**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

🙠🙠🙠 🕮 🙢🙢🙢



**DATA WAREHOUSE**

**Đề tài: Dự báo thời tiết**

Gíao viên hướng dẫn : Nguyễn Đức Công Song

Sinh viên thực hiện : Phạm Thùy Tiên

Lớp : DH20DTC

Mssv : 20130432

Ngày… Tháng … Năm 2023

Catalog

[1. Kiến trúc hệ thống 2](#_Toc2511)

[2. Mô tả dữ liệu đầu vào (Datafeed Specification) 2](#_Toc19237)

[2.1 Overview 3](#_Toc17986)

[2.1.1 Data Feed Specification 3](#_Toc30779)

[2.1.2 Contact Information 3](#_Toc5716)

[2.1.3 References 3](#_Toc1580)

[2.2 Transaction Definition 3](#_Toc26023)

[2.2.1 Overview 3](#_Toc20959)

[2.2.2 Data Feed Process Flow 3](#_Toc26155)

[2.2.3 Validation Guidelines 6](#_Toc29595)

[2.2.4 Attribute Specification 6](#_Toc11452)

[2.2.5 Code Value for weather data 7](#_Toc23483)

[2.2.6 Data Source Extraction and Exception Handling Procedures 7](#_Toc908)

[2.3 Feed Architecture 7](#_Toc25051)

[2.3.1 Overview 7](#_Toc10519)

[2.3.2 Transport Mechanism 8](#_Toc17047)

[2.3.3 Feed Characteristics 8](#_Toc10095)

[2.3.4 Data/Process Controls 9](#_Toc10256)

[2.3.5 Checksum Definition (Not Applicable) 9](#_Toc3869)

[2.3.6 Control File Format (Not Applicable) 9](#_Toc31703)

[2.3.7 FTP Setup (Not Applicable) 9](#_Toc31220)

[2.3.8 Operations Interface (Not Applicable) 9](#_Toc29934)

[2.4 SLA Negotiation (Not Applicable) 9](#_Toc7553)

[Appendix A – Attribute Specification 9](#_Toc10997)

[3. Cấu trúc CSDL của hệ thống 10](#_Toc12801)

[ Weather database 10](#_Toc32718)

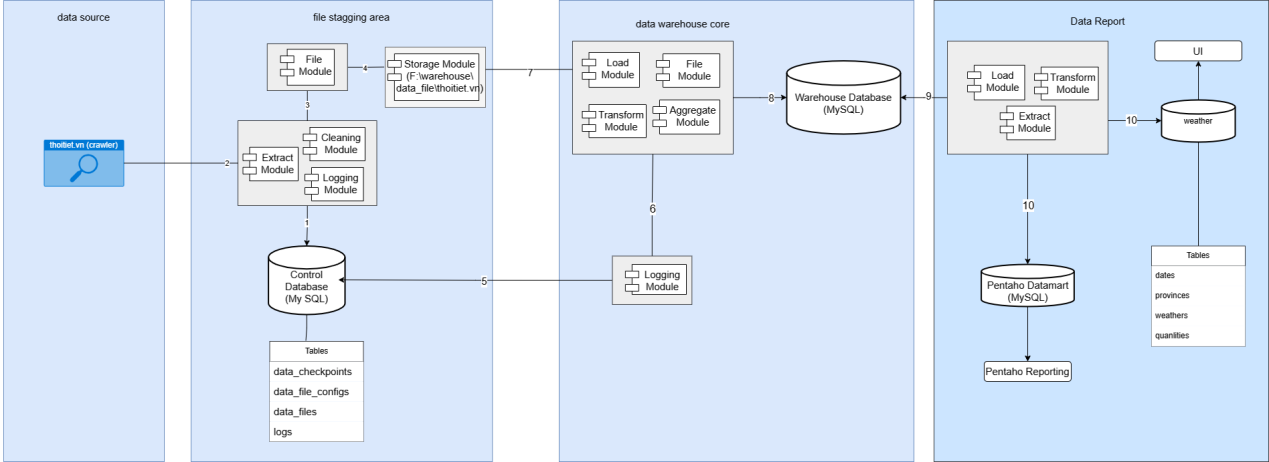
[ Data Warehouse Database 11](#_Toc23047)

[4. Sơ đồ luồng thực hiện 12](#_Toc31535)

[Sơ đồ ở đây 12](#_Toc22573)

1. **Kiến trúc hệ thống**

**[Sơ đồ ở đây](https://drive.google.com/file/d/1x2P0vhfs2qnmkrbgUi2hlD-n2VkJsqD7/view?usp=sharing)**

****

1. **Mô tả dữ liệu đầu vào (Datafeed Specification)**

**Data Feed Specification**

**Thoitiet.vn**

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Author** | **Changes** |
| 08/10/2023 | 1 | Phạm Thùy Tiên | Initialize |

**2.1 Overview**

**2.1.1 Data Feed Specification**

This document provides a detailed description for getting data from the website thoitiet.vn using crawl and ETL methods from the application server.

