

Formation MS-BI

Intégration de données avec SSIS

2022 - 2023

Amaury LAVERGNE

TechLead
Microsoft Data Analytics



lavergne@protonmail.com

[1lavergne@github.io](https://github.com/1lavergne)



Présentation de SSIS

Qu'est-ce qu'un ETL ?

QU'EST-CE QU'UN ETL ?

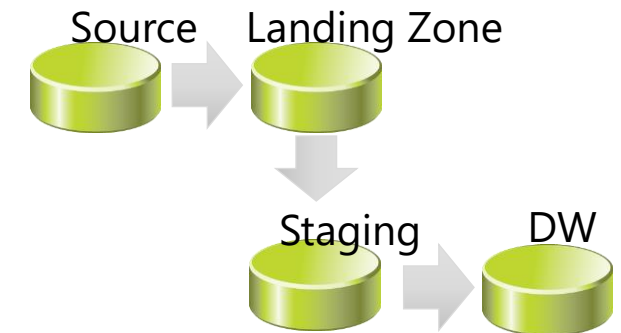
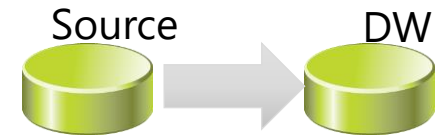
- Un ETL est une solution permettant d'effectuer des synchronisations massives d'information d'une source de données vers une autre.
- Un ETL effectue les actions suivantes :
 - › Extract → extrait la donnée depuis des sources hétérogènes
 - › Transform → traite la données en fonction de règles de gestion
 - › Load → charge la données à la destination voulue



Qu'est-ce qu'un ETL ?

ARCHITECTURES COMMUNES

- ETL en **une seule** étape :
 - › Les données sont transférées directement de la source à l'entrepôt de données
 - › Transformations et validations lors de l'extraction
- ETL en **deux** étapes :
 - › Les données sont stockées dans une étape intermédiaire
 - › Transformations et validations à la volée ou lors de l'extraction
- ETL en **trois** étapes :
 - › Les données sont extraites rapidement dans une zone tampon puis passent par une zone de traitement
 - › Les transformations et la validation peuvent se produire tout au long du flux de données



Qu'est-ce que SSIS ?

SQL SERVER BI

SQL Server BI est un environnement complet de traitement et de visualisation des données.

- Regroupe traditionnellement les trois outils :
 - › SQL Server Integration Services
 - › SQL Server Analysis Services
 - › SQL Server Reporting Services
- Progressivement complété / remplacé par la **Data Platform** (SQL Server, Azure Data Factory ...) et de la **Power Platform** (Power BI, Power Apps ...)

Qu'est-ce que SSIS ?

SQL SERVER BI

Microsoft **SQL Server Integration Services** est une plateforme qui permet de créer des solutions de transformation et d'intégration de données.

- › C'est l'**ETL** distribué par Microsoft
- › Installé en temps que composant de SQL Server
- › Moteur de flux de contrôle:
 - › Ressources d'exécution et support opérationnel pour les flux de données
 - › Permet l'exécution successive de de tâches indépendantes
- › Moteur de flux de données:
 - › Architecture 'pipeline' pour le traitement de flux de données en mémoire
 - › Actions de transformations successives appliquées à un même flux de données

Qu'est-ce que SSIS ?

PROJET ET PACKAGES

- SSIS peut être déployé selon de modes :
 - › Mode projet : plusieurs packages sous formes de fichiers déployés dans un seul projet.
 - › Mode package : les packages sont déployés et gérés individuellement dans un catalogue stocké en base.

Qu'est-ce que SSIS ?

ENVIRONNEMENT

- Les packages SSIS sont développés l'environnement **SQL Server Data Tools**
 - › Même outils que pour SSAS et SSRS
 - › Installation indépendante ou intégré à Microsoft Visual Studio 2017
 - › Peut-être connecté à un gestionnaire de version (Azure Dev OPS, Git...)
- A partir de Visual Studio 2019, SSIS est disponible sous forme d'extension à installer dans l'IDE.



RELEASE 15.8.0

Microsoft SQL Server Data Tools

Qu'est-ce que SSIS ?

ENVIRONNEMENT

- L'interface de **SSDT** se compose des éléments suivants :
 - › Explorateur de solutions
 - › Volet Propriétés
 - › Espace de conception des flux de contrôle
 - › Espace de conception des flux de données
 - › Onglet Paramètres
 - › Espace de conception des gestionnaires d'événements
 - › Explorateur de packages
 - › Volet gestionnaires de connexions
 - › Volet variables
 - › Boîte à outils SSIS

Qu'est-ce que SSIS ?

SQL SERVER MANAGEMENT STUDIO

- SQL Server Management Studio (SSMS) est un environnement intégré pour la gestion des infrastructures SQL (SQL Server, Azure SQL Database).
- SSMS fournit des outils permettant de configurer, de superviser et d'administrer des instances de SQL Server et des bases de données.
- SSMS permet de Utilisez SSMS pour déployer, superviser et mettre à niveau les composants de la couche Données (base de données, SSIS, SSAS), ainsi que pour créer des requêtes et des scripts.

