**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙢🙠🕮🙢🙠**



**ĐỒ ÁN MÔN DATA WAREHOUSE.**

**Đề tài: Bán Laptop.**

**Giảng viên hướng dẫn:** Nguyễn Đức Công Song

**Sinh viên thực hiện:** Huỳnh Ngọc Quang Minh – 21130445

**Nhóm**: 4

**Chuyên ngành:** Công nghệ thông tin

**Tp. Hồ Chí Minh, tháng 11 năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙢🙠🕮🙢🙠**



**ĐỒ ÁN MÔN DATA WAREHOUSE.**

**Đề tài: Bán Laptop.**

**Giảng viên hướng dẫn:** Nguyễn Đức Công Song

**Sinh viên thực hiện:** Huỳnh Ngọc Quang Minh – 21130445

**Nhóm**: 4

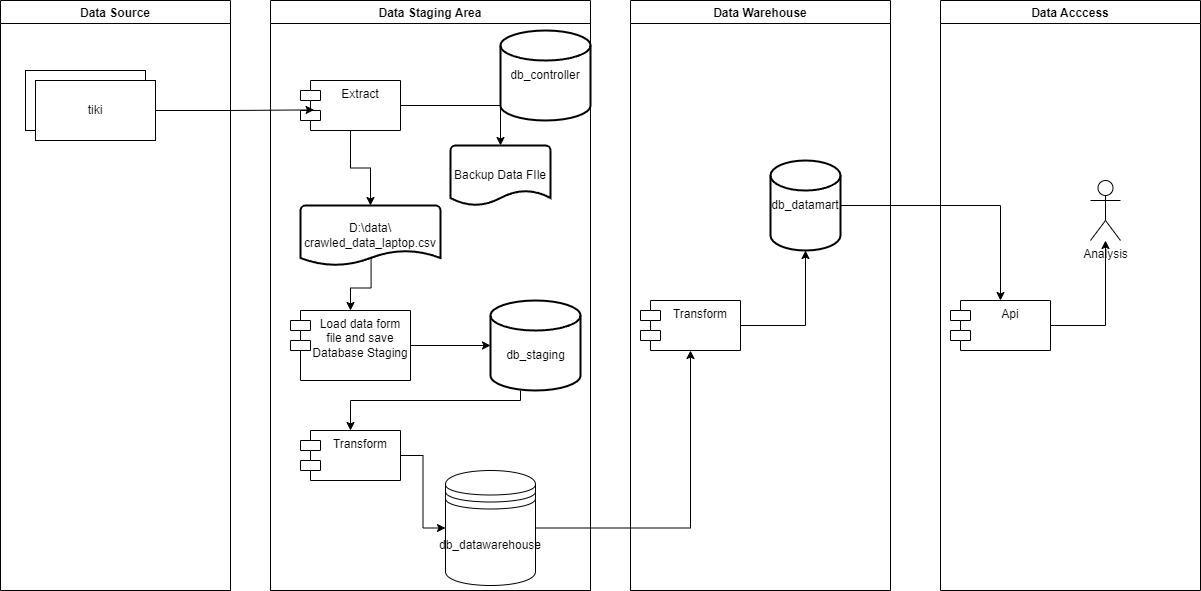
**Chuyên ngành:** Công nghệ thông tin

**Tp. Hồ Chí Minh, tháng 11 năm 2024**

MỤC LỤC

# OVERVIEW

## Current DW Architecture Overview



Link: [System Architecture](https://drive.google.com/file/d/10roI7XmV6CnCVxjEaGYMUifz05hFA71R/view?usp=drive_link)

## Propose DW Architecture Overview

* Các process phải thực hiện nhiệm vụ xác định:
* Lấy dữ liệu từ API về máy tính và lưu vào file csv.
* Load data từ file csv vào database staging.
* Tiến hành transform data.
* Thực hiện load dữ liệu từ database staging vào database datawarehouse.
* Phân tích và thống kê dữ liệu sau đó load lên data marts.
* Các process hoạt động độc lập với nhau.
* Dữ liệu được lấy vào lúc 7h sáng mỗi ngày
* Áp dụng cơ chế type 2: add row

# REFERENCE

|  |  |  |  |
| --- | --- | --- | --- |
| **SN** | **Tên tài liệu** | **Mô tả** | **Link** |
| 1 | Tổng hợp | Tài liệu tổng hợp chứa thông tin chi tiết về toàn bộ dự án. | [link](https://drive.google.com/drive/folders/1yBsZHg61Ut-BbRiXBBalJz1oe21XK3qu?usp=sharing) |
| 2 | Data Mapping | Mô tả cách mapping dữ liệu | [link](https://docs.google.com/spreadsheets/d/1JZGHxOiI8izGNTOvHk4isc-vfP5ABxjLwwODv_Gj4Xs/edit?usp=sharing) |
| 3 | Data feed specification | Tài liệu về cấu trúc cơ sở dữ liệu, bao gồm thiết kế bảng, quan hệ, và chi tiết về các trường dữ liệu chính. | [link](https://docs.google.com/spreadsheets/d/1cSdb6vve6z68qy6HMXFJ9Wo3JmK90kQSk1MXXp0ETII/edit?usp=sharing) |

