Les entrepôts de données pour les nuls... ou pas !

Cécile Favre*, Fadila Bentayeb*, Omar Boussaid*, Jérôme Darmont*,
Gérald Gavin**, Nouria Harbi*, Nadia Kabachi**, Sabine Loudcher*

Université de Lyon
*ERIC - Lyon 2
{prenom.nom}@univ-lyon2.fr
**ERIC - Lyon 1
{prenom.nom}@univ-lyon1.fr

Résumé. Dans cet article, nous portons notre regard sur l'aide à la décision du point de vue des systèmes décisionnels au sens des entrepôts de données et de l'analyse en ligne. Après avoir défini les concepts qui sous-tendent ces systèmes, nous nous proposons d'aborder les problématiques de recherche qui leur sont liées selon quatre points de vue : les données, les environnements de stockage, les utilisateurs et la sécurité.

1 Introduction

Le processus décisionnel ou les systèmes décisionnels au sens des entrepôts de données sont nés d'un besoin exprimé par les entreprises qui n'était pas satisfait par les systèmes traditionnels de bases de données. En intégrant la technologie des entrepôts de données (*data warehouses*), le processus décisionnel apporte une réponse au problème de la croissance continuelle des données pouvant être de formats différents. De plus, il supporte efficacement les processus d'analyse en ligne (*On-Line Analytical Processing* - OLAP) (Chaudhuri et Dayal, 1997; Chaudhuri et al., 2011).

L'entreposage de données est donc né dans les entreprises. Ainsi, les "grands comptes" sont les principaux utilisateurs de ces technologies qui font partie intégrante de l'entreprise comme outil d'aide à la décision (le terme de *Business Intelligence* est aussi largement utilisé). Nous pouvons citer les secteurs de la grande distribution, des banques et des assurances, ainsi que ceux de l'automobile et des institutions médicales. Mais bien au-delà, l'entreposage de données suscite de plus en plus d'intérêt, avec une ouverture vers des entreprises plus petites mais qui peuvent tirer parti aujourd'hui de ces outils. Notons aussi que plusieurs domaines d'application ont vu le jour autour du Web, des systèmes d'informations géographiques, des flux de données, etc. Le Web est par ailleurs devenu une source de données à part entière.

Dans cet article, nous nous attachons à aborder la thématique de l'aide à la décision au travers du prisme de ces systèmes décisionnels en exposant leur fonctionnement, en faisant état des travaux de recherche réalisés. Mais il s'agit aussi de tenter de cerner les enjeux des recherches futures dans ce domaine par rapport à l'évolution du contexte actuel, et ce aux niveaux technologique et économique en particulier avec le succès de l'informatique dans le

nuage (*Cloud Computing*) et des outils libres (*Open Source*) entre autres. En effet la prolifération des outils libres et la possibilité de délocaliser les données dans le nuage ouvre un accès à ce processus décisionnel à un plus grand nombre d'utilisateurs et crée de nouveaux verrous scientifiques.

Cet article est organisé de la façon suivante. Dans un premier temps, nous définissons les concepts clés du domaine des entrepôts de données et de l'analyse en ligne dans la section 2. Nous abordons ensuite les quatre volets qui nous apparaissent cruciaux, à savoir les données (section 3), les environnements de stockage de ces données (section 4), les utilisateurs (section 5) et la sécurité (section 6), en détaillant pour chacun de ces volets les tendances qui se dessinent pour l'avenir. Nous concluons finalement dans la section 7.

2 L'informatique décisionnelle dans tous ses états

2.1 Préambule

Contrairement à certains processus fondés uniquement sur l'utilisation d'outils logiciels, un processus décisionnel est un projet qui se construit. Il doit s'insérer dans un cadre pouvant prendre en compte des données, des informations et des connaissances. L'approche d'entreposage de données ("data warehousing") constitue un champ de recherche important dans lequel de nombreux problèmes restent à résoudre. Les entrepôts de données sont généralement intégrés dans un système d'aide à la prise de décision où l'on distingue deux espaces de stockage : l'entrepôt de données et les magasins de données. Une architecture du processus décisionnel est représentée dans la Figure 1 (Bentayeb et al., 2009).

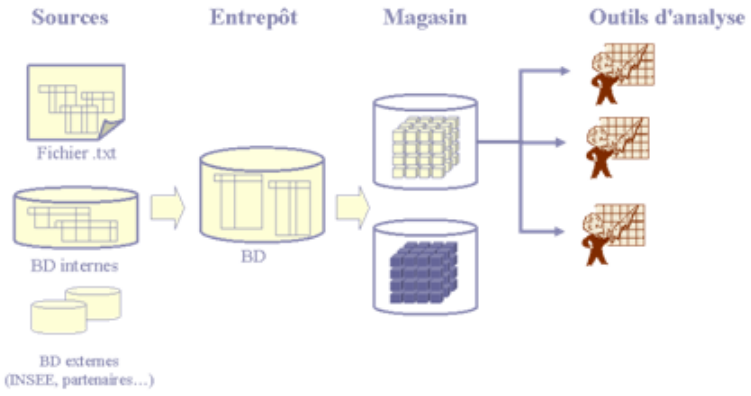

FIG. 1 – *Architecture générale d'un système décisionnel.*

Plusieurs auteurs ont défini le concept d'entrepôt de données. Selon Inmon (1996), c'est une collection de données orientée sujets, intégrée, non volatile et en mode de lecture seule, importée à partir de sources de données hétérogènes et stockée à différents niveaux de granularité dans un but de prise de décision. Ainsi, un entrepôt de données est généralement vu comme

un espace de stockage centralisé regroupant dans un format homogène les données issues de différentes sources, qui peuvent faire l'objet de transformations et d'historisation, à des fins d'analyse pour la prise de décision. Un magasin de données peut constituer un extrait de l'entrepôt, où les données sont préparées de manière spécifique pour faciliter leur analyse et leur exploitation par un groupe d'utilisateurs, en fonction par exemple d'une orientation métier.

Finalement, les possibilités d'analyse des données sélectionnées sont très variées. Elles dépendent des besoins des utilisateurs et font appel à des techniques différentes :

- le *reporting* avec la construction de tableaux de bord, d'indicateurs, de graphiques ;
- la navigation multidimensionnelle dans les données avec la technologie OLAP ;
- la fouille dans les données à l'aide des méthodes de *Data Mining*.

2.2 Modélisation et alimentation de l'entrepôt

2.2.1 Modélisation multidimensionnelle

Les modèles multidimensionnels ont pour objectif de proposer un accès aux données intuitif et très performant. Pour cela, les données sont organisées autour des faits que l'on cherche à analyser, caractérisés à l'aide d'indicateurs (appelés mesures) qui sont des données normalement numériques et additives, permettant de mesurer l'activité modélisée. Ces faits sont décrits par un ensemble d'axes d'analyse, ou dimensions, d'où le terme de modèle multidimensionnel.

