

Name	Shubham Golwal
UID no.	2020300015
Experiment No.	08

AIM:	To study various ETL tools used in data warehouses.
-------------	---

Program 1

PROBLEM STATEMENT:	To study various ETL tools used in data warehouses between Talend Open Studio and Oracle Data Integrator.
---------------------------	---

Theory:	<p>Talend Open Studio:</p> <p>Talend Open Studio is an open-source data integration software that enables users to integrate, transform, and manage large volumes of data from various sources. It provides a graphical interface and drag-and-drop tools for building complex data integration workflows without requiring users to write code. Some of the key features of Talend Open Studio include:</p> <ol style="list-style-type: none"> 1. Data Integration: Talend Open Studio supports various data integration techniques, such as Extract-Transform-Load (ETL), Extract-Load-Transform (ELT), and Extract-Load-Update (ELU). 2. Data Mapping: Talend Open Studio provides a graphical interface and visual mapping tools for designing and managing data integration workflows. 3. Data Quality: Talend Open Studio offers data profiling, cleansing, and matching capabilities, which helps users ensure data accuracy and consistency. 4. Cloud Integration: Talend Open Studio supports cloud-based data integration, enabling users to manage and process data from various cloud-based services. 5. Real-time Data Streaming: Talend Open Studio supports real-time data streaming and processing, which helps users to manage and analyze large volumes of data in real-time. 6. Development Environment: Talend Open Studio provides an intuitive development environment that enables users to design, develop, and deploy data integration workflows.
----------------	---

Talend Open Studio supports a wide range of data sources, including databases, flat files, cloud services, and enterprise applications. It also offers a vast library of connectors and pre-built components, enabling users to integrate data quickly and easily from various sources.



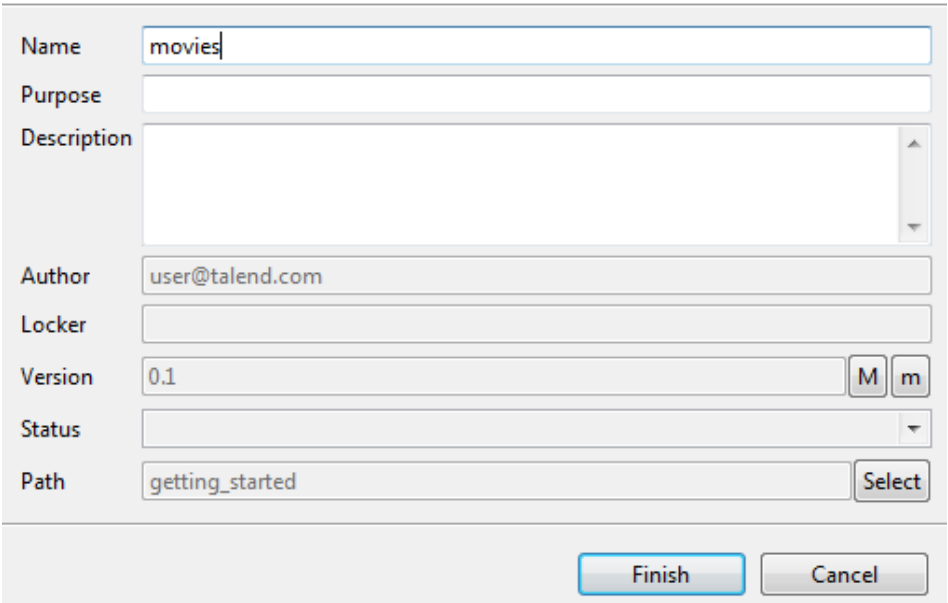
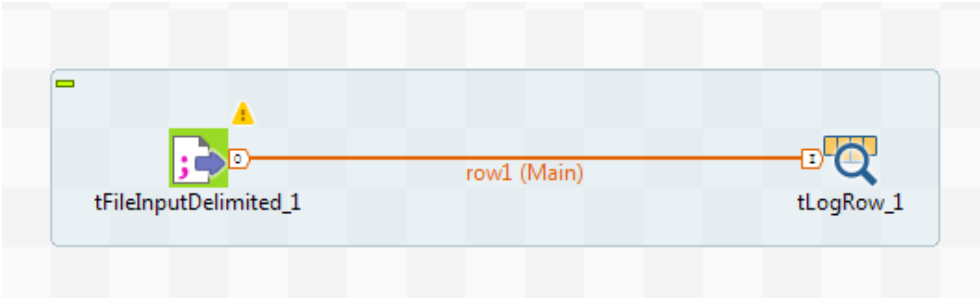
Overall, Talend Open Studio is a powerful and flexible data integration tool that can help organizations to streamline their data integration processes and improve data quality. It is suitable for organizations of all sizes and can be used for various use cases, including data warehousing, business intelligence, and data migration.

