Enterprise Analytics Solution for Everyone (EASE)

EASE Data Set Standards and SFTP Setup Parameters

Document ID: SD-110

Version 9

December 29, 2022

GEORGIA DEPARTMENT
OF COMMUNITY HEALTH

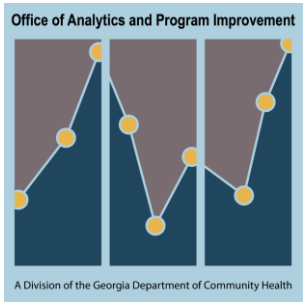


Table of Contents

Table of Contents....2
Revision Record....2
EASE Data Set Standards and Guidelines....3
Overview....3
Data Set Standards and Guidelines....4
EASE SFTP Setup....8
Instructions....8

Revision Record

Revision	Date	By	Description
Draft 1	November 18, 2021	Daniel Sklar	Initial draft
Version 1	January 26, 2022	Daniel Sklar	Updates based on feedback
Version 2	February 18, 2022	Daniel Sklar	Updated background, acceptable file formats; added SFTP setup instructions
Version 3	March 4, 2022	Daniel Sklar	Changed to use the term “Data Partner”; added information on public/private pair keys
Version 4	March 13, 2022	Daniel Sklar	Clarified the number of SSH keys required
Version 5	March 23, 2022	Daniel Sklar	Updates based on feedback for SHBP data acquisition
Version 6	April 28, 2022	Daniel Sklar	Clarified file extension to use, added file naming convention for history files, clarified control totals file content
Version 7	May 4, 2022	Daniel Sklar	Clarified to mention Gzip has a .gz extension, update of file naming conventions for sorting purposes
Version 8	June 15, 2022	Daniel Sklar	Added custom endpoint host names to the instructions on setting up SFTP
Version 9	December 29, 2022	Daniel Sklar	Removed S# from file naming convention except when a file is submitted as multiple files

EASE Data Set Standards and Guidelines

Overview

Background

As the Georgia Department of Community Health (DCH) continually looks for ways to cut costs, increase contracting efficiencies, and enhance quality controls, we have made a strategic decision to leverage statewide contracts through the Georgia Department of Administrative Services to support deployment of our data warehouse project, Enterprise Analytics Solution for Everyone (EASE). EASE is creating a new technology and data warehousing ecosystem that uses best-in-class, cloud-based services and automation to meet our growing reporting and analytics demands with speed and agility. It also prioritizes cyber security and data preservation due to the highly confidential nature of information the Department manages.

The EASE project began on November 1, 2021 and is requesting data sets from Data Partners (organizations that provide data sets to the EASE data warehouse, including DCH vendors, partners, state agencies, and others) to load into the EASE solution.

Purpose

This document contains the standards and guidelines DCH is requesting with all Data Partners for the EASE solution.

These standards and guidelines provide consistency in the data sets and the processes for both the Data Partners and DCH. The document will also help DCH be consistent in its communications with all Data Partners.

Important

While the EASE solution is being developed, Data Partners who are currently submitting data sets to other DCH vendors (for example, to IBM Watson Health for the legacy data warehouse) should continue to provide those files in the same way and format as currently submitted until there is an official communication from DCH to no longer submit those files.

The EASE project is separate from the existing legacy systems.

Interface Control Document (ICD)

DCH will create and maintain an Information Control Document (ICD) for each data set (or group of related data sets). DCH will work with each Data Partner to create the ICD and ensure understanding and agreement.

Any exceptions to these standards and guidelines for a data set will require exception approval by DCH and will be documented in the ICD.

Data Set Standards and Guidelines

Data sets and control files

Data Partners should provide data in a data set, along with an associated control file.

A data set consists of one type of data record. Data records of a different type (e.g., different data fields) should be provided in a separate data set.

For each data set, an associated control file should also be provided. The control file will contain one or more values that will help DCH confirm full receipt of the data set. For example, the control file may provide a record count for the associated data set.