# FUNCTIONAL SPECIFICATION

## Impact Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **SN** | **Criteria** | **Evaluation** | **Tên** |
| 1 | Extract modules | Lấy dữ liệu từ API về máy tính và lưu vào file csv. |  |
| 2 | Load Data Into Staging | Load dữ liệu từ file dữ liệu vào database staging |  |
| 3 | Load Data to Datawarehouse | Load dữ liệu từ database staging vào database datawarehouse | Huỳnh Ngọc Quang Minh |
| 4 | Load Data to Data Mart |  |  |

## Assumptions

Trong quá trình phát triển và triển khai chức năng này, một số giả định quan trọng đã được đưa ra để đảm bảo tính nhất quán và hiệu quả.

*Extract modules:*

*Load Data to Datawarehouse:*

* Có data trong database staging
* Có database controller

*Cấu hình hệ thống:*

## Exception Handling

### Normal Condition/ Exception catching

|  |
| --- |
| **[Load Data to Datawarehouse failed]**   * Ghi log vào database controller: nội dung lỗi, thời gian, process nào * Gửi thông báo Email: nội dung lỗi, thời gian chạy chương trình, thời gian kết thúc, process nào |

* **Normal Condition Checking**

|  |  |  |  |
| --- | --- | --- | --- |
| **SN** | **Criteria** | **Value** | **Action** |
| 4 | LoadDataWarhouse | RUNNING | * Không chạy process * Ghi Log vào db\_controller * Gửi Email “Process is already running” |
| READY | * Chuyển Process sang status RUNNING * Load dữ liệu sang Datawarehouse |
| FAILED | * Ghi Log vào db\_controller * Gửi Email “Failed” và thông tin lỗi |
| SUCCESS | * Cập nhật Process warehouse sang SUCCESS * Thêm Process datamart vào, với status READY * Ghi Log vào db\_controller * Gửi Email “SUCCESS” với số lượng dữ liệu đã load |

