Community Experience Distilled

Getting Started with Talend Open Studio for Data Integration

Develop system integrations with speed and quality using Talend Open Studio for Data Integration

Forewords by Yves de Montcheuil, VP of Marketing, Talend
Olivier Carbone, Knowledge Manager, Talend

Jonathan Bowen

[PACKT] PUBLISHING open source community experience distilled

Getting Started with Talend Open Studio for Data Integration

Develop system integrations with speed and quality using Talend Open Studio for Data Integration

Jonathan Bowen

[PACKT] PUBLISHING
open source
community experience distilled

BIRMINGHAM - MUMBAI

Getting Started with Talend Open Studio for Data Integration

First published: November 2012

Production Reference: 1251012

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-84951-472-9

www.packtpub.com

Cover Image by Dean Taylor (dean.taylor@bt.com)

Credits

Author
Jonathan Bowen

Reviewers
Mark Chapman
Carbone Olivier
Philip Yurchuk

Acquisition Editor
Mary Nadar

Lead Technical Editor
Azharuddin Sheikh

Technical Editors
Veronica Fernandes
Ankita Meshram

Copy Editors
Insiya Morbiwala
Aditya Nair
Alfida Paiva
Laxmi Subramanian

Project Coordinator
Yashodhan Dere

Proofreader
Maria Gould

Indexer
Rekha Nair

Production Coordinator
Melwyn D’sa

Cover Work
Melwyn D’sa

Foreword

Talend's open source approach shatters the traditional proprietary model by supplying open, innovative, and powerful software solutions with the flexibility to meet the needs of all the organizations. By publishing the code of its core modules under the GNU Public License or the Apache License, Talend offers the developer community the ability to improve products and make enhancements that can benefit everyone.

Contributions from the community are critical to the success of Talend's products. They take many different forms: code contributions, extensions such as connectors and components, documentation and tutorials, community support and help, and more.

Documentation is probably one of the most important aspects of the usability experience. Without good learning and reference materials, the best software is impossible to fully master. Contributions from the community to this usability experience are hence very important and they participate to the broad adoption effort.

It should be no surprise then that Talend is supportive of initiatives such as this book. By providing an angle enriched by his real-life experience, the author guides the user of Talend Open Studio through a learning experience that is complementary to the user documentation provided by Talend.

Enjoy!

Yves de Montcheuil

VP of Marketing,
Talend

Foreword

As with any software, there are many different ways to learn how to properly use Talend Open Studio for Data Integration. There are many helpful resources now available online, thanks to the Talend company and the community. Undoubtedly, it may be difficult to choose which training path to take: Which tutorial shall I begin with? Which topics in the forum will be of use? Have I understood the explanatory video?

The method described by Jonathan Bowen in this book is straightforward. It is based on hands-on examples. There's no need for previous knowledge and anyone can try to perform the exercises.

Throughout the chapters, you will discover the main features of Talend Open Studio and learn the best practices. Take your time when setting up the technical environment and when solving all of the exercises. Learning will be easier that way.

Interested readers should consider studying for Talend Certification. Keep upgrading your skills with online resources as described in the last chapter.

Olivier Carbone

Knowledge Manager,
Talend

About the Author

Jonathan Bowen is an E-commerce and Retail Systems Consultant and has worked in and around the retail industry for the past 20 years. His early career was in retail operations, then in the late 1990s he switched to the back office and has been integrating and implementing retail systems ever since.

Since 2006, he has worked for one of the UK's largest e-commerce platform vendors as Head of Projects and, later, Head of Product Strategy. In that time he has worked on over 30 major e-commerce implementations.

Outside of work, Jonathan, like many parents, has a busy schedule of sporting events, music lessons, and parties to take his kids to, and any downtime is often spent catching up with the latest tech news or trying to record electronic music in his home studio.

You can get in touch with Jonathan at his website: `www.learnintegration.com`.

Acknowledgement

I am grateful to the editorial team at Packt – Theresa, Mary, Yashodhan, Azhar in particular – for helping with the production of this book. Their advice and guidance has been critical in getting it published. I would also like to thank the technical reviewers – Mark Chapman, Olivier Carbone, and Philip Yurchuk for their feedback. My friend and colleague, Dean Taylor, provided the book's cover photograph. Nice one Dean!

