Università degli Studi di Modena e Reggio Emilia
Facoltà di Ingegneria “Enzo Ferrari” di Modena

Corso di Laurea in Ingegneria Informatica (270/04)

Sistemi open source di integrazione di dati: MOMIS e Pentaho Data Integration a confronto

Relatore:

Dott. Ing. Laura Po

Candidato:

Marco Maria Santese

Obiettivo

- Comparare le funzionalità di due sistemi di integrazione di dati open source: MOMIS e Pentaho Data Integration
- Fornire un'analisi dei tempi per la creazione del processo di integrazione
- Mostrare i pregi e i difetti dei due sistemi di integrazione, analizzando le funzionalità messe a disposizione

MOMIS

- **MOMIS** (**M**ediator envir**O**nment for **M**ultiple **I**nformation **S**ources) Software sviluppato dalla collaborazione del DBGroup e DataRiver che sfrutta la semantica presente nelle sorgenti informative per integrare sorgenti dati eterogenee
- Processo Semi-Automatico per l'integrazione
- Approccio GAV (Global As View)
- Basato sul DataBase lessicale WordNet (Princeton University, *George A. Miller*)

PENTAHO Data Integration

- Precedentemente chiamato Kettle: "Kettle Extraction, Transport, Transformation and Loading Environment " : è un modulo del software PENTAHO che consente di estrarre, trasformare e caricare (ETL) i dati da una qualsiasi fonte
- Ambiente di sviluppo grafico

Test

Il test per analizzare il funzionamento di questi due sistemi di integrazione di dati è stato tratto dalla guida Oracle®*:*

“Data Integrator: Getting Started with an ETL Project”.

Obiettivo del test

Partendo dai dati contenuti in:

- *src_customer* (*tabella* contenente i clienti)
- *trg_city (tabella* delle città «ammesse»*)*
- *src_sales_person* (file *Excel* contenente i dati dei venditori)
- *src_age_group (* file *Excel* contenente il range dell’età dei clienti)

Popolare una tabella di destinazione

- *trg_customer*

Sorgenti

Tabella CLIENTI

SRC_CUSTOMER			
CUST_ID	NUMERIC (10)	<PK>	NOT NULL
DEAR	NUMERIC (1)		NULL
LAST_NAME	VARCHAR (50)		NULL
FIRST_NAME	VARCHAR (50)		NULL
ADDRESS	VARCHAR (100)		NULL
CITY_ID	NUMERIC (10)		NULL
PHONE	VARCHAR (50)		NULL
AGE	NUMERIC (3)		NULL
SALES_PERS_ID	NUMERIC (10)		NULL

TRG_CITY			
CITY_ID	NUMERIC(10)	<pk>	not null
REGION_ID	NUMERIC(10)	<fk>	not null
CITY	VARCHAR(50)		null
POPULATION	NUMERIC(10)		null

Tabella Città Ammesse

Venditori

SRC_SALES_PERSON			
SALES_PERSON_ID	NUMERIC (10)	<PK>	NOT NULL
FIRST_NAME	VARCHAR (50)		NULL
LAST_NAME	VARCHAR (50)		NULL
HIRE_DATE	DATE		NULL

