TÉCNICO LISBOA

Análise e Integração de Dados

Relatório de Projeto

Grupo 25

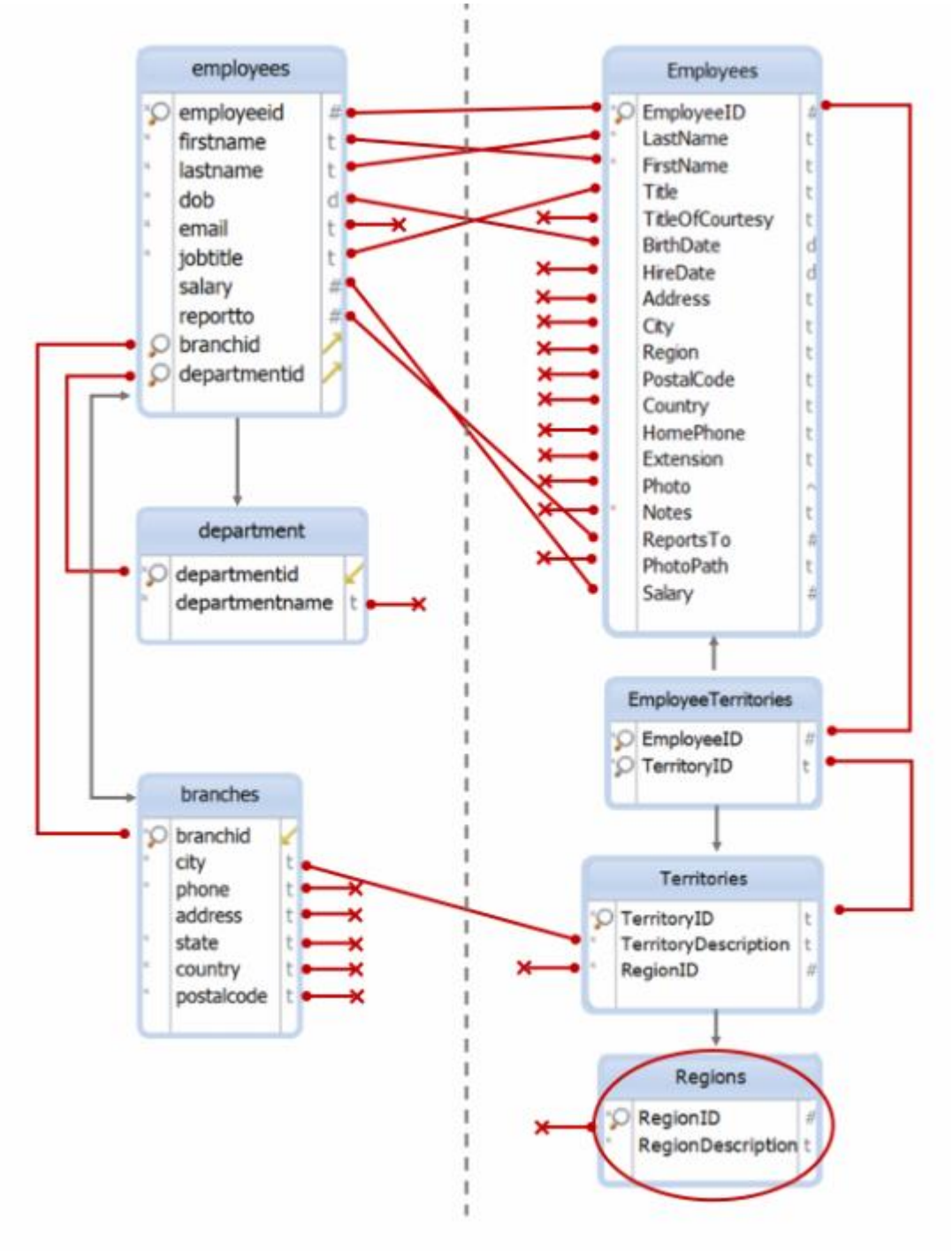
76991 - Rúben Vines
78328 - Pedro Duarte
79175 - Beatriz Portugal

Table of Contents

1. Data Integration

1.1. Schema Matching

Company Tables Northwind Tables



1.2. Common Mediated Schema

After identifying the employee related common fields in the both databases, views were created for each company.

“Northwind” & “Company” views over all its employees:

northwind_employees(employeeid, lastname, firstname, title, birthdate, reportto, salary, territoryid, city);
company_employees(employeeid, lastname, firstname, title, birthdate, reportto, salary, branchid, city);

Common Mediated Schema:
all_employees(employeeid, firstname, lastname, jobtitle, dob, reportto, salary, branchid, city);

Note: ‘city’ field in both cases is referent to the employee’s current workplace address, not its home address. In the case of *northwind*’s company, one employee can be registered in many different cities at the same time.

1.3. Schema Mapping

northwind_employees view creation:

```
create or replace view northwind_employees(employeeid,
lastname, firstname, title, birthdate, reportto, salary,
territoryid, city) as
     (select e.EmployeeID, e.LastName, e.firstName, e.Title,
e.BirthDate, e.ReportsTo, e.Salary, et.TerritoryID,
t.TerritoryDescription
     from Employees as e, EmployeeTerritories as et,
Territories as t
     where e.EmployeeID = et.EmployeeID and t.TerritoryID =
et.TerritoryID);
```

northwind_employees resulting table:

employeeid	lastname	firstname	title	birthdate	reportto	salary	territoryid	city
1	Davolio	Nancy	Sales Representative	1948-12-08 00:00:00	2	2954.55	06897	Wilton
1	Davolio	Nancy	Sales Representative	1948-12-08 00:00:00	2	2954.55	19713	Neward
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	01581	Westboro
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	01730	Bedford
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	01833	Georgetow
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	02116	Boston
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	02139	Cambridge
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	02184	Braintree
2	Fuller	Andrew	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	40222	Louisville
3	Leverling	Janet	Sales Representative	1963-08-30 00:00:00	2	3119.15	30346	Atlanta
3	Leverling	Janet	Sales Representative	1963-08-30 00:00:00	2	3119.15	31406	Savannah
3	Leverling	Janet	Sales Representative	1963-08-30 00:00:00	2	3119.15	32859	Orlando
3	Leverling	Janet	Sales Representative	1963-08-30 00:00:00	2	3119.15	33607	Tampa
4	Peacock	Margaret	Sales Representative	1937-09-19 00:00:00	2	1861.08	20852	Rockville
4	Peacock	Margaret	Sales Representative	1937-09-19 00:00:00	2	1861.08	27403	Greensboro
4	Peacock	Margaret	Sales Representative	1937-09-19 00:00:00	2	1861.08	27511	Cary
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	02903	Providence
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	07960	Morristown
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	08837	Edison
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	10019	New York
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	10038	New York
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	11747	Mellvile
5	Buchanan	Steven	Sales Manager	1955-03-04 00:00:00	2	1744.21	14450	Fairport
6	Suyama	Michael	Sales Representative	1963-07-02 00:00:00	5	2004.07	85014	Phoenix
6	Suyama	Michael	Sales Representative	1963-07-02 00:00:00	5	2004.07	85251	Scottsdale
6	Suyama	Michael	Sales Representative	1963-07-02 00:00:00	5	2004.07	98004	Bellevue
6	Suyama	Michael	Sales Representative	1963-07-02 00:00:00	5	2004.07	98052	Redmond
6	Suyama	Michael	Sales Representative	1963-07-02 00:00:00	5	2004.07	98104	Seattle
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	60179	Hoffman Estates
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	60601	Chicago
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	80202	Denver
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	80909	Colorado Springs
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	90405	Santa Monica
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	94025	Menlo Park
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	94105	San Francisco
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	95008	Campbell
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	95054	Santa Clara
7	King	Robert	Sales Representative	1960-05-29 00:00:00	5	1991.55	95060	Santa Cruz
8	Callahan	Laura	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	19428	Philadelphia
8	Callahan	Laura	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	44122	Beachwood
8	Callahan	Laura	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	45839	Findlay
8	Callahan	Laura	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	53404	Racine
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	03049	Hollis
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	03801	Portsmouth
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	48075	Southfield
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	48084	Troy
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	48304	Bloomfield Hills
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	55113	Roseville
9	Dodsworth	Anne	Sales Representative	1966-01-27 00:00:00	5	2333.33	55439	Minneapolis

company_employees view creation:

```
create or replace view company_employees(employeeid, lastname,
firstname, title, birthdate, reportto, salary, territoryid,
city) as
     (select e.employeeid, e.lastname, e.firstname,
e.jobtitle, e.dob, e.reportto, e.salary, b.branchid, b.city
     from company.employees as e, company.branches as b
     where e.branchid = b.branchid);
```

company_employees resulting table:

employeeid	lastname	firstname	title	birthdate	reportto	salary	branchid	city
1001	Gupta	Ravi	CEO	1969-12-03	1001	850000.00	18	Santa Fe
1002	charan	Ram	Director	1985-02-20	1001	650000.00	18	Santa Fe
1003	jha	Ramesh	President	1977-02-11	1002	625000.00	1	Bangalore
1004	Patra	Maheswar	Vice President	1979-12-16	1003	215000.00	1	Bangalore
1005	Diane	Murphy	Sr. Manager	1984-01-11	1004	115000.00	2	Pune
1006	Parhihar	Ashok	Accountant	1983-04-03	1005	38000.00	2	Pune
1007	William	Patterson	Sales Manager	1988-12-20	1005	65000.00	2	Pune
1008	Gahoi	Sam	Reporting Manager	1986-11-23	1005	75000.00	2	Pune
1009	Agarwal	Rakul	Team Leader	1990-09-08	1008	45000.00	1	Bangalore
1010	Thakur	Hari	Sales Rep	1991-07-13	1007	15000.00	1	Bangalore
1011	Lohan	Addision	Sales Rep	1994-06-19	1007	15000.00	1	Bangalore
1012	Julie	Fothine	Sales Rep	1992-08-22	1007	15000.00	1	Bangalore
1013	Patt	Steve	Sales Rep	1993-02-24	1007	15000.00	1	Bangalore
1014	Jain	Tilak	Software Engineer	1990-03-17	1008	35000.00	2	Pune
1015	Guru	vanky	Software Engineer	1989-04-09	1008	35000.00	2	Pune
1016	Loui	Baldev	Software Engineer	1991-12-02	1008	25000.00	1	Bangalore
1017	Gerard	Hernandez	Software Engineer	1989-11-14	1008	25000.00	1	Bangalore
1018	Pamela	Catty	Sales Rep	1992-10-17	1007	18000.00	2	Pune
1019	Larry	Bott	Sales Rep	1993-07-21	1007	25000.00	2	Pune
1020	Barry	Jones	Sales Rep	1988-05-30	1007	22000.00	2	Pune
1021	Andy	Fixter	Sales Rep	1989-04-29	1007	21000.00	2	Pune
1022	Poddi	Marsh	Sales Rep	1993-02-25	1007	19000.00	2	Pune
1023	Perry	King	Admin	1984-01-07	1005	55000.00	1	Bangalore
1024	Kushwah	Nishi	Network Engineer	1989-05-14	1008	17000.00	1	Bangalore
1025	Yoshimi	Kato	Network Engineer	1991-03-12	1008	25000.00	2	Pune

Note: “northwind” database was in use so “company” was accessed via prefix.

all_employees view creation:

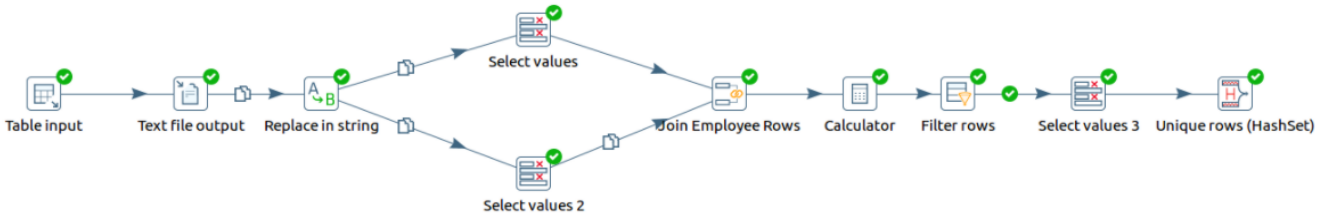
```
create or replace view all_employees(employeeid, firstname,
lastname, jobtitle, dob, reportto, salary, branchid, city) as
     (select employeeid, firstname, lastname, title,
birthdate, reportto, salary, branchid, city from
company_employees)
     union
     (select employeeid, firstname, lastname, title,
birthdate, reportto, salary, territoryid, city from
northwind_employees);
```

all_employees resulting table:

employeeid	firstname	lastname	jobtitle	dob	reportto	salary	branchid	city
1001	Ravi	Gupta	CEO	1969-12-03 00:00:00	1001	850000	18	Santa Fe
1002	Ram	charan	Director	1985-02-20 00:00:00	1001	650000	18	Santa Fe
1003	Ramesh	jha	President	1977-02-11 00:00:00	1002	625000	1	Bangalore
1004	Maheswar	Patra	Vice President	1979-12-16 00:00:00	1003	215000	1	Bangalore
1005	Murphy	Diane	Sr. Manager	1984-01-11 00:00:00	1004	115000	2	Pune
1006	Ashok	Parhihar	Accountant	1983-04-03 00:00:00	1005	38000	2	Pune
1007	Patterson	William	Sales Manager	1988-12-20 00:00:00	1005	65000	2	Pune
1008	Sam	Gahoi	Reporting Manager	1986-11-23 00:00:00	1005	75000	2	Pune
1009	Rakul	Agarwal	Team Leader	1990-09-08 00:00:00	1008	45000	1	Bangalore
1010	Hari	Thakur	Sales Rep	1991-07-13 00:00:00	1007	15000	1	Bangalore
1011	Addision	Lohan	Sales Rep	1994-06-19 00:00:00	1007	15000	1	Bangalore
1012	Fothine	Julie	Sales Rep	1992-08-22 00:00:00	1007	15000	1	Bangalore
1013	Steve	Patt	Sales Rep	1993-02-24 00:00:00	1007	15000	1	Bangalore
1014	Tilak	Jain	Software Engineer	1990-03-17 00:00:00	1008	35000	2	Pune
1015	vanky	Guru	Software Engineer	1989-04-09 00:00:00	1008	35000	2	Pune
1016	Baldev	Loui	Software Engineer	1991-12-02 00:00:00	1008	25000	1	Bangalore
1017	Hernandez	Gerard	Software Engineer	1989-11-14 00:00:00	1008	25000	1	Bangalore
1018	Catty	Pamela	Sales Rep	1992-10-17 00:00:00	1007	18000	2	Pune
1019	Bott	Larry	Sales Rep	1993-07-21 00:00:00	1007	25000	2	Pune
1020	Jones	Barry	Sales Rep	1988-05-30 00:00:00	1007	22000	2	Pune
1021	Fixter	Andy	Sales Rep	1989-04-29 00:00:00	1007	21000	2	Pune
1022	Marsh	Poddi	Sales Rep	1993-02-25 00:00:00	1007	19000	2	Pune
1023	King	Perry	Admin	1984-01-07 00:00:00	1005	55000	1	Bangalore
1024	Nishi	Kushwah	Network Engineer	1989-05-14 00:00:00	1008	17000	1	Bangalore
1025	Kato	Yoshimi	Network Engineer	1991-03-12 00:00:00	1008	25000	2	Pune
1	Nancy	Davolio	Sales Representative	1948-12-08 00:00:00	2	2954.55	06897	Wilton
1	Nancy	Davolio	Sales Representative	1948-12-08 00:00:00	2	2954.55	19713	Neward
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	01581	Westboro
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	01730	Bedford
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	01833	Georgetow
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	02116	Boston
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	02139	Cambridge
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	02184	Braintree
2	Andrew	Fuller	Vice President, Sales	1952-02-19 00:00:00	NULL	2254.49	40222	Louisville
3	Janet	Leverling	Sales Representative	1963-08-30 00:00:00	2	3119.15	30346	Atlanta
3	Janet	Leverling	Sales Representative	1963-08-30 00:00:00	2	3119.15	31406	Savannah
3	Janet	Leverling	Sales Representative	1963-08-30 00:00:00	2	3119.15	32859	Orlando
3	Janet	Leverling	Sales Representative	1963-08-30 00:00:00	2	3119.15	33607	Tampa
4	Margaret	Peacock	Sales Representative	1937-09-19 00:00:00	2	1861.08	20852	Rockville
4	Margaret	Peacock	Sales Representative	1937-09-19 00:00:00	2	1861.08	27403	Greensboro
4	Margaret	Peacock	Sales Representative	1937-09-19 00:00:00	2	1861.08	27511	Cary
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	02903	Providence
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	07960	Morristown
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	08837	Edison
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	10019	New York
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	10038	New York
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	11747	Mellvile
5	Steven	Buchanan	Sales Manager	1955-03-04 00:00:00	2	1744.21	14450	Fairport
6	Michael	Suyama	Sales Representative	1963-07-02 00:00:00	5	2004.07	85014	Phoenix
6	Michael	Suyama	Sales Representative	1963-07-02 00:00:00	5	2004.07	85251	Scottsdale
6	Michael	Suyama	Sales Representative	1963-07-02 00:00:00	5	2004.07	98004	Bellevue
6	Michael	Suyama	Sales Representative	1963-07-02 00:00:00	5	2004.07	98052	Redmond
6	Michael	Suyama	Sales Representative	1963-07-02 00:00:00	5	2004.07	98104	Seattle
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	60179	Hoffman Estates
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	60601	Chicago
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	80202	Denver
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	80909	Colorado Springs
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	90405	Santa Monica
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	94025	Menlo Park
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	94105	San Francisco
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	95008	Campbell
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	95054	Santa Clara
7	Robert	King	Sales Representative	1960-05-29 00:00:00	5	1991.55	95060	Santa Cruz
8	Laura	Callahan	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	19428	Philadelphia
8	Laura	Callahan	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	44122	Beachwood
8	Laura	Callahan	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	45839	Findlay
8	Laura	Callahan	Inside Sales Coordinator	1958-01-09 00:00:00	2	2100.5	53404	Racine
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	03049	Hollis
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	03801	Portsmouth
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	48075	Southfield
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	48084	Troy
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	48304	Bloomfield Hills
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	55113	Roseville
9	Anne	Dodsworth	Sales Representative	1966-01-27 00:00:00	5	2333.33	55439	Minneapolis

1.4. Transformation - Duplicated Job Titles

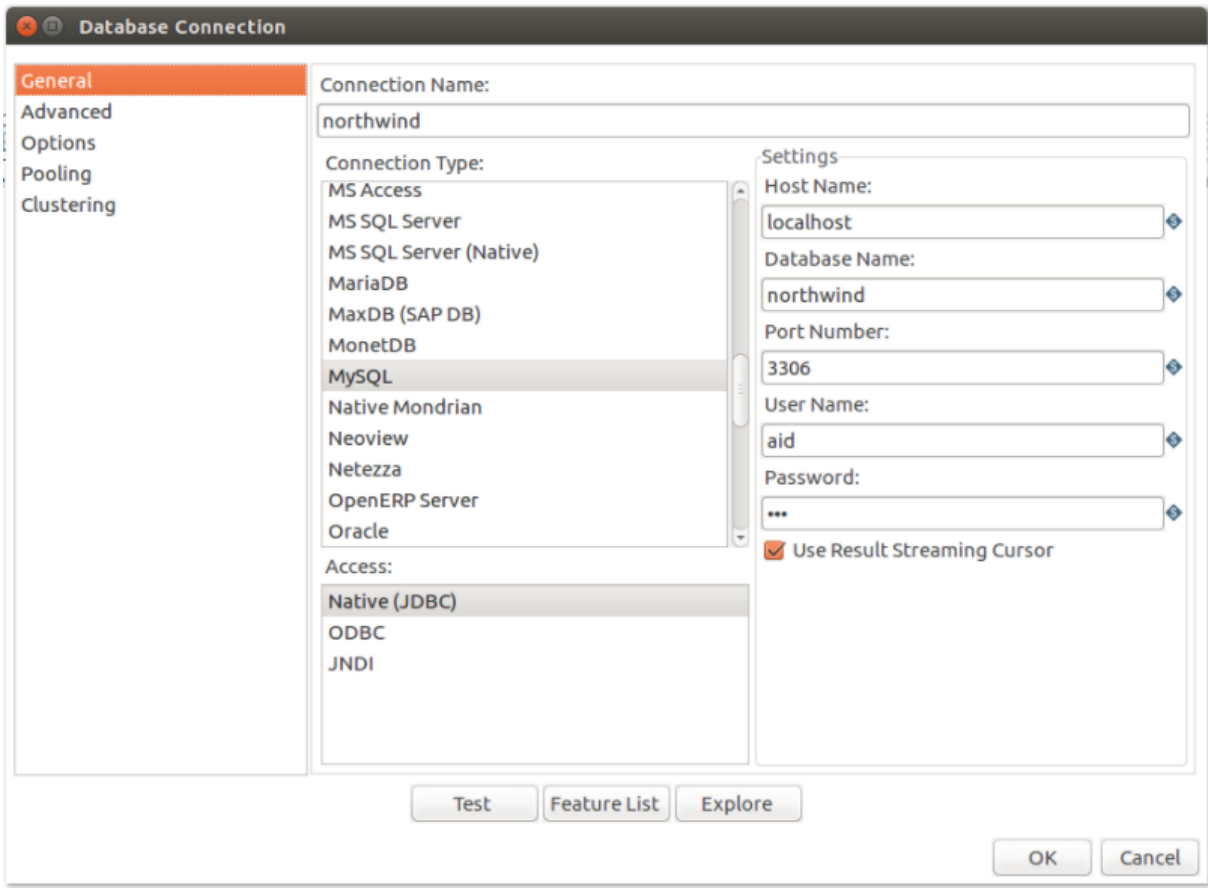
To exeute this transformation, Jaro measure was used in order to calculate the distance between job titles.
Final Transformation overview:



Step-by-step development:

View > Transformations > Database connections > New

Database connection config:



Design > Input > Table Input

Table input config:

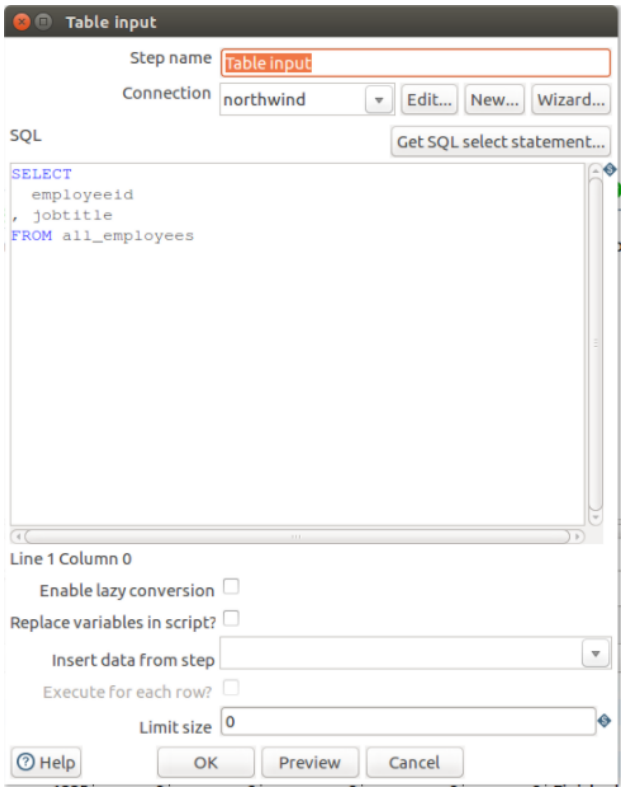


Table input preview (sample):

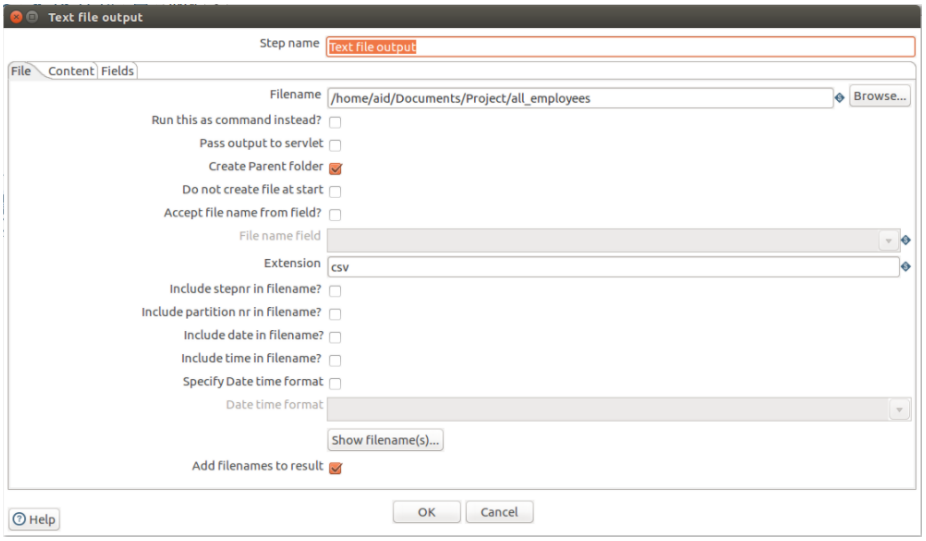
Examine preview data

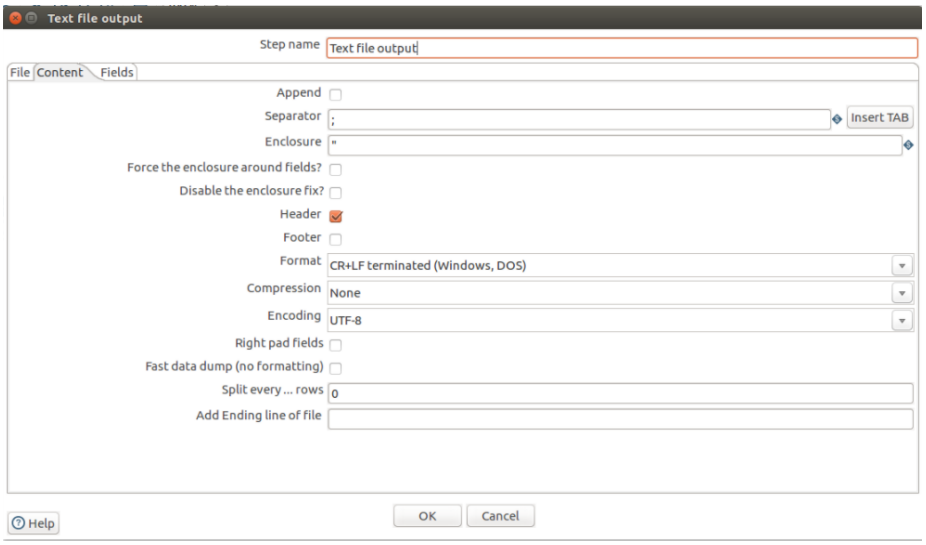
Rows of step: Table input (74 rows)

