Introduzione a Pentaho Data Integration (Kettle)

Data Warehousing (2015/2016)

Andrea Battistini – 5695559

Informatica Magistrale - UniFI

31 maggio 2016

Pentaho Data Integration

- **Pentaho Data Integration** (PDI) è un software per l'ETL (Extract, Transform, Load) ed è noto anche con il nome **Kettle**
- E' disponibile nella versione commerciale:
  - www.pentaho.com
- Oppure nella versione open source:
  - community.pentaho.com
- Un ottimo tutorial per iniziare è "Getting Started with PDI" disponibile all'indirizzo (lingua inglese):
  - https://help.pentaho.com/Documentation/6.1/0J0/0C0
  - cliccare quindi su "Introduction" e seguire i vari passi

Prerequisiti

- Tutta la suite Pentaho (tranne Pentaho Mobile App) necessita di Oracle Java 7 o 8
- Per scaricare Java 7:
  - http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase7-521261.html#jdk-7u80-oth-JPR
- Per scaricare Java 8:
  - http://www.oracle.com/technetwork/java/javase/downloads/index.html
- Prestare attenzione alle varie versioni di Java installate, all'ordine con cui sono state installate e a quale sia effettivamente quella di default disponibile a linea di comando
- E' indifferente installare JDK o JRE

Installazione (Community Edition)

- Collegarsi a:
  - http://community.pentaho.com/projects/data-integration/
- Cliccare il link “Data Integration 6.1 Stable” posto in fondo alla pagina
- Avviene il redirect verso:
  - https://sourceforge.net/projects/pentaho/files/latest/download
- Si avvia automaticamente lo scaricamento del file “pdi-ce-6.1.0.1-196.zip” (781 MB circa)
- Scompattare lo zip nella directory desiderata

Spoon: l'ambiente grafico

- Trovare (non cliccare) il file "Spoon.bat" che serve ad avviare l'ambiente grafico dove creare il progetto di ETL e che si trova (nel mio caso sul sistema operativo Windows) in:
  - C:\Inst\kettle\pdi-ce-6.1.0.1-196\data-integration
- Creare, se lo si desidera, un collegamento sul desktop per semplificare l'avvio