Ce modèle de base correspond au modèle en étoile (Kimball et al., 2000; Chaudhuri et Dayal, 1997). Citons l'exemple classique de faits concernant des ventes, dont les mesures sont la quantité commandée et le prix correspondant. Les dimensions (clients, produits concernés, dates, etc.) sont des descripteurs des faits de vente. Ainsi, pour un client donné, un produit, une date, nous disposons de la quantité commandée et du prix correspondant.

Si l'on considère une implémentation en relationnel (ROLAP), les faits seront dans une table (table de faits) et chacune des dimensions sera dans une table respectivement (tables de dimension), chacune étant reliée à la table des faits. Les avantages de ce modèle sont la facilité de navigation, grâce à la table de faits centrale, de bonnes performances en raison du faible nombre de jointures à effectuer pour l'analyse sur une dimension donnée et des agrégations faciles des mesures.

La modélisation en flocons est une première variante du modèle en étoile. Il consiste à décomposer les dimensions d'un modèle en étoile en des hiérarchies explicites, chacun des niveaux de la hiérarchie correspondant à une table dans une implémentation ROLAP. Cette modélisation permet de réduire le volume de stockage et autorise des analyses par paliers sur la dimension hiérarchisée. En revanche, les jointures nécessaires pour accéder aux données entraînent une dégradation des performances.

Finalement, la modélisation en constellation consiste à fusionner plusieurs modèles en flocons, permettant le partage de certaines dimensions par plusieurs ensemble de faits.

2.2.2 Alimentation

L'alimentation d'un entrepôt de données est une phase essentielle dans le processus d'entreposage. Elle se déroule en plusieurs étapes : extraction, transformation, chargement et rafraîchissement des données, qui sont prises en charge par le processus d'ETL (*Extracting, Transforming and Loading*). Ce processus constitue la phase de migration des données de production

dans le système décisionnel après qu'elles ont subi des opérations de sélection, de nettoyage et de reformatage dans le but de les homogénéiser. Cette phase constitue une étape importante et très chronophage dans la mesure où on l'estime à environ 80% du temps de mise en place de la solution décisionnelle. Ainsi cette phase fait l'objet de nombreux travaux de recherche, en terme de modélisation, d'automatisation du processus (Simitsis et al., 2010; Jovanovic et al., 2012; Papastefanatos et al., 2012; Akkaoui et al., 2011; Muñoz et al., 2009).

2.3 Analyse en ligne

L'analyse en ligne constitue un autre aspect du processus d'entreposage des données. Codd (1993) a défini l'OLAP comme "l'analyse dynamique d'une entreprise qui est requise pour créer, manipuler, animer et synthétiser l'information des modèles d'analyse de données. Cela inclut la capacité à discerner des relations nouvelles ou non anticipées entre les variables, la capacité à identifier les paramètres nécessaires pour traiter des grosses quantités de données, la création d'un nombre illimité de dimensions". Un système OLAP est un dispositif muni d'opérateurs spécifiques permettant l'analyse en ligne des données. Il est également considéré comme un serveur d'applications pouvant traiter directement les données d'un entrepôt ou pouvant être utilisé comme un outil d'exploration de données grâce à une navigation interactive. Les applications OLAP permettent entre autres de travailler sur des données historiques pour étudier les tendances ou les prévisions d'une activité, ou de travailler sur des données récapitulatives pour créer de l'information stratégique pour la prise de décision. L'analyse en ligne peut aussi bien s'appliquer aux données de l'entrepôt qu'à celles d'un magasin de données. Généralement, elle est plutôt effectuée sur une collection de données encore plus fine appelée cube de données.

2.3.1 Cubes de données

Le modèle multidimensionnel permet d'organiser les données selon des axes représentant des éléments essentiels de l'activité d'une entreprise. Trois niveaux de représentation des données sont définis dans le processus décisionnel : l'entrepôt qui regroupe des données transversales à l'ensemble des métiers de l'entreprise, le magasin de données qui est une représentation verticale des données portant sur un métier particulier et enfin le cube de données (ou hypercube). Le cube correspond à une vue métier où l'analyste choisit les mesures à observer selon certaines dimensions. Un cube est une collection de données agrégées et consolidées pour résumer l'information et expliquer la pertinence d'une observation. Le cube de données est exploré à l'aide de nombreuses opérations qui permettent sa manipulation.

2.3.2 Opérateurs OLAP

De manière générale, il existe deux classes d'opérations. La première, liée à la structure des données, permet de la manipuler pour mettre en relief la pertinence de certaines informations. Les opérations de manipulation des données multidimensionnelles permettent de réorienter la vue multidimensionnelle ou d'en changer l'agencement en agissant sur la position des membres des dimensions et des mesures : rotation (*rotate*), permutation (*switch*) , division (*split*), emboitement (*nest*), enfoncement (*push*) et retrait (*pull*) . La deuxième classe d'opérations est liée à la granularité des données. Ces opérations agrègent et résument les données

ou les détaillent et permettent une analyse par paliers : agrégation (*roll up*), forage vers le bas (*drill down*). Dans ce cas, on a recours à une opération d'agrégation qui est appliquée sur la (ou les) mesure(s) étudiée(s) (somme, moyenne, max, min, etc.). Ces deux derniers opérateurs sont largement évoqués dans les travaux de recherche contrairement à ceux de la première catégorie. En effet, ils se basent sur les hiérarchies et soulèvent donc les problèmes de complexité des hiérarchies à modéliser (Malinowski et Zimányi, 2004) et d'additivité des données (Mazón et al., 2009).

2.4 Un point sur le positionnement par rapport aux bases de données classiques

Généralement, le processus décisionnel est basé sur un entrepôt de données qui constitue son élément central. Il est alors intéressant de comprendre ce qu'est ce concept de stockage des données et de le positionner par rapport aux bases de données classiques.

La règle-clef du développement d'une base de données traditionnelle est d'optimiser le traitement efficace d'un ensemble de transactions. En effet, les bases de données classiques sont dites transactionnelles car elles sont conçues pour des opérations quotidiennes. Ces transactions nécessitent des données détaillées et actualisées. Elles lisent ou mettent à jour des enregistrements accessibles par leur identifiant. Elles sont conçues pour refléter une sémantique plutôt opérationnelle en minimisant les conflits et en garantissant la persistance des données avec un minimum de redondance et un maximum de contrôle d'intégrité. Les requêtes visent un nombre relativement peu important d'enregistrements. Le but est de mettre à jour les données pour garder une trace des événements de l'entreprise. Ces bases de données sont qualifiées alors de production. Elles sont orientées vers des applications de type OLTP (*On-Line Transactional Processing*).

