International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 47 No. 3 Jan. 2026, pp. 332-352

http://www.ijias.issr-journals.org/

Intégration de sources de données hétérogènes dans un entrepôt de données dans un environnement minier

[heterogeneous data sources Integration in the data warehouse in mining environment]

Edouard Ngoy MUSHAME[1], John Tshomba KALUMBU[2], Sage Ngoie MBAYO[3], and Luc Lumanji MBUNGA[4]

[1]Section sciences de base, Institut Supérieur des Techniques Appliquées de Kolwezi, ISTA - Kolwezi, RD Congo

[2]Département de phytotechnie, Faculté des sciences agronomiques de l'Université de Lubumbashi, Lubumbashi, RD Congo

[3]Philosophiae Doctor, IS, Université de l'État libre, South Africa

[4]Section sciences de base, Institut Supérieur des Techniques Appliquées de Kolwezi, ISTA - Kolwezi, RD Congo

ABSTRACT: Easily accessible and useful data provides businesses and customers with a solid foundation to make real-time decisions or revise them as necessary. Besides your personnel and your products, data is one of your company's most valuable assets. As new information and communications technologies, digital transformation and smart technologies continue to reshape the industrial and mining landscape, data is essential to remaining competitive. For business leaders, it is imperative to fully understand data management tools and their impact on the industry. Organizations today have a multitude of information from most data sources: IoT, mining software, ERP, partners, websites, applications, cloud, customer feedback in the field, etc. These heterogeneous sources of varied formats are very often stored in independent silos. This heterogeneity combined with impressive volumes of data makes decision-making increasingly difficult. The objective of data integration is therefore to bring together all this data from different sources in order to allow decision-makers to have an overview.

KEYWORDS: data sources, databases, business intelligence, information systems, mining software.

RESUME: Des données facilement accessibles et utiles représentent, pour les entreprises ainsi que les clients, une base solide à partir de laquelle ils peuvent prendre des décisions en temps réel ou les réviser, le cas échéant. Outre votre personnel et vos produits, les données sont l'un des biens les plus précieux de votre entreprise. Alors que les nouvelles technologies de l'information et de la communication, la transformation numérique et les technologies intelligentes continuent de remodeler le paysage industriel et minier, les données sont essentielles pour demeurer compétitif. Pour les chefs d'entreprise, il est impératif de bien comprendre les outils de gestion de données et leur impact sur l'industrie. Les organisations disposent aujourd'hui d'une multitude d'informations provenant de nombreuses sources de données: IoT, Logiciels miniers, ERP, partenaires, sites internet, applications, cloud, retours clients sur le terrain, etc. Ces sources hétérogènes et de formats variés sont très souvent stockées dans des silos autonomes. Cette hétérogénéité combinée avec des volumes de données impressionnants rend la prise de décision de plus en plus difficile. L'objectif de l'intégration de données est donc de rassembler toutes ces données provenant de différentes sources afin de permettre aux décideurs d'avoir une vision d'ensemble.

MOTS-CLEFS: sources des données, bases de données, informatique décisionnelle, systèmes d'information, logiciels miniers.

Corresponding Author: Edouard Ngoy MUSHAME

1 INTRODUCTION

Les entreprises collectent des données à partir de sources disparates dans une variété de formats. Pour donner un sens à ces données et en extraire de la valeur, les entreprises utilisent différents outils d'intégration de données qui rassemblent les données en un seul endroit pour les transformer et les charger dans un référentiel de données centralisé comme un entrepôt de données ou un Data Mart, cette pratique d'organiser et présenter les données n'épargne pas les industries minières en général et en particulier celles de la République Démocratique du Congo. Tirant parti de différentes technologies, telles que ETL, l'intégration d'API, la virtualisation des données, l'EDI et autres, l'intégration des données permet aux entreprises d'accéder, de déplacer, de manipuler et de synchroniser plus facilement les données collectées à l'intérieur et à l'extérieur de l'entreprise. Cependant, la valeur obtenue grâce au processus dépend de la bonne sélection d'un outil et des méthodes d'intégration de données qui offre la combinaison optimale de performances, de facilité d'utilisation et d'évolutivité.

L'intégration de données est un processus qui consiste à combiner des données provenant de sources hétérogènes et à les présenter dans un format unifié. Ceci comprend:

- Consolider les données d'une grande variété de systèmes sources avec des formats disparates, tels que des systèmes de fichiers, des API, des bases de données, etc.
- Nettoyage des données en supprimant les doublons, les erreurs, etc.
- Catégorisation des données en fonction des règles métier.
- Le transformer dans le format requis afin qu'il puisse être utilisé pour le reporting ou l'analyse.

L'intégration des données est utilisée dans divers processus de gestion des données tels que la migration de données, l'intégration d'applications, la gestion des données de base, etc. Cependant, maintenant, le processus d'intégration des données n'est pas seulement limité aux utilisateurs informatiques. En raison de la disponibilité d'outils d'intégration de données faciles à utiliser, de plus en plus d'utilisateurs métier prennent la tête de l'optimisation des processus métier et stimuler l'agilité commerciale [1]. Suite à tout ce qui précède, les objectifs de cette recherche se place dans la théorie et la thématique de d'abord comprendre les bonnes pratiques de l'intégration des données dans les environnements miniers et comment développer et mettre en place une solution d'intégration de données, nous n'allons pas exclure le fait que les entreprises minières font partie de larges champs dans lesquels une diversité des sources de données est au rendez-vous.

2 ETAT DE L'ART SUR L'INTEGRATION DES DONNEES

L'intégration des données est un ensemble de pratiques, d'outils et de procédures architecturales qui permettent aux entreprises d'utiliser, de combiner et de tirer parti de tous les types de données. En plus de consolider les données provenant de systèmes disparates, le processus garantit que les données sont nettoyées et exemptes d'erreurs afin d'optimiser leur utilité pour l'entreprise [2], [3].

Les données intégrées sont particulièrement utiles pour les entreprises ayant un environnement diversifié et distribué, avec un éventail de sources de données et de ressources générant des informations. Dans ce cas, les données sont souvent cloisonnées et déconnectées des autres données métier, ce qui prive l'entreprise d'une vue unifiée de ses activités [3].

L'intégration des données permet à l'entreprise d'atteindre son plein potentiel. Les décisions importantes reposent sur des informations précises et les nouvelles technologies qui dépendent des données nettoyées peuvent être mises en œuvre et optimisées, ce qui aide l'entreprise à innover et à prospérer.

2.1 BREF HISTORIQUE DE L'INTEGRATION DE DONNEES

La combinaison des différentes sources de données pose un problème depuis que les systèmes de gestion ont commencé à collecter des données. Ce n'est qu'au début des années 1980 que les informaticiens ont commencé à concevoir des systèmes prenant en charge l'interopérabilité des bases de données hétérogènes ou différentes. L'un des premiers systèmes d'intégration des données a été développé par l'Université du Minnesota en 1991: son objectif était d'assurer l'interopérabilité des milliers de bases de données sur la population. Le système utilisait une approche d'entreposage de données permettant d'extraire, de transformer et de charger des données de sources disparates dans un schéma de visualisation pour les rendre compatibles.

Dans les années qui ont suivi, différents défis sont apparus, notamment en ce qui concerne la qualité, la gouvernance, la modélisation et, surtout, l'isolation ou le cloisonnement des données.

Les données intégrées sont devenues l'impératif des entreprises au début des années 2010, avec l'avènement de l'Internet des Objets (IoT). Soudain, un large éventail de terminaux, d'applications et de plateformes généraient d'énormes volumes de données, submergeant les entreprises. Le Big Data a fait son apparition et les entreprises ont dû trouver un moyen d'exploiter le potentiel de toutes ces informations. Aujourd'hui, les sociétés de toutes tailles et de tous secteurs utilisent l'intégration pour extraire de la valeur des données stockées dans toutes les applications et plateformes de l'entreprise [3].

2.2 TYPES D'INTÉGRATION DE DONNEES [4]

L'intégration des données peut être réalisée de plusieurs façons. Communément appelées méthodes, techniques, approches ou types d'intégration de données, il existe 5 façons différentes d'intégrer vos données.

- Intégration de données par lots

Dans ce type d'intégration de données, les données passent par le processus ETL par lots à des moments programmés (hebdomadaires ou mensuels). Elles sont *extraites* de sources disparates, *transformées* en une vue cohérente et normalisée, puis *chargées* dans un nouveau magasin de données, tel qu'un entrepôt de données ou plusieurs marts de données. Cette intégration est surtout utile pour l'analyse des données et la veille stratégique, car un outil de veille stratégique ou une équipe d'analystes peut simplement observer les données stockées dans l'entrepôt.

- Intégration des données en temps réel

Dans ce type d'intégration de données, les données entrantes ou en continu sont intégrées aux enregistrements existants en quasi temps réel par le biais de pipelines de données configurés. Les entreprises utilisent des pipelines de données pour automatiser le mouvement et la transformation des données, et les acheminer vers la destination ciblée. Les processus d'intégration des données entrantes (en tant que nouvel enregistrement ou mise à jour/application des informations existantes) sont intégrés dans le pipeline de données.

- Consolidation des données

Dans ce type d'intégration de données, une copie de tous les ensembles de données sources est créée dans un environnement ou une application de transit, les enregistrements de données sont ensuite consolidés pour représenter une vue unique, puis finalement déplacés vers une source de destination. Bien que ce type soit similaire à l'ETL, il présente quelques différences essentielles telles que:

- La consolidation des données se concentre davantage sur des concepts tels que le nettoyage et la normalisation des données et la résolution des entités, tandis que l'ETL se concentre sur la transformation des données.
- Alors que l'ETL est une meilleure option pour le big data, la consolidation des données est un type plus approprié pour relier les enregistrements et identifier de manière unique les principaux actifs de données, tels que le client, le produit et l'emplacement.
- Les entrepôts de données aident principalement à l'analyse des données et à la BI, tandis que la consolidation des données est également utile pour améliorer les opérations commerciales, comme l'utilisation du dossier consolidé d'un client pour le contacter ou créer des factures, etc.

- La virtualisation des données

Comme son nom l'indique, ce type d'intégration de données ne crée pas réellement une copie des données ou ne les déplace pas vers une nouvelle base de données avec un modèle de données amélioré. Il introduit plutôt une couche virtuelle qui se connecte à toutes les sources de données et offre un accès uniforme comme une application frontale.

Comme elle ne dispose pas de son propre modèle de données, la couche virtuelle a pour but d'accepter les demandes entrantes, de créer des résultats en interrogeant les informations requises dans les bases de données connectées et de présenter une vue unifiée. La virtualisation des données réduit le coût de l'espace de stockage et la complexité de l'intégration, puisque les données semblent intégrées mais résident séparément dans les systèmes sources.

- Fédération de données

La fédération de données est similaire à la virtualisation des données et est souvent considérée comme son sous-type. Encore une fois, dans la fédération de données, les données ne sont pas copiées ou déplacées vers une nouvelle base de données, mais un nouveau modèle de données est conçu qui représente une vue intégrée des systèmes sources.

Il fournit une interface frontale d'interrogation et, lorsque des données sont demandées, il les extrait des sources connectées et les transforme en modèle de données amélioré avant de présenter les résultats. La fédération de données est utile lorsque les modèles de données sous-jacents des systèmes sources sont trop différents et doivent être mis en correspondance avec un modèle plus récent afin d'utiliser les informations plus efficacement.