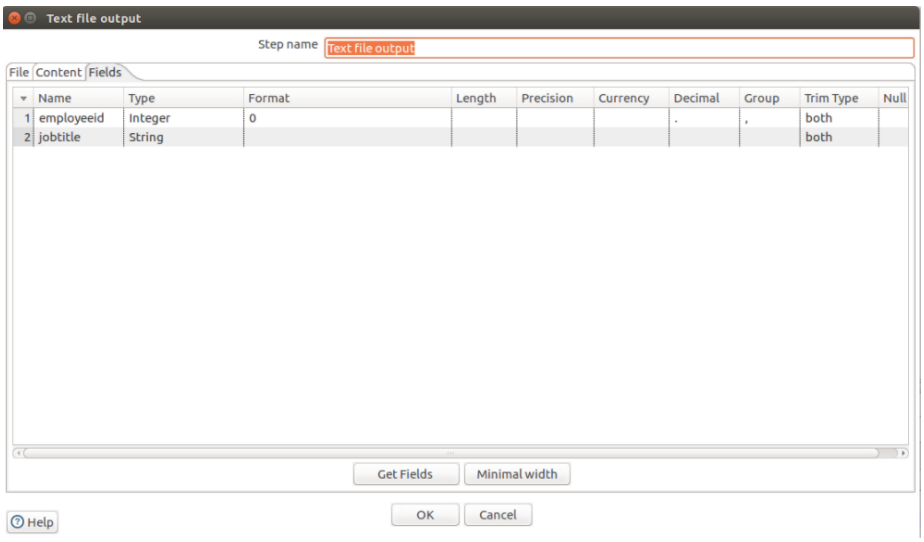
	employeeid	jobtitle
1	1001	CEO
2	1002	Director
3	1003	President
4	1004	Vice President
5	1005	Sr. Manager
6	1006	Accountant
7	1007	Sales Manager
8	1008	Reporting Manager
9	1009	Team Leader
10	1010	Sales Rep
11	1011	Sales Rep
12	1012	Sales Rep
13	1013	Sales Rep
14	1014	Software Engineer
15	1015	Software Engineer
16	1016	Software Engineer
17	1017	Software Engineer
18	1018	Sales Rep
19	1019	Sales Rep
20	1020	Sales Rep
21	1021	Sales Rep
22	1022	Sales Rep
23	1023	Admin
24	1024	Network Engineer
25	1025	Network Engineer
26	1	Sales Representative
27	1	Sales Representative
28	2	Vice President, Sales

Design > Output > Text file output

Text file output config:





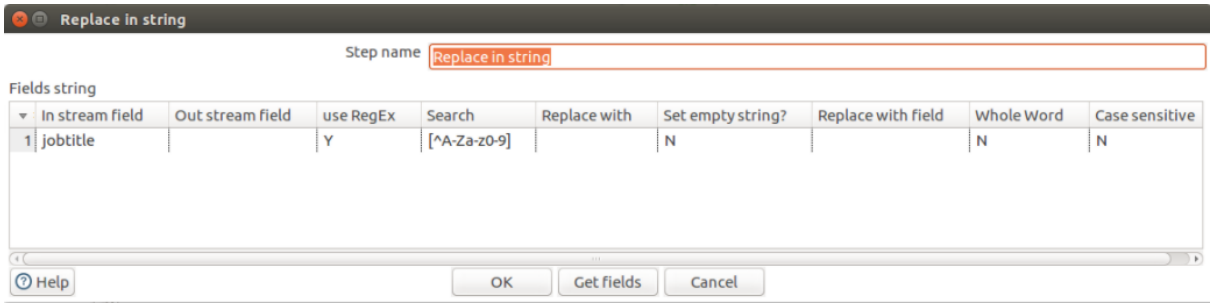


Text file output preview (sample):

	A	B
1	employeeid	jobtitle
2	1001	CEO
3	1002	Director
4	1003	President
5	1004	Vice President
6	1005	Sr. Manager
7	1006	Accountant
8	1007	Sales Manager
9	1008	Reporting Manager
10	1009	Team Leader
11	1010	Sales Rep
12	1011	Sales Rep
13	1012	Sales Rep
14	1013	Sales Rep
15	1014	Software Engineer
16	1015	Software Engineer
17	1016	Software Engineer
18	1017	Software Engineer
19	1018	Sales Rep
20	1019	Sales Rep
21	1020	Sales Rep
22	1021	Sales Rep
23	1022	Sales Rep
24	1023	Admin
25	1024	Network Engineer
26	1025	Network Engineer
27	1	Sales Representative
28	1	Sales Representative
29	2	Vice President, Sales
30	2	Vice President, Sales
31	2	Vice President, Sales
32	2	Vice President, Sales

Design > Transform > Replace in string

Replace in string config:



Replace in string preview (sample):

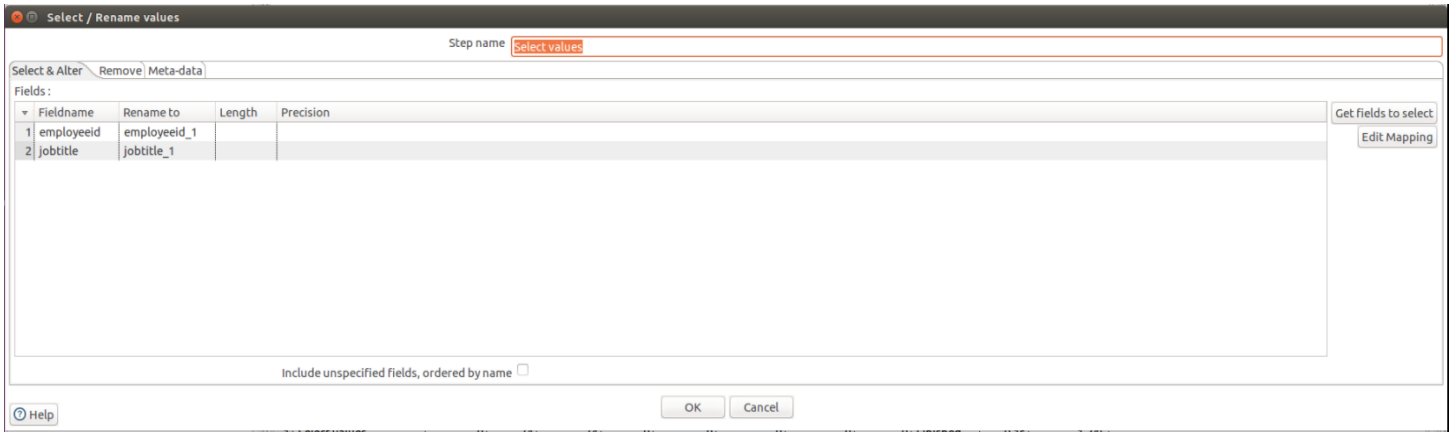
Examine preview data

Rows of step: Replace in string (74 rows)

	employeeid	jobtitle
1	1001	CEO
2	1002	Director
3	1003	President
4	1004	VicePresident
5	1005	SrManager
6	1006	Accountant
7	1007	SalesManager
8	1008	ReportingManager
9	1009	TeamLeader
10	1010	SalesRep
11	1011	SalesRep
12	1012	SalesRep
13	1013	SalesRep
14	1014	SoftwareEngineer
15	1015	SoftwareEngineer
16	1016	SoftwareEngineer
17	1017	SoftwareEngineer
18	1018	SalesRep
19	1019	SalesRep
20	1020	SalesRep
21	1021	SalesRep
22	1022	SalesRep
23	1023	Admin
24	1024	NetworkEngineer
25	1025	NetworkEngineer
26	1	SalesRepresentative
27	1	SalesRepresentative
28	2	VicePresidentSales
29	2	VicePresidentSales
30	2	VicePresidentSales
31	2	VicePresidentSales
32	2	VicePresidentSales
33	2	VicePresidentSales
34	2	VicePresidentSales

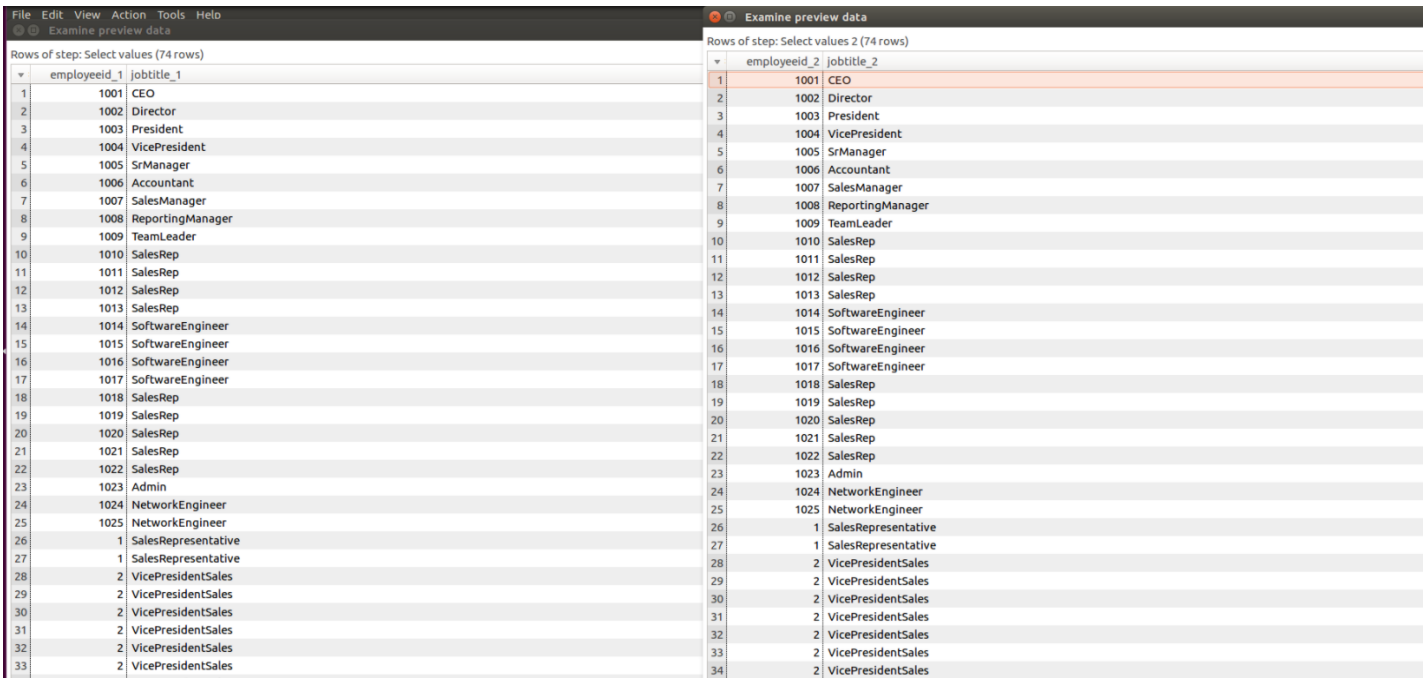
Design > Transform > Select values

Select values config:



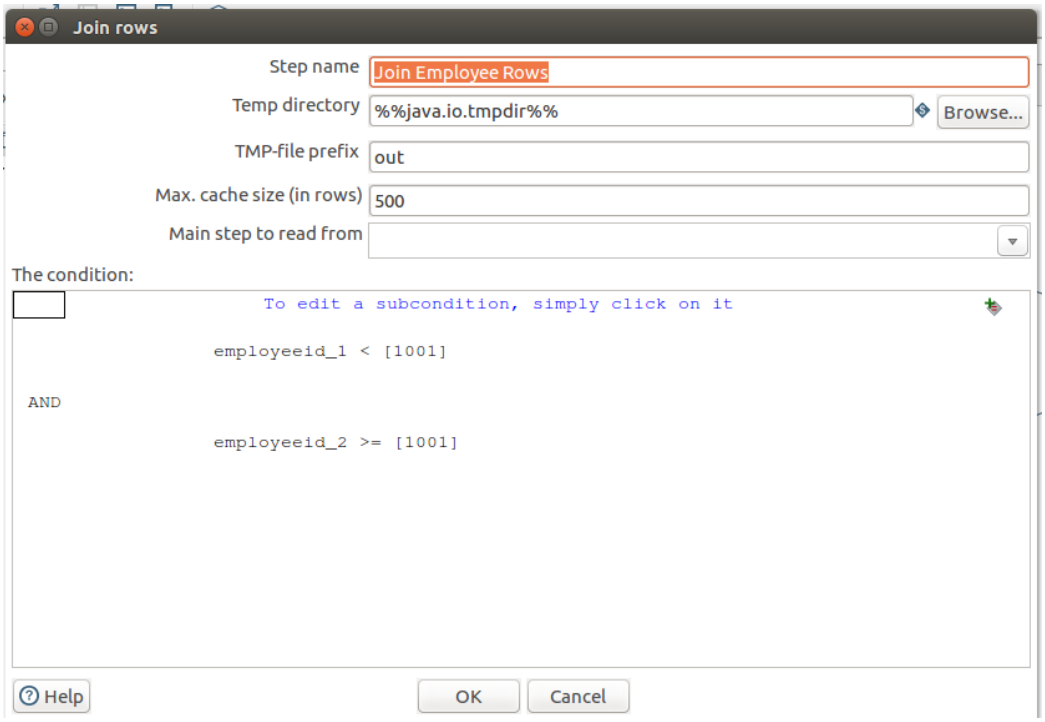
Note: the same step was done for "Select values 2’ where all the fields were renamed instead to their homonyms with the identifier ‘_2’ after the corresponding name.

Select values and Select values 2 previews (samples):



Design > Joins > Join rows

Join rows config:



Conditions: ID's within the same company don't need to be compared.

