : status, event\_name, location with status is "EI" and event\_name is "insert author", Send email to newsofgame2023@gmail.com with title is "Error" and message is "Error insert ….", then Delete the control that was created and finally, stop. Otherwise, proceed to the next step.

### 2.3.4 module 4 – Save data from dbnew to dbmart

When retrieving data from the 'dbnew' database diagram and saving it to the 'dbnew' database diagram:

* Check if any process is running. If this condition is false proceed to the next step. Otherwise, insert 1 row in the log table : status, event\_name, location, and send email to newsofgame2124@gmail.com with title is "Error" and message is "There is a process running" and then, stop
* Every time a list being added, check if the list is empty. If this condition is false proceed to the next step. Otherwise, Insert 1 row in the log table : status, event\_name, location with status = …… data table is empty, and send email to newsofgame2023@gmail.com with title is "Error" and message is "Table ….. with data in the list is empty" and then, stop
* Every time a insert being used to insert data to table, Check if the log status is successful. If this condition is false, Insert 1 row in the log table : status, event\_name, location with status is "EI" and event\_name is "insert author", Send email to newsofgame2023@gmail.com with title is "Error" and message is "Error insert ….", then Delete the control that was created and finally, stop. Otherwise, proceed to the next step.

## 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: Source 🡪 Import

### 2.6.2 Error Code

### 2.6.3 Exception Handling Procedure (Not Applicable)

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

* Operator can do when this 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” request
  + 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.

### 3.3.1 Staging schema: dbcontroller

**Data Location:**

~database/dbcontroller.db – in gcpdw0

STAGING schema: dbcontroller.proccesscontrol table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | INT |  |  |  | 12 |
|  | name | Varchar(1000) |  |  |  | Get Data from web |
|  | description | TEXT |  |  |  | Get data from source and save in CSV file |
|  | status | Varchar(500) |  |  |  | SC |
|  | config\_id | INT |  |  |  | 15 |
|  | created\_at | TIMESTAMP |  |  |  | 2023-12-14 13:49:07 |
|  | updated\_at | TIMESTAMP |  |  |  | 2023-12-14 13:49:07 |

STAGING schema: dbcontroller.configs table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | INT |  |  |  | 12 |
|  | name | Varchar(1000) |  |  |  | url Source Riot News |
|  | description | TEXT |  |  |  | liên kết nguồn Riot News |
|  | source\_path | Varchar(1000) |  |  |  | https://www.leagueoflegends.com/vi-vn/news/ |
|  | destination | Varchar(500) |  |  |  | File csv |
|  | columns | TEXT |  |  |  | 1 |
|  | separator | Varchar(500) |  |  |  | , |
|  | location | Varchar(1000) |  |  |  | module1/src/dao/GetDataFromSource |
|  | format | Varchar(255) |  |  |  | Riot Games |
|  | created\_at | TIMESTAMP |  |  |  | 2023-12-01 15:07:39 |
|  | updated\_at | TIMESTAMP |  |  |  | 2023-12-06 14:43:05 |
|  | created\_by | Varchar(250) |  |  |  | La Thị Thúy Hảo |
|  | updated\_by | Varchar(250) |  |  |  | La Thị Thúy Hảo |

STAGING schema: dbcontroller.files table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | INT |  |  |  | 2 |
|  | config\_id | INT |  |  |  | 15 |
|  | name | Varchar(1000) |  |  |  | output\GetDataWeb\_20231213\_151036 |
|  | column\_name | Varchar(1000) |  |  |  | Tiêu đề, Mô tả, Tác giả, Thời gian, Liên kết, Danh mục, Ảnh, Nội dung bài viết, Nguồn |
|  | data\_format | Varchar(500) |  |  |  | .csv |
|  | file\_timestamp | TIMESTAMP |  |  |  | 2023-12-13 15:10:36 |
|  | destination | Varchar(500) |  |  |  | module 1 |
|  | dir\_save | Varchar(500) |  |  |  | output |
|  | dir\_archive | Varchar(500) |  |  |  | data\_archive |
|  | note | TEXT |  |  |  |  |
|  | status | Varchar(250) |  |  |  | SC |
|  | created\_at | TIMESTAMP |  |  |  | 2023-12-13 15:10:37 |
|  | updated\_at | TIMESTAMP |  |  |  | 2023-12-13 15:10:37 |
|  | created\_by | Varchar(255) |  |  |  | La Thị Thúy Hảo |
|  | updated\_by | Varchar(255) |  |  |  | La Thị Thúy Hảo |

