Introduction aux ETL

Clément Agret

CY Cergy Paris Université

20 mars 2024

Plan de la présentation

Présentation Générale

- Qu'est ce qu'un ETL
  - Extract : Extraction des données d'une ou de plusieurs sources
  - Transform : Transformation des données collectées
  - Load : Chargement des données dans la cible
- Types d'ETL existants
  - Via du code
  - ETL
  - ELT

Présentation Générale

- Via du code
  - Possibilité de faire tout ce que l'on veut
  - Possibilité d'utiliser des procédures stockées
  - Difficultés dans la maintenance
  - Compétences avancées en code
  - Compétences avancées en SQL
  - Je ne vous apprend rien

Présentation Générale

- ELT
  - Fonctionnalités de travail collaboratif classiques
  - Le travail de transformation des données est assumé par les SGBD
  - Interfaces pour la maintenance et la reprise des développements
  - Peu de couts de machine en plus
  - Les serveurs cibles et sources sont indisponibles pendant le traitement des données
  - Aujourd'hui beaucoup moins considérés
  - Exemple : SUNOPSIS (Oracle Data Integrator)

Présentation Générale

- ETL
  - Fonctionnalités de travail collaboratif classiques
  - Le travail de transformation est assuré par un serveur indépendant
  - Possibilité de faire travailler les bases en ELT si besoin est
  - Interfaces pour la maintenance et la reprise des développements
  - Possibilité de dimensionner le serveur d'ETL
  - Supervision intégrée
  - Ajout d'un serveur d'ETL
  - Formation spécifique selon l'outil
  - Exemple : Talend

Présentation Générale

- Quelques ETL du marché
  - Open Source :
    - TALEND
    - PENTAHO
  - Incontournables :
    - Informatica
    - Datastage
  - Divers :
    - Sunopsis (Oracle data integrator)
    - Integration services (Microsoft)
- Dans le cadre d'un entretien, ca peut passer pour de la culture générale... En tous cas ca montre une certaine ouverture d'esprit

Introduction aux ETL

Les ETL (**Extract, Transform, Load**) sont des processus utilisés pour extraire des données à partir de différentes sources, les transformer selon les besoins et les charger dans une destination cible.

Composants des ETL

- **Extract (Extraction)** : Collecte des données à partir de différentes sources.
- **Transform (Transformation)** : Nettoyage, reformatage et enrichissement des données.
- **Load (Chargement)** : Chargement des données transformées dans la destination cible.

Exemples d'outils ETL

- Talend
- Apache NiFi
- Pentaho Data Integration (Kettle)
- Apache Hop (Hop Orchestration Platform)
- StreamSets

Exemples d'outils ETL

- Talend
- Apache NiFi
- Pentaho Data Integration (Kettle)
- Apache Hop (Hop Orchestration Platform)
- StreamSets

Conclusion

Les ETL jouent un rôle essentiel dans le processus d'intégration et d'analyse des données dans les environnements informatiques modernes.

Introduction à Apache Hop

- Apache Hop, anciennement connu sous le nom de **Hop Orchestration Platform**, est un outil d'intégration de données open source.
- Il offre des fonctionnalités puissantes pour l'extraction, la transformation et le chargement (ETL) de données, ainsi que pour la gestion des workflows et des métadonnées.

Téléchargement et Installation

- Apache Hop est disponible en téléchargement gratuit sur le site officiel du projet : `https://hop.apache.org/`.
- Pour l'installer, téléchargez le package approprié pour votre système d'exploitation et suivez les instructions fournies dans la documentation.

Prise en Main

- Une fois installé, lancez Apache Hop en exécutant le fichier exécutable ou en utilisant la commande appropriée selon votre système d'exploitation.
- Vous serez accueilli par une interface graphique conviviale où vous pouvez commencer à créer vos premiers workflows.

Création d'un Workflow Basique

- Lancez Apache Hop.
- Cliquez sur "Nouveau" pour créer un nouveau projet.
- Ajoutez un nouveau workflow à votre projet.
- Double-cliquez sur le workflow pour l'ouvrir dans l'éditeur graphique.
- Utilisez les composants disponibles pour ajouter des étapes d'extraction, de transformation et de chargement à votre workflow.
- Connectez les étapes en utilisant les flèches pour définir le flux de données.
- Une fois le workflow configuré, enregistrez-le et exécutez-le pour tester son fonctionnement.

Qu'est-ce qu'Apache Hop ?

- **Définition :** Plateforme d'orchestration de données et de génie des données.

Qu'est-ce qu'Apache Hop ?

- **Définition :** Plateforme d'orchestration de données et de génie des données.
- **Objectif :** Simplifier l'orchestration des données et des métadonnées.

Qu'est-ce qu'Apache Hop ?

- **Définition :** Plateforme d'orchestration de données et de génie des données.
- **Objectif :** Simplifier l'orchestration des données et des métadonnées.
- **Simplicité et Complexité :** Les tâches simples sont faciles, les tâches complexes sont possibles.

Avantages d'Apache Hop

- **Conception Visuelle :** Permet une approche centrée sur les objectifs plutôt que sur le processus.

Avantages d'Apache Hop

- **Conception Visuelle :** Permet une approche centrée sur les objectifs plutôt que sur le processus.
- **Productivité :** Augmente la productivité des développeurs en réduisant le besoin d'écriture de code.

Architecture

- **Flexibilité :** Au cœur de Hop se trouve le moteur Hop, compact et puissant.

Architecture

- **Flexibilité :** Au cœur de Hop se trouve le moteur Hop, compact et puissant.
- **Plugins :** Fonctionnalité étendue avec environ 400 plugins. Personnalisation facile en ajoutant ou supprimant des plugins.

Architecture

- **Flexibilité :** Au cœur de Hop se trouve le moteur Hop, compact et puissant.
- **Plugins :** Fonctionnalité étendue avec environ 400 plugins. Personnalisation facile en ajoutant ou supprimant des plugins.
- **Adaptabilité :** Conçu pour tout scénario - IoT, gros volumes de données, local, cloud, OS nu, conteneurs, et Kubernetes.

Runtimes Flexibles

- **Développement Visuel :** Les workflows et pipelines sont créés dans Hop Gui, un environnement de développement visuel.

Runtimes Flexibles

- **Développement Visuel :** Les workflows et pipelines sont créés dans Hop Gui, un environnement de développement visuel.
- **Moteurs d'Exécution :** Peuvent s'exécuter sur le moteur Hop natif (localement ou à distance) et sur Apache Spark, Apache Flink, et Google Dataflow via Apache Beam.

Runtimes Flexibles

- **Développement Visuel :** Les workflows et pipelines sont créés dans Hop Gui, un environnement de développement visuel.
- **Moteurs d'Exécution :** Peuvent s'exécuter sur le moteur Hop natif (localement ou à distance) et sur Apache Spark, Apache Flink, et Google Dataflow via Apache Beam.
- **Opérations sur les Données :** Support de centaines d'opérations pour lire, écrire, combiner, enrichir, nettoyer et manipuler les données.