Oracle Data Integrator:

Oracle Data Integrator (ODI) is a comprehensive data integration platform that provides a powerful set of tools for managing and transforming data from various sources. It enables users to extract, load, and transform data from various sources, including databases, flat files, and enterprise applications.

Some of the key features of Oracle Data Integrator include:

1. **Data Integration:** Oracle Data Integrator offers a comprehensive set of data integration tools that enable users to extract, load, and transform data from various sources.
2. **Data Mapping:** Oracle Data Integrator provides a graphical interface and visual mapping tools that allow users to design and manage complex data integration workflows.
3. **Data Quality:** Oracle Data Integrator offers a comprehensive set of data quality tools that enable users to profile, cleanse, and match data from various sources.
4. **Cloud Integration:** Oracle Data Integrator provides cloud-based data integration capabilities, which enable users to manage and process data from various cloud-based services.
5. **Real-time Data Streaming:** Oracle Data Integrator supports real-time data streaming and processing, which allows users to manage and analyze large volumes of data in real-time.
6. **ETL Capabilities:** Oracle Data Integrator provides Extract-Transform-Load (ETL) capabilities, which enable users to extract data from various sources, transform it, and load it into target systems.
7. **Development Environment:** Oracle Data Integrator provides a comprehensive development environment that enables users to design, develop, and deploy data integration workflows.

	<p>Overall, Oracle Data Integrator is a powerful and comprehensive data integration platform that can help organizations to streamline their data integration processes and improve data quality. It is suitable for organizations of all sizes and can be used for various use cases, including data warehousing, business intelligence, and data migration.</p>
Output	<p><u>Talend Open Studio working:</u></p> <p>Reading movies information from a CSV file</p> <p>New Job</p> <p> It is inadvisable to leave the purpose blank.</p>   <p>Adding and linking components</p>  <p>Preparing the movies metadata</p>

New Delimited File

File - Step 1 of 4

Add a Metadata File on repository
Define the properties

Name	movies
Purpose	Centralize metadata of movies.csv
Description	Metadata of file movies.csv
Author	talend.com
Locker	
Version	0.1
Status	
Path	Select

< Back

Next >

Finish

Cancel

New Delimited File

File - Step 2 of 4

Add a Metadata File on repository
Define the path of the file and the format settings

File Settings	
Server	localhost 127.0.0.1
File	C:/getting_started/input_data/movies.csv
Format	WINDOWS
File Viewer	
movieID;title;releaseYear;url;directorID 315;Apt Pupil;1998;http://us.imdb.com/Title?Apt+Pupil+(1998);26 1294;Ayn Rand: A Sense of Life;1998;http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Life+(1997);12 1679;B. Monkey;1998;http://us.imdb.com/M/title-exact?B%2E+Monkey+(1998);124 1649;Big One, The;1998;http://us.imdb.com/Title?Big+One,+The+(1997);122 362;Blues Brothers 2000;1998;http://us.imdb.com/M/title-exact?Blues+Brothers+2000+(1998);86 1645;Butcher Boy, The;1998;http://us.imdb.com/M/title-exact?imdb-title-118804;134 1650;Butcher Boy, The;1998;http://us.imdb.com/M/title-exact?imdb-title-118804;134	

< Back

Next >

Finish

Cancel

New Delimited File

File - Step 3 of 4

Add a Metadata File on repository
Define the setting of the parse job

File Settings		Rows To Skip
Encoding	US-ASCII	If any rows must be ignored, specify the following parameters
Field Separator	Semicolon	Header <input checked="" type="checkbox"/> 1
Row Separator	Standard EOL	Footer <input type="checkbox"/>
Corresponding Character	"\"	<input type="checkbox"/> Skip empty row
Corresponding Character	"\n"	Limit Of Rows
Escape Char Settings		If the number of lines must be limited, specify this number
<input type="radio"/> CSV		Limit <input type="checkbox"/>
<input checked="" type="radio"/> Delimited		
Escape Char	Empty	
Text Enclosure	Empty	
<input type="checkbox"/> Split row before field		

