

New

Intégration de données avec Talend Data Fabric

Concevoir, transformer, déployer et superviser des pipelines d'intégration de données avec Talend Data Fabric

 Présentiel ou en classe à distance



3 jours (21 h)

Prix inter : 2.490,00 € HT
Forfait intra : 7.690,00 € HT

Réf.: BD525

La formation **Talend Data Fabric** permet de maîtriser une plateforme de référence pour l'**intégration de données** et la conception de **pipelines ETL**. Elle couvre la création de flux avec **Talend Studio**, les transformations de données, le **Data Quality**, l'orchestration des traitements et le déploiement dans **Talend Cloud**. Une formation idéale pour concevoir, automatiser et superviser des pipelines de données fiables, évolutifs et adaptés aux projets de **data engineering** en entreprise.

A qui s'adresse cette formation ?



Pour qui

- Intégrateurs, data engineers



Prérequis

- Connaissances en SQL et ETL de base
- Disposez-vous des connaissances nécessaires pour suivre cette formation ? Testez-vous !

Programme

1 - Présentation de Talend Data Fabric et de son architecture

- Positionnement de Talend Data Fabric, édité par Qlik, et présentation des composants de la suite : Talend Studio, Talend Cloud, Talend Management Console (TMC) et Data Quality
- Architecture client-serveur : Studio installé sur le poste du développeur, référentiel de projets et environnement d'exécution avec Cloud Engine ou Remote Engine
- Modèle de génération de code Java : transformation d'un job graphique en fichier JAR exécutable
- Principaux cas d'usage : alimentation d'un data warehouse ou d'un data lake, migration de données et synchronisation entre systèmes

Lancer Talend Studio et se connecter au workspace de formation

Créer un projet et explorer les perspectives Data Integration et Data Quality

Identifier les principales sections du référentiel : Job Designs, Metadata, Contexts, Routines et Documentation

Importer un projet de démonstration et exécuter un job d'exemple afin d'en observer les résultats dans la console

Configurer les paramètres de Talend Studio : mémoire JVM, dépôts d'artifacts et niveau de journalisation

2 - Composants Talend et création d'un premier job ETL

- Comprendre les notions de job, de sous-job, de master job et de joblet réutilisable
- Découvrir les principales familles de composants Talend : File, Database, Processing, Logs et Errors, Orchestration et Cloud

- Utiliser les différents liens entre les composants : Main, Lookup, Reject, Output, Iterate, OnSubjobOK, OnSubjobError et Run If
- Définir les schémas d'entrée et de sortie et comprendre leur propagation entre les composants
- Maîtriser le cycle de développement d'un job Talend : conception, aperçu des données, exécution, débogage et export
- Mettre en place la journalisation et la gestion des logs avec tLogRow, tLogCatcher, tStatCatcher, tDie, tWarn et tFlowMeter

3 - Connecter Talend aux différentes sources de données

- Centraliser les connexions, les schémas et les requêtes à l'aide des métadonnées partagées
- Connecter Talend aux bases de données relationnelles MySQL, PostgreSQL, Oracle, SQL Server et DB2 avec les composants génériques tDBInput, tDBOutput et tDBRow
- Lire et écrire différents formats de fichiers avec tFileInputDelimited, tFileInputJSON, tFileInputXML, tFileInputExcel et tFileOutputDelimited, ainsi qu'au format Parquet
- Connecter Talend à des API REST avec tRESTClient et tRESTRequest et gérer l'authentification Basic, par clé API ou OAuth 2
- Intégrer des données issues des services cloud Amazon S3, Azure Blob Storage, Azure Data Lake Storage, Snowflake, BigQuery et Databricks Delta Lake
- Appliquer les bonnes pratiques de connexion : pooling et gestion sécurisée des credentials par contexte

Créer des métadonnées partagées pour une base PostgreSQL source et une base de données cible

Configurer un composant tFileInputDelimited afin de lire un fichier CSV contenant un référentiel de produits

Interroger une API REST avec tRESTClient et analyser la réponse JSON avec tExtractJSONFields

Joindre les trois sources à l'aide d'un composant adapté et écrire les données consolidées dans la base cible

Tester le flux ETL sur un échantillon, consulter les logs et corriger les éventuelles erreurs de schéma ou de typage

4 - Transformations avancées avec tMap

- Présentation du composant tMap, élément central des transformations de données dans Talend
- Création de mappings et de conversions : copie directe, expressions Java et conversion des types
- Mise en oeuvre de jointures dans tMap : inner join, left outer join, all matches, first match et unique match
- Utilisation des variables tMap et des sorties multiples conditionnelles pour orienter les lignes selon des règles métier
- Présentation des composants de traitement : tFilterRow, tReplace, tConvertType, tExtractRegexFields et tSplitRow
- Agrégation, tri et déduplication avec tAggregateRow, tSortRow et tUniqRow
- Restructuration des données avec tNormalize, tDenormalize et tPivot

Mettre en oeuvre un tMap avec une jointure left outer entre une table de ventes et une table de produits

Définir une variable tMap permettant de calculer un taux de marge à partir de plusieurs colonnes

Produire deux sorties conditionnelles distinguant les ventes rentables des ventes à perte

Agréger les ventes par catégorie de produits avec tAggregateRow afin de calculer le chiffre d'affaires et la marge moyenne

Trier les résultats par chiffre d'affaires décroissant

Écrire les données obtenues dans deux fichiers Parquet distincts

5 - Data Quality et nettoyage des données

- Présentation des dimensions de la qualité des données : complétude, validité, cohérence, unicité et conformité
- Profiling des données dans le module Data Quality : counts, valeurs distinctes, distributions, analyse de colonnes, patterns et tables
- Validation des données et application de règles avec tSchemaComplianceCheck, tIntervalMatch et tPatternCheck
- Standardisation des données avec tStandardizeRow, tReplaceList et tNormalize
- Détection des doublons avec tFuzzyMatch et tMatchGroup
- Présentation des stratégies de matching : distance de Levenshtein, méthode Jaro-Winkler et Soundex
- Gestion des rejets et reporting sur la qualité des données : isolement des lignes invalides avec les composants Reject et production de rapports avant et après traitement

Réaliser une analyse de profiling sur une table client : complétude des champs, patterns des adresses e-mail et des numéros de téléphone, et nombre de valeurs distinctes par pays

Identifier les anomalies telles que les adresses e-mail invalides, les numéros de téléphone mal formatés et les doublons potentiels

Concevoir un job avec tPatternCheck afin de valider les adresses e-mail à l'aide d'une expression régulière et de diriger les données invalides vers les rejets

Standardiser les numéros de téléphone et les codes pays avec tStandardizeRow

Détecter les doublons avec tMatchGroup et orienter les groupes vers un fichier d'arbitrage

6 - Conception avancée et automatisation des flux Talend

- Utilisation des contextes Talend pour paramétrer les jobs : variables d'environnement, credentials, chemins d'accès et URL
 - Création de groupes de contextes DEV, TEST et PROD et chargement dynamique avec tContextLoad
 - Conception de sous-jobs et de master jobs avec tRunJob, passage de paramètres entre jobs parents et enfants et création d'un job d'orchestration
 - Centralisation de la gestion des erreurs avec tDie, tWarn et tLogCatcher
 - Pilotage conditionnel des traitements avec les triggers OnSubjobOK, OnSubjobError, OnComponentOK et Run If
 - Itération sur des fichiers ou des enregistrements avec tFileList, tFlowTolterate et tIterateToFlow
- Atelier

Définir un groupe de contextes DEV, TEST et PROD contenant les credentials de la base, les chemins des fichiers et les URL des API

Créer un master job appelant trois sous-jobs avec tRunJob et transmission des paramètres

Gérer les erreurs avec OnSubjobError et centraliser les logs avec tLogCatcher

Externaliser les credentials dans un fichier .properties chargé avec tContextLoad

Mesurer les performances de chaque étape avec tStatCatcher et tFlowMeter, ou à l'aide d'un gestionnaire de secrets tel que HashiCorp Vault ou AWS Secrets Manager

Exécuter le master job successivement dans les trois contextes afin de comparer les comportements

7 - Déploiement et supervision avec Talend Cloud et TMC

- Présentation de Qlik Talend Management Console (TMC), portail central d'administration de Talend Cloud
 - Organisation des projets, environnements DEV, TEST et PROD, et workspaces
 - Présentation des moteurs d'exécution : Cloud Engine, exécuté dans l'environnement Talend, et Remote Engine, exécuté dans l'environnement du client
 - Publication d'un job avec Build for Cloud et Cloud Pipelines
 - Création de tâches pour encapsuler un job, paramétrer les contextes et sélectionner le moteur d'exécution
 - Planification et orchestration des traitements : déclencheurs CRON, déclencheurs sur fichier et plans d'enchaînement de tâches
 - Monitoring et alerting : tableau de bord des exécutions, logs, durées et notifications par e-mail ou webhook en cas d'échec
- Atelier

Publier le master job dans Talend Cloud avec Build for Cloud, puis effectuer un Push vers TMC

Créer une tâche dans TMC, l'associer à un Remote Engine et configurer les contextes d'exécution

Planifier une exécution quotidienne avec un déclencheur CRON

Configurer une alerte par e-mail et un webhook en cas d'échec du traitement

Déclencher manuellement la tâche, consulter les logs et les métriques d'exécution dans TMC

Mettre en place un plan reliant deux tâches avec une dépendance d'enchaînement conditionnelle



Les objectifs de la formation

- Concevoir et automatiser des flux ETL avec Talend Data Integration
- Transformer, enrichir et nettoyer des données à l'aide de Talend
- Déployer, exécuter et superviser des jobs Talend
- Connecter Talend à différentes sources de données et applications



Evaluation

- Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.



Les points forts de la formation

- Une pédagogie orientée pratique s'appuyant sur la conception, le déploiement et la supervision de pipelines Talend inspirés de cas d'usage réels, tels que l'intégration de données multi-sources, la qualité des données et l'orchestration des traitements
- Une approche complète du cycle de vie ETL permettant de maîtriser toutes les étapes d'un projet, de la conception des flux dans Talend Studio jusqu'à leur déploiement sur Talend Cloud, avec le suivi de l'exécution, le monitoring et la gestion des alertes
- Une formation basée sur les fonctionnalités actuelles de Talend Data Fabric intégrant les principaux composants de la plateforme, notamment Talend Studio, Qlik Talend Management Console et Remote Engine, afin d'acquérir des compétences directement applicables en entreprise.



Dates et villes 2026 - Référence BD525



Dernières places disponibles



Session garantie

A distance

du 5 oct. au 7 oct.

du 23 nov. au 25 nov.

Paris

du 5 oct. au 7 oct.

du 23 nov. au 25 nov.