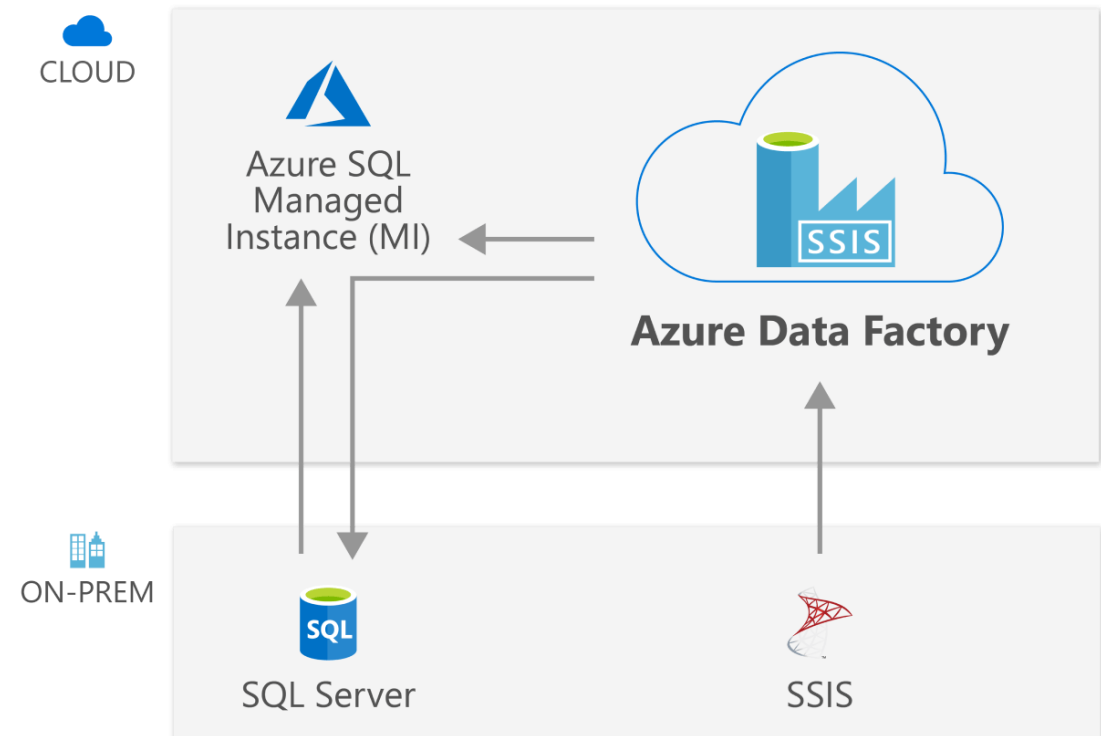
The logo for Microsoft SQL Server Management Studio (SSMS) is displayed on a dark rectangular background. The word "Microsoft" is in a smaller, white, sans-serif font at the top. Below it, "SQL Server Management Studio" is written in a larger, white, sans-serif font.

Microsoft
SQL Server Management Studio

Qu'est-ce que SSIS ?

AZURE DATA FACTORY

- Azure Data Factory est le « successeur » de SSIS
 - › Solution d'intégration « Low-code » déployé sur le cloud managé Azure
 - › Permet la migration des packages SSIS vers Azure
 - › ETL autonome avec plus de 90 connecteurs intégrés



Ressources

- Installer SSDT :
 - › <https://docs.microsoft.com/fr-fr/sql/ssdt/download-sql-server-data-tools-ssdt>
- Installer SSMS
 - › <https://docs.microsoft.com/fr-fr/sql/ssms/download-sql-server-management-studio-ssms>
- Documentation SSIS
 - › <https://docs.microsoft.com/fr-fr/sql/integration-services/docs.microsoft.com/fr-fr/sql/integration-services>

Structure d'un projet SSIS

Projet SSIS



PROJET

- On utilise Visual Studio / SSDT pour créer un projet SSIS.
 - › Un projet est contenu dans une solution ; une solution peut contenir plusieurs projets.
 - › Habituellement le projet SSIS, le projet SQL et les éventuels projets SSAS / SSRS sont contenus dans la même solution.
- Un projet SSIS se compose des éléments suivants :
 - › Les paramètres du projet
 - › Les connexions partagées : pour lire ou écrire de la données
 - › Les packages : un package est un ensemble de tâches
 - › Les composants de package : des parties de packages réutilisées dans des packages complets