OLAP, autrement dit l'analyse en ligne, est une démarche orientée "aide à la décision". Les données sont stockées dans un entrepôt de données, où elles sont historisées, résumées et consolidées. Le volume de données des entrepôts est important et va de centaines de gigaoctets à des téraoctets, voire même encore davantage de nos jours. Les entrepôts de données stockent des données collectées au cours du temps, en provenance de plusieurs bases de données opérationnelles. Le temps de réponse doit être court. Pour cela, il est nécessaire d'agréger les données afin d'apporter des réponses rapides à des requêtes pouvant être posées à de multiples niveaux. Il est nécessaire d'optimiser les requêtes les plus fréquemment utilisées afin d'améliorer les temps de réponse. Divers travaux se sont intéressés à cette question de l'optimisation de performances qui est cruciale dans ce contexte d'analyse en ligne. Un entrepôt de données vise à répondre à un utilisateur en termes d'informations et non en termes d'applications (Franco, 1997). Ainsi les systèmes transactionnels et les systèmes d'analyse en ligne ne peuvent coexister dans un même environnement de base de données de par leurs caractéristiques différentes (Codd, 1993), même si un entrepôt de données peut être stocké de manière relationnelle.

2.5 Outils

Le domaine des entrepôts de données est né dans les entreprises. Et c'est aujourd'hui un secteur en pleine expansion avec de nombreux projets décisionnels qui se construisent. La question de la mesure du retour sur investissement se pose alors. Le recours à des technologies de type "*Open Source*" peut constituer une alternative au coût de mise en place de tels projets.

Les outils proposés actuellement sont de plus en plus nombreux également et il est souvent difficile de s'y retrouver. L'objectif n'est pas ici d'en faire une liste exhaustive. Notons d'ailleurs, l'intérêt d'un éventuel travail qui viserait à recenser et positionner (cartographier) tous ces outils, un tel travail étant pour le moment inexistant à notre connaissance, malgré son intérêt indéniable.

Nous pouvons distinguer les outils selon ce qu'ils couvrent comme fonctionnalités. Citons par exemple les deux ETL *Open Source* les plus connus : Kettle (Pentaho Data Integration) et Talend.

Mentionnons également les moteurs OLAP tels que Mondrian (Open Source) qui permettent, à partir d'en entrepôt stocké dans un système de gestion de bases de données relationnelles, de pouvoir construire les cubes de données, qui peuvent être ensuite interfacés avec des applications de visualisation (telles que JPivot, Pentaho Analyzer, Pentaho Analysis Tool, Geo Analysis Tool, etc.)

Nous pouvons également citer d'autres outils connus qui sont dédiés au reporting tels que JasperSoft (Open Source), QlikView, BusinessObject.

3 Des données à tous les niveaux

3.1 Complexité des données

Les entrepôts de données et l'OLAP sont des technologies relativement bien maîtrisées quand il s'agit de données "simples". Cependant, la communauté scientifique s'accorde pour dire que, avec l'avènement du Web et la profusion des données multimédias (son, image, vidéo, texte...), les données sont de plus en plus hétérogènes, diverses et qu'elles sont devenues complexes. L'avènement des données complexes a remis en cause le processus d'entreposage et d'analyse des données ; il a induit l'émergence de nouveaux problèmes de recherche comme l'intégration des données complexes dans les entrepôts, le stockage, la représentation ou la modélisation, l'analyse en ligne et la fouille de telles données.

L'informatique décisionnelle tente de s'adapter aux données complexes depuis plusieurs années. De nombreuses adaptations ou évolutions pourraient être citées. Par exemple, les opérateurs OLAP, comme celui d'agrégation (souvent basé sur la somme ou la moyenne), sont définis pour des données classiques (numériques) et ils deviennent inadaptés quand il s'agit de données complexes, par exemple composées de textes, d'images, de sons ou de vidéos. Plusieurs équipes de recherche travaillent sur ce problème clef d'agrégation des données complexes, par exemple textuelles (Ravat et al., 2008), ou images (Jin et al., 2010). D'autres équipes travaillent sur l'association des Systèmes d'Information Géographique, des entrepôts de données et de l'analyse OLAP pour créer le SOLAP (*Spatial OLAP*) (Bédard et Han, 2009). Les données spatiales sont une forme de données complexes. En effet, dans un cube de données spatiales, les dimensions et les mesures peuvent contenir des composantes spatiales ou géométriques. Un autre exemple de données complexes est celui des flux de données (*data stream*). Dans ces flux, les analystes souhaitent détecter des changements dynamiques par une analyse en ligne. On parle de fouille de flots de données multidimensionnelles, d'*OLAPing Stream Data* ou de *Stream cube* (Han et al., 2005). Enfin, le *XOLAP* (ou *XML OLAP*) cherche à faire des analyses OLAP sur des documents XML tout en tenant compte de leurs spécificités (hiérarchies multiples, imbriquées, incomplètes...) (Wang et al., 2005).

Ces déclinaisons de l'OLAP sont des exemples d'adaptation des entrepôts de données et de l'OLAP aux différents types de données, mais elles ne portent souvent que sur la structure des données et non pas sur leur contenu. Une autre spécificité des données complexes réside dans la sémantique qu'elles véhiculent. Par conséquent, un nouveau problème émerge : comment prendre en compte la sémantique contenue dans les données complexes pour la modélisation et l'analyse ? Le recours à des solutions telles que les ontologies constitue une issue prometteuse explorée dans différents travaux (Cao et al., 2006; Selma et al., 2012).

3.2 Volume des données

Parallèlement à cette problématique de la sémantique des données, la question du volume de ces données peut également poser problème au niveau de leur requêtage en terme de performance. En effet, les requêtes décisionnelles s'appliquent sur de très grandes quantités de données. Elles nécessitent pourtant des temps de réponse ne dépassant pas quelques secondes ou quelques minutes. Il existe plusieurs techniques traitant le problème de l'amélioration des performances des requêtes avec un souci constant de l'optimisation en utilisant des techniques issues des bases de données : la matérialisation des vues, l'indexation, la fragmentation, etc. (Aouiche et Darmont, 2009; Benkrid et Bellatreche, 2011) (se basant souvent sur l'exploitation d'algorithmes de fouille de données)

La production croissante de données, le partage des informations entre utilisateurs, la diffusion des données via les réseaux engendrent de très gros volumes de données disponibles et intéressantes à analyser. L'expression anglaise *Big Data* est utilisée pour désigner des données dont le volume est tel qu'il devient difficile de les stocker, de les interroger, de les modéliser, de les analyser et de les visualiser avec les outils et architectures informatiques existants, du fait également de leur manque de structure.

En effet, la prolifération de très grandes quantités de données, produites principalement par le Web, notamment par les grands acteurs d'Internet et les réseaux sociaux, engendre des évolutions technologiques qui posent de réels problèmes scientifiques. Les volumes de données à très grandes échelles nécessitent des moyens de stockage appropriés (Agrawal et al., 2011). L'utilisation de nouvelles unités de mesures de stockage, telles que les peta-octets voire les zeta-octets sont aujourd'hui des réalités. Outre le stockage, l'exploitation de telles données soulève également de nouveaux challenges scientifiques. De nombreux travaux de recherche proposent aujourd'hui des solutions de gestion de données à très grande échelle. Disposer en ligne de plus en plus de données historisées pour l'analyse est un besoin réel pour les grands acteurs d'Internet ainsi que pour d'autres entreprises, entrainant une expansion des bases de données orientées analyse, tels que les entrepôts de données. L'informatique dans le nuage tente d'apporter des réponses à ces problèmes.