Runtimes Flexibles

- **Développement Visuel :** Les workflows et pipelines sont créés dans Hop Gui, un environnement de développement visuel.
- **Moteurs d'Exécution :** Peuvent s'exécuter sur le moteur Hop natif (localement ou à distance) et sur Apache Spark, Apache Flink, et Google Dataflow via Apache Beam.
- **Opérations sur les Données :** Support de centaines d'opérations pour lire, écrire, combiner, enrichir, nettoyer et manipuler les données.
- **Modes de Traitement :** Traitement par lots, en continu, ou un hybride des deux, selon le moteur et la fonctionnalité utilisés.

Cas d'utilisation

- **Traitement de Grandes Données :** Chargement dans des bases de données via des environnements parallèles.

Cas d'utilisation

- **Traitement de Grandes Données :** Chargement dans des bases de données via des environnements parallèles.
- **Entrepôt de Données :** Support des Dimensions à Changement Lent, Capture des Données de Changement, et création de clés de substitution.

Cas d'utilisation

- **Traitement de Grandes Données :** Chargement dans des bases de données via des environnements parallèles.
- **Entrepôt de Données :** Support des Dimensions à Changement Lent, Capture des Données de Changement, et création de clés de substitution.
- **Intégration de Données :** Combinaison de bases de données relationnelles, fichiers, et bases de données NoSQL (Neo4j, MongoDB, Cassandra, etc.).

Cas d'utilisation

- **Traitement de Grandes Données :** Chargement dans des bases de données via des environnements parallèles.
- **Entrepôt de Données :** Support des Dimensions à Changement Lent, Capture des Données de Changement, et création de clés de substitution.
- **Intégration de Données :** Combinaison de bases de données relationnelles, fichiers, et bases de données NoSQL (Neo4j, MongoDB, Cassandra, etc.).
- **Migration de Données :** Entre différentes bases de données et applications.

Cas d'utilisation

- **Traitement de Grandes Données :** Chargement dans des bases de données via des environnements parallèles.
- **Entrepôt de Données :** Support des Dimensions à Changement Lent, Capture des Données de Changement, et création de clés de substitution.
- **Intégration de Données :** Combinaison de bases de données relationnelles, fichiers, et bases de données NoSQL (Neo4j, MongoDB, Cassandra, etc.).
- **Migration de Données :** Entre différentes bases de données et applications.
- **Profilage et Nettoyage :** Analyse et purification des données.

Concepts Fondamentaux

- **Métadonnée :** Au cœur de Hop, base de toutes interactions et définitions dans l'architecture de données.

Concepts Fondamentaux

- **Métadonnée :** Au cœur de Hop, base de toutes interactions et définitions dans l'architecture de données.
- **Pipelines :** Collections de transformations s'exécutant en parallèle, connectées par des sauts.

Concepts Fondamentaux

- **Métadonnée :** Au cœur de Hop, base de toutes interactions et définitions dans l'architecture de données.
- **Pipelines :** Collections de transformations s'exécutant en parallèle, connectées par des sauts.
- **Workflows :** Séquences d'actions exécutées séquentiellement, également connectées par des sauts.

Concepts Fondamentaux

- **Métadonnée :** Au cœur de Hop, base de toutes interactions et définitions dans l'architecture de données.
- **Pipelines :** Collections de transformations s'exécutant en parallèle, connectées par des sauts.
- **Workflows :** Séquences d'actions exécutées séquentiellement, également connectées par des sauts.
- **Projets :** Collections logiques de code Hop et de configurations, organisées en environnements spécifiques (dev, uat, prd).

Action

- Une **Action** est une opération effectuée dans un *Workflow*.

Action

- Une **Action** est une opération effectuée dans un *Workflow*.
- **Exécution :** Séquentielle par défaut, avec option pour parallélisme.

Action

- Une **Action** est une opération effectuée dans un *Workflow*.
- **Exécution :** Séquentielle par défaut, avec option pour parallélisme.
- **Résultat :** Renvoie un code vrai ou faux, influençant le flux du *Workflow*.

Hop

- Un **Hop** connecte les *Actions* dans un *Workflow* ou les *Transformations* dans un *Pipeline*.

Hop

- Un **Hop** connecte les *Actions* dans un *Workflow* ou les *Transformations* dans un *Pipeline*.
- **Fonctionnement :** Dans les *Workflows*, basé sur le statut de sortie des *Actions* ; dans les *Pipelines*, transfère les données entre *Transformations*.

Pipeline

- Les **Pipelines** effectuent le travail concret sur les données : lecture, modification, enrichissement, nettoyage, et écriture.

Pipeline

- Les **Pipelines** effectuent le travail concret sur les données : lecture, modification, enrichissement, nettoyage, et écriture.
- **Orchestration :** Les *Pipelines* sont orchestrés via d'autres *Pipelines* et/ou *Workflows*.

Transformation

- Une **Transformation** est une unité de travail dans un *Pipeline*.

Transformation

- Une **Transformation** est une unité de travail dans un *Pipeline*.
- **Opérations typiques :** Lecture de données, exécution de recherches/jointures, enrichissement, nettoyage, etc.

Transformation

- Une **Transformation** est une unité de travail dans un *Pipeline*.
- **Opérations typiques :** Lecture de données, exécution de recherches/jointures, enrichissement, nettoyage, etc.
- **Exécution :** Toutes les *Transformations* s'exécutent en parallèle, déplaçant les données traitées à travers des *Hops*.

