MASTER EN INGÉNIEUR DE GESTION À FINALITÉ SPÉCIALISÉE EN DATA SCIENCE

Conception d'une solution de Business Intelligence
étude d'un cas réel

Lopez Ruiz, Lucas

Award date:
2019

Awarding institution:
Universite de Namur

Conception d'une solution de Business Intelligence :

étude d'un cas réel

Lucas LOPEZ RUIZ

Directeur : Prof. I. JURETA

Mémoire présenté
en vue de l'obtention du titre de
Master 120 en Ingénieur de gestion,
à finalité spécialisée en Data Science

ANNEE ACADEMIQUE 2018-2019

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

Avant toute chose, je tiens à remercier les personnes qui ont facilité la mise en œuvre de ce mémoire.

Merci au Dr Ivan Jureta, promoteur principal de ce mémoire, de m'avoir transmis les exigences et l'ensemble de données propres à l'entreprise rendant mon étude de cas possible. Je le remercie également de m'avoir enseigné des concepts-clés au travers de son cours d'*Ingénierie des exigences et d'analyse des décisions* donné à l'Université de Namur. Son expérience de chercheur qualifié du FNRS (Fonds National de la Recherche Scientifique) m'a permis d'acquérir une connaissance appliquée de certains des concepts abordés dans ce mémoire.

Merci également au Dr Isabelle Linden, co-promotrice de ce mémoire, pour les outils théoriques et pratiques qu'elle m'a fournis au travers de son cours de *Business Intelligence* donné à l'Université de Namur. Son approche consistant à illustrer chacun des éléments théoriques à l'aide d'une étude de cas m'a permis d'avoir une compréhension plus approfondie et appliquée de la matière.

Enfin, je remercie également le Dr Caroline Herssens, le Dr Stéphane Faulkner et le Dr Joseph Gillain (collaborateurs de l'entreprise BStorm de Louvain-la-Neuve), qui m'ont amené à approfondir mes connaissances en matière de Business Intelligence grâce aux divers projets réalisés lors de mon stage en entreprise.

Table des matières

Introduction

Dans un monde où la quantité de données à disposition croît sans cesse et où les acteurs économiques ont des exigences de plus en plus nombreuses, les technologies de l'information et de la communication se sont multipliées. Ce mémoire se penche sur la Business Intelligence (BI ci-après), qui a pour objectif principal de soutenir les choix stratégiques des décideurs au travers d'indicateurs de performance et de divers outils d'analyse.

Il est important de préciser que ce travail s'adresse à des personnes ayant déjà une connaissance de la BI. Les différents modèles, théories et méthodes qui seront présentés sont néanmoins régulièrement remis dans leur contexte afin de rappeler ou expliquer au lecteur (à la lectrice) certains des avantages, inconvénients et caractéristiques qui leur sont propres. Un nombre important d'éléments lui seront donc sans doute déjà familiers, mais il reste fondamental de préciser les aspects théoriques sur lesquels se base l'étude de cas qui suit.

On accordera ici une grande importance aux concepts liés aux visualisations déployées au sein des solutions BI. La revue de la littérature mettra ainsi l'accent sur des aspects qu'il est nécessaire de prendre en compte lorsque l'on modélise ces visualisations : critères pour scinder les informations entre les différents dashboards, densité visuelle (proportion de pixels à utiliser), utilisation d'éléments familiers tels que des pictogrammes, ornements visuels, aspects naturels (bords arrondis, structures semblables aux éléments de la vie quotidienne…), types de figures (histogrammes, tableaux…), éléments prétraités de manière inconsciente, éléments destinés aux utilisateurs ayant des besoins particuliers (ex. : daltoniens)…

Ces aspects peuvent parfois sembler anodins mais ne le sont en fait pas du tout. L'interaction qui s'établit entre l'utilisateur et les visualisations est primordiale, et la façon dont il comprend et retient l'information est directement liée à bon nombre des critères susmentionnés.

Notons que le choix de ce sujet a été motivé par le fait que, bien que la littérature qui y est liée soit relativement développée en matière de reporting opérationnel (et fasse l'objet de nombreux débats), il n'en va pas de même pour la littérature liée à la BI. La recherche académique sur le sujet est peu abondante, en dépit du fait qu'elle présente un grand intérêt pour l'industrie. Il sera donc notamment question d'analyser certains des critères qui font débat

en matière de visualisations opérationnelles, en s'assurant qu'ils soient compatibles avec la nature stratégique de la BI.

Un cas d'étude permettra ensuite non seulement de mettre en application les aspects détaillés dans la revue de la littérature, mais également de se pencher sur d'autres aspects liés aux visualisations tels que les indicateurs clés de performance à afficher sur les interfaces.

Aborder la problématique des visualisations n'implique évidemment pas, quoi qu'il en soit, de négliger les autres aspects essentiels que sont l'ingénierie des exigences et le traitement de données.

Bien que l'ingénierie des exigences soit abordée dans la revue de la littérature, il est important de préciser que le choix fait ici est celui d'analyser les exigences et le traitement de données de manière principalement appliquée, c'est-à-dire directement au travers de l'étude de cas. Ce mémoire n'a en effet pas pour objectif de confronter les différents points de vue et théories liés à ces problématiques, et c'est pourquoi il a été décidé de considérer un paradigme de BI bien spécifique et de l'appliquer directement au cas d'étude.

En résumé, il sera question de détailler les aspects liés à l'interaction avec le client tant dans la revue de la littérature que dans l'analyse de cas : les visualisations et l'interaction interfaces-utilisateur occuperont la majeure partie de cette revue de la littérature et une grande partie de l'analyse de cas ; l'ingénierie des exigences sera abordée dans la revue de la littérature et occupera, elle aussi, une grande partie de l'analyse de cas. Les aspects davantage liés au traitement des données seront, eux, explicités en détail au travers du cas d'étude uniquement.

Notons que le cas pratique se base sur une problématique réelle. Elle a trait au domaine du transport routier de marchandises et, plus précisément, aux *freight brokers*. Comme précisé par la suite, ces derniers jouent le rôle d'intermédiaire entre des expéditeurs ayant une cargaison à envoyer et des transporteurs à la recherche de biens à véhiculer.

Ce mémoire se décline en cinq chapitres : mise en contexte des processus de la BI et revue de la littérature (chapitre 1), présentation détaillée de l'étude de cas (chapitre 2), ingénierie des exigences et modélisation des entrepôts de données (chapitre 3), extraction, transformation et chargement des données et modélisation multidimensionnelle (chapitre 4) et, enfin, déploiement des visualisations (chapitre 5).

Chapitre 1 : mise en contexte et revue de la littérature

1.1 Mise en contexte des processus de la Business Intelligence

Bien que ce mémoire soit destiné à des personnes ayant déjà des connaissances en Business Intelligence, il est utile, dans un premier temps, de mentionner quelques-uns des fondements sur lesquels se basent la revue de la littérature et l'analyse qui suivent. L'objectif n'est pas ici de s'étendre sur le sujet, mais plutôt de **rappeler brièvement quelques définitions de base** et de **mettre en évidence certains des choix et des paradigmes qui sont adoptés par la suite**.

Le terme Business Intelligence, ou intelligence décisionnelle, fait référence au processus qui vise à transformer les données opérationnelles à disposition en informations et connaissances utiles à la prise de décisions (Golfarelli, Rizzi et Cella, 2004). Il s'agit d'une prise de décisions principalement stratégique, voire tactique. La BI fait ainsi référence aux théories, méthodes, modèles et technologies logicielles qui permettent de collecter, transformer, stocker, structurer et analyser les données opérationnelles (Vaisman et Zimanyi, 2014).

L'analyse qui suit va donc se pencher sur 5 grandes étapes bien connues du déploiement d'une solution BI : analyser les exigences du business et la qualité des données brutes/opérationnelles à disposition, faire des choix en matière de transformation des données, définir des modèles qui puissent incorporer ces données, construire des interfaces et, enfin, analyser le comportement des utilisateurs qui en font usage (Vaisman et Zimanyi, 2014).

