International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 32 No. 3 Apr. 2021, pp. 381-400

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

Concevoir et déployer un système analytique des ressources humaines

[Design and deploy an analytical human resources system]

Mohamed Baslam, Youssef Fakir, and Bouzekri Moustaid

Computer Sciences Department, University of Science and and Technics, Sultan Moulay Slimane University, Beni-Mellal, Morocco

ABSTRACT: The purpose of this article is to present a methodological approach about the construction of a decision-making system for the implementation of a number of descriptive and predictive techniques using the query tools (SQL OLAP or MDX) and the process of Data Mining. The field of human resources (HR) is chosen as an application area. This field is currently an important subject in both the business and the scientific research world. Also, Human Resource analytics (HR analytics) is one of the most important emerging and disparate technologies of HR for the coming years. Ralph Kimball's approach is used to build this system with Microsoft technologies.

KEYWORDS: Data warehouse, multidimensional, data mining, Human Resource analytics, SQL Server, classification, SQL OLAP, MDX, Ralph Kimball.

RESUME: Cet article a pour objet de présenter une démarche méthodologique pour la construction d'un système décisionnel visant la mise en application d'un certain nombre de techniques à caractère descriptive, prédictive et prévisionnel faisant recours aux outils de requête (SQL OLAP ou MDX) et au processus du Data Mining. Le domaine des ressources humaines est pris comme domaine d'application, ce qui inscrit ce travail dans la discipline de l'analytique ressources humaines. En effet, on assiste actuellement à un essor de l'analytique ressources humaines aussi bien dans le monde des entreprises qu'universitaire; où il est l'objet d'enseignements et de recherches. Aussi, ce sujet constitue l'une des technologies prometteuses émergentes et disparates les plus importantes des technologies RH pour les années à venir. L'approche de Ralph Kimball est utilisée pour la construction de ce système avec les technologies de Microsoft.

MOTS-CLEFS: Entrepôts de données, magasins de données, multidimensionnel, data mining, data warehouse, data mining, analytique, ressources humaines, SQL Server, classification, SQL OLAP, MDX, Ralph Kimball.

1 INTRODUCTION

Les systèmes décisionnels ont connu et connaissent aujourd'hui un développement très important. Ils permettent de collecter, organiser et analyser les données d'une organisation pour aider la prise de décision. Ils manipulent de très importants volumes stockés dans un entrepôt de données. Pour analyser ces données de manière multidimensionnelle et interactive, les traitements OLAP (On line Analytical Processing) ont été définies à l'instar des traitements transactionnels OLTP (online transaction processing). L'OLAP offre la possibilité d'agréger, de visualiser, d'explorer des données à l'aide des opérateurs. L'ensemble de ce processus est désigné par le terme d'entreposage et comprend plusieurs ph ases telles que l'intégration, la structuration, la restitution et l'analyse en ligne des données. A l'heure actuelle les entrepôts de données et l'OLAP sont des technologies relativement bien maîtrisées quand il s'agit des données simples et bien structurées.

Corresponding Author: Bouzekri Moustaid

Cependant, les logiciels d'aide à la décision du marché ne répondent que partiellement aux attentes des décideurs lorsque ceux-ci souhaitent analyser les données sachant que la réactivité demandée aux entreprises est beaucoup plus importante et les temps des décisions devront être plus courts. Aussi avec l'avènement du web 2.0, le web social, le web 3.0 ou le web sémantique, de nouveaux territoires de veille et d'analyse de données ont apparu aux entreprises en passant d'un reporting descriptif à un reporting analytique et stratégique qui sera beaucoup plus marqué par le temps réel produisant ainsi plus de valeur en liant les données internes de l'entreprise avec les données ouvertes du web. De nouveaux problèmes ont été émergés au niveau de la modélisation, l'intégration des données dans l'entrepôt, l'analyse en ligne et la fouille de données. Par conséquent l'informatique décisionnelle demeure un thème de recherche pour lequel la communauté scientifique accorde plus d'importance.

Le domaine des ressources humaines est pris comme domaine d'application pour les raisons suivantes:

- Il constitue l'une des technologies prometteuses qui fait intervenir un grand nombre d'acteurs (décideurs, opérationnels et personnel et postes) et traite une masse importante de données de différentes sources (documents textuels de format libre, des messages, des bases de données, réseaux sociaux, ….)
- Il est considéré le parent pauvre du décisionnel par apport au domaine de gestion des clients sachant qu'un résonnement équivalent à celui mené sur le capital client peut été décliné aisément sur les ressources humaines de l'entreprise. Cette carence est accentuée au niveau des solutions analytiques basée sur l'intelligence d'affaires permettant l'amélioration de la prise de décision à travers la mise en application d'un certain nombre de techniques à caractère descriptive, prédictive et prévisionnel dont le processus du data mining occupe une place primordiale.

2 ETAT DE L'ART

2.1 LES SYSTÈMES DÉCISIONNELS

Les systèmes décisionnels sont destinés à recueillir des données opérationnelles de gestion des activités de l'entreprise et les présenter sous un format pour analyser et modifier le comportement de l'entreprise d'une façon intelligente.

Bien que ces systèmes datent des années soixante-dix avec l'apparition des systèmes d'aide à la décision sur mesure, le concept d'infocentre lancé par IBM et les systèmes d'interrogation des données qui sont développés tels que Focus, Datatrieve et Nomad , ils connaissent un développement important depuis les années quatre-vingt avec l'émergence du concept du data warehouse (DW) ou entrepôts de données (ED), induisant ainsi la notion d'une base de données unique pour centraliser les données. L'extraction et le croisement des données des différents systèmes opérationnels puis le chargement dans l'entrepôt de données, ont donné la naissance à des outils dédiés à ces tâches appelés les ETL (Extract, Transform, Load). Aussi, au tour du DW et les ETL, on retrouve au premier plan les outils de restitution c'est-à-dire les requêteurs, les outils de reporting et les outils de présentation (OLAP) (On Line Analysis Processing) ou de stockage multidimensionnel (Multidimensional OLAP) qui permettent l'analyse interactive de données de l'entrepôt de données. Au-delà de ces outils, des méthodes de data Mining se sont répandus permettant l'extraction de règles et des modèles à partir des données.

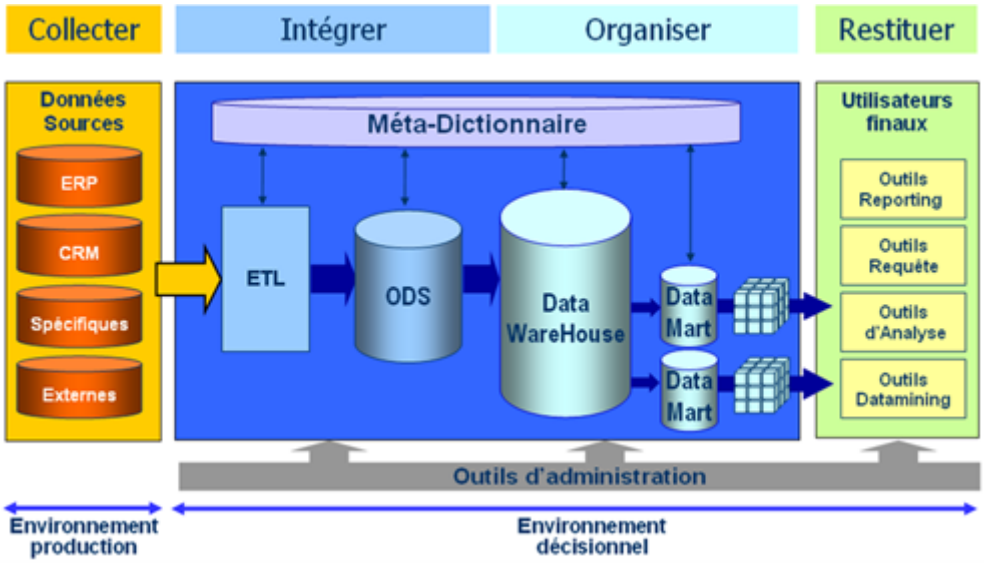

Fig. 1. Schéma d'entrepôt de données

Afin d'améliorer le requêtage et l'analyse de ces données entreposées à l'instar des traitements transactionnels des systèmes opérationnels OLTP, des processus de traitement analytique ont été développés dénommés OLAP (On-Line Analytical Processing »). Le terme OLAP a été inventé par E.F Codd, l'inventeur des bases de données relationnelles, dans un livre blanc à la demande de la compagnie Arbor Software en 1993, en introduisant un ensemble de règles que doivent assurer les ED pour supporter l'analyse OLAP. Il peut être défini comme 'le processus interactif de création, de gestion, d'analyse et de compte rendu sur les données. Ralph Kimball définit le concept OLAP comme "Activité globale de requêtage et de présentation de données textuelles et numériques contenues dans l'entrepôt de données; style d'interrogation et de présentation spécifiquement dimensionnel".

2.2 ENTREPÔT DE DONNÉES, MAGASINS DE DONNÉES ET CUBE DE DONNÉES:

Les entrepôts de données « data warehouse » constituent une solution adéquate pour construire un système d'aide à la décision. Plusieurs définitions ont été données pour le concept d'entrepôt de données. On retient les définitions des pères de l'entrepôt de données Ralph Kimball et W.H. Inmon. Ce dernier le définit comme "une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour supporter un processus d'aide à la décision"].