4 Environnement de stockage

L'informatique décisionnelle (*Business Intelligence*) a beaucoup évolué depuis une trentaine d'années passant d'une discipline exclusivement réservée à un groupe d'utilisateurs, les décideurs, pour se démocratiser en délocalisant la prise de décision du haut de la pyramide au plus proche du terrain pour une meilleure réactivité. L'enjeu est de disposer de la bonne

information afin de délivrer la bonne connaissance à la bonne personne. Cela passe par le déploiement d'un environnement de stockage qui doit permettre de rendre accessible, de mettre en forme et de présenter les informations clés aux différents utilisateurs concernés afin de faciliter la prise de décision.

4.1 Au-delà du relationnel, les entrepôts continuent

Comme nous l'avons vu, l'architecture d'un système décisionnel est généralement vue comme une architecture à trois niveaux :

- les sources d'information qui correspondent à l'ensemble des bases de données de production et sites dont sont extraites les informations ;
- l'entrepôt qui contient l'ensemble des données extraites de ces sources ;
- les magasins extraits de l'entrepôt et dédiés aux différentes classes de décideurs.

Les sources d'informations utiles aux décideurs peuvent être stockées sur des sites de nature diverse (sites Web, bases de données...). Cependant, avec l'avènement des données très volumineuses, peu ou pas structurées (*Big Data*), le monde traditionnel des bases de données relationnelles, support des entrepôts de données, n'est plus adapté pour gérer et traiter ces grandes masses de données de type texte, image, etc. provenant du Web, des publications sur les média sociaux, les logs des serveurs Web et des applications, etc. Pour faire face à ces énormes volumes de données, de nouvelles technologies sont apparues comme Hadoop, MapReduce ou les bases de données NoSQL (*Not only SQL*) (Cattell, 2011; Leavitt, 2010). Pour autant, est-ce que l'émergence de ces nouvelles technologies *Big Data* signe la fin des entrepôts de données ? Nous pensons que les bases de données NoSQL n'ont pas la vocation de remplacer les bases de données relationnelles, mais de les compléter selon les besoins des entreprises en proposant une alternative pour adapter le fonctionnement des bases de données à des besoins spécifiques.

Le terme NoSQL fait en fait référence à une diversité d'approches, classées en quatre catégories de bases de données : les bases de données orientées colonnes (comme MonetDB [1]) , les bases de données orientées graphes (comme Neo4J [2]), les bases de données orientées clé/valeur (comme Riak [3]) et les bases de données orientées documents (comme MongoDB [4]). Les différents systèmes de gestion de bases de données qui supportent les bases de données NoSQL sont destinés à manipuler de gigantesques bases de données pour des sites Web tels que Google, Amazon, ou Facebook. En abandonnant les propriétés ACID (Atomicité, Cohérence, Isolation et Durabilité) des bases de données relationnelles, les bases de données NoSQL permettent une montée en charge élevée et assurent une grande performance.

L'architecture décisionnelle "traditionnelle" avec sa base de données centralisée n'est donc plus l'unique architecture de référence. En effet, nous pensons que dans un contexte *Big Data*, il est important de mettre en place d'autres architectures décisionnelles, notamment pour la prise en compte à la fois de données peu ou pas structurées et le passage à l'échelle.

1. http ://www.monetdb.org/
2. http ://neo4j.org/
3. http ://basho.com/
4. http ://www.mongodb.org/

4.2 Jusque dans les nuages

Les solutions de bases de données orientées analyses doivent vérifier les mêmes propriétés que celles des environnements dans le nuage, à savoir : fiabilité, évolutivité, sécurité, bonne performance, tolérance aux pannes, capacité de fonctionner dans un environnement hétérogène, flexibilité de requêtes...

Les problèmes liés aux entrepôts de données et à l'analyse en ligne (OLAP) sont à réétudier dans le cadre des environnements de *Cloud Computing* et cela augure des perspectives prometteuses de combinaisons de ces deux technologies. Entreposer des données à très grande échelle suppose des moyens de traitements à grande échelle également. Le *Cloud Computing* offre ces moyens grâce à une association de plusieurs clusters regroupant un très grand nombre d'ordinateurs. Cependant, le recours à une telle infrastructure n'est pas gratuit. Il fonctionne selon un modèle de facturation à l'utilisation. Ceci engendre un ensemble de problèmes scientifiques à étudier pour mettre au point des approches techniquement et économiquement viables.

Par ailleurs, l'un des points cruciaux à prendre en charge porte sur la virtualisation des données. C'est un problème ouvert. La répartition, la réplication et la distribution des données à travers les nœuds des clusters nécessitent des modèles de données appropriés aux environnements du *Cloud*. Ceux-ci doivent permettre à l'utilisateur de ne voir que ses données.

Un autre point crucial à considérer porte sur les traitements des données. Il existe déjà des travaux dans la littérature, dont certains préconisent des approches basées sur les bases de données parallèles privilégiant les performances (Abouzeid et al., 2009). D'autres sont plus favorables à des solutions utilisant le paradigme de *MapReduce*, mettant en avant son adéquation avec des traitements répartis sur des données distribuées (Stonebraker et al., 2010). Cependant, *MapReduce* est plutôt adapté pour les données non structurées et s'illustre par sa congruence à des environnements tels le *Cloud*. Cependant, le traitement de requêtes réparties sur plusieurs nœuds ainsi que l'équilibrage des charges (requêtes) et des données sur les différents nœuds sont de réels challenges. L'apparition de nouveaux nœuds peut impliquer des changements de stratégie de répartition, de réplication et de distribution des données et des traitements. Ceci demeure un problème ouvert. La conception de démarches, utilisant les deux techniques de parallélisation et de partitionnement des données, constitue certainement une perspective prometteuse pour les entrepôts de données dans le *Cloud*.

Construire des entrepôts de données sur le *Cloud* devrait tenir compte des contraintes de ce dernier et plus particulièrement de la tarification de l'usage des ressources. Il s'agit de la notion d'élasticité qui constitue un argument financier convaincant. L'utilisateur peut demander des ressources selon ses préférences. Il peut avoir besoin soit de hautes performances avec des prix élevés, soit de basses performances avec un prix moindre. Du fait de l'hétérogénéité des ressources, il faut lui laisser la possibilité de louer des ressources sur mesure. Pour cela, il faut définir des métriques pour mieux évaluer et décider des performances des ressources à utiliser. Ces objectifs sont également des challenges à relever même si le déploiement d'un entrepôt sur le *Cloud* doit être totalement automatisé.