I have had two significant lucky breaks in my technical career: firstly, when I joined STS Systems, a retail systems vendor (thanks Peter!) and secondly, when I joined Fresca, a UK e-commerce platform provider (thanks Sarah, Gavin, and Justin!). Both experiences had a huge impact on me and greatly influenced my career path and technology skills. This book is, in large part, the result of those influences. I've also been fortunate to work with many fantastic people over the years that have contributed to my technical education in one way or another. A big thanks to you all!

I'd like to thank my parents and family for their support and encouragement, not just while writing this book, but since day one!

Finally, this book is dedicated to the three most important people in my life – Tanya, William, and Rose.

About the Reviewers

Mark Chapman is the Pre-Sales Manager for Talend in the UK, the leading Open Source Integration software vendor. He joined Talend in October 2009 as a Technical Consultant and is now the primary hands-on Pre-sales Manager for Integration (ETL and ESB), Data Quality (DQ), Master Data Management (MDM), and Business Process Management (BPM) for the Enterprise sector.

Mark has over 25 years of business experience in the Information Technology industry, the past 18 years as a consultant for Enterprise Master Data Management software vendors including Search Software America (now Informatica), Datactics, Datanomics (now Oracle), and Talend. Mark works with organizations from all sectors and sizes to understand how EDM technologies encompassing ETL, AI, DQ, MDM, and BPM can be deployed to meet the needs of the business.

`http://uk.linkedin.com/in/markvchapman`

Olivier Carbone began his career in the world of training and in 2006, he joined the open source software vendor, Talend. He worked on the team that produced the beta version of Talend Open Studio and subsequently was responsible for producing Talend training tools. He also worked in Talend's marketing team.

In 2008, he was appointed as Customer Care Manager at Talend. His role was to ensure that the services from Talend were on par with the customers' expectations. Continuing to satisfy his customers, he took the role of the Head of the Professional Services (training + expertise + support) and has enjoyed managing teams of French consultants.

In September 2011, he came back to R&D to work on a government project linked to e-learning and to build the internal social networks.

Olivier also belongs to the social network (`http://www.apprendre2point0.org`) dedicated to the impact of technology on our way to learn/work. Since 2007, he has invested a lot of time to write the first pages of this story and define the rules of practice.

He aims to share with us his experiences on his blog at `http://ocarbone.free.fr/blog/`.

Philip Yurchuk is an independent Enterprise E-commerce and Web Application Consultant. He started his career at NASA's Jet Propulsion Laboratory developing satellite mission planning software. At Boeing, he developed and managed systems software and web applications. He then began consulting, starting with an engagement managing the development of a high volume consumer web application for Toyota. Since then, he has worked as a project manager and developer on several e-commerce sites for major retail brands. He has a bachelors degree in Computer Science from Rensselaer Polytechnic Institute. Outside of work, he is a typical geek who enjoys movies, music, books, board games, and blogging at `http://philip.yurchuk.com`.

www.PacktPub.com

Support files, eBooks, discount offers, and more

You might want to visit `www.PacktPub.com` for support files and downloads related to your book.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`http://PacktLib.PacktPub.com`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across Packt's entire library of books.

Why Subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print and bookmark content
- On demand and accessible via web browser

Free Access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view nine entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

We've all been there. Your boss drops you an e-mail saying:

> *Good news, we've just bought system X, which is going to make our lives a lot easier. First though, we need to hook it up to system Y for daily product and inventory feeds and system Z to post the financials back for invoicing. Should be easy, right? It's going to be live in two months. Any problems, please let me know. Oh....if you can get some extracts for the data warehouse at the same time, that would be great too.*

What to do? Well, you could ask your senior developer to code some integration jobs from scratch, but they might be hard to maintain, particularly if he/she left the company. In addition, you know he/she is working flat out on another important project. Alternatively, you could ask your boss if you can invest in a proprietary integration suite, with a legion of highly paid consultants. That will certainly do the job, but the budget, and timeline might not stretch to this.

