Elaboration d'un Data Warehouse à partir d'un Data Lake

Rabah TIGHILT FERHAT

Institut de Recherche en Informatique de Toulouse (IRIT), Université Toulouse 1 Capitole
2 Rue du Doyen-Gabriel-Marty
31042 Toulouse

rabah.tighilt.ferhat@gmail.com

MOTS-CLÉS : *Réservoir de données, Entrepôt de données, Données massives, Processus décisionnel, Métadonnées, Extraction-Transformation-Chargement.*

KEYWORDS: *Data Lake, Data Warehouse, Big Data, Decision support system, Metadata, Extracting-Transforming-Loading.*

ENCADREMENT. *Gilles ZURFLUH*

1. Contexte

Le développement des bases de données massives (Big Data) pose plusieurs problèmes. Nous citons par exemple : la gestion des données très variées pour fournir de la connaissance. De nouveaux systèmes sont récemment apparus comme une solution à ce problème (Hai et al., 2016). Il s'agit des systèmes appelés « Data Lake » ou « réservoir de données ». Un Data Lake (DL) est un référentiel de stockage et d'exploration de grandes quantités de données brutes peu ou pas structurées permettant d'acquérir de la connaissance (Chessell et al., 2014). Les DL intègrent des données dans leur format d'origine à partir de sources de type Big Data (Hai et al., 2016), (Walker et al., 2015). Généralement, les données d'un Data Lake sont décrites par des métadonnées et organisées d'une certaine manière pour qu'elles soient facilement accessibles à tout moment et à tout utilisateur autorisé à effectuer des activités analytiques (Terrizzano et al., 2015). Notre travail s'intègre dans ce contexte et concerne particulièrement l'élaboration d'un ED à partir d'une source de type Data Lake.

Un ED est une base de données conçue pour supporter des processus d'aide à la décision. Dans un ED, les données sont souvent organisées selon un modèle multidimensionnel (en étoile, en flocon ou en constellation). Dans ce type de modèles, seules les données structurées (nombres et chaînes de caractères) sont utilisées. Or, dans un Data Lake, les données sont généralement complexes (textes brutes, images, fichiers audio et vidéo, etc.). D'où l'intérêt d'adapter les ED pour prendre en considération les caractéristiques d'un DL notamment la variété des données.

Pour illustrer notre travail, nous utilisons le cas des dossiers médicaux partagés (DMP) inspiré de l'application nationale gérée actuellement par la Caisse Primaire d'Assurance Maladie (CPAM). Un dossier médical contient les données de santé et le parcours de soins d'une personne (assuré social). Il permet au médecin traitant et aux autres professionnels de santé (médecins, infirmiers, dentistes, protection civile, etc.), d'accéder et de partager des informations médicales concernant cette personne. Il doit permettre aussi aux médecins de visualiser l'historique médical d'une personne. En outre, le DMP permet aux analystes de la CPAM de traiter les données des patients ou de les confronter avec celles d'autres malades dans le but d'en avoir une meilleure connaissance des dépenses de santé.

2. État de l'art

Dans le contexte des ED, les auteurs de (Bala et al., 2013) ont proposé une approche de modélisation des Big Data dans un système décisionnel. Ils s'intéressent particulièrement à repenser et à adapter les processus ETL (Extracting, Transforming, Loading) à l'aide du modèle MapReduce pour la capture des données massives dans un ED. Par ailleurs, pour construire des ED massifs, les auteurs (Chevalier et al., 2015) ont défini des règles pour traduire un modèle multidimensionnel en étoile, en deux modèles physiques NoSQL, un modèle orienté colonnes et un modèle orienté documents. Les liens entre faits et dimensions ont été traduits sous la forme d'imbrications. (Dehdouh et al., 2015) traitent de l'implantation des entrepôts de données de grande taille (Big Data). Le processus d'implantation repose sur une architecture à 3 niveaux : conceptuel, logique et physique. Ils proposent trois approches permettant de construire des ED volumineux en utilisant le modèle NoSQL orienté colonnes. Les trois approches utilisent les spécificités des systèmes Big Data en privilégiant certains choix de stockage des faits et des dimensions. (Li et al., 2010) ont étudié les mécanismes d'implantation d'une base de données relationnelle dans le système HBase ; le but est d'élaborer un entrepôt de données NoSQL orienté colonnes. La méthode proposée est basée sur des règles de correspondance entre un schéma relationnel et un schéma HBase. Les relations entre les tables (clés étrangères) sont traduites par l'ajout des familles de colonnes contenant des références.

