Pitney Bowes Software

Fixed-Length File Scanner

User Guide

Spectrum™ Technology Platform

Data Discovery Module

Version 8

April 2013

Global IDs Product Suite® Fixed-Length File Scanner Manual

Copyright © 2001 – 2013 Global IDs Inc.

All rights reserved. Printed in USA.

This software and documentation contain proprietary information of Global IDs Inc., they are provided under a license agreement containing restrictions on use and disclosure and is also protected by copyright law. Reverse engineering of the software is prohibited. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopy, recording, or otherwise) without prior consent of Global IDs Inc.

Use, duplication, or disclosure of the Software by the U.S. Government is subject to the restrictions set forth in the applicable software license agreement as provided in DFARS 227.7202-1 (a) and 227.7702-3 (a) (1995), DFARS 252.227-7013(c)(1)(ii) (OCT 1988), FAR 12.212(a) (1995), FAR 52.227-19, or FAR 52.227-14(ALT III), as applicable.

This information in this document is subject to change without notice. If you find any problems in the documentation, please report them to us in writing. Global IDs Inc. does not warrant that this documentation is error free.

Global IDs Product Suite is the registered trademark of the Global IDs Inc. in the United States and in jurisdictions throughout the world. All other company and product names may be trade names or trademarks of their respective owners.

Global IDs products contain GUI Components from JIDE™ Software, yFiles from yWorks™, JGo: GoDiagram™ from Northwoods Software Corporation, StelsEngine from J-Stels™ Software, EasyCharts™ from ObjectPlanet, Install4j from ej-technologies GmbH, Nmap™ - Free Security Scanner For Network Exploration & Security Audits(originally written by Gordon Lyon). Java™ is a registered trademark of Sun Microsystems and Apache Derby™ from Apache Software Foundation. All rights reserved.

DISCLAIMER: Global IDs Inc. provides this documentation ‘as is’ without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of non-infringement , merchantability, or use for a particular purpose. The information provided in this documentation may include many technical inaccuracies or typographical errors. Global IDs could make improvements and/or changes in the products described in this documentation at many time without notice.

Contents

List of Figures 4
About Fixed-Length Files and Fixed-Length File Scanner 6
Scanning Fixed-Length files 6
Working with scanned Fixed-Length files 7
Discovering Fixed-Length file metadata 7
Loading Fixed-Length files 9
Profiling columns 11
Finding Patterns 13
Profiling Domains 15

List of Figures

Figure 1: Scanning fixed-length files.....6
Figure 2: Viewing fixed-length files.....7
Figure 3: Discovering metadata.....8
Figure 4: Discovering metadata wizard.....8
Figure 5: Viewing fixed-length files.....9
Figure 6: Loading fixed-length files - step 2.....10
Figure 7: Loading fixed-length files - step 3.....11
Figure 8: Profiling columns - step 2.....12
Figure 9: Profiling columns - step 3.....13
Figure 10: Finding patterns.....14
Figure 11: Profiling domains - step 2.....16
Figure 12: Profiling domains - step 3.....16

Fixed-Length File Scanner

What it includes:

- About Fixed-Length Files and Fixed-Length File Scanner
- Scanning Fixed-Length files
- Working with scanned Fixed-Length files
    - Discovering Fixed-Length file metadata
    - Loading Fixed-Length file
    - Profiling Fixed-Length file
    - Pattern Mining Fixed-Length file

About Fixed-Length Files and Fixed-Length File Scanner

This kind of file contains data that have fixed length. Fixed - Length File Scanner application enables you to scan for all fixed-length files in your data landscape and load their contents as tables in the database.

Scanning Fixed-Length files

This topic teaches you to find out fixed-length files located in any system in your network.
To scan Fixed-Length Files:

- Right click a system from the left hierarchy pane and select **Scan Fixed Length Files** option. **Scan Fixed Length Files** wizard appears
- On Step 1 of the wizard:
  - Select a source
  - Specify a directory you would like to scan or select appropriate checkbox to scan the entire system.
  - Select the appropriate check box to:
    - scan compressed files
    - scan fixed-length files in emails

Figure 1: Scanning fixed-length files

Click **Next**.

- On Step 2 of the wizard:
  - Select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing scanned fixed-length files

Selecting the system from the hierarchy pane displays all scanned fixed-length files in the system. From the list view you can find the file name, file location, file type, time when each file was last modified, file size, and information on whether metadata has been discovered.