STAGING schema: dbcontroller.logs table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 12 |
|  | event\_name | Varchar(1000) |  |  |  | Get Data from web |
|  | event\_type | Varchar(1000) |  |  |  | get data from source |
|  | status | Varchar(500) |  |  |  | SC |
|  | location | Date |  |  |  | module 1 |
|  | created\_at | Text |  |  |  | 2023-12-13 15:10:40 |

### 3.3.2 Staging schema: dbstaging

**Data Location:**

~database/dbstaging.db – in gcpdw0

STAGING schema: dbstaging.staging table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | title | Varchar(500) |  |  |  | /nhật ký phát triển: Thay Đổi Trang Bị Mùa 2024 |
|  | description | Text |  |  |  | Hậu trường phía sau một số thay đổi về trang bị dành cho Mùa Giải 2024 của chúng tôi. |
|  | author\_name | Varchar(255) |  |  |  | Riot Endstep, Riot Rose7ea |
|  | timeUp | Timestamp |  |  |  | 29/11/2023 16:00:00 |
|  | url | Varchar(1000) |  |  |  | <https://www.leagueoflegends.com/vi-vn/news/dev/dev-2024-item-changes/> |
|  | category\_name | Varchar(500) |  |  |  | Cập Nhật Trò Chơi |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | content | Text |  |  |  | sức mạnh và cuối cùng bị tước đi hoàn toàn khả năng dồn sát thương tăng dần. Biểu Tượng (Lưu ý: Biểu Cảm Taliyah Tinh Thể Pha Lê sẽ được bổ sung trong phiên bản 13.19.)…. |
|  | source\_name | Varchar(255) |  |  |  | Riot Games |
|  | status | Varchar(500) |  |  |  | SC |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |

### 3.3.3 Staging schema: dbnew

**Data Location:**

~database/dbnew.db – in gcpdw0

STAGING schema: dbnew.game\_newsFact table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | title | Varchar(500) |  |  |  | Faker trở thành tuyển thủ đầu tiên đạt 1.000 trận thắng ở LCK |
|  | author\_id | SERIAL |  |  |  | 101 |
|  | description | Text |  |  |  | Faker đã trở thành tuyển thủ đầu tiên đạt 1.000 trận thắng ở LCK trong trận đấu giữa T1 và DRX. Faker đã có một màn trình diễn xuất sắc trong trận đấu này, giúp T1 giành chiến thắng với tỷ số 2-0. |
|  | url | Varchar(1000) |  |  |  | <https://www.leagueoflegends.com/vi-vn/news/dev/dev-2024-item-changes/> |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | content | Text |  |  |  | Các DJ đang soạn bài gì vậy nhỉ? Chào mừng đến với bản cập nhật đầu tiên của Remix Rumble. Đây là bản cập nhật vô cùng quan trọng vậy nên chúng tôi sẽ tập trung vào việc giảm sức mạnh của một vài tộc hệ đang hưởng các đặc quyền (Jazz, Đồng Quê, Fan Cứng, v.v) từ bản cập nhật mở màn cũng như "nâng đỡ" những tướng đang yếu thế (Yasuo, Yone, Viego, v.v).  ……….. |
|  | source | Text |  |  |  | Riot Games |
|  | category\_id | SERIAL |  |  |  | 5 |
|  | isDelete | INT(11) |  |  |  | 0 |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | updated\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | created\_by | Varchar(225) |  |  |  | Huy |
|  | update\_by | Varchar(225) |  |  |  | Huy |

STAGING schema: dbnew.authorsDim table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name | Varchar(500) |  |  |  | Rodger “Riot Prism” Caudill |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | updated\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | created\_by | Varchar(225) |  |  |  | Huy |
|  | update\_by | Varchar(225) |  |  |  | Huy |

STAGING schema: dbnew.categoriesDim table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name | Varchar(500) |  |  |  | Cập nhật trò chơi |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | updated\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | created\_by | Varchar(225) |  |  |  | Huy |
|  | update\_by | Varchar(225) |  |  |  | Huy |

