Introduction to
Talend Open Studio for Data Integration

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

BALCCON2K16

talend

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

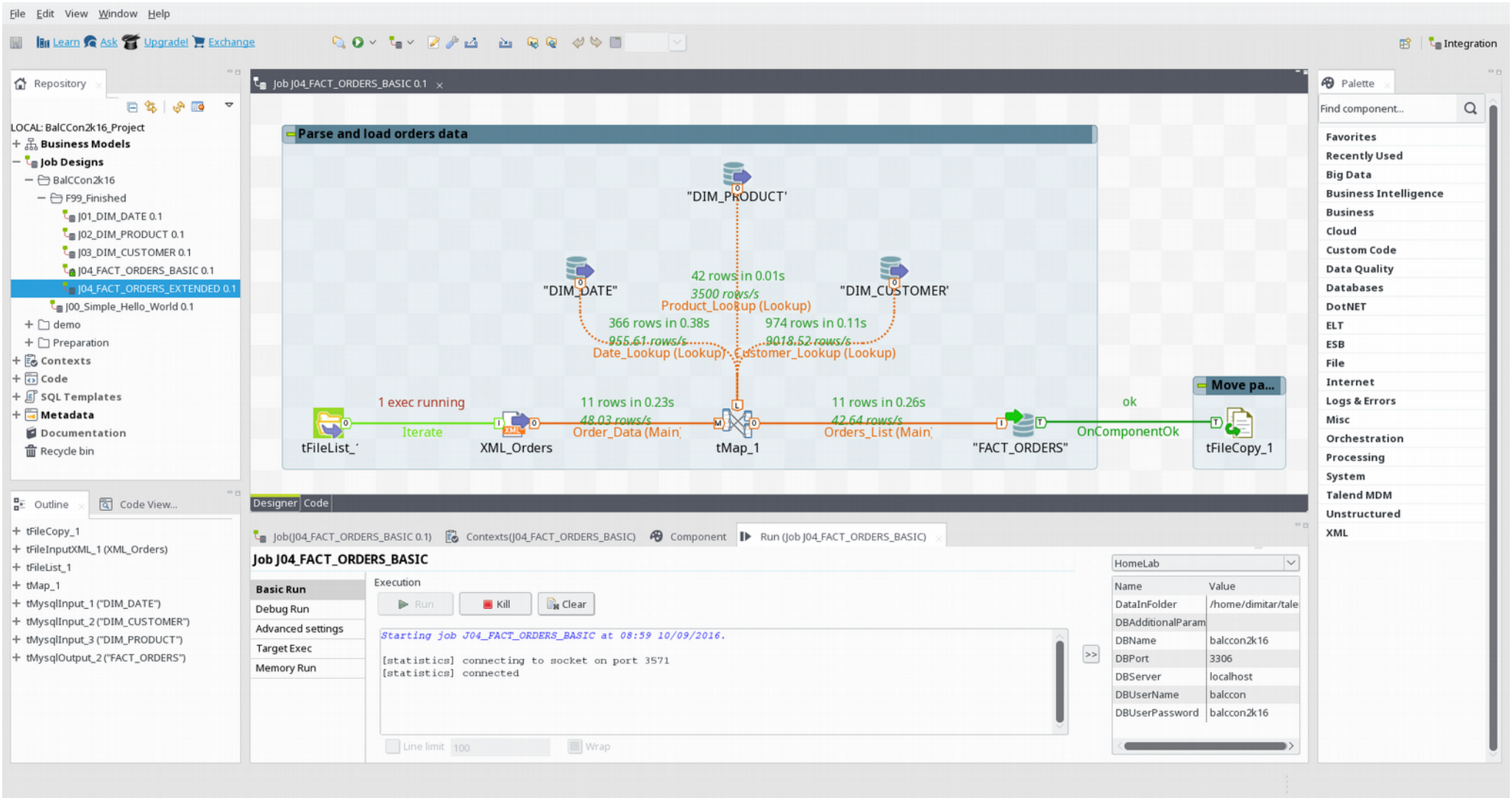
- **Project**

  Logical grouping of one or more jobs

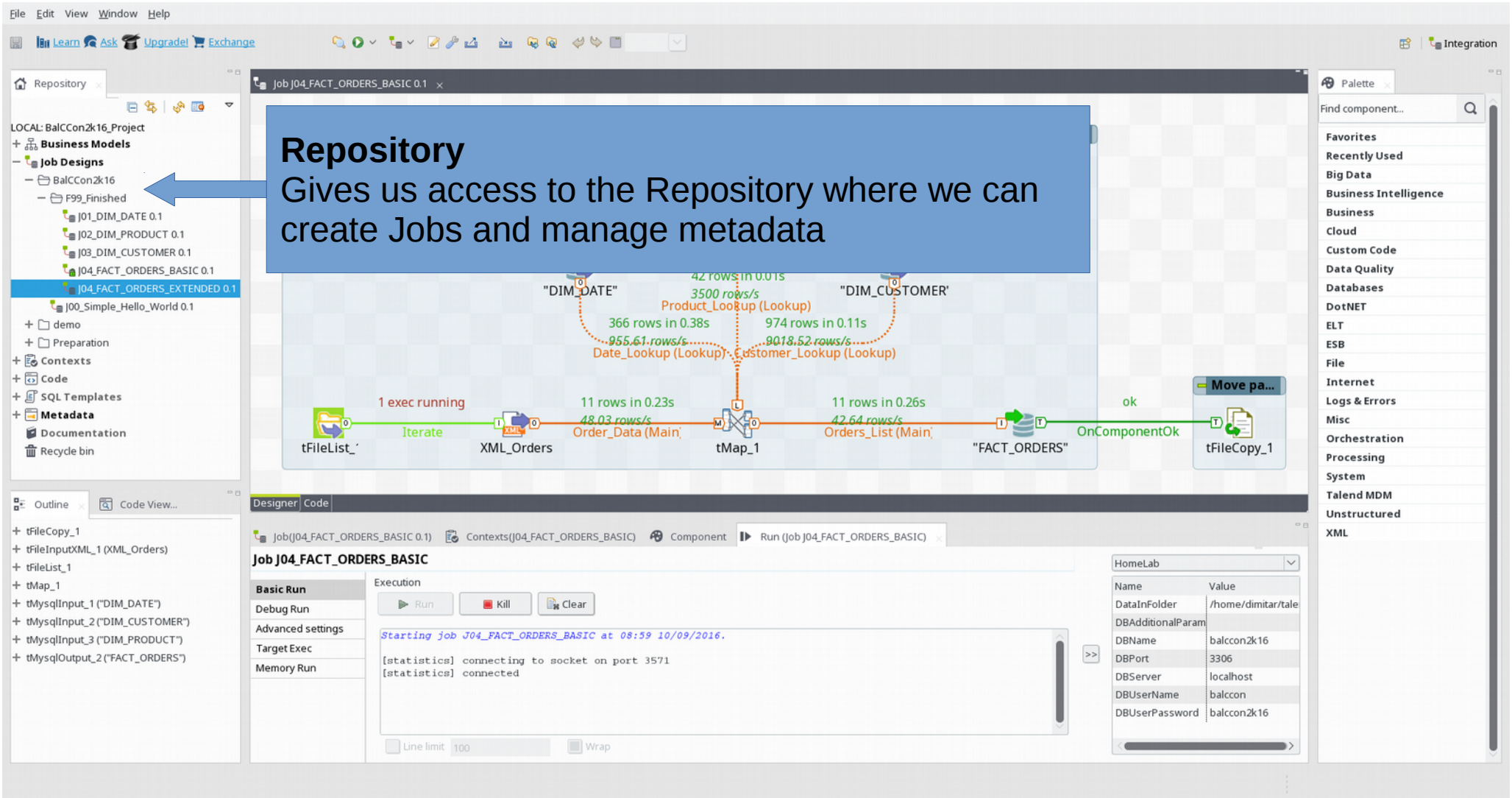
