MASTER EN INGÉNIEUR DE GESTION À FINALITÉ SPÉCIALISÉE EN ANALYTICS & DIGITAL BUSINESS

Le modèle Data Vault 2.0 est-il le mieux adapté lors du déploiement d'un Data Warehouse?

Mellaerts, Amaury

Award date:
2020

Awarding institution:
Universite de Namur

Le modèle Data Vault 2.0 est- il le mieux adapté lors du déploiement d'un Data Warehouse ?

Amaury MELLAERTS

Directeur: Prof. S. Faulkner

Mémoire présenté
en vue de l'obtention du titre de
Master 120 en ingénieur de gestion, à finalité spécialisée
en Analytics & Digital Business

ANNEE ACADEMIQUE 2019-2020

Université de Namur, ASBL
Faculté des Sciences économiques, sociales et de gestion – Département des Sciences de gestion

Rempart de la Vierge 8, B-5000 Namur, Belgique, Tel. +32 [0]81 72 48 41/49 58, Fax +32 [0]81 72 48 40

AVANT-PROPOS

Je remercie toutes les personnes, de près ou de loin, qui m'ont apporté de l'aide, durant à la réalisation de ce mémoire.

Je tiens aussi à remercier tout particulièrement le directeur de ce mémoire, M. Stéphane FAULKNER pour sa disponibilité, sa confiance, son suivi et sa relecture qui m'ont permis d'affiner ce travail.

Je remercie la société AkaBI, et notamment son CEO M. Jonathan DESMET pour son aide et sa coopération. Et également Messieurs Badreddine BEN AISSA et Lucas VANOVERBERGHE pour leurs disponibilités et leurs expertises lors de leurs interviews, qui ont été indispensable quant à la réussite de ce mémoire.

Pour terminer, je témoigne de toute ma gratitude bien évidemment à mon papa, Frédéric MELLAERTS, ma tante, Kathy MELLAERTS et, ma marraine, Vanessa MALINOWSKI, pour leur soutien, leur encouragement et la chance qu'ils m'ont donnée de réaliser les études que je souhaitais. Merci également à ma famille et mes amis pour leur aide morale qui m'a été précieuse tout au long de mon cursus universitaire.

"Un Data Warehouse ne crée pas de valeur lui-même mais la valeur vient de l'utilisation des données dans la Warehouse."
Parzinger and Frolick, 2001.

Table des matières

1 Introduction

1.1 Contexte

Dans le monde actuel, les données et informations provenant de toutes sortes de sources (web, cloud, data warehouse (DW),...) jouent un rôle fondamental dans la vie quotidienne des entreprises. Ces données représentent les fondations du processus de prise de décision. Pouvant être utile pour différents aspects du business, il est donc crucial pour l'organisation de choisir un modèle d'Enterprise Data Warehousing (EDW) qui convient à ses besoins et qui puisse stocker les données ainsi que donner un accès à un ensemble d'analyse permettant de faciliter les prises de décisions.

Existant un nombre important de modèles de Data Warehousing, il est donc difficile de faire un choix lorsque ce dernier se présente. Les techniques les plus largement déployées sont celle de Kimball (Kimball et Ross 2013) dans lequel les dimensions sont standardisées et classées dans des dimensions "mères" qui, une fois l' "extraction, transformation et loading" (ETL) exécutées (ibid.) sont réutilisables dans de multiples tables de fait. La seconde est celle d'Inmon (B. Inmon 2005) appelée la normalisation via un diagramme d'entité-association. Cette dernière technique est la forme basique du data warehousing basée sur les associations entre les attributs afin de déterminer les types de structures des entités. Chacun de ces modèles possède ses avantages mais un problème similaire se pose lorsque le système d'approvisionnement de la DW est analysé de plus près. La transformation des données, aussi nommé nettoyage, qui consiste à détecter et supprimer les erreurs et les inconsistances dans les données afin qu'elles soient pleinement utilisables dans leurs dimensions, provoque une perte dans le nombre, la qualité et la fiabilité des données (Rahm et Hai Do 2000). C'est dans l'optique de minimiser l'impact de l'inconvénient cité plus haut que le modèle Data Vault (DV) rentre en jeu.

1.2 Problème

La littérature actuelle offre de nombreuses introductions à ce type de modèle. Cependant, la comparaison du point de vue théorique et pratique quant aux avantages et inconvénients de ce modèle semble moins abondante. Le concept 2.0 commençant seule-

ment à se développer dans les entreprises belges, il est intéressant d'en présenter ce qui le constitue basé sur l'ouvrage "Building a scalable data warehouse with Data Vault 2.0" de Daniel Linstedt et Michael Olschimke (2016), tout en incrémentant d'informations provenant d'autres sources telles que des articles scientifiques, des journaux, . . . , ainsi que des interviews d'experts dans le domaine. Cette étude dresse un rapport et une évaluation objective sur le modèle de Dan Linstedt ainsi qu'une analyse sur sa structure actuelle et future. La question de recherche de ce mémoire peut être formulée comme telle : *"Le modèle Data Vault 2.0 est- il le mieux adapté lors du déploiement d'une Data Warehouse ?"*. Cette question générale peut être subdivisée en 3 sous-questions auxquelles cette étude apporte des questions :

- Quelles sont les éléments clés de la Data Vault 2.0 à respecter afin d'assurer la réussite du projet ? Est-ce que la méthodologie proposée est adaptée et optimale ?
- Quelles sont les avantages et inconvénients à utiliser la Data Vault 2.0 ?
- Dans quel cas, la Data Vault 2.0 est le meilleur choix lors de la décision de la mise en place d'un data warehouse ?

1.3 Approche

Ce mémoire abordera donc la Data Vault 2.0 d'un point de vue théorique, en commençant par un état de l'art et les bases du DW. Pour ensuite aborder la DV 2.0 plus en détail, en décrivant l'architecture, les composantes, la méthodologie et la modélisation de cette technique. Cette recherche se finira sur une analyse de cas grâce à l'aide d'experts DV 2.0 de chez Partena Professional et Dieteren Auto dans laquelle, l'objectif sera de faire ressortir les avantages et inconvénients remarqués sur le terrain afin d'apporter une approche pratique à l'analyse théorique réalisée et obtenir des conseils d'implémentation résultant d'une application réelle et journalière du concept DV 2.0.

2 Revue de la littérature

2.1 Le Data Warehousing

Le Data Warehousing est un concept qui englobe l'ensemble des méthodes, techniques et outils afin d'apporter une aide et un source de savoir supplémentaire (via l'analyse des données) aux membres d'une entreprise pour améliorer la performance lors de la prise de décision mais aussi pour améliorer la qualité des sources d'information.(Golfarelli et Rizzi 2009)

2.2 Introduction à la Data Warehouse

Suite à l'avènement et l'utilisation grandissante, à partir des années 60, des Database Management System (DBMS) et des applications online permettant l'échange et le partage de données entre les ordinateurs, l'idée de centraliser ces données et de les rendre opérationnelles pour toute l'entreprise commença à naître dans les esprits (Foote 2018). Il fallut tout de même attendre 1992 avant de voir apparaître le concept de DW par W.H. Inmon dans deux publications *"Rdb/VMS : Developing the Data Warehouse"* (W. H. Inmon et Kelley 1993) avec Chuck Kelley et *"Building the Data Warehouse"* (B. Inmon 2005).

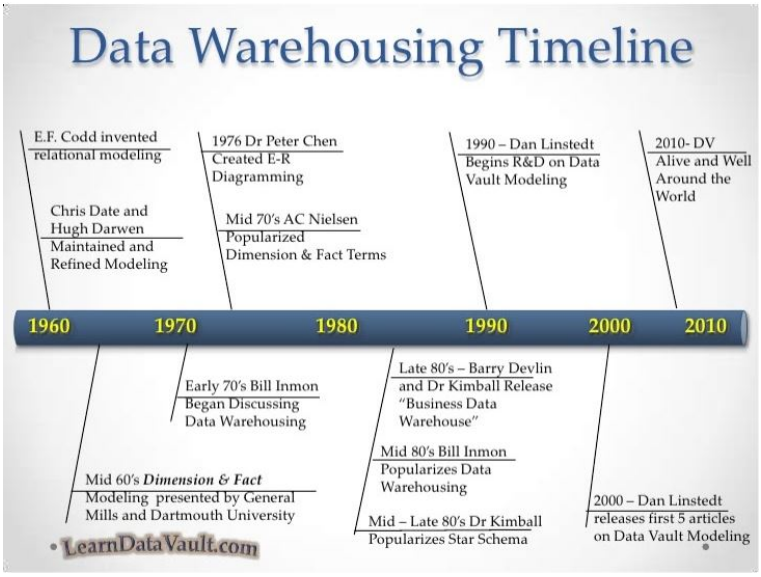

FIGURE 1 – La ligne du temps de la Data Warehousing (HOLDINGS 2011)

Le terme de data warehouse a été, au cours du temps, repris par de nombreux experts et a donc vu sa définition variée selon l'auteur. D'après Barry Delvin (DEVLIN 2000), la DW est le stock constant d'informations obtenu lorsqu'une sélection de sources de natures différentes est complètement accessible pour les end-users avec une approche reconnaissable et utilisable dans une perspective d'entreprise. Selon Michael Brackett (BRACKETT 1996), c'est *"un référentiel de données historiques cohérentes, facilement accessibles et manipulables à des fins d'aide à la décision"*. Quant à l'un des pères fondateurs de ce concept, W.H. Inmon (B. INMON 2005), il l'a initialement défini comme *"une collection de données orientée-objet, intégrée, variant dans le temps et non-volatile offrant un support dans le processus de management de prise de décision"*.

Basé sur la vision d'Inmon (ibid.), une DW possède 4 propriétés (B. INMON 2005 ; ORACLE p. d.[b]) :

- Orientée-objet
- Intégrée
- Variant dans le temps

- Non-volatile

Les DWs ayant pour but d'aider les data analystes dans le processus de décision, les données sont donc organisées selon les besoins et les visions du ou des business de l'entreprise. Il est important de noter que le concept "orienté-objet" dépend aussi de l'utilisation attendue des données, soit "orienté-application", soit en support du processus décisionnel. Dans le premier cas, les données sont détaillées et basées sur les exigences fonctionnelles, tandis que, dans le cas des DWs, les données incluses servent uniquement à des fins d'analyse et d'aide à la prise de décision. Le concept d'intégration permet d'assurer une consistance lorsque les données provenant de sources hétérogènes sont extraites. Par consistance, on entend consolider la base de données en s'assurant que le format des données introduites soit non-conflictuelle et assez similaire via des conventions de nommage standardisées, des structures d'encodages, etc. Une des caractéristiques principales de l'analyse décisionnelle est de pouvoir prédire le futur sur base de l'historique des données. Les DW peuvent avoir une dimension temporelle sur différentes périodes de temps : journalière, mensuelle, annuelle, etc. Cette dimension est utilisée dans l'index structurel de la DW et permet de représenter les données sous forme d'une suite de mises à jour qui garde les versions précédentes tout en ajoutant la nouvelle et offrant donc une traçabilité de l'évolution de la donnée au cours du temps. Cela peut être utile en cas de recherche de tendances pour le business mais aussi pour analyser les changements de comportements des consommateurs. La dernière caractéristique est la non-volatilité des données. En d'autres mots, une fois que la donnée est introduite dans la DB, elle ne changera plus. Cela permet de renforcer et d'assurer que l'analyse de ce qui a eu lieu au cours du temps soit correcte et que les données ont été inchangées depuis leurs insertions. (B. Inmon 2005 ; Brackett 1996)

Pour terminer cette introduction à la DW, il est important de faire la distinction entre les deux types de base de DW auxquels on peut être confronté : les "Relational Database" (RDB) ou les "Multi-Dimensional Database" (MDDB) (Humphries et al. 1999). La RDB (Christensson 2017) a pour but d'être une base de données dites "relationnelle". C'est une DB qui stocke les données dans un format structuré en utilisant des lignes et des colonnes qui permettent une localisation et un accès à une donnée précise facilités. Ce type de DW est

dit relationnel car les valeurs de chaque tableau sont relié aux autres, ainsi que les tableaux entre eux. Cette structure rend donc possible l'implémentation et l'application de requêtes sur différentes tables simultanément. Les requêtes et l'accès à la DB via le RDB peuvent se faire à l'aide du langage SQL (Structured Query Language) sur des applications telles que MySQL, Microsoft SQL, etc. Mais la raison pour laquelle ce type n'est pas considéré comme la meilleure solution en DW face à des demandes complexes, est dû à la manière dont les requêtes s'exécutent et vont rechercher les données. C'est-à-dire qu'à chaque fois qu'une requête est lancée, elle devra, pour recueillir les données souhaitées, parcourir parfois un million de records, ce qui n'est pas optimal d'un point de vue performance. La MDDB (Rouse 2005) est, quant à elle, un type de DB permettant l'application des cubes OLAP (Online Analytical Processing). Contrairement aux RDB, où l'utilisation de requêtes SQL est nécessaire pour extraire les informations nécessaires, une MDDB permet de poser des questions plus concrètes telles que "combien de contrats d'assurances vie ont été contracté au cours des 6 derniers mois dans la province de Namur ?". Ces questions, faisant appel à différentes dimensions (temporelle, géographique, etc.), sont liés à la synthèse des opérations et des tendances. Le concept d'une MDDB est basé sur l'idée d'un cube de données (nommé cube OLAP) pour représenter la multidimensionnalité des données exploitables par l'utilisateur. Prenons l'exemple d'un détaillant de matériaux, "ventes" peut être visualisée dans les dimensions du modèle du produit, de la géographie, du temps ou dans une autre dimension. Dans ce cas, "ventes" est appelé l' attribut de mesure du cube de données et les autres dimensions sont considérées comme l' attribut d'objets. Il est aussi possible pour le gestionnaire de la DB de définir des hiérarchies et des niveaux dans les dimensions (par exemple, dans la dimensionnalité géographique, il est possible d'avoir une division hiérarchique pour différencier le siège social des bureaux régionaux).

2.3 L'environnement Data Warehouse en entreprise ou EDW

L'EDW est une forme dérivée des DW ordinaires qui étaient appelés "Decision Support Systems (DSS)". Cette forme initiale de la DW offrait un accès rapide à l'information requise afin de supporter le processus de prise d'aide à la définition durant les années 60 (Power 2007). Un DSS (Linstedt et Olschimke 2016) est composé d'une DB de modèles

analytiques alimentés par un ensemble de données provenant des systèmes sources de l'entreprise. Ces systèmes sources sont en réalité tous les systèmes opérationnels disponibles au sein de l'organisation. Les données brutes sont donc regroupées dans la base de données renfermant les modèles analytiques ou directement dans le système (K. C. Laudon et J. P. Laudon 2019). Afin de permettre l'analyse des données provenant de sources et des formats différents, l'utilisation d'outils ETL (Extract, Transform, Load) est primordiale et fait partie d'un processus de préparation des données nécessaires afin d'assurer la pertinence des données extraites. Un exemple de DSS est présenté dans la figure 2 . La DB de modèles analytiques est chargé via l'ETL avec des données provenant de 4 sources. Les données sont ensuite agrégées par le processus ETL ou lorsque le business user interagit avec les données via des requêtes spécifiques et plus complexes en fonction des besoins de l'organisation (Golfarelli et Rizzi 2009).