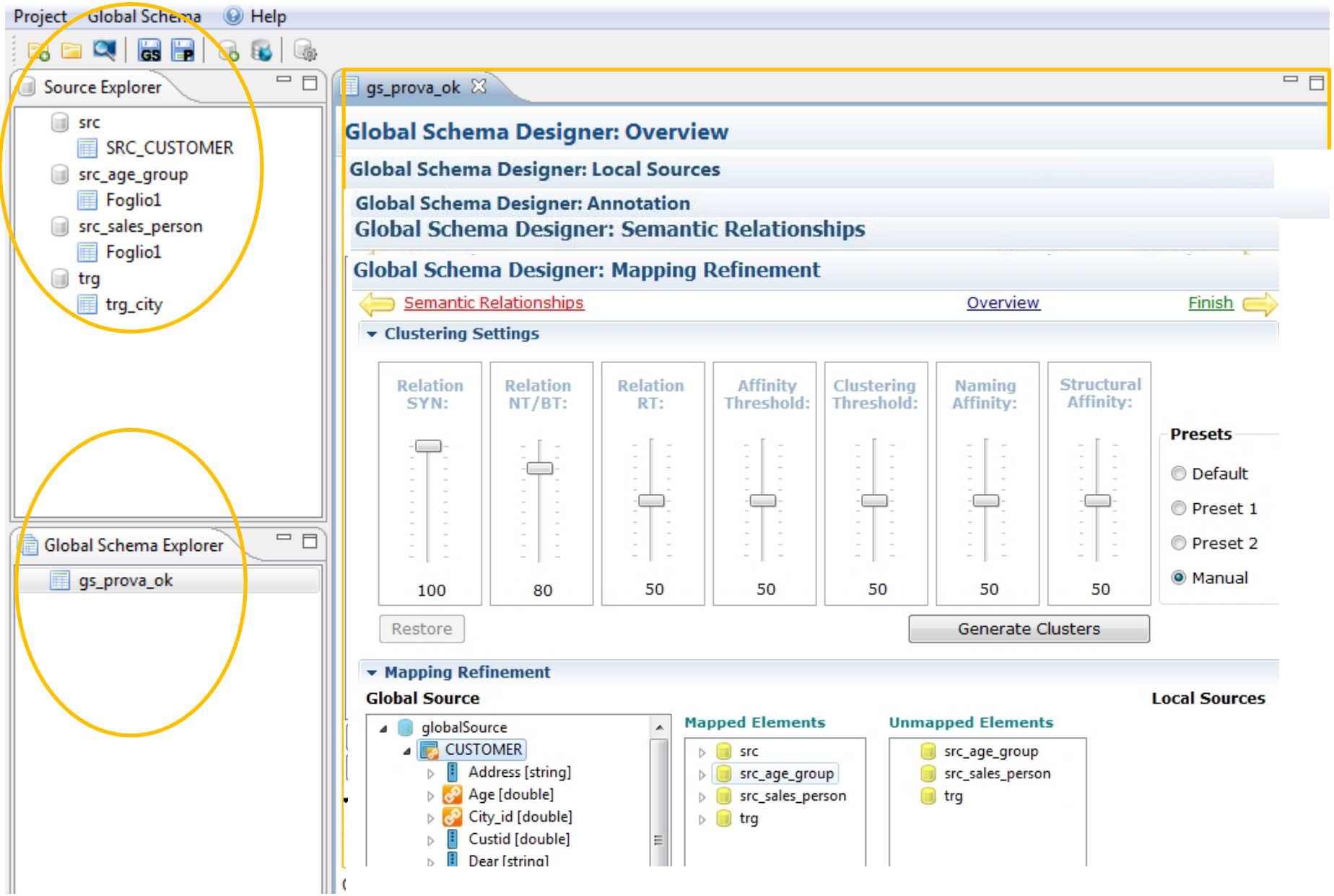
Età Ammesse

SRC_AGE_GROUP		
AGE_MIN	NUMERIC (3)	<PK> NOT NULL
AGE_MAX	NUMERIC (3)	<PK> NOT NULL
AGE_RANGE	VARCHAR (50)	NULL

Destinazione

TRG_CUSTOMER			
CUST_ID	NUMERIC (10)	<PK>	NOT NULL
DEAR	VARCHAR (10)		NULL
CUST_NAME	VARCHAR (100)		NULL
SALES_NAME	VARCHAR (100)		NULL
ADDRESS	VARCHAR (100)		NULL
CITY_NAME	VARCHAR (50)		NULL
PHONE	VARCHAR (50)		NULL
AGE	NUMERIC (3)		NULL
SALES_PERS_ID	NUMERIC (10)		NULL

Test effettuato con MOMIS



Continua test MOMIS

Analizziamo il passo di Mapping Refinement:

- Tramite la generazione automatica dei Cluster ottengo:

Mapping Table: trg_city

trg_city(globalSource)	Foglio1(src_age_group)	Foglio1(src_sales_person)	SRC_CUSTOMER(src)	trg_city(trg)
AGE_MIN	AGE_MIN			
AGE_RANGE	AGE_RANGE			
Address			Address	
Age	AGE_MAX		Age	
Dear			Dear	
First_name		FIRST_NAME	First_name	
HIRE_DATE		HIRE_DATE		
LAST_NAME		LAST_NAME	Last_name	
Phone			Phone	
SALES_PERSON_ID		SALES_PERSON_ID	Sales_pers_id , Cu...	region_id
city				city
city_id			City_id	city_id
population				population