- **Job**

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

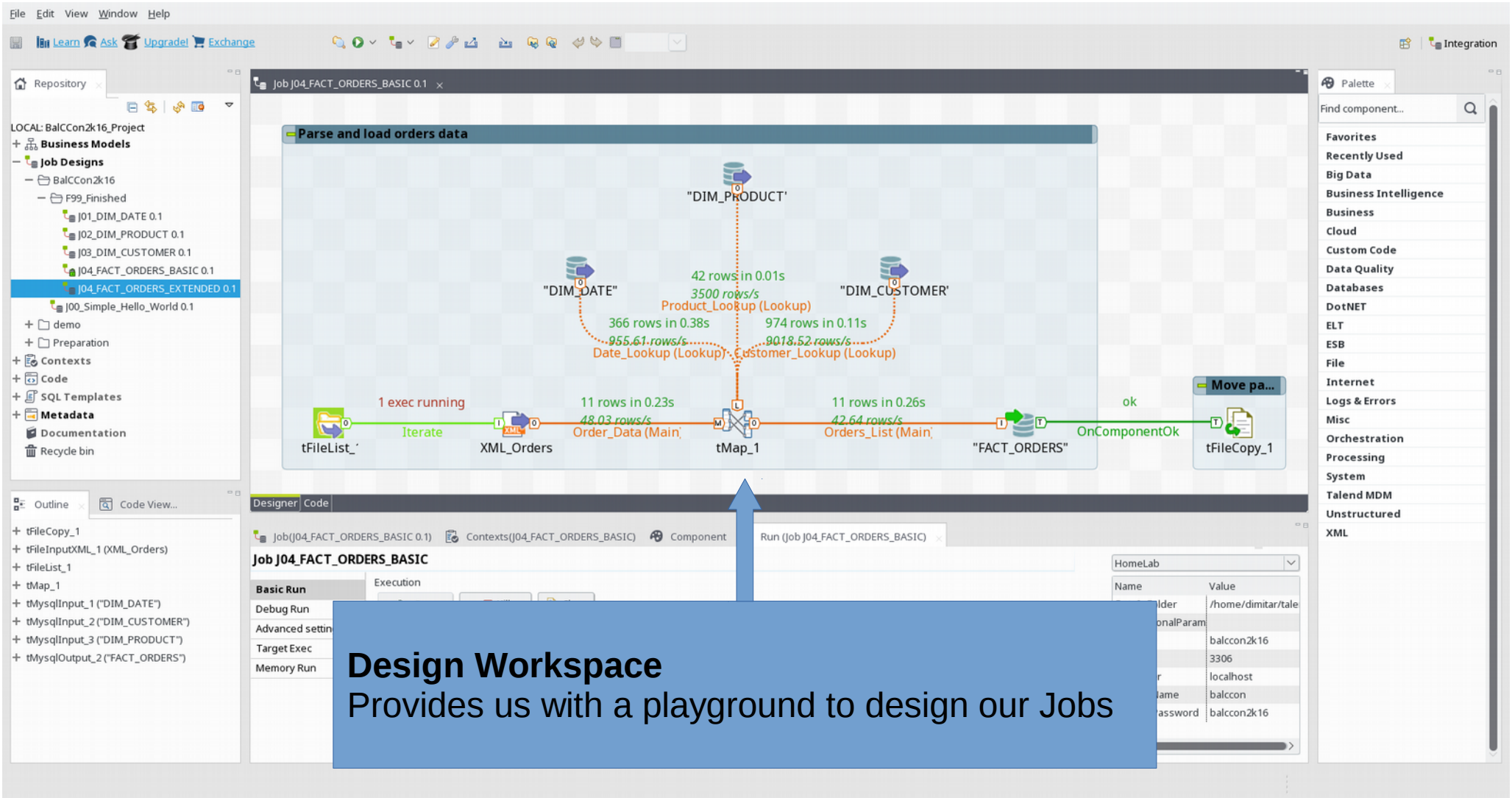
General look and feel



General look and feel