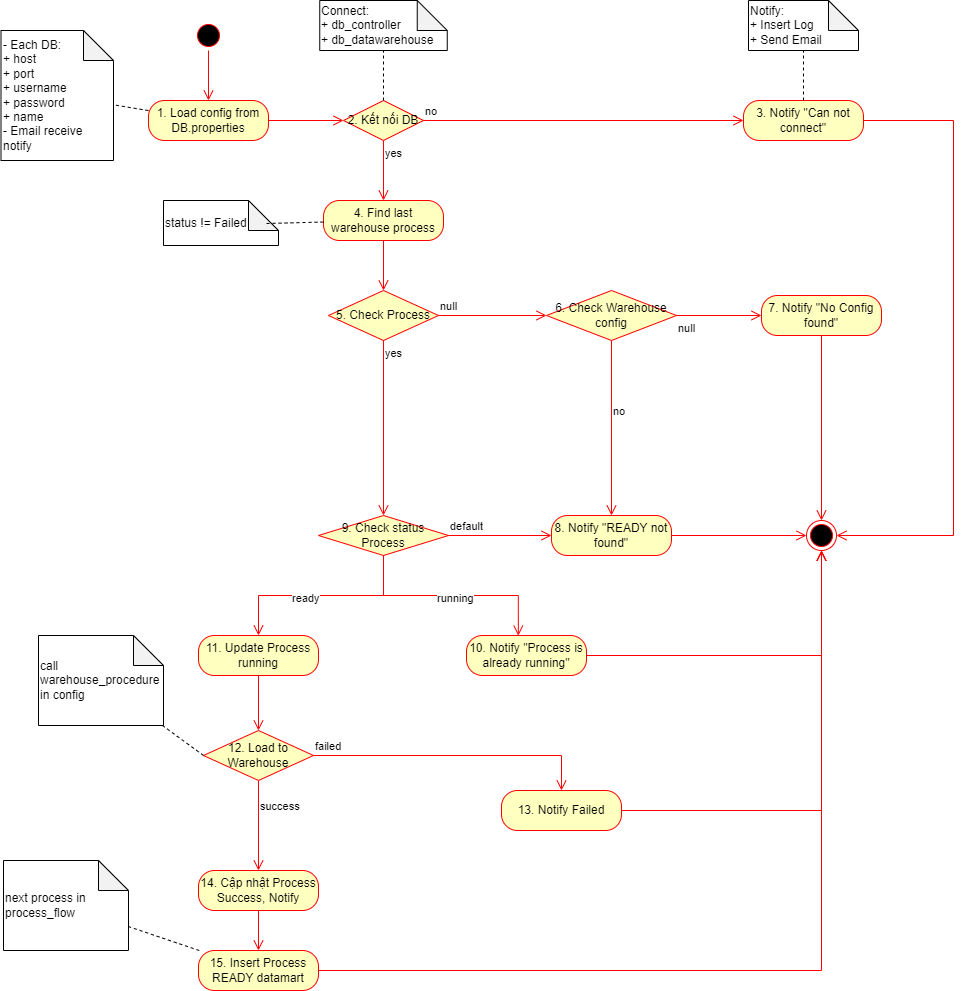
* **Exception/Error**

|  |  |  |  |
| --- | --- | --- | --- |
| **SN** | **Type** | **Exception** | **Handling** |
| 1 | Data error | **[LoadToDatawarehouse]**  Kiểu dữ liệu không khớp. | - Ghi lại Log  - Gửi Email  - Dừng chương trình |
| 2 | Configuration Error | **[LoadToDatawarehouse]** Không load được File DB.properties  Dữ liệu trong DB.properties không phù hợp  Không kết nối được database | - Gửi email thông báo lỗi |
| 3 | Structure error | **[LoadToDatawarehouse]**  Không có cột dữ liệu | - Ghi lại Log  - Gửi Email  - Dừng chương trình |
|  |  |
| 4 | Process error  (System interrupt) | **[LoadToDatawarehouse]** Không có Procedure | - Ghi lại Log  - Gửi Email  - Dừng chương trình |

## How to Run Job manually

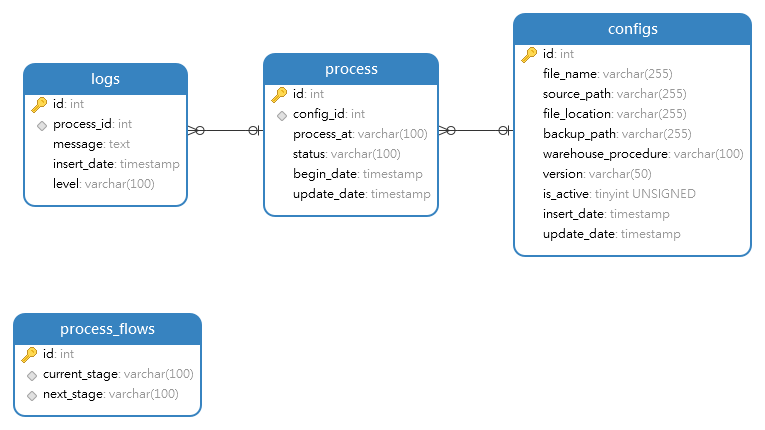
## Work flow

Work flow load data to datawarehouse:



## Data Structure

### Database controller:



|  |
| --- |
| Revisions:  **[2024-12-06]**  Initial |
| SET NAMES utf8mb4;  SET FOREIGN\_KEY\_CHECKS = 0;  -- ----------------------------  -- Table structure for config  -- ----------------------------  DROP TABLE IF EXISTS `config`;  CREATE TABLE `config` (  `id` int NOT NULL,  `file\_name` varchar(255) DEFAULT NULL,  `file\_path` varchar(255) DEFAULT NULL,  `file\_encoding` varchar(255) DEFAULT NULL,  `crawl\_frequency` int NOT NULL,  `data\_size` int NOT NULL,  `retry\_count` int NOT NULL,  `timeout` int NOT NULL,  `dw\_source\_port` int NOT NULL,  `staging\_source\_port` int NOT NULL,  `source\_path` varchar(255) DEFAULT NULL,  `destination\_path` varchar(255) DEFAULT NULL,  `backup\_path` varchar(255) DEFAULT NULL,  `file\_type` enum('CSV', 'JSON', 'XML') DEFAULT NULL,  `delimiter` varchar(255) DEFAULT NULL,  `dw\_source\_host` varchar(255) DEFAULT NULL,  `dw\_source\_password` varchar(255) DEFAULT NULL,  `dw\_source\_username` varchar(255) DEFAULT NULL,  `staging\_source\_host` varchar(255) DEFAULT NULL,  `staging\_source\_password` varchar(255) DEFAULT NULL,  `staging\_source\_username` varchar(255) DEFAULT NULL,  `columns` TEXT DEFAULT NULL,  `tables` varchar(255) DEFAULT NULL,  `note` varchar(255) DEFAULT NULL,  `notification\_emails` varchar(255) DEFAULT NULL,  `create\_time` datetime(6) DEFAULT NULL,  `update\_time` datetime(6) NOT NULL DEFAULT CURRENT\_TIMESTAMP(6) ON UPDATE CURRENT\_TIMESTAMP(6),  `created\_by` varchar(255) DEFAULT NULL,  `updated\_by` varchar(255) DEFAULT NULL,  `version` varchar(255) DEFAULT NULL,  `is\_active` bit(1) NOT NULL,  PRIMARY KEY (`id`) USING BTREE  ) ENGINE = InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_unicode\_ci ROW\_FORMAT=DYNAMIC;  DROP TABLE IF EXISTS `logs`;  CREATE TABLE `logs` (  `id` int NOT NULL AUTO\_INCREMENT,  `id\_config` int NOT NULL,  `count` int NOT NULL,  `log\_level` enum('DEBUG', 'ERROR', 'INFO', 'WARNING') DEFAULT NULL,  `status` enum('FAILURE\_EXTRACT', 'PROCESSING', 'READY\_EXTRACT', 'SUCCESS\_EXTRACT', 'FAILURE\_ LOAD\_DATA', 'SUCCESS\_LOAD\_DATA', 'FAILURE\_TRANSFORM, 'SUCCESS\_TRANSFORM) DEFAULT NULL,  `create\_time` datetime(6) NOT NULL DEFAULT CURRENT\_TIMESTAMP(6) ON UPDATE CURRENT\_TIMESTAMP(6),  `update\_time` datetime(6) DEFAULT NULL,  `created\_by` varchar(255) DEFAULT NULL,  `destination\_path` varchar(255) DEFAULT NULL,  `error\_message` varchar(255) DEFAULT NULL,  `location` varchar(255) DEFAULT NULL,  `stack\_trace` varchar(255) DEFAULT NULL,  PRIMARY KEY (`id`) USING BTREE,  FOREIGN KEY (`id\_config`) REFERENCES `config` (`id`) ON DELETE CASCADE ON UPDATE CASCADE  ) ENGINE = InnoDB AUTO\_INCREMENT=15 DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_unicode\_ci ROW\_FORMAT=DYNAMIC; |