Join rows preview (sample):

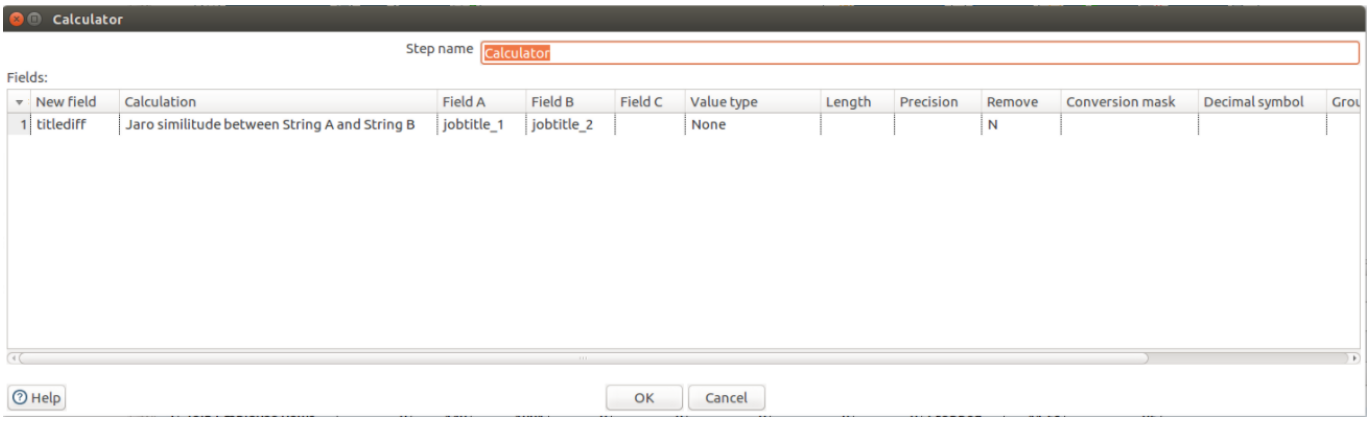
Examine preview data

Rows of step: Join Employee Rows (1000 rows)

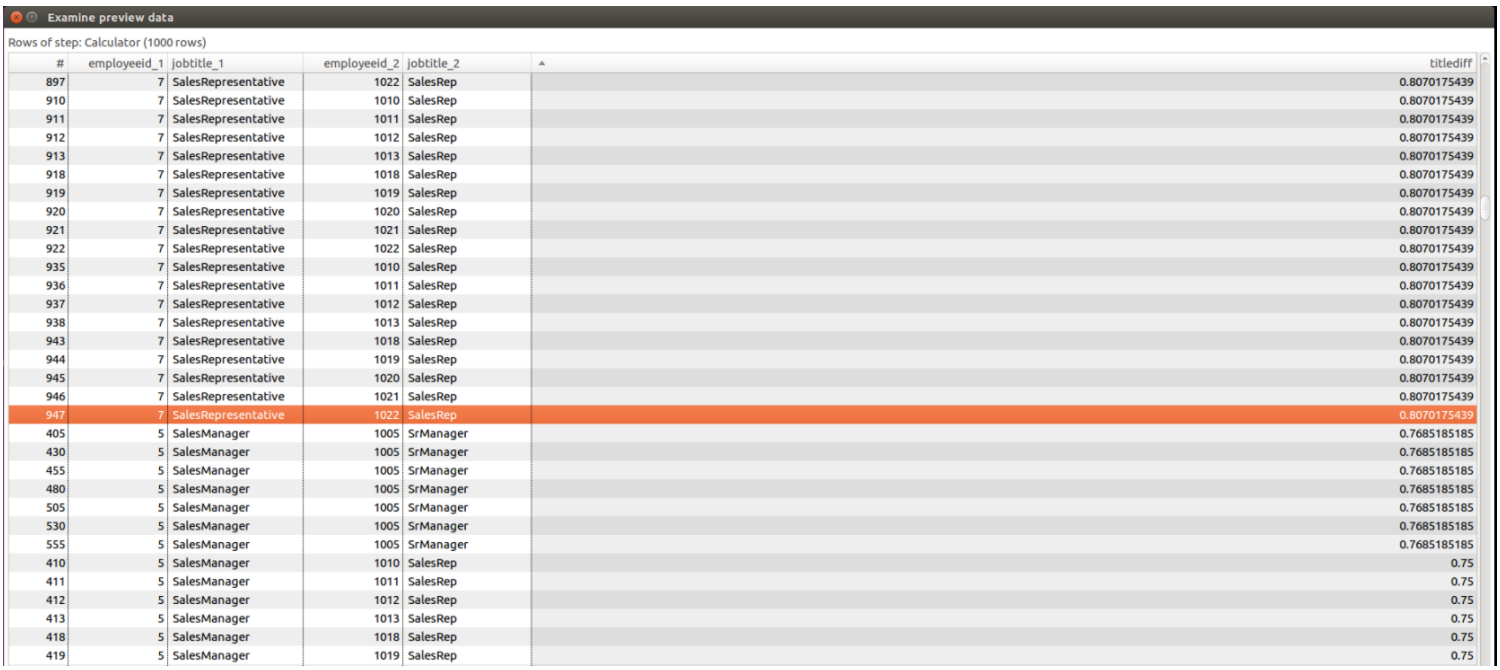
	employeeid_1	jobtitle_1	employeeid_2	jobtitle_2
1	1	SalesRepresentative	1001	CEO
2	1	SalesRepresentative	1002	Director
3	1	SalesRepresentative	1003	President
4	1	SalesRepresentative	1004	VicePresident
5	1	SalesRepresentative	1005	SrManager
6	1	SalesRepresentative	1006	Accountant
7	1	SalesRepresentative	1007	SalesManager
8	1	SalesRepresentative	1008	ReportingManager
9	1	SalesRepresentative	1009	TeamLeader
10	1	SalesRepresentative	1010	SalesRep
11	1	SalesRepresentative	1011	SalesRep
12	1	SalesRepresentative	1012	SalesRep
13	1	SalesRepresentative	1013	SalesRep
14	1	SalesRepresentative	1014	SoftwareEngineer
15	1	SalesRepresentative	1015	SoftwareEngineer
16	1	SalesRepresentative	1016	SoftwareEngineer
17	1	SalesRepresentative	1017	SoftwareEngineer
18	1	SalesRepresentative	1018	SalesRep
19	1	SalesRepresentative	1019	SalesRep
20	1	SalesRepresentative	1020	SalesRep
21	1	SalesRepresentative	1021	SalesRep
22	1	SalesRepresentative	1022	SalesRep
23	1	SalesRepresentative	1023	Admin
24	1	SalesRepresentative	1024	NetworkEngineer
25	1	SalesRepresentative	1025	NetworkEngineer
26	1	SalesRepresentative	1001	CEO
27	1	SalesRepresentative	1002	Director
28	1	SalesRepresentative	1003	President

Design > Transform > Calculator

Calculator config:



Calculator preview (sample):

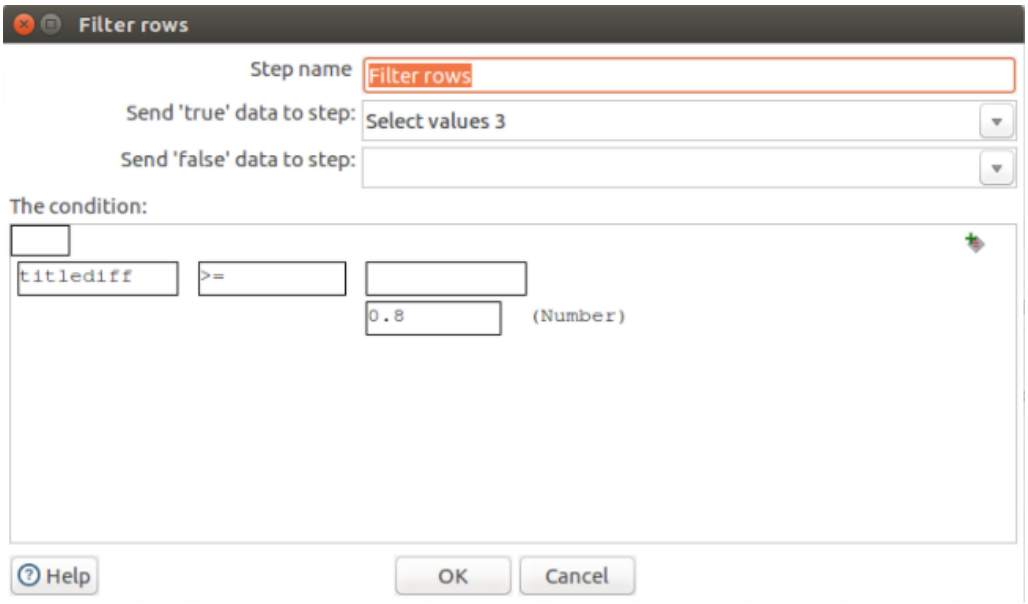


#	employeeid_1	jobtitle_1	employeeid_2	jobtitle_2	titlediff
897	7	SalesRepresentative	1022	SalesRep	0.8070175439
910	7	SalesRepresentative	1010	SalesRep	0.8070175439
911	7	SalesRepresentative	1011	SalesRep	0.8070175439
912	7	SalesRepresentative	1012	SalesRep	0.8070175439
913	7	SalesRepresentative	1013	SalesRep	0.8070175439
918	7	SalesRepresentative	1018	SalesRep	0.8070175439
919	7	SalesRepresentative	1019	SalesRep	0.8070175439
920	7	SalesRepresentative	1020	SalesRep	0.8070175439
921	7	SalesRepresentative	1021	SalesRep	0.8070175439
922	7	SalesRepresentative	1022	SalesRep	0.8070175439
935	7	SalesRepresentative	1010	SalesRep	0.8070175439
936	7	SalesRepresentative	1011	SalesRep	0.8070175439
937	7	SalesRepresentative	1012	SalesRep	0.8070175439
938	7	SalesRepresentative	1013	SalesRep	0.8070175439
943	7	SalesRepresentative	1018	SalesRep	0.8070175439
944	7	SalesRepresentative	1019	SalesRep	0.8070175439
945	7	SalesRepresentative	1020	SalesRep	0.8070175439
946	7	SalesRepresentative	1021	SalesRep	0.8070175439
947	7	SalesRepresentative	1022	SalesRep	0.8070175439
405	5	SalesManager	1005	SrManager	0.7685185185
430	5	SalesManager	1005	SrManager	0.7685185185
455	5	SalesManager	1005	SrManager	0.7685185185
480	5	SalesManager	1005	SrManager	0.7685185185
505	5	SalesManager	1005	SrManager	0.7685185185
530	5	SalesManager	1005	SrManager	0.7685185185
555	5	SalesManager	1005	SrManager	0.7685185185
410	5	SalesManager	1010	SalesRep	0.75
411	5	SalesManager	1011	SalesRep	0.75
412	5	SalesManager	1012	SalesRep	0.75
413	5	SalesManager	1013	SalesRep	0.75
418	5	SalesManager	1018	SalesRep	0.75
419	5	SalesManager	1019	SalesRep	0.75

All values below a ‘titlediff’ value of 0.8 seem to be non-duplicate values.

Design > Flow > Filter rows

Filter rows config:



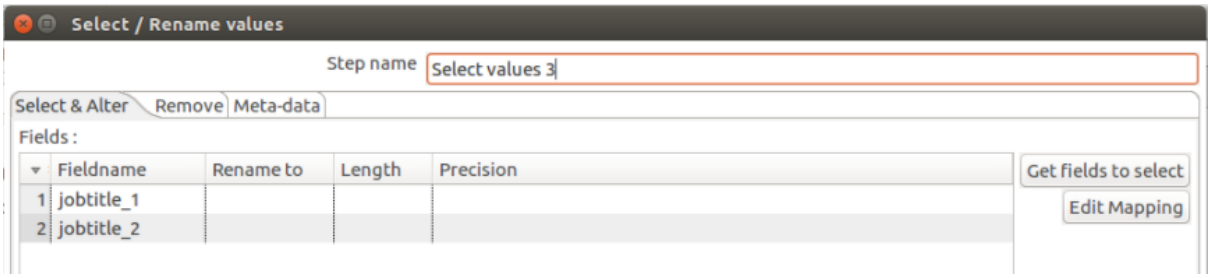
Filter rows preview (sample):

#	employeeid_1	jobtitle_1	employeeid_2	jobtitle_2	titlediff
1	1	SalesRepresentative	1010	SalesRep	0.8070175439
2	1	SalesRepresentative	1011	SalesRep	0.8070175439
3	1	SalesRepresentative	1012	SalesRep	0.8070175439
4	1	SalesRepresentative	1013	SalesRep	0.8070175439
5	1	SalesRepresentative	1018	SalesRep	0.8070175439
6	1	SalesRepresentative	1019	SalesRep	0.8070175439
7	1	SalesRepresentative	1020	SalesRep	0.8070175439
8	1	SalesRepresentative	1021	SalesRep	0.8070175439
9	1	SalesRepresentative	1022	SalesRep	0.8070175439
10	1	SalesRepresentative	1010	SalesRep	0.8070175439
11	1	SalesRepresentative	1011	SalesRep	0.8070175439
12	1	SalesRepresentative	1012	SalesRep	0.8070175439
13	1	SalesRepresentative	1013	SalesRep	0.8070175439
14	1	SalesRepresentative	1018	SalesRep	0.8070175439
15	1	SalesRepresentative	1019	SalesRep	0.8070175439
16	1	SalesRepresentative	1020	SalesRep	0.8070175439
17	1	SalesRepresentative	1021	SalesRep	0.8070175439
18	1	SalesRepresentative	1022	SalesRep	0.8070175439
26	3	SalesRepresentative	1010	SalesRep	0.8070175439
27	3	SalesRepresentative	1011	SalesRep	0.8070175439
28	3	SalesRepresentative	1012	SalesRep	0.8070175439
29	3	SalesRepresentative	1013	SalesRep	0.8070175439
30	3	SalesRepresentative	1018	SalesRep	0.8070175439
31	3	SalesRepresentative	1019	SalesRep	0.8070175439
32	3	SalesRepresentative	1020	SalesRep	0.8070175439

Values of ‘titlediff’ ordered in an ascending order starting at 0.8.