Or you can take the new junior developer who joined your company a couple of weeks ago, dust off your business analyst and testing skills, and get the job done on time, on budget with Talend Open Studio for Data Integration.

Getting Started with Talend Open Studio for Data Integration is an introductory guide to solving this problem and many others like it.

What this book covers

Chapter 1, *Knowing Talend Open Studio*, introduces the reader to Talend Open Studio for Data Integration and what it can be used for. It also covers the installation of Talend Open Studio for Data Integration.

Chapter 2, *Working with Talend Open Studio*, introduces some common concepts the reader will come across when using Talend Open Studio for Data Integration, including creating a workspace to contain integration jobs, a tour of the Talend Open Studio for Data Integration interface, and use of metadata and schemas. We'll also build a simple "hello world" job.

Chapter 3, *Transforming Files*, gets into the detail of Talend Open Studio for Data Integration integrations and looks at using Talend Open Studio for Data Integration to transform files from one format to another.

Chapter 4, *Working with Databases*, looks at databases—how to get data out and how to get data in.

Chapter 5, *Filtering, Sorting, and Other Processing Techniques*, introduces common data operations: filtering, sorting, and aggregating.

Chapter 6, *Managing Files*, shows how to manage files during integration jobs. We'll look at renaming, moving, copying, and deleting files; how to timestamp a file; connecting to remote servers to FTP files; and zipping and unzipping files.

Chapter 7, *Job Orchestration*, will look at more complex integrations and how "one-shot" tasks can be combined to form multi-step jobs. We'll create subjobs and link them together using "if/then" logic. Integrations often produce temporary files, so we'll look at ways to clean up afterwards.

Chapter 8, *Managing Jobs*, covers the process of packaging, deploying, and scheduling jobs in a live environment.

Chapter 9, *Global Variables and Contexts*, looks at contexts and we explore how the same job can be used in different environments. We introduce dynamic variables, allowing our integration jobs to run flexibly, based on the current runtime information, rather than introducing complex, hardcoded routines.

Chapter 10, *Worked Examples*, brings together all of the knowledge from previous chapters in a series of worked examples. A real-life integration project is explored and developed to illustrate the use of Talend Open Studio for Data Integration "in the wild".

Appendix A, *Installing Sample Jobs and Data*, details how to obtain and use the sample data files required to follow the job development examples in the book. All of the jobs created throughout the book are also provided for reference.

Appendix B, *Resources*, highlights some resources and further reading to expand your knowledge of Talend Open Studio for Data Integration.

What you need for this book

The hardware and software requirements for this book are:

- A computer running Windows, Linux, or Mac OS with Java installed
- Talend Open Studio for Data Integration
- A text file/XML editor
- A MySQL database instance

Who this book is for

This book is for developers, business analysts, project managers, business intelligence specialists, system architects, and consultants who need to undertake integration projects. The book assumes a certain level of technical aptitude and readers should be comfortable with some of the following concepts and technologies:

- Relational database management systems with some SQL (structured query language) experience
- XML
- Java
- File Transfer Protocol (FTP)
- Programming flow and logic

Conventions

In this book, you will find a number of styles of text that distinguish between different kinds of information. Here are some examples of these styles, and an explanation of their meaning.

Code words in text are shown as follows: "Create a file delimited metadata for the `currencies.csv` file."

A block of code is set as follows:

```
String datestamp=TalendDate.getDate("YYYYMMDD");

globalMap.put("dateStamp",datestamp);
```

Any command-line input or output is written as follows:

```
sh [file name].sh
```

New terms and **important words** are shown in bold. Words that you see on the screen, in menus or dialog boxes for example, appear in the text like this: "Go to the **Debug Run** tab and click on **Traces Debug**".

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or may have disliked. Reader feedback is important for us to develop titles that you really get the most out of.

To send us general feedback, simply send an e-mail to `feedback@packtpub.com`, and mention the book title through the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide on `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files for all Packt books you have purchased from your account at `http://www.packtpub.com`. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you would report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/support`, selecting your book, clicking on the **errata submission form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website, or added to any list of existing errata, under the Errata section of that title.

Piracy

Piracy of copyright material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works, in any form, on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors, and our ability to bring you valuable content.