Projet SSIS

PARAMÉTRAGE PROJET

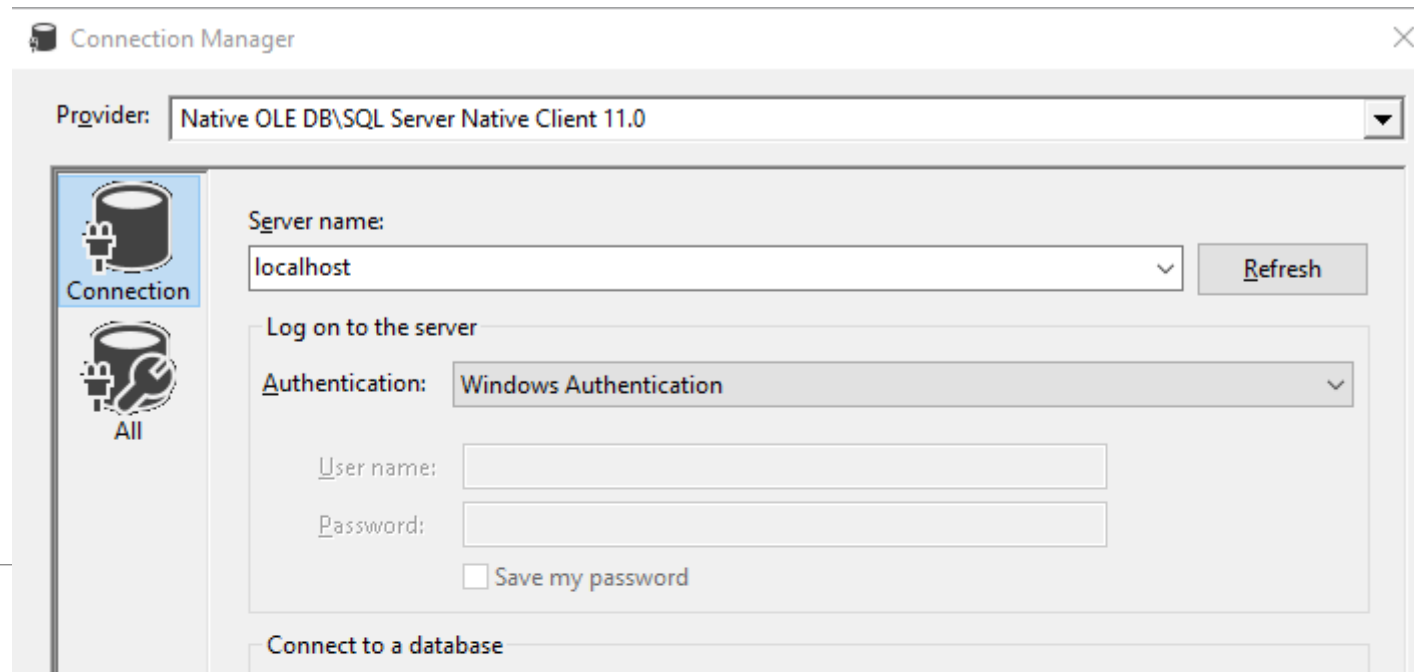
- Les paramètres projets sont :
 - › Définis dans un seul fichier XML
 - › Partagés par tout le projet
 - › Accessibles par tous les objets du projet uniquement en lecture seule

Name	Data type	Value	Sensitive	Required	Description
 DbConnectionString	String	*****	True	True	
 SourceFilePath	String	D:\OneDrive - UMANIS\Tra...	False	True	

Projet SSIS

CONNEXIONS PARTAGÉES

- Une connexion partagée correspond à la **définition de la méthode de connexion** à une source de données.
 - › Enregistrée dans un fichier XML
 - › Accessible par tous les packages du projet



Implémenter une solution SSIS

PACKAGE

- Un package est l'élément qui sera exécuté dans l'ETL. Le package contient l'ensemble des actions à appliquer pour transformer la données.
 - › Ces actions sont appelées des « tâches »
 - › Les tâches sont organisées dans un flux de contrôle
 - › Les packages peuvent être exécutés de manières successives ou indépendantes

Projet SSIS

PACKAGE



- Chaque package est autonome et dédié à une action ou un ensemble d'actions précises.
- Un package se compose :
 - › D'un flux de contrôle
 - › D'un ou plusieurs flux de données
 - › De connexions de données
 - › De variables
 - › De paramètres (projet et package)
 - › D'un gestionnaire d'évènements

Implémenter une solution SSIS
















FLUX DE CONTRÔLE

- Le flux de contrôle est la « zone d'exécution » du package.
- Il se compose de plusieurs tâches liées entre elles de manière logique.
- Les tâches peuvent être de différentes natures :
 - › Appelle l'exécution d'autres packages.
 - › Exécuter des scripts C# ou VB
 - › Exécuter des requêtes SQL
 - › Traiter un flux de données (« Data Flow »)
 - › ...

▲ Favorites

-  Tâche de flux de données
-  Tâche d'exécution de requêtes SQL

▲ Common

-  Tâche de profilage des données
-  Tâche de script
-  Tâche de service Web
-  Tâche de système de fichiers
-  Tâche de traitement SQL Server Analysis Services
-  Tâche d'exécution de package
-  Tâche d'exécution de processus
-  Tâche d'expression
-  Tâche d'insertion en bloc
-  Tâche du système de fichiers Hadoop
-  Tâche Envoyer un message
-  Tâche FTP
-  Tâche Hadoop Hive
-  Tâche Hadoop Pig
-  Tâche XML

Implémenter une solution SSIS

FLUX DE CONTRÔLE

- Les connexions entre deux tâches dépendent du résultat de l'exécution de la première tâche :
 - › Contrainte : succès, échec, achèvement
 - › Expression : évaluation d'une expression logique
 - › Contrainte et/ou expression
- Une tâche qui dépend de plusieurs tâches est enclenché :
 - › ET : lorsque toutes les contraintes de précedence sont validé
 - › OU : dès qu'une contrainte de précedence est validée

Éditeur de contrainte de précedence

Une contrainte de précedence définit le flux de travail entre deux exécutoires. Elle peut être basée sur une combinaison des résultats d'exécution et l'évaluation d'expressions.

