Google Developer Student Clubs

Data Engineering Workshop

Tawfik Y. Tawfik

LinkedIn

Data Engineering Workshop

GDSC

Session #3: Introduction to Talend Open Studio

- Who didn’t attend Session #2

Talend Open Studio

Introduction

- Talend Open Studio is a free open source ETL tool for Data Integration and Big Data. It is an Eclipse based developer tool and job designer. You just need to Drag and Drop components and connect them to create and run ETL or ETL Jobs. The tool will create the Java code for the job automatically and you need not write a single line of code.

Talend Open Studio

Introduction

- Some important benefits which Talend Open Studio offers are as below –
    - Provides all features needed for data integration and synchronization with 900 components, built-in connectors, converting jobs to Java code automatically and much more.
    - The tool is completely free, hence there are big cost savings.
    - In last 12 years, multiple giant organizations have adopted TOS for Data integration, which shows very high trust factor in this tool.
    - The Talend community for Data Integration is very active.
    - Talend keeps on adding features to these tools and the documentations are well structured and very easy to follow.

Talend Open Studio

Outline for Today

- Components
- Using Connections in a Job
- Metadata
- Using Routines
- Mapping a Data Flow (tMap)
- Additional Concepts

* There are a lot of concepts and learnings in Talend, We will see the basics only and you must continue learning.

Talend Open Studio

Components

- Adding Components.
  - (Drag & Drop – Typing on the Design Workspace)
- Settings Tab.

Talend Open Studio

Connections

- **Row Connections:** Many types, we will talk about Main and Lookup
  - **Main:** This type of row connection is the most used connection. It passes on data flows from one component to the other, iterating on each row and reading input data according to the component properties setting (schema).
  - **Lookup:** This row connection connects a sub-flow component to a main flow component (which should be allowed to receive more than one incoming flow). This connection is used only in the case of multiple input flows.

Talend Open Studio

Connections – Cont.

- **Trigger Connection:** Trigger connections define the processing sequence, so no data is handled through these connections.

Any Questions?

Feel free to ask and interact

Talend Open Studio

Metadata

- The Metadata folder in the Repository tree view stores reusable information on files, databases, and/or systems that you need to create your Jobs.
- Various corresponding wizards help you store these pieces of information that can be used later to set the connection parameters of the relevant input or output components and the data description called "schemas" in a centralized manner in Talend Studio.
- The procedures of different wizards slightly differ depending on the type of connection chosen.

Talend Open Studio

Metadata – Cont.

- Database metadata.
- File Delimited metadata.
- File Excel metadata.
- Etc.

Break Time

00:00

Talend Open Studio

Mapping a Data Flow – Need your attention ☺

- Mapping components are advanced components which require more detailed explanation than other Talend components. The Map Editor is an "all-in-one" tool allowing you to define all parameters needed to map, transform and route your data flows via a convenient graphical interface.

- **tMap:** Allows to do data transformation, fields concatenation, field filtering using constraints, data rejection.

Talend Open Studio

Using Routines

- Routines are complex Java functions, generally used to factorize code. They therefore optimize data processing and improve Job capacities.
- **Two Types:**
  - **System Routines:**
    - Several system routines are provided. They are classed according to the type of data which they process: numerical, string, date...
  - **User Routines:**
    - These are routines which you have created or adapted from existing routines.

Talend Open Studio

Additional Info

- 1. See your code
- 2. Adding a note.
- 3. Activating/Deactivating tasks.
- 4. Importing/Exporting jobs
- 5. Documenting a job.

Talend Open Studio

How to study Talend

- **Suggested Plan:**
  - **1.** Read the definitions from the documentation. (Intro to Talend Documentation)
  - **2.** Watch these two courses on YouTube. (C1 – C2)
  - **3.** Apply anything you're learning and build a complete project.

Practice and Fun!

We're almost done

I wish you all enjoyed.

We’ll have a feedback form, and questions form now.

Next?: Understand the dataset, work on the project step by step with us, create folder with your **formal name** on the GitHub repo. to add your work. Follow Up.

Next Session we will start working on the project.

Thanks

GDSC

Data Engineering Workshop