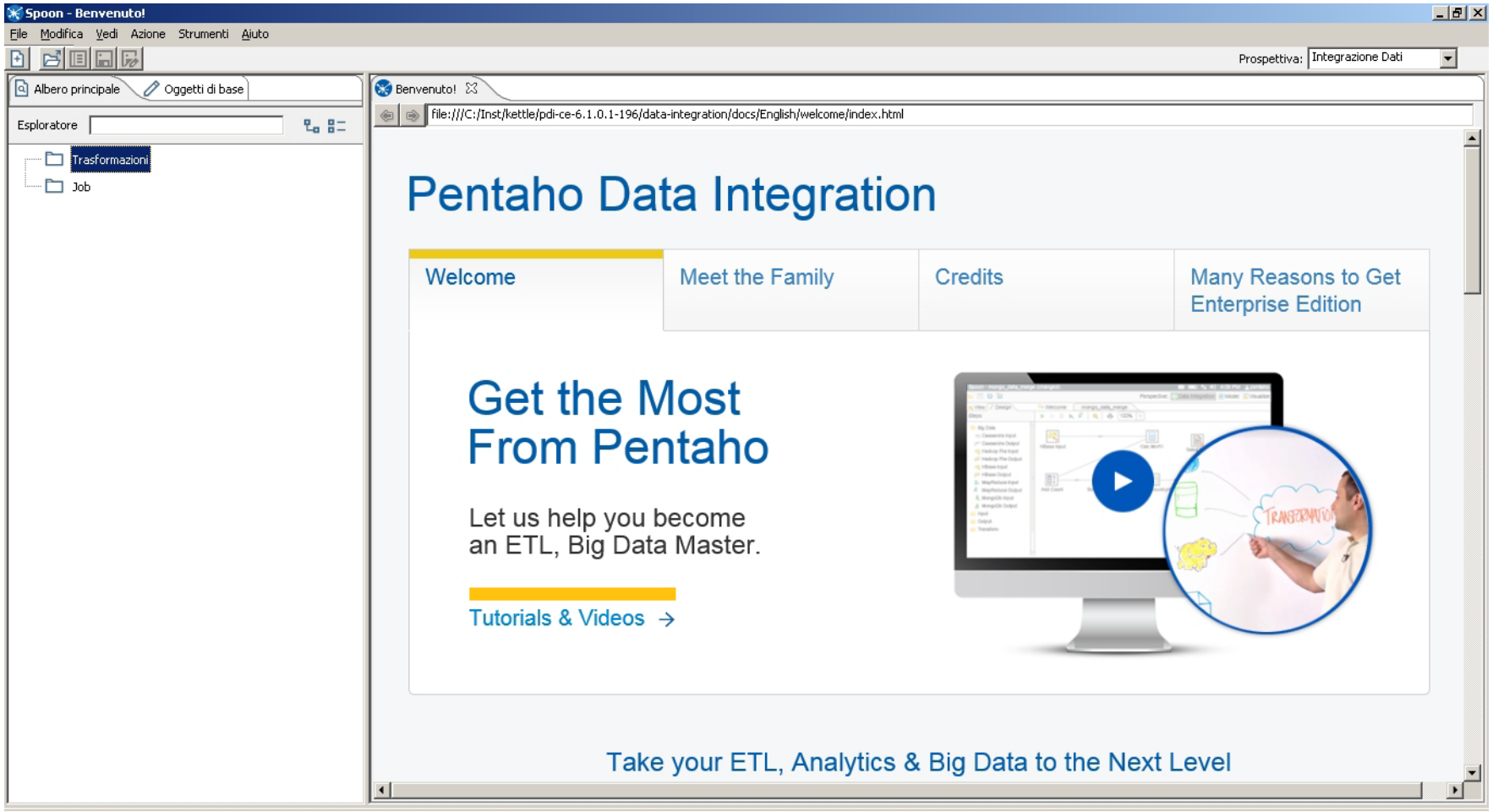
In caso di PC con "poca" RAM può essere necessario effettuare una modifica al file di avvio come indicato nella slide successiva

Nota tecnica sulla RAM

- Avendo un PC con Windows 7 e solo 2 GB di RAM è necessario modificare i parametri di avvio della JVM (Java Virtual Machine), che sono troppo alti, all'interno del file “Spoon.bat”
- Cercare la riga (verso la fine del file):
  - `if "%PENTAHO_DI_JAVA_OPTIONS%"=="" set PENTAHO_DI_JAVA_OPTIONS="-Xms1024m" "-Xmx2048m" "-XX:MaxPermSize=256m"`
- Modificarla dimezzando i valori e ottenere quindi:
  - `if "%PENTAHO_DI_JAVA_OPTIONS%"=="" set PENTAHO_DI_JAVA_OPTIONS="-Xms512m" "-Xmx1024m" "-XX:MaxPermSize=256m"`

Avvio di Spoon

- Cliccare su Spoon.bat e attendere (anche 30 s.)

