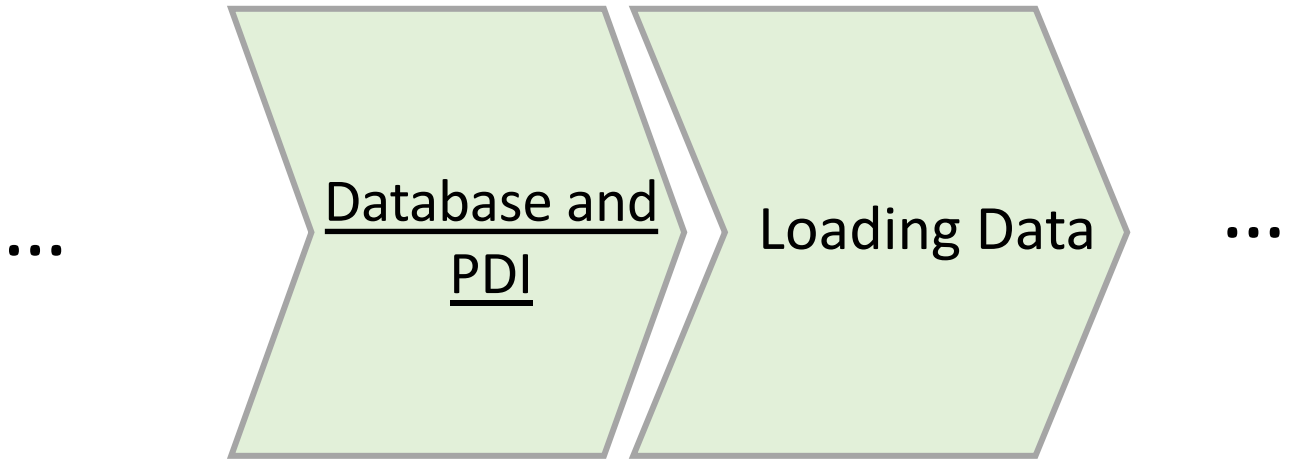
DATA ANALYTICS
(Data Warehouse)
Pentaho Data Integration