Pipeline

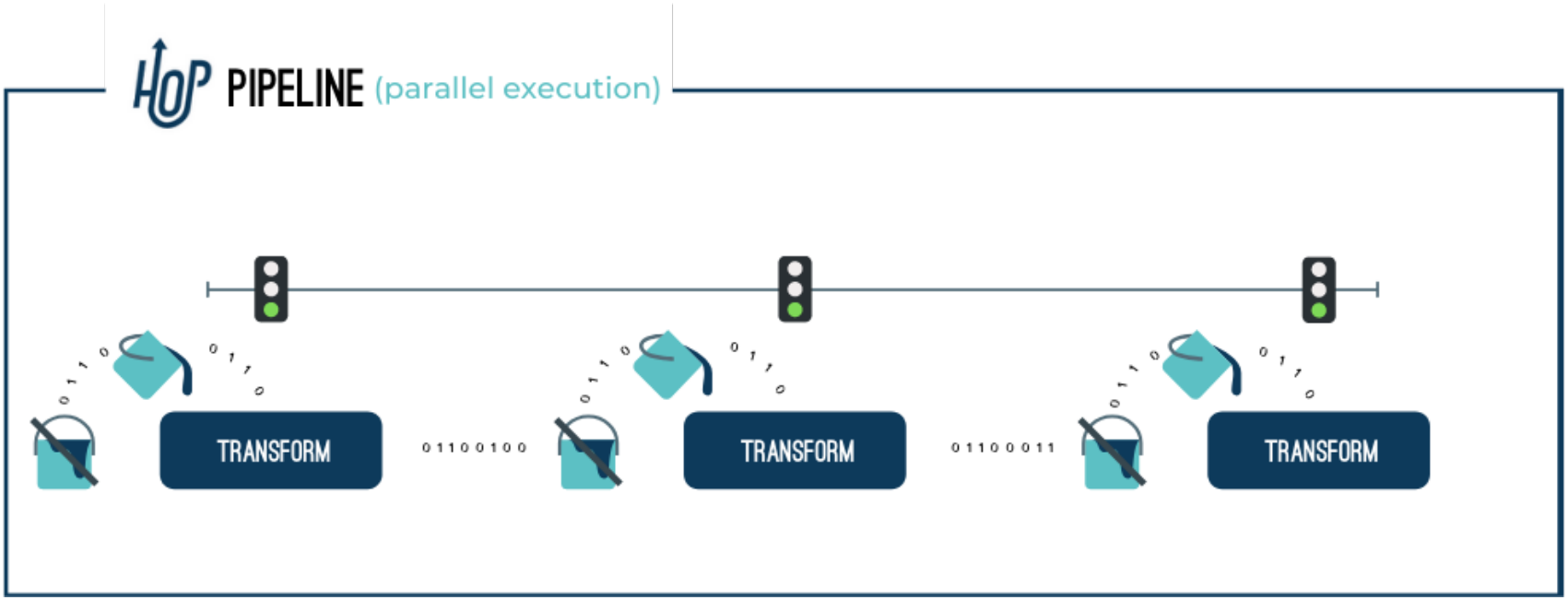


Figure – Exemple de pipeline dans Apache Hop

Types d'Éléments

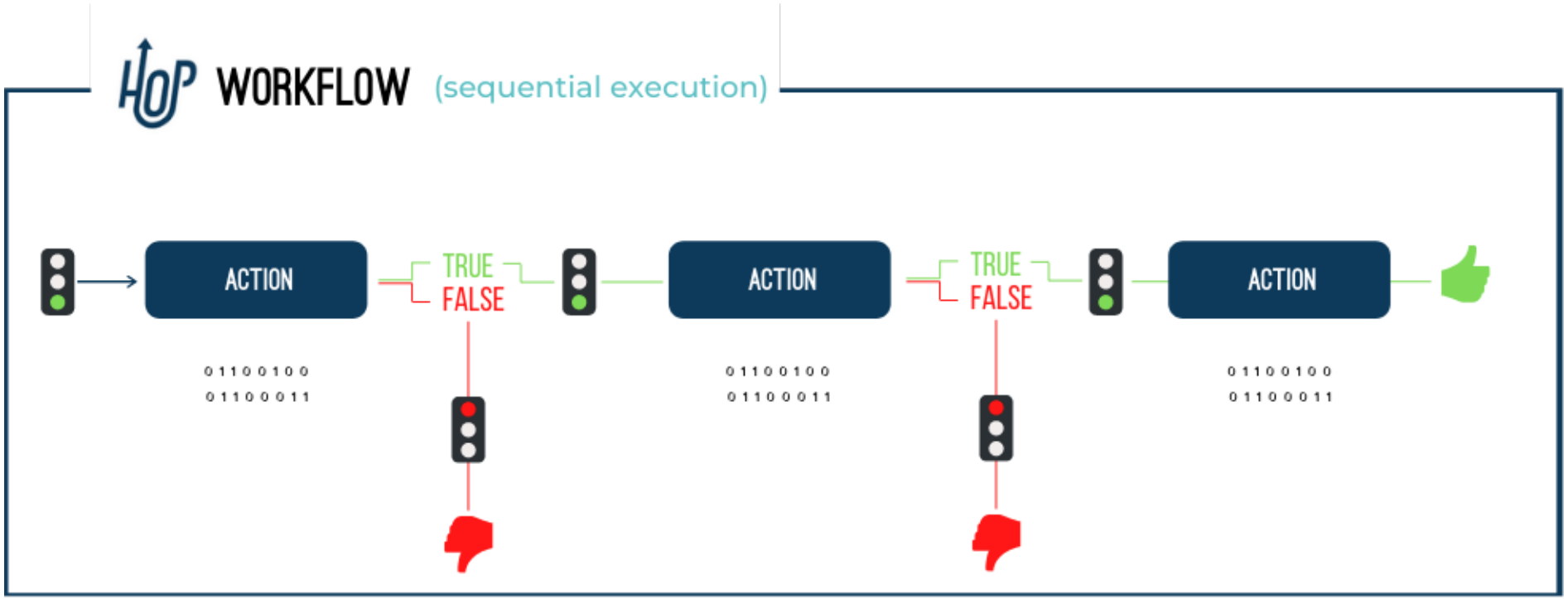


Figure – Exemple de workflow dans Apache Hop

Projets et Environnements

Projet

Les Projets Hop sont un regroupement conceptuel de configurations, de variables, d'objets de métadonnées, de workflows et de pipelines. Les projets peuvent hériter des métadonnées des projets parent. Un projet contient un ou plusieurs environnements où la configuration réelle est définie.

Environnement

Les Environnements Hop sont des instances de projets qui contiennent les configurations d'exécution réelles et d'autres objets de métadonnées pour un projet donné.

Exemples d'Environnements

Exemple : l'environnement 'dev' pour le projet 'Sales' spécifie de lire à partir de l'hôte '10.0.0.1' pour la connexion à la base de données 'clients'.

