Chapitre 5
Alimenter l'entrepôt de données avec SSIS

1. Découverte de SSIS

Au cours des chapitres précédents, vous avez appris à modéliser un entrepôt de données. L'idée était de faire abstraction des sources de données disponibles dans votre société. Au cours de ce chapitre, vous allez apprendre et comprendre comment va se réaliser la remontée des données du système source vers un entrepôt de données. La principale difficulté est que celui-ci dispose d'une modélisation dimensionnelle conforme, très éloignée de la structure de vos données actuelles.

Dans la gamme SQL Server, l'outil qui va permettre de réaliser le chargement de ces données est SQL Server Integration Services (SSIS).

SSIS a deux aspects :

– Un aspect classique avec une logique de flux de tâches, organisées par des règles de précédence. Cet aspect est appelé **Flux de contrôles**.
– Un aspect plus spécifique au décisionnel, avec une logique purement ETL. Cet aspect est appelé **Flux de données**.

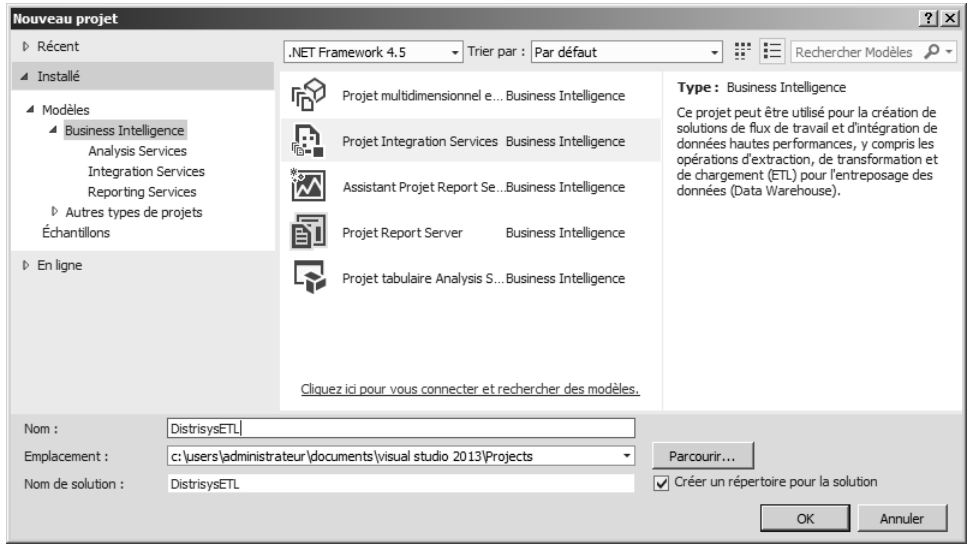
On peut utiliser SSIS sans pour autant faire de l'ETL. Par exemple, vous pouvez vous servir de SSIS pour exécuter des tâches de maintenance de bases de données, pour lancer une suite de batch un peu complexe ou pour réaliser de la réplication de données.

Toutefois, SSIS est aussi un ETL. Le monde de l'ETL a ses codes et ses règles issues de ces quinze dernières années. L'objectif du chapitre, au-delà de la compréhension de ce qu'est l'outil SSIS, est de vous faire découvrir certaines de ces pratiques bien spécifiques au monde du décisionnel. Des pratiques auxquelles SSIS est assez bien adapté.

Un peu comme pour toute la gamme SQL Server, le développement des flux se fera sous SQL Server Data Tools (SSDT). On utilisera en revanche SQL Server Management Studio pour l'administration et l'exploitation.

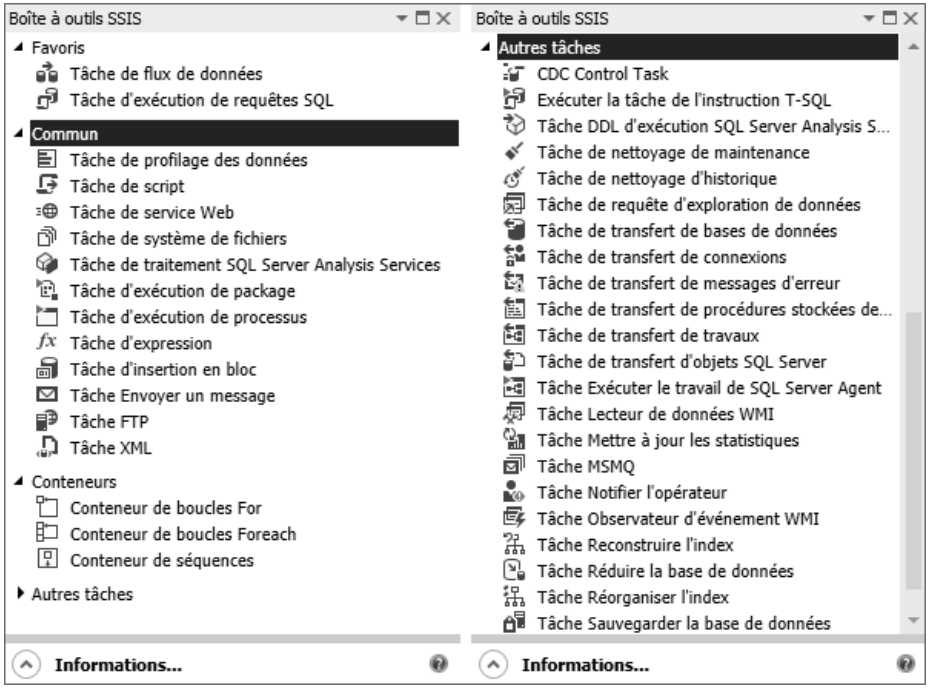
Découvrons ensemble dès à présent l'interface de développement :

- Ouvrez SSDT.
- Cliquez dans la barre de menu sur **Fichier - Nouveau - Projet**.
- Dans la fenêtre **Nouveau projet**, sélectionnez **Projet Integration Services**, puis saisissez le nom et l'emplacement du projet comme ci-dessous :



Le projet s'ouvre par défaut sur l'onglet **Flux de contrôle** d'un package vide. Un package est un fichier au format XML à l'extension **.dtsx**.

- Sur le côté gauche, ouvrez la boîte à outils SSIS pour découvrir les objets du flux de contrôle disponibles.



Boîte à outils SSIS

Les tâches disponibles donnent une assez bonne idée du rôle que l'on pourrait faire jouer à SSIS et de ses possibilités : connexion à un service web, exécution de requête SQL, exécution d'application, écriture et exécution de scripts, connexion à un serveur FTP, tâche de traitement de SSAS, tâche de sauvegarde de la base de données…

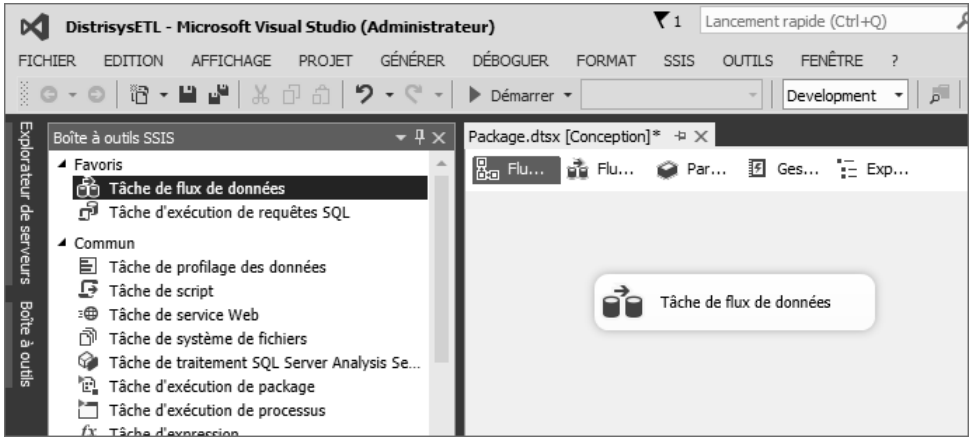
Dans un flux décisionnel, les tâches de flux de contrôle vont avoir des fonctions de support et d'orchestration, mais ce ne sont pas ces tâches qui vont faire à proprement parlé le chargement des données.

Remarque

Attention, dans le monde du décisionnel, un entrepôt de données ne se charge pas avec de simples requêtes SQL. Vous verrez que les exigences de traçabilité et de maintenance de tels flux sont trop élevées pour que des requêtes SQL remplissent ce rôle correctement.

Le chargement de données va se réaliser avec la tâche de flux de données. Découvrons cet aspect du produit :

- Glissez et posez la tâche de flux de données dans la zone de travail centrale.

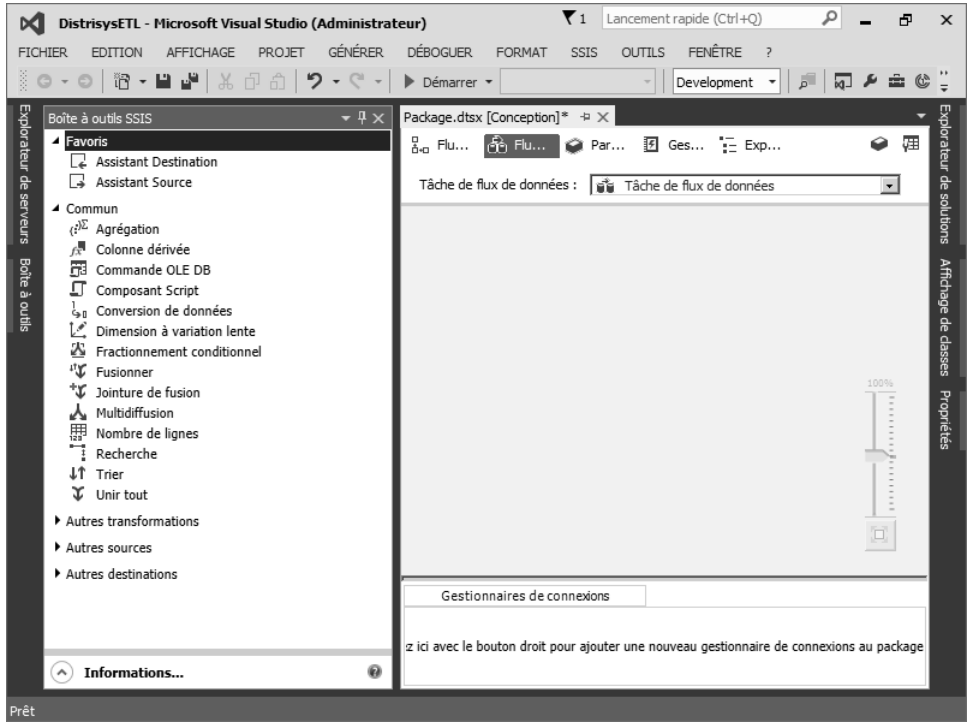


Ajout d'une tâche de flux de données

- Puis double cliquez sur la tâche de flux de données pour accéder à l'onglet **Flux de données**.

Vous noterez que la barre d'outils propose maintenant de nouvelles tâches organisées autour de trois thématiques :

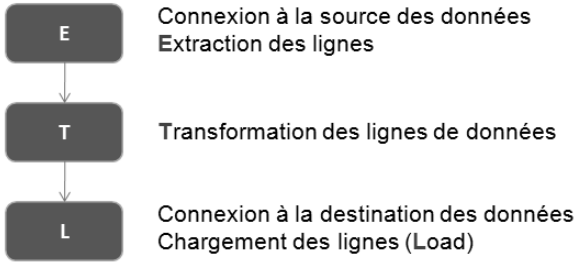
– Les tâches Sources
– Les tâches Transformations
– Les tâches Destinations



La boîte à outils de l'interface de flux de données de SSIS

En faisant glisser la tâche de flux de données, vous avez basculé l'interface en mode véritablement ETL. L'acronyme **ETL** signifie que le flux va être organisé en trois grandes phases :

– La phase **E** signifie qu'une tâche va se connecter à une source, pour en **Extraire** des lignes de données.
– La phase **T** signifie que ces lignes vont passer par des tâches de **Transformation** pour subir des tests, des validations ou des modifications.
– La phase **L** signifie que ces lignes, une fois traitées et transformées, vont être chargées (**Load** en anglais) dans la base de données destination.



Représentation schématique du déroulement d'un flux ETL

L'ensemble de ces phases va se dérouler uniquement en mémoire, d'où des gains de performance qui peuvent être substantiels par rapport au SQL, si on exploite correctement l'outil.

La barre d'outils à gauche organise les tâches disponibles dans SSIS par ces trois grandes phases ETL.

Dans la partie suivante, nous réaliserons un premier flux ETL pour comprendre le fonctionnement de SSIS.