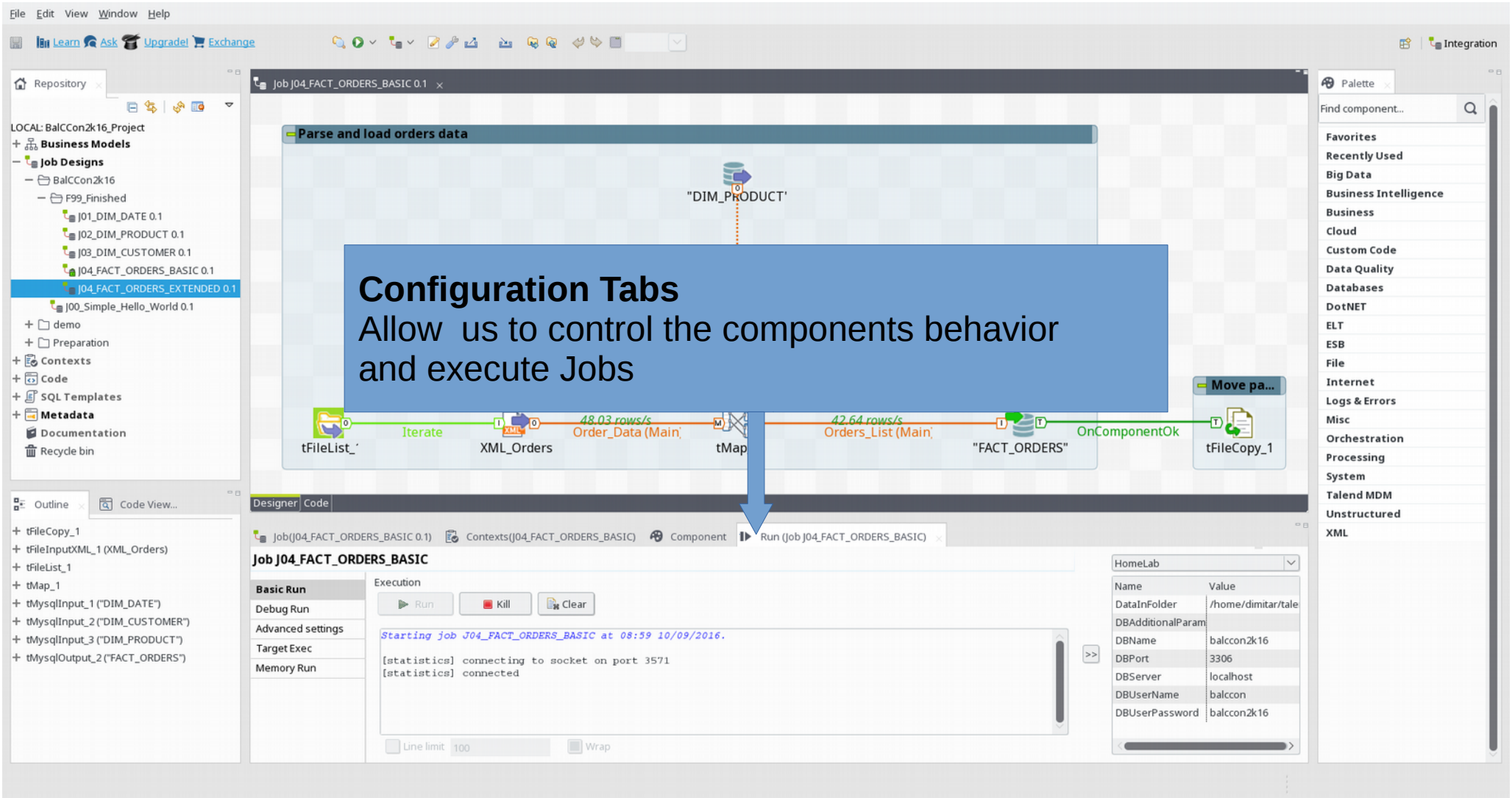


General look and feel

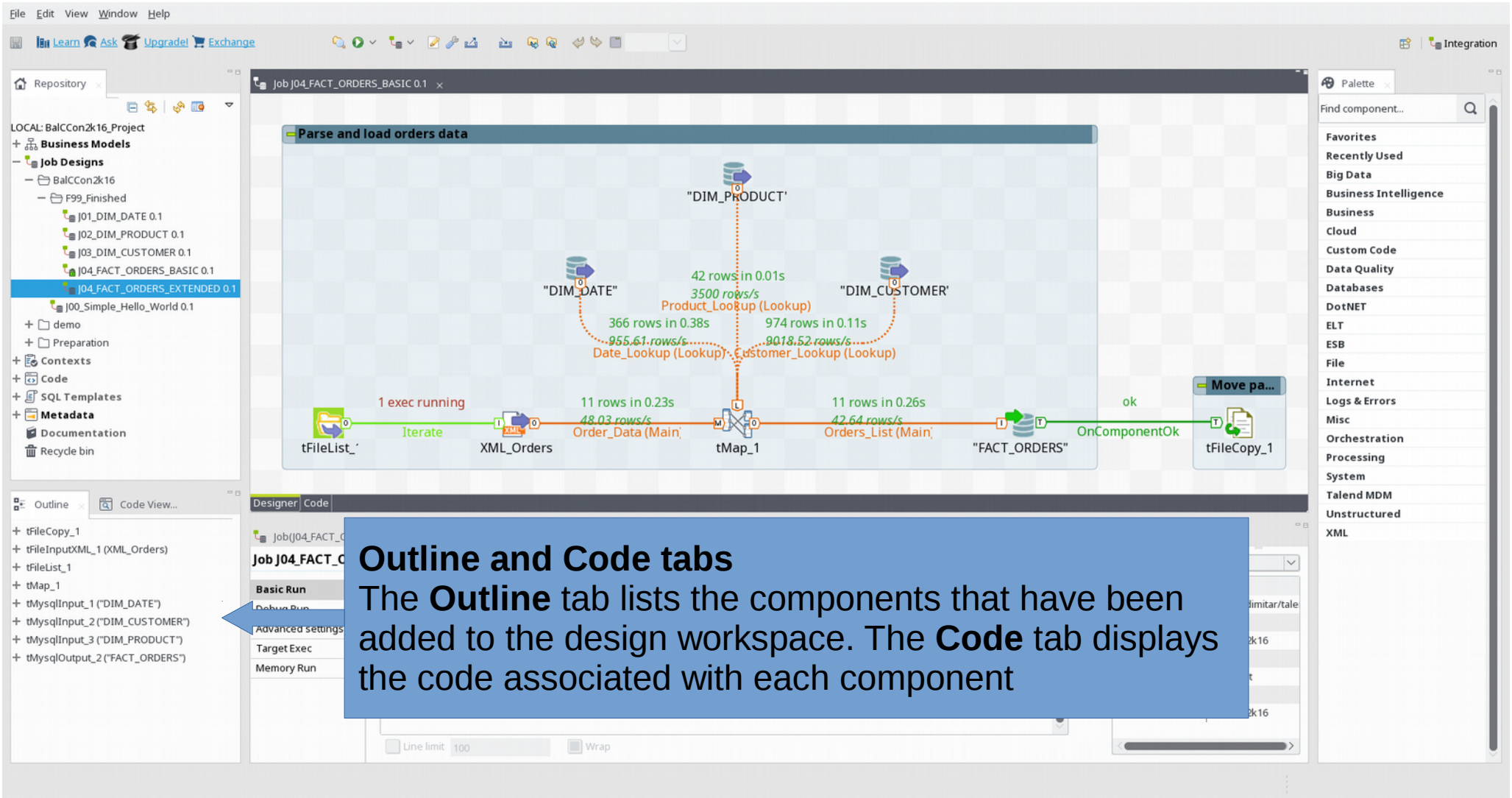


Design Workspace
Provides us with a playground to design our Jobs