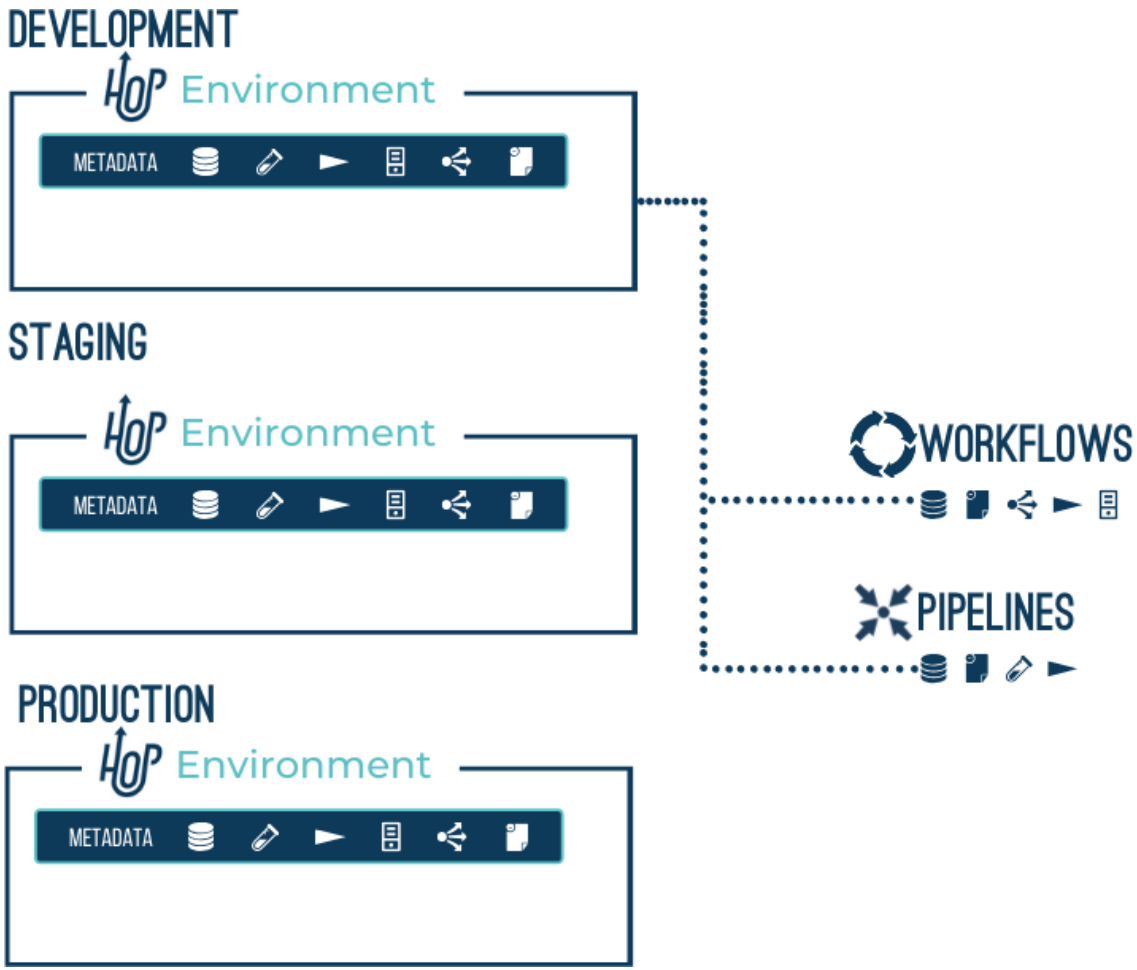


Figure – Exemple d'environnement dans Apache Hop

Télécharger et Installer Hop

Hop est conçu pour être simple et accessible.
Pour commencer avec Hop :

Téléchargement

1. Rendez-vous sur la page de téléchargements.

Prérequis

2. Hop nécessite Java. Utilisez Java version 11.
   - Consultez Adoptium.net pour Java 11.

Installation

3. Décompressez Hop dans le dossier de votre choix.
   - Exécutez Hop via les scripts dans le dossier décompressé.

Pipelines

Un résumé rapide sur les pipelines, transformations, et sauts :

Pipeline

- Chaîne de **transformations** traitant les données.
- Forme un DAG (Graphe Acyclique Dirigé), sans boucles.

Transformation

- Opération de base dans un pipeline.
- Lit, traite, ou écrit des données.

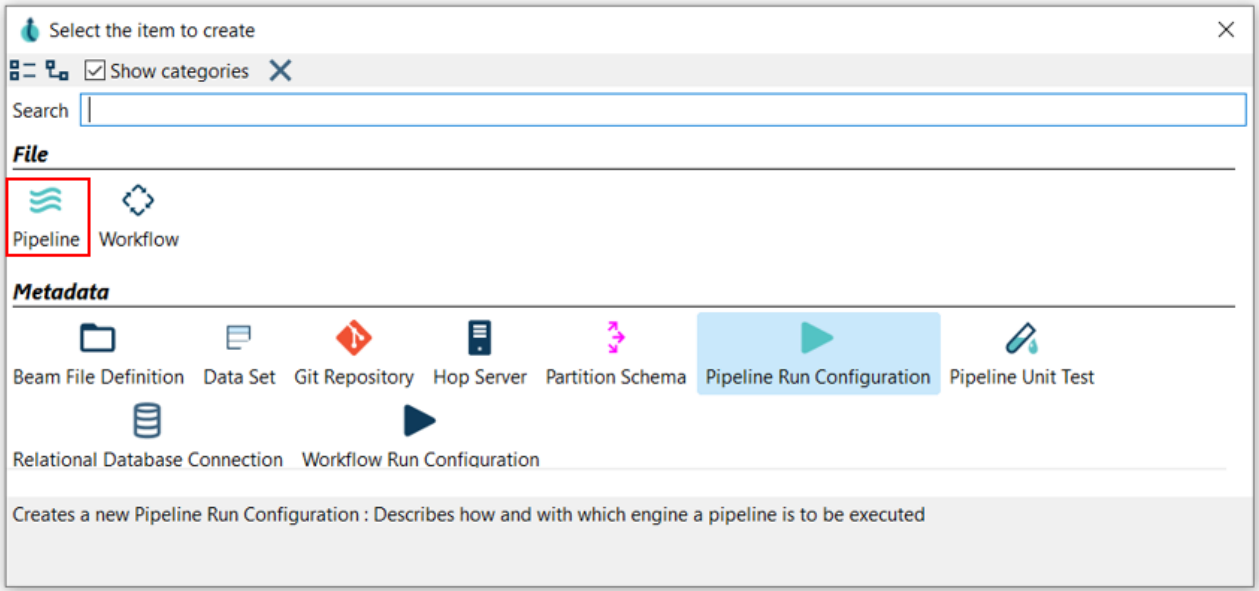
Saut

- Transfère des données entre transformations.

Création d'un Pipeline

Il existe deux façons de créer un pipeline :

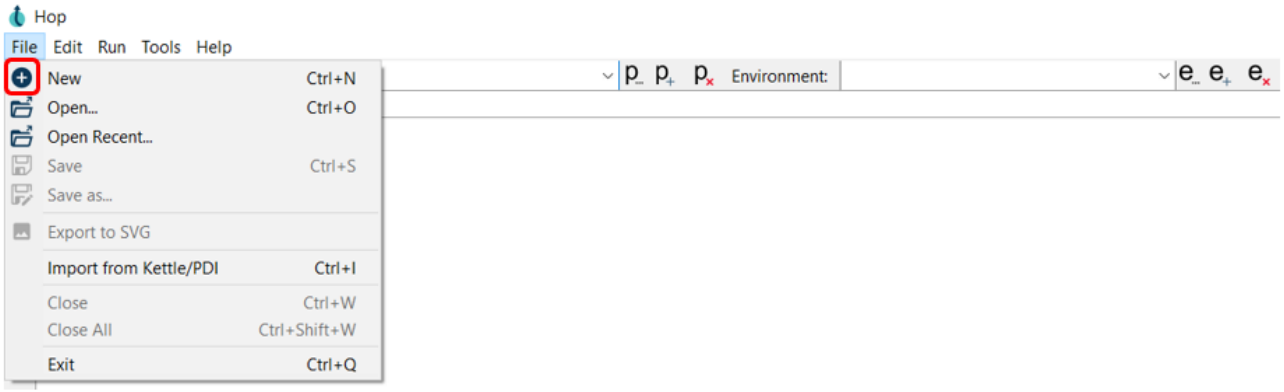
- Cliquez sur l'option *Nouveau* dans la barre d'outils horizontale et sélectionnez l'option *Pipeline*.



Création d'un Pipeline

Il existe deux façons de créer un pipeline :

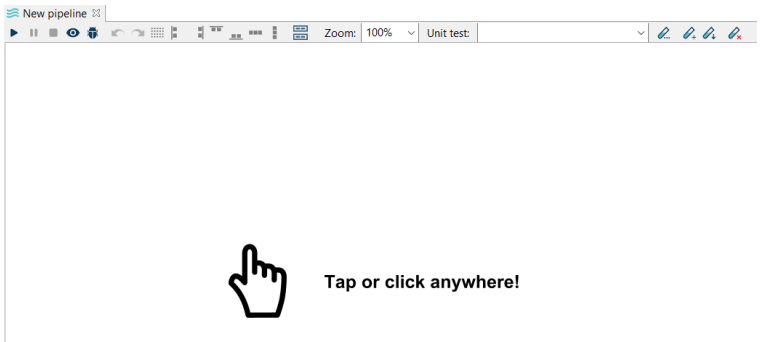
- Allez dans *Fichier* → *Nouveau* → *Pipeline*.



