Windows Installation Guidelines for Pentaho Data Integration (Community Edition)

This document contains simple installation guidelines and two guidelines from external websites for installation of the community edition of Pentaho Data Integration on Windows. If your installation did not work with the simple installation guidelines below, I encourage you to begin your installation again following one of the external guidelines. Both guidelines provide details about Java installation (JRE and JDK) and Windows environment variables. You can also find other guidelines and video installation guidelines if you search under "Pentaho Data Integration Installation Guidelines for Windows". Hitachi provides a white paper with installation instructions as another option for Windows installation. Make sure that you install the community edition, not the trial version of the commercial edition of Pentaho Data Integration. The commercial version has a limited free usage period.

Simple Installation Guidelines

To install Pentaho, you should follow the steps below. It is highly recommended that you use the community edition from SourceForge as the instructions in this document follow the community edition interface. The simple guidelines assume that you installed Java before installing PDI. If you have not installed Java, you should use one of the external installation guidelines.

- The latest PDI version in June 2022 on SourceForge is 9.3.0. You should be able to use this version or a newer version to complete the tutorial and assignment.
- You can find the community edition download for version 9.3.0 and other versions at sourceforge.net/projects/pentaho/files/.
- Unzip the downloaded zip file to any folder.
- Windows users should copy the folder data-integration to the folder C:\Program Files\Pentaho. Mac and Linux users may move the file to any folder.

To ensure that the installation worked, you should launch Pentaho Data Integration.

- Run the file Spoon.bat by double clicking on it. You may want to create a shortcut to the spoon.bat file so starting data integration is easier. If you get a permission error or cannot execute the bat file, you should right click and select "Run as Administrator". For Mac and Linux users, run the Spoon.sh from terminal (./spoon.sh).
- If Pentaho Data Integration does not launch after trying the simple guidelines, you should try one of the external guidelines, Another document on the course website contains resolutions of common installation and execution problems with Pentaho Data Integration.

Pentaho Data Integration (PDI) Installation Guide - Easy yet powerful ETL tool

From: https://ai.plainenglish.io/pentaho-data-integration-installation-guide-easy-yet-powerful-etl-tool-80930cff46c6

Shravankumar Suvarna

April 10, 2020

Building data-pipelines using simple graphical user-interface without writing a single line of code.

This is a step-by-step installation guide for Pentaho Data Integration. Well, why Pentaho Data Integration aka 'Kettle'? Pentaho Data Integration (PDI) is an ETL (Extract, Transform, Load) tool to manage data ingestion pipelines. As we generate more and more data via various sources and formats, it gets difficult to manage the data pipelines for better decision making.

PDI is a useful tool to manage such pipelines seamlessly. I'll be writing a series of blogs explaining the end-to-end process of creating configurable data ingestion pipelines for managing a variety of data structures and formats. We will start with the installation process first and end with deployment.

PDI comes in two editions - Enterprise and Community. We will be using community version in this blog series so that everyone can follow along.

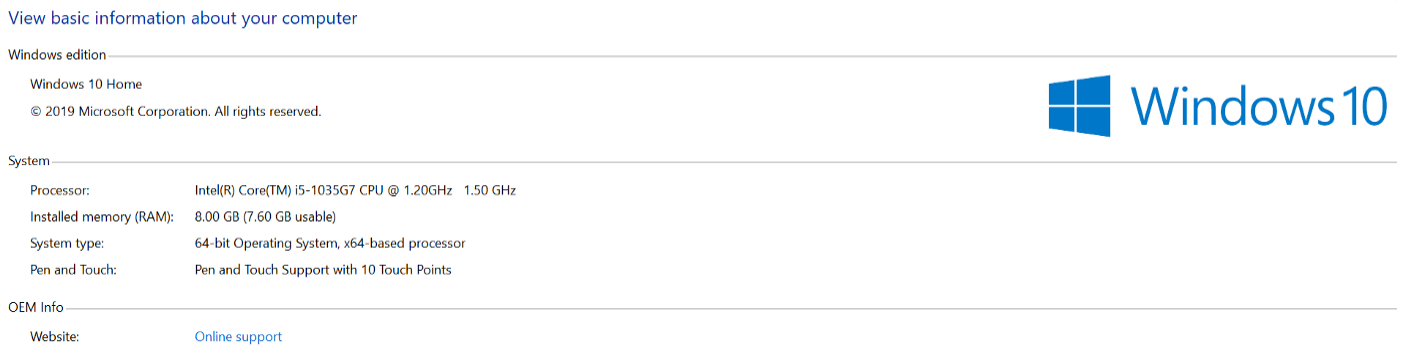
So let's cut the introduction and start with the actual process.

Please note, since I use a Windows laptop some steps will be specific to Windows OS users only. However, the majority of the same is OS independent.

Prerequisite:

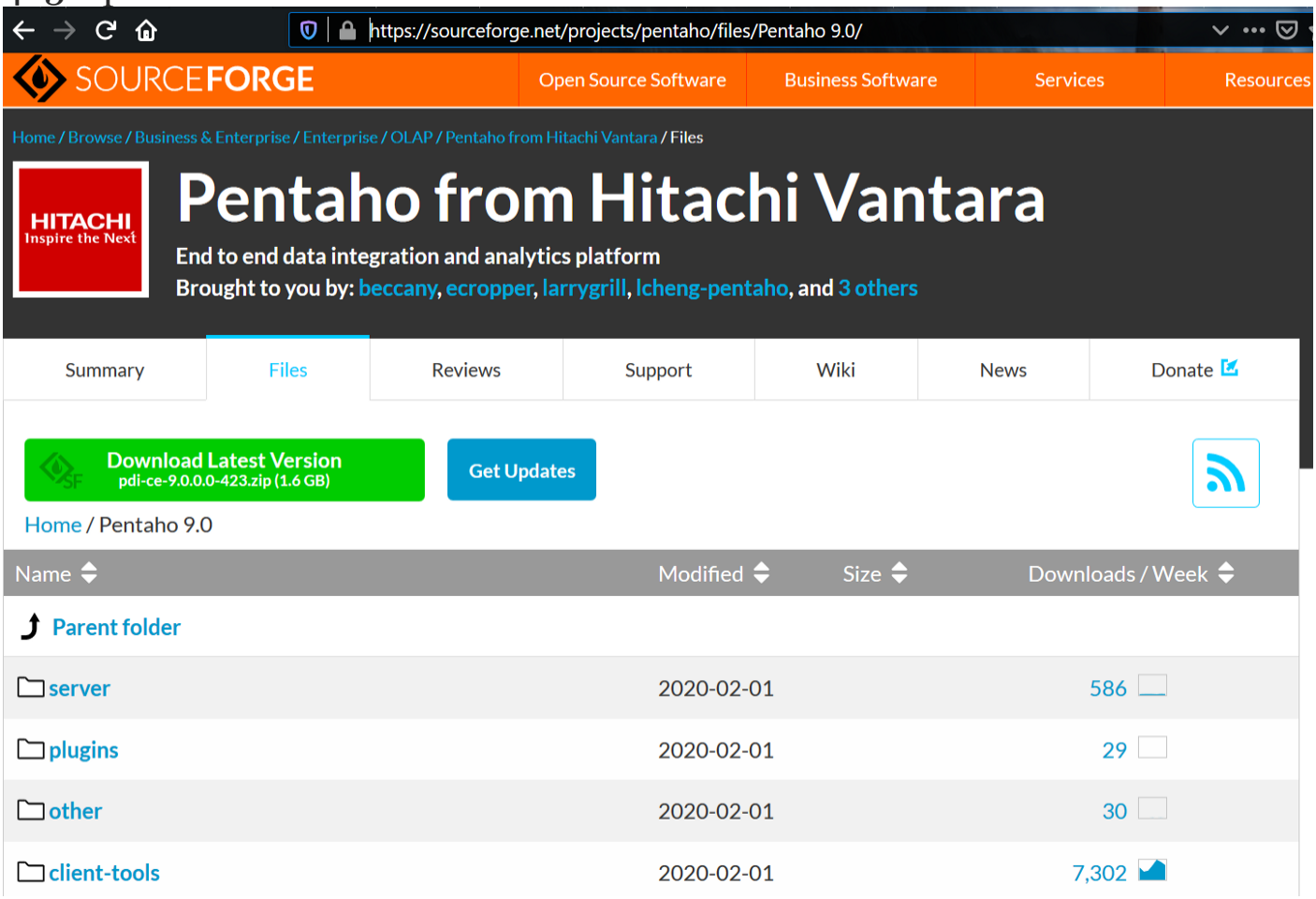
- **Processor:** Intel EM64T or AMD64 Dual-Core
- **RAM:** 8 GB with 2 GB dedicated to PDI - It can work on a 4GB RAM system as well. However, it is recommended to have an 8GB RAM system.
- **Disk Space:** 20 GB free after installation
- **Screen size**: 1280x 960 - Easy to view to PDI UI