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 (HTTP), feed characteristics (format, naming convention, frequency, etc), data/process controls (control counts, checksums, etc.), and operations interface

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

**2.1.2 Contact Information**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Contact Email** | **Contact Phone** | **Role** |
| Phạm Thùy Tiên | 20130432@st.hcmuaf.edu.vn | 0393326958 | Owner |

**2.1.3 References**

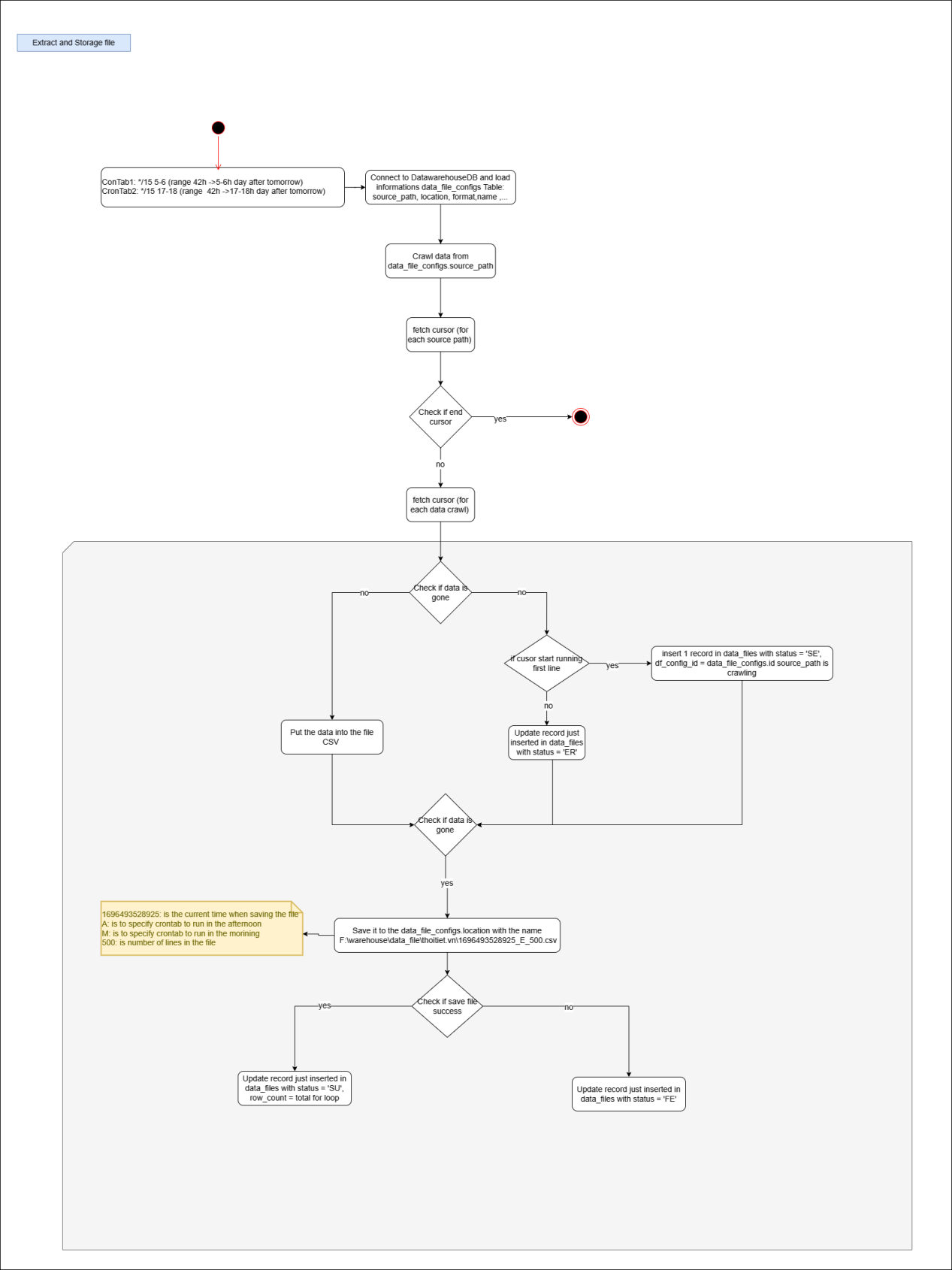
|  |  |  |  |
| --- | --- | --- | --- |
| **Document** | **Version** | **Date** | **Author** |
| Data Warehouse Systems design and Implementation | first |  | Alejandro Vaisman Esteban Zimanyi |
| The data warehouse toolkit | second |  | Ralph Kimball Margy Ross |