Création d'un Pipeline

Il existe deux façons de créer un pipeline :

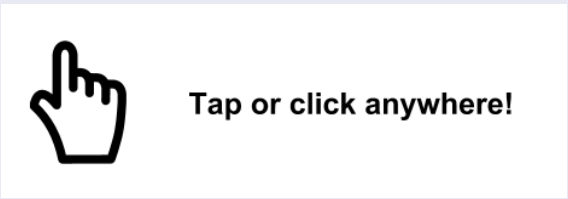
- Une fois votre nouveau pipeline créé, vous verrez la boîte de dialogue ci-dessous.



Ajout et Connexion des Transformations

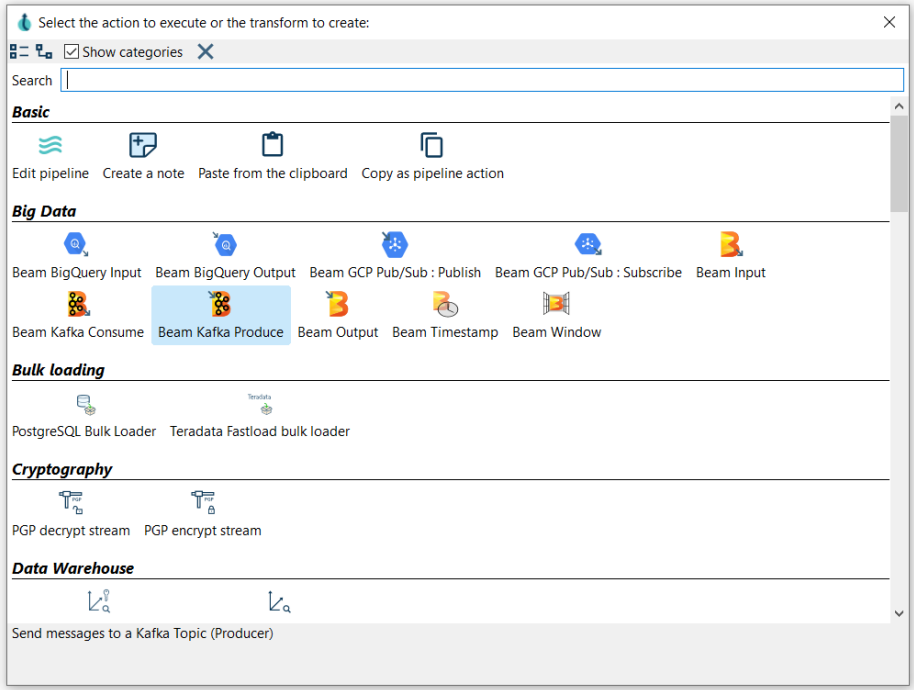
Ajout des Transformations

Maintenant, vous êtes prêt à ajouter la première transformation. Cliquez n'importe où dans le canevas du pipeline, la zone où vous verrez l'image ci-dessous.



Ajout et Connexion des Transformations

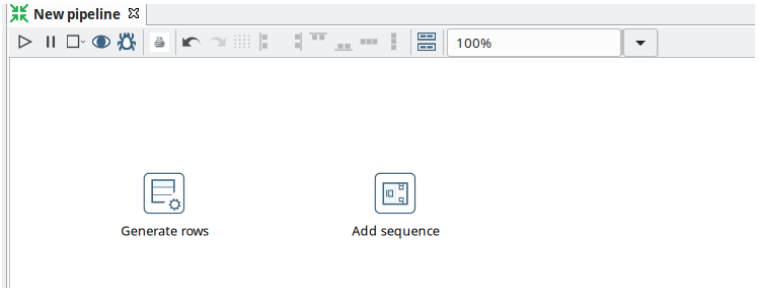
Après avoir cliqué, vous verrez la boîte de dialogue contextuelle. C'est cette boîte de dialogue que vous utiliserez chaque fois que vous aurez besoin de sélectionner des transformations à ajouter à votre pipeline.



Utilisez la zone de recherche dans cette boîte de dialogue pour trouver les transformations dont vous avez besoin. Cliquez ou utilisez les touches fléchées et appuyez sur Entrée pour ajouter une transformation à votre pipeline.

Ajout et Connexion des Transformations

Pour l'instant, ajoutez une transformation de génération de lignes et une transformation d'ajout de séquence à votre pipeline.



Création d'un Saut

Il existe plusieurs façons de créer un saut :

- Shift-drag : tout en maintenant la touche Shift enfoncée sur votre clavier. Cliquez sur une transformation, tout en maintenant enfoncé le bouton principal de votre souris, faites glisser vers la deuxième transformation. Relâchez le bouton principal de la souris et la touche Shift.

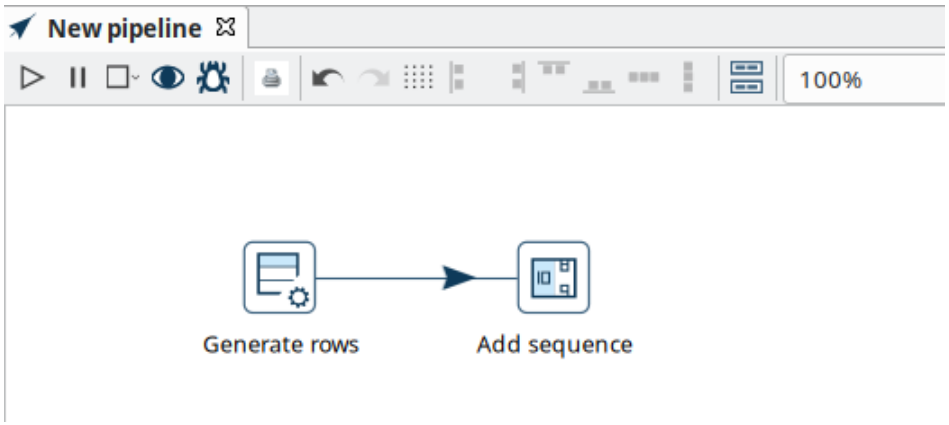


Figure – Création d'un saut dans Apache Hop

Création d'un Saut

Il existe plusieurs façons de créer un saut :

- Scroll-drag : cliquez sur une transformation avec le bouton de défilement de votre souris enfoncé, faites glisser vers la deuxième transformation. Relâchez le bouton de défilement.