General look and feel



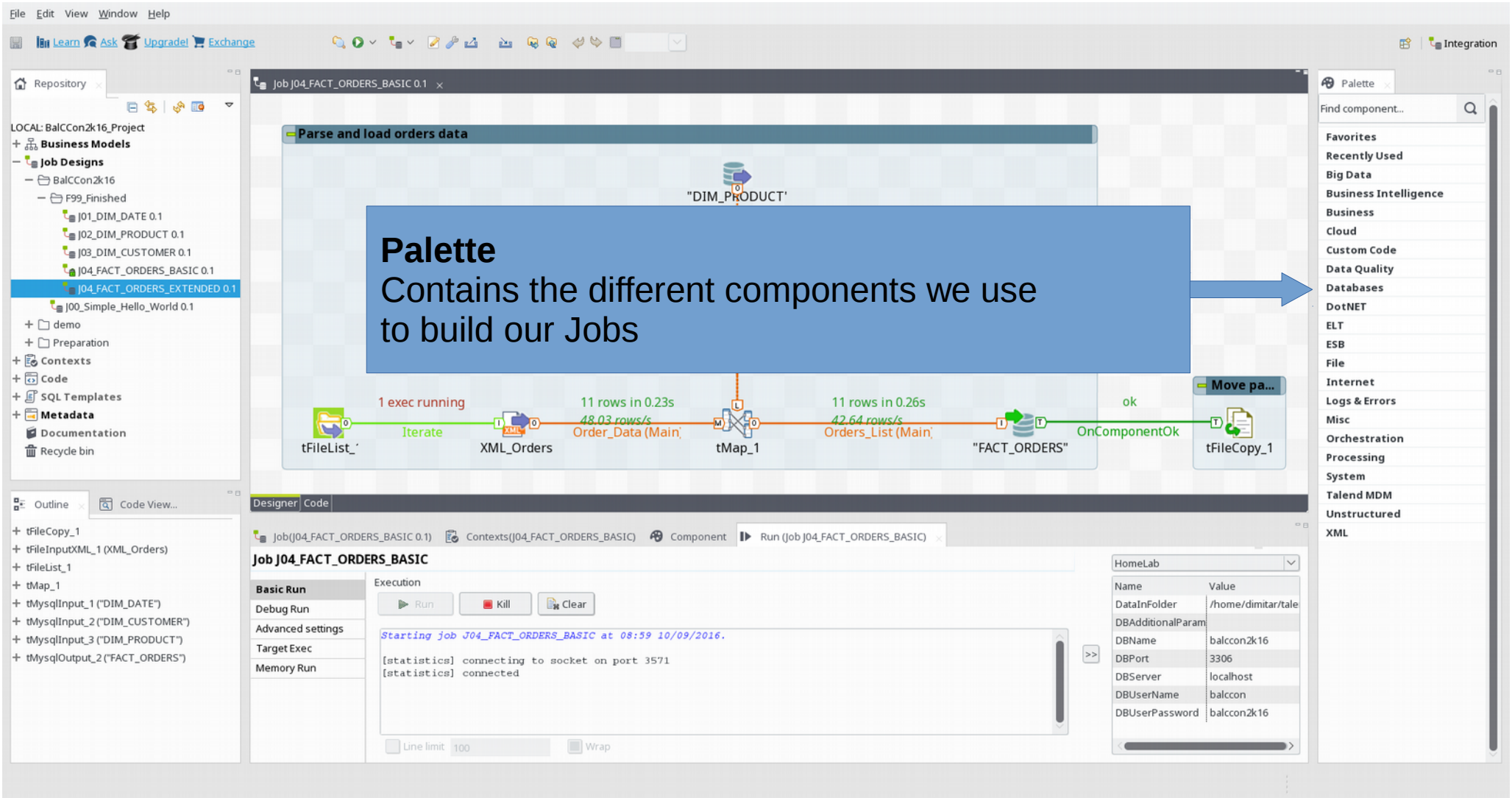
General look and feel



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