Cette analyse vise en grande partie à étudier les clivages observés dans la littérature liée aux visualisations comme, par exemple, ceux qui ont trait à la question de l'efficacité et du caractère mémorisable des interfaces ((Bendoly, 2016) ; (Janes, Succi et Sillitti, 2013)). On aura néanmoins compris que l'on ne peut envisager de développer des visualisations sans vouer une attention toute particulière à l'analyse des exigences des utilisateurs et au traitement de données.

Notons que le paradigme d'architecture BI qui va être adopté par la suite est celui de la ***two-layer architecture***. Ce dernier parvient à mettre en œuvre les grandes étapes mentionnées ci-dessus au travers de la collection de sources de données hétérogènes (et des exigences qui y

sont liées), de l'extraction-transformation-chargement de ces données dans les structures BI (*Extraction-Transform-Load* ou ETL), de la conception du *data warehouse* (DW) et des *data marts* (DM) et de l'étude des interfaces d'analyse (Vaisman et Zimanyi, 2014). Certains des avantages et inconvénients de ce type d'architecture seront mentionnés par la suite.

La *figure* 1.1 construite ci-dessous récapitule les étapes propres à l'architecture à double couche utilisée par la suite : la première couche est celle des sources et la deuxième celle du data warehouse.

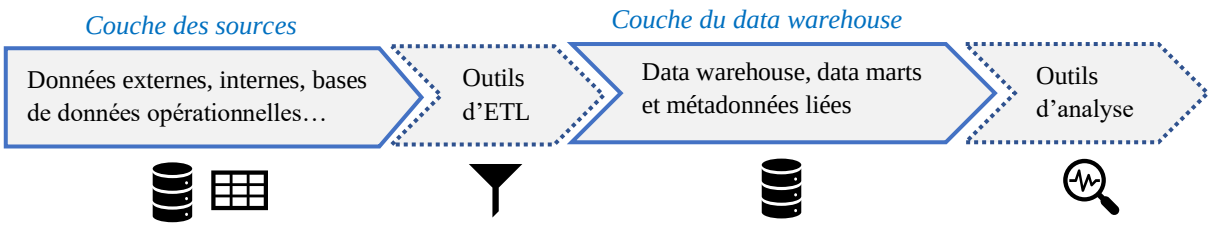


Figure 1.1 : l'architecture **double couche** (architecture choisie)

La mise en œuvre de l'architecture double couche s'accompagne d'une ingénierie des exigences consciencieuse. Lorsqu'il s'agit d'analyser les exigences, de très nombreux auteurs et acteurs du monde professionnel avancent très souvent deux approches : **l'approche traditionnelle et l'approche « agile »**.

La comparaison de ces deux approches pourrait faire l'objet de longues discussions, mais attachons-nous ici à souligner que l'étude de cas qui suit repose sur une méthode traditionnelle, séquentielle et linéaire, et explicitons quelques-uns des avantages et inconvénients de cette méthode. Le cas d'étude ne laisse en fait pas réellement d'autre choix que de recourir à cette méthode. Sa nature statique (les exigences et les données ne sont ici transmises qu'à une seule reprise) ne se prête en effet pas aux processus agiles et à leur volonté de s'adapter aux changements ((Lindvall et al., 2002) ; (Highsmith, 2002) ; (Beck et al., 2001) ; (Mirza et Datta, 2019)).

Pour mieux appréhender la suite du travail, insistons sur le fait que l'inconvénient principal de la méthode traditionnelle est globalement qu'elle est peu adaptée aux changements.

Les fournisseurs de solutions qui y ont recours ont souvent tendance à vouloir recueillir, dès le départ, un maximum d'informations auprès du client ((Mirza et Datta, 2019) ; (Cho, 2008)). Les analystes cherchent en effet à écrire d'emblée une documentation des plus complètes afin de permettre aux développeurs d'avoir « toutes » les cartes en main pour se lancer dans les phases suivantes du projet. Ceci a par exemple l'avantage de rassurer ces derniers et de les conforter dans leurs choix (Cho, 2008). L'idée est en théorie d'avancer étape par étape (définition des besoins, développement, tests, etc.) en évitant au maximum de devoir faire marche arrière. Dans ce cas-ci, on parle souvent d'une méthode en « cascade », appelée méthode *waterfall* en anglais (Mirza et Datta, 2019).

Le problème lié à ce type de fonctionnement est que la mise en place d'une documentation complète des exigences demande souvent un temps considérable et s'avère très complexe, principalement parce que le client n'est généralement pas en mesure de fournir toutes ses exigences dès le départ ((Beck et al., 2001) ; (Cho, 2008) ; (Mirza et Datta, 2019) ; (Williams et Cockburn, 2003)).

L'approche séquentielle et le manque de souplesse qui caractérisent les méthodes traditionnelles impliquent donc une résistance significative aux changements. Pour tenter de pallier ce problème, les méthodes agiles ont pour objectif de satisfaire le client en interagissant régulièrement avec lui, en précisant ses besoins tout au long du projet et en lui exposant l'avancée du projet ((Lindvall et al., 2002) ; (Beck et al., 2001) ; (Highsmith, 2002)). Il s'agirait donc de procéder par « itérations », ce qui n'est pas possible dans le cadre du cas d'étude qui nous occupe.

Ces propos mériteraient néanmoins d'être approfondis dans un autre contexte que celui de ce mémoire, notamment parce que les constats dressés ici s'avèrent plus nuancés dans la pratique. Cela étant dit, rappelons que d'autres aspects essentiels ayant trait à l'ingénierie des exigences sont étudiés en profondeur par la suite.

1.2 LE LIEN ENTRE EXIGENCES, DONNÉES ET VISUALISATIONS

La suite de ce chapitre se penche sur la littérature relative à la conception des visualisations. Notons que les constats dressés ici sont bien entendu liés à la BI, mais sont également régulièrement applicables à d'autres domaines de l'IT qui ont pour but de fournir des visualisations à l'utilisateur final, comme le reporting opérationnel par exemple. Cette section est capitale, puisqu'elle détaille de nombreux critères nécessaires à l'analyse effectuée ultérieurement.

Une illustration vaut mille mots. Une interface vaut mille illustrations (Shneiderman, 2003).

Shukla et Dhir (2016) expliquent que les visualisations permettent de présenter les données de manière intelligible à l'utilisateur final. Ces derniers rappellent également que la BI permet une compréhension et une interprétation des données des plus rapides et des plus utiles à la prise de décisions, et ce, malgré une quantité parfois très importante de données. L'utilisateur peut ainsi être en mesure d'identifier visuellement d'éventuelles opportunités, menaces, forces et faiblesses inhérentes à son business et d'adapter les stratégies en conséquence ((Shukla et Dhir, 2016) ; (Few, 2006)). Les interfaces et les rapports doivent donc permettre de communiquer ce qui doit l'être en un minimum de temps ((Bendoly, 2016) ; (Janes, Succi et Sillitti, 2013)).

Contrairement à certaines idées reçues, concevoir des visualisations est loin d'être chose aisée et il est donc important de souligner certains des aspects auxquels il faut prêter attention lorsque l'on se lance dans une telle tâche. Notons par exemple que Few (2006) et Bendoly (2016) expliquent que les visualisations peuvent évidemment s'avérer être une mine d'informations lorsqu'elles sont correctement mises à profit mais que, dans la pratique, les développeurs ont souvent tendance à concevoir leurs visualisations de manière à ce qu'elles incorporent autant de données que possible et qu'elles démontrent des compétences techniques et graphiques avancées qui impressionnent le client : les visualisations sont parfois plus « esthétiques » qu'utiles, ergonomiques et stratégiques. Les acteurs de la BI mettraient d'ailleurs régulièrement la charrue avant les bœufs en choisissant leurs outils de visualisation avant même d'avoir appréhendé la logique des données ((Bendoly, 2016) ; (Zhu, 2007)).

Il ne peut pas exister de définition universelle de ce que serait une interface idéale (ex. : archétypes de tableaux de bord/dashboards) et ce, pour la simple raison que chaque projet a un contexte et des exigences qui lui sont propres. Il n'en demeure pas moins que certains constats peuvent être dressés ((Zhu, 2007) ; (Bendoly, 2016) ; (Froese et Tory, 2016)). Cherchons dès lors à établir **le lien qui existe entre exigences, données et visualisations**.

