TALEND_ETL LEGNAIA_FURLA_TESI_BREGATA

Nome progetto	Talend Open Studio per Furla
Documento	Talend_ETL legnaia_FURLA_TESI_BREGATA v1.0
Oggetto	-
Status	Confidenziale
Versione	01.00.00
Data emissione	1/04/2019
Data revisione	-
Autore	Bregata
Destinatario	-

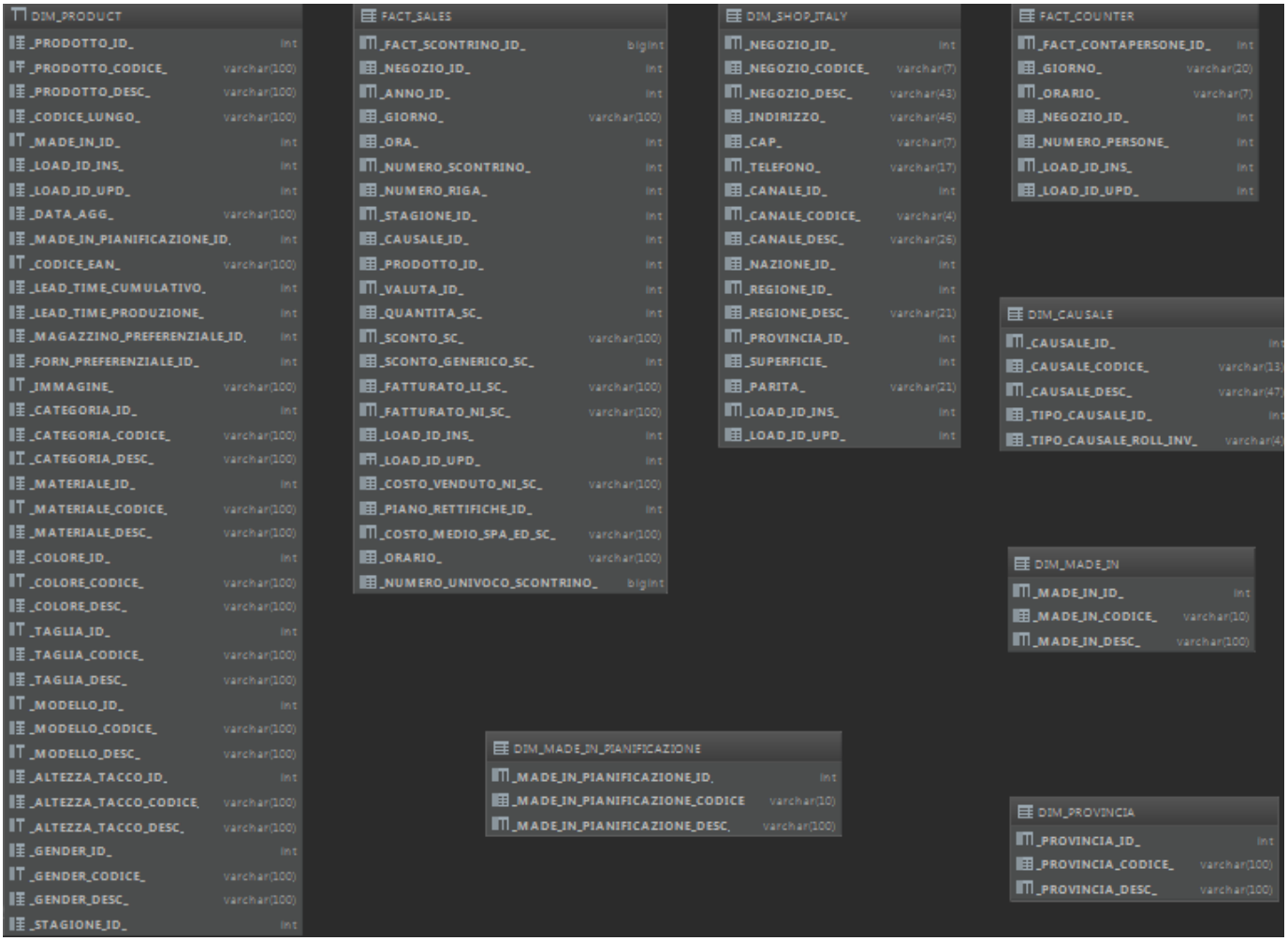
Mediamente Consulting Srl
www.mediamenteconsulting.it
Via Piovola, 138 50053 Empoli (Firenze)
Part. IVA e Cod. Fisc. 06302280489
info@mediamenteconsulting.it
Tel. 0571 9988 | Fax 0571 993366

SOMMARIO

OBIETTIVO

Furla è già un nostro cliente su effettuiamo attività legate al CPM e all'ETL per la parte delle Vendite.

Si vuole proporre una fase di ETL che riguarda l'utilizzo di strumenti di Advanced Analytics per ottenere una FACT_SALES più specifica con dati che seguono il sistema ACID, che attualmente il cliente non conosce e che potrebbero produrre un potenziale vantaggio economico e competitivo.



DATI DI PARTENZA

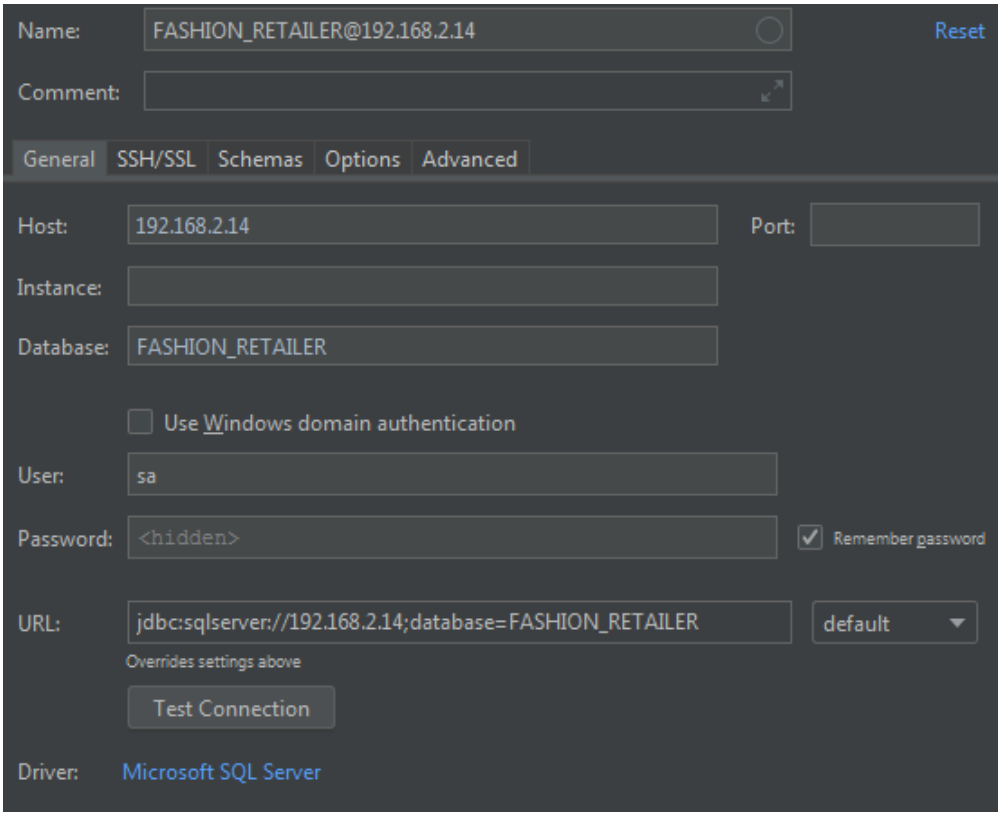
Come dimensioni sono state create le seguenti tabelle ma prive di dati all'interno, perché andremo noi a riempirle tramite l'uso di Talend:

DATABESE TABLE	METADATA	NAME METADATA
DIM_CAUSALE	FILE: .csv	Causale
DIM_MADE_IN	FILE: .csv	Made_In
DIM_MADE_IN_PIANIFICAZIONE	FILE: .csv	Made_In_Pianificazione
DIM_PRODOTTO	FILE: .csv	Prodotto1, Prodotto2, Prodotto3,
DIM_SHOP	FILE: .excel	Negozio
FACT_SCONTRINO	FILE: .csv	Scontrino
DIM_PROVINCIA	FILE: .csv	Provincia
FACT_CONTAPERSONE	FILE: .csv	Contapersone

SOLUZIONE

DATA INGESTION

I dati sono stati estratti dai cubi TM1 del cliente ed importati su SQL server tramite esportazione/importazione di file csv.,excel e nel caso della DIM_SCONTRINO direttamente dal database già creato in precedenza. Il database SQL Server è installato sulla macchina 192.168.2.14 e si trova sullo schema LEGNAIA.
La visione su DATAGRIP è illustrasta nella seguente immagine.



