Hitachi Vantara | Pentaho+

Pentaho GenAI Plugin Suite

e

USER GUIDE DOCUMENTATION
HTML Parser

Pentaho Professional Services
2024

Table of Contents

Revision History

Below is a list of edits, updates, and refinements made to this document, detailing the progression and refinement of its content over time.

Date	Version	Author	Major Revisions
18-Sept-2024	1.0	Pentaho Professional Services	First Version

1. Introduction

1.1 Purpose

This document is a plugin user guide document. The purpose of the document is to provide end users with information regarding the plugin. The user guide document will have the following sections:

- **Plugin Description**: Information about the plugins and UI view of the plugin.
- **Compatibility Matrix**: Details about the supported versions and minimum requirements for the usage of the plugin.
- **Installation Steps**: The process to install the plugin.
- **Parameters**: The details about the plugin parameters that a user can configure including metadata injection and error handling.
- **Plugin Usage**: High level information about using the plugin, including a brief description of the samples.
- **Support**: Details about the plugin support.

1.2 Audience Scope

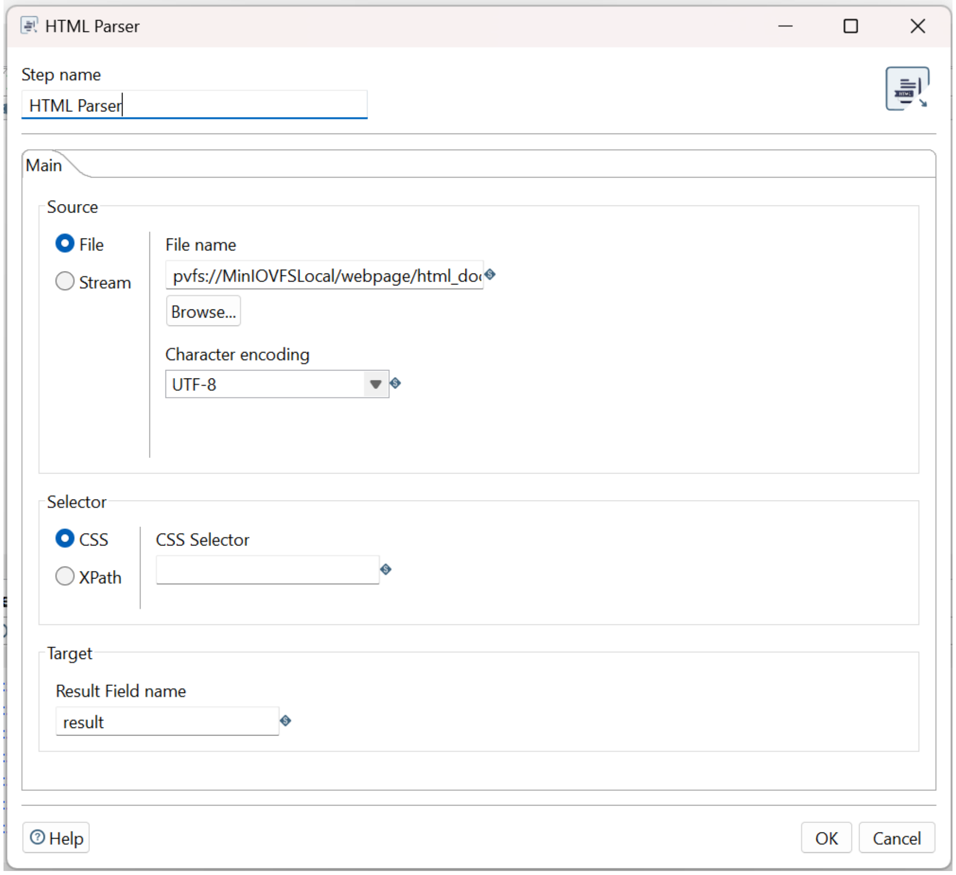
This user guide document is intended for the Pentaho EE customers who are using the GenAI suite.

2. Plugin Description

2.1 Plugin Metadata

Attribute	Description / Value
Plugin ID	HtmlParser
Plugin Name	**HTML Parser**
Plugin Version	**1.0.0**
Plugin Icon	HTML
Plugin Description	A utility plugin that allows you to extract desired text out of HTML or XML. Useful for cleaning data for the sake of NLP like sentiment analysis and SEO like keyword analysis. It accepts data from the stream and files.
Plugin Type	Pentaho Data Integration Step
Plugin Category (Spoon)	Transform

2.2 Plugin UI View



HTML Parser

Step name: HTML Parser

Main

Source

● File
○ Stream

File name: pvfs://MinIOVFSLocal/webpage/html_do

Character encoding: UTF-8

Selector

● CSS
○ XPath

CSS Selector:

Target

Result Field name: result

Figure 1: View of the HTML Parser plugin in PDI Spoon Interface.