Nombre d'auteurs veillent à rappeler que, pour choisir les techniques de visualisation les plus appropriées, le développeur doit avant tout capturer la logique des données du projet. Janes, Succi et Sillitti (2013), Wattenberg et Fisher (2004) et Tufte (2001) soulignent ainsi que la structure des visualisations doit refléter celle des données. Ce constat peut sembler évident de prime abord mais il l'est apparemment moins dans la pratique, pour des raisons telles que le manque de communication entre les parties prenantes, le manque de ressources et de temps à disposition et le peu de connaissances des clients au regard des systèmes de Business Intelligence (Froese et Tory, 2016).

Janes, Succi et Sillitti (2013), dans la lignée de chercheurs tels que Basili et Rombach (1994, 2010), se sont penchés sur le modèle appelé *Goal-Question-Measurement* (GQM). Ce modèle et l'analyse stratégique qui en découle ont pour objectif de schématiser une structure de données des plus optimales pour un projet en cours et de la mettre en relation avec une structure de visualisations qui la traduirait au mieux. Le choix est fait ici de présenter ce modèle afin de pouvoir l'utiliser lors de l'analyse des exigences fonctionnelles de BrokerRoad.

Pour déterminer les données à utiliser dans les visualisations, le Goal-Question-Measurement fait appel à des techniques génériques d'ingénierie des exigences et se base, comme son nom l'indique, sur 3 niveaux d'analyse ((Janes, Succi et Sillitti, 2013) ; (Basili et al., 2010)) :

1. Objectif *(niveau conceptuel)* : quel est l'objet d'étude ?
2. Questions *(niveau opérationnel)* : quels sont les aspects de l'objet d'étude qui permettent de déterminer si l'objectif est atteint ? Quelles questions se poser ?
3. Mesures *(niveau quantitatif)* : quelles données collecter pour répondre aux questions ?

Si plusieurs objectifs stratégiques, voire tactiques, sont à formuler, il s'agit alors de hiérarchiser ces derniers (Basili et al., 2010) : objectif principal, sous-objectifs, etc.

La figure qui suit a été élaborée ici afin d'illustrer de manière visuelle l'exemple du producteur de vin proposé par Janes, Succi et Sillitti (2013) et démontre, de manière simplifiée, comment exprimer une hiérarchie d'objectifs à deux paliers à l'aide du modèle GQM (l'objectif principal et un des objectifs secondaires) : il s'agit de détailler les objectifs, les questions qu'ils engendrent et les mesures permettant de donner réponse à ces dernières. La structure de données peut alors être mise en parallèle avec une structure de visualisations adéquate (cf. *figure* 1.2).

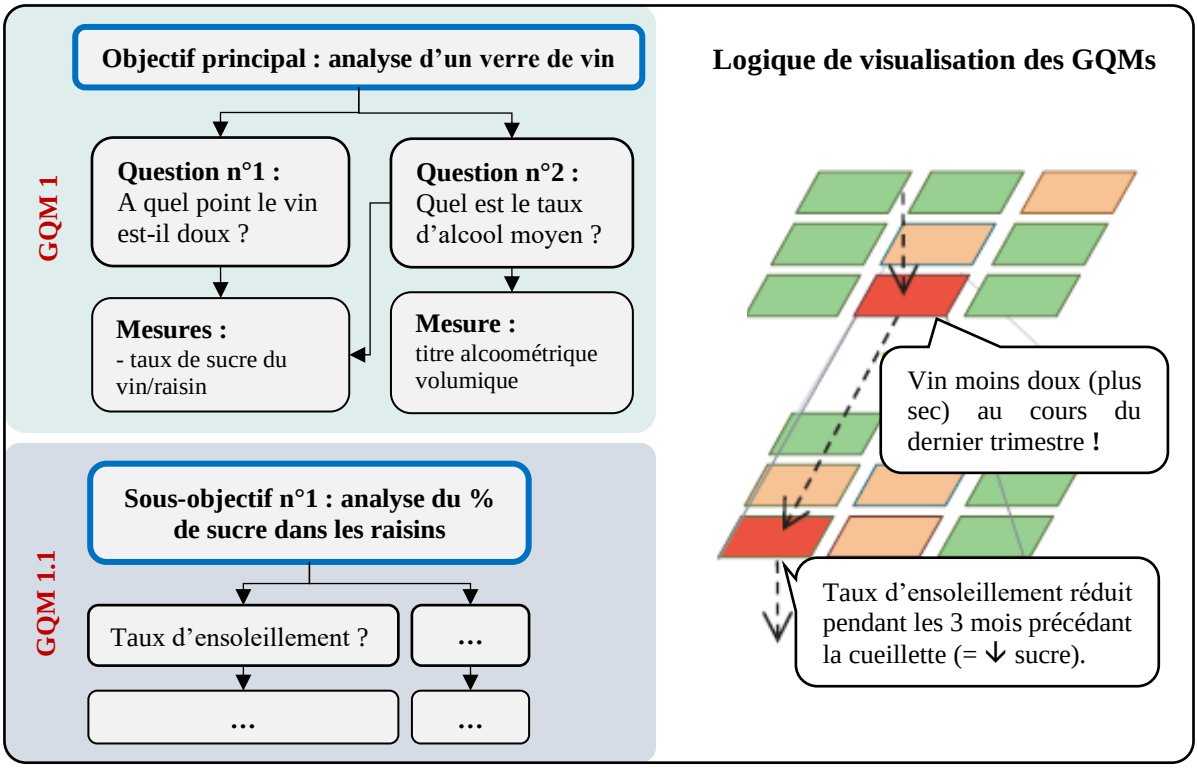


Figure 1.2 : élaboration d'un modèle GQM

1.3 LES VISUALISATIONS ET LEUR COMPLEXITÉ

1.3.1 Les différents paradigmes

Remarquons que **deux types de scénarios de conception de visualisations** sont évoqués dans la littérature : **les types *pull* et *push*** (Janes et Succi, 2009). En type pull, l'utilisateur emploie un dashboard, ou d'autres outils, pour obtenir les informations qu'il recherche. Il s'agit par exemple d'outils de BI proposant des fonctionnalités avancées : paradigme multidimensionnel, prédictions, filtres, tris… En type push, le dashboard cherche à capturer l'attention de l'utilisateur pour lui transmettre des informations importantes. On peut par

exemple citer le cas d'un dashboard placé dans les lieux de passage d'un hôpital en vue d'informer le personnel d'éventuelles anomalies ou problèmes (Weiner, Balijepally et Tanniru, 2015). Notons cependant que ce deuxième type d'interface a un caractère plus opérationnel que stratégique et est dès lors moins applicable à la Business Intelligence. On est donc ici dans le type pull.

Il est ensuite utile de remarquer que **différents débats animent la recherche** relative aux visualisations depuis de nombreuses années. C'est par exemple le cas du *junk chart debate* qui oppose le paradigme de la *junk chart*, à savoir une interface où l'on inclurait un maximum de données et d'éléments visuels au vu des pixels disponibles, au paradigme de la version épurée. C'est aussi le cas du *task-oriented chart debate* dans lequel certains pensent que les visualisations doivent être destinées à une tâche bien spécifique et prédéfinie, alors que d'autres considèrent qu'un tel point de vue restreint beaucoup trop l'étendue de l'analyse stratégique. Sachant cela, analysons maintenant chacune de ces controverses, pour ensuite se pencher sur des constatations importantes qui sont moins sujettes à débat.