Options de contrainte

Opération d'évaluation : Expression et contrainte

Valeur : Succès

Expression : ...

Contraintes multiples

Si la tâche unique comporte plusieurs contraintes, vous pouvez choisir la manière dont celles-ci interagissent afin de contrôler son exécution.

☒ Opérateur logique AND. Toutes les contraintes doivent avoir la valeur True.



☐ Opérateur logique OR. Une contrainte doit avoir la valeur True.

Implémenter une solution SSIS
















FLUX DE CONTRÔLE – LES TÂCHES

- Tâche d'exécution de requêtes SQL
 - › Exécute une requête SQL (en lecture ou en écriture) sur une connexion
 - › La requête peut-être paramétrée ; le résultat peut-être stocké dans une variable
- Tâche de script
 - › Exécute un script en Visual C# ou en Visual Basic
 - › Le script peut accéder aux variables en lecture et en écriture
- Tâche d'exécution de processus
 - › Exécute une instruction en ligne de commande Windows
- Tâche d'expression
 - › Evalue une expression logique SSIS
 - › L'expression accède aux variables en lecture et en écriture

▲ Favoris

-  Tâche de flux de données
-  Tâche d'exécution de requêtes SQL

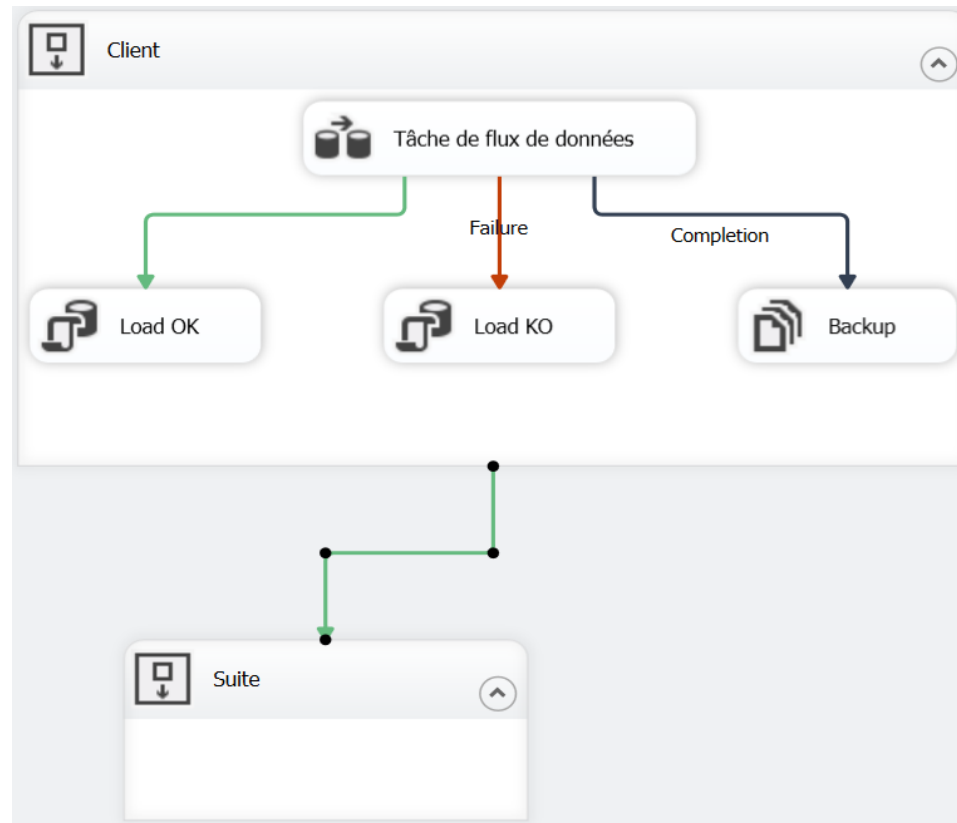
▲ Commun

-  Tâche de profilage des données
-  Tâche de script
-  Tâche de service Web
-  Tâche de système de fichiers
-  Tâche de traitement SQL Server Analysis Services
-  Tâche d'exécution de package
-  Tâche d'exécution de processus
-  Tâche d'expression
-  Tâche d'insertion en bloc
-  Tâche du système de fichiers Hadoop
-  Tâche Envoyer un message
-  Tâche FTP
-  Tâche Hadoop Hive
-  Tâche Hadoop Pig
-  Tâche XML

