

# Data integration made easy with Talend Open Studio for Data Integration

Dimitar Zahariev  
BI / DI Consultant  
[dimitar@zahariev.pro](mailto: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*

*[https://en.wikipedia.org/wiki/Data\\_integration](https://en.wikipedia.org/wiki/Data_integration)*

# 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

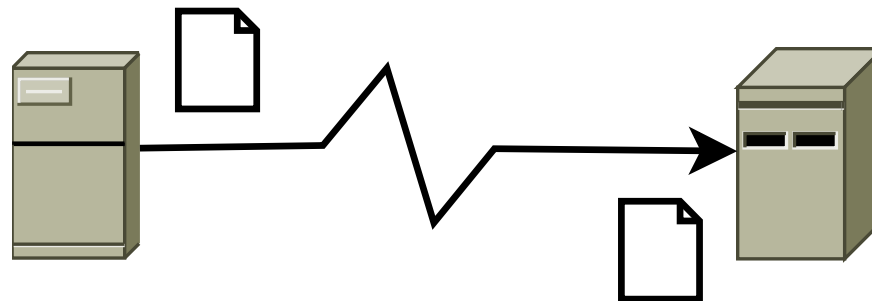
- **Data synchronization**

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

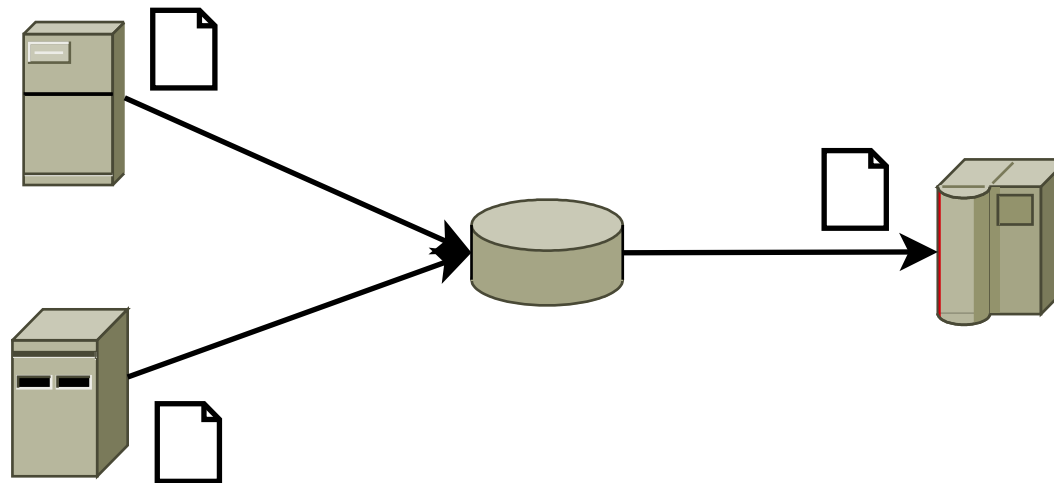
# A simple scenario



Two systems – one source and one target, exchanging only plain text files, but with different structures.

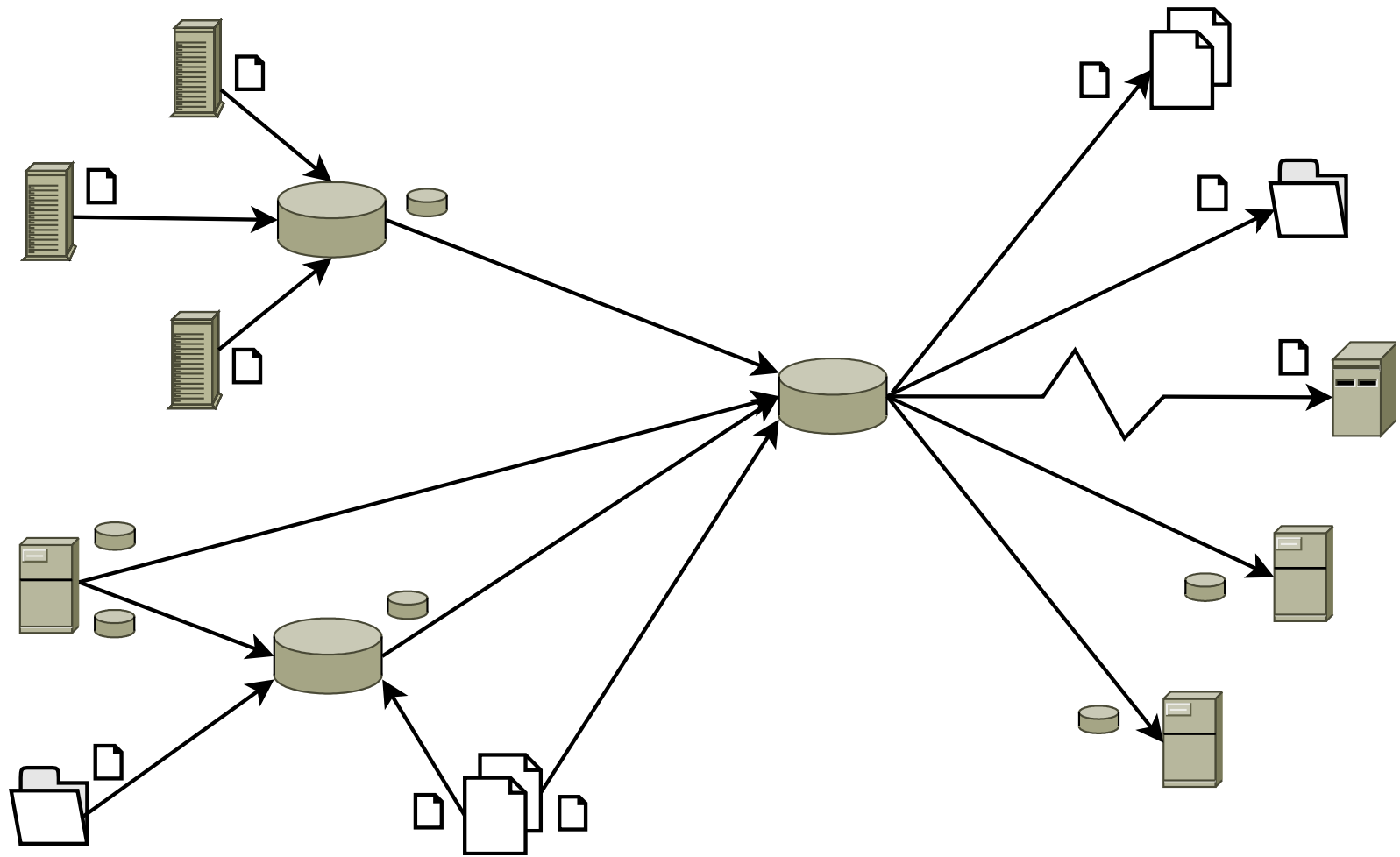


# A bit more complex scenario



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.

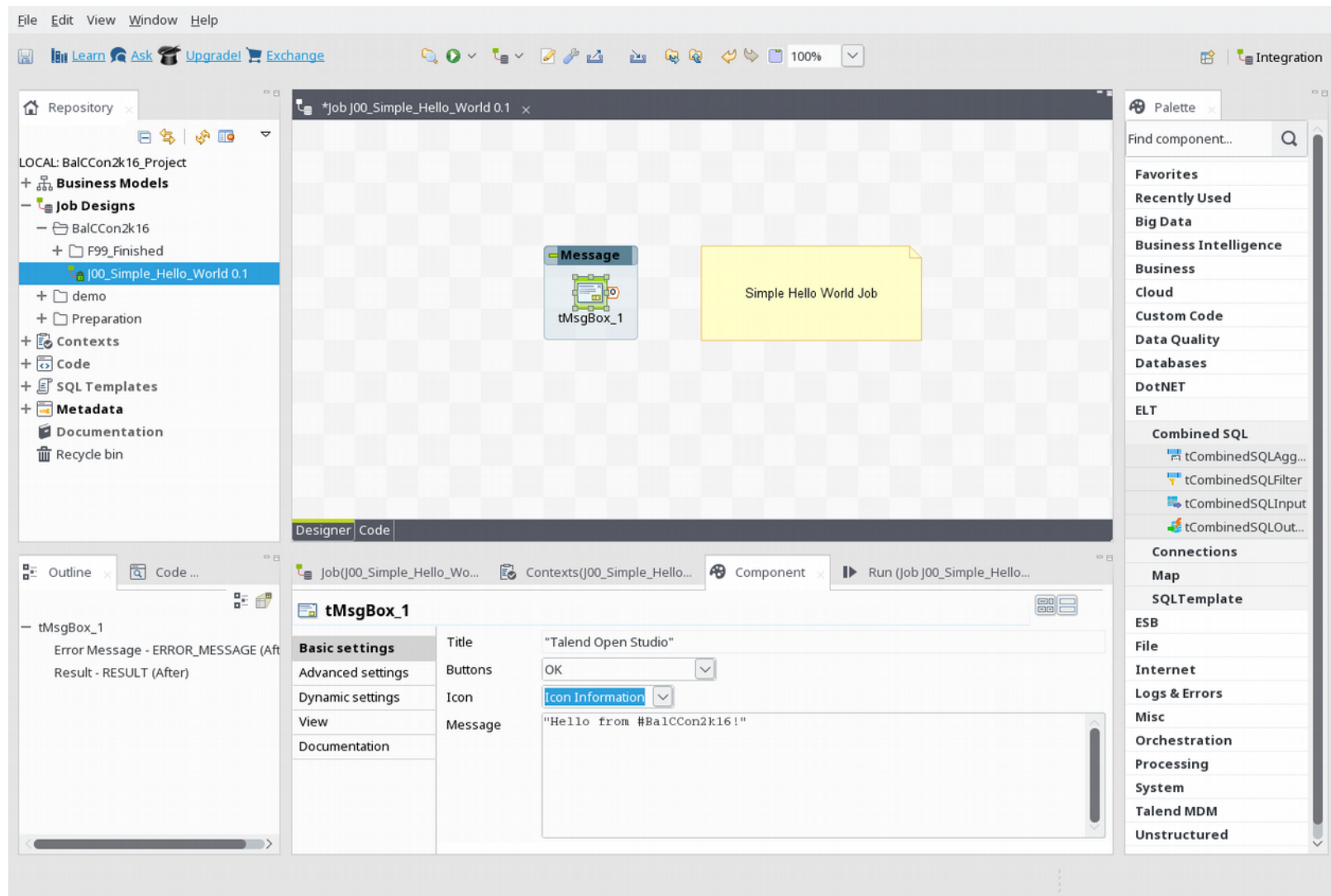
# What it does?

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

The screenshot displays the Talend Open Studio for Data Integration interface. The main workspace shows a job titled "Job J04\_FACT\_ORDERS\_BASIC 0.1" with a diagram titled "Parse and load orders data". The diagram illustrates a data flow from a file source (tFileList\_) through an XML parser (XML\_Orders) and a map component (tMap\_1) to a database target (FACT\_ORDERS). The tMap\_1 component is configured with three lookups: DIM\_DATE, DIM\_PRODUCT, and DIM\_CUSTOMER. The execution progress is shown with green bars and text indicating the number of rows processed and the time taken for each step.

**Job J04\_FACT\_ORDERS\_BASIC 0.1**

**Parse and load orders data**

1 exec running

Iterate

XML\_Orders

11 rows in 0.23s  
48.03 rows/s  
Order\_Data (Main)

tMap\_1

42 rows in 0.01s  
3500 rows/s  
Product\_Lookup (Lookup)

366 rows in 0.38s  
955.61 rows/s  
Date\_Lookup (Lookup)

974 rows in 0.11s  
9018.52 rows/s  
Customer\_Lookup (Lookup)

11 rows in 0.26s  
42.64 rows/s  
Orders\_List (Main)

FACT\_ORDERS

OnComponentOk

Move pa...

tFileCopy\_1

ok

**Job J04\_FACT\_ORDERS\_BASIC**

Execution

Run Kill Clear

Starting job J04\_FACT\_ORDERS\_BASIC at 08:59 10/09/2016.

[statistics] connecting to socket on port 3571  
[statistics] connected

Line limit 100 Wrap

**HomeLab**

Name	Value
DataInFolder	/home/dimitar/tale
DBAdditionalParam	
DBName	balcon2k16
DBPort	3306
DBServer	localhost
DBUserName	balcon
DBUserPassword	balcon2k16

# 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/balcon2k16>
- Articles on the subject (*they will increase with time*)  
<https://zahariev.pro/category/talend>

# Thank you!

Dimitar Zahariev  
BI / DI Consultant  
[dimitar@zahariev.pro](mailto:dimitar@zahariev.pro)  
@shekeriev

BALCCON2K16