How to check if you meet the above requirements? You can right-click on This PC to choose properties. You check the system configurations.



Step-1: PDI Download

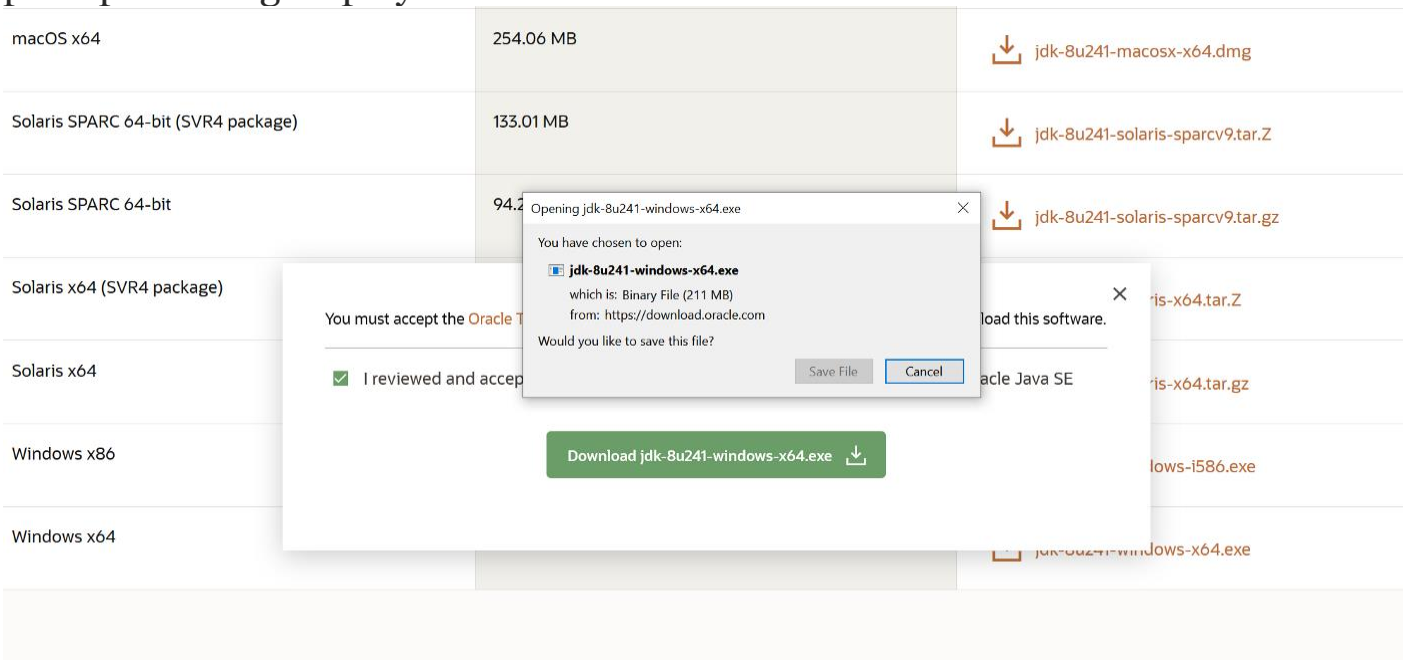
Download the PDI-CE from SourceForge *link* at the time of writing this blog. PDI's latest version is 9.0, you can download the latest stable version as per your requirements. The file name is 'pdi-ce-9.0.0.0–423.zip'.



Click on the green button with title 'Download Latest Version' to download the .zip file

Step-2: Java Download

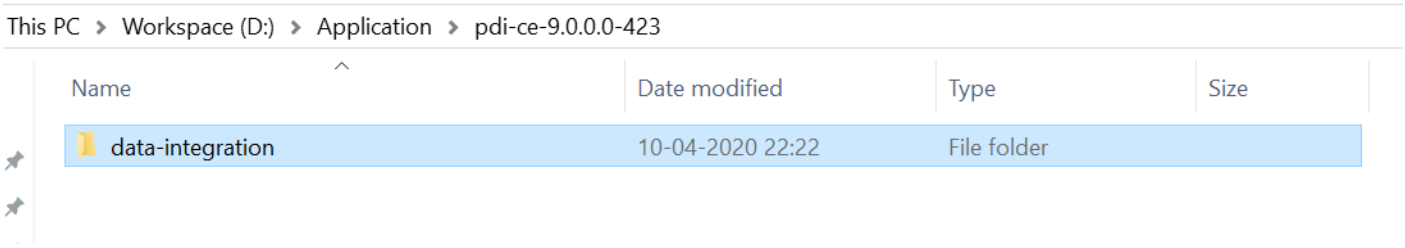
Download Java SE Development Kit 8 from the *official website*. Since PDI is built using Java as a programming language in the back-end. Download the version as shown in the below image. You will be prompted to sign-up by Oracle with basic information.