Design > Transform > Select values

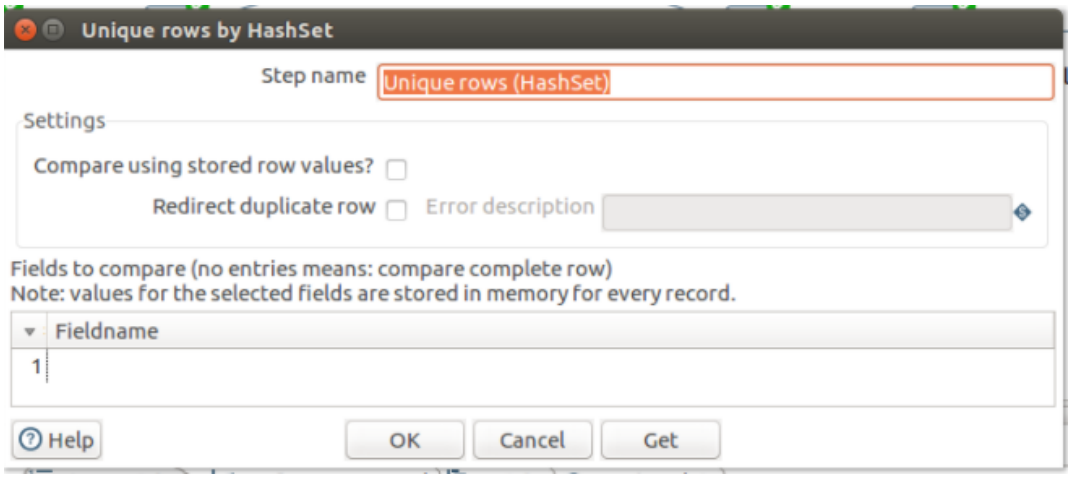
Select Values config:



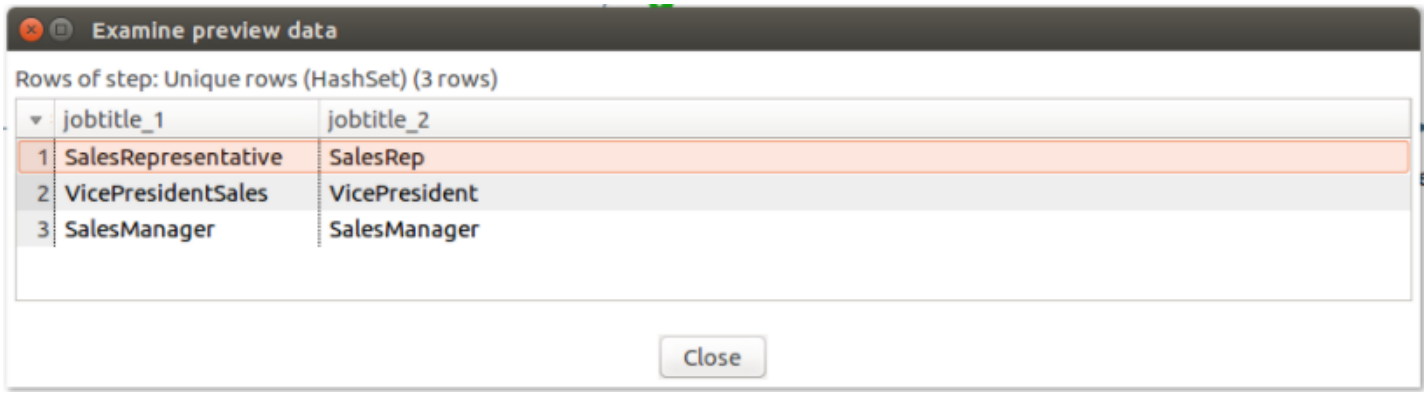
Auxiliar step before retrieving the final result.

Design > Transform > Unique rows (HashSet)

Unique rows config:



Final result:


Examine preview data
Rows of step: Unique rows (HashSet) (3 rows)
jobtitle_1
jobtitle_2
1 SalesRepresentative SalesRep
2 VicePresidentSales VicePresident
3 SalesManager SalesManager
Close

2.

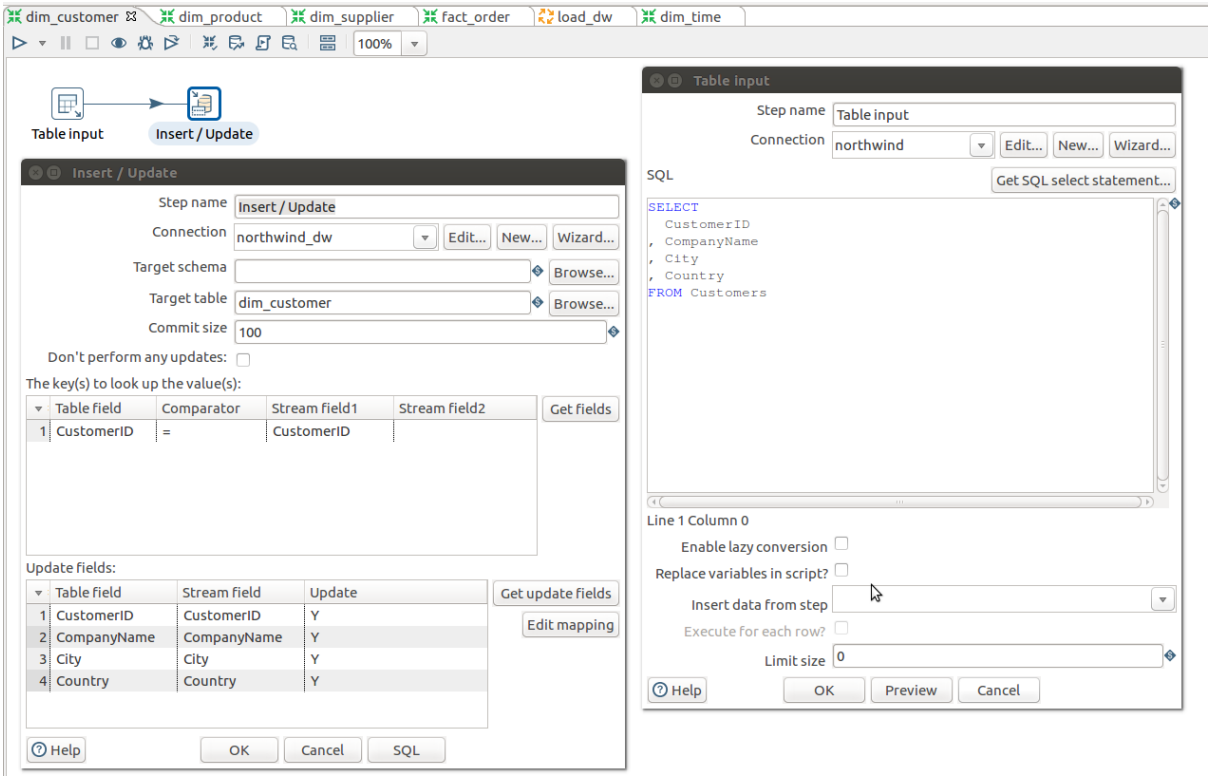
2.1. Data Warehouse Tables - SQL Instructions

```
CREATE TABLE dim_customer (
     CustomerID VARCHAR(5),

     CompanyName VARCHAR(40),
     City VARCHAR(15),
     Country VARCHAR(15),

     PRIMARY KEY (CustomerID)
);

CREATE TABLE dim_product (
     ProductKey INT,

     ProductID INT,
     ProductName VARCHAR(40),
     CategoryName VARCHAR(15),
     VERSION INT,
     DATE_FROM DATETIME,
     DATE_TO DATETIME,

     PRIMARY KEY (ProductKey)
);


CREATE TABLE dim_supplier (
     SupplierID INT,

     CompanyName VARCHAR(40),
     City VARCHAR(15),
     Country VARCHAR(15),

     PRIMARY KEY (SupplierID)
);

CREATE TABLE dim_time (
     TimeID DATETIME,

     DayID INT,
     MonthID INT,
     MonthName VARCHAR(4),
```

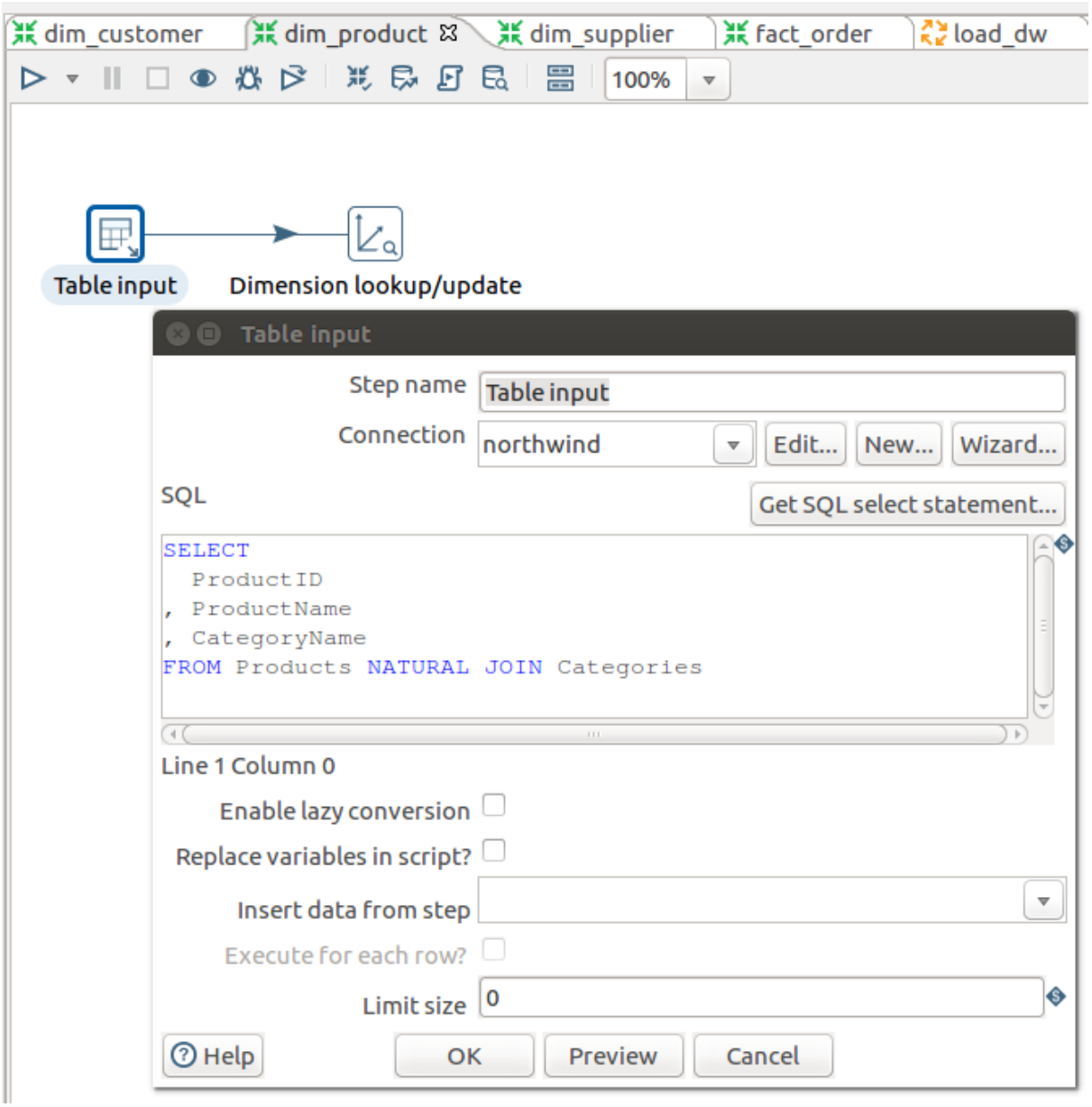
```
     YearID INT,

     PRIMARY KEY (TimeID)
);

CREATE TABLE fact_order (
     OrderID INT,
     ProductID INT,

     Sales DOUBLE,
     Quantity INT,

     CustomerID VARCHAR(5),
     SupplierID INT,
     ProductKey INT,
     TimeID DATETIME,

     PRIMARY KEY (OrderID,ProductID),
     FOREIGN KEY (CustomerID) REFERENCES dim_customer
(CustomerID),
     FOREIGN KEY (SupplierID) REFERENCES dim_supplier
(SupplierID),
     FOREIGN KEY (ProductKey) REFERENCES dim_product
(ProductKey),
     FOREIGN KEY (TimeID) REFERENCES dim_time (TimeID)
);
```