Implémenter une solution SSIS

FLUX DE CONTRÔLE – LES CONTENEURS

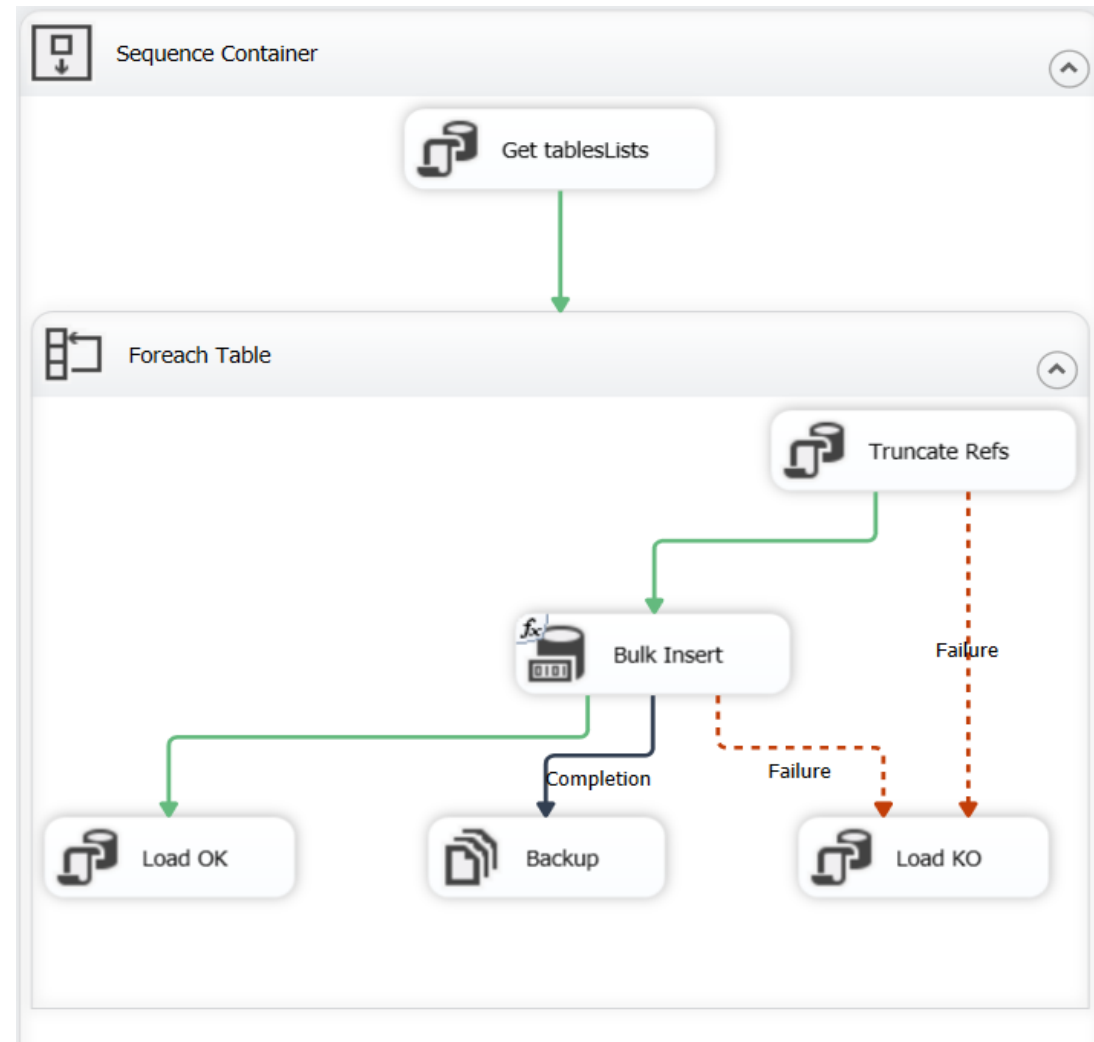
- Les conteneurs permettent de partitionner le flux de contrôle.



Implémenter une solution SSIS

FLUX DE CONTRÔLE – LES BOUCLES

- Les boucles permettent de parcourir plusieurs fois la même séquence de tâches.
 - › Boucle **FOR** : incrémente une variable
 - › Boucle **FOR-EACH** : assigne à une variable les valeurs successives d'une liste
























Implémenter une solution SSIS

FLUX DE DONNÉES

- Les flux de données (« **Data Flow** ») sont une partie essentielle de SSIS.
- Ils permettent de gérer le transfert et la transformation des données d'une source à une destination.
 - › Une ou plusieurs sources de données
 - › Des tâches de transformations liées entre elles logiquement
 - › Une ou plusieurs destinations de données