2.3 CAS D'UTILISATION RELATIFS À L'INTEGRATION DES DONNEES

Lorsque des données sont générées, elles peuvent être intégrées et utilisées pour obtenir des insights en temps réel qui profitent à l'entreprise. Une entreprise établie dans plusieurs régions peut obtenir une vue consolidée sur l'ensemble de ses opérations pour comprendre ce qui fonctionne et ce qui ne fonctionne pas. Une vue unique de l'entreprise facilite la compréhension des causes et des effets, ce qui permet d'effectuer des rectifications en temps réel et de réduire les risques.

Grâce à l'intégration des données, les entreprises peuvent:

- Optimiser l'analytique: accéder aux données des systèmes opérationnels, les mettre en attente ou les extraire (entreposage de données), puis les transformer et les livrer à l'entreprise sous la forme d'une analytique fiable.
- Favoriser la cohérence entre les applications opérationnelles: assurer la cohérence de la base de données entre les applications (intra-entreprise et interentreprises), de manière bidirectionnelle et unidirectionnelle.
- Partager des données en dehors de l'entreprise: fournir des données fiables aux parties externes comme les clients, les fournisseurs et les partenaires.
- Orchestrer les services de données: déployer toutes les fonctionnalités d'intégration des données d'exécution en tant que services de données pour garantir rapidité et précision.
- Assurer la migration et la consolidation des données: répondre aux besoins de mouvement et de transformation dans le cadre de la migration et de la consolidation des données, par exemple, lors du remplacement d'applications héritées ou de la migration vers de nouveaux environnements [5], [6].

3 INTEGRATION D'APPLICATIONS VS INTEGRATION DE DONNEES

Intégration d'applications est un autre concept fréquemment utilisé dans les organisations ainsi. Il est important de faire la différence entre l'intégration d'applications et l'intégration de données, d'autant plus que les deux se complètent souvent pour réaliser des opérations transparentes [7].

Alors que l'intégration d'applications vise à permettre aux applications logicielles de fonctionner ensemble en partageant des données, l'intégration de données se concentre sur la consolidation et l'harmonisation des données provenant de sources disparates à des fins d'analyse et de prise de décision. Encore une fois, nous avons un tableau ci-dessous pour résumer l'intégration d'applications par rapport à l'intégration de données:

Tableau 1. Comparatif entre intégration d'applications et de données

Elément de comparaison	Intégration d'applications	Intégration des données
Définition	Connecter et coordonner les applications logicielles et les systèmes pour partage de données et l'automatisation des processus.	Combiner des données provenant de diverses sources dans une vue unifiée et précise pour l'analyse et la prise de décision.
Domaine	Permettre aux applications de fonctionner ensemble de manière transparente.	Consolidation des données et l'harmonisation à partir de sources multiples, en se concentrant sur le mouvement et la transformation des données.
Objectif commercial	Améliorer l'efficacité des processus métier, automatiser les flux de travail et améliorer l'expérience utilisateur grâce à des interactions transparentes avec les applications.	Fournir une vue globale des données dans toute l'organisation, prenant en charge la prise de décision, le reporting et l'analyse basés sur les données
Flux de données	Gérer les flux de données et de processus entre les applications, garantissant une communication et une collaboration en temps réel.	Implique des processus d'extraction, de transformation et de chargement de données, entre autres.
Cas d'usage	Intégrer le SAP aux logiciels de contrôle d'usine (PI System ou SCADA), connecter le SharePoint de l'entreprise au système de messagerie Outlook, etc.	Création d'entrepôts de données centralisés, consolidation des données clients, fusion des données pour le reporting financier, etc.
Outils et technologies	Middleware, API, files d'attente de messages, ESB, plates-formes d'intégration et passerelles API.	intégration des données et Outils ETL, entrepôts de données, lacs de données et Systèmes de gestion de bases de données.

4 PROCESSUS INTEGRATION DE DONNEES EN ENTREPRISE

Le processus d'intégration des données est la méthode par laquelle une organisation combine des données provenant de plusieurs plateformes et ensembles de données différents pour créer une architecture numérique cohérente et globale. Dans ce processus, les étapes de l'intégration des données ne sont pas aussi importantes que le résultat. Certaines organisations choisissent d'intégrer tous les ensembles de données de l'entreprise en un seul. D'autres choisissent de n'intégrer qu'un ensemble de données particulier et important. Quel que soit le choix, ce processus correspond aux étapes définies et aux résultats proposés pour cette combinaison d'ensembles de données.

Le processus d'intégration des données a un impact direct sur des domaines industriels tels que les logiciels de sécurité, les applications commerciales et l'utilisation des données collectées. Cela facilite le suivi des processus à l'échelle de l'entreprise, les

transformations numériques globales et une meilleure connaissance de la manière d'utiliser les données collectées sur une période donnée.

4.1 DIAGRAMME DE FLUX DE TRAVAIL

Le processus d'intégration des données de votre entreprise doit être flexible pour soutenir la croissance continue de l'entreprise. Il existe de nombreuses possibilités de plans d'intégration de données. Ce processus doit néanmoins suivre un ensemble d'étapes bien spécifiques comme présentés sur le diagramme ci-dessous:

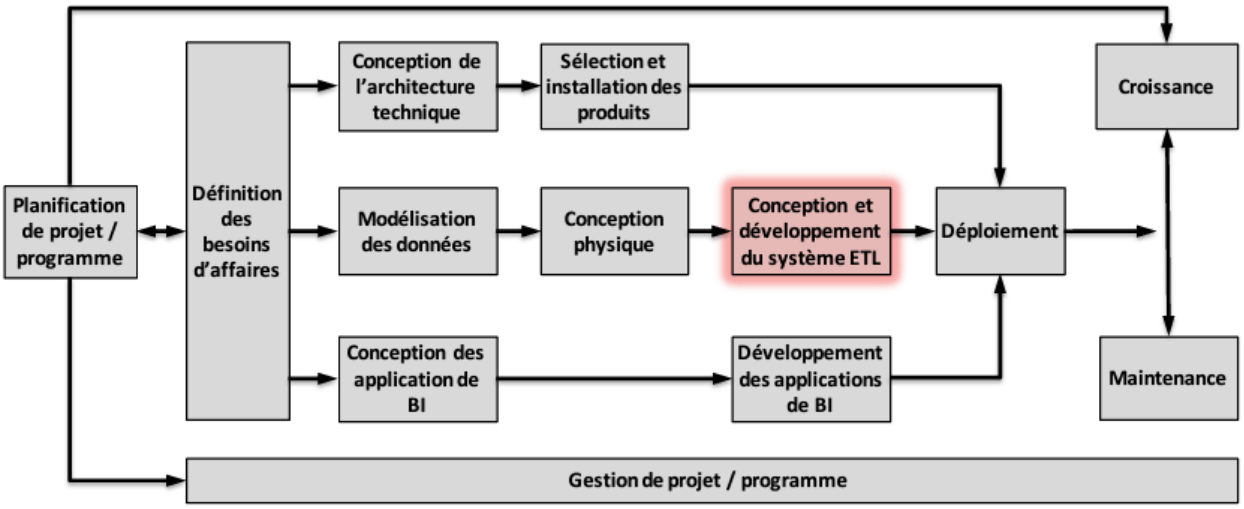


Fig. 1. Diagramme de Flux du travail dans l'intégration des données

4.2 APPROCHES INTÉGRATION DE DONNÉES

La technologie d'intégration de données a évolué à un rythme rapide au cours de la dernière décennie. Initialement, Extract, Transform, Load (ETL) était la seule technologie d'intégration de données disponible pour le traitement par lots. Cependant, à mesure que les entreprises continuaient d'ajouter davantage de sources à leur écosystème de données, le besoin de technologies d'intégration de données en temps réel s'est fait sentir [5].

ETL: Cette approche (Extract, Transform and Load), permet justement:

- Intégration et livraison des données en lot ainsi que
- Transformations appliquées sur les données

Enterprise Information Integration:

- Fédération des données provenant de plusieurs sources
- Accès en temps réel aux données
- Données structurées ou semi – structurées

Enterprise Application Integration:

- Processus d'intégrations de données d'applications
- Basé sur les échanges sur un bus commun.

Dans nos recherches nous allons nous appuyer sur les intégrations par l'approche des ETL, nous allons également démontrer ces approches en utilisant les outils de Business Intelligence de Microsoft SQL Server, en occurrence SSIS et SSRS et pour le reporting nous allons ajouter l'usage de Power BI.

5 PRESENTATION DE SQL SERVER INTEGRATION SERVICES: SSIS

Avant de présenter SSIS, nous allons revenir sur les notions de bases de données ainsi que le système de gestion de bases de données.

5.1 NOTIONS DE BASES DE DONNÉES

Une base de données est un ensemble d'informations qui est organisé de manière à être facilement accessible, géré et mis à jour. Elle est utilisée par les organisations comme méthode de stockage, de gestion et de récupération de l'informations [6].

Les données sont organisées en lignes, colonnes et tableaux et sont indexées pour faciliter la recherche d'informations. Les données sont mises à jour, complétées ou encore supprimées au fur et à mesure que de nouvelles informations sont ajoutées. Elles contiennent généralement des agrégations d'enregistrements ou de fichiers de données, tels que les transactions de vente, les catalogues et inventaires de produits et les profils de clients. Généralement, l'administrateur de la base de données régule les accès des utilisateurs afin de contrôler leurs actions et d'analyser les usages. Pour garantir la cohérence des données et l'intégralité des transactions, toutes les transactions réalisées sur une base de données doivent répondre aux exigences de la conformité ACID:

- Le principe d'Atomicité garantit la bonne exécution de la transaction. Les transactions de base de données, comme les atomes, peuvent être décomposées en plus petites parties. Si une partie d'une transaction échoue, toute la transaction sera annulée.
- La propriété de Cohérence signifie que seules les données qui suivent des règles prédéfinies peuvent être écrites dans la base de données.
- L'isolement fait référence à la capacité de traiter simultanément plusieurs transactions de manière indépendante.
- La durabilité requiert de rendre les défaillances invisibles pour l'utilisateur final. Les données sont sauvegardées une fois la transaction terminée, même en cas de panne de courant ou de défaillance du système [7].

5.2 NOTIONS DE SYSTEME DE GESTION DE BASE DE DONNEES (SGBD)

Une base de données et un système de gestion de base de données, abrégé en SGBD, forment un système de base de données (ce dernier terme est toutefois souvent appelé simplement « base de données »). De manière générale, un SGBD est un **logiciel** qui définit le **modèle d'un système de base de données** et constitue ainsi une composante indispensable à la création, à la gestion et à l'utilisation d'une base de données. L'utilisateur peut uniquement ajouter et lire l'ensemble de données désiré après avoir installé et paramétré le système de gestion de base de données correspondant. Des interfaces spécifiques à l'application et un langage de base de données adapté permettent les **accès en écriture et en lecture** ainsi que les **fonctionnalités d'administration** générales. Le langage de ce type le plus connu est SQL (**S**tructured **Q**uery **L**anguage).

Le système de gestion de base de données est le composant le plus important d'un système de base de données. Sans SGBD, il serait impossible d'administrer, commander ou contrôler la base de données. Par ailleurs, le logiciel gère tous les accès à la base de données en lecture et en écriture. Pour décrire les fonctions et les exigences des opérations d'un système de gestion de base de données, on utilise fréquemment l'acronyme ACID pour atomicity, consistency, isolation et durability (atomicité, cohérence, isolation et durabilité). Chacun des termes composant l'acronyme ACID couvre les principales exigences applicables à un SGBD:

- L'atomicité est la propriété « tout ou rien » du SGBD qui implique que la transaction complète ne peut être exécutée correctement que si les interrogations sont valides et arrivent dans le bon ordre.
- La cohérence implique que la base de données reste stable même en cas de transaction réussie ce qui nécessite une vérification constante de toutes les transactions.
- L'isolation est l'exigence imposant que les transactions soient indépendantes les unes des autres ce qui est souvent garanti par des fonctionnalités de blocage.
- La durabilité signifie que l'ensemble des données du SGBD doivent être enregistrées durablement, même après la réalisation d'une transaction avec succès. Cela s'applique tout particulièrement en cas d'erreur système ou de panne du SGBD. Cette durabilité est notamment assurée par des journaux de transaction qui documentent l'ensemble des processus dans le SGBD [8], [9], [10], [11].