Questions

You can contact us at `questions@packtpub.com` if you are having a problem with any aspect of the book, and we will do our best to address it.

1
Knowing Talend Open Studio

Ever since the *second* computer system came along, integrating systems has been a key part of the work of IT teams.

Today's IT landscape is increasingly complex, with **enterprise resource planning** (**ERP**), **customer relationship management** (**CRM**), finance, warehousing, human resources, and e-business systems, both within and outside the enterprise, all needing to exchange data. The real-time nature of business today and the fast pace of business change add to the need to have a set of tools and skills that make the business of integrating systems quick and easy. New systems come along all the time, but it is also a requirement to respond quickly to new business opportunities that drive system integrations. Company takeovers and mergers, new markets and customers, new suppliers, and joint ventures are commonplace events that all require data to be exchanged on a one-off or regular basis to make them work.

As you might expect, for such a critical systems-development activity, there is no end of options to choose from to fulfill the need. From complex multi-million dollar integration suites from the major systems vendors to humble, yet powerful, scripting languages such as Perl, there is something for every budget and taste. So what is Talend Open Studio for Data Integration and why should you consider it for your next integration project?

What Talend Open Studio is

Talend Open Studio for Data Integration is an open source graphical development environment for creating and deploying custom integrations between systems. It comes with over 600 pre-built connectors that make it quick and easy to connect databases, transform files, load data, move, copy and rename files, and connect individual components in order to define complex integration processes.

Talend Open Studio for Data Integration is a code generator, and so does a lot of the "heavy lifting" for you. As such, it is a suitable tool for experienced developers and non-developers alike. Talend Open Studio for Data Integration is easy to use and reduces the time taken to develop integrations from weeks and months to days or even hours.

Integration jobs are created from components that are configured rather than coded and jobs can be run from within the development environment or executed as standalone scripts.

Use cases

Some common use cases for Talend Open Studio for Data Integration include:

- **Data migration from one database to another**: This is a common scenario when new systems are implemented or existing systems are upgraded. Data has to be populated into the new or upgraded system and database schemas may be subtly or completely different, requiring some modification of the data prior to loading. Data migrations tend to be "one-off" activities, not integrations that are deployed on an ongoing basis. The Studio facilitates data migrations through its many database connectors and actions.
- **Regular file exchanges between systems**: The humble flat file is still a cornerstone of many systems integrations. Their low-tech approach makes them particularly suitable for batch processes when real-time data flows are unnecessary. File exchanges will often require some form of file transformation, either data content, data format, or both. The Studio has the ability to manage many different file formats and, with its file management capabilities such as FTP and archiving (zipping), is able to facilitate a full end-to-end file exchange process.
- **Data synchronization**: Enterprises often have multiple data repositories of the same data. For example, data about customers might reside in the CRM system, the finance system, and the distribution system. They will probably have similar but different data models across these systems and every time a change is made in one, the same change needs to be made in the others—typically a time-consuming and manual process. The Studio can be used to keep the data in sync across systems with jobs that automate and transform the data transfer.
- **ETL** (**Extract, Transform, and Load**): A key component process of a data warehouse or business intelligence system, ETL processes extract data from operational systems, transform the data, applying a series of rules or functions, and load the data into a database or data warehouse system.

History of Talend Open Studio

Talend was founded in 2005 and is an open source software vendor providing solutions for data integration, data quality, master data management, enterprise service bus, and business process management.

Talend's first product, Talend Open Studio for Data Integration, was launched in 2006, under the name Talend Open Studio, and has since been downloaded over 20 million times. Talend has continued to develop its product portfolio and has added complementary tools that provide a single platform for application, data, and process integration. The Talend Open Studio brand has since been adopted across the range of Talend's products.

Benefits of Talend Open Studio

An obvious question to ask is "Why should I use Talend Open Studio above other similar products? What can it do for me?" Talend Open Studio for Data Integration offers a number of benefits:

- The Studio is open source, free to download and use, with access to the source code, allowing users to extend the product to their particular needs if required.
- The Studio is a great productivity-booster. It's easy to learn and quick to develop with. Even novice developers will be building complex integrations in no time.
- The Studio's pre-built components handle many common and not-so-common tasks. Developers can focus on the end-to-end process, rather than the low-level technical details.
- Talend has an active and open user community. Practical, problem-solving advice is easy to access.