En ce qui concerne le premier débat, les auteurs opposés au **paradigme de la junk chart** soulignent qu'il est impératif d'afficher les informations de la manière la plus claire possible, en évitant les distractions et artifices, et ce, quel que soit le potentiel des outils de visualisation à disposition. Ce point de vue s'appuie notamment sur certaines recherches relatives aux facteurs de la compréhension et de la rétention des informations ((Cleveland et McGill, 1984) ; (Kosslyn, 1989) ; (Few, 2011) ; (Rust, Thompson et Hamilton, 2006)). Les défenseurs de la junk chart considèrent, quant à eux, que c'est précisément en forçant l'utilisateur à faire davantage d'efforts cognitifs pour appréhender les visualisations qu'on lui permettra de retenir plus d'informations. Certaines études neurobiologiques tendent d'ailleurs à confirmer ce point de vue ((Borgo et al., 2012) ; (Brady et al., 2008) ; (Bateman et al., 2010) ; (Hullman, Adar et Shah, 2011)).

La plupart des auteurs s'accordent toutefois à dire que la quantité d'éléments affichés sur les interfaces est loin d'être le seul facteur à prendre en compte (Borkin et al., 2013) : des aspects tels que les types de graphiques, les couleurs, les formes, le ratio données/pixels, l'agencement des figures, la familiarité de l'utilisateur avec les figures… sont également des facteurs primordiaux (cf. sous-section 1.3.2).

En ce qui concerne le second débat, Bertin (1983), Casner (1991) et Nowell, Schulman et Hix (2002) sont de ceux qui considèrent que chaque visualisation doit être conçue pour répondre à une tâche ou une question bien spécifique et, dès lors, que les développeurs doivent se concentrer davantage sur les tâches à effectuer que sur les interactions qui existent au sein de la globalité des données : il s'agit du **paradigme task-oriented**. Ce point de vue est cependant contesté par des chercheurs tels que Tweedie (1997), qui considèrent que les visualisations doivent, par définition, être polyvalentes puisque la dynamique des interfaces doit permettre de varier l'analyse et de la rendre utile à une diversité de tâches.

Froese et Tory (2016) nuancent ces propos en tirant le constat suivant : les utilisateurs doivent pouvoir consulter différents sous-ensembles d'informations et de KPIs répondant chacun à différentes questions bien spécifiques, tout en ayant la possibilité d'adapter quelque peu les interfaces en fonction des besoins du moment. Les auteurs voient là une opportunité de répondre à des questions ciblées tout en étant en mesure de s'adapter, avec agilité, aux exigences stratégiques du moment (sans devoir forcément faire appel aux développeurs de l'interface). Rappelons que les outils de BI ont pour objectif de permettre aux décideurs d'avoir une vue d'ensemble sur les activités de leur entreprise (Shukla et Dhir, 2016).

1.3.2 Les détails qui font la différence

Lorsque l'on prête attention à certains « détails » des visualisations, on peut remarquer que rien ne doit être laissé au hasard. Et c'est précisément là que réside une des complexités majeures des visualisations : **des choses qui peuvent sembler banales à première vue font finalement la différence**.

Borkin et al. (2013) font partie des auteurs qui se sont largement penchés sur le sujet et ils ont dressé différents constats. Les interfaces sont, selon eux, plus efficaces si elles présentent des pictogrammes et des images familières, si elles sont colorées (plus de 6-7 couleurs), visuellement denses, enjolivées/agrémentées (le *data-ink* ratio doit être relativement bas) et si elles ne semblent pas trop banales pour les utilisateurs (des figures moins courantes, telles que des *heat maps*, doivent accompagner des figures plus traditionnelles, telles que les histogrammes et les nuages de points). Notons que beaucoup d'études s'accordent à dire que ces aspects seraient souvent presque aussi déterminants que les compétences et les connaissances des utilisateurs (Isola et al., 2011).

Ces différents constats font écho à la thèse de la *junk chart* puisqu'on cherche ici à agrémenter les figures de données et d'éléments visuels. Aussi remarque-t-on une volonté d'afficher certains éléments familiers pour l'utilisateur, tels que des pictogrammes : ⚠, 🚚, 🧺… Borkin et al. (2013) affirment également qu'il existe une corrélation entre la mémorisation de l'information et le caractère « naturel » des visualisations. Ils soulignent par exemple que des diagrammes colorés tels que les *heat maps* sont à privilégier car ils ont une apparence plus naturelle que d'autres. Bar et Neta (2006) affirment pour leur part que l'utilisation de bords/angles arrondis est pertinente car les utilisateurs auraient des émotions plus positives à leur égard, notamment parce qu'ils ont une apparence plus naturelle.

Ware (2012) insiste lui sur la notion de *pre-attentive processing*. L'idée est que certaines propriétés graphiques sont traitées rapidement par les individus avant même qu'ils en aient conscience. Il s'agit principalement de tirer profit des formes, des couleurs, des mouvements et des positions spatiales. La *figure* 1.3 ci-dessous en présente un exemple. Dans le premier cas, la détection des chiffres 3 et 4 est facilitée par les nuances de gris plutôt que par les symboles « 3 » et « 4 » en eux-mêmes. Dans le deuxième cas, ce sont les couleurs qui font la différence. Dans le troisième cas, on remarque que, lorsque plusieurs signaux sont combinés, ils sont traités de manière séquentielle : on analyse les couleurs et puis les formes, ou inversement (Janes, Succi et Sillitti, 2013).

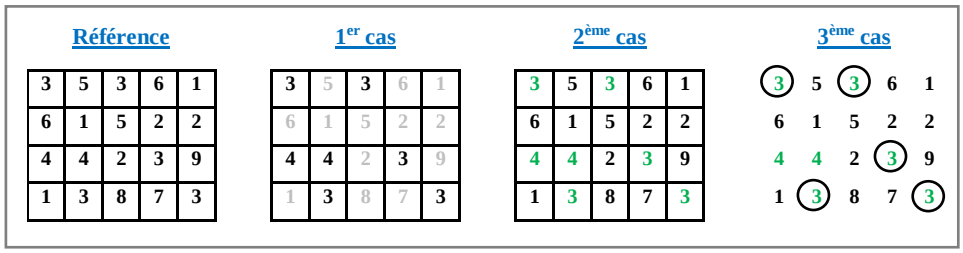


Figure 1.3 : illustration du *pre-attentive processing*

Notons aussi qu'il faut prendre en considération les utilisateurs ayant des besoins spécifiques, tels que les daltoniens. Ce faisant, il est utile de prêter attention à des critères tels que l'épaisseur des lignes et les nuances de gris (Janes, Succi et Sillitti, 2013).

Enfin, remarquons qu'il est fréquemment recommandé d'éviter la 3D tant la lecture peut être biaisée par le chevauchement des données et les erreurs de perspective (Bendoly, 2016). Il est également important, quelle que soit la visualisation adoptée, de permettre à l'utilisateur de

consulter les données sur lesquelles cette visualisation est basée, et ce, typiquement au travers d'un tableau croisé que l'on peut consulter de manière optionnelle (Janes, Succi et Sillitti, 2013).

Le ***tableau* 1.1** présenté ci-dessous récapitule les éléments étudiés précédemment. Il était important de le construire car il fait office de **checklist pour la suite de l'analyse**.

Tableau 1.1 : synthèse de certains des critères à considérer pour les visualisations

CRITÈRES À CONSIDÉRER	DESCRIPTION	CHECKBOX	RÉFÉRENCES PRINCIPALES UTILISÉES
Pictogrammes familiers		☑	(Borkin et al., 2013) (Isola et al., 2011)
Couleurs	• Couleurs naturelles de préférence • > 6-7 couleurs pour l'ensemble de la page	☑	
Densité visuelle	• Ratio pixels utilisés/pixels disponibles 'élevé'	☑	
Embellissement, « ornements » visuels	• Ratio pixels dédiés à l'affichage des données/pixels utilisés plus 'faible'	☑	(Borkin et al., 2013) (Few, 2011) (Isola et al., 2011) (Kosslyn, 1989)
% de figures peu courantes	• Eléments courants (scatter plots, histogrammes…) **+** moins courants (heat maps, arbres, diagrammes…)	☑	(Borkin et al., 2013) (Isola et al., 2011)
Aspects naturels	• Bords arrondis, couleurs naturelles, structures semblables aux éléments de la vie quotidienne (jauges similaires aux cadrans de voiture, arbres, classements…)	☑	(Bar et Neta, 2006) (Borkin et al., 2013)
Eléments prétraités de manière inconsciente et rapide	• Formes, couleurs, mouvements et positions spatiales	☑	(Janes, Succi et Sillitti, 2013) (Ware, 2012)
Eléments destinés aux utilisateurs ayant des besoins particuliers	• Ex. : daltoniens → épaisseur des lignes, nuances de gris…	☑	(Janes, Succi et Sillitti, 2013)
Absence de 3D	• Eviter les problèmes de perspective, le chevauchement des données…	☑	(Bendoly, 2016)
Personnalisation des interfaces	• Possibilité d'adapter les interfaces aux besoins stratégiques du moment (agilité) en fonction des outils et des compétences de l'utilisateur	☑	(Froese et Tory, 2016)
Données sous-jacentes aux visualisations	• Procurer une vue détaillée des données au travers d'un tableau croisé	☑	(Janes, Succi et Sillitti, 2013)