File format

Data sets and control files can be submitted in one of the following formats.

- CSV (comma separated values) format. The RFC 4180 standard in strict form should be followed. See https://www.ietf.org/rfc/rfc4180.txt for specifics. The UTF-8 character set should be used. Values containing commas should be in quotes (e.g., "Red, White, and Blue"). These files should have a .csv file extension.
- Pipe delimited (|) format, if the pipe character is not used elsewhere in the file. These files should have a .txt or .psv file extension.
- Fixed-width format. These files should have a .txt file extension.
- JSON. These files should have a .json file extension.
- XML. These files should have a .xml file extension.

Data set layout

The data set file should have two types of records if in CSV and pipe delimited formats:

- Header Record – The header record is the first record of the file containing column heading information; it will occur only once.
- Detail Records – The remainder of the file should be detail records with data corresponding to the fields listed in the header record

There should be no Trailer record or any additional data in the data set.

Data that does not correspond to the Header Record (fields and in field order) should be in a separate data set.

Reference files

Reference files are data sets and follow the same standards as any other data set. These data sets contain code sets and their descriptions, and they may include effective date ranges and other descriptive fields. Each code and description within a code set should be unique.

A separate reference file (data set) should be provided for each code set.

File naming conventions

Files should follow standard file naming conventions.

Type	File naming convention	Notes
Data set file	*VendorAbbr_Category_ContentDesc_Version_Date_Sequence.extension* Example if one file for submission: ACME_PRVDR_V1_20211105.csv Example if file split into multiple files for submission: ACME_PRVDR_V1_20211105_S1.csv ACME_PRVDR_V1_20211105_S2.csv If a corrected file is sent: *VendorAbbr_ContentDesc_Version_Date_Sequence_Correction.extension* Example for the first corrected file sent: ACME_PRVDR_V1_20211105_C1.csv or ACME_PRVDR_V1_20211105_S1_C1.csv	• *VendorAbbr* = DCH supplied • *Category* = e.g., PRVDR, CLMMDCL, CLMPHRMCY • *Version* = V*n*, where there are different versions of the file layout over time; initial version V1 • *Date* = Through date or as of date of the data content, in YYYYMMDD format • *Sequence* = skipped if only one file is submitted, S*n* (starting with *n*=1 if data set is submitted in multiple files) • *Extension* = .csv, .txt, .json, or .xml, depending on the file format (see File Format earlier in document) • *Correction* = C*n* (starting with *n*=1 for the first correction)
Reference data set file	*VendorAbbr_Category*_RFRNC_*ContentDesc_Version_Date_Sequence.extension* Example if one file for submission: ACME_PRVDR_RFRNC_PRVDRTYPECD_V1_20211105.csv Example if file split into multiple files for submission: ACME_PRVDR_RFRNC_PRVDRTYPECD_V1_20211105_S1.csv ACME_PRVDR_RFRNC_PRVDRTYPECD_V1_20211105_S2.csv	• *VendorAbbr* = DCH supplied • *Category* = e.g., PRVDR, CLMMDCL, CLMPHRMCY • RFRNC (indicates file contains reference data) • *ContentDesc* = Description of data contained in the file • *Version* = V*n*, where there are different versions of the file layout over time • *Date* = Through date or as of date of the data content, in YYYYMMDD format • *Sequence* = skipped if only one file is submitted, S*n* (starting with *n*=1 if data set is submitted in multiple files) • *Extension* = .csv, .txt, .json, or .xml, depending on the file format (see File Format earlier in document)