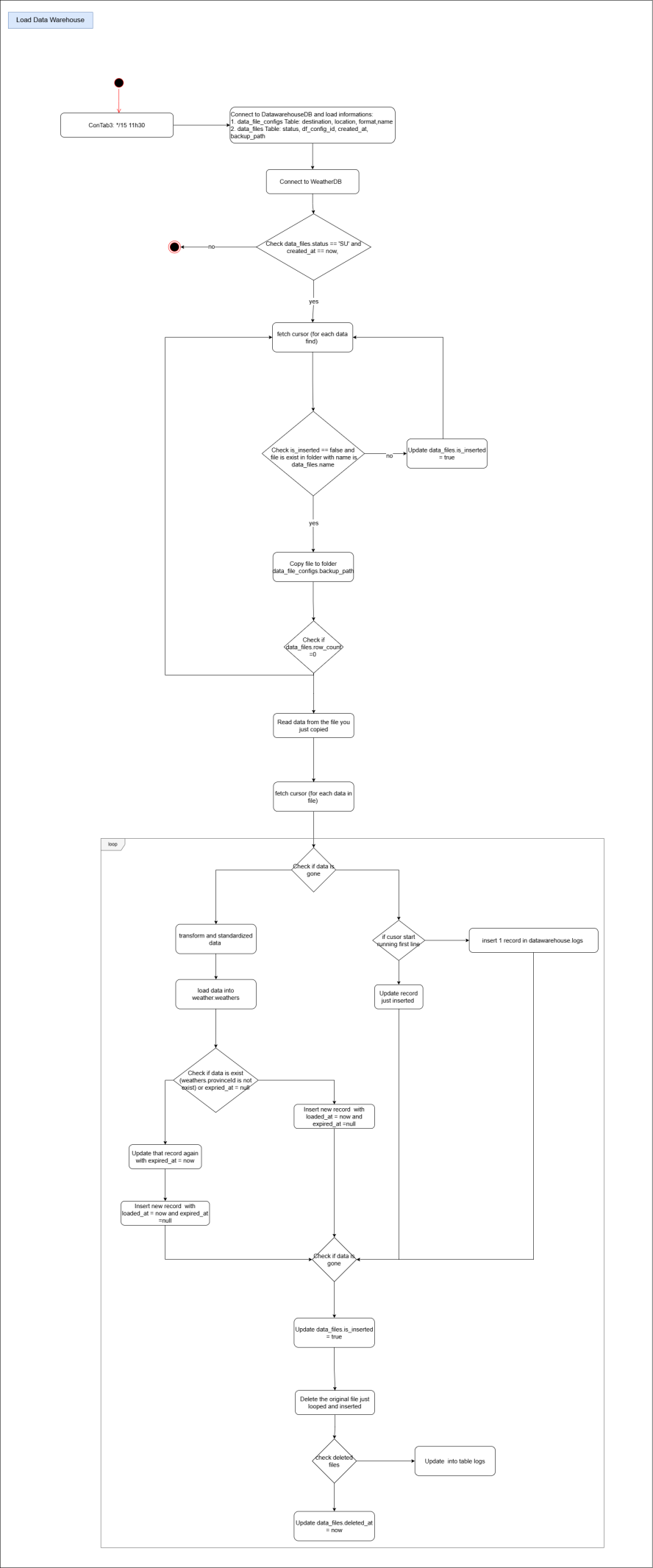
**2.2 Transaction Definition**

**2.2.1 Overview**

**2.2.2 Data Feed Process Flow**

**[Sơ đồ ở đây](https://drive.google.com/file/d/1x2P0vhfs2qnmkrbgUi2hlD-n2VkJsqD7/view?usp=sharing)**

****

****

**2.2.3 Validation Guidelines**

**2.2.4 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

**2.2.4.1 Weather Data Attribute Specification**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Field Name | Format | Mask | Edit Rules | Description | Example | Unit |
| Date | Date |  |  | Ngày dự đoán thời tiết |  |  |
| Time | Time |  |  | Giờ dự báo tời tiết |  |  |
| Province | Text |  |  | Tỉnh dự báo thời tiết | Kiên Giang |  |
| Wards | Text |  |  | Xã dự báo thời tiết | Bãi Thơm |  |
| District | Text |  |  | Huyện dự báo thời tiết | Phú Quốc |  |
| Temperature | double |  |  | Nhiệt độ thời tiết | 16 độ | độ C |
| Feeling | double |  |  | Cảm giác nhiệt độ thời tiết | Như 25 độ | Độ C |
| status | Text |  |  | Trạng thái của thời tiết | Nhiều mây |  |
| Humidity | double |  |  | Độ ẩm thời tiết | 83% | % |
| Vision | double |  |  | Tầm nhìn của thời tiết | 10 km | Km |
| Wind\_speed | double |  |  | Sức gió của thời tiết | 1.16 km/h | Km/h |
| Stop\_point | double |  |  | Điểm ngừng của độ ẩm | 21 độ | Độ C |
| Uv\_index | double |  |  | Chỉ số tia uv trong không khí | 0.29 |  |
| Airquality | Text |  |  | Chất lượng không khí của thời tiết | tốt |  |
| Last\_update\_time | integer |  |  | Giờ cập nhật gần nhất từ source | 6 phút trước | phút |
| Breadcrumb |  |  |  |  |  |  |
| Url | Text |  |  | Đường dẫn tuyệt đối đến source | https://thoitiet.vn/lai-chau/muong-te/bum-nua |  |
| Path | Text |  |  | Đường dẫn tương đối đến source | /lai-chau/muong-te/bum-nua |  |
| Dtrequest | Text |  |  | Request timestamp (YYY-MM-DD hh:mm:ss[+-]oooo) | [09/Jun/2011:00:02:48 -0700] |  |
| Request | Text |  |  | Full request string (method+path+args+protocol) | GET /lai-chau/muong-te/bum-nua |  |
