

TALEND DATA INTEGRATION





Agenda

- ❑ Introduction à Talend Data integration
- ❑ **Prise en main de Talend Big Data**



2.

PRISE EN MAIN

DE TALEND

OPEN STUDIO FOR Big DATA



Objectifs

- ❑ Avantages de Talend pour Big Data
- ❑ Les principaux composants de Talend BD
- ❑ Déposer des fichiers sur HDFS
- ❑ Ecrire et lire des fichiers sur HDFS
- ❑ Lire des tables Hive
- ❑ Création des tables Hive
- ❑ Gérer des métadonnées dans Hive avec TOS For Big Data
- ❑ Importer et exporter des données avec Sqoop



Avantages du TOS For Big Data

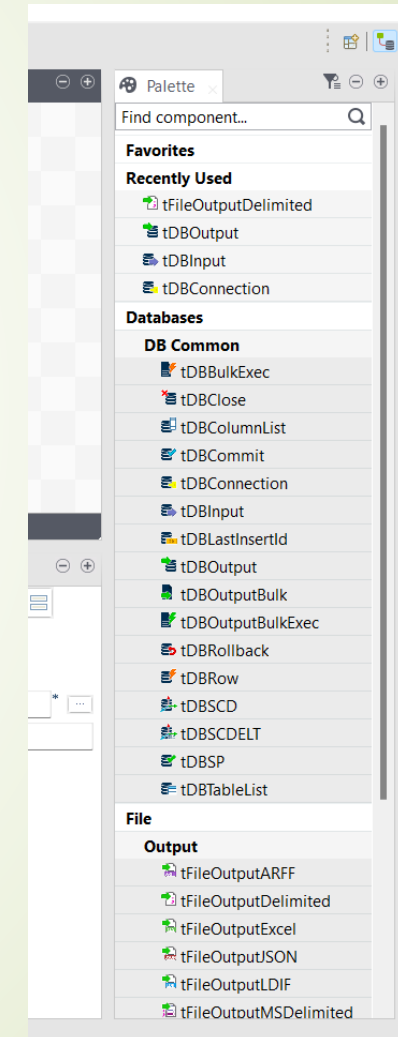
- ❑ Conception plus rapide : un environnement de développement graphique simple et facile à utiliser
- ❑ Meilleure collaboration : communauté active pour améliorer et créer les composants Big Data
- ❑ Portabilité : le TOS peut être utilisé sur n'importe quel OS
- ❑ Scalabilité : l'évolution des jobs Talend est facile.
- ❑ Personnalisable : un développeur peut ajouter des nouveaux composants.

Les composants du TOS For Data Integration



Les composants du TOS For Big Data

- ❑ Une pièce fonctionnelle permettant d'effectuer une seule opération
- ❑ Les composants sont classifiés par besoin fonctionnel ou technique dans la palette
- ❑ C'est un extrait de code java généré lors de l'exécution d'un job
- ❑ Talend fournit plus de 900 composants



Les composants du TOS For Big Data

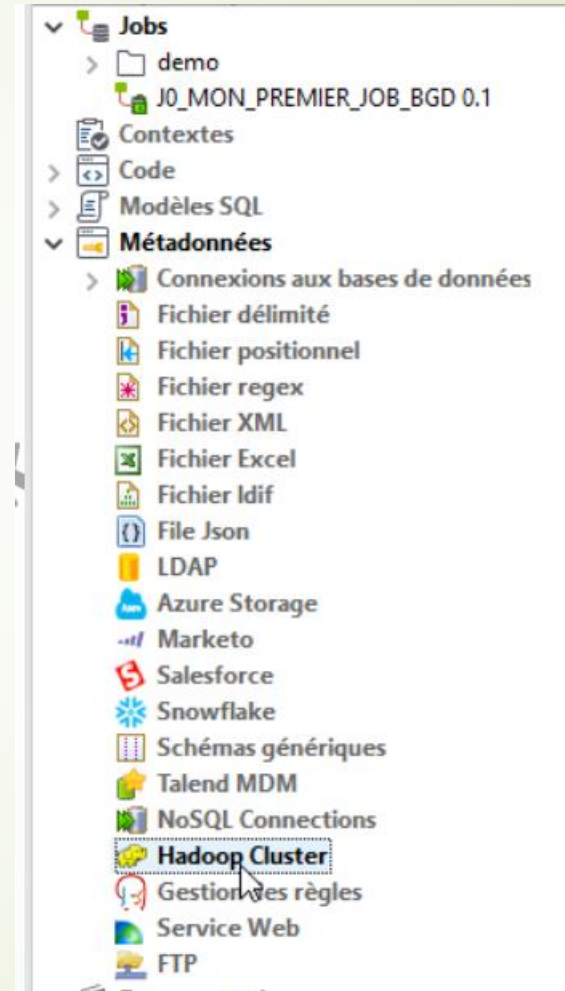
Type de composants	Description	Exemples
Connexion	<ul style="list-style-type: none">✓ Configure et initialise une connexion à un environnement ou à une base de données.✓ Pas de connecteurs d'entrée ou sortie	tHDFSConnection, tHiveConnexion, tHiveClose ...
Flux entrée	<ul style="list-style-type: none">✓ Lire à partir d'un fichier ou une table et met à disposition des lignes des données pour le composant qui succède✓ Pas de connecteur d'entrée	tHDFSInput, tHiveInput, tHbaseInput
Flux en Sortie	<ul style="list-style-type: none">✓ Ecrire le flux dans un fichier ou une table✓ Pas de connecteur de sortie	tHDFSOutput, tHiveOutput, tHbaseOutput ...
Traitement des flux	<ul style="list-style-type: none">✓ Transformation, jointure, agrégation, tri, filtrage ...✓ Connecteurs en entrée et en sortie	tMap, tSortRow, tAggregateRow ...

Les composants du TOS For Data Integration

Type de composants	Description	Exemples
Traitement des fichiers	✓ Lister, copier, supprimer, renommer des fichiers • Existent pour HDFS et Local	tHDFSList, tFileList, tFileCopy, tHDFSCopy ...
Composants Java	✓ Composants personnalisables en écrivant du code java	Java, tJavaRow, tJavaFlex
Orchestration	✓ Orchestration des composants dans un job Talend : lancement d'un job, iteration ...	TPostJob, tPreJob, tRunJob, tLoop ...

Déclaration des Métadonnées

- ✓ Talend fournit des connecteurs permettant d'extraire les métadonnées des fichiers, base de données ...
- ✓ Ces métadonnées sont mise à disposition au développeur dans l'onglet référentiel projet
- ✓ Les métadonnées peuvent être utilisés par plusieurs jobs

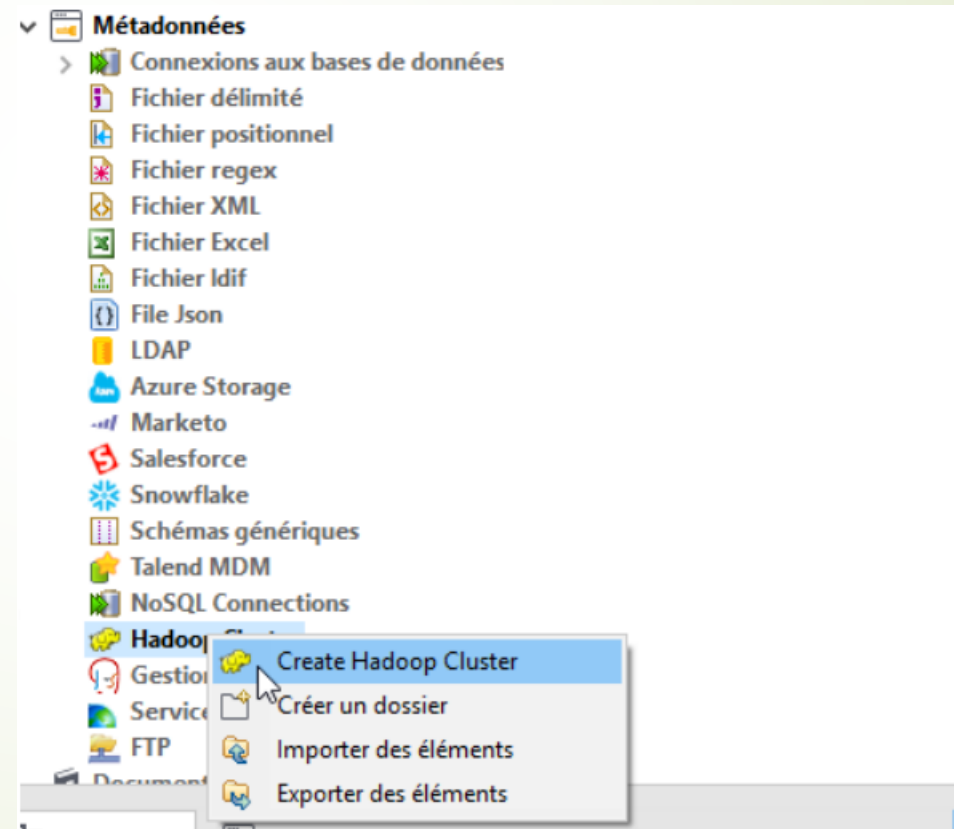




Lab !

