

TALEND BIG DATA





Agenda

- ☐ Introduction à Talend Big Data
- ☐ Prise en main de Talend Big Data



Objectif du cours

- ☐ Introduction à Talend Big Data
- ☐ Architecture du Talend Open Studio For Big Data
- ☐ Installation du Talend Open Studio For Big Data
- ☐ Lancer le premier Job Talend Big Data



Talend



talend

Les produits Talend



Les produits Talend





Talend Big Data

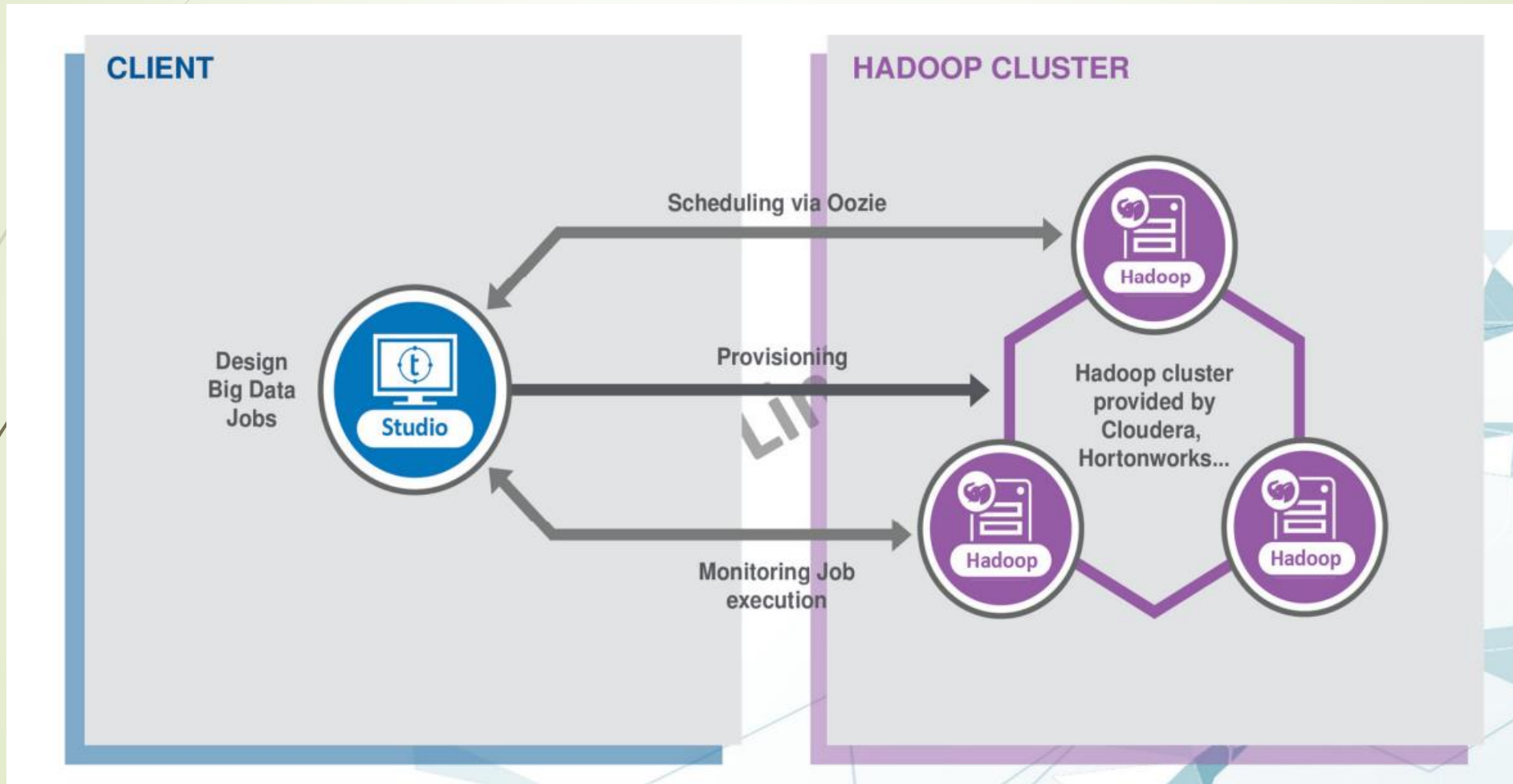
- ✓ Talend Big Data a été conçu pour simplifier le développement, l'intégration et la gestion des flux data
- ✓ Talend Big Data élimine la nécessité pour les utilisateurs d'affronter la complexité liée au développement et à la maintenance de code Java.
- ✓ Talend génère le code natif et optimisé pour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



Talend Big Data

- ✓ 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

Architecture Talend for Big Data





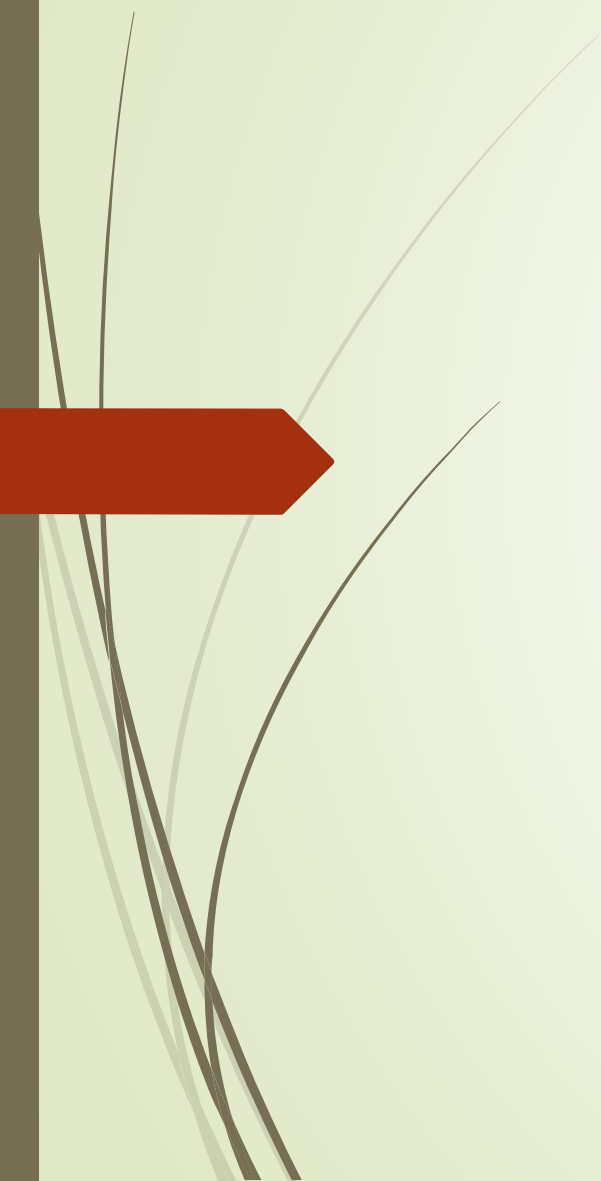
Architecture Talend for Big data

- ✓ Dans le studio Talend , vous créer et exécutez des job Big Data
Tirant parti du cluster Hadoop afin de gérer de grandes volumes de données.
- ✓ Une fois lancés, ces Jobs sont envoyés, déployés et exécutés sur le cluster Hadoop.
- ✓ Oozie, un système d'ordonnancement de workflows, est intégré dans le Studio, à travers lequel vous pouvez déployer, ordonnancer et exécuter des Jobs Big Data dans un cluster Hadoop et monitorer le statut d'exécution, ainsi que les résultats des Jobs.



Architecture Talend for Big data

- ✓ Le lancement des jobs peut se faire sur la machine de développeur sans passer par oozie
- 



LAB

Lab 0 : Installation du TOS For Data integration

Télécharger le TOS For Data Integration version 6.4.1 à partir du lien :

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

Installer Java 8 :

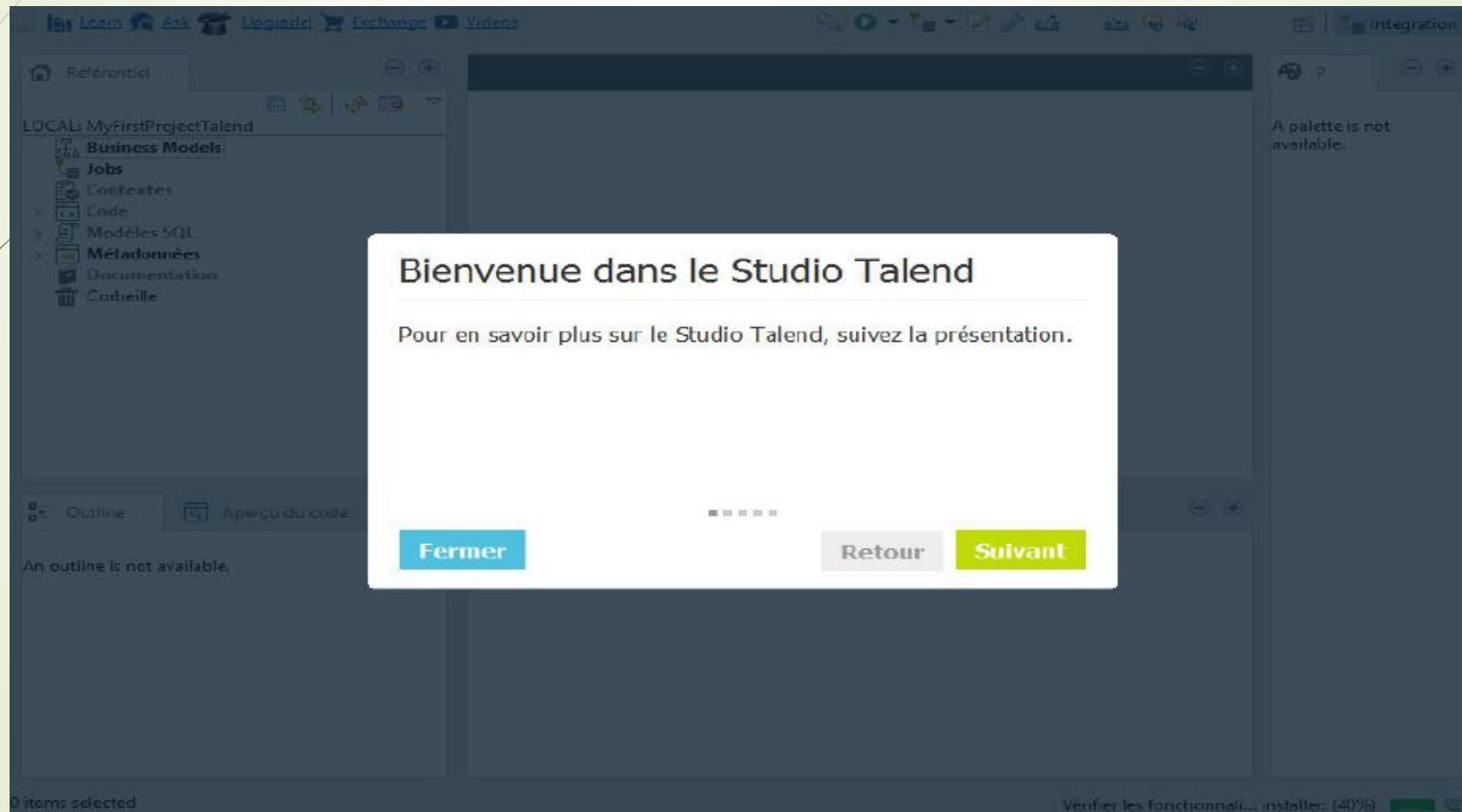
<https://www.oracle.com/technetwork/java/javase/downloads/index.html>

Lab 0 : Installation du TOS For Big Data

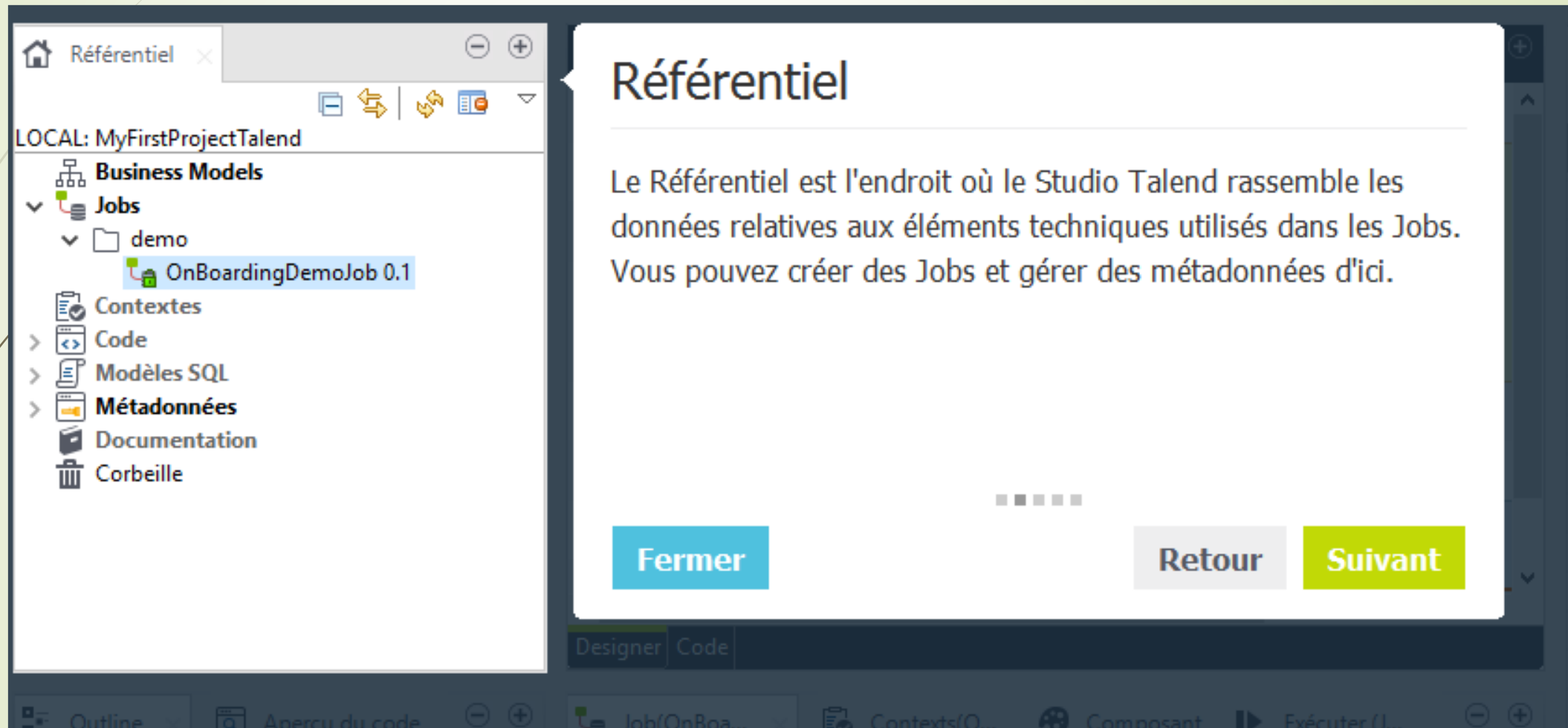
Lancer le TOS For Data Integration en exécutant TOS_DI-win-x86_64.exe

about_files	02/08/2022 00:30	Dossier de fichiers	
configuration	02/08/2022 01:01	Dossier de fichiers	
features	02/08/2022 00:31	Dossier de fichiers	
p2	02/08/2022 00:31	Dossier de fichiers	
plugins	02/08/2022 00:47	Dossier de fichiers	
temp	02/08/2022 00:59	Dossier de fichiers	
TOS_BD-macosx-cocoa.app	02/08/2022 00:30	Dossier de fichiers	
workspace	02/08/2022 00:59	Dossier de fichiers	
.eclipseproduct	02/08/2022 00:30	Fichier ECLIPSEPR...	1 Ko
license.txt	02/08/2022 00:30	Document texte	1 Ko
NOTICE.txt	02/08/2022 00:30	Document texte	27 Ko
TOS_BD-linux-gtk-x86.sh	02/08/2022 00:30	Shell Script	1 Ko
TOS_BD-linux-gtk-x86_64	02/08/2022 00:30	Fichier	73 Ko
TOS_BD-linux-gtk-x86_64.ini	02/08/2022 00:30	Paramètres de con...	1 Ko
TOS_BD-macosx-cocoa.ini	02/08/2022 00:30	Paramètres de con...	1 Ko
TOS_BD-win-x86_64.exe	02/08/2022 00:30	Application	305 Ko
TOS_BD-win-x86_64.ini	02/08/2022 00:30	Paramètres de con...	1 Ko

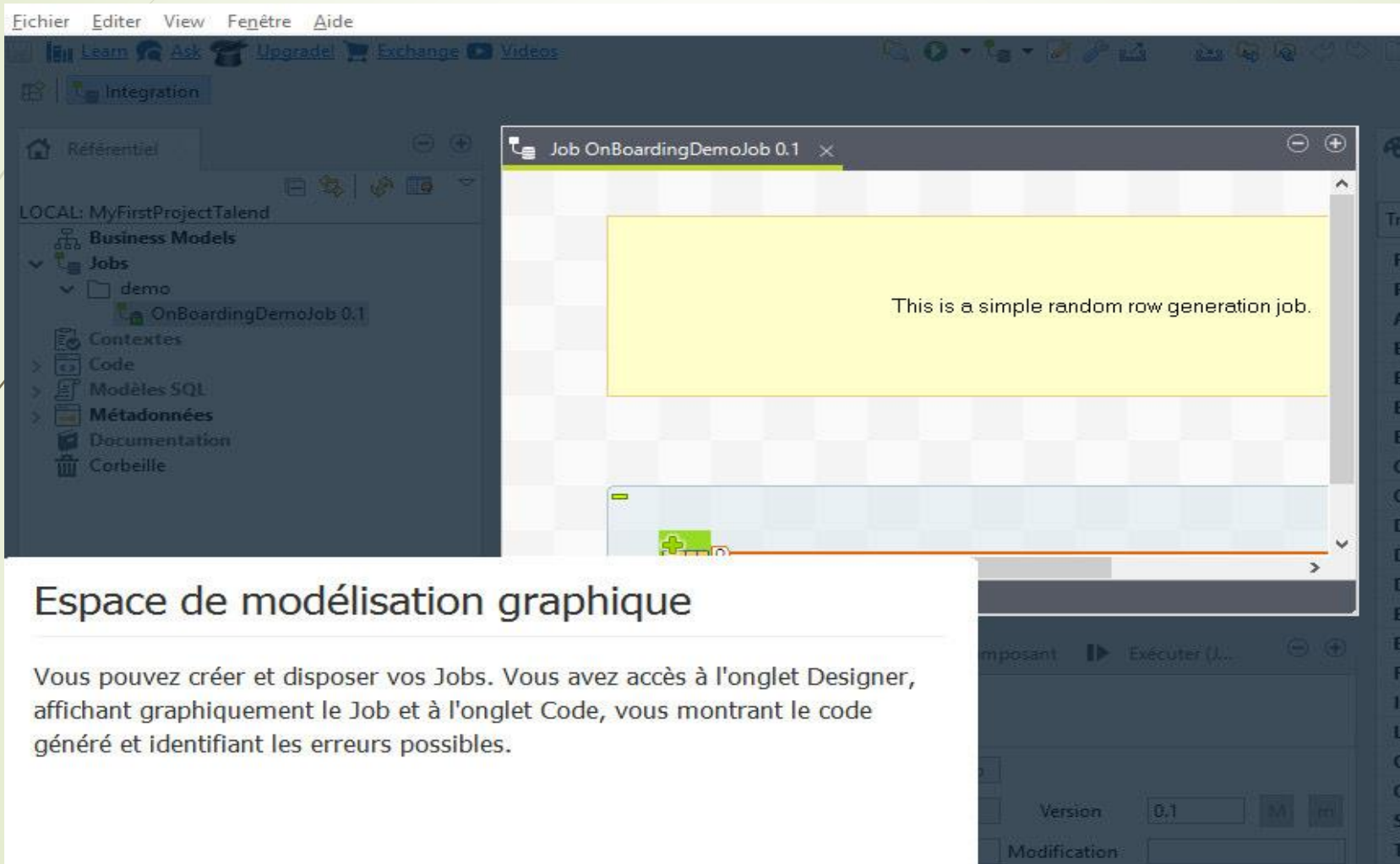
Lab 0 : Installation du TOS For Big Data



Lab 0 : Installation du TOS For Big Data



Lab 0 : Installation du TOS For Big Data



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 Big Data

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 main workspace displays a 'simple random row generation job.' diagram. Overlaid on this is a configuration window titled 'OnBoardingDemoJob 0.1'. The 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...)'.

Lab 0 : Installation du TOS For Big Data

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 <input type="text"/> M <input type="text"/> m
Stats & Logs	Création	<input type="text"/>	Modification	<input type="text"/>
Version	Objectifs	Used for on-boarding p	Statut	<input type="text"/>
	Description	A simple row generation job		

Trouver un cor

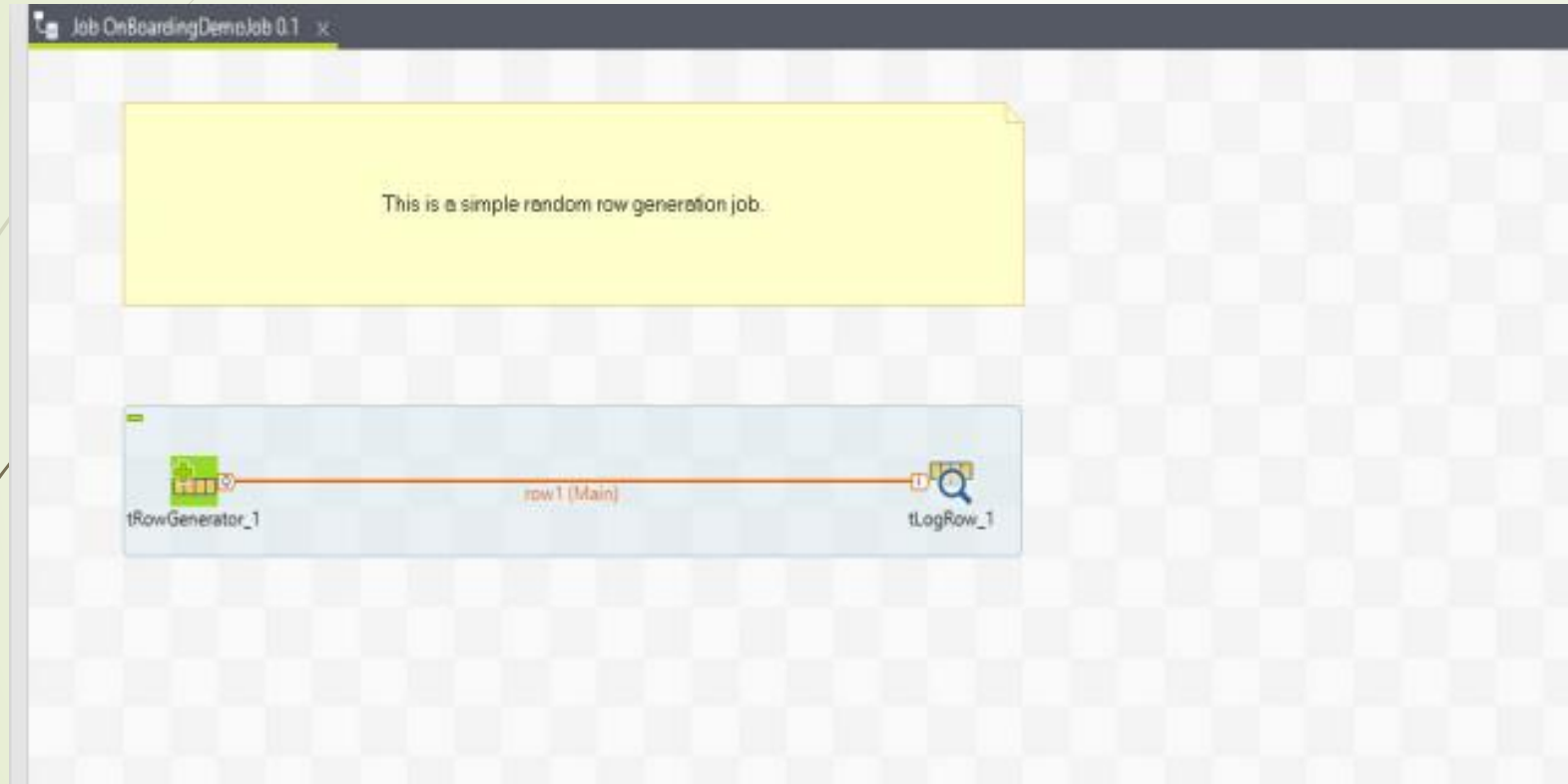
- Favoris
- Récemment util...
- Applications Mé...
- Bases de données
- Big Data
- Business Intellig...
- Business
- Cloud
- Code Utilisateur
- Databases
- Divers
- DotNET
- ELT
- ESB
- Fichier
- Internet
- Logs & Erreurs
- Orchestration
- Qualité de donn...
- Système
- Talend MDM
- Transformation
-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



Prise en main du TOS for Big Data

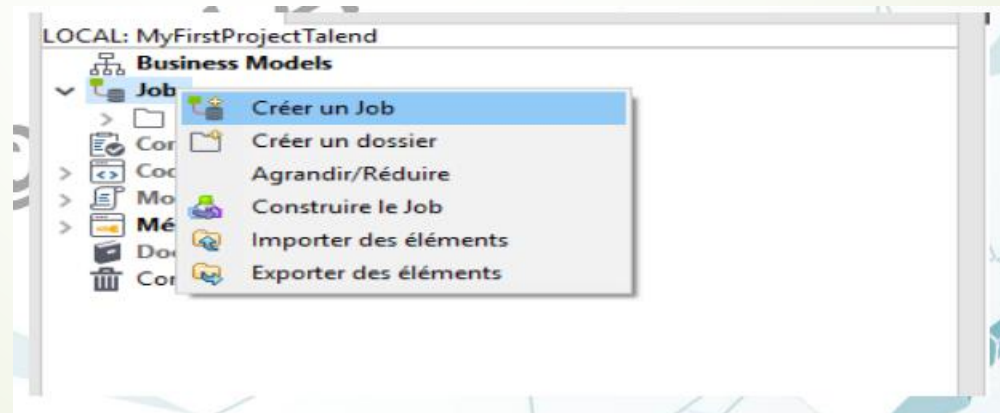
- ✓ Le TOS For Big Data est basé sur Eclipse
- ✓ Le TOS For Big Data vous permet de développer en mode graphique des traitements java.
- ✓ Un **job** Talend est un ensemble des composants (ou modules) liés entre eux pour traiter un flux de données.

Prise en main Talend Big Data



Mon premier Job Talend Big Data

- ✓ Si vous utiliser Virtual Box, ouvrir le port 50010 pour le datanode
 - ✓ Démarrer votre cluster HDP
 - ✓ Sur le TOS, sélectionner la référence « Job »
- Avec le bouton droit choisir « créer un job »



Mon premier Job Talend Big Data

- ✓ 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 »

The screenshot shows the 'Nouveau Job' dialog box in Talend Studio. The title bar says 'Nouveau Job'. Below the title bar, there is a warning icon and text: 'Il est recommandé de ne pas laisser le champ Description vide.' (It is recommended not to leave the Description field empty). The dialog contains the following fields:

- Nom: JO_MON_PREMIER_JOB_BGD
- Objectif: Mon premier job big data
- Description: (empty text area)
- Créé par: user@talend.com
- Verrouillé par: (empty text field)
- Version: 0.1 (with 'M' and 'm' buttons)
- Statut: (empty dropdown menu)
- Chemin d'accès: (empty text field with a 'Sélectionner' button)

At the bottom right, there are two buttons: 'Finish' (highlighted with a blue border) and 'Cancel'.



Mon premier Job Talend Big Data

- ✓ Positionner le curseur au niveau de l'espace designer et taper « trow »
- ✓ Choisir « tRowGenerator »
- ✓ Ce composant nous permet de générer des lignes de données qu'on va les stocker dans un fichier HDFS

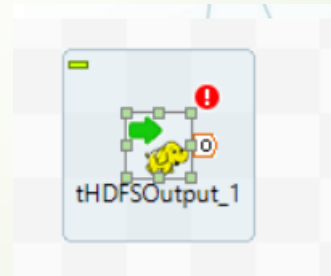


Mon premier Job Talend Big Data


- ✓ Positionner le curseur au niveau de l'espace designer et taper « trow »
- ✓ Choisir « tRowGenerator »
- ✓ Ce composant nous permet de générer des lignes de données qu'on va les stocker dans un fichier HDFS
- ✓ Ajouter deux colonnes FirsName et LastName

Ajout du composant tHDFSOutput

- ✓ Taper thdfsoutput dans le designer
- ✓ Choisir le composant « tHDFSOutput »
- ✓ Ce composant sert à écrire dans des fichiers HDFS
- ✓ Il faut configurer ce composant pour se connecter à HDFS



Ajout du composant tHDFSOutput

 Ce composant tHDFSOutput requiert l'installation d'au moins un Jar externe. Installer...

Type de propriété Built-In ▼

Schéma Built-In ▼ Modifier le schéma ...

☐ Utiliser une connexion existante

Version

Distribution Amazon EMR ▼ Version EMR 5.5.0 (Apache 2.7.3) ▼

Connexion

URI du NameNode Amazon EMR
Apache
Cloudera
HortonWorks
MapR
Pivotal HD
Custom - Unsupported

☒ Utiliser le nom d'utilisateur

Authentification

Utilisateur

Nom de fichier "" *

Type de fichier

Type Fichier texte *

Action Create ▼

Séparateur de lignes "\n" * Séparateur de champs "," *

☐ Encodage personnalisé

☐ Compression

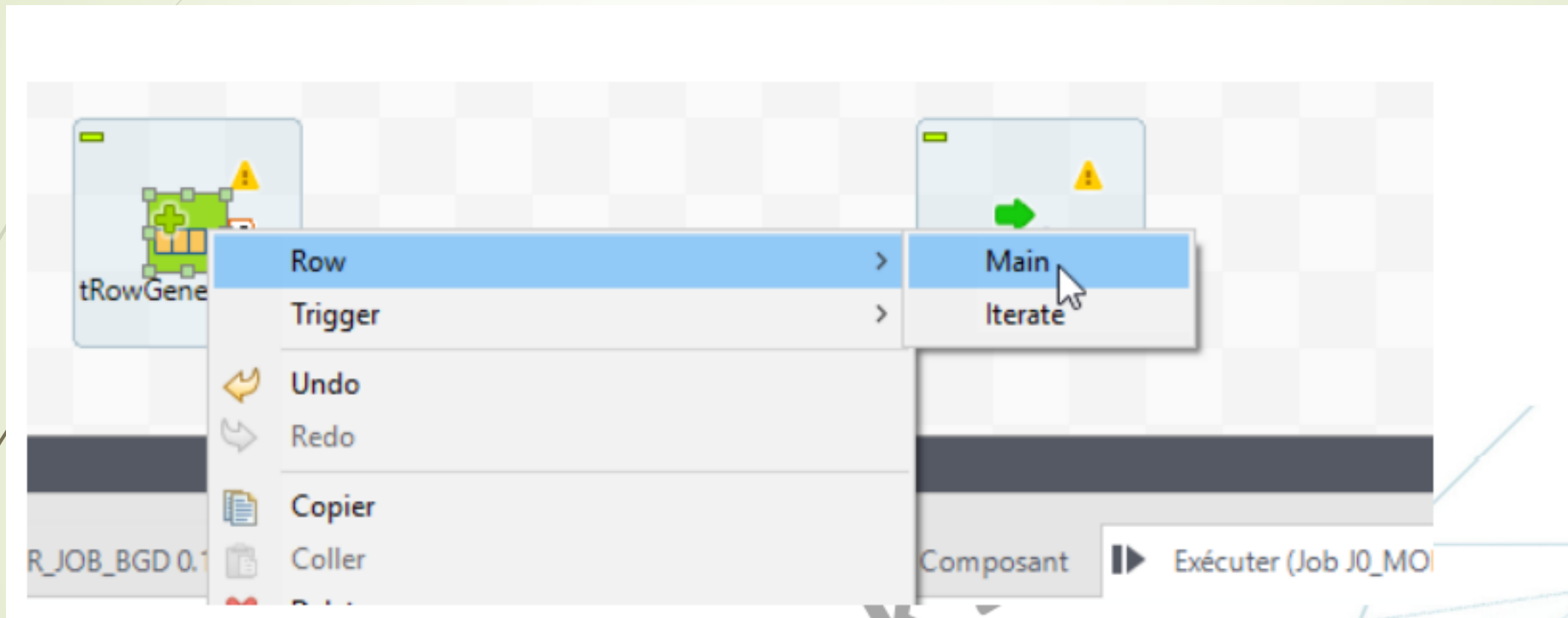
☐ Compresser les données



Configurer le tHDFSoutput

- ✓ Choisir la version de votre Horthonworks
- ✓ URI NamdeNode=hdfe://sandbox.hortonworks.com:8020
- ✓ Laisser le user par default « anonymous »
- ✓ Choisir un nom à votre fichier HDFS

Lier les deux composants



Exécuter votre premier job







”