D'après cette définition, les données sont:

- Intégrées: Les données proviennent de différentes sources souvent structurées et codées de façon différente. L'intégration assure une représentation uniforme, cohérente et transparente. Cela permet de résoudre les problèmes d'hétérogénéité des systèmes de stockage, des modèles de données et de la sémantique de données.
- Orientées sujet: les données sont regroupées et organisées par sujet ou thème d'analyse, ce qui permet de rassembler toutes les informations utiles sur le sujet pour la prise de décision.
- Non volatiles: Les données d'un ED sont généralement utilisées en mode consultation. Elles peuvent être interrogées mais ne sont ni modifiées, ni supprimées (sauf dans les cas de rafraichissement de l'ED).
- Historisées: l'entrepôt garde une trace de l'historique des données. Les données d'un entrepôt sont identifiées par des périodes temporelles spécifiques.

De son coté, Ralph Kimball a fourni une définition plus simple d'un entrepôt de données : « un entrepôt de données est une copie des données transactionnelles d'une entreprise structurée de manière spécifique pour l'interrogation et l'analyse. »

Il en découle de ce qui précède qu'un entrepôt de données (ED) est une structure de données persistante, hébergeant les données consolidées qui sont extraites des diverses sources des systèmes d'information opérationnels (SIO) pour l'interrogation et l'analyse.

MAGASIN DE DONNÉES

Les données de l'entrepôt peuvent être réparties par classe de décideurs dans des espaces de stockages nommés magasins de données ou « data mart ». Ce dernier est un extrait de l'entrepôt de données qui contient tout ou partie de données de l'entrepôt selon le besoin des décideurs.

CUBES DE DONNÉES

Le décideur a besoin d'effectuer des requêtes agrégeant les données par rapport à plusieurs combinaisons de dimensions. Typiquement, ces requêtes mettent en jeu des fonctions agrégatives (ou statistiques) appliquées sur des mesures selon les diverses dimensions. Le concept de cube de données a été introduit par en vue de pré-calculer tous les agrégats en combinant un ensemble de dimensions pour répondre efficacement aux requêtes OLAP. Le cube ou l'hypercube de données est donc une structure multidimensionnelle présentée comme le résultat d'une requête combinant les Group-By selon toutes les combinaisons de dimensions. Les cellules du cube contiennent les données du sujet d'analyse et les arêtes du cube représentent les axes d'analyse. Le résultat de chaque Group-By est appelé un cuboïde. L'extrait des cubes de données à deux dimensions (ligne et colonne) est appelé une table multidimensionnelle (TM). Les cubes de données fournissent une aide non négligeable lorsqu'il s'agit d'interroger des entrepôts de données car ils sont exploités par différents outils d'analyse. Il est en particulier possible, pour le décideur, de naviguer dans les données grâce à la technologie OLAP en utilisant l'algèbre des cubes.

2.3 MODÉLISATION DES DONNÉES MULTIDIMENSIONNELLES

La modélisation multidimensionnelle est une approche de modélisation des données dédiée aux systèmes décisionnels. Elle permet de structurer les données sous une forme standardisée pour pouvoir analyser la performance de l'entreprise. Cette performance peut se matérialiser au travers d'un ensemble d'indicateurs mis en relation avec des dimensions d'analyse. Deux concepts ont émergé : le concept de fait et le concept de dimension. Un fait représente un sujet d'analyse, caractérisé par une ou plusieurs mesures, qui sont des indicateurs. Ce fait est analysé selon des axes d'observation. Les dimensions ou les axes sont composés d'attributs, appelés paramètres. Elles peuvent présenter des hiérarchies qui offrent la possibilité de réaliser des analyses à différents niveaux de granularité (niveaux de détail) pour restreindre ou accroître les niveaux de détail de l'analyse.

Il n'existe à ce jour aucun consensus sur la méthodologie de conception de l'entrepôt, comme le cas de la conception des bases de données relationnelles. Les concepteurs des entrepôts de données ou des magasins de données débutent souvent leurs travaux par la modélisation dimensionnelle relationnelle constituée des tables de faits et des tables de dimensions ou la modélisation physique par des cubes. Par exemple l'approche du Ralph Kimball commence la conception par le modèle dimensionnel correspondant au modèle logique de données pour une implémentation relationnelle.

Au niveau du modèle logique de données, le schéma de l'entrepôt de données peut être présenté sous différents schémas en étoile, en flocon ou en constellation. Si le modèle de données est constitué d'un fait et ses dimensions associées, alors le schéma s'appellera schéma en étoile (star schéma). Ce schéma est celui adopté par Ralph Kimball. Le schéma en flocon (snowflake) consiste à décomposer les dimensions du modèle en étoile en sous hiérarchies. La modélisation en flocon induit donc une normalisation des dimensions. Une généralisation possible du schéma en étoile ou en flocon est le schéma en constellation (fact constellation) qui est constitué de plusieurs faits et dimensions partagées entre les faits.

Aussi, à ce niveau plusieurs modèles sont utilisés pour implanter les schémas multidimensionnels :

- Modèle R-OLAP (Relational - On Line Analytical Processing): ce modèle est le plus courant. Il se base sur l'implantation des schémas multidimensionnels dans un environnement relationnel. Chaque fait et chaque dimension du modèle multidimensionnel conceptuel sont transformés en tables relationnelles.
- Modèle M-OLAP (Multidimensional - On Line Analytical Processing): ce modèle permet de stocker les données sous une forme nativement multidimensionnelle (dans des cubes de données, des matrices ou des vecteurs à n dimensions). Ce modèle offre des temps optimisés d'accès aux données et d'exécution des requêtes d'analyse. Par contre, ce modèle nécessite le recours à des systèmes de gestion des données multidimensionnelles.
- Modèle H-OLAP (Hybrid - On Line Analytical Processing): ce modèle réunit les avantages des deux modèles M-OLAP et R-OLAP. Il est utilisé surtout dans les outils commerciaux (Oracle Application Server, Microsoft Analysis Services). Il consiste à stocker les données détaillées dans des tables relationnelles (comme le modèle R-OLAP), tandis qu'il stocke les données agrégées sous une forme multidimensionnelle (comme le modèle MOLAP).

2.4 LANGAGE DE MANIPULATION DES DONNÉES MULTIDIMENSIONNELLES:

Dans ce domaine, la communauté scientifique a effectué de nombreux travaux. Une proposition importante est celle de l'opérateur cube qui propose différents opérateurs algébriques dans le contexte ROLAP et définit plusieurs opérateurs dans un langage algébrique et dans un langage graphique.

Dans le domaine commercial, de nombreux outils de manipulation ont été développés à savoir:

- Les outils spécifiques OLAP manipulent directement les concepts de l'approche multidimensionnelle (fait, dimension, mesure, …) et visualisent les données sous forme de tranches de cube de données sur lesquelles des opérateurs multidimensionnels peuvent s'appliquer.
- Les requêteurs graphiques très utilisés par les décideurs, permettent de manipuler et restituer les données sous forme de tableaux et de graphiques (Crystal Report, Discoverer d'Oracle, Explorer de Business Objects,...),
- Les SGBD relationnels (Oracle, MS SQL Server…), étendent le langage d'interrogation SQL en particulier l'opérateur GROUP BY en intégrant les commandes GROUPING SETS, ROLLUP et CUBE qui permettent l'expression de calculs de sous totaux en une seule requête. Ces options sont des instructions de SQL OLAP et sont incluses dans SQL3.
- Les serveurs OLAP utilisent le langage MDX doté d'une syntaxe de type SQL et s'applique sur des cubes de données. Le langage MultiDimensionnel eXpressions (MDX) est mis au point à la fin des années 90 par Microsoft pour SQL Serveur est actuellement implémenté par d'autres éditeurs au sein de leurs moteurs OLAP et tend à devenir un standard. MDX permet de naviguer dans les bases de données multidimensionnelles au moyen des requêtes contenant les objets (dimensions, hiérarchies, niveaux, membres et cellules) afin de présenter les résultats sous forme de tableaux.

MDX ressemble à SQL par ses mots clé SELECT, FROM, WHERE, mais: SQL construit des vues relationnelles et MDX construits des vues multidimensionnelles des données

Structure générale d'une requête MDX

Un prototype de requête MDX est donné par la syntaxe suivante :

SELECT [<axis_specification>

[, <spécification_des_axes>...]]

FROM [<spécification_d_un_cube>]

[WHERE [<spécification_de_filtres>]]

Il y'a deux axes, les colonnes (columns) et les rangées (row). On spécifie l'axe colonnes et les rangées avec les mots clés respectivement columns et row. La cellule est précédée par le mot clé Measures.

2.5 CYCLE DE VIE DE RALPH KIMBALL

L'approche globale de l'implémentation d'entrepôts de données par le cycle de vie est illustrée par la figure ci-dessous, ce schéma représente la succession des tâches nécessaires à la conception, au développement et au déploiement d'entrepôt.

Le cycle de vie de Ralph Kimball décrit la succession des tâches nécessaires à la conception, au développement et au déploiement d'entrepôts de données.

Le cycle de vie d'un logiciel représente toutes les étapes de son développement et de sa maintenance. Il se divise en trois grands axes: l'introduction, le développement et la production. Brièvement, le processus comprend trois axes principaux:

- L'axe technique: architecture technique et sélection des produits;
- L'axe des données: modélisation dimensionnelle et zone temporaire de traitement;
- L'axe d'analyse: spécification et réalisation d'applications d'analyse.