5.3 SQL SERVER INTEGRATION SERVICES

SQL Server Integration Services (SSIS) est un ETL (Extract Transform Load). Il permet de se connecter à n'importe quelle source de données (Excel, fichier plat csv, XML, base de données, etc.). SSIS offre la possibilité de collecter des données, de les transformer en données exploitables par les outils d'analyse. Ce sont ces données exploitables qui vont alimenter une ou plusieurs bases de données dédiées (bases de données relationnelles ou multidimensionnelles).

SQL Server Integration Service est la version améliorée de Data Transformation Service (DTS), présent dans les versions de SQL Server antérieures à 2005. SQL Server Integration Services est une plateforme qui permet de créer des solutions de transformation et d'intégration de données au niveau de l'entreprise. Grâce à SQL Server Integration Services, vous pouvez résoudre des problèmes d'intégration ou d'échange de données complexes en copiant ou en téléchargeant des fichiers, en envoyant des messages électroniques en réponse à des événements, en mettant à jour des entrepôts de données, en nettoyant, en vérifiant l'intégrité / la qualité des données et, enfin, en explorant des données et en gérant des données et des objets SQL Server.

Les packages peuvent fonctionner en mode autonome ou de concert avec d'autres packages en fonction des réalités des batches de chargement ou d'échange de données à mettre en œuvre. Avec SQL Server Integration Services, il est possible d'extraire et de transformer des données à partir d'un large éventail de sources, comme par exemple, des fichiers de données XML, des fichiers plats et des sources de données relationnelles, puis les charger dans une ou plusieurs destinations [9], [12].

5.4 FONCTIONNALITÉS PRINCIPALES DE SSIS [13], [14], [15]

- Sources et destinations de données multiples:

  Connexion OLE DB donc toutes bases de données, Excel, Fichier plat (CSV), XML, etc.

- Transformations de données:

  Agrégation, Filtre, Colonne dérivée, Conversions, etc.

- Flux de contrôles:

  Tâche d'exécution de packages SSIS, Tâche d'insertion en bloc (BULK INSERT), Tâche d'exécution de requêtes SQL, Tâche de script (VB, C#), etc.

- Taches de maintenance:

  Gestion des événements, Reconstruction d'index, Envoi de mails, Nettoyage d'historiques, etc.

6 DATA TRANSFORMATION ENTRE DEUX SOURCES DE DONNEES DIFFERENTES

Nous allons démontrer cette expérience sur l'intégration de données en utilisant la synchronisation avec un composant *Merge Join Transformation* entre les données issues d'une source Excel et une table SQL Server, pour l'exemple.

Le modus operandi *de notre exemple d'implémentation sera effectué avec l'outil de développement Visual Studio ou dans un autre cas échéant vous pouvez utiliser Business Intelligence Development Studio (BIDS) dont l'interface graphique est sensiblement la même. Nous allons dans ces exemples nous insérer sur les opérations SQL d'insertions et de mises à jour de données, notons qu'une opération d'insertions et de mises-à-jour est appelée* UPSERT *(ou* UPDATE-INSERT*).*

6.1 PRÉSENTATION DES BESOINS

Admettons que l'on dispose d'une source de données Excel (Employee-Database.xlsx*) et que l'on souhaite pouvoir intégrer les contenus de cette base de données avec une table* Employee *stockée au sein d'une base de données SQL Server* Employee_Database_staging*.*

Voici le contenu du fichier Excel:

ID	Firstname	Lastname	Age
10235	Jean	Kasongo	32
10235	Idriss	Mujinga	28
10235	Sakadie	Ilunga	52
10235	Serge	Mwamba	36
10235	Yacinte	Longwangwa	38
10235	Edouard	Mushame	27
10235	Matembo	Toto	38
10235	Syntiche	Monga	23
10235	Ruth	Kanam	19
10235	Arianne	Mujinga	29
10235	Sandra	Kangolombo	25
10235	Stephane	Malamba	29
10235	Pax	Kaulu	31
10235	Idriss	Banza	32
10235	Russelle	Yolamu	30
10235	Christelle	Kasongo	22
10235	Jacques	Kyabula	38
10235	Fifi	Masuka	45
10235	Freddy	Kakonde	

Fig. 2. Les informations stockées au sein d'une feuille Excel appelée EEs

La table Employee *de la base de données SQL Server, quant à elle, possède une structure assez simple:*

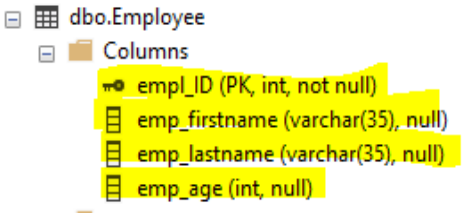


Fig. 3. On suppose qu'elle est vierge de tout enregistrement

6.2 CHOIX DE LA METHODE DE SYNCHRONISATION

La stratégie de synchronisation choisie sera celle incrémentielle. Il est, bien sûr, possible d'adopter une solution plus « brute » qui consiste à tronquer la table de destination avant d'y insérer le contenu de la source de données (Excel, dans notre cas). Cette solution possède principalement 2 limites qui peuvent être problématiques dans le monde réel:

- La troncature n'est pas supportée sur une table concernée par une contrainte d'intégrité référentielle de type FK (foreign key, ou clé étrangère).
- Si l'on est amené à travailler avec une dizaine de millions d'enregistrements, le fait de fréquemment insérer tous les enregistrements « from scratch » peut avoir un impact non-négligeable au niveau des performances. Sans parler de problèmes liés aux blocages (toute opération d'insertion ou de mise-à-jour étant transactionnellement bloquante).

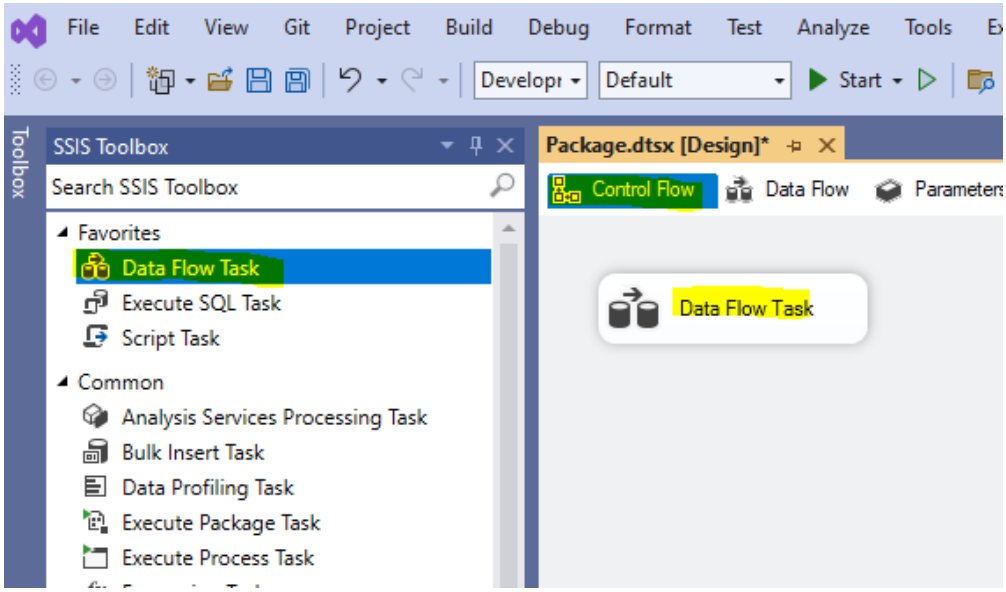
Quant aux méthodes de transformation, il en existe diverses méthodes de traiter les données à synchroniser, avec notamment l'utilisation de 2 types de composants de *data flows* SSIS:

- Merge Join Transformation, qui permet de réaliser une réunion (LEFT, FULL ou INNER) de deux jeux de données préalablement triés afin d'y générer le résultat de leur fusion.
- Lookup Transformation, qui permet d'effectuer des opérations de recherche via l'équijointure de colonnes d'entrée à des colonnes d'un jeu de données de référence.

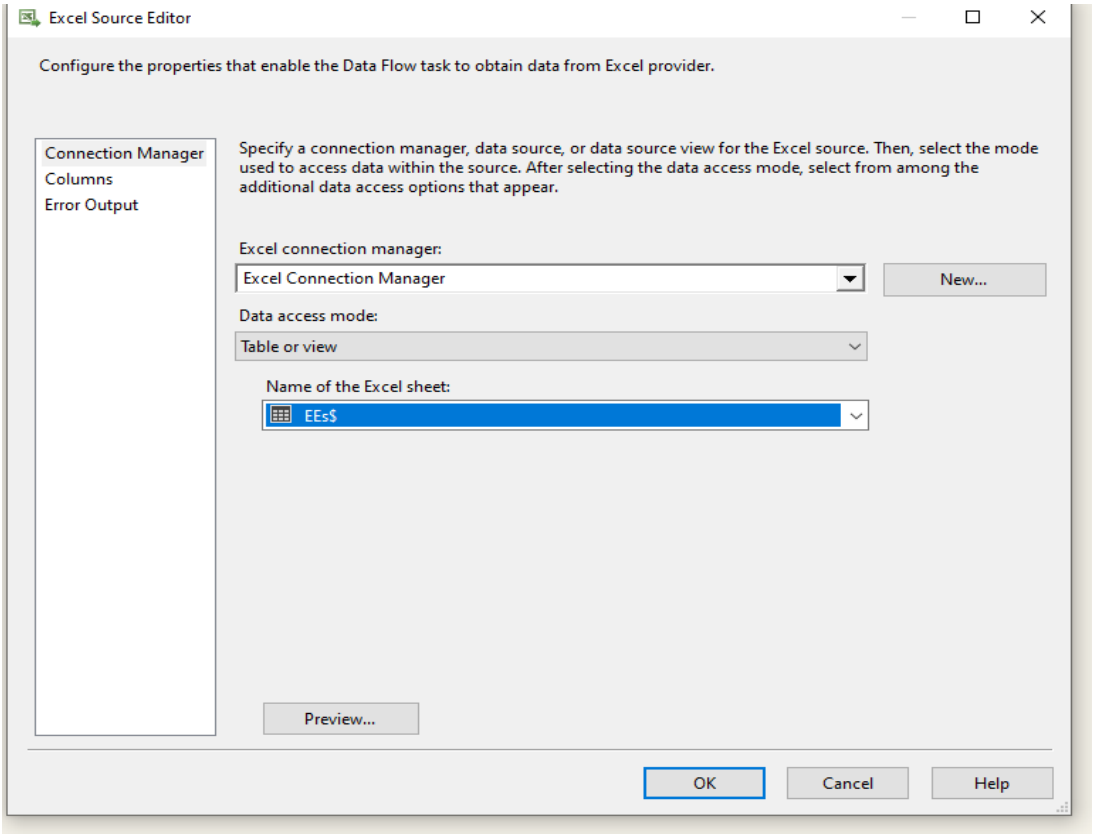
Pour illustrer ces pratiques, nous allons opter d'utiliser la méthode *Merge Join Transformation*.

6.3 PROCÉDURES PAS À PAS

- Au sein de l'onglet Control Flow:
  - Faites glisser un composant Data Flow Task:

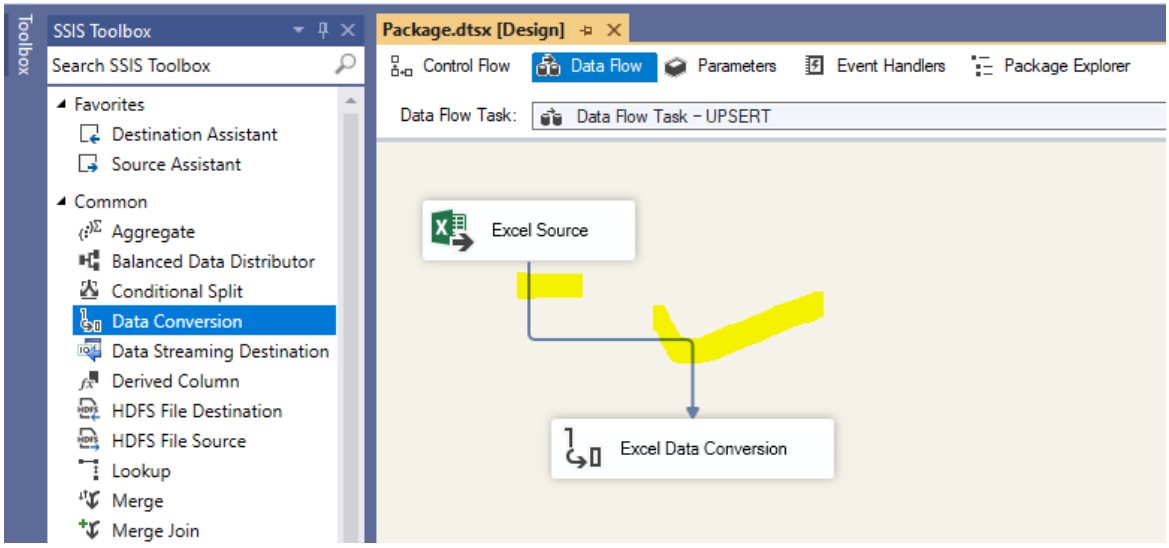


- Après avoir renommé le composant en Data Flow Task – UPSERT (en cliquant dessus, ou en effectuant un clic-droit>Rename), on va double-cliquer sur ledit composant pour accéder à la définition de ses flux de données (onglet Data Flow).
- Au sein du Data Flow de Data Flow Task – UPSERT:
  - Faites glisser (ou double-cliquez sur) le composant Excel Source, situé dans la SSIS Toolbox, dans la zone de design:
  - Renommez le composant Excel Source en Excel Source – Employees:
- Au sein des propriétés du composant Excel Source – Employees (accessible via clic-droit>Edit… ou double-clic), spécifiez le nouveau gestionnaire de connexion Excel:
  - Choisissez la feuille du fichier Excel à utiliser (EEs, dans notre cas):

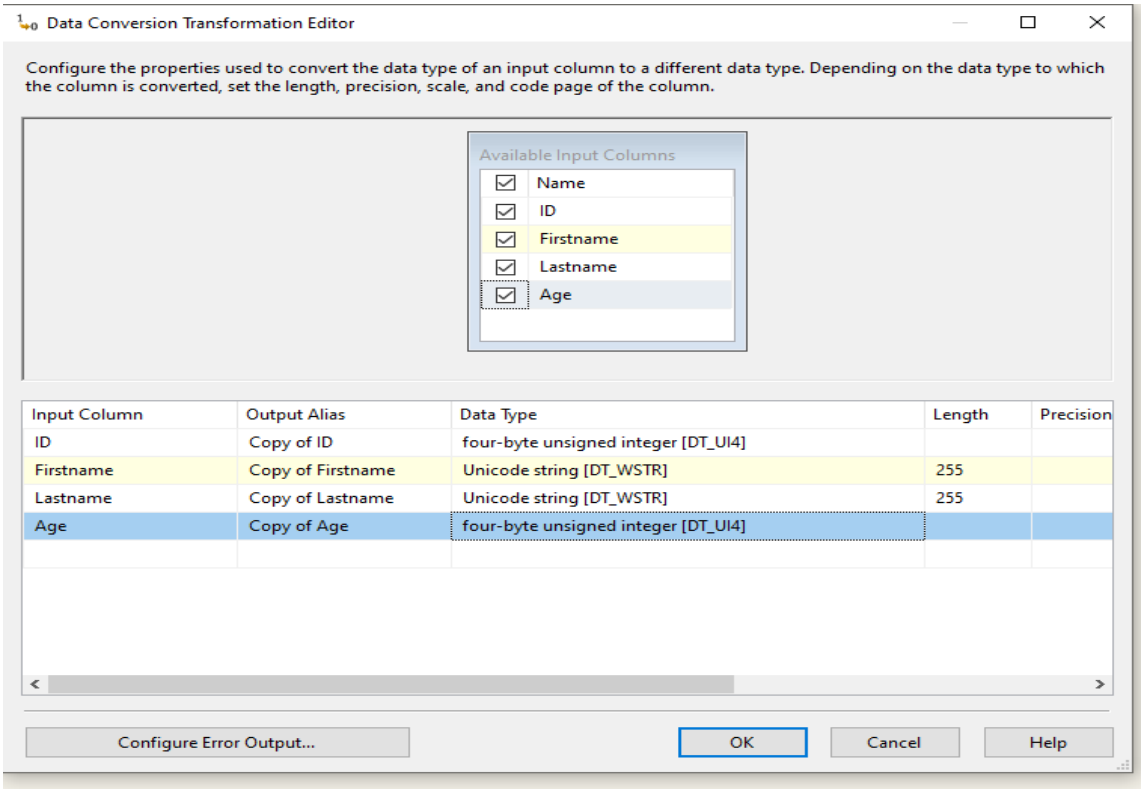


Par la suite nous allons valider les modifications effectuées pour le composant *Excel Source* en pressant sur *OK*.

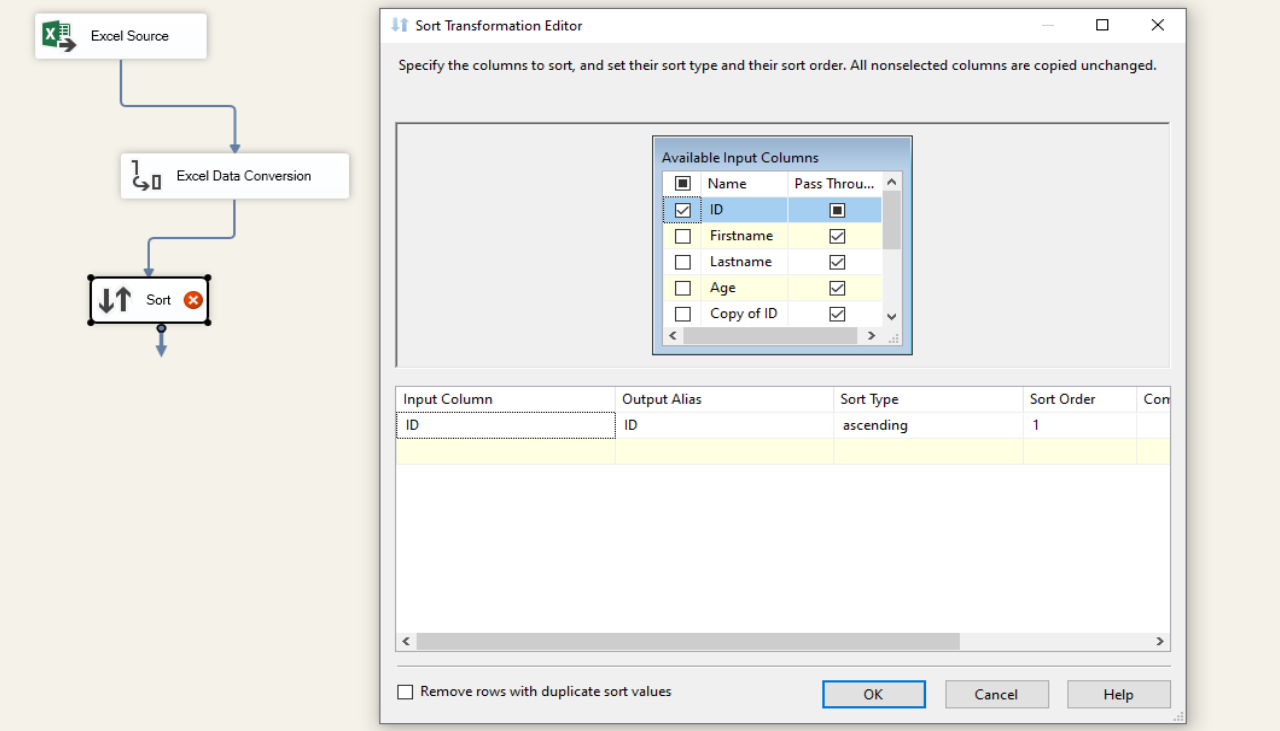
- Toujours au sein de la zone de design *Data Flow* de *Data Flow Task – UPSERT,* faites glisser (ou double-cliquez sur) un composant *Data Conversion*
- Renommez le composant Data Conversion en Excel Data Conversion
- Placez une contrainte de précédence de succès allant de Excel Source – Employees vers Excel Data Conversion:



- Au sein des propriétés du composant Excel Data Conversion, spécifiez les types de données de chaque colonne, en tenant compte de celles de la future destination (i.e., la table Employee de SQL Server):



- Par la suite nous allons Faire glisser (ou double-cliquez sur) un composant *Sort:*
- Placez une contrainte de précédence de succès allant d'Excel Data Conversion vers Sort:
- Dans les propriétés de Sort, sélectionnez Excel Data Conversion.ID afin que le tri soit effectué en fonction de la colonne d'IDs convertie, et assurez-vous que les autres colonnes converties soient les seules traversées (section Pass Through):



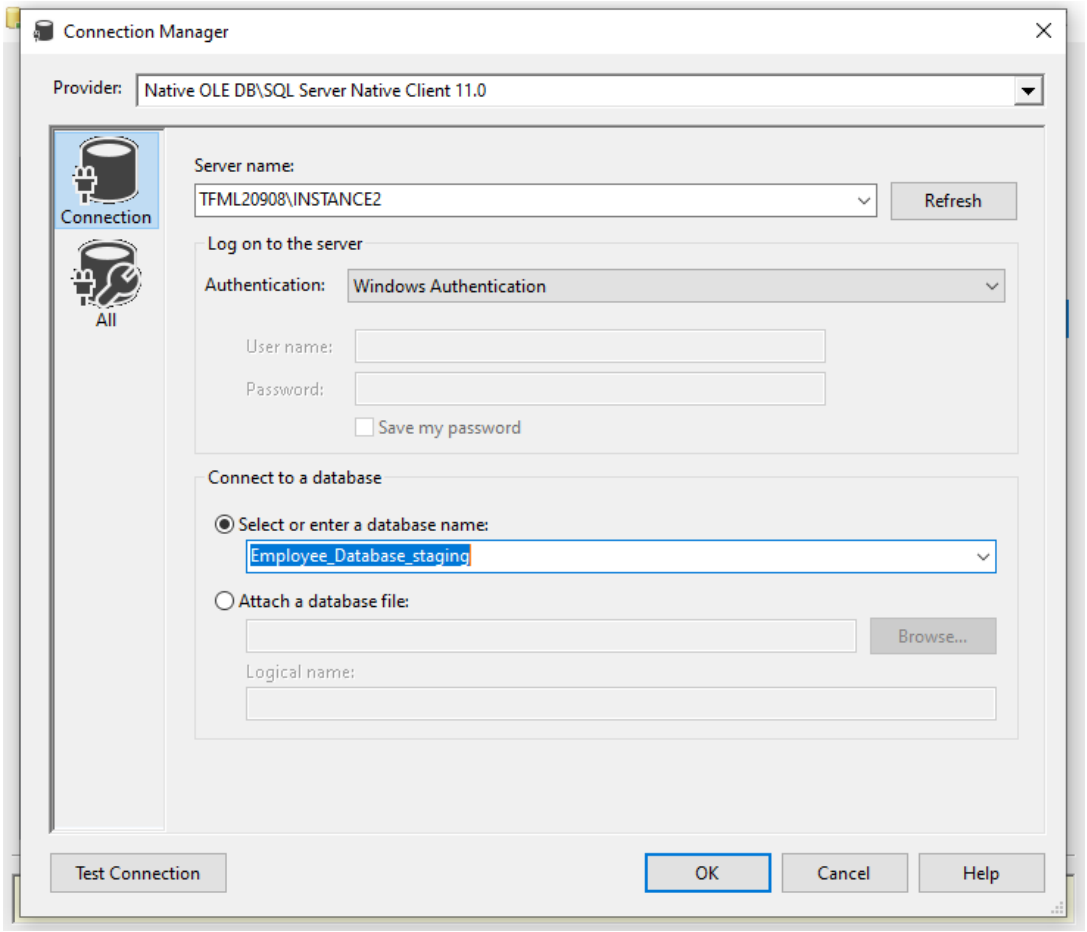
- Suivez les mêmes étapes allant de l'implémentation d'Excel Source à Sort, mais en remplaçant Excel Source par OLE DB Source qui pointera vers la base de destination: MaBase:

  Avec:

- Configuration d'OLE DB Source (renommé en OLE DB Source – Employés) comme suit:

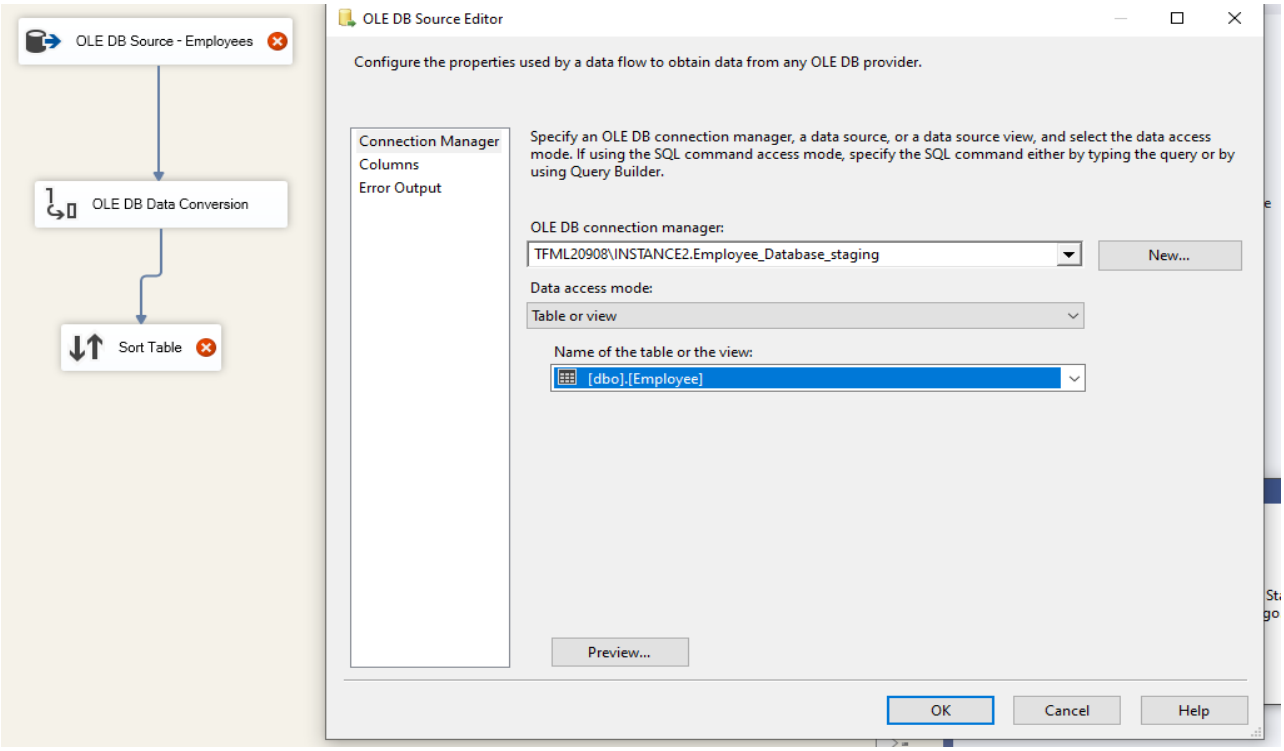
- Onglet Connection Manager:

Si aucun gestionnaire de connexion pointant vers [Employee_Database_staging] n'a été préalablement configuré, alors cliquez sur le bouton NEW afin de créer un nouveau gestionnaire de connexion puis, dans la section Connection, spécifiez les informations de connexion de la table [Employee] de la base de données [Employee_Database_staging] située sur une INSTANCE2 SQL Server située sur la machine appelée *TFML20908*. Pour ce faire, cliquez sur NEW afin d'afficher, par la suite, la boîte modale suivante où vous pourrez remplir vos informations:

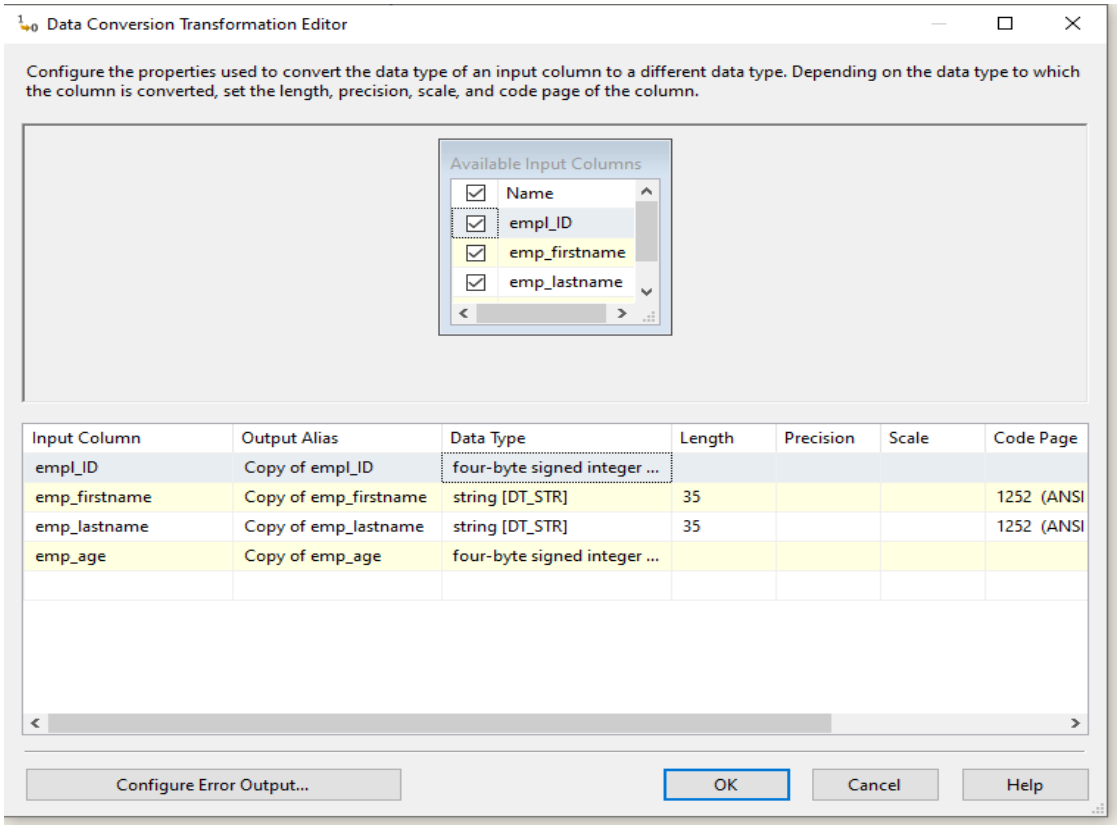


Vous pouvez cependant utiliser un compte Windows (en le supposant idéalement dédié aux tâches SSIS).

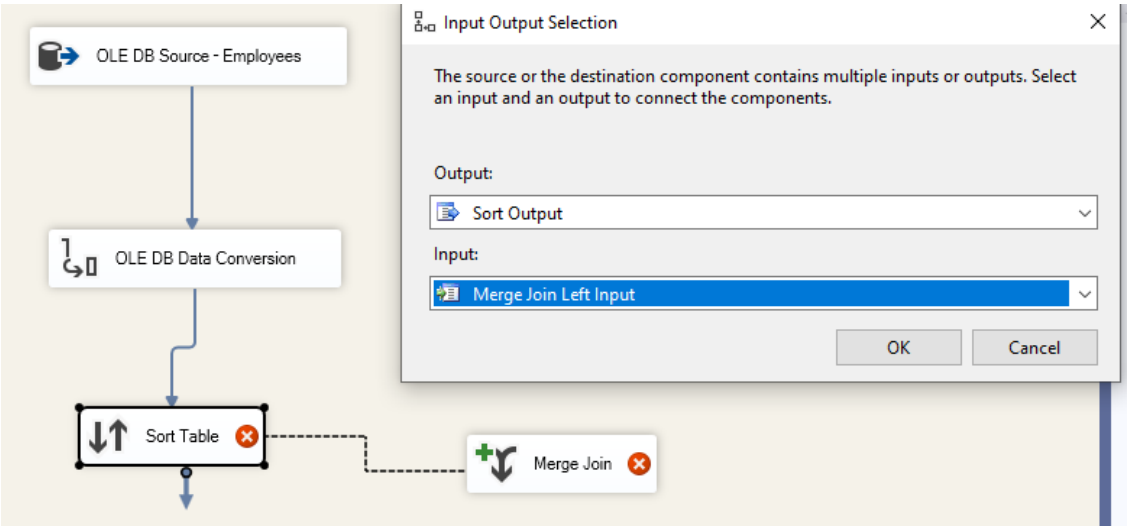
Validez et sélectionnez la table Employee*:*



- Configuration du Data Conversion (renommé en OLE DB Data Conversion):

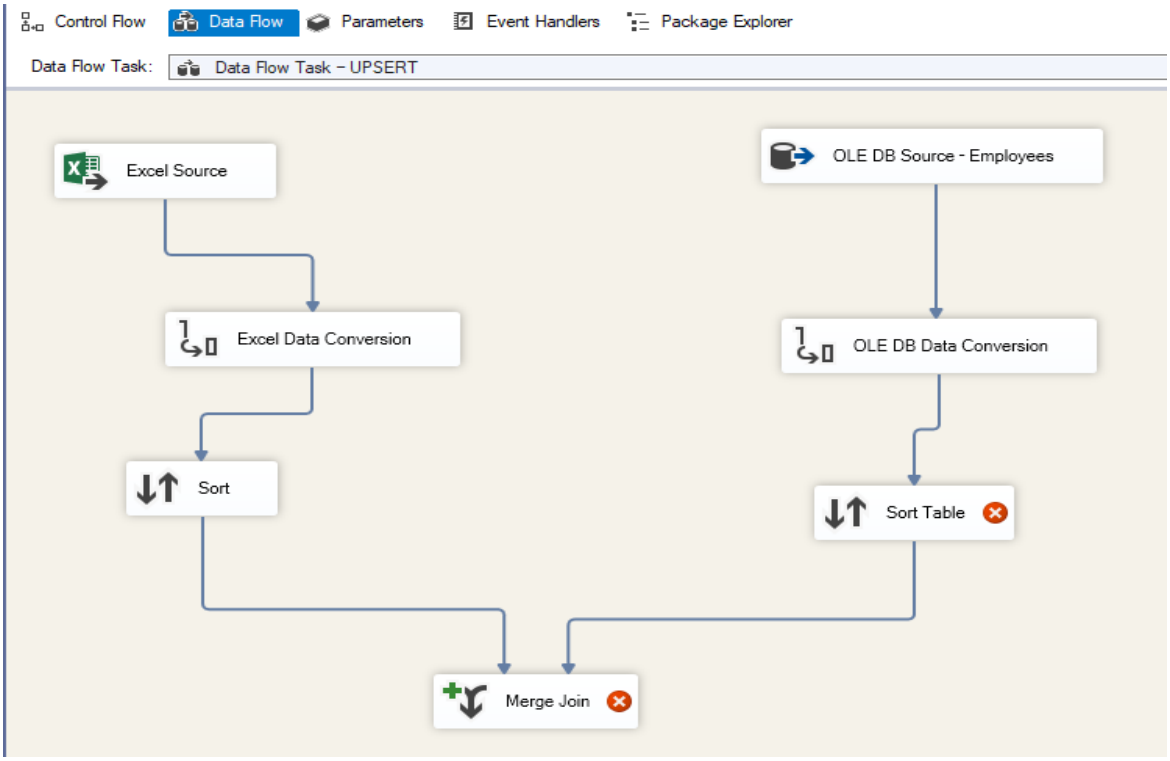


- Configuration du Sort (renommé en Sort Table) de la même façon que pour la version Excel.
  - Faites glisser un composant Merge Join dans le canevas:
  - Placez une contrainte de précédence de succès allant de *Sort* vers *Merge Join* avec *Left Join* comme type de jointure:

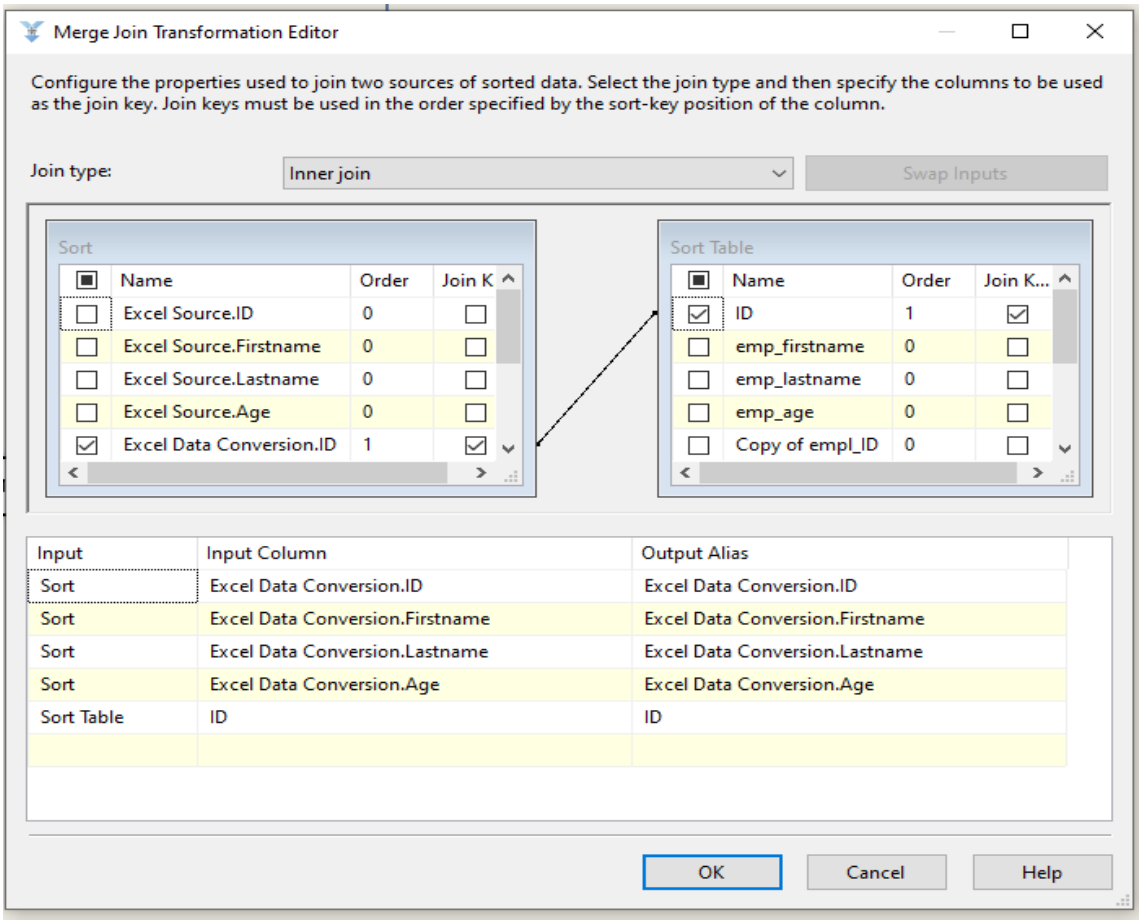


L'utilisation d'une jointure de type Left Join permet de ne sélectionner que les lignes de données de la source de données qui ne sont pas présents dans la table de destination.

Placez également une contrainte de précédence allant de Sort Table vers Merge Join:



- Configurez le Merge Join comme suit:



- Faites glisser un composant Conditional Split et placez une contrainte de précédence allant de Merge Join vers ledit composant.

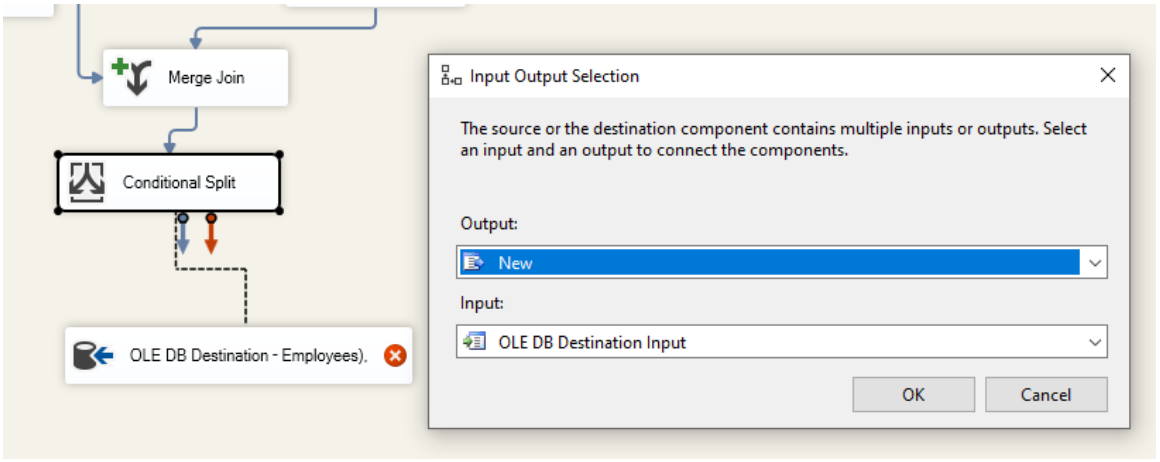
  La sortie par défaut a été renommée en *Existing* (au lieu de *Conditional Split Default Output*) pour des raisons de lisibilité.

Le *Merge Join* avec option « *Left Join* » récupère tout le contenu de l'Excel – que les données soient présentes ou non dans la table *Employes* – puis remplace, dans le jeu de résultats, les lignes Excel absentes dans ladite table par des valeurs nulles (*NULL*).

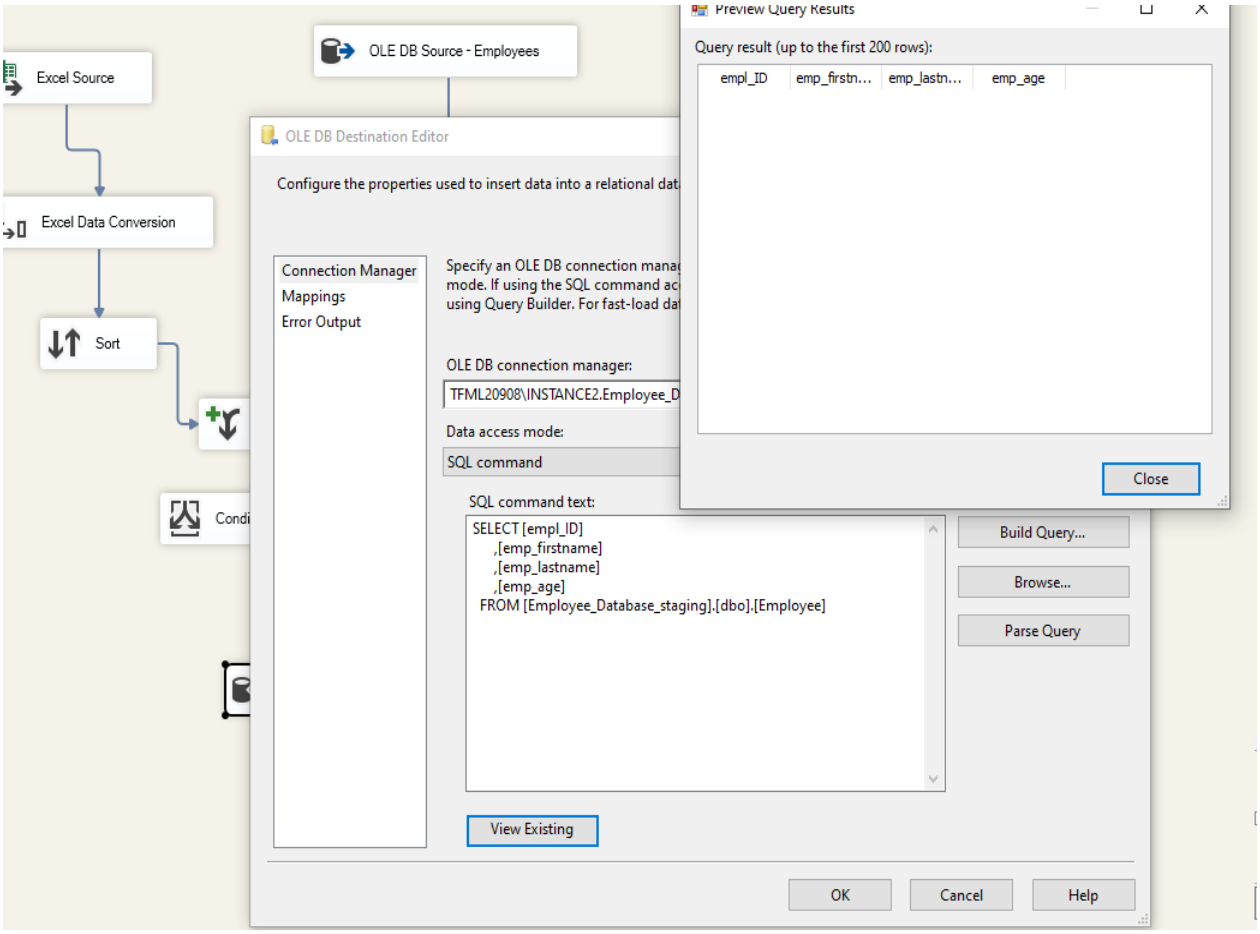
Le but du *Conditional Split* est donc, ici, de récupérer les lignes de données de l'Excel qui sont en *NULL*s dans la table *Employes,* puis de « splitter » le nouveau jeu de résultats en 2 parties: l'une pour les insertions (*New*) et l'autre pour les mises-à-jour (*Existing*).