Installing Talend Open Studio

Before we can begin, we need to install the Studio. Talend provides installation guides and other material on its wiki at the following URL:

`http://www.talendforge.org/wiki/doku.php?id=doc:installation_guide`

We will also cover the basic installation instructions here.

Prerequisites

The Studio is a cross-platform application, running on Windows, Linux, and Mac OS. A list of hardware and software prerequisites can be found at `http://www.talend.com/docs/community/prerequisites.html`.

As a minimum, you will need a supported operating system, Java, and of course, the Studio itself.

Installation guide

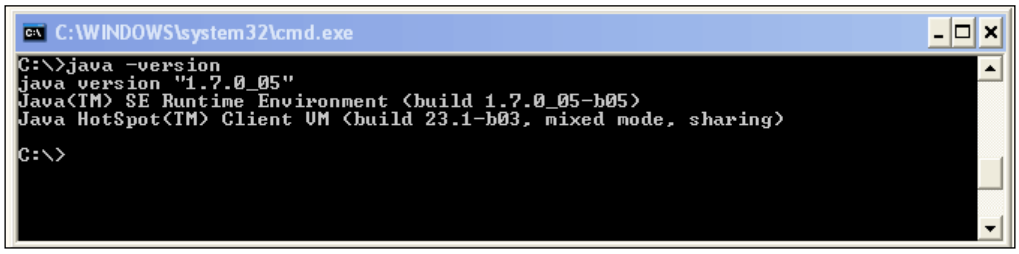
The installation process for the Studio is essentially the same across all supported operating systems. We will show how to complete the installation on Windows, but you can follow the same steps on other platforms.

Follow the instructions given to install the Studio on Windows:

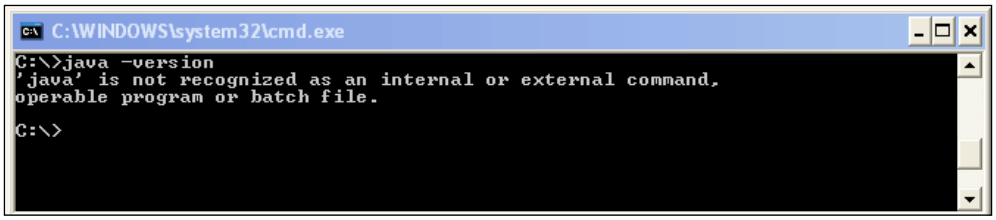
1. Check to see if Java is installed on your computer by opening a command window and running the following command:

   ```
   java -version
   ```

2. If Java is present, you will see a message showing which version is installed, as shown in the following screenshot:



   In the preceding screenshot, you can see that Version 1.7.0_05 of Java is installed. If Java is not present, you will get an error message, as shown in the following screenshot:



3. If you need to install Java, visit the following URL to download a Java installer:

   `http://www.oracle.com/technetwork/java/javase/downloads/index.html`

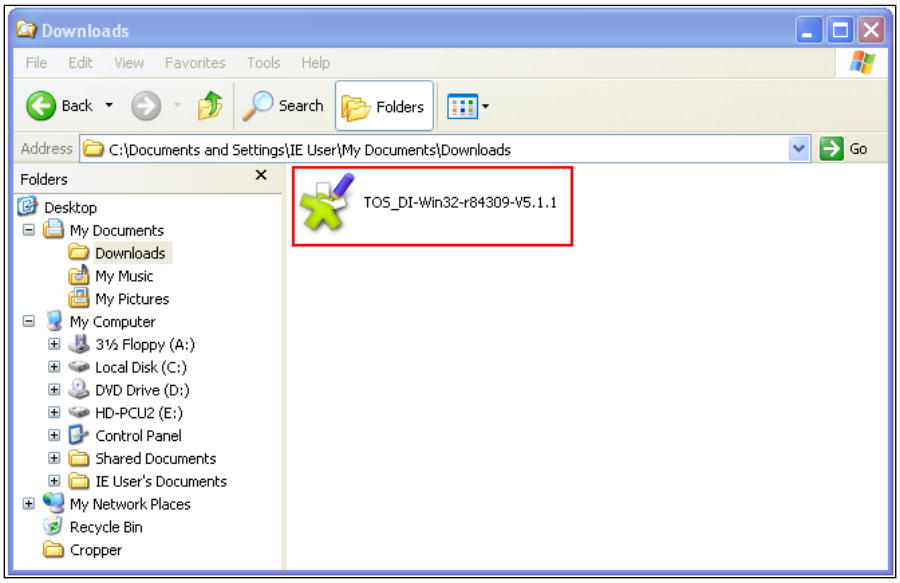
   There are various versions of the Java Standard Edition JDK for different operating systems. Choose the appropriate version for your computer and download the installer to your computer.