Luca Cinelli, PhD

luca.cinelli@unical.it

Outline



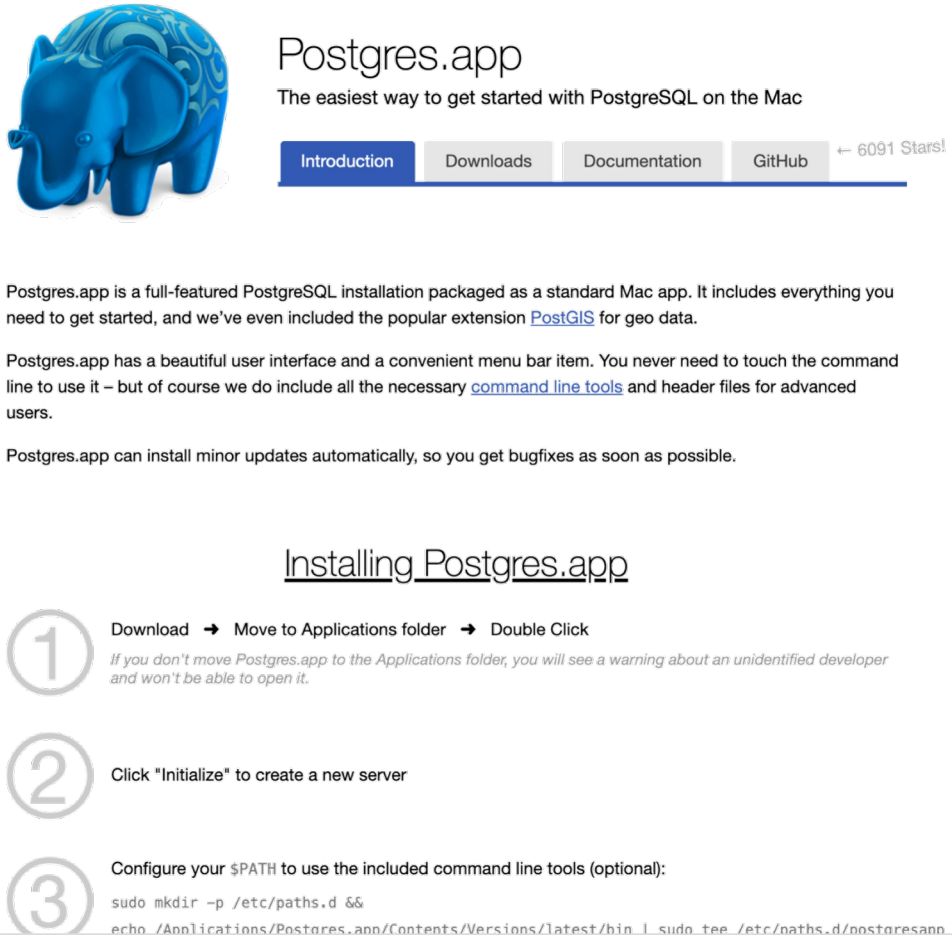
Setup PostgreSQL

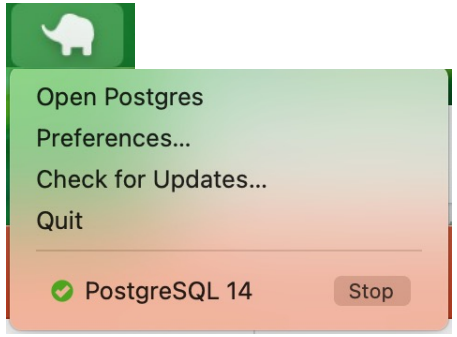
- https://www.postgresql.org/

- Database engine
- Graphical Interface: PGAdmin

Setup PostgreSQL

- https://www.postgresapp.com



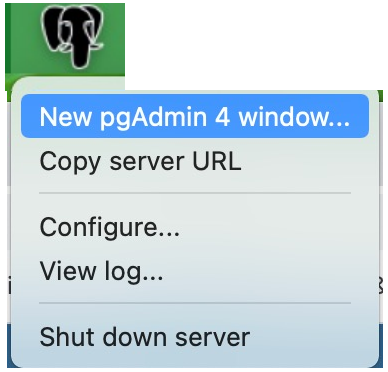


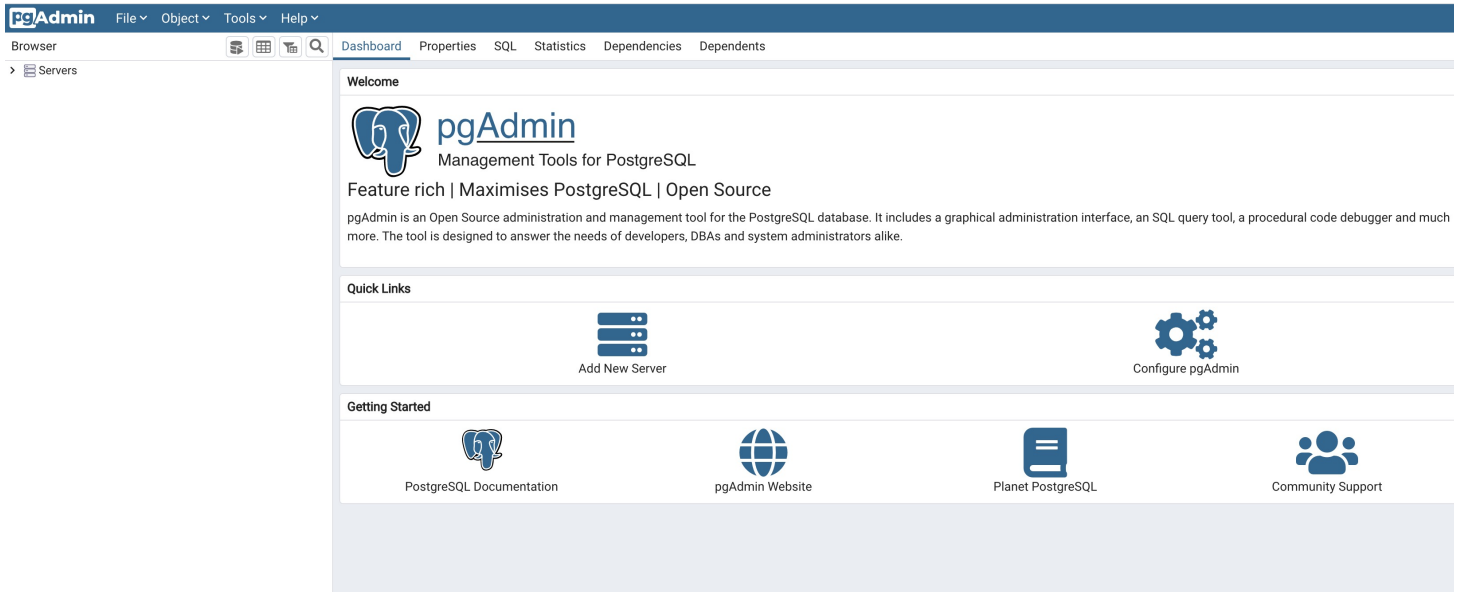
- Database engine
- Graphical Interface: PGAdmin

Setup PostgreSQL

- https://www.postgresapp.com

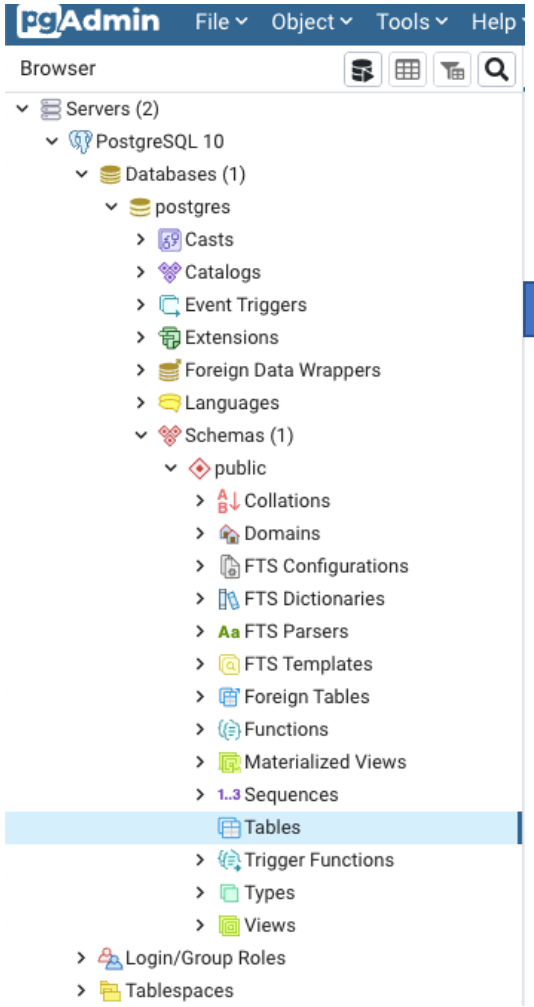
- Database engine
- Graphical Interface: PGAdmin (http://127.0.0.1:51765/browser/)