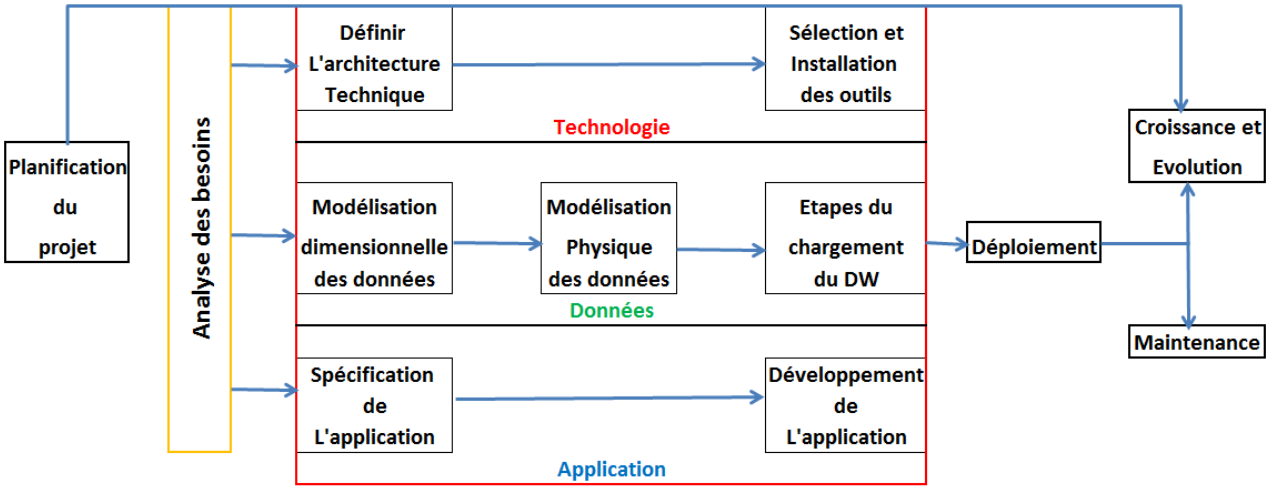

Fig. 2. Cycle de vie dimensionnel

NB: Le développement de l'étude de cas qui sera entamée dans la partie qui suit sera axé sur l'approche de Ralph Kimball en suivant les principales étapes de son cycle de vie.

2.6 DATA MING

Le data mining, ou fouille de données, ou forage de données constitue une composante fondamentale des systèmes décisionnels. Il permet de découvrir des connaissances cachées dans les données. Plusieurs définitions du data mining sont proposées dans , on retient quelques unes:

« L'extraction d'informations originales auparavant inconnues et potentiellement utiles à partir de données » (Frawley et Piateski-Shapiro)

« Un processus d'aide à la décision ou les utilisateurs cherchent des modèles d'interprétation dans les données » (Kamran Parsaye)

« L'exploration et l'analyse par des moyens automatiques ou semi automatiques d'un large volume de données afin de découvrir des tendances ou des régles » (Michael J. A. Berry)

En bref, le data mining est l'art d'extraire des informations (ou même des connaissances) à partir des données.

Les concepts de fouille de données et d'extraction de connaissances à partir de données sont parfois confondus et considérés comme synonymes. Mais, formellement on considère la fouille de données comme une étape centrale du processus d'extraction de connaissances des bases de données (ECBD) Knowledge Discovery in Databases ou KDD en anglais). Ce dernier est un processus de traitement composé de plusieurs étapes. Il débute par l'intégration de données brutes et se termine par l'interprétation des résultats obtenu.

La méthode standardisée CRISP-DM ' Cross-Industry Standard Process for Data Mining' découpe le processus de data mining en six phases principales :

- Connaissance du Métier
- Connaissance des Données
- Préparation des Données
- Modélisation
- Évaluation
- Déploiement

Les principales techniques de data mining et d'analyse des données se répartissent en deux grandes familles: les méthodes descriptives (exploratoires ou non supervisées) et les méthodes prédictives (explicatives ou supervisées).

Les méthodes prédictives ou explicatives visent à estimer la valeur d'une variable dite variable à expliquer, cible, réponse, dépendante ou endogène d'un individu ou d'un objet en fonction de la valeur d'un certain nombre d'autres variables du même individu, indiquées comme variables explicatives (dites encore variables indépendantes, de contrôle ou exogènes . On distingue deux grandes opérations: le classement (ou discrimination) et la prédiction (ou la régression). Elles se distinguent par la nature de la variable à expliquer: qualitative dans le cas du classement ou scoring, quantitative dans le cas de la prédiction.

Les techniques descriptives ou exploratoires permettent de mettre en évidence des informations présentes cachées par le volume de données c'est le cas des techniques de classification, règles d'associations, ….

La classification est l'une des méthodes du Data Mining la plus fréquemment utilisée. Elle est désignée par plusieurs noms tels que la classification automatique, apprentissage non supervisée (dans le domaine de la reconnaissance de forme), segmentation ou typologie ou analyse typologique (en Marketing), nosologie en médecine et partition en théorie de graphe. Les anglo-saxons parlent de clustering. Elle consiste à regrouper des objets (individus ou variables) ayant des caractéristiques similaires, en un nombre de groupes, de classes ou segments ou clusters.

Les méthodes de classification peuvent être divisées en trois grandes catégories: la classification par partition (partitionnement), la classification hiérarchique et la classification floue. Les méthodes de partionnement consistent en général à diviser les données d'un ensemble en k classes disjointes. Généralement le nombre de classes est fixé au départ. Un individu est affecté à la classe dont il est le plus proche. Les méthodes hiérarchiques regroupent deux à deux les éléments les plus proches de sorte à former de nouveaux éléments que l'on regroupe à leur tour. Dans la classification floue, les classes peuvent avoir plusieurs objets en commun (classes 'empiétantes' ou 'recouvrantes' et que chaque objet a une certaine probabilité d'appartenir à une classe donnée.

Les méthodes de classification sont fondées principalement sur la notion de distance ou celle de densité.

La fonction de distance D doit obéir aux règles suivantes [14]:

$D(A, B) > 0$

$D(A, A) = 0$ (Identité)

$D(A, B) = D(B, A)$ (Commutativité)

$D(A, B) <= D(A, C) + D(C, B)$ (Inégalité du triangle)

La notion de densité nécessite la définition du voisinage à l'aide d'une fonction de distance [14]. En effet, deux points sont dans un même voisinage s'ils sont à une distance inférieure à un seuil donné ξ et qu'un voisinage est dense s'il contient plus d'un nombre fixé K de points. Il en découle donc:

Voisinage (p, q) $\Leftrightarrow$ distance (p, q) <= ξ

Dense (p) $\Leftrightarrow$ Nombre (Voisinage (p)) >=K

Un individu est affecté à la classe dont il est le plus proche au sens d'une distance ou d'un indice de similarité.

Les techniques de classification font appel à une démarche algorithmique. Il existe plusieurs familles d'algorithmes de classification: les algorithmes conduisent directement à la production des partitions, les algorithmes ascendants ou agglomératifs et les algorithmes descendants ou divisifs.

Les principales méthodes de partitionnement sont les suivantes (Stéphane Tufféry 2017):

- Centres mobiles, K-means et nuées dynamique
- k-medois, k-modes, k-prototypes
- Réseaux de Kohonen
- Méthodes basées sur la notion de densité

ALGORITHMES DE DATA MINING DE MICROSOFT SQL SERVER

Microsoft fournit aussi plusieurs algorithmes pour les solutions d'exploration de données. Chacune étant adapté à un type de tâche et permet de créer un type de modèle. Ces algorithmes sont des implémentations de certaines méthodologies les plus utilisées dans l'exploration de données qui peuvent être personnalisés. Les types d'algorithmes d'Analysis Services par tâche sont les suivants :

Tableau 1. Types d'algorithmes d'Analysis Services

Tâches	Algorithmes Microsoft à utiliser
Prédiction d'un attribut discret.	Algorithme MDT (Microsoft Decision Trees) Algorithme MNB (Microsoft Naive Bayes) Algorithme de gestion de clusters Microsoft Algorithme MNN (Microsoft Neural Network)
Prédiction d'un attribut continu	Algorithme MDT (Microsoft Decision Trees) Algorithme MTS (Microsoft Time Series) Algorithme MLR (Microsoft Linear Regression)
Prédiction d'une séquence	Algorithme MSC (Microsoft Sequence Clustering)
Recherche de groupes d'éléments communs dans des transactions	Algorithme Microsoft Association Algorithme MDT (Microsoft Decision Trees)
Recherche de groupes d'éléments similaires	Algorithme de gestion de clusters Microsoft Algorithme MSC (Microsoft Sequence Clustering)

ALGORITHME DE GESTION DE CLUSTERS

Cet algorithme permet de créer des clusters et d'attribuer des points de données aux clusters. Il prend en charge plusieurs paramètres permettant ainsi sa personnalisation à plusieurs applications. Ces paramètres affectent le comportement, les performances et la précision du modèle d'exploration de données résultant. Ces principaux paramètres sont:

CLUSTERING_METHOD, CLUSTER_COUNT, CLUSTER_SEED, MINIMUM_SUPPORT, SAMPLE_SIZE, MAXIMUM_INPUT_ATTRIBUTES;

Elle fournit deux méthodes:

- K-means: Elle de type hard clustering. Un point de données peut appartenir à un seul cluster et qu'une probabilité unique est calculée pour l'appartenance de chaque point de données à ce cluster. Le clustering K-means repose sur la minimisation des différences entre les éléments d'un cluster et la maximisation de la distance entre les clusters. Elle se base sur le calcul des distances euclidiennes au carré entre les enregistrements de données dans un cluster et le vecteur qui représente la moyenne du cluster, puis converge vers un jeu final de k clusters lorsque cette somme atteint sa valeur minimale. L'algorithme K-means est généralement utilisé pour créer des clusters d'attributs continus.
- EM (Expectation Maximization): Elle est de type soft clustering. Un point de données appartient toujours à plusieurs clusters et qu'une probabilité est calculée pour chaque combinaison point de données/cluster. Cet algorithme affine de manière itérative un modèle de cluster initial pour l'adapter aux données. L'algorithme EM est l'algorithme par défaut utilisé dans les modèles de clustering Microsoft. L'implémentation Microsoft propose deux options: EM évolutif et EM non évolutif. Dans la méthode EM évolutif, les 50 000 premiers enregistrements sont utilisés pour entamer l'analyse initiale. Si l'opération réussit, le modèle se limite à ces données. Si l'adaptation du modèle échoue avec 50 000 enregistrements, 50 000 autres enregistrements sont lus. Dans la méthode EM non évolutif, le dataset est lu en sa totalité quelle que soit sa taille. Cette méthode peut créer des clusters plus précis, mais les besoins en mémoire peuvent être importants.