Type	File naming convention	Notes
History data set file	*VendorAbbr_Category*_HSTRY_*Version_DateStart_DateEnd_Sequence.extension* Examples: ACME_PRVDR_HSTRY_H1_20110101_20151231.csv ACME_PRVDR_HSTRY_H2_20160101_20190831.csv ACME_PRVDR_HSTRY_H3-V1_20190901_20211231.csv In the third example, the historical layout version H3 is the same as V1 for regular data set layout	• *VendorAbbr* = DCH supplied • *ContentDesc* = Description of data contained in the history file • *Version* = version of historical layout H*n*, where there may be different versions of the history file layout over time; initial version H1 If a historical version layout is the same as one of the regular feed layouts, note this by appending -V*n* , where *n* is the number of the same layout (example: H1-V1) • *DateStart* = Start date of the data content, in YYYYMMDD format • *DateEnd* = Through date of the data content, in YYYYMMDD format • *Sequence* = skipped if only one file is submitted, S*n* (starting with *n*=1 if data set is submitted in multiple files) • *Extension* = .csv, .txt, .json, or .xml, depending on the file format (see File Format earlier in document)
Control file	*DataSetFileName*_CTRL.csv Example: ACME_PRVDR_V1_20211105_CTRL.csv Example if data file split into multiple files for submission with multiple control files: ACME_PRVDR_V1_20211105_S1_CTRL.csv	• *DataSetFileName* = File name (without the extension) of the associated data set; see the file naming conventions above • CTRL (indicates file is a control file)

File compression

Files should be submitted unarchived (no ZIP file permitted) and as separate files.

Files may be provided compressed in Gzip format (.gz extension) if desired.

File destination

DCH will provide the destination location and work with the Data Partner to set up connection to the DCH AWS SFTP server. DCH has three separate STFP environments available for use:

- Development environment: ease-dev.sftp.gadch.com
- QA environment: ease-qa.sftp.gadch.com
- Production environment: ease.sftp.gadch.com

Data set content

DCH will specify if a full replacement file, historical data, snapshots, or a file containing only records with changes (e.g., records with changes in data values, new records) should be provided in the data set.

Data set field layout

Field/element definitions should be provided by the Data Partner and will be documented in the ICD. The data field names, field descriptions, data types, and maximum lengths should be provided, and for fixed-width formatted files, start and end positions should also be provided. Additional attributes of each field may also be requested.

Cleansing of data in data set

Data should be provided as clean as possible by the Data Partner prior to submitting the files.

Data should be provided with leading and trailing spaces trimmed.

Data should adhere to industry standards (e.g., Office of the National Coordinator for Health Information Technology (ONC), Centers for Medicare & Medicaid Services (CMS)) as much as possible.

Contact information

Data Partners should provide and maintain the names and contact information for the individuals who are able to address any questions with the files. This should include both individuals:

- Responsible for the transmission of the files
- Knowledgeable about the content.

The contact information will be used to obtain and assign security credentials to the destination location and folders to deposit the files. The information will also be used to contact the appropriate individual if there is a problem with or question related to the files.

Snowflake option

The EASE solution utilizes Snowflake in Amazon Web Services (AWS). If the data set is available in Snowflake, please discuss with DCH as a possible option to sharing data.

Control totals file

The control totals file should be provided in CSV format. A control totals file should contain one Header Record and one Detail Record.

```
Data_File_Name,File_Date,Environment,Record_Count
DatasetFileName,mm/dd/yyyy,EEEE,nnnn
```

A control totals file has the following content in the Detail Record, where:

- *DatasetFileName* is the name of the associated data set (e.g., ACME_PRVDR_V1_20211105.CSV)
- *mm/dd/yyyy* is the date the dataset was created
- *EEEE* specifies the environment the data set is intended for:
  - DEV
  - QA
  - PROD
- *nnnn* is a number related to the associated data set (in this case, the record count of the associated data set). The record count should not include the header record in the count.

Additional control total values may also be added in addition to record count, depending on the data content of the data set. These values will be added at the end of the Detail Record, and a field name will be added to the Header Record.

EASE SFTP Setup

Instructions

Introduction

EASE is using the AWS Transfer Family service to provide our Data Partners with a secure process to transfer files to DCH using Secure File Transport Protocol (SFTP).

