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

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**DATA FEED SPECIFICATION**

**ĐỀ TÀI**

**KIỂM TRA SỐ XỐ**

Giảng viên: Nguyễn Công Đức Song

Đặng Minh Tiến

Nhóm 13:

19130011 – Nguyễn Kỳ Anh

19130154 – Võ Chí Nguyên

19130257 – Phạm Anh Tuấn

**Ngày 30 Tháng 11 Năm 2022**

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Author | Changes |
| 7/9/2022 | 1.0 | 13rd Group | First Change |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Mục lục

[1.0 Overview 4](#_Toc120395361)

[1.1 Data Feed Specification 4](#_Toc120395362)

[1.2 Contact Information 4](#_Toc120395363)

[1.3 References 4](#_Toc120395364)

[2.0 Transaction Definition 5](#_Toc120395365)

[2.1 Overview 5](#_Toc120395366)

[2.2 Data Feed Process Flow 5](#_Toc120395367)

[2.3 Validation Guidelines 5](#_Toc120395368)

[2.4 Attribute Specification 5](#_Toc120395369)

[2.4.1 Lotto check 5](#_Toc120395370)

[Attribute Specification 5](#_Toc120395371)

[2.5 Code Values 6](#_Toc120395372)

[2.6 Data Source Extraction and Exception Handling Procedures 7](#_Toc120395373)

[2.6.1 Special Extract Procedure: Source 🡪 Import 7](#_Toc120395374)

[2.6.2 Error Code 7](#_Toc120395375)

[2.6.3 Exception Handling Procedure 7](#_Toc120395376)

[2.6.4 Special Design Consideration (Not Applicable) 7](#_Toc120395377)

[3.0 Feed Architecture 7](#_Toc120395378)

[3.1 Overview 7](#_Toc120395379)

[3.2 Transport Mechanism 8](#_Toc120395380)

[3.3 Feed Characteristics 8](#_Toc120395381)

[3.4 Data/Process Controls 10](#_Toc120395382)

[3.5 Checksum Definition (Not Applicable) 10](#_Toc120395383)

[3.6 Control File Format 10](#_Toc120395384)

[3.7 FTP Setup 11](#_Toc120395385)

[3.8 Operations Interface 11](#_Toc120395386)

[4.0 SLA Negotiation (Not Applicable) 11](#_Toc120395387)

[Appendix A – Attribute Specification 12](#_Toc120395388)

[Appendix B – Attribute specification for each relational table 13](#_Toc120395389)

[Appendix C – Error Codes & Exception Handling 13](#_Toc120395390)

[Appendix D – Raw log file example 13](#_Toc120395391)

# 1.0 Overview

## 1.1 Data Feed Specification

Tài liệu này cung cấp chi tiết về mô tả của trang web xổ số.

Phần đầu tiên của tài liệu cung cấp định nghĩa dữ liệu nằm trong nguồn cấp dữ liệu và chi tiết về thuộc tính. Đặc tả thuộc tính bao gồm định nghĩa, định dạng và các luật quy định.

Phần kế tiếp mô tả kiến trúc của tiến trình chạy. Bao gồm: cơ chế chuyển đổi (FTP) , feed characteristics (format, naming convention, frequency, etc), data/process controls (control counts, checksums, etc.), and operations interface

Sau khi hoàn thành việc test, lịch trình triển khai sẽ được triển khai và data feed được trích xuất.

## 1.2 Contact Information

Các thành viên bên dưới được giao nhiệm vụ cho từng khu vực cho từng data feed.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Contact Email** | **Contact Phone** | **Role** |
| Nguyễn Kỳ Anh | [19130011@st.hcmuaf.edu.vn](mailto:19130011@st.hcmuaf.edu.vn) | 0915039849 | Thành viên |
| Võ Chí Nguyên | [19130154@st.hcmuaf.edu.vn](mailto:19130154@st.hcmuaf.edu.vn) | 0819778801 | Thành viên |
| Phạm Anh Tuấn | [19130257@st.hcmuaf.edu.vn](mailto:19130257@st.hcmuaf.edu.vn) | 0773506505 | Thành viên |

## 1.3 References

|  |  |  |  |
| --- | --- | --- | --- |
| **Document** | **Version** | **Date** | **Author** |
| Lotto Check | 0.1 | 14/9 | Nhóm 13 |
|  |  |  |  |
|  |  |  |  |

# 2.0 Transaction Definition

## 2.1 Overview

## 2.2 Data Feed Process Flow

## 2.3 Validation Guidelines

Đối với các trường bắt buộc được xác thực bởi các trigger sẵn có hoặc được tạo trong Mysql hoặc các database server qui định.