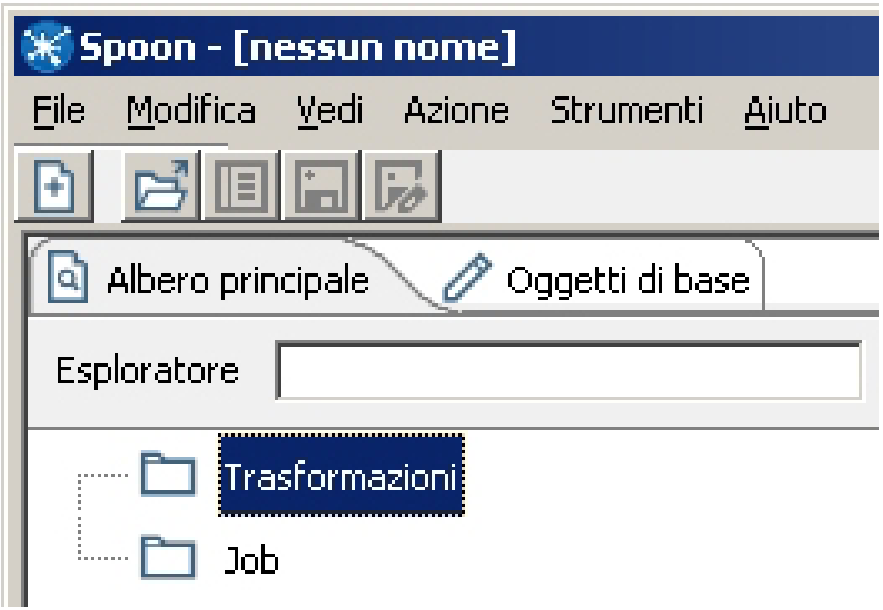


Trasformazioni e Job

- PDI permette la creazione di due tipi di entità:
  - **Trasformazioni**
    - Descrivono il flusso dei dati inteso come lettura da una sorgente, trasformazione di ciò che è stato letto e caricamento/scrittura sul target di destinazione
    - Ad esempio, le “unità” di base con cui si lavora (sia in lettura che in scrittura) sono: righe di testo, record di tabelle su database, campi di file testuali e/o tabelle
  - **Job**
    - Permettono la gestione del flusso logico delle attività e di coordinare l’esecuzione di più trasformazioni (ma anche, ricorsivamente, di altri job)
    - L’obiettivo è impostare il flusso di esecuzione e/o verificare le condizioni necessarie per eseguire determinate azioni
    - Esempi di azioni possono essere: verificare l’esistenza di un file, inviare una mail, eseguire uno script SQL

Trasformazioni e Job: nuovo progetto

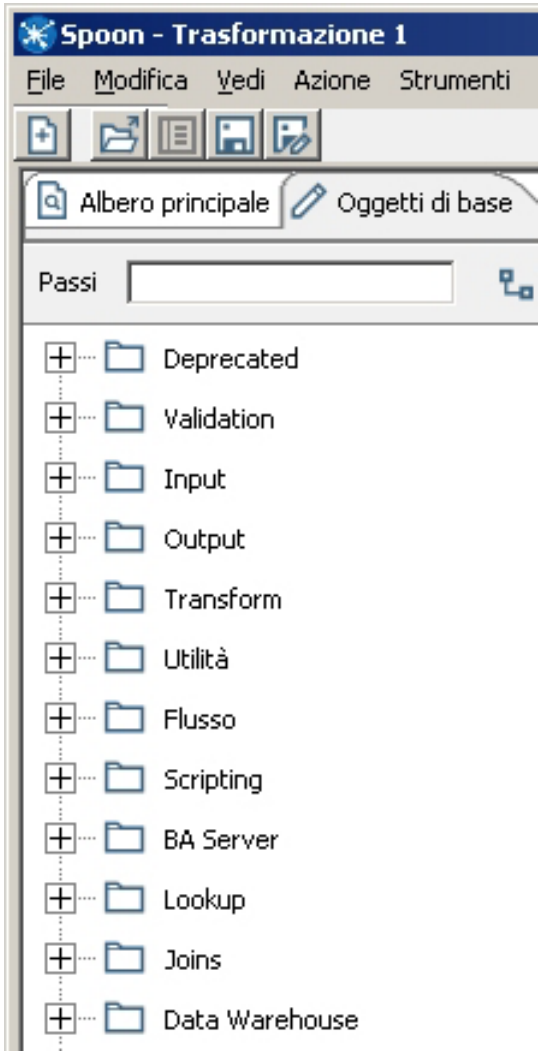
- Per avviare un nuovo progetto si può:
  - selezionare *File* > *Nuovo* > *Trasformazione*
  - selezionare *File* > *Nuovo* > *Job*

