Giuseppe Berio, Michel Dubois

INF2204

« Systèmes d'information décisionnels et entrepôts de données »

Sujet d'étude

Le travail demandé porte sur un cas pratique de mise en œuvre d'intégration de données sur la suite Microsoft SQLServer.

Cas pratique

La PME MoreMovies a récemment décidé d'acquérir trois magasins spécialisés dans la vente et la location de films et de produits dérivés. Auparavant, ces trois magasins avaient trois propriétaires différents et s'appelaient respectivement MovieMegaMart, BuckBoaster et MetroStarlet. Dans chacun de ces magasins, les clients sont identifiés par un code enregistré sur une carte magnétique personnelle. Seuls les clients de ces magasins peuvent y acheter ou louer les produits. De plus, ces magasins gèrent les locations et les ventes à travers une identification électronique des produits. Pour MoreMovies, il devient difficile de traiter les données provenant de chaque magasin car chacun de ceux-ci dispose de son propre système d'information, qui lui est spécifique. MoreMovies souhaite intégrer les données provenant des 3 magasins dans un système unique (entrepôt de données), de manière à effectuer des études des ventes ou locations suivant les produits proposés, les magasins, les clients (homme, femme, tranche d'âge, ...).MoreMovies n'arrive pas à faire des prévisions fiables sur la base de ces données comme, par exemple, évaluer la possibilité d'étendre à l'ensemble des magasins la vente de gadgets ou encore proposer des livres concernant les films ou les secteurs de cinéma. On aimerait par exemple pouvoir faire des analyses fiables concernant les ventes et locations de films, ce qui n'est pas facilement réalisable dans l'état actuel du système. Il vous est demandé de réaliser les étapes initiales pour la construction d'un entrepôt pour MoreMovies. Les sources de données, bases de données relationnelles ACCESS, qui vous sont fournies représentent des **états observés** de chaque système d'information. Aucun document expliquant ces bases n'est disponible (dans la réalité les documents sont peut-être disponibles mais pas fiables). Il est à préciser que les objectifs de MoreMovies ne sont pas forcément atteignables mais ils peuvent l'être partiellement. Vous analyserez les schémas logiques de chaque source (comprenant seulement des tables) et les données mémorisées pour :

1. Produire **pour chaque source un schéma conceptuel** (un diagramme de classes UML) faisant des transformations de schéma d'enrichissement, de normalisation, d'abstraction, de spécialisation, de généralisation des schémas sources ; vérifier que le schéma conceptuel proposé est « faisable » pas rapport aux données disponibles, utilisant des **requêtes appropriées** ; indiquer des **mesures de la qualité des données** telles que

: volume des données, pourcentage de données manquantes, recherche des données aberrantes, ... ; les mesures de qualité peuvent être utiles pour la reconceptualisation des sources et, notamment, pour l'identification de contraintes ; ces mesures seront utiles pour mieux conceptualiser et programmer l'échange de données (flux d'extraction ou d'échange) entre les sources et l'environnement ETL où les sources seraient réimplantées.