3 EXEMPLE D'APPLICATION: CAS DE L'ENTREPÔT DE DONNÉES DE GESTION DES RESSOURCES HUMAINES

On s'intéresse à appliquer notre approche au domaine des ressources humaines. Pour cela on a défini des besoins. On a choisi la démarche de Ralph Kimball pour la conception, le développement et le déploiement de l'entrepôt de données relatif à la gestion des ressources humaines. La solution technique choisie pour la mise en œuvre est celle de Microsoft car elle fournit toute la panoplie d'outils pour bâtir un système décisionnel dans les règles de l'art (Microsoft Visual Studio 2012) à savoir l'ETL, le serveur Olap et les outils de reporting et de data mining.

3.1 ANALYSE DES BESOINS

Les informations traitées sont extraites des données opérationnelles de la gestion quotidienne du personnel qui sont enregistrées au niveau d'une base de données de gestion de la paie et des carrières. Il y a donc tout intérêt à assurer la

cohérence et la continuité au sein de l'ensemble des composantes du système d'information des ressources humaines. Les besoins exprimés se présentent comme suit:

- Evaluer les composantes de la rémunération et de la masse salariale selon différentes dimensions (mois, affectation, fonction, ….).
- Identifier les caractéristiques du personnel concerné par le départ à la retraite dans les cinq prochaines années pour préparer un plan prévisionnel de recrutement.
- Catégorisation du personnel par nombre d'enfant et grade pour préparer des actions sociales à leurs profits (des aides, des colonies de vacances)

Ces questions peuvent être résolues par des requêtes SQL de type regroupement multiple ou des requêtes MDX ou celles de natures exploratoires qui font appel à des techniques du Data Mining qui offrent plusieurs algorithmes répondant à cette problématique.

3.2 MODÈLE LOGIQUE DES DONNÉES

La conception du schéma de l'entrepôt relatif à la gestion des ressources humaines selon l'approche étoile a produit le modèle logique de données sus-après composé de 6 tables:

-Table de fait:

SAL_RET : contient tous les mesures qu'on veut analyser.

-Cinq tables de dimension:

Table Personnel : représente le personnel

Table Grade : représente l'ensemble des grades du personnel

Table Affectation : représente les affections

Table Fonctions : représente les fonctions occupées par le personnel

Table Temps : contient les mois, les trimestres, les semestres et les années

3.3 SCHÉMA DE L'ENTREPÔT

Les différentes tables de dimension citées précédemment ainsi que la table de fait sont placées sur le schéma en étoile ci-dessous. La table de fait est au centre, entourée par les tables de dimension.

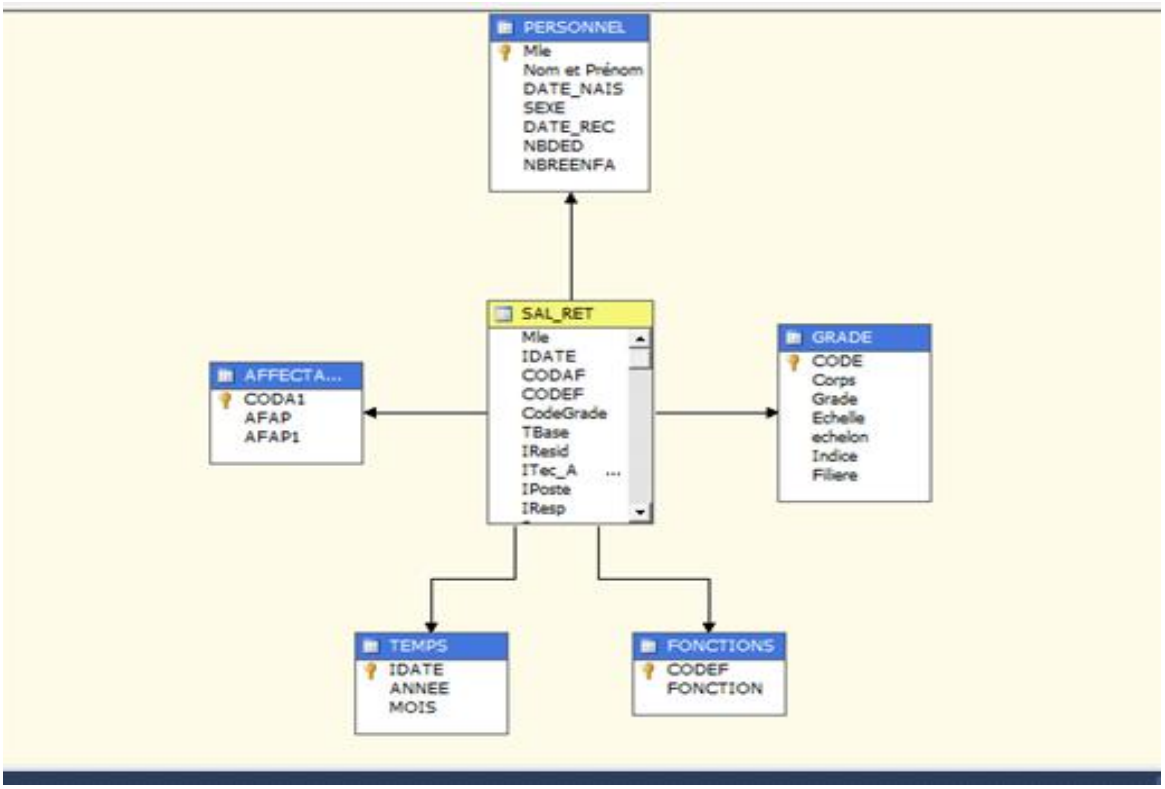

Fig. 3. Le schéma en étoile

3.4 ARCHITECTURE TECHNIQUE

Les outils utilisés sont consignés dans le tableau suivant:

Tableau 2. Outils de la solution

Outils	Fonctions
SQL Server	SGBD: permet l'implémentation de l'entrepôt de données et l'écriture de requêtes SQL OLAP
SQL Server Intégration Services (SSIS)	ETL (Extract, Transform, Load) ou ETC (Extraction /Transformation/Chargement)
SQL Server Analysis Services (SSAS)	-Création et gestion des structures multidimensuelles -Requêtes MDX -Exploration de données (Data Mining)
Excel	Préparation des données

3.5 ARCHITECTURE DU SYSTÈME

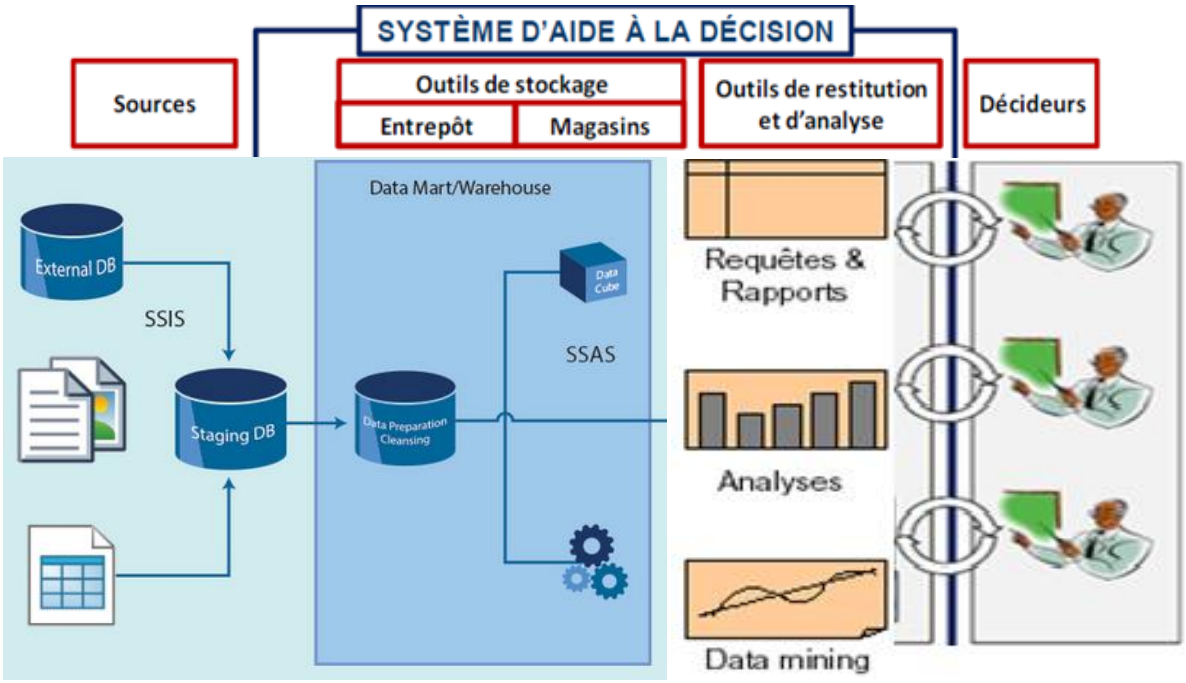

Fig. 4. Architecture du système

Le système fournit aux décideurs un ensemble d'outils informatiques constituant de véritables interfaces entre les utilisateurs (décideurs avec différents niveaux de responsabilité) et les sources des données pour simplifier l'accès aux données, de masquer l'hétérogénéité des sources et de faciliter l'interrogation des données aux décideurs.