File Name	File Path	File Type	Last Modified	File Size (in KB)	Metadata Discov...	Load Started At	Load Ended At	Loaded row count	# Rejected Rows
FixedLength1.txt	D:\Fixed Length...	Flat File	2012 03 16, 09:...	99.03	✓	2012 03 16, 05:...	2012 03 16, 05:...	1004	0

Figure 2: Viewing fixed-length files

Working with scanned Fixed-Length files

You can discover fixed-length file metadata and load data in the scanned files in a database.

Discovering Fixed-Length file metadata

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Fixed Length File
- Discover Fixed Length File Metadata

Fixed length file metadata can be discovered in two ways:

1. Select individual file from the list view to open **Discover Metadata** dialog box.
2. Right click on the desired system and select **Discover Fixed Length File Metadata** option.

To discover metadata of fixed length files using the first process:

- Select a system and ensure that fixed length files have been scanned. The scanned files can be viewed in the list view.

- Select a file from the list view. If metadata of the file is not discovered, **Discover Metadata** dialog box pops up.
- Ensure that **Yes** option is selected in the **Discover File Metadata** field.
- Select the appropriate option in the **Consider first row as column header** field.

Figure 3: Discovering metadata

- Click **OK** to discover the metadata.

To discover metadata of fixed length files using the second process:

- Right click the desired system from the left hierarchy pane and select **Discover Fixed Length File Metadata** option. **Discover Fixed Length File Metadata** wizard appears.
- On Step 1 of the wizard:
  - Select the file(s) to discover their metadata.
  - Check the **Consider first row as column header** option if appropriate.

Figure 4: Discovering metadata wizard

  - Click **Next**

- On Step 2 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**

- You can also add time restrictions by clicking on the **Add Time Restrictions** link.

Click **Finish**.

Viewing fixed-length file metadata

You can view metadata of the selected file under **Metadata** tab. Select a file from the list view and the metadata will be displayed in the **Metadata** tab below the list view.

File Name	File Path	File Type	Last Modified	File Size (in KB)	Metadata Discov...	Load Started At	
FixedLength1.txt	C:\Documents a...	Flat File	2008 09 02, 11:...	99.03	☐	N/A	N/
FixedLength2.txt	C:\Documents a...	Flat File	2008 09 02, 11:...	99.03	☑	2012 06 26, 09:...	20

Metadata | Sample Data

Column Name	Data Type	Column Size	Decimal Digits	Start Index	End Index	Date Format	Encoding	Description
Column 1	VARCHAR	16	0	1	16		ASCII	
Column 2	VARCHAR	79	0	17	95		ASCII	
Column 3	NUMBER	5	0	96	100		ASCII	

Figure 5: Viewing fixed-length files

Loading Fixed-Length files

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Fixed Length File
- Discover Fixed Length File Metadata
- Load Fixed Length Files

Fixed length files can be loaded to a database for future reference.

To load Fixed-Length files:

- Right click on a system from the left hierarchy pane and select **Load Fixed Length File** option. Load Fixed Length file wizard appears.
- On Step 1 of the wizard:
  - Choose a storage location. The following storage locations are available - Staging Area**,** Operational Data Store, Reference Data Hub, Data Warehouse, or specify a database of your choice (which has already been scanned).

Click **Next.**

- On Step 2 of the wizard:
  - Select fixed-length files you want to load.
  - Select **Consider all data types as text** checkbox to load the entire content of the file as text.

Figure 6: Loading fixed-length files - step 2

Click **Next**

- On Step 3 of the wizard you can:
  - Change the table name generated if required by clicking on the table name.

Figure 7: Loading fixed-length files - step 3

Click **Next**

- On Step 4 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link.

Click **Finish**.

Viewing loaded fixed-length files

Selecting the desired system from the hierarchy pane displays load start time, load end time, loaded row count and number of rejected rows. The loaded files are added to the left hierarchy pane.

File Name	File Path	File Type	Last Modified	File Size (in KB)	Metadata Discov...	Load Started At	Load Ended At	Loaded row count	# Rejected Rows
FixedLength1.txt	D:\Fixed Length...	Flat File	2012 03 16, 09:...	99.03	✓	2012 03 16, 05:...	2012 03 16, 05:...	1004	0

Profiling columns

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Fixed Length File
- Discover Fixed Length File Metadata
- Load Fixed Length Files
- Profile Columns

After fixed-length files are loaded in a database as tables, the columns can be profiled. This topic teaches you to profile columns of loaded fixed-length files.