2. Réaliser pour chaque schéma conceptuel correspondant à une source, un **schéma logique** (relationnel) ; en fonction de l'outil choisi au point 1, cela peut se faire d'une manière semi-automatique.
3. Conceptualiser les **flux d'échanges**, précisant des **mappings** entres chaque schéma « tel quel » d'une source et le schéma de la source reconceptualisé (produit au points 1 et 2).
4. Produire **un schéma conceptuel intégré** (visualisé comme diagramme de classes UML) à l'aide de **correspondances inter-schémas** (vous pouvez aussi explorer les capacités de SQL Developer pour outiller l'intégration de schémas).
5. Produire un **schéma intégré logique** (relationnel) à partir du schéma conceptuel intégré défini au point 1 ; en fonction de l'outil choisi au point 1, cela peut se faire d'une manière semi-automatique.
6. Conceptualiser les **flux d'intégration**, précisant des **mappings** entres chaque schéma reconceptualisé des sources et le schéma intégré (produit au point 4 et 5).
7. Proposer un **schéma dimensionnel conceptuel basé sur le schéma conceptuel intégré**, avec les dimensions {*produit, magasin, période, client*} et indicateurs d'intérêt (*CA, nombre de clients, nombreventes, nombredelocations*) sur les ventes et les locations réalisées ; les hiérarchies, doivent s'organiser en 2 ou 3 niveaux ; la dimension *produit* devra nécessairement distinguer entre *type* de produit (film, gadget etc .) ; la dimension *client* devra nécessairement comporter l'information sur l'*âge* des clients ; suivez l'approche « supply driven » pour le concevoir, vous basant sur le schéma intégré produit au point 3.
    a. **Rappel.** Un schéma dimensionnel conceptuel décrit un ensemble de mesures/observations (dans un – ou plusieurs - « cube » -s-), et chaque niveau des hiérarchies portées par chaque dimension selon laquelle ces mesures peuvent être obtenues (faits inférés),
    b. Vous pouvez dessiner ce schéma sans utiliser des outils dédiés ou avec des outils dédiés (par exemple, SQLDeveloper) sachant que ces outils dédiés ne serviront qu'à garantir la cohérence entre le schéma dimensionnel et le schéma rolap mais il sera nécessaire de « reporter manuellement » ces schémas suivant les consignes de l'environnement propre à SQLServer.
8. Produire un **schéma logique rolap** suivant le schéma dimensionnel conceptuel et complétez le avec des choix de **modélisation physique tels que des index, de partitions, de vues matérialisées** (sans faire référence à une plateforme technologique particulière).
    a. **Rappel.** Un schéma rolap est en effet le schéma proposant une (ou plusieurs) table(s) de faits liée(s) aux tables de dimension via le mécanisme classique de la clé étrangère,

b. Le schéma rolap logique peut avoir 2 formes de base : **étoile et flocon**. Il faudra donc décider si le schéma proposé sera étoile ou flocon. Bien entendu, le schéma peut assumer la forme en **constellation** en cas de plusieurs tables de faits.

Outils utilisables : Vous pouvez utiliser tout outil permettant la modélisation (par exemple, Visual Paradigm pour une modélisation centrée UML) ou utiliser des représentations visuelles, sachant qu'en fonction de l'outil choisi, certains choses seront à faire manuellement. Vous pouvez aussi faire communiquer plusieurs outils si vous pensez cela utile. Par exemple, on peut créer des échanges entre Visual Paradigm et SQL Developer suivant ces étapes : création d'un diagramme de classes (conceptuel), génération automatique du code pour Oracle, importation du code Oracle dans SQL Developer, visualisation des tables et visualisation du modèle ER. Également, il est possible de faire communiquer SQL Server et SQL Developer via l'importation de scripts relationnels ou la génération du code (relationnel) pour SQL Sever. Cependant, la liaison entre SQL Developer et SQL Server pour la partie dimensionnelle n'est pas fonctionnelle car SQL Server ne possède pas l'instruction « create dimension » ; par conséquent, un schéma dimensionnel conçu en SQL Developer ne sera utile que pour son « dessin » et pour garantir la cohérence avec le schéma rolap, probablement à refaire dans l'environnement SQL Server.

Sous Microsoft SQL Server :

9. Générer toutes les bases de données correspondantes aux schémas logiques conçus aux points 2 et 5 suivant les consignés appropriées.
10. Alimenter, mettant en œuvre les flux nécessaires d'échange de données, conceptualisés au point 3, chaque base de données correspondant à une source, générée au point 9, à partir des données mémorisées dans les sources, avec les nettoyages et standardisations nécessaires à l'aide de l'ETL (SQL Server Integration Services). Ensuite, à l'aide de l'ETL, réaliser les flux d'intégration de données, conceptualisés au point 6, entre les bases correspondantes à chaque source et la base sous-jacente le schéma intégré, générée au point 9. Alimenter la base sous-jacente le schéma intégré exécutant les flux d'intégration.
11. Réaliser le schéma logique rolap et le modèle physique (pour le modèle physique, **il faudra retrouver au sein de SQL Server, si possible, les mêmes mécanismes d'optimisation indiqués** ; si des mécanismes ne seraient pas utiles ou disponibles il faudra reporter les justification techniques dans le rapport final.) conçus au point 8 suivant les consignes appropriées. Alimenter les tables du schéma rolap à partir de la base (alimentée) sous-jacente le schéma intégré, schéma à l'aide de l'ETL (SQL Server Integration Services). De plus vous créerez une solution SQL Server Analysis Services multi-dimensionnelle HOLAP ou une solution SQL Server Analysis Services tabulaire.
12. Écrire **les requêtes** permettant de répondre aux questions suivantes :

    Quels sont les 5 films représentant les plus forts montants de ventes mensuelles ?