Cette architecture est constituée des couches suivantes: la couche d'acquisition, la couche de stockage et la couche de restitution.

- La couche d'acquisition ou la couche ETL est assurée par l'outil IIS (intégration services)
- La couche de stockage est constituée d'un data warehouse ou data mart utilisant les serveurs SQL server et OLAP SSAS (Server Analysis Services) assurant des espaces de stockage utilisés pour la restitution et l'analyse de données multidimensionnelles.
- La couche de restitution permet de restituer aux décideurs les données contenues dans le data warehouse sous forme de rapport ou de tableau de bords ou des résultats d'exploration de data mining.

3.6 CHARGEMENT DU DATAWAREHOUSE

Les opérations effectuées lors des étapes de chargement du Data Warehouse sont:

- L'extraction des données depuis des fichiers Excel, crées pour récupérer les données à partir de l'application de gestion des ressources humaines.
- Préparation des données: Formatage des données extraites depuis des fichiers selon les structures des données du DataWarehouse, le nettoyage (remplissage des valeurs manquantes), et la conversion de types.
- Le chargement des données dans le Data Warehouse.

Pour l'Intégration Services, il faut définir:

- La tâche de flux de données (Data Flow Task)
- Sources de flux de données à déplacer: Excel.
- Transformation du flux de données: Sélection des transformations à appliquer aux données.
- Destination du flux de données pour le stockage: OLE DB.

3.7 DÉPLOIEMENT

Le déploiement de l'entrepôt de données peut être effectué par des commandes de requêtes à l'aide des langages SQL OLAP et MDX et avec des navigateurs.

LANGAGE SQL OLAP

Les évaluations des salaires par affectation et fonction en utilisant les opérateurs Group By Rollup, Group by cube et Grouping Set se font par les requêtes suivantes:

```
select  a.CODAF , f.CODEF, sum(brut) as salaires from FONCTIONS f, AFFECTATION a,SAL_RET s
where  f.CODEF=s.CODEF and a.CODAF=s.CODAF and a.CODAF like '12%'
group by rollup (a.CODAF , f.CODEF)
ORDER BY a.CODAF DESC, f.CODEF desc
```

```
select  a.CODAF , f.CODEF, sum(brut) as salaires from FONCTIONS f, AFFECTATION a,SAL_RET s
where  f.CODEF=s.CODEF and a.CODAF=s.CODAF and a.CODAF like '12%'
group by  GROUPING SETS ((a.CODAF) ,( f.CODEF))
ORDER BY a.CODAF DESC, f.CODEF desc
```

```
select  a.CODAF , f.CODEF, sum(brut) as salaires from FONCTIONS f, AFFECTATION a,SAL_RET s
where  f.CODEF=s.CODEF and a.CODAF=s.CODAF and a.CODAF like '12%'
group by cube (a.CODAF , f.CODEF)
ORDER BY a.CODAF DESC, f.CODEF desc
```

REQUÊTE MDX

Le langage MDX s'applique sur les cubes. Ces structures doivent donc être créées en deux phases:

- Spécification des sources de données qui alimentent le cube: elle consiste à créer une source de données et la vue de source de données (DSV).
- Créer le cube en indiquant les mesures et les dimensions.

Les composantes ainsi créées sont:

- Source de données: Masse salaire.ds
- Vue de sources de données: Masse salaire.dsv
- Cube: Masse salaire.cube

Pour afficher les données du cube dans le projet, il est nécessaire de déployer le projet au sein d'une instance d'Analysis Services, puis de traiter le cube et ses dimensions. Le déploiement entraîne la création des objets définis dans (au sein de) l'instance et le traitement des objets dans une instance entraîne la copie des données à partir des sources de données sous-jacentes dans les objets du cube et le calcul des agrégats dans la structure déployée.

Le schéma suivant récapitule l'ensemble des actions du déploiement qui sont réalisées avec succès:

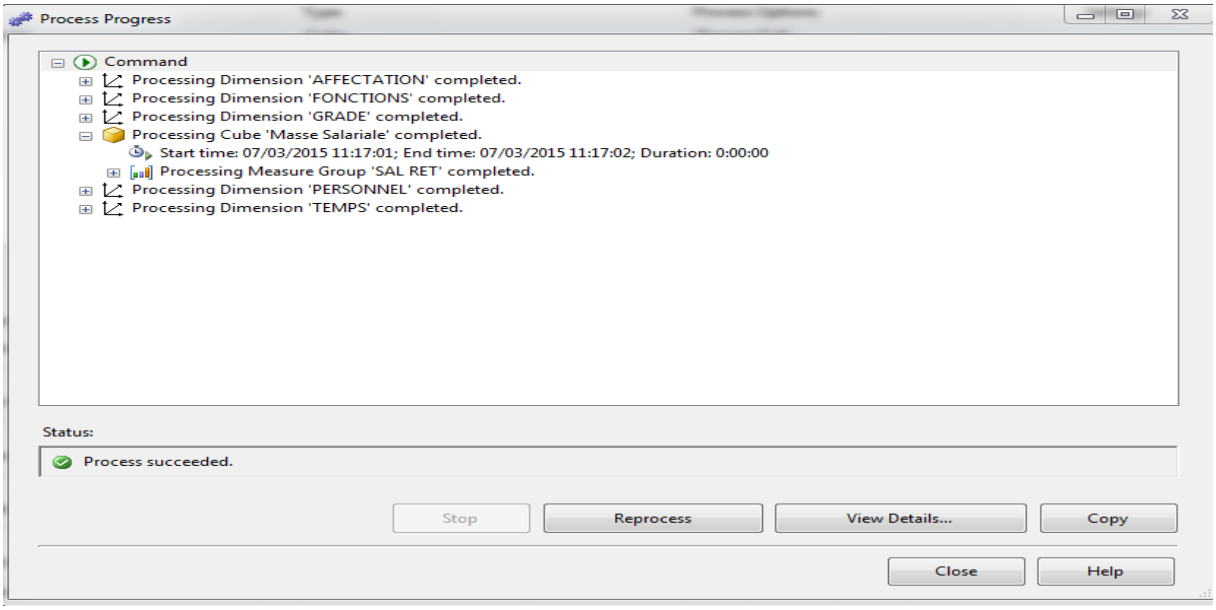

Fig. 5. Actions de déploiement

SCHÉMA DU CUBE CRÉE

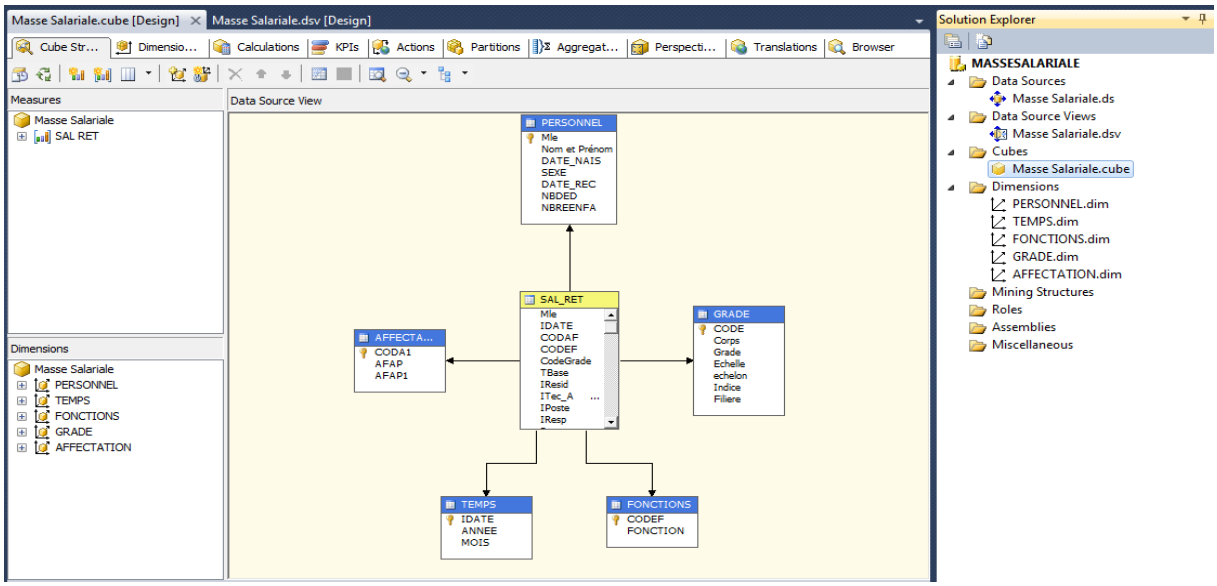

Fig. 6. Schéma du cube

Une fois le cube est déployé, on procède à l'écriture des requêtes MDX qui permettent le calcul des composantes des allocations famillliales, le salaire brut par rapport aux dimensions codeaf, codegr et codef:

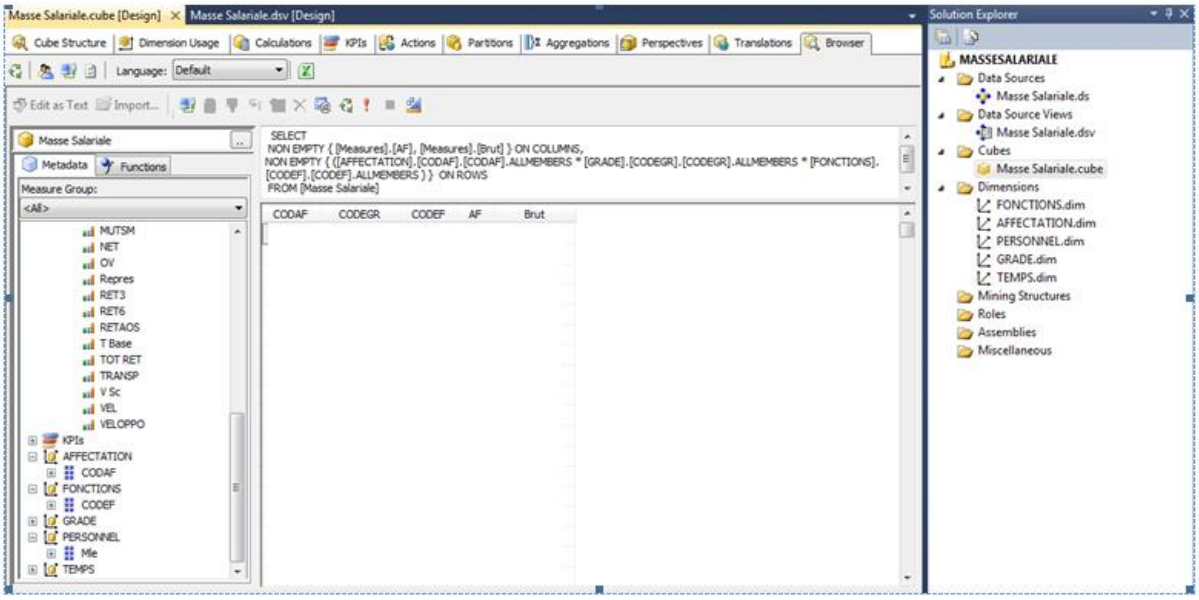

Fig. 7. Editeur de requête MDX

NAVIGATEUR

Il est donc possible de naviguer dans le cube grâce à l'onglet Navigateur. Cet onglet permet d'afficher les données du cube selon une présentation similaire à celle d'un tableau croisé dynamique. Grâce à un glisser-déplacer, les mesures prennent place dans l'espace central d'un tableau croisé dynamique. Les attributs de dimension se retrouvent soit en ligne, soit en colonne ou soit en filtre.

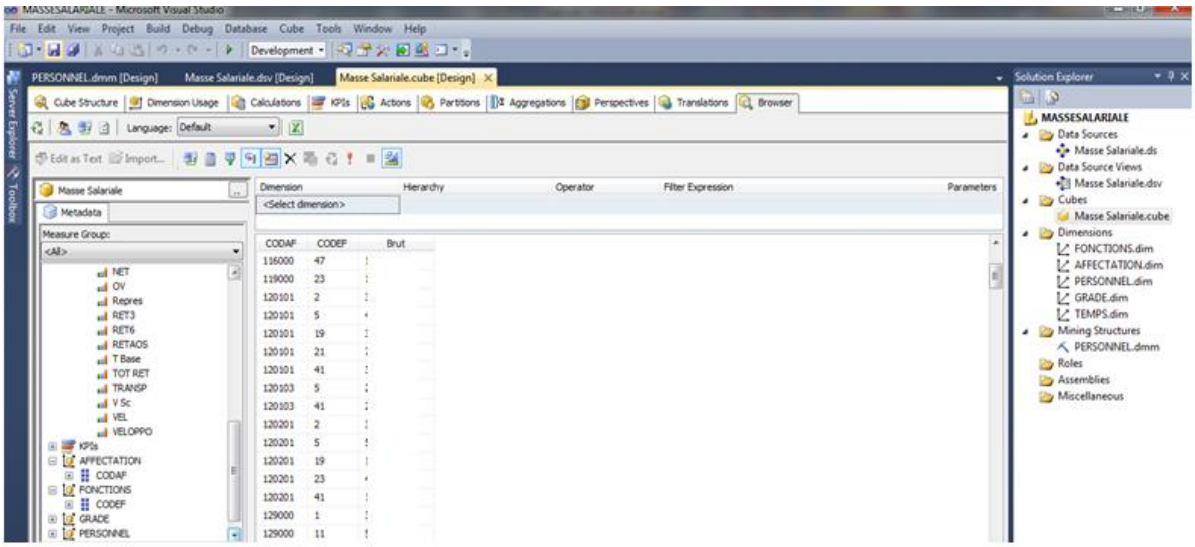

Fig. 8. Navigateur de cube

3.8 CRÉATION DU MODÈLE DE DATA MINING

La création d'un modèle de Data Mining comprend les étapes suivantes:

- Définition de la source de données c'est-à-dire la connexion qui identifie le serveur et la base de données ou le fichier dans lequel résident les données à analyser.
- Spécification de la vue de la source de données: c'est la source de données opérationnelle (Table, vue, fichier, …)

- Création du modèle: un clic droit sur Structures d'exploration de données permet de choisir l'algorithme qui convient.

Pour notre cas, L'analyse des critères qui caractérisent la population des employés concernés par le départ à la retraite dans les cinq prochaines années est un problème de classification pour lequel les algorithmes Naïve Bayes, Decision Tree et Clusters sont particulièrement adaptés. L'algorithme choisi est celui de Clustering qui permet de regrouper les employés possédant des caractéristiques similaires.

LES COMPOSANTES AINSI CRÉÉES

- Source de données: Masse salaire.ds
- Vue de sources de données: Masse salaire.dsv
- VueComplAnalyse1.dmm

La vue de source est basée sur une requête de jointure de la table de fait centrale avec les tables dimensionnelles nécessaires à l'exploration des données, l'âge de l'employé est déterminé par la fonction datediff (yy, date_naissance, getdate ()).

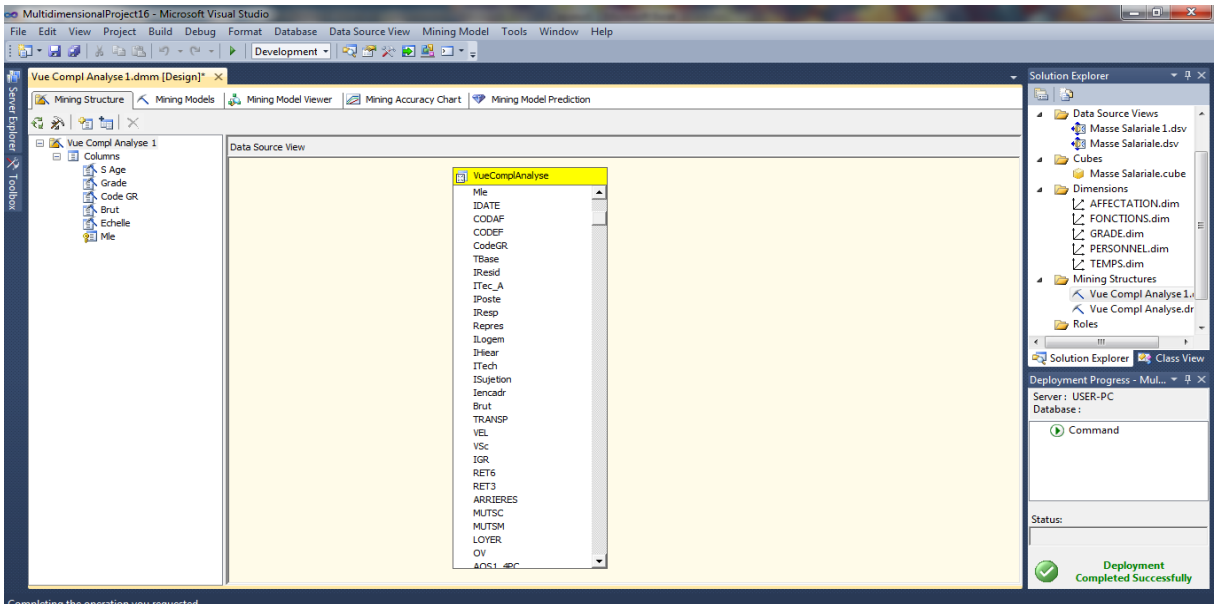

Fig. 9. Vue de la source

ETAPES DE CRÉATION DU MODÈLE DE DM

- Spécification des types de tables contenant les données à analyser: La vue VueComplAnalyse1

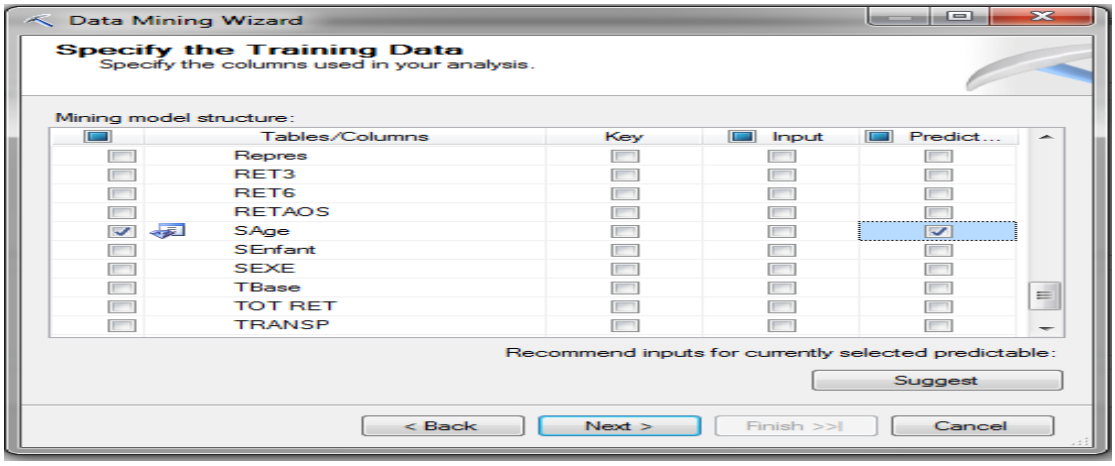

Fig. 10. Choix des données du modèle

- Spécification de la clé dans la source et du champ à prévoir (SAge) et les colonnes en entrées. En ce qui concerne les colonnes en entrée, on demande à l'assistant de suggérer les champs les plus susceptibles d'entrer dans le processus prédictif. Le bouton Suggérer permet de lister ces champs.