| Method | Text |  |  | Request method | GET |  |
| Protocols | Text |  |  | Request protocol | HTTP/1.0 |  |
| Status | integer |  |  | Request status | 200 |  |
| Host | text |  |  | Application server | app11-tomcat |  |
| Server | text |  |  | WAP site hostname | m.godanriver.com |  |
| Ip | Text |  |  | Client IP address | 74.82.68.33 |  |

**2.2.5 Code Value for weather data**

|  |  |  |
| --- | --- | --- |
| **Field Name** | **code** | **List Of Code** |
| date | dd/mm/yy |  |
| Time | hh/mm/ss |  |
| Airquality |  | **G, R, M, P, VP** |

**2.2.6 Data Source Extraction and Exception Handling Procedures**

**2.2.6.1 Special Extract Procedure:**

**2.2.6.2 Error Code**

Resolution 1: No find source (404)

**2.2.6.3 Exception Handling Procedure (Not Applicable)**

**2.2.6.4 Special Design Consideration (Not Applicable)**

**2.3 Feed Architecture**

## **2.3.1 Overview**

The key components of the feed architecture are:

**Transport Mechanism**.

Data will flow indirectly from website thoitiet.vn through proccess crawl data and ELT and load into Staging

**Feed Characteristics**.

Log file :  {milisecondtime}\_{A,M}\_{total record}.csv

Frequency :  Daily

Content :  Contains data from **previous** day

Estimated Table size :  ~ 130MB/day

Pre-processing :  no

Estimated monthly size req :  130\*30 + 1400=5300MB/month

**SOURCE -> IMPORT**

Website thoitiet.vn : crawling -> F:\warehouse\data\_file\thoitiet.vn

**Data/Process Controls**.

**Operations Interface.**

The procedures used for the two sets of operation's staff to manage the feed processes. The procedures will typically identify templates for processing procedures, escalation procedures for problems, and required contact data.

**Operations Interface.**

The procedures used for the two sets of operation's staff to manage the feed processes. The procedures will typically identify templates for processing procedures, escalation procedures for problems, and required contact data.