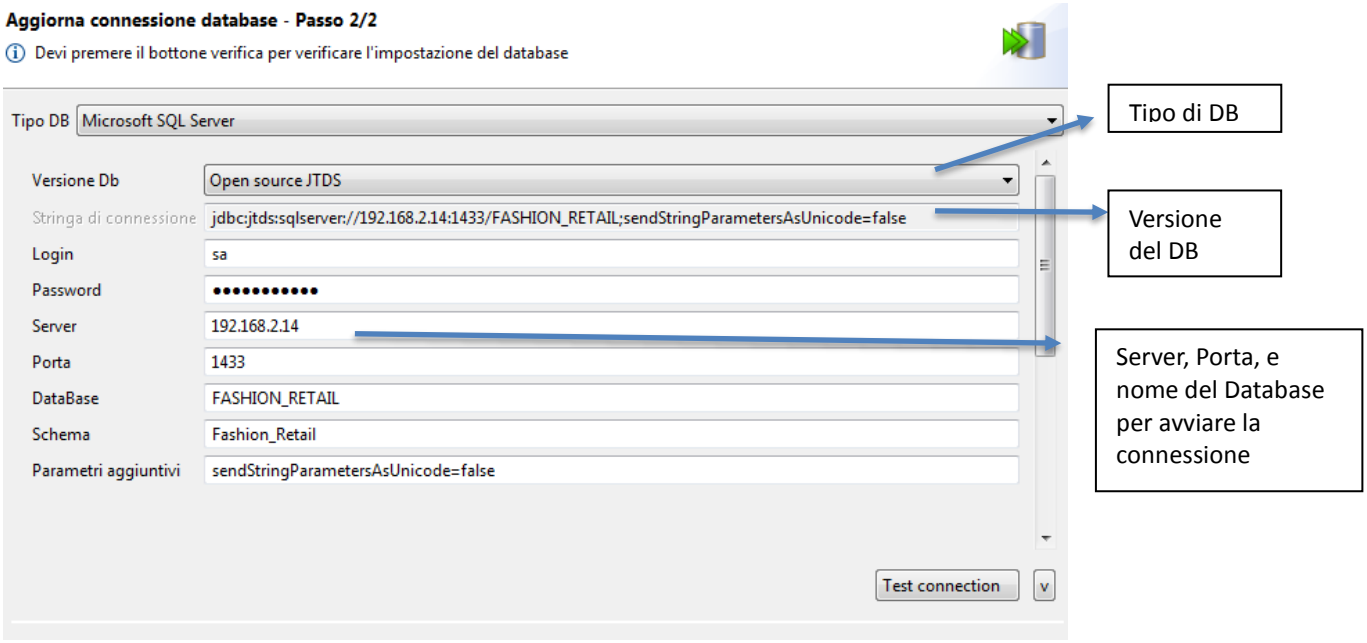
METADATI

Tramite le funzionalità del software ETL Talend Open Studio è stato possibile importare vari tipi di file per costruire un Nuovo database più Semplice, Ma utile per valutare tutto il reparto vendite di LEGNAIA.
Per usare il database su Talend ho bisogno di creare una connessione al database Legnaia dove poi costruirò le mie Dimensioni e la consegunte Fact table.
Per fare ciò, ho bisogno dei Metadati, ma cosa sono?
Tutte le informazioni associate a un documento possono essere organizzate in tre macrocategorie: metadati descrittivi, strutturali e gestionali.
I **metadati descrittivi** sono quelli che descrivono il contenuto del documento. Variano ovviamente in base al contenuto. Usare un metadato per il numero di protocollo e mantenerlo vuoto in tutti i documenti perché non protocollati non avrebbe senso. Questi metadati vanno adattati al contenuto che vogliamo descrivere. Per alcune tipologie di documenti fiscali alcuni metadati sono obbligatori, al fine di conservare le giuste informazioni per la mappatura del documento e per il suo immediato reperimento in caso di necessità. Una fattura emessa che non abbia il destinatario rimane un pezzo di carta inutile, e se questo destinatario c'è va mappato tramite apposita etichetta di descrizione.
I **metadati strutturali** si riferiscono invece alla persistenza fisica e logica del documento. Possono descrivere dove è stato localizzato in una cartella, o il suo codice identificativo interno, oppure contenere l'informazione dell'integrità nel tempo del contenuto. La loro mappatura permette di tenere sotto controllo dove fisicamente risiedono i file, e in che condizioni si trovano in quanto a struttura logica.

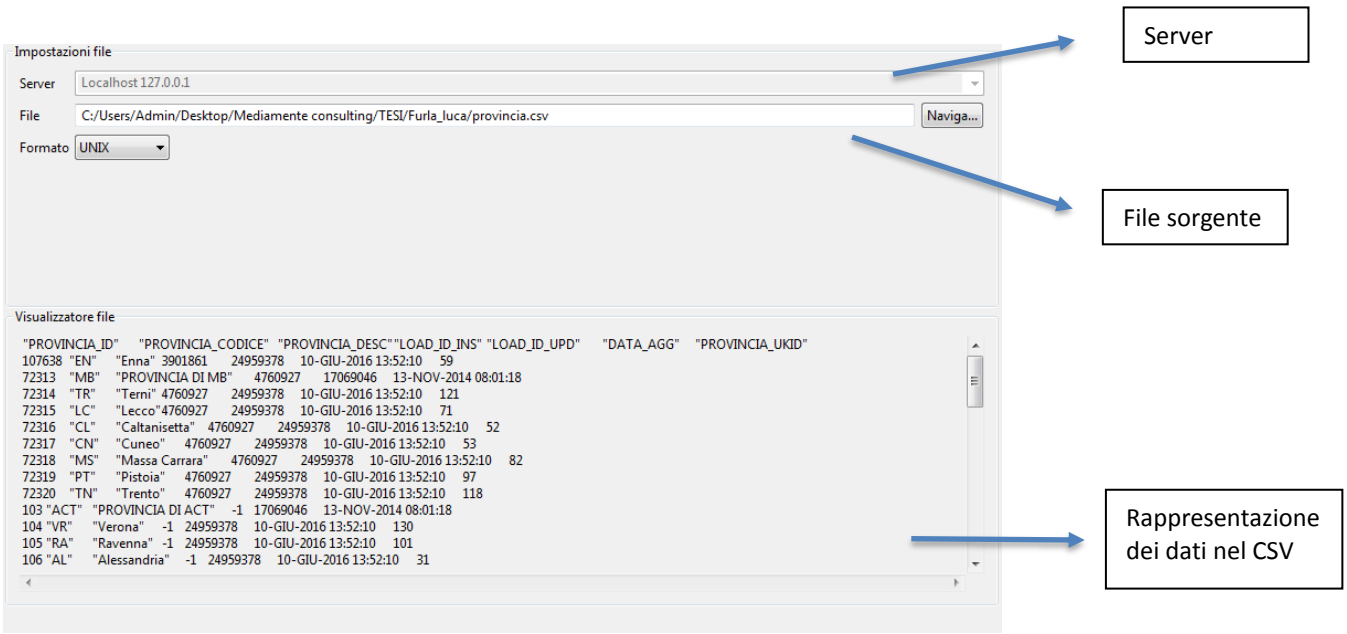
Infine i **metadati gestionali**, che come dice il nome sottendono alla gestione dei documenti e alla loro amministrazione. Possono contenere informazioni sui diritti d'accesso, su quando e come procedere allo scarto quando sarà il momento, su chi sia il custode di quei file.

Insomma, per essere molto sintetici possiamo dire che i metadati costituiscono un corredo di informazioni ai documenti informatici, e che servono alla loro descrizione e amministrazione. Un punto di partenza attraverso il quale potremo analizzare in seguito più approfonditamente quali sequenze di metadati sia possibile utilizzare, quali sono quelli obbligatori e per quali tipi di documenti vadano utilizzati.