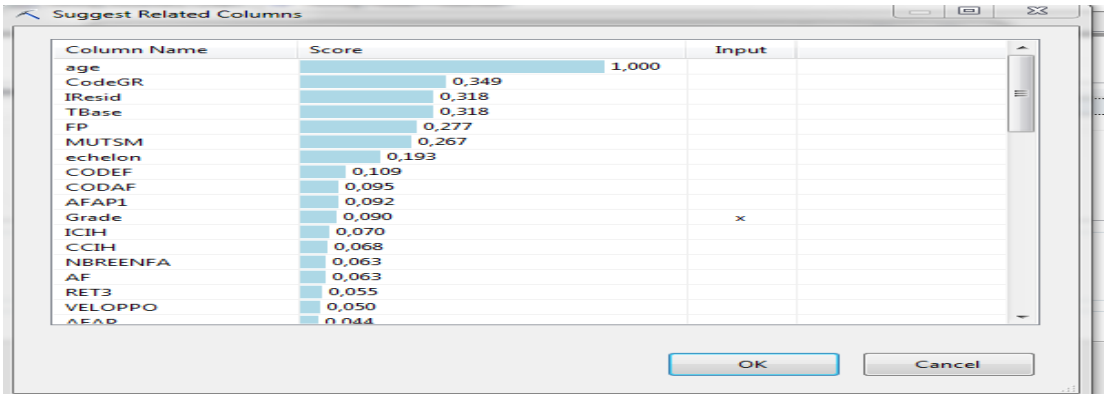

Fig. 11. Bouton de suggestion des colonnes d'entrées

- Choix des colonnes en entrée:

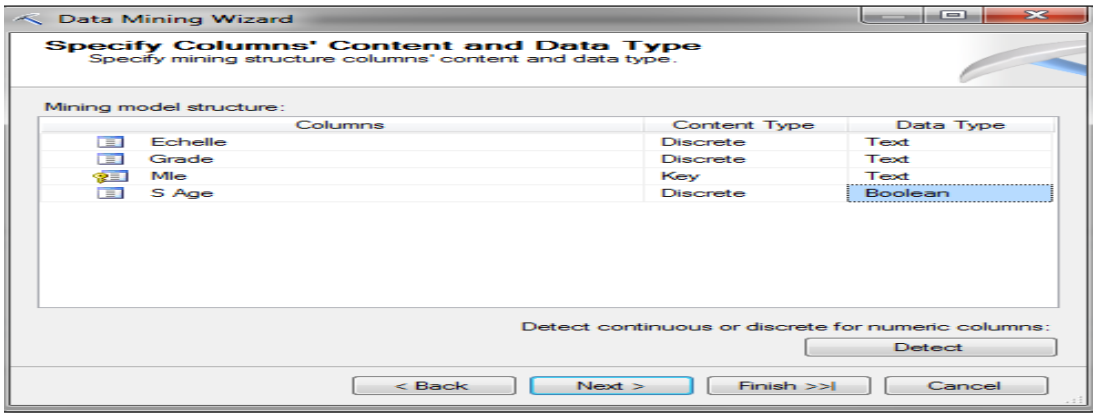

Fig. 12. Les colonnes d'entrées

- Déploiement du modèle: Une fois, les structures sont crées, on procéde au déploiement

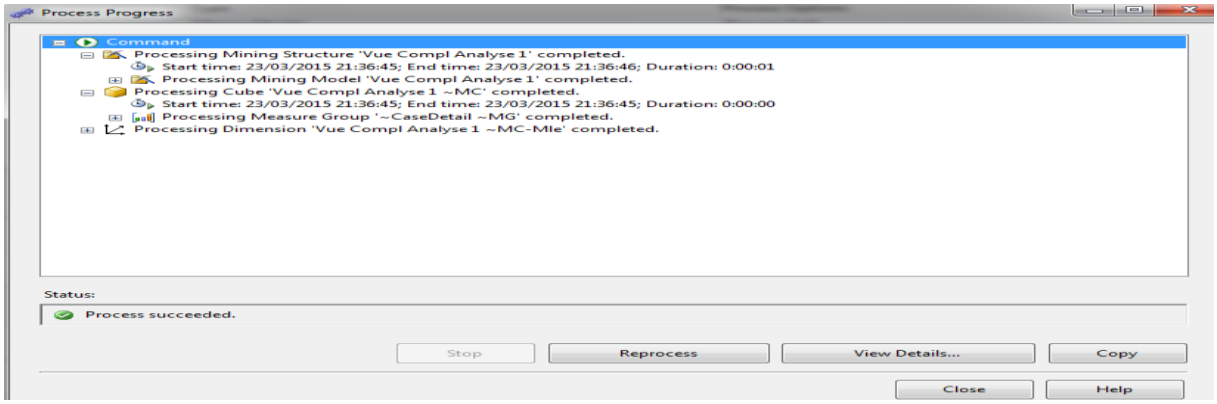

Fig. 13. Processus de déploiement

NAVIGUER DANS LE MODÈLE CLUSTERS

- Le diagramme *cluster* permet d'établir des relations entre des groupes homogènes.
- Les lignes qui relient les clusters sont plus denses si les liens entre clusters sont étroits.
- Le curseur à gauche du diagramme permet d'appliquer un filtre afin de cacher les liens les moins forts.
- Dans le diagramme ci-après, le cluster 1 contient l'effectif le plus grand du personnel concerné par le départ voir le profit des clusters. Un lien entre les clusters 5 et 3 apparaît comme très étroit. Les deux groupes sont formés par le personnel appartenant principalement à l'échelle 9. La comparaison entre les deux clusters est possible par l'onglet Cluster Discrimination.

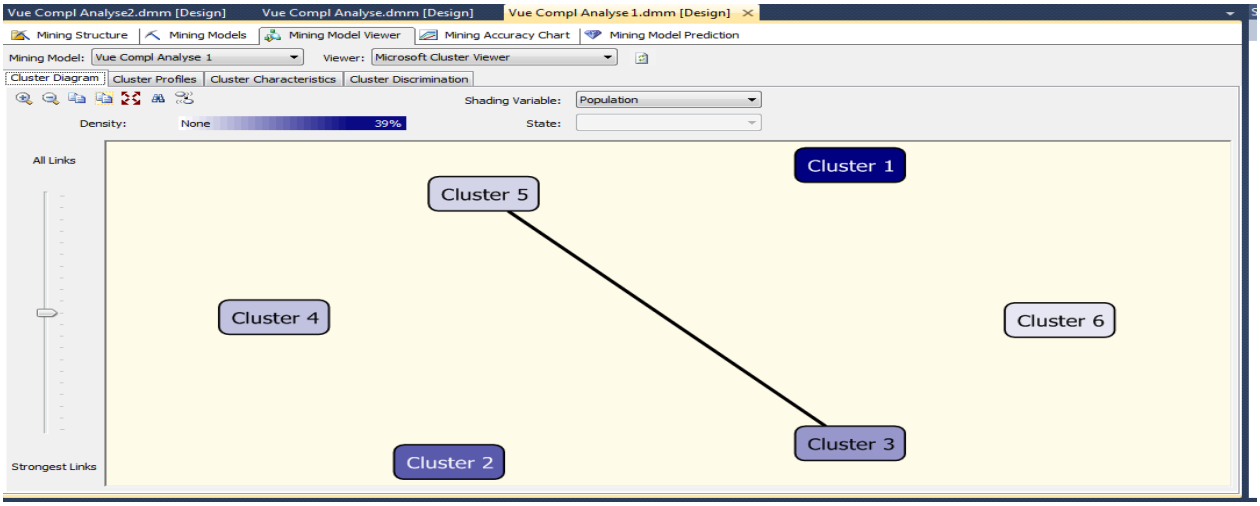

Fig. 14. Diagramme de cluster

Les caractéristiques des différents clusters est visualisées par l'onglet cluster profiles:

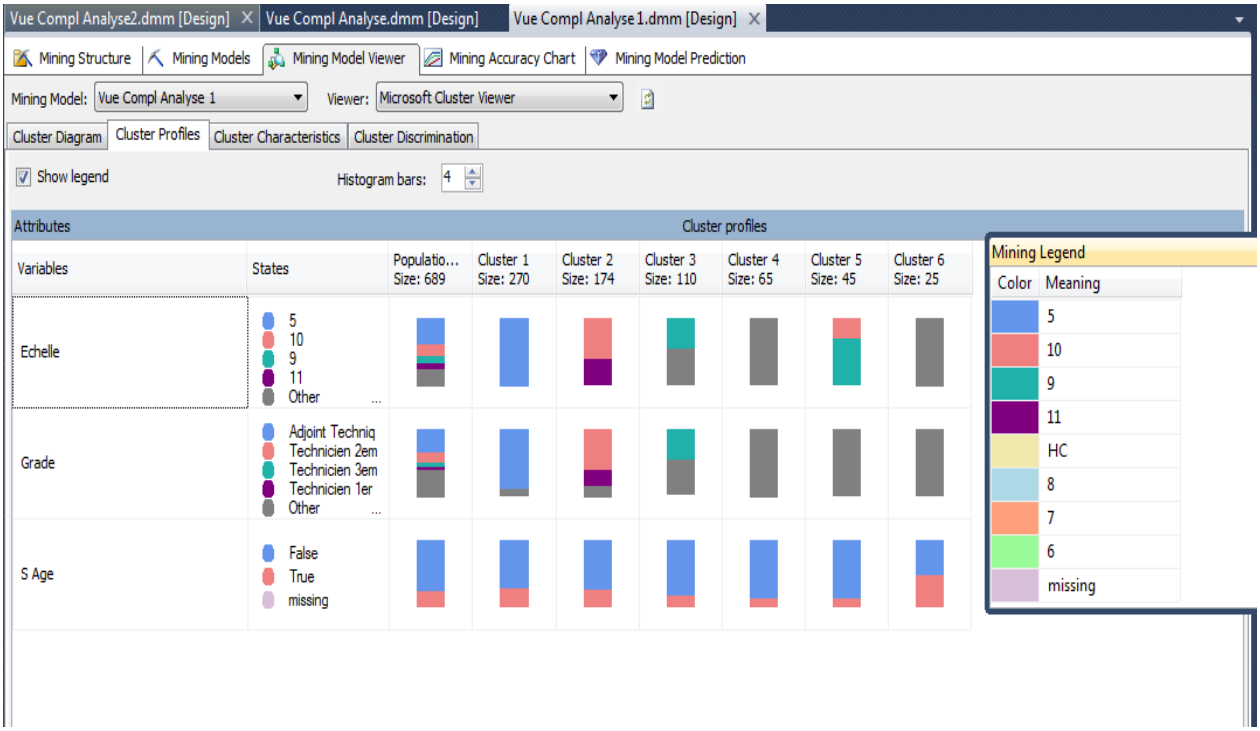

Fig. 15. Caractéristiques des clusters

La répartition des effectifs concernés par le départ à la retraite en fonction des attributs en indiquant les supports et les probabilités est fournie par l'outil Microsoft Generic Content Tree Viewer:

MODEL_CATALOG	MultidimensionalProject25			
MODEL_NAME	Vue Compl Analyse 1			
NODE_CAPTION	Modèle de cluster			
NODE_DESCRIPTION	Tout			
NODE_RULE				
	ATTRIBUTE_NAME	ATTRIBUTE_VALUE	SUPPORT	PROBABILITY
	Echelle	Manquant	0	0,00
	Echelle	10	121	0,18
	Echelle	11	67	0,10
	Echelle	5	272	0,40
	Echelle	6	25	0,04
	Echelle	7	28	0,04
	Echelle	8	35	0,05
	Echelle	9	82	0,12
	Echelle	HC	59	0,08
	S Age	Manquant	0	0,00
	S Age	False	515	0,75
	S Age	True	174	0,25
	Grade	Manquant	0	0,00
	Grade	Adjoint administratif 1er grade	9	0,01
	Grade	Adjoint administratif 2eme grade	22	0,03
	Grade	Adjoint administratif 3eme grade	11	0,02
NODE_DISTRIBUTION	Grade	Adjoint administratif 4eme grade	31	0,05
	Grade	Adjoint Technique de 2eme grade	6	0,01
	Grade	Adjoint Technique de 3eme grade	14	0,02
	Grade	Adjoint Technique de 4eme grade	241	0,35
	Grade	Administrateur 1er grade	12	0,02
	Grade	Administrateur 2eme grade	12	0,02
	Grade	Administrateur 3eme grade	14	0,02
	Grade	Ingenieur d'Etat 1.Grade	5	0,01
	Grade	Ingenieur d'Etat G.P	6	0,01
	Grade	Ingenieur en Chef 1er Gr.	15	0,02
	Grade	Ingenieur en Chef Gr.Pr.	31	0,05
	Grade	Redacteur 2eme grade	2	0,00
	Grade	Redacteur 3eme grade	31	0,05
	Grade	Redacteur 4eme grade	14	0,02
	Grade	Technicien 1er grade	43	0,06
	Grade	Technicien 2eme grade	105	0,15
	Grade	Technicien 3eme grade	50	0,07
	Grade	Technicien 4eme grade	12	0,02
NODE_SUPPORT	689			
MSOLAP_NODE_SCORE	0,662518313			

Fig. 16. Résultats dans Generic Content Tree Viewer

NB:

- Avec l'outil Mining Model Prédiction on aura les résultats de modélisation en affectant à chaque agent son cluster, ce qui permet de préparer un plan de recrutement pour remplacer le personnel concerné et d'évaluer l'impact sur la masse salariale.
- Excel est l'outil privilégié des « data scientist ». Microsoft a introduit au niveau d'excel les outils nécessaires à la mise en œuvre d'un projet de data mining. Une macro complémentaire nommé add-in de data mining est intégrable dans excel. Elle est téléchargeable à partir du site internet de Microsoft. Elle est installée et testée.

4 CONCLUSION

Bien longtemps, l'informatique décisionnelle a été réservée aux grandes organisations qui étaient les seules à tirer parti d'investissements lourds aussi bien en termes d'équipes de projet qu'en termes d'infrastructures matérielles et logicielles pour le reporting financiers et d'analyse marketing. Dès le début des années quatre-vingt-dix avec l'apparition du nouveau type d'organisation des données appelées hypercubes OLAP, plusieurs progrès ont été enregistrés en matière de restitution de l'information, des interfaces au profit des managers pour accéder à leurs données et le développement des applications analytiques proposant à l'entreprise un schéma analytique standard. Aussi, les éditeurs de SGBD ont fait évoluer leurs moteurs par greffage d'outils pour supporter le décisionnel. Ainsi, Microsoft introduisait depuis SQL Server 2000, le composant décisionnel appelé Analysis Services qui peut être le germe de business intelligence et depuis lors les nouvelles versions de SQL Server ont été enrichies d'outils à tous les niveaux de la chaine de fabrication des systèmes décisionnels. C'est ainsi que s'inscrit ce présent article qui a présenté une vision méthodologique de construction et de déploiement d'un datawarehouse et un état de l'art des outils de la solution décisionnelle de Microsoft tout en démontrant leurs applicabilités à la gestion des ressources humaines qui peut être considérée le parent pauvre du décisionnel à ce jour. Aussi, un résonnement équivalent à celui mené

sur le capital client a été décliné aisément sur les ressources humaines constituant ainsi un mode opératoire permettant de réaliser le déploiement des fonctions de business intelligence au service du capital humain de l'entreprise.

REFERENCES

[1] Burquier, B. (2007). BUSINESS INTELLIGENCE AVEC SQL SERVER 2005 Mise en œuvre d'un projet décisionnel Dunod. Dunod.

[2] Chris Date. (2000). Introduction aux bases de données 7 éme édition Traduction de Martine Chalmond, Nora et Frédiric Cuppens. Addison-Wesley Longman, Inc.

[3] Emmanuel Ferragu. (2013). Modélisation des systèmes d'information décisionnels Techniques de modélisation conceptuelle et relationnelle des entrepôts de données. Vuibert.

[4] ESPINASSE, B. (2015). Entrepôts de données: Introduction au langage MDX.

[5] FANTINI, S. (2010). Business Intelligence avec SQL Server 2008 R2. Editions ENI.

[6] Frada Burstein, C. W. (2008). Handbook on Decision Support Systems 1: Basic Themes (International Handbooks on Information Systems). Springer; 2008th edition (January 11, 2008).

[7] Gardarin, G. (1999). Internet / Intranet et bases de données Data Web, Data Media, Data Warehouse, Data Mining. Eyrolles.

[8] Gille, M. G. (2010). SIRH Système d'information des ressources humaines. Dunod.

[9] GODIN, R. (2012). Systèmes de gestion de bases de données par l'exemple. Loze Dion.

[10] Jac Fitz-Enz, J. R. (2014). Wiley Predictive Analytics for Human Resources. Wiley.

[11] Jim Gray, A. B. (1996). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. 12th International Conference on Data Engineering (ICDE'96) IEEE.

[12] Lignerolles, J.-M. F. (2001). Piloter l'entreprise grâce au data warehouse. Eyrolles.

[13] Ludovic Merville. (2018). Migration de la chaîne décisionnelle du calcul des taux d'usure et proposition d'une méthodologie de migration du logiciel SAS vers R.

[14] Microsoft. (2018). *Data Mining Algorithms (Analyses Services Data Ming).* Récupéré sur Data Mining Algorithms: https://docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=asallproducts-allversions.

[15] Peña-Ayala, A. (2014). Educational Data Mining: Applications and Trends (Studies in Computational Intelligence (524)). Springer; 2014th edition.

[16] Ralph Kimball, L. R. (2000). Concevoir et déployer un data warehouse Guide de conduite de projet. Eyrolles.

[17] Ralph Kimball, M. R. (2002). Entrepôts de données Guide pratique de Modélisation dimensionnelle Deuxième édition Traduction de Claude Raimond 2003. Vuibert.

[18] René Lefébure, G. V. (2001). Data mining Gestion de la Relation Client Personnalisation de sites web. Eyrolles.

[19] Salles, M. (2015). Décision et système d'information Systèmes d'information avancés - Volume 2. ISTE Group.

[20] Tufféry, S. (2005). Data Mining et statistique décisionnelle: l'intelligence dans les bases de données. Editions Technip.

[21] Tufféry, S. (2017). Data mining et statistique décisionnelles: l'intelligence des données. Technip.

[22] ZhaoHui Tang, J. M. (2005). *Data Mining with SQL Server 2005.* John Wiley & Sons.