



Module 5

Architectures, Features, and Details of Data Integration Tools

Lesson 4: Pentaho Data Integration






Lesson Objectives

- List major features of Pentaho Data Integration
- Gain familiarity with Pentaho features for jobs and transformations
- Gain experience with Pentaho on the practice exercise and assignment



Pentaho Products

- Platform for data integration, business analytics, and big data
- Open core business model
- Pentaho Data Integration 
- Pentaho Business Analytics 
- Pentaho Big Data Analytics 















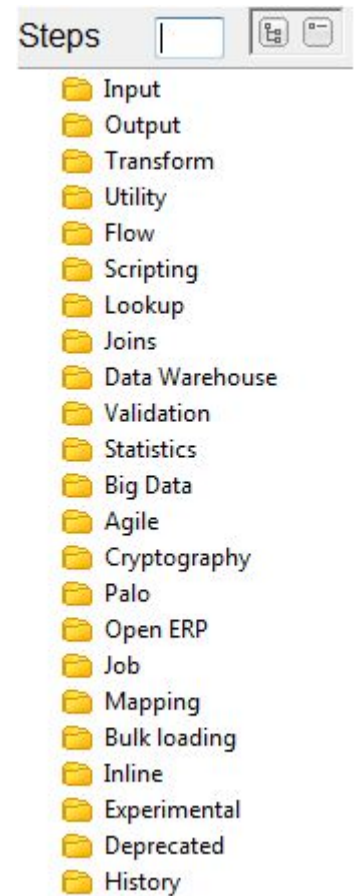
Pentaho Data Integration

- Editions
 - Subscription service from Pentaho website
 - Community edition: Kettle
- Basic concepts
 - Transformation with data flow among steps and hops
 - Job with data flow among transformations and external entities
- Tools:
 - Spoon: graphical design of transformations and jobs
 - Pan and Kitchen: execution of transformations and jobs