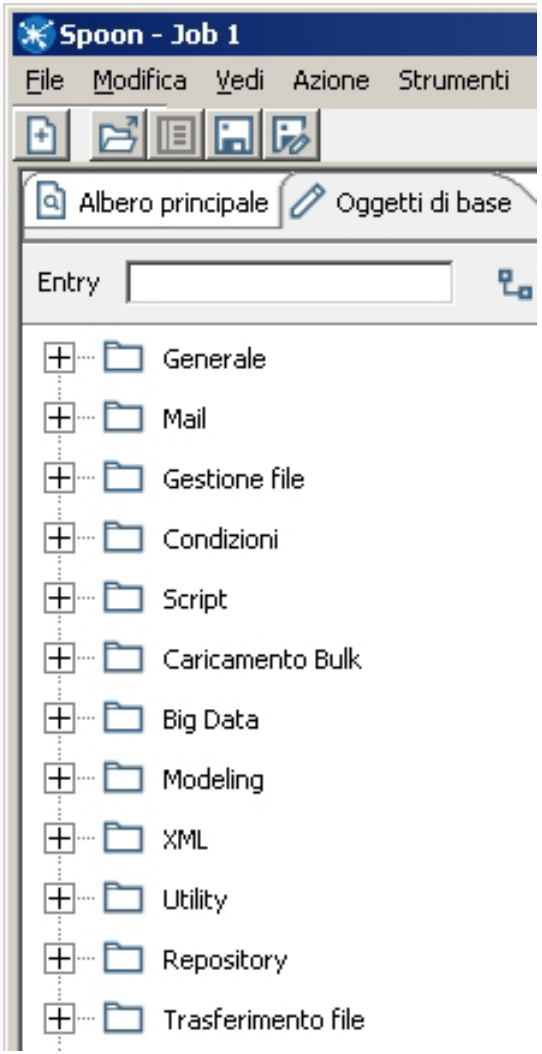


- In alternativa:
  - effettuare doppio click sul tipo di entità desiderata

- Una volta aperto il nuovo progetto, sulla sinistra, saranno visibili i relativi “Oggetti di base”

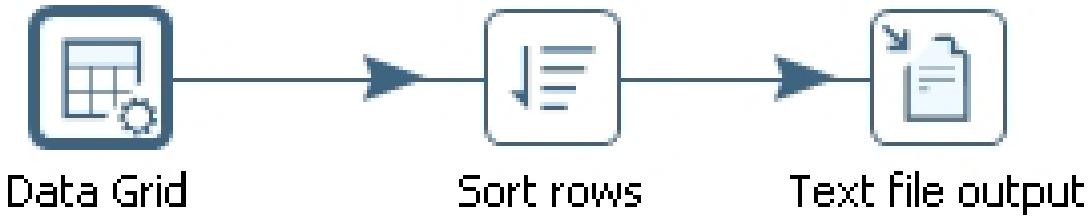
Trasformazioni e Job: oggetti di base





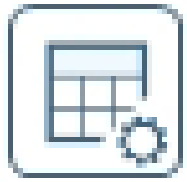
Esempio1

- Questo esempio realizza un processo di ETL molto semplice composto di tre passi:
  - Generazione di righe, come se provenissero da un database, mediante l'oggetto di base:
    Data Grid (classe *Input*)
  - Ordinamento di tali righe mediante:
    Sort rows (classe *Transform*)
  - Scrittura del risultato in un file di testo mediante:
    Text file output (classe *Output*)