Step-3: Extract PDI .zip file and Install Java

1. PDI

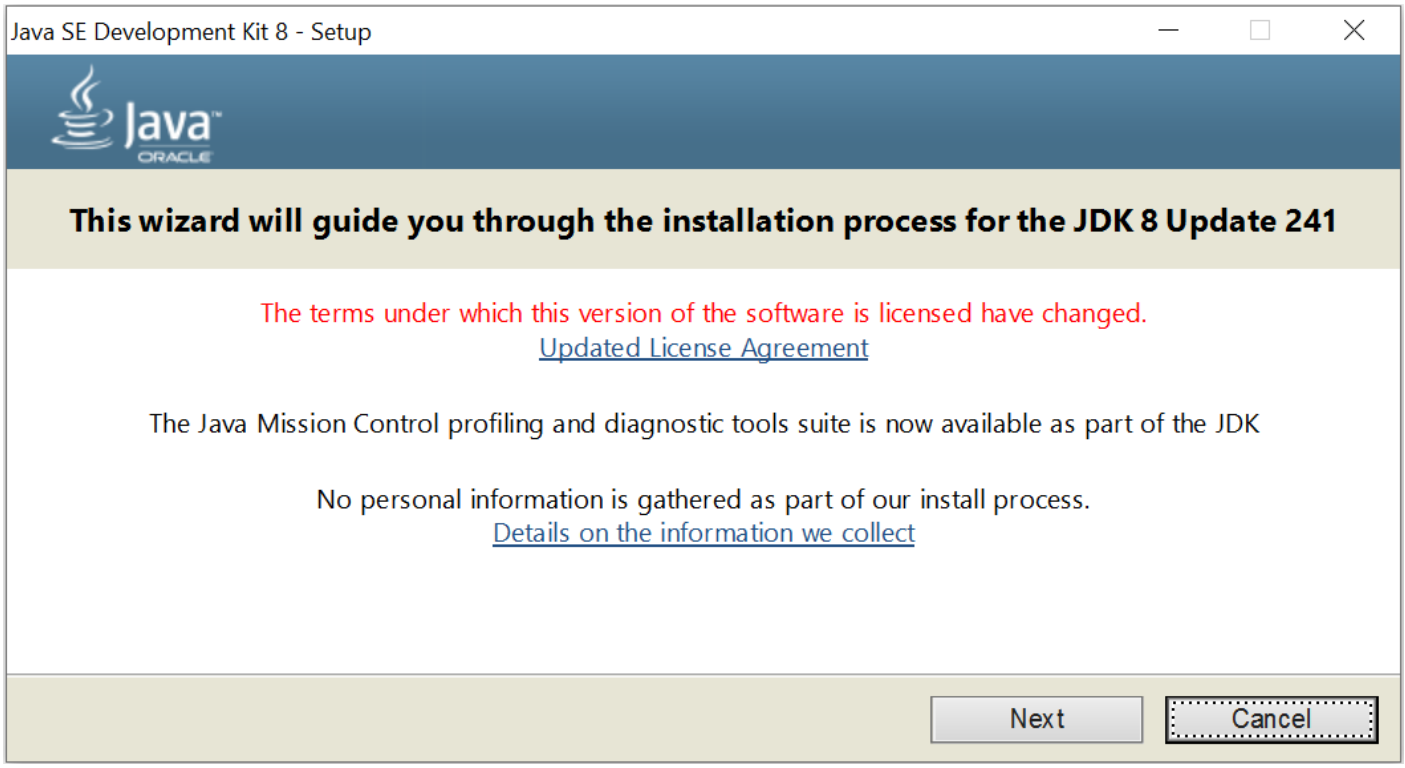
Extract the PDI .zip file in a setup folder. It is recommended to store it in the non C drive (Since the size of the file is more than 1GB). Ideally, I create a folder 'Application' in 'D' drive and store all the third-party applications in the same. Let's go with the same approach here. There's no executable file (.exe) that we need to run to install PDI, just the extraction of .zip file. Easy!



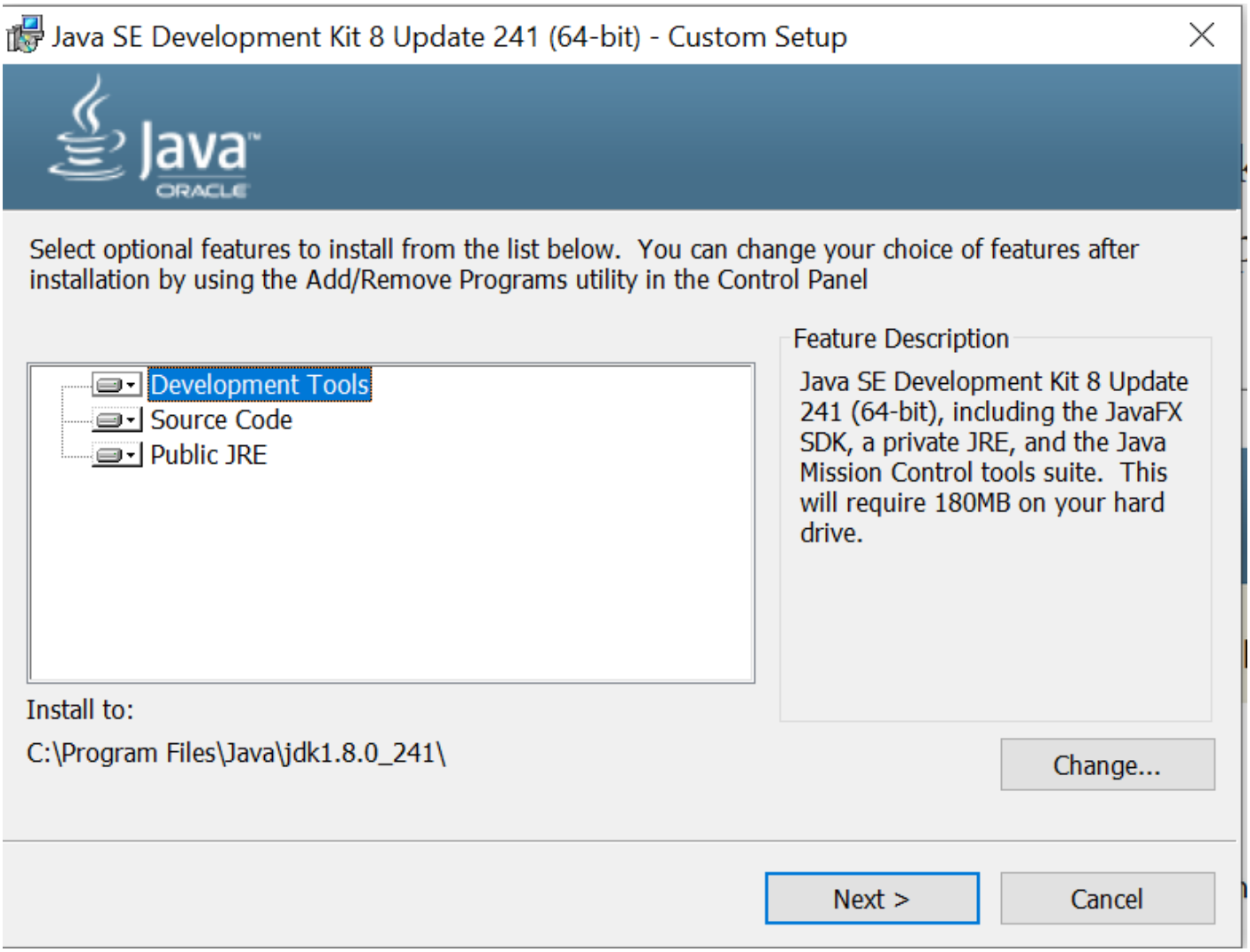
You will see the data-integration folder post-extraction