Quel est, par produit, le montant mensuel des ventes ?
Quel est l'âge moyen des clients (femmes, hommes) qui louent des films ?
Quels sont, par magasin, les films les plus loués ?

En nombre de locations de film mois, quels sont les mois pour lesquels une baisse de ce nombre de plus de 15% par rapport au mois précédent est constatée ?

Pour « écrire » ces requêtes, vous pouvez utiliser au choix :

a. SQL Server Management Studio connecté au moteur SQL Server Database Engine ou connecté à SQL Server Analysis Services (multidim ou tabulaire),
b. Excel connecté à SSAS (SQL Server Analysis Services),
c. PowerBI Desktop connecté à SSAS (SQL Server Analysis Services) ou connecté à SSDE,
d. SQL Server Reporting Services connecté à SSAS (SQL Server Analysis Services) ou à SSDE.

Les points 9,10, 11 et 12 **doivent** se baser sur Microsoft SQL Server (en particulier les services d'intégration et la gestion de métadonnées). Les étapes 1, 2, 3, 4, 5, 6, 7 et 8 doivent être décrites précisément et complètement (diagrammes UML, autres visualisations utilisées, correspondances, démarche suivie pour intégrer et nettoyer les données, conceptualisation des flux).

Rendus

1. **Document d'étape :**

   les points 1, 2, 3, 4, 5, 6, 7 et 8 ne dépendent pas de l'utilisation de Microsoft SQL Server ; ils feront notamment l'objet d'une étude décrite dans un **document d'étape techniquement précis, complet et détaillé sur les points** 1, 2, 3, 4, 5, 6, 7 et 8. Ce document est à déposer en format pdf sur l'espace de cours, **pour le 08/05/2023** au plus tard.

2. **Rapport final et présentation :**
   a. L'ensemble du travail réalisé fera l'objet d'un **rapport final complet** (reprenant le contenu du document d'étape, éventuellement amélioré) techniquement précis et détaillé, à remettre **pour le 25/5/2023** au plus tard, en document pdf à déposer sur l'espace de cours ;
   b. D'une présentation orale supportée par un **document de présentation** dans la semaine des contrôles. La présentation sera faite lors de la soutenance qui aura lieu le **30 et 31 mai 2023.**

Accès à SQL Server :

L'accès à SQL Server (et aux environnements connectés) se fait à travers la plateforme décisionnelle (et via le bureau à distance si vous êtes en distanciel). Mot de passe et login vous seront distribués au moment venu.

Formation des groupes de projet

Le travail à accomplir doit être fait un groupe. Il est conseillé d'utiliser TEAMS comme outil de communication et de coordination projet. Pour des questions d'organisation, il est obligatoire de composer **des trinômes (un maximum de 4 binômes est accepté, 2 binômes par promotion mais vous pouvez aussi former des trinômes mixtes et dans ce cas, il peut y avoir qu'un seul binôme). Il ne sera pas possible de conduire un travail individuel dans tous les cas. En cas de dépassement des seuils indiqués ci-avant, les responsables choisiront au hasard les binômes et des trinômes seront reformés automatiquement. Il est donc vivement conseillé, d'établir une liste binômes/trinômes pour tous les étudiants et faire des choix par vous-mêmes. Les étudiants n'appartenant à aucun trinôme (binôme) seront automatiquement affectés à un binôme existant ou seront inclus dans un nouveau trinôme/binôme**

La date limite pour former vous-mêmes les groupes de projet est le **07/3/2023.** Ensuite, les enseignants formeront les groupes projet. Pour former un trinôme/binôme il est suffisant d'envoyer un mail au plus tard le 7 mars 2023 à giuseppe.berio@univ-ubs.fr.