Esempio1: Data Grid

- Dal pannello Oggetti di base, doppio click su *Input* > *Data Grid*:
  - appare un oggetto Data Grid nell'area di disegno
- Doppio click sull'oggetto Data Grid creato:
  - per visualizzare la relativa Finestra Impostazioni
- Dal tab *Meta* specificare i seguenti campi:
  - Nome = nome; Tipo = String; Lunghezza = 30
  - Nome = cognome; Tipo = String; Lunghezza = 30
  - Nome = eta; Tipo = Integer;
- Passare al tab *Dati* che mostra una griglia con i campi appena specificati e permette l'introduzione dei dati che verranno generati come flusso di input
- Premere OK per salvare le configurazioni effettuate e chiudere la finestra



Data Grid

Esempio1: Text file output

- Dal pannello Oggetti di base, doppio click su *Input* > *Text file output*:
  - appare un oggetto Text file output nell'area di disegno.
- Notare che il tool ha generato automaticamente l'arco orientato (hop) che va dall'oggetto Data Grid all'oggetto Text file output
- Doppio click sull'oggetto Text file output creato:
  - per visualizzare la relativa Finestra Impostazioni

Text file input

- Dal tab *File* specificare:
  - Nome file = ${Internal.Transformation.Filename.Directory}\out1
- Usare CTRL+SPAZIO per ottenere l'elenco delle variabili predefinite:
  - quella utilizzata sopra indica la cartella dove risiede il file della trasformazione che stiamo componendo

Esempio1: Text file output (2)

- Dal tab *Contenuto* specificare:
  - Separatore = ,
- Sul tab *Campi*:
  - Premere *Preleva campi* per popolare automaticamente la lista dei campi che verranno scritti in output
  - Premere *Larghezza minima* per configurare i campi in modo da occupare il minor spazio possibile
    - Notare come venga annullato l'attributo Lunghezza e venga impostato il trimming dei dati
    - Specificare una Lunghezza e impostare il trimming non “sembra” produrre alcun effetto (per ovvi motivi)
- Premere OK per salvare le configurazioni effettuate e chiudere la finestra.

Esempio1: salvataggio ed esecuzione

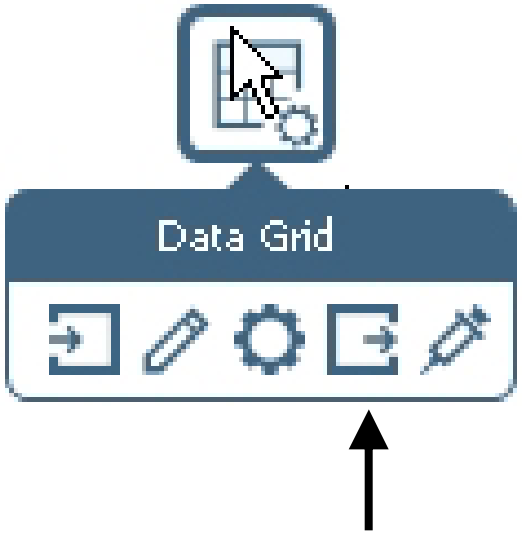
- Salvare la trasformazione mediante *File* > *Salva*
- (altrimenti il salvataggio verrà richiesto obbligatoriamente prima dell'esecuzione)

- Possiamo fare una prima verifica del buon funzionamento del processo premendo il pulsante di esecuzione presente nella toolbar sopra all'area di disegno.

Esempio1: Sort rows

- Per concludere, dal pannello Oggetti di base, doppio click su *Transform* > *Sort rows*:
  - appare un oggetto Sort rows nell'area di disegno.