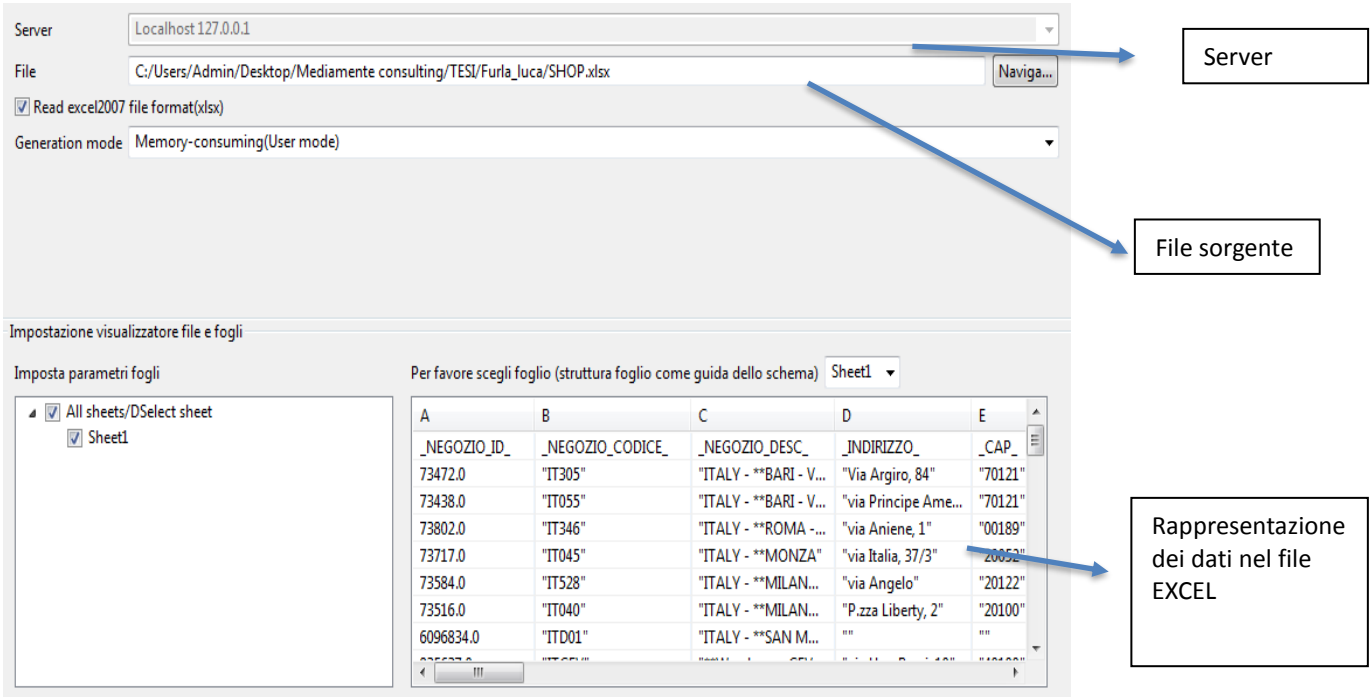
- Connessione al DB



- File .csv



- File excel

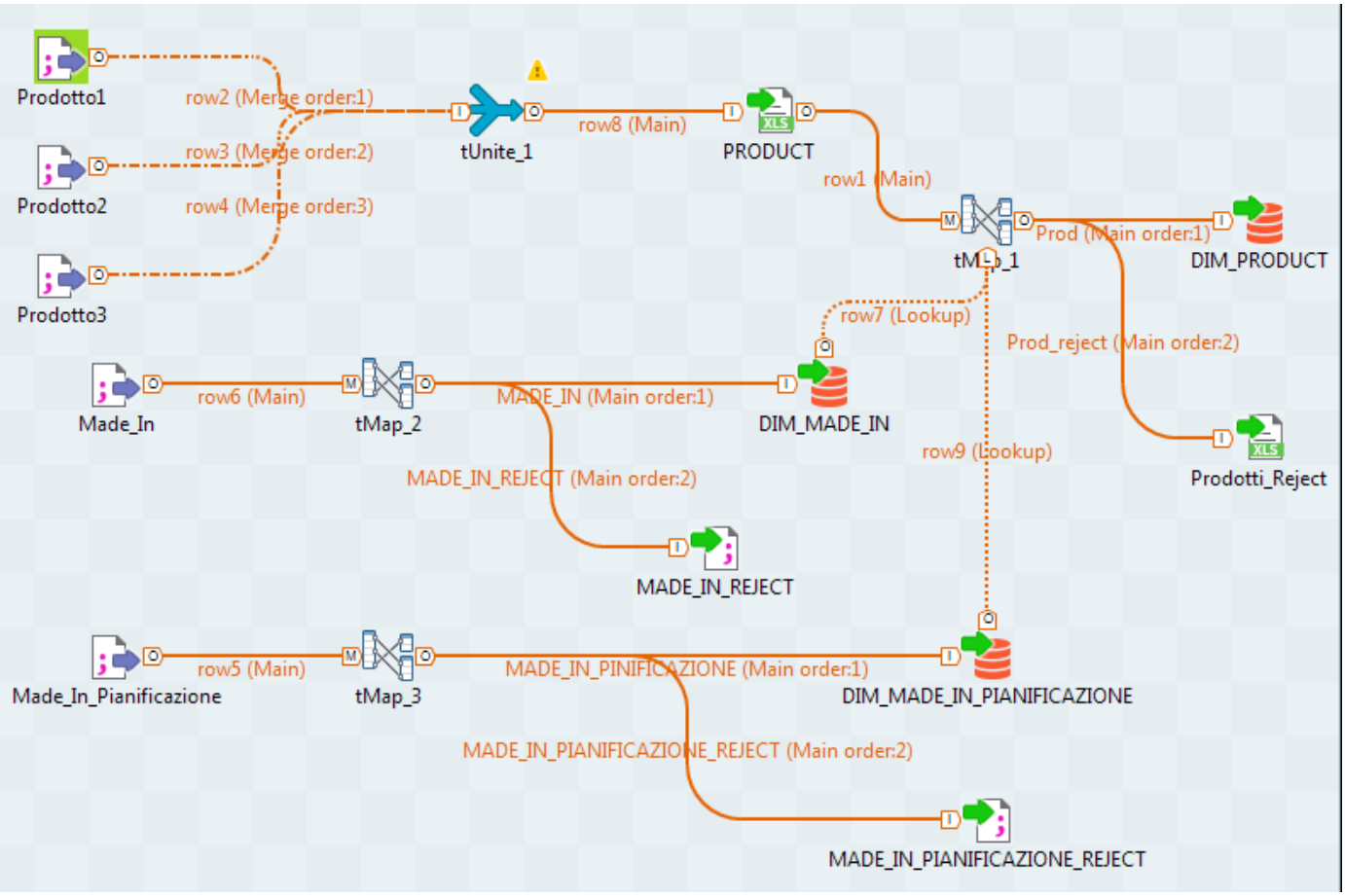


DIM_PRODUCT

I Metadati, una volta importati, devono essere rielaborati per la creazione del database sul server. Concettualmente sono state effettuate le seguenti quattro operazioni:

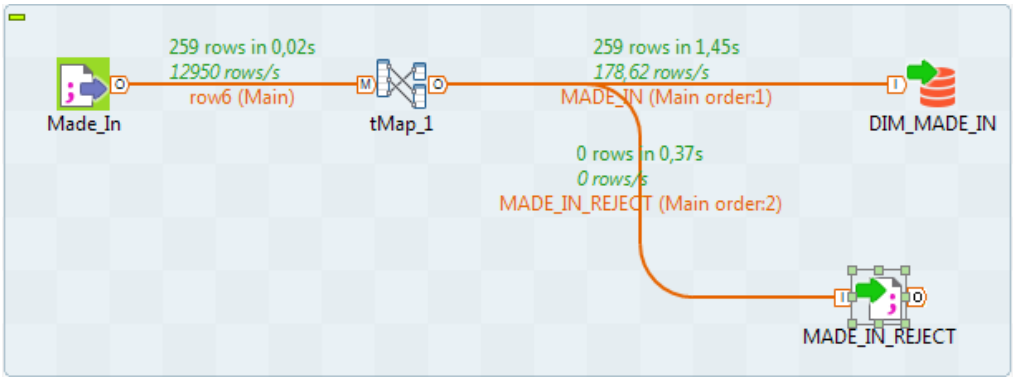
1. Importazione file .csv o Excel.
2. Unione dei file tramite lo strumento tUnite, se neccessario.
3. Cambiamento e mappatura dei nomi, lunghezza o tipo degli attributi o unione di tabelle grazie alle chiavi primarie tramite lo strumento tMap.
4. Creazione della tabella in SQL SERVER tramite lo strumento tDBOutput(tMSSQLOutput)

Per svolgere le seguenti operazioni è necessario creare un nuovo JOB, che conterrà i vari metodi di importazione.
Di seguito vediamo un esempio completo descritto dalla creazione della DIM_PRODUCT, dopo aver fatto un Inner Join con le tabelle DIM_MADE_IN_PIANIFICAZIONE e DIM_MADE_IN create precedentemente.
Per ogni tabella creata si è creato un Output dove vengono caricati i file rigettati per contrastare eventuali Errori nel caricamento.

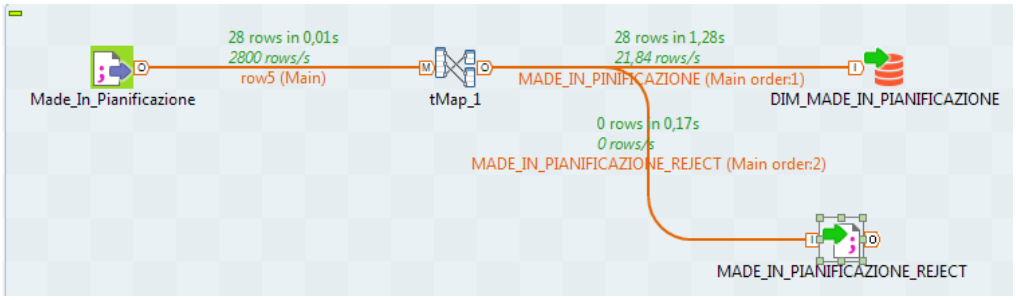


Un'altra specifica usata per controllare e prevenire gli errori possibili nel caricamento nell'ETL, è separare ogni Dimensione in un job differente.