## **2.3.2 Transport Mechanism**

This process is flat file to database data transfer.

## **2.3.3 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.

**Data Size:** Approximately: ~130MB/Day

**Data Location:** F:\warehouse\data\_file\thoitiet.vn in Staging

**Data Frequency:** Daily

**Delivery Location:** Is where the process of crawling data and saving files takes place

STAGING schema: data\_file\_configs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Sample value** |
| id | varchar | 500 |  |  | 1 |
| name | text |  |  |  | thoitiet.vn |
| code | varchar | 1000 |  |  | 123a |
| description | varchar | 1000 |  |  | trang web thời tiết 63 tỉnh thành việt nam |
| source\_path | varchar | 255 |  |  | https://thoitiet.vn/ |
| location | varchar | 255 |  |  | F:\warehouse\data\_file\thoitiet.vn |
| format | text |  |  |  | csv |
| separator | varchar | 1000 |  |  | tinh nhiet do đá ádasd |
| columns | timestamp |  |  |  | ádasd tinh nhiet do đá ádasd ádasd |
| destination | timestamp |  |  |  | {  "username":"root",  "password":"",  "port":8080,  "ip":"localhost",  "type":"mysql",  "name":"datawarehouse"  } |
| created\_at | int | 11 |  |  | 2023-10-08 20:12:01 |
| updated\_at | int | 11 |  |  | 2023-10-08 20:12:01 |
| created\_by | int | 11 |  |  | 1 |
| updated\_by | int | 11 |  |  | 2 |

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

{milisecondtime}\_{A,M}\_{total record}.csv

Where:

milisecondtime is the current time when saving the file  
A: is to specify crontab to run in the afternoon  
M: is to specify crontab to run in the morining  
total record is number of lines in the file

**Delivery Schedule:** Daily / Month-End delivery

## **2.3.4 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** |
| --- | --- | --- |
| SE | Start extract | Work extract file |
| SU | Status success | Report success |
| EF | Extract fail | Report extract fail |

## **2.3.5 Checksum Definition (Not Applicable)**

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

## **2.3.6 Control File Format (Not Applicable)**

## **2.3.7 FTP Setup (Not Applicable)**

## **2.3.8 Operations Interface (Not Applicable)**

## **2.4 SLA Negotiation (Not Applicable)**

# **Appendix A – Attribute Specification**

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

| **Attribute Format** | **Description** | **Example** |
| --- | --- | --- |
| String(x) | Character data of length x characters. Be sure to allow for the maximum length possible for the attribute. | '123 Willow Road' is a character data with a length of 15 characters. |
| Number(x,y) | Numeric data with total length of x characters with y characters after the decimal point. | 12.45 would be Number(4,2), 1123 would be Number(4) |
| Date | Date data. The length and format will be dependent upon the mask used. The default mask is MM/DD/YYYY. | 02/28/2002 is February 28th, 2002 |

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

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| MM | Month in numeric format | 01 is January |
| MON | Month in string format | JANUARY |
| DD | Day in numeric format | 23 is the 23rd day in the month |
| YYYY | Year including century | 2002 |
| YY | Year excluding century | 02 |
| HH | Hour in 24 hour format | 23 is 11pm |
| SS | Second in minute | 05 is second 05 |

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

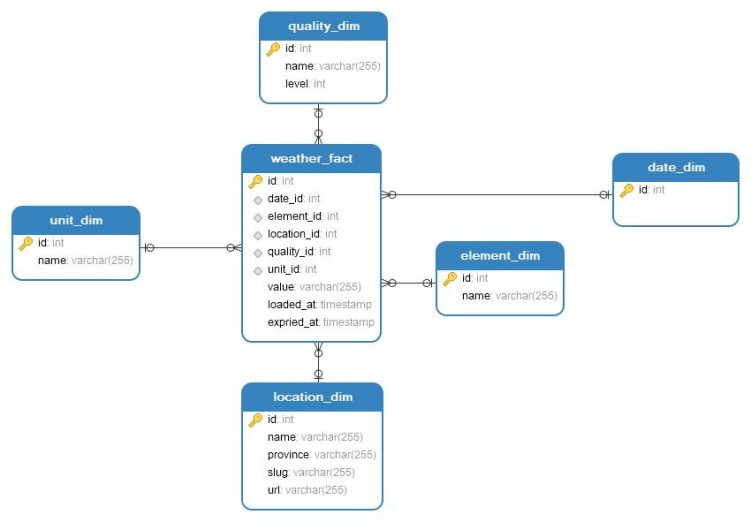
| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| INITCAP | All words in attribute province are capitalized | Hồ Chí Minh |