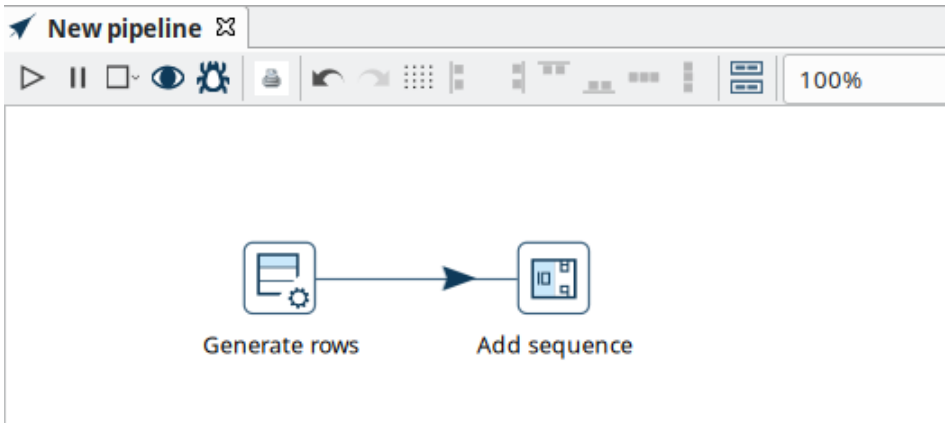


Figure – Création d'un saut dans Apache Hop

Création d'un Saut

Il existe plusieurs façons de créer un saut :

- Cliquez sur une transformation dans votre pipeline pour ouvrir la boîte de dialogue contextuelle. Cliquez sur le bouton "Créer un saut" et sélectionnez la transformation à laquelle vous voulez créer le saut.

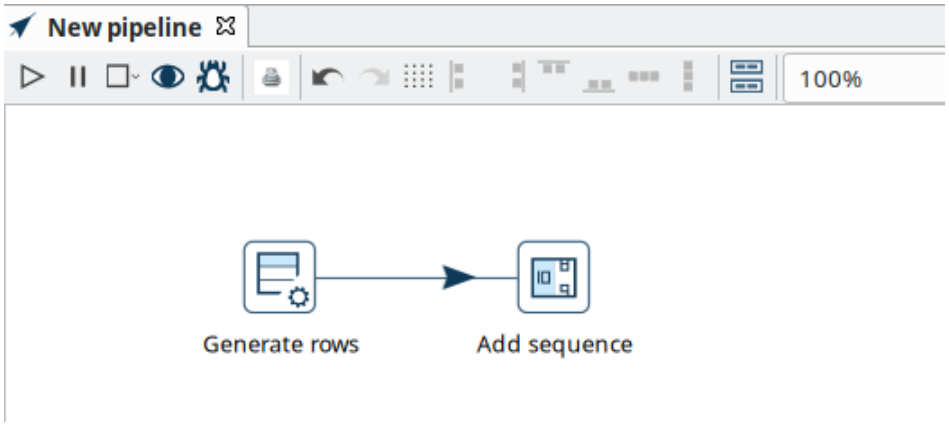


Figure – Création d'un saut dans Apache Hop

Exécution de votre Pipeline

L'exécution d'un pipeline pour voir comment il se comporte peut être effectuée de l'une des manières suivantes :

- En utilisant l'icône Exécuter.
- Sélectionnez Exécuter et cliquez sur Démarrer l'exécution dans la barre d'outils.
- Appuyez sur F8.

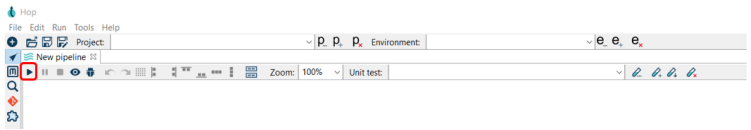


Figure – Exécution d'un pipeline dans Apache Hop (1)

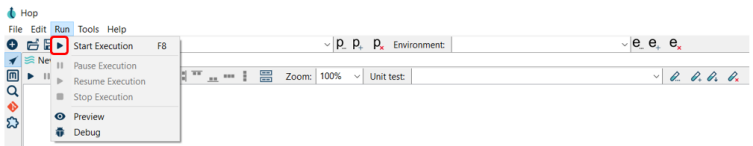


Figure – Exécution d'un pipeline dans Apache Hop (2)

Exécution de votre Pipeline (suite)

Vous verrez la boîte de dialogue des Options d'exécution.

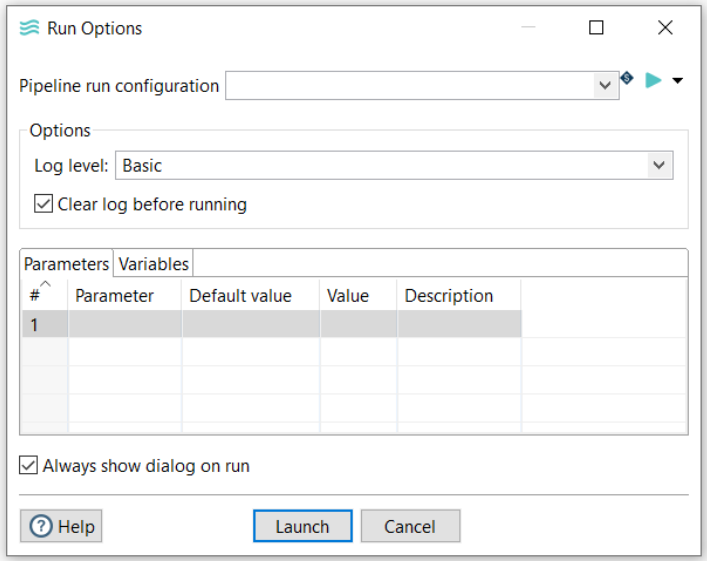


Figure – Options d'exécution dans Apache Hop

Une configuration d'exécution "locale" est créée lorsque vous démarrez Hop Gui pour la première fois. Vérifiez les configurations d'exécution disponibles pour d'autres moteurs sur lesquels exécuter vos pipelines. Assurez-vous que votre configuration est sélectionnée et appuyez sur Lancer.

Exécution de votre Pipeline (suite)

Après chaque exécution, les résultats d'exécution sont affichés dans le panneau en bas de votre fenêtre. Les résultats d'exécution contiennent deux onglets :

- Mesures de transformation
- Journalisation

Figure – Résultats d'exécution dans Apache Hop

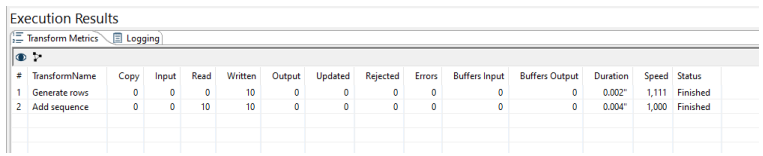


Figure – Journalisation dans Apache Hop

Exécution de votre Pipeline (suite)

Pour des informations plus détaillées, consultez la page Exécuter, Prévisualiser et Déboguer un Pipeline.