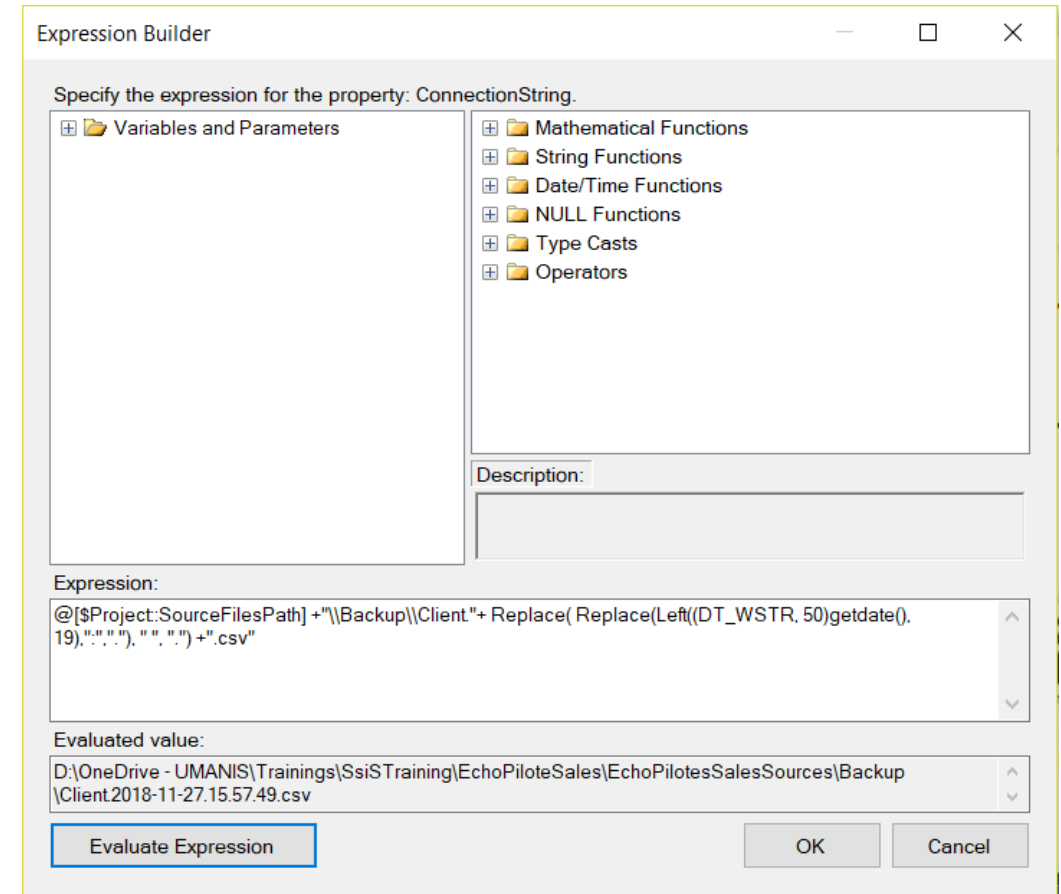
Common

-  Agrégation
-  Colonne dérivée
-  Commande OLE DB
-  Composant Script
-  Conversion de données
-  Destination de diffusion des données en continu
-  Destination du fichier HDFS
-  Destination ODBC
-  Dimension à variation lente
-  Distributeur de données équilibrées
-  Fractionnement conditionnel
-  Fusionner
-  Jointure de fusion
-  Multidiffusion
-  Nombre de lignes
-  Recherche
-  Source du fichier HDFS
-  Source OData
-  Source ODBC
-  Trier
-  Unir tout

Implémenter une solution SSIS

LES EXPRESSIONS

- Les expressions établissent des valeurs dynamiquement :
 - › Propriété, critère de fractionnement conditionnel
 - › Valeur de colonne dérivée
 - › Contrainte de précedence
- Elles sont basées sur la syntaxe d'expression SSIS :
 - › Peut inclure des variables et des paramètres
 - › Saisissez des expressions à l'aide de l'Expression Builder








Implémenter une solution SSIS

LES VARIABLES ET PARAMÈTRES

- Une variable permet de stocker en mémoire une valeur. Elle peut être de différents types :

- › String
- › Integer
- › Object
- › Date

Name	Scope	Data type	Value	Expression	
 GetTablesList	LoadR...	String	select distinct
 sourceFile	LoadR...	String	path...		...
 tableName	LoadR...	String	Toto		...
 tablesList	LoadR...	Object	System.Object		...
 TruncateTable	LoadR...	String	truncate table ...	"truncate table " + @[User::tableName]	...

- Il existe des variables :
 - › Systèmes : avec des valeurs prédéfinies correspondants au système
 - › Packages : avec des valeurs prédéfinies correspondants au package en cours
 - › Utilisateurs : avec des valeurs assignées lors de l'exécution

Implémenter une solution SSIS



LES VARIABLES ET PARAMÈTRES

- La valeur d'une variable utilisateur est définie avec une tâche :
 - › Tâche d'expression
 - › Tâche de script (C-Sharp, VB)
 - › Tâche SQL Server
 - › ...
- Les variables peuvent être consultées par tous les éléments du projet :
 - › Les tâches du flux de contrôles
 - › Les tâches du flux de données
 - › Les expressions
 - › ...

Implémenter une solution SSIS

LES VARIABLES ET PARAMÈTRES

- Paramètres de package
 - › Accessibles comme des variables mais en lecture seule
 - › Initialisés à l'appel du package
- Paramètres de projet
 - › Partagés par tout le projet
 - › Accessibles comme des variables mais en lecture seule
 - › Initialisés par un fichier ou au lancement du projet

Name	Data type	Value	Sensitive	Required	Description
 DbConnectionString	String	*****	True	True	
 SourceFilePath	String	D:\OneDrive - UMANIS\Tra...	False	True	

Implémenter une solution SSIS

LES CONNEXIONS

- Une connexion à une source ou à une destination de données se compose de :
 - › Un connecteur / fournisseur (ADO.NET, OLE DB, fichier plat ...)
 - › Une chaîne de connexion
 - › Des droits d'authentification
- Une connexion peut être déclarée au niveau du projet ou du package :
 - › Les gestionnaires de connexions au niveau du projet peuvent être partagés entre plusieurs packages
 - › Les gestionnaires de connexions au niveau du package n'existent que dans ce package

Implémenter une solution SSIS

LES CONNEXIONS

- Certains connecteurs sont natifs :
 - › SQL Server
 - › OLE DB
 - › ...
- D'autres connecteurs doivent être installés sur le serveur qui exécute l'instance SSIS :
 - › Excel
 - › MySQL
 - › ...