Preview / Output

☒ Set heading row as column names [Refresh Preview](#)

movieID	title	releaseYear	url	directorID
315	Apt Pupil	1998	http://us.imdb.com/Title?Apt+Pupil+(1998)	26
1294	Ayn Rand: A Sense of Life	1998	http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Life+(1997)	123
1679	B. Monkey	1998	http://us.imdb.com/M/title-exact?B%2E+Monkey+(1998)	124
1648	Bis Opa, The	1998	http://us.imdb.com/Title?Bis+Opa,+The+(1997)	122

[Export as context](#) [Revert Context](#)

< Back Next > Finish Cancel

New Delimited File

File - Step 4 of 4

Add a Schema on repository
Define the Schema

Name	movies_schema
Comment	

Schema

Click to update schema preview

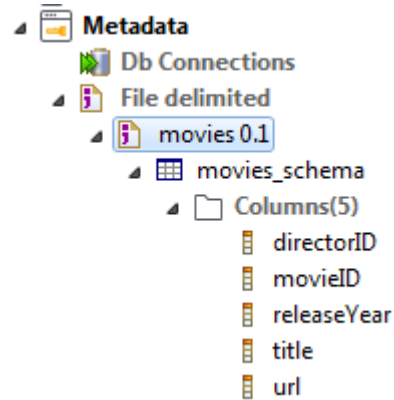
[Guess](#)

Description of the Schema

Column	Key	Type	<input checked="" type="checkbox"/> N...	Date Pattern (Ctrl+S...	Length	Precision	Default	Comment
movieID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
title	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		72	0		
releaseYear	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
url	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		120	0		
directorID	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		3	0		



< Back Next > Finish Cancel



Configuring and executing your Job

The job designer shows a flow from **tFileInputDelimited_1** to **tLogRow_1** via **row1 (Main)**.

Designer | **Code**

Job(movies 0.1) | Contexts(movies) | Component | Run (Job movies)

tFileInputDelimited_1

Basic settings

Property Type: Repository | DELIM:file_movies

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File name/Stream: "C:/getting_started/input_data/movies.csv" *

Row Separator: "\n" | Field Separator: "," *

☐ CSV options

Header: 1 | Footer: 0 | Limit: *

Schema: Repository | DELIM:file_movies - metadata *

☐ Skip empty rows | ☐ Uncompress as zip file | ☐ Die on error

Job(movies 0.1)
Contexts(movies)
Component
Run (Job movies)

tLogRow_1

Basic settings
Advanced settings
Dynamic settings
View
Documentation

Schema
Built-In
Edit schema
Sync columns

Mode
☐ Basic
☐ Table (print values in cells of a table)
☒ Vertical (each row is a key/value list)

Title printing mode
☐ Print unique name
☒ Print label
☐ Print unique name and label

☒ Print content with log4j

Job(movies 0.1)
Contexts(movies)
Component
Run (Job movies)

Job movies

Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run

Execution
Run
Kill
Clear

```

[statistics] connected
-----
#1. tLogRow_1
-----
key      value
-----
movieID  315
title    Apt Pupil
releaseYear  1998
url      http://us.imdb.com/Title?Apt+Pupil+(1998)
directorID  26
-----

#2. tLogRow_1
-----
key      value
-----
movieID  1294
title    Ayn Rand: A Sense of Life
releaseYear  1998
url      http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Life+(1997)
directorID  123
-----

```

Filtering the movies information

Preparing directors file metadata

File - Step 1 of 4

Add a Metadata File on repository
Define the properties



Name	<input type="text" value="directors"/>		
Purpose	<input type="text" value="Centralize the metadata of directors info"/>		
Description	<input type="text" value="Metadata of the directors dataset"/>		
Author	<input type="text" value="user@talend.com"/>		
Locker	<input type="text"/>		
Version	<input type="text" value="0.1"/>	<input type="button" value="M"/>	<input type="button" value="m"/>
Status	<input type="text"/>		
Path	<input type="text"/>		<input type="button" value="Select"/>

< Back

Next >

Finish

Cancel

File - Step 2 of 4



Add a Metadata File on repository
Define the path of the file and the format settings

File Settings

Server

File

Format

File Viewer

- 1, Gregg Araki
- 2, P.J. Hogan
- 3, Alan Rudolph
- 4, Alex Proyas
- 5, Alex Sichel
- 6, Alex Zamm
- 7, Alfonso Cuarón
- 8, Alfred Hitchcock
- 9, Allison Anders
- 10, Andrew Davis
- 11, Andrew Niccol
- 12, Antoine Fuqua
- 13, Barbet Schroeder
- 14, Barbet Schroeder
- 15, Barry Levinson
- 16, Barry Sonnenfeld