La construction de modèle de coûts est également un objectif important du fait que la construction d'entrepôts de données sur le *Cloud* ne porte pas seulement sur des aspects techniques, la dimension économique représente un point crucial. Dans les environnements de *Cloud*, les vitesses de communication (via LAN) peuvent être irrégulières selon la proximité des nœuds les uns des autres et l'architecture des réseaux. Ceci peut avoir un impact sur les

transferts de très grands volumes de données qui peuvent s'exprimer en tera-octets, voire en peta-octets. Ceci nécessite alors des techniques de compression de données.

Une partie des problèmes de recherche classiques qui se posent encore dans le domaine des entrepôts de données trouve une nouvelle expression lorsque l'on se situe dans le nuage. Faut-il continuer de dénormaliser les modèles physiques dans un cadre NoSQL pour bénéficier de meilleures performances, demeurer dans un environnement SQL qui garantit l'intégrité des données, ou encore tenter de travailler intégralement en mémoire vive ? L'élasticité est-elle la réponse à tous les problèmes de performance, ou ne vaut-il pas mieux adapter des techniques d'optimisation bien connues (index, vues matérialisées...) pour minimiser le coût en ressources (et donc, monétaire) des requêtes dans le nuage (Nguyen et al., 2012) ? Doit-on inclure dans la notion d'élasticité la prise en compte des données situationnelles (Pedersen, 2010) et les problèmes d'intégration des données qui en découlent ? De plus, travailler au moins en partie à partir de données situationnelles impose d'accepter une perte de contrôle sur les données du système décisionnel, notamment sur leur fiabilité et leur pérénité, et donc de se contenter d'analyses de tendances plutôt que d'historiques avérés (Middelfart, 2012). Evaluer ce degré de contrôle est donc important. L'étude de Kandel et al. (2012) constitue un point de départ tout à fait intéressant pour ces réflexions.

D'autres problèmes sont davantage liés au paradigme de l'informatique dans le nuage et aux usages décisionnels plus personnels et collaboratifs qu'il permet. Par exemple, classiquement, l'investissement (en général très important) dans un système décisionnel doit être effectué a priori par les entreprises. En revanche, dans le nuage, la construction d'un système décisionnel peut être incrémentale, collaborative et exploiter au mieux le paiement à la demande (Darmont et al., 2012). Il est tout à fait possible de "partir petit", voire de "rester petit", d'adaptant à la cible des utilisateurs. Nous abordons alors à présent ce volet utilisateurs si crucial pour des systèmes qui, par définition, sont centrés utilisateurs.

5 Des entrepôts pour tous : utilisateurs à tous les étages

L'informatique décisionnelle, en raison des architectures matérielles, logicielles et des compétences requises, n'a longtemps été accessible qu'aux grandes entreprises. Pourtant, les besoins en décisionnel existent dans de plus petites structures, que ce soient des PME (Petites et Moyennes Entreprises) ou TPE (Très Petites Entreprises), des ONG (Organisations Non Gouvernementales), des associations, des communautés en ligne ou même de simples citoyens (les indignés espagnols ont, par exemple, exprimé une forte demande de données publiques ouvertes). Pour ce type d'utilisateurs, des solutions bon marché, légères, faciles à utiliser, flexibles et rapides, sont nécessaires (Grabova et al., 2010). Avec l'avènement de l'informatique dans le nuage, le décisionnel à la demande (*cloud BI*, *personal BI*, *self-service BI*, *on-demand BI* ou encore *collaborative BI*, dans la terminologie anglo-saxonne encore non standardisée, avec *BI* pour *Business Intelligence*) sous forme de service est devenu possible et accessible avec un simple navigateur Web depuis une tablette ou un smartphone. Ce nouveau type de services en ligne doit permettre à des utilisateurs non-experts de prendre des décisions éclairées en enrichissant le processus décisionnel par des données situationnelles, c'est-à-dire très ciblées, de portée limitée dans le temps et pertinentes pour un petit groupe d'utilisateurs (Abello et al., 2013), soit typiquement des données glânées sur le Web.

A l'heure où l'informatisation tend à diminuer les relations inter-personnes, dans la mesure où beaucoup de ces relations se transforment en relations homme-machine, le besoin d'"humaniser" les systèmes se fait ressentir pour permettre le processus d'aide à la décision. Cette humanisation nécessite de rendre l'interaction système-utilisateur plus personnelle, afin d'assurer l'adaptation de l'informatique aux utilisateurs, avec pour objectif de répondre à leurs propres besoins. Ceci passe donc initialement par une conception de l'entrepôt de données où les utilisateurs finaux sont considérés.

5.1 Implication de l'utilisateur dans le processus décisionnel

L'un des points clés de l'entreposage de données réside dans la conception du schéma de l'entrepôt. En effet, les possibilités d'analyse sont conditionnées par ce dernier. Il est donc important que les utilisateurs soient impliqués dans la conception de l'entrepôt pour une bonne prise en compte de leurs besoins d'analyse.

Dans un second temps, pour permettre un processus décisionnel centré utilisateurs, la prise en compte de leurs préférences et de leurs caractéristiques à travers un profil constitue une piste intéressante. Dans l'exploitation des données, il s'agit alors de proposer la personnalisation du système (visualisation des données, par exemple), et la recommandation, par rapport à une aide à la navigation dans les données (Aligon et al., 2011), qui permet à terme une aide à la décision. En effet, par rapport au volume considérable de données, l'accès à une information pertinente devient un enjeu crucial pour l'utilisateur. Mais au-delà de cet aspect, il s'agit aussi pour l'utilisateur d'avoir l'impression que le système informatique ait été fait pour lui et qu'il s'adresse à lui "personnellement".

Par ailleurs, si les outils méthodologiques et technologiques permettant de mettre en œuvre des solutions décisionnelles à la demande existent depuis quelques années (entrepôts de données Web, de documents, de données XML, logiciels ETL et OLAP libres, systèmes de gestion de bases de données en mémoire vive... (Grabova et al., 2010)), le tout premier service a été le prototype Google Fusion Tables (Gonzalez et al., 2010). Ce dernier permet d'intégrer des données privées et situationnelles dans un tableur simple, de les visualiser, de les analyser et de les partager de façon très intuitive. Les applications en ligne de nombreux éditeurs de solutions décisionnelles proposent désormais également ce type de fonctionnalités. De plus, il a été proposé d'étendre le principe de fusion de tables à des cubes de données (*fusion cubes* dont le schéma et les instances peuvent être modifiés à la volée et qui intègrent des données situationnelles ainsi que les métadonnées décrivant leur provenance et leur qualité (Abello et al., 2013).