Chapitre 2 : présentation détaillée de l'étude de cas

L'objectif de ce chapitre est de présenter l'étude de cas qui est au cœur de l'analyse. Bien que cette dernière ne soit pas une fin en soi, elle est un support primordial au déploiement des outils de BI. Il est donc nécessaire d'avoir une compréhension détaillée du domaine d'application, des données, des enjeux et des exigences si l'on souhaite mettre en place des outils stratégiques réellement efficaces. Il s'agit de **prendre le temps de fournir des éléments qui faciliteront l'analyse effectuée par la suite** et, par la même occasion, de transmettre au lecteur (à la lectrice) quelques informations qui seront susceptibles d'enrichir sa connaissance du marché étudié.

Cette étude de cas porte sur les *freight brokers.* Ces derniers jouent donc le rôle d'intermédiaire entre des expéditeurs ayant une cargaison à envoyer et des transporteurs à la recherche de biens à véhiculer. Pour des raisons d'anonymat, l'entreprise qui a fourni les données utiles à l'analyse est ici fictivement appelée BrokerRoad.

La première section de ce chapitre explicite les caractéristiques de l'industrie dans laquelle BrokerRoad évolue ainsi que son processus métier. La seconde section décrit, quant à elle, les données qui sont mises à disposition et leur potentiel stratégique. Et, enfin, la troisième section souligne les enjeux et exigences qui caractérisent le marché.

Les constats dressés ici, ainsi que la revue de la littérature du chapitre précédent, constituent donc le terreau nécessaire au déploiement des outils de BI. L'ingénierie des exigences, le traitement de données (ETL, data warehouse…) et les outils d'analyse décisionnelle reposeront ainsi sur des bases solides.

2.1 L'ACTIVITÉ DE L'ENTREPRISE

L'étude de cas est liée au **transport routier de marchandises** aux USA et au Mexique. Cette industrie rapporte chaque année plusieurs centaines de milliards de dollars aux entreprises des USA et est donc d'une importance considérable. A titre d'illustration, notons qu'environ 3 billions de tonnes-kilomètres ont été enregistrés aux USA en 2016 (tonnes-kilomètres = tonnes

transportées x kilomètres parcourus) : des camions chargés en moyenne de 25 tonnes de marchandises parcourent donc au total environ 120 milliards de kilomètres par an. En comparaison, on dénombre 251 milliards de tonnes-km au Mexique et « à peine » 151 milliards en France (OECD, 2017).

Le cas qui nous occupe est centré sur un acteur bien spécifique du transport de marchandises : le ***freight broker***. Il établit le lien entre les entreprises qui veulent expédier des marchandises et les transporteurs, en échange d'une commission (FMCSA, 2017).

Notons que, contrairement à ceux que l'on appelle *freight forwarders* (qui jouent aussi un rôle d'intermédiaire entre les expéditeurs et les transporteurs), les freight brokers n'entrent en principe jamais en contact physique avec les marchandises. Alors que les freight forwarders proposent certains services supplémentaires tels que l'entreposage et le dispatching, les freight brokers ne manipulent pas les biens et ne sont dès lors normalement pas responsables des éventuels dommages causés à ces derniers ((Masterslogistical, n.d.) ; (FMCSA, 2017)).

Mentionnons à titre d'information qu'il existe des fournisseurs de logistique proposant des services encore plus complets : on les appelle souvent 3PLs ou Third Party Logistics. Ils sont une forme d'externalisation de la chaîne logistique de l'entreprise et offrent des solutions qui incluent des services tels que l'entreposage, le cross-docking (c.-à-d. : le fait de réagencer les colis et de les faire passer des quais d'arrivée aux quais de départ sans véritablement les stocker), la distribution et aussi le transport (Masterslogistical, n.d.).

On l'aura compris, le broker ne dispose pas réellement d'installations logistiques, mais il fait appel à un réseau étendu de partenaires pour relier les expéditeurs aux transporteurs. D'une part, ceci évite aux expéditeurs d'entreprendre des démarches pénibles pour trouver des transporteurs appropriés, fiables, rapides et abordables et, d'autre part, permet aux transporteurs de trouver des cargaisons à transporter à moindre effort (Welna, n.d.).

L'entreprise qui nous intéresse ici, **BrokerRoad**, est basée au Texas. Ce freight broker est actif dans l'industrie du transport routier effectué sur de longues distances aux USA et au Mexique, avec de grands colis et en Full Truck Load (FTL). Le terme FTL fait référence à la notion d'expédition complète : l'expédition occupe la totalité du camion en accaparant tout l'espace de la remorque ou en ayant atteint la charge utile maximale, c.-à-d. le poids maximal de marchandises autorisé. A titre d'information, notons que la Full Truck Load s'oppose à la

Less than Truck Load (LTL) qui, elle, fait référence à un chargement partiel (Freightquote.com, 2016).

Les termes *event* et *shipment* utilisés par la suite font référence à un record particulier de la base de données. Chaque record a trait à une transaction qui a pour but de lier un seul expéditeur à un seul transporteur pour assurer le transport d'une cargaison spécifique avec un type de véhicule spécifique. Cette livraison est acheminée d'un point A à un point B mais peut faire l'objet d'arrêts intermédiaires. Les biens transportés peuvent être des *hazardous materials*, c'est-à-dire des matières dangereuses qui nécessitent une manutention particulière. Les véhicules utilisés sont souvent des *dry vans* (camions à remorque non réfrigérée), des *reefers* (camions à remorque réfrigérée) et des *flatbeds* (camions à remorque ouverte, c.-à-d. qui n'ont ni paroi, ni toit) (Freightquote.com, 2017).

Le **processus métier** étant quelque peu complexe, il est question ici d'en présenter les grandes étapes. Tout en restant précis, il s'agit donc de ne pas entrer dans des détails trop techniques et de respecter le caractère confidentiel de certaines informations. La ***figure* 2.1** construite à la fin de cette section et les explications qui suivent visent ainsi à résumer ces étapes.