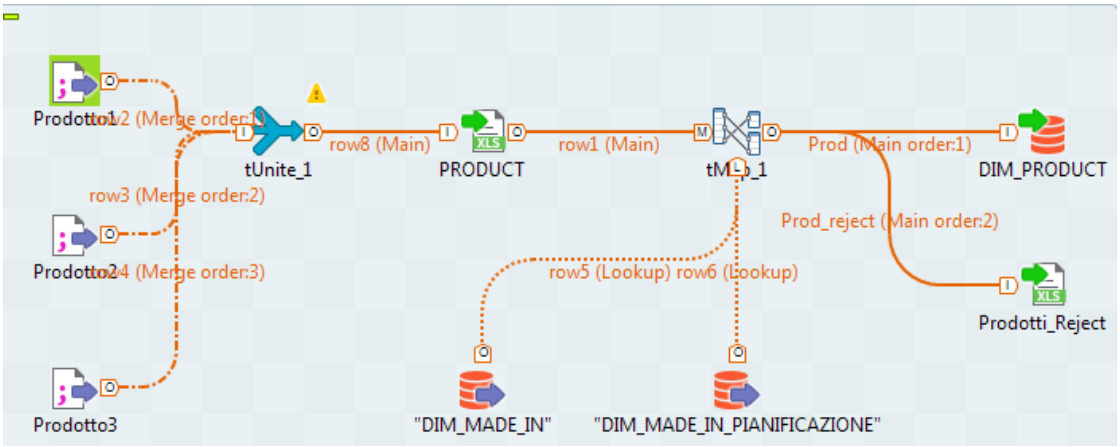
DIM_MADE_IN:



DIM_MADE_IN_PIANIFICAZIONE:

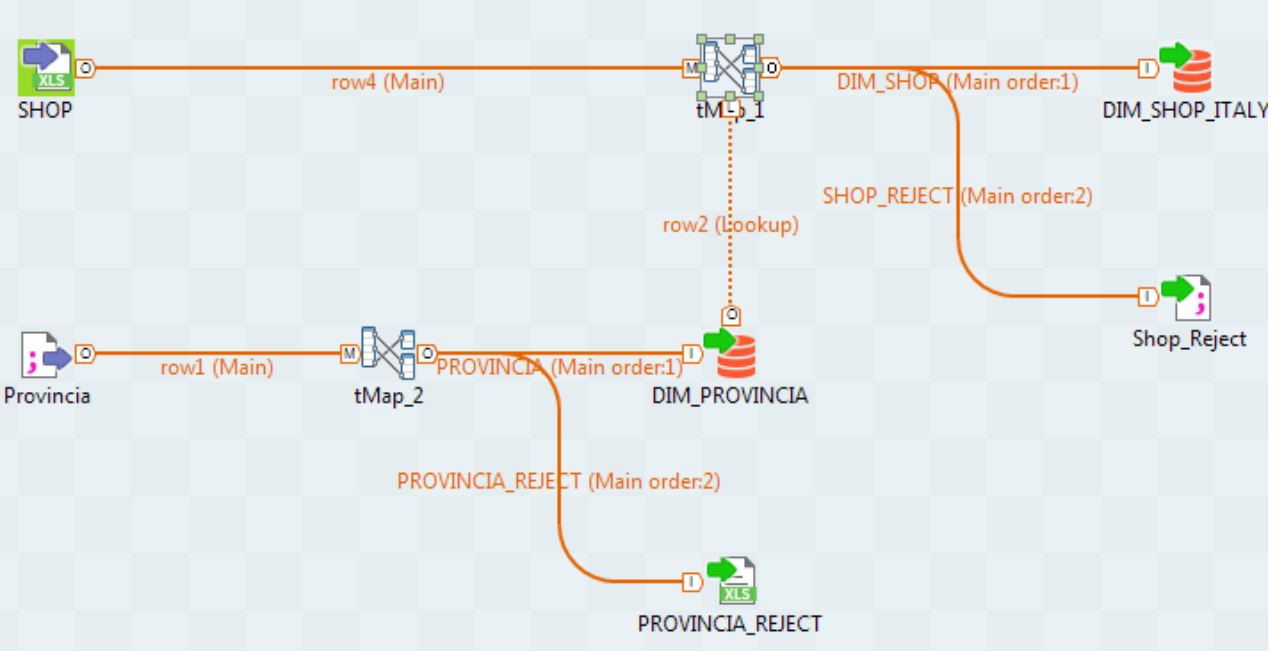


Per poi unirle Tramite L'import dei metadati direttamente dal DB sottoforma di Input.



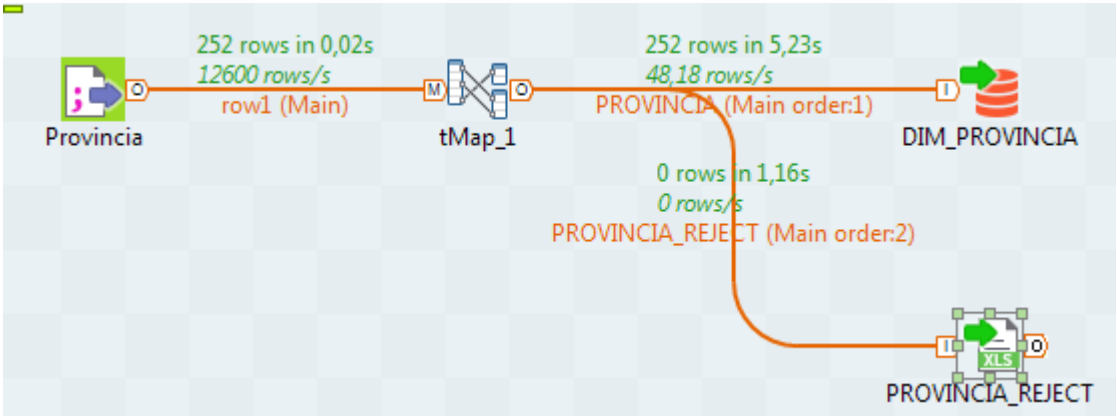
DIM_SHOP_ITALY

Per la creazione della DIM_SHOP, si eseguirà lo stesso principio elencato in precedenza, facendo un Inner Join con la tabella DIM_PROVINCIA creata precedentemente.
Come in precedenza, per ogni tabella si è crea un Output dove vengono caricati i file rigettati per contrastare eventuali Errori nel caricamento.

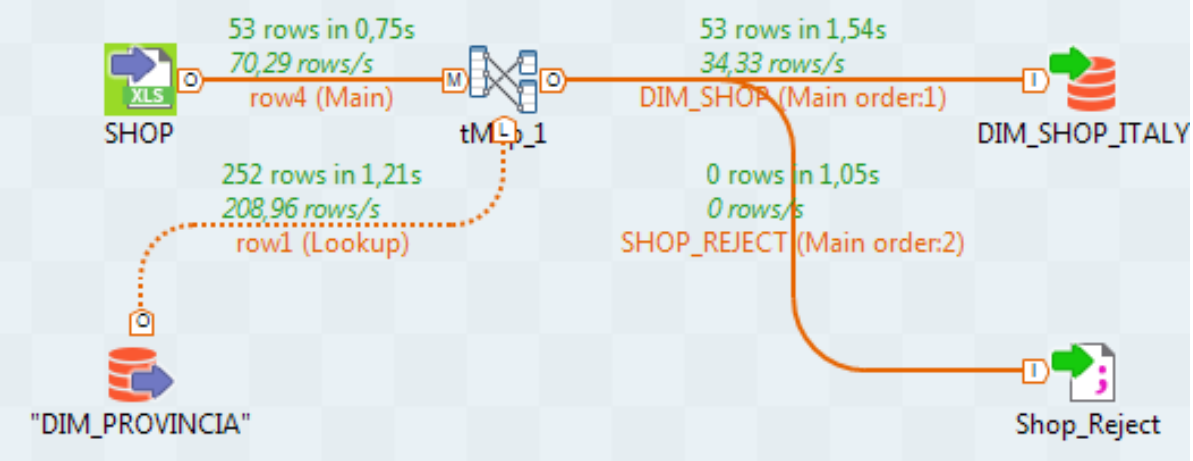


Separiamo ogni Dimensione in un job differente.

DIM_PROVINCIA:



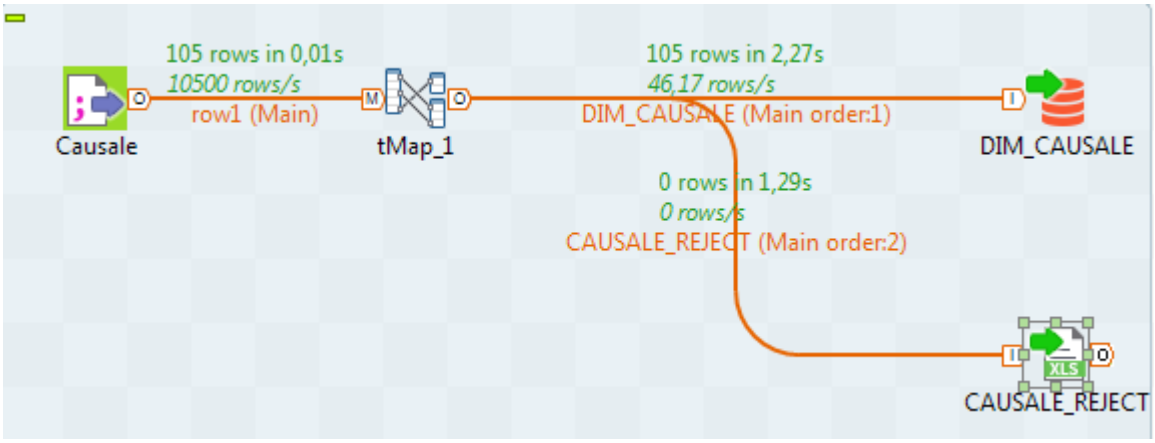
Per poi unirle Tramite L'import dei metadati direttamente dal DB sottoforma di Input.