2. Java

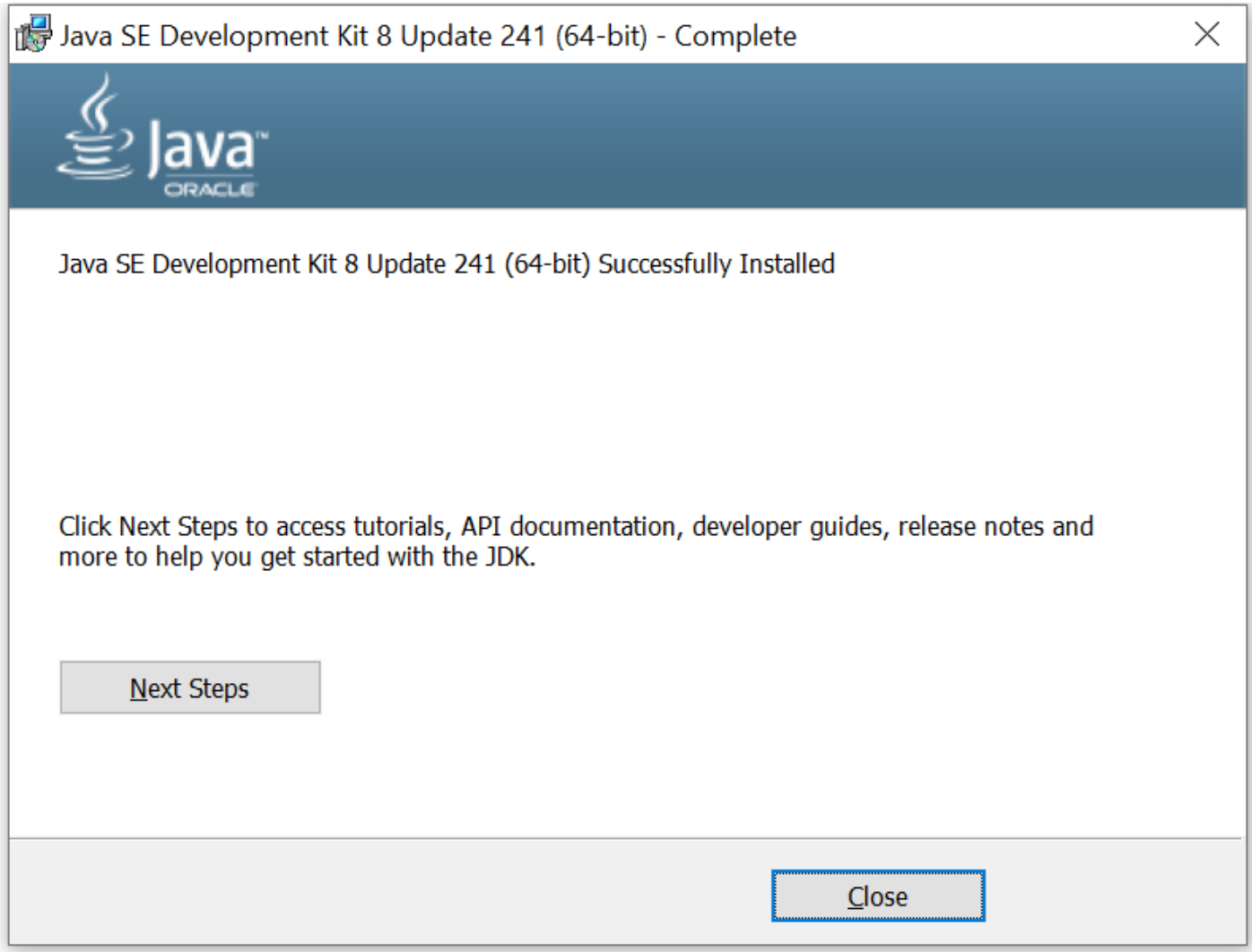
Installation of Java is a simple process. We just need to double click the windows installer .exe file, agree to the terms and click next multiple times. Go with the default unless you need specific customizations.