- Eliminiamo tutti gli archi visto che probabilmente il tool ha collegato l'oggetto di ordinamento come successore dell'output testuale:
  - click destro su ogni freccia e scegliere *Cancella hop*
- Eliminati gli archi ricostruiamo la sequenza di nostro interesse (riposizionando eventualmente gli oggetti per una migliore visualizzazione):
  - posizionare il mouse su Data Grid: apparirà il widget menù su cui scegliamo Connettore Output
  - spostare il mouse sopra a Sort rows e cliccare nuovamente per completare il collegamento
  - ripetere i due passi precedenti per collegare Sort rows a Text file output

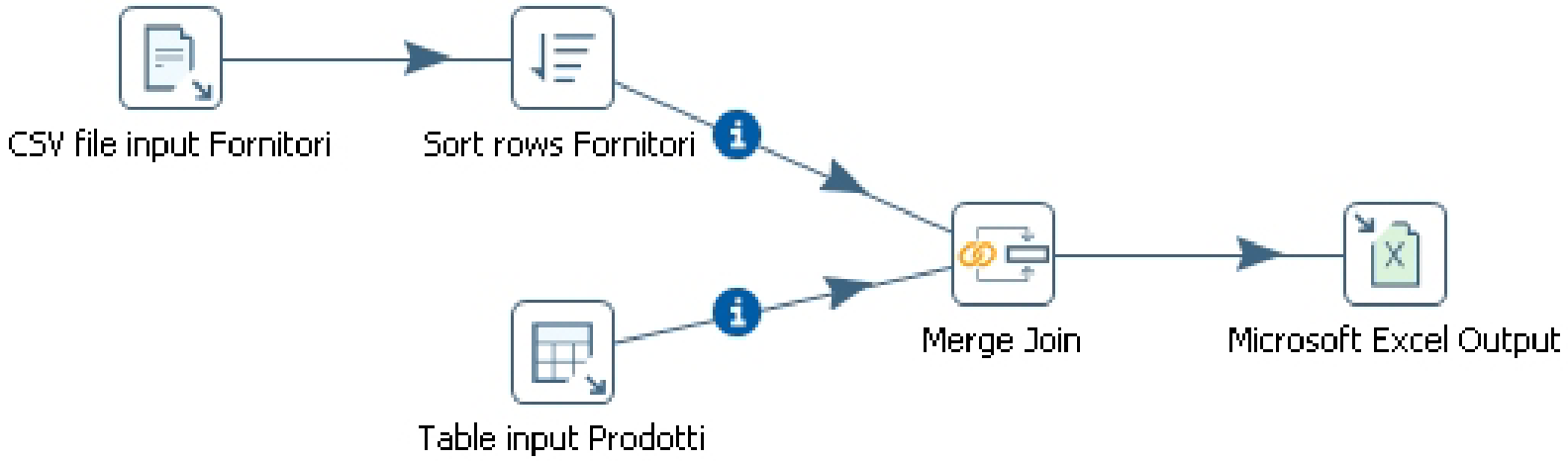


Esempio1: Sort rows ed esecuzione

- Doppio click sull'oggetto Sort rows creato per visualizzare la relativa Finestra Impostazioni
- Premere *Preleva campi* per popolare automaticamente la griglia dove configurare l'ordinamento dei dati
  - per eliminare un campo o cambiarne la posizione, usare le voci del menù contestuale (click destro su una riga della griglia)
  - lasciare solo il campo “eta” (con Ascendente S)
- Premere OK per salvare le configurazioni effettuate e chiudere la finestra.
- Eseguire la trasformazione (salvataggio richiesto)