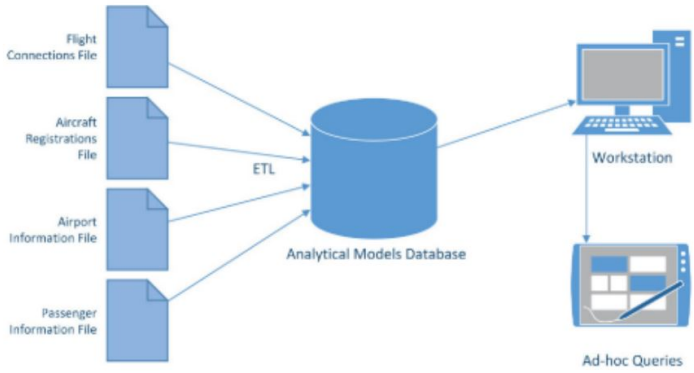

Figure 2 – Exemple d'un DSS (Linstedt et Olschimke 2016)

Alors qu'un DSS va uniquement se concentrer sur un seul domaine d'analyses, l'EDW essaye d'apporter une vue plus globale et horizontale dans laquelle l'ensemble des données de l'entreprise et leurs business rules sont visibles. Cela permet une représentation des données qui fournit aux business users la possibilité d'accès à tous les domaines d'analyses disponibles (GeekInterview 2007). Afin d'assurer la clarté du concept d'EDW et de marquer ses différences avec les DSS, 5 attributs doivent être respectés (B. Inmon 2005) :

l'unicité, la multiplicité des domaines, la normalisation, l'implémentation stratégique et l'extensibilité. Le premier attribut d'un EDW est qu'il ne doit posséder qu'une seule version de la vérité. Dans une organisation, il y a souvent plusieurs systèmes opérationnels, voire même plusieurs DW. La présence de disparités entre les données stockées est donc possible dû à des délais de synchronisation ou d'autres types d'erreurs, tel que des inputs manuels incorrects. Par exemple, un même client pourrait suite à une mise à jour non synchronisée posséder deux adresses de facturations différentes dans différentes tables de la DB. L'EDW doit donc assurer que les données qui sont stockées et qui alimentent le processus de décision soient cohérentes et qu'elles fassent partie d'une seule version de la vérité. Le deuxième attribut est qu'un Enterprise Data Warehouse doit avoir plusieurs domaines. Une entreprise est composée de différents départements qui possèdent chacun des attentes particulières concernant les données et la façon dont elles doivent être traitées et analysées pour qu'elles soient exploitables dans le cadre de leur environnement de travail. Chaque domaine offre donc des données qui sont pertinentes en fonction du type d'utilisateur. Afin d'assurer le respect de cet attribut, les données brutes sont nettoyées et chargées dans la DB après un passage dans le processus ETL. Cela permet ensuite le développement de data marts qui sont des référentiels conçus pour contenir des données qui offrent une assistance et un support aux départements auxquels ils sont dédiés. Il existe deux types de data marts : les dépendants, qui ont comme source le DW, et les indépendants, qui sont capables de récupérer les données nécessaires directement dans les systèmes opérationnels de l'organisation.

Le troisième attribut est qu'un EDW doit avoir une conception normalisée. En fonction du point de vue, cet attribut peut-être discutable lors d'un déploiement d'un DW, car les DBs normalisées et dénormalisées présentent chacune leurs propres avantages et inconvénients. Il existe certainement des DWs qui ont été conçues en suivant des modèles dénormalisés, comme, par exemple, les schémas en étoiles ou en flocons de neige (voir figure 3). Cependant, les DBs normalisées ont plus de succès car elles offrent avant tout une plus grande flexibilité. Le quatrième attribut est qu'un EDW doit être implémenté en tant qu'environnement stratégique. Étant donné que la DW représente les bases des

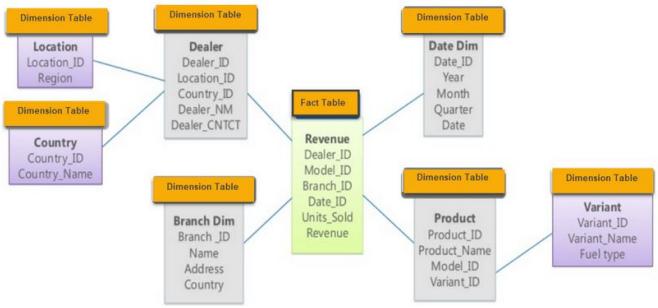

FIGURE 3 – Représentation d'un DB suivant un schéma en flocons de neige (BESENYEI p. d.)

prises de décisions stratégiques, l'infrastructure doit être à la hauteur. Les DWs ne sont pas seulement des atouts pour l'entreprise d'un point de vue décisionnel, ils aident aussi à enrichir les données présentes dans les systèmes opérationnels de l'organisation. Enfin, un EDW doit être extensible sur plusieurs dimensions. L'entreprise a pour objectif principal de croître et donc elle doit être capable de gérer la croissance des données ainsi que la complexité croissante des processus qui vont de pair avec l'évolution de l'organisation.

2.4 L'architecture Data Warehouse

En IT, l'architecture des DW est composée de modèles, normes, règles qui fixent un cadre de travail afin d'assurer une collecte, un stockage et une mise à disposition des données dans l'entreprise efficiente et régulée. (B. INMON 2005 ; HUMPHRIES et AL. 1999) Ci-dessus, une représentation du modèle d'architecture de Murtaza (LAPLUEA p. d.) est présentée, elle visualise une architecture à trois couches (basées sur la représentation d'Inmon) (B. INMON 2005 ; ABRAMSON p. d.), un des types d'architecture possible en data warehouse. La seconde forme possible étant l'architecture à deux couches (basées sur la représentation de Kimball), composée uniquement de la couche staging et la couche data warehouse, la couche data mart n'est donc pas utilisée dans ce cas. Suite à leur passage dans l'ETL, les données sont cohérentes, intégrées (B. INMON 2005 ; LAPLUEA p. d.) et, donc, prêtes à être chargé dans l'EDW. Sur base des données stockées dans l'EDW, il est possible de créer un data mart pour chaque service de l'entreprise qui nécessite une utilisation de

données dans son processus décisionnel. Turban et al. (2007) ont qualifié cette approche comme étant idéale car elle offre une image globale et fiable à l'entreprise.

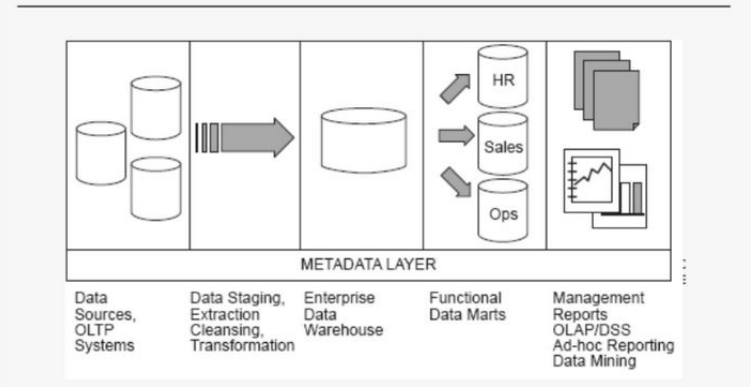

FIGURE 4 – Architecture en 3 couches selon (LAPLUEA p. d.)

2.5 La différence entre l'approche "top-down" et "bottom-up"

Dans la théorie, il existe deux types d'approches lors de la conception d'un DW. La première est nommée approche "top-down", de William H. Inmon, qui offre une vue consistante et définitive de l'information stocké dans la DW. Pour ensuite, utiliser cette information afin de créer les data marts (EDUCBA p. d.). La seconde approche est nommée approche "bottom-up" et a été introduite par Ralph Kimball. Cette approche consiste à apporter des rapports utilisables rapidement en créant directement les data marts en premier (BANGUY 2017).

2.6 Les datamarts

Un DM est un conteneur de données construit afin d'aider un domaine spécifique de l'entreprise (marketing, finance, ...) dans le processus d'aide à la décision. Alors que

Inmon	Kimball
Commence par la conception du modèle de DW	Commence par la conception du modèle dimensionnel pour les datamarts
Architecture composé d'un staging area permanent, d'un DW et de datamarts dépendants	Architecture qui consiste en un staging area et de datamarts, le DW physique n'existe pas
Le DW est orienté entreprise et les datamarts sont orientés processus	Les datamarts contiennent les données atomiques et agrégées
Le DW contient les données atomiques ; les datamarts les données agrégées	Les datamarts peuvent fournir une vue entreprise ou processus
Le DW utilise un modèle normalisé de toute l'entreprise; les datamarts utilisent des données dimensionnelles orientés sujet	Les datamarts sont implémentés de façon incrémentale et intégrée en utilisant les dimensions conformes
Les utilisateurs peuvent effectuer des requêtes sur le DW et les datamarts	

FIGURE 5 – Tableau comparatif entre le modèle de Inmon et Kimball selon (LAPLUEA p. d.)

leur construction et leur contrôle sont axés sur un seul service, les DMs effectuent leur extraction de données dans différentes sources comme : les sources externes, un DW central ou un système opérationnel interne de l'organisation (FENTAW 2014) pour ensuite centraliser et faciliter l'accès aux données pertinentes pour les utilisateurs du DM. Il existe deux types de datamarts, dépendants et indépendants, distinguable de par leur façon d'obtenir les données, appelé processus d'Extraction-Transformation-Transport (ETT). Ce processus est beaucoup plus simple pour les datamarts dépendants, puisque les données qui le peuplent, ont été préalablement nettoyer à l'aide d'un processus ETL lors du chargement dans la data warehouse. La mission du processus ETT, dans ce cas, consiste uniquement à s'assurer que les données soient correctement liées à leur(s) datamart(s) correspondant(s). Les datamarts indépendants quand à eux, recueillent leurs données via différentes sources externes ou systèmes opérationnels. Malgré cette indépendance au data warehouse, ces derniers doivent tout de même respecter et traiter les données en suivant les aspects du processus ETL. (ORACLE p. d.[a])

2.7 L'avenir du Data Warehouse

Depuis de nombreuses années, les entreprises ont investi des millions afin d'améliorer les performances de leurs EDWs, puisque ces derniers constituent à présent les fondations de leur business. Dans l'avenir, les EDWs occuperont un rôle majeur dans le développement des technologies de nouvelle génération telles que l'IA, le machine learning ou l' "Internet-

Of-Thing" (Trends et Applications 2019).

Avec l'avènement des services clouds, la DW va de plus en plus s'orienter vers ce type de service étant donné que son accessibilité, sa polyvalence et sa capacité à fournir les données en temps réel tout en gardant une trace historique représentent des atouts non négligeables pour les organisations (ibid.).

La DW va devenir de plus en plus analytique car les compagnies ont reconnu que le pouvoir d'analyse d'une DW est une capacité indispensable pour chaque aspect du business. Cela est rendu possible grâce à la combinaison de la DW et de la data science. Un exemple concret de ce que cette combinaison est capable d'apporter dans le futur est qu'en se basant sur les données concernant l'absentéisme dans une entreprise et en appliquant une analyse orientée "data science", il est possible de savoir quels événements incite le plus les employés à prendre des congés (vacances scolaires, jours fériés, compétitions de sport majeurs, etc. . .)(ibid.).

L'utilisation de la DW va requérir des teams de plus en plus réduits dû à l'autonomisation grandissante des environnements DW(ibid.). La DW va commencer à fournir des données au sein de lacs de données (Rouse 2015) (Hadoop, Spark,...). Un lac de données est *"un référentiel de stockage qui conserve une grande quantité de données brutes dans leur format natif jusqu'à ce qu'elles soient nécessaires."*(ibid.) Contrairement à la DW traditionnelle dans laquelle les données sont hiérarchisées et contenues dans des fichiers, un lac de données présentera les données dans une architecture à plat. Toutes les données sont chargées au sein du lac accompagné d' *"un identifiant unique et marquée au moyen d'un jeu de balises de métadonnées étendues"*. (ibid.) Les lacs de données apportent une plus grande flexibilité à la DW et une facilité à l'analyse des données peut importe la question métier. Sur base du scope et des besoins des différents métiers, le lac de données permet de rechercher les informations s'appliquant à la question posée et restreindre l'ensemble de données afin de faciliter l'analyse finale et assurer une réponse cohérente et bénéfique pour le business. Son utilisation est similaire à celle d'un datamart à l'exception que dans le data lake, contrairement aux DMs, les données stockées sont brutes et ne sont pas encore passées par l'ETL.

La place et l'importance de la DW dans l'optimisation de la customer experience continueront de se renforcer(Trends et Applications 2019). Créée et construite afin

d'améliorer la compréhension des besoins du consommateur, la DW offre à présent des données sur l'historique des clients ainsi que sur leur démographie et se verra renforcer dans le futur par un flux de données en temps réel qui fournira un nouveau type de service et des réponses afin d'améliorer l'expérience du consommateur.

3 Data Vault 2.0

3.1 Introduction à la Data Vault 2.0

Le Data Vault fait partie des techniques de modélisation disponibles dans le monde de la Business Intelligence. Son nom réel est "Common Foundational Warehouse Architecture" (Linstedt et Olschimke 2016). C'est une méthodologie de modélisation de données hybrides offrant la possibilité de visualiser l'historique des données de plusieurs sources hétérogènes. Conçu en 1990 et publié 2000 dans le domaine public, par Dan Linstedt, il définit sa création comme suit : *"Le Data Vault est un suivi historique axé sur les détails et un ensemble de tables normalisées liées de manière unique qui prennent en charge un ou plusieurs domaines fonctionnels de l'entreprise. il s'agit d'une approche hybride englobant le meilleur de la race entre la 3e forme normale (3nf) et le schéma en étoile.* (Linstedt 1990)

Data Vault 2.0 est composé de 4 concepts principaux (Linstedt et Olschimke 2016), qui combinés avec le savoir du business tel que la CMMI (Capability Maturity Model), Six Sigma, TQM (Total Quality Management) et le PMP (Project Management Professional) assure le bon déroulement des projets EDW sous DV 2.0 :

- La modélisation - *"Changements dans le modèle pour la performance et l'adaptabilité"*
- La méthodologie - *"Utilisation des meilleures pratiques des méthodes Scrums et Agile"*
- L'architecture - *"Incluant les systèmes NoSQL et Big-Data"*
- L'implémentation - "Basé sur des modèles, automatisations et de génération CMMI niveau 5"

La Data Vault 2.0 (Vos 2014) est une approche globale prenante en charge la modélisation (déjà couverte par la Data Vault 1.0) mais surtout une vision de l'architecture de la DW en couches d'intégration "end-to-end" et une nouvelle optique de la méthodologie offrant la possibilité d'utiliser d'autres concepts méthodologiques tel que Scrum/Agile, Six Sigma,... La figure, ci-dessous, offre une vue des ajouts apportés au concept Data Vault durant le passage à cette nouvelle version. À gauche, les concepts initiaux de la version 1.0 et à droite, en rouge, les concepts supplémentaires fournis par la version 2.0.

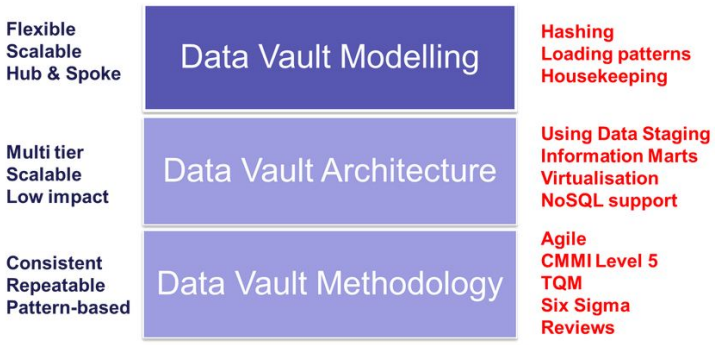

FIGURE 6 – Comparaison entre les versions 1.0 et 2.0 du modèle Data Vault.(Vos 2014)

Un de ces 4 concepts n'est pas réellement influencé par le passage à cette nouvelle génération de DV, l'implémentation. Étant donné que les concepts ont déjà été définis précisément en DV 1.0. Cependant, trois changements majeurs influençant la méthodologie, l'architecture et la modélisation du flux ETL (ibid.) se doivent d'être développé.

Tout d'abord, l'utilisation de "Hash Key". Elle fait maintenant partie des 3 alternatives afin de créer une clé primaire dans notre DW (Vos 2014 ; LINSTEDT 2018b). La clé primaire est donc le résultat d'un hachage plutôt qu'une séquence de nombre ou une clé business générer par l'outil ETL. Ce qui permet donc un chargement parallèle plus élevé qu'en DV 1.0. Le remplacement des clés des hubs, des links et des "Surrogate Keys" par des hash keys augmentant considérablement les performances de la DW, tout en réduisant la complexité des recherches des objets dans la DB.

Ensuite, L' "intégrité réferentielle" (IR), quant à elle, a été désactivée suite à l'utilisation

des hash keys. L'outil ETL n'étant plus apte à effectuer l'IR, cette dernière doit être traitée en interne. On nomme donc ce traitement le "soft IR". Teradata définit d'ailleurs ce concept de soft IR de la manière suivante : *"Le Soft IR fournit une définition déclarative pour une relation référentielle, mais il ne force pas la relation. La mise en vigueur de la relation référentielle déclarée est laissée sous la responsabilité de l'utilisateur à l'aide de la méthode de son choix."*(Teradata p. d.)

Enfin, L'intégration de DB non structurée NoSQL (Linstedt et Olschimke 2016) est rendue possible grâce à l'indépendance de la plateforme DV 2.0. NoSQL peut être utilisé dans chacune des couches de la DW (la couche "staging area", la couche "Data Warehouse" et la couche "datamart", qui seront décrits plus en profondeur dans la partie 3.2. L'architecture).

