Guide de démarrage rapide avec DataStudio Online Edition

www.data.fr

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73

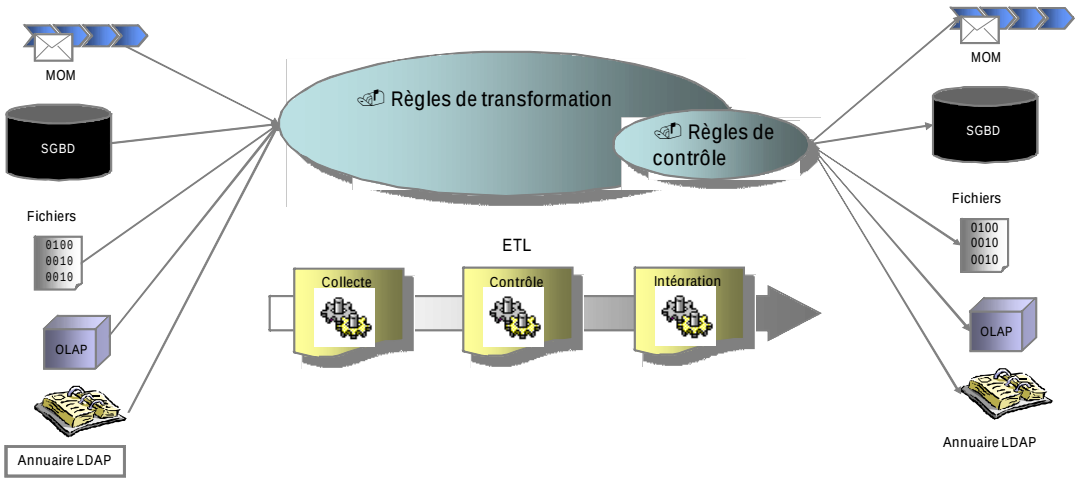
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

Introduction

Ce document vient en complément des films de démonstration disponibles sur le site web de data www.data.fr.

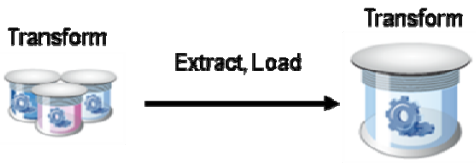
L'ETL

ETL est un sigle qui signifie E pour **Extraction** qui consiste en la lecture et l'interprétation de données depuis un support et un format d'enregistrement, T pour **Transformation** qui consiste en une ou plusieurs opérations sur les données extraites comme l'agrégation, le filtrage, la transcodification etc. et L pour **Loading** qui consiste au chargement de ces données lues et transformées vers un support et un format d'enregistrement souvent différent de ceux de la lecture.



La philosophie de DATA

La philosophie de DataStudio est d'utiliser les moteurs de transformation des bases de données et donc le langage SQL pour réaliser la partie **Transformation** de l'ETL, DataStudio se chargeant de faire l'**Extraction** et le **Loading** depuis ou vers les supports et formats de données. Les opérations d'ETL dans l'ordre sont donc E-L-T puisque DataStudio transfère les données nécessaires à la transformation avant d'exécuter la requête SQL.



La plupart des utilisations d'un ETL sont pour intégrer des données dans une base de données. Cette base peut donc aussi être la base de transformation.

La richesse et la simplicité du **langage SQL** permet toutes les transformations simplement et très efficacement même quand le volume de données se compte en millions de ligne. En plus, de nombreux exemples de requêtes et de nombreuses documentations sont disponibles sur le web pour les débutants.

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

Aide contextuelle

Pour une prise en main rapide, DataStudio offre une aide contextuelle complète avec de nombreux exemples d'utilisation. A tout moment l'appui sur la touche F1 affiche l'aide correspondant à l'écran actif.

Lors de la mise au point d'un traitement ETL dans le « Script editor » l'aide contextuelle (appui sur la touche F1) affichera la documentation de la fonction sous le curseur de texte.

Premiers pas

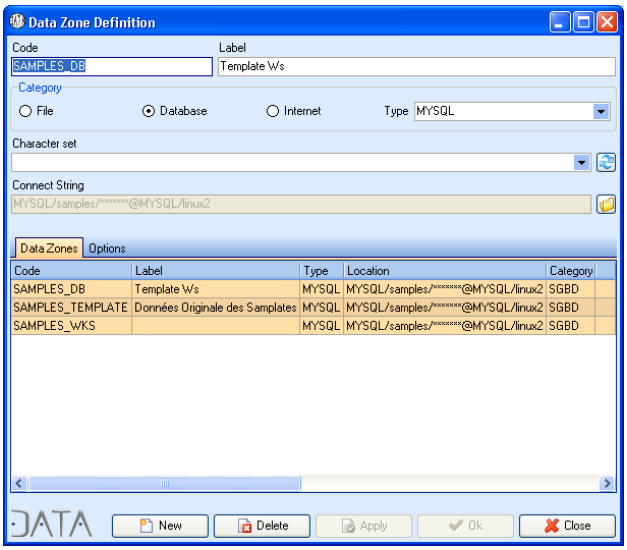
Le référencement

La première étape du traitement ETL de DataStudio est le référencement de la donnée. Elle permet à DataStudio de connaître les caractéristiques d'une source de données afin de pouvoir extraire ou charger correctement les données.

Une fois référencées, les données sont utilisables de la même façon que des tables dans une requête SQL.

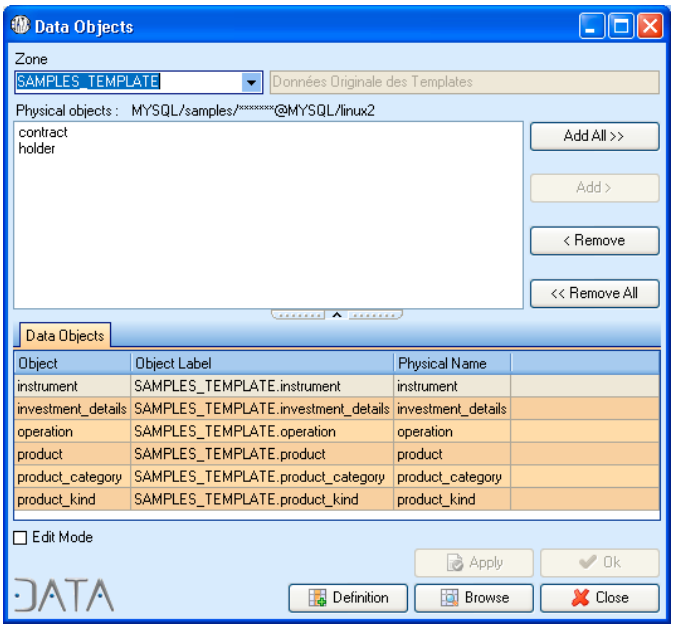
DataStudio distingue 2 niveaux de référencement, la **DataZone** et le **DataObjet**.

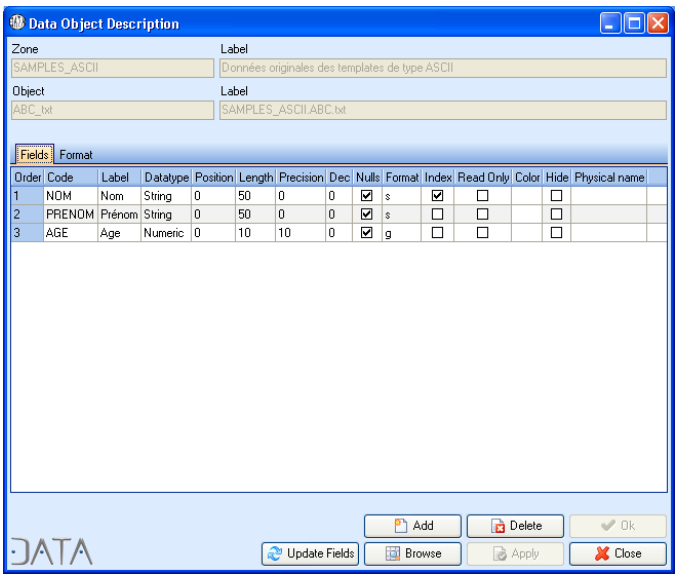
La DataZone est un nom logique qui correspond à un regroupement d'objets de même support et format. Par exemple : un répertoire de fichier ASCII, un répertoire de fichier Excel, un user d'une base de données Oracle, une URL LDAP.



Le DataObjet est un élément d'une DataZone. C'est un nom logique qui correspond à un accès de type « table » à un objet physique. De ce fait, un DataObjet pourra être utilisé dans une requête de la même façon qu'une table. Par exemple : Un fichier ASCII, une feuille Excel, une table Oracle, un tableau de propriétés LDAP.

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324





Le référencement se fait à l'aide des écrans correspondant aux menus :

- Component/Data/Data Zone
- Components/Data/Data Object.

Le traitement des données

Les données étant référencées, le traitement ETL ne consiste plus alors qu'à écrire au minimum une requête qui lit les données depuis l'objet source et écrit tout ou partie des données éventuellement transformées par la requête dans la destination. Par exemple :

Insert into TABLE_DESTINATION
select * from @ZONE_FICHIER.FICHIER1

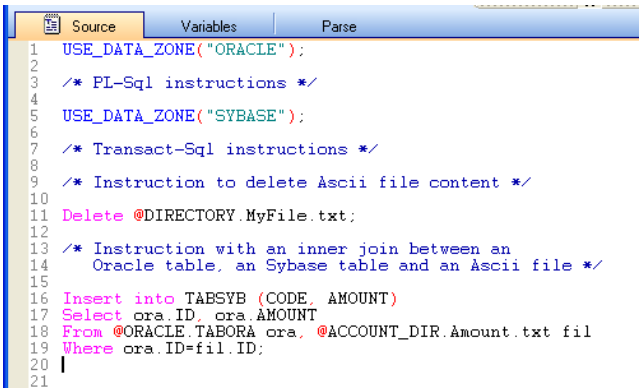
44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

Cette requête permet de lire les données du fichier 1 et de les insérer dans la table d'une base de données notée TABLE_DESTINATION.

Cette requête utilise la notation spécifique représentant des DataObjets (@ZONE_FICHIER.FICHIER1) dans les requêtes.

Dans un traitement ETL, il est possible d'exécuter plusieurs requêtes successives pour obtenir un traitement particulier.

Il existe de nombreuses fonctions spéciales pour tous les traitements impossibles en SQL (cf. documentation touche F1).



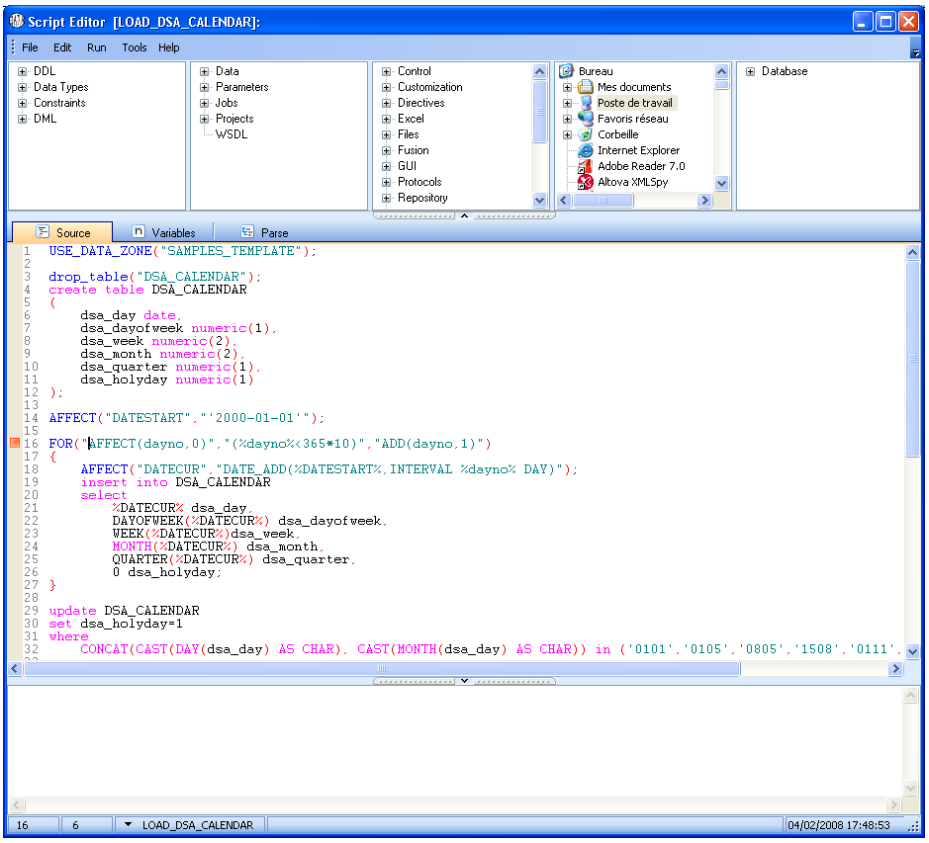
Exécution des traitements

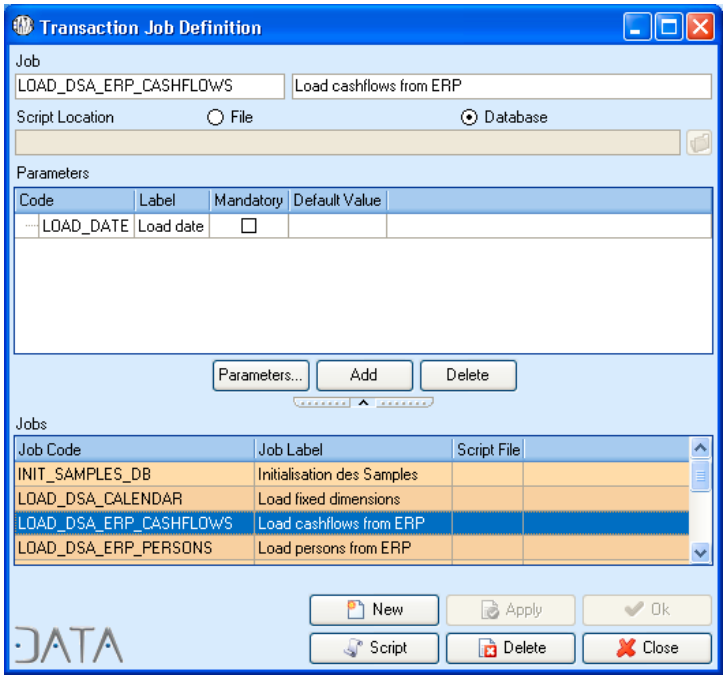
L'ETL DataStudio permet 3 niveaux de regroupement de traitement :

Le job de Transaction: C'est l'élément que l'on peut éditer avec le « Script Editor » 

pour décrire les traitements à la requête près. Les instructions sont enchaînées dans l'ordre et au gré des éventuelles conditions. La liste des jobs du référentiel est gérée par l'écran du menu Component/Transaction

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

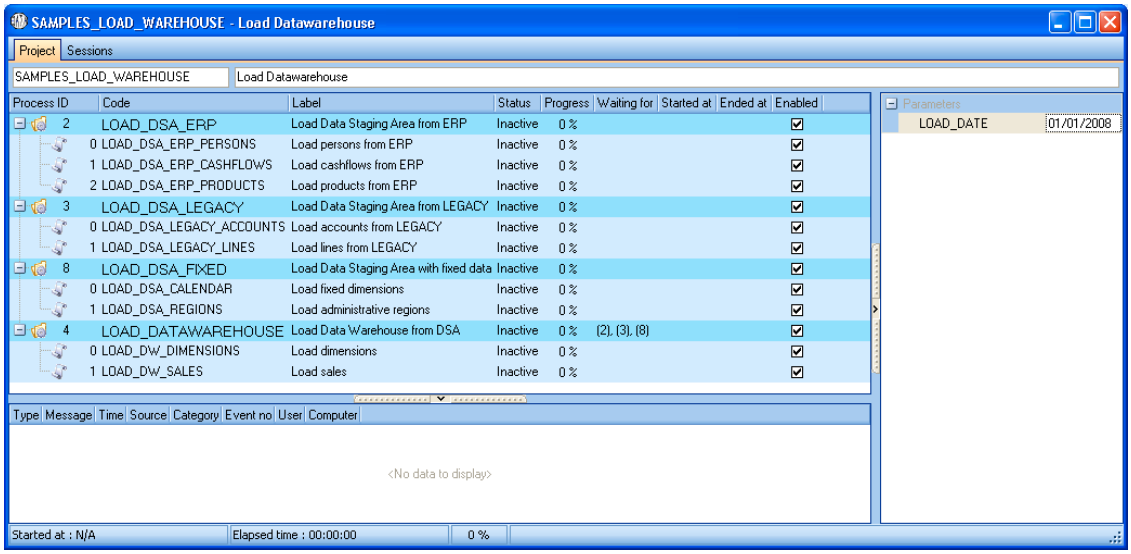




Le Folder : Il contient une liste de jobs à exécuter dans l'ordre. Par défaut, les folders s'exécutent en parallèle avec bien sûr la possibilité de pouvoir synchroniser un folder par rapport à un autre. Il est ainsi facile de construire un véritable arbre de traitement en quelques clicks de souris.

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

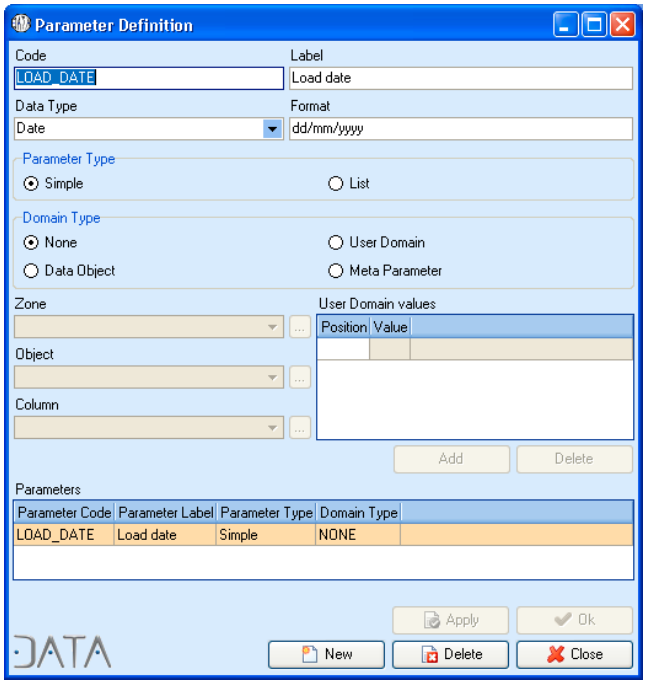
Le Projet : C'est l'élément visible et exécutable depuis l'extérieur de DataStudio. Il contient les folders et les règles de synchronisation entre eux (les waiting for). Il permet également de renseigner les éventuels paramètres d'entrée du traitement.



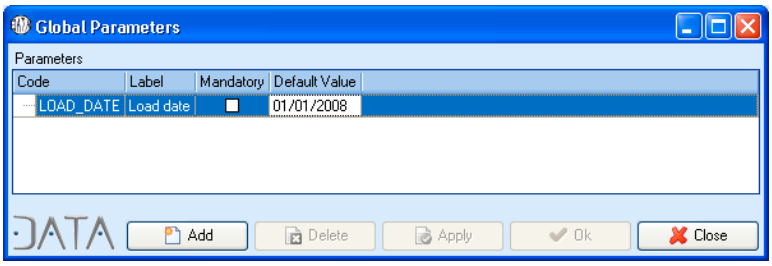
Les paramètres

DataStudio permet de définir des paramètres associés à des valeurs Global, User ou Projet. Si plusieurs valeurs sont définies pour le même paramètre, l'ordre décroissant des priorités est Projet, User, Global. Il est aussi possible de définir à l'intérieur d'un script des paramètres locaux. Ils sont alors prioritaires sur tous les précédents.

Dans la version Business de DataStudio, les paramètres peuvent être passés en ligne de commande à DataStudio Serveur. Ceci permet d'intégrer les traitements ETL à la chaîne des traitements informatiques de production.



44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324



Les écrans de gestion des paramètres sont :

- Menu Component/Parameters : définition des paramètres,
- Menu Administration/Settings/Global Parameters : définition de valeurs globales,
- Menu Administration/Settings/User Parameters : definition de valeur user,
- Ecran des projets pour la saisie des valeurs de projet.

Utilisation des autres modules de la plate-forme

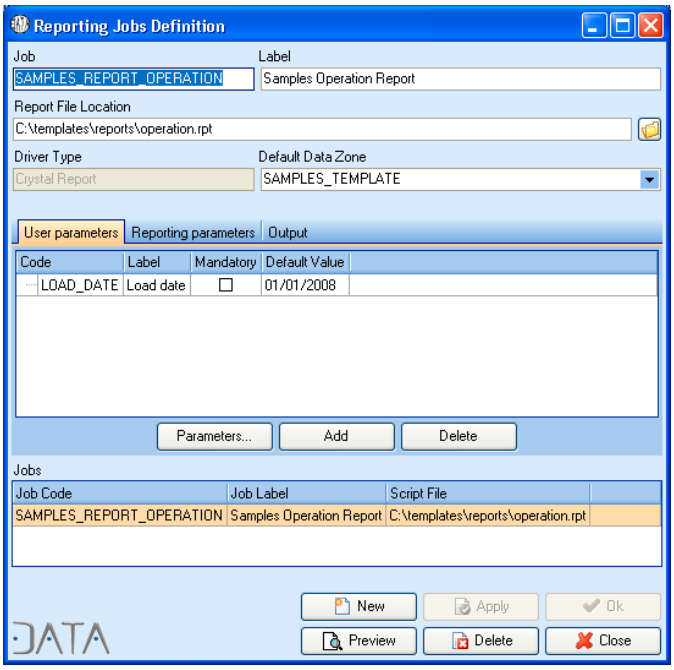
Reporting

DataStudio possède des jobs de type reporting qui consistent en une intégration du moteur de reporting de BO Crystal Reports en runtime.

Les rapports sont conçus avec le designer de Crystal Reports et exécutés avec DataStudio qui passe aux rapports les paramètres, les connexions (Le designer de Crystal Reports doit être acquis séparément).

Le reporting intégré à l'ETL permet de déporter le traitement des données qui seront affichées dans le rapport au niveau de l'ETL. Ceci permet tout type de traitement avec la garantie de la performance. L'outil de reporting est alors utilisé pour ce qu'il fait le mieux, la restitution graphique des données dans le respect des relations entre les données à rapporter.

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324



L'écran de gestion du reporting est :

- Component/Report

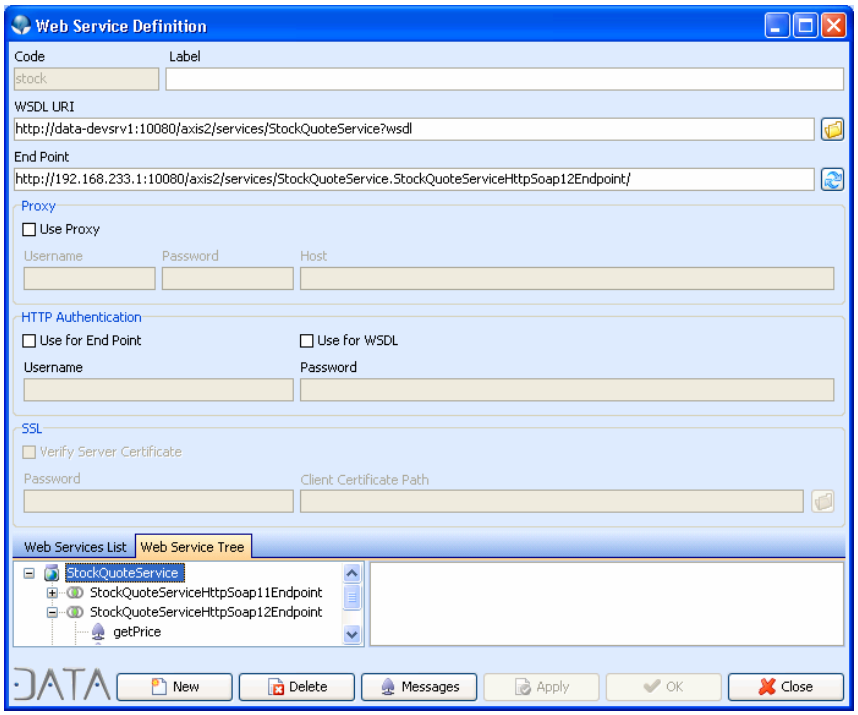
Web Services

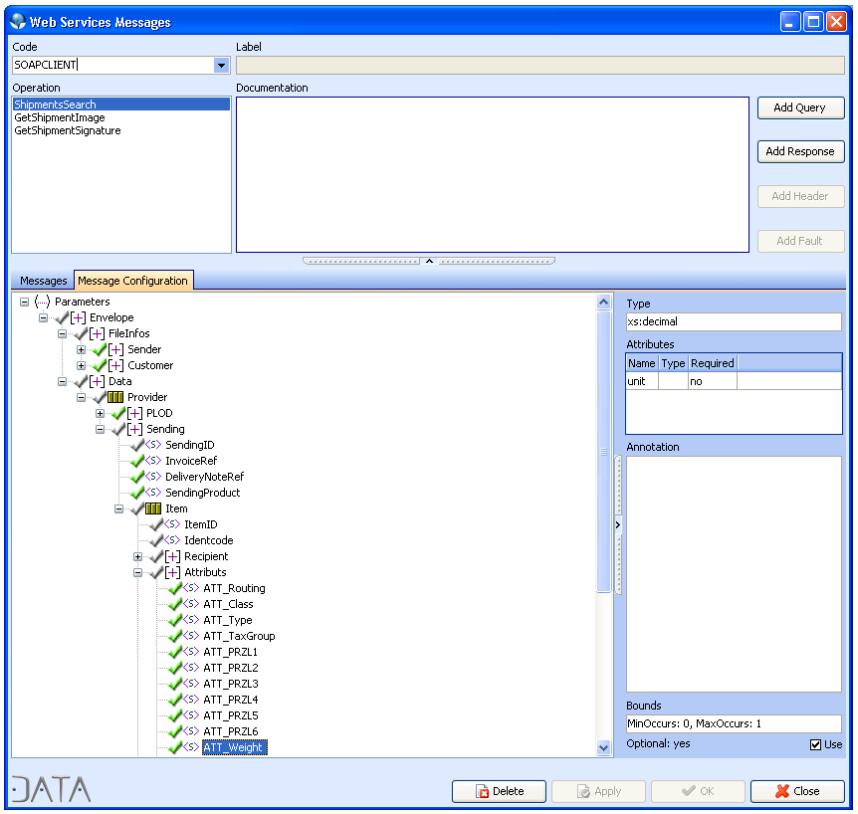
DataStudio est une plate-forme de gestion de Web Services complète.

Les Web Services sont définis graphiquement puis intégrés à un job en simple drag&drop. Il ne reste plus qu'à relier les paramètres d'appel avec des données gérées par le DataStudio.

Il est possible d'implémenter tout web service défini au préalable par un fichier WSDL ou d'invoquer tout web service à partir de sa description WSDL.

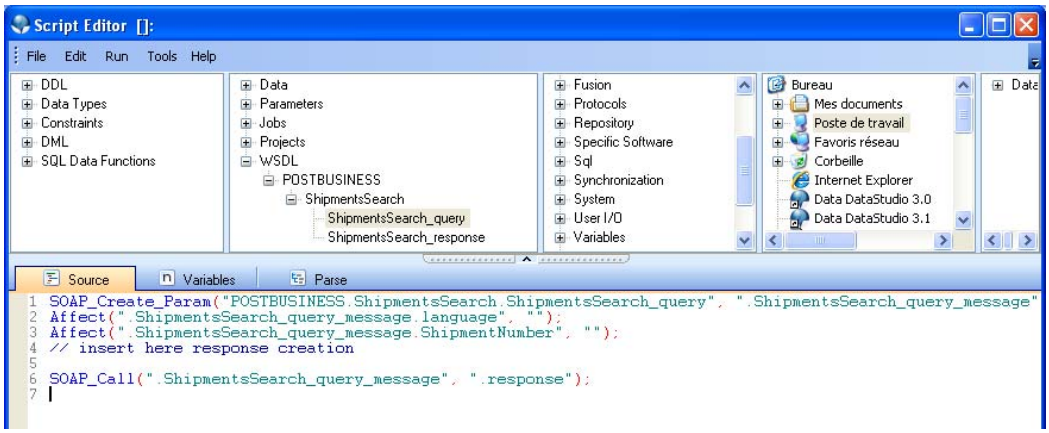
44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324





44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73

Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324



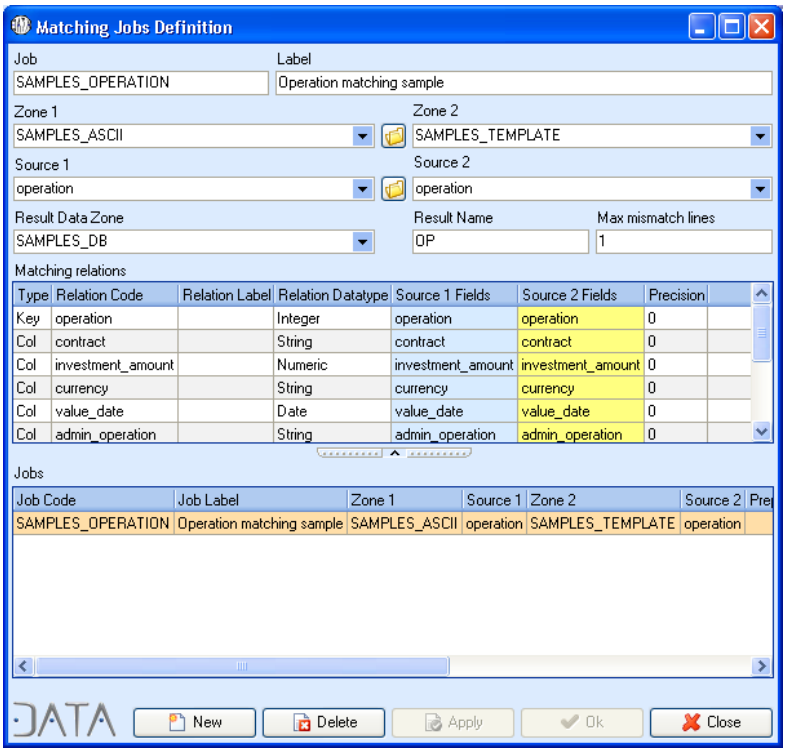
Les écrans de gestion des Web services sont :

- Component/Web Services/Web Service Definition,
- Component/Web Services/Web Service Message,
- Component/Web Services/Web Service Project Mapping.

Matching

DataStudio possède un module de Matching qui permet de rapprocher deux DataObjets et d'en déterminer les écarts sans écrire de ligne de script.

Le résultat est envoyé dans une table dynamique pour pouvoir être exploité par l'ETL.



Les écrans de gestion des Matching sont :

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

- Component/Matching,
- Display/Matching Results.

La version DataStudio Business Edition

Cette version est faite pour la mise en production des traitements ETL.

L'architecture est la suivante :

Elle dispose des fonctionnalités supplémentaires suivantes :

- un mode serveur batch et un mode service,

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324

- un nombre illimité de référentiels (Workspace),
- un module de gestion des configurations et du versioning des traitements (gestion des déploiements),
- un scheduler,
- le module de gestion du référentiel (choix de la base, import, export, sauvegarde, performance de l'accès),
- binaires pour Linux et Unix,
- un module de centralisation des log,
- un fonctionnement indépendant de toute connexion à internet.

Formation

Pour exploiter au mieux les possibilités de la plate-forme DataStudio, Data, centre de formation agréé propose une formation à l'ETL DataStudio. Cette formation donc éligible à la formation professionnelle est d'une durée de 3 jours sur site ou chez DATA.

44 bis boulevard Félix Faure
92320 Châtillon - France
Tél. : 33 (0) 1 57 19 59 38 - Fax : 33 (0)1 57 19 59 73
Société par actions simplifiée au capital de 38113 euros
RCS Nanterre B 434 535 324