Pitney Bowes Software

# HDFS File Scanner

## User Guide

Spectrum™ Technology Platform

Data Discovery Module

Version 8

April 2013

Global IDs Product Suite® HDFS 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 HDFS File Scanner ..... 6
Scanning HDFS Files ..... 6
Discovering metadata of files ..... 9

# List of Figures

*Figure 1: Scanning HDFS files* ..... *7*
*Figure 2: Hierarchy pane* ..... *8*
*Figure 3: Viewing scanned files* ..... *8*
*Figure 4: Structured Files tab* ..... *9*
*Figure 5: Discovering metadata* ..... *10*
*Figure 6: File metadata discovered* ..... *10*

# HDFS File Scanner

## What it includes:

- About HDFS File Scanner
- Scanning Hadoop File System
- Discovering metadata of structured files

# About HDFS File Scanner

Using **HDFS File Scanner** you can scan all files on Hadoop Distributed File System (HDFS) located across your enterprise data landscape and discover their metadata. Both version 1 and 2 of HDFS can be scanned.

# Scanning HDFS Files

This topic teaches you to find out files present in HDFS servers located in any system in your network.

To do so:

- Right click the root of the left hierarchy pane and select **Scan Hadoop File System** option. **Scan Hadoop File System** wizard appears.
- On Step 1 of the wizard:
  - Provide a connection name.
  - Provide the host name or IP address.
  - The default port is displayed. You can change it if required.
  - Click **Test Connection** link. This will check connectivity to the HDFS server. If the connection is successful, a dialog box will appear with the message "**Connection Succeeded**".
  - Specify a directory or folder to scan or select appropriate checkbox to scan the entire system.
  - Specify the types of files to scan. The following file types can be scanned.
    - PDF (.pdf)
    - Flat File (.txt)
    - Rich Text Format (.rtf)
    - Excel (.xls and .xlsx)
    - Power Point (.ppt and .pptx)
    - Word (.doc and .docx)
    - Access 2007 Database File (.accdb)
    - Access (.mdb)
    - Hypertext Markup Language File (.htm and .html)
    - Comma Delimited (.csv and .txt)
    - Pipe Delimited (.txt)
    - Tab Delimited (.tsv, .tab and .txt)
    - Java Source Code File (.java)
    - C++ Source Code File (.cpp and .cc)
    - C/C++ Source Code File (.c)
    - VBScript File (.vb)
    - Java Server Page (.jsp)
    - Hypertext Preprocessor File (.php)
    - Python Postscript File (.py)
    - C# Source Code File (.cs)

- Text (.txt)

By default both these file types are scanned.

- Select appropriate checkbox to:
  - scan compressed files
  - scan subdirectories

**Figure 1: Scanning HDFS 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 files

Once the HDFS server is scanned, it appears in the left hierarchy pane as a system.

Figure 2: Hierarchy pane

You will be able to view the files in the HDFS server from the **Scanned Files** tab if you select it from the left hierarchy pane. The **Scanned Files** tab is divided into 2 sections. The scanned files of the selected HDFS server appear in a separate explorer pane. These files are categorized by the directories under which they reside. You can view this explorer pane from the **Scanned Files** tab on the right display pane if you select a system/HDFS server from the left hierarchy pane. In the explorer pane, you can select a directory from the hierarchy on the left to view all scanned files along with their path, size and last modified date in a list view on the right. You can double click a file name from this list to view its content. Please note, you will be able to view the content only if the file is residing in your local system.

| File Path | File Name | File Size (in KB) | Last Modified |
|---|---|---|---|
| /Sample_Files | BIRDS.tsv | 717.58 | 2013 03 13, 16:53:27 |
| /Sample_Files | COVERS.csv | 3.29 | 2013 03 13, 16:53:28 |
| /Sample_Files | CUSTOMER.tsv | 6202.68 | 2013 03 13, 16:53:29 |
| /Sample_Files | Categories.tsv | 1390.54 | 2013 03 13, 16:53:27 |
| /Sample_Files | DATA_SRC.csv | 53.38 | 2013 03 13, 16:53:29 |
| /Sample_Files | FACET.tsv | 4350.63 | 2013 03 13, 16:53:30 |
| /Sample_Files | FOOD_DES.tsv | 1006.41 | 2013 03 13, 16:53:30 |
| /Sample_Files | LOCATIONS.csv | 1.26 | 2013 03 13, 16:53:30 |
| /Sample_Files | ORDER_DETAILS.tsv | 1448.71 | 2013 03 13, 16:53:30 |
| /Sample_Files | ORDER_INFO.txt | 145.09 | 2013 03 13, 16:53:30 |
| /Sample_Files | PERFORMANCE.txt | 160957.69 | 2013 03 13, 16:53:56 |
| /Sample_Files | PERSONS.txt | 385.16 | 2013 03 13, 16:53:56 |
| /Sample_Files | PRODUCT.tsv | 452.2 | 2013 03 13, 16:53:57 |
| /Sample_Files | PUBLICATIONS.tsv | 862.79 | 2013 03 13, 16:53:57 |
| /Sample_Files | SALES.csv | 452.18 | 2013 03 13, 16:54:08 |
| /Sample_Files | SALESDATABASE.txt | 1779.42 | 2013 03 13, 16:54:08 |
| /Sample_Files | SUPPLIERS.csv | 3.63 | 2013 03 13, 16:54:08 |
| /Sample_Files | TERRITORIES.csv | 3.25 | 2013 03 13, 16:54:08 |
| /Sample_Files | VERNACULARS.txt | 9176.4 | 2013 03 13, 16:54:09 |

Figure 3: Viewing scanned files

You can view the structured files like CSV, TSV etc. from the **Structured Files** tab if you select a system (HDFS server) from the left hierarchy pane. The **Structured Files** tab is divided into 2 sections. The scanned structured files of the selected HDFS server appear in a separate explorer pane. These files are categorized by the directories under which they reside. You can view this explorer pane from the **Structured Files** tab on the right display pane if you select a system/HDFS server from the left hierarchy pane. In the explorer pane, you can select a directory from the hierarchy on the left to view all scanned structured files along with their row count, column count, whether metadata has been discovered, file owner and path in a list view on the right. If you select a file from this list, you will be able to view its data (a sample set) and metadata from appropriate tabs at the bottom of the right display pane.

**Figure 4: Structured Files tab**

# Discovering metadata of 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 Hadoop File System
- Discover File MetaData

This topic teaches you to discover the metadata of the scanned files.

To do so:

- Right click the desired HDFS 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 files whose metadata you want to discover.
  - Select appropriate checkbox for considering the first row of the files as column header. Click **Next**.

Figure 5: Discovering metadata

- 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 file metadata

Once metadata of the files is discovered, the **Metadata Discovered** column in the **Structured Files** tab will be checked and the row count and column count columns will be populated. You can select any file from the list to view its metadata from the **Metadata** tab at the bottom the **Structured Files** tab.

| File Name | Row Count | Column count | Metadata Discove... |
|---|---|---|---|
| Files | | | |
| SALES.csv | 4248 | 18 | ✓ |
| PRODUCT.tsv | 4418 | 18 | ✓ |
| CUSTOMER.tsv | 50000 | 13 | ✓ |
| PUBLICATIONS.tsv | 3355 | 14 | ✓ |

Figure 6: File metadata discovered

# Index


## D

Discovering metadata of files .....9

## H

Hadoop Distributed File System.....6

## S

Scanning HDFS Files..... 6

## V

Viewing file metadata ..... 10

Viewing scanned files ..... 7