



1. Скачать отсюда и запустить Pentaho DI. Pentaho DI требует установку Java 8. Попробуйте скачать архив и распаковать его. Необходимо запустить spoon.sh для Linux/Mac и spoon.bat для Windows. Видео по установке Pentaho DI на примере Windows 10

Выполним

https://gist.github.com/wavezhang/ba8425f24a968ec9b2a8619d7c2d86a6?permalink_comment_id=3541602 (ссылка для скачивания Jdk)

 jdk-8u271-windows-x64.exe

 jre-8u401-windows-x64.exe

<https://www.oracle.com/java/technologies/downloads/#java8-windows> (скачивание jre)

<https://www.hitachivantara.com/pentaho/pentaho-plus-platform/data-integration-analytics/pentaho-community-edition.html>
(ссылка для скачивания Пентаго)

pdi-ce-9.4.0.0-343.zip

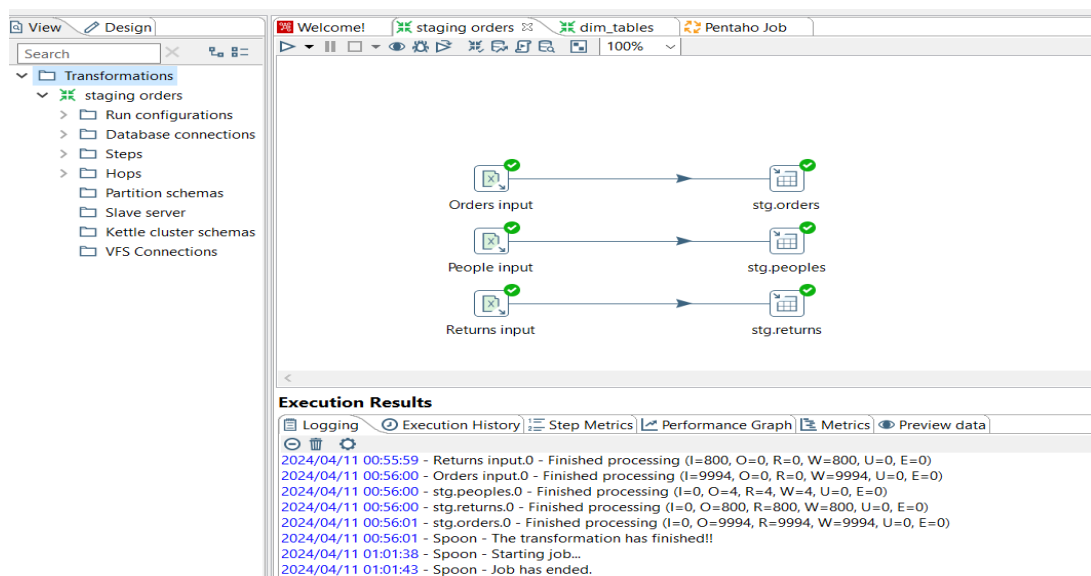
Pentaho Data Integration (Base Install)

Важно установить переменные как в инструкции

JAVA_HOME	C:\Program Files\Java\jdk1.8.0_271\
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PENTAO_JAVA	C:\Program Files\Java\jre1.8.0_271\bin\java.exe
PENTAO_JAVA_HOME	C:\Program Files\Java\jre1.8.0_271\

2. Скачать примеры Pentaho jobs для Staging и Dimension Tables.