Implémenter une solution SSIS

LES CONNEXIONS – SOURCES DE DONNÉES

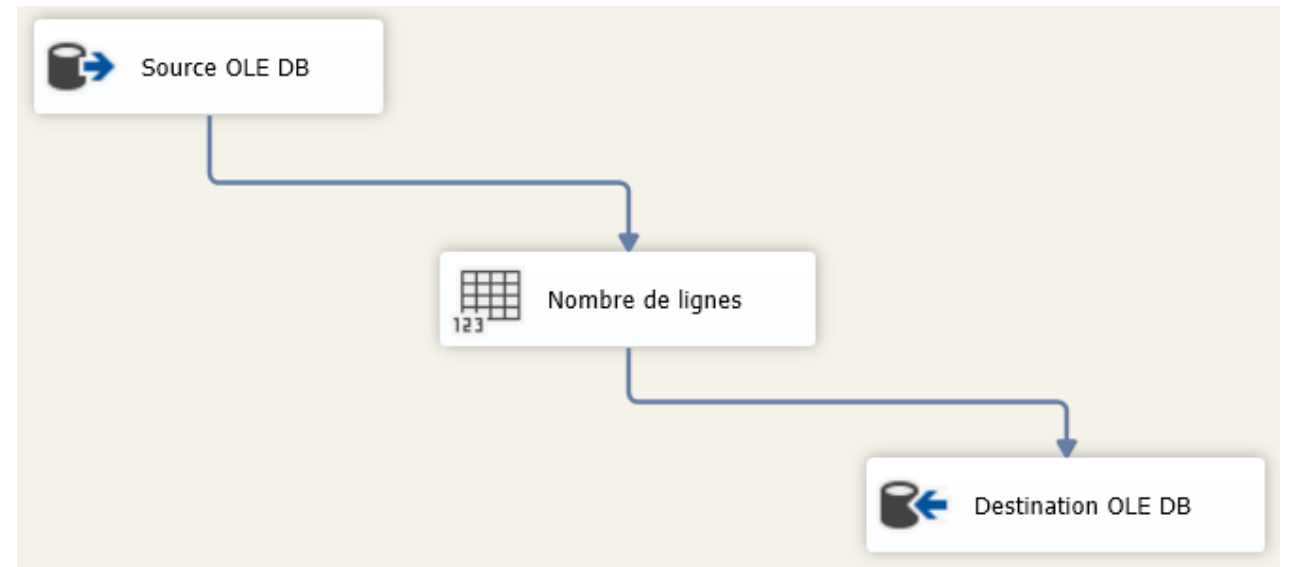
- La source des données pour un flux de données:
 - › Gestionnaires de connexions
 - › Synchronisation des données entre la source et le flux
- Nombreuses sources prises en charge:
 - › Fichier plat
 - › Base de données
 - › Source personnalisée
 - › ...



Implémenter une solution SSIS

LES CONNEXIONS – DESTINATION DE DONNÉES

- Point de terminaison pour un flux de données :
 - › Gestionnaires de connexions
 - › Synchronisation des données entre le flux et la destination
- Types de destination multiples:
 - › Base
 - › Fichier
 - › SQL Server Analysis Services
 - › Rowset
 - › ...



Déboguage

Débogage et résolution de problèmes

DÉBOGAGE

- Débogage pendant le développement
 - › Observer le nombre de lignes et les résultats des tâches
 - › Afficher les événements dans la fenêtre sortie et l'onglet résultats de progression/exécution
 - › Exécution pas-à-pas du package
 - › Suivre les valeurs des variables
 - › Afficher les données dans le flux de données
 - › Points d'arrêts (Breakpoints)
- Débogage dans l'environnement de production
 - › Afficher les journaux d'exécution du package
 - › Créer un fichier de vidage

Débogage et résolution de problèmes

SUIVI D'EXÉCUTION

- L'exécution de package est une séquence d'événements générés par des tâches et des conteneurs
- Pendant le débogage, les événements sont affichés:
 - › Dans l'onglet résultats de progression/exécution
 - › Dans la fenêtre de sortie

Débogage et résolution de problèmes

SUIVI D'EXÉCUTION – POINTS D'ARRÊTS

- Les points d'arrêts (« Breakpoints ») permettent d'interrompre l'exécution en fonction de conditions d'arrêts :
 - › Événement
 - › Compteur
- La valeur d'une variable peut être consultée lorsque l'exécution est interrompue par un point d'arrêt :
 - › Fenêtre « Local » : affiche toutes les variables dans la portée courante
 - › Fenêtre « Watch » : affiche les variables sélectionnées

Débogage et résolution de problèmes

SUIVI D'EXÉCUTION – VISIONNEUSE DE DONNÉES

- La visionneuse de données permet de visualiser la donnée en transit dans un flux de données.
 - › Activer la visionneuse de données sur les chemins de flux de données
 - › Afficher les données au fur et à mesure qu'elles passent par le flux de données
 - › Copier les données pour plus d'investigations