4. Once the installer is downloaded, click on the executable file to run it. Follow the instructions on the installer as it progresses.
5. Now that Java is installed, we can download and install the Studio. Start by going to the Talend download page at the following URL:

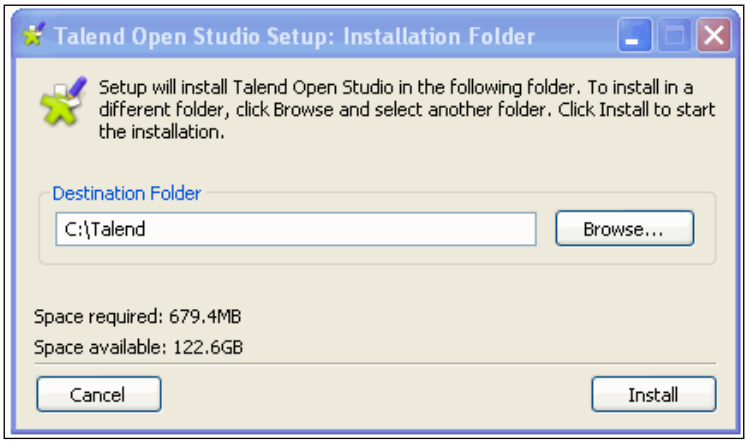
   `http://www.talend.com/download.php`

6. Choose the **Data Integration** tab and click on the **Download** button for **Talend Open Studio for Data Integration,** as shown in the following screenshot:

7. Once it has downloaded, double-click on the executable to extract the Studio files as shown in the following screenshot:



8. Follow the installation instructions on-screen. You will be prompted to choose an installation directory. Enter an appropriate location such as `C:\Talend`, as shown in the following screenshot:



Once the installation is complete, you can start the Studio and start to develop jobs. See *Chapter 2, Working with Talend Open Studio*, for details on how to start the Studio.

Other useful software

In order to follow the sample jobs throughout the book, you may wish to install some additional software.

Text editor

A decent text editor will be very useful to view CSV and XML files. There are hundreds of text editors—both free and paid-for—and here are some recommendations if you don't already have a favorite:

- If you are using a Linux operating system, you will probably have at least one good text editor installed as part of your distribution. gedit (`http://projects.gnome.org/gedit/`) is the official text editor of the GNOME project and will do the job admirably.
- Windows users can download Notepad++ (`http://notepad-plus-plus.org/`), which really is a double-plus compared to the default Notepad application that Windows provides.
- Mac users can pick up TextWrangler from `http://www.barebones.com/products/TextWrangler/`.

MySQL

Chapter 4, *Working with Databases*, focuses on using the Studio to extract from and insert data into a relational database system. The Studio supports many different database systems, but for the examples in this book, we have chosen to use MySQL.

MySQL is the most popular open source relational database and is used by many large-scale applications and websites. It is free to use and there are a number of tools you can use to administer databases. To follow the examples as they are, use MySQL. However, if you have another preferred database you wish to use, it should not be too difficult to modify the job examples to incorporate other database components instead of the illustrated MySQL components.