Lab1: Prise en main de métadonnées Fichier délimité

Création d'une métadonnée Hadoop Cluster



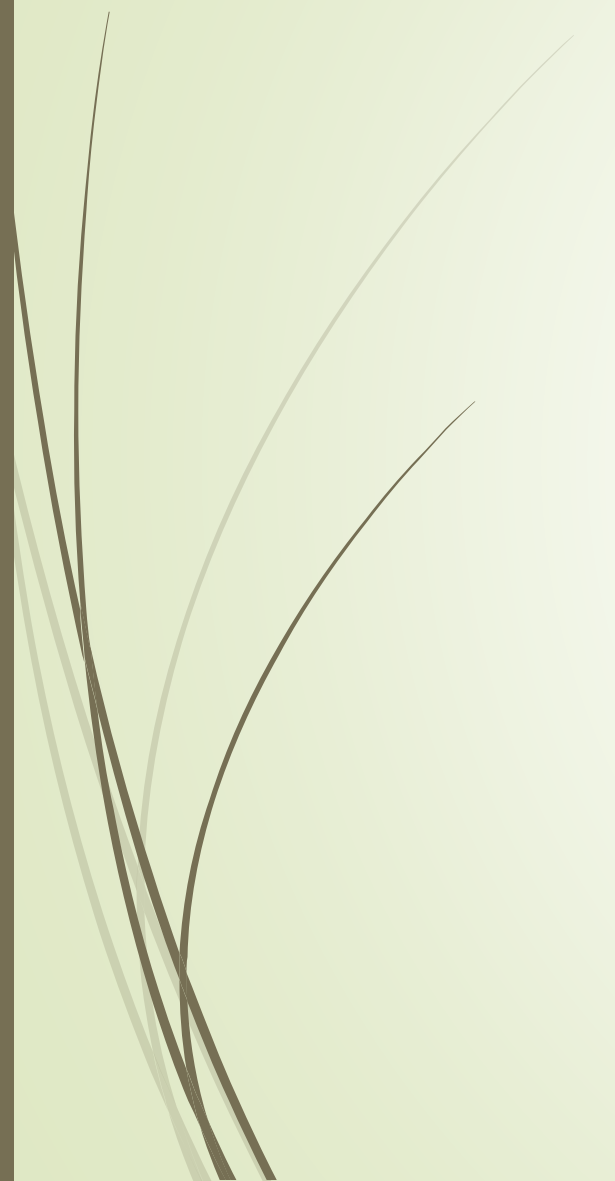


Lab 1: Prise en main de métadonnées Hadoop






Lab1: Prise en main de métadonnées Fichier délimité




Lab1: Prise en main de métadonnées Fichier délimité

 Hadoop Cluster Connection

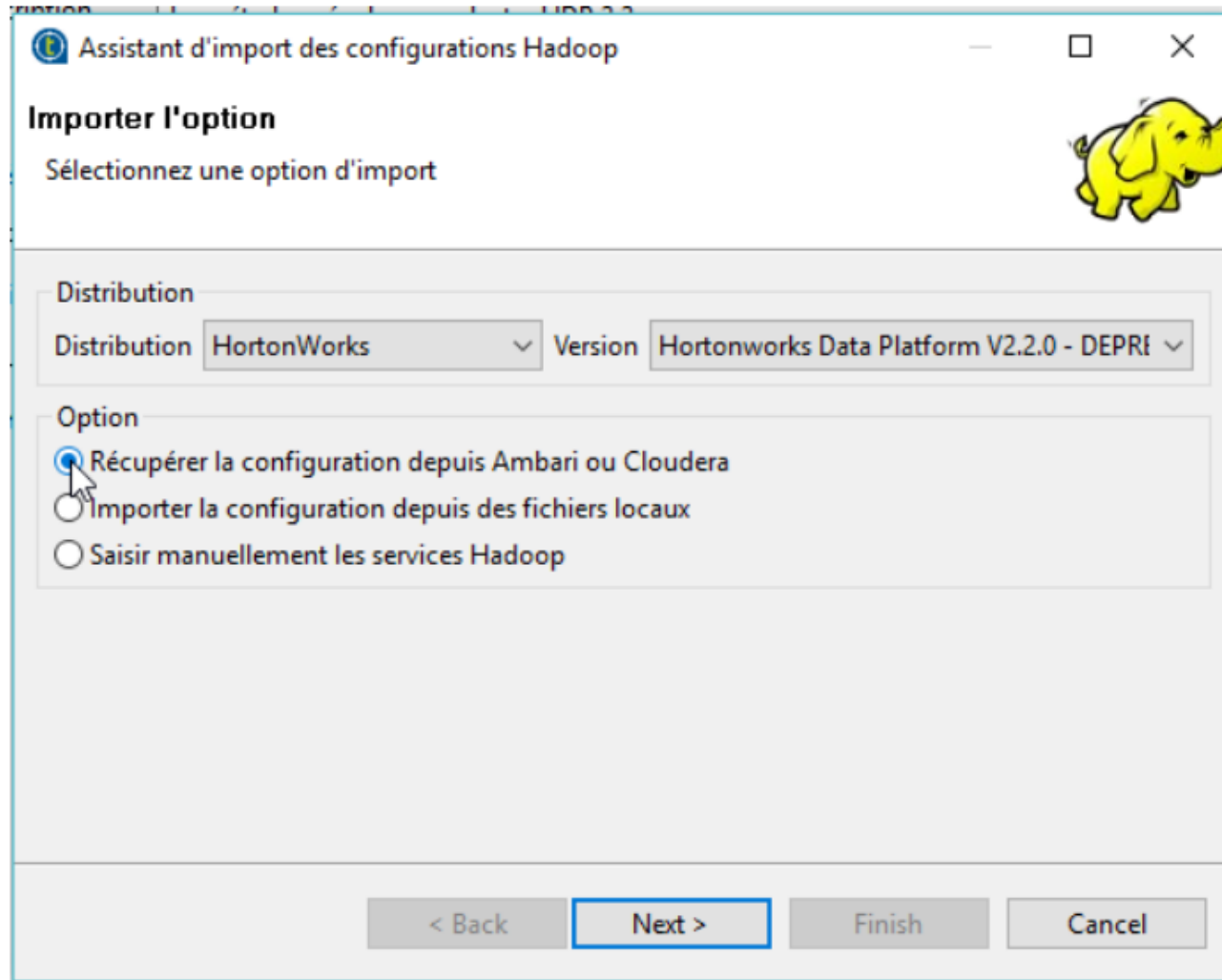
New Hadoop Cluster Connection on repository - Step 1/2

Define the properties



Nom	<input type="text" value="HDP_2_2"/>
Objectif	<input type="text" value="Metadonnee de mon cluster HDP 2.2"/>
Description	<input type="text" value="La métadonnée de mon cluster HDP 2.2"/>
Créé par :	<input type="text" value="user@talend.com"/>
Verrouillé par :	<input type="text"/>
Version	<input type="text" value="0.1"/> <input type="button" value="M"/> <input type="button" value="m"/>
Statut	<input type="text"/>
Chemin d'accès	<input type="text"/> <input type="button" value="Sélectionner"/>

Lab1: Prise en main de métadonnées Fichier délimité



The screenshot shows a window titled "Assistant d'import des configurations Hadoop". The main heading is "Importer l'option" with the instruction "Sélectionnez une option d'import". A yellow cartoon elephant icon is in the top right corner. Under the "Distribution" section, "Distribution" is set to "HortonWorks" and "Version" is set to "Hortonworks Data Platform V2.2.0 - DEPRE". Under the "Option" section, three radio buttons are listed: "Récupérer la configuration depuis Ambari ou Cloudera" (which is selected), "Importer la configuration depuis des fichiers locaux", and "Saisir manuellement les services Hadoop". At the bottom, there are four buttons: "< Back", "Next >" (highlighted with a blue border), "Finish", and "Cancel".

Assistant d'import des configurations Hadoop

Importer l'option

Sélectionnez une option d'import

Distribution

Distribution HortonWorks Version Hortonworks Data Platform V2.2.0 - DEPRE

Option

☒ Récupérer la configuration depuis Ambari ou Cloudera

☐ Importer la configuration depuis des fichiers locaux

☐ Saisir manuellement les services Hadoop

< Back Next > Finish Cancel

Lab1: Prise en main de métadonnées Fichier délimité

Saisissez les informations d'authentification Ambari

URI Ambari (avec port)

Utilisateur

Mot de passe

☐ Customize SSL truststore

TrustStore type

Mot de passe du TrustStore

Fichier TrustStore

Discovered clusters

Sandbox

- ☒ OOOZIE
- ☒ STORM
- ☒ WEBHCAT
- ☒ HDFS
- ☒ HIVE
- ☒ MAPREDUCE2
- ☒ TEZ
- ☒ HBASE
- ☒ YARN

Filtre des propriétés Hadoop (hive.exec.post.hooks;hive.exec.pre.hooks;net.topa...)

Lab1: Prise en main de métadonnées Fichier délimité

Mettre
un user
hadoop

Hadoop Cluster Connection

New Hadoop Cluster Connection on repository - Step 2/2

Define the connection parameters

Version

Distribution HortonWorks Version Hortonworks Data Platform V2.2.0 - DEPRECATED

Connection

Namenode URI hdfs://sandbox.hortonworks.com:8020

Resource Manager sandbox.hortonworks.com:8050

Resource Manager Scheduler sandbox.hortonworks.com:8030

Historique du Job sandbox.hortonworks.com:10020

Répertoire de préparation /user

☒ Utiliser le nom d'hôte du nœud de données

Authentication

☐ Enable kerberos security

Username root

Propriétés Hadoop (Vide) ☒ Utiliser les propriétés Spark (Vide)

☒ Utiliser les configurations Hadoop personnalisées

Vérifier les services

Exporter en tant que contexte Revenir au contexte précédent

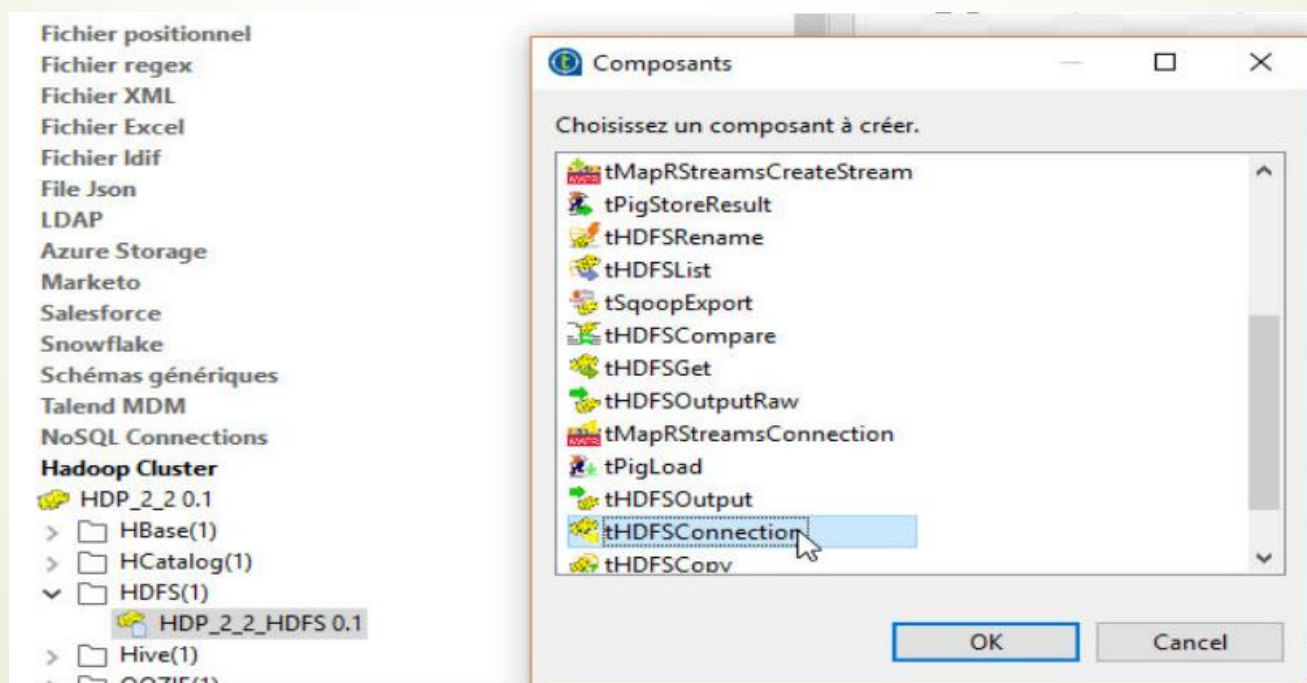
Help < Back Next > Finish Cancel


Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ L'objectif de ce lab est la création d'un job Talend qui dépose des fichiers locaux sur HDFS
- ❑ Télécharger l'archive Data_accounting.zip
- ❑ Dézipper l'archive dans un répertoire local.
- ❑ Ici par exemple, on choisi le répertoire C:\Data_Talend
- ❑ Créer un nouveau job Talend nommé J001_HDFS_PUT_FILES

Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ Avec un drag and drop, créer une connexion HDFS à partir de métadonnées HDFS





Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ Ajouter le composant tHDFSPut Configurer le composant pour déplacer les fichier .csv dans le répertoire HDFS
`/mydata/talendData/csv`



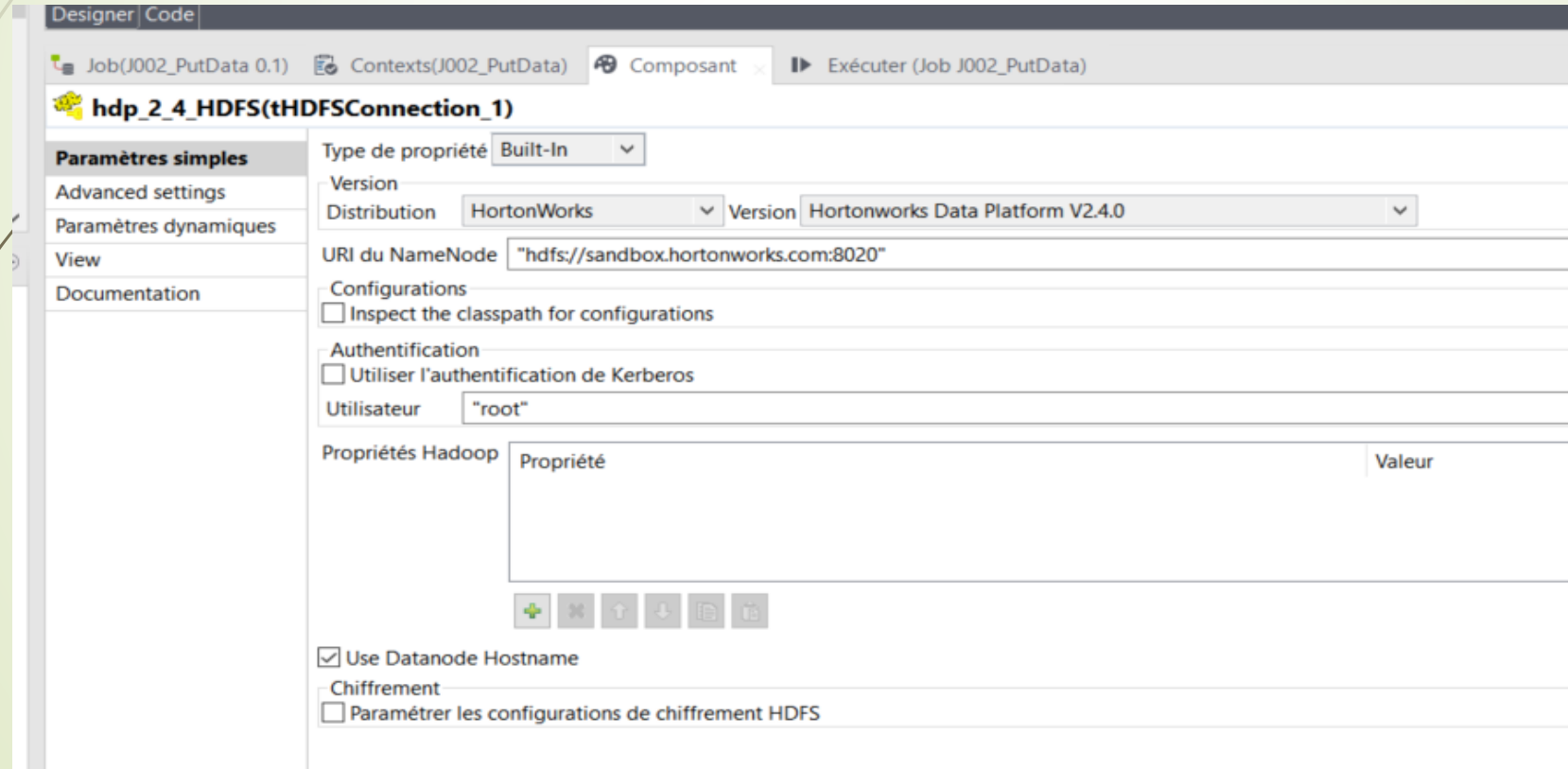
Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ Lancer le job et vérifier le résultat avec la commande

```
hdfs dfs -ls /mydata/talendData/txt
```


Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ Déposer le composant tHDFSConnection



The screenshot shows the Hadoop Designer interface for configuring a component named **hdp_2_4_HDFS(tHDFSConnection_1)**. The interface includes tabs for **Designer**, **Code**, and **Composant**, and a button to **Exécuter (Job J002_PutData)**.

The configuration panel on the right shows the following settings:

- Type de propriété:** Built-In
- Version:** HortonWorks
- Distribution:** HortonWorks
- Version:** Hortonworks Data Platform V2.4.0
- URI du NameNode:** "hdfs://sandbox.hortonworks.com:8020"
- Configurations:**
 - ☐ Inspect the classpath for configurations
- Authentication:**
 - ☐ Utiliser l'authentification de Kerberos
- Utilisateur:** "root"
- Propriétés Hadoop:** A table with columns **Propriété** and **Valeur**.

Below the table, there are icons for adding, deleting, and moving properties. At the bottom, there are checkboxes for **Use Datanode Hostname** (checked) and **Chiffrement** (unchecked).

Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ Déposer le composant tHDFSPut

The screenshot shows the Talend Designer interface with the 'Designer' tab selected. The main workspace displays the configuration for the component 'hdp_2_4_HDFS(tHDFSPut_1)'. The left sidebar contains a tree view with 'Job(J002_PutData 0.1)', 'Contexts(J002_PutData)', and 'Composant'. The top bar shows 'Exécuter (Job J002_PutData)'. The configuration panel on the right includes the following settings:

- Paramètres simples** (selected in the left sidebar):
 - ☒ Utiliser une connexion existante: Liste des composants: tHDFSConnection_1 - hdp_2_4_HDFS *
 - Répertoire local: "C:/Data/Accounting/DC/Comptes/020000"
 - Répertoire HDFS: "/mydata/talendData/txt/"
 - Overwrite file: Toujours ▼
 - ☐ Use Perl5 Regex Expressions as Filemask (Unchecked means Glob Expressions)
- Fichiers**:
 - Masque de fichier: "*.csv"
 - Nouveau nom: ""
- ☒ Arrêter en cas d'erreur

At the bottom of the configuration panel, there are icons for adding, removing, and moving components.

Lab2: Création d'un job qui charge des fichiers sur HDFS

- ❑ Relier les deux composants et relancer votre job

The screenshot displays the Apache Hadoop IDE interface. At the top, a visual workflow shows two components labeled 'hdp_2_4_HDFS' connected by a green line with the label 'ok' and 'OnSubjobOk'. Below this, the 'Designer' tab is active, showing the job configuration for 'Job(J002_PutData 0.1)'. The 'Exécution simple' (Simple Execution) tab is selected in the left sidebar. The main area shows the execution log for 'Job J002_PutData', which includes the following text:

```
Démarrage du job J002_PutData à 08:35 08/08/2022.  
[statistics] connecting to socket on port 3511  
[statistics] connected  
[WARN ]: org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform...  
using builtin-java classes where applicable  
[WARN ]: org.apache.hadoop.hdfs.shortcircuit.DomainSocketFactory - The short-circuit local reads feature cannot  
be used because UNIX Domain sockets are not available on Windows.  
[statistics] disconnected  
Job J002_PutData terminé à 08:35 08/08/2022. [Code sortie=0]
```

Lab3: Lecture d'un fichier HDFS

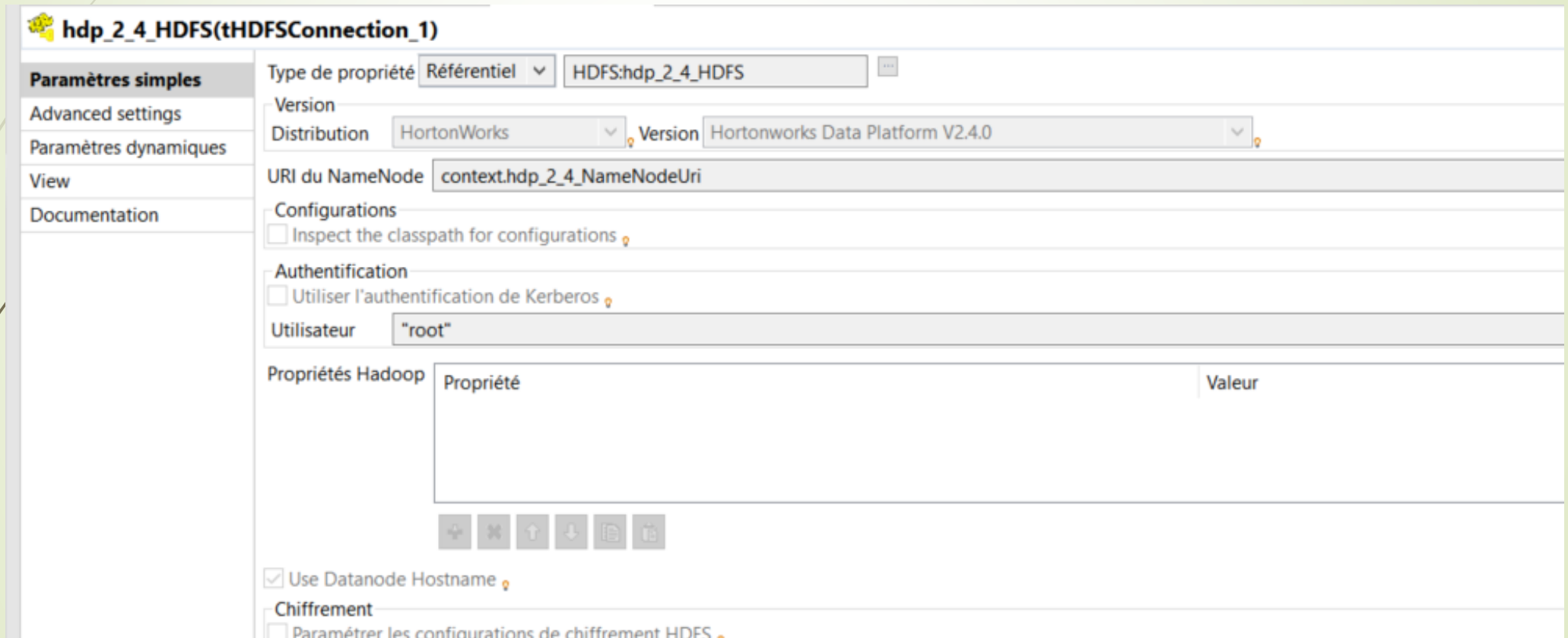
- ❑ Pour lire un fichier HDFS, il faut commencer par identifier ses métadonnées :
- ❑ Format de fichier : texte ou autre
- ❑ Schéma de fichier : les noms de colonnes et leurs types
- ❑ Le séparateur des lignes : « \n »
- ❑ Le séparateur des colonnes : « ; » Nous essayons de lire le fichier
- ❑ Nous essayons de lire le fichier /mydata/Data/txt/Comptes.csv

Lab3: Lecture d'un fichier HDFS

- ❑ Pour lire un fichier HDFS, il faut commencer par identifier ses les métadonnées des fichier.
- ❑ L'objectif de ce lab est de lire le fichier Personnes.txt

Lab3: Lecture d'un fichier HDFS

- ❑ Etape 1 : création une connexion




The screenshot shows the configuration interface for HDP 2.4 HDFS, titled "hdp_2_4_HDFS(tHDFSConnection_1)". The left sidebar contains a menu with the following items: "Paramètres simples" (selected), "Advanced settings", "Paramètres dynamiques", "View", and "Documentation". The main configuration area is divided into several sections:

- Type de propriété:** A dropdown menu set to "Référentiel" and a text field containing "HDFS:hdp_2_4_HDFS".
- Version:** A section with a "Distribution" dropdown set to "HortonWorks" and a "Version" dropdown set to "Hortonworks Data Platform V2.4.0".
- URI du NameNode:** A text field containing "context.hdp_2_4_NameNodeUri".
- Configurations:** A checkbox labeled "Inspect the classpath for configurations" which is currently unchecked.
- Authentification:** A checkbox labeled "Utiliser l'authentification de Kerberos" which is currently unchecked.
- Utilisateur:** A text field containing "root".
- Propriétés Hadoop:** A table with two columns: "Propriété" and "Valeur". The table is currently empty.
- Use Datanode Hostname:** A checkbox which is checked.
- Chiffrement:** A checkbox labeled "Paramétrer les configurations de chiffrement HDFS" which is currently unchecked.

At the bottom of the "Propriétés Hadoop" section, there are six icons: a plus sign, a minus sign, an up arrow, a down arrow, a document icon, and a trash can icon.

Lab3: Lecture d'un fichier HDFS

□ Etape 2 : Configuration du composant tHDFSInput

 **hdp_2_4_HDFS(tHDFSConnection_1)**

Paramètres simples

Advanced settings

Paramètres dynamiques

View

Documentation

Type de propriété: Référéntiel ▼ HDFS:hdp_2_4_HDFS

Version

Distribution: HortonWorks ▼ Version: Hortonworks Data Platform V2.4.0 ▼

URI du NameNode: context.hdp_2_4_NameNodeUri

Configurations

☐ Inspect the classpath for configurations

Authentification

☐ Utiliser l'authentification de Kerberos

Utilisateur: "root"

Propriétés Hadoop

Propriété	Valeur
-----------	--------

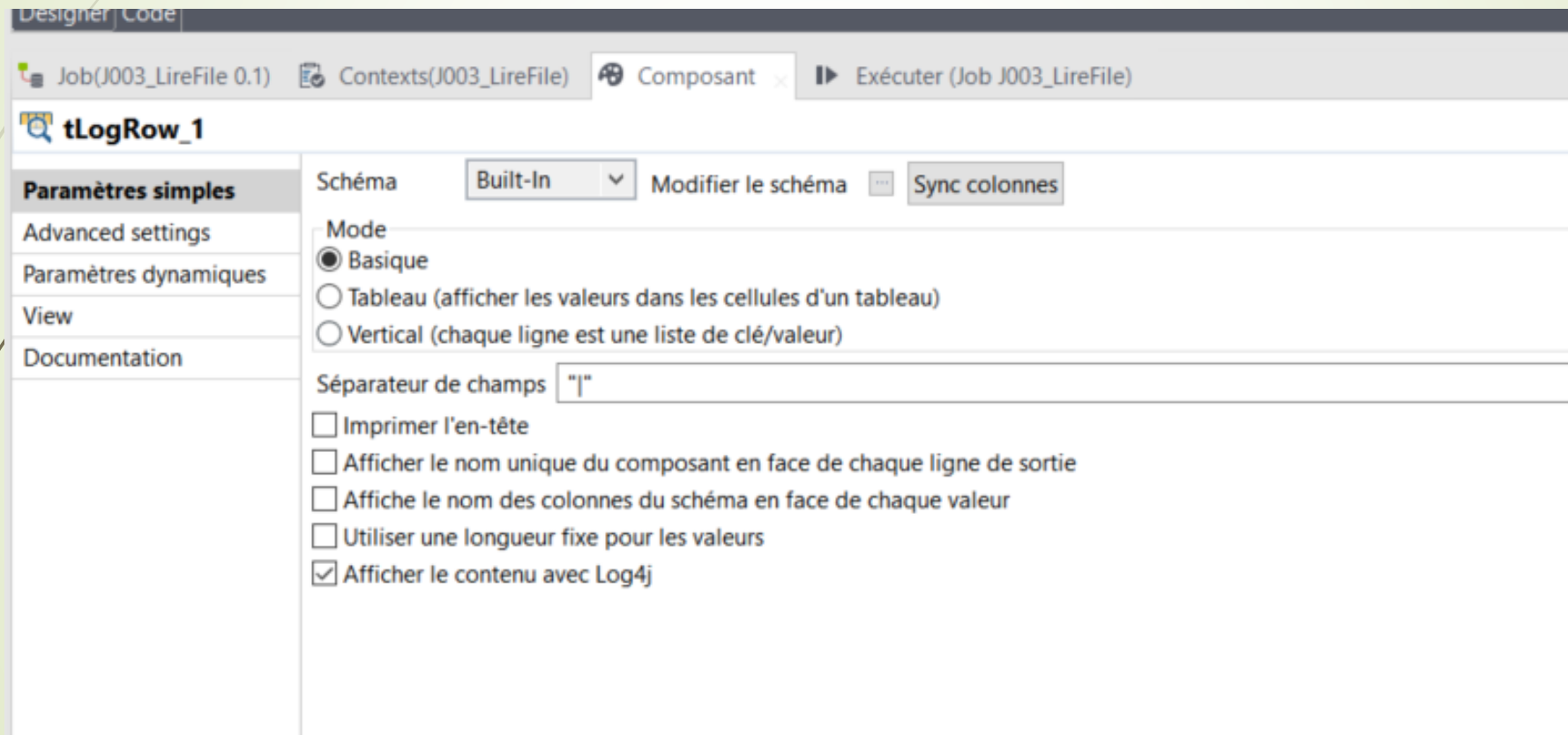
☒ Use Datanode Hostname

Chiffrement

☐ Paramétrer les configurations de chiffrement HDFS

Lab3: Lecture d'un fichier HDFS

- ❑ Etape 3 : Ajouter et configurer le composant tLogRow



Lab3: Lecture d'un fichier HDFS

❑ Etape 4 : Lancer votre job et vérifier le résultat

The screenshot displays a job execution environment. At the top, a workflow diagram shows three components: 'hdp_2_4_HDFS', 'tHDFSInput_1', and 'tLogRow_1'. A green line connects 'hdp_2_4_HDFS' to 'tHDFSInput_1' with the label 'ok' and 'OnSubjobOk'. A red line connects 'tHDFSInput_1' to 'tLogRow_1' with the label '100 rows in 1,45s' and 'row1 (Main)'. Below the diagram, the 'Designer' tab is active, showing the job configuration for 'Job J003_LireFile'. The 'Exécution simple' tab is selected, displaying a list of names in a scrollable area. The list includes names like Grover, Reagan; Jimmy, Garfield; Bill, Madison; George, Harrison; Grover, Adams; Millard, Fillmore; Zachary, Kennedy; Richard, Nixon; Millard, Madison; Rutherford, Tyler; Rutherford, Clinton; Woodrow, Jackson; Millard, Johnson; Theodore, Polk; Woodrow, Quincy; Harry, Ford; John, Roosevelt; Zachary, Harrison; George, Garfield; Millard, Hoover; Andrew, Reagan; James, Johnson; Dwight, Arthur; and Andrew, Grant. The list ends with '[statistics] disconnected' and 'Job J003_LireFile terminé à 08:43 08/08/2022. [Code sortie=0]'. On the right, a 'Default' table shows the configuration for 'hdp_2_4_Name...' and 'hdp_2_4_User'.

Nom	Valeur
hdp_2_4_Name...	hdfs://sandbo...
hdp_2_4_User	root



Attention !

- ✓ Nous avons utilisé les composants tMap, tAgregateRow et tSort pour nettoyer et transformer nos données
- ✓ Ces composants sont couteux en mémoire lorsque la volumétrie est importante
- ✓ Ces composants s'exécutent seulement sur la machine client
- ✓ En Big Data, il est conseillé de remplacer ces composants par Hive ou Pig

Utilisation de Hive avec Talend



Utilisation de Hive avec Talend

- ❑ Talend propose quelques composants Hive pour traiter la données

Nom du composant	Rôle
tHiveConnection	Ouvrir une connexion Hive
tHiveClose	Fermer une connexion Hive après son utilisation
tHiveCreateTable	Création des tables Hive : interne et externe
tHiveInput	Lire une table ou une requête hive
tHiveRow	Lancer une requête Hive
tHiveLoad	Charger le contenu d'un fichier dans une table Hive

Utilisation de Hive avec Talend

- ❑ Les trois composants tELTHiveInput et tELTHiveOutput doivent être utilisés ensemble

Nom du composant	Rôle
tELTHiveMap	Mapper les données entre tELTHiveInput et tELTHiveOutput. Possibilité d'utiliser des fonctions de transformation Hive
tELTHiveInput	Extraire des données à partir d'une table Hive et les passer en entrée d'un tELTHiveMaP
tELTHiveOutput	Charger les données en sortie d'un tELTHiveOutput dans une table Hive

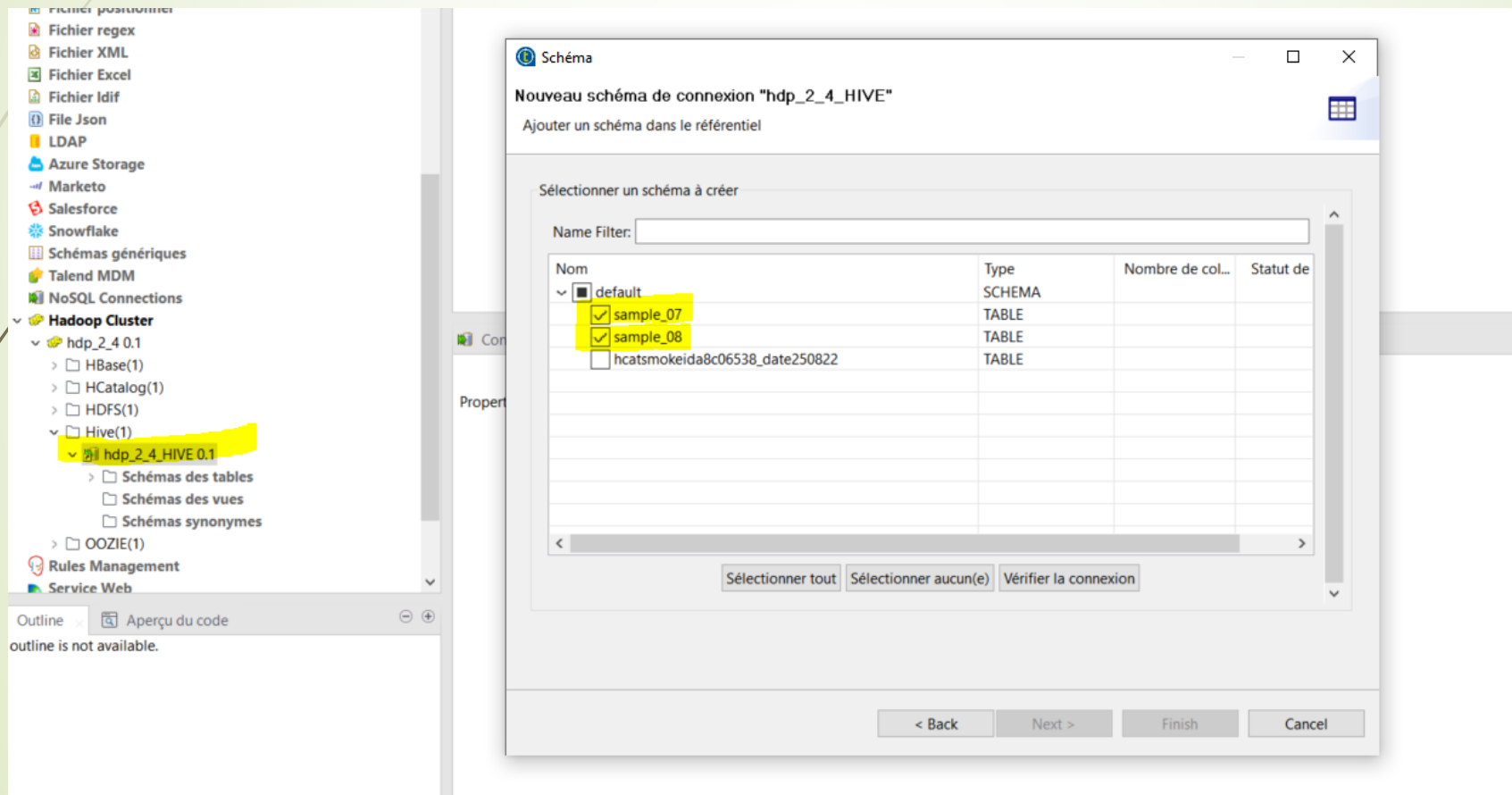


Lab4: Lire une table Hive

- ❑ L'objectif de ce Lab est de lire et d'afficher les données de la table default.sample_07.

Lab4: Lire une table Hive

❑ Etape 1 : définir le schéma de la table



The screenshot shows the Talend Studio interface. On the left, the 'Hadoop Cluster' tree is expanded, showing the 'hdp_2_4_HIVE 0.1' schema selected. The 'Schéma' dialog box is open, titled 'Nouveau schéma de connexion "hdp_2_4_HIVE"'. It contains a table of available schemas:

Nom	Type	Nombre de col...	Statut de
default	SCHEMA		
<input checked="" type="checkbox"/> sample_07	TABLE		
<input checked="" type="checkbox"/> sample_08	TABLE		
<input type="checkbox"/> hcatsmokeida8c06538_date250822	TABLE		

At the bottom of the dialog box, there are buttons: 'Sélectionner tout', 'Sélectionner aucun(e)', 'Vérifier la connexion', '< Back', 'Next >', 'Finish', and 'Cancel'.

Lab4: Lire une table Hive

❑ Etape 2 : Créer un job nommé J001_LireTableHive

Créer la connexion en utilisant le composant tHiveConnexion

The screenshot shows the configuration interface for the **hdp_2_4_HIVE(tHiveConnection_1)** component. The interface is divided into a left sidebar with tabs and a main configuration area on the right.

Left Sidebar Tabs:

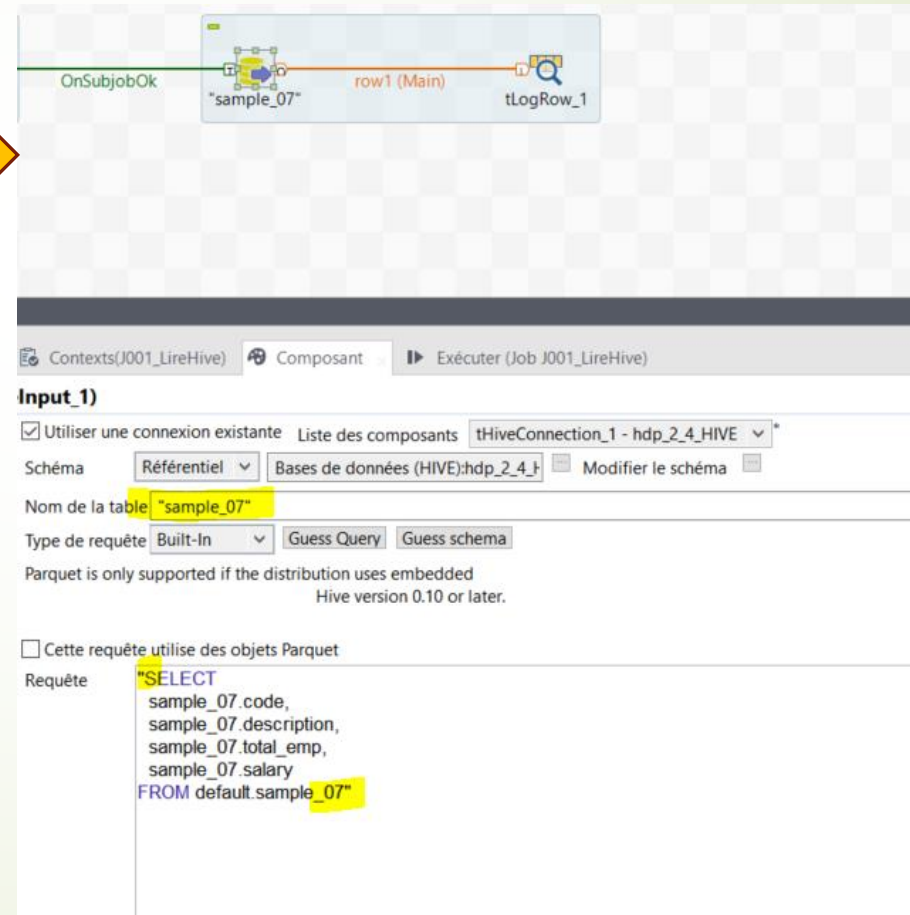
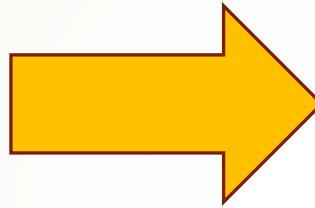
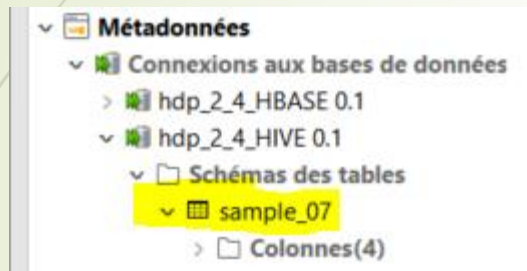
- Paramètres simples (selected)
- Advanced settings
- Paramètres dynamiques
- View
- Documentation

Main Configuration Area:

- Type de propriété:** Built-In
- Version:** HortonWorks (Distribution), Hortonworks Data Platform V2.4.0 (Version)
- Connexion:**
 - Mode de connexion:** Standalone
 - Serveur de Hive:** Hive 2
 - Hôte:** 192.168.56.101
 - Port:** 10000
 - Database:** default
 - Utilisateur:** root
 - Mot de passe:** ****
- Paramètres JDBC supplémentaires:** ""
- Configurations:**
 - ☐ Inspecter le chemin de classe pour les configurations
- Authentification:**
 - ☐ Utiliser l'authentification Kerberos
- Cryptage:**
 - ☐ Utiliser l'encodage SSL
- Propriétés Hadoop:**
 - ☐ Set Resource Manager
 - ☐ Configurer l'URI du NameNode
 - ☐ Set resourcemanager scheduler address
 - ☐ Set jobhistory address
 - ☐ Configurer l'utilisateur Hadoop
 - ☐ Use datanode hostname
- ☐ Utiliser ou enregistrer une connexion partagée à une base de données
- ☐ Configuration de Hbase

Lab4: Lire une table Hive

Etape 3 : faire un drag and drop de la table la table Sample_07



Lab4: Lire une table Hive

Etape 4 : Ajouter le composant tLogRow et relier les trois composants
Exécuter votre job.

The screenshot displays the Talend Studio interface. At the top, a job diagram shows three components: 'hdp_2_4_HIVE', 'sample_07', and 'tLogRow_1'. The 'sample_07' component is connected to 'tLogRow_1' with a label '823 rows in 2.46s'. Below the diagram, the 'Job J001_LireHive' is selected. The 'Exécution' (Execution) tab is active, showing a list of data rows. The first few rows are:

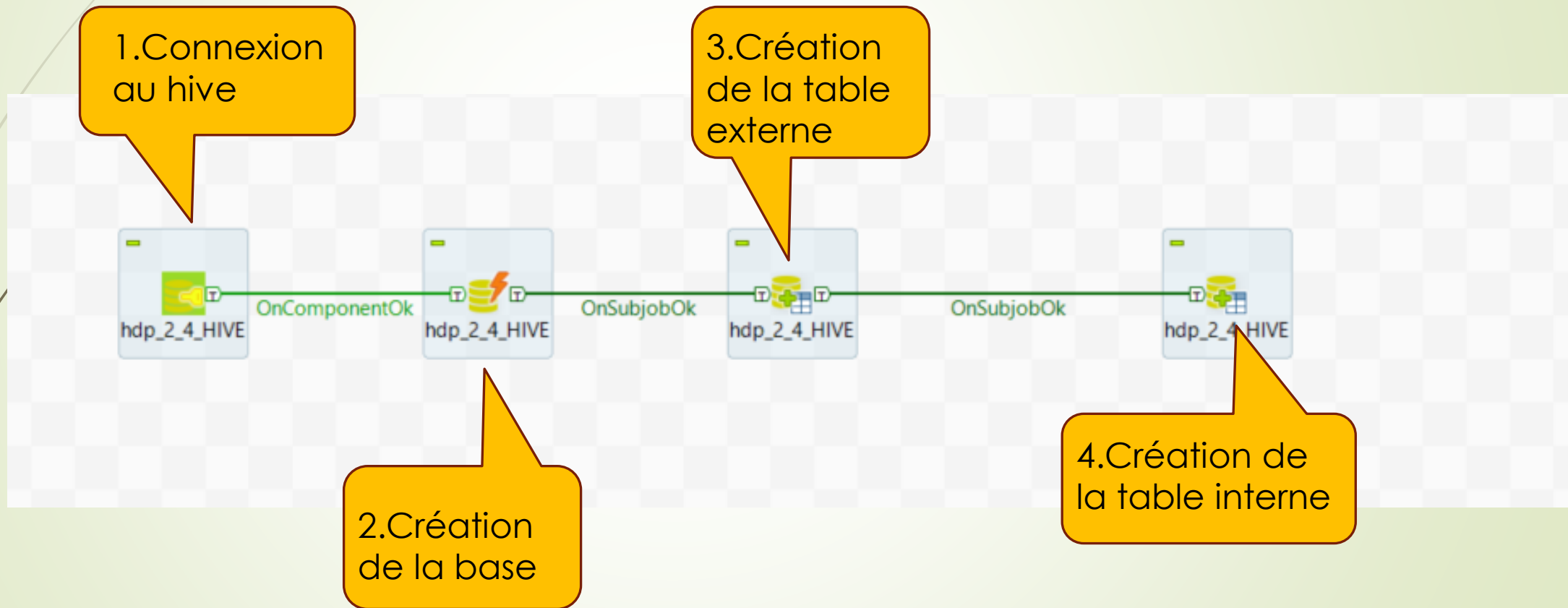
Nom	Valeur
53-4011	Locomotive engineers 41760 63180
53-4012	Locomotive firers 580 49660
53-4013	Rail yard engineers, dinkey operators, and hostlers 4950 40510
53-4021	Railroad brake, signal, and switch operators 23120 54530
53-4031	Railroad conductors and yardmasters 37540 61480
53-4041	Subway and streetcar operators 6600 47740
53-4099	Rail transportation workers, all other 5210 38730
53-5011	Sailors and marine oilers 32520 34050
53-5021	Captains, mates, and pilots of water vessels 30540 62720
53-5022	Motorboat operators 3250 36570
53-5031	Ship engineers 13710 61680
53-6011	Bridge and lock tenders 4750 38120
53-6021	Parking lot attendants 131860 19320
53-6031	Service station attendants 93140 19720
53-6041	Traffic technicians 6550 40150
53-6051	Transportation inspectors 24130 57050
53-6099	Transportation workers, all other 46720 34330
53-7011	Conveyor operators and tenders 45580 29020
53-7021	Crane and tower operators 45720 42940
53-7031	Dredge operators 1910 38320
53-7032	Excavating and loading machine and dragline operators 68040 36990
53-7033	Loading machine operators, underground mining 2770 41980
53-7041	Hoist and winch operators 3220 39190
53-7051	Industrial truck and tractor operators 630700 29760
53-7061	Cleaners of vehicles and equipment 336210 20870
53-7062	Laborers and freight, stock, and material movers, hand 2363440 23840
53-7063	Machine feeders and offbearers 143140 25260
53-7064	Packers and packagers, hand 798450 20320
53-7071	Gas compressor and gas pumping station operators 4230 44590
53-7072	Pump operators, except wellhead pumpers 10400 40660
53-7073	Wellhead pumpers 15780 37680
53-7081	Refuse and recyclable material collectors 126270 31410
53-7111	Shuttle car operators 2660 41300
53-7121	Tank car, truck, and ship loaders 14870 35820
53-7199	Material moving workers, all other 43840 33170

The job execution status is 'Job J001_LireHive terminé à 09:59 08-08-2022. [Code sortie=0]'. The 'Exécution' tab also shows a table with 'Nom' and 'Valeur' columns.

Lab5: Lire une table Hive

- ❑ Créer un nouveau Job J3_HIVE_INIT
- ❑ Ajouter une connexion Hive dans ce Job
- ❑ Ajouter un composant tHiveRow pour créer la base « talend_hive_db »
- ❑ Ajouter le composant tHiveCreateTable Configurer ce composant pour créer la table « ext_comptes » : table externe qui pointe vers le répertoire HDFS
"/mydata/talendData/txt »
- ❑ Ajouter un composant tHiveCreateTable pour créer une table interne Comptes
(numero String, Type String)

Lab5: Solution



Lab5: Solution

Job(J002_HIVE_INIT 0.1) Contexts(J002_HIVE_INIT) Composant x Exécuter (Job J002_HIVE_INIT)

hdp_2_4_HIVE(tHiveConnection_1)

Paramètres simples
Advanced settings
Paramètres dynamiques
View
Documentation

Type de propriété Built-In

Version

Distribution HortonWorks Version Hortonworks Data Platform V2.4.0

Connexion

Mode de connexion Standalone * Serveur de Hive Hive 2 *

Hôte "sandbox.hortonworks.com" Port "10000"

Database "default"

Utilisateur "root" * Mot de passe ****

Paramètres JDBC supplémentaires ""

Configurations

☐ Inspecter le chemin de classe pour les configurations

Authentification

☐ Utiliser l'authentification Kerberos

Cryptage

☐ Utiliser l'encodage SSL

Propriétés Hadoop

☐ Set Resource Manager

☐ Configurer l'URI du NameNode

☐ Set resourcemanager scheduler address

☐ Set jobhistory address

☐ Configurer l'utilisateur Hadoop

☐ Use datanode hostname

☐ Utiliser ou enregistrer une connexion partagée à une base de données

Configuration de Hbase

☐ Stocké par HBase

Lab5: Solution

Job(J002_HIVE_INIT 0.1) Contexts(J002_HIVE_INIT) Composant x Exécuter (Job J002_HIVE_INIT)

hdp_2_4_HIVE(tHiveRow_1)

Paramètres simples

☒ Utiliser une connexion existante Liste des composants tHiveConnection_1 - hdp_2_4_HIVE *

Advanced settings Schéma Built-In v Modifier le schéma ... Nom de la table ""

Paramètres dynamiques Type de requête Built-In v Guess Query

View Parquet is only supported if the distribution uses embedded Hive version 0.10 or later.

Documentation ☐ Cette requête utilise des objets Parquet

Requête "CREATE DATABASE IF NOT EXISTS TALEND_HIVE_DB"

Lab5: Solution

Job(J002_HIVE_INIT 0.1) Contexts(J002_HIVE_INIT) Composant x Exécuter (Job J002_HIVE_INIT)

hdp_2_4_HIVE(tHiveCreateTable_1)

Paramètres simples

Advanced settings

Paramètres dynamiques

View

Documentation

☒ Utiliser une connexion existante

Liste des composants tHiveConnection_1 - hdp_2_4_HIVE

Schéma Built-In Modifier le schéma

Créer la table

Nom de la table "TALEND_HIVE_DB.ext_comptes"

Action sur la table Create table if not exists

Format TEXTFILE

☐ Configurer les partitions

☒ Configurer l'emplacement du fichier ☐ Use S3 endpoint (External table)

URI path "hdfs://sandbox.hortonworks.com:8020/mydata/talendData/txt"

Format de ligne

☒ Set delimited row format

☒ Champ ";" ☐ Echappement

☐ Élément de collection

☐ Map Key

☐ Ligne

☐ Arrêter en cas d'erreur

Lab5: Solution

Job(J002_HIVE_INIT 0.1) Contexts(J002_HIVE_INIT) Composant Exécuter (Job J002_HIVE_INIT)

hdp_2_4_HIVE(tHiveCreateTable_2)

Paramètres simples

Advanced settings

Paramètres dynamiques

View

Documentation

☒ Utiliser une connexion existante

Liste des composants: tHiveConnection_1 - hdp_2_4_HIVE

Schéma: Built-In Modifier le schéma

Créer la table

Nom de la table: "comptes"

Action sur la table: Create table if not exists

Format: TEXTFILE

☐ Configurer les partitions

☐ Configurer l'emplacement du fichier

Format de ligne

☒ Set delimited row format

☒ Champ: ";" ☐ Echappement

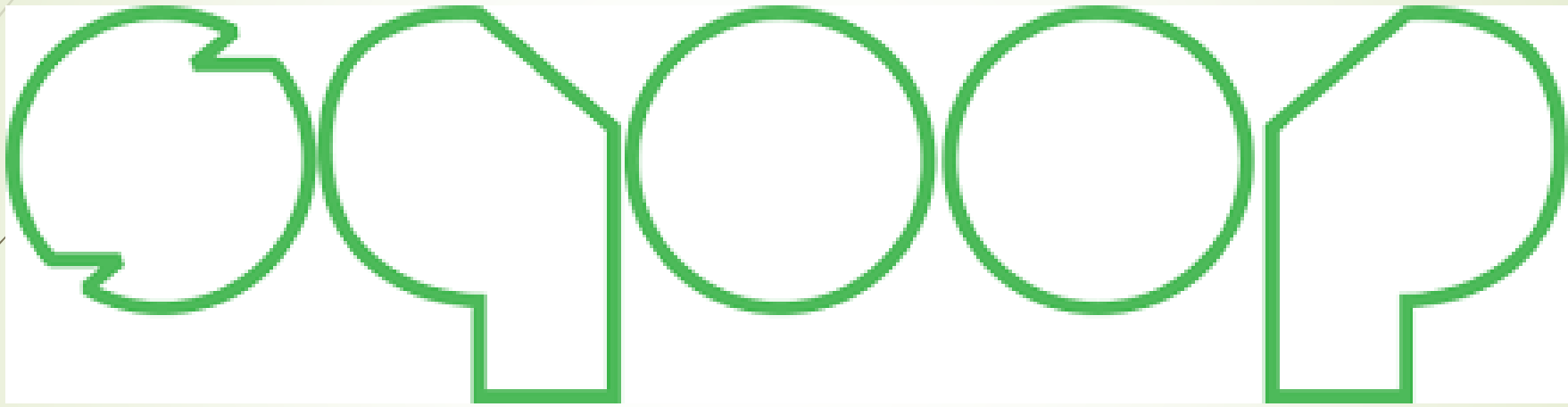
☐ Élément de collection

☐ Map Key

☐ Ligne

☐ Arrêter en cas d'erreur

Apache Sqoop





Apache Sqoop

- ❑ Sqoop est un outil conçu pour transférer des données entre Hadoop et des entrepôts de données structurés externes tels que les SGBDR et les Data Warehouses.
- ❑ Sqoop permet un échange de données JDBC entrant et sortant avec Hadoop et son écosystème.
- ❑ La source de données fournit le schéma et Sqoop génère et exécute des instructions SQL à l'aide de JDBC ou d'autres connecteurs.

Les commandes Sqoop



```
C:\_ 
```



Les commandes Sqoop

- ❑ Sqoop fournit une interface en ligne de commande.
- ❑ Dans la commande Sqoop il faut simplement fournir des informations de base telles que :
 - L'adresse de la source.
 - Les détails d'authentification de la source.
 - La destination.
- ❑ Sqoop prendra en charge la partie restante !

La commande import

- ❑ La syntaxe de la commande sqoop import est la suivante :

`sqoop import(generic-args) (import-args)`

les arguments generic-args doivent précéder les arguments imports-args.

Les arguments import-args peuvent être entrés dans n'importe quel ordre.

La syntaxe de la commande sqoop export est la suivante :

`sqoop import(generic-args)(export-args)`



La commande export

- ❑ La commande sqoop export exporte un ensemble de fichiers de HDFS vers des tables SGBDR.
- ❑ La table cible devrait déjà exister dans la base de données.
- ❑ La commande sqoop export prépare les requêtes INSERT avec un ensemble de données d'entrée, puis les jouent sur la base de données.




sqoop export(generic-args)(export-args)

Lab – prise en main de sqoop

```
sandbox login: root
root@sandbox.hortonworks.com's password:
Last login: Mon Aug  8 15:59:44 2022 from 192.168.56.101
[root@sandbox ~]#
[root@sandbox ~]# su sqoop
[sqoop@sandbox root]$
```

Lab – prise en main de sqoop

- ❑ Déposer la base sales_database.sql au répertoire suivant :

/home/sqoop/					
Nom	Taille	Date de modification	Droits	Propriét...	
 		14/03/2016 15:49:37	rw-r-xr-x	root	
 sales_database.sql	205 KB	18/11/2018 14:49:36	rw-r--r--	root	

Lab – prise en main de sqoop

- ❑ Création de la base et l'utilisateur de la base;

```
[sqoop@sandbox ~]$ mysql -t < sales_database.sql
[sqoop@sandbox ~]$ mysql
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 501
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use Abortedb;
[sqoop@sandbox ~]$ mysql
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 502
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use sales_db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> create user sqoop_dba identified by 'sqoopdba';
ERROR 1396 (HY000): Operation CREATE USER failed for 'sqoop_dba'@'%'
mysql> create user sales_dba identified by 'salesdb';
Query OK, 0 rows affected (0.00 sec)
```

Lab – prise en main de sqoop

- ❑ Importer la table customers de la base sales_db:

The screenshot shows the configuration interface for a Sqoop job named "tSqoopImport_1". The interface is divided into several sections:

- Mode:** ☐ Use Commandline, ☒ Utiliser API Java
- Propriété Hadoop:** Référentiel, HCdp_2_4
- Version:** Distribution: HortonWorks, Version: Hortonworks Data Platform V2.4.0
- Configuration:**
 - URI du NameNode: context.hdp_2_4_NameNodeUri
 - Gestionnaire de ressources: context.hdp_2_4_ResourceManager
 - ☒ Set resourcemanager scheduler address: context.hdp_2_4_ResourceManagerScheduler
 - ☒ Set jobhistory address: context.hdp_2_4_JobHistory
 - ☒ Set staging directory: context.hdp_2_4_StagingDirectory
 - ☒ Use Datanode Hostname
- Authentification:**
 - ☐ Utiliser l'authentification Kerberos
 - Utilisateur Hadoop: context.hdp_2_4_User
- Propriété JDBC:** Built-In
- Arguments communs:**
 - Connexion: jdbc:mysql://192.168.56.101:3306/sales_db?noDatetimeStringSync=true
 - Utilisateur: context.sales_db_Login
 - ☐ Le mot de passe est stocké dans un fichier
 - Mot de passe: context.sales_db_Password
 - Jar du pilote: Nom du Jar, mysql-connector-java-5.1.30-bin.jar
 - Nom de classe: com.mysql.jdbc.Driver
- Importer les arguments de contrôle:**
 - Nom de la table: customers
 - Format du fichier: Fichier Parquet
 - ☒ Supprimer le répertoire cible
 - ☐ Ecrire après
 - ☐ Compresser
 - ☐ Direct
 - ☐ Sélectionner les colonnes





”