3.2 L'architecture

L'architecture DV 2.0 se base sur l'architecture à 3 couches, voire partie 2.4., ainsi que l'approche "top-down" de Inmon, et permet d'assurer l'extensibilité en y apportant les modifications suivantes (Linstedt et Olschimke 2016 ; Lans 2012a ; Lans 2012b) :

- Une distinction entre les hard et les soft business rules.
- Une couche nommée "staging area", dont le but est d'assurer le bon format et type de données au sein de l'EDW.
- Une couche nommée "data warehouse".
- Un ou plusieurs datamarts, aussi considérés comme la dernière couche de cette architecture. Le nombre de DMs variant en fonction des besoins particuliers de l'entreprise.
- Trois types de coffres (dit "vault" en anglais) optionnels faisant partie de la couche "data warehouse" :
  — Le "Metric Vault" qui permet d'obtenir et de garder une trace des informations d'exécution.
  — Le "Business Vault" qui renferme toutes les informations liées à l'application des business rules.

— L' "Operational Vault" qui recueille les données provenant du ou des systèmes opérationnels de l'entreprise.

Chacune de ces modifications apportées fera l'objet d'un développement plus en détails dans les prochaines sections.

De manière plus globale, l'architecture DV 2.0 est composée de 3 niveaux (EDUCBA p. d.).Le niveau inférieur constitué des sources de données, la staging area, de la data warehouse et des data marts. Le niveau intermédiaire composé uniquement de cubes multidimensionnels OLAP. Le niveau supérieur, dans lequel se trouve toutes les composantes front-end à destination du client ainsi que les outils nécessaire au reporting, à l'analyse et au data mining. Par exemple, l'outil de reporting "Tableau" est un type d'outils qui se situerait dans le niveau supérieur d'une architecture DV 2.0.

Seuls les deux premiers niveaux feront le sujet d'un développement théorique. Le niveau supérieur étant un niveau variant en fonction de chaque entreprise et de leurs affinités avec les différents outils disponibles sur le marché, l'étude réalisée ne se portera pas sur eux.

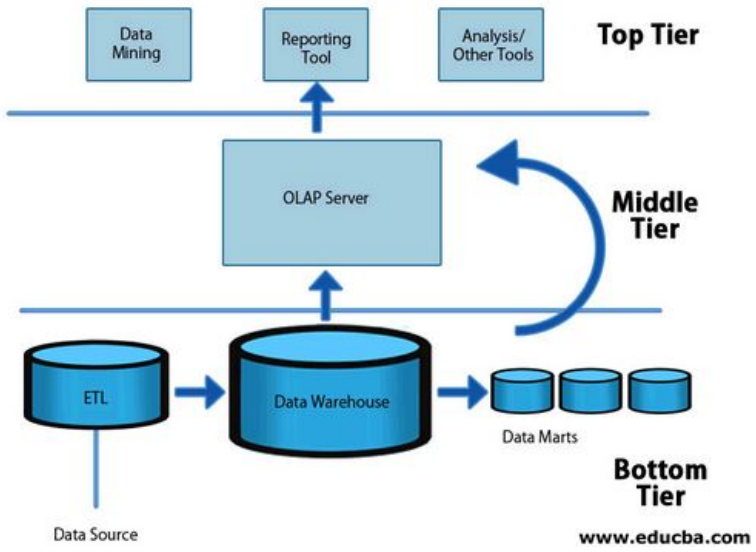

FIGURE 7 – Illustration des 3 niveaux de l'architecture Data Vault 2.0 selon (EDUCBA p. d.)

3.2.1 Définition des business rules (hard vs. soft rules)

De manière générale, les business rules sont les règles qui modifient et assurent que les données entrantes sont en concordances avec les attentes du business. En Data Vault 2.0, il existe 2 types de business rules : les règles business “soft” et “hard” (Linstedt et Olschimke 2016 ; Linstedt et al. 2017). En séparant l’interprétation des données stockées de l’alignement des règles, l’agilité de l’EDW et de l’équipe qui la gère peuvent s’accroître fortement.

Les règles “hard” se définissent comme “toute règle qui ne change pas le contenu de champs ou grains individuels”(Linstedt et al. 2017). Ces règles business “hard” sont les règles techniques qui permettent d’aligner les domaines de données, en d’autres elles permettent d’assurer la concordance entre les domaines brutes qui arrivent et les données attendues par le domaine (Linstedt et Olschimke 2016). Ce type de règles peut être employé dans différents cas (ibid.).En cas d’alignement de type de donnée. Par exemple, une source contenant des strings à un format plus long que celui attendu dans la table destination, alors un alignement de type via une règle “hard” préalablement définie sera nécessaire. En cas de normalisation/dénormalisation des tables. En cas de déduplication des données lors du chargement dans la base de données. Ces règles agissent uniquement sur les types de données et la normalisation lors du chargement dans l’EDW. En aucun cas, une règle “hard” ne modifiera la valeur chargée (arrondissements, conversions,...) dans la table.

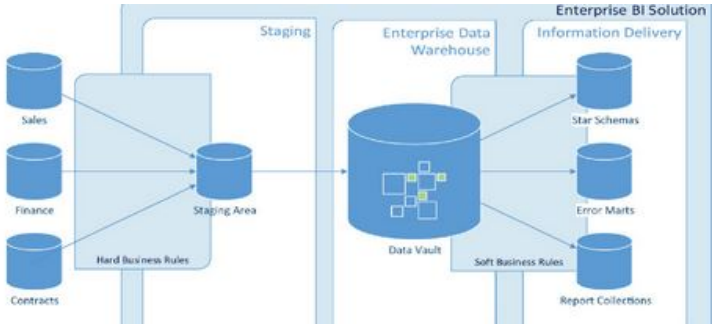

Figure 8 – Représentation visuelle du moment d’implémentation des règles “hard” et “soft” (Linstedt et Olschimke 2016).

Les règles “soft” se définissent comme l’opposé des règles “hard”, “toute règle qui change, interprète la donnée ou modifie son grain.” (Linstedt et al. 2017) Ce sont les règles “soft”, mise en place en collaboration avec les business users, qui vont permettre de donner un sens à la donnée et donc la transformer en information utilisable dans le processus décisionnel. Comme exemples d’utilisations courantes de ce type de règle, il y a (Linstedt et Olschimke 2016) : la concaténation des noms de champs, le calcul des ventes mensuelles ou encore la consolidation des données (un exemple concret est la consolidation des données clients brutes en vue consolidée qui permet au business d’avoir une vue d’ensemble et un accès plus rapide à la donnée lors de reportings).

Dans les DW conventionnels, au plus tôt les règles sont appliquées dans le flux de données, c’est-à-dire au moment de l’ETL, au plus la qualité des données se verra impactée de manière positive. Mais la dépendance à ces règles mises en place à une couche inférieure rendra le système moi agile en cas de changements ou améliorations provenant du business. La séparation en deux types de règles, via Data Vault, offrent l’assurance de données chargées sans erreur (règles “hard”) à la source, tout en minimisant la dépendance avec les couches supérieures en ne faisant intervenir les règles “soft”, plus tard dans le processus(Linstedt et Olschimke 2016 ; Linstedt et al. 2017).

3.2.2 Les 3 types de couches

L’architecture DV 2.0 peut être utiliser lors de deux types de situations bien distinctes qui définissent la présence ou non d’une staging area (Linstedt et Olschimke 2016 ; Linstedt p. d.) : le “Batch Loading” ou chargement par lots qui consiste à charger les données de manière séparée dans la staging area pour ensuite alimenter la DW et le “Real-Time Loading” ou chargement en temps réel. Dans ce cas, la DW s’alimente directement dans l’ESB de la compagnie. Dans le cadre de ce travail, afin de ne pas sortir du scope des experts intervenants, seul le batch loading sera abordé.

La première couche est la "staging area". Son rôle principal est de charger les données brutes provenant des différentes sources du système afin de rendre le système plus performant. Les opérations réalisées dans cette couche ne consistent pas uniquement à charger des données de tables provenant de la DB de production, elle peut aussi extraire

l'information contenue dans des feuilles Excel, des fichiers CSV, XML, cobol ou "flat" tout en effectuant des requêtes SQL qui peuvent être servies pour appliquer les règles "hard" ainsi que le contenu NoSQL (Linstedt et Olschimke 2016). *"La staging area ne contient pas de données historiques"*(Vos 2014). En d'autres mots, les données qui atterrissent dans cette couche ne vont y rester que temporairement, le temps de les introduire dans la data warehouse. En pratique, le chargement des lots est programmé de manière journalière, hebdomadaire, ... en fonction des besoins du business et du système afin que ce dernier puisse apporter les rapports nécessaires à la prise de décision.

D'après Dan Linstedt (Linstedt p. d.), voici les 6 raisons pour lesquelles la staging area est un atout majeur dans le développement d'un système DV 2.0. L'**adaptabilité** qui permet de partitionner ou non la structure en fonction des besoins spécifiques de l'entreprise ou l'ajout de activités si nécessaire. Par exemple, un des lots programmés durant la nuit n'a pas pu être accompli à 100 pourcents suite à une erreur survenue. Il est possible de rectifier cette erreur, en relançant le chargement de ce lot, parallèlement à celui déjà programmé à cet instant. Il est important de faire remarquer, qu'en règle générale, chaque lot est planifié et chargé dans la staging area de manière isolée. La **flexibilité** qui permet d'assimiler et d'introduire tout nouveau type de modification rapidement au sein du système (nouvelles sources, colonnes, ...). Le **dynamisme** qui offre la possibilité d'ajouter de nouveaux types de flux qui n'ont pas été programmés (erreurs dans le chargement ou sur demande spécifique du business). La possibilité de **redémarrer le système en cas d'erreurs**. La **planification** permet de diminuer les dépendances de timing dans le système. Puisque les chargements des lots de données sont planifié à l'avance et restent assez fixes au cours du temps, le système est donc plus flexible et s'adapte plus facilement en cas d'ajout/modification dans le planning. La possibilité de **sauvegarder et restaurer** les différents cycles de chargements.

En Data Vault 2.0, le "data warehouse" représente la seconde couche du système. Les données ayant été nettoyées dans la couche précédente à l'aide des règles "hard", seulement, sont stockées en respectant la granularité avec laquelle les données ont été chargés. D'après Linstedt (Linstedt et Olschimke 2016), le but de cette couche est de tenir un compte-rendu de toutes les données historiques, variant dans le temps sans pour autant être accédé directement par les end-users. Les informations contenues sont dites orientées fonction ("function-oriented") puisqu'elles sont ensuite redirigées vers un ou

plusieurs datamarts, qui représentent les différents départements de l'organisation et donc les différentes fonctions possibles que l'information peut prendre.

Enfin, la dernière couche est, comme son nom l'indique, composée des data marts, précédemment définis dans la section 2.6 Les Data Marts. La couche Data Mart est la seule couche en contact direct avec les end-users. Contrairement à la DW, les informations contenues sont dites orientées sujet ("subject-oriented") car elles doivent correspondre aux besoins du end-user. Cette couche suit, généralement, le modèle en étoile afin de former une source de données propres et prêtes à être utilisées à des fins de reporting ou dans des cubes OLAP (Linstedt et Olschimke 2016). L'objectif final de cette couche est de transformer les données en informations exploitables par le business dans la prise de décision(Graziano 2016a).

3.2.3 Les 3 types de "vault"

Contrairement aux couches précédemment étudiées, les 3 "vaults" ne possèdent pas de couche propre à eux (Linstedt et Olschimke 2016). Un "vault" est une extension de la couche DW dont le but est de faciliter la présentation aux end-users.

Premièrement, le "metric vault" est une couche, qui comme définit introduit au début de cette section 3.2, a pour but de fournir à l'équipe en charge du système des informations concernant la performance : les exécutions réalisées (historique des opérations, etc.), le CPU, les RAM utilisées, ...) (ibid.).

Ensuite, sur base de la définition donnée par Graziano K. (2016), la "Business Vault est une extension de la data warehouse qui applique les règles business "soft", des dénormalisations, des calculs et d'autres fonctions d'assistance aux requêtes afin de faciliter l'accès des utilisateurs et la génération de rapports. Les tables business vault doivent être actualisées une fois que leurs tables Raw Vault dépendantes sont actualisées"(Graziano 2016b). Le but de ce "vault" est d'obtenir une vue consolidée et pré-chargé avant d'introduire les données dans le datamart afin d'alléger le processus de chargement (Linstedt et Olschimke 2016).

Finalement, l' "Operational Vault" diffère par rapport aux deux "vault" précédents, puisqu'il possède la capacité d'extraire les données qui proviennent directement de sys-

tèmes opérationnels(Linstedt et Olschimke 2016). Le but de ce “vault” est “d’offrir des opérations plus rapides” (Linstedt 2010) ainsi que de ne plus dépendre des chargements de lots de données, source interne au système, en permettant d’obtenir des données détenues par des web-services (Linstedt et Olschimke 2016 ; Linstedt 2010).

3.3 La méthodologie

Dans cette section, la méthodologie et les différents modèles et frameworks utilisés au cours du cycle de vie d’un projet Data Vault seront abordés. Ce cycle de vie est divisé en 3 étapes : planification management, exécution et revue améliorations (Linstedt et Olschimke 2016). Chacune de ces étapes renferme un ou plusieurs modèles décrits ci-dessous.

3.3.1 Planification management

Le premier modèle utilisé dans cette étape porte le nom de CMMI (en anglais Capability Maturity Model Integration), créé en 1991 grâce à la collaboration du SEI (Software Engineering Institut), le DOD (Département de la Défense américaine) et la société Mitre (organisation des technologies américaine) . Depuis plus de 20 ans, ce modèle a pour but d’aider toute organisation, peu importe la taille de celle-ci. L’objectif de ce framework est d’apporter une structure et une méthodologie aux entreprises afin de créer et améliorer leurs projets, en rationalisant les coûts, en réduisant les erreurs, donc l’obligation de revenir en arrière en les corrigeant et en améliorant les gestions du temps et de la qualité générale du projet. Le modèle CMMI possède 3 variantes, chacune spécifique aux besoins de l’entreprise qui y a recours : le développement, l’acquisition ou le déploiement de services. Dans le cadre d’un projet Data Vault, la compagnie se tournera vers le CMMI pour le développement dans le but d’optimiser de planifier, manager et contrôler les activités de l’équipe. Comme son nom l’indique, ce modèle va permettre de définir la situation du projet vis-à-vis de deux indicateurs : la capacité et de la maturité du projet. Le premier indicateur offre une représentation continue des performances et des améliorations qui restent à faire pour les différentes parties du processus, en les analysant de manière individuelle. L’indicateur de capacité est divisé en 6 niveaux (allant du niveau 0 = incomplet, au niveau

5 = optimisation). Chaque niveau est complémentaire au précédent et améliore la capacité du projet en y ajoutant de nouvelles fonctions ou exigeant une rigueur plus accrue. L'indicateur de maturité, contrairement à celui de la capacité, possède un champ d'application plus globale. En d'autres mots, plutôt que d'analyser chaque zone du processus individuellement, ce dernier va traiter des ensembles de composantes du processus. Similairement au premier indicateur, la maturité est mesurée en 5 niveaux complémentaires (allant du niveau 1 = initiale, au niveau 5 = optimisation). (Linstedt et Olschimke 2016 ; Balázs 2017 ; Alexandre p. d. ; CMMI p. d.)