6.4 INSERTION DES DONNEES

- Faites glisser un composant OLE DB Destination (renommé en OLE DB Destination – Employees), puis placez une contrainte de précédence entre Conditional Split et ledit composant, avec New comme option de sortie:

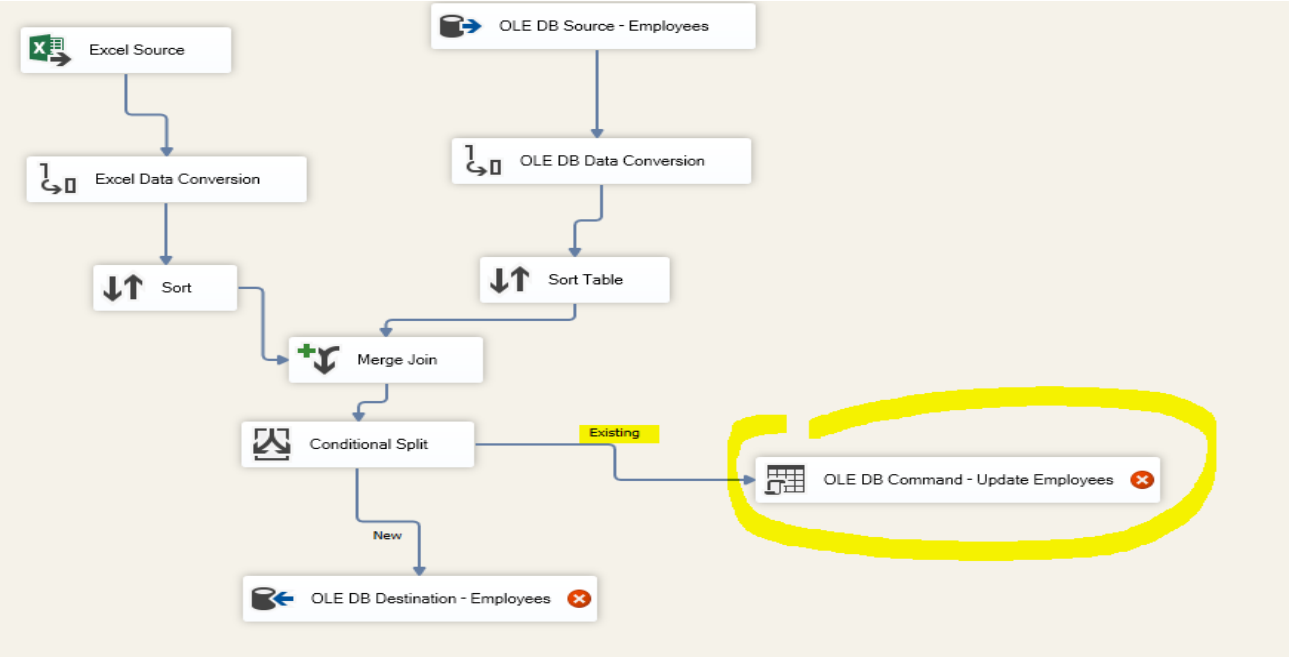


- Configurez OLE DB Destination – Employees comme suit:
  - Optez pour le même gestionnaire de connexion qu'OLE DB Source – Employees, puis spécifiez la table à utiliser: Employee. Dans Data Access Mode selection SQL Command pour spécifier les colonnes à joindre de la table Employee:



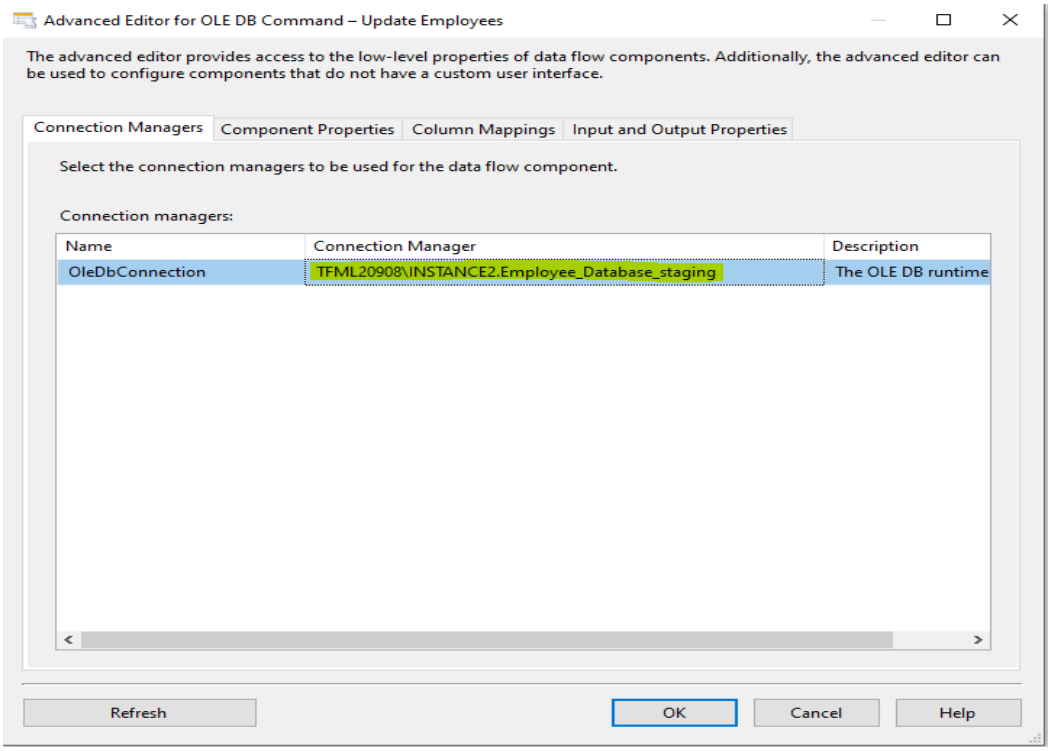
6.5 MISE-À-JOUR DE DONNEES

- Toujours au sein du canevas de Data Flow de Data Flow Task – UPSERT, faites glisser (ou double-cliquez sur) un composant OLE DB Command (renommé en OLE DB Command – Update Employees):
- Placez une contrainte de précédence de succès allant de Conditional Split vers OLE DB Command – Update Employés. L'option de sortie Existing est à choisir:

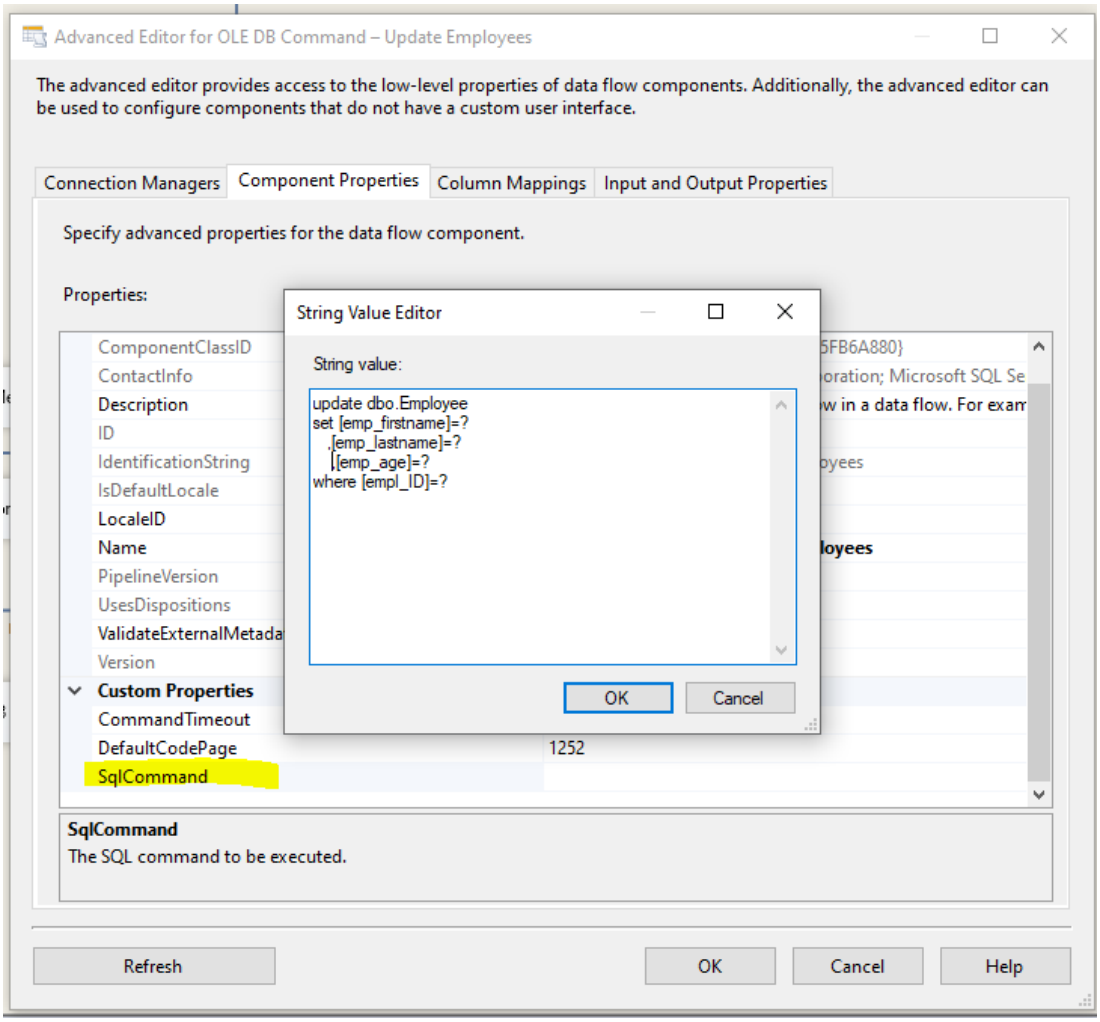


- Au sein des propriétés du composant OLE DB Command – Employees…

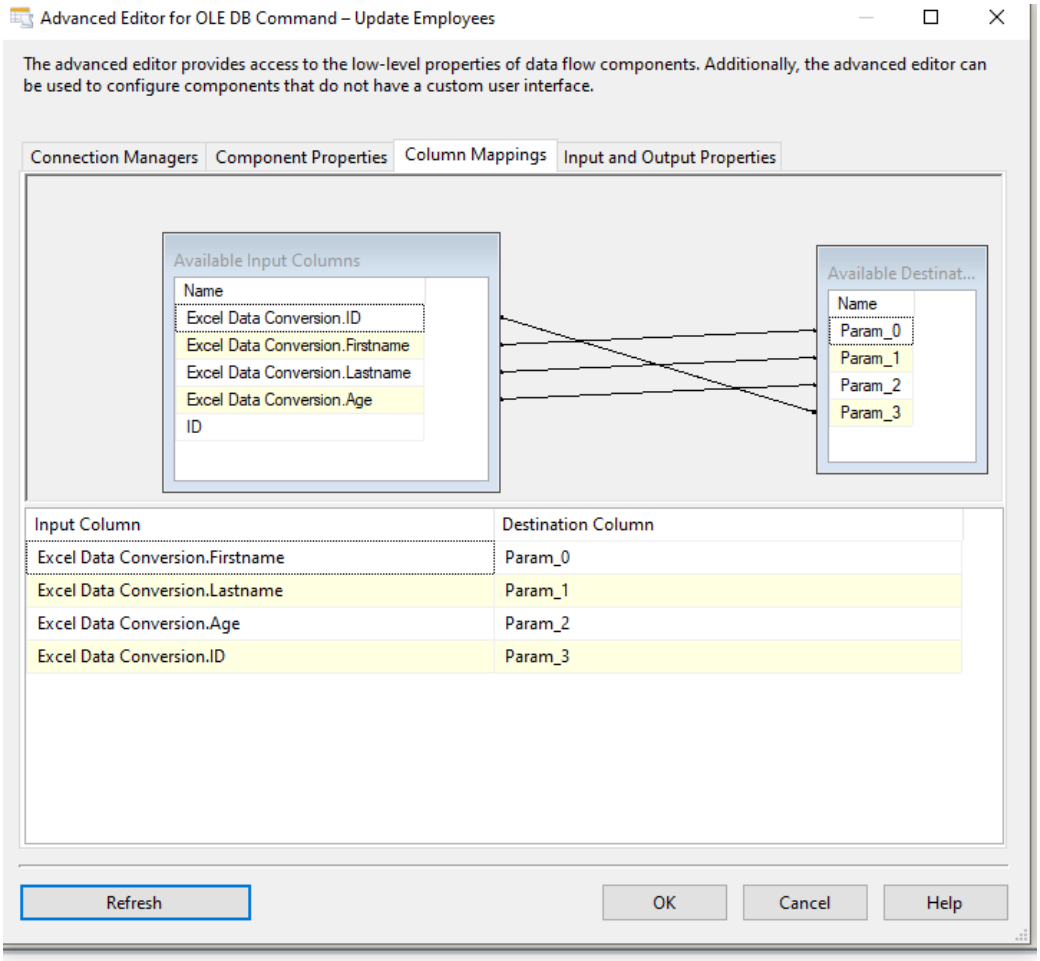
  - Dans la section Connection Manager, spécifiez le gestionnaire de connexion déjà créé lors de l'implémentation du traitement des insertions pour accéder à la table Employee de la base [Employee_Database_staging] de l'instance SQL Server située sur la machine TFML20908:



- Dans la section Component Properties, spécifiez le code SQL paramétré de mise-à-jour de la table Employee dans la zone dédiée à SqlCommand:



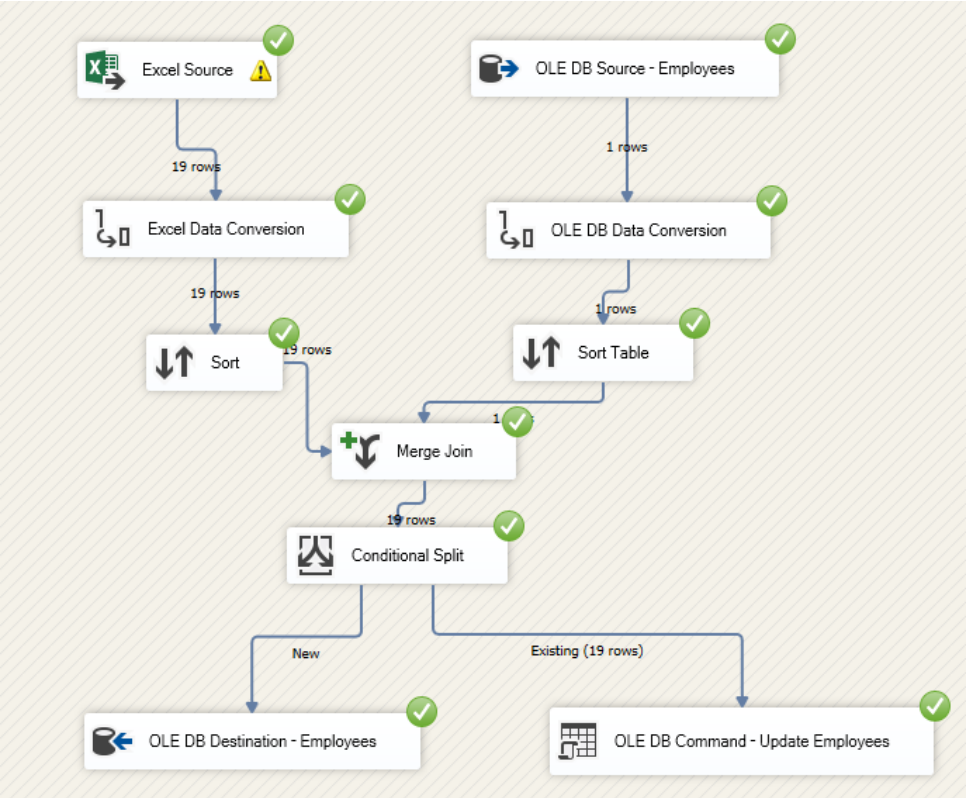
- Dans la section Column Mappings, faites les correspondances entre chaque colonne convertie du fichier Excel et chaque paramètre, sachant que:
  - Param_0 correspond au premier « ? » de notre code SQL spécifié précédemment. Et donc, au paramètre de emp_firstname.
  - Param_1, à celui d'emp_lastname.
  - Param_2, à celui d'emp_age.
  - Param_3, à celui d'empl_ID.



6.6 TESTS DE NOTRE PACKAGE

Dans l'onglet Control Flow ou Data Flow Task – UPSERT, pressez sur F5 pour exécuter les flux de travaux.

Si tout va bien, le résultat sera le suivant:



7 CONCLUSION

L'intégration des données hétérogènes en entreprises, figure aujourd'hui parmi les principaux enjeux des entreprises. Nous l'avons décrit plus haut que la plupart des entreprises de nos jours, traitent et utilisent les donnes issues de différentes sources de données, il est parfois bien compliqué de parvenir à « tout faire marcher » ensemble afin de choisir des décisions dans l'idéale de cas.

Cette difficulté à réunir des éléments très hétérogènes et ne partageant bien souvent que très peu de standards, met en évidence des problématiques de type technologique, économique et organisationnel auxquelles les entreprises doivent faire face. Le volume de données de plus en plus important ainsi que les exigences en matière de confidentialité de prise de décisions, d'accès facile a ces données et de sécurité transforment un peu plus la question de l'intégration des données en un véritable défi et une solution importante pour les entreprises.

Dans cet article, nous avons paraphrasés de cette hypothèse et avons chuté par présenter une étude cas bien illustrée pour une intégration de données issues d'une source de données EXCEL vers un système de gestion de base données sous Microsoft SQL Server, avec les outils de Business Intelligence, nous notons qu'avec une telle solution pour une entreprise, nous pouvons faciliter les accès aux données au moments opportun et faciliter également la prise de décisions.

REFERENCES

[1] Hani Zitout, «Talend Open Studio - Le guide complet pour l'intégration de données» *Editions ENI,* pp. 180–325, 2023.
[2] Anonymous, what is enterprise integration. [Online] Available: https://www.sap.com/canada-fr/products/technology-platform/what-is-enterprise-integration/data-integration.html (October 16, 2023).
[3] Jill Dyché and Evan Levy, Customer Data Integration: Reaching a single version of the truth, 2nd Ed. Wiley India Pvt. Limited, 2006.
[4] Alice LaPlante, building a unified data Infrastructure: Access, Gaven and Share all data with greater consistency and control, O'Reilly Report.
[5] Sheth A., Kashyap V. So Far (Schematically) yet So Near (Semantically), In Proceedings of IFIP DS-5 Conference Semantics of Interoperable Databases Systems, (Nov. 16-20. Lorne, Australia), 1992, pp. 272-301.
[6] Tom Yu, EII, ETL, EAI: Why, What and How, 2005.
[7] Christian SOUTOU et Frederic Brouard, Modélisation des bases de données: UML et les méthodes entité-association, Eyrolles Edition, 2017.
[8] Urban S.D. A Semantic Framework for Heterogeneous Database Environments In Proceedings of RIDE-IMS'91 Interoperability in Multidatabase Systems (April 7-9, Kyoto, Japan), 1991, pp. 156-163.

[9] Dupont Y. Resolving Fragmentation Conflicts in Schema Integration. In Entity-Relationship Approach - ER'94, P. Loucopoulos Ed., LNCS 881, Springer-Verlag, 1994, pp. 513-532.
[10] Lawrence Miller – CISSP, *Oracle Autonomous Database,* Editions DUMMIES 3eme Edition Spéciale.
[11] Anonymous, Explication du système de gestion de bases données. [Online] Avalaible: https://www.ionos.fr/digitalguide/hebergement/aspects-techniques/systeme-de-gestion-de-base-de-donnees-sgbd/ (October 17, 2023).
[12] Microsoft, Microsoft SSIS. [Online] Available: https://www.next-decision.fr/editeurs-bi/etl/microsoft-ssis (October 17, 2023).
[13] Andersson M. Extracting an Entity Relationship Schema from a Relational Database Through Reverse Engineering. In Entity-Relationship Approach - ER'94, P. Loucopoulos Ed., LNCS 881, Springer-Verlag, 1994, pp. 403-419.
[14] Lakshmanan L.V.S., Sadri F., Subramanian I.N. On the Logical Foundation of Schema Integration and Evolution in Heterogeneous Database Systems. In Deductive and Object-Oriented Databases, Ceri S., Tanaka K., Tsur S. (Eds.), LNCS 760, Springer-Verlag, 1993, pp. 81-100.
[15] Saltor F., Castellanos M.G., Garcia-Solaco M. Overcoming Schematic Discrepancies in Interoperable Databases, In Proceedings of IFIP DS-5 Conference Semantics of Interoperable Databases Systems, (Nov. 16-20. Lorne, Australia), 1992, pp. 184-198.
[16] Sheth, A., & Larson, J. A. (1990). Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys, 22* (3), 183–236.
[17] Wiederhold, G. (1992). Mediators in the Architecture of Future Information Systems. *IEEE Computer, 25* (3), 38–49.
[18] Özsu, M. T., & Valduriez, P. (2011). *Principles of Distributed Database Systems* (3rd ed.). Springer. ISBN: 978-1441988331.
[19] Papazoglou, M. P., & Georgakopoulos, D. (2003). Service-Oriented Computing. *Communications of the ACM, 46* (10), 25–28.
[20] Hohpe, G., & Woolf, B. (2003). Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley.
[21] Singh, M. P., & Huhns, M. N. (2005). Service-Oriented Computing: Semantics, Processes, Agents. Wiley. ISBN: 978-0470091482.
[22] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983). *Machine Learning: An Artificial Intelligence Approach.* Springer. ISBN: 978-3540093381.
[23] Sadeghi, A. R., Wachsmann, C., & Waidner, M. (2015). Security and Privacy Challenges in Industrial Internet of Things. *Proceedings of the 52nd Annual Design Automation Conference* (pp. 1–6). ACM. DOI: 10.1145/2744769.2747942.
[24] Garlan, D., & Shaw, M. (1993). An Introduction to Software Architecture. *Advances in Software Engineering and Knowledge Engineering, 2,* 1–39. DOI: 10.1142/9789812798039_0001.
[25] Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley. ISBN: 978-1118530801.