You will see the above screen post clicking the .exe file



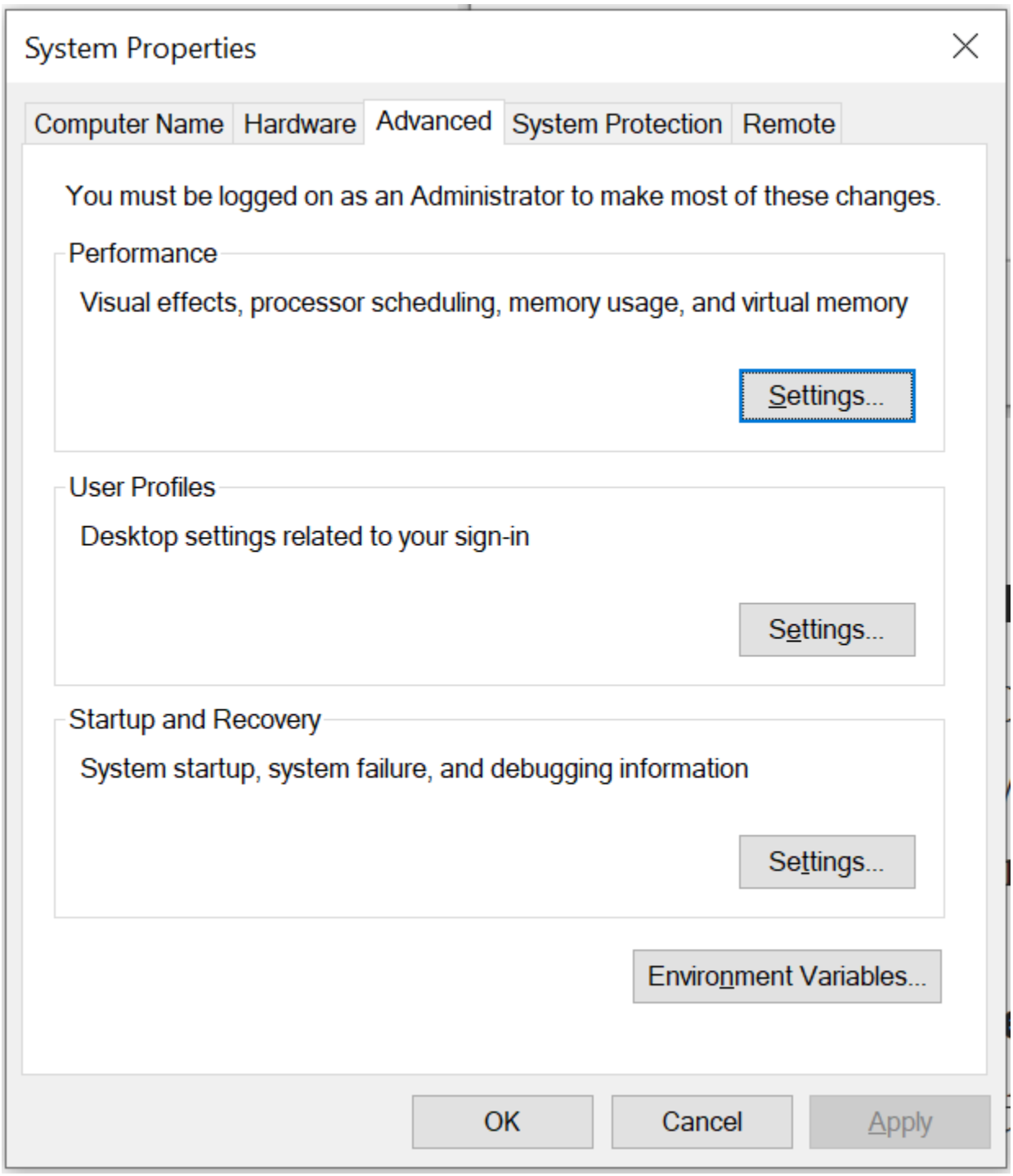
Please choose Development Tools option



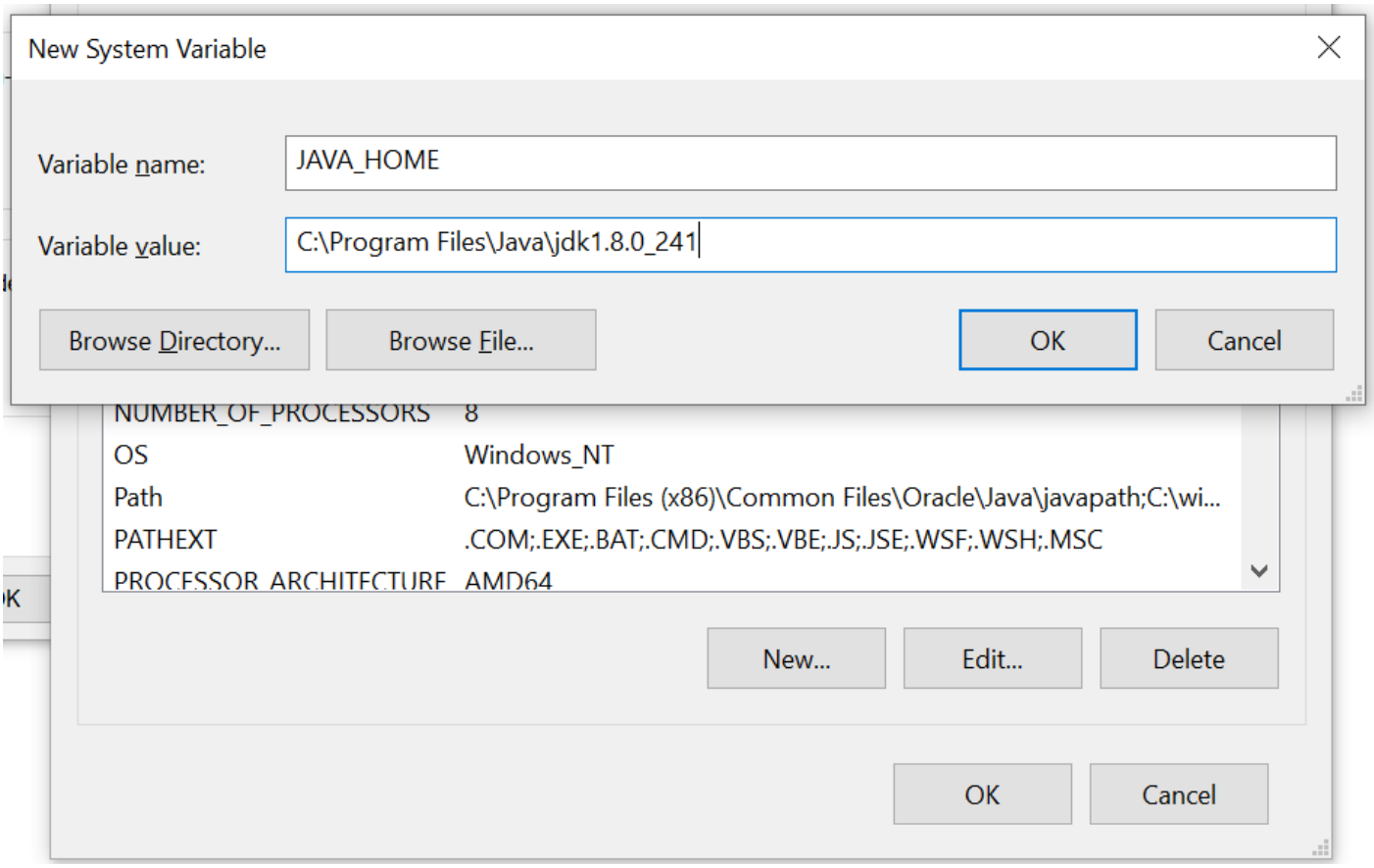
Successfully Installed JDK and JRE

Step-4: Add Environment Variables

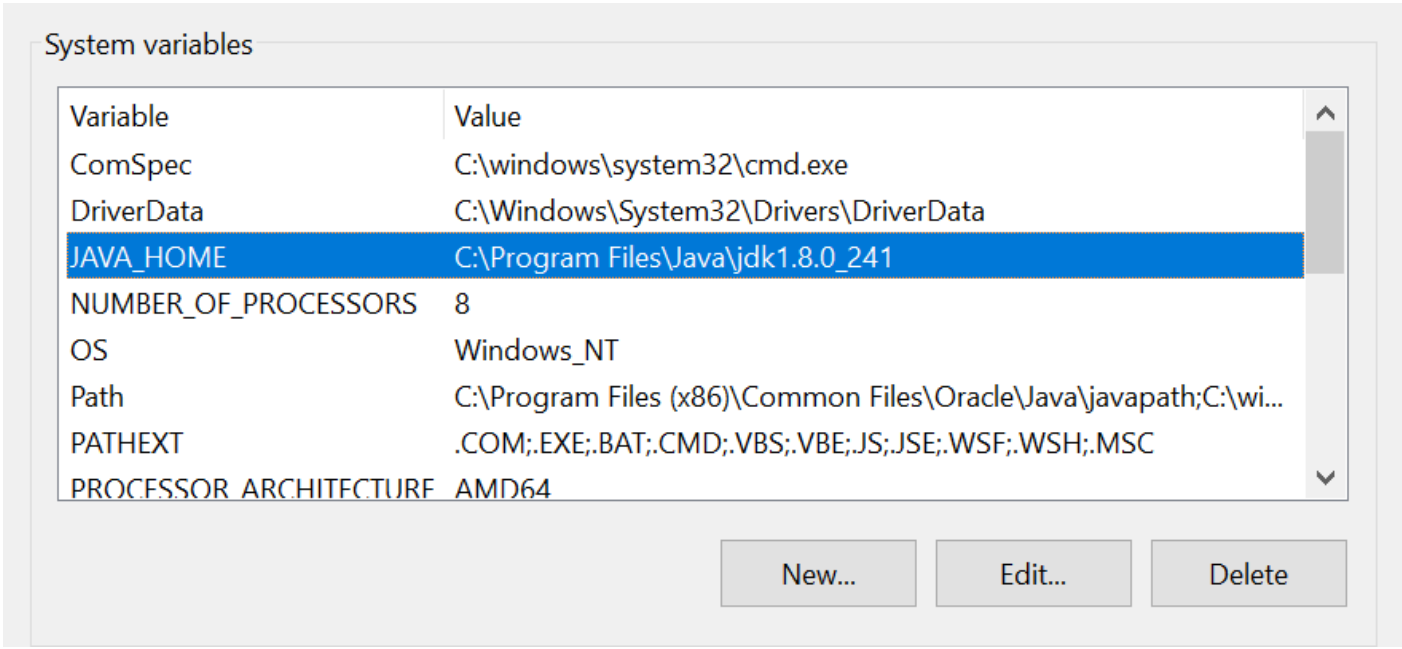
We need to add the JAVA_HOME path on our environment variables. Now, this one is specific for Windows users only. We will have to go to start menu and type 'environment variables'. Once you select the option, the below screen will be displayed, click on the 'Environment Variables' button.



We need to edit the system variables. Please make sure you have administrator rights

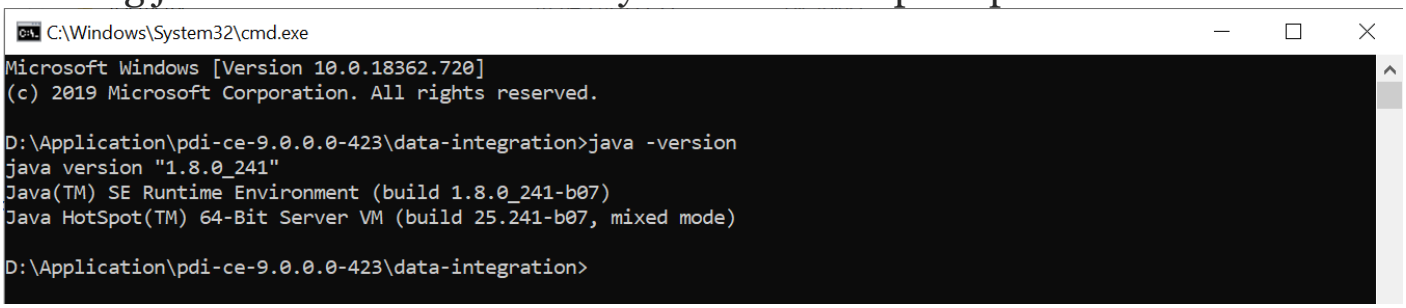


We need to provide the path to JDK file in variable value.



You can add JAVA_HOME variable by clicking on the New button

You can perform a quick check if the Java was installed properly by running java -version command in your command prompt.



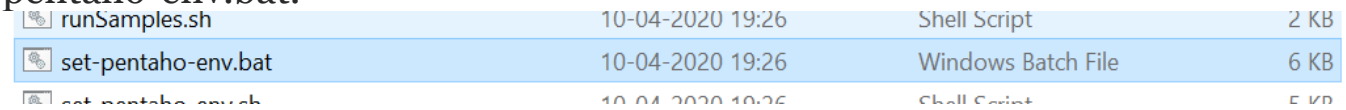
Please make sure the java version is above 1.8

