Pitney Bowes Software

Delimited File Scanner

User Guide

Spectrum™ Technology Platform

Data Discovery Module

Version 8

April 2013

Global IDs Product Suite® Delimited 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 Delimited Files and Delimited File Scanner 6
Scanning Delimited files 6
Working with scanned Delimited files 7
Discovering Delimited file metadata 7
Loading Delimited files 9
Profiling columns 11
Finding Patterns 14
Profiling Domains 16
Index 18

List of Figures

Figure 1: Scanning delimited files 6
Figure 2: Viewing delimited files 7
Figure 3: Discovering metadata 8
Figure 4: Discovering metadata - step 1 8
Figure 5: Viewing metadata 9
Figure 6: Loading delimited file – step 2 10
Figure 7: Loading delimited file - step 3 10
Figure 8: Viewing loaded files 11
Figure 9: Profiling columns - step 2 12
Figure 10: Profiling columns - step 3 12
Figure 11: Viewing profile information 13
Figure 12: Finding patterns 14
Figure 13: Viewing patterns 15
Figure 14: Profiling domains - step 2 16
Figure 15: Profiling domains - step 3 17

Delimited File Scanner

What it includes:

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

About Delimited Files and Delimited File Scanner

This kind of file contains delimiters in their data. Delimiters supported by the Data Discovery Module are pipe, comma and tab.

Delimited File Scanner application enables you to scan all the delimited files in your data landscape and load their content as tables in the database.

Scanning Delimited files

This topic teaches you to find out delimited files located in any system in your network.

To scan Delimited Files:

- Right click the desired system from the left hierarchy pane and select **Scan Delimited Files** option. **Scan Delimited Files** wizard appears.
- On Step 1 of the wizard:
  - Select a **source**
  - Specify a directory or select appropriate checkbox to scan the entire system
  - Choose Delimited file types you want to scan. By default it finds all file types.
  - Select the appropriate check box to:
    - scan compressed files
    - scan delimited files in Emails

Figure 1: Scanning delimited 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 delimited files

By selecting the desired system from the hierarchy pane you can view all delimited files located in the system along with information on file path, file type, last modified time, 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
new.csv	C:\Documents a...	Comma Delimited	2012 01 19, 08:...	13.59	✓	2012 01 19, 03:...	2012 01 19, 03:...	109	0

Figure 2: Viewing delimited files

Working with scanned Delimited files

You can discover delimited file metadata and load data in the scanned files in a database.

Discovering Delimited 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 Delimited Files
- Discover Delimited File Metadata

Delimited 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 Delimited File Metadata** option

To discover metadata of delimited files using the first process:

- Select a system and ensure that delimited 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 delimited files using the second process:

- Right click the desired system from the left hierarchy pane and select **Discover File Metadata** option. **Discover 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 - step 1

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 delimited file metadata

Select the desired system from hierarchy pane and corresponding file from the list view on the right to find metadata information under **Metadata** tab. To edit the metadata of each column select columns in **Metadata** tab and click on **Edit**.

Figure 5: Viewing metadata

Loading Delimited 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 Delimited Files
- Discover Delimited File Metadata
- Load Delimited File

Delimited files can be loaded to a database for future reference.

To load Delimited files:

- Right click on the desired system from the left hierarchy pane and select **Load Delimited File** option. **Load Delimited 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 a database of your choice (which has already been scanned).

Click **Next**.

- On Step 2 of the wizard:

- Select delimited 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 delimited file – 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 delimited file - 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 information

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
new.csv	C:\Documents a...	Comma Delimited	2012 01 19, 08:...	13.59	✓	2012 01 19, 03:...	2012 01 19, 03:...	109	

Figure 8: Viewing loaded files

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 Delimited Files
- Discover Delimited File Metadata
- Load Delimited File
- Profile Columns

After delimited files are loaded in a database as tables, the columns can be profiled. This topic teaches you to profile columns of loaded delimited 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 9: 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 the sample set and the 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 10: 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

After profiling task is complete you can view profiling information by selecting the delimited file from hierarchy pane and clicking on **Profile** tab.

Figure 11: Viewing profile information

The Display pane under Profile tab is segmented into two parts:

1. Table view
2. Graphical View

The table view displays column name, documented data type, inferred data type, primary key details, uniqueness of data in each column and time when columns were profiled for the selected Excel sheet.

The graphical view displays coverage map, data composition, values, length, frequencies and value clusters. While Coverage Map displays the distribution of different types of data values in each profiled column, Data Composition, Value, Length, Frequencies, Numeric Clusters and Character Distribution tabs display the suggested information on columns selected from the list view.

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 Delimited Files
- Discover Delimited File Metadata
- Load Delimited File
- Profile Columns
- Find Patterns

This topic teaches you to pattern mine loaded delimited files.

To find patterns:

- Right click the desired delimited file from the left hierarchy pane and select **Find Patterns.** 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 12: 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 delimited files

After pattern mining is complete you can view pattern information by selecting the delimited file from hierarchy pane and clicking on **Patterns** tab.

Figure 13: Viewing patterns

The list view under **Patterns** tab displays the most occurring pattern in each column of the selected file. Selecting individual columns from the list view displays detailed pattern information and character distribution under the respective tabs.

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 Delimited Files
- Discover Delimited File Metadata
- Load Delimited File
- Profile Columns
- Profile Domains

This topic teaches you to **profile domains** in delimited files.

To profile domains:

- Select the desired file from the left hierarchy pane and select **Profile Domains** icon from the application toolbar. You can also click Systems menu under Actions menu 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 14: Profiling domains - step 2

Click **Next**

- 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 15: 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 delimited files

Select the delimited file from the hierarchy pane and corresponding **Domains** tab to view the nearest domain match for each column. The list view also displays match score for each column. You can also view information on data type, classification and domain matched on details in this table.

In the **Domain Details** table you can view all domains that match with a particular column along with their match score and match type. You can identify each domain as **Verified Match or False Positive** by selecting the appropriate check box. You can also associate or disassociate a domain to the selected column by clicking on **Associate/Disassociate** button.

A graphical view at the bottom right displays the match score of all matching domains of a column. You can click on the graph and select an area to zoom it. Alternatively you can click on the edge of the horizontal and vertical sliders on the X and Y axis to zoom in and out of an area.

Index

S

Scanning Delimited files.....6

V

Viewing scanned delimited files.....7

W

Working with scanned Delimited files

Discovering Delimited file metadata 7

Finding Patterns..... 7

Loading Delimited files..... 7

Profiling columns 7

Profiling Domains 7