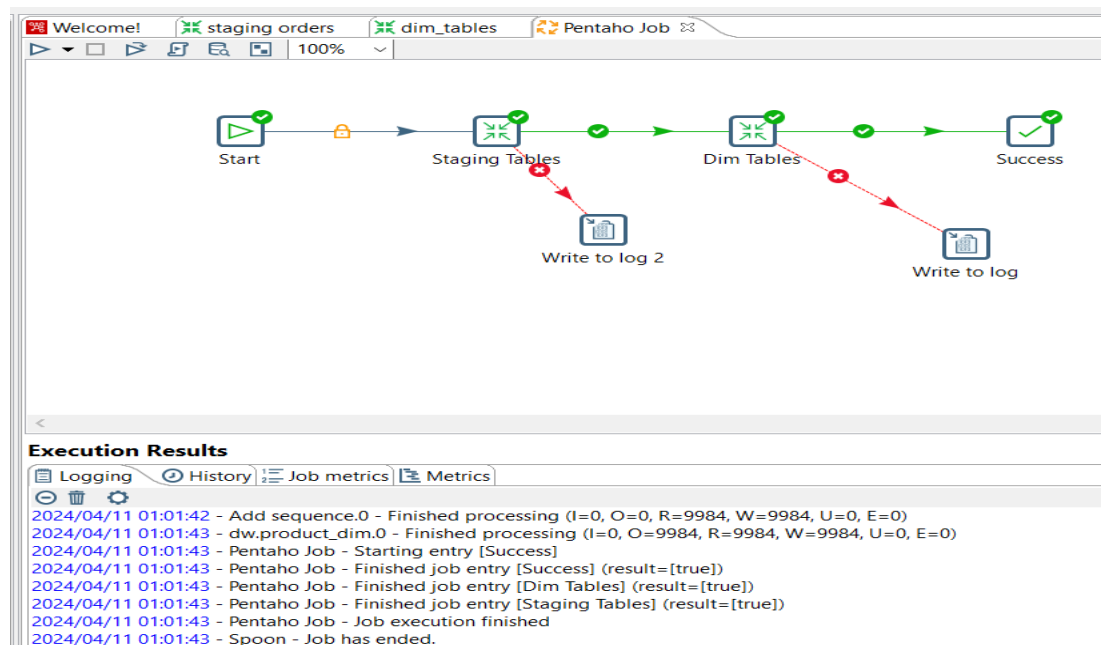
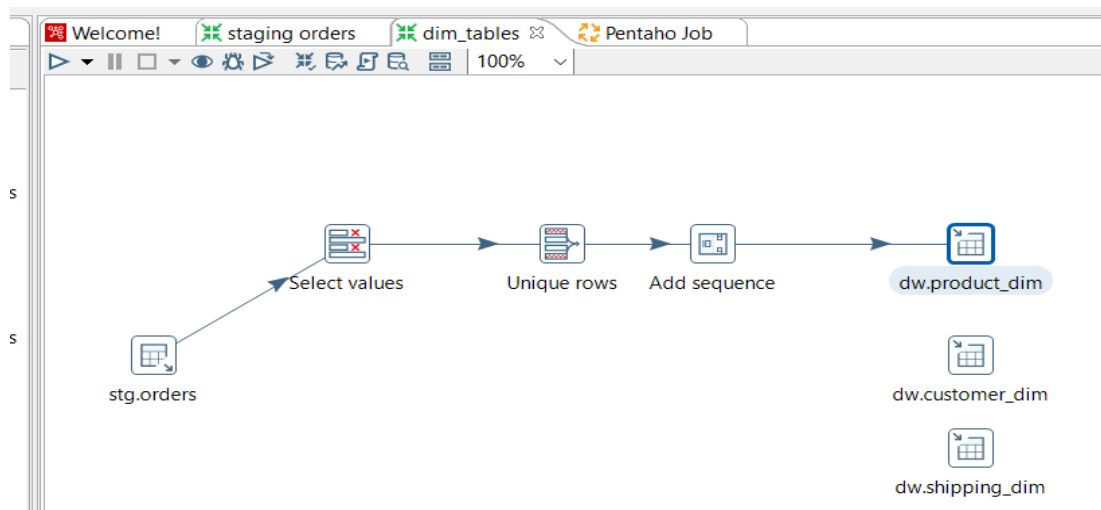
Выполним



The screenshot displays the Pentaho Data Integration (PDI) software interface. The top pane shows a job design with three input nodes (Orders input, People input, Returns input) connected to three staging table nodes (stg.orders, stg.peoples, stg.returns). The bottom pane shows the 'Execution Results' tab, which contains a log of the job's execution. The log indicates that the job was successfully completed on 2024/04/11 at 01:01:43.