### Staging table

|  |
| --- |
| Revisions:  **[2024-12-06]**  Initial |
| //Table: staging.product\_staging  CREATE TABLE if not exists product\_staging (  id INT AUTO\_INCREMENT PRIMARY KEY, -- ID tự tăng của bảng staging  product\_id BIGINT, -- ID sản phẩm  sku VARCHAR(50), -- Mã SKU  product\_name TEXT, -- Tên sản phẩm  price DECIMAL(15, 2), -- Giá bán hiện tại  original\_price DECIMAL(15, 2), -- Giá gốc  brand\_name VARCHAR(255), -- Tên thương hiệu  discount\_value DECIMAL(15, 2), -- Giá trị giảm giá  thumbnail\_url VARCHAR(500), -- URL ảnh thumbnail  short\_description TEXT, -- Mô tả ngắn  image\_urls TEXT, -- URL của nhiều hình ảnh, phân tách bằng dấu chấm phẩy (;)  color\_options VARCHAR(255), -- Các màu sản phẩm, phân tách bằng dấu chấm phẩy (;)  size\_options VARCHAR(100), -- Các kích thước (nếu có)  rating\_average DECIMAL(3, 2), -- Đánh giá trung bình  review\_count INT, -- Số lượng đánh giá  discount\_rate DECIMAL(5, 2), -- Tỷ lệ giảm giá (%)  quantity\_sold INT, -- Số lượng đã bán  url\_key VARCHAR(255), -- Khóa URL sản phẩm  url\_path VARCHAR(500), -- Đường dẫn URL sản phẩm  short\_url VARCHAR(500), -- URL rút gọn  product\_type VARCHAR(50), -- Loại sản phẩm (vd: configurable)  created\_at TIMESTAMP -- Thời gian tạo  );  //Table: staging. staging\_tiki  CREATE TABLE if not exists staging\_tiki (  id INT AUTO\_INCREMENT PRIMARY KEY, -- ID tự tăng của bảng staging  product\_id TEXT, -- ID sản phẩm  sku TEXT, -- Mã SKU  product\_name TEXT, -- Tên sản phẩm  price TEXT, -- Giá bán hiện tại  original\_price TEXT, -- Giá gốc  brand\_name TEXT, -- Tên thương hiệu  discount\_value TEXT, -- Giá trị giảm giá  thumbnail\_url TEXT, -- URL ảnh thumbnail  short\_description TEXT, -- Mô tả ngắn  image\_urls TEXT, -- URL của nhiều hình ảnh, phân tách bằng dấu chấm phẩy (;)  color\_options TEXT, -- Các màu sản phẩm, phân tách bằng dấu chấm phẩy (;)  size\_options TEXT, -- Các kích thước (nếu có)  rating\_average TEXT, -- Đánh giá trung bình  review\_count TEXT, -- Số lượng đánh giá  discount\_rate TEXT, -- Tỷ lệ giảm giá (%)  quantity\_sold TEXT, -- Số lượng đã bán  url\_key TEXT, -- Khóa URL sản phẩm  url\_path TEXT, -- Đường dẫn URL sản phẩm  short\_url TEXT, -- URL rút gọn  product\_type TEXT, -- Loại sản phẩm (vd: configurable)  created\_at DATETIME DEFAULT CURRENT\_TIMESTAMP  ); |