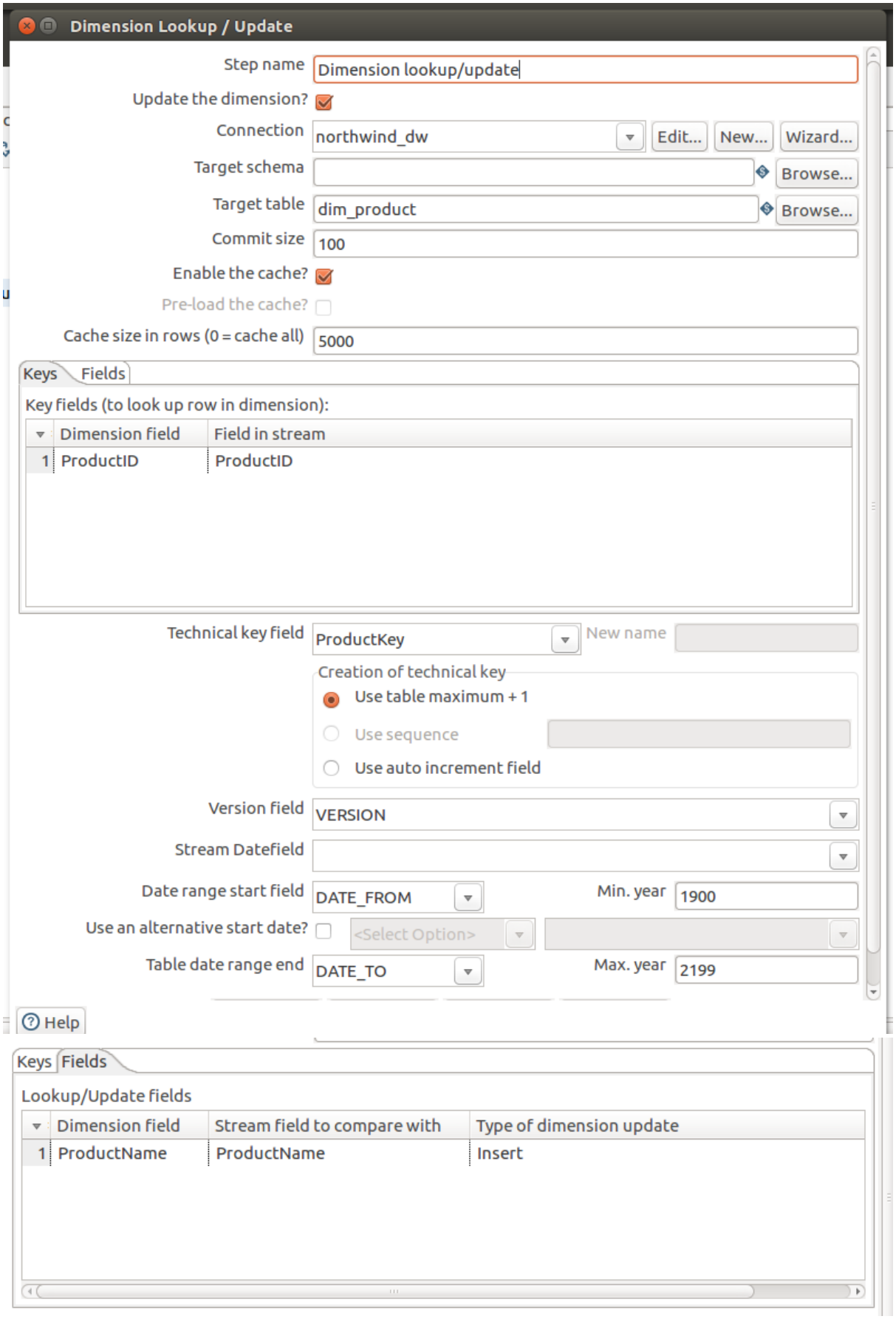
2.2. Transformation - ETL process in PDI

dim_customer:

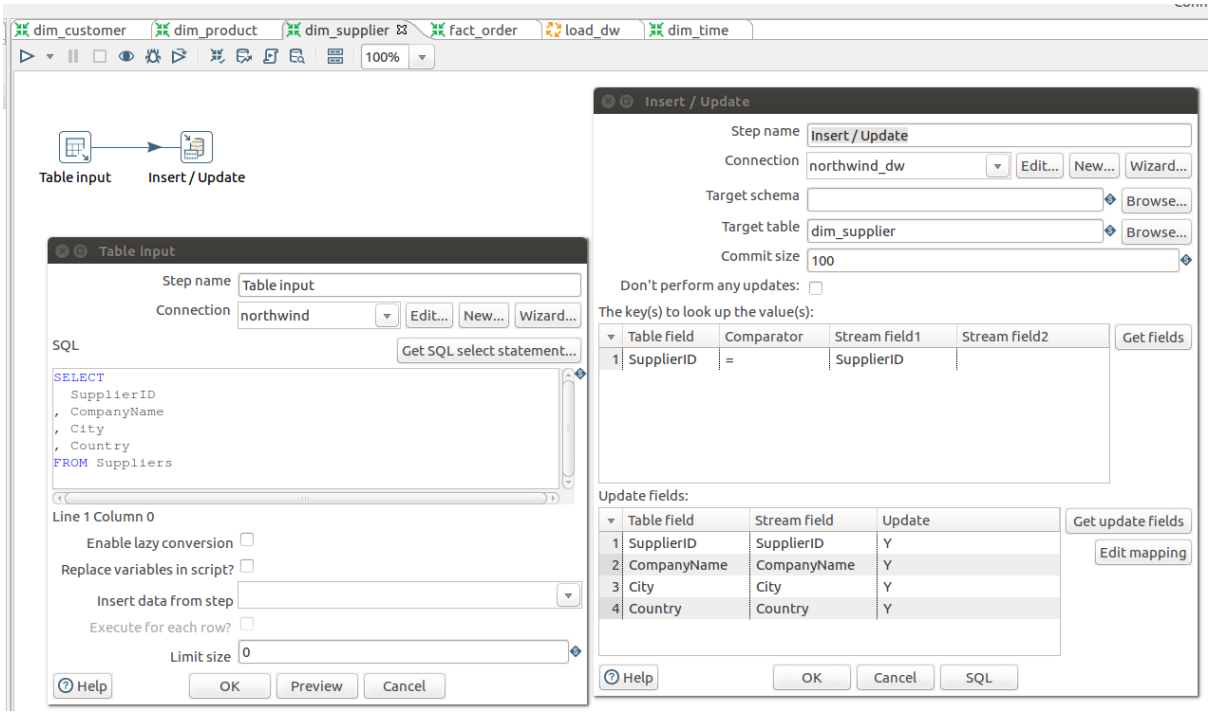

dim_customer
dim_product
dim_supplier
fact_order
load_dw
dim_time
100%
Table input
Insert / Update
Insert / Update
Step name
Insert / Update
Connection
northwind_dw
Edit...
New...
Wizard...
Target schema
Browse...
Target table
dim_customer
Browse...
Commit size
100
Don't perform any updates:
The key(s) to look up the value(s):
Table field
Comparator
Stream field1
Stream field2
Get fields
1 CustomerID = CustomerID
Update fields:
Table field
Stream field
Update
Get update fields
Edit mapping
1 CustomerID CustomerID Y
2 CompanyName CompanyName Y
3 City City Y
4 Country Country Y
Help
OK
Cancel
SQL
Table input
Step name
Table input
Connection
northwind
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT
CustomerID
, CompanyName
, City
, Country
FROM Customers
Line 1 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel

dim_product:

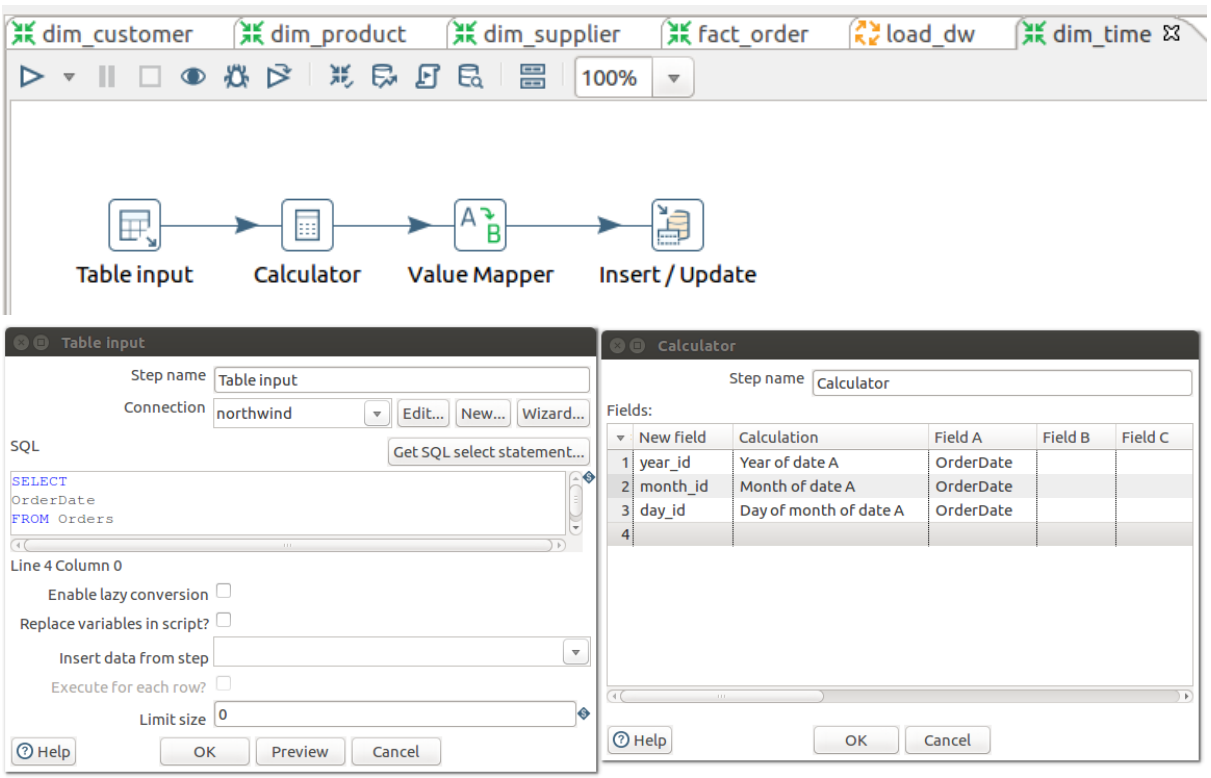

dim_customer
dim_product
dim_supplier
fact_order
load_dw
100%
Table input
Dimension lookup/update
Table input
Step name
Table input
Connection
northwind
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT
ProductID
, ProductName
, CategoryName
FROM Products NATURAL JOIN Categories
Line 1 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel

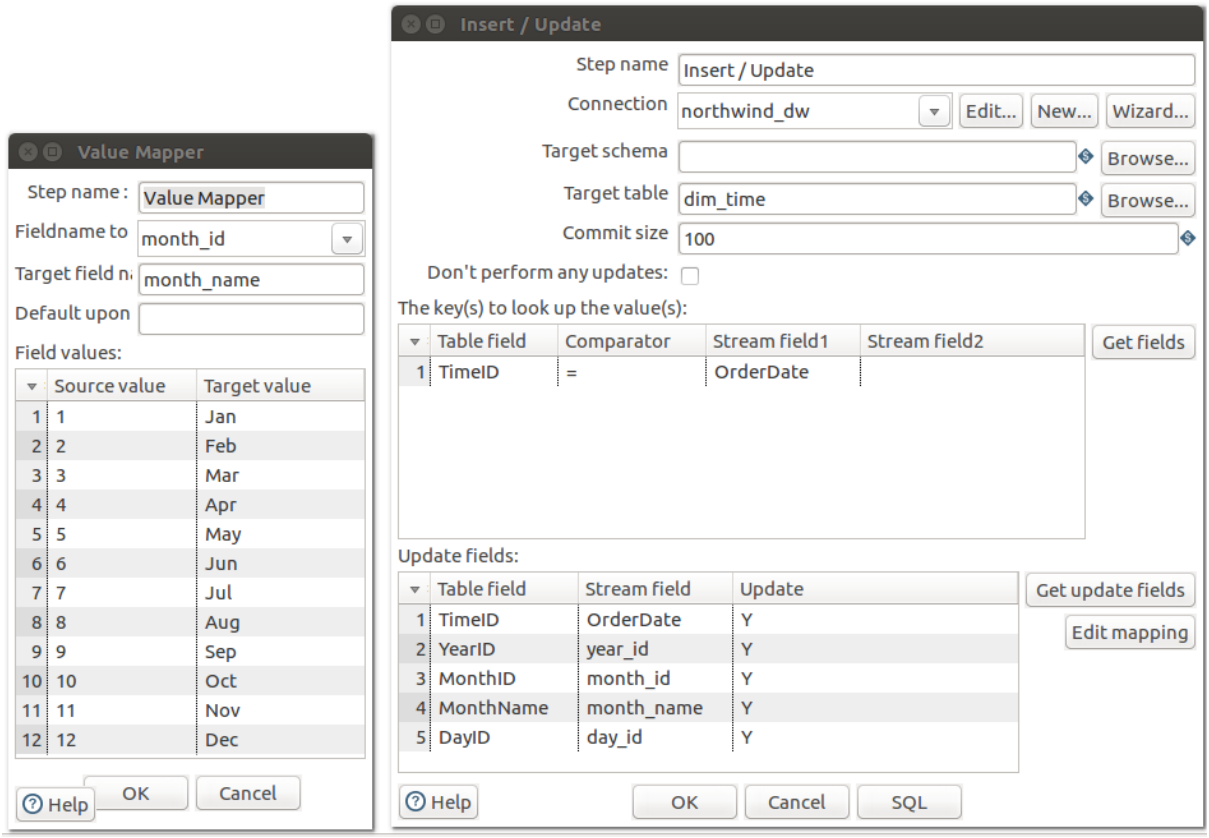


dim_supplier:


dim_customer
dim_product
dim_supplier
fact_order
load_dw
dim_time
100%
Table input
Insert / Update
Table input
Step name
Table input
Connection
northwind
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT
SupplierID
, CompanyName
, City
, Country
FROM Suppliers
Line 1 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel
Insert / Update
Step name
Insert / Update
Connection
northwind_dw
Edit...
New...
Wizard...
Target schema
Browse...
Target table
dim_supplier
Browse...
Commit size
100
Don't perform any updates:
The key(s) to look up the value(s):
Table field
Comparator
Stream field1
Stream field2
Get fields
1 SupplierID = SupplierID
Update fields:
Table field
Stream field
Update
Get update fields
Edit mapping
1 SupplierID SupplierID Y
2 CompanyName CompanyName Y
3 City City Y
4 Country Country Y
Help
OK
Cancel
SQL

dim_time:


dim_customer
dim_product
dim_supplier
fact_order
load_dw
dim_time
100%
Table input
Calculator
Value Mapper
Insert / Update
Table input
Step name
Table input
Connection
northwind
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT
OrderDate
FROM Orders
Line 4 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel
Calculator
Step name
Calculator
Fields:
New field
Calculation
Field A
Field B
Field C
1 year_id Year of date A OrderDate
2 month_id Month of date A OrderDate
3 day_id Day of month of date A OrderDate
4
Help
OK
Cancel


Value Mapper
Step name : Value Mapper
Fieldname to month_id
Target field n month_name
Default upon
Field values:
Source value | Target value
1 | Jan
2 | Feb
3 | Mar
4 | Apr
5 | May
6 | Jun
7 | Jul
8 | Aug
9 | Sep
10 | Oct
11 | Nov
12 | Dec
Help
OK
Cancel
Insert / Update
Step name Insert / Update
Connection northwind_dw
Edit...
New...
Wizard...
Target schema
Browse...
Target table dim_time
Browse...
Commit size 100
Don't perform any updates:
The key(s) to look up the value(s):
Table field | Comparator | Stream field1 | Stream field2
1 | TimeID | = | OrderDate
Get fields
Update fields:
Table field | Stream field | Update
1 | TimeID | OrderDate | Y
2 | YearID | year_id | Y
3 | MonthID | month_id | Y
4 | MonthName | month_name | Y
5 | DayID | day_id | Y
Get update fields
Edit mapping
Help
OK
Cancel
SQL

fact_order:

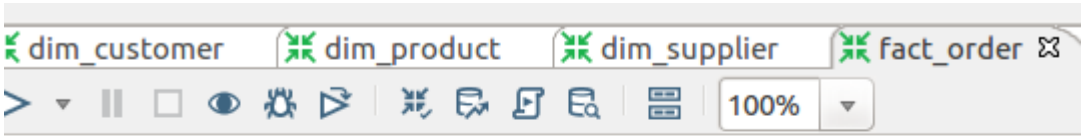
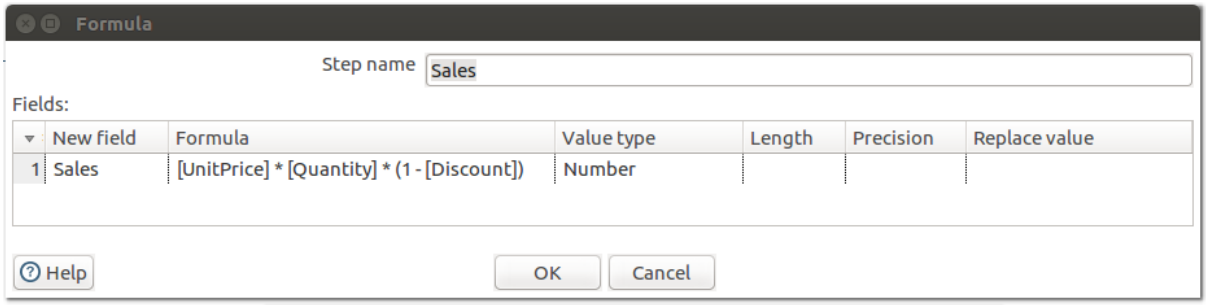

dim_customer
dim_product
dim_supplier
fact_order
100%


Table input
Sales
Database lookup
Insert / Update


Formula
Step name
Sales
Fields:
New field
Formula
Value type
Length
Precision
Replace value
1
Sales
[UnitPrice] * [Quantity] * (1 - [Discount])
Number
Help
OK
Cancel

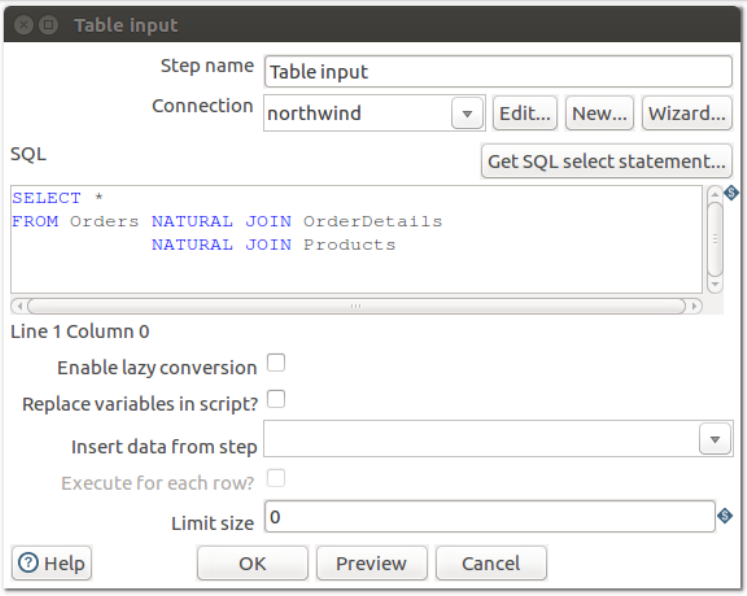
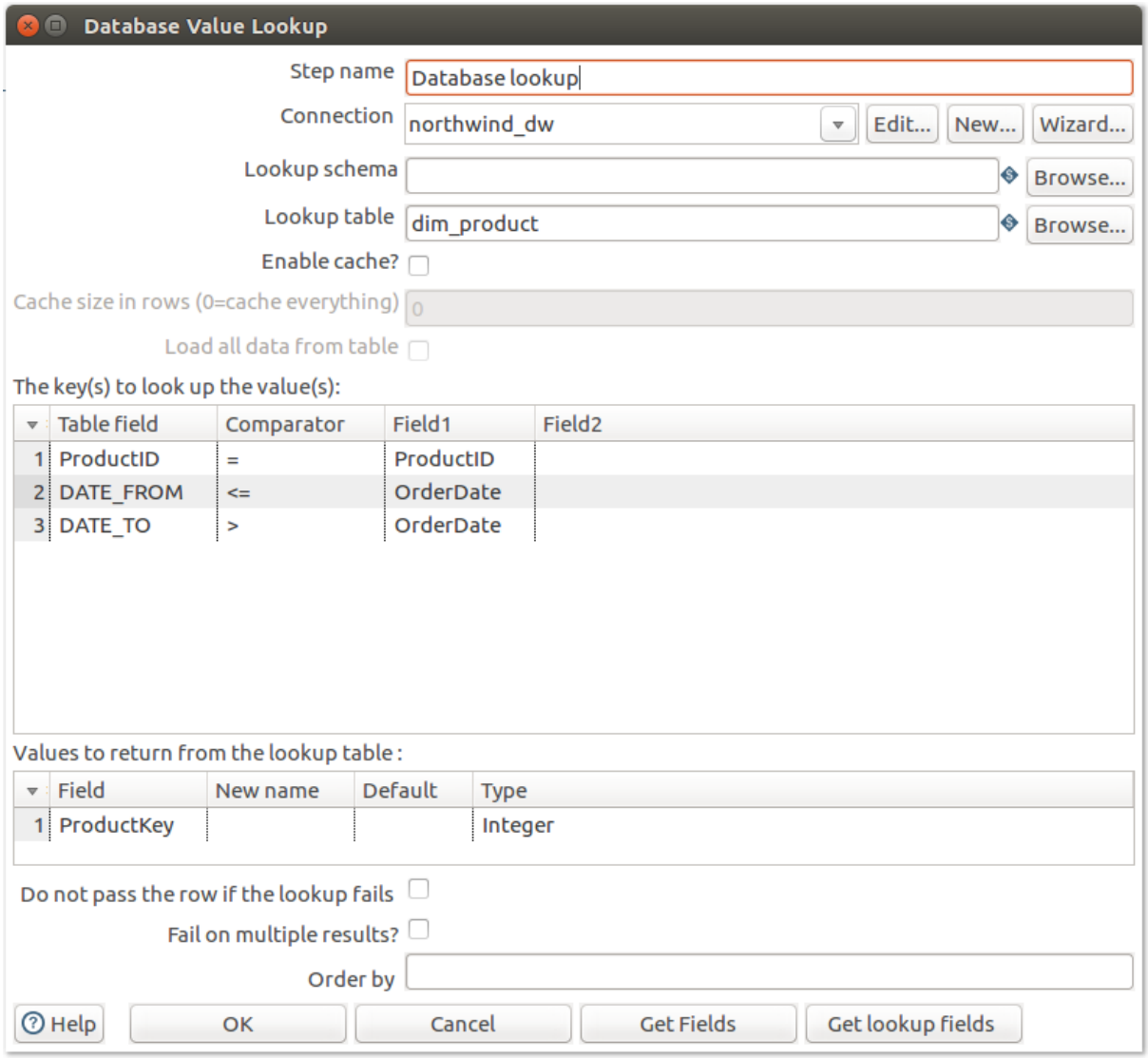
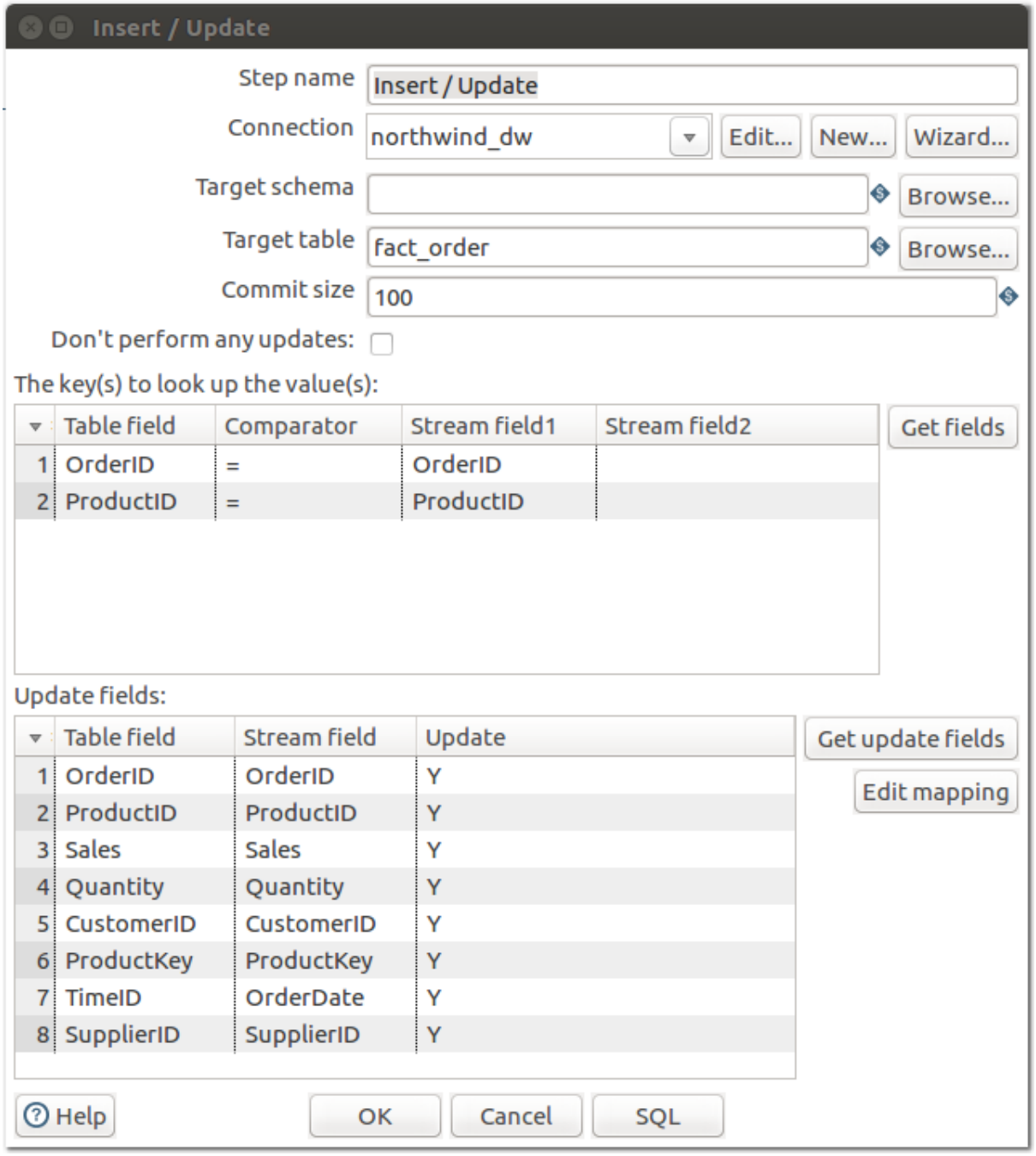


Table input
Step name
Table input
Connection
northwind
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT *
FROM Orders NATURAL JOIN OrderDetails
NATURAL JOIN Products
Line 1 Column 0
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
Limit size
0
Help
OK
Preview
Cancel


Database Value Lookup
Step name
Database lookup
Connection
northwind_dw
Edit...
New...
Wizard...
Lookup schema
Browse...
Lookup table
dim_product
Browse...
Enable cache?
Cache size in rows (0=cache everything)
0
Load all data from table
The key(s) to look up the value(s):
Table field
Comparator
Field1
Field2
1 ProductID = ProductID
2 DATE_FROM <= OrderDate
3 DATE_TO > OrderDate
Values to return from the lookup table :
Field
New name
Default
Type
1 ProductKey Integer
Do not pass the row if the lookup fails
Fail on multiple results?
Order by
Help
OK
Cancel
Get Fields
Get lookup fields



2.3. XML Cube Definition

```
<Schema name="northwind_dw">
  <Cube name="Orders" visible="true" cache="true" enabled="true">
    <Table name="fact_order">
    </Table>
    <Dimension type="StandardDimension" visible="true"
foreignKey="CustomerID" highCardinality="false" name="Customer">
      <Hierarchy name="Customer Hierarchy" visible="true" hasAll="true"
allMemberName="All Customers" primaryKey="CustomerID">
        <Table name="dim_customer">
        </Table>
```

```
        <Level name="Country" visible="true" column="Country" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="City" visible="true" column="City" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Company Name" visible="true" column="CompanyName"
type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" visible="true"
foreignKey="ProductKey" highCardinality="false" name="Product">
      <Hierarchy name="Product Hierarchy" visible="true" hasAll="true"
allMemberName="All Products" primaryKey="ProductKey">
        <Table name="dim_product">
        </Table>
        <Level name="Product Category" visible="true" column="CategoryName"
type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never">
        </Level>
        <Level name="Product Name" visible="true" column="ProductName"
type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="TimeDimension" visible="true" foreignKey="TimeID"
highCardinality="false" name="Time">
      <Hierarchy name="Time Hierarchy" visible="true" hasAll="true"
allMemberName="All Years" primaryKey="TimeID">
        <Table name="dim_time">
        </Table>
        <Level name="Year" visible="true" column="YearID" type="Integer"
uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
        </Level>
        <Level name="Month" visible="true" column="MonthName"
ordinalColumn="MonthID" type="String" uniqueMembers="false"
levelType="TimeMonths" hideMemberIf="Never">
        </Level>
        <Level name="Day" visible="true" column="DayID" type="Integer"
uniqueMembers="false" levelType="TimeDays" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" visible="true"
foreignKey="SupplierID" highCardinality="false" name="Supplier">
      <Hierarchy name="Supplier Hierarchy" visible="true" hasAll="true"
allMemberName="All Suppliers" primaryKey="SupplierID">
        <Table name="dim_supplier">
        </Table>
```

```
        <Level name="Company Name" visible="true" column="CompanyName"
type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never">
        </Level>
        <Level name="City" visible="true" column="City" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Country" visible="true" column="Country" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Measure name="Sales" column="Sales" datatype="Numeric"
formatString="$ #,###.00" aggregator="sum" visible="true">
    </Measure>
    <Measure name="Quantity" column="Quantity" datatype="Integer"
formatString="#,###" aggregator="sum" visible="true">
    </Measure>
  </Cube>
</Schema>
```

2.4. MDX Query

```
WITH MEMBER Measures.PercentSales AS
     (Measures.Sales, Customer.Country.CurrentMember)  /
               (Measures.Sales,
     Customer.Country.CurrentMember.Parent),
     FORMAT_STRING = '#0.00%'

SELECT {Measures.Sales, Measures.PercentSales} ON COLUMNS,
     ORDER(CROSSJOIN(Time.Year.Members,
Customer.Country.Members), PercentSales, DESC) ON ROWS
FROM Orders
```

Year	Country	Sales	PercentSales
1998	USA	$ 92,633.67	21.02%
	Germany	$ 77,557.32	17.60%
	Austria	$ 45,000.65	10.21%
	Brazil	$ 44,835.77	10.18%
	UK	$ 22,623.54	5.13%
	Venezuela	$ 20,667.61	4.69%
	Ireland	$ 20,402.12	4.63%
	Sweden	$ 20,398.23	4.63%
	France	$ 18,722.18	4.25%
	Belgium	$ 16,083.68	3.65%
	Canada	$ 11,525.55	2.62%
	Switzerland	$ 9,147.12	2.08%
	Spain	$ 8,028.60	1.82%
	Italy	$ 6,843.80	1.55%
	Argentina	$ 6,302.50	1.43%
	Mexico	$ 4,544.90	1.03%
	Denmark	$ 4,516.09	1.02%
	Norway	$ 3,976.75	0.90%
	Portugal	$ 2,691.70	0.61%
	Finland	$ 2,257.00	0.51%
	Poland	$ 1,865.10	0.42%

1997	Germany	$ 117,320.16	19.01%
	USA	$ 114,845.26	18.61%
	Austria	$ 57,401.84	9.30%
	France	$ 45,263.38	7.34%
	Brazil	$ 41,941.19	6.80%
	Canada	$ 31,298.06	5.07%
	Sweden	$ 27,163.68	4.40%
	UK	$ 27,074.10	4.39%
	Venezuela	$ 26,404.92	4.28%
	Denmark	$ 25,192.54	4.08%
	Ireland	$ 20,454.41	3.31%
	Switzerland	$ 18,380.82	2.98%
	Mexico	$ 14,349.28	2.33%
	Finland	$ 13,437.29	2.18%
	Belgium	$ 11,434.48	1.85%
	Italy	$ 7,946.42	1.29%
	Spain	$ 6,978.40	1.13%
	Portugal	$ 6,474.52	1.05%
	Argentina	$ 1,816.60	0.29%
	Poland	$ 1,207.85	0.20%
	Norway	$ 700.00	0.11%

1996	USA	$ 38,105.68	18.31%
	Germany	$ 35,407.15	17.02%
	Austria	$ 25,601.35	12.30%
	Brazil	$ 20,148.82	9.68%
	France	$ 17,372.76	8.35%
	Venezuela	$ 9,738.10	4.68%
	UK	$ 9,273.68	4.46%
	Ireland	$ 9,123.38	4.38%
	Canada	$ 7,372.68	3.54%
	Sweden	$ 6,933.23	3.33%
	Belgium	$ 6,306.70	3.03%
	Mexico	$ 4,687.90	2.25%
	Switzerland	$ 4,164.72	2.00%
	Finland	$ 3,115.76	1.50%
	Spain	$ 2,976.20	1.43%
	Denmark	$ 2,952.40	1.42%
	Portugal	$ 2,306.14	1.11%
	Norway	$ 1,058.40	0.51%
	Italy	$ 979.94	0.47%
	Poland	$ 459.00	0.22%

2.5. PRD Report

```
SELECT Measures.Sales ON COLUMNS,
     NON EMPTY ORDER(Product.[Product Name].Members,
Measures.Sales, DESC) ON ROWS
FROM Orders
```

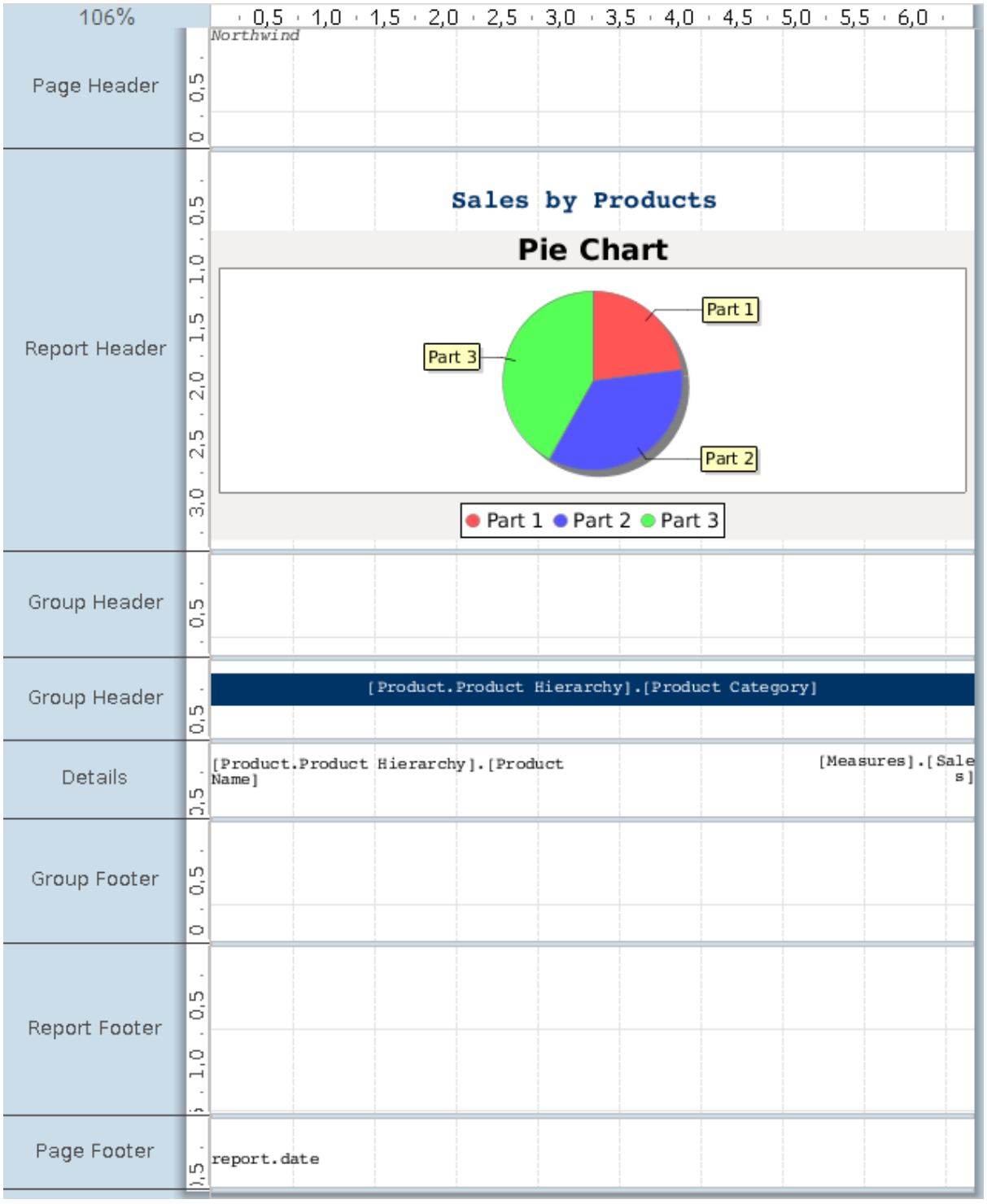


Fig. 2.5.1. Design mode

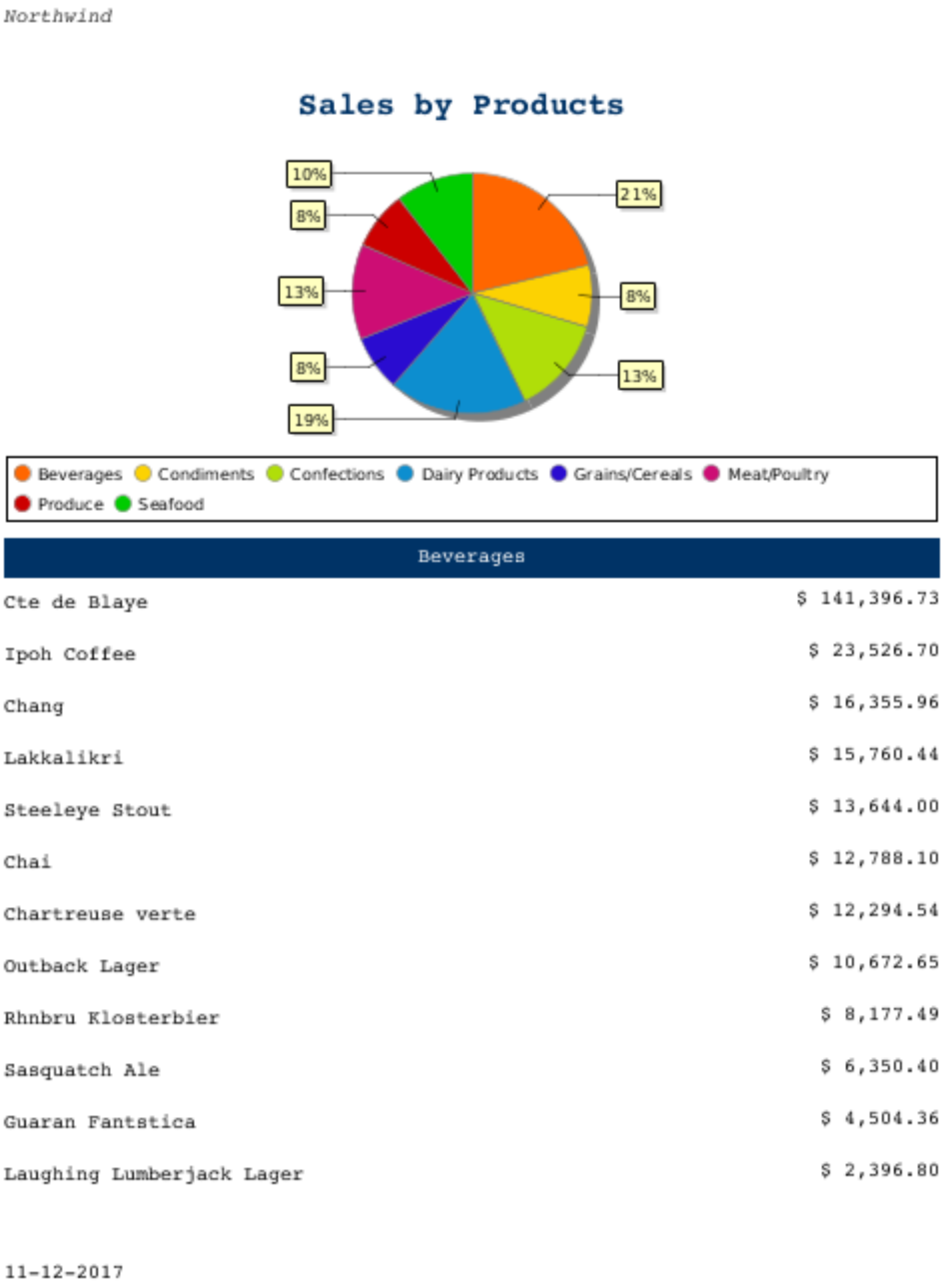


Fig. 2.5.2. Preview mode