When Filepath is specifi

Finish

Cancel

File - Step 3 of 4Add a Metadata File on repository
Define the setting of the parse job

File Settings		Rows To Skip	
Encoding	UTF-8	If any rows must be ignored, specify the following parameters	
Field Separator	Comma	Header	<input type="checkbox"/>
Row Separator	Standard EOL	Footer	<input type="checkbox"/>
	Corresponding Character	<input type="checkbox"/> Skip empty row	
Escape Char Settings		Limit Of Rows	
<input type="radio"/> CSV		If the number of lines must be limited, specify this number	
<input checked="" type="radio"/> Delimited		Limit	
Escape Char	Empty		
Text Enclosure	Empty		
<input type="checkbox"/> Split row before field			

Preview Output

☐ Set heading row as column names Refresh Preview

Column 0
1, Gregg Araki
2, P.J. Hogan
3, Alan Rudolph
4, Alex Proyas
5, Alex Sichel
6, Alex Zamm

Export as context

Revert Context

< Back

Next >

Finish

Cancel

File - Step 4 of 4Add a Schema on repository
Define the SchemaName directors_schema
Comment

Schema

Click to update schema preview

Guess

Description of the Schema

Column	Key	Type	<input checked="" type="checkbox"/> N..	Date Pattern (Ctrl+Sp...	Length	Precision	Default	Comment
directorID	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
directorName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		20	0		

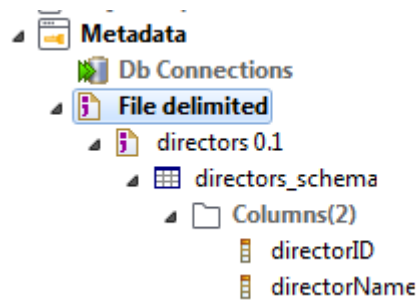


< Back

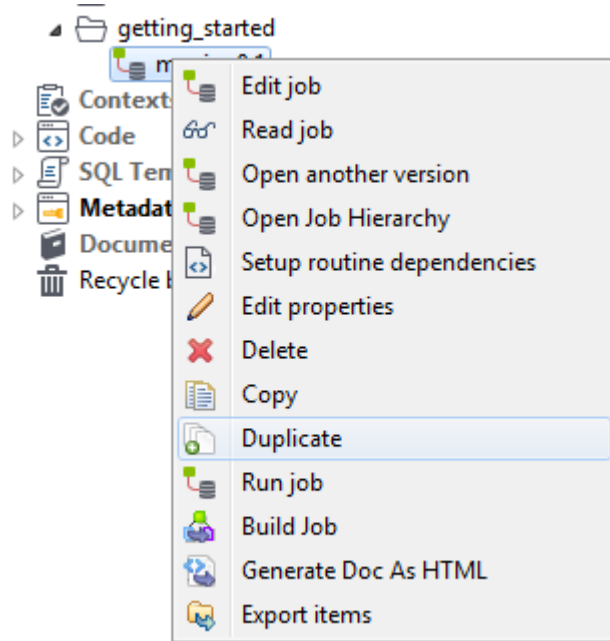
Next >

Finish

Cancel



Duplicating the existing Job



Input new name

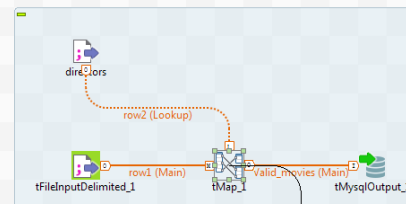
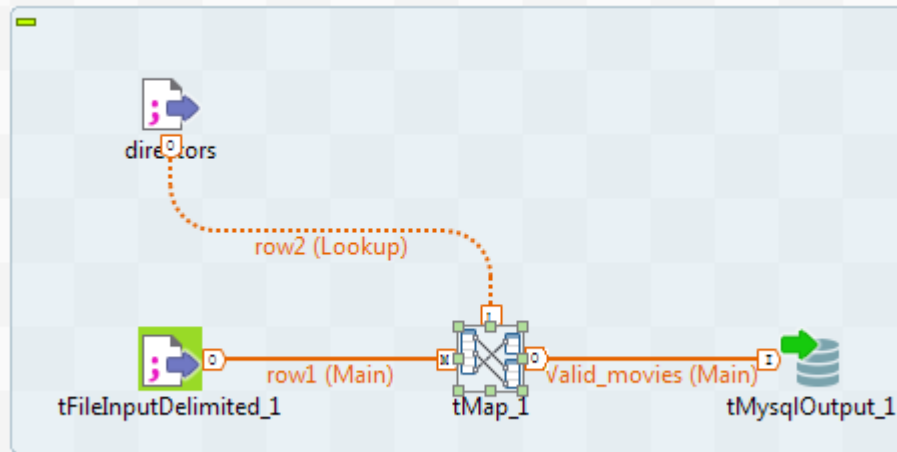
OK Cancel