D'un point de vue technique, Essaidi (2010) a proposé une plateforme décisionnelle à la demande. Toutefois, si cette plateforme est bien disponible en tant que service dans le nuage (en mode SaaS : *Software as a Service*), l'intégration dynamique de données situationnelles n'est pas mentionnée. Thiele et Lehner (2011) proposent une solution à ce problème en combinant des données existantes chez l'utilisateur à des services Web qui créent de nouveaux contenus à partir de sources externes. Ainsi, le processus habituel d'ETL est conduit par l'utilisateur lui-même, de façon interactive. Toutefois, il n'y a aucune garantie quant à la qualité et à l'intégrité des données recueillies. Pour cela, il est toutefois possible d'utiliser les travaux de Jörg et Dessloch (2009), qui garantissent l'intégrité d'un entrepôt quand les données sources sont fournies avec une faible latence, comme c'est le cas pour des données situationnelles.

Enfin, l'aspect collaboratif du décisionnel est apparu dès 2007, avec l'annotation de cubes pour modéliser et permettre le partage de l'expertise des utilisateurs d'OLAP (Cabanac et al., 2007). Une architecture décisionnelle collaborative a ensuite été proposée par Berthold et al. (2010), qui inclut des fonctionnalités dites sociales afin d'enrichir le processus de décision grâce aux opinions d'experts. Une dernière approche répartit des magasins de données dans une architecture pair à pair (Golfarelli et al., 2012). Bien que le processus de décision soit amélioré dynamiquement grâce au partage de connaissances dans toutes ces approches, l'intégration de données situationnelles à la volée n'y est pas envisagée.

5.2 La visualisation pour aider l'utilisateur à décider

La phase d'analyse de données est bien évidemment cruciale par rapport à l'aide à la décision et au pilotage. Ainsi, la production de tableaux de bord et la visualisation interactive de l'information constituent des étapes phares, d'autant plus que l'exploration de données massives est un problème difficile, en particulier pour l'œil humain.

Nous pouvons distinguer deux types de travaux de recherche dans ce domaine : les travaux sur la visualisation elle-même et les besoins émergents par rappport aux nouveaux supports de communication.

Le premier porte sur l'amélioration de la visualisation par des algorithmes. Ainsi, on a vu se développer la combinaison de l'analyse en ligne avec des techniques de fouille de données (Messaoud et al., 2006). Et les chercheurs spécialisés en visualisation commencent à s'intéresser au domaine de l'OLAP. Citons en particulier la possibilité de navigation OLAP en 3D (Sureau et al., 2009).

Parallèlement, l'évolution technologique en matière de support modifie considérablement le rapport des utilisateurs à la visualisation de données. Il est nécessaire de considérer l'adaptation d'outils d'analyse aux nouveaux supports de diffusion. Selon K. Bornauw [5], "si nous parvenons à relever ce défi de la visualisation des données et à la rendre conviviale et accessible depuis n'importe quel appareil (ordinateur, smartphone, tablette,..), elle sera non seulement économiquement utile, mais également agréable à l'utilisateur de systèmes d'information décisionnels, qui se verra soulagé du fardeau des modèles complexes et douloureux d'exploration des données".

Si les outils traditionnels tels que Cognos ou BusinessObjects sont encore d'actualité, on a vu émerger de nouveaux outils comme Spotfire, QlikView, Tableau 7. Et l'usage de nouveaux supports a nécessité le développement d'applications spéciales par les terminaux mobiles (Roambi-ESX pour Ipad, Yellowfin, etc.). En effet, le déploiement des applications de Business Intelligence sur des terminaux mobiles complique la problématique de visualisation.

Dès l'avènement des premiers téléphones intelligents, la question de l'accès au système d'information de l'entreprise depuis tout lieu et à tout instant s'est posée. D'ici 2014, les accès internet seront majoritairement mobiles. Les applications mobiles transforment la communication et donc l'organisation même des entreprises, commente Benoit Herr [6], l'auteur de l'étude. L'essor des téléphones intelligents et autres tablettes modifient les usages de la *Business Intelligence* puisque l'accès à distance au système d'information depuis son terminal portable est devenu réel. Ces possibilités d'accès à distance de données à fort potentiel stratégique reposent

5. Kris Bornauw , BI Expert, EoZen, Groupe SQLI,- www.eozen.com, 2012.

6. Proginov, "Cloud, SaaS et mobilité : nouveaux outils, nouveaux usages". Mars 2012, Journal Solutions & Logiciels, N28.

bien évidemment la question de la sécurité, que nous nous proposons d'aborder dans la section suivante.

6 Sécurité

Chaque jour de nouvelles vulnérabilités sont découvertes sur tous les types de composants d'un système d'information classique, et aussi décisionnel a fortiori. Lorsqu'elles sont exploitées par des individus malveillants, elles risquent de perturber gravement le système d'information décisionnel : indisponibilité (partielle ou totale, temporaire ou prolongée), pertes de données, vol d'informations confidentielles, pertes d'exploitation, la liste n'est malheureusement pas exhaustive... La protection du système d'information décisionnel est une lutte incessante. Elle exige des administrateurs système et réseau en charge de la maintenance informatique de s'astreindre à :

– surveiller les menaces qui pèsent sur les systèmes d'information
– mettre en œuvre rapidement les parades permettant de réduire les possibilités d'attaque

Pour cela, il faut définir le périmètre de surveillance : systèmes d'exploitation ou applications, et ceci pour les équipements réseaux, serveurs, postes de travail, et périphériques. Nous constatons que la veille technologique, dans le domaine de la sécurité, concerne jusqu'à présent le suivi des nouvelles technologies disponibles sur le marché, mais concerne également le suivi des alertes de sécurité ou plus précisément des nouvelles vulnérabilités découvertes sur les systèmes informatiques.

Renforcer la sécurité des systèmes d'information décisionnelle consiste pour la plupart des acteurs à ajouter des équipements supplémentaires : serveurs, pare-feux... ou à complexifier et à sophistiquer la gestion des accès ... Nous sommes persuadés que la sécurité doit aussi être intégrée dans la phase de conception, dans les mécanismes d'architecture des entrepôts de données pour imposer des méthodes et des outils. Cette démarche permet de pallier à d'éventuelles défaillances des dispositifs mis en place au niveau des infrastructures et des systèmes de détection d'intrusions (IDS).

Les systèmes d'information décisionnels sont souvent stockés sur des machines virtuelles différentes pour des raisons de volumétrie et d'optimisation. La communication entre les différentes machines est très vulnérable. Cette faille doit être supprimée par des moyens de communication naturellement sécurisés. En considérant ces machines virtuelles comme des parties indépendantes, des primitives cryptographiques peuvent permettre de sécuriser les communications. Basée sur la cryptographie asymétrique, la signature numérique (parfois appelée signature électronique) est un mécanisme permettant de garantir l'intégrité d'un document électronique et d'en authentifier l'auteur, par analogie avec la signature manuscrite d'un document papier. Un mécanisme de signature numérique doit permettre au lecteur d'un document (une couche) d'identifier l'expéditeur (une couche) qui a apposé sa signature. Il doit garantir que le document n'a pas été altéré entre l'instant où l'auteur l'a signé et le moment où le lecteur le consulte. La confidentialité des données peut être assurée classiquement par des cryptosystèmes symétriques. Le problème qui se pose est le stockage de la clé privée. Il s'agit en effet de prémunir les systèmes contre des attaques visant à la recouvrer. Aucune solution ne permet de se prémunir totalement contre ce risque. Cependant, on assiste depuis quelques années à l'émergence de cryptosystèmes complètement homomorphiques. Ces cryptosystèmes permettent de faire des calculs sur des valeurs encryptées sans avoir à les décrypter. Ils peuvent

donc grandement limiter l'usage de la clé privée. Toutefois, ces cryptosystèmes nécessitent de grosses ressources et ne sont pas encore opérationnels en pratique.