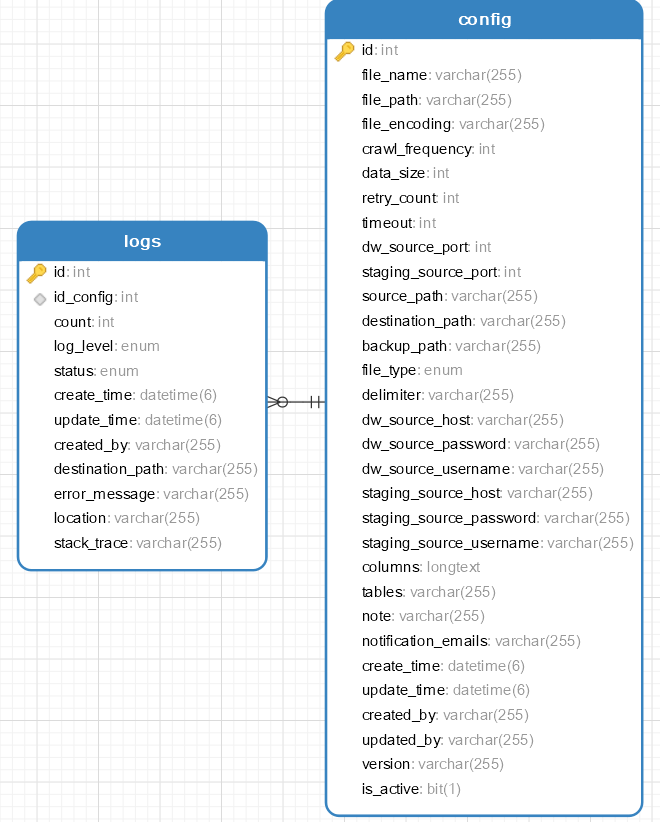
### Target Table

|  |
| --- |
| Revisions:  **[2024-12-06]**  Initial |
| create database if not exists datawarehouse;  use datawarehouse;  CREATE TABLE IF NOT EXISTS product\_dim (  id INT AUTO\_INCREMENT PRIMARY KEY,  product\_id BIGINT UNIQUE, -- Đảm bảo product\_id là duy nhất  sku VARCHAR(50),  product\_name TEXT,  brand\_name VARCHAR(255),  price DECIMAL(15, 2),  original\_price DECIMAL(15, 2),  discount DECIMAL(15, 2),  rating\_average DECIMAL(3, 2),  review\_count INT,  discount\_rate DECIMAL(5, 2),  quantity\_sold INT,  image\_urls TEXT,  product\_type VARCHAR(50),  thumbnail\_url VARCHAR(500),  url\_key VARCHAR(255),  url\_path VARCHAR(500),  short\_url VARCHAR(500),  short\_description TEXT,  create\_time TIMESTAMP DEFAULT CURRENT\_TIMESTAMP, -- Thời gian tạo  dt\_ex DATE DEFAULT '9999-12-31', -- Ngày hết hạn mặc định là 31/12/9999  isDelete BIT DEFAULT 0, -- Đánh dấu sản phẩm đã xóa  delete\_time TIMESTAMP NULL, -- Thời gian xóa sản phẩm  action\_time TIMESTAMP DEFAULT CURRENT\_TIMESTAMP ON UPDATE CURRENT\_TIMESTAMP -- Thời gian hành động  );  CREATE TABLE IF NOT EXISTS color\_dim (  color\_id INT AUTO\_INCREMENT PRIMARY KEY,  color\_name VARCHAR(100) UNIQUE  );  CREATE TABLE IF NOT EXISTS size\_dim (  size\_id INT AUTO\_INCREMENT PRIMARY KEY,  size\_value VARCHAR(50) UNIQUE  );  CREATE TABLE IF NOT EXISTS product\_color\_bridge (  product\_id BIGINT,  color\_id INT,  PRIMARY KEY (product\_id, color\_id),  FOREIGN KEY (product\_id) REFERENCES product\_dim(product\_id) ON DELETE CASCADE,  FOREIGN KEY (color\_id) REFERENCES color\_dim(color\_id) ON DELETE CASCADE  );  CREATE TABLE IF NOT EXISTS product\_size\_bridge (  product\_id BIGINT,  size\_id INT,  PRIMARY KEY (product\_id, size\_id),  FOREIGN KEY (product\_id) REFERENCES product\_dim(product\_id) ON DELETE CASCADE,  FOREIGN KEY (size\_id) REFERENCES size\_dim(size\_id) ON DELETE CASCADE  ); |

