

Mis à jour le 18/07/2025

S'inscrire

# Formation Pentaho ETL Fundamentals

3 jours (21 heures)

## Présentation

Maîtrisez Pentaho Data Integration (PDI) dans sa globalité grâce à cette formation complète, structurée et résolument orientée pratique. De la modélisation de flux simples jusqu'à l'automatisation avancée de traitements, vous apprendrez à construire, orchestrer et optimiser vos processus ETL pour répondre aux besoins d'intégration de données dans un contexte métier exigeant.

Vous commencerez par découvrir l'environnement Spoon, les notions de transformation, de job, de step, ainsi que les connexions aux sources de données (fichiers, bases SQL, APIs). Objectif : acquérir les bases pour manipuler et fiabiliser vos premiers flux.

Vous aborderez ensuite les enchaînements de traitements, le contrôle de flux, le passage de variables et la planification, pour créer des workflows robustes et dynamiques adaptés à votre SI.

Un module est dédié aux bonnes pratiques : structuration des projets, optimisation des performances, gestion des erreurs et audit des exécutions. Vous apprendrez à documenter, versionner et sécuriser vos développements.

Comme pour toutes nos formations, celle-ci vous sera présentée avec les toutes dernières actualisations de [Pentaho ETL Fundamentals](#).

## Objectifs

- Comprendre l'architecture de Pentaho Data Integration, son rôle dans la chaîne décisionnelle, et ses cas d'usage typiques en intégration, migration et automatisation de flux de données
- Savoir concevoir, structurer et maintenir des transformations et des jobs Pentaho robustes, en manipulant efficacement les steps, les connexions aux sources, les conditions, les regroupements et les jointures
- Maîtriser les mécanismes de contrôle de flux, de paramétrage, de gestion des erreurs, de logging et de planification, afin de construire des pipelines ETL fiables, modulaires et maintenables

- Être capable d'intégrer et d'orchestrer des traitements complexes, tout en garantissant performance, qualité des données, auditabilité et traçabilité des processus d'intégration
- Appliquer les bonnes pratiques de structuration, de nommage, de documentation, de versioning et de sécurisation dans une logique projet, afin de devenir rapidement opérationnel en environnement professionnel

## Public visé

- Data analyste
- Data Base Administrator

## Pré-requis

- Connaissances de base en manipulation de données
- Maîtrise des fondamentaux SQL

## Programme de la formation Pentaho ETL Fundamentals

### Introduction à Pentaho Data Integration

- Architecture globale : BI Server, PDI, BA Server
- Place de PDI dans la chaîne BI
- Navigation dans Spoon
- Création et ouverture de projets
- Structure d'un projet ETL dans Pentaho
- Transformations vs Jobs
- Steps et Hops
- Variables et paramètres

### Manipulation des données dans les transformations

- Fichiers plats : CSV, Excel, JSON, XML
- Bases de données : MySQL, PostgreSQL, SQL Server, Oracle, etc.
- Autres sources : Web Services, API, MongoDB, etc.
- Table input / output
- Row filter, Calculator, String operations
- Join Rows, Merge, Sort, Group by
- Gestion des erreurs dans les steps
- Flux d'erreurs dédiés
- Utilisation de "Abort", "Dummy", "Null if" pour contrôler les flux

### Construction de Jobs

- Différences entre job et transformation
- Steps spécifiques aux jobs : Start, Success, Transformation, Job, etc.
- Contrôle de flux : if, switch, while, success/failure
- Appel conditionnel de jobs ou transformations
- Gestion des événements
- Utilisation de variables d'environnement
- Passer des paramètres entre jobs/transformations
- Planification via cron, script, carte PDI ou outils externes

## Bonnes pratiques de conception ETL

- Conventions de nommage pour jobs, transformations, steps
- Organisation des projets : dossiers, logique métier, techniques
- Ajout de notes, de metadata
- Gestion de la documentation projet dans Spoon
- Utilisation de transformations en streaming
- Limitation des jointures coûteuses
- Utilisation de buffers et partitionnement

## Requêtage et traitement des données avancé

- Requêtes SQL avancées dans "Table Input"
- Étapes de Lookup
- Pivot/dépivot de données
- Step "Modified Java Script Value"
- Utilisation de conditions complexes
- Appel de fonctions personnalisées
- Conversion de formats
- Nettoyage des données : remplacement, trim, null

## Intégration avec l'environnement d'entreprise

- Exportation des jobs .kjb et transformations .ktr
- Déploiement avec carte PDI (Carte Pan) ou ligne de commande
- Logging et surveillance
- Configuration du logging (Step, Job, Transformation)
- Analyse des résultats d'exécution
- Audit des performances et de la qualité des données
- Exécution depuis shell / scripts batch
- Intégration avec Jenkins, Talend ou Apache NiFi
- Pentaho Server : exécution planifiée et monitoring centralisé

## Sécurité, versioning et collaboration

- Accès aux fichiers et bases de données
- Chiffrement des mots de passe dans PDI
- Utilisation de Pentaho Repository avec sécurité
- Travail collaboratif sur jobs et transformations

- Versioning avec Git
- Bonnes pratiques de gestion de versions

## Étude de cas et atelier pratique

- Construction d'un flux ETL complet :
- Ingestion de fichiers CSV
- Intégration en base de données
- Nettoyage et transformation
- Génération de rapports
- Validation étape par étape
- Utilisation des "preview", "debug", "breakpoint"
- Analyse du temps d'exécution
- Application des bonnes pratiques de performance
- Ajout de logs et d'audits

## Sociétés concernées

Cette formation s'adresse à la fois aux particuliers ainsi qu'aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

## Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

## Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

## Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

## Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte des compétences.

## Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.