Workflows

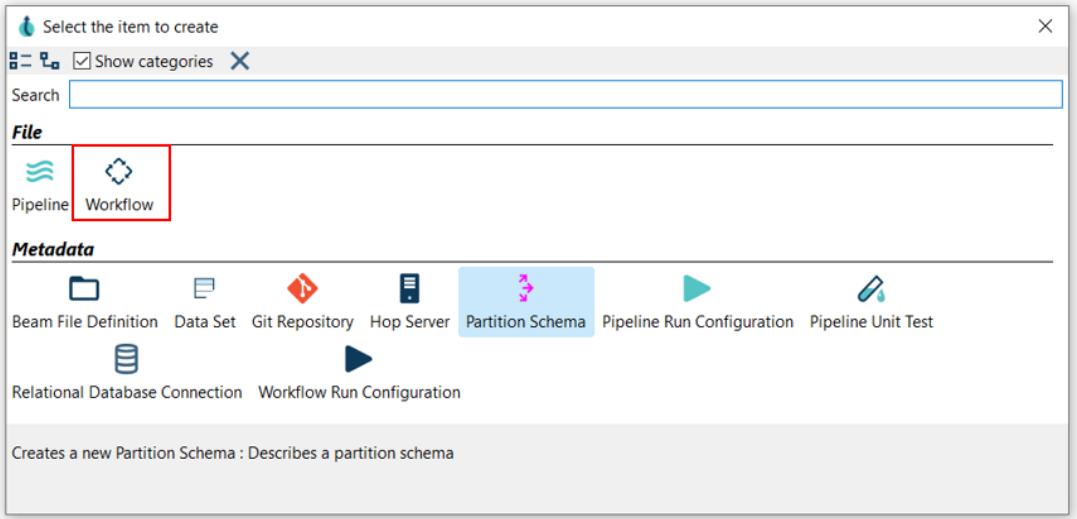
Dans les concepts, nous avons parcouru les workflows, les actions et les sauts. Rappelons-nous :

- Un workflow est par défaut un processus séquentiel qui a un point de départ et un ou plusieurs points d'arrivée.
- Une action est un élément de fonctionnalité du workflow qui exécute des pipelines déjà implémentés.
- Un saut dans un workflow peut connecter conditionnellement des actions et déterminer quelle action le workflow doit exécuter ensuite.

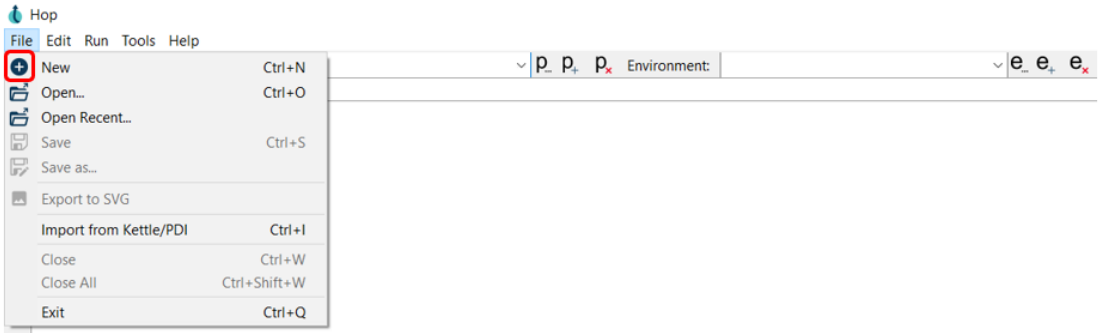
Création d'un Workflow

Il existe deux façons de créer un workflow :

- Cliquez sur l'option Nouveau dans la barre d'outils horizontale et sélectionnez l'option Workflow.

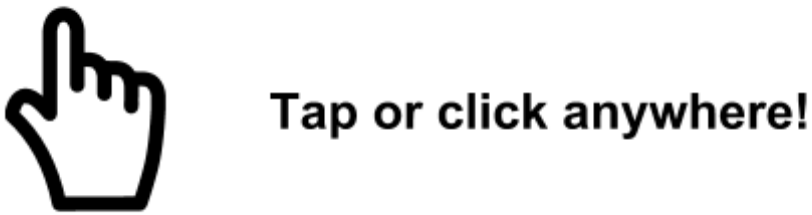


- Allez dans Fichier → Nouveau → Workflow.



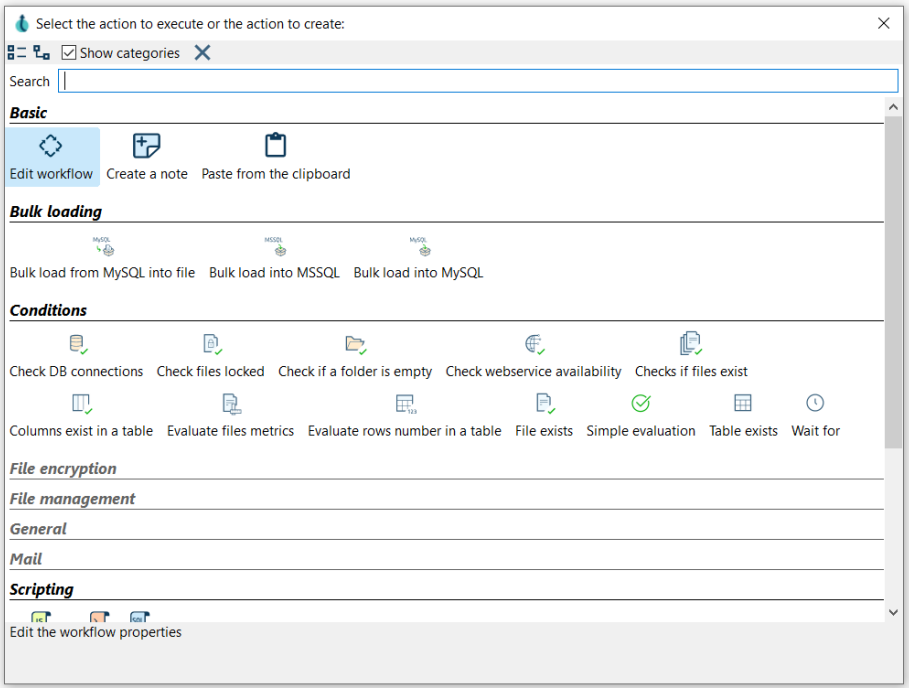
Ajout et Connexion des Actions

Maintenant, vous êtes prêt à ajouter la première action. Cliquez n'importe où dans le canevas du workflow, la zone où vous verrez l'image ci-dessous.



Ajout et Connexion des Actions

Après avoir cliqué, vous verrez une boîte de dialogue. Utilisez la zone de recherche dans cette boîte de dialogue pour trouver les actions dont vous avez besoin. Cliquez ou utilisez les touches fléchées et appuyez sur Entrée pour ajouter une action à votre workflow.
Pour l'instant, ajoutez une action Pipeline à votre workflow.



Création d'un Saut

La création d'un saut entre deux actions dans un workflow ou pipeline peut se faire de plusieurs manières :

1. **Shift-drag :**
   - Maintenez la touche *Shift* et cliquez sur une action.
   - Faites glisser vers la deuxième action et relâchez.

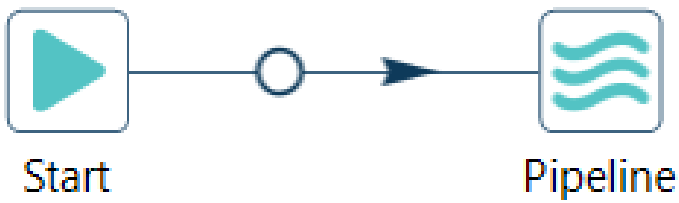


Figure – Ajout d'une action Pipeline dans Apache Hop

Création d'un Saut

La création d'un saut entre deux actions dans un workflow ou pipeline peut se faire de plusieurs manières :

1. **Shift-drag :**
   - Maintenez la touche *Shift* et cliquez sur une action.
   - Faites glisser vers la deuxième action et relâchez.
2. **Scroll-drag :**
   - Utilisez le bouton de défilement de la souris pour cliquer et glisser d'une action à l'autre.