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:

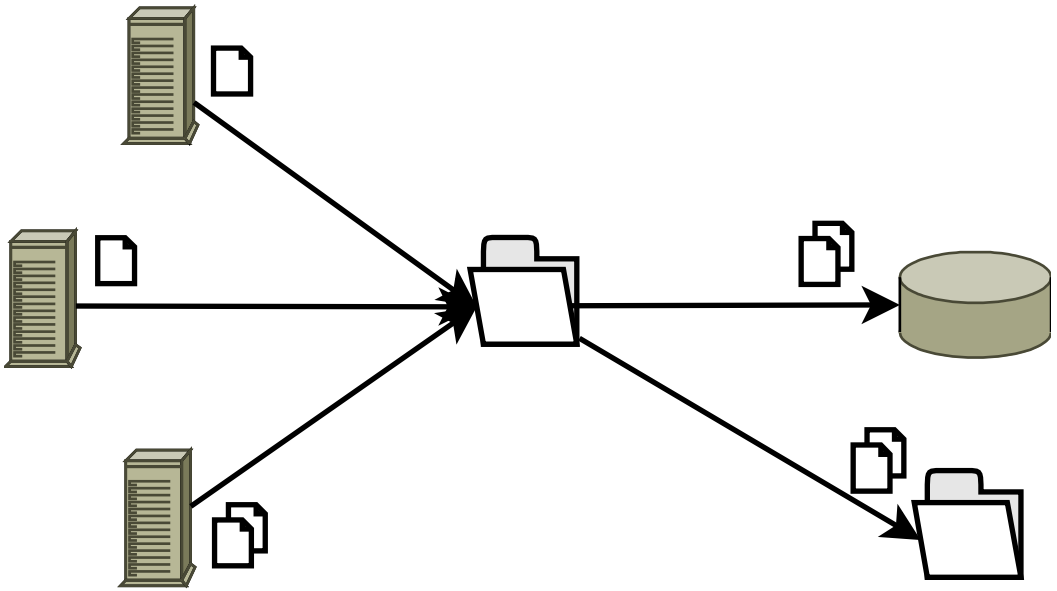
- Badges
- Stickers
- T-shirts
- Hats
- and etc.