Débogage et résolution de problèmes

LOG

- SSIS embarque un système de gestion des LOGs pour enregistrer le déroulé d'une exécution.
 - › Windows Event Log
 - › Text file
 - › XML file
 - › SQL Server
 - › SQL Server Profiler

Débogage et résolution de problèmes

LOG

- Log Schema

- › StartTime
- › EndTime
- › DataCode
- › Computer
- › Operator
- › MessageText
- › DataBytes
- › SourceName
- › SourceID
- › ExecutionID

- Log Events

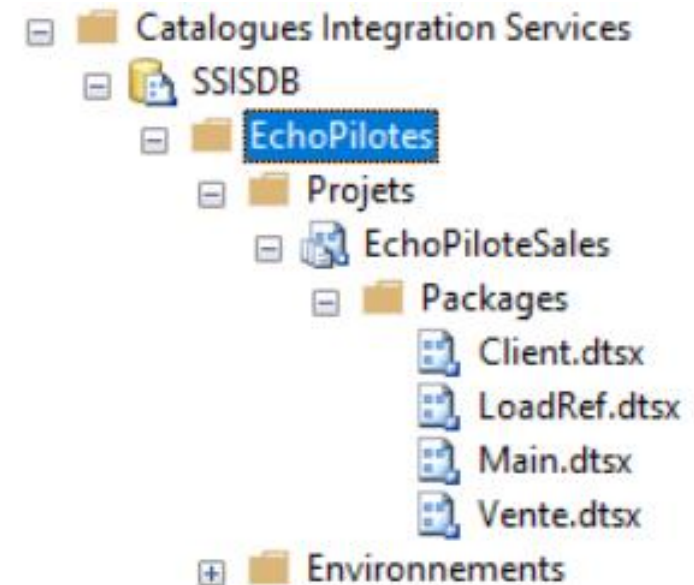
- › OnError
- › OnExecStatusChanged
- › OnInformation
- › OnPipelinePostComponentCall
- › OnPipelinePostEndOfRowset
- › OnPipelinePostPrimeOutput
- › OnPipelinePreComponentCall
- › OnPipelinePreEndOfRowset
- › OnPipelinePrePrimeOutput
- › OnPipelineRowsSent
- › OnPostExecute
- › OnPreExecute
- › OnPreValidate
- › OnProgress
- › OnQueryCancelled
- › OnTaskFailed
- › OnVariableChangedValue
- › OnWarning
- › Diagnostic
- › DiagnosticEX

Déploiement

Déploiement et exécution

DÉPLOIEMENT

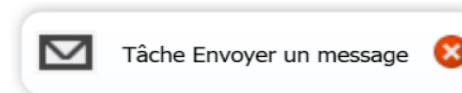
- Déploiement en mode en package
 - › Packages stockés sous forme de fichiers
 - › Exécutés avec 'dtexec'
- Déploiement en mode SQL
 - › Packages stockés dans un catalogue en base SQL Server
 - › Appelés par le SQL Agent



Usages complémentaires

Usages complémentaires

- Ordonnancement des packages :
 - › La listes des packages peut être stockées en base et chaque package est appelé dynamiquement.
- Traitement SSAS
 - › Utiliser la tâche «Tâche de traitement SQL Server AnalysisServices » pour traiter un cube SSAS (en intégralité ou partiellement).
 - › La tâche «Tâche DDL d'exécution SQL Server AnalysisServices » permet d'exécuter des scripts XMLA sur un cube (création dynamique de partitions, traitement des partitions rechargées ...).
- Envoie de mails
 - › La tâche «Tâche Envoyer un message » permet d'envoyer des mails à partir d'un server SMTP (suivi de l'alimentation, rapport de fin de traitement, ...).
 - › Le server SMTP doit être configuré indépendamment de SSIS.



Bonne pratiques

Pas de look-up en FULL ... C'est tout

Bonnes pratiques

RÉUTILISER L'EXISTANT

- Capitaliser sur les éléments existants :
 - › Utiliser des templates de packages
 - › Utiliser des connexions partagées (plutôt que des connexions de packages)
- Effectuer en amont ce qui peut l'être :
 - › Privilégier une transformation dans la requête SQL source plutôt que dans une tâche du Data Flow
 - › Faire les jointures en SQL plutôt que d'utiliser la tâche de look-up

Bonnes pratiques

OPTIMISER LES PERFORMANCES DU DATA-FLOW

- Optimiser les requêtes:
 - › Sélectionnez uniquement les lignes et les colonnes dont vous avez besoin
- Évitez le tri inutile:
 - › Utiliser les données pré-triées si possible
 - › Définir la propriété «IsSorted», le cas échéant
- Configurer les propriétés des tâches de flux de données:
 - › Taille du tampon
 - › Emplacement de stockage temporaire
 - › Parallélisme
 - › Mode optimisé

Amaury LAVERGNE

TechLead
Microsoft Data Analytics
JEMS Ouest

lavergne@protonmail.com

[1lavergne@github.io](https://github.com/1lavergne)