L'utilisation du schéma Scrum sert à appliquer la méthode agile habilitant tous projets Data Vault à être plus flexible afin d'augmenter leurs chances de succès. Cette méthode porte le nom d'une stratégie de rugby dans laquelle une équipe réalise un objectif fixé grâce à l'effort individuel de chaque membre. Scrum est composé de 4 phases. Tout d'abord, la planification et le staging, durant lequel l'équipe et le business s'alignent sur les attentes et les différentes activités à réaliser afin de faire aboutir le projet. Ces activités portent le nom de "user stories", sont stockées dans ce que l'on nomme le product backlog (catalogue des tâches à accomplir) et sont classées par ordres de priorités. Au plus une user story est importante, au plus elle devra être détaillée afin que lorsqu'un membre de l'équipe décide de la prendre en charge cette activité soit parfaitement compréhensible et réalisable de manière autonome. Ensuite, la troisième phase, le développement, se base sur l'approche itérative. Lors du développement, des périodes de 2 à 4 semaines nommées Sprints seront mises en place. Au cours d'un sprint, chaque membre se verra attribuer plusieurs "user stories" qu'il devra avoir terminé à la fin de cette période. Dans l'optique de pouvoir déceler éventuellement les problèmes de développement ou de s'adapter aux modifications demandées par le business, le "Scrum Master", personnage central de cette méthodologie dont le rôle est de gérer le bon déroulement du projet ainsi que d'effectuer la liaison entre l'équipe et le business, réalise des "daily standup". Quotidiennement, chaque membre de l'équipe fait un point sur son avancement et partage ses résultats ou ses obstacles avec le reste de ses collègues. De cette manière, l'équipe et le business corrigent les erreurs au fur et à mesure de l'avancement du projet. Finalement, le déploiement ("release" en anglais) est la phase durant laquelle l'équipe délivre une nouvelle version du système au business pour être ensuite mis en production si les attentes sont remplies. (Linstedt et Olschimke

2016 ; Tochi 2014 ; Advanced Development Methods 2003)

3.3.2 Exécution

Dans cette étape, l'approche utilisée est le cycle de vie traditionnel d'un développement de softwares selon le modèle en cascade. Ce modèle étant le premier modèle de processus introduit, il reste assez simple à comprendre. Divisé en 5 phases dont l'output de chaque représente l'input nécessaire au démarrage de la phase suivante, le modèle en cascade est dit "séquentiel". La première phase à réaliser est la "définition des exigences" du projet. L'équipe Data Vault doit recueillir toutes les exigences du business et techniques via des interviews, des observations de l'utilisation de la base de données par le business, l'analyse des documents renfermant les règles, politiques de données appliquées, des réunions avec les personnages-clés de l'entreprise (managers, analystes) ou encore le développement d'un prototype. Le design est la deuxième phase dans laquelle l'architecture de la data warehouse est défini, les différentes couches, les modules ainsi que les tables, leurs noms et colonnes sont expliqués en détail. L'infrastructure sera très souvent représentée dans un schéma entité-relation, ce qui permet de mettre en évidence les potentielles faiblesses du projet. La troisième phase consiste à l'implémentation et au "unit-testing" (test des modules individuellement). L'avant-dernière phase est constituée de l'implémentation et du test du système dans sa globalité. Lorsque les modules ont passé les tests avec succès dans la phase précédente, ils sont intégrés dans le système qui sera lui-même de nouveau testé. Pour finir, le système est rendu opérationnel et mis à disposition des end-users. L'équipe Data Vault s'assure de la maintenance en cas d'erreurs trouver lors de l'utilisation par le business. (Linstedt et Olschimke 2016 ; Yu et al. 2012 ; Sharma 2019)

3.3.3 Revue améliorations

Le modèle Six Sigma est le premier modèle employé dans un projet Data Vault lors de l'étape de revue améliorations. "Six Sigma est simplement une méthode pour résoudre efficacement un problème. L'utilisation de Six Sigma réduit la quantité de produits défectueux fabriquée ou de services fournis, résultant en une augmentation des revenus et une plus grande satisfaction client" (GoLeanSixSigma 2012). Ce modèle est composé de trois élé-

ments indispensables à son bon fonctionnement. Tout d'abord, un engagement et un suivi de la direction de l'entreprise. Un encadrement bien défini permet à la direction de garder un contrôle et s'assurer que les activités de chaque membre du projet se déroulent sans soucis. Ensuite, l'implication des différents utilisateurs du système (employées, clients, etc.) est le second élément nécessaire à l'utilisation du Six Sigma. Effectivement, le fait d'incorporer dans le processus d'amélioration et d'inciter les end-users du système à faire part de leurs feedbacks aide à pouvoir détecter les erreurs plus rapidement et optimiser le système. Finalement, le troisième élément est l'approche DMAIC (Define-Measure-Analyze-Improve-Control). La première étape est donc de définir le problème ainsi que de décrire précisément son effet sur le système et l'utilisation par le end-user afin que ce dernier puisse être parfaitement compris par le membre de l'équipe qui prendra en charge sa résolution. Ensuite, il faut mesurer les performances avant et après la résolution du problème afin de savoir si l'amélioration est un succès ou pas. L'analyse est la troisième étape de cette approche qui vise à chercher et trouver l'origine du problème. Pour déboucher sur le développement et l'intégration de l'amélioration, dans cette étape les membres de l'équipe vont développer la solution qui offre les performances les plus optimales. Enfin, la dernière étape se charge de contrôler et de s'assurer de la bonne application du modèle Six Sigma. Cette approche étant un cycle, si les attentes ne sont pas respectées, il suffit de redémarrer ce processus. (Linstedt et Olschimke 2016 ; GoLeanSixSigma 2012 ; Radha Krishnan et Arun Prasath 2013)

Le dernier modèle rencontré est le "Total Quality Management" (TQM) modèle. Basé sur un suivi continu, le TQM constitue l'ensemble des pratiques auxquelles l'entreprise (ou plus spécifiquement l'équipe Data Vault en fonction de l'organisation et des end-users du système mis en place) doit avoir recours afin d'atteindre voire de dépasser les exigences du client. Dans ce modèle, il est impératif que l'ensemble des acteurs, peu importe leurs fonctions au sein de la compagnie, fassent partie du processus continu de recherche de qualité supérieure. Lui aussi divisé en 5 phases, le TQM se compose de la préparation au cours de laquelle l'équipe va préparer toute l'implémentation, d'un point de vue ressources techniques et humaines pour augmenter les chances d'aboutissement

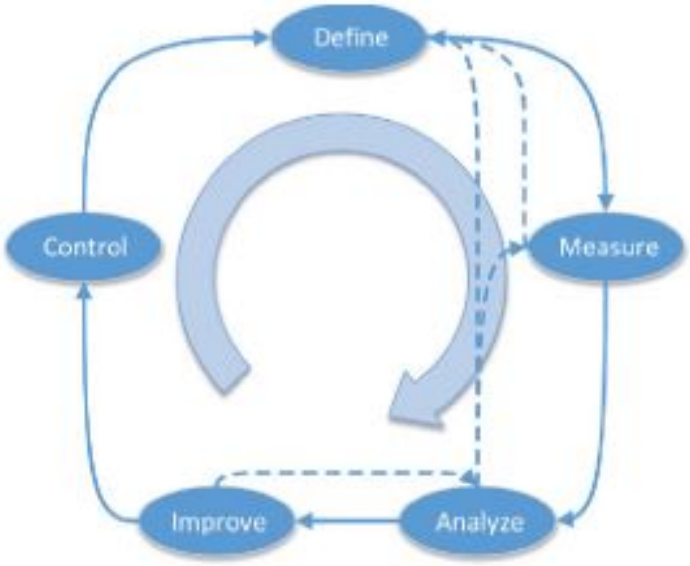

FIGURE 9 – Représentation de l'approche DMAIC (LINSTEDT et OLSCHIMKE 2016).

du projet. Cette phase aboutit sur la planification du projet. Suivie de l'évaluation, durant laquelle l'équipe Data Vault va recueillir des informations concernant l'entreprise afin de mieux connaître le contexte et les particularités de l'environnement de développement. La quatrième phase est l'implémentation du projet, les pratiques à utiliser sont partagées et répandues dans l'organisation. Ce qui aboutit sur la phase d'établissement d'un réseau au sein de la compagnie. Tous les acteurs sont impliqués dans un effort commun afin d'assurer ce suivi et cette recherche continue d'améliorations (gage de qualité). (LINSTEDT et OLSCHIMKE 2016 ; ARIKKÖK 2012 ; SCIENCEDIRECT p. d.)

3.4 La modélisation

Lors de sa création dans les années 90 par Dan Linstedt, le modèle Data Vault a été inspiré des réseaux complexes que l'on retrouve dans la nature, tels que le cerveau humain ou les serveurs peer-to-peer tel que utorrent. Le point commun entre tous ces réseaux est qu'ils sont composés de liens, de noeuds reliés entre eux. Afin de reproduire ce type de réseaux complexes et de le reproduire dans un modèle orienté business applicable à la DW,

le Data Vault se compose de 3 types de composants ainsi que leurs définitions, selon Dan Linstedt (LINSTEDT 2018a) :

— • Un hub est “une liste unique de business keys”.
— • Un link est “une liste unique de relations entre deux ou plusieurs business keys”.
— • Un satellite est une table de dimension temporelle qui renferme les attributs d’une business key ou d’une relation(LINSTEDT et OLSCHIMKE 2016 ; LINSTEDT 2018a).

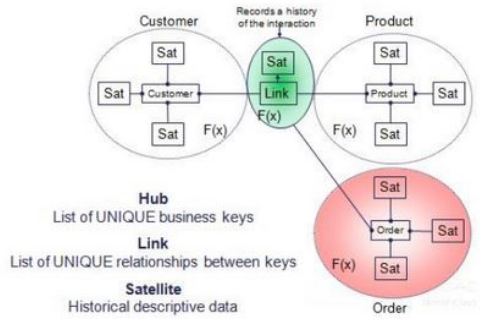

FIGURE 10 – Exemple de la structure d’un modèle Data Vault. (MAKINGDATAMEANINGFUL 2018)

Le terme business key représente ce que le business utilise afin de tracker, localiser et identifier des informations tout en étant indépendant au système opérationnel, unique et avec une tendance à subir des modifications très faible (LINSTEDT et OLSCHIMKE 2016 ; LINSTEDT 2018a).

3.4.1 Les hubs

Les hubs sont des tables contenant des business keys, éléments centraux du modèle, à la localisation des données dans la data warehouse. Ils stockent également des métadonnées additionnelles telles que la source d’enregistrement ou le numéro de séquence. Son rôle est de réagir à chaque nouvelle business key entrante lors du chargement de la warehouse en utilisant les métadonnées afin de connaître la source (“record source”), la date et l’heure du chargement de cette business key mais aussi généré une clé de hachage, nécessaire au référencement par rapport aux links et aux satellites. Grâce à ces clés, le business est donc capable de tracer les informations dans l’ensemble du système (LINSTEDT et

Olschimke 2016 ; Graziano 2011a ; MakingDataMeaningful 2018). Kent Graziano (2011) a défini l'objectif des hubs comme étant "de fournir un point d'intégration souple de données brutes qui n'est pas modifié par le système source, mais est censé avoir la même signification sémantique" (Graziano 2011a).

La structure d'une table hub est composée de 4 éléments obligatoires (Linstedt et Olschimke 2016 ; MakingDataMeaningful 2018) : une clé de hachage qui représente la clé primaire de ce hub (une séquence de nombres générée par la DB), une ou des business keys (un string pouvant contenir tout type de données),une date de chargement (une date et une heure) et une source d'enregistrement (un string).

Figure 11 – Structure d'un hub. (Graziano 2011b)

La clé de hachage a pour rôle d'augmenter les performances du système. Puisqu'elle est la clé primaire dans le hub, et donc la clé étrangère dans les links et satellites qui référence à cet hub, offrant une vitesse de recherche de données dans la DW plus rapide. Chaque business key stocké dans un hub, doit avoir sa clé de hachage unique. L'algorithme de hachage recommandé en Data Vault est celui de Ronald Rivest datant de 1991. Nommé MD5 ou "Message Digest 5", cet algorithme renvoie l'empreinte numérique de l'objet dans la base de données (Linstedt et Olschimke 2016 ; Rouse 2017).

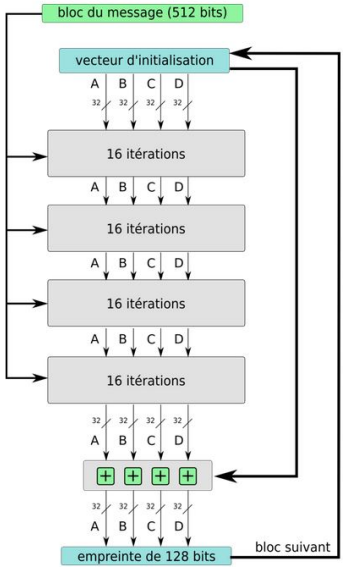

FIGURE 12 – Représentation de l'algorithme de hachage MD5. (WIKIPÉDIA 2020)

3.4.2 Les links

Les links contiennent les relations entre les business keys contenues dans les hubs, ils portent aussi le nom d' "entité associative" (GRAZIANO 2011a) . En d'autres mots, "les links capturent et enregistrent les relations passées, présentes et futures entre les éléments au plus petit niveau de granularité possible" (LINSTEDT et OLSCHIMKE 2016). Ils représentent le ciment du modèle Data Vault, puisqu'ils permettent d'enregistrer ces liaisons, sans devoir se soucier de la nature de la relation (many to many, one to many). À chaque interaction entre deux ou plusieurs business keys, une nouvelle entité link est générée. Par exemple une transaction dans le système, une nouvelle relation demandée par le business entre deux business keys, etc. Les tables links ont deux rôles majeurs(LINSTEDT et OLSCHIMKE 2016 ; GRAZIANO 2011a ; MAKINGDATAMEANINGFUL 2018) qui sont de stocker toutes les connexions entre les objets de la datawarehouse qui interviennent dans le business process (hiérarchies, achats, ventes, marketing,...)(LINSTEDT et OLSCHIMKE 2016) et de permettre la flexibilité et l'adaptabilité du modèle. Le modèle est capable de s'adapter assez rapidement en fonction des changements constants de le business et ses besoins spécifiques "sans perte d'audibilité et de conformité" (GRAZIANO 2011a).

La structure d'une table link se compose de 4 éléments (Linstedt et Olschimke 2016 ; Graziano 2011a ; MakingDataMeaningful 2018) : une clé primaire permettant l'identification du link dans la base de données, générée de manière similaire à celle d'un hub, une date de chargement, une source d'enregistrement et au minimum 2 clés étrangères de hubs ou d'autres links, si ce n'est pas le cas alors ce link est considéré comme invalide.

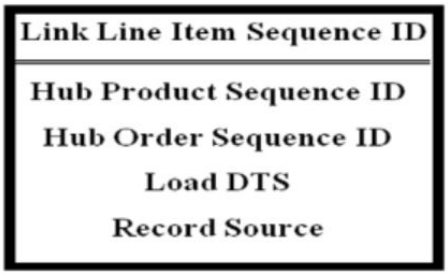

Figure 13 – Structure d'un link. (Graziano 2011a)

La structure d'un link ne peut jamais être compromise, par exemple en contenant des clés business naturelles (qui doivent se trouver impérativement dans les hubs), au risque de mettre en péril la flexibilité du modèle et amener à revoir la structure du modèle dans le futur (Graziano 2011a).

3.4.3 Les satellites

Les satellites sont les entités du modèle Data Vault qui donnent un sens et un contexte aux business keys et aux relations que l'on retrouve dans l'EDW "à un moment précis dans le temps ou sur une période de temps" (ibid.). Un satellite aura toujours un seul et unique parent (hub ou link), il est donc identifié par la clé de son parent ainsi que la date et l'heure du changement. Le but principal d'un satellite est d'être une table de dimension temporelle qui renferme les informations nécessaires à la description du contexte, soit en extrayant les données nécessaires dans le système comme description, soit en les entrant manuellement ou bien en y introduisant des éléments calculés dans le process. Des exemples concrets de données descriptions sont la nature d'un contrat assurance, les caractéristiques du bien vendus (couleur, taille,...), etc. (Linstedt et Olschimke 2016 ; Graziano 2011a ;

Graziano 2011b).

La structure d'un satellite se compose de 4 éléments obligatoires en plus des données descriptives qu'il contient (Linstedt et Olschimke 2016 ; Graziano 2011a) : une clé primaire composée de deux parties (la clé de hachage du parent, qui sera aussi la clé étrangère du hub ou link lié à ce satellite, et la date de chargement du satellite), une date de chargement final et une source d'enregistrement.

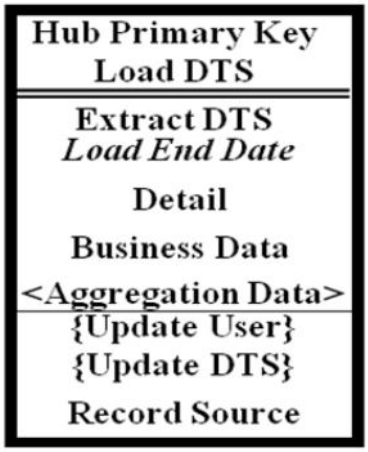

Figure 14 – Structure d'un satellite (Graziano 2011a)

.