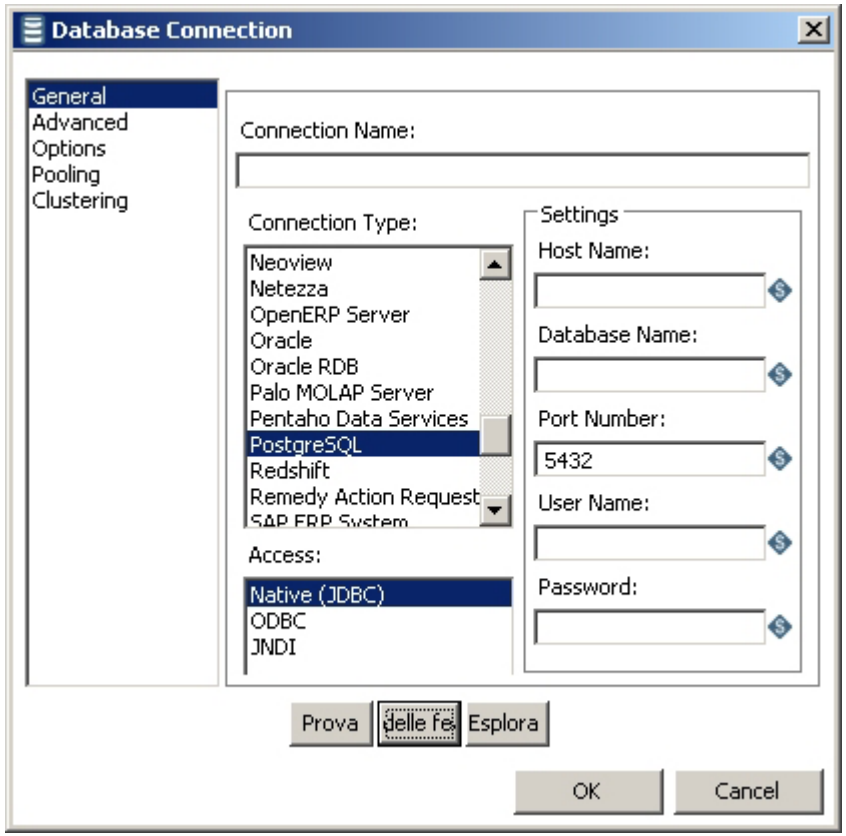
Esempio2

- Questo processo di ETL prevede la lettura di dati da file (Fornitori) e da tabelle su database (Prodotti). Tra Prodotti e Fornitori esiste una relazione 1:N e pertanto si effettua una Join prima di scrivere l'output su un foglio Excel
- Notare che la *Merge Join* avviene tra input di tipologia diversa (file/tabella)



Esempio2: connessione al database

- Effettuiamo *File* > *Nuovo* > *Trasformazione*
- Creiamo una connessione al database mediante *File* > *Nuovo* > *Connessione Database*:
  - Scegliamo PostgreSQL
  - Connection Name = conn
  - Host Name = localhost
  - Database Name = kettledb
  - Port Number = 5432
  - User Name = …
  - Password = …
  - Premere *Prova* per verifica
  - Premere OK per chiudere



Esempio2: CSV file input

- Inserire un oggetto *Input* > *CSV file input* e modificarne le impostazioni come segue:
  - Nome del passo = CSV file input Fornitori
  - Nome del file = (premere *Apri* e scegliere il file)
  - Delimitatore = ;
  - Premere *Preleva campi*
    - verificare che siano presenti tutti i campi desiderati
  - Premere OK per chiudere la Finestra Impostazioni

Nei prossimi passi non indicheremo più come eliminare/introdurre gli hop: tale problematica è già stata affrontata nell'Esempio1

Esempio2: Sort rows

- Inserire un oggetto *Transform* > *Sort rows* e modificarne le impostazioni come segue:
  - Nome del passo = Sort rows Fornitori
  - Premere *Preleva campi*
    - lasciare solo il campo CodiceFornitore (con Ascendente S)
  - Premere OK per chiudere la Finestra Impostazioni

- L'ordinamento è reso necessario da Merge Join che "pretende" flussi di input ordinati per funzionare correttamente

Esempio2: Table input

- Inserire un oggetto *Input* > *Table input* e modificarne le impostazioni come segue:
  - Nome del passo = Table input Prodotti
  - Connessione = conn (scelto dalla tendina)
  - SQL =
    SELECT * FROM prodotti ORDER BY cod_fornitore
  - Premere OK per chiudere la Finestra Impostazioni