STAGING schema: dbnew.homeAggregate table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name\_category | Varchar(500) |  |  |  | Cập nhật trò chơi |
|  | title | Varchar(500) |  |  |  | SKT vô địch valorant |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | description | Text |  |  |  | thức gia nhập DWG KIA. Theshy là một trong những top laner xuất sắc |
|  | name\_author | Varchar(500) |  |  |  | prism |
|  | day\_up | Timestamp |  |  |  | 14/12/2023 18:36:40 |

STAGING schema: dbnew.detailNewAggregate table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name\_category | Varchar(500) |  |  |  | Cập nhật trò chơi |
|  | title | Varchar(500) |  |  |  | SKT vô địch valorant |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | description | Text |  |  |  | thức gia nhập DWG KIA. Theshy là một trong những top laner xuất sắc |
|  | name\_author | Varchar(500) |  |  |  | prism |
|  | day\_up | Timestamp |  |  |  | 20/10/2022 8:00:00 |
|  | content | text |  |  |  | Một chân trời mới, một đỉnh cao mới! Chào mừng đến với cập nhật giữa mùa của ĐTCL! Trong bài viết này, chúng tôi sẽ liệt kê su viết Tổng Quan Lối Chơi Chân Trời Mới.  ……. |

### 3.3.4 Warehouse schema: dbmart

**Data Location:**

~database/dbmart.db – in gcpdw0

WAREHOUSE schema: dbmart.game\_news table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | title | Varchar(500) |  |  |  | Faker trở thành tuyển thủ đầu tiên đạt 1.000 trận thắng ở LCK |
|  | author\_id | SERIAL |  |  |  | 101 |
|  | description | Text |  |  |  | Faker đã trở thành tuyển thủ đầu tiên đạt 1.000 trận thắng ở LCK trong trận đấu giữa T1 và DRX. Faker đã có một màn trình diễn xuất sắc trong trận đấu này, giúp T1 giành chiến thắng với tỷ số 2-0. |
|  | url | Varchar(1000) |  |  |  | <https://www.leagueoflegends.com/vi-vn/news/dev/dev-2024-item-changes/> |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | content | Text |  |  |  | Các DJ đang soạn bài gì vậy nhỉ? Chào mừng đến với bản cập nhật đầu tiên của Remix Rumble. Đây là bản cập nhật tộc hệ đang hưởng các đặc quyền (Jazz, Đồng Quê, Fan "nâng đỡ" những tướng đang yếu thế (Yasuo, Yone, Viego, v.v).  ……….. |
|  | source | Text |  |  |  | Riot Games |
|  | category\_id | SERIAL |  |  |  | 5 |
|  | isDelete | INT(11) |  |  |  | 0 |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | updated\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | created\_by | Varchar(225) |  |  |  | Huy |
|  | update\_by | Varchar(225) |  |  |  | Huy |

WAREHOUSE schema: dbmart.authors table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name | Varchar(500) |  |  |  | Rodger “Riot Prism” Caudill |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | updated\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | created\_by | Varchar(225) |  |  |  | Huy |
|  | update\_by | Varchar(225) |  |  |  | Huy |

WAREHOUSE schema: dbmart.categories table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name | Varchar(500) |  |  |  | Cập nhật trò chơi |
|  | created\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | updated\_at | Timestamp |  |  |  | 14/12/2023 18:36:40 |
|  | created\_by | Varchar(225) |  |  |  | Huy |
|  | update\_by | Varchar(225) |  |  |  | Huy |

WAREHOUSE schema: dbmart.homeAggregate table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name\_category | Varchar(500) |  |  |  | Cập nhật trò chơi |
|  | title | Varchar(500) |  |  |  | SKT vô địch valorant |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | description | Text |  |  |  | thức gia nhập DWG KIA. Theshy là một trong những top laner xuất sắc |
|  | name\_author | Varchar(500) |  |  |  | prism |
|  | day\_up | Timestamp |  |  |  | 14/12/2023 18:36:40 |