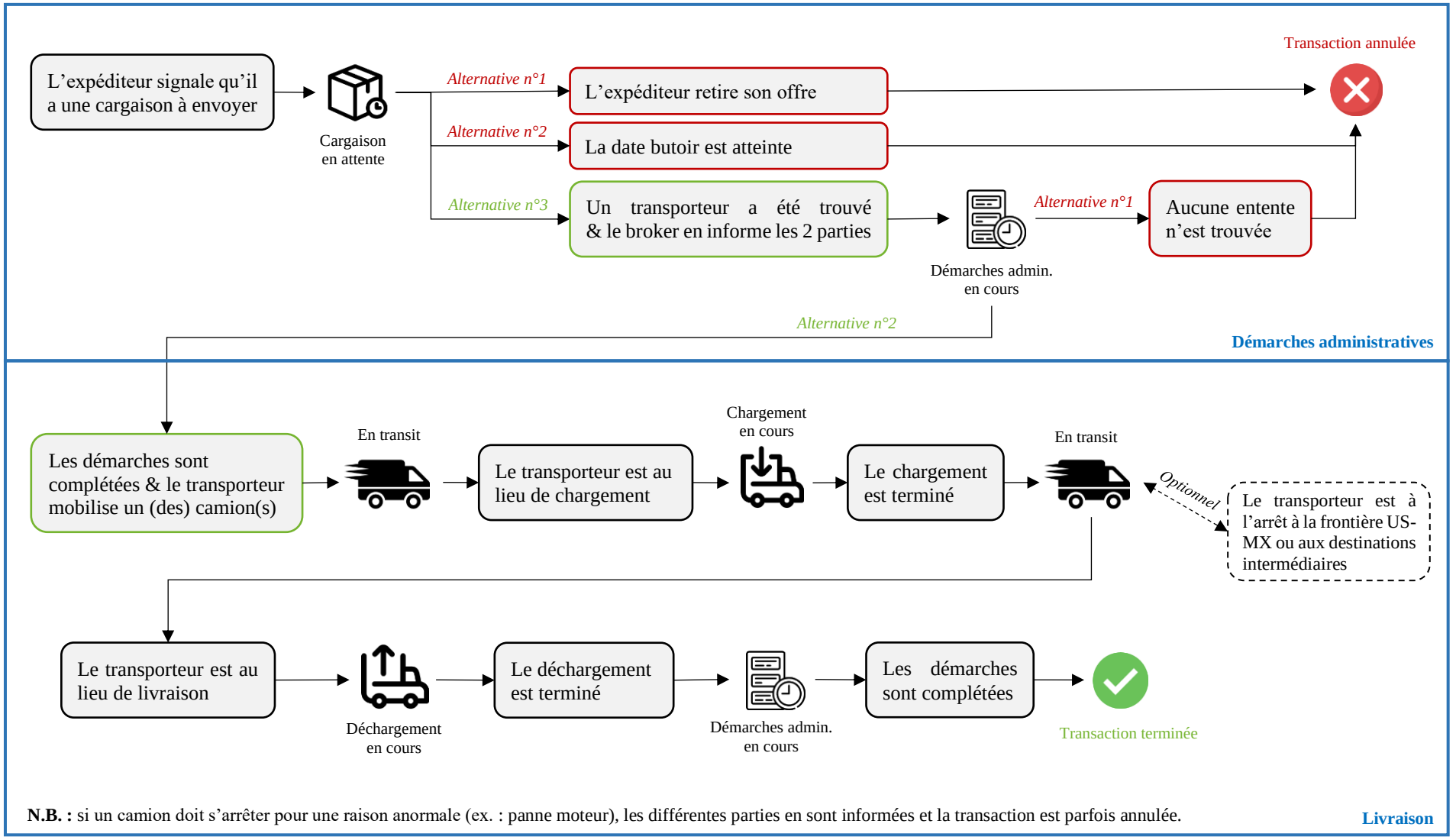
Dans un premier temps, l'expéditeur fournit des informations relatives à la cargaison à transporter sur la plateforme/marketplace de BrokerRoad (types de biens, valeur…). Il mentionne également la date butoir avant laquelle les marchandises doivent être acheminées (date d'expiration). Tant qu'un transporteur n'a pas été trouvé, l'expéditeur est en droit de retirer son offre de la plateforme.

Une fois qu'une offre est postée, BrokerRoad va, au travers de sa plateforme ou d'interventions d'une autre nature (ex. : appels téléphoniques), trouver un transporteur qui accepte la mission aux conditions les plus avantageuses. Différentes offres sont alors proposées et on procède parfois à des enchères. Une fois le transporteur trouvé, BrokerRoad informe les deux parties prenantes de la décision et attend certaines informations administratives en retour. Il arrive parfois que la transaction soit annulée à ce stade si les parties ne s'accordent pas sur les modalités.

Une fois les démarches de nature administrative réglées, le transporteur envoie son camion vers le lieu de chargement ; après quoi il informe les autres parties du moment d'arrivée

sur ledit lieu, du moment auquel le chargement a été complété et du moment auquel le camion s'est remis en mouvement. Sur son trajet, le camion peut éventuellement effectuer une (des) livraison(s) intermédiaire(s) et être amené à traverser la frontière américano-mexicaine. Enfin, le transporteur donne des informations à propos du moment où le camion est arrivé à la destination finale et du moment où la cargaison a été déchargée. Notons que la transaction est indiquée comme étant complétée lorsque les démarches administratives sont terminées. D'autre part, en cas de problème (panne moteur ou autre), les différentes parties sont informées du fait que le camion est à l'arrêt (*halted*) et la transaction est alors parfois annulée.

Soulignons déjà que **de nombreux champs de la base de données ne sont pas complétés**, soit parce que le personnel n'a pas reçu les informations requises, soit parce qu'il n'a pas pris soin de les encoder. Cette carence en informations a bien évidemment un impact sur l'analyse et sur les dashboards.



Pictogrammes : (flaticon.com, 2019)

Figure 2.1 : processus métier de BrokerRoad

2.2 LES DONNÉES À DISPOSITION

Le domaine d'activité et le processus métier de l'entreprise ayant été étudiés, il est nécessaire de se pencher sur les données, qui sont mises à disposition au travers d'un fichier Excel.

Les **différents enregistrements** recensent toute une série d'informations liées aux diverses transactions mais ne sont pas, à proprement parler, directement issus des tables de la base de données opérationnelle. Ils sont en quelque sorte le fruit d'une sélection partielle de transactions (lignes du fichier) et d'attributs (colonnes du fichier) extraits de l'application de BrokerRoad. Il n'existe par ailleurs aucune clé primaire : il n'est pas possible de respecter la contrainte d'unicité car, bien que des numéros d'expédition soient attribués aux records, ceux-ci ne sont en fait renseignés que pour les transactions pour lesquelles un transporteur a déjà été trouvé. Les transactions aux stades préliminaires n'ont donc pas de numéro d'expédition.

On dénombre **3 678 entrées et 60 attributs**. Les 3678 entrées ont été capturées sur une période allant du 22/05/17 au 14/01/19. Les 60 attributs ne sont également qu'un sous-ensemble des données opérationnelles de BrokerRoad : il était en effet important de prendre en considération le caractère confidentiel des données et de concentrer l'analyse sur une problématique spécifique. Notons que le fait que le nombre de transactions à disposition soit assez limité et que les attributs ne soient pas densément peuplés (vu les nombreux champs anormalement incomplets) a un impact sur les visualisations développées par la suite.

Les 60 variables sont décrites en détail en ***annexe* A-1**. La ***figure* 2.2** construite à la fin de cette section a pour objectif de résumer les grandes catégories qui émergent de ces variables.

Les données ont un **potentiel stratégique important** puisqu'elles permettent tout particulièrement de répondre à des questions d'ordre financier et spatio-temporel : revenus, coûts, analyse dans le temps, analyse géographique, etc.

Par exemple, lorsqu'il est question des aspects financiers, nombre de questions sont envisageables. En veillant à analyser les informations sous le prisme du temps et de l'espace, on peut fournir de nombreuses informations et statistiques (minimum, maximum, moyenne, série temporelle, etc.) au niveau désiré de granularité/précision : quelles sont les commissions

empochées par BrokerRoad ? Quels sont les coûts supportés par les expéditeurs ? Quels prix les transporteurs fixent-ils ? Combien d'expéditeurs et de transporteurs prennent part à l'ensemble des transactions ? Quelles parties prenantes recourent le plus souvent aux services de BrokerRoad ? A quelle fréquence postent-elles des offres sur la plateforme ? Quels intervenants et quels types de cargaisons génèrent le plus de revenus ou de commissions pour BrokerRoad ? Quels transporteurs sont les plus onéreux/abordables ? Les expéditeurs ont-ils souvent tendance à recourir aux mêmes transporteurs ? …

Ces questions sont donc analysées sous le prisme du temps et de l'espace. Ces deux critères entraînent d'ailleurs bien d'autres questions dans leur sillage : quelles sont les origines et destinations les plus populaires ? Les parties prenantes sont-elles souvent liées aux mêmes origines et/ou destinations ? Quels sont les temps/délais nécessaires pour mener à bien les différentes tâches (chargement, déchargement, etc.) ? …

Il s'agit de répondre à ces questions au travers de l'approche multidimensionnelle de la BI : l'information contenue dans les données est conceptuellement explorée au travers d'un **cube OLAP**. Rappelons que le cube OLAP (*OnLine Analytical Processing*) s'oppose à l'OLTP (*OnLine Transaction Processing*) puisqu'il permet d'adopter une posture stratégique en analysant les métriques sous le prisme de différentes dimensions (Vaisman et Zimanyi, 2014). Dans le cas qui nous occupe, les cases du cube représentent les métriques liées à chacune des expéditions (commissions, prix, etc.), tandis que les arêtes multiples représentent des dimensions telles que les emplacements, les parties prenantes et le temps. Ces aspects sont bien évidemment largement détaillés par la suite.