## Job Implementation

# Database Architecture

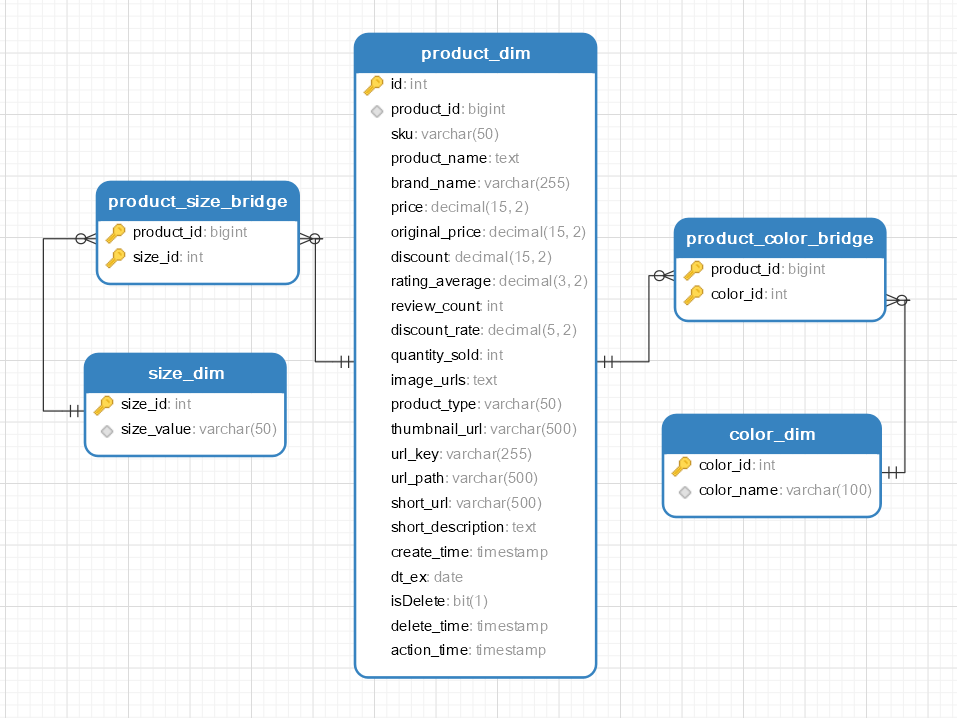
## Control



## Staging



## Data warehouse



# Setup Processes

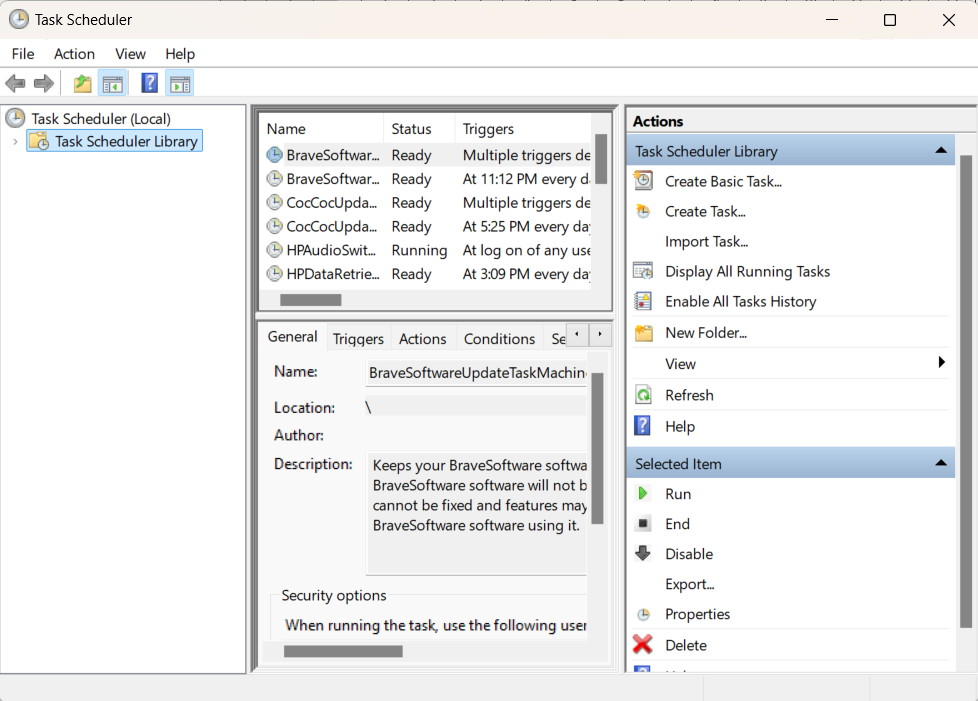
## Chuẩn bị.

Sử dụng hệ điều hành window vào tác vụ task scheduler để thực hiện.