****Below process is optional****

There is another batch file which lets you set up the environment-specific for PDI. However, I was able to run the spoon.bat file without following the below process. Please observe in case you are not able to run the file by following the above steps.

Where is this batch file located?

It's within the extracted data-integration folder with the name set-pentaho-env.bat.



This is within the folder data-integration

What does this batch file do?

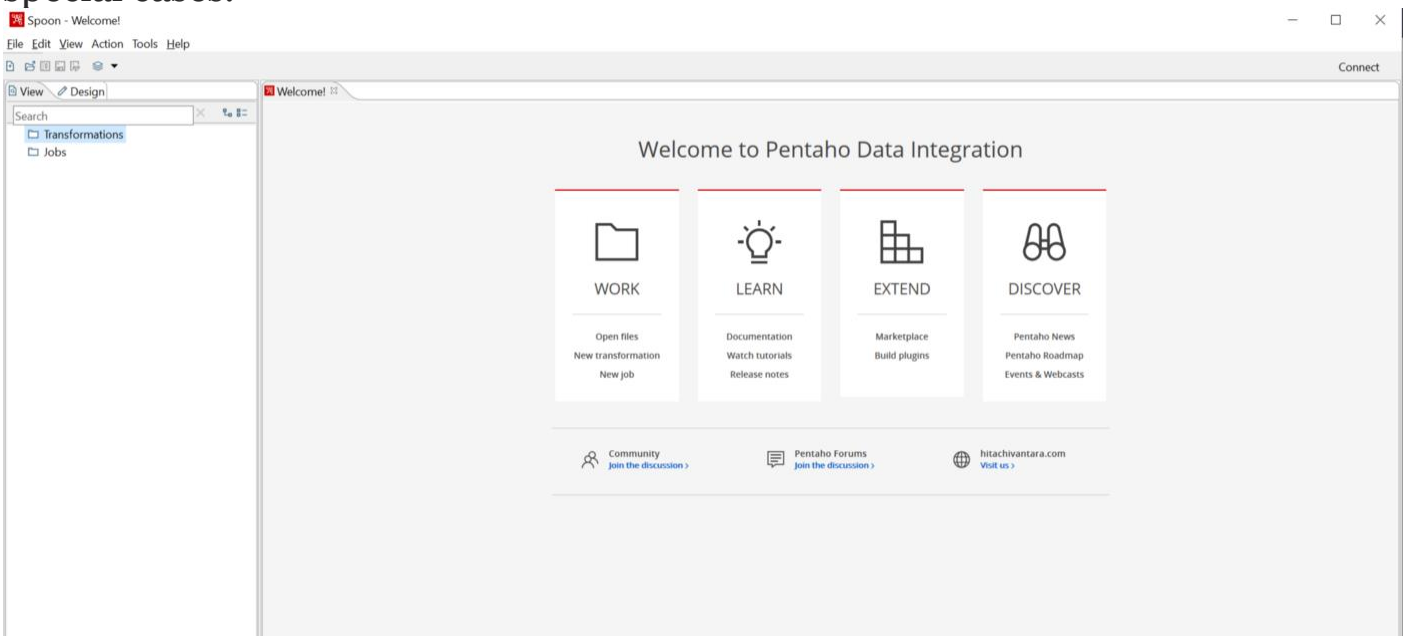
Well, it looks in well-known locations to find a suitable Java then sets two environment variables for use in other .bat files. The two

environment variables are:
* _PENTAHO_JAVA_HOME — absolute path to Java home
* _PENTAHO_JAVA — absolute path to Java launcher (e.g. java.exe)

Step-5: Open Spoon - UI

Spoon is an important component of PDI. Spoon.bat file helps you open the user-friendly UI; which allows you to simply create complex data ingestion pipelines using drag-drop widgets.