## 2.4 Attribute Specification

Phần này cung cấp các định nghĩa ngắn gọn về từng thuộc tính, bao gồm định dạng trường và tiêu chí chỉnh sửa. Để biết thêm thông tin về các thuật ngữ được sử dụng cho định nghĩa thuộc tính

### 2.4.1 Lotto check

### Attribute Specification

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Field Name** | **Format** | **Mask** | **Edit Rules** | **Description** | **Example** |
|  | id\_province | text |  | != null | Identify number of province | 68 |
|  | name\_province | text |  | != null | Name of province | Kiên Giang |
|  | id\_prize | text |  | != null | Identify number of prize | giai8 |
|  | name\_prize | text |  | != null | Name of prize | Giải tám |
|  | value\_prize | interger |  | >0 | Money of prize | 100000 |
|  | number | interger |  | >0 | Figure of number | 93 |
|  | created\_date | date | YYYY-MM-DD | != null | The date that figure of number is showed | 2022-09-18 |
|  | updated\_date | date | YYYY-MM-DD |  | Ngày cập nhật dữ iệu | 9999-12-31 |
|  | status | text |  | != null | Trạng thái của dòng dữ liệu | ER |
|  | natural\_key | text |  | != null | Khóa tự nhiên | 1fn2gh3abc |
|  | id\_config | integer |  | >0 | Mã định danh của dòng dữ liệu trong bảng config | 1 |
|  | server\_name | text |  | != null |  | 103.97.126.21 |
|  | url | text |  | != null |  | https://www.minhngoc.net.vn/xo-so-mien-nam.html |
|  | user\_name | text |  | != null |  | cineqoon |
|  | password | text |  | != null |  | 0819778801DEV |
|  | path\_upload | text |  | != null | Đường dẫn đăng file lên ftp | C:\Users\admin\Desktop\FTP-Upload\ |
|  | path\_download | text |  | != null | Đường dẫn tải file từ ftp | C:\Users\admin\Desktop\FTP-Down\ |
|  | id\_contactor | integer |  | >0 | Mã định danh của người tải dữ liệu | 1 |
|  | name | text |  | != null | Tên của người tải dữ liệu | Nguyen Ky Anh |
|  | email | text |  | != null | Email của người tải dữ liệu | 19130011@st.hcmuaf.edu.vn |
|  | id\_log | integer |  | >0 | Mã định danh của dòng dữ liệu trong bảng log | 1 |

## 2.5 Code Values

Code values là từ viết tắt cho chuẩn giá trị được sử dụng cho 1 thuộc tính. Chúng chỉ được sử dụng khi 1 thuộc tính có miền giá trị đã biết. Nó cung cấp 1 kỹ thuật ảnh hưởng hơn cho lưu trữ tập lớn của việc lặp lại các giá trị. Chúng được sử dụng cho mọi thứ từ các loại và trạng thái của giao dịch.

Đối với bất kỳ thuộc tính nào được xác định là LOV (List of Values) do nhà cung cấp cung cấp, danh sách đầy đủ các giá trị được cung cấp cùng với các mô tả được liên kết với giá trị mã đó.

Duy trì code values có thể là 1 vấn đề - cả 2 bên đều phải duy trì đồng bộ. Điều này được hỗ trợ bởi quy trình vận hành và phiên bảng feed.

**List of Values (LOV)**

## 2.6 Data Source Extraction and Exception Handling Procedures

### 2.6.1 Special Extract Procedure: Source 🡪 Import

Chuyển các file .csv thuộc folder ngày đó mà được lưu từ FPT server vào trong database staging. Trong database staging, dữ liệu được luôn được cập nhật theo từng ngày. Đối với dữ liệu trùng và không có sự thay đổi thì không được cập nhật, với dữ liệu mới hoàn toàn được được ghi vào các bảng tương ứng và dữ liệu trùng mã (id) thì được cập nhật lại các trường mới tương ứng và ngày cập nhật sẽ được thay đổi.

### 2.6.2 Error Code

Cách 1: (xem Appendix C)

### 2.6.3 Exception Handling Procedure

Cách 1: (xem Appendix C)

### 2.6.4 Special Design Consideration (Not Applicable)

# 3.0 Feed Architecture

## 3.1 Overview

Các thành phần chính của feed architecture là:

**Transport Mechanism**.

Dữ liệu được lấy trực tiếp từ link <https://www.minhngoc.net.vn/xo-so-mien-nam.html> bằng ngôn ngữ lập trình java.