1. **Cấu trúc CSDL của hệ thống**

* **Weather database**

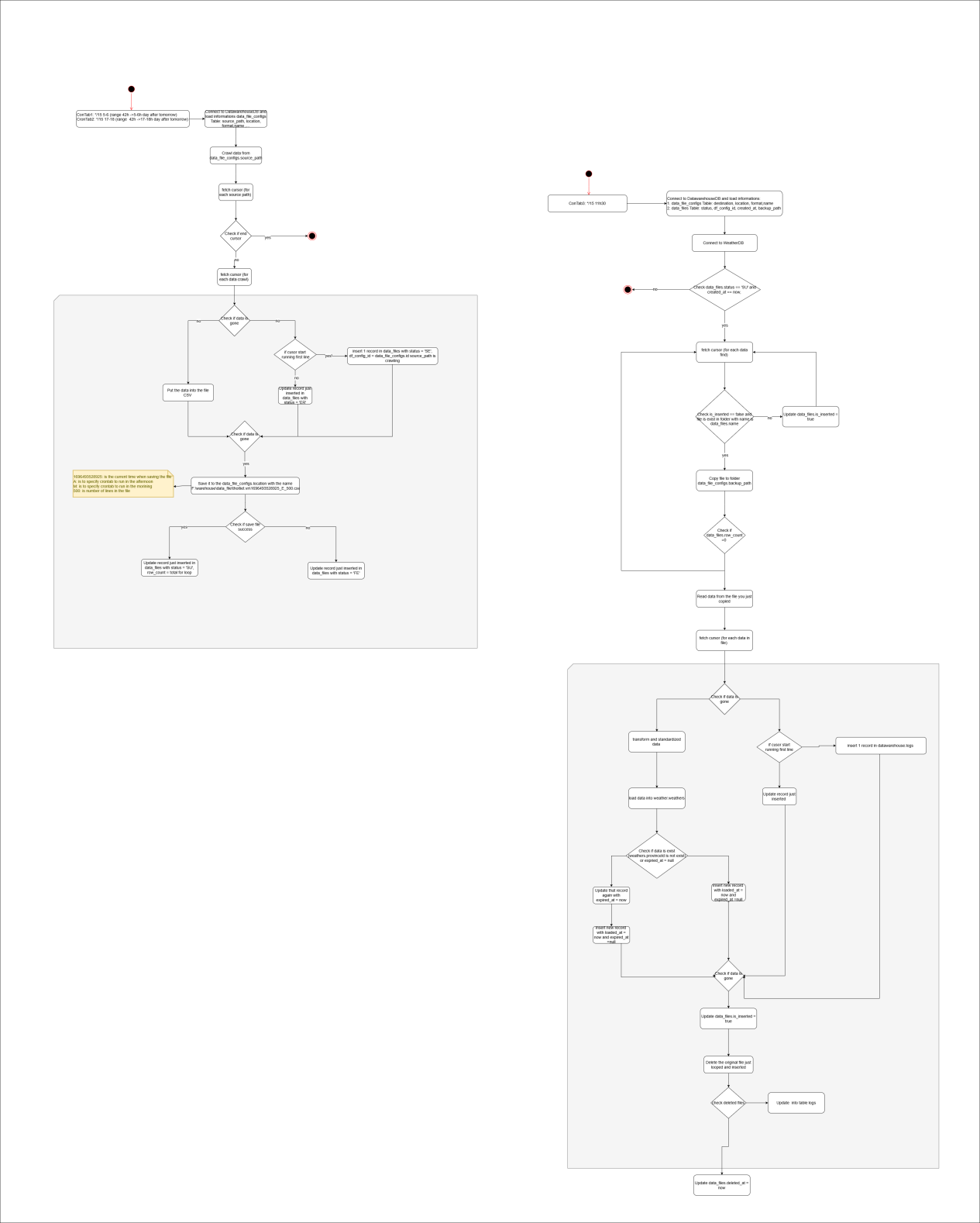
****

* **Data Warehouse Database**



1. **Sơ đồ luồng thực hiện**

**[Sơ đồ ở đây](https://drive.google.com/file/d/1x2P0vhfs2qnmkrbgUi2hlD-n2VkJsqD7/view?usp=sharing)**

****