DRAFT

Data Feed Specification

XXX

**Rev. 0.3**

Revision History

| Date | Version | Author | Changes |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[1.0](#_heading=h.tyjcwt) Overview 4

[1.1](#_heading=h.3dy6vkm) Data Feed Specification 4

[1.2](#_heading=h.1t3h5sf) Contact Information 4

[1.3](#_heading=h.4d34og8) References 4

[2.0](#_heading=h.2s8eyo1) Transaction Definition 5

[2.1](#_heading=h.17dp8vu) Overview 5

[2.2](#_heading=h.lnxbz9) Data Feed Process Flow 5

[2.3](#_heading=h.35nkun2) Validation Guidelines 5

[2.4](#_heading=h.1ksv4uv) Attribute Specification 5

[2.4.1](#_heading=h.44sinio) Tomcat access logs (wap-site\_access\_log.YYYY-MM-DD\_appN-servername.txt) Attribute Specification 5

[2.5](#_heading=h.3j2qqm3) Code Values 6

[2.6](#_heading=h.1y810tw) Data Source Extraction and Exception Handling Procedures 6

[2.6.1](#_heading=h.4i7ojhp) Special Extract Procedure: Source 🡪 Import 6

[2.6.2](#_heading=h.2xcytpi) Error Code 6

[2.6.3](#_heading=h.1ci93xb) Exception Handling Procedure (Not Applicable) 6

[2.6.4](#_heading=h.3whwml4) Special Design Consideration (Not Applicable) 6

[3.0](#_heading=h.2bn6wsx) Feed Architecture 6

[3.1](#_heading=h.qsh70q) Overview 6

[3.2](#_heading=h.3as4poj) Transport Mechanism 7

[3.3](#_heading=h.1pxezwc) Feed Characteristics 7

[3.4](#_heading=h.2p2csry) Data/Process Controls 9

[3.5](#_heading=h.147n2zr) Checksum Definition (Not Applicable) 9

[3.6](#_heading=h.3o7alnk) Control File Format (Not Applicable) 9

[3.7](#_heading=h.23ckvvd) FTP Setup (Not Applicable) 9

[3.8](#_heading=h.ihv636) Operations Interface (Not Applicable) 9

[4.0](#_heading=h.32hioqz) SLA Negotiation (Not Applicable) 10

[Appendix A – Attribute Specification 11](#_heading=h.1hmsyys)

[Appendix B – Attribute specification for each relational table 12](#_heading=h.41mghml)

[Appendix C – Error Codes & Exception Handling 12](#_heading=h.2grqrue)

[Appendix D – Raw log file example 12](#_heading=h.vx1227)

# 1.0 Overview

## 1.1 Data Feed Specification

Tài liệu này cung cấp dữ liệu kết quả xổ số lấy từ các nguồn dữ liệu từ các trang:

<https://xosodaiphat.com/>

<https://xoso.me/>

<https://xsmn.me/>

Phương pháp lấy dữ liệu từ các script

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, checksums, 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** |
| --- | --- | --- | --- |
| Lương Hữu Luân | 19130131@st.hcmuaf.edu.vn | 0914067629 |  |
| Hồ Thạch Thảo | 19130210@st.hcmuaf.edu.vn | 0377134894 |  |
| Nguyễn Văn Hiếu | 19130074@st.hcmuaf.edu.vn | 0798631497 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## 1.3 References

| **Document** | **Version** | **Date** | **Author** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 2.0 Transaction Definition

## 2.1 Overview

## 2.2 Data Feed Process Flow

## 2.3 Validation Guidelines

For the required fields the two following validations will be done.

## 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 lottery logs (lottery-result\_YYYY-MM-DD\_source-data.csv) Attribute Specification

|  | **Field Name** | **Format** | **Mask** | **Edit Rules** | **Description** | **Example** |
| --- | --- | --- | --- | --- | --- | --- |
|  | Thứ | text |  |  | Thời gian phát hành(Thứ) | Thứ 4 |
|  | Ngày | text |  |  | Thời gian phát hành(Ngày) | 14/09/2022 |
|  | Tỉnh | text |  |  | Tỉnh phát hành vé số | TP.HCM |
|  | Giải | text |  |  | Giải loại gì?(ĐB,1,2,3,..8) | ĐB |
|  | Kết quả | text |  |  | Kết quả | 123456 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## 2.5 Code Values

Code values are abbreviations for the standard values used for an attribute. They are only used whenever an attribute has a known domain of values. It provides a more efficient technique for storing large sets of repeating values. They are used for everything from transaction types and statuses to large sets of values such as manufacturer names.

For any attribute defined as a vendor-supplied LOV (List of Values), the full list of values are provided for that list along with descriptions associated with that code value.

Maintenance of code values can be a problem – both parties must remain in sync. This is supported by operations procedures and feed versioning (any change to feed characteristics will increment a feed version attribute).