Execution Results

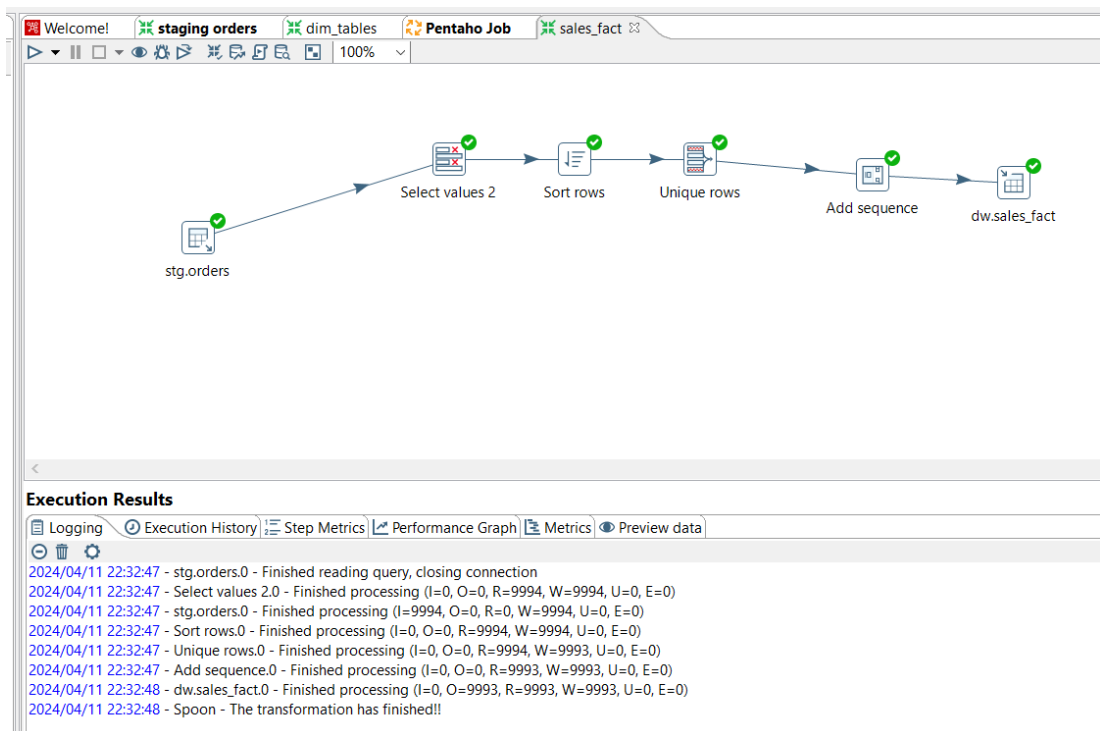
- 2024/04/11 00:55:59 - Returns input.0 - Finished processing (I=800, O=0, R=0, W=800, U=0, E=0)
- 2024/04/11 00:56:00 - Orders input.0 - Finished processing (I=9994, O=0, R=0, W=9994, U=0, E=0)
- 2024/04/11 00:56:00 - stg.peoples.0 - Finished processing (I=0, O=4, R=4, W=4, U=0, E=0)
- 2024/04/11 00:56:00 - stg.returns.0 - Finished processing (I=0, O=800, R=800, W=800, U=0, E=0)
- 2024/04/11 00:56:01 - stg.orders.0 - Finished processing (I=0, O=9994, R=9994, W=9994, U=0, E=0)
- 2024/04/11 00:56:01 - Spoon - The transformation has finished!!
- 2024/04/11 01:01:38 - Spoon - Starting job...
- 2024/04/11 01:01:43 - Spoon - Job has ended.



3. Создайте еще одну трансформацию, в которой создать

sales_fact таблицу.

Выполнить



Использовал базу данных Postgres, доступ выдан в начале курса

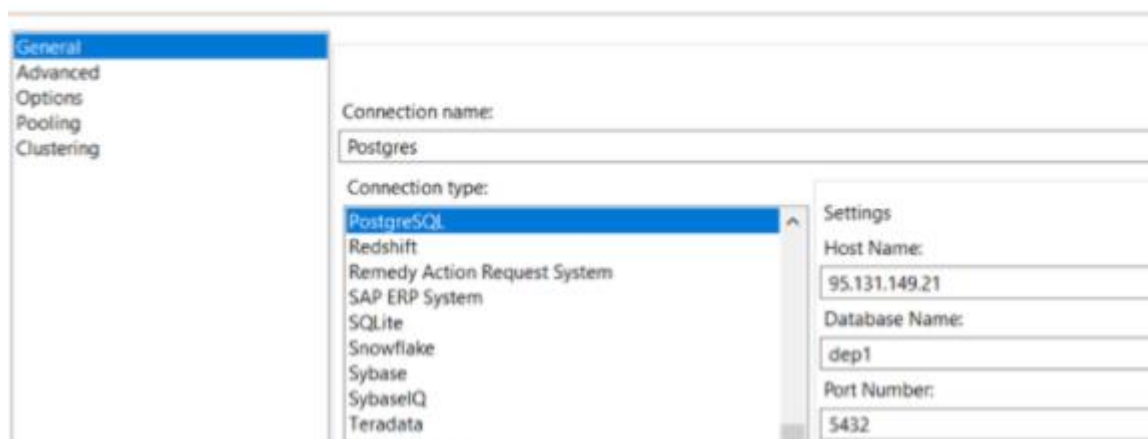
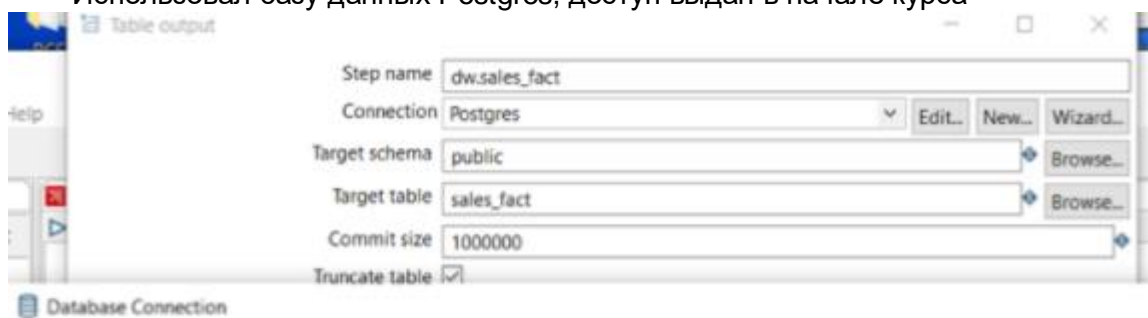