Service managed authentication

We ask our Data Partners to generate their own SSH keys and only share the public key with DCH to complete their account configuration.

Upon receipt of the public keys, DCH will forward the following for set up and connection validation:

- User name
- A test file to upload for connection validation.

Embedded here is a test file that can be used:

DCH requires different SFTP public keys for authentication for production and non-production SFTP environments. Therefore, we request one SFTP public key for authenticating to our SFTP production server and a different SFTP public key for our SFTP development and/or QA server. Data Partners may provide additional keys beyond these requirements if it is a requirement of the Data Partner.

For more information on generating SSH keys, please refer to these instructions:
https://docs.aws.amazon.com/transfer/latest/userguide/key-management.html#sshkeygen

Setup

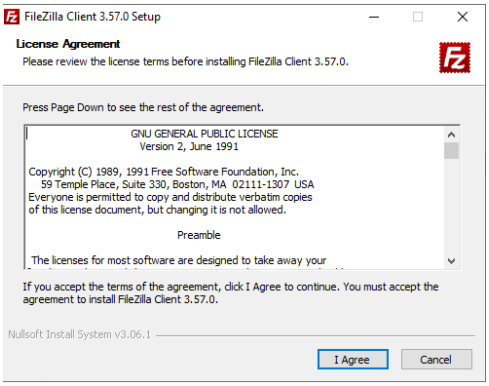
Follow the steps below to perform the setup.

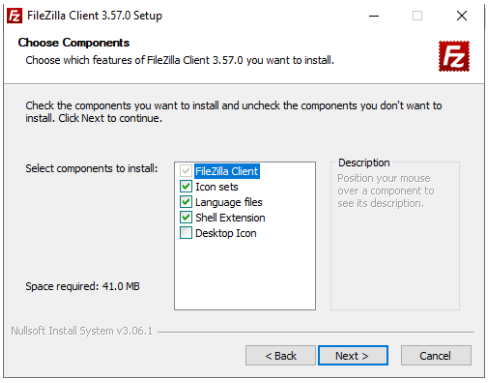
The steps use FileZilla as an example. You can use your own SFTP client tools. For Data Partners with an established set of tools and processes, jump to Step 3, which contains the Site Configuration information.

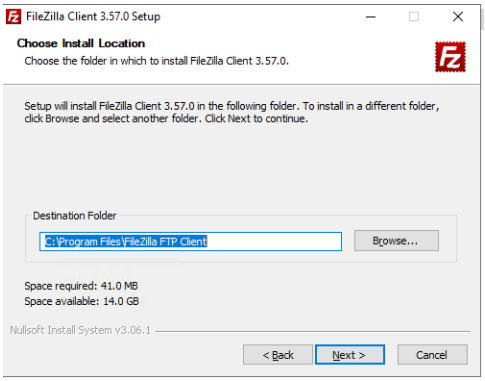
Step 1 – Download and Install SFTP Client Tools

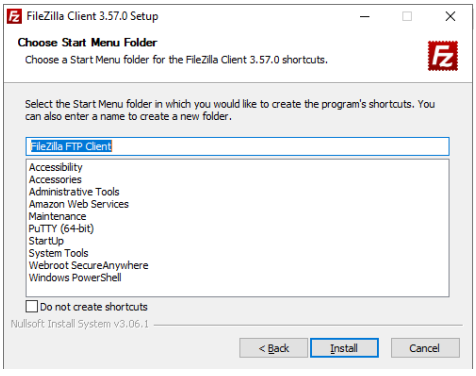
Download and install latest version of FileZilla Client Tool from
https://filezilla-project.org/download.php?platform=win64