Les questions relatives à la sécurité et à la confidentialité des données sur le *Cloud* ont été les premières préoccupations des fournisseurs et des usagers du *Cloud*. Il en est de même dans le cas des entrepôts de données dans le *Cloud*. Différents scénarios peuvent être envisagés : soit la soustraction des données sensibles de l'analyse à partir du *Cloud* ; soit l'encryptage de celles-ci. Des travaux commencent à émerger portant sur l'analyse des données encryptées. Ces questions représentent sans doute des pistes de recherche intéressantes. Cependant, elles ne sont pas les seules préoccupations, les nombreux problèmes cités ci-dessus montrent également la diversité des pistes de recherche que suscite cette nouvelle problématique des entrepôts dans le *Cloud*. Celle-ci souffre aujourd'hui d'un manque de conceptualisation du fait de son émergence récente.

Ainsi, les problèmes de sécurité intrinsèques au stockage de données dans le nuage demeurent : espionnage de la part du fournisseur de service ou d'un sous-traitant, garantie de disponibilité des données, croisements incontrôlés de données... (Chow et al., 2009). Il existe cependant des pistes de recherche prometteuses, notamment au niveau de l'anonymisation des données qui, même cryptées, restent interrogeables et utilisables dans certains traitements. Stocker des données volontairement altérées, mélangées dans le Cloud peut être aussi une possibilité pour assurer la confidentialité des données. Cette solution soulève également des questions au niveau du cryptage et décryptage pour les interrogations. De plus, le calcul multi-parties permet à des individus distincts de construire de façon collaborative un résultat d'analyse commun sans pour autant dévoiler leurs sources de données. Ces techniques de cryptographie ne sont toutefois pas encore assez matures pour permettre l'analyse en ligne ou la fouille de données, ni pour un déploiement à l'échelle du nuage. Ce dernier soulève des problèmes de temps de traitement qui pousse à ne sécuriser que certaines données : les plus sensibles, les plus récentes ...

7 Conclusion

Dans cet article, nous avons présenté le domaine de l'aide à la décision au travers du prisme des entrepôts de données et de l'analyse en ligne. Ainsi, l'aide à la décision apparaît ainsi dans ce domaine comme la proposition de méthodes et d'outils permettant aux décideurs de naviguer dans les données consolidées dédiées à l'analyse.

Après avoir présenté les concepts fondateurs de ce domaine, nous nous sommes penchés sur quatre aspects pouvant être considérés comme structurants par rapport à la recherche dans ce domaine, à savoir : les données, les environnements de stockage de ces données, les utilisateurs et la sécurité. Par ailleurs, ce travail a permis de synthétiser les problèmes ouverts de ce domaine, qui se posent dans un nouveau contexte économique et technologique. Ce contexte est fortement corrélé avec l'émergence du *Cloud*, des outils *Open Source* qui modifient en profondeur le rapport des utilisateurs aux données et à leur analyse, posant de réels problèmes de sécurité. L'aide à la décision du point de vue des entrepôts de données et de l'analyse est amenée à évoluer en fonction de ce nouveau contexte, assurant aux professionnels du domaine un développement d'activité croissant et, aussi, un avenir scientifique prometteur avec des verrous identifiables nombreux, comme nous avons pu le constater.

Références

Abello, A., J. Darmont, L. Etcheverry, M. Golfarelli, J.-N. Mazon, F. Naumann, T.-B. Pedersen, S. Rizzi, J. Trujillo, P. Vassiliadis, et G. Vossen (2013). Fusion cubes : Towards self-service business intelligence. *International Journal of Data Warehousing and Mining 9*(2).

Abouzeid, A., K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, et A. Rasin (2009). Hadoopdb : an architectural hybrid of mapreduce and dbms technologies for analytical workloads. *Proc. VLDB Endow. 2*(1), 922–933.

Agrawal, D., S. Das, et A. El Abbadi (2011). Big data and cloud computing : current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology*, EDBT/ICDT '11, New York, NY, USA, pp. 530–533. ACM.

Akkaoui, Z. E., E. Zimányi, J.-N. Mazón, et J. Trujillo (2011). A model-driven framework for etl process development. In *14th International Workshop on Data Warehousing and OLAP, Glasgow, United Kingdom (DOLAP 2011)*, pp. 45–52. ACM.

Aligon, J., M. Golfarelli, P. Marcel, S. Rizzi, et E. Turricchia (2011). Mining preferences from olap query logs for proactive personalization. In *15th International Conference on Advances in Databases and Information Systems, Vienna, Austria (ADBIS 2011)*, Volume 6909 of *Lecture Notes in Computer Science*, pp. 84–97. Springer.

Aouiche, K. et J. Darmont (2009). Data mining-based materialized view and index selection in data warehouses. *J. Intell. Inf. Syst. 33*(1), 65–93.

Bédard, Y. et J. Han (2009). *Geographic Data Mining and Knowledge Discovery*, Chapter Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. Taylor & Francis.

Benkrid, S. et L. Bellatreche (2011). Une démarche conjointe de fragmentation et de placement dans le cadre des entrepôts de données parallèles. *Technique et Science Informatiques 30*(8), 953–973.

Bentayeb, F., O. Boussaid, C. Favre, F. Ravat, et O. Teste (2009). Personnalisation dans les entrepôts de données : bilan et perspectives. In *5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2009), Montpellier*, Volume B-5 of *RNTI*, Toulouse, pp. 7–22. Cépaduès.

Berthold, H., P. Rösch, S. Zöller, F. Wortmann, A. Carenini, S. Campbell, P. Bisson, et F. Strohmaier (2010). An architecture for ad-hoc and collaborative business intelligence. In *Proceedings of the EDBT/ICDT Workshops*.

Cabanac, G., M. Chevalier, F. Ravat, et O. Teste (2007). An annotation management system for multidimensional databases. In *9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg, Germany*, Volume 4654 of *LNCS*, pp. 89–98. Springer.

Cao, L., J. Ni, et D. Luo (2006). Ontological engineering in data warehousing. In *8th Asia-Pacific Web Conference (APWeb 2006), Harbin, China*, Volume 3841 of *Lecture Notes in Computer Science*, pp. 923–929. Springer.

Cattell, R. (2011). Scalable sql and nosql data stores. *SIGMOD Rec. 39*(4), 12–27.

Chaudhuri, S. et U. Dayal (1997). An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec. 26*(1), 65–74.