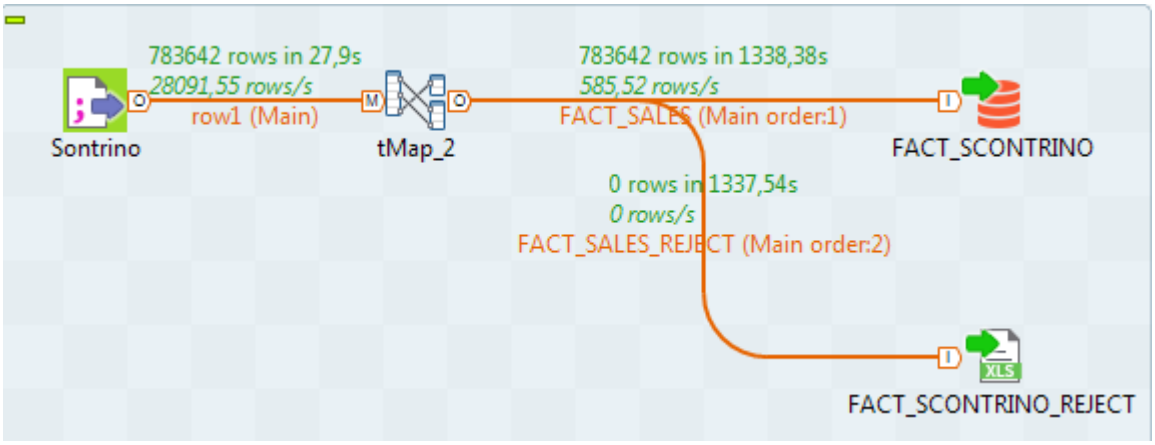
FACT_SALES

Dopo aver creato la DIM_SHOP e la DIM_PRODUCT, abbiamo tutto il neccessario per costruirci la nostra FACT_SALES, facendo un Inner Join tra la tabella DIM_CAUSALE, derivante dai metadati del file csv e la DIM_SCONTRINO, anch'essa da file delimitato.
Creiamo la tabella nel Database MSSQL e la colleghiamo tramite il tMap alla FACT_SALES, che caricheremo nel DB FASHION_RETAILER.

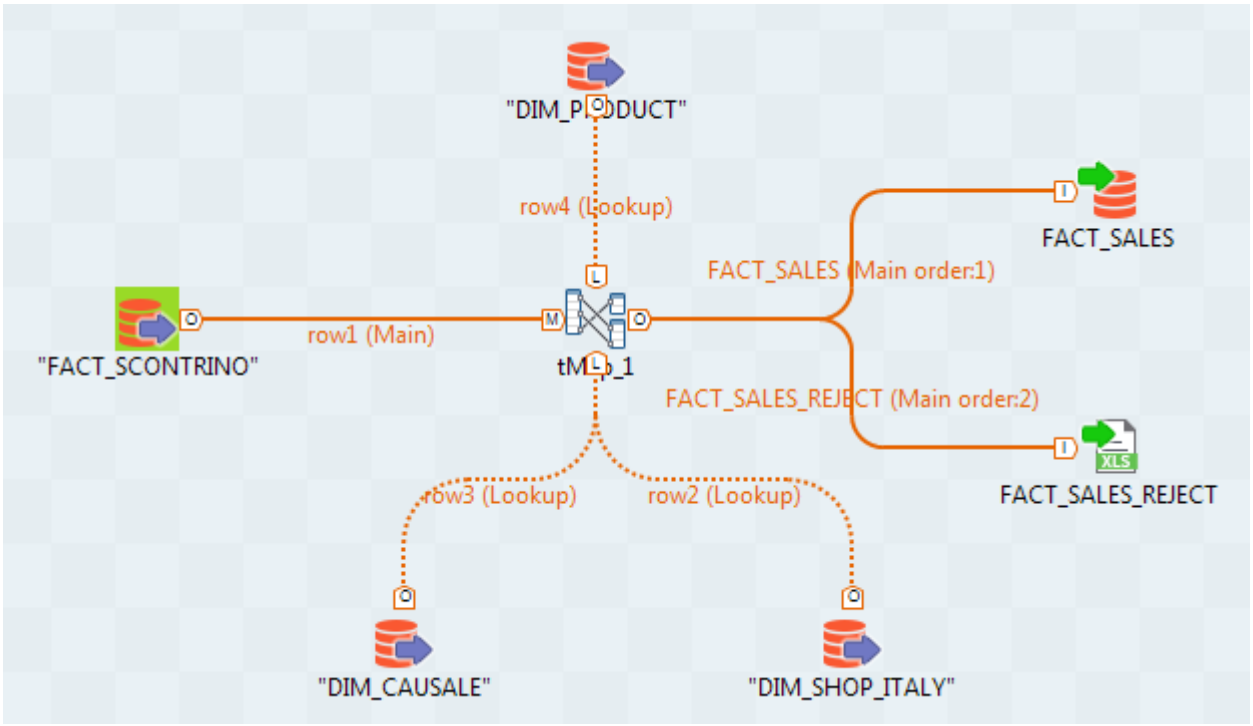
DIM_CAUSALE:



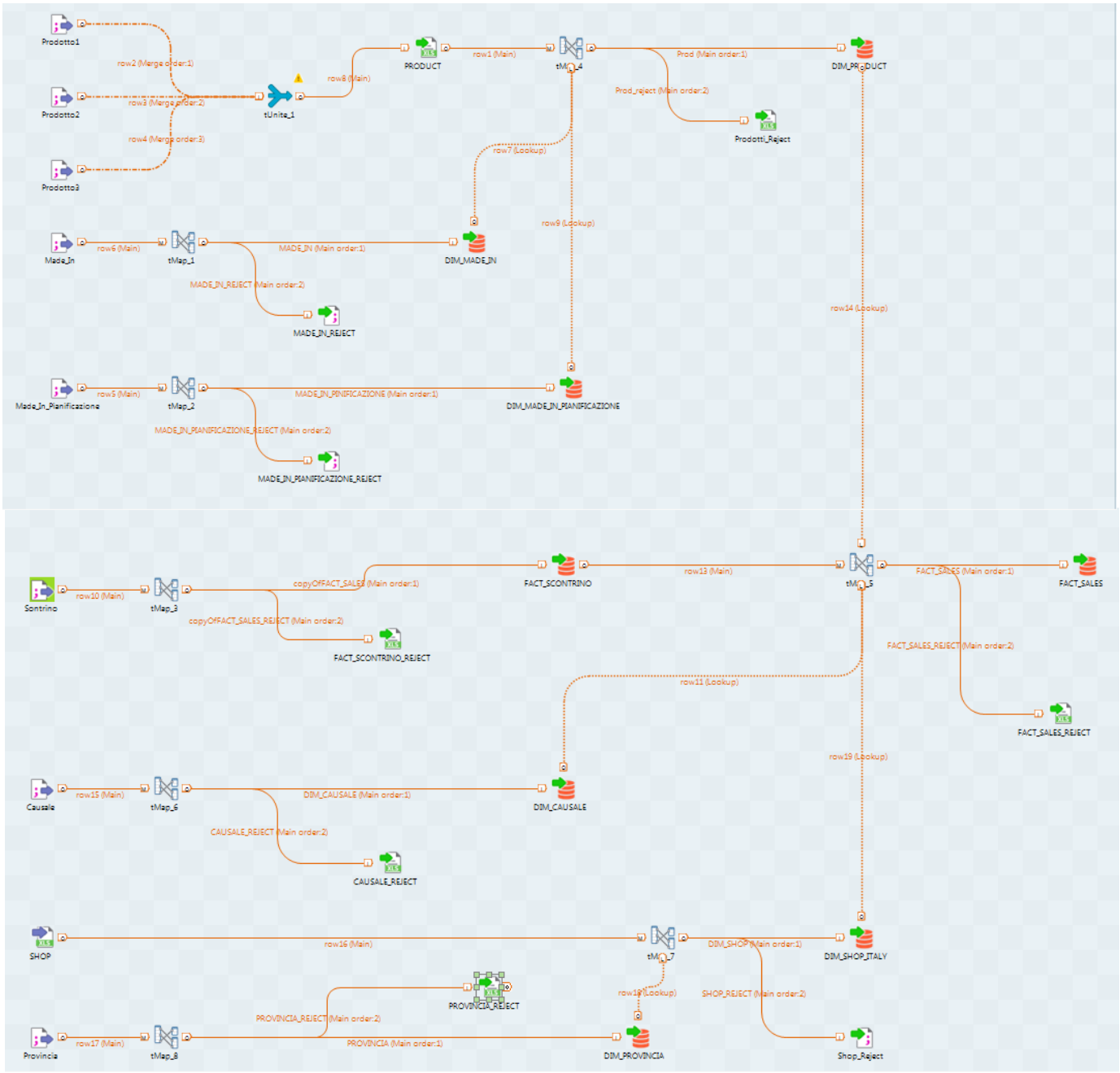
FACT_SCONTRINO:



FACT_SALES:

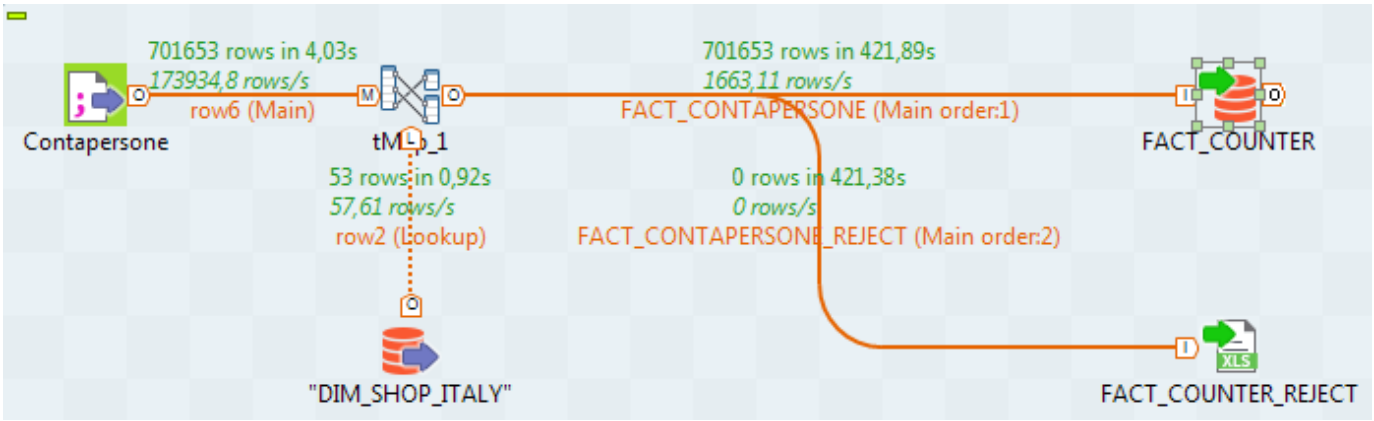


LO SCHEMA GENERALE DELLA FACT_SALES OTTENUTA è IL SEGUENTE:

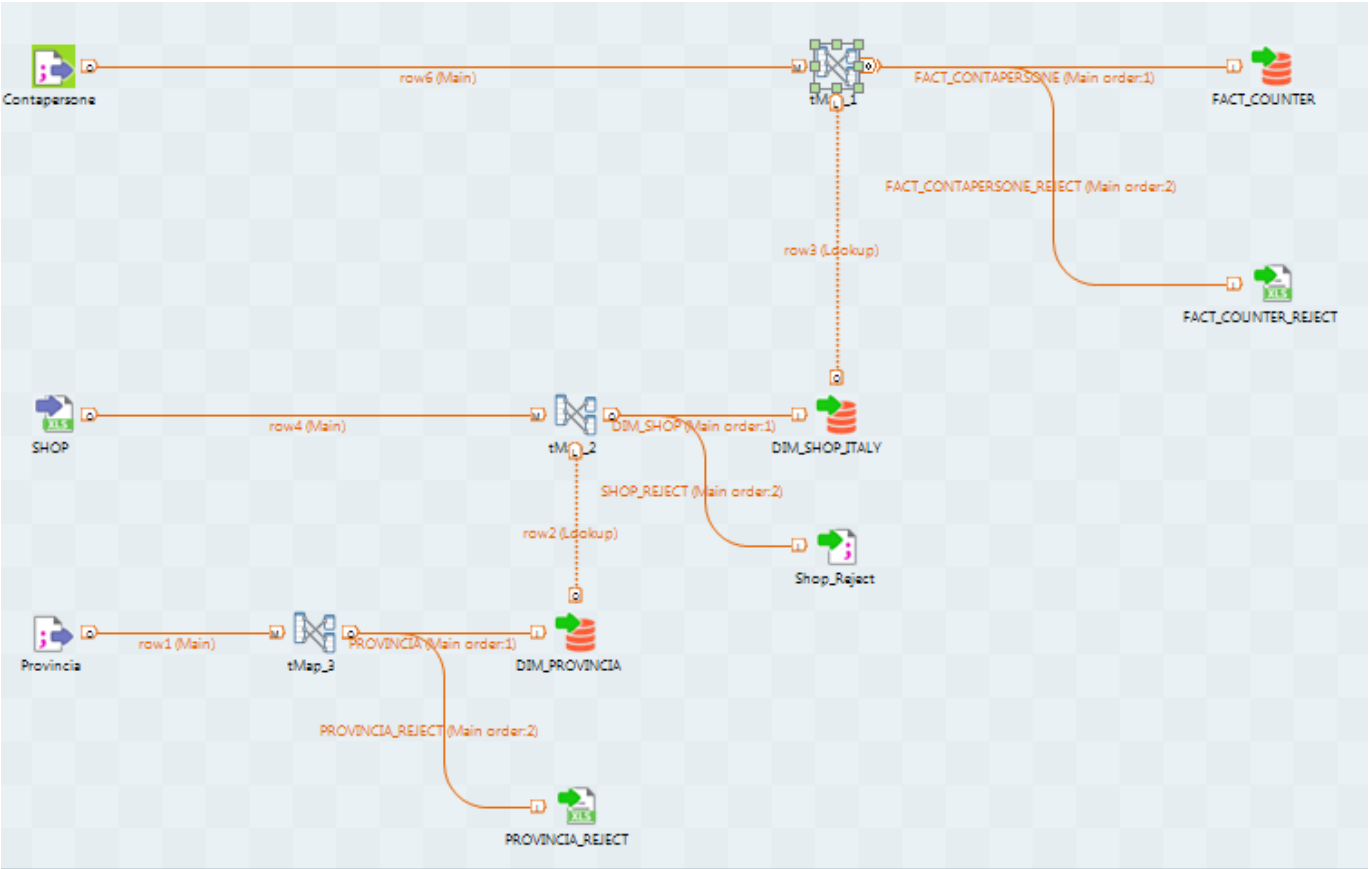


FACT_COUNTER

L'ULTIMA TABELLA DA Creare e caricare nel Database FASHION_RETAILER è la FACT counter, facendo un Inner Join tra la tabella DIM_SHOP già creata, e i record derivante dai metadati del file csv Contapersone.



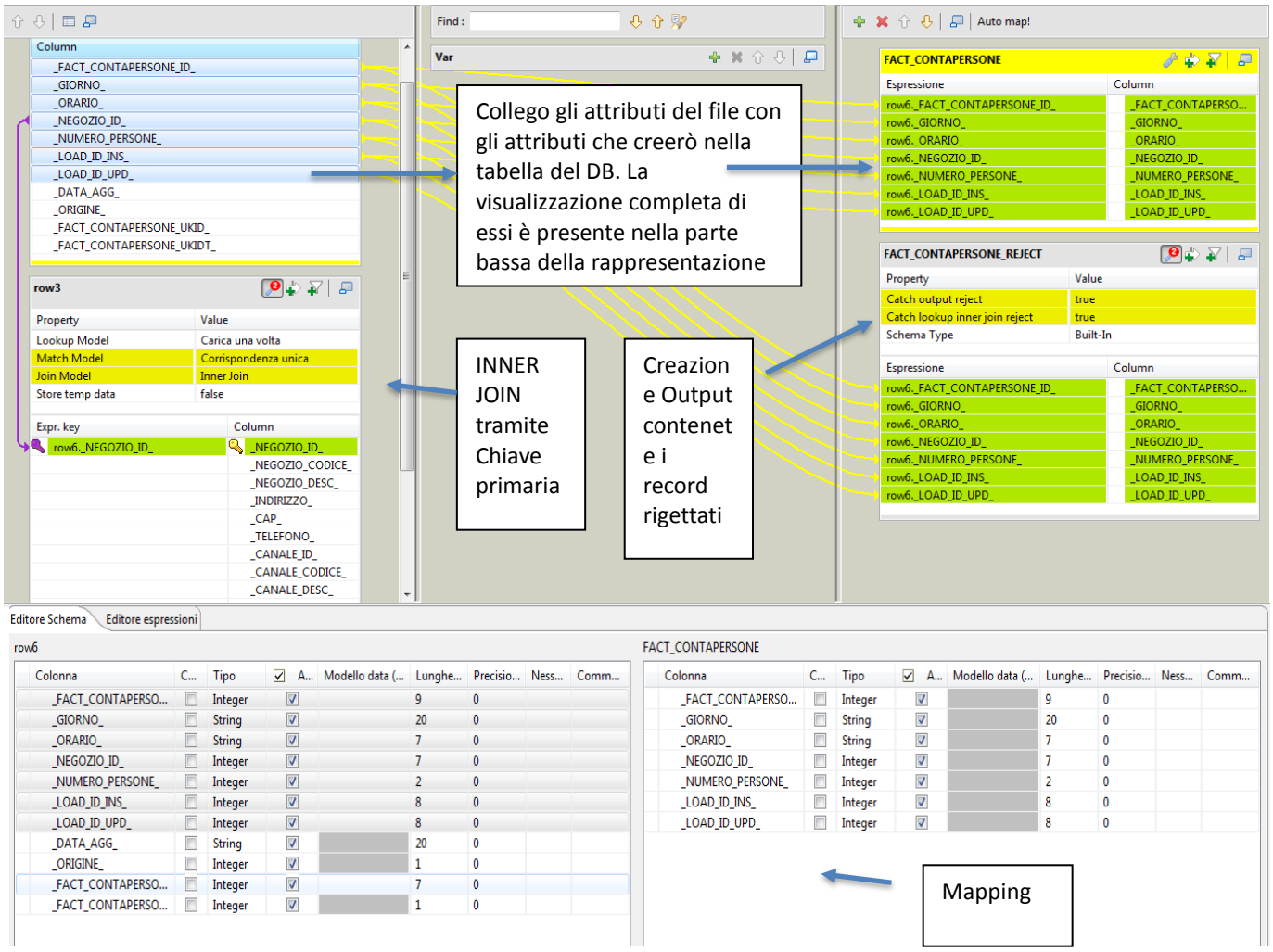
Lo schema generale è il seguente:



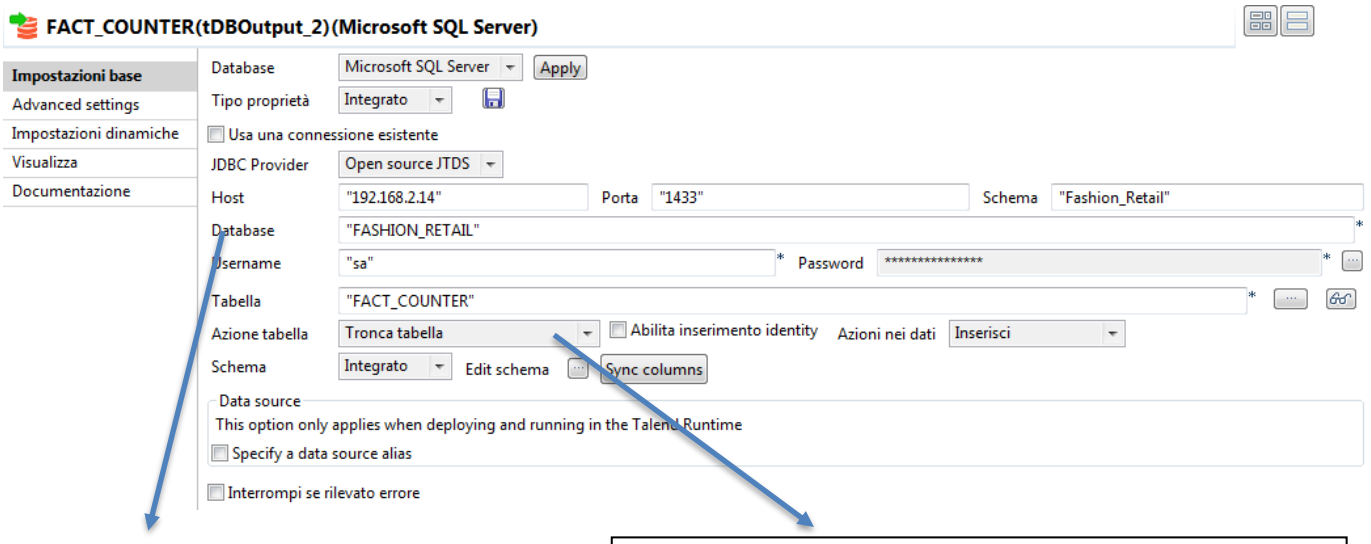
PALETTE UTILIZZATE

- **tMap**

Function	**tMap** is an advanced component, which integrates itself as plugin to *Talend Studio*.
Purpose	**tMap** transforms and routes data from single or multiple sources to single or multiple destinations.

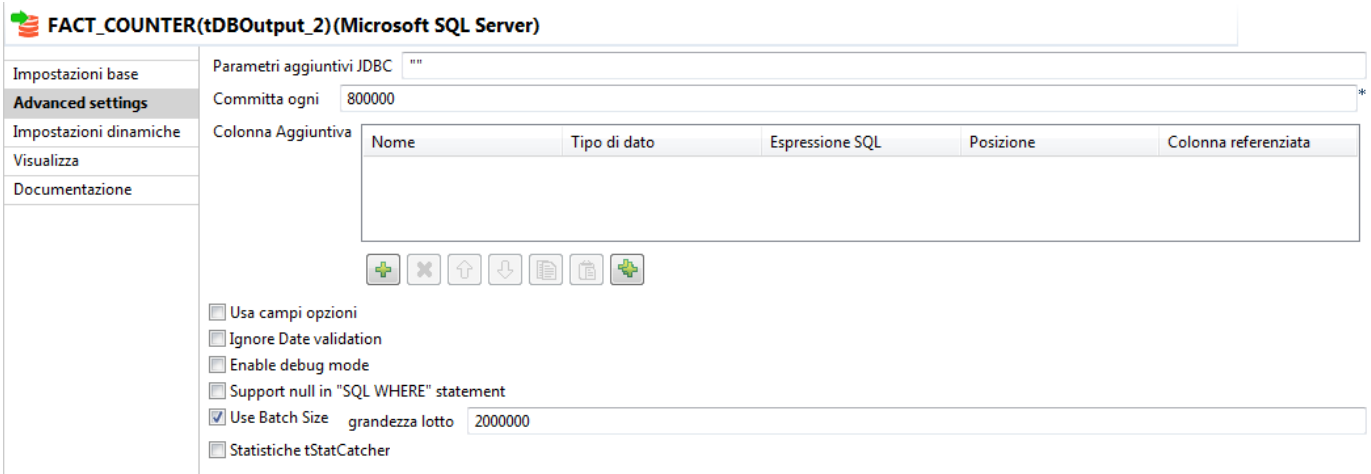


- **tMSSQLOutput**



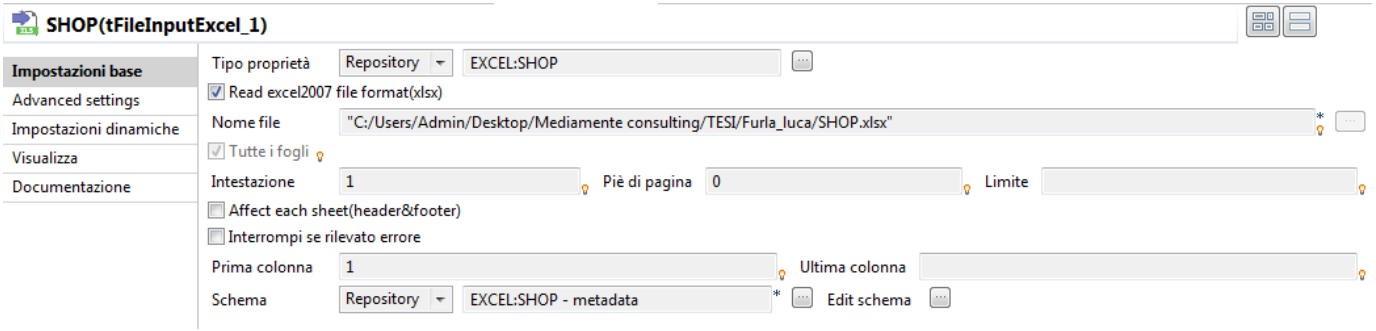
Dati della connessione riferita al DB (MSSQL Server) Legnaia facendo riferimento allo schema corretto dove è presente la tabella vuota nel DB da riempire. Usare una proprietà di tipo Integrato per avere uno schema sempre aggiornato

Selezionare il nome della Tabella appartenete allo schema scelto su cui scriverò i dati.
ATTENZIONE! È molto importante scegliere l'azione Tronca Tabella per velocizzare il processo rimuovendo tutte le righe da una tabella o da partizioni specificate di una tabella senza registrare le eliminazioni delle singole righe