MySQL Community Server can be downloaded from the following URL:

```
http://dev.mysql.com/downloads/mysql/
```

Installation instructions for various operating systems can be found at the following URL:

```
http://dev.mysql.com/doc/refman/5.1/en/installing.html
```

Once you have installed the MySQL server, download and install the client tools, which you can use to administer the database, view data, and so on. The MySQL Workbench can be downloaded from `http://www.mysql.com/downloads/workbench/`.

MySQL Workbench documentation, including installation instructions, can be found at `http://dev.mysql.com/doc/workbench/en/`.

Readers who wish to use other database systems can find a full list of supported databases at `http://www.talendforge.org/components/`.

The list includes Oracle, DB2, MS SQL, Postgres, SQLite, and Sybase, among others. TOS also supports the JBDC API to connect to, and a relational database that supports this protocol.

Sample jobs and data

Each chapter of the book contains a number of example jobs that we will construct in a systematic manner. Readers are encouraged to follow the steps in order to get the most out of the book and consolidate their learning as they go. However, you can download and import the full set of example jobs if you wish.

Additionally, some jobs rely on database data and file-based data sources to work correctly. Again, these data sources can be downloaded and installed prior to working through the examples.

Appendix A, *Installing Sample Jobs and Data*, gives full instructions on downloading and installing the example jobs and data files.

Note that some sample data files may have their encoding changed as they are downloaded, unzipped, and copied from one location to another. As a result you may occasionally get some encoding errors notified in the Studio. If this happens, open the offending file and ensure it is saved with the UTF-8 encoding.

Summary

Welcome to Talend Open Studio for Data Integration! In this chapter, we learned what the Studio is and what it can be used for. We walked through installing the Studio on your computer (along with some additional useful software).

Our next step is to log on to the Studio, become familiar with the Studio working environment, and create a simple job to illustrate the development workflow. All of this will be covered in *Chapter 2*, *Working with Talend Open Studio*.

2
Working with Talend Open Studio

Now that we have the Studio installed, we can start to build integration jobs. However, before we dive straight in with some complex developments, let's get familiar with the working environment, get ourselves organized, and start with something simple.

In this chapter, we will:

- Learn how to log on to the Studio
- Get a tour of the Studio's environment and find out the different elements that make up the Studio tool
- Learn how to create a new project and a new job
- Learn about metadata—what it is and how it is used in the Studio

Studio definitions

Let's start with a few definitions to make everything clear:

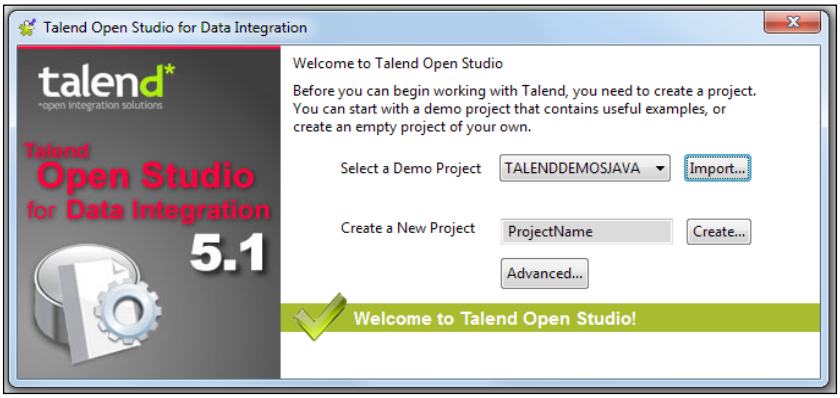
- A *workspace* is a directory on your computer that contains one or more projects
- A *project* is a logical grouping of one or more jobs
- A *job* is a group or one or more components that, when executed, implement a data flow or integration process

We will create each of these as we work through the chapter.

Starting the Studio

The Studio is a cross-platform development tool and supports Windows, Linux, and Mac OS in both 32-bit and 64-bit versions. To start the Studio, go to the directory where the Studio was installed, and double-click on the executable appropriate for your operating system.

1. When you start the Studio for the first time, you will be presented with a license notification. Click on **Accept** to proceed. We will then see the first-time start up screen and we are presented with a few options at this point. We can:
    - Import a demo project
    - Create a new project
    - Change some basic settings



2. We will start by amending some settings. Click on **Advanced**. You will see the following screen: