Chapitre 3
Les bases de l'ETL

1. Présentation et installation de la solution de développement de ce livre

Une partie des exemples présentés dans cet ouvrage sont disponibles en téléchargement. Afin de pouvoir les exploiter, il est indispensable de disposer de l'environnement matériel et logiciel ci-dessous. Si ce n'est pas le cas, il est tout à fait possible de parcourir l'ouvrage et alors seule la partie pratique en sera limitée.

- une machine 4 cœurs, 8 Go de RAM, disque SSD
- SQL Server 2017 Developer/Enterprise Edition
- Visual Studio SSDT 2015

Cet ouvrage ne couvre pas les étapes d'installation de SQL Server ni de Visual Studio. Des références externes pourront être utilisées.

2. Les bases avant une première implémentation

2.1 Généralités

Avant de procéder au moindre développement, il est nécessaire de fixer quelques points de définition. Généralement, Integration Services est qualifié d'outil d'extraction, de transformation et de chargement de données, ce qui est juste, mais cette description simpliste limite la portée réelle de cette solution.

Un outil d'ETL complet se doit de fournir des fonctionnalités d'orchestration, c'est-à-dire d'ordonnancement des tâches, pour traiter tout le cycle d'intégration des données. Par exemple, avant de charger un fichier, il est fréquent de devoir le copier depuis un serveur, pour éventuellement, une fois les données récupérées, notifier tel ou tel opérateur, ou lancer un traitement externe.

Il doit également fournir un système de journalisation des exécutions qui permet de suivre la bonne avancée des traitements et de tracer les éventuelles erreurs, que ce soit pendant la phase de développement ou bien lorsque la solution est en production.

2.2 Le flux de contrôle : généralités

2.2.1 Définition

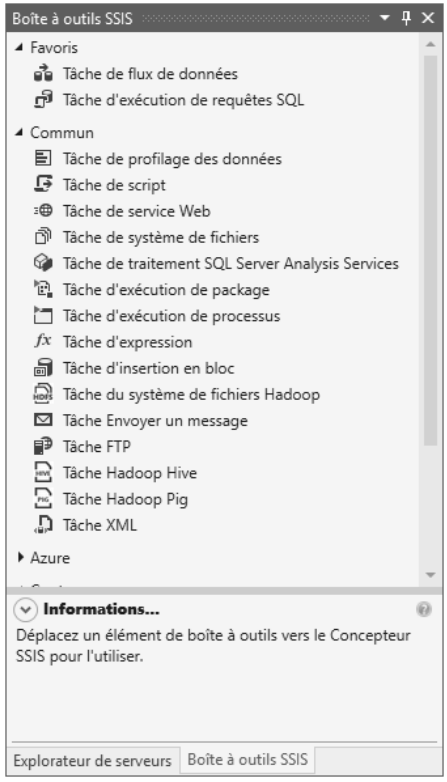
Le flux de contrôle (Control Flow) est ce que l'on peut désigner comme le point d'entrée, la couche externe d'un lot SSIS. C'est à partir de ce dernier que tout ce qui doit être exécuté dans un lot le sera. Il encapsule d'autres éléments comme l'indispensable flux de données (Data Flow), le Gestionnaire d'Évènements, les Paramètres. C'est lui qui offre une vision complète sur tout ce que réalise un lot SSIS en termes d'exécution. Il est donc le premier objet auquel les développeurs ont à faire. Il est possible de le comparer à une fonction Main() des langages de développement comme le C#.

Les principaux objets qui le composent sont appelés des Tâches et ce n'est pas par hasard, car le flux de contrôle peut également être perçu comme un ordonnanceur de tâches.

C'est tout naturellement que l'on trouve dans le flux de contrôle les structures de code que l'on peut voir dans la plupart des langages comme les boucles ou les expressions conditionnelles. Pour ces dernières, il ne faut pas parler de Tâches, mais respectivement de Conteneurs et de Contraintes de précédence.

2.2.2 Le concepteur de flux de contrôle

Dans Visual Studio, la boîte à outils, qui se trouve très probablement sur la gauche de l'écran, permet de visualiser deux des trois types d'objets mentionnés précédemment : les Tâches ainsi que les Conteneurs. Si la boîte à outils n'est pas visible, elle est accessible via le menu **SSIS - Boîte à outils SSIS** de la barre de navigation principale de Visual Studio.

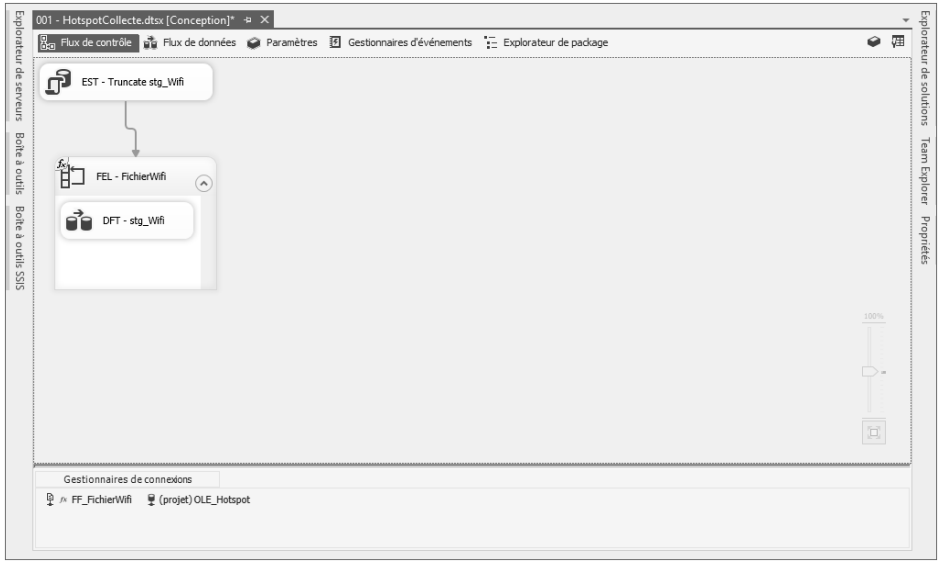


Le contenu de la boîte à outils change en fonction de la section où l'on se trouve dans le lot SSIS. Il faut ici veiller à être dans la section flux de contrôle pour avoir la même chose que sur l'image ci-dessus. C'est le premier onglet de la fenêtre de conception du lot SSIS.

La classification au travers des répertoires **Favoris**, **Commun**, **Azure**, **Conteneurs**, **Autres Tâches** et **Parties de package** permet une meilleure lisibilité. Le contenu de **Favoris** et **Commun** n'est pas immuable et il est tout à fait possible de l'adapter en fonction des habitudes de développement.

Dans le bas de la boîte à outils, une aide donne une brève description de la fonction de l'objet mis en surbrillance par un simple clic. Elle offre également la possibilité de chercher des exemples d'utilisation au travers de MSDN en cliquant sur **Rechercher des exemples**.

Les objets de la boîte à outils sont déplacés vers l'espace de conception au moyen d'un double clic ou bien d'un glisser-déposer. Il est possible de faire glisser plusieurs fois le même objet, dans ce cas Visual Studio va suffixer le nom par défaut d'un incrément. Il est conseillé de ne pas conserver les noms fournis par défaut, mais de décrire au moyen d'un texte court la fonction de chacune des tâches qui sont utilisées.



Le flux de contrôle ci-dessus propose un aperçu complet de tous les types d'objets utilisables et dont la fonctionnalité sera détaillée par la suite : Tâche (EST - Truncate stg Wifi, DFT - stg Wifi), Conteneur (FEL - FichierWifi) et Contrainte de précédence (les flèches). Il est accessible dans les fichiers de la solution téléchargeable sous le projet Chapitre 3 et son nom est 001 - HotspotCollecte.dtsx (se référer à la section Présentation et installation de la solution de développement de ce chapitre).

Pour des raisons de simplicité de lecture ainsi que par convention, il est préférable de débuter chaque développement SSIS par en haut et de le compléter verticalement vers le bas.

2.3 Le flux de données : généralités

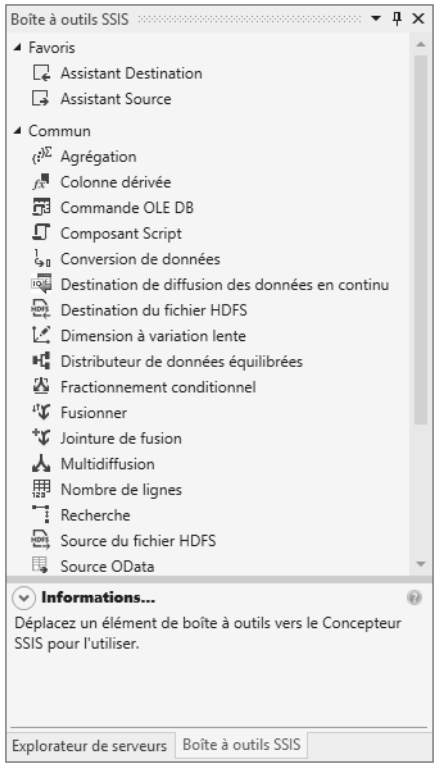
2.3.1 Objectifs

Le flux de données permet de faire transiter les données d'un point à un autre, d'une forme vers une autre, tout en réalisant des transformations sur ces dernières si nécessaire. Dit comme cela, il est facile de se rendre compte que cette définition du flux de données colle parfaitement à celle qui a été donnée de l'ETL au début de ce chapitre. Il faut donc considérer le flux de données comme la partie ETL de SSIS.

Le flux de données constitue une tâche particulière du flux de contrôle. Un flux de contrôle aura donc la possibilité d'exécuter plusieurs flux de données. Cette remarque met en exergue l'une des grandes différences qui existe entre flux de données et contrôle de flux de contrôle : le premier fait transiter des données tandis que le second ordonnance et exécute les tâches responsables de la gestion des données. Le flux de contrôle ne gère pas l'extraction, la transformation et le chargement des données directement : ce n'est que l'ordonnanceur de l'ETL. Cette distinction aura son importance lorsque la distinction entre les flèches visibles dans le flux de données et celles du flux de contrôle sera faite.

2.3.2 Le concepteur de flux de données

Tout comme pour le flux de contrôle, la réalisation d'un flux de données passe par l'utilisation d'une boîte à outils et d'une fenêtre de conception. Comme évoqué précédemment, le contenu de la boîte à outils sera fonction de l'endroit où l'on se trouve dans Visual Studio. Après s'être assuré de se trouver dans l'éditeur de flux de données, second onglet de la fenêtre de développement du lot SSIS, voici le contenu de la boîte à outils.



Ici, il n'est plus question de Tâche ou de Conteneur, mais uniquement de Composant. Ces derniers sont de quatre types : Azure, Source, Transformation et Destination. Comme pour le flux de contrôle, les sections Favoris et Commun sont personnalisables en fonction des habitudes de développement. Hormis pour le répertoire Azure, il faut identifier dans Source, Transformation et Destination, les grands principes de l'ETL.