4 Méthodologie

Pour rappel, l'objectif de cette thèse est d'étudier l'implémentation de la Data Vault 2.0 lors d'un projet de développement d'un EDW. Il est ici question d'analyser si les principes de ce modèle sont respectés en pratique, ainsi que de mettre en évidence les motivations de choisir ce framework. Dans cette optique, le choix d'une étude qualitative par entretiens individuels semble être le plus adapté afin d'analyser le déroulement en pratique du déploiement d'un data warehouse grâce à la Data Vault 2.0 . En effet, il est question ici d'étudier de manière précise et complète les pratiques de la gestion d'un projet DV 2.0, les avantages de ces pratiques ainsi que les éventuels désavantages rencontrés par nos experts. L'entretien individuel offre l'opportunité au répondant de pouvoir détailler le

fond de sa pensée sans devoir faire attention à l'opinion de ces pairs, ce qui permet un approvisionnement en résultats plus riches et plus complets à cette recherche.

4.1 Recueil d'informations

Un guide d'entretien est réalisé dans le but d' en assurer le bon déroulement (voir annexe 1). Le modèle semi-directif a été utilisé lors de l'élaboration de ce guide et s'inspirent des différents éléments de l'architecture, méthodologie et modélisation abordées précédemment. L'utilisation de ce type d'entretien garantit au chargé d'étude que chaque sujet d'intérêt sera discuté, sans pour autant entraver le chemin de pensées du répondant en l'enfermant dans une structure de questionnement rigide (Malhotra et al. 2011).

La base d'échantillonnage se compose de deux personnes travaillant dans le secteur de la consultance en Business Intelligence, pour la société de référence Akabi. Cette dernière a été sélectionné en appliquant un critère de convenience sampling. Le critère de sélection utilisé lors de la sélection des candidats est le purposive sampling et plus précisément l'expert sampling. Effectivement, ce critère sélectionne les répondants selon un critère d'expertise par rapport à un sujet (Etikan et al. 2016). Lors de cette étude, l'expertise réside dans leur expérience avec le modèle DV 2.0 et son application sur le terrain. Ceux-ci doivent participer ou avoir participé à un projet dans ce domaine. S'il avait été possible d'avoir une base de candidats potentiels plus grande (via des entreprises diversifiées), sélectionner les candidats de manière hétérogène selon un critère hiérarchique aurait enrichi l'analyse quant à l'influence de ce critère sur la perception de ce modèle.

Pour l'évaluation de la réussite des projets, un questionnaire a été transmis aux deux répondants à la fin de l'entretien. Basé sur sept questions reprises de l'étude de Pedro Serrador (Serrador et Pinto 2015), ce questionnaire offre la possibilité d'analyser le projet selon deux axes : son efficacité globale ainsi que son succès vis-à-vis de toutes les parties prenantes du projets(voir annexe 2).

Le premier projet porte sur le domaine de la mutualité chez PARTENA, est composé d'une équipe de 7 personnes. L'expert interviewé occupe le poste de consultant/développeur depuis un an et demi. Le second projet, quant à lui, porte sur le domaine automobile chez D'Ieteren. L'équipe est constituée d'un effectif de dix personnes. L'intervenant est

développeur et fait parti de la team depuis 4 ans.

4.2 Analyse des interviews

Une analyse thématique a été réalisée afin d'analyser les informations pertinentes obtenues lors des entretiens individuels(voir annexes 3 et 4). Le guide d'entretien ayant été construit dans le but de pouvoir aborder toutes les parties importantes de la DV 2.0, comme la méthodologie, l'architecture ou encore la modélisation. L'analyse ne se porte que sur le fond des interviews et non la forme, les unités d'analyse employées sont les phares et les paragraphes. Connaissant à l'avance les éléments à étudier, une grille d'analyse a été mise en place et structuré autour de ces éléments. Le dernier point de l'analyse des résultats qui reste à présenter est la triangulation. La triangulation fait référence au mécanisme de comparaison des informations obtenues via l'étude.Malhotra, Décaudin et al. (2011) décrit trois stratégies dans livre : la triangulation entre les sources, les théories, les méthodes ou encore les chercheurs. La triangulation entre chercheurs ne fait pas partie des stratégies utilisables étant donné le caractère personnel de ce travail. Le choix d'avoir recours seulement à une étude qualitative écarte automatiquement la triangulation par les méthodes et les théories s58. En conclusion, la triangulation entre les sources est la seule appliquée.

5 Résultats

Dans cette section, les résultats obtenus grâce à l'étude qualitative sont présentés. Ces résultats sont abordés par rapport aux différents thèmes du framework Data Vault qui occupent un rôle-clé : son architecture, sa méthodologie ainsi que sa modélisation. L'analyse est divisée en 2 parties. Premièrement une analyse inter-projet est menée afin d'étudier les éléments de la Data Vault 2.0 qui sont employées véritablement en pratique. Les avantages et inconvénients seront aussi analysés dans cette partie. Ensuite, une seconde analyse intra-groupe est réalisée afin d'évaluer si la bonne implémentation de la Data Vault participe au succès du projet.

5.1 Analyse intergroupe

Cette première partie des résultats s'intéresse aux éléments rencontrés en pratique lors de l'implémentation de la Data Vault.

Pour l'architecture, chacun des projets respecte la structure en 3 couches d'Inmon (Linstedt et Olschimke 2016). C'est-à-dire composé d'une staging area, d'un data warehouse et de datamarts propre à un framework Data Vault. Concernant les règles "hard" et "soft", même si les deux intervenants se sont exprimés avec des termes assez différents, il est évident que leur fond était similaire. En pratique les règles "hard" servent à deux rôles : définir les frontières d'accès à la staging area ainsi que la gouvernance technique de l'équipe data vault. La staging area, quant à elle, a une définition similaire selon les deux experts mais la technologie et la répartition d'équipe ne le sont pas. Pour commencer, la staging area représente le stockeur des données brutes en transit de la source jusqu'au data warehouse. Le projet de chez Partena diffère de celui de chez Dieteren, car alors que dans le premier, une partie de l'équipe est strictement dédié à nourrir la staging area, dans le second projet, chaque membre de l'équipe est capable de gérer l'entièreté du flux (de l'extraction jusqu'aux datamarts). Au sujet des datamarts, chaque équipe crée les datamarts de manière orienté-business. Pour terminer, uniquement la "vault" dite business est représenté dans les deux projets. Pourtant, chez Dieteren, ce "business vault" porte le nom de Business Domain View (BDV). Leur nom est différent mais leur objectif similaire : apporter une couche de logique supplémentaire sur le data warehouse renfermant éventuellement les règles "soft" ou des fonctions logiques.

Ensuite autour de la méthodologie, le premier modèle "CMMI" n'est visiblement pas employé. Tandis que le prochain est celui le mieux suivi à la lettre par les deux répondants. Des sprints de deux semaines sont organisés au cours du quel, chaque membre de l'équipe se verra attribuer des tâches à réaliser nommé user-story. Nos deux projets, malgré leur utilisation impeccable du framework scrum, s'éloignent sur un point. Alors que Dieteren n'a parlé que de choisir en fonction de la priorité par rapport aux besoins du business, chez Partena, les user-stories sont hiérarchisées de la plus importante à la plus petite tâche : "Epic", "Features", "User-Story" et "Tasks". Cette hiérarchie fonctionne comme des poupées russes, c'est-à-dire que chaque élément plus petit fait d'office parti d'un

ensemble plus grand dont l'objectif est fixé à un plus long terme. Le modèle en cascade en data vault 2.0 insiste sur le test des modules au fur et à mesure du développement, pourtant chez Dieteren, d'après Mr. Vanoverberghe, cette pratique ne cesse de s'estomper. Contrairement à Partena, où un framework de test de module automatisé est actuellement en développement. La seconde grande différence entre ces deux projets est que, chez Partena, les membres de la team sont en charge du test de ses modules et du passage en production. Tandis qu'avant de pouvoir passer en production, chez Dieteren, une confirmation de passage doit être octroyer par le business après avoir poussé le module dans l'environnement de Quality and Acceptance (QA). Finalement, les modèles Six-Sigma et TQM se retrouvent dans les deux projets étant donnés que l'ensemble des acteurs est impliqué dans le processus de développement et feedbacks. Il est important de noter que l'implication est plus poussée chez Partena, puisque le business participe même au sprint donc peut donner son avis en temps réel sur la planification du sprint et les attentes réelles du business.

La modélisation est le dernier point à aborder. Chacun des répondants possédait un avis positif sur le Data Vault et sa modélisation mais avait tout de même, certaines craintes. Cette manière de modéliser est très répétitive et peut donc être rapidement automatisé afin de pouvoir s'adapter très agilement aux changements de taille de la DB, aux changements du business, ... Pourtant, le fait est que l'automatisation accélérera le splittage propre à la Data Vault, ce qui engendrera un nombre important de nouveaux links, hubs, satellites. Ce qui est problématique puisque les développeurs doivent garder une certaine connaissance et compréhension du mapping entre les objets. Mr. Vanoverberghe qualifia à juste terme la modélisation comme étant à "double-tranchant".

5.2 Analyse intragroupe

Cette section passe en revue les projets individuellement et analyse leur implémentation en Data Vault 2.0. Pusiqu'une analyse globale est effectuée préalablement, il est ici question de n'aborder que les spécificités de chaque projet.

Le projet de Mr. Ben Aissa, Partena, semble être beaucoup plus tourné vers la version 2.0 de la Data Vault, pourtant il ne semble pas que le modèle CMMI soit employé. C'est la seule divergence enregistrée mais il est possible qu'en nature du poste de Mr. Ben

Aissa, développeur, le fait de connaitre la capacité et la maturité du projet ne lui soit pas destiné mais plutôt à des postes plus élevées, tels que scrum master ou chef de projet. Les réponses au questionnaire (voir annexe 2) permettent d'évaluer que ce premier projet est un succès pour les différents acteurs mais aussi que l'efficacité est assurée. Effectivement, sur l'échelle allant de 0 à 4, 4 étant la note de respect et de satisfaction maximum, les réponses sont toutes 3 sauf un 2 pour le respect des délais.

Le second projet analysé, grâce à Mr. Vanonverberghe, est chez Dieteren. Âgé de presque 10 ans, ce projet data vault a donc fait ses débuts avec la version 1.0. Les divergences avec le modèle 2.0 dans sa globalité viennent donc du fait, que le projet data vault de cette compagnie est un hybride entre l'ancienne et la nouvelle vision de Dan Linstedt. Pas de CMMI et une implication moindre du business dans le déroulement du framework scrum ainsi que dans le processus d'améliorations sont le résultat de choix propre à l'entreprise. Cela semble bien marché niveau succès, puisque les notes montrent une réussiste supérieur à l'autre projet avec une note maximale de 4 concernant le respect du budget et le restant des 3. À l'exception du 2 pour la satisfaction de la team.

5.3 Constats

Durant la section 1.2, la question de recherche principale a été énoncé mais aussi séparée en trois sous-questions reprises ci-dessous et accompagnées des réponses basées sur les résultats obtenus.

- *Quelles sont les éléments clés de la Data Vault 2.0 à respecter afin d'assurer la réussite du projet ? Est-ce que la méthodologie proposée est adaptée et optimale ?*
  Comme vu dans la section 5.2, il semblerait qu'au plus une compagnie implémente et utilise la data vault pendant longtemps, au moins son coup sera important. Une seconde observation, toujours dans section 5.2, peut être notée. En regardant les notes pour la satisfaction en tant que membre de l'équipe, on note que Partena a une note plus élevée. Cela évoque éventuellement une corrélation positive entre l'implication de chaque acteur dans le projet, peu importe qu'ils soient du business ou non, et le succès du projet.

- *Quelles sont les avantages et inconvénients à utiliser la Data Vault 2.0 ?* La section 5.1 a permis de mettre en exergue les avantages suivants : la capacité à s'adapter, l'automatisation facilitée et la répétitivité dans les requêtes qui facilite la prise en main. Pour les désavantages, il y a : la taille imposante d'un tel projet (tant au niveau ressources technologiques qu'humaines) et le temps d'implémentation.

- *Dans quel cas, la Data Vault 2.0 est le meilleur choix lors de la décision de la mise en place d'un data warehouse ?* La réponse à cette dernière question fait à nouveau référence à la section 5.1., la Data Vault 2.0 est un choix à prendre en compte si votre secteur professionnel est un environnement très véloce, dans lequelle, les données sont perpétuellement en mouvement. Il est important de bien analyser le cas de la compagnie avant de prendre la décision car l'équipe devra être expérimentée, les infrastructures à jour mais aussi être conscient que de la patience sera requise due à la complexité de l'implémentation.

6 Limitations

Durant ce mémoire, l'étude qualitative était de type exploratoire. Cela permet d'offrir un nouveau point de vue à la recherche. Ce qui engendre donc souvent la succession d'une étude quantitative dans le but de confirmer ou non les corrélations observées durant la recherche. Deux types de validités sont à prendre compte lors de la mise en place d'une étude qualitative ou quantitative. La validité interne représente la capacité de l'étude d'affirmer qu'une relation observée est causale. La validité externe quant à elle se définit comme la généralisation hors de l'étude de la causalité identifiée durant l'étude, c'est-à-dire dans un autre contexte, à une autre période ou avec d'autres individus (Johnson 1997). Dans le cas de cette thèse, la triangulation permet de couvrir la validité internet (Johnson 1997). Suite au covid-19 et les difficultés à obtenir une réponse de différents intervenants, le nombre d'experts interrogés n'est que de deux, cela correspond tout de même à une forme de diversification d'utilisation de sources. Deux corrélations ayant été découverte durant l'analyse, la validité externe se doit être vérifiée dans le futur à l'aide de la logique

de réplication, avec un échantillon d'experts plus étoffés et pas uniquement évoluant dans le domaine de la Data Vault.

7 Conclusions

À cet instant de la recherche, il est bon de rappeler la question d'étude sur laquelle se penche ce mémoire : "Le modèle Data Vault 2.0 est- il le mieux adapté lors du déploiement d'une Data Warehouse ?". Il peut donc être conclu que le modèle Data Vault 2.0 peut-être le mieux lors d'un déploiement, uniquement après une étude et une analyse complète de la compagnie, son secteur et ses infrastructures. Il est important de comprendre qu'une grande partie du succès d'un projet data warehouse, dépend de l'analyse et du choix de la technologie et du modèle fait en amont de l'implémentation. Le déploiement d'un data warehouse à l'aide de la data vault est conseillé pour toutes les entreprises possédant différents départements avec un flux de données entrant très actif et qui nécessite donc une mise à jour récurrente de ces données.

Références

Abramson, I (p. d.). *Data Warehouse : The Choice of Inmon vs. Kimball. IAS Inc.* [Slydes Powerpoint][Consulté le 8 janvier 2020]. url : http://www.scribd.com/doc/253618546/080827Abramson-Inmon-vs-Kimball#scribd.

Advanced Development Methods, Inc. (2003). *Scrum Methodology : Incremental, Iterative Software Development from Agile Processes*. [Consulté le 21 mai 2020]. url : https://itq.ch/pdf/SCRUM_methodology.pdf.

Alexandre, F. (p. d.). *Introduction au CMMI*. [Consulté le 18 mai 2020]. url : www.mit.bme.hu.

Arikkök, M. (2012). *TOTAL QUALITY MANAGEMENT*. [PDF][Consulté le 24 mai 2020]. url : https://www.researchgate.net/publication/312054032_TOTAL_QUALITY_MANAGEMENT.

Balázs, S. (2017). *Introduction Basics of CMMI Project Management*. [PDF][Consulté le 18 mai 2020]. url : www.mit.bme.hu.

Banguy, Z. (2017). *2 approches pour construire un Data Warehouse (DW)*. [Consulté le 15 mai 2020]. url : https://www.aerow.group/a16u1509/.

Besenyei, E. (p. d.). *Star and SnowFlake Schema in Data Warehousing*. [Consulté le 29 mai 2020]. url : https://www.eandbsoftware.org/star-and-snowflake-schema-in-data-warehousing/.

Brackett, M. H. (1996). *The data warehouse challenge : taming data chaos*. New-York : John Wiley & Sons.

Christensson, P. (2017). *RDBMS Definition*. [Consulté le 7 janvier 2020]. url : https://techterms.com.

CMMI, Institute (p. d.). *CMMI Levels of Capability and Performance*. [Consulté le 18 mai 2020]. url : https://cmmiinstitute.com/learning/appraisals/levels.

Devlin, B. (2000). *Data warehouse : from architecture to implementation*. Reading, MA : Addison-Wesley.

EDUCBA (p. d.). *Data Warehouse Architecture : Different Types of Layers And Architecture*. [Consulté le 15 mai 2020]. URL : https://www.educba.com/data-warehouse-architecture/.

ETIKAN, I. et AL. (2016). *Comparison of convenience sampling and purposive sampling*. T. 5.1. American journal of theoretical et applied statistics, p. 1-4.

FENTAW, A. E. (2014). *Data Vault Modelling An Introductory Guide*. Helsinki : Metropolia University of Applied Sciences.

FOOTE, K. D. (2018). *A Brief History of the Data Warehouse*. [Consulté le 23 octobre 2019]. URL : https://www.dataversity.net/brief-history-data-warehouse/.

GEEKINTERVIEW (2007). *What is Enterprise Data Warehouse*. [Consulté le 8 janvier 2020]. URL : http://geekinterview.com/data-warehouse/data-types/what-is-enterprise-data-warehouse.html.

GOLEANSIXSIGMA (2012). *The Basics of Lean Six Sigma*. [PDF][Consulté le 22 mai 2020]. URL : http://www.goleansixsigma.com/wp-content/uploads/2012/02/The-Basics-of-Lean-Six-Sigma-www.GoLeanSixSigma.com_.pdf.

GOLFARELLI, M. et S. RIZZI (2009). *Data warehouse design : modern principles and methodologies*. Emeryville, CA : McGraw-Hill.

GRAZIANO, K. (2011a). *Introduction to Data Vault Modeling*. [PDF][Consulté le 17 mai 2020]. URL : https://kentgraziano.files.wordpress.com/2012/02/introduction-to-data-vault-modeling.pdf.

— (2011b). *Introduction to Data Vault Modeling*. [Slides en ligne][Consulté le 17 mai 2020]. URL : https://www.slideshare.net/kgraziano/introduction-to-data-vault-modeling.

— (2016a). *Building an Information Mart With Your Data Vault*. [Consulté le 16 mai 2020]. URL : https://www.vertabelo.com/blog/data-vault-series-building-an-information-mart-with-your-data-vault/.

— (2016b). *The Business Data Vault*. [Consulté le 16 mai 2020]. URL : https://www.vertabelo.com/blog/data-vault-series-the-business-data-vault/.

Holdings, E. (2011). *Operational Data Vault*. [Consulté le 23 octobre 2019]. url : https://www.slideshare.net/dlinstedt/operational-data-vault.

Humphries, M. et al. (1999). *Data warehousing : architecture and implementation*. Upper Saddle River, NJ : Prentice Hall, p. 114-115.

Inmon, B. (2005). *Building the Data Warehouse (5th ed.)* Indianapolis : John Wiley & Sons.

Inmon, W. H. et C. Kelley (1993). *Rdb/Vms, developing the data warehouse*. New-York : Wiley.

Kimball, R. et M. Ross (2013). *The Data Warehouse Toolkit (3rd ed.)* Indianapolis : John Wiley & Sons.

Lans, R. F. (2012a). *Data Architecture - A Primer for the Data Scientist : Big Data, Data Warehouse and Data Vault*. Amsterdam : Morgan Kaufmann. Chap. 2.5.4 The Operational Data Store. url : https://doi.org/10.1016/C2011-0-07129-6.

— (2012b). *Data Architecture - A Primer for the Data Scientist : Big Data, Data Warehouse and Data Vault*. Amsterdam : Morgan Kaufmann. Chap. 7.5.2 Strategy 2 : Developing a New Business Intelligence System with Data Virtualization. url : https://doi.org/10.1016/C2011-0-07129-6.

Lapluea, T. (p. d.). *"Why Enterprise Data Warehouse ?" : Data Warehousing Development at Offco Ltd.* [Consulté le 8 janvier 2020]. url : https://www.academia.edu/208553/_Why_Enterprise_Data_Warehouse_Data_Warehousing_Development_at_Offco_Ltd..

Laudon, K. C. et J. P. Laudon (2019). *Essentials of management information systems*. Harlow, England : Paerson.

Linstedt, D. (1990). *Data Vault Basics | Accelerated Business Intelligence*. [Consulté le 5 mai 2020]. url : https://danlinstedt.com/solutions-2/data-vault-basics/.

— (2010). *Accelerated Business Intelligence*. [Consulté le 16 mai 2020]. url : https://danlinstedt.com/allposts/datavaultcat/introduction-to-the-operational-data-vault/.

Linstedt, D. (2018a). *Data Vault Data Modeling Specification v 2.0.2*. [PDF][Consulté le 16 mai 2020]. url : https://danlinstedt.com/wp-content/uploads/2018/06/DVModelingSpecs2-0-1.pdf.

— (2018b). *DV2 Sequences, Hash Keys, Business Keys – Candid Look*. [Consulté le 6 mai 2020]. url : https://danlinstedt.com/allposts/datavaultcat/dv2-keys-pros-cons/.

— (p. d.). *Accelerated Business Intelligence*. [Consulté le 15 mai 2020]. url : https://danlinstedt.com/allposts/datavaultcat/data-vault-and-staging-area/.

Linstedt, D. et al. (2017). *Data Architecture - A Primer for the Data Scientist : Big Data, Data Warehouse and Data Vault*. John Wiley & Sons. Chap. Chapter 4.3 Introduction to Data Vault Architecture - Hard and Soft Business Rules, p. 152-153.

Linstedt, D. et M. Olschimke (2016). *Building a scalable data warehouse with Data Vault 2.0*. Waltham, MA : Morgan Kaufmann.

MakingDataMeaningful (2018). *Data Vault : Hubs, Links, and Satellites With Associated Loading Patterns*. [Consulté le 17 mai 2020]. url : https://makingdatameaningful.com/data-vault-hubs-links-and-satellites-with-associated-loading-patterns/.

Malhotra, Naresh et al. (2011). *Etudes marketing (6è éd.)* Paris : Pearson Éducation.

Malhotra, Naresh, Jean-Marc Décaudin et al. (2011). *Etudes marketing (6è éd.)* Paris : Pearson Éducation.

Oracle (p. d.[a]). *Oracle8i Data Warehousing Guide*. [Documentation en ligne][Consulté le 15 mai 2020]. url : https://docs.oracle.com/cd/A87860_01/doc/server.817/a76994/marts.html.

— (p. d.[b]). *Orcale9i Data Warehousing Guide : Data Warehousing Concepts*. [Consulté le 24 octobre 2019]. url : https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.html.

Power, D. J. (2007). *A Brief History of Decision Support Systems*. [Consulté le 7 janvier 2020]. url : https://dssresources.com/history/dsshistory.html.

Radha Krishnan, B. et K. Arun Prasath (2013). *SIX SIGMA CONCEPT AND DMAIC IMPLEMENTATION*. T. 3. International Journal of Business Management Research (IJBMR), p. 111-114.

Rahm, E. et H. Hai Do (2000). *Data Cleaning : Problems and Current Approaches*. Germany : University of Liepzig.

Rouse, M. (2005). *What is multidimensional database (MDB) ?* [Consulté le 7 janvier 2020]. url : from%20https://searchoracle.techtarget.com/definition/multidimensional-database.

— (2015). *Data lake (lac de données)*. [Consulté le 5 mai 2020]. url : https://www.lemagit.fr/definition/Data-lake-lac-de-donnees.

— (2017). *What is MD5 ?* [Consulté le 18 mai 2020]. url : https://searchsecurity.techtarget.com/definition/MD5.

Sciencedirect (p. d.). *Total Quality Management*. [Consulté le 24 mai 2020]. url : https://www.sciencedirect.com/topics/computer-science/total-quality-management.

Serrador, Pedro et Jeffrey K Pinto (2015). « Does Agile work ?—A quantitative analysis of agile project success ». In : *International Journal of Project Management* 33.5, p. 1040-1051.

Sharma, L. (2019). *WaterFall Model in Software Developement Life Cycle : SDLC*. [Consulté le 22 mai 2020]. url : https://www.toolsqa.com/software-testing/waterfall-model/.

Teradata (p. d.). *Referential Integrity*. [Documentation en ligne][Consulté le 6 mai 2020]. url :https://docs.teradata.com/reader/m~O~fVLqvU~MIZ5ZcaXIhg/EJ7TMoDXMHoWLZWkBI5E3w.

Tochi, I (2014). *Characteristics of Agile SCRUM*. [Consulté le 21 mai 2020]. url : Https://www.researchgate.net/publication/268811634_Agile_SCRUM_Methodology_-_A_Project_Management_Framework.

Trends et Applications (2019). *Rethinking The Future of Data Warehousing*. [Database][Consulté le 8 janvier 2020]. url : https://www.dbta.com/DBTA-Downloads/WhitePapers/Rethinking-the-Future-of-Data-Warehousing-8918.aspx.

Vos, R. (2014). *Data Vault 2.0 - Introduction and (technical) differences with 1.0*. [Consulté le 5 mai 2020]. URL : http://roelantvos.com/blog/data-vault-2-0-introduction-and-technical-differences-with-1-0/.

Wikipédia (2020). *MD5*. [Consulté le 18 mai 2020]. URL : https://fr.wikipedia.org/wiki/MD5.

Yu, B. et al. (2012). *Software Development Life Cycle AGILE vs Traditional Approaches*. T. 37. IPCSIT, p. 162-163.

9 Annexes

Plan des annexes

Annexe 1: Questionnaire Interview

1. Topo générale de la discussion

- **Introduction**

Présentation: bienvenue et remerciements, présentation de l'animateur et du déroulement de l'entretien (objectif global, durée,...), absence "d'évalution" (pas de bonnes ou mauvaises réponses).

Aspects déontologiques: principe de l'anonymat, autorisation d'enregistrer, utilisation des données à des fins universitaires.

- **Phase d'échauffement (mettre à l'aise et présenté l'interviewé)**

Tout d'abord, pouvez-vous vous présenter brièvement? (études, employeur, client/domaine (banque, sécurité sociale, ...), année(s) d'expérience, ...)

De manière générale, quelles sont les caractéristiques d'une implémentation Data Vault 2.0?

2. Discussion spécifique

- **Phase de centrage sur le sujet**

En bref, pouvez-vous décrire la gestion d'un projet Data Vault 2.0? (brièvement: aborder la Data Vault 2.0 en en tirant un portrait général)

Dans combien de projets Data Vault avez-vous été impliqué? Quel(s) en étaient les secteurs d'applications/ domaines professionnels? Quel rôle occupiez-vous?

- **Phase d'approfondissement**

- **Architecture**

Quels sont les différentes composantes de la Data Vault et les rôles respectifs?

Règles business Hard/Soft: Quand et comment sont-elles définies au sein de votre projet?

Staging Area: Quelles sont les personnes impliquées dans le chargement de données de la staging area? Comment se déroule cette étape au sein de votre équipe? Quelle programme utilisez-vous?

Data Mart Area: Combien de data marts possèdent votre projet? Comment définissez-vous les data marts à créer? Suivez-vous un procédé particulier? Quelles sont les avantages et inconvénients de l'utilisation de data marts?

Les 3 "Vault" optionnels (metric, operational et business): Avez-vous recours à l'utilisation d'un de ces vault optionnels dans votre projet? Si oui, lequel et pourquoi? Quelles sont les avantages que cela apporte à votre projet?

- **Méthodologie**

Quelles sont les modèles de méthodologie liés à la Data Vault qui sont appliqués pour le développement du projets?

<u>Planification & Management</u>

CMMI: Comment mesurez-vous à quelle niveau de capacité et de maturité se trouve votre projet? Qu'apporte de telles informations pour la suite du déroulement du projet?

SCRUM: Comment appliquez-vous la méthodologie scrum au sein de votre équipe? (Product backlog, user-story, hiérarchie des tâches, daily-standup,...) Quels sont les avantages et inconvénients de ce framework? Combien de user-story par membre d'équipe par sprint? Quelle est la durée moyenne d'un sprint?

Exécution

Modèle en cascade: Comment avez-vous recueilli les exigences du business? Quels sont vos rapports avec les end-users? Quand est-il du unit-testing et du system-testing au sein de votre projet? Comment le/les respectez-vous?

Revue & améliorations

SIX-Sigma/TQM: Comment se déroule le processus de feedbacks et d'améliorations du projet? Quels sont les acteurs impliqués? Les actions prises? Et le type d'échange renfermant les proposition d'améliorations (emails, meetings,...)

- **Modélisation**

Décrivez les 3 composantes d'une structure data vault ainsi que leurs rôles? (hubs, links, satellites) Qu'apporte cette manière de modéliser?

3. Conclusion

Récapitulatif éventuel des idées clés

Quels sont les avantages/inconvénients de la Data Vault 2.0 dans une implémentation d'EDW? Quels sont les éléments jugés clés?

Idées supplémentaires? Critiques?

Remerciements

Annexe 2: Mesure de succès des projets

Mesure d'efficacité

1. Comment le projet respecte-il les objectifs de budget?

	0	1	2	3	4	
Ne respecte pas du tout						Respecte parfaitement

2. Comment le projet respecte-il les délais?

	0	1	2	3	4	
Ne respecte pas du tout						Respecte parfaitement

3. Comment le projet respecte-il les objectifs d'exigences du projet?

	0	1	2	3	4	
Ne respecte pas du tout						Respecte parfaitement

Mesure de succès des parties prenantes

1. Comment évaluez-vous la satisfaction des sponsors/parties prenantes concernant le projet?

	0	1	2	3	4	
Très mauvais						Excellent

2. Comment évaluez-vous la satisfaction du client concernant le projet?

	0	1	2	3	4	
Très mauvais						Excellent

3. Comment évaluez-vous la satisfaction de la Team concernant le projet?

	0	1	2	3	4	
Très mauvais						Excellent

4. Comment évaluez-vous la satisfaction des utilisateurs finaux concernant le projet?

	0	1	2	3	4	
Très mauvais						Excellent

Réponses recues

Questions (dans l'ordre d'apparition ci-dessus)	Lucas Vanoverberghe	Badreddine Ben Aissa
1.	4	3
2.	3	2
3.	3	3
1.	3	3
2.	3	3
3.	2	3
4.	3	3

Annexe 3: Interview des Experts Data Vault

A: Tout d'abord, bienvenu à vous. Merci de m'accorder votre temps. Donc moi, je me présente je m'appelles Amaury Mellaerts et je suis en Master 2 à l'université de Namur en Data Analytics et business. Donc… euh… ici le but de ce mémoire sur lequel je travaille c'est surtout de pouvoir confronter la théorie et la pratique dans le data vault. Je vais vous demander 30 à 45 minutes de votre temps pour justement vous demander votre avis et votre expérience sur le sujet. Donc il n y aura pas d'évaluation par rapport à vos réponses. Il n'y a pas de bonnes et de mauvaises réponses. Toutes les réponses que vous allez donner seront uniquement utilisées dans un cadre universitaire et cela restera anonyme si vous le souhaitez. Mais d'abord, je dois vous demander si vous acceptez que j'enregistre votre entretien… Vous m'entendez?

B: Euh oui

L: Oui pas de problème pour moi donc euh voilà.

B: Oui pas de soucis

A: Chacun votre tour, je vais vous demander de vous présenter brièvement, votre parcours les etudes que vous avez faites, votre employeur actuel, le domaine et le projet dans lequel vous travaillez. Et, euh, vos années d'expériences dans le domaine de la BI et plus particulièrement la data vault.

L: D'accord, euh, vas-y Badreddine.

B: Ok pour moi, je suis Badreddine Ben Aissa, je suis ingénieur informatique, je travaille avec Akabi et je suis un consultant chez Partena professionnel depuis euh novembre 2018. Et en faite, mon expérience avec le data vault a commencé avec partena, donc j ai presque un an et demi maintenant en data vault et 3 ans en tant que consultant bi.

A: Ok. Et vous Lucas?

L: Donc moi, Lucas Vanoverbergh, je suis également consultant chez akabi comme Badreddine, j'ai d'abord fait des études en bachelier informatique et système et puis j'ai fait un master en computer science à l'université de mons et donc mtn je suis en mission chez Dieteren Auto depuis 4 ans. Et depuis 4 ans, on travaille avec data vault comme data warehouse principale. Donc je participe au développement des flux ETL, au développement de datamart, l'automatisation et de mise en place de données pour le business en utlisant la méthodologie data vault.

A: Ok parfait. Donc en bref, est-ce que vous pourriez me décrire la facon dont vous percevez l'exécution et la gestion d'un projet Data Vault. Donc me tirer la grande photo du Data Vault.

B: Je t'en prie Lucas.

L: D'abord pour clarifier ta question, c'est quoi que tu appelles un projet data vault? Est-ce que tu veux dire que c'est le projet de mettre en place, de créer le data vault en lui même. Ou bien d'utiliser

le data vault, donc la méthodologie et la database que tu as mis en place pour créer un projet business. Parce qu'en fait c'est deux choses différentes.

A: D'accord, en fait. Moi le cadre de ce mémoire, c'est plus l'analyse que ca soit de la méthodologie, l'architecture et de la modélisation. C'est vraiment du moment ou l'on décide de choisir la data vault comme modèle jusqu'à la livraison.

L: Oui et quand tu dis livraison du projet. C'est livré un data warehouse, data vault qui ne répond pas encore à un besoin business ou alors un projet business qui utilise ce data vault.

A: C'est plus.. Euh.. Comment qui soit utile pour le business.

L: D'accord, est-ce que tu inclus dedans le fait qu'il faut développer un flux etl, une structure, choisir une technologie, etc. Ou tu considères que ca c'est deja acquis.

A: Non, non. Enfait rien n'est acquis. Enfin vis-à-vis du flux etl, moi j'en parle avec tout ce qui est la staging area. Donc voilà, moi c'est plus. Moi j'aimerais juste savoir en gros si vous on vous parle d'un projet data vault. À partir du moment où on décide d'utiliser la technologie et celui où le end-user va l'utiliser, ce qui vous vient à l'esprit en fait.

B: Donc un projet data vault from scratch quoi.

A: C'est ca.

L: Ok, donc on va dire tous les aspects. Et donc en gros, ta question de base finalement quelles seront les grandes d'un tel projet.

A: C'est ca. C'est voir si vous avez les mêmes grandes étapes que moi j'ai lu dans la théorie et ensuite j'aborderai les différentes étapes plus précisement et de manière plus approfondie.

L: D'accord ouais.. Euh.. grosse questio. Et bien, je pense que dans un premier temps déjà analyser le besoin. Si tu parles de data vault, c'est que tu as déjà fait une étude pour savoir quelles sont les besoins en terme de data, de data intelligence et est-ce que data vault est appriori une bonne méthodologie qui a l air de pouvoir répondre à ce besoin. Pour déjà voir quel est le product case, quel est le besoin de ton business, quelle est la vitesse de réaction dont tu as besoin, quelles la structure des données de ton entreprise, quelles sont tes applications sources et voir sur base de ca est ce que le fait de fonctionner en data vault ca pourrait s'intégrer ou pas? Donc si la réponse est oui après tu te lances dans la mise en place. Donc je pense qu'il faut déjà cette première phase d'alignment de communication, de stratégie avec ton entreprise avec ton entreprise pour voir si ca s'intègre bien dans la stratégie de data que ton entreprise a.

A: Ok.

L: Après il y aura une phase d'architecutre donc je pense où il faut avoir une idée concrète de globablement ce à quoi va ressembler ta stack technologique BI. Donc tu vas te mettre avec des architectes et faire une proposition d'architecture sur quels sont les grands éléments que tu auras probablement un gros data warehouse centrale qui utilisera la modélisation data vault, tu vas pouvoir

importer des données provenant des sources. Donc il va falloir voir comment tu peux mettre en place les intéractions entre tes sources et ton data warehouse qui va avoir la gouvernance de ramener les data, qui va avoir le leadership, comment on va ramener les data, est-ce quelles vont être poussées dedans, est-ce qu'on va aller les chercher pour les ramener? Quelles technologies ils va falloir utiliser pour extraire les données, quelle technologie on devra utiliser pour la database, quelle technologie on devra utilser pour après publier les données et les mettre à disposition. Donc c'est les choix technologiques et l'architecture globale de tout. Ca me parait être une grosse deuxième phase, le développement et la mise en place à proprement dite. Donc si tu as ton plan d'architecture comment tu vas fonctinner? Après il faut commencer à mettre en place des poc, faire des implémentations. Tu mets en place une database, des outils ETL. Tu testes ces outils, tu crèes les environnements. Donc tu dois pouvoir créer différents environnements par exemple, developpement, test, tu as prod pour tester des développements et avoir le droit à l'erreur avant de monter d'environnement. Euh… Voilà donc c'est un peu une grosse d'implémentation et après il y a une phase de test, de validation. Pour vérifier que tout ca fonctionne correctement que ca soit techniquement ou conjointement avec le business, donc il y à la fois le test des technologies, des jobs qui sont automatisés et le test des données business qui sont mis à disposition donc validées avec les personnes compétentes que les données sont correctes, soient à jour ect.

A: Ok, euh..

L: Et après il y a une phase de mise en prod, déploiement et maintenance de l'existant. Donc là on met en place la sécurité et l'accès des utilisateurs. Avec tout qui fonctionne au quotidien, il faut mettre en place du monitoring, du lobbying, etc. Même si tout ca fait parti de l architecture initial, il faut prévoir le monitoring de sécurité dedans. Donc toutes ces phases sont pas forcément linéaire l'une après l'autre, c'est des choses qu'on peut faire en parallèle mais je pense que ce sont les gros points qui rentre dans un projet de mise en place d'un data vault. En gros, c'est un projet IT avec une méthodologie de modélisation des données particulières.

A: Ok merci beaucoup pour ta réponse. Vous auriez quelque chose à rajouter badreddine?

B: Oui en fait, euh, je pense que pour un projet data vault, il faut aussi mettre en place, ca fait aussi parti de l'architecture mais le framework data vault c'est un point très important en fait dans le projet. Parce que pour générer toutes les entités data vault, il faut aussi un bon framework qui va faciliter après le travail des développeurs. Et un point que Lucas avait mentionné au début, il faut faire une bonne analyse avant de commencer. Est-ce qu'il faut vraiment mettre en place un projet data vault, parce qu'il y a d'autres solutions pour creer un data warehouse. Il y a plusieurs facons, il y a plusieurs solutions, plusieurs méthodes donc il faut bien analyser si on a besoin de vraiment utiliser data vault. Parce qu'un projet data vault, c'est un projet lourd qui prend vraiment beaucoup de ressources que ce soit en fait, euh, ressources développeurs, humaines ou bien serveurs et logiciels quoi.

A: Ok, d'accord. Maintenant, je vais un peu plus aborder l'architecture en détails avec les points les plus importants qui selon moi devait être aborder en fonction de mon travail effectué. Tout d'abord ce sont pour les business rules que ce soit hard et soft. La facon dont vous les définissez que ce soit pour les types de business hard ou les softs et quels moments vous allez les implémenter et les utiliser dans ce processus et dans votre data warehouse?

L: Donc je ne sais pas si toi ca te parle Badreddine?

B: Non, ca n'est pas très claire.

A: Donc en fait, tout ce qui est les business rules "soft", c'est toutes celles qui sont plutôt définies par le business. On dit que les hard ce sont celles qui sont directement utiliser dans la staging area pour assurer que le type de données qui va être entrer dans le data warehouse soit correcte.

L: D'accord donc on peut dire que soft c'est un peu une contrainte qui émane du business et hard ca va etre plutot une contrainte technique de format, par exemple pour que les données puissent rentrer quoi.

A: Ok donc on va dire que les softs, vous les définissez au début du projet ou alors elles évoluent aussi en fonction du projet et des besoins du business qui vont se voir modifier au cours du temps?

L: D'accord ouais, j'ai compris la question. Je ne sais pas Badreddine, tu veux que je répondes ou tu réponds en premier..

B: Ouais euh, pour les business rules soft, je dirais peut-être que ce sont toutes des demandes du business, toutes les règles, les mesures qu'on va calculer c'est ca, et mettre en place dans le datamart.

A: Oui c'est ca, pour les softs vous avez compris. Et pour ce qui est des hard, donc euh, de la source au loading de packs dans le data warehouse. Comment vous définissez les règles hard en fait?

B: Les règles hard, tout ce qui est les nommages, les règles de nommages de colonnes, les nettoyage de données, c'est ca?

A: Oui, c'est bien ca.

B: Ok et du coup, la question de base, c'est?

A: La question de base c'est quand est-ce que vous allez decider des hard et des soft même si là j'ai compris que les softs ca venait en fonction des besoins du business. Les hard rules, j'aimerais savoir si vous êtes obligés de faire ca au tout début du projet ou alors si c'est aussi evolutif parce que les sources que vont utiliser le data warehouse vont aussi evoluer et euh.. Et s'améliorer avec le temps.

B: Mmh, je dirais que.. Je vais commencer par les hard, tout ce qui est business rules hard en fait ca fait parti de l architecture, donc avant de commencer le projet, on a mis en place toutes les règles de nommages, toutes les règles qu'on va utiliser pour le nettoyage en fait des données. Et après bien sur, au fur et à mesure, on peut changer quelques règles mais on va les mettre en place avant de commencer le projet en fait. Et ca, ca fait parti aussi du framework du data vault, on va le mettre en place avant même de commencer le projet et ca va faciliter le travail après. Pour les business rules

soft, ca aussi ca dépend les demande business de préférence on va les recevoir avant de commencer le projet. Et après, ca dépend la facon et les méthodes de travail, on peut les modifier aussi au fur et à mesure mais de préférence on doit les recevoir avec l analyse aussi.

A: Ok, vous auriez quelques choses à rajouter Lucas?

L: Oui, alors pour les hard rules et les soft rules aussi, je ne dirais pas que c'est avant de commencer le projet, parce qu en soit la conception c'est une partie à part entière du projet, mais en tout cas c'est avant de commencer l'implémentation.

B: En fait, quand j ai dit le projet, je voulais dire le developpement, avant de commencer le dev.

A: Merci pour la précision.

L: Effectivement dans la phase de création de l'architecture, de conception. Tu as la mise en place d une gouvernance technique qui va être la mise en place de tes hard rules. Ca va être un ensemble de règles et de bonnes pratiques que les développeurs vont devoir respecter pour avoir une ligne de conduite et un framework unifié et standardisé pour la totalité du data warehouse. Donc ca veut dire que tous les types de données par exemple vont etre définis donc on va restreindre tous les types qu'on peut utiliser. Pour définir la règle des nettoyages des gens, on va dire est-ce qu'on veut obligatoirement retirer les blancs. Est-ce qu'on veut encoder toutes les chaines en une seule suite, etc. quelles sont les règles d'encodages et de nettoyage pour les données qu'on va stocker. Comme ca tous les développement qui vont suivre, vont suivre ces règles de nettoyages et d'encodage. Après pour les soft rules, évidemment les hard rules peuvent évoluer aussi hein dans n'importe quel projet, on se rend compte de chose au fur et à mesure de l'avancement. On a jamais tout qui est parfait dés le début et ne changera jamais. Donc ca ca peut évoluer aussi, on peut évidemment adapter selon comment les procédures sont construites pour permettre l'évolution et le changement des règles de nettoyage par la suite. Et pareil pour les soft rules et encore plus d ailleurs. C'est qu'il a des demandes qui varient constamment donc les soft rules changent tout le temps, selon les besoins du business, les initiatives, selon les directives du management et selon la manière dont le business fonctionne vraiment sur le terrain au jour le jour. Donc il y a pleins de choses qui font que les règles binaires de calcul, les choses qu'on peut monitorer qu'on regarde et change au quotidien. Et donc les soft rules durant toute la durée du projet. C est là qu'on vient intégré des méthodes agiles pour pouvoir intégrer ce changement constant de besoins et l'intégrer dans le développement.

A: Merci beaucoup, je pense que pour ce point là c'est assez et vos réponses sont assez claires. Du coup, maintenant j'aimerais bien parler aussi de la staging area et du flux ETL pour savoir si toute l'équipe de votre projet est impliqué ou pas sur la staging area? Comment ca se déroule? Et aussi le programme que vous utiliser pour le flux ETL?

L: Oui, aussi pour être sur aussi. J'ai remarqué qu'il y avait plusieurs personnes qui avaient des définitions différentes de ce qu'était la staging area donc c'est quoi en ton sens la staging area?

A: Moi la staging area, c'est quand on vient chercher les données des sources extérieurs au data warehouse et c'est là qu'on va effectuer l'ETL, le nettoyage de données..

L: Ok, donc pour toi le nettoyage de données ca fait parti de la staging area.

A: Oui c'est ca.

L: La staging area, c'est une partie que tu vois comme à l'intérieur de ton data warehouse?

A: Euh oui c'est ca. Enfin non, désolé, je me suis mal exprimé. Vu que ca respecte le modèle en 3 couches d'Inmon. J'ai ma staging area, qui est ma première couche, j'ai ma couche data warehouse et seulement après ma couche datamart.

L: D'accord ok.

B: Donc là où on stocke les données brutes, c'est ca?

A: Oui c'est ca.

L: Brutes donc on les stocke plus le nettoyage qui vient par dessus.

A: Oui:

L: D'accord. Et du coup, j'ai oublié le reste de la question. Donc la staging area?

A: M'expliquer comment ca se déroule au sein de l'équipe, le programme utilisé et est-ce que tous les membres de l'equipe sont impliqués ou bien seulement une partie de l'équipe?

L: Pour mon cas, chez Dieteren, faut bien visualiser le flux, donc les systèmes sources, un outils d'extraction qui vient chercher les données dans ces systèmes. Un outil de transfert qui va transférer ces données extraites jusqu'à un endroit qui rentre dans la gouvernance de notre équipe BI qui permet de récuperer les données et de les ingérer dans la staging et après il y a la phase de nettoyage et d'empilement de données. En tout cas ici c'est le cas, en général c'est pareil sauf que tu as peut-être un ou deux steps en moi. Ca dépend quel équipe gère quelle partie, en tout cas dans mon équipe BI tout le monde est capable de créer les steps de staging, de sourcer les nouvelles données depuis un nouveau système source. Donc en fait, dans notre cas, c'est nous même l'équipe qui geront le data vault qui allont dans la plus part des cas extraire les données dans les systèmes sources pour les ramener chez nous parce qu il n y a pas forcément en face une équipe capable de le faire. Dans pas mal de cas, c'est nous qui mettont en place un programme d extraction, comme via des outils comme IBM data manager. La c'est un ETL, de la suite IBM chronos mais qui est veillisant mais on avait ca pour extraire les données des B2, un mainframe, des bases de données qui utilisent des technologies qui ont deja pas mal d'ancienneté mais qui sont toujours utilisé dans l'entreprise. Où alors on passe par un etl custome en powershell donc là on va nous même développer un script selon la technologie du système source extraire des fichiers de la maniere dont c'est possible de le faire avec ce système pour rattrapier ces fichiers jusqu'à chez nous pour les ingérer dans la staging. Où alors, on va utiliser maintenant de plus en plus des technologies cloud donc là on va pour le cas de dieteren utiliser le cloud Azure et se baser sur du data factory pour grace à un connecteur aller se connecter au système

source et périodiquement extraire les données. Dans d'autres systèmes plus modernes, il y a souvent une équipe dédié et donc il va falloir collaborer avec cette equipe pour savoir comment on fait pour que les données de ce système puisse sortir et être injecter dans ce data vault. Donc là, ca va être souvent la mise en place d'un flux FTP, interroger une API pour extraire les données et les envoyer par SFTP sur un serveur. Il y a une équipe ESB qui est chargé de faire les connexions entre les systèmes et donc eux ils font faire les transerts de fichier entre les systèmes sources et le data warehouse en utilisant des protocoles pour le transfert et la reception du fichier. Et donc quand nous on a les fichiers, on peut à l'aide d'un package ETL d'ingestion fait avec cognos data manager ou avec un script powershell maison ou avec data factory. Donc en tout cas, chez nous la staging, ce sont des tables en database qui contiennent de la donnée brut qu'on recoit de la source mais qui sont passer dans tout ces transerts. Et derriere, on vient commencer à faire le cleaning de ces données. Dans la staging, on essaye de garder en fait de faire en sorte qu'on injecte pas de mauvaises choses. Le but, c'est que le type des données, il n'y ai pas de restriction au niveau de la staging, pour que ca soit le plus proche possible du type qu'on recoit du système source, pour que ca soit ingérer comme il faut dans la staging et c'est dans le step suivant qu'on commence à faire du cleaning. Donc on va caster les champs vers les types targets qu'on a définit pour notre data vault, on va retirer les blancs, etc comme ca on va commencer le nettoyage pour standardiser nos champs selon la gouvernance que l'on a mis en place. Ca si ca pète entre guillemets, ca pète uniquement après cette staging et on a les données brutes prêtes et qui sont toujours là en cas d erreur. Je ne sais pas si ca répond à ta question.

A: Si si ca répond, c'est même plus que complet donc merci pour ta réponse. Du coup, chez partena, ca se passe de la même manière ou ca se diférencie un peu?

B: Chez partena, c'est un peu différent en fait. Pour partena, il y a deux équipes. La première équipe elle fait toujours l'extraction des données et le stockage des données brutes dans la stage area, donc eux ils font seulement le extract and load sans rien modifier. Et en général, on utilise le delta, donc à chaque fois, il charge les nouvelles records, chaque nuit on fait ca. On stocke seulement ces nouvelles lignes dans la stage area sans rien modifier, sans faire le nettoyage, … Donc c'est les données brutes, et la deuxième équipe elle travaille sur la partie extraction de données et nettoyage depuis le stage area vers le data vault. Donc là où on utilise le framework pour changer les données dans data vault en faisant le nettoyage, les règles de nommages, … Voilà tout ce qu'il faut pour charger les données, en ajoutant les metadata, dans le data vault. Donc chez partena il y a deux équipes, une équipe qui travaille sur la stage area, ils font seulement extract and load sans transformation. La deuxième équipe, elle travaille sur le data vault, là où ils utilisent le framework pour creer toutes les entités data vault. Et pour les outils en fait, on utilise SSIS.

A: Ok.

L: Pardon, j'ai peut-être oublié de donner les outils également. Donc je t avais dit il y a effectivement, descript powershel, pour l extraction et gestion de fichier, Azure Data Factory, IBM Cognos Data Manager. Et pour la staging en elle même, ce sont des tables dans un SQL server, ils y a des serveurs sql unpremice et on migre maintenant vers des sql server dans le cloud et Azure.

A: D'accord, ok parfait. Maintenant, j'aimerais bien aborder tout ce qui est datamarts et savoir le nombre de datamarts que vous avez chacun dans chacun de vos projets. Mais aussi le procédé qui a précédé le fait d'avoir autant de datamarts. Et comment le nombre de datamarts est défini, si c'est en fonction du nombre de différents secteurs de la société, le client, pour lequel vous travaillez? Ou bien si ca dépend d'autres facteurs?

L: Oui alors pour ma mission chez Dieteren, le data vault existait déjà bien avant que je commence ma mission. Il y a un peu près 10 ans. Mais il y a en gros, un datamart par sujet business, dieteren a à la fois la vente des véhicules neufs, la vente des véhicules d'occasion, il y a l'après vente donc tout ce qui est entretien, réparation carroserie, mécanique, commandes de pièces, etc. Il y a les leasings, tout ce qui est aspect financier, contrat de leasing. Les immatriculations donc voir quels sont les parts de marcher par rapport aux immatriculations, etc. donc il y a plusieurs departements comme ca, plusieurs business. Cela évolue en fonction des années mais globablement par sujet, il y a un datamart.

A: Est-ce que vous voyiez des avantages ou inconvénients à l'utilisation particulière de datamarts par rapport aux méthodes traditionnelles?

L: C'est pas facile d'y répondre pour moi car la seule expérience pour moi que j'ai eu c'est avec cette stack data vault, donc je n'ai pas une conception très clair de comment fonctionnerait autrement. Parce que pour moi data vault c'est juste une face de modéliser un data warehouse centralisé où l'on collecte toutes les données ensemble d'une manière agile ou entre guillemets, c'est très facile d'ajouter de nouvelles données et de connecter à d'existantes. C'est juste pour moi une modélisation de ca et par dessus on vient de toute facon construire une logique business qui aurait pu etre construit par dessus n'importe quel autre systeme de modelisation. Donc c'est juste que ca va faciliter ta gestion parce que tu crées juste des nouveaux links,hubs, satelittes pour venir ingérer tes nouveaux systèmes mais après toute la complexité d'implémenter les règles business se fait par dessus. T'as un peu la simplicité du systématisme de data vault qui fait que tu fais tous le temps des joints de la meme facon pour aller rechercher tes données dedans. Donc une fois que tu as compris la solution data vault, c'est assez facile de creer les query pour obtenir les données, mais toute la difficulté c'est de connecter ton besoin business à la manière d y répondre en allant rechercher les données qu'il faut dans tes sources et donc dans ton data vault. Donc pour moi, je ne suis pas convaincu qu'il y ai un gros avantages à utiliser data vault au niveau des datamarts puisque.. Si ce n'est que tu as deja ces connections qui sont fait de manière systématique mais si non c'est très laborieux car en data vault les données sont splitté en pleins de table. De part le fait qu'il fait séparer les clés business entre elles, ca crée des query très

longues et des fonctions très complexes pour créer le datamart. Voilà avec mon expérience, je ne sais pas dire si c'est mieux ou pas mais j ai pas l'impression que cela change énormément de choses.

A: D'accord, je note, c'est un avis très intéressant et je ne m'attendais à une réponse comme ca donc c'est vraiment chouette. Concernant Partena, ca fonctionne de la même manière ou?

B: Oui chez partena ca se passe aussi comme ca, tous les datamarts sont orientés business. Donc pour chaque business, on a un datamart mais parfois c'est par request. Donc mais le point en fait de data vault, moi j'ai travaillé avant chez IPM. On a travaillé avec une autre architecture, de Kimball. Et maintenant, chez partena, c'est plus Inmon. Donc pour data vault, on doit créé tout le dw qui contient tous les différents business et au-dessus, on doit créer les datamarts, orientés sujet-business. Donc pour moi en faite, le data vault, il prend plus de temps et il ajoute plus de complexité pour créer le datamart.

A: D'accord…

L: J'ai aussi un peu songé, mais c'est vrai que dans certains cas, on applique le data vault plus par automatisme mais parfois ca irait plus vite de skiper la couche data vault, en construisant directement le datamart. Et autre truc, il y a quand même une chose commune entre tous les sujets business, les départements. Par exemple, chez Dieteren, c'est la liste de concessionaires, elle est la même pour toute la boite. Il y aussi de plus en plus une unification des données clients, on a ce que l'on appelle les golden customers qui sont considéré comme la seule vérité pour toute l'emprise.Donc de plus en plus on a des choses commune, on va donc construire une couche que l'on appelle BDV, business domain view, dans laquelle on peut mettre les données communes.En fait, on a ca sous forme de fonctions logiques dans lesquelles on a implémenté les régles business et après on peut les réutiliser dans tous les datamarts. C'est un peu une couche intermédiaire où l'on a regroupé la logique commune pour éviter justement ces sillots.

B: En fait, chez Partena, on a une couche similaire mais nous on appelle ca business vault. Là, on fait toutes les régles business, les fonctions, les calcules. Du coup, on peut réutiliser ca plusieurs fois dans différents datamarts.

A: Dans le projet partena, je sais que sur base de la théroie, on a trois types de vault. On a la metric, l'operational et le business vault. Est-ce que vous utilisez autre chose que le business vault?

B: Non, on a uniquement la couche business vault.

A: Et du coté de chez Dieteren?

L: Moi, je ne suis pas familié avec ces concepts. Donc on a une couche BDV suivi d'une couche datamart.

A: Ok, du coup vous avez une couche similaire à la business vault?

L: Oui mais elle est purement virtuellement. Enfin non pas vraiment, on essaye de ne pas copier les données que ce n'est pas nécessaire. Quand c'est possible d'implémenter une sorte de view, de fonction alors on le fait. Mais effectivement, cela est une couche business.

A: Merci pour ces réponses concernant l'architecture. Maintenant ce sera plus des questions sur la méthodologie. Donc il y a trois grandes familles, on va dire, de méthodes dans data vault. On a la planification & le management, l'exécution et la revue & améliorations. Tout d'abord pour entamer le sujet de planification et management, si je vous parle de Capability Maturity Model Integration et la méthode scrum, est-ce que cela vous parle?

B: Oui chez partena, on utilise scrum. Pour tout ce qui est planification, on travaille avec scrum.

A: Justement en parlant de scrum, quand est- il de la technologie utilisée, les products backlogs, les user-story et la facon dont vous faites la hierarchisation des user-stories..

B: Chez Partena, on utilise TFS pour tout ce qui est la gestion du sprint. Les backlogs chez partena, on a les epic, les features dans chaque features, il y a plusieurs user-story et pour finir plusieurs tasks. Donc avant de commencer le sprint, on fait un sprint planning où l'on checke les capacités de chaque développeur. Après, on va voir qui va travailler sur quelle user-story mais ca dépend aussi des priorités. En général, le sprint chez partena, c'est 2 semaines.

A: Pour ce qui est des user-stories, c'est combien de user-stories par sprints?

B: Cela dépend des tâches de chacun, mais chez partena en moyenne c'est 16 par sprint. Mais ca dépend des disponibilités de l'équipe.

A: Est-ce qu'il y a des avantages et inconvénients justement sur cette méthode scrum, sur base de votre expérience passée dans d'autres entreprises? Est-ce que c'est justement une solution adapté à un projet aussi imposant qu'un projet data vault?

B: En fait que ca soit pour un projet data vault, ou la bi en générale. Pour moi, c'est très important d'utiliser les méthodes agiles comme scrum. Du coup, comme on a parlé les règles business, ca peut changer dans le temps. Donc imagines, par exemple, sans scrum, tu travailles sur un projet aussi imposant que data vault et après deux mois les besoins du business changent, donc tu as perdu 2 mois de travail. Alors qu'avec scrum, à chaque fin de sprint, tu fais une petite démo au business et ainsi tu sais directement avoir des retours et t'adapter directement aux changements. Avec un petit review, pour voir ce qui a bien fonctionné et ce qui n'a pas fonctionné. Du coup, toutes les deux semaines, on a les feedbacks du business, du coup c'est un échange interactif. À chaque fois, le business est courant d'où on en est.

A: Et du côté de Dieteren, ca se passe comment?

L: Pour chez nous, c'est le meme cas que dans beaucoup d'entreprises, et je pense que c'est le cas de chez partena. Juste un petit truc, chez Dieteren, ce n'est pas vraiment la data vault 2.0, on a commencé avec le 1.0 et on a juste commencé à utiliser les méthodologies pour le faire évoluer. Nous aussi, on a

l'agilité et on fonctionne en scrum. Il y a pleins de granularités différentes dans les demandes qui arrivent donc il faut souvent une phase d'analyse pour formuler des vraies user-stories. On utilise JIRA pour le ticketing. Je pense que l'utilisation de scrum est bonne dans une équipe de développement et globalement elle nous permet de mieux définir les tâches. Pour l'instant ca fonctionne bien comme ca, donc on continue.

A: Donc du coup, vos deux projets utilisent scrum mais si je vous parle de mesuer la capacité et la maturité du projet, ca ne vous parle pas beaucoup?

L: Non

B: Non

A: Ok ce n'est pas grave, c'est que vous ne l'utilisez pas donc ca sort du scope de l'interview. Vous avez déjà aborder le sujet sur comment vous échangiez les informations avec le business. Mais dans le modèle waterfall, il y a le unit-test et le système-test. Et j'aimerais savoir si vous appliquez le test d'unité ou meme le test du système dans sa globalité? Et comment vous réalisez ca au sein de votre équipe.

L: Pour le testing, je pense qu'il faut que ca soit confier à des gens qui ont été sensibiliser à l'importance de ca et de mettre en place des automatismes pour détecter la qualité. Dans une équipe bi, il y a souvent deux types de personnages, ceux qui testent régulièrement et ceux qui passe cette étape.

A: Et pour Partena, ca se passe comment?

B: On a mis en place un framework pour le unit- test mais il est encore en phase de dev. Chez partena, chaque développeur fait ses tests en dev avec avoir effectué ses tâches. Si l'environnement test accepte les unit test.

A: On va entamer la dernière partie qui parle de la revue et des améliorations. On va parler des modèles SIX-Sigma/TQM. J'aimerais savoir si chez vous, vous essayez d'inclure tout le monde dans ce processus de feedback? Est-ce que vous avez inclus le business pour répondre plus facilement à leur besoin.

B: Je pense que l'implication des business, c'est important. Car si le business est impliqué, c'est un point important dans la qualité du projet final. L'implication du business est un point très important pour la qualité final du produit.

A: Et chez Dieteren?

L: Oui, pareil. Avant de faire rentrer une story dans le sprint, on s'assure de la disponibilité d'au moins une personne business, Comme ca, on est sur qu'en cas de questions, validations, on a un contact. Chez nous, après l'environnement dev, tout est poussé en environnement quality acceptance, où le business confirme son accord à pousser en prod.

A: Pour cette section, on a finit, je vous remercie pour vos réponses. On va entamer le dernier point avec l'une des plus petites questions concernant la modélisation, on a les hubs, les links, satellites. Et j'aimerais avoir votre avis sur les avantages et inconvénients de ces 3 composantes.

L: Je pense que c'est à double tranchant. Je pense que Data Vault est génial si tu arrives à mettre en place un framework automatisé qui va t'aider à le générer automatiquement. Car le data vault a une forme très répétitive autant que tu es l'outil adapté car data vault splitent en pleins de tables. Le fait d'avoir pleins de hubs, de links et de satellites permet de rester agile puisqu'en cas de modifications nécessaires, il suffit de relier le nouveau composant dans l'architecture actuel.

A: Et chez partena, vous en pensez quoi?

B: L'avantage du data vault, c'est une DB où tu as centralisé toutes les données de centraliser mais pour cela il faut avoir un bon framework pour pouvoir générer les nouveaux links, hubs et satellites qui arrivent avec une nouvelles sources.

L: Cela t'oblige que pour chaque champs, tu dois te poser et savoir où il est mapper dans le data vault.

A: Ok, du coup, je pense qu'on a fait le tour de la question et ma question de conclusion viens d'être aborder. Donc est-ce que vous auriez enventuellement des points que vous voudriez aborder ou bien d'éventuelles critiques sur la facon dont l'entretien c'est déroulé?

L: Là tout de suite, non.

B: Non.

A: Ca a été assez complet, on est resté plus longtemps que prévu donc c'est chouette. Du coup, je vous remercie pour votre temps et les réponses très complètes que vous m'avez apportez. Et j'espère que je pourrais vous rencontacter après pour vous dire que ce mémoire aura été une réussite.

L: Pas de problème, je l'espere pour toi et bonne chance.

B: Bonne chance Amaury et bon courage.

L: Bon courage.

Légende:

A: Amaury Mellaerts (Interviewer)

L: Lucas Vanoverberghe (Akabi/Dieteren)

B: Badreddine Ben Aissa (Akabi/Partena)

Annexe 4: Tableau d'analyse thématique des interviews

Thèmes	Lucas Vanoverberghe	Badreddine Ben Aissa
• **Architecture** **1. Hard/Soft régles** **2. Staging Area** **3. Datamart Area** **4. 3 "Vault" Optional**	1. Utilisation des hard et soft business rules dans son projet. (Hard = gouvernance technique mise en place avant l'implémentation) (Soft= demander par le business). 2. Récupérer les données de manière brute pour les stocker avant nettoyage et envoie à la data warehouse (Utilisation de IBM Cognos pour le flux ETL). Requiert que chaque membre soit impliqué. 3. Un datamart par sujet business (orienté-business). Pas d'avantages et inconvénients. 4. Utilisation d'un concept similaire au "business vault", nommé BDV (Business Domain View). Il occupe une place similaire dans leur architecture.	1. Utilisation des hard et soft business rules dans son projet. (Hard = partie de l'architecture, règle de changement de type et de nommage) (Soft= émane du business). 2. Extraire et stocker les données brutes. Une partie de l'équipe est uniquement dédié à cette tâche. (Utilisation de SSIS). 3. Datamarts orientés-business 4. Utilisation d'un business vault
• **Méthodologie** **1. CMMI** **2. SCRUM** **3. Modèle en cascade(unit-test/system-test)** **4. Six Sigma/TQM**	1. Non-utilisé 2. Implémenté dans l'équipe. Sprint de 2 semaines, et prioriser des user-stories. Application complète de ce modèle. Designation de points de contact dans le business . (JIRA) 3. Utilisation décroissante des unit-tests. Emploi	1. Non-utilisé 2. Implémenté dans l'équipe. Sprint de 2 semaines, et hiérarchisation des user-stories. Application complète de ce modèle. Agilité importante dans le suivi du projet par le business (TFS). Offre la possibilité de réagir

	d'un environnement dit QA (Quality acceptance), en attente de confirmation de business avant la production. Phase d'analyse primordiale due aux différences de granularités. Conscience de l'importance de ces tests. 4. Implication pas autant pousser que la théorie.	plus rapidement au changement business. 3. Implémentation d'un framework de test. Chaque individu est en charge de tester ses modules. 4. Échange interactif avec le business. Démo à chaque fin de sprint des avancements actuels du projet.
• **Modélisation**	1. Modélisation très répétitive, utilisation d'un framework automatisé (double-tranchant)	1. DB centralisé dont la réussite dépend de la capacité d'adaptation du framework qui génère les links, hubs, satellites.