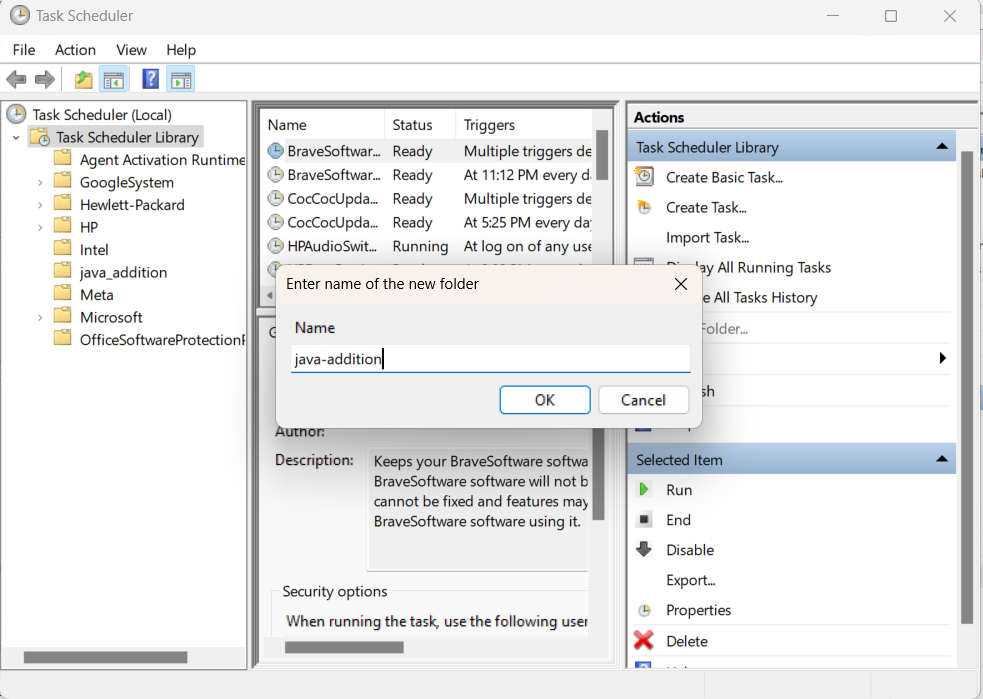
Khởi chạy Laragon

## Setup process

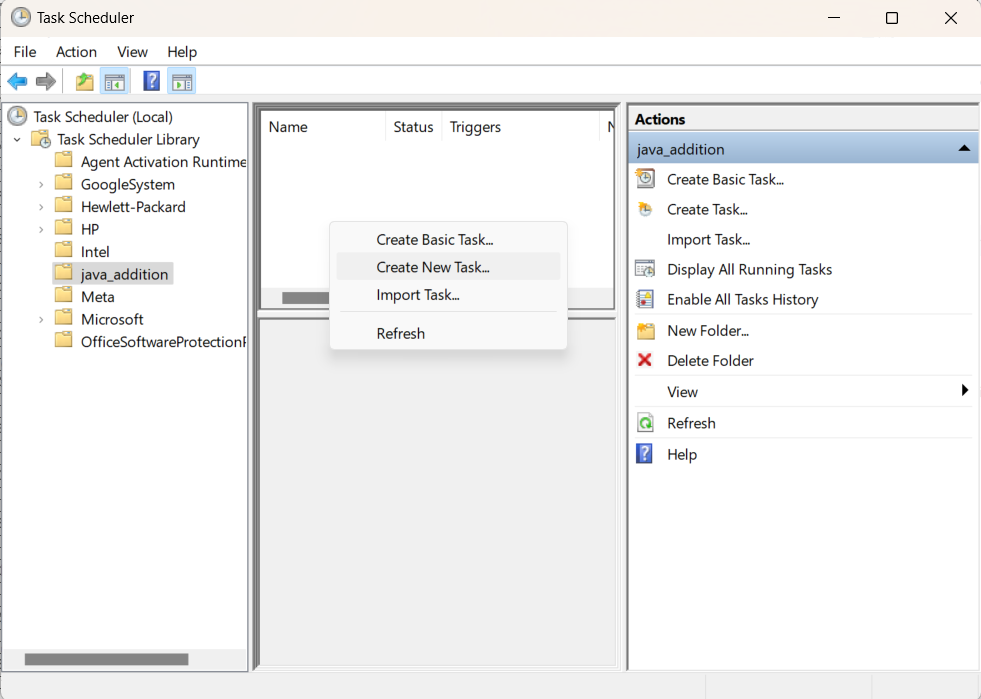
* Bước 1: Xác định đường dẫn file jar: D:\workspace\Project\data-warehouse-2024\Module1\_CrawlData\target\demo-0.0.1-SNAPSHOT.jar
* Bước 2: Khởi chạy task scheduler