3. Compatibility Matrix

3.1 Supported Systems

Platform	Version
Pentaho+ Data Platform	Pentaho Data Integration version 10.2.x and higher
Java	JDK version 11 or higher.

4. Installation

4.1 Download Plugin

To download the latest version of the plugin, contact your Pentaho account manager or Pentaho Support team to provide you with the plugin.

- The plugin file *(in zip file)*: pdi-step-parse-html-plugin.zip
- The plugin file *(without zip file):* pdi-step-parse-html-plugin/

4.2 Install Plugin

To use these plugins, follow the standard procedure for installing PDI plugins.

For Standard PDI Installation (Spoon):

- Unzip the plugin file *(if provided in zip format)*.
  - For Windows, use “extract all” to unzip.
  - For Unix/Mac, use the command: unzip pdi-step-parse-html-plugin.zip -d <download path>
- Place the extracted plugin folder/files in the plugins/ directory of your Pentaho Data Integration installation.
- Restart Pentaho Data Integration.
- Access the plugins via the PDI interface under the appropriate plugin categories.

For Pentaho dock-maker Installation:

- Place the extracted plugin folder/files in generatedFiles/fileOverride/plugins location.
- Stop and Start docker-compose.yml to restart pentaho data integration.

For more information, follow the official plugin documentation for installing pentaho plugins.

5. Plugin Parameters

5.1 Configurations

Tab: Main

There is only one main tab in this plugin. Use the main tab to perform all the tasks required for this plugin.

Section 1: Source

File

Select this option when working with a single file from the local file system or VFS file system.

Option	Description
File name	The absolute path of the location of the file. Use the "browse" button to choose a file from the local or virtual file systems. If you are sending URI as a filename (e.g.: file:///C:/Users/sample.txt), ensure that all the path separators are back-slash (/) instead of forward-slash (\\).
Character Encoding	Select the character encoding of the file. By default, it is "UTF-8".

Stream

Select this option when working data from an input stream *(previous step).* The input stream can be either a file path *(File)* or raw text *(Binary Datum).*

Option	Description
Format	There are two format options for users to choose from: • **File**: Select this option if you are sending absolute file paths from the input stream. • **Binary Datum**: Select this option if you want to send a list of raw text.
Field name	Select the field that you need to process. Ensure that the field content and format *(File/Binary Datum)* should match. Mismatch in the field and format will result in incorrect parsing.
Character encoding field name	Select the character encoding from the input field. Each input row can have a different encoding field value. By default, it is set to "UTF-8".

Section 2: Selector

Select between CSS or XPath languages to extract text based on a query. This is an optional section. By default, all the text values will be extracted from HTML files.

Option	Description
CSS Selector	Select this option to parse based on a CSS tag. Leaving this field blank will result in all tags being removed and being returned the extracted text content. For more on CSS Selector syntax, follow this document.
XPath Selector	Select this option in order parse based on a XPath expression. Leaving this field blank will result in all tags being removed and being returned the extracted text content. For more on XPath Selector syntax, follow this document.

Section 3: Target

Option	Description
Result Field Name	The result field name. This field will return the HTML parsed output. By default, it is set to “result”. User can have a different field name of their choice or use a Pentaho Environment Variable.

5.2 Environment Variable Support

This step supports Pentaho Environment Variables for all the step components. The support for variable substitution is indicated by $ ($) next to the step component. Missing the $ symbol indicates no support.

5.3 Metadata Injection Support

This step supports metadata injection.

5.4 Error Handling Support

This step does not support Error Handling flow of PDI. All errors and exceptions will be thrown in the log and would result in the execution to stop. Future upgrades of the step will support this.

5.5 List of Supported File Format

It supports file formats of various documents *(Text, HTML, XML, etc.)*.

6. Plugin Usage

The downloaded plugin file is bundled with samples of using this step plugin. Locate the samples directory inside the <pdi installation>/data-integration/plugins/pdi-step-parse-html-plugin/samples.

7. Support

For more information, support, or to inquire about licensing and usage, please contact:

Pentaho Support Team	https://support.pentaho.com/hc/en-us