Gathering rejected movies information and saving processing results to a database

tmysqlou

- tELTMysqlOutput - ELT/Map/MySQL
- tMysqlOutputBulk - Databases/MySQL
- tMysqlOutput - Databases/MySQL
- tMysqlOutputBulkExec - Databases/MySQL
- Note - Misc

Inserts or updates lines into a MySQL table

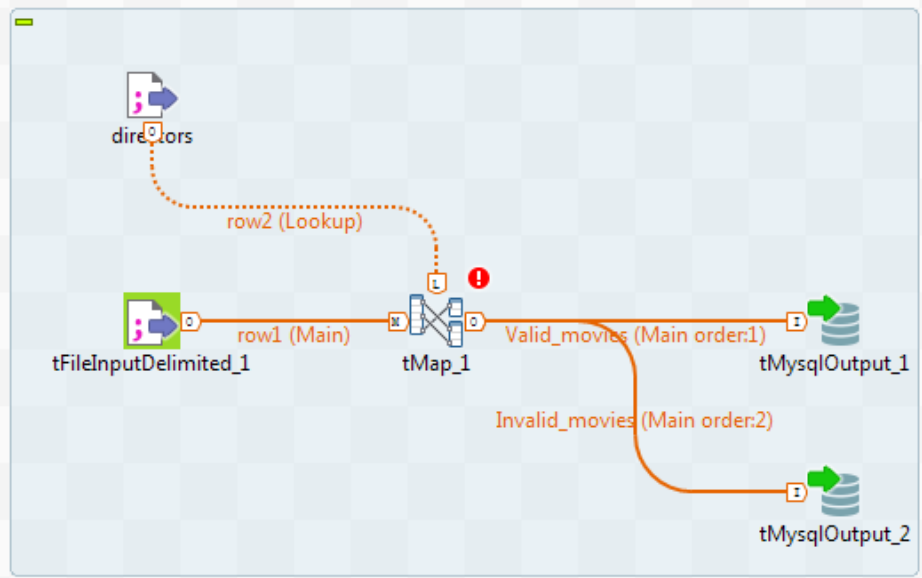


es_to_db 0.1

s_to_db 0.1

- tAmazonOracleCommit - Cloud/AmazonRDS/Oracle
- tHSQLdbOutput - Databases/HSQLdb
- tGoogleDrivePut - Cloud/Google Drive
- tMSSqlCommit - Databases/MS_SQL_Server
- tPostgresPlusSCDELT - Databases/PostgresPlusBusir
- tLDAPRenameEntry - Databases/LDAP
- tMSSqlOutputBulk - Databases/MS_SQL_Server
- tSocketOutput - Internet
- tAccessOutputBulk - Databases/Access
- tConvertType - Processing

Commits all transactions not already committed in a selected Amazon RDS Oracle connection



Feature	Talend Open Studio	Oracle Data Integrator (ODI)
Cost	Free and open source	Proprietary software with paid licenses
Data integration	Supports various data integration techniques and tools	Provides comprehensive data integration and transformation
Data mapping	Provides visual mapping tools and graphical user interface	Offers comprehensive mapping capabilities and graphical tools
Data quality	Supports data profiling, cleansing, and matching capabilities	Provides comprehensive data quality features and functions
Real-time data streaming	Supports real-time data streaming and processing	Provides real-time data integration and transformation
Cloud integration	Offers cloud-based data integration capabilities	Provides cloud-based data integration and migration
Development	Provides drag-and-drop interface and intuitive development	Offers a comprehensive development environment
Integration with ETL	Provides Extract-Transform-Load (ETL) capabilities	Offers ETL capabilities
Scalability	Highly scalable and supports large-scale data integration	Offers highly scalable and flexible integration capabilities

Conclusion:

To create a data warehouse, a lot of data from various sources is required. This data available at different sources may not fit our design. Hence, ETL tools come to rescue from this problem. ETL tools

extract data from various sources and transform them into required format making our work easier and faster. They also load the formatted data into data warehouse.

References:

<https://www.talend.com/products/talend-open-studio/>

<https://www.dremio.com/resources/guides/adv-types-etl-tools/>