

Introduction to 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

- General definitions
- Business case
- Demo

General definitions

Just to be sure that we are on the same page

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

General look and feel

The screenshot displays the Talend Studio interface for a job named **J04_FACT_ORDERS_BASIC 0.1**. The main workspace shows a data flow diagram titled **Parse and load orders data**. The process starts with **tFileList_1** (1 exec running), followed by an **Iterate** component, then **XML_Orders** (11 rows in 0.23s, 48.03 rows/s). This feeds into **tMap_1**, which performs three lookups: **"DIM_DATE"** (366 rows in 0.38s, 955.61 rows/s), **"DIM_PRODUCT"** (42 rows in 0.01s, 3500 rows/s), and **"DIM_CUSTOMER"** (974 rows in 0.11s, 9018.52 rows/s). The output of **tMap_1** goes to **"FACT_ORDERS"** (11 rows in 0.26s, 42.64 rows/s), which then connects to **tFileCopy_1** (ok) via an **OnComponentOk** trigger.

The left sidebar shows the **Repository** with a tree view of **Business Models** and **Job Designs**. The **Job Designs** tree is expanded, showing **J04_FACT_ORDERS_EXTENDED 0.1** selected. The **Outline** pane at the bottom left lists the components in the job: **tFileCopy_1**, **tFileInputXML_1 (XML_Orders)**, **tFileList_1**, **tMap_1**, **tMySQLInput_1 ("DIM_DATE")**, **tMySQLInput_2 ("DIM_CUSTOMER")**, **tMySQLInput_3 ("DIM_PRODUCT")**, and **tMySQLOutput_2 ("FACT_ORDERS")**.

The bottom section shows the **Job J04_FACT_ORDERS_BASIC** execution details. The **Basic Run** tab is active, displaying the **Execution** log. The log shows the job starting at 08:59 10/09/2016, with statistics indicating a connection to a socket on port 3571. The **HomeLab** tab is also visible, showing a table of configuration values:

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

General look and feel

The screenshot displays the Talend Studio interface with a job design for 'Job J04_FACT_ORDERS_BASIC 0.1'. The job flow includes components like tFileList, Iterate, XML_Orders, tMap_1, and tFileCopy_1. A blue callout box highlights the 'Repository' on the left, stating: 'Repository Gives us access to the Repository where we can create Jobs and manage metadata'. The bottom panel shows the 'Basic Run' tab with execution logs and a 'HomeLab' configuration table.

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

Job J04_FACT_ORDERS_BASIC 0.1

Basic Run

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

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

General look and feel

The screenshot displays the Talend Design Workspace interface for a job named "Job J04_FACT_ORDERS_BASIC 0.1". The main canvas shows a data flow diagram with the following components and data:

- Parse and load orders data** (Job Title)
- Input Components:**
 - tFileList_1**: 1 exec running
 - Iterate**: 11 rows in 0.23s, 48.03 rows/s
 - XML_Orders**: 11 rows in 0.23s, 48.03 rows/s
- Lookup Components:**
 - "DIM_DATE"**: 366 rows in 0.38s, 955.61 rows/s
 - Product_Lookup (Lookup)**: 42 rows in 0.01s, 3500 rows/s
 - "DIM_CUSTOMER"**: 974 rows in 0.11s, 9018.52 rows/s
 - Date_Lookup (Lookup)**: 366 rows in 0.38s, 955.61 rows/s
 - Customer_Lookup (Lookup)**: 974 rows in 0.11s, 9018.52 rows/s
- Map Component:**
 - tMap_1**: 11 rows in 0.26s, 42.64 rows/s
- Output Component:**
 - "FACT_ORDERS"**: 11 rows in 0.26s, 42.64 rows/s
- Move Component:**
 - tFileCopy_1**: Move pa...

The interface includes a left sidebar with a **Repository** tree showing project structure (LOCAL: BalCon2k16_Project, Business Models, Job Designs, BalCon2k16, F99_Finished, J01_DIM_DATE 0.1, J02_DIM_PRODUCT 0.1, J03_DIM_CUSTOMER 0.1, J04_FACT_ORDERS_BASIC 0.1, J04_FACT_ORDERS_EXTENDED 0.1, J00_Simple_Hello_World 0.1, demo, Preparation, Contexts, Code, SQL Templates, Metadata, Documentation, Recycle bin). The bottom left shows an **Outline** of job components (tFileCopy_1, tFileInputXML_1 (XML_Orders), tFileList_1, tMap_1, tMysqlInput_1 ("DIM_DATE"), tMysqlInput_2 ("DIM_CUSTOMER"), tMysqlInput_3 ("DIM_PRODUCT"), tMysqlOutput_2 ("FACT_ORDERS")). The bottom right shows a **HomeLab** table with configuration details.