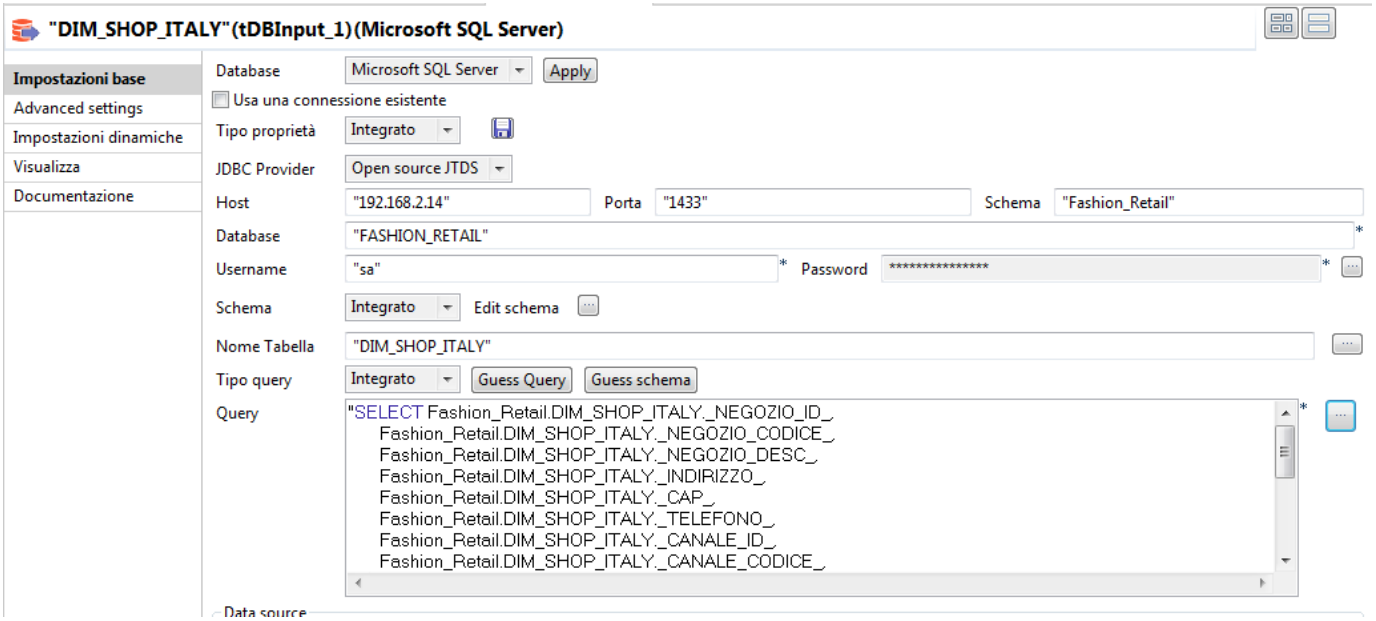
Le impostazioni base dell'Output sono identiche a quelle di Default. Abbiano però le Impostazioni Avanzate che cambiano. Infatti, dopo vari tentativi, siamo arrivati a stabilire come Batch Size una grandezza del lotto a 2000000 che committa ogni 800000 record, arrivando alla massima efficienza per quanto riguarda il nostro DB.

- **tFileInputExcel**

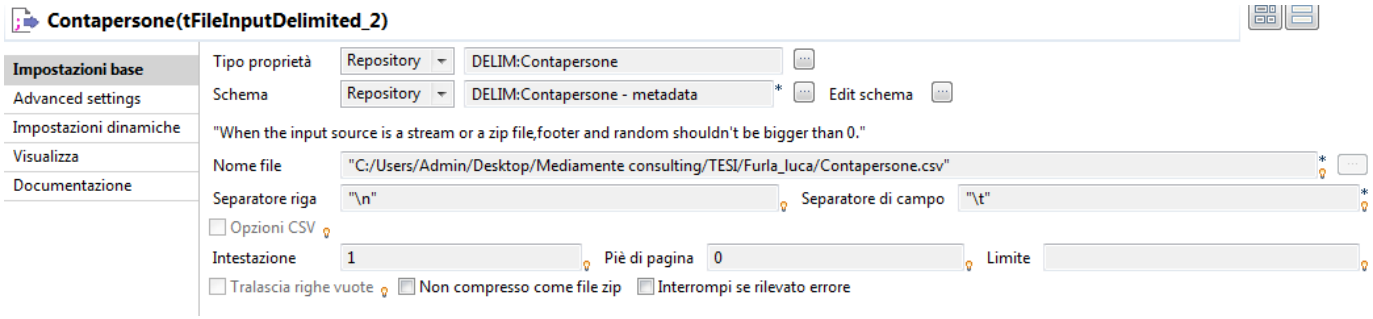


Usare la REPOSITORY cosi nel caso di un cambiamento nei metadati il programma si aggiorna automaticamente.

- **tMSSQLinput**

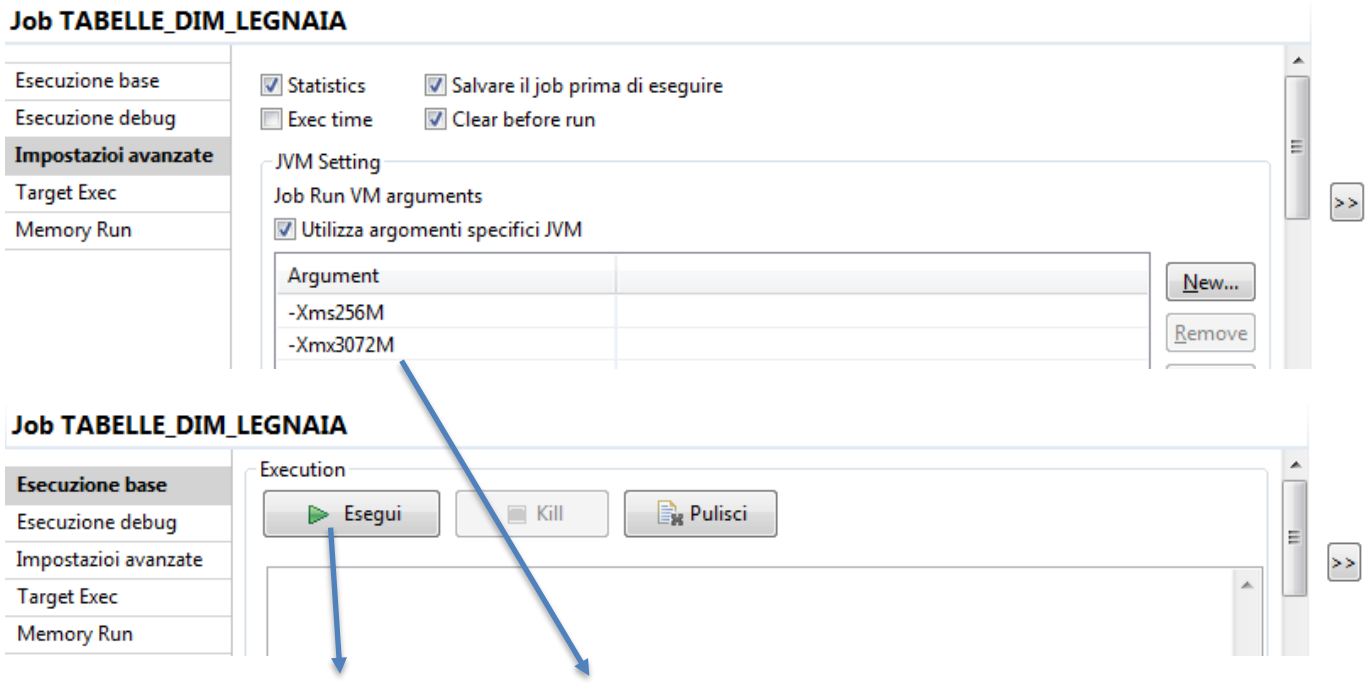


- **tFileInputDelimited**



Usare la REPOSITORY cosi nel caso di un cambiamento nei metadati il programma si aggiorna automaticamente.
Fare attenzione ai separatori di campo dei dati che aiutano la separazione degli attributi e la creazione delle tabelle

RUNNING – CARICAMENTO TABELLE



Usa 3 GB di RAM per aver migliori prestazioni durante l'esecuzione del job, inizializzata tramite il pulsante esegui nella sezione "Esecuzione base"

A fine esecuzione si vedrà il seguente schema con le relative statistiche per ogni passaggio.

RIASSUNTO

CAMPO	DESCRIZIONE
ARGOMENTO	ETL, METADATI, CARICAMENTO DA FILE A DB, DATA INGESTION
STRUMENTO	TALEND OPEN STUDIO, DATAGRIP
HOST	LOCALHOST
SERVER	192.168.2.14
PORTA	1433
JDBC PROVIDER	OPEN SOURCE JTDS
TIPO DB	MICROSOFT SQL SERVER
DATABASE	FASHION_RETAILER
SCHEMA	Fashion_Retailer
QUERY TOPICS	WHERE CLAUSE
COMPONENTI USATE	TMSSQLCONNECT, TMAP, TFILEINPUTEXCEL, TFILEINPUTDELIMITED, TMSSQLINPUT, TMSSQLOUTPUT, TMEMORIZEROWS, TJAVAFLEX, TLOGROWS

TABELLA	SCHEMA	DESCRIZIONE
DIM_CAUSALE	Fashion_Retailer	CAUSALE
DIM_PROVINCIA	Fashion_Retailer	PROVINCIA
DIM_MADE_IN	Fashion_Retailer	LUOGO DI FABBRICAZIONE
DIM_MADE_IN_PIANIFICAZIONE	Fashion_Retailer	SEDE DI FABBRICAZIONE
DIM_SHOP_ITALY	Fashion_Retailer	NEGOZIO
DIM_PRODUCT	Fashion_Retailer	PRODOTTO
FACT_SALES	Fashion_Retailer	VENDITE OTTIMIZZATE
FACT_COUNTER	Fashion_Retailer	CONTAPERSONE