Chaudhuri, S., U. Dayal, et V. Narasayya (2011). An overview of business intelligence technology. *Commun. ACM 54*(8), 88–98.

Chow, R., P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, et J. Molina (2009). Controlling data in the cloud : Outsourcing computation without outsourcing control. In *First ACM Cloud Computing Security Workshop (CCSW 2009), Chicago, IL, USA*, pp. 85–90.

Codd, E. (1993). Providing olap (on-line analytical processing) to user-analysts : an it mandate. Technical report, E.F. Codd and Associates.

Darmont, J., T.-B. Pedersen, et M. Middelfart (2012). Cloud intelligence : What is really new ? Panel.

Essaidi, M. (2010). ODBIS : towards a platform for on-demand business intelligence services. In *Proceedings of the EDBT/ICDT Workshops*, Lausanne, Switzerland.

Franco, J. M. (1997). *Le Data Warehouse, le Data Mining*. Eyrolles.

Golfarelli, M., F. Mandreoli, W. Penzo, S. Rizzi, et E. Turricchia (2012). OLAP query reformulation in peer-to-peer data warehousing. *Information Systems 5*(32), 393–411.

Gonzalez, H., A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, et J. Goldberg-Kidon (2010). Google fusion tables : web-centered data management and collaboration. In *2010 ACM International Conference on Management of Data (SIGMOD 2010), Indianapolis, USA*, pp. 1061–1066.

Grabova, O., J. Darmont, J.-H. Chauchat, et I. Zolotaryova (2010). Business intelligence for small and middle-sized enterprises. *SIGMOD Record 39*(2), 39–50.

Han, J., Y. Chen, G. Dong, J. Pei, B. Wah, J. Wang, et Y. Cai (2005). Stream Cube :An Architecture for Multidimensional analysis of Data Streams. *Distributed and Parallel Databases 18*, 173–187.

Inmon, W. H. (1996). *Building the Data Warehouse*. John Wiley & Sons.

Jin, X., J. Han, L. Cao, J. Luo, B. Ding, et C. X. Lin (2010). Visual cube and on-line analytical processing of images. In *19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada*, pp. 849–858.

Jörg, T. et S. Dessloch (2009). Near real-time data warehousing using state-of-the-art etl tools. In *Enabling Real-Time Business Intelligence – Third International Workshop (BIRTE 2009), Lyon, France*, Volume 41 of *LNBIP*, pp. 100–117. Springer.

Jovanovic, P., O. Romero, A. Simitsis, et A. Abelló (2012). Integrating etl processes from information requirements. In *14th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2012), Vienna, Austria*, Volume 7448 of *Lecture Notes in Computer Science*, pp. 65–80. Springer.

Kandel, S., A. Paepcke, J. M. Hellerstein, et J. Heer (2012). Enterprise data analysis and visualization : An interview study. *IEEE Trans. Vis. Comput. Graph. 18*(12), 2917–2926.

Kimball, R., L. Reeves, M. Ross, et W. Thornthwaite (2000). *Concevoir et déployer un data warehouse*. Eyrolles.

Leavitt, N. (2010). Will nosql databases live up to their promise ? *Computer 43*(2), 12–14.

Malinowski, E. et E. Zimányi (2004). Olap hierarchies : A conceptual perspective. In *16th International Conference on Advanced Information Systems Engineering (CAiSE 2004), Riga, Latvia*, Volume 3084 of *Lecture Notes in Computer Science*, pp. 477–491. Springer.

Mazón, J.-N., J. Lechtenbörger, et J. Trujillo (2009). A survey on summarizability issues in multidimensional modeling. *Data Knowl. Eng. 68*(12), 1452–1469.

Messaoud, R. B., O. Boussaid, et S. L. Rabaséda (2006). A multiple correspondence analysis to organize data cubes. In *Databases and Information Systems IV - Selected Papers from the Seventh International Baltic Conference, DB&IS 2006, July 3-6, 2006, Vilnius, Lithuania*, Volume 155 of *Frontiers in Artificial Intelligence and Applications*, pp. 133–146. IOS Press.

Middelfart, M. (2012). Analytic lessons : in the cloud, about the cloud. Indistrial keynote.

Muñoz, L., J.-N. Mazón, et J. Trujillo (2009). Automatic generation of etl processes from conceptual models. In *12th International Workshop on Data Warehousing and OLAP (DOLAP 2009), Hong Kong, China*, pp. 33–40. ACM.

Nguyen, T.-V.-A., L. d'Orazio, S. Bimonte, et J. Darmont (2012). Cost models for view materialization in the cloud. In *Workshop on Data Analytics in the Cloud (EDBT-ICDT/DanaC 12), Berlin, Germany*.

Papastefanatos, G., P. Vassiliadis, A. Simitsis, et Y. Vassiliou (2012). Metrics for the prediction of evolution impact in etl ecosystems : A case study. *J. Data Semantics 1*(2), 75–97.

Pedersen, T. B. (2010). Research challenges for cloud intelligence : invited talk. In *2010 EDBT/ICDT Workshops, Lausanne, Switzerland*.

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2008). A top keyword extraction method for olap document. In *International Conference on Data Warehousing and Knowledge Discovery (DAWAK 2008)*, Volume 5182, pp. 55–64. Springer Verlag, LNCS.

Selma, K., B. IlyèS, B. Ladjel, S. Eric, J. StéPhane, et B. Michael (2012). Ontology-based structured web data warehouses for sustainable interoperability : requirement modeling, design methodology and tool. *Comput. Ind. 63*(8), 799–812.

Simitsis, A., D. Skoutas, et M. Castellanos (2010). Representation of conceptual etl designs in natural language using semantic web technology. *Data Knowl. Eng. 69*(1), 96–115.

Stonebraker, M., D. Abadi, D. J. DeWitt, S. Madden, E. Paulson, A. Pavlo, et A. Rasin (2010). Mapreduce and parallel dbmss : friends or foes ? *Commun. ACM 53*(1), 64–71.

Sureau, F., F. Bouali, et G. Venturini (2009). Optimisation heuristique et génétique de visualisations 2d et 3d dans olap : premiers résultats. In *5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2009), Montpellier*, Volume B-5 of *RNTI*, Toulouse, pp. 65–78. Cépaduès.

Thiele, M. et W. Lehner (2011). Real-time BI and situational analysis. In *Business Intelligence Applications and the Web : Models, Systems and Technologies*, pp. 285–309. Hershey, PA : IGI Global.

Wang, H., J. Li, Z. He, et H. Gao (2005). OLAP for XML data. In *Proceedings of the* 1^{st} *International Conference on Computer and Information Technology (CIT2005)*, Shangai, China, pp. 233–237. IEEE Computer Society.

Entrepôts de données et aide à la décision

Summary

In this paper, we present the background regarding decisional processes in terms of data warehousing and OLAP. We present the main related concepts and the research challenges according to four points of view: data, storage, users and security.