Yes - we don't have to write a single line of code to process the unstructured files; everything happens using simple widgets. However, we can write Javascript or Java code to create our widgets to handle special cases.



If you can see the above welcome screen on your PC; then yes, we have successfully installed PDI on your machine.

Conclusion

This is the beginning of a journey. I will walk you through the entire data ingestion process via a series of blog posts. This will cover from reading excel/csv files, performing calculation/transformation and loading the same in the database.

Install Pentaho Data Integration (CE) on Windows - A Detailed Step-by-Step Tutorial

February 25, 2019 / 4 comments

From: https://piezobytes.blogspot.com/2019/02/install-pentaho-data-integration-on-windows.html

Install Pentaho Data Integration (CE) on Windows - A Detailed Step-by-Step Tutorial

Hi fellow Devs! Hope you had a wonderful day. Today's tutorial is all about how to download and install Pentaho Data Integration (Community Edition) in Windows laptop/desktop.

For those who are new to ETL process and PDI/Kettle, check my blog post:

Tutorial on Pentaho Data Integration.

When I was new to Pentaho, I was having difficulty in installing particularly the configuration part. I have searched internet and learned those things by trial and error. Hence, I thought it will be worth sharing my experience.

The basic requirements are:

- Pentaho Data Integration Community Edition
- JRE 1.8 and above (Java Runtime Environment)
- JDK 1.8 and above (Java Development Kit)

- Windows 7 and above (Though PDI can be installed in Linux or Mac OS as well, the scope of this post is limited to Windows Operating System)

Step-1: Downloading the Pentaho Data Integration (PDI/Kettle) Software

The first step is to download the PDI Community Edition from the Official Sourceforge download page. The recent version is 8.2 at the time of writing this post and the download file is about 1.1 GB. The files will be downloaded as a zip file in the name 'pdi-ce-8.2.0.0.-342.zip'.

Step-2: Extracting the zip file

Extract the downloaded zip file which will be in the Downloads folder . Right click the file and choose the 'Extract Here' if you want it to get extracted in the downloads folder.

If you want to choose a different folder, then right click and select 'Extract Files...' option and give the destination folder path. The default name of the extracted folder would be 'data-integration'.

Step-3: Checking JRE version

Next step is to check the Java Runtime Environment version in your system. First check if your machine has Java installed. If not download it from official Java download page.

If it is already installed, check for JRE versions. To do this,

Go to C:\Program Files\Java\ (in case if it is 32 bit Windows OS)

(or)

check it in C:\Program Files (x86)\Java (in case of 64 bit OS).

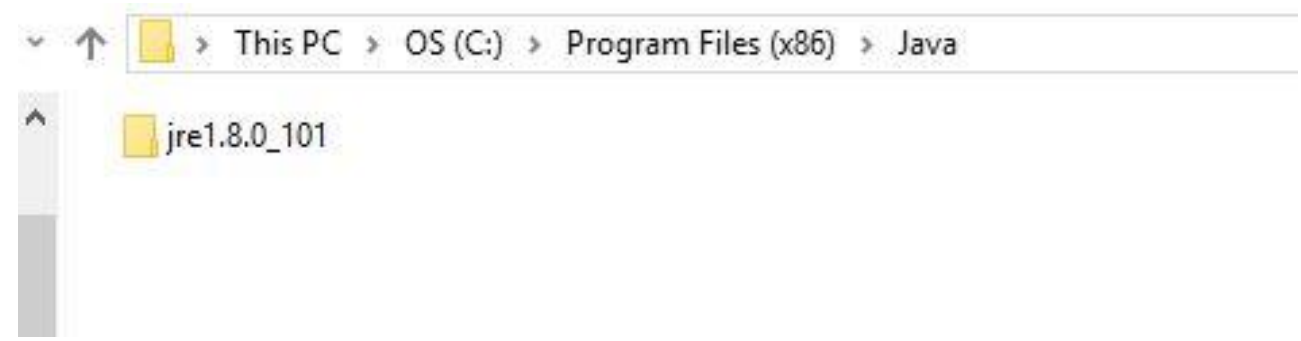


Fig 1. Screenshot showing the path containing JRE

There will be a folder within this folder path, starting with 'jre' followed by version number. This version should be 1.8 or higher. If this is the case, then we are good to proceed with the next step.

If your PC/Laptop does not have JRE, we need to download the JRE 1.8 or higher version from the official Oracle JRE download page and install it. The recent version at the time of writing this post is '1.9'. We need to download the correct file according to OS type (32-bit/64-bit) and also select the file ending with '.exe' extension as it does not require any extracting tools.

Once the file is downloaded, run the file and install JRE.

Step-4: Checking JDK version

The next step is to check the version of JDK in your Windows PC. This step is like the previous step.

Go to C:\Program Files\Java\

In the same folder, there will be a folder with a name similar to 'jdk1.8.0_191'. If you can find this folder, then we are good with this step. If the folder is not there, we have to download the JDK from the official Oracle JDK download page.

Like the previous step, we need to download the correct file according to our OS architecture type (32-bit/64-bit) and there will be multiple formats of the file will be available. Choose the file that ends with '.exe' extension and install the JDK by running the file after the download.

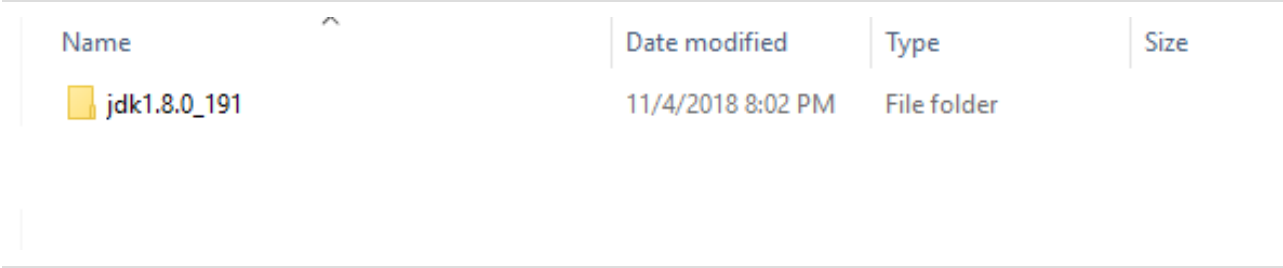


Fig 2. Screenshot showing the path containing JDK folder

Step-4 Setting the Environment Variables

The final step is to configure the environment variables to point to the JRE folder path.

