**Collège de Bois de Boulogne**

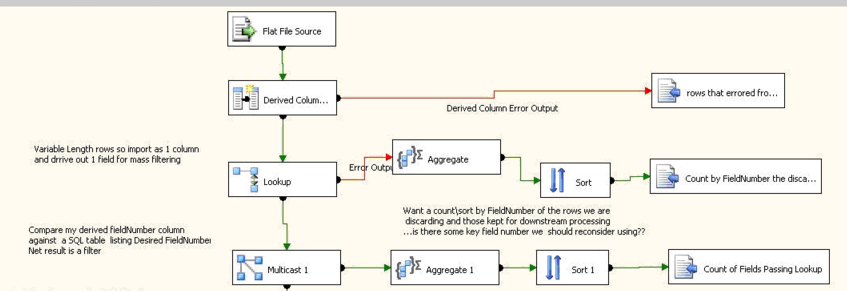
**420-BD4-BB (ingestion, capture de données)**

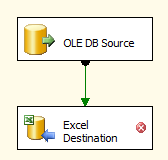
**Tp1: capture et nettoyage de données**

1. Format: individuel
2. Datasets: on a à notre disposition plusieurs sources de données sous des formats différents. Ceux que nous avons reçus sont aux formats csv, SQL et JSON. Il est important de savoir que les champs disponibles dans chacune des sources peuvent ne pas être les mêmes et que des données peuvent ne pas être présentes. De plus, il se peut que des données soient présentes et qu'une transformation sera nécessaire. Par exemple, il se peut que le genre d'un client soit mâle ou femelle dans une source mais décrit comme m ou f dans une autre source.

**Objectif**: dans la source de destination, on devra définir le schéma (modèle) approprié.

1. Pour faciliter la mise en place du processus, on devra en premier décrire le processus à l'aide d'une notation et d'un workflow. Dans ce cadre, on utilisera par exemple le symbolisme présent au niveau de SSIS. Quelques-uns des symboles sont les suivants:





1. Les objectifs du TP sont les suivants:
   1. Mettre en place un processus de capture et de transformation de données. Le schéma général peut être exprimé par



* 1. Développer le code nécessaire pour arriver à répondre aux besoins du client. Le choix de l'approche est basé sur un processus développé en python
     1. On peut utiliser des modules dédiés aux différents types de source. Voir les liens proposés
  2. Les sources considérées sont:
     1. Table Oracle/MySQL/SQL server
     2. Flux JSON
     3. Fichier Excel, csv

1. La destination sera une table MySQL ou autre éditeur approprié.

**Références**

1. Voir les liens portant sur les différents modules python qui peuvent être utilisés pour ce processus ETL.