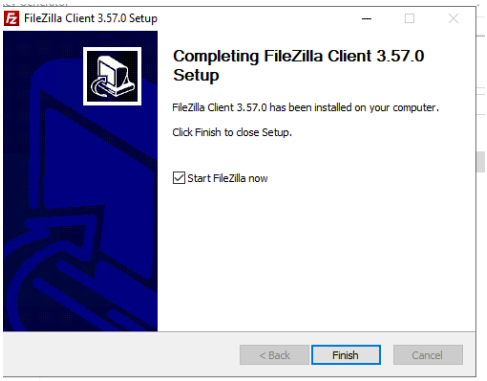
Install the downloaded file







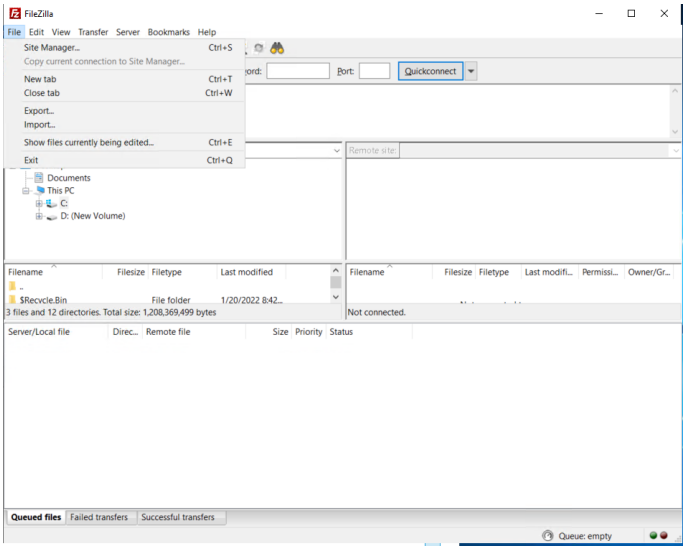




Other SFTP Client Tools can be utilized as well

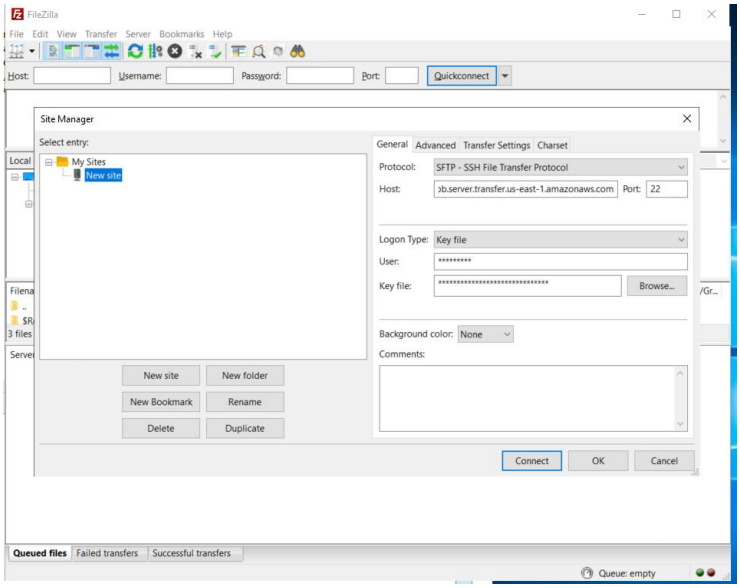
Step 2 – Configure SFTP to access Georgia DCH Environment

Open FileZilla Client and Go to **File → Site Manager**



Step 3 – Site Configuration

Select 'New Site' and provide below protocols



Protocol: SFTP – SSH File Transfer Protocol
Host:

Environment	SFTP Endpoint	Custom Endpoint
Development	s-44ef74fb13b14885a.server.transfer.us-east-1.amazonaws.com	ease-dev.sftp.gadch.com
QA	s-5e6df82e517b412bb.server.transfer.us-east-1.amazonaws.com	ease-qa.sftp.gadch.com
Production	s-b3507fd3fe7245258.server.transfer.us-east-1.amazonaws.com	ease.sftp.gadch.com

Port: 22
Logon Type: Key File
User: *Username provided
Key File: Browser and select the private key

Click ‘Connect’