**List of Values (LOV)**

## 2.6 Data Source Extraction and Exception Handling Procedures

### 2.6.1 Special Extract Procedure: Source 🡪 Import

Select the listed fields from source application server logs and load into Staging. stg\_wap\_site\_access\_log is an archive for all records and is partitioned by day. In Staging, the data is aggregated for monthly report and kept for three months.

### 2.6.2 Error Code

Resolution 1: No pass through (see Appendix C)

### 2.6.3 Exception Handling Procedure (Not Applicable)

### 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 the Tomcat access logs file of the application server in ~home/file\_xfer to STAGING using PostGreSQL by call PERT scripts.

**Feed Characteristics**.

Log file : lottery-result\_YYYY-MM-DD\_source-data.csv

Frequency : Mỗi ngày lúc 4h30

Content : Kết quả xổ số của ngày đó

Estimated Table size : ~ 130MB/day

Pre-processing : extract zip file (extracted file size: ~1.4GB)

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

* **SOURCE 🡪 IMPORT**

~tomcat/tomcat/logs in DW1 : 🡪 Staging.stg\_wap\_site\_access\_log

**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.

## Transport Mechanism

Chương trình này được viết bằng java lấy dữ liệu từ trang … lưu thành … bỏ vào …

## 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: ~130MB/Day

**Data Location:**

~home/file\_xfer/logs – in DW1

**Data Frequency:** Daily

**Delivery Location:** This process is flat file to database data transfer from the application server log to

STAGING schema: file\_configuration table

| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
| --- | --- | --- | --- | --- | --- | --- |
|  | id\_config | integer |  |  |  | 1 |
|  | source | varchar(50) |  |  |  | https://xosodaiphat.com/ |
|  | source\_local | varchar(50) |  |  |  | output/data.csv |
|  | ftp\_server | inet |  |  |  | 74.82.68.33 |
|  | username | varchar(50) |  |  |  | user1 |
|  | password | varchar(50) |  |  |  | 123456 |

STAGING schema: file\_logs table

| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
| --- | --- | --- | --- | --- | --- | --- |
|  | id\_file\_logs | integer |  |  |  | 1 |
|  | id\_config | integer |  |  |  | https://xosodaiphat.com/ |
|  | file\_name | varchar(50) |  |  |  | output/data.csv |
|  | date | datetime |  |  |  | 10-09-2022 00:00:00 |
|  | state | varchar(10) |  |  |  | finish |
|  | id\_contact | integer |  |  |  | 1 |

STAGING schema: contact table

| No. | **Field Name** | **Format** | **Length** | **Mask** | **Edit Rules** | **Description(sample, value )** |
| --- | --- | --- | --- | --- | --- | --- |
|  | id\_contact | integer |  |  |  | 1 |
|  | full\_name | varchar(50) |  |  |  | Nguyễn Văn A |
|  | phone\_number | varchar(11) |  |  |  | 0123456789 |
|  | email | varchar(50) |  |  |  | nguyenvana@gmail.com |

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

wap-site\_access\_log.YYYY-MM-DD\_app{N}-username.txt.gz

Where:

YYYY-MM-DD is the date of the file.

app{N} is application name: app3, app4, app6, app8, app9, app10, app11

servername: tomcat (Tomcat server)

**Delivery Schedule:** Daily / Month-End 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** |
| --- | --- | --- |
| ER | Extract ready | Work extract file |
| T1 | Transform data step 1 | Work transform data step 1 |
| T2 | Transform data step 2 | Work transform data step 2 |
| T3 | Transform data step 3 | Work transform data step 3 |
| T4 | Transform data step 4 | Work transform data step 4 |
| T5 | Transform data step 5 | Work transform data step 5 |
| LR | Load ready | Work Load data |
| SU | Status success | Report success |
| EF | Extract fail | Report extract fail |

## 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)

Since data is acquired via direct access from internal system, SLA is *not applicable* for this data feed.

# 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 |
| Bit | Binary field. Normally used as a switch to indicate status. May also be specified as Number(1). | Only values 0 or 1. |

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 |
| MI | Minute in hour | 59 is minute 59 |
| SS | Second in minute | 05 is second 05 |
| / | Value separator | 01/02/02 |
| - | Value separator | 01-02-02 |
| B | Single blank/space character |  |

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

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| NONE | No preference | Mr Jones home address |
| UPPER | All upper case | MR JONES HOME ADDRESS |
| LOWER | All lower case | mr jones home address |
| INITCAP | All words are capitalized | Mr Jones Home Address |

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

| **Mask Character** | **Description** | **Example** |
| --- | --- | --- |
| N | Numeric digit (0-9) | NNNNN or N(5) |
| C | Alphanumeric digit (A-Z, 0-9) | NN-CC-NN |
| A | Alphabetic digit (A-Z, a-z) | AAAA-NNN |
| X | Uppercase alphabetic digit (A-Z) | ZZZ or Z(3) |
| x | Lowercase alphabetic digit (a-z) | Xxxxxxx or Xx(6) |
| . | Period | XX.XX |
| - | Hyphen | AAA-NN |
| / | Field separator | Can be any value |
| B | Single blank/space character | AAABNNN |

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