- Open My Computer (or) press Ctrl+E.
- In the left side pane, right click 'Computer' for Win 7 (or) 'This PC' for Win 10 and select 'Properties' option. A new window will appear showing the Processor, RAM capacity, Computer Name, etc.
- Alternatively, navigating to Control Panel\System and Security\System can bring the same window.
- On the left side pane, click the 'Advanced system settings' and it will bring the System Properties window. Go to 'Advanced' Tab and select 'Environment Variables' button.
- In the Environment variables window, click the new button.
- Give 'PENTAHO_JAVA' in upper case as variable name and folder path pointing to java.exe under the jre folder as variable value and click Ok. Please refer the screenshot below.

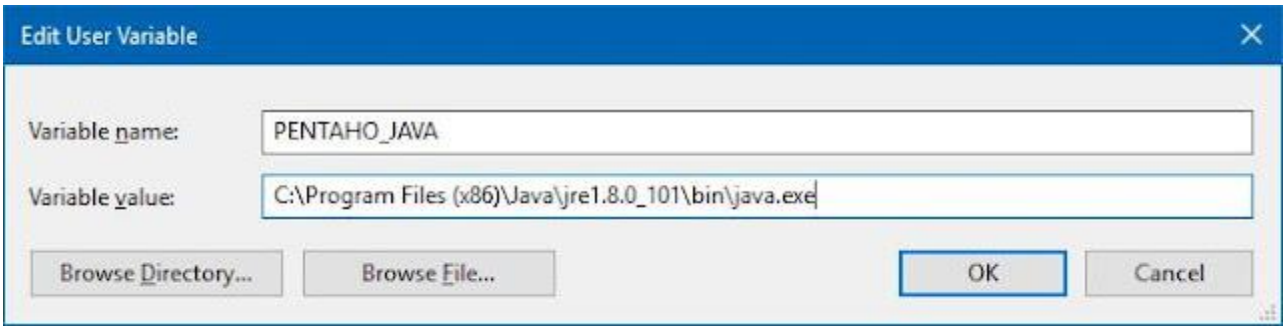


Fig 3. Creating Environment variable - PENTAHO_JAVA

Again, create a new variable in the same way, but set the variable name as 'PENTAHO_JAVA_HOME' as variable name and jre folder path as variable value like in the screenshot given below.

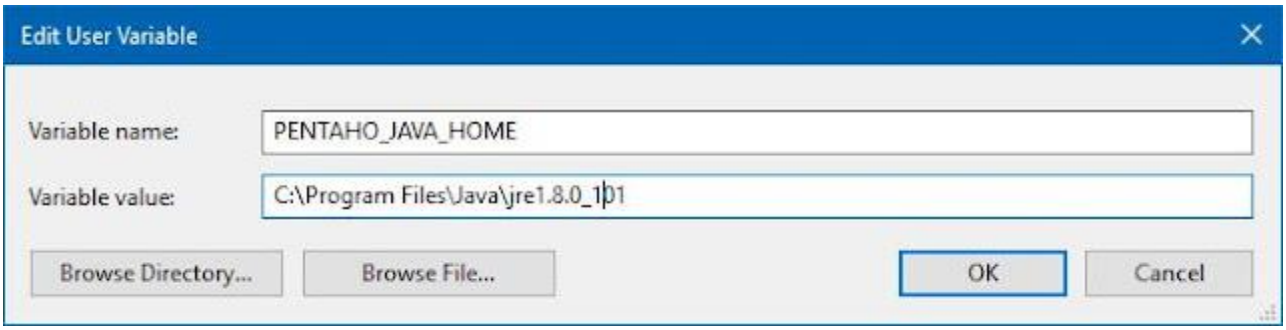


Fig 4. Creating Environment variable - PENTAHO_JAVA_HOME

Finally, create another variable in the name JAVA_HOME and the variable value would be the path where the jdk folder is available (see Step-4).

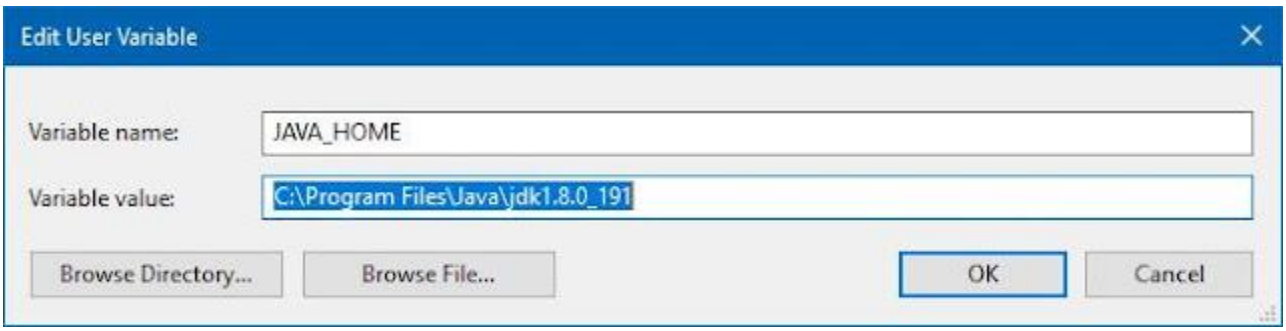


Fig 5. Creating Environment variable - JAVA_HOME

Note:

In the above step, example is given for 64-bit version of Windows. If your machine is 32-bit OS, replace the 'C:\Program Files (x86)\Java' string with 'C:\Program Files\Java' in the variable value for both the environment variables.

Restart your machine. Go to 'data-integration' folder which we extracted in the Step-2 and search for the file 'spoon.bat'. Double click it and PDI will open.

You can right click the spoon.bat file and select send to --> Desktop (create shortcut). This helps you to open PDI right from the desktop instead of going to data-integration folder every time.

It may be slow to start for the first time but it will not take more than 3-4 minutes. If you feel difficulty in any of the above steps, please let me know in the comments area. Happy Learning.

Troubleshooting

For Win 7 32-bit, if PDI did not open after configuring as per step-4. Try this method.

Go to data-integration folder, select spoon.bat file. Right-click it and select rename. Change the file extension '.bat' to '.txt'. Now we can edit it in text editor like Notepad. Search for 'xmx256m' and replace it with 'xmx1024m' and save the file.

Again, change the file extension back to '.bat' from '.txt'. Now double-click the spoon.bat file. It will open PDI aka Kettle.