Dữ liệu sẽ được lưu lại thành file .csv, được tải lên FPT server ( https://free02.123host.vn:2222/CMD\_FILE\_MANAGER?path=%2FDW ), khi sử dụng thì được lưu từ FPT server và chuyển đổi vào trong database staging, sau đó được chuyển qua database data\_warehouse và cuối cùng là chuyển qua database data\_mart, từ đây dữ liệu được lấy để hiển thị lên front-end.

**Feed Characteristics**.

Log file : DD/MM/YYYY/lotto.csv

Frequency : hằng ngày

Content : chứa dữ liệu sổ xố

Estimated Table size : ~ 956 KB/ngày

Pre-processing : không cần giải nén trước khi sử dụng

Estimated monthly size req : 956 \* 30 = 28680 KB/tháng

* **SOURCE 🡪 IMPORT**

~FTP : 🡪 DD-MM-YYYY/file\_name.csv

**Data/Process Controls**.

**Operations Interface.**

Các quy trình được sử dụng cho hai nhóm được các thành viên vận hành để quản lý các quy trình nguồn cấp dữ liệu. Các quy trình này thường sẽ xác định các mẫu cho quy trình xử lý, quy trình báo cáo sự cố và dữ liệu liên hệ được yêu cầu.

## Transport Mechanism

Đây là tiến trình chuyển đổi từ file .csv vào database staging.

## 3.3 Feed Characteristics

**File Format:**

Các file được chứa trong các thư mục có tên là ngày mà được extract với định dạng: DD-MM-YYYY. Trong các file, các trường được ngăn cách bởi dấu phẩy (,).

**Data Size:** Khoảng: ~956MB/Ngày

**Data Location:**

Có thể lấy file từ FTP và lưu về local từng máy để dễ dàng sử dụng.

**Data Frequency:** Hằng ngày

**Delivery Location:** Tiến trình này chuyển đổi từ file .csv vào database staging từ url (<https://www.minhngoc.net.vn/xo-so-mien-nam.html>)

STAGING schema: tổng hợp các cột của các bảng (các cột trùng đã được lọc)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
|  | id\_province | varchar | 20 |  | !=null | 68 |
|  | name\_province | varchar | 100 |  | != null | Kiên Giang |
|  | id\_prize | varchar | 10 |  | !=null | giai8 |
|  | name\_prize | varchar | 100 |  | != null | Giải tám |
|  | value\_prize | int | 11 |  | >0 | 100000 |
|  | number | int | 11 |  | >0 | 93 |
|  | created\_date | date |  | YYYY-MM-DD | != null | 2022-09-18 |
|  | updated\_date | date |  | YYYY-MM-DD |  | 9999-12-31 |
|  | status | varchar | 10 |  | != null | ER |
|  | natural\_key | varchar | 255 |  | != null | 1fn2gh3abc |
|  | id\_config | int | 11 |  | >0, !=null | 1 |
|  | server\_name | varchar | 100 |  | != null | 103.97.126.21 |
|  | url | text |  |  | != null | https://www.minhngoc.net.vn/xo-so-mien-nam.html |
|  | user\_name | varchar | 100 |  | != null | cineqoon |
|  | password | varchar | 100 |  | != null | 0819778801DEV |
|  | path\_upload | varchar | 150 |  | != null | C:\Users\admin\Desktop\FTP-Upload\ |
|  | path\_download | varchar | 150 |  | != null | C:\Users\admin\Desktop\FTP-Down\ |
|  | id\_contactor | int | 11 |  | >0, !=null | 1 |
|  | name | varchar | 100 |  | != null | Nguyen Ky Anh |
|  | email | varchar | 255 |  | != null | 19130011@st.hcmuaf.edu.vn |
|  | id\_log | int | 11 |  | >0, !=null | 1 |

**Naming Convention:** Quy ước đặt tên cho data feed:

DD-MM-YYYY/file\_name.csv

Giải thích:

DD-MM-YYYY là ngày trích xuất ra file.

file\_name là tên file

csv là định dạng theo csv

**Delivery Schedule:** Hằng ngày

## 3.4 Data/Process Controls

Data and process controls được đặt ở nơi để bảo đảm rằng the data feed chính xác, hoàn thành và đúng lúc.

