

Formation Apache Hop : orchestrer et automatiser vos flux de données ETL

Réf. SF34308

Maîtrisez Apache Hop pour extraire, transformer et charger vos données depuis des sources hétérogènes grâce à des pipelines visuels et des workflows automatisés.

3 j Durée	2 400 € HT Inter / participant	5 250 € HT Intra (≤12 pers.)	0 Sessions à venir
--------------	-----------------------------------	---------------------------------	-----------------------

Objectifs pédagogiques

- ✓ Identifier le positionnement d'Apache Hop dans l'écosystème ETL et configurer son environnement de travail
- ✓ Concevoir des pipelines de transformation et des workflows d'orchestration
- ✓ Configurer les connexions aux sources et cibles de données hétérogènes
- ✓ Appliquer les opérations de manipulation de flux : tri, filtrage, jointure, calculs et comparaison
- ✓ Utiliser les variables et les boucles pour enrichir et itérer sur les flux de données
- ✓ Diagnostiquer les erreurs, paralléliser les exécutions et automatiser les traitements en production

Public visé & prérequis

Public visé Administrateurs de base de données, Développeurs, Ingénieurs Big Data

Prérequis- Maîtrise du langage SQL (requêtes, jointures, types de données) - Connaissance des concepts de bases de données relationnelles

Programme détaillé

Découverte d'Apache Hop et mise en place de l'environnement

- Positionnement d'Apache Hop : outil ETL open source et son rôle dans l'écosystème d'intégration de données
- Architecture de la plateforme : distinction entre pipelines (transformation) et workflows (orchestration)
- Historique et communauté : origines du projet, évolution depuis Kettle/Pentaho et gouvernance Apache
- Installation et configuration : mise en place de l'environnement de développement et premiers repères dans l'interface

Activités pratiques :

- Installation d'Apache Hop et configuration d'un premier projet
- Navigation dans l'interface graphique et exploration des composants disponibles

Conception de pipelines et workflows

- Pipelines de transformation : création d'un flux de données de bout en bout, de la source à la cible
- Workflows d'orchestration : enchaînement et ordonnancement des pipelines pour automatiser un processus complet
- Exécution et supervision : lancement d'un pipeline et d'un workflow, lecture des résultats d'exécution
- Gestion des flux de données : appréhender le parcours des données à travers les étapes de transformation

Activités pratiques :

- Conception d'un pipeline simple d'extraction et de chargement de données
- Élaboration d'un workflow orchestrant plusieurs pipelines séquentiels

Connexion aux sources et cibles de données

- Métadonnées de connexion : configuration et gestion centralisée des accès aux bases de données
- Connecteurs source et cible : panorama des systèmes supportés (bases relationnelles, fichiers, APIs)
- Jointures entre sources : liens et associations entre données issues de systèmes différents

- Alimentation en Insert/Update : stratégies d'insertion et de mise à jour des données cibles

Activités pratiques :

- Configuration de connexions vers plusieurs bases de données
- Mise en place d'un flux d'alimentation avec gestion Insert/Update

Manipulation et transformation des flux de données

- Tri et ordonnancement : classement ascendant ou descendant des enregistrements d'un flux
- Dédoublage et filtrage : séparation d'un flux en sous-ensembles selon des critères métier
- Extraction de champs : récupération d'informations à partir de chaînes de caractères
- Remplacement et calculs : substitution de valeurs et opérations arithmétiques sur les données du flux
- Produit cartésien et jointures hétérogènes : croisement de données issues de sources différentes
- Comparaison de flux : détection de différences entre deux jeux de données

Activités pratiques :

- Transformation d'un flux brut : tri, dédoublage et filtrage selon des règles métier
- Jointure de données issues de deux sources hétérogènes avec comparaison des résultats

Enrichissement des flux et gestion des variables

- Génération de logs : traçabilité des opérations et suivi de l'exécution des flux
- Variables et paramètres : création et récupération de variables (dates, numériques, alphanumériques) pour dynamiser les traitements
- Résultat de flux : exploitation de la sortie d'un pipeline dans un workflow ou un autre pipeline
- Propriétés d'ordonnancement : configuration des propriétés d'un flux et de son orchestrateur

Activités pratiques :

- Paramétrage dynamique d'un pipeline à l'aide de variables de dates et de chemins
- Chaînage de deux pipelines en exploitant le résultat du premier comme entrée du second

Itérations et boucles

- Problématiques d'itération : cas d'usage nécessitant un traitement répétitif sur des jeux de données
- Boucles paramétrées : utilisation du composant « Copie lignes vers résultat » pour transmettre des données entre itérations
- Boucles avec récupération : combinaison des composants « Copie lignes vers résultat » et « Récupération lignes depuis le résultat » pour des traitements par lots

Activités pratiques :

- Mise en place d'une boucle traitant un ensemble de fichiers de manière itérative
- Construction d'un workflow itératif avec passage de paramètres entre chaque tour de boucle

Exploitation, automatisation et mise en production

- Gestion des erreurs : détection, capture et traitement des erreurs dans les pipelines et workflows
- Débogage et alertes : analyse des traces d'exécution et mise en place de notifications en cas d'anomalie
- Parallélisation : exécution simultanée de plusieurs flux pour optimiser les temps de traitement
- Import/export des développements : portabilité des pipelines et workflows entre environnements
- Automatisation des tâches : planification et déclenchement automatique des traitements
- Documentation et normes : bonnes pratiques de nommage, commentaires et gestion des reprises sur erreur

Activités pratiques :

- Exécution parallèle de plusieurs flux de données avec gestion des erreurs
- Automatisation d'un workflow complet avec planification et alertes en cas d'échec

Pourquoi choisir Sparks Formation ?

20 ans d'expertise IT, des formateurs experts et des sessions garanties contre l'annulation.

20 ans
d'expertise

1 068+
formations

11 178
apprenants formés/an

+67
Net Promoter Score

Avis des participants

Note moyenne des participants de la filière Data Engineering & Big Data : 4,9/5 — 13 avis

Arnaud M. — That was really useful for me, even though I am a PDA. The trainer was very patient with me (I needed more explanation as I was the only...

Thierry R. — Formation idéale pour appréhender la technologie nosql Bon support de formation. Bonne pédagogie

Votre formateur

Formation N. — Formatrice Data & SAP

Formation intervient sur des formations portant sur le data engineering et le Big Data. Parmi les formations qui lui sont confiées figurent notamment « Semarchy xDI » et « Data...

Ils nous font confiance

Cette année notre filière Data Engineering & Big Data a accompagné plus de 73 entreprises dont :

SOPRA STERIA CAPGEMINI TECHNOLOGY SERVICES AMADEUS SAS GROUPAMA SUPPORTS & SERVICES AUBAY FRAIKIN FRANCE LBC FRANCE
ORANGE BUSINESS SERVICES SA ACATUS INFORMATIQUE ACENSI SANA +63 autres

Financement & certifications

Cette formation est éligible aux financements OPCO et FNE-Formation. Nos équipes vous accompagnent dans le montage de votre dossier.



Réservez votre place dès maintenant

20 ans d'expertise · 1 000+ formations · NPS +67 · réponse sous 24h ouvrées

0 805 950 800

demande@sparks-formation.com