WAREHOUSE schema: dbmart.detailNewAggregate table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id | SERIAL |  |  |  | 1 |
|  | name\_category | Varchar(500) |  |  |  | Cập nhật trò chơi |
|  | title | Varchar(500) |  |  |  | SKT vô địch valorant |
|  | image | Varchar(1000) |  |  |  | [https://images.contentstack.io/v3/assets/blt73](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [1acb42bb3d1659/blt8e70830e317b0e48/6567](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg)  [dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg](https://images.contentstack.io/v3/assets/blt731acb42bb3d1659/blt8e70830e317b0e48/6567dd4504116d4bb82a1a40/tft-patch-13-24-banner.jpg) |
|  | description | Text |  |  |  | thức gia nhập DWG KIA. Theshy là một trong những top laner xuất sắc |
|  | name\_author | Varchar(500) |  |  |  | prism |
|  | day\_up | Timestamp |  |  |  | 20/10/2022 8:00:00 |
|  | content | text |  |  |  | Một chân trời mới, một đỉnh cao mới! Chào mừng đến với cập nhật giữa mùa của ĐTCL! Trong bài viết này, chúng tôi sẽ liệt kê tất cả những thay đổi cho hệ thống hiện có, nhưng để nắm được suy nghĩ đằng sau cập nhật giữa mùa này, vui lòng tới đọc bài viết Tổng Quan Lối Chơi Chân Trời Mới.  … |

**Naming Convention:** The naming convention for the data feed is:

GetDataWeb\_yyyyMMdd\_HHmmss.csv

Where:

yyyyMMdd is the date of the file.

HHmmss is the time of the file.