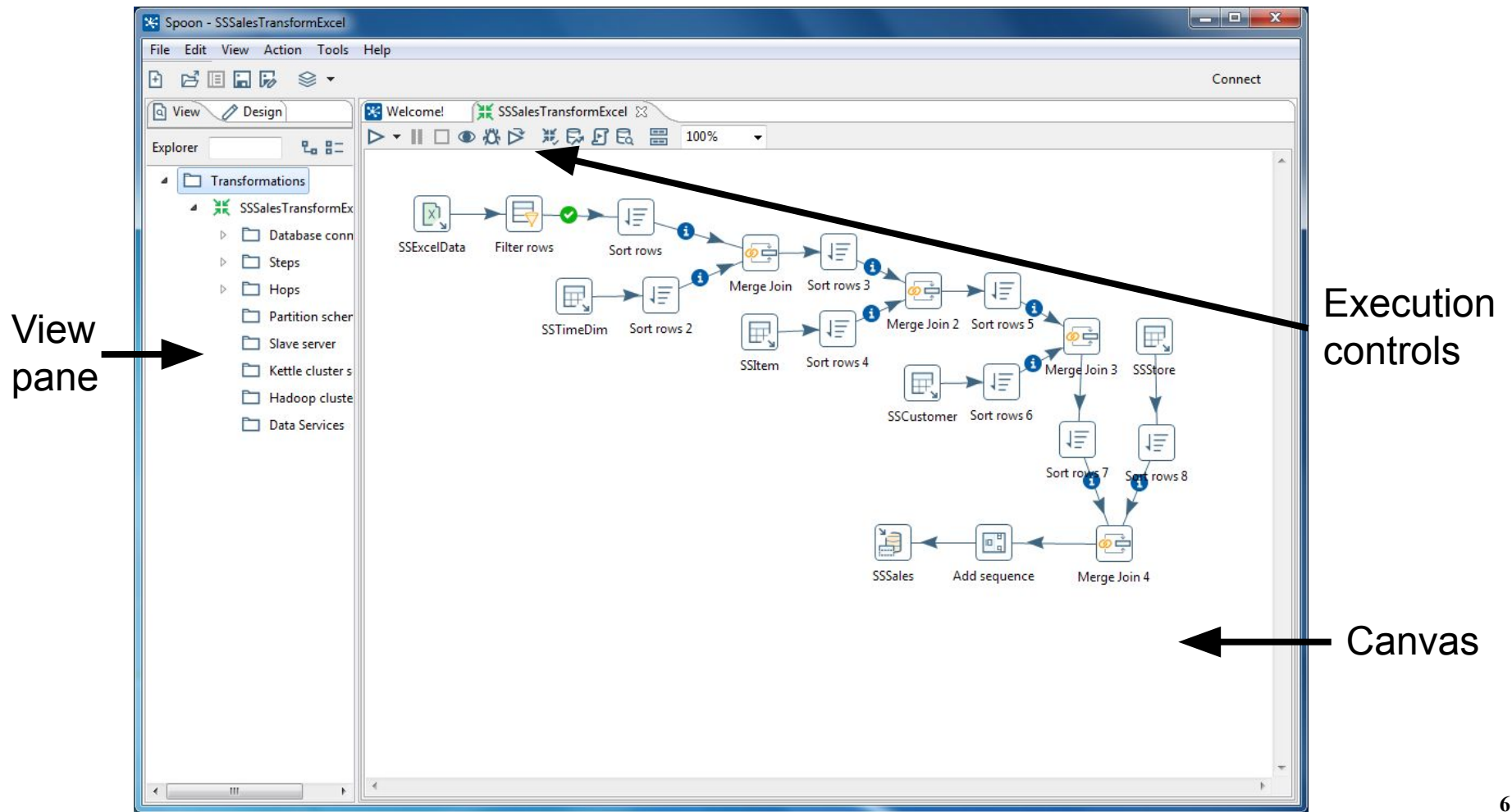


Transformations

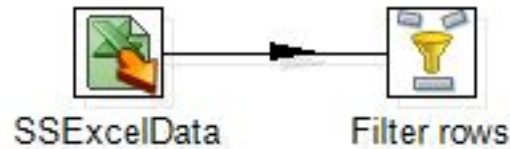
- Step: process in a data flow
 - Input/Output  
 - Transform: sort, split, concatenate, ...  
 - Flow: filter rows  
 - Lookup: existence of rows, tables, files, ...  
 - Join: merge join, multiway merge, ...  
 - Validation: credit card, mail, data  
- Hop: directed connection between steps
- Database connections
- Distributed processing: partition, cluster, ...



Spoon IDE



Example Transformations



Microsoft Excel input

Step name: SSEXcelData

Files | Sheets | Content | Error Handling | Fields | Additional output fields

Spread sheet type (engine): Excel 97-2003 XLS (XLS)

File or directory: Add Browse...

Regular Expression:

Exclude Regular Expression:

Selected files:

#	File/Directory
1	C:\Users\Training\Downloads\Pentaho\SSEXcelSourceNew.xls

Accept filenames from previous steps: ☐

Accept filenames from:

Step to read filenames from:

Field in the input to use as:

Show filename(s)...

OK Preview rows Cancel

Filter rows

Step name: Filter rows

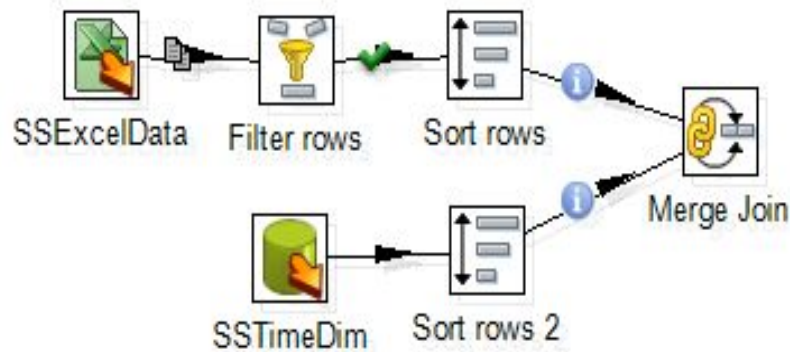
Send 'true' data to step:

Send 'false' data to step:

The condition:

OK Cancel

Merge Join Step



Merge Join

Step name: Merge Join

First Step: Sort rows

Second Step: Sort rows 2

Join Type: INNER

Keys for 1st step:

#	Key field
1	Day
2	Month
3	Year

Get key fields

Keys for 2nd step:

#	Key field
1	TIMEDAY
2	TIMEMON...
3	TIMEYEAR

Get key fields

Help OK Cancel

Summary

- Prominent open source tools (Talend and Pentaho)
- Community and subscription editions
- Supports specification of transformations and steps and transformation execution
- Use Pentaho for exercise and assignment



Talend versus Pentaho

- Pentaho advantages
 - Incremental execution
 - Easier to export
 - Easier reuse of database connections
- Talend advantages
 - More compact specification especially for multiple joins and not null checks
 - HTML documentation generation