Таблица Sales_fact







	order_id	customer_id	customer_name	ship_date	ship_mode	product_id	region	country	city	state	quantity	sales	profit	discount
1	1016-102800	DP-13000	Darren Powers	16-01-07 00:00:00.000	Standard Class	OFF-PA-10000174	Central	United States	Houston	Texas	2	16,448	5,5512	
2	1016-106054	JO-15145	Jack O'Brian	16-01-07 00:00:00.000	First Class	OFF-AR-10002399	South	United States	Athens	Georgia	3	12,78	5,2398	
3	1016-112326	PO-19195	Phillina Ober	16-01-08 00:00:00.000	Standard Class	OFF-BI-10004094	Central	United States	Naperville	Illinois	2	3,54	-5,487	
4	1016-112326	PO-19195	Phillina Ober	16-01-08 00:00:00.000	Standard Class	OFF-LA-10003223	Central	United States	Naperville	Illinois	3	11,784	-4,2717	
5	1016-112326	PO-19195	Phillina Ober	16-01-08 00:00:00.000	Standard Class	OFF-ST-10002743	Central	United States	Naperville	Illinois	3	272,736	-64,7748	
6	1016-130813	LS-17230	Lycoris Saunders	16-01-08 00:00:00.000	Second Class	OFF-PA-10002005	West	United States	Los Angeles	California	3	19,44	9,3312	
7	1016-167199	ME-17320	Maria Etezadi	16-01-10 00:00:00.000	Standard Class	FUR-CH-10004063	South	United States	Henderson	Kentucky	9	2 573,82	746,4078	
8	1016-167199	ME-17320	Maria Etezadi	16-01-10 00:00:00.000	Standard Class	OFF-AR-10001662	South	United States	Henderson	Kentucky	2	5,48	1,4796	
9	1016-167199	ME-17320	Maria Etezadi	16-01-10 00:00:00.000	Standard Class	OFF-BI-10004632	South	United States	Henderson	Kentucky	2	609,98	274,491	
10	1016-167199	ME-17320	Maria Etezadi	16-01-10 00:00:00.000	Standard Class	OFF-FA-10001883	South	United States	Henderson	Kentucky	4	31,12	0,3112	
11	1016-167199	ME-17320	Maria Etezadi	16-01-10 00:00:00.000	Standard Class	OFF-PA-10000955	South	United States	Henderson	Kentucky	1	6,54	3,0084	


Main options Database fields



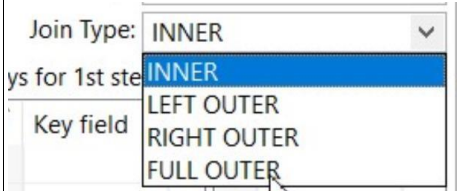

Fields to insert:

#	Table field	Stream field	
1	order_id	order_id	
2	customer_id	customer_id	
3	customer_n...	customer_na...	
4	ship_date	ship_date	
5	ship_mode	ship_mode	
6	product_id	product_id	
7	region	region	
8	country	country	
9	city	city	
1..	state	state	
1..	quantity	quantity	
1..	sales	sales	
1..	profit	profit	
1..	discount	discount	
1..	sales_PK	sales_PK	

4. Выявить 8-10 подсистем в ETL Pentaho DI и написать небольшой отчет, в котором приложить print screen компонента (ETL подсистемы) и написать про его свойства. Результат сохраните в Git.

Компонент	Свойства
<p>Select values</p>  <p>Select values</p>	<p>Выбор, удаление, переименование, изменение типов. Измените название, тип, длину и точность метаданных одного или нескольких полей.</p>
<p>Calculator</p>  <p>Calculator</p>	<p>Используется для вычислений в таблице. Доступно множество функций. Ниже на скриншоте изображены функции</p>
<p>Table input</p>  <p>stg.orders</p>	<p>Используется для считывания информации из базы данных с использованием подключения и SQL. Требуется подключение к БД.</p>
<p>Add sequence</p>  <p>Add sequence 2</p>	<p>Позволяет добавить последовательность в поток, где последовательность - это отдельное целочисленное значение с определенным начальным значением и значением приращения. Есть возможность сгенерировать последовательность из базы данных, либо с помощью. Т.е генерирование нумерации строк (последовательности чисел). Можно использовать для создания первичного ключа</p>
<p>Excel input и Text file input</p>   <p>Microsoft Excel inputText file input</p>	<p>Позволяет считывать данные из одного или нескольких файлов Excel и других форматов. Требуется указывать путь к файлу и файловой системе</p>

	<div>Select the calculation type</div> <div>Filter: <input type="text"/></div> <div>Select the calculation type to perform</div> <div><div>-</div><div>Set field to constant value A</div><div>Create a copy of field A</div><div>A + B</div><div>A - B</div><div>A * B</div><div>A / B</div><div>A * A</div><div>SQRT(A)</div><div>100 * A / B</div><div>A - (A * B / 100)</div><div>A + (A * B / 100)</div><div>A + B * C</div><div>SQRT(A*A + B*B)</div><div>ROUND(A)</div><div>ROUND(A , B)</div><div>STDROUND(A)</div><div>STDROUND(A , B)</div><div>CEIL(A)</div><div>FLOOR(A)</div><div>NVL(A , B)</div><div>Date A + B Days</div><div>Year of date A</div><div>Month of date A</div><div>Day of year of date A</div><div>Day of month of date A</div><div>Day of week of date A</div><div>Week of year of date A</div><div>ISO8601 Week of year of date A</div><div>ISO8601 Year of date A</div><div>Byte to hex encode of string A</div><div>Hex to byte decode of string A</div><div>Char to hex encode of string A</div><div>Hex to char decode of string A</div><div>Checksum of a file A using CRC-32</div><div>Checksum of a file A using Adler-32</div><div>Checksum of a file A using MD5</div><div>Checksum of a file A using SHA-1</div><div>Levenshtein Distance (source A and target B)</div></div>																		
<div>Unique rows</div> <div></div> <div>Unique rows 2</div>	<div>Позволяет удалять повторяющиеся строки из входного потока и фильтрует только уникальные строки в качестве входных данных для этого шага.</div> <div>Оставляет только уникальные строки по выбранному столбцу, группе столбцов или по всем столбцам</div> <div><div>Unique rows</div><div>Step name Unique rows</div><div>Settings<div>Add counter to output? <input type="checkbox"/> Counter field <input type="text"/></div><div>Redirect duplicate row <input type="checkbox"/> Error description <input type="text"/></div></div><div>Fields to compare on (no entries means: compare complete row)<table><tr><th>#</th><th>Fieldname</th><th>Ignore case</th></tr><tr><td>1</td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr></table></div><div><div>Help</div><div>OK</div><div>Cancel</div><div>Get</div></div></div>	#	Fieldname	Ignore case	1														
#	Fieldname	Ignore case																	
1																			

<p>Table output</p>  <p>Table output</p>	<p>Позволяет загружать данные в таблицу базы данных. Этот шаг предоставляет параметры конфигурации для целевой таблицы. Требуется настраивать подключение к БД. Также важно ставить флаг</p> <p>Truncate table <input checked="" type="checkbox"/></p> <p>Чтобы данные перезаписывались при каждом запуске</p>
<p>Merge join</p>  <p>Merge join</p>	<p>Слияние таблиц. Позволяет объединить наборы данных с данными, полученными на двух разных этапах ввода. Варианты объединения включают: INNER, LEFT OUTER, RIGHT OUTER, and FULL OUTER.</p> <p>Об этом предупреждает Pentaho. Доступны все типы соединения</p> 
<p>Group by</p>  <p>Group by</p>	<p>Позволяет группировать данные и рассчитывать значения для определенной группы полей.</p>

