Data integration made easy with Talend Open Studio for Data Integration

Dimitar Zahariev
BI / DI Consultant
dimitar@zahariev.pro
@shekeriev
Disclaimer

Please keep in mind that:

- I’m not related in any way to Talend
- Everything stated from now on is my personal opinion and it doesn’t reflect in any way the position of my employer or other related parties
Agenda

What is data integration?
Aren’t there any tools?
What is Talend Open Studio for Data Integration?
What is data integration?

It has something to do with data ...
What Wikipedia says about data integration?

“Data integration involves combining data residing in different sources and providing users with a unified view of these data.”

Wikipedia

What are the main aspects to consider?

We should pay attention to the following:

- **Source**
  Where data is coming from and in what format

- **Transportation**
  How we will get the data from and to the end points

- **Processing**
  What we are expected to do with the data

- **Target**
  Where and under what format we are expected to put the processed data
Data integration use cases

Typical data integration use cases:

- File exchange
  Batch processing of different file formats

- Data migration
  A one-time process that moves and transforms data between two systems

- Data synchronization
  A repeatable process that keeps the data in sync across many systems

- ETL (extract, transform, load)
  A key component process of a data warehouse or business intelligence systems
A simple scenario

Two systems – one source and one target, exchanging only plain text files, but with different structures.
Two source systems and two targets – a database and another system. Input data is coming as plain text files and it has to be stored in the database and delivered as text files to the third system.
And what about this?
What can we do to meet the requirements?

Any of the following will do the job:

- Write our own solution from end to end
  Time consuming solution which can be difficult to support or extend also

- Use a combination of existing tools and own code
  Leads to a fragmented solution with limited options to scale

- Go out and pick up an integration solution
  Usually the most robust choice
Aren’t there any tools?

Yes, there are but ...
Data integration tools classification

Based on their level of completeness/coverage:
- Integration packages
- Development environments
- Complete suites

Based on their support terms:
- Community based
- Subscription levels
Integration engines

Integration engines have the following specifics:

- Do not offer UI capabilities
- Intended for easy embedding
- Distributed as libraries
- Offer fewer functionalities
Integration solutions

Integration solutions usually are:

- Offered in different packages
- Richer in terms of functionality
- Convenient graphical UI
- Complementary tools
- Enterprise ready
What is Talend Open Studio for Data Integration?

Besides being a great tool it is also ...
What is Talend Open Studio for Data Integration?

Talend Open Studio is:

• An open source graphical environment
• It allows us to rapidly develop data integration processes
• It makes data integration a manageable process

Which in turn allows us to focus on the process rather than the technical details.
What it does?

Typical usage of Studio includes but is not limited to:

- Data migration
- Data synchronization
- Data exchange
- and etc.

Talend Studio addresses both the needs of ETL for analytics and ETL for operational integration equally well.
General look and feel
Main definitions

• Workspace
  Local directory that stores one or more projects

• Project
  Logical grouping of one or more jobs

• Job
  The smallest executable unit. It is a group of one or more components. Typically implements a data flow or integration process
Main components of the Studio

• Repository
  Gives us access to the Repository where we can create Jobs and manage metadata

• Design Workspace
  Provides us with a playground to design our Jobs

• Configuration Tabs
  Allow us to control the components behavior and execute Jobs

• Outline and Code Tabs

• Palette
  Contains the different components we use to build our Jobs
Main sections of the Repository

• Job Designs
  Stores Jobs we work on. Furthermore Jobs can be organized into folders

• Contexts
  Contains sets of global or job-specific variables

• Metadata
  Holds descriptive information about our data sources and targets grouped by type
Talend Open Studio for Data Integration in action
What skills do we need to possess?

It is good to be comfortable to some extent with:

- General file structures (CSV, XML, and etc.)
- Relational Databases and SQL
- Data warehousing methodology and techniques
- General Java knowledge
- Protocols like SMTP, FTP, and etc.

... and of course willingness to learn.
Why we should use it?

Here are some points to consider:

• It is open source
• It is easy to learn
• It is quick to develop with
• It is easy to install
• It has a huge amount of components
• It is backed by a solid community
Resources

Useful stuff to help us on our journey with Talend
Official resources

A short list of helpful resources:

- Software and documentation
  
  http://www.talend.com/download/talend-open-studio#t4

- Talend knowledge base
  
  https://help.talend.com/display/HOME/Knowledge+Base

- Talend community site
  
  https://www.talendforge.org/

- Talend demo project *(available within the studio)*
Additional resources

A very good book on the subject:

- Getting Started with Talend Open Studio for Data Integration
  
  *by Jonathan Bowen*

Resources prepared by me:

- Pre-Built Linux VMs with Talend installed for VirtualBox
  
  https://zahariev.pro/balccon2k16

- Articles on the subject *(they will increase with time)*
  
  https://zahariev.pro/category/talend
Thank you!