Từ khóa để điều khiền là:

| **Control** | **Description** | **Implementation** |
| --- | --- | --- |
| ER | Hoàn thành việc trích xuất file | Kết thúc script 1 |
| EN | Chưa hoàn thành việc trích xuất file do dữ liệu còn thiếu | Trong script 1 |
| UPFI | Tải file lên FTP thành công | Trong script 2 |
| SAVE | Lưu file từ FTP thành công | Trong script 2 |
| SU | Hoàn thành việc chuyển dữ liệu từ file vào database staging | Kết thúc script 2 |
| FI | Chuyển dữ liệu từ database staging vào database data warehouse và xóa tất cả dữ liệu trong database staging | Kết thúc script 3 |
| END | Hoàn thành việc chuyển dữ liệu từ database data warehouse vào database data mart | Kết thúc script 4 |

## Checksum Definition (Not Applicable)

## 3.6 Control File Format

Dựa vào ngày hôm đó mà đặt tên thư mục chứa file trích xuất với định dạng DD-MM-YYYY

## 3.7 FTP Setup

Sau khi dữ liệu được trích xuất từ các nguồn trên các url thành các file .csv trong thư mục ngày hôm đó, hệ thống sẽ chuyển thư mục này lên FTP server. FPT server được thành lập với các trường bên dưới:

| **Name** | **Value** |
| --- | --- |
| Link | https://free02.123host.vn:2222/ |
| DNS | 103.97.126.21 |
| Tài khoản | cineqoon |
| Mật khẩu | 0819778801DEV |

## 3.8 Operations Interface

Đối với mỗi thành viên sẽ được giao nhiệm vụ và sẽ cung cấp tên, số điện thoại và địa chỉ mail để liên lạc.

# 4.0 SLA Negotiation (Not Applicable)

Vì dữ liệu được lấy thông qua truy cập trực tiếp từ hệ thống nội bộ, SLA không áp dụng cho nguồn cấp dữ liệu này.

# Appendix A – Attribute Specification

Các giá trị bên dưới được sử dụng để định nghĩa định dạng thuộc tính.

| **Attribute Format** | **Description** | **Example** |
| --- | --- | --- |
| String(x) | Chuỗi kí tự có độ dài x kí tự. Cho phép đỗ dài tối đa. | '123 Willow Road' là 1 chuỗi kí tự với độ dài là 15 kí tự. |
| Number(x,y) | Dữ liệu số với x kí tự là phần nguyên và y kí tự là phần thập phân. | 12.45 là Number(12,45) |
| Date | Dữ liệu ngày phụ thuộc vào luật của Mysql. Mặc định là YYYY-MM-DD. | 01-01-2022 là ngày 1 tháng 1 năm 2022 |
| Bit | Trường nhị phân. Thường được sử dụng làm cờ. Có thể là Number(1). | Giá trị chỉ 0 hoặc 1. |

Đối với định nghĩa của thuộc tính ngày, được mô tả với các giá trị bên dưới:

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| MM | Định dạng số tháng | 01 is January |
| MON | Định dạng chuỗi tháng | JANUARY |
| DD | Định dạng ngày | 23 is the 23rd day in the month |
| YYYY | Định dạng năm | 2002 |
| YY | Định dạng năm | 02 |
| HH | Định dạng giờ | 23 is 11pm |
| MI | Định dạng phút | 59 is minute 59 |
| SS | Định dạng giây | 05 is second 05 |
| / | Giá trị ngăn cách | 01/02/02 |
| - | Giá trị ngăn cách | 01-02-02 |
| B | Kí tự trống |  |

Đối với định nghĩa của thuộc tính chuỗi kí tự, được mô tả với các giá trị bên dưới:

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| NONE | No preference | Mr Jones home address |
| UPPER | Viết hoa | MR JONES HOME ADDRESS |
| LOWER | Viết thường | mr jones home address |
| INITCAP | Viết hoa chữ cái đầu mỗi từ | Mr Jones Home Address |

Đối với định nghĩa của kí tự, với các giá trị:

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| N | Số tờ 0 - 9 | NNNNN or N(5) |
| C | Mã alpha 0 – 9, A – Z | NN-CC-NN |
| A | Mã alpha A – Z, a – z | AAAA-NNN |
| X | Mã alpha viết hoa (A-Z) | ZZZ or Z(3) |
| x | Mã alpha viết thường (a-z) | Xxxxxxx or Xx(6) |
| . | Dấu chấm | XX.XX |
| - | Dấu gạch | AAA-NN |
| / | Kí tự chia | Có thể là 1 số giá trị |
| B | Kí tự trống | AAABNNN |

Các luật thay đổi định nghĩa trong thuộc tính của data feed là:

| **Edit Rule** | **Description** | **Example** |
| --- | --- | --- |
| created\_date | Thay đổi ngày cập nhật mỗi khi dòng đó có trường thay đổi | 2022-11-25 → 2022-11-26 |

# Appendix B – Attribute specification for each relational table

# Appendix C – Error Codes & Exception Handling

Error Codes: Hệ thống không cập nhật được dữ liệu liệu mới.

Exception Handling: Dựa vào các exception của MySQL để đưa ra ngoại lệ hoặc sử dụng kỹ thuật lập trình để bỏ qua bước cập nhật.

# Appendix D – Raw log file example