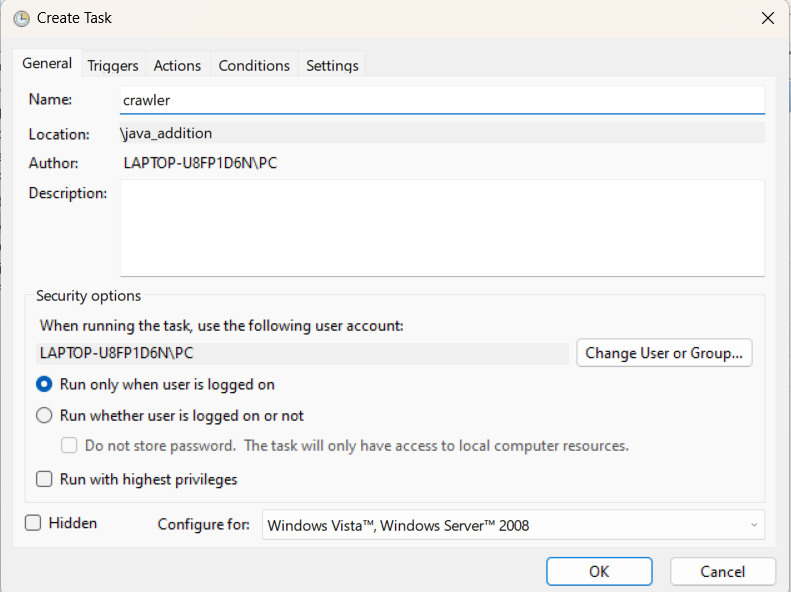
* Bước 3: Tạo thư mục java-addition để làm việc. Nhấp chuột phải chọn new folder.



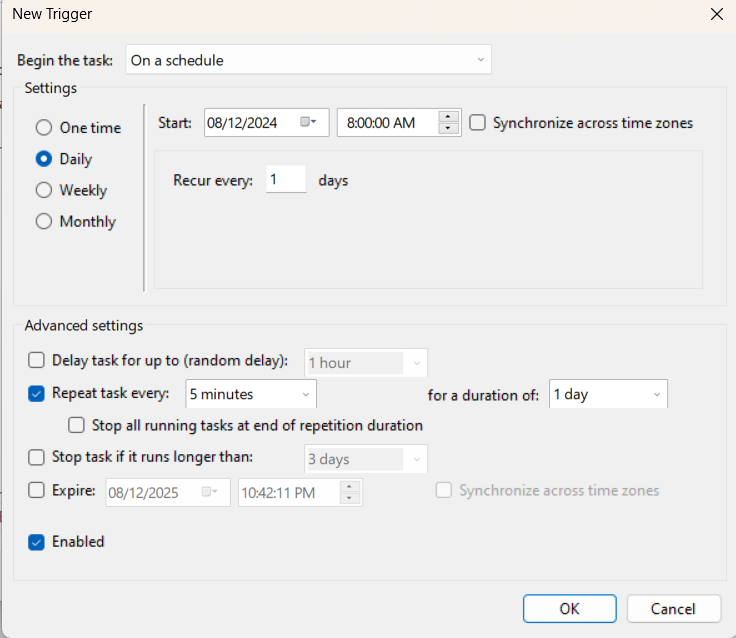
* Bước 3: Nhấp chuột phải vào khung trống và chọn create new task

****

* Bước 4: Ở tab general đặt tên task



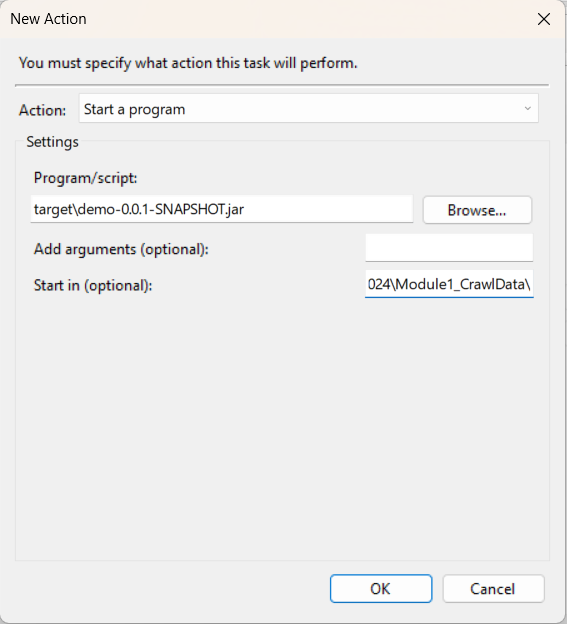
* Bước 5 Chuyển qua tab trigger nhấn new trigger cài đặt thời gian chạy cho file. Ở đây file crawl data chạy lúc 8h sáng mỗi ngày.



* Bước 6 Chuyển qua tab action

Action = Start a program

Program/script : Đường dẫn file đã ghi ở mục chuẩn bị



* Bước 7: Bấm OK để hoàn tất setup.
* Hướng dẫn: <https://youtu.be/ON2uRHlt_GI?si=j6dTjuB3u24s6mNJ>