A notre connaissance, peu de travaux ont proposé des approches pour le stockage et l'exploration des Data Lake. Les auteurs de (Walker et al., 2015) ont proposé un système de stockage et de gestion des Data Lake. Le remplissage et l'interrogation des Data Lake reposent sur une gestion des métadonnées pour les données structurées

(relationnelles) et semi- structurées (XML, CSV, JSON, etc.). Pour les données non structurées (textes, images, fichiers audio et vidéo, etc.), les auteurs ont défini une « Mesure de gravitation » pour attribuer une densité à ces données. Dans (Hai et al., 2016), il a été proposé un système appelé « Constance », qui permet de stocker et de traiter des données dans un Data Lake. Constance se focalise particulièrement sur la gestion des métadonnées explicites et implicites. L'interrogation de données se base sur un langage de requêtes par mots-clés. De plus, Constance fournit une interface qui permet aux utilisateurs de contrôler la qualité des données en choisissant des métriques de qualité des données définies au préalable.

Il apparait à travers cet état de l'art que peu de travaux ont étudié la construction d'un ED à partir d'une source de données massives. Dans les travaux les plus proches de notre problématique notamment les travaux de (Bala et al., 2013), (Chevalier et al., 2015), (Dehdouh et al., 2015) et (Li et al., 2010) ; généralement, le schéma de données Big Data est défini préalablement (i.e. avant de commencer la saisie). Dans un Data Lake, le schéma n'est connu que partiellement. Ceci est dû au fait que les traitements décisionnels ne sont pas connus au moment du stockage.

3. Problématique

Notre problématique consiste à proposer un outil capable d'extraire les données d'un Data Lake puis de les restructurer dans le but d'effectuer des traitements décisionnels. En effet, les Data Lake ont deux problèmes majeurs, structurel et sémantique : (1) Problème structurel : le schéma de données dans un DL n'est connu que partiellement. L'absence du schéma est dû au fait que les traitements décisionnels ne sont pas définis préalablement, (2) Problème sémantique : lors de la création des DL, la sémantique des données est peu connue ; cela est en raison de l'absence d'une gestion efficace des métadonnées. Par conséquent, l'interrogation des DL s'avère être difficile. Ainsi, nous visons à proposer une approche pour la construction d'un ED à partir d'un Data Lake en résolvant à la fois le problème structurel et le problème sémantique des données. Nous proposons une approche pour la construction d'un ED à partir d'un Data Lake.

Notre problématique générale se résume à travers les trois questions suivantes :

– Comment obtenir le schéma des données brutes ?

– Comment relier les éléments du schéma sur un plan sémantique ?

– Comment restructurer les données brutes dans un ED ?

4. Actions réalisées

Comme actions réalisées, nous avons (1) commencé l'implantation de l'étude de cas présentée dans la section contexte, sur un cluster de 3 machines. Chaque machine est de type Intel Core i5, 8 Go de RAM et 2 To de disque. L'une de ces machines est configurée pour agir comme maitre et les autres comme esclaves, (2) fait un état de

l'art sur les systèmes de gestion des données massives, en particulier les systèmes Data Lake.

5. Actions futures

Nous allons proposer des solutions aux sous problèmes considérés dans la section problématique. D'une part, nous visons à pouvoir extraire le schéma de données en nous basant sur les métadonnées explicites et implicites. Pour ceci, le travail (Hai et al., 2016) peut nous être utile. D'autre part, nous allons proposer un outil permettant de restructurer les données en fonction des besoins des décideurs.

Bibliographie

Chessell, Mandy, Scheepers, Ferd, Nguyen, Nhan, et al.Governing and managing big data for analytics and decision makers. IBM Redguides for Business Leaders, 2014.

Terrizzano, Ignacio G., Schwarz, Peter M., Roth, Mary, et al. Data Wrangling : The Challenging Yourney from the Wild to the Lake. In : CIDR. 2015.

Hai, R., Geisler, S., Quix, C. (2016, June). Constance : An intelligent data lake system. In Proceedings of the 2016 International Conference on Management of Data (pp. 2097-2100). ACM.

Walker, Coral et Alrehamy, Hassan. Personal data lake with data gravity pull. In : Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on. IEEE, 2015. p. 160-167.

Bala, Mahfoud et Alimazighi, Zaia. Modélisation de processus ETL dans un modèle MapReduce. In : Conférence Maghrébine sur les Avancées des Systèmes Décisionnels (ASD'13). 2013. p. 1-12.

Chevalier, M., M. E. Malki, A. Kopliku, O. Teste, et R. Tournier (2015). Entrepôts de données multidimensionnelles nosql. EDA.

Dehdouh, Khaled, et al. "Using the column oriented NoSQL model for implementing big data warehouses." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015.

Li, Chongxin. Transforming relational database into HBase : A case study. In : Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on. IEEE, 2010. p. 683-687.