- Anche per le tabelle su database è necessario l'ordinamento affinché la Merge Join possa funzionare correttamente

Esempio2: Merge Join

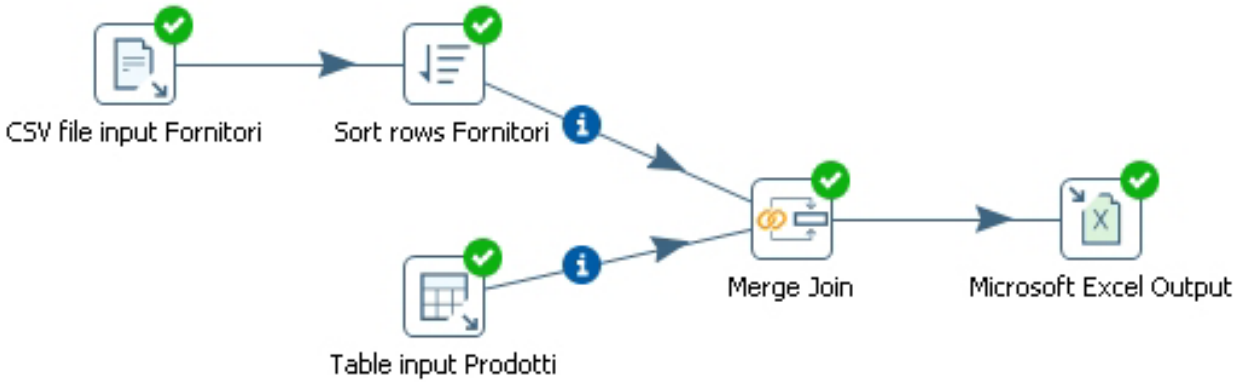
- Inserire un oggetto *Joins* > *Merge Join* e modificarne le impostazioni come segue:
  - Primo passo = Sort rows Fornitori
  - Secondo passo = Table input Prodotti
  - Tipo di join = INNER
  - Chiavi per il 1° passo = CodiceFornitore
  - Chiavi per il 2° passo = cod_fornitore
  - (usare eventualmente i pulsanti *Preleva campi chiave*)
  - Premere OK per chiudere la Finestra Impostazioni
- Sottolineiamo nuovamente la “potenza” del tool che permette la Join tra fonti di tipo diverso e non solo tra tabelle di database

Esempio2: Microsoft Excel Output

- Inserire un oggetto *Output* > *Microsoft Excel Output* e modificarne le impostazioni
- Dal tab *File*:
  - Nome del file = ${Internal.Transformation.Filename.Directory}\ExOUT
- Dal tab *Campi:*
  - Premere Preleva campi
  - Scegliere/Modificare l'elenco dei campi da produrre in uscita
- <u>Notare</u> che esiste anche l'oggetto *Microsoft Excel Writer* che possiede però un insieme di impostazioni diverse e organizzate in altro modo

Esempio2: esecuzione

- Eseguire la trasformazione (tasto F9):
  - se l'esecuzione è andata a buon fine si dovrebbero osservare i vari segni di spunta verdi nell'area di disegno…



  - …e dei risultati di esecuzioni simili ai questi:

	Nome passo	Copia n.	Letti	Scritti	Input	Output	Aggiornati
1	Table input Prodotti	0	0	19	19	0	0
2	CSV file input Fornitori	0	0	5	6	0	0
3	Sort rows Fornitori	0	5	5	0	0	0
4	Merge Join	0	24	19	0	0	0
5	Microsoft Excel Output	0	19	19	0	19	0

Conclusioni

- Con questa breve introduzione a Pentaho Data Integration abbiamo voluto fornire le conoscenze di base per costruire semplici processi di ETL
- Come abbiamo visto durante la presentazione gli oggetti di base per comporre le trasformazioni sono innumerevoli e per trarre il massimo profitto da ognuno di essi è necessario studiarne a fondo le caratteristiche e le relative impostazioni