Name	Value
Folder	/home/dimitar/tale
PersonalParam	balcon2k16
3306	localhost
r	balcon
name	balcon
password	balcon2k16

Design Workspace
Provides us with a playground to design our Jobs

General look and feel

The screenshot displays the Talend Studio interface with a job named "Job J04_FACT_ORDERS_BASIC 0.1". The job is configured with the following components and flow:

- Parse and load orders data** (Job Title)
- tFileList_1** (Starts the job)
- Iterate** (Loop component)
- XML_Orders** (XML Input/Output component)
- tMap** (Map component)
- Order_Data (Main)** (Data Table component)
- Orders_List (Main)** (Data Table component)
- "FACT_ORDERS"** (Data Table component)
- OnComponentOk** (Event-driven component)
- tFileCopy_1** (File Copy component)

A blue box highlights the **Configuration Tabs** section, stating: "Allow us to control the components behavior and execute Jobs".

The **Job J04_FACT_ORDERS_BASIC** configuration is shown below:

Basic Run	Execution
Debug Run	Run Kill Clear
Advanced settings	Starting job J04_FACT_ORDERS_BASIC at 08:59 10/09/2016.
Target Exec	[statistics] connecting to socket on port 3571
Memory Run	[statistics] connected

The **HomeLab** configuration is shown on the right:

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

General look and feel

The screenshot displays the Talend Studio interface with a job titled "Job J04_FACT_ORDERS_BASIC 0.1". The design workspace shows a flow starting from a "tFileList_" component, followed by an "Iterate" loop containing an "XML_Orders" component. This leads to a "tMap_1" component, which is connected to three lookup components: "DIM_DATE", "DIM_PRODUCT", and "DIM_CUSTOMER". These lookups feed into a "FACT_ORDERS" table output component, which is then connected to a "tFileCopy_1" component. Performance metrics are visible on the flow lines, such as "11 rows in 0.23s" and "48.03 rows/s".

On the left, the "Repository" pane shows a tree structure of projects and jobs, with "J04_FACT_ORDERS_EXTENDED 0.1" selected. Below it, the "Outline" tab lists the components in the job: "tFileCopy_1", "tFileInputXML_1 (XML_Orders)", "tFileList_1", "tMap_1", and three "tMySQLInput" components for the dimensions and the fact table.

On the right, the "Palette" pane provides a search bar and a categorized list of components for selection.

A blue callout box is overlaid on the bottom right of the design workspace, containing the following text:

Outline and Code tabs
The **Outline** tab lists the components that have been added to the design workspace. The **Code** tab displays the code associated with each component

General look and feel

The screenshot displays the Talend Studio interface with a job named "Job J04_FACT_ORDERS_BASIC 0.1". The job design is visible in the Designer view, showing a sequence of components: tFileList_1, Iterate, XML_Orders, tMap_1, "FACT_ORDERS", and tFileCopy_1. A blue box with the text "Palette Contains the different components we use to build our Jobs" is overlaid on the job design, with an arrow pointing to the Palette on the right. The Palette lists various components categorized by type, including Favorites, Recently Used, Big Data, Business Intelligence, Business, Cloud, Custom Code, Data Quality, Databases, DotNET, ELT, ESB, File, Internet, Logs & Errors, Misc, Orchestration, Processing, System, Talend MDM, Unstructured, and XML.

The job execution status is shown in the bottom panel, indicating that the job is running. The execution log shows the following details:

- Starting job J04_FACT_ORDERS_BASIC at 08:59 10/09/2016.
- [statistics] connecting to socket on port 3571
- [statistics] connected

The job execution progress bar shows the following details:

- 1 exec running
- 11 rows in 0.23s
- 48.03 rows/s
- Order_Data (Main)
- 11 rows in 0.26s
- 42.64 rows/s
- Orders_List (Main)
- ok
- OnComponentOk
- tFileCopy_1

The job execution status is also shown in the bottom panel, indicating that the job is running.

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

Building blocks of a data warehouse

- **Dimensions**

A dimension is a structure that categorizes facts and measures in order to enable users to answer business questions. Commonly used dimensions are people, products, place and time. Historical changes in dimensions are usually tracked by SCD management methodologies referred to as Type 0 through 6.

- **Facts**

A fact is a value or measurement, which represents a fact about the managed entity or system.

Wikipedia

Business case

What is the problem and how to deal with it?

The customer

LinuxGoods.rs is a local Serbian on-line shop for Linux and Unix related merchandise like:

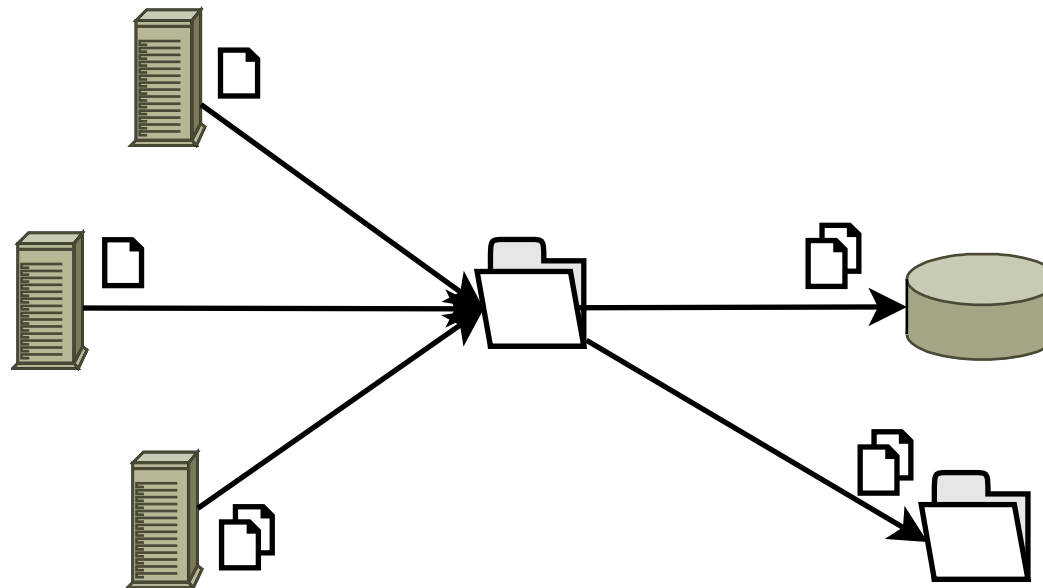
- Badges
- Stickers
- T-shirts
- Hats
- and etc.

The case

As their business was growing they began to realize that there had to be a way to analyze what is going on. It would allow them to keep the trend.

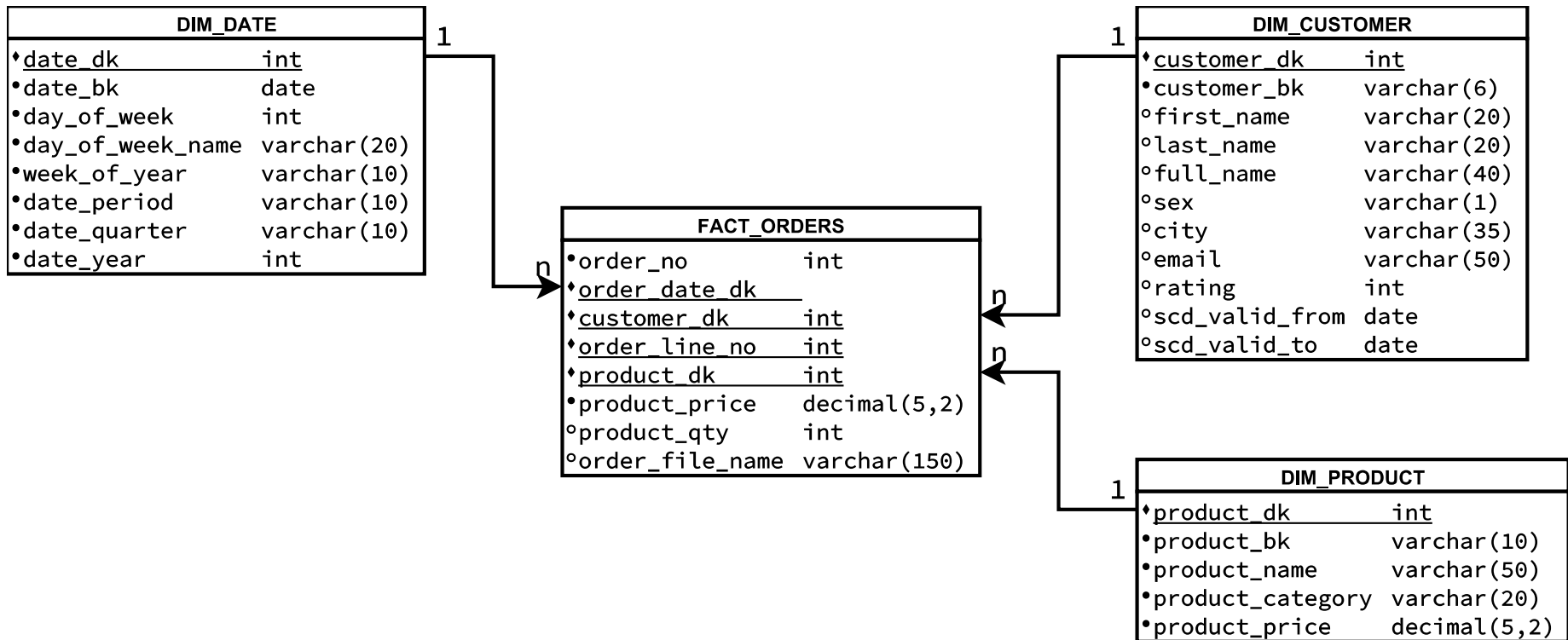
So they decided to build a small data warehouse to meet their growing need for analytical overview of the business.

The landscape



Three source systems and one target – a database. Input data is coming in three forms - plain text files, excel files, and XML files. Part of the processed files should be moved in another folder for archiving purposes.

The solution



Demo

Talend Open Studio for Data Integration in action

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
@shekeriev

BALCCON2K16