server name: gcpdw0

**Delivery Schedule:** Daily delivery

## 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

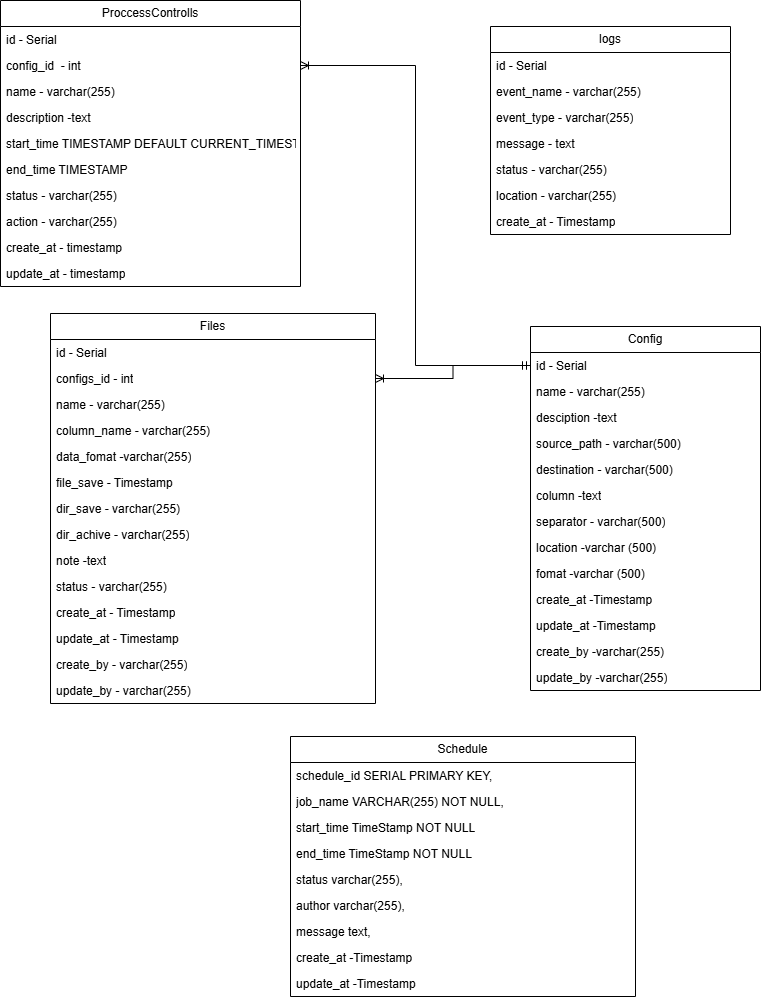
## 5.1. Overview

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



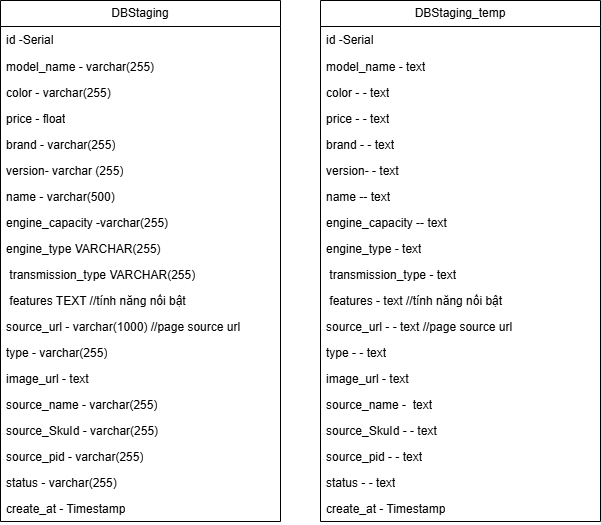
## In there:

## - 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

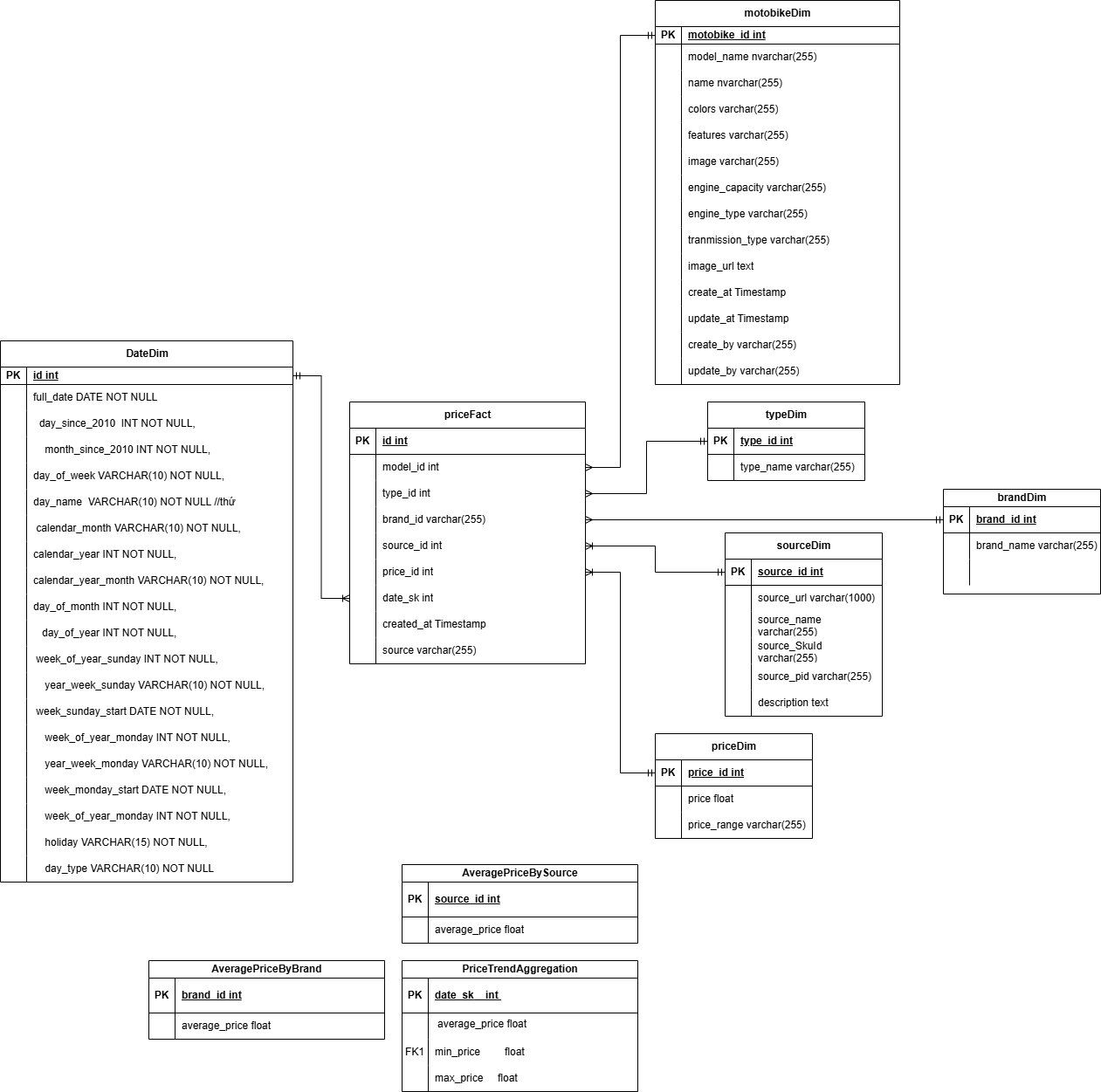


In there:

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

mss.csv

## 5.4. Database dbwarehouse:



In there:

- Dbnew schema: typeDim table is a place to store data taken from the staging table in dbstaging, processed to get data related to the type of motobike.