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.

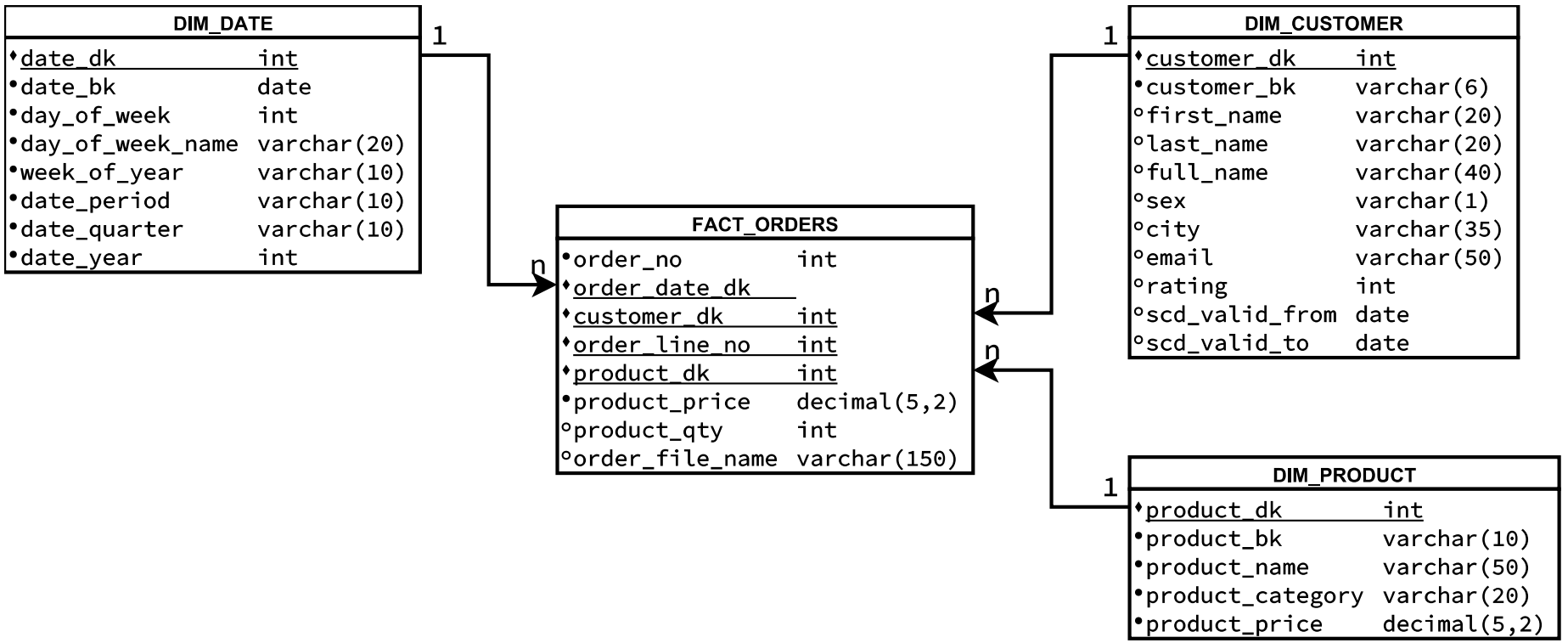
So they decided to build a small data warehouse to meet their growing need for analytical overview of the business.

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/balccon2k16

- 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

talend