Formation Apache HOP

Atol CD
un acteur incontournable sur **Apache HOP**

Présentation

Apache HOP est un fork de Pentaho Data Integration (Hitachi Vantara), qui descend lui-même de l'ETL Kettle [1] créé en 2005 par Matt Casters [2].

Atol CD a mis en œuvre Pentaho Data Integration (PDI) pendant **plus de 15 ans** chez ses clients, avec un nombre conséquent de projets (plusieurs milliers de Jours/Hommes réalisés). Plus de 600 personnes ont été formées à cet ETL par nos formateurs experts, en les rendant rapidement autonomes pour la conception et la création de traitements d'intégration de données, dans de nombreux domaines : administration, collectivités, monde agricole, industrie, santé, banque…

Atol CD a ainsi été pendant de nombreuses années l'intégrateur de référence en France sur Pentaho, avec notamment l'organisation d'un évènement phare : le **Pentaho Day** [3], dont l'objectif était de présenter des retours d'expérience variés, notamment sur Pentaho Data Integration.

En 2021, au moment du fork de PDI vers Apache HOP, Atol CD a poursuivi et renforcé son soutien à la communauté, en effectuant le portage de ses plugins GIS de PDI à Apache HOP : https://www.atolcd.com/actualite/les-plugins-gis-pour-pdi-disponibles-pour-apache-hop
https://github.com/atolcd/hop-gis-plugins

C'est dans ce contexte que le pôle Data a souhaité proposer une formation Apache HOP, en s'appuyant sur ses retours d'expérience projet.

Tout comme le programme de formation PDI (désormais abandonné), le **programme de formation Apache HOP** est conçu pour une prise en main immédiate de l'ETL, et une application immédiate sur vos projets

Si vous souhaitez découvrir les fonctionnalités principales d'Apache HOP, nous vous invitons à consulter le webinar donné par Sylvain Decloix, responsable du pôle Data Atol CD : [Webinar] Apache HOP : la plate-forme d'intégration de données open source

Notes :

(1) Atol CD a commencé en 2006 la mise en œuvre de projets d'intégration de données avec l'ETL Kettle. En 2008, Atol CD publie le livre blanc "Les ETL open source - une alternative réelle aux ETL propriétaires", qui compare Kettle 3 et Talend 2. Ce livre blanc est encore disponible au téléchargement (même s'il est bien sûr complètement obsolète)

(2) Pendant plusieurs années, le pôle Data Atol CD a organisé le "Pentaho Day", l'évènement de référence de la communauté Pentaho en France. En 2019, Matt Casters en était l'invité principal. Son intervention est toujours disponible sur la chaîne Youtub Atol https://www.youtube.com/watch?v=ykkonp7we3E

(3) A découvrir : quelques liens sur les éditions du Pentaho Day : 2017, 2018, 2019.

Objectifs de la formation

La formation Apache HOP est essentiellement **axée sur la pratique**, à l'issue de celle-ci les participants auront :

- compris les principes généraux et les cas d'utilisations de l'ETL Open Source Apache HOP
- appréhendé l'architecture et le mode de fonctionnement de HOP
- intégré les bonnes pratiques de développement
- modélisé des traitements de données avec HOP, via des travaux pratiques de mise en œuvre de pipelines et workflows.
- manipulé les briques de transformation essentielles
- manipulé et pris connaissance de briques et concepts avancés

Déroulement de la formation / pré-requis

La formation se déroule sur **2 jours**, à distance ou en présentiel, avec mise à disposition des éléments suivants par Atol CD :

- Un package de formation contenant :
  - Apache HOP en dernière version stable
  - une base de données SQLite avec l'ensemble des données nécessaires aux travaux pratiques
  - des fichiers additionnels (CSV, XML, Excel)
- Un support de formation (pdf)
- Les solutions des exercices effectués en travaux pratiques

Le nombre de participants est limité à **6 personnes maximum**, aussi bien en mode distanciel que présentiel.

- Chaque participant dispose d'un PC (Windows ou Linux) avec au minimum 8 Go de RAM et 500 Mo d'espace disque.
- Chaque participant possède des connaissances dans le domaine des bases de données relationnelles (SGBDR) et maîtrise les bases du langage SQL.

Programme de formation

Journée 1

	➔ ETL : présentation des concepts et des cas d'utilisations ➔ Présentation de l'architecture et des fonctionnalités de Apache HOP
TP 01 **Installation et prise en main du client de conception graphique HOP GUI**	➔ installation de HOP et du package de formation ➔ configuration JVM et hop-gui.bat ➔ vue d'ensemble et prise en main du studio : métadonnées, explorateur de fichier, options, logs et paramétrages ➔ présentation des types de traitement HOP : pipelines et workflows ➔ présentation des types de données et concepts java associés
TP 02 **Création d'un projet HOP**	➔ concept de mise en œuvre ➔ gestion des environnements (dev, test, prod) ➔ définition des fichiers de configuration et des variables

	➔ principes de déploiement ➔ création du projet de formation
TP 03 **Manipulation du Database Explorer**	➔ définition d'une connexion à une base de données ➔ exploitation du contenu d'une base de données ➔ editeur SQL ➔ paramétrages avancés
TP 04 **Extraction de données depuis une base de données**	➔ récupération de données depuis une base ➔ modification de la structure du flux (transtypage, renommage des colonnes, tri) ➔ export vers un fichier CSV
TP 05 **Extraction de données en mode parallèle**	➔ comprendre le fonctionnement de la parallélisation dans un pipeline (distribution vs copie) ➔ export vers plusieurs fichiers en parallèle : CSV, XML, Excel ➔ revue des diverses options de génération des fichiers
TP 06 **Mise à jour d'une base de données**	➔ chargement des données dans une table ➔ définition du schéma de correspondance ➔ revue des différentes méthodes : insertion, mise à jour, insertion/mise à jour (upsert) ➔ gestion des rejets d'insertion
TP 07 **Enrichissement d'un flux avec des étapes de recherche (lookup)**	➔ mise en œuvre des étapes "recherche dans flux" et "recherche dans base", avec présentation des modalités d'utilisation (mise en cache des données - impact sur les performances) ➔ ajout d'étapes de calcul ➔ ajout de constantes

Journée 2

TP 08 et 09 **Filtrage des données d'un flux**	➔ mise en place de filtres pour exclure les données inutiles ➔ mise en place de l'étape "validation de données" , permettant la validation des données à partir de fichiers ou tables de référence.
TP 10 **Redirection conditionnelle des données**	➔ effectuer le routage des données sur des cibles différentes, en fonction des valeurs d'un ou plusieurs champs du flux ➔ assignation de valeurs littérales à des plages numériques
TP 11 **Création d'un workflow**	➔ comprendre et mettre en place un workflow dans apache HOP ➔ configurer plusieurs pipelines dans le workflow ➔ gérer les erreurs ➔ définir les notifications par mail ➔ planifier le déclenchement automatique d'un workflow, via "hop-run"
TP 12 **Exercice de synthèse**	➔ pour ce TP, le formateur décrit les attendus du traitement à réaliser (sur des données UNEDIC). ➔ les participants effectuent la conception et mise en œuvre complète du traitement ➔ l'objectif est de vérifier la bonne compréhension de l'ensemble des concepts et des étapes de base Apache HOP
TP 13 **Opérations d'agrégations de données et calculs avancés**	➔ mise en œuvre d'étapes d'agrégation de données sur un flux de données (équivalent du "group by" en SQL) ➔ présentation des différentes étapes

	➔ mise en œuvre de calculs avancés avec l'étape javascript rhino
TP 14 **Dénormalisation et normalisation de données**	➔ présentation du concept de "Pivot" de données ➔ présentation des étapes "dénormalisation" et "normalisation" ➔ exemple de traitement de dénormalisation ➔ génération d'un fichier excel dénormalisé, avec extraction dynamique en SQL (passage de paramètre) ➔ exécution en ligne de commande (hop-run) avec passage de paramètre
Finalisation de la formation, avec de nombreux conseils pratiques issus de nos projets Apache HOP	➔ revue des "samples" ➔ conseils pour la lisibilité et la maintenabilité des traitements ➔ bonnes pratiques de déploiement ➔ l'approche DataOps avec Apache HOP : versionnement Git, conteneurisation Docker ➔ présentation du concept d'injection de métadonnées : ce concept permet une configuration automatique de tout ou partie d'un pipeline HOP (il a été mis en œuvre sur des projets clients)
	➔ Questions ouvertes des participants ➔ Clôture de la formation

Après la formation

La formation Apache HOP que nous dispensons a pour objectif de permettre une mise en œuvre immédiate de l'ETL par les participants, sur des sujets concrets de traitements de données. Les retours des participants indiquent que la formation permet d'atteindre avec efficience cet objectif !
Néanmoins, tout ne peut pas être abordé en formation, notamment des points spécifiques, tels que la mise en œuvre de récupération de données au travers d'API, la gestion du SSL, le parsing avancé de données XML ou JSON, etc…
Pour cette raison, nous proposons en complément de cette formation un contrat d'assistance au travers de notre plate-forme de support.
Ce contrat repose sur des carnets de tickets d'intervention, chaque ticket donnant droit à 1h d'intervention (1 ticket = 85 € HT). Ce contrat de support permet de procurer des réponses à toute question de faisabilité dès les premiers travaux engagés à l'issue de la formation.

Tarifs

Formation Apache HOP à distance

2 jours : 1 190 € HT / personne

Pour des membres d'une même organisation : nous consulter (tarif personnalisé, possibilité d'intervention sur site, sous conditions).

Carnet d'assistance :

- 10 tickets : 850 € HT
- 20 tickets : 1 650 € HT (remise de 3%)
- 50 tickets : 3 950 € HT (remise de 7%)