- Dbnew schema: dateDim table is a place to store data taken from the staging table in dbstaging, processed to get data related to the date and time of a row of data in warehouse

- Dbnew schema: priceFact table contains data taken from the staging table in dbstaging, processed to get data related to general price and information of the motobikes.

- Dbnew schema: sourceDim table contains data taken from the staging table in dbstaging, processed to get data related to crawled source. This table is designed to expand to crawl multiple data source .

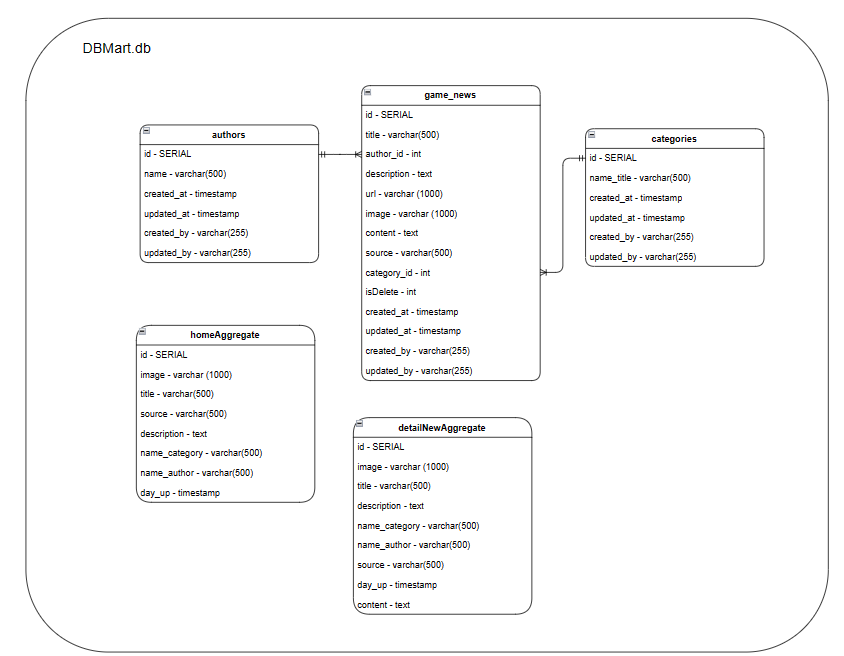
- Dbnew schema: motobikeDim table contains data taken from the staging table in dbstaging, processed to get data related to motobike.

- Dbnew schema: AveragePriceByBrandAggregate table contains data taken from the priceFact table in dbwarehouse itself, processed to only contain necessary information for the getting average price by brand

- Dbnew schema: AveragePriceBySourceAggregate table contains data taken from the priceFact table in dbwarehouse itself, processed to only contain necessary information for the getting average price by source (this serve for client want to know the best price supplier for a motobike)

- Dbnew schema: PriceTrendAggregate table contains data taken from the priceFact table in dbwarehouse itself, processed to only contain the trend of inscreasing or decreasing price by the time

## 5.5. Database dbmart:



## In there:

## - Dbmart schema: authors table contains data taken from the authorsDim table in dbnew, containing new data or updated edited data related to news authors.

## - Dbnew schema: categories table contains data taken from the categoriesDim table in dbnew, containing new data or updated edited data related to the categories contained in the news.

## - Dbnew schema: game\_news table contains data taken from the game\_newsFact table in dbnew, containing new data or updated edited data related to general information of the news.

## - Dbnew schema: homeAggregate table contains data taken from the homeAggregate table in dbnew itself, containing new data or updated edited data needed for the news website.

## - Dbnew schema: detailNewsAggregate table contains data taken from the detailNewsAggregate table in dbnew itself, containing necessary information and detailed content of each new news or updated edited data for the news website.

~END~