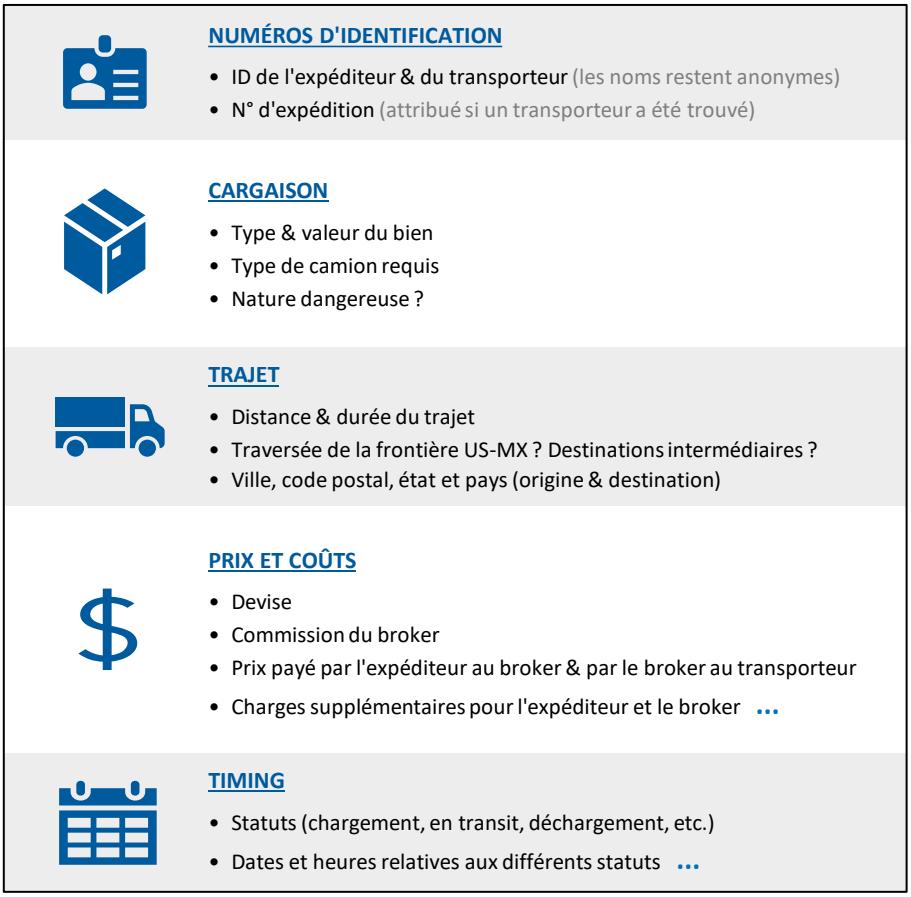


Figure 2.2 : catégorisation générique des variables à disposition

2.3 LES ENJEUX ET EXIGENCES DU BUSINESS

Le domaine d'activité de BrokerRoad et les données à disposition ayant ainsi été explorés, concluons ce chapitre en étudiant de plus près quelques-uns des enjeux majeurs de cette industrie, ainsi que certaines des exigences qui y sont liées. Les constats établis ici ont pour but d'enrichir le référentiel d'exigences construit par la suite.

Il est crucial de chercher à comprendre les grands enjeux du marché des freight brokers puisque, de toute évidence, **le fait de ne pas appréhender correctement les exigences de ce business aboutirait à des visualisations peu pertinentes**…

Rappelons ainsi que l'industrie de la logistique est d'une importance considérable aux USA et que nombre d'acteurs du monde académique et commercial se penchent sur les facteurs clés qui l'influencent. C'est à cet égard que de nombreux auteurs tels que Johnson et Schneider

(1995) et Huang et al. (2011) ont pris soin de souligner l'impact que les changements législatifs et technologiques ont eu et continuent d'avoir sur le marché des freight brokers et sur les exigences de ces derniers.

Du point de vue législatif, notons simplement que ces auteurs expliquent qu'au cours des années 80, cette industrie a connu un essor considérable grâce à la suppression des quotas de compagnies de transport. Devant une quantité sans cesse croissante de compagnies de transport et d'expéditeurs ne sachant où donner de la tête pour envoyer leurs cargaisons, les brokers se sont multipliés par milliers ((Johnson et Schneider, 1995) ; (Huang et al., 2011) ; (Bleggi et Zhou, 2017)).

Huang et al. (2011) insistent également sur le rôle que le **facteur technologique** a joué et continue de jouer dans la gestion opérationnelle et stratégique des transactions. Ces auteurs tirent d'ailleurs un constat applicable à la majorité des brokers, BrokerRoad y compris. Ils expliquent que, malgré l'évolution des outils technologiques et des possibilités qui en découlent, de nombreuses transactions restent chaotiques dans la pratique. De nombreux intervenants ne fournissent en effet pas les informations utiles au processus…

Dans bien des cas, les expéditeurs et les transporteurs sont des entreprises peu sophistiquées avec peu de moyens technologiques à disposition : pas d'ERP (*Enterprise Ressource Planning*), pas de TMS (*Transport Management System*, pseudo-équivalent de l'ERP dans le domaine du transport) ... De plus, le fait que beaucoup d'échanges s'opèrent encore par téléphone et par mails dégrade la qualité des données car les employés ne prennent pas suffisamment soin d'encoder les informations dans le système. Enfin, malgré l'avancée technologique qui les caractérise, les plus grandes entreprises ne fournissent que trop souvent des données incomplètes, soit parce qu'elles sont réticentes à fournir les informations requises, soit parce qu'elles peinent à les fournir (Huang et al. ; 2011). Ces différents facteurs expliquent une grande partie des **lacunes observées dans la base de données de BrokerRoad**.

Introduisons maintenant certains des **indicateurs clés de performance** (KPIs) les plus importants.

Nombre d'indicateurs stratégiques ont déjà été mentionnés plus ou moins directement dans la section précédente. D'autres indicateurs importants sont également évoqués dans la littérature, mais beaucoup d'entre eux ne sont pas calculables au travers des données de

BrokerRoad, du moins pas au travers de celles qui sont mises à disposition. Bien que certains de ces indicateurs ne soient tout bonnement pas applicables au cas de BrokerRoad, d'autres sont en fait exclus du fichier Excel.

Les quelques constats qui suivent visent donc à poser le contexte de la problématique du choix des KPIs (qui est largement approfondie dans les chapitres ultérieurs) ainsi que certaines des limites qui y sont inhérentes.

Une analyse stratégique idéale impliquerait d'étudier les dimensions financières et spatio-temporelles, mais aussi de se pencher sur d'autres indicateurs de performance. Il n'est pas seulement question de commissions, de *On Time Pickup*, de *On Time Delivery* ou autres, mais également d'indicateurs tels que l'impact environnemental et la consommation d'énergie (Crainic, Damay et Gendreau, 2007). Ces données sont manquantes dans le cas qui nous occupe et ne peuvent donc pas être inclues dans les visualisations. On ignore par ailleurs dans quelle mesure ces informations pourraient être susceptibles d'intéresser BrokerRoad.

Mackenzie et al. (2009) insistent, eux, sur l'importance des feedbacks : les expéditeurs et les transporteurs peuvent ainsi exprimer leur degré de satisfaction quant aux performances et établir des ratings. Ces auteurs insistent également sur l'utilité des données extérieures aux acteurs, telles que les vitesses moyennes des camions, les conditions de circulation et les conditions météorologiques. Autant de critères qui peuvent, par exemple, expliquer certains des délais et certaines des annulations de transactions.

En *annexe* A-2, on retrouve la liste dans laquelle Mackenzie et al. (2009) répertorient les données que les différents acteurs doivent, selon eux, idéalement fournir afin de pouvoir construire les indicateurs clés de performance. On aura néanmoins compris que l'échange d'informations est bien plus complexe dans la pratique : informations manquantes, problèmes d'intégration des données liés au fait que les différents acteurs possèdent des systèmes de données très hétérogènes, etc.

Notons, à titre d'information, que c'est en vue d'éviter ce genre de problèmes que l'on fait régulièrement appel à des *data workers*, qui ont pour fonction d'acquérir les données au travers d'interfaces telles que les *Application Programming Interfaces* (APIs), de les intégrer et de les revendre aux parties prenantes (Project44.com, 2019). Cela a bien évidemment un coût financier, mais pas seulement : en créant des schémas de données uniques et fédérés, le data

worker détermine la SSoT (*single source of truth*), à savoir la référence absolue pour la prise de décisions et l'administration des données. Il n'est dès lors pas toujours facile de déterminer à qui confier cette tâche dont dépend l'intégrité des données opérationnelles et des data warehouses.

Chapitre 3 : ingénierie des exigences et modélisation du data warehouse

3.1 L'INGÉNIERIE DES EXIGENCES

Le chapitre précédent a mis en évidence un grand nombre d'exigences inhérentes au business de BrokerRoad. Les exigences formulées par BrokerRoad (cf. section 2.2), ainsi qu'une analyse des données à disposition et de la littérature académique et professionnelle, ont permis de comprendre certains des facteurs et indicateurs utiles à l'analyse. Le chapitre précédent ayant été relativement détaillé, il n'est pas question ici de répéter ce qui a déjà été dit, mais bien d'incorporer ces constats dans des ***frameworks* qui facilitent le traitement de données et la création des visualisations**.

Les informations (questions stratégiques et données) n'ayant été transmises qu'à une seule reprise par BrokerRoad, il n'est pas réellement possible de procéder de manière agile. Il s'agit alors d'effectuer une analyse traditionnelle, linéaire et séquentielle (cf. section 1.1).

Pour résumer les résultats de l'ingénierie des exigences, le choix a été fait de recourir à l'approche **GQM** présentée précédemment et de la combiner à une spécification hybride des exigences, appelée ***analysis- and source-driven approach***. Cette technique, chère notamment à Vaisman et Zimanyi (2014), permet de déterminer la structure du DW à mettre en place en combinant les questions stratégiques du client (*analysis-driven*) et le potentiel des données opérationnelles à disposition (*source-driven*). Il est à noter qu'on ne peut parfois pas donner une réponse aux questions du client, et ce, principalement parce que les données opérationnelles nécessaires ne sont pas disponibles. Il aurait par exemple été impossible, si le client en avait fait la demande, d'analyser les conditions météorologiques impactant ou non les expéditions.

Notons que la 2ème technique d'analyse n'a pas été explicitée dans la revue de la littérature parce qu'elle fait clairement référence à la structure des data warehouses, sujet que l'on a choisi de ne pas aborder dans la revue de la littérature.

La logique du modèle Goal-Question-Measurement ayant déjà été illustrée au travers de la *figure* 1.2, il n'a pas été jugé opportun d'avoir à nouveau recours à une figure ici : une liste

des constats dressés à l'aide du modèle semble plus explicite et adéquate. En suivant une logique quelque peu « adaptée » du modèle, on parvient à dégager un objectif principal et des sous-objectifs (niveau conceptuel), des requêtes représentatives des exigences fonctionnelles (niveau opérationnel) et des métriques utiles pour répondre à ces requêtes (niveau quantitatif). Ce modèle a priori peu complexe permet donc, lorsqu'il est répété de manière rigoureuse, de dresser la liste suivante :

- **Objectif principal** : maximiser les commissions
- **Sous-objectifs principaux**
  1. Etudier les profils des expéditeurs et des transporteurs
  2. Analyser les stratégies financières et logistiques mises en place pour augmenter leur impact positif (si succès) ou limiter leur impact négatif (si échec)
  3. Analyser à la volée (*on the fly*) la progression globale actuelle des expéditions
- **Exemples de requêtes représentatives**
  - **1.a.** Commissions/dépenses liées aux expéditeurs par type de cargaison/type d'équipement/zone géographique/période
  - **1.b.** Commissions/revenus liés aux transporteurs par type de cargaison/type d'équipement/zone géographique/période
  - **1.c.** Expéditeurs/transporteurs ayant le plus souvent recours aux services de BrokerRoad par zone géographique/période **…**
  - **2.a.** Commissions/revenus/dépenses par type de cargaison/zone géographique/période
  - **2.b.** Budget max. des clients vs prix effectivement payé par zone géographique/période
  - **2.c.** Prix ayant remporté les enchères vs prix effectivement facturé par le transporteur par zone géographique/période
  - **2.d.** Origines et destinations les plus pratiquées par période
  - **2.e.** Durée et distance des trajets par client/transporteur/type de trajet
  - **2.f.** Délais pour mener à bien les différentes étapes du processus par client/transporteur/type de cargaison/type d'équipement/zone géographique/période **…**
  - **3.a.** Statut global des expéditions en cours par client/transporteur/type de trajet
  - **3.b.** Valeur des cargaisons en cours d'acheminement **…**

Notons que les données opérationnelles dont on dispose permettent effectivement de répondre aux exigences fonctionnelles décrites ci-dessus. On a ainsi dressé des besoins

d’analyse (méthode GQM et approche orientée analyse) pour ensuite les confronter au potentiel des données opérationnelles (approche orientée sources). Le *tableau* 3.1 ci-après (qui se réfère à la liste ci-dessus) a pour but de former un tout cohérent qui puisse synthétiser les faits, mesures et dimensions qui émergent de cette approche hybride : on y voit les fondements du DW de BrokerRoad. Le tableau étant a priori suffisamment explicite, le choix est fait ici de ne pas le commenter, mais des précisions sont apportées en section 3.3. Les exigences non-fonctionnelles sont, quant à elles, abordées en section 5.1.

Tableau 3.1 : spécification des exigences orientée analyse et sources

DIMENSIONS	HIÉRARCHIES ET NIVEAUX	CARDINALITÉS	EXEMPLES DE SCÉNARIOS D'ANALYSE (LIÉS À LA LISTE DES REQUÊTES)										
		Cf. section 3.3	1			2						3	
			A	B	C	A	B	C	D	E	F	A	B
Expéditeur		(1,n)	✓		✓					✓	✓	✓	
Transporteur		(1,n)		✓	✓					✓	✓	✓	
Cargaison		(1,n)	✓	✓		✓					✓		✓
Equipement (remorque)		(1,n)	✓	✓							✓		
Type de trajet (livraison(s) intermédiaire(s), traversée de la frontière ?)		(1,n)								✓		✓	
Devise		(1,n)	✓	✓		✓	✓	✓					✓
Emplacement	*Lieu : Ville → Etat → Pays*	(1,n) x 2	✓	✓	✓	✓	✓	✓	✓		✓		
Date	*Calendrier : Jour → Mois → Trimestre → Année*	(1,n) x 15	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Temps (hh:mm:ss)		(1,n) x 15									✓	✓	✓
Statut de l'expédition		Dim. dégénérée										✓	✓
MESURES	**DESCRIPTIONS**												
Valeur de la cargaison													✓
Durée du trajet										✓	✓	✓	
Distance du trajet									✓	✓			
Budget max. de l'expéditeur	*Mesure non-additive : (: Avg /+)*						✓						
Prix payé par l'expéditeur			✓				✓						
Coûts suppl. à charge de l'expéditeur			✓										
/ Dépenses totales de l'expéditeur	*Mesure dérivée : prix payé + coûts additionnels*		✓		✓	✓							
Offre du transporteur qui a remporté d'éventuelles enchères	*Mesure non-additive : (: