

## Домашняя работа – установка Pentaho

Выполнила Вереина М.С., АДЭУ-211

The screenshot shows the 'Connection name' field set to 'Postgres'. Under 'Connection type', 'PostgreSQL' is selected from a list that includes Oracle, Oracle RDB, Palo MOLAP Server, Pentaho Data Services, PostgreSQL, Redshift, Remedy Action Request System, SAP ERP System, SQLite, Snowflake, SparkSQL, Sybase, SybaseIQ, Teradata, UniVerse database, Vertica, Vertica 5+, and dBase III, IV or 5. Under 'Access', 'Native (JDBC)' is selected from a list that also includes ODBC and JNDI. The 'Settings' panel on the right contains the following fields: 'Host Name' (localhost), 'Database Name' (postgres), 'Port Number' (5432), 'Username' (postgres), and 'Password' (masked with dots).

Рис. 1

Скачали дистрибутив PostgreSQL и установили соответствующее соединение (рис. 1)

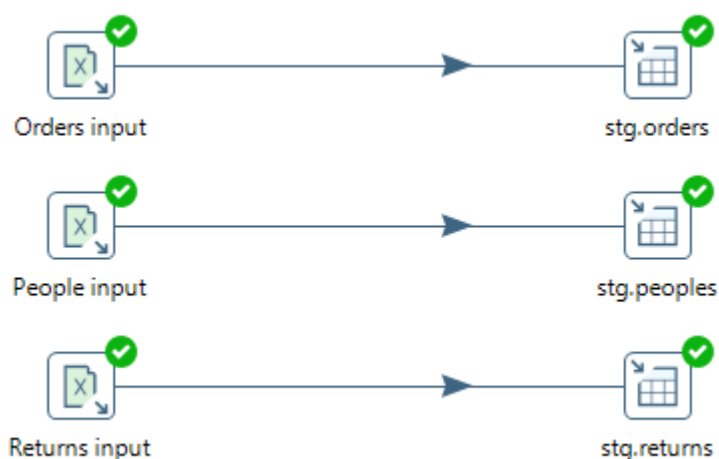


Рис. 2

Импортировали файл, в котором находились данные о заказах, людях, которые их сделали, и возвратах (рис. 2)

Далее делаем объединение всех трех датасетов с помощью компонента Join rows (Рис. 3) Видим, что среди информации о заказах появились имена покупателей

☒ First rows
 ☐ Last rows
 ☐ Off

Ship Date	Ship Mode	Customer ID	Customer Name	Segment
2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer
2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer
2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer
2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer
2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer
2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer



 Join rows

Рис. 3

## Индивидуальное задание

 CSV file input

Step name:

Filename:

Delimiter:

Рис. 1.1

Импортировали файл в рабочее пространство Pentaho DI (рис. 1.1)

#	Name	Type
1	response_date	Date
2	user_id	Integer
3	nps_rating	Integer
4	source	String
5	nps_group	String

Рис. 2.1

Изменили тип данных для правильной сортировки (рис. 2.1)

#	response_date	user_id	nps_rating	source	nps_group
1	Fri Nov 06 00:00:00 MSK 2020	11037	7	email	passive
2	Thu Dec 24 00:00:00 MSK 2020	34434	9	email	promoter
3	Thu Dec 03 00:00:00 MSK 2020	49547	8	email	passive
4	Sun Oct 04 00:00:00 MSK 2020	13821	7	email	passive
5	Fri Oct 23 00:00:00 MSK 2020	29407	9	email	promoter
6	Thu Dec 03 00:00:00 MSK 2020	13041	10	email	promoter
7	Mon Oct 12 00:00:00 MSK 2020	27166	8	email	passive

Рис. 3.1

Предпросмотр получившегося датасета (рис. 3.1)

Далее отсортировали «чистые» данные по возрастанию рейтинга клиента (рис. 4.1)

Fields :							
#	Fieldname	Ascending	#	response_date	user_id	nps_rating	nps_group
1	nps_rating	Y	1	Mon Dec 14 00:00:00 MSK 2020	38198	0	detractor
			2	Thu Oct 22 00:00:00 MSK 2020	24825	0	detractor
			3	Wed Nov 04 00:00:00 MSK 2020	44531	0	detractor
			4	Mon Dec 28 00:00:00 MSK 2020	35548	0	detractor
			5	Tue Nov 24 00:00:00 MSK 2020	18867	0	detractor
			6	Thu Dec 10 00:00:00 MSK 2020	12202	0	detractor
			7	Sat Nov 21 00:00:00 MSK 2020	21774	0	detractor

Рис. 4.1


		response_date	user_id	nps_rating	source	nps_group
		14.12.2020	38 198,00	0	email	detractor
		22.10.2020	24 825,00	0	email	detractor
		04.11.2020	44 531,00	0	email	detractor
		28.12.2020	35 548,00	0	email	detractor
		24.11.2020	18 867,00	0	email	detractor
		10.12.2020	12 202,00	0	email	detractor

Рис. 5.1

Выгрузка полученного датасета в формате Excel в выбранный репозиторий