Il est demandé de produire un **diagramme de GANTT** pour montrer le planning de projet et le travail individuel ; ce diagramme doit être inclut dans le document d'étape (voir section RENDUS).

Calcul de la note de projet

La note de projet est une note individuelle calculée sur la base de plusieurs notes. Ce calcul est basé sur le principe que si vous n'obtenez pas des résultats satisfaisants aux épreuves individuelles, votre contribution au projet est insatisfaisante, peu importe les livrables rendus, étant donné une faible maitrise de concepts et techniques nécessaires au projet et des techniques.

NI1 : note individuelle (GBerio) /20 (si absent NI1=0)

NI2 : note individuelle (MDubois) /20 (si absent NI2=0)

NII : note individuelle intermédiaire /10 = (NI1+NI2)/4

Note : note individuelle calculée /20

NP : note des 2 livrables / 20 (si non déposé(s) NP=0) – Veuillez noter que cette note est une note par groupe de projet, sauf si contrairement indiqué par les membres de groupe ou par évidence contraire constatée lors de la soutenance.

NS : évaluation soutenance ([0%,100%]) (si absent NS=0%)

Calcul de N (les seuils s1,s2,s3,s4 sont recalculés chaque année ; à titre d'exemple, l'année passée s1=4,43,s2=3,8,s3=2,8,s4=0) :

N=NP*NS si s1<=NI<=10

N=NP*NS*0,9 si s2<=NI<s1

N=NP*NS*0,8 si s3<=NI<s2

N=NP*NS*0,7 si s4<=NI<s3

N=NP*NS*0=0 si NI<s4.

Critères d'évaluation des livrables.

Livrable 1

Le livrable 1 doit contenir les éléments suivants (chaque élément manquant ou non approprié impliquera une décote de la note) :

Descriptif techniquement correct, des analyses des sources : vous pouvez inclure une analyse de qualité à priori (exemple, valeurs non nulles, aberrantes, non à jour etc. – pour cette analyse, il est possible d'utiliser des requêtes ACCESS ou toute autre moyen) qui est, de toute manière nécessaire pour pouvoir justifier les transformations de schéma sur la base d'un raisonnement hypothétique.

Justification de chaque transformation appliquée, supportée par un **raisonnement hypothétique (normalement basé sur des requêtes visualisant des données non respectant certaines contraintes ou des données interprétables plus précisément ou différemment par rapport au schéma)**, si la source contient de données, pour obtenir le schéma de données extraites, pour chaque source.

Liste des correspondances interschéma écrites, étant chaque correspondance écrite dans le format standard :

relation+WCI+WCP (sauf relation = disjoint). Tout format ne respectant pas le standard, ne sera pas pris en compte.

Pour chaque schéma de données extraites, représentation du schéma correspondant, non conflictuel.

Représentation du schéma intégrée et de chaque règle utilisée pour intégrer les schémas non conflictuels.

Représentation du schéma dimensionnel, construit sur la base du schéma intégré, faisant apparaitre, les mesures, les dimensions, les niveaux, le(ou les) cubes nécessaires.

Représentation du schéma rolap (avec précision si schéma étoile ou flocon).

Indication des éléments du modèle physique (en particulier index bitmap qui pourraient être introduits, vues matérialisées pour stocker les pré-calculs, les partitions (pour la (ou les) table(s) de faits principalement).

Livrable 2

Le livrable 2 doit contenir les éléments suivants (chaque élément manquant ou non approprié impliquera une décote de la note) :

Descriptif conforme au schéma standard de chaque flux d'extraction. Ce descriptif doit précision des opérations éventuelles demandant un calcul de similarité.

Descriptif conforme au schéma standard des flux d'intégration, Ce descriptif doit précision des opérations éventuelles demandant un calcul de similarité et une fusion de données.

Descriptif conforme au schéma standard des flux de chargement (dans l'entrepôt).

Descriptif de l'approche suivie pour réaliser les extractions, intégrations, chargements successifs ; ce descriptif peut demander de décrire les changements apportés aux schémas de

données extraites, intégrées et de l’entrepôt pour permettre la réalisation des l’approche choisie.

Pendent la soutenance tous ces éléments seront discutés dans le contexte d’une discussion sur l’aboutissement des flux.