To profile columns:

- Right click the desired file from the left hierarchy pane and select **Profile Columns** option. **Profile Columns** wizard appears
- On Step 1 of the wizard:
  - Choose the columns you would like to profile.

Click **Next**

- On Step 2 of the wizard allows you to:
  - Change default sample size/set
  - Select appropriate checkbox to find
    - value frequencies (by default 20 most occurring and 20 least occurring values are found out)
    - statistical values like standard deviation, median etc.
    - length clusters
  - Check re-profile Profiled Columns to profile columns that have already been profiled.

Figure 8: Profiling columns - step 2

Click **Next**

- On Step 3 of the wizard:
  - You need to specify parameters for pattern mining if you want to find out regular expression patterns. Choose **sample set** and **maximum length clusters** (by default 8). Maximum length clusters option means that if a column has more than 8 length clusters, the column will not be eligible for pattern mining. If you do not want to perform pattern mining, uncheck the checkbox next to **Find Patterns**.

Figure 9: Profiling columns - step 3

Click **Next**

- On Step 4 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**

Viewing profile information

Column profiling provides you information on the data composition of columns, their value range, length clusters, uniqueness of data, frequency distribution, character distribution, inferred data type and other details that are crucial for in-depth data analysis. It also helps in detecting data type outliers, frequency outliers, length outliers, pattern outliers and value outliers. These results get displayed in the **Profile** tab on the right (display pane) when the loaded file is selected from the hierarchy pane.

Finding Patterns

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Fixed Length File
- Discover Fixed Length File Metadata
- Load Fixed Length Files
- Profile Columns
- Find Patterns

This topic teaches you to pattern mine fixed-length files that have been loaded into a database.

To find patterns:

- Right click the desired fixed-length file from the left hierarchy pane and select **Find Patterns** option. **Find Patterns** wizard appears.
- On Step 1 of the wizard:
  - Select columns to find their patterns.

Click **Next**.

- On Step 2 of the wizard you can:
  - change default sample size
  - change maximum length cluster
  - select appropriate checkbox to
    - find patterns for ID and natural key columns only
    - find patterns for profiled columns only
    - rerun pattern mining for columns whose patterns are known

Figure 10: Finding patterns

Click **Next**

- On Step 3 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**

Viewing patterns in fixed-length files

Find Patterns finds out the pattern in which data is composed in each column of a given table. For Example, data in Zip Code column of a table is in 5 digit pattern (pattern may be expressed as 99999 or regular expression [0-9]{5}). The results get displayed in the **Patterns** tab on the right (display pane) when the loaded file is selected from the hierarchy pane.

> **Note**: The software provides two alternatives to pattern mining. You have the option of finding out patterns for columns during statistical profiling (column profiling) or you can find out patterns separately for the columns of your choice. In both cases, the results get displayed in the **Patterns** tab on the right.

Profiling Domains

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Fixed Length File
- Discover Fixed Length File Metadata
- Load Fixed Length Files
- Profile Domains

This topic teaches you to **profile domains** in fixed-length files.

To profile domains:

- Right click the desired file from the left hierarchy pane and select **Profile Domains** option. **Profile Domains** wizard appears.
- On Step 1 of the wizard:
  - select columns to profile their domains
  - Click **Next**
- On Step 2 of the wizard you can:
  - choose **domains** (from a list of pre-defined domains bundled with the software) that you would like to include in this process. However, selecting all the pre-defined domains is recommended.

Figure 11: Profiling domains - step 2

- On Step 3 of the wizard:
  - change the default sample set
  - change the match threshold
  - choose from the options – whether to profile domains for identifying columns or non-identifying columns.
  - choose **Rerun Domain profiling for already domain-profiled columns** option to re-profile columns on which domain profiling has already been run.
  - choose **Find more accurate match percentage (takes longer)** option to find multiple domain matches for a column.

Figure 12: Profiling domains - step 3

- On Step 4 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time

- Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
- Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
- You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**

Viewing domains in fixed-length files

You can view the domain having the highest match to a column from the **Domains** tab on the right. The **Domains** tab also displays domain match details and graphical representation of matching domains for each column matched.

Index

D

Discovering Fixed-Length file metadata.....7

F

Finding Patterns.....13

L

Loading Fixed-Length files.....9

P

Profiling columns.....11

Profiling Domains.....15

S

Scanning Fixed-Length files 6

V

Viewing domains in fixed-length files 17

Viewing fixed-length file metadata 9

Viewing loaded fixed-length files 11

Viewing patterns in fixed-length files 15

Viewing profile information..... 13

Viewing scanned fixed-length files 7