

# TALEND DATA INTEGRATION





# Agenda

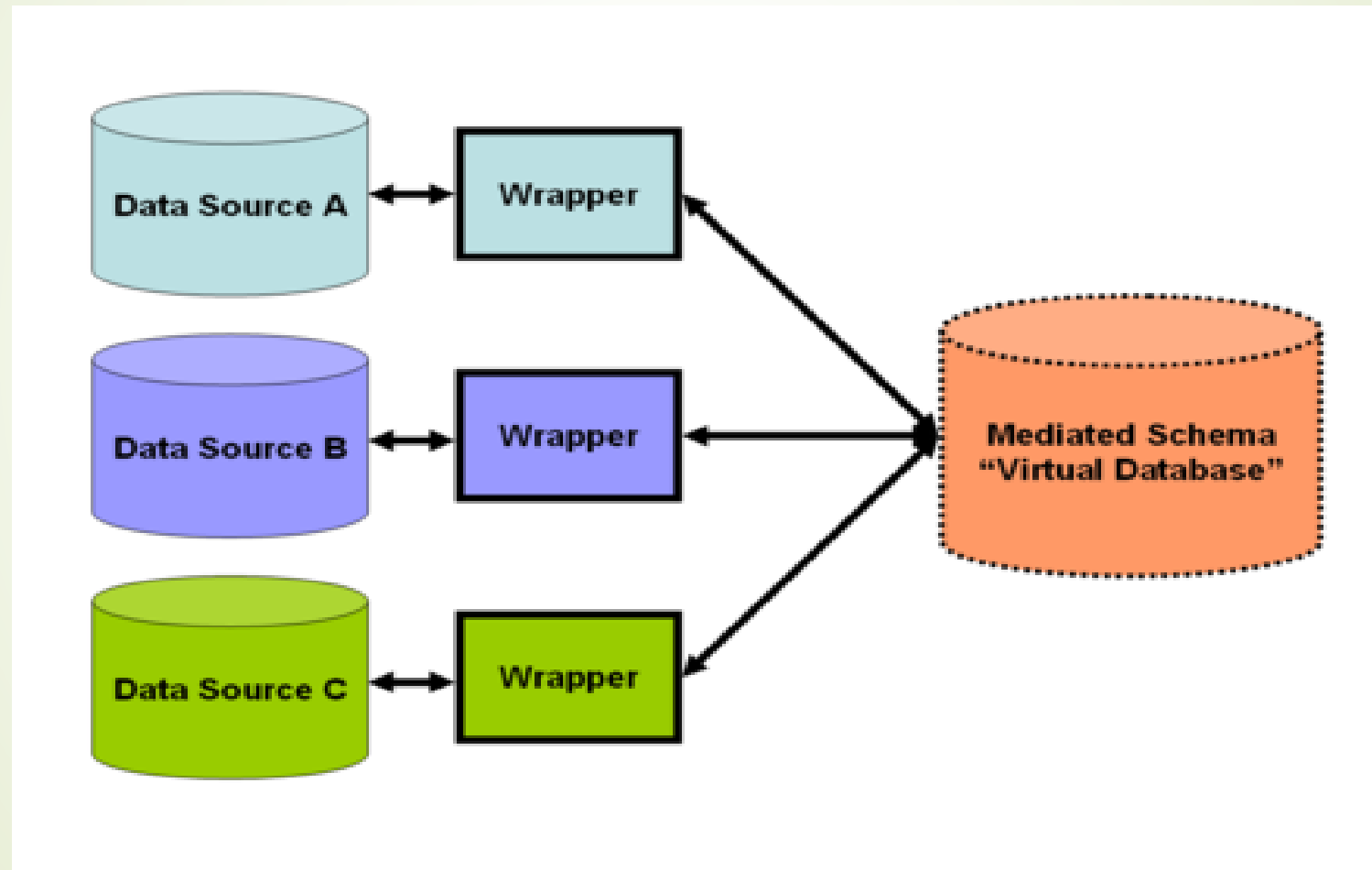
- ❑ Introduction à Talend Data integration
- ❑ Prise en main de Talend Data integration
- ❑ Utilisation avancée de Talend Data intégration



# Objectif du cours

- ❑ Introduction à l'intégration des données
- ❑ La plateforme Talend
- ❑ Architecture du Talend Open Studio For Data integration
- ❑ Installation du Talend Open Studio For Data integration
- ❑ Lancer le premier Job Talend Data Integration

# Intégration des données

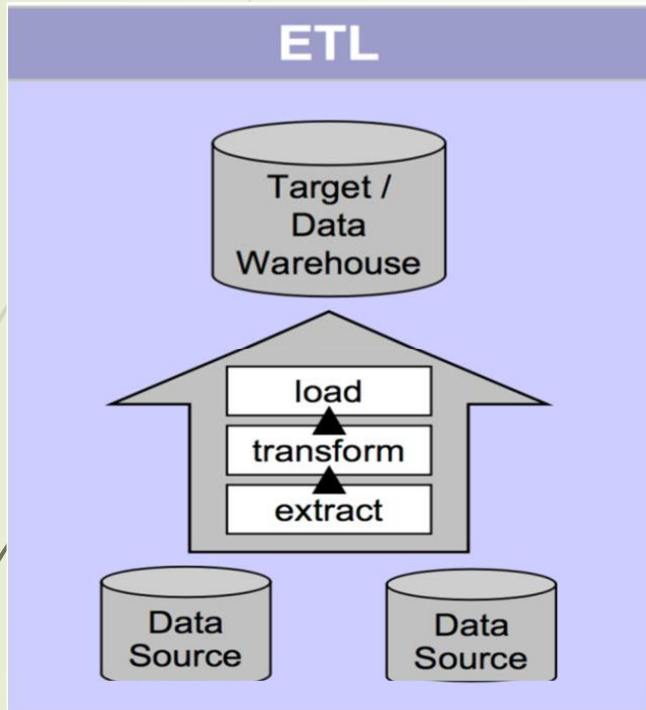




# Intégration des données

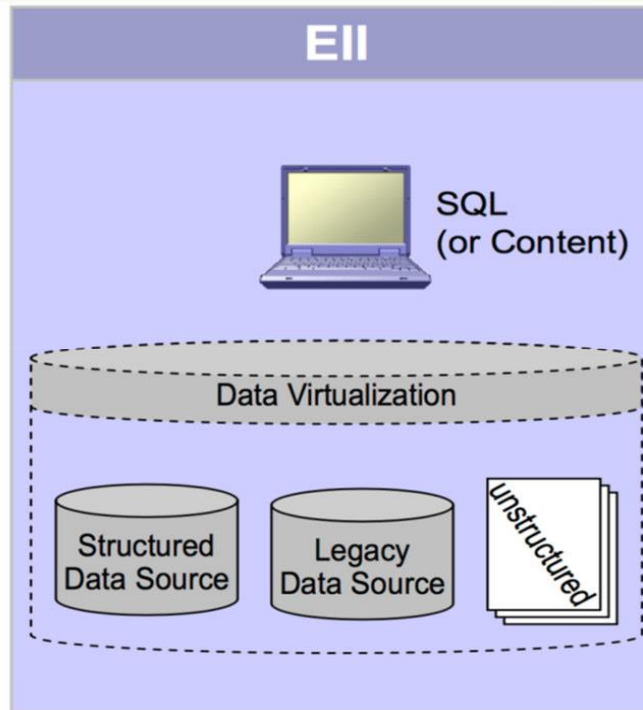
- ✓ Pourquoi l'intégration des données ?
- ✓ Sources diverses et différentes
- ✓ Sources sur différentes plateformes et OS
- ✓ Applications legacy utilisant des BD et autres technologies obsolètes
- ✓ Historique de changement non-préservé dans les sources
- ✓ Qualité de données douteuse et changeante dans le temps
- ✓ Structure des systèmes sources changeante dans le temps
- ✓ Incohérence entre les différentes sources
- ✓ Données dans un format difficilement interprétable ou ambigu.

# Intégration des données



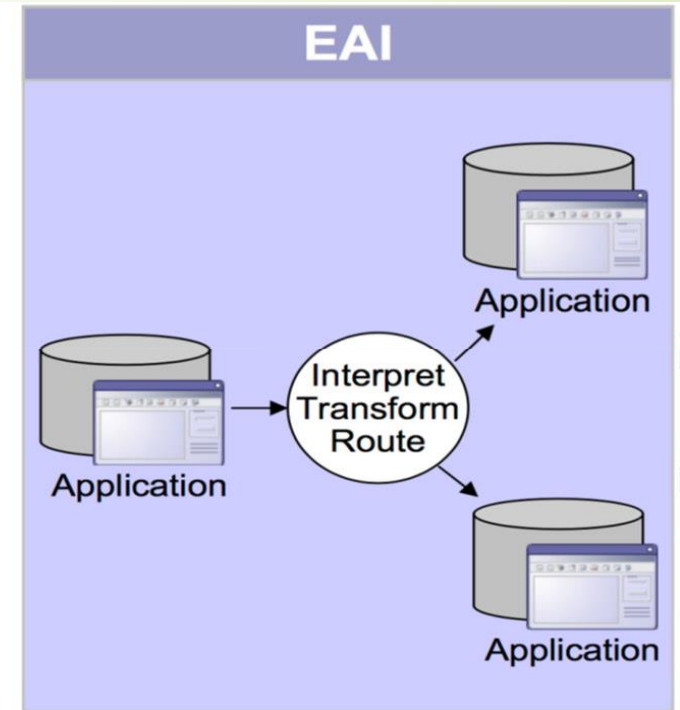
## Extract, Transform and Load

- Intégration et livraison des données en lot
- Transformations appliquées sur les données



## Enterprise Information Intergration

- Fédération de données provenant de plusieurs sources
- Accès temps-réel aux données
- Données structurées ou semi-structurées



## Enterprise Application Intergration

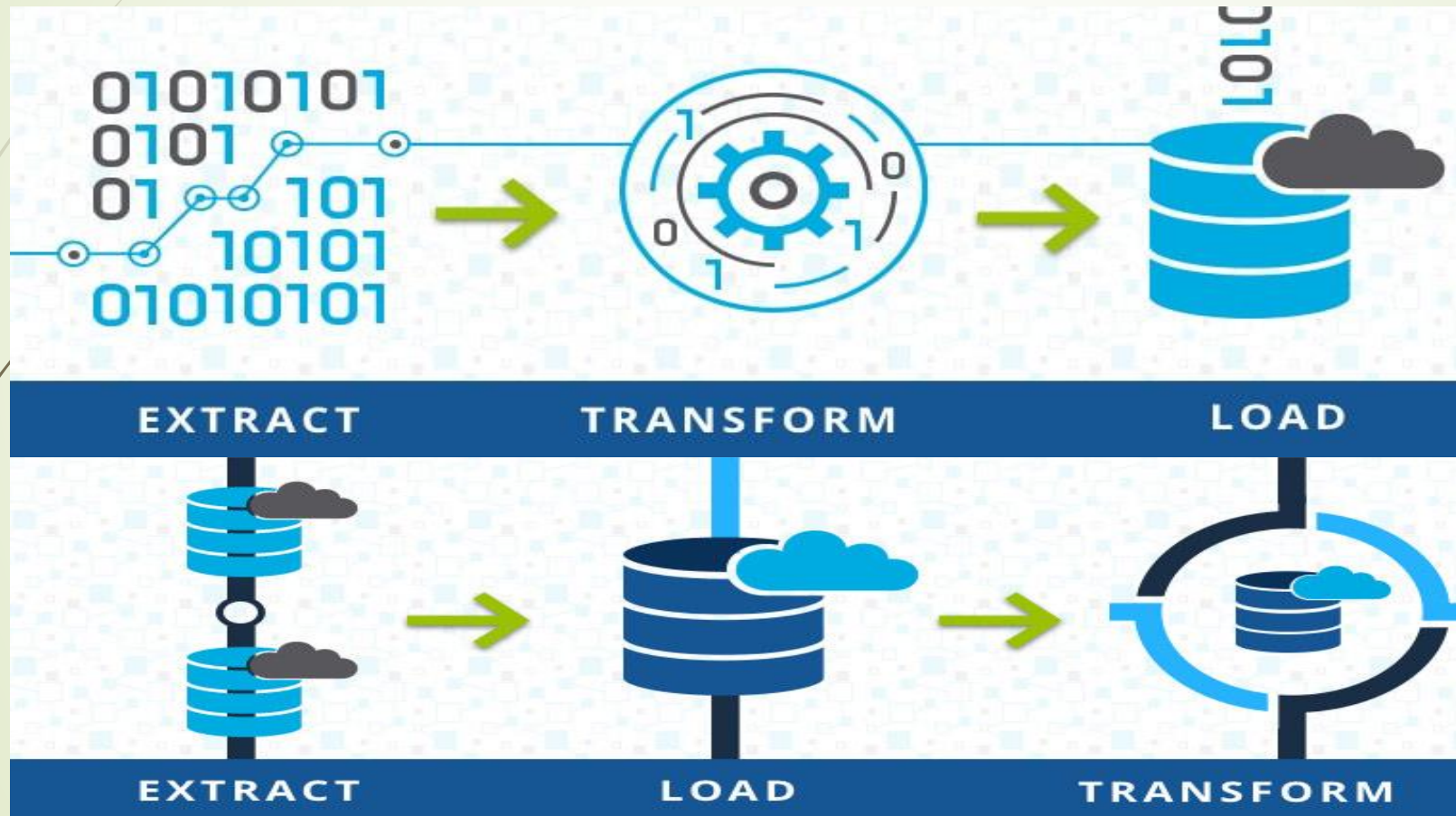
- Processus d'intégration des données d'applications
- Basé sur l'échange de messages sur un bus commun

# Intégration des données

	ETL/ELT	EII	EAI
Flot des données	Unidirectionnelle	bidirectionnelle	bidirectionnelle
Mouvement des données	Lots planifiés	Au moment de la requête	Déclenché par la transaction
Fréquence	Journalier à Mensuel	Temps-réel	Quasi temps réel
Transformation/ Agrégation des données	Grande capacité	Moyenne capacité	Faible capacité
Volumes des donnés	Grand	Moyen	Faible



# ETL / ELT





# ETL / ELT

Comparison Parameters	ETL	ELT
Ease of adoption to the tool	ETL is a well-developed process used for over 20 years, and ETL experts are easily available.	ELT is a new technology, so it can be difficult to find experts and develop an ELT pipeline
Data size	ETL is better suited for dealing with smaller data sets that require complex transformations.	ELT is better suited when dealing with massive amounts of structured and unstructured data.
Order of the process	Data transformations happens after extraction in the staging area. After transformation, the data is loaded into the destination system.	Data is extracted, loaded into the target system, and then transformed.
Transformation process	The staging area is located on the ETL solution's server.	The staging area is located on the source or target database.
Load time	ETL load times are longer than ELT because it's a multi-stage process: (1) data loads into the staging area, (2) transformations take place, (3) data loads into the data warehouse.	Data loading happens faster because there's no waiting for transformations and the data only loads one time into the target data system.



# Les étapes d'un système ETL/ELT

- ✓ Déterminer les données nécessaires à l'application Métier (BI, Data science ...)
- ✓ Déterminer les sources internes et externes contenant ces données
- ✓ Définir les règles d'extraction des données cibles
- ✓ Définir les règles de transformation et de nettoyage des données
- ✓ Planifier les agrégations des données
- ✓ Charger les données dans le modèle de l'application cible



# Extraction des données

- ✓ Identifier les sources de données et leurs structures
- ✓ Décider, pour chaque source, quel outil pourra être le meilleur à l'interroger (Sqoop, Talend,, script ...)
- ✓ Choisir, pour chaque source, la fenêtre temporelle durant laquelle sera faite l'extraction
- ✓ Déterminer le plan d'ordonnancement des tâches d'extraction
- ✓ Déterminer comment gérer les exceptions



# Extraction des données

## ❑ Extraction complète :

- ✓ Capture l'ensemble des données à un certain instant (snapshot de l'état opérationnel)
- ✓ Utilisée dans deux cas :
  - Chargement initial des données
  - Rafraichissement complet des données
- ✓ Peut être très couteuse en temps d'exécution (plusieurs heures/jours)



# Extraction des données

- ✓ Extraction incrémentale :
- ✓ Capture uniquement de données qui ont été changées ou ont été ajoutées depuis la dernière extraction
- ✓ Peut être faite de deux façons :
  - Extraction temps réel
  - Extraction différée (bach)



# Transformation des données

Révision de format :

Ex: Changer le type ou la longueur de champs individuels

Décodage de champs:

Consolider les données de sources multiples

- Ex: ['homme', 'femme'] vs ['M', 'F'] vs [1,2]

Traduire les valeurs cryptiques

- Ex: 'AC', 'IN', 'SU' pour les statuts actif, inactif et suspendu.





# Transformation des données

Pré-calcul des valeurs dérivées:

Ex: profit calculé à partir de ventes et coûts.

Découpage de champs complexes:

Ex: extraire les valeurs prénom, secondPrénom et nomFamille à partir d'une seule chaine de caractères nomComplet



# Transformation des données

Fusion de plusieurs champs:

Ex: information d'un produit

- ✓ Source 1: code et description;
- ✓ Source 2: types de forfaits;
- ✓ Source 3: coût.

Conversion de jeu de caractères

Conversion de dates

Pré-calcul des agrégations



# Chargement des données

## ❑ Chargement initial:

- ✓ Fait une seule fois lors de l'activation de l'entrepôt de données
- ✓ Les indexes et contraintes d'intégrité référentielle (clé étrangères) sont normalement désactivés temporairement
- ✓ Peut prendre plusieurs heures.

# Chargement des données

## ☐ Chargement incrémental:

- ✓ Se fait une fois le chargement initial complété
- ✓ Tient compte de la nature des changements
- ✓ Peut être fait en temps-réel ou en lot.

## ☐ Rafraîchissement complet:

- ✓ Employé lorsque la volumétrie rend le chargement incrémental trop complexe
- ✓ Ex: lorsque plus de 20% des enregistrements ont changé depuis le dernier chargement.



Talend



talend

# Les produits Talend





# Les produits Talend





# Talend Data integration

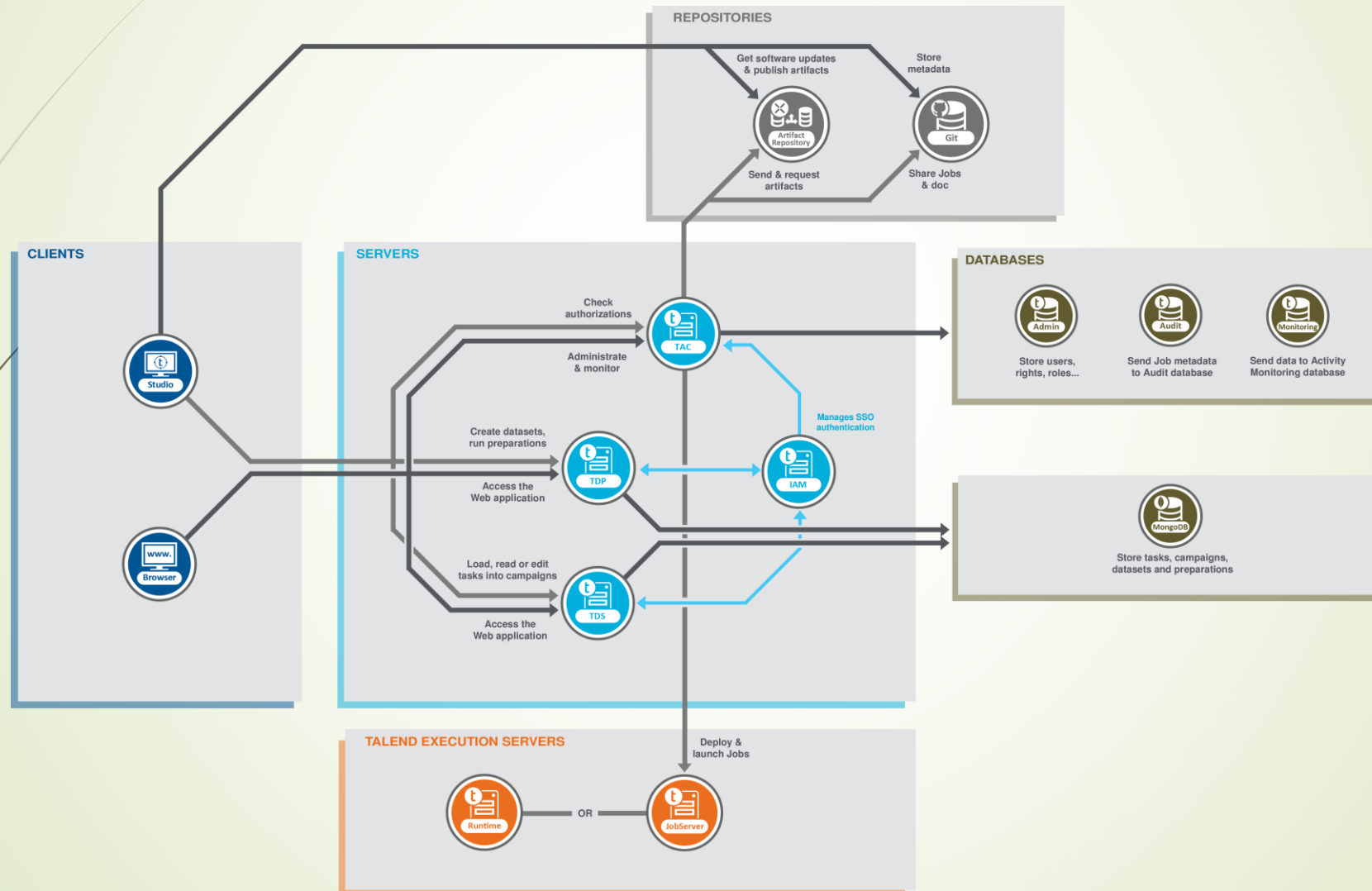
- ✓ Talend Data a été conçu pour simplifier le développement, l'intégration et la gestion des flux data
- ✓ Talend Data integration élimine la nécessité pour les utilisateurs d'affronter la complexité liée au développement et à la maintenance de code Java.
- ✓ Talend génère le code natif et optimisé pour charger, transformer, enrichir et nettoyer les données à l'intérieur sans stockage supplémentaire ou de frais lié au calcul

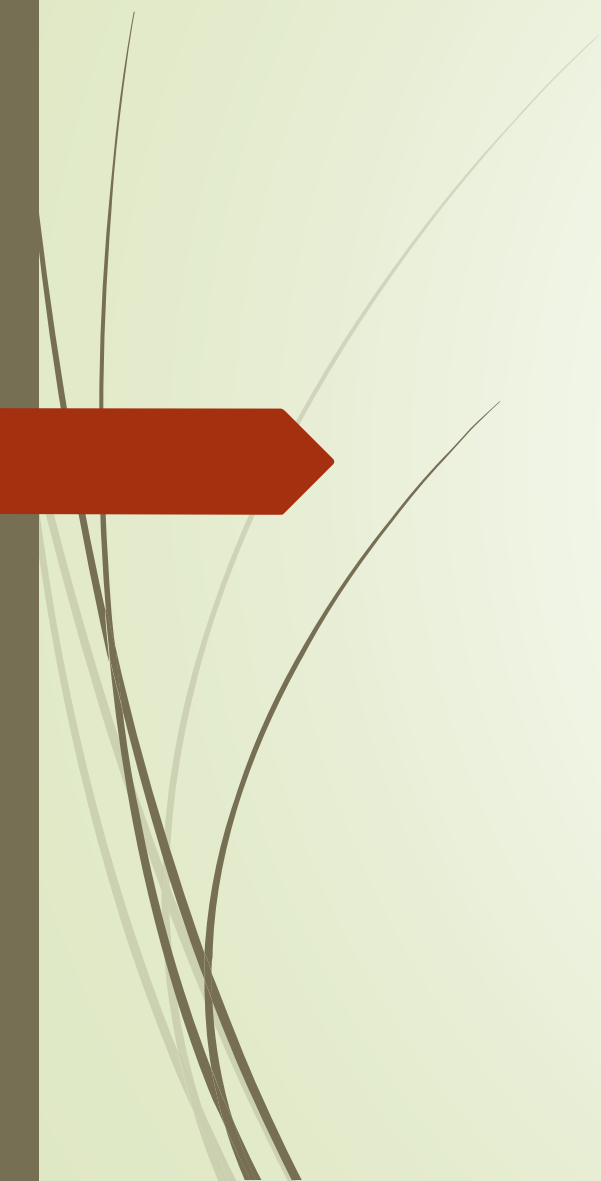


# Talend Data integration

- ✓ En plus des produits payants, Talend offre aux développeurs des produits Open Source
- ✓ Talend Open Studio For Big Data est le produit gratuit de Talend pour développer des applications Big Data
- ✓ TOS For Data Integration est l'outil gratuit pour l'intégration des données dans des environnements Data

# Architecture Talend for data integration





LAB

# Lab 0 : Installation du TOS For Data integration

Télécharger le TOS For Data Integration version 7.3.1 à partir du lien :

<https://sourceforge.net/projects/talend-studio/files/Talend%20Open%20Studio/7.3.1/>

**Installer Java 8 :**

<https://www.oracle.com/technetwork/java/javase/downloads/index.html>



# Lab 0 : Installation du TOS For Data integration

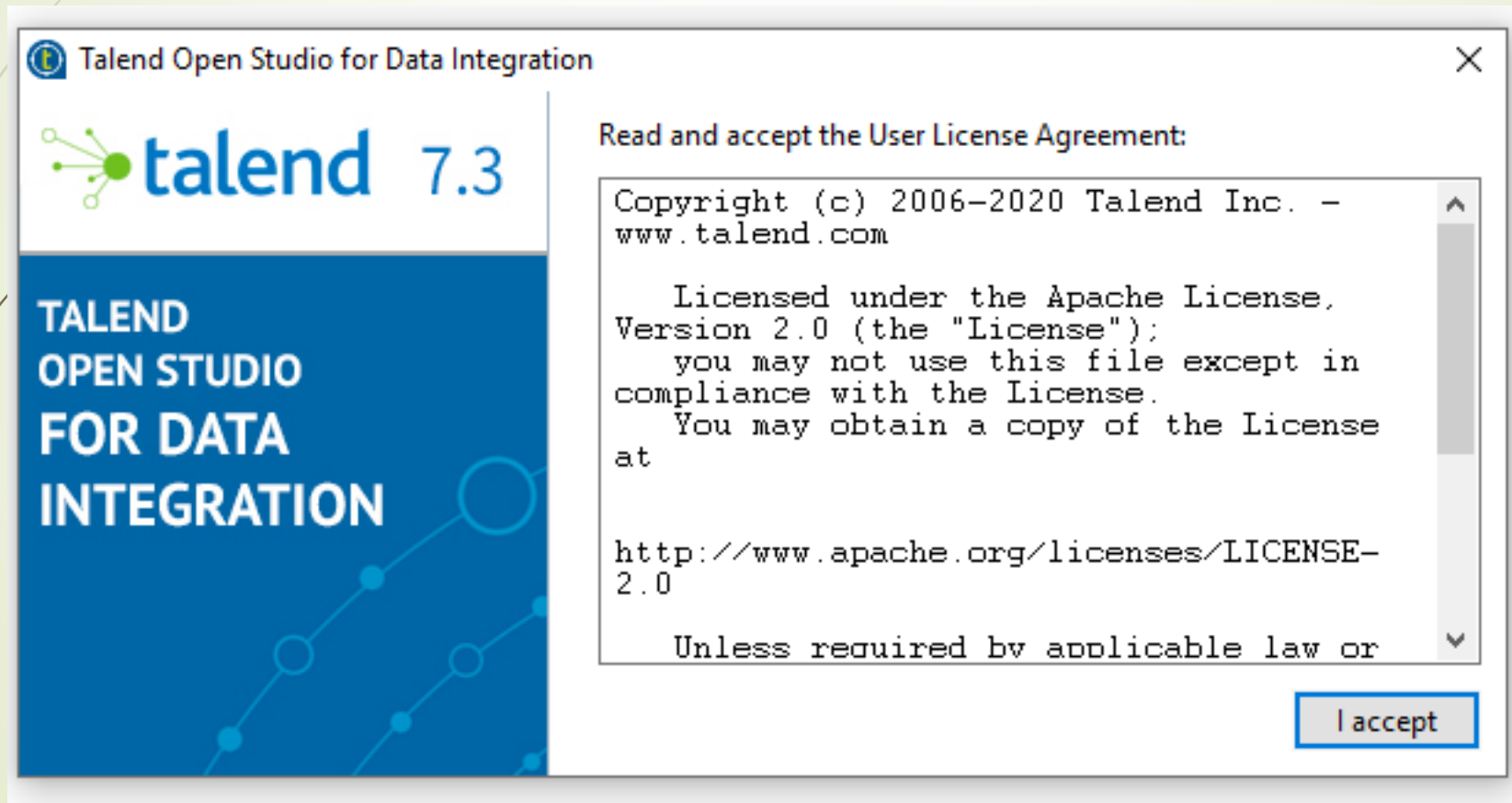
Lancer le TOS For Data Integration en exécutant TOS\_DI-win-x86\_64.exe

Ce PC > OS (C:) > Talend > TOS\_DI-Win32-20200219\_1130-V7.3.1

Nom	Modifié le	Type	Taille
about_files	23/01/2021 19:35	Dossier de fichiers	
configuration	14/07/2022 23:56	Dossier de fichiers	
features	23/01/2021 19:35	Dossier de fichiers	
p2	23/01/2021 19:35	Dossier de fichiers	
plugins	23/01/2021 19:39	Dossier de fichiers	
temp	23/01/2021 19:41	Dossier de fichiers	
TOS_DI-macosx-cocoa.app	23/01/2021 19:35	Dossier de fichiers	
workspace	27/01/2021 20:08	Dossier de fichiers	
.eclipseproduct	19/02/2020 14:18	Fichier ECLIPSEPR...	1 Ko
license.txt	19/02/2020 14:18	Document texte	1 Ko
NOTICE.txt	19/02/2020 14:18	Document texte	14 Ko
TOS_DI-linux-gtk-x86.sh	19/02/2020 14:18	Shell Script	1 Ko
TOS_DI-linux-gtk-x86_64	20/08/2014 08:34	Fichier	73 Ko
TOS_DI-linux-gtk-x86_64.ini	19/02/2020 14:18	Paramètres de con...	1 Ko
TOS_DI-macosx-cocoa.ini	19/02/2020 14:18	Paramètres de con...	1 Ko
TOS_DI-win-x86_64.exe	05/05/2015 06:41	Application	305 Ko
TOS_DI-win-x86_64.ini	19/02/2020 14:18	Paramètres de con...	1 Ko
Uninstall-TOS_DI-Win32-20200219_1130-...	23/01/2021 19:39	Application	57 Ko

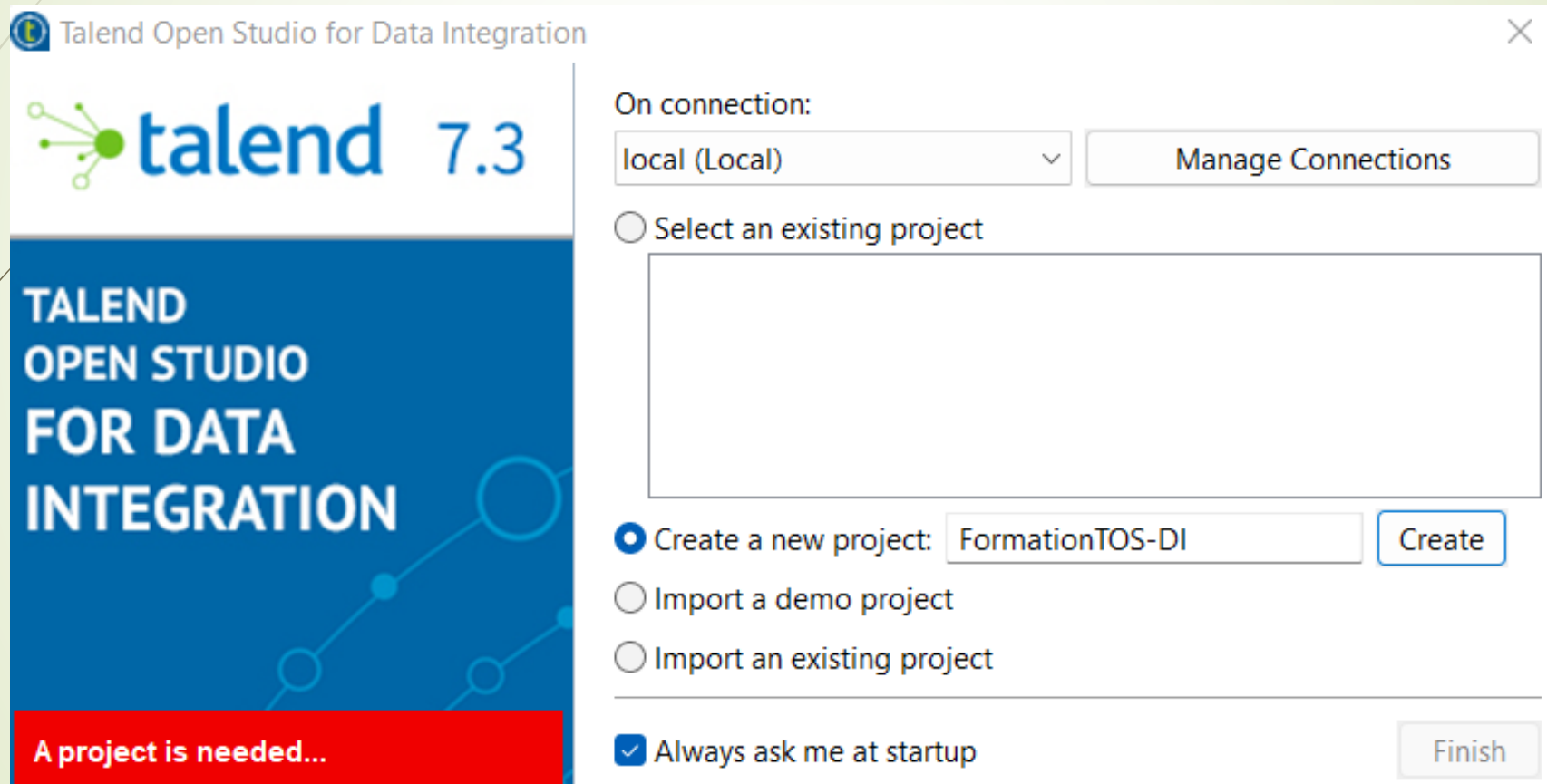
# Lab 0 : Installation du TOS For Data integration

Accepter la licence



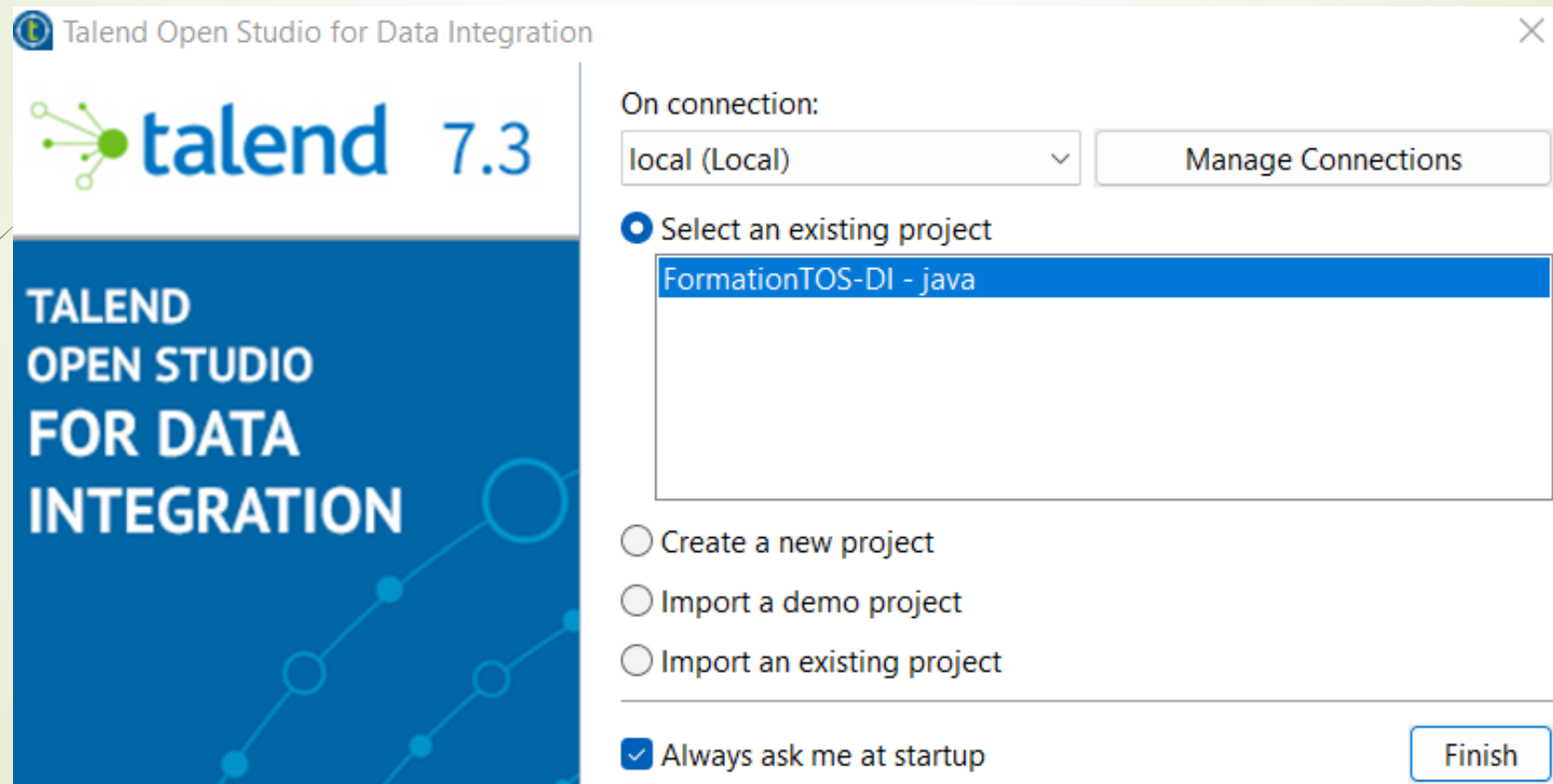
# Lab 0 : Installation du TOS For Data integration

Créer un nouveau

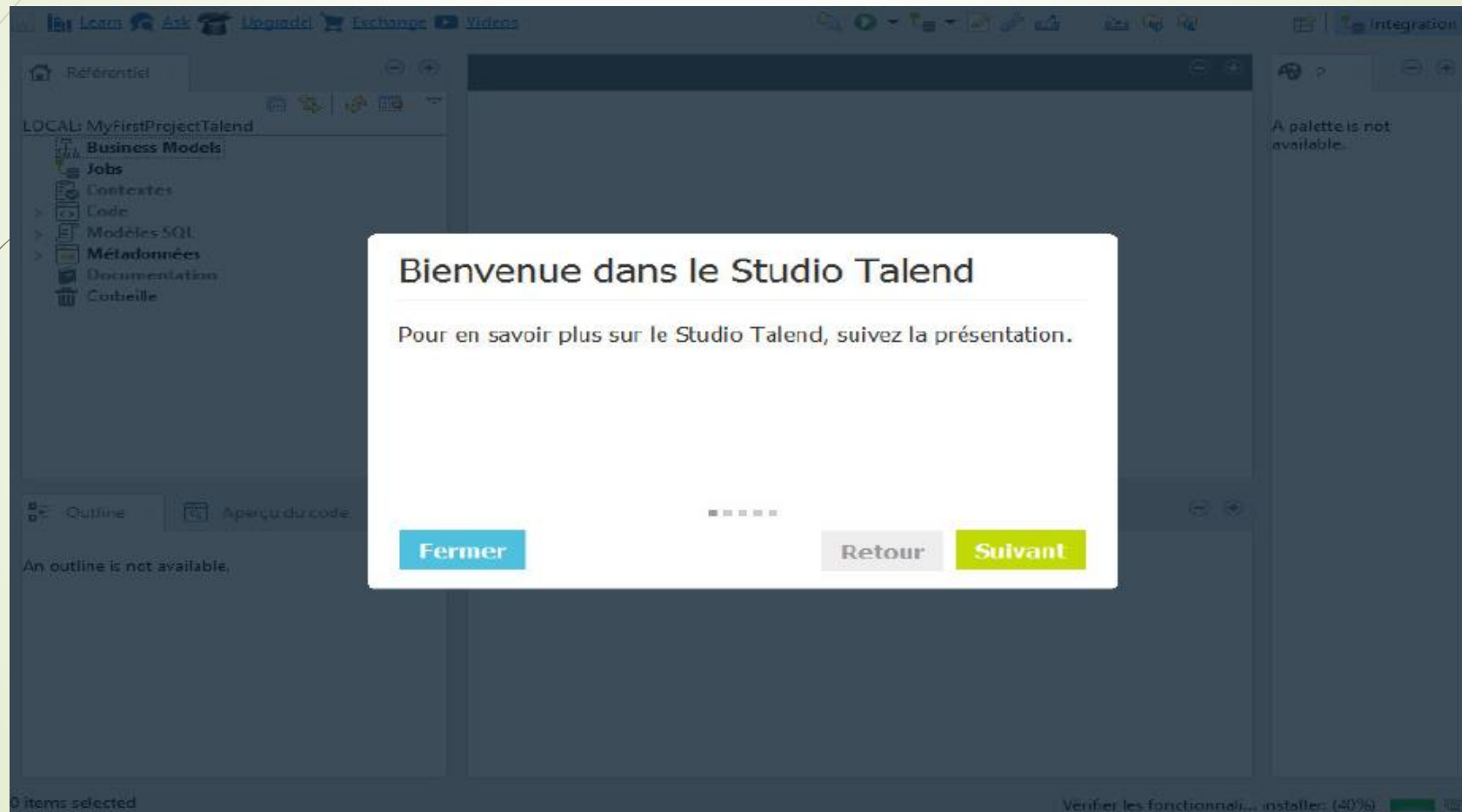


# Lab 0 : Installation du TOS For Data integration

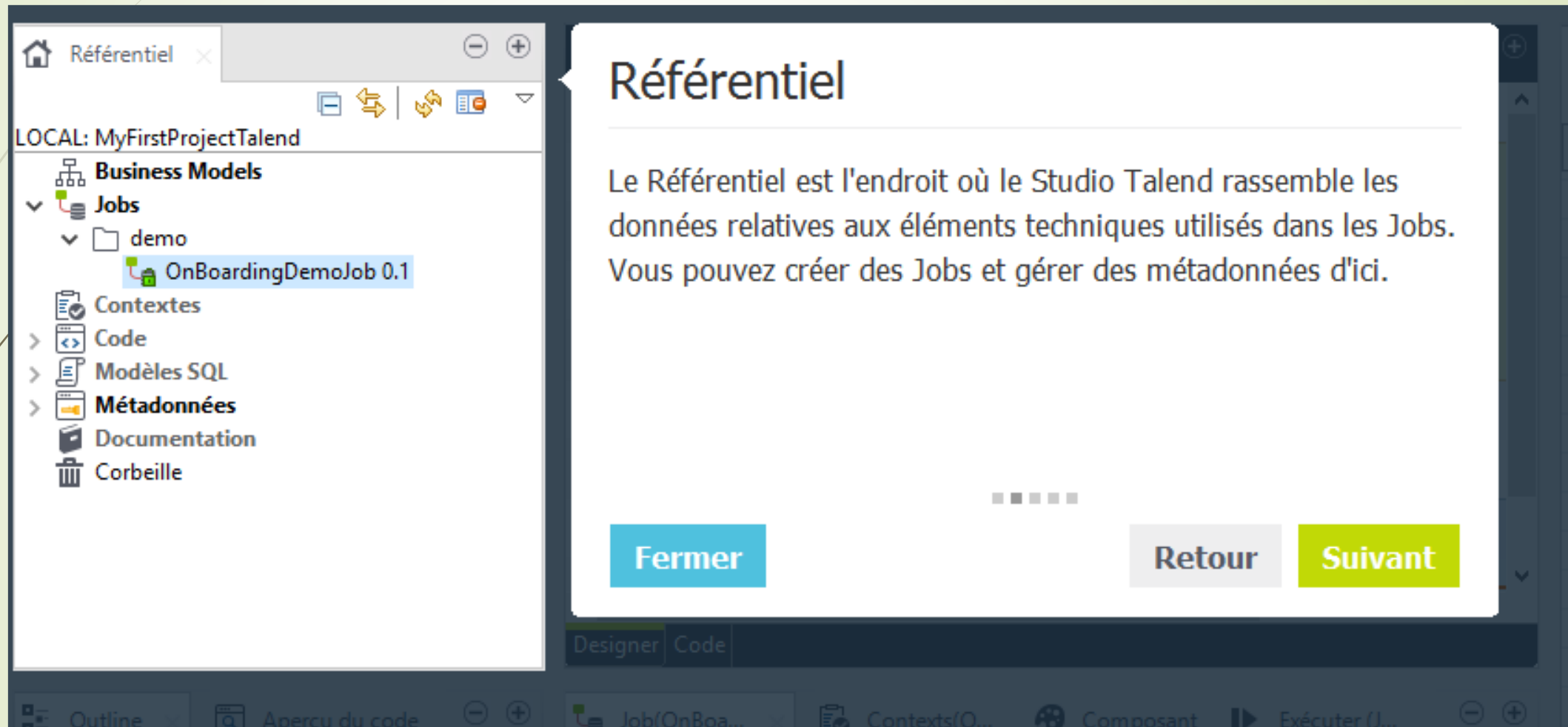
Choisir le répertoire local du projet



# Lab 0 : Installation du TOS For Data integration

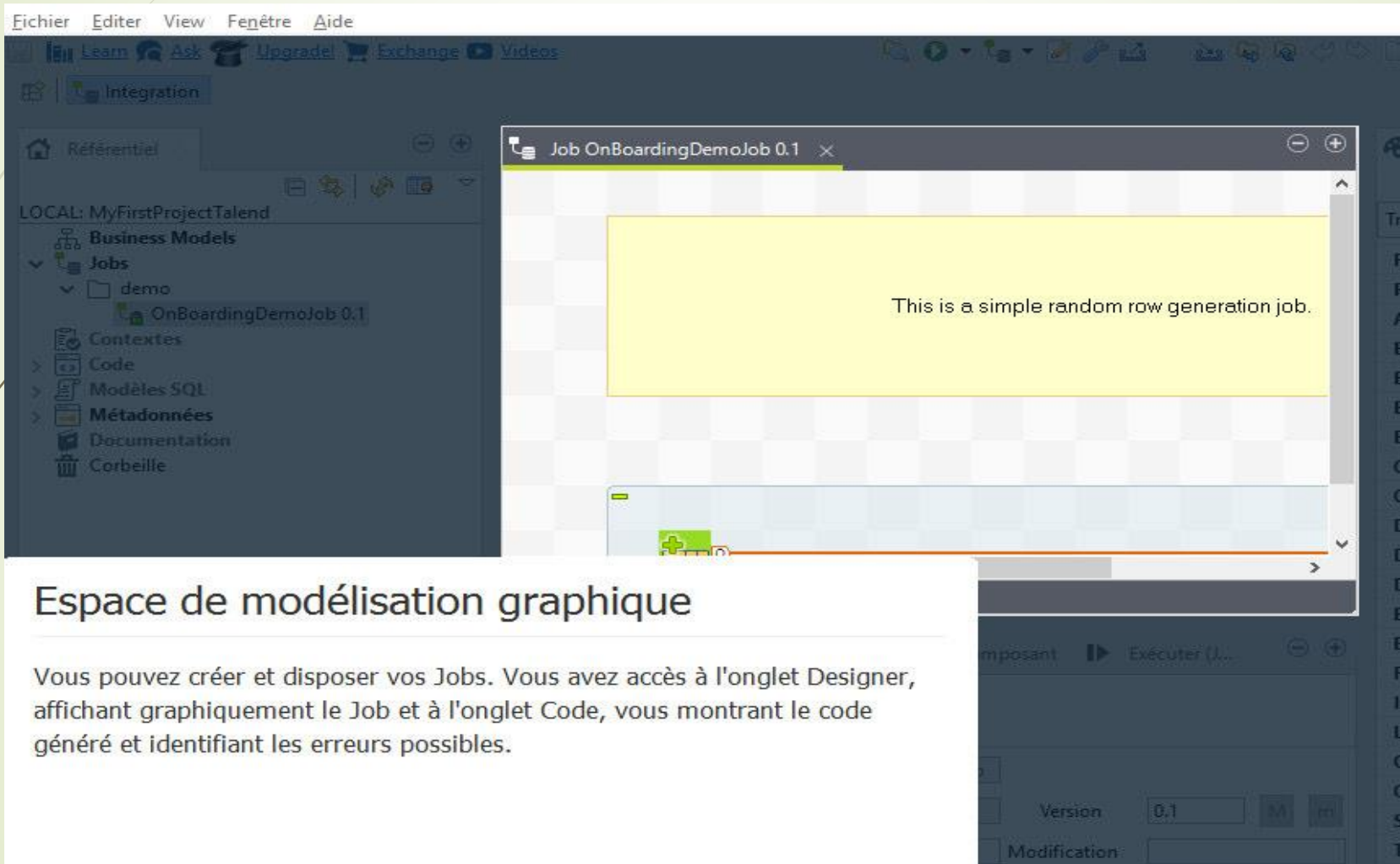


# Lab 0 : Installation du TOS For Data integration





# Lab 0 : Installation du TOS For Data integration



**Espace de modélisation graphique**

Vous pouvez créer et disposer vos Jobs. Vous avez accès à l'onglet Designer, affichant graphiquement le Job et à l'onglet Code, vous montrant le code généré et identifiant les erreurs possibles.

# Lab 0 : Installation du TOS For Data integration

## Onglet de configuration

Chaque onglet ouvre une vue affichant les propriétés de l'élément sélectionné dans l'espace de modélisation graphique. Ces propriétés peuvent être modifiées pour configurer des paramètres relatifs à un composant particulier ou au Job entier.

L'onglet Exécuter vous permet d'exécuter votre Job. [Sélectionnez cet onglet](#) et cliquez sur le bouton Exécuter pour essayer.

Fermer

Retour

Suivant

The screenshot shows the Talend Studio interface. On the left, the 'Outline' pane lists 'tLogRow\_1' and 'tRowGenerator\_1'. The 'Aperçu du code' pane is also visible. The main workspace displays a 'simple random row generation job.' diagram. Overlaid on the workspace is a configuration window titled 'OnBoardingDemoJob 0.1'. This window has a sidebar with tabs: 'Main' (selected), 'Extra', 'Stats & Logs', and 'Version'. The 'Main' tab contains the following fields:

Nom	OnBoardingDemoJob		
Auteur	user@talend.com	Version	0.1
Création		Modification	
Objectifs	Used for on-boarding p		Statut
Description	A simple row generation job		

At the bottom of the configuration window, there are buttons for 'Retour' and 'Suivant'. The background interface also shows a toolbar with icons for 'Contexts(O...', 'Composant', and 'Exécuter (J...'. The 'Exécuter (J...' button is highlighted in yellow.

# Lab 0 : Installation du TOS For Data integration

The screenshot displays the Talend OnBoarding Demo Job 0.1 interface. On the left, a 'Palette' window is open, explaining its purpose: 'La Palette contient différents composants techniques à utiliser pour construire vos Jobs, groupés en familles. Un composant est un connecteur préconfiguré utilisé pour effectuer une opération d'intégration de données spécifique. Il peut minimiser le code manuel requis pour utiliser des sources hétérogènes.' Below this text are five dots and two buttons: 'Essayer' (blue) and 'Retour' (grey). To the right of the palette is a vertical list of categories for finding components, including 'Favoris', 'Récemment util...', 'Applications Mé...', 'Bases de données', 'Big Data', 'Business Intellig...', 'Business', 'Cloud', 'Code Utilisateur', 'Databases', 'Divers', 'DotNET', 'ELT', 'ESB', 'Fichier', 'Internet', 'Logs & Erreurs', 'Orchestration', 'Qualité de donn...', 'Système', 'Talend MDM', 'Transformation', 'Unstructured', and 'XML'. At the bottom, the 'OnBoardingDemoJob 0.1' configuration panel is visible, showing fields for 'Nom' (OnBoardingDemoJob), 'Auteur' (user@talend.com), 'Version' (0.1), 'Création', 'Modification', 'Objectifs' (Used for on-boarding p), 'Statut', and 'Description' (A simple row generation job).

**Palette**

La Palette contient différents composants techniques à utiliser pour construire vos Jobs, groupés en familles. Un composant est un connecteur préconfiguré utilisé pour effectuer une opération d'intégration de données spécifique. Il peut minimiser le code manuel requis pour utiliser des sources hétérogènes.

.....

**Essayer** **Retour** **Suivant**

**OnBoardingDemoJob 0.1**

Main	Nom	OnBoardingDemoJob		
Extra	Auteur	user@talend.com	Version	0.1 M m
Stats & Logs	Création		Modification	
Version	Objectifs	Used for on-boarding p		
	Description	A simple row generation job		

**Finder un composant**

- Favoris
- Récemment util...
- Applications Mé...
- Bases de données
- Big Data
- Business Intellig...
- Business
- Cloud
- Code Utilisateur
- Databases
- Divers
- DotNET
- ELT
- ESB
- Fichier
- Internet
- Logs & Erreurs
- Orchestration
- Qualité de donn...
- Système
- Talend MDM
- Transformation
- Unstructured
- XML

# Prise en main du TOS for data integration

The screenshot displays the Talend Open Studio for Data Integration (7.3.1.20200219\_1130) interface. The main workspace shows a job named "Job Test 0.1" with a single component, "tRowGenerator\_1", connected to "tLogRow\_1". A note above the components states: "This a simple random generation job". The interface includes a Repository pane on the left, a Palette on the right, and a Properties pane at the bottom.

**Repository:** LOCAL: FormationTOS-DI

- Business Models
- Job Designs
  - Test 0.1
- Contexts
- Code
- SQL Templates
- Metadata
  - Documentation
  - Recycle bin

**Palette:**

- note
- Favorites
- Recently Used
  - tDBOutput
  - tDBInput
  - tDBConnection
- Business Intelligence
  - OLAP Cube
  - Palo
- Business
  - Salesforce
    - tSalesforceEinsteinOutput...
- Cloud
  - Amazon
  - Salesforce
- Databases
  - DB Common
  - DB Specifics
- ELT
- Misc
  - Note

**Properties:** Properties not available.



## Prise en main du TOS for data integration

- ✓ Le TOS For Data Integration est basé sur Eclipse
- ✓ Le TOS For Big Data vous permet de développer en mode graphique des traitements java.
- ✓ Un **job** Talend est un ensemble des composants (ou modules) liés entre eux pour traiter un flux de données.

# Mon premier Job Talend Data Integration

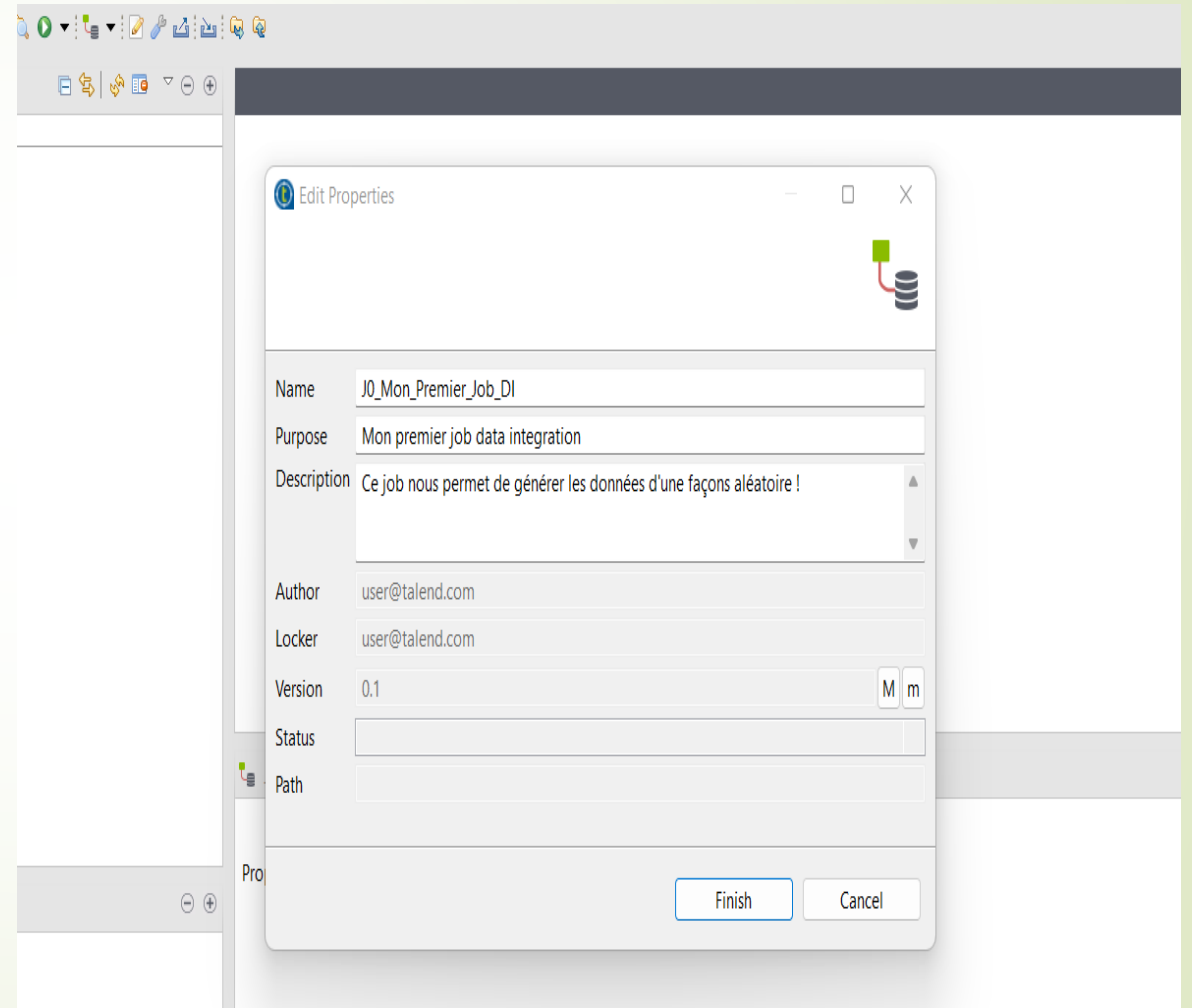
Sur le TOS, sélectionner la référence « Job »  
et avec le bouton droite choisir « créer un Job »





# Mon premier Job Talend Data Integration

- ✓ Donner un nom à votre nouveau job
- ✓ Les champs Objectifs et description sont optionnels
- ✓ Ces deux champs sont importants pour le cycle de vie de votre job
- ✓ Appuyer sur « Finish »



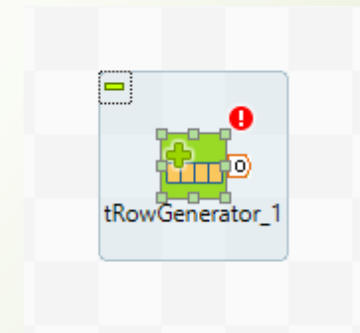
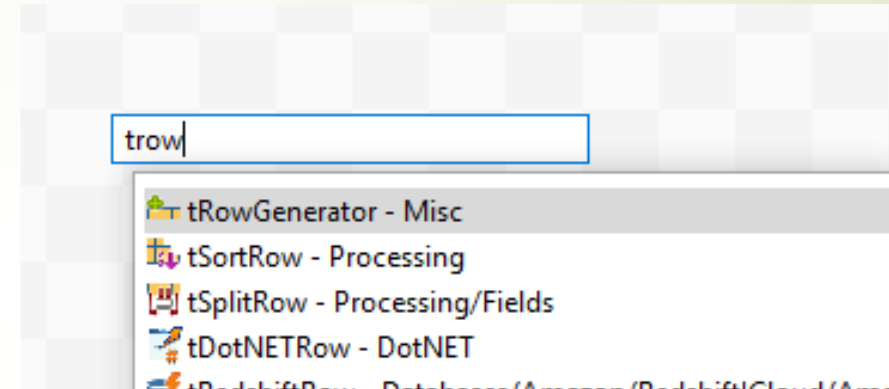
# Mon premier Job Talend Data Integration

Positionner le curseur au  
niveau de l'espace  
designer et taper  
« trow »

Choisir

« tRowGenerator »

Ce composant nous  
permet de générer des  
lignes de données



# Configuration de tRowGenerator

Talend Open Studio for Data Integration - tRowGenerator - tRowGenerator\_1

Schema		Functions		Preview
Column	Type	Functions	Environment vari...	Preview

+ × ↑ ↓ [Icon] [Icon] [Icon] [Icon] Columns ▼ Number of Rows for RowGenerator 100

Function parameters Preview

Parameter	Value	Comment
-----------	-------	---------

OK Cancel

# Ajouter la colonne FirstName

Talend Open Studio for Data Integration - tRowGenerator - tRowGenerator\_1

Schema		Functions		Preview
Column	Type	Functions	Environment vari...	Preview
FirstName	String	TalendString.get...	length=>6 ;	

+ ✖ ⬆ ⬇ 📄 📋 🔄 🔄 💾 Columns ▼ Number of Rows for RowGenerator 100

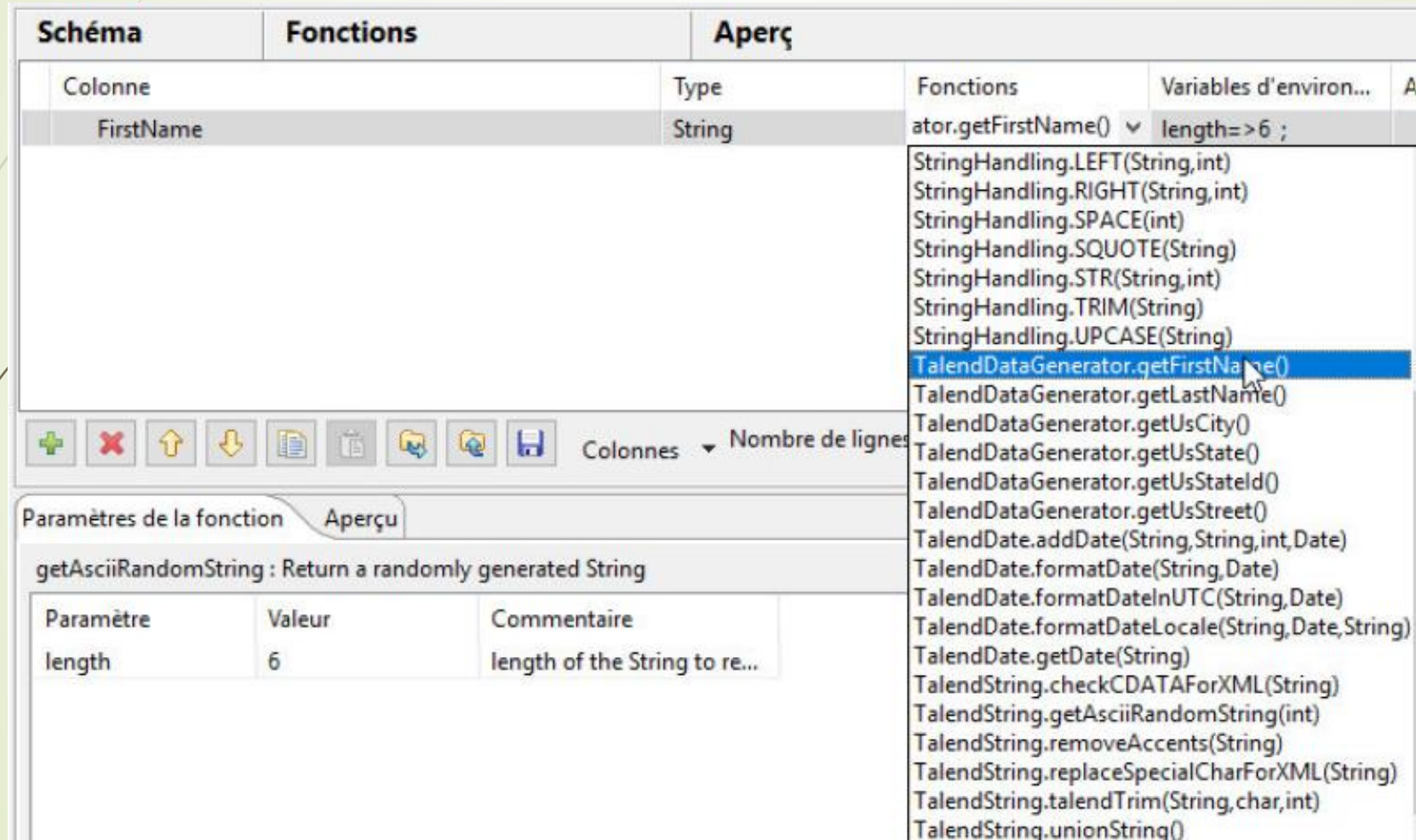
Function parameters Preview

getAsciiRandomString : Return a randomly generated String

Parameter	Value	Comment
length	6	length of the String ...

OK Cancel

# Choisir la fonction getFirstName()



The screenshot shows the Talend Studio interface with the 'Fonctions' (Functions) tab selected. The 'Colonne' (Column) is 'FirstName' and the 'Type' is 'String'. The 'Fonctions' list is open, showing various functions. 'TalendDataGenerator.getFirstName()' is highlighted. The 'Paramètres de la fonction' (Function Parameters) tab is also visible, showing the 'length' parameter set to 6.

Colonne	Type	Fonctions	Variables d'environ...
FirstName	String	ator.getFirstName() ▾	length=>6 ;

StringHandling.LEFT(String,int)  
StringHandling.RIGHT(String,int)  
StringHandling.SPACE(int)  
StringHandling.SQUOTE(String)  
StringHandling.STR(String,int)  
StringHandling.TRIM(String)  
StringHandling.UPCASE(String)  
**TalendDataGenerator.getFirstName()**  
TalendDataGenerator.getLastName()  
TalendDataGenerator.getUsCity()  
TalendDataGenerator.getUsState()  
TalendDataGenerator.getUsStatId()  
TalendDataGenerator.getUsStreet()  
TalendDate.addDate(String,String,int,Date)  
TalendDate.formatDate(String,Date)  
TalendDate.formatDateInUTC(String,Date)  
TalendDate.formatDateLocale(String,Date,String)  
TalendDate.getDate(String)  
TalendString.checkCDATAForXML(String)  
TalendString.getAsciiRandomString(int)  
TalendString.removeAccents(String)  
TalendString.replaceSpecialCharForXML(String)  
TalendString.talendTrim(String,char,int)  
TalendString.unionString()

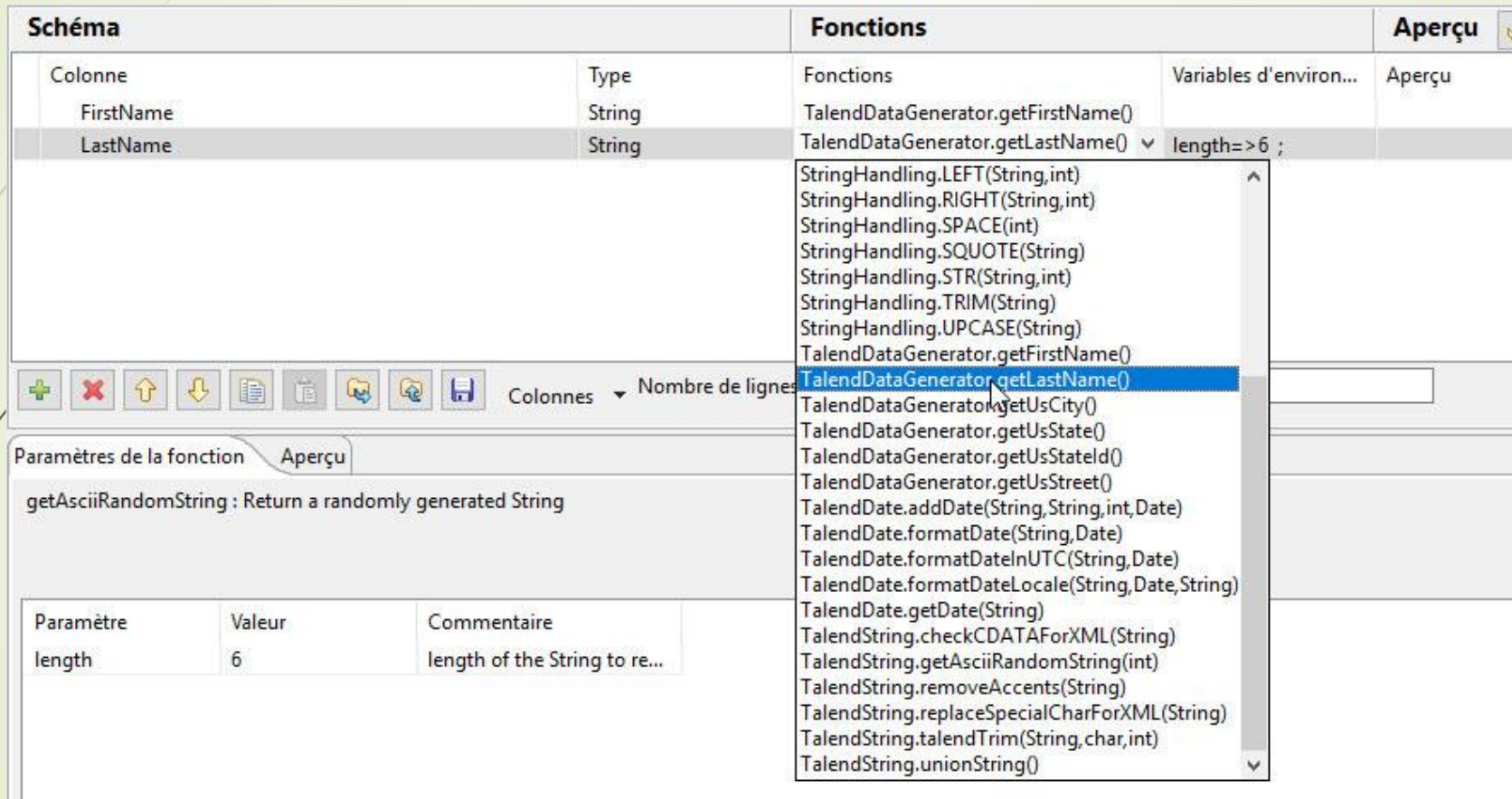
Paramètres de la fonction    Aperçu

getAsciiRandomString : Return a randomly generated String

Paramètre	Valeur	Commentaire
length	6	length of the String to re...



# Choisir la fonction getLastName()



The screenshot displays the Talend Studio interface with the 'Fonctions' (Functions) dropdown menu open. The 'Schéma' (Schema) tab shows a table with two columns: 'FirstName' and 'LastName', both of type 'String'. The 'Paramètres de la fonction' (Function Parameters) tab shows the 'length' parameter set to 6. The 'Fonctions' dropdown menu lists various functions, with 'TalendDataGenerator.getLastName()' highlighted.

Colonne	Type
FirstName	String
LastName	String

Paramètre	Valeur	Commentaire
length	6	length of the String to re...

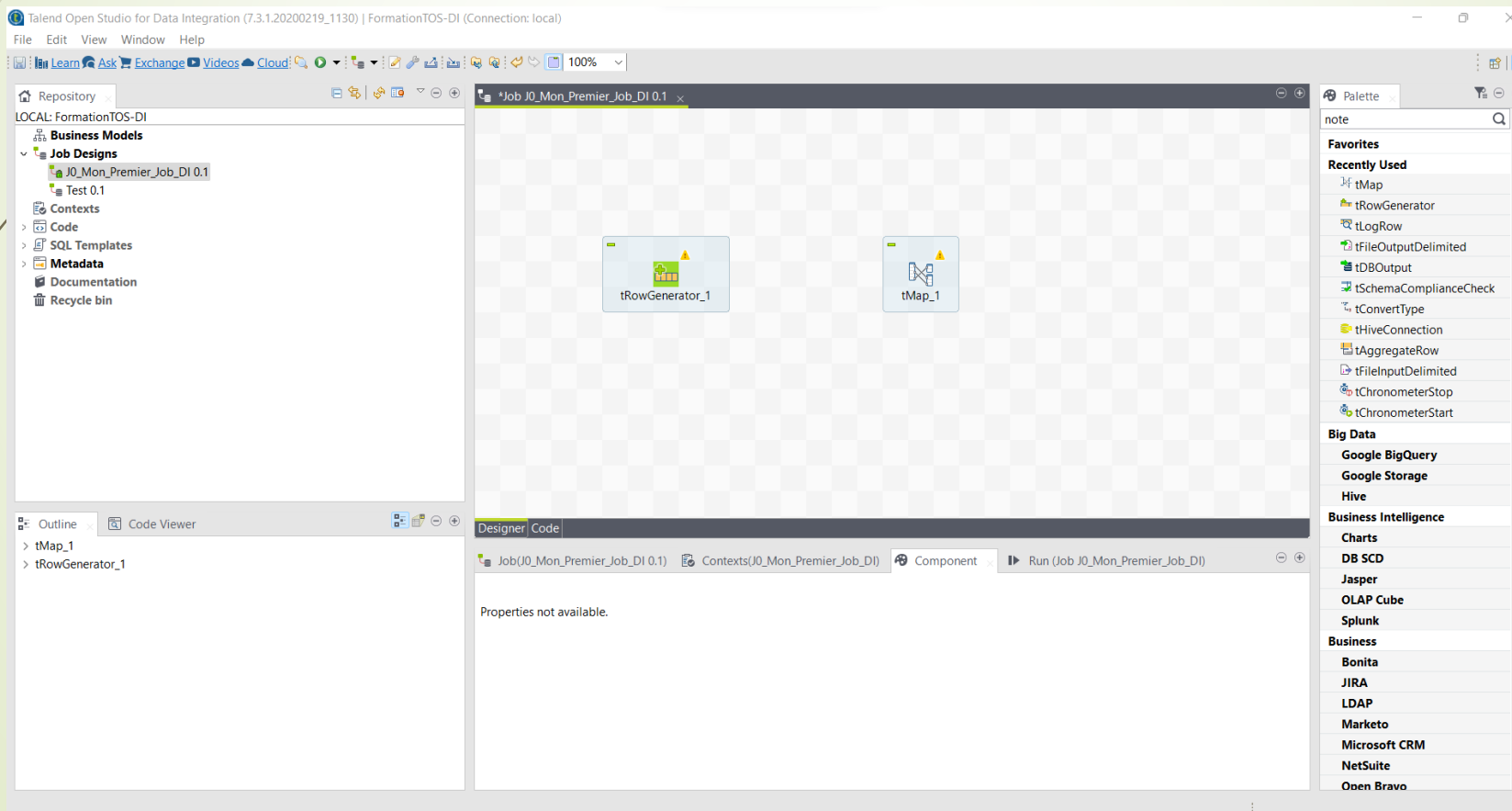
Functions list (from top to bottom):

- TalendDataGenerator.getFirstName()
- TalendDataGenerator.getLastName()
- StringHandling.LEFT(String,int)
- StringHandling.RIGHT(String,int)
- StringHandling.SPACE(int)
- StringHandling.SQUOTE(String)
- StringHandling.STR(String,int)
- StringHandling.TRIM(String)
- StringHandling.UPCASE(String)
- TalendDataGenerator.getFirstName()
- TalendDataGenerator.getLastName()
- TalendDataGenerator.getUsCity()
- TalendDataGenerator.getUsState()
- TalendDataGenerator.getUsStateId()
- TalendDataGenerator.getUsStreet()
- TalendDate.addDate(String,String,int,Date)
- TalendDate.formatDate(String,Date)
- TalendDate.formatDateInUTC(String,Date)
- TalendDate.formatDateLocale(String,Date,String)
- TalendDate.getDate(String)
- TalendString.checkCDATAForXML(String)
- TalendString.getAsciiRandomString(int)
- TalendString.removeAccents(String)
- TalendString.replaceSpecialCharForXML(String)
- TalendString.talendTrim(String,char,int)
- TalendString.unionString()

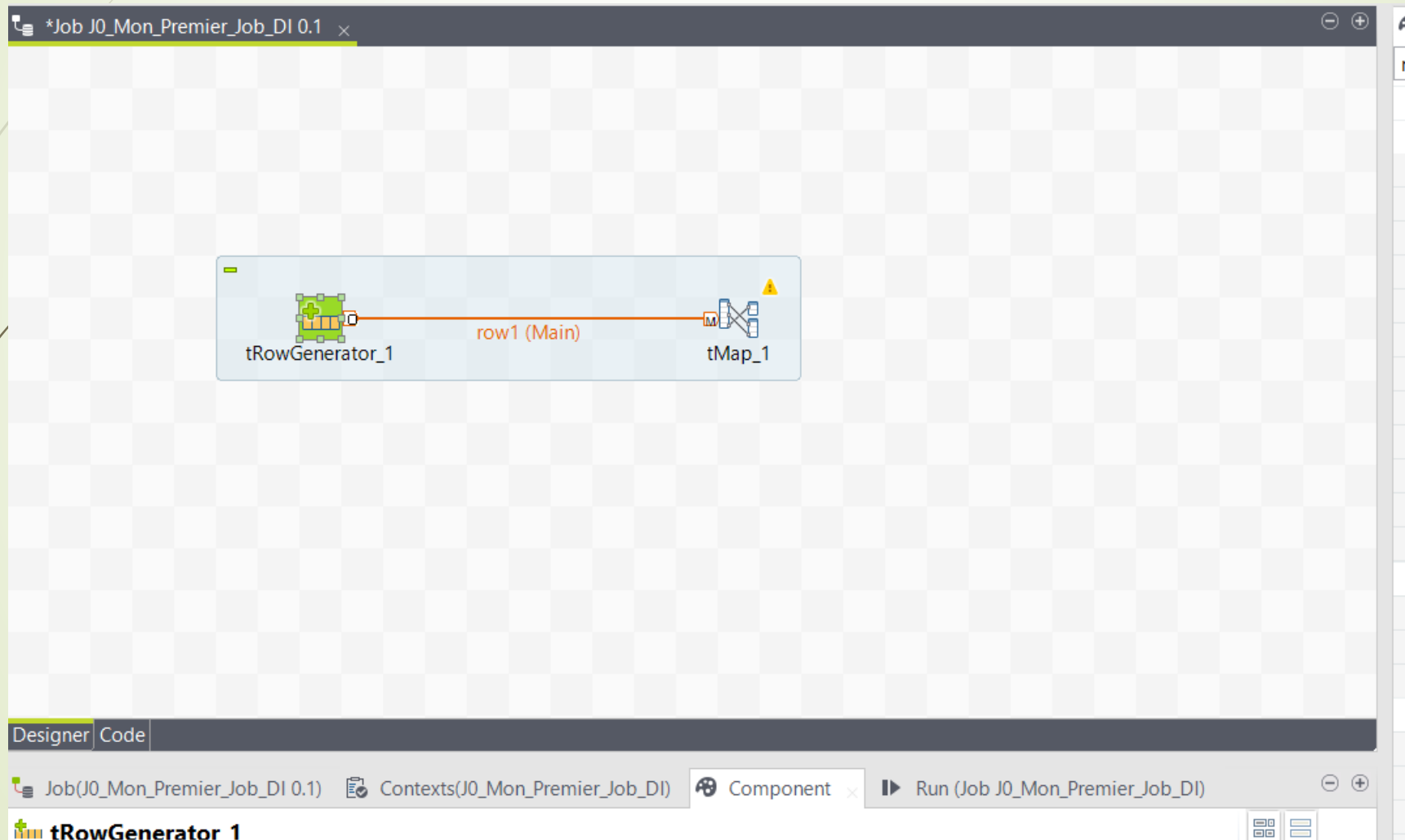


# Ajout du composant Tmap

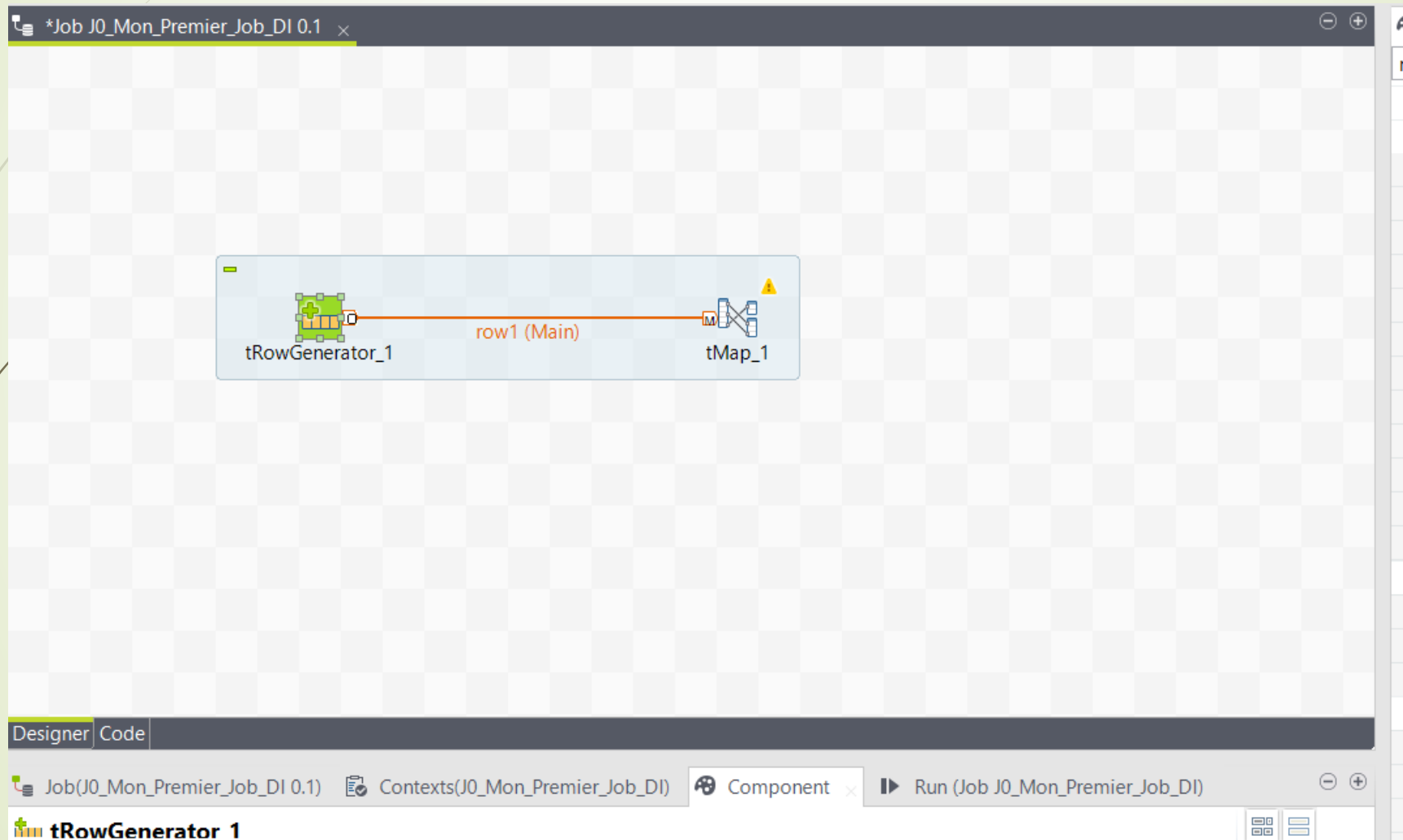
Taper Tmap dans le designer  
choisir le composant «Tmap »



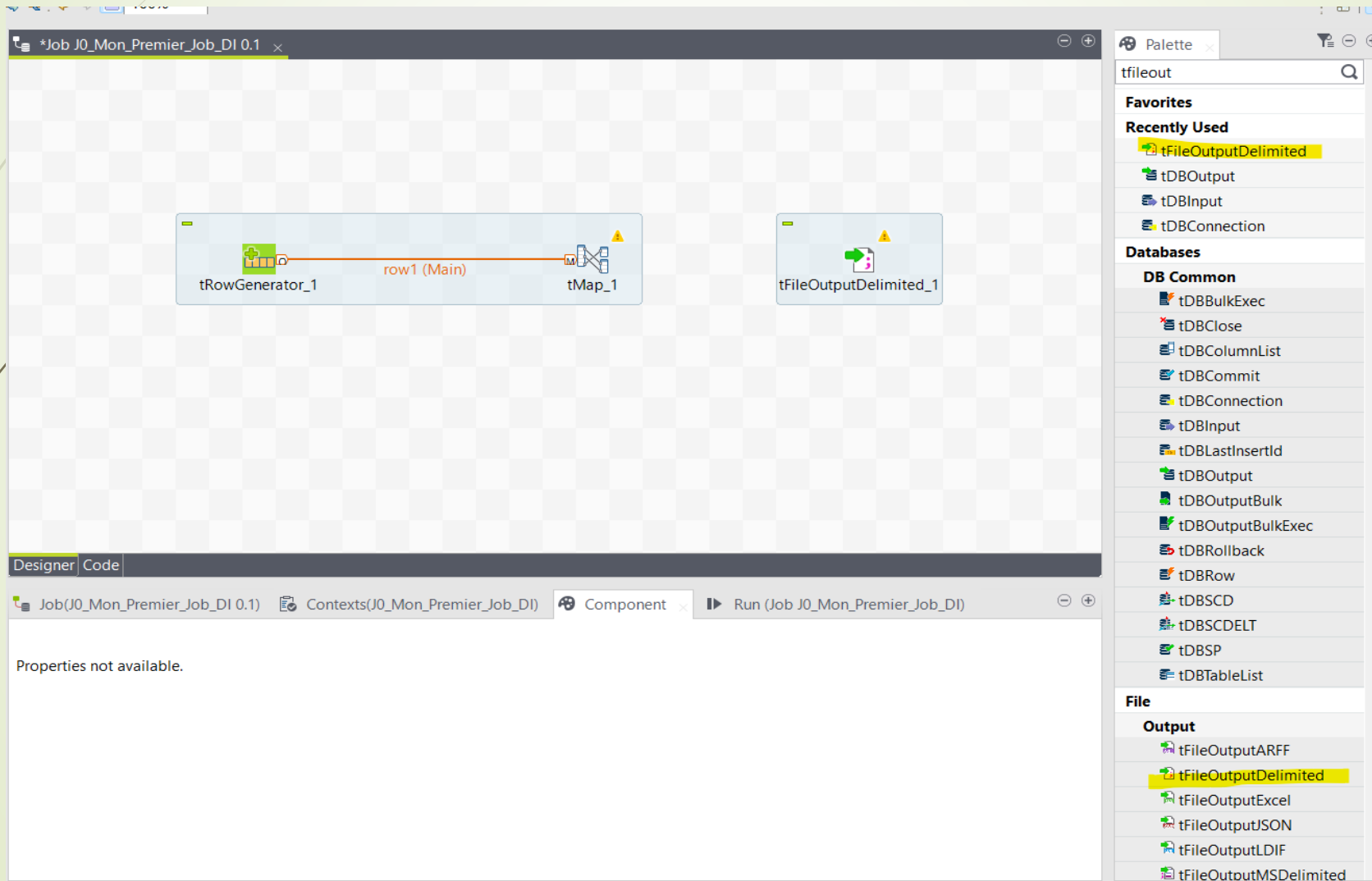
# Lier les deux composants «TRowGenerator et Tmap



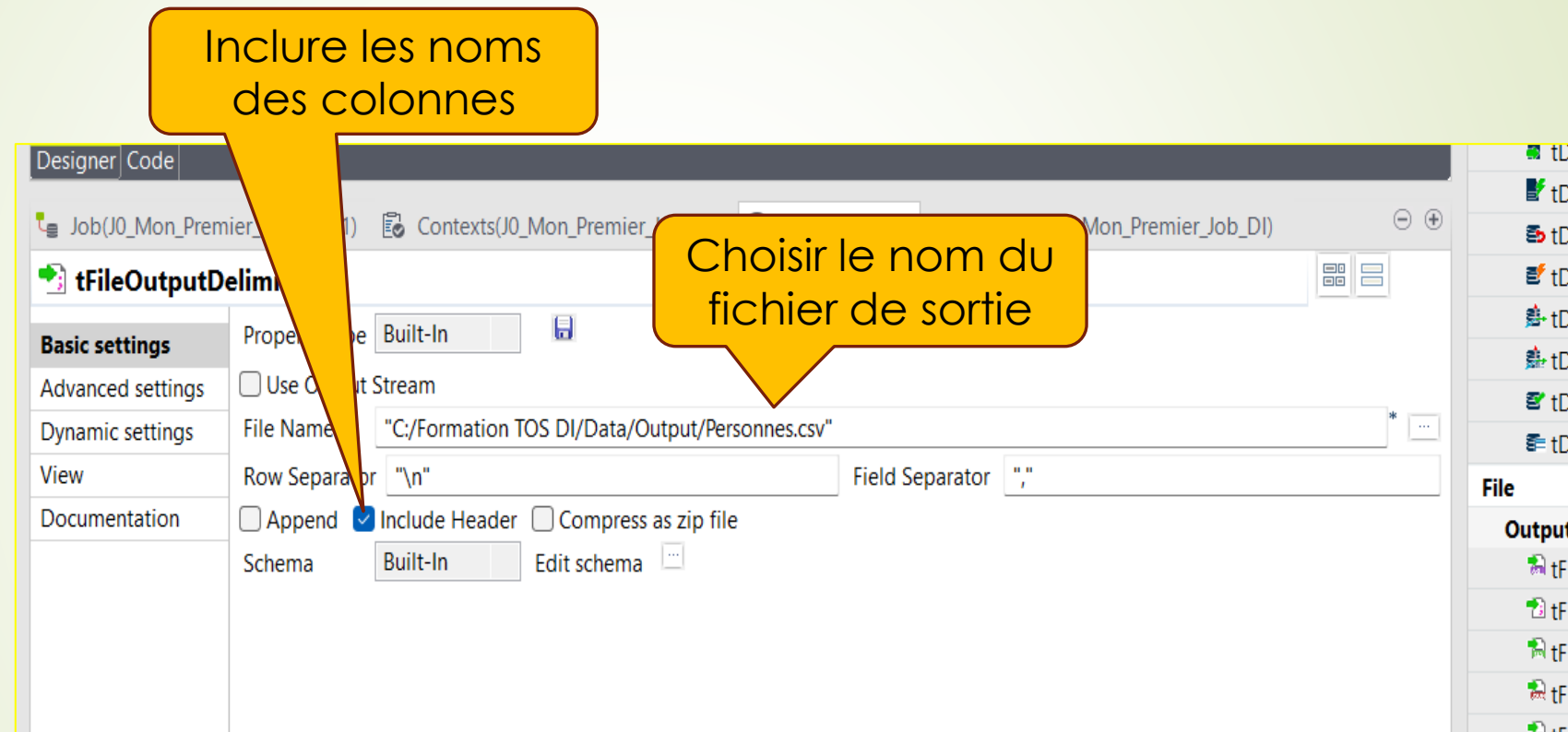
# Lier les deux composants «TRowGenerator et Tmap



# Ajout du composant TfileOutputdelimited



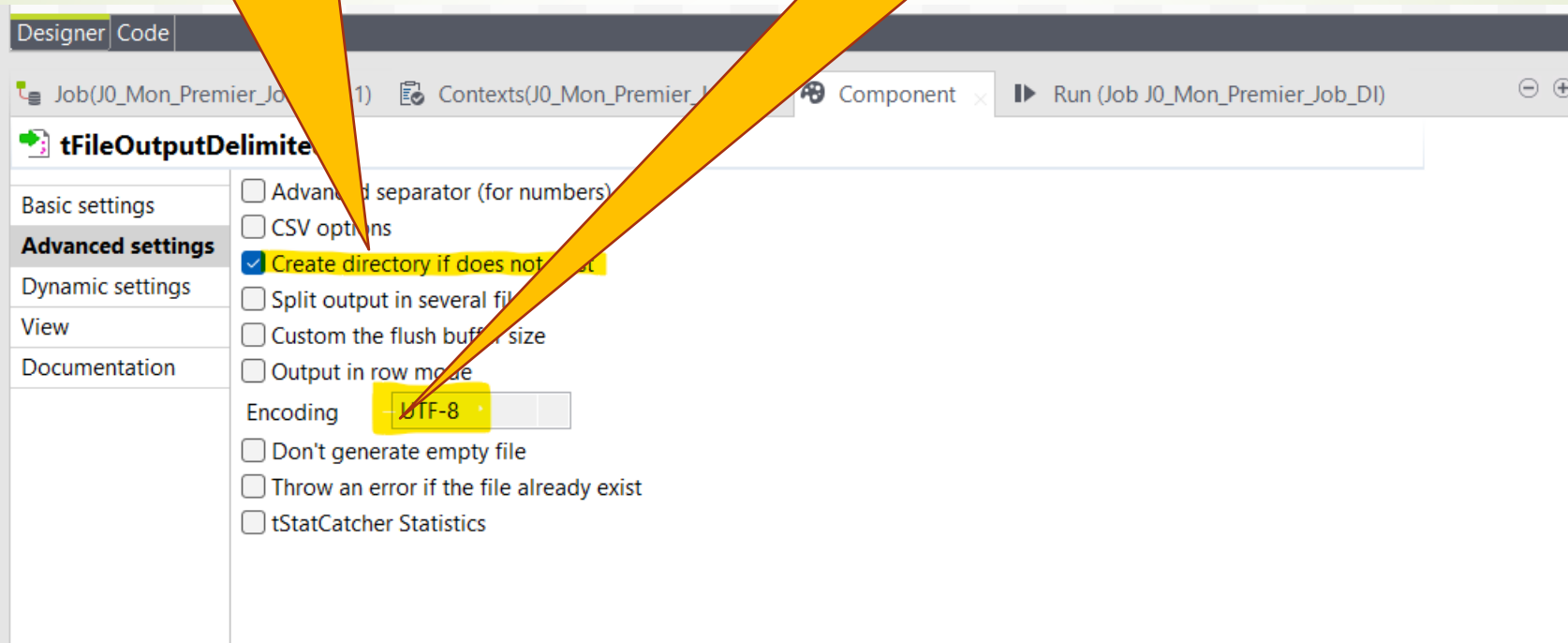
# Configuration TfileOutputdelimited



# Configuration TfileOutputdelimited

création du répertoire

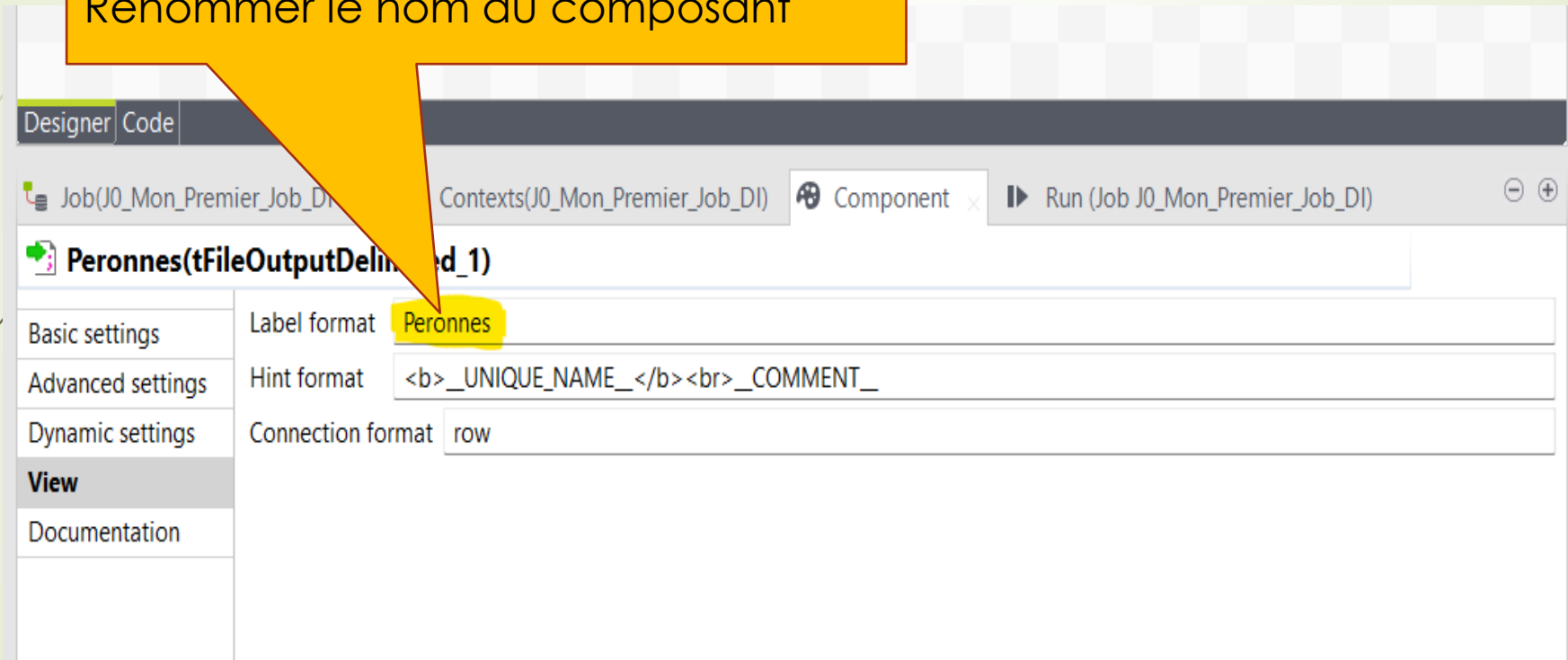
Encodage du fichier





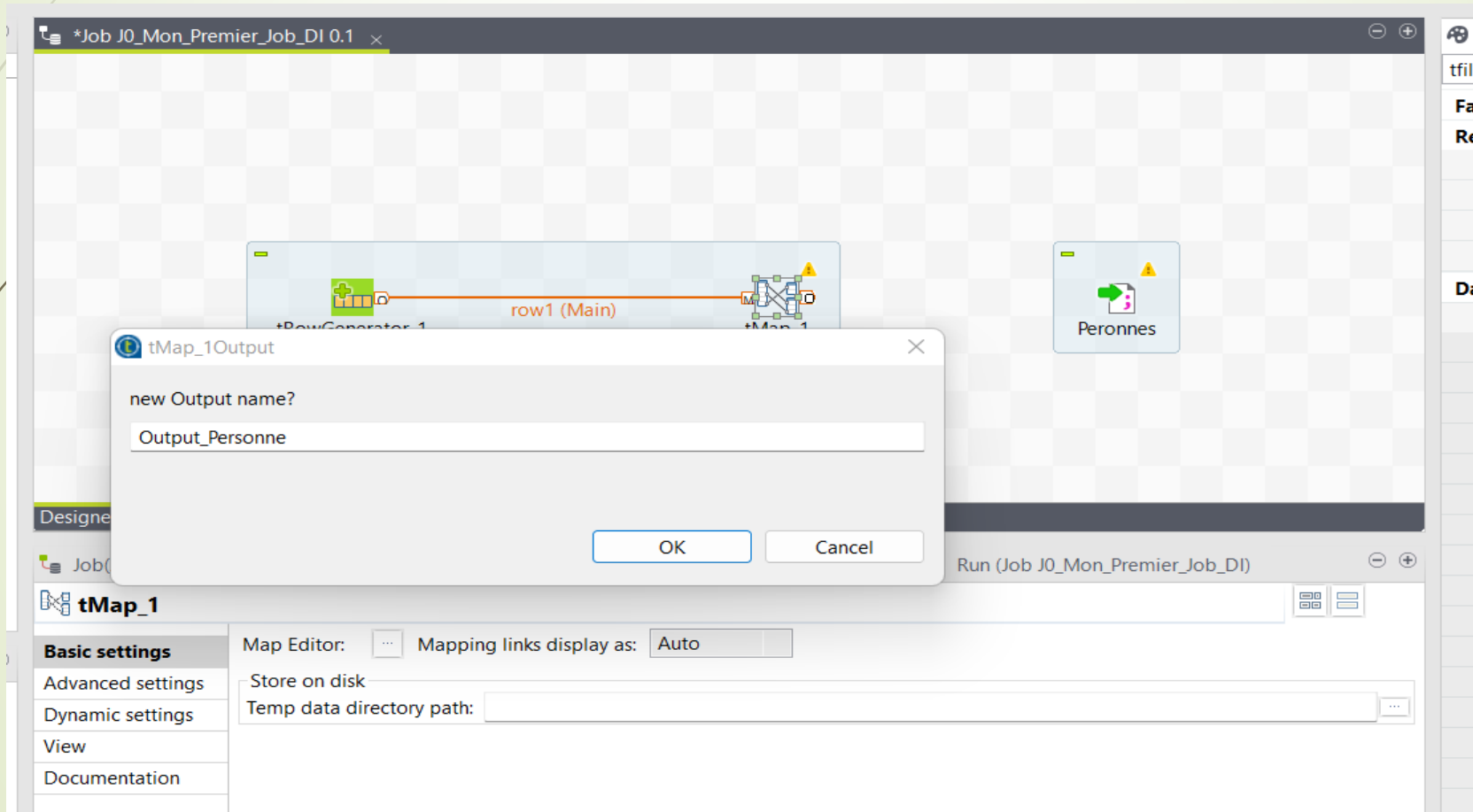
# Renommer le nom du Tfileoutputdelimited

Renommer le nom du composant



# Lier les deux composants

Cliquer sur le composant Tmap et le donner un nom



# Lier les deux composants

Mapper les deux champs entres eux

The screenshot shows the Talend Open Studio interface for mapping two components: 'row1' and 'Output\_Personne'. The 'row1' component has two columns: 'FirstName' and 'LastName'. The 'Output\_Personne' component has two columns: 'FirstName' and 'LastName'. The mapping is shown in the 'Var' panel, where 'row1.FirstName' is mapped to 'Output\_Personne.FirstName' and 'row1.LastName' is mapped to 'Output\_Personne.LastName'. The 'Schema editor' and 'Expression editor' tabs are visible at the bottom, showing the column details for both components.

**row1**

Column
FirstName
LastName

**Output\_Personne**

Expression	Column
row1.FirstName	FirstName
row1.LastName	LastName

**Schema editor**

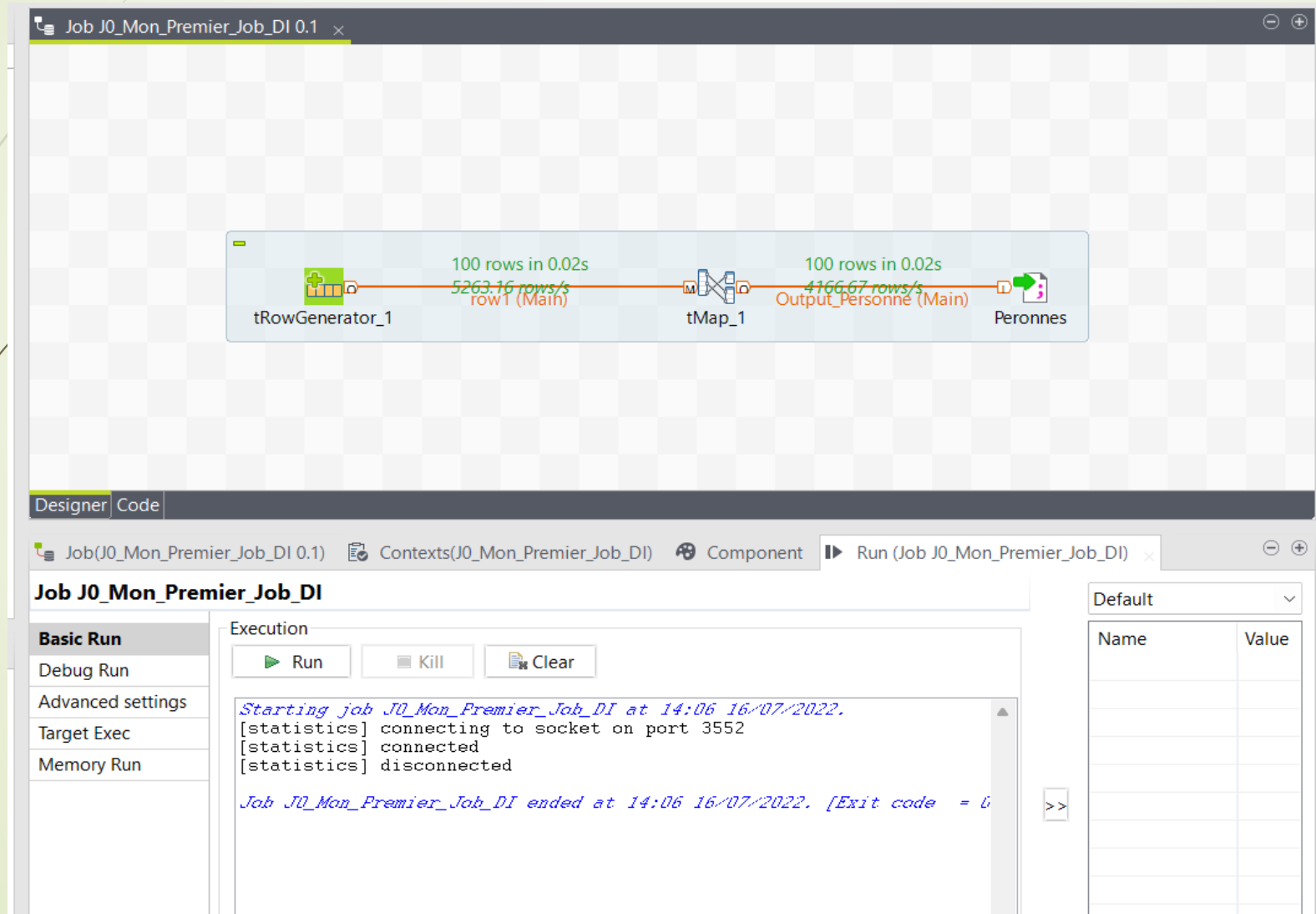
Column	K...	Type	<input checked="" type="checkbox"/> N.	Date Pattern (Ctrl+...)	Length	Precision	Default	Comment
FirstName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					
LastName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					

**Output\_Personne**

Column	K...	Type	<input checked="" type="checkbox"/> N.	Date Pattern (Ctrl+...)	Length	Precision	Default	Comment
FirstName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					
LastName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>					

Apply Ok Cancel

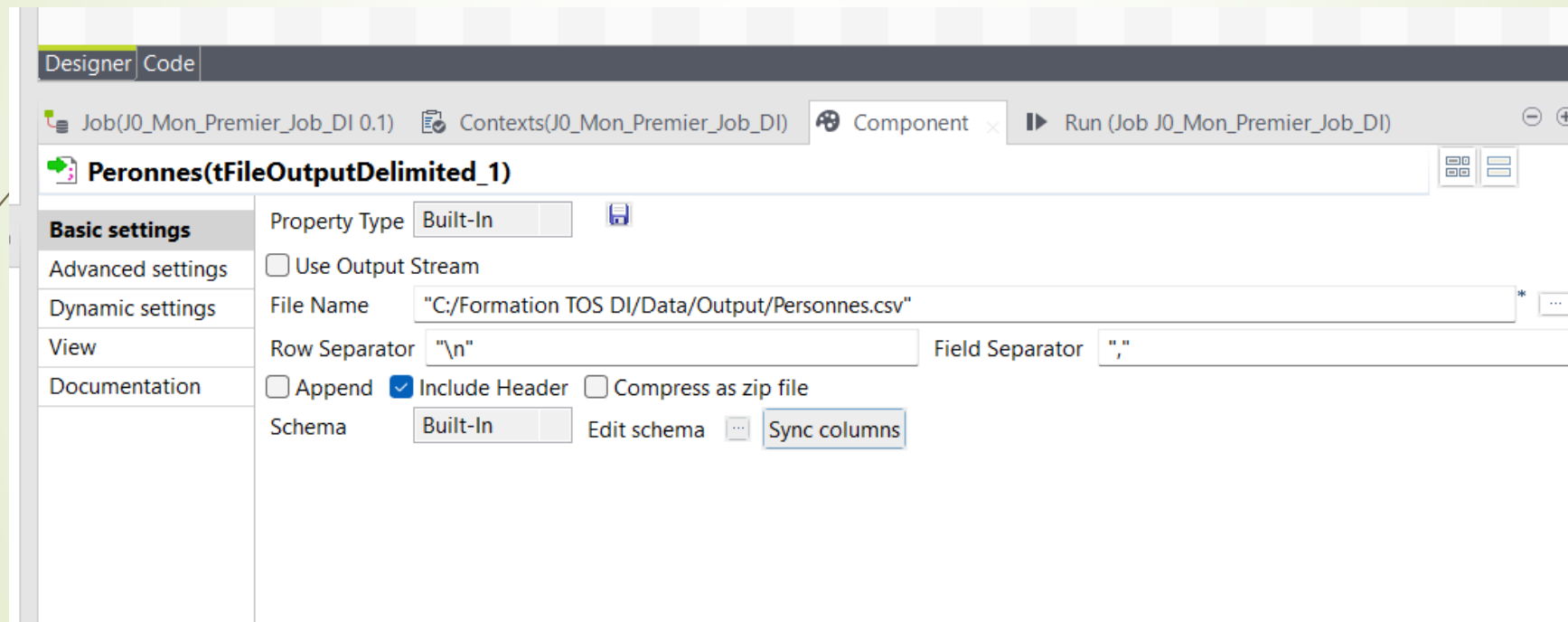
# Exécuter votre premier job





# Vérification du résultat

Vérifiez que le fichier est correctement créé dans le répertoire que vous avez choisi.







”