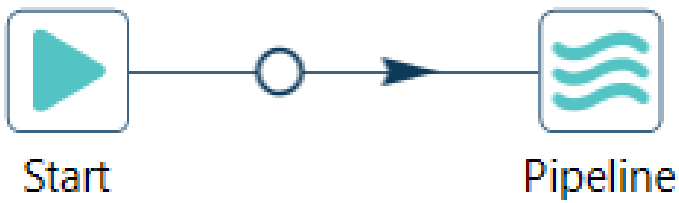


Figure – Ajout d'une action Pipeline dans Apache Hop

Création d'un Saut

La création d'un saut entre deux actions dans un workflow ou pipeline peut se faire de plusieurs manières :

1. **Shift-drag :**
   - Maintenez la touche *Shift* et cliquez sur une action.
   - Faites glisser vers la deuxième action et relâchez.
2. **Scroll-drag :**
   - Utilisez le bouton de défilement de la souris pour cliquer et glisser d'une action à l'autre.
3. **Création via dialogue :**
   - Cliquez sur une action et, dans la boîte de dialogue, choisissez 'Créer un saut' vers une autre action.

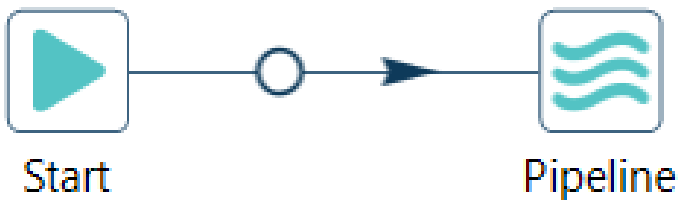


Figure – Ajout d'une action Pipeline dans Apache Hop

Sauvegarde de votre Workflow

Une fois votre workflow créé et configuré, il est temps de le sauvegarder :

- Cliquez sur l'option Enregistrer dans la barre d'outils ou allez dans Fichier → Enregistrer.

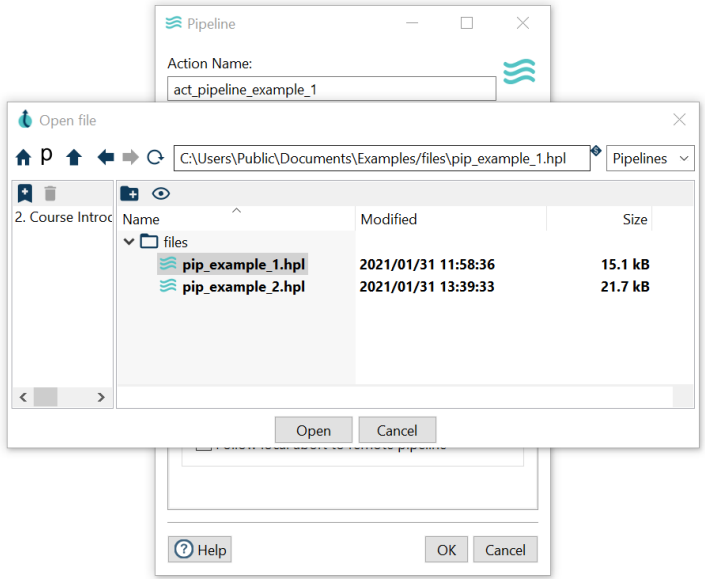


Figure – Enregistrement d'un Workflow dans Apache Hop

- Sélectionnez la configuration d'exécution du workflow. Choisissez la configuration d'exécution "locale" qui est disponible par défaut dans votre installation Hop et cliquez sur "Lancer".

Conclusion

- Vous avez maintenant une bonne compréhension de ce qu'est Apache Hop et comment créer vos premiers workflows et pipelines.

Conclusion

- Vous avez maintenant une bonne compréhension de ce qu'est Apache Hop et comment créer vos premiers workflows et pipelines.
- Il y a encore beaucoup à découvrir dans Apache Hop. Voici quelques sujets que vous voudrez peut-être explorer :

Conclusion

- **Pipelines** : examine de plus près les différents aspects de la création et de l'exécution des pipelines, et contient la liste complète des transformations à votre disposition.

Conclusion

- **Pipelines** : examine de plus près les différents aspects de la création et de l'exécution des pipelines, et contient la liste complète des transformations à votre disposition.
- **Workflows** : examine de plus près les différents aspects de la création et de l'exécution des workflows, et contient la liste complète des actions à votre disposition.

Conclusion

- **Pipelines** : examine de plus près les différents aspects de la création et de l'exécution des pipelines, et contient la liste complète des transformations à votre disposition.
- **Workflows** : examine de plus près les différents aspects de la création et de l'exécution des workflows, et contient la liste complète des actions à votre disposition.
- **Bonnes Pratiques** : couvre un certain nombre de choses auxquelles vous voudrez peut-être réfléchir lors de l'utilisation d'Apache Hop.

Conclusion

- **Pipelines** : examine de plus près les différents aspects de la création et de l'exécution des pipelines, et contient la liste complète des transformations à votre disposition.
- **Workflows** : examine de plus près les différents aspects de la création et de l'exécution des workflows, et contient la liste complète des actions à votre disposition.
- **Bonnes Pratiques** : couvre un certain nombre de choses auxquelles vous voudrez peut-être réfléchir lors de l'utilisation d'Apache Hop.
- **Projets** : explique comment travailler avec des projets et des environnements.

Conclusion

- **Pipelines** : examine de plus près les différents aspects de la création et de l'exécution des pipelines, et contient la liste complète des transformations à votre disposition.
- **Workflows** : examine de plus près les différents aspects de la création et de l'exécution des workflows, et contient la liste complète des actions à votre disposition.
- **Bonnes Pratiques** : couvre un certain nombre de choses auxquelles vous voudrez peut-être réfléchir lors de l'utilisation d'Apache Hop.
- **Projets** : explique comment travailler avec des projets et des environnements.
- **VFS** : explique comment accéder aux ressources dans les 3 principales plateformes cloud : AWS, Azure et GCP.

Conclusion

- **Pipelines** : examine de plus près les différents aspects de la création et de l'exécution des pipelines, et contient la liste complète des transformations à votre disposition.
- **Workflows** : examine de plus près les différents aspects de la création et de l'exécution des workflows, et contient la liste complète des actions à votre disposition.
- **Bonnes Pratiques** : couvre un certain nombre de choses auxquelles vous voudrez peut-être réfléchir lors de l'utilisation d'Apache Hop.
- **Projets** : explique comment travailler avec des projets et des environnements.
- **VFS** : explique comment accéder aux ressources dans les 3 principales plateformes cloud : AWS, Azure et GCP.
- **Journalisation** : explique comment configurer Hop pour votre niveau de journalisation désiré et la plateforme cible.