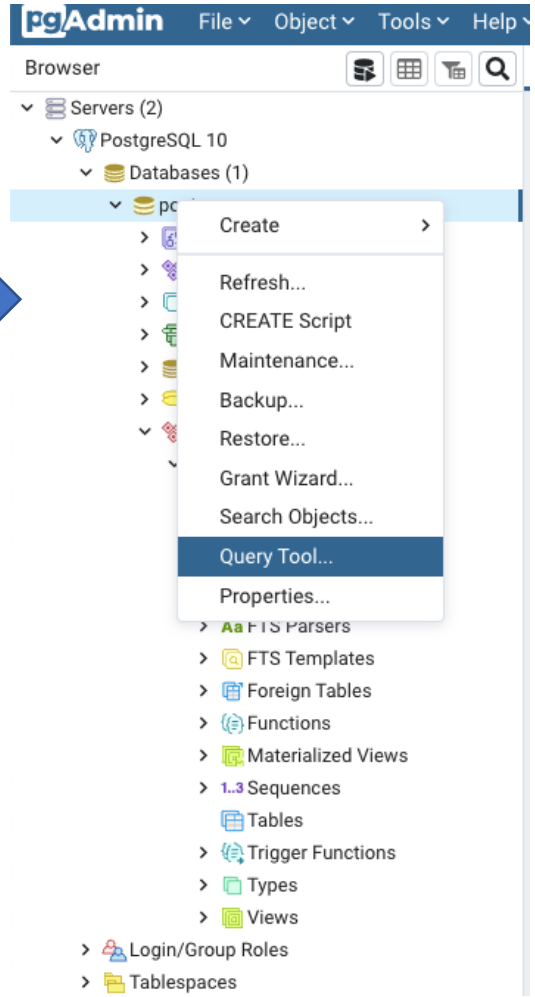


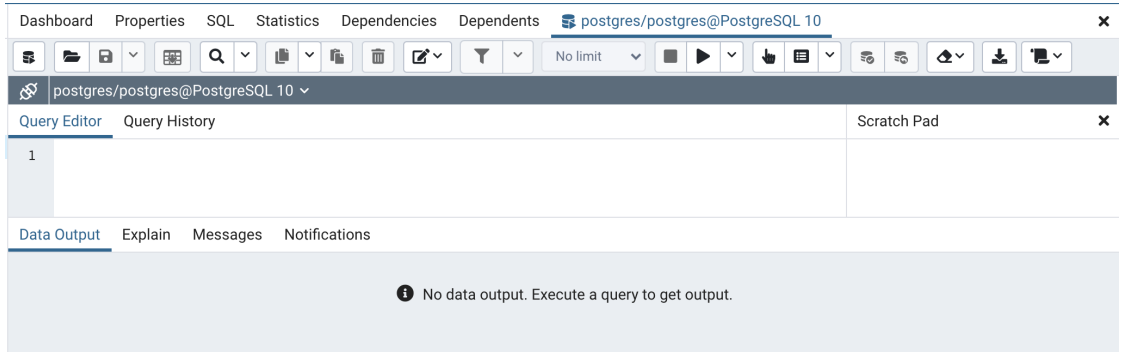


Creating tables in PostgreSQL

http://127.0.0.1:60624/browser/

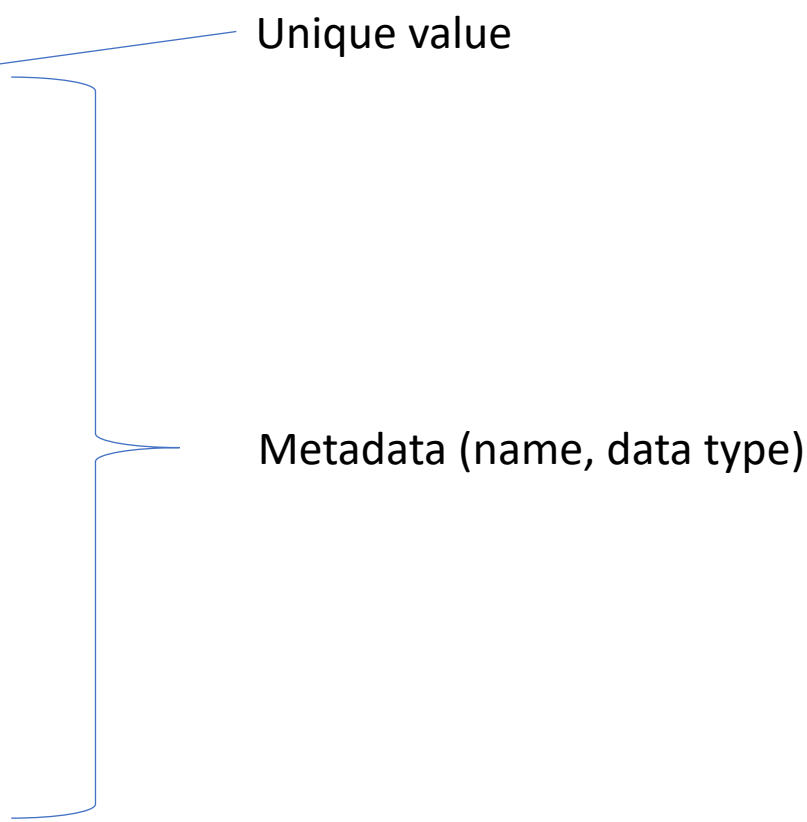




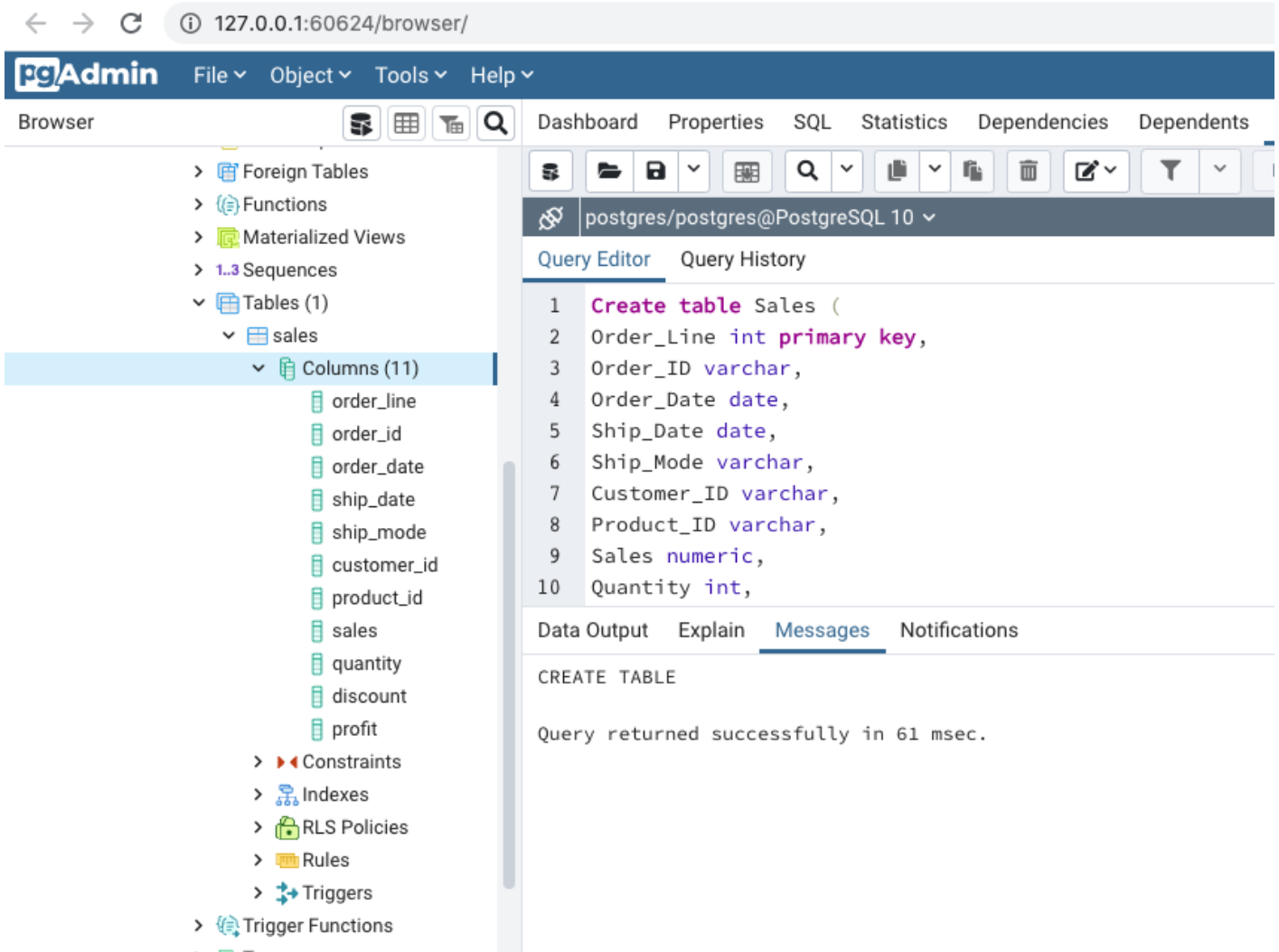


Creating tables in PostgreSQL – Sales Table schema example

```
Create table Sales (
Order_Line int primary key,
Order_ID varchar,
Order_Date date,
Ship_Date date,
Ship_Mode varchar,
Customer_ID varchar,
Product_ID varchar,
Sales numeric,
Quantity int,
Discount numeric,
Profit numeric
);
```



Creating tables in PostgreSQL – Sales Table schema example

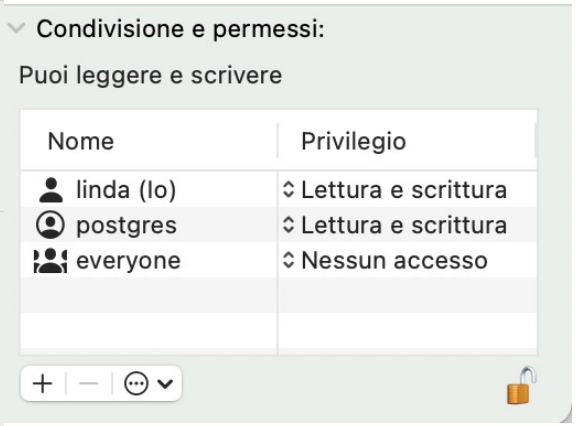


Creating tables in PostgreSQL – Sales Table schema example

Input

SalesforSQL

Order Line	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Product ID	Sales	Quantity	Discount	Profit
7973	CA-2017-166142	2020-01-01	2020-01-05	Standard Class	MM-17260	OFF-BI-10004094	26.55	3	0	13.01
7974	CA-2017-166142	2020-01-01	2020-01-05	Standard Class	MM-17260	FUR-TA-10004607	310.44	3	0.3	-48.78
7975	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	OFF-AR-10001177	6.56	2	0	1.9
7976	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	OFF-LA-10000634	7.83	3	0	3.6
7977	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	TEC-AC-10004708	41.9	2	0	8.8
7978	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	FUR-CH-10002084	664.15	6	0.1	88.55
7979	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	OFF-PA-10004519	8.96	2	0	4.39
7980	US-2017-128951	2020-01-01	2020-01-03	First Class	RS-19420	OFF-AP-10002191	179.94	3	0	50.38
7981	US-2017-128951	2020-01-01	2020-01-03	First Class	RS-19420	FUR-TA-10004575	872.94	3	0	157.13
7982	US-2017-128951	2020-01-01	2020-01-03	First Class	RS-19420	OFF-PA-10003177	12.96	2	0	6.22
7983	US-2017-156909	2020-01-02	2020-01-04	Second Class	SF-20065	FUR-CH-10002774	71.37	2	0.3	-1.02
7984	CA-2017-109778	2020-01-02	2020-01-07	Standard Class	VM-21685	OFF-AR-10003759	2.91	2	0.2	0.91
7985	US-2017-152842	2020-01-02	2020-01-09	Standard Class	NF-18385	FUR-CH-10004218	242.35	3	0.2	15.15
7986	CA-2017-139948	2020-01-03	2020-01-08	Standard Class	SW-20455	FUR-FU-10002597	7.9	2	0.2	2.17
7987	US-2017-105046	2020-01-03	2020-01-09	Standard Class	BE-11335	TEC-PH-10004536	269.98	2	0	67.5
7988	US-2017-105046	2020-01-03	2020-01-09	Standard Class	BE-11335	OFF-PA-10004353	99.9	5	0	47.95
7989	US-2017-105046	2020-01-03	2020-01-09	Standard Class	BE-11335	FUR-FU-10004848	39.08	4	0	14.46
7990	CA-2017-126662	2020-01-03	2020-01-07	Standard Class	AB-10255	TEC-CO-10004202	479.98	2	0.2	90
7991	CA-2017-142342	2020-01-03	2020-01-05	Second Class	AJ-10795	OFF-PA-10004609	32.4	5	0	15.55
7992	CA-2017-142342	2020-01-03	2020-01-05	Second Class	AJ-10795	OFF-EN-10002592	57.9	5	0	28.95
7993	CA-2017-142342	2020-01-03	2020-01-05	Second Class	AJ-10795	OFF-ST-10002957	10.56	2	0	0
7994	CA-2017-142342	2020-01-03	2020-01-05	Second Class	AJ-10795	FUR-BO-10002613	1194.17	5	0.15	210.74
7995	US-2017-142573	2020-01-04	2020-01-09	Standard Class	ML-17410	FUR-TA-10001932	801.6	5	0.5	-448.9
7996	US-2017-142573	2020-01-04	2020-01-09	Standard Class	ML-17410	FUR-CH-10004218	161.57	2	0.2	10.1
7997	US-2017-142573	2020-01-04	2020-01-09	Standard Class	ML-17410	OFF-PA-10000246	16.1	2	0.2	5.23
7998	US-2017-142573	2020-01-04	2020-01-09	Standard Class	ML-17410	OFF-BI-10003350	7.66	4	0.7	-6.12



Copy the input in the folder:
PROGRAMS/PostgreSQL/NN/**data/dataset**
Where:
- PROGRAMS is where was installed PostgresSQL
- NN is the PostgresSQL version
Create the subfolders data/datasets if don't exist

Creating tables in PostgreSQL – Sales Table data example

- SalesforSQL.csv
- Change the location as per your installation directory
- COPY sales from '/Library/PostgreSQL/14/**data/dataset**/SalesforSQL.csv' delimiter ',' csv header;
  - Example Linux: COPY sales from '/var/lib/postgresql/13/data/dataset/SalesforSQL.csv' delimiter ','csv header;
  - Example Windows: COPY sales from 'C:\Program Files\PostgreSQL\14\data\dataset\SalesforSQL.csv' delimiter ',' csv header;
  - Select and run only this line

- To Check if the data has been correctly imported, run the select command
  - SELECT * FROM sales;

Creating tables in PostgreSQL – Sales Table data example

Query Editor | Query History

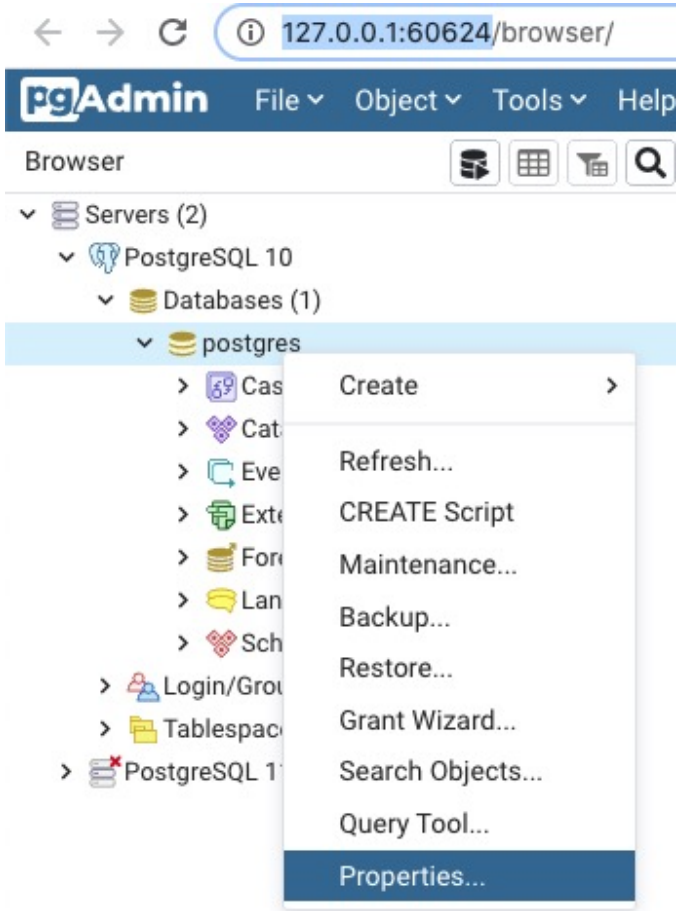
```
1  Create table Sales (
2  Order_Line int primary key,
3  Order_ID varchar,
4  Order_Date date,
5  Ship_Date date,
6  Ship_Mode varchar,
7  Customer_ID varchar,
8  Product_ID varchar,
9  Sales numeric,
10 Quantity int,
11 Discount numeric,
12 Profit numeric
13 );
14
15
16 COPY sales from '/Library/PostgreSQL/10/data/dataset/SalesforSQL.csv' delimiter ',' csv header;
17
18 SELECT * FROM sales;
19
```

Data Output | Explain | Messages | Notifications

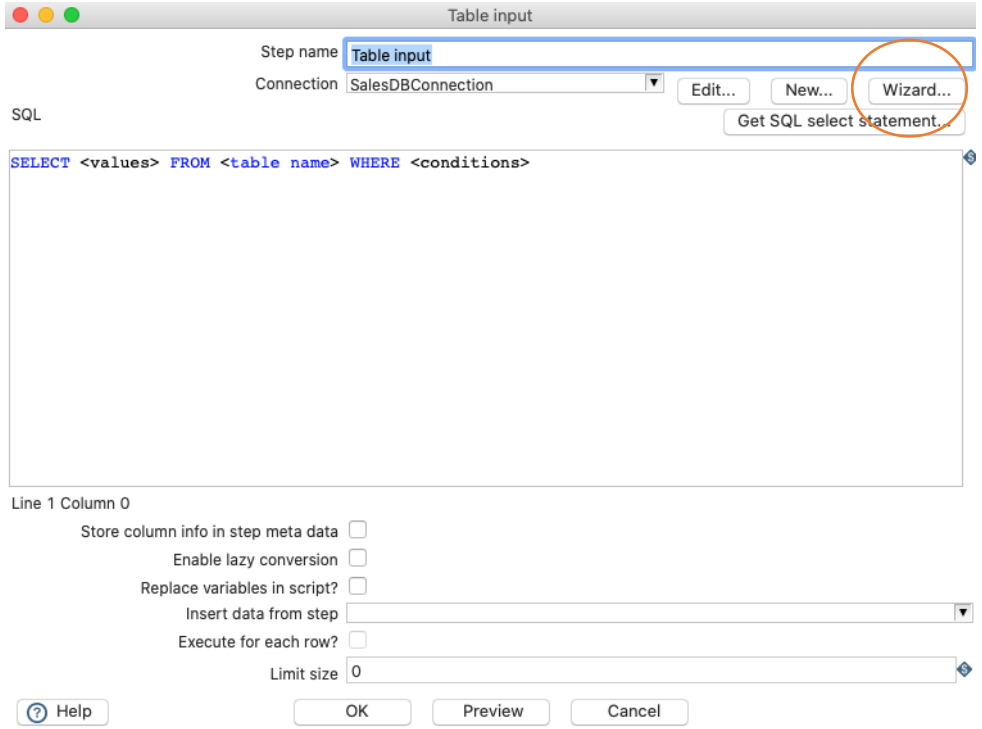
	order_line [PK] integer	order_id character varying	order_date date	ship_date date	ship_mode character varying	customer_id character varying	product_id character varying	sales numeric	quantity integer	discount numeric	profit numeric
1	7973	CA-2017-166142	2020-01-01	2020-01-05	Standard Class	MM-17260	OFF-BI-10004094	26.55	3	0	13.01
2	7974	CA-2017-166142	2020-01-01	2020-01-05	Standard Class	MM-17260	FUR-TA-10004607	310.44	3	0.3	-48.78
3	7975	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	OFF-AR-10001177	6.56	2	0	1.9
4	7976	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	OFF-LA-10000634	7.83	3	0	3.6
5	7977	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	TEC-AC-10004708	41.9	2	0	8.8
6	7978	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	FUR-CH-10002084	664.15	6	0.1	88.55
7	7979	CA-2017-164378	2020-01-01	2020-01-04	Second Class	MM-18055	OFF-PA-10004519	8.96	2	0	4.39
8	7980	US-2017-128951	2020-01-01	2020-01-03	First Class	RS-19420	OFF-AP-10002191	179.94	3	0	50.38
9	7981	US-2017-128951	2020-01-01	2020-01-03	First Class	RS-19420	FUR-TA-10004575	872.94	3	0	157.13

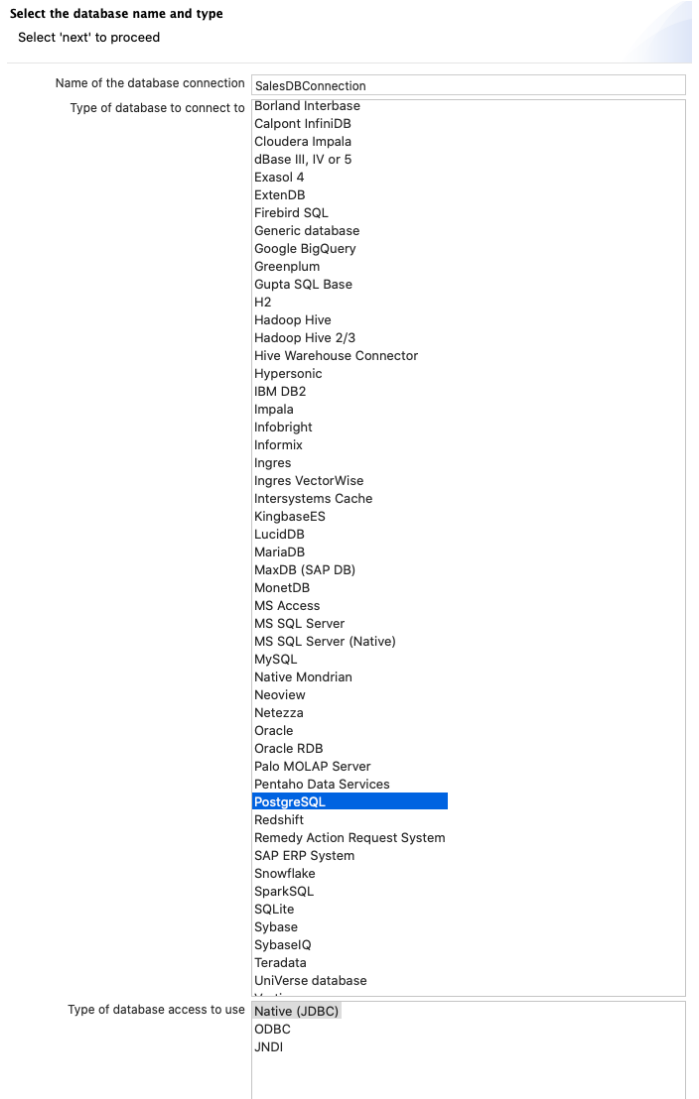
Connect to database: information

- **Host server address/location**: 127.0.0.1 take a look to your url http://127.0.0.1:60624/
- **Name database**: postgres
- **Password**
- **User**: postgres

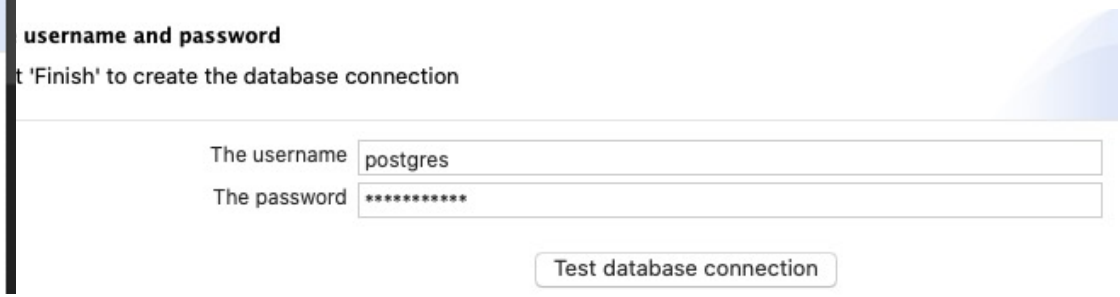


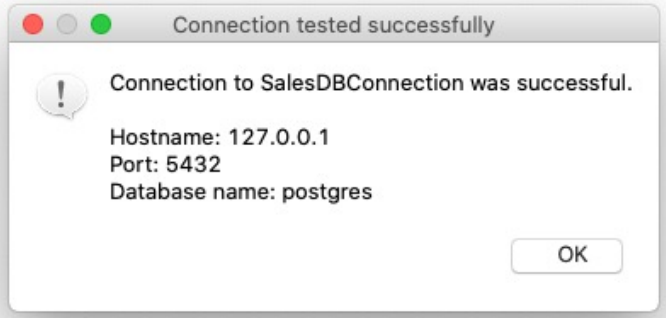
Import data from SQL database in PDI





Import data from SQL database in PDI





Query data from SQL database in PDI

- SELECT <values> FROM <table name> WHERE <conditions>

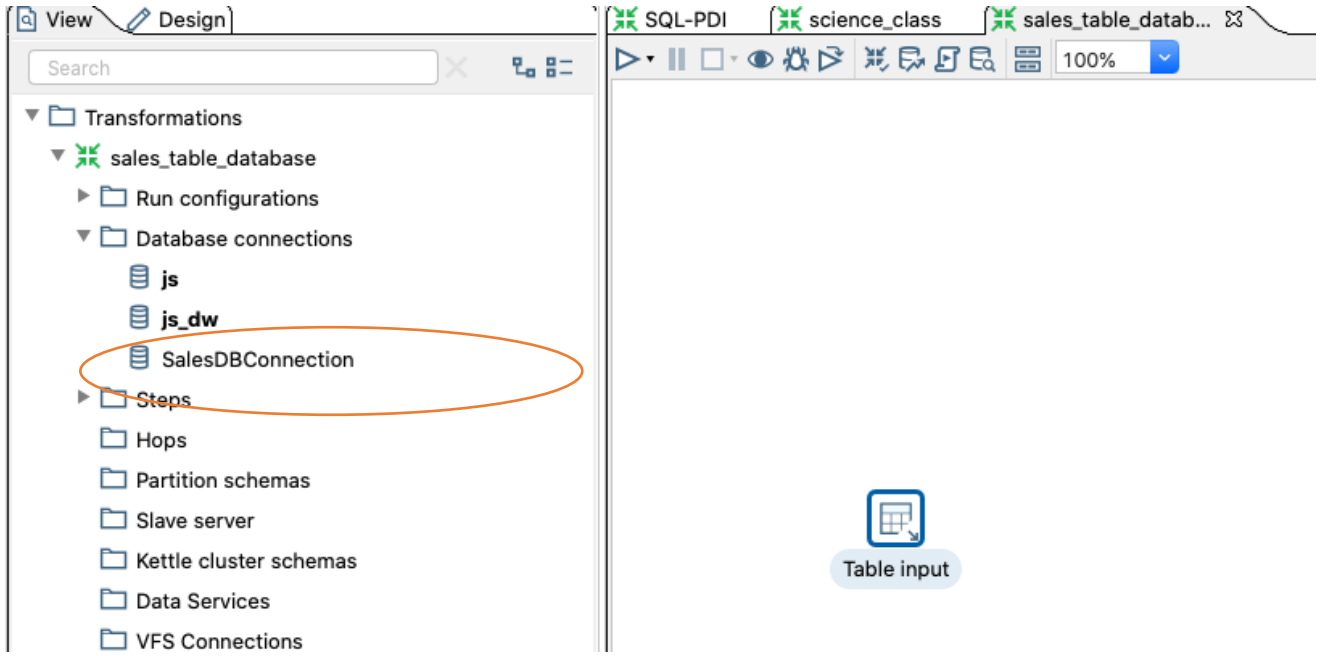
- SELECT * FROM sales

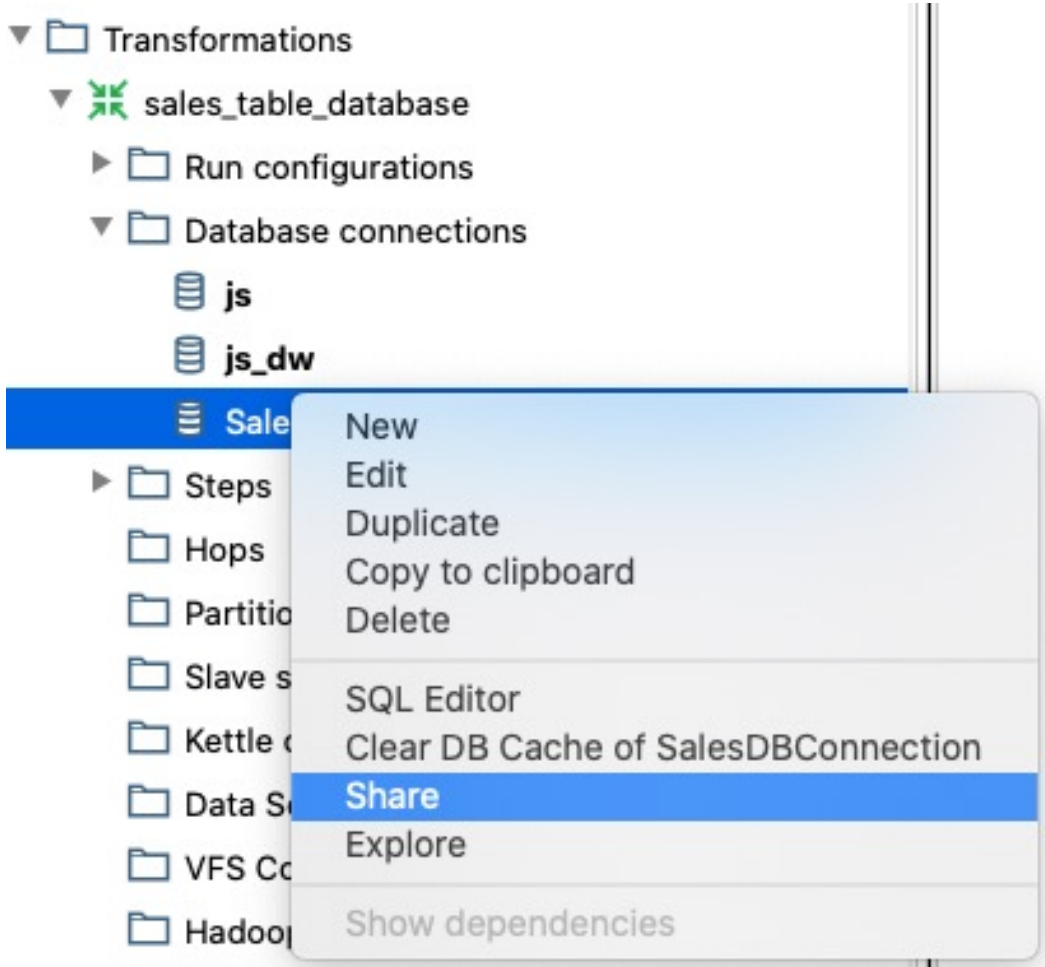
sales_table_database.ktr

Rows of step: Table input (1000 rows)

#	order_line	order_id	order_date	ship_date	ship_mode	customer_id	product_id	sales	quantity	discount	profit
1	7973	CA-2017-166...	2020/01/01 00:00:0...	2020/01/05 00:00:0...	Standard Class	MM-17260	OFF-BI-10004094	26.55	3	0.0	13.01
2	7974	CA-2017-166...	2020/01/01 00:00:0...	2020/01/05 00:00:0...	Standard Class	MM-17260	FUR-TA-10004607	310.44	3	0.3	-48.78
3	7975	CA-2017-164...	2020/01/01 00:00:0...	2020/01/04 00:00:0...	Second Class	MM-18055	OFF-AR-10001...	6.56	2	0.0	1.9
4	7976	CA-2017-164...	2020/01/01 00:00:0...	2020/01/04 00:00:0...	Second Class	MM-18055	OFF-LA-10000634	7.83	3	0.0	3.6
5	7977	CA-2017-164...	2020/01/01 00:00:0...	2020/01/04 00:00:0...	Second Class	MM-18055	TEC-AC-10004...	41.9	2	0.0	8.8
6	7978	CA-2017-164...	2020/01/01 00:00:0...	2020/01/04 00:00:0...	Second Class	MM-18055	FUR-CH-10002...	664.15	6	0.1	88.55
7	7979	CA-2017-164...	2020/01/01 00:00:0...	2020/01/04 00:00:0...	Second Class	MM-18055	OFF-PA-100045...	8.96	2	0.0	4.39
8	7980	US-2017-128...	2020/01/01 00:00:0...	2020/01/03 00:00:0...	First Class	RS-19420	OFF-AP-100021...	179.94	3	0.0	50.38
9	7981	US-2017-128...	2020/01/01 00:00:0...	2020/01/03 00:00:0...	First Class	RS-19420	FUR-TA-10004575	872.94	3	0.0	157.13
10	7982	US-2017-128...	2020/01/01 00:00:0...	2020/01/03 00:00:0...	First Class	RS-19420	OFF-PA-100031...	12.96	2	0.0	6.22
11	7983	US-2017-156...	2020/01/02 00:00:0...	2020/01/04 00:00:0...	Second Class	SF-20065	FUR-CH-10002...	71.37	2	0.3	-1.02
12	7984	CA-2017-109...	2020/01/02 00:00:0...	2020/01/07 00:00:0...	Standard Class	VM-21685	OFF-AR-10003...	2.91	2	0.2	0.91
13	7985	US-2017-152...	2020/01/02 00:00:0...	2020/01/09 00:00:0...	Standard Class	NF-18385	FUR-CH-10004...	242.35	3	0.2	15.15
14	7986	CA-2017-139...	2020/01/03 00:00:0...	2020/01/08 00:00:0...	Standard Class	SW-20455	FUR-FU-10002...	7.9	2	0.2	2.17
15	7987	US-2017-105...	2020/01/03 00:00:0...	2020/01/09 00:00:0...	Standard Class	BE-11335	TEC-PH-10004...	269.98	2	0.0	67.5
16	7988	US-2017-105...	2020/01/03 00:00:0...	2020/01/09 00:00:0...	Standard Class	BE-11335	OFF-PA-100043...	99.9	5	0.0	47.95
17	7989	US-2017-105...	2020/01/03 00:00:0...	2020/01/09 00:00:0...	Standard Class	BE-11335	FUR-FU-10004...	39.08	4	0.0	14.46
18	7990	CA-2017-126...	2020/01/03 00:00:0...	2020/01/07 00:00:0...	Standard Class	AB-10255	TEC-CO-10004...	479.98	2	0.2	90.0
19	7991	CA-2017-142...	2020/01/03 00:00:0...	2020/01/05 00:00:0...	Second Class	AJ-10795	OFF-PA-100046...	32.4	5	0.0	15.55
20	7992	CA-2017-142...	2020/01/03 00:00:0...	2020/01/05 00:00:0...	Second Class	AJ-10795	OFF-EN-10002...	57.9	5	0.0	28.95
21	7993	CA-2017-142...	2020/01/03 00:00:0...	2020/01/05 00:00:0...	Second Class	AJ-10795	OFF-ST-10002957	10.56	2	0.0	0.0
22	7994	CA-2017-142...	2020/01/03 00:00:0...	2020/01/05 00:00:0...	Second Class	AJ-10795	FUR-BO-10002...	1194.17	5	0.15	210.74
23	7995	US-2017-142...	2020/01/04 00:00:0...	2020/01/09 00:00:0...	Standard Class	ML-17410	FUR-TA-10001932	801.6	5	0.5	-448.9
24	7996	US-2017-142...	2020/01/04 00:00:0...	2020/01/09 00:00:0...	Standard Class	ML-17410	FUR-CH-10004...	161.57	2	0.2	10.1
25	7997	US-2017-142...	2020/01/04 00:00:0...	2020/01/09 00:00:0...	Standard Class	ML-17410	OFF-PA-100002...	16.1	2	0.2	5.23
26	7998	US-2017-142...	2020/01/04 00:00:0...	2020/01/09 00:00:0...	Standard Class	ML-17410	OFF-BI-10003350	7.66	4	0.7	-6.12
27	7999	US-2017-142	2020/01/04 00:00:0	2020/01/09 00:00:0	Standard Class	ML-17410	FUR-CH-10000	311.98	3	0.2	-42.9

Global database connection





SQL-PDI connection: a simple example

1. Create a new table in postgresqlby running this command in PGAdmin

```
create table science_class(
Enrollment_no INT,
Name VARCHAR,
Science_Marks INT
);
```

2. Insert some sample values

```
insert into science_class values (1,'Popeye',33);
insert into science_class values (2,'Olive',54);
insert into science_class values (3,'Brutus',98);
```

3. Query: select * from science_class

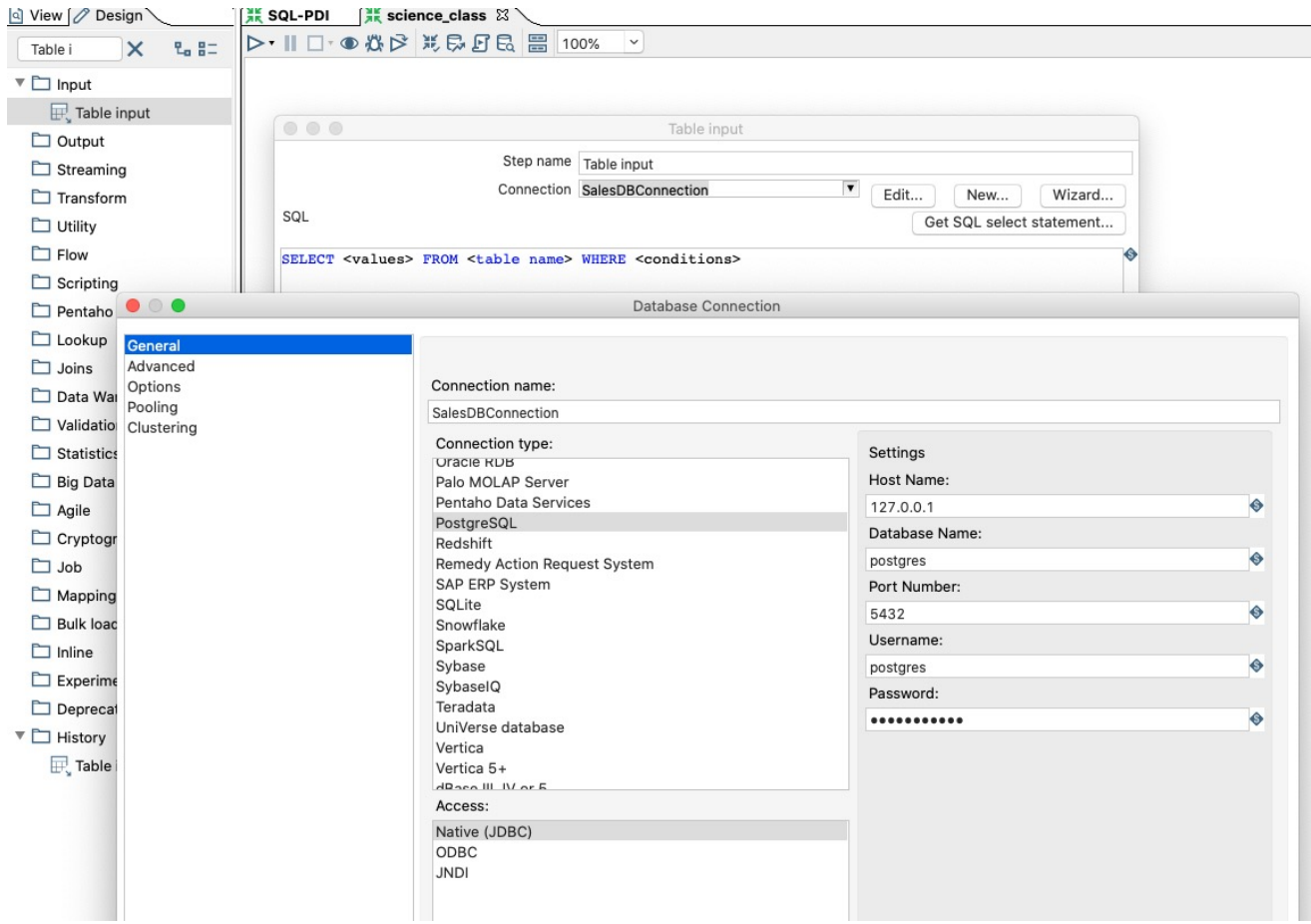
SQL-PDI connection: task examples

1. **Read:**
   - Retrieve all data from the table ‘Science_Class’
2. **Read with conditions**
   - Retrieve the name of students who have scored more than 60 marks
3. **Update**
   - Update the marks of Popeye to 45
4. **Insert**
   - Insert a new row with “Wimpy” who has scored 75 marks
5. **Delete**
   - Delete the record of Wimpy

SQL-PDI connection: Read example

Retrieve all data from the table ‘Science_Class’

- Input table -> reuse database connection
- SELECT * FROM science_class



SQL-PDI connection: Read with conditions example

- Retrieve the name of students who have scored more than 60 marks
- SELECT * FROM science_class WHERE science_mark > 60

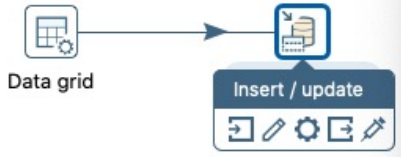
enrollment_no	name	science_marks
3	Brutus	98

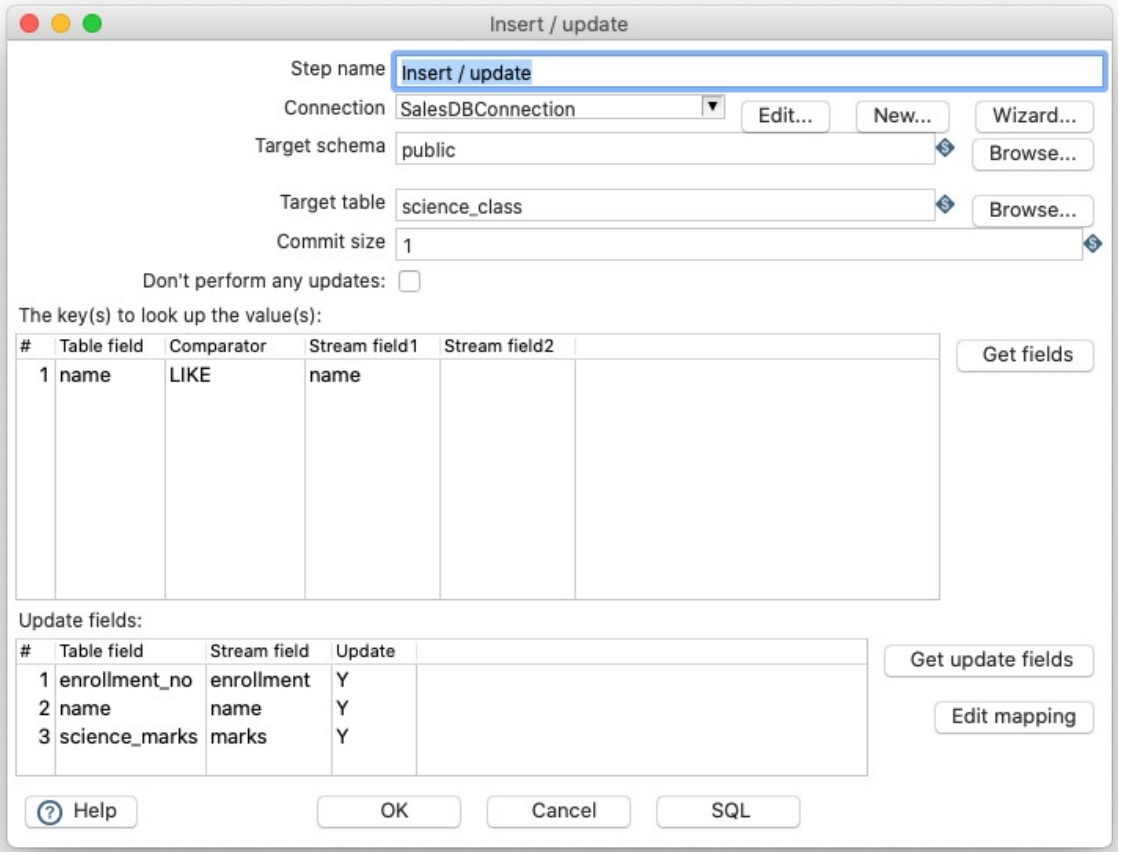
science_class_read.ktr

SQL-PDI connection: Update and Insert example

- Update the marks of Popeye to 45
- Insert a new row with “Wimpy” who has scored 75 marks

science_class_update_insert.ktr





Data Output | Explain | Messages | Notifications

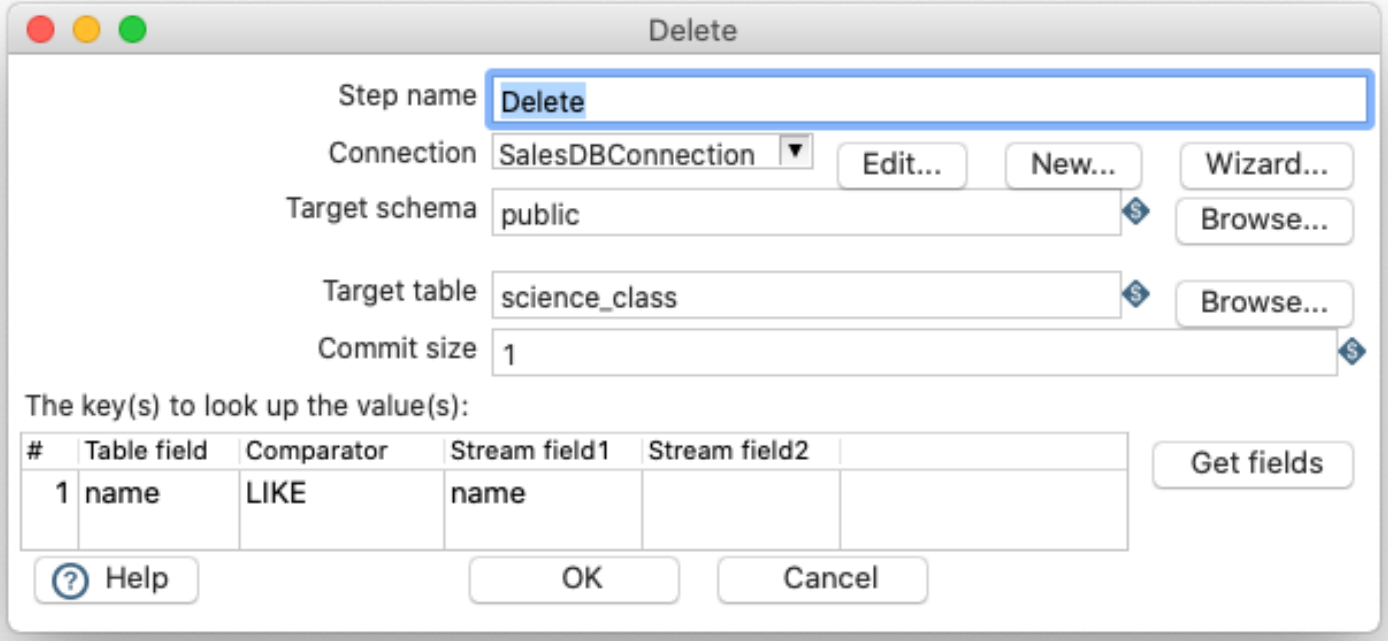
	enrollment_no integer	name character varying	science_marks integer
1	1	Popeye	33
2	2	Olive	54
3	3	Brutus	98

	enrollment_no integer	name character varying	science_marks integer
1	2	Olive	54
2	3	Brutus	98
3	1	Popeye	45
4	4	Wimpy	75

SQL-PDI connection: Delete example

- Delete the record of Wimpy

	enrollment_no integer	name character varying	science_marks integer
1	2	Olive	54
2	3	Brutus	98
3	1	Popeye	45



science_class_delete.ktr