Non rispetta le specifiche

- Imposto gli attributi di Join e correggo la Join Function calcolata da Momis in automatico

Join Function about: CUSTOMER

Join Function

```
src_age_group.Foglio1 full outer join src.SRC_CUSTOMER on 0=0  full outer join
src_sales_person.Foglio1 on (((src_sales_person.Foglio1.SALES_PERSON_ID) =
(src.SRC_CUSTOMER.Sales_pers_id)
= (src.SRC_CUSTOMER.City_id)) OR
```

Phone		
Sales_pers_id		
city_name		
customer_name		
max	f	AGE_MAX
min	f	AGE_MIN
seller_name		

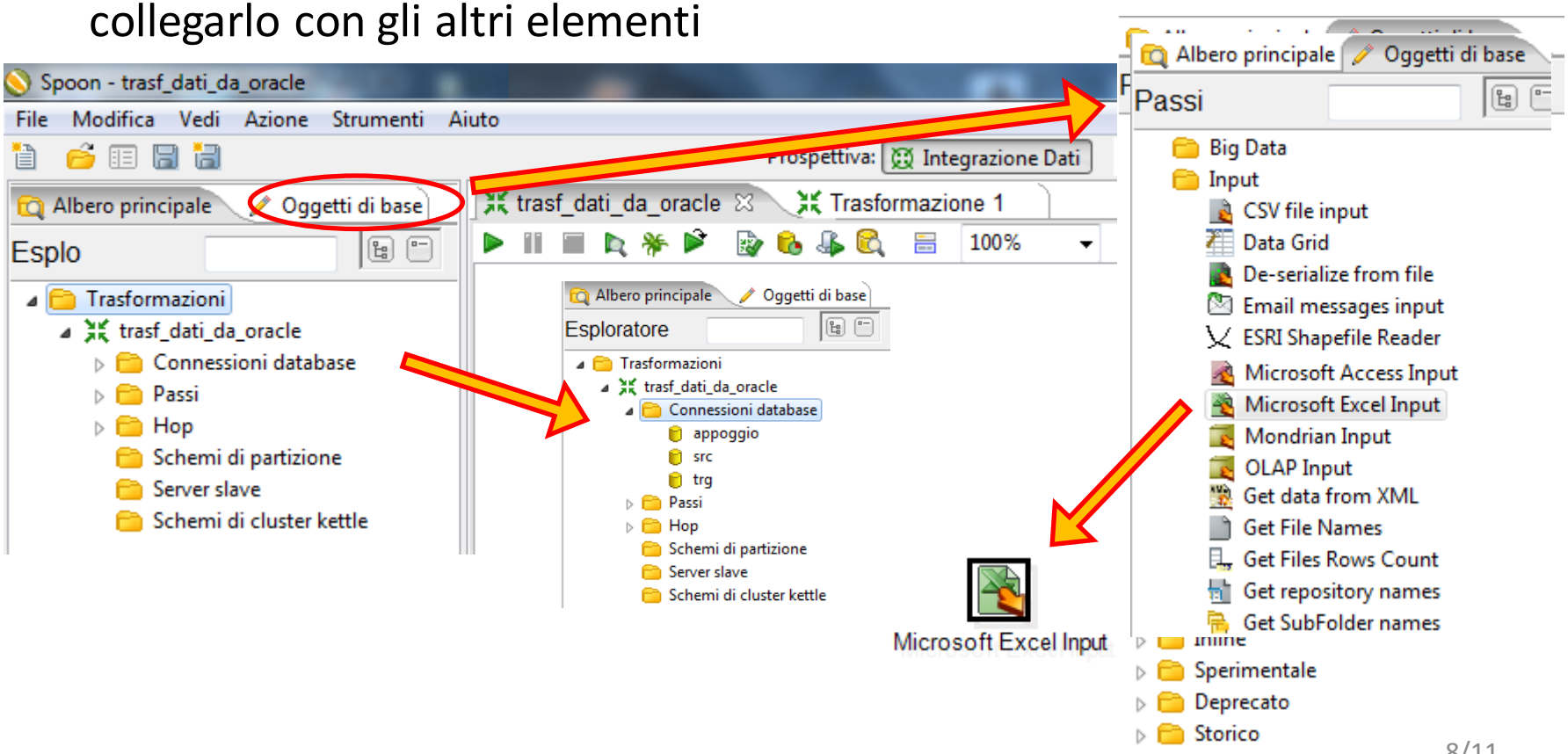
Join Function about: CUSTOMER

Join Function

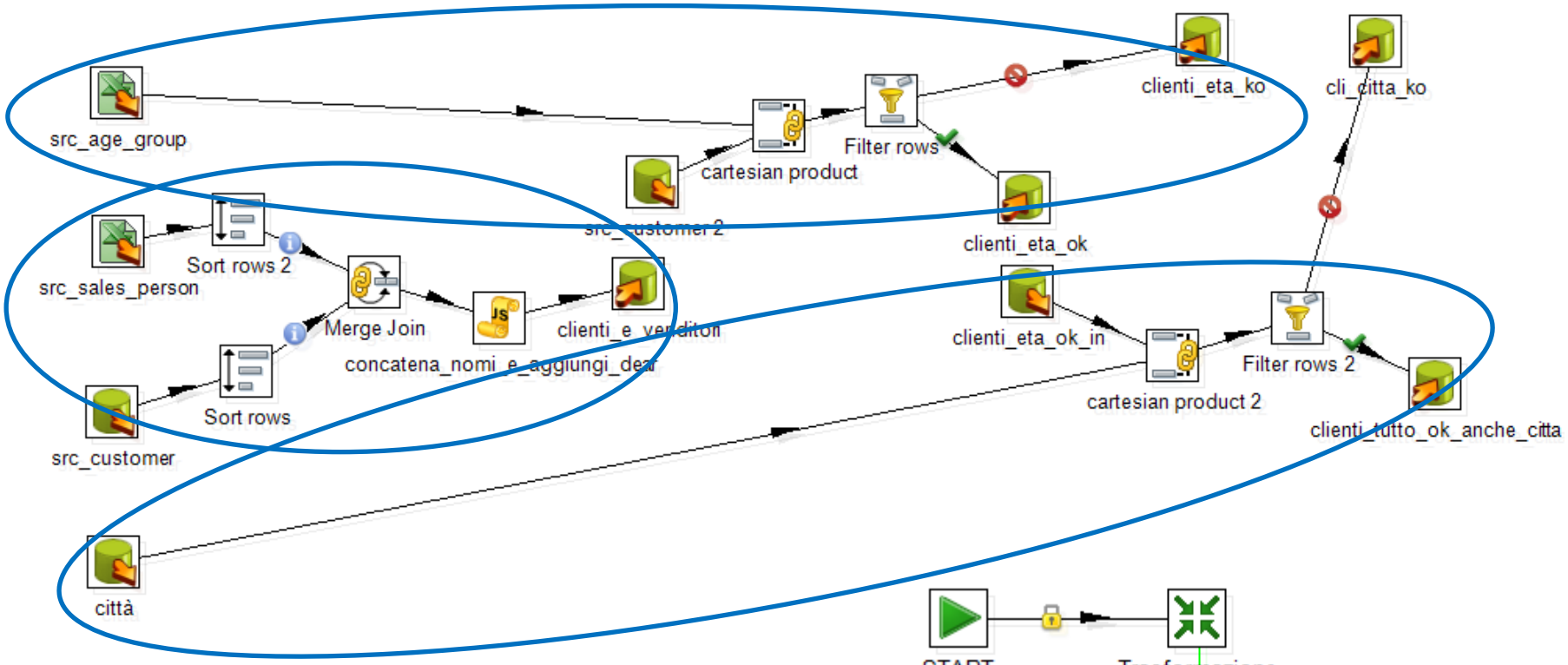
```
src.SRC_CUSTOMER inner join src_sales_person.Foglio1 on
(((src_sales_person.Foglio1.SALES_PERSON_ID) = (src.SRC_CUSTOMER.Sales_pers_id)))
inner join trg.trg_city on
(( (src.SRC_CUSTOMER.City_id)=(trg.trg_city.city_id) ) ) inner join src_age_group.Foglio1
on
((src.SRC_CUSTOMER.Age between src_age_group.Foglio1.AGE_MIN and
src_age_group.Foglio1.AGE_MAX ))
```

Test eseguito con Pentaho D.I.

- L'integrazione di dati da sorgenti eterogenee è stata ottenuta tramite una serie di trasformazioni
- La prima azione da effettuare è quella di aggiungere al progetto le connessioni alle varie sorgenti
- Selezionare «oggetti di base», espandere la classe di nostro interesse, selezionare il passo per poi trascinarlo nello spazio di lavoro e collegarlo con gli altri elementi



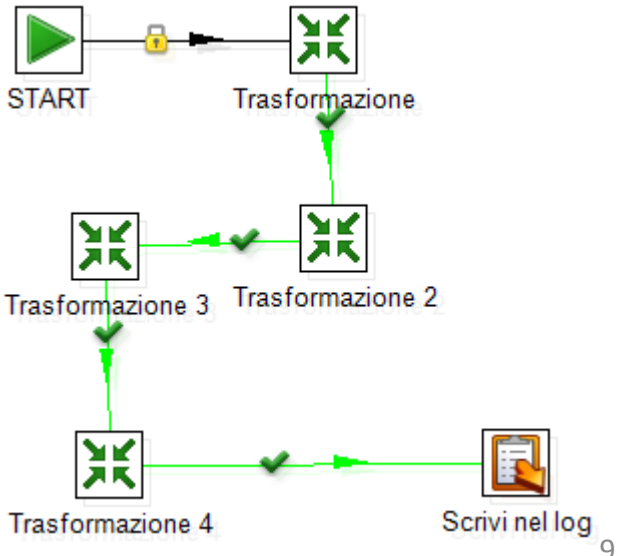
Continua test Pentaho D.I.



Azioni effettuate:

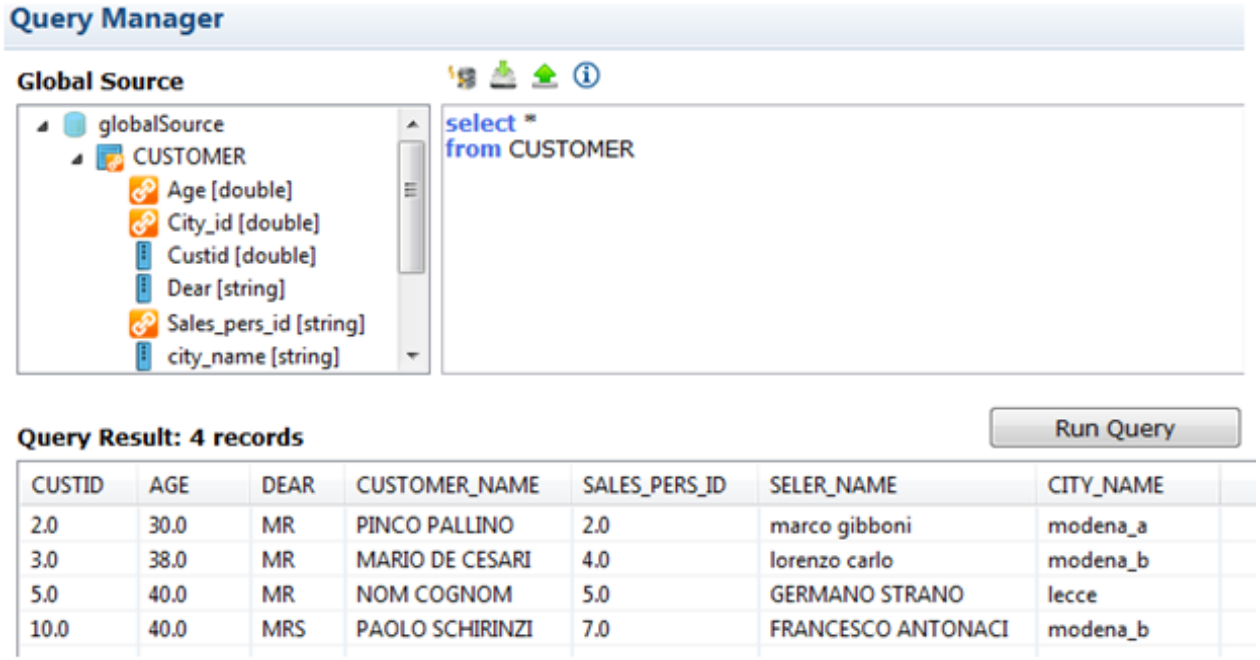
- 3 Join
- 6 Select
- 3 Tabelle ausiliarie per dati parziali
- 1 Script Java
- 2 Filtri

Organizzando in un Job tutte le trasformazioni



Risultato Integrazione

MOMIS



CUSTID	AGE	DEAR	CUSTOMER_NAME	SALES_PERS_ID	SELER_NAME	CITY_NAME
2.0	30.0	MR	PINCO PALLINO	2.0	marco gibboni	modena_a
3.0	38.0	MR	MARIO DE CESARI	4.0	lorenzo carlo	modena_b
5.0	40.0	MR	NOM COGNOM	5.0	GERMANO STRANO	lecce
10.0	40.0	MRS	PAOLO SCHIRINZI	7.0	FRANCESCO ANTONACI	modena_b

Pentaho Data Integration

```
select * from CUSTOMER
```

Results | Messages

Custid	Address	City_id	Phone	Age	city	dear_v	cust_name	sales_pers	SALES_PERSON_ID
3	VIA MARE	41126	895555	38	modena_b	MR	lorenzo carlo	MARIO DE CESARI	4
10	VIA MANO	41126	23133	40	modena_b	MRS	FRANCESCO ANTONACI	PAOLO SCHIRINZI	7
5	VIA STRANA	73100	832510858	40	lecce	MR	GERMANO STRANO	NOM COGNOM	5

Conclusioni:
Vantaggi e Svantaggi

- **MOMIS**

(+) Processo di integrazione guidato

(+) Non necessita di conoscenze relative alla Programmazione

(-) Le fasi del processo non possono essere modificate

- **Pentaho Data Integration**

(+) Supporto a molte sorgenti di dati

(+) Estrema libertà nella definizione del processo

(-) Processo di integrazione completamente manuale

(-) L'utente deve studiare a fondo lo strumento prima di poterlo usare ed inoltre necessita di conoscenze di programmazione

Fine

Conclusioni

MOMIS

- Indirizzato a degli utenti che vogliono arrivare ad un risultato in pochi passaggi
- Integrazione di dati semi-automatica
- Intervenire manualmente solo per aggiustare e/o migliorare il risultato dell'integrazione

Pentaho Data Integration

Utile per chi ha la necessita di:

- Intergare dati partendo da sorgenti molto varie, senza voler utilizzare altri software per fare un primo passo di "omogeneizzazione "
- Effettuare studi di Business Intelligence (**BI**)

	MOMIS	PENTAHO D. I.
Facilità d'utilizzo	✓ ✓ ✓	✓
Funzionalità	✓	✓ ✓ ✓
Conoscenze necessarie per l'utilizzo	✓ ✓	✓ ✓ ✓
Forum di Supporto		✓ ✓
Tutorial	✓ ✓ ✓	✓
Materiale on-line	✓	✓ ✓ ✓

Funzioni disponibili	MOMIS	PENTAHO D.I.
Importazione Risorse	SI (DB2, Microsoft SQLServer , MySQL , Oracle , PostgreSQL Database, Sorgenti JDBC / JDBC-ODBC , File Excel, File CSV, Altre risorse tramite WEB)	SI (oltre 25 DB oltre che varie tipologie di file)
Utilizzo Risorse Remote	SI	SI
Plug-in per introdurre nuove funzionalità	NO	SI
Progetto dimostrativo	SI	NO
Integrazione basata su relazioni semantiche	SI (WordNet)	NO (Weka *)
Possibilità di interrogare il risultato dell'integrazione	SI (strumento integrato)	NO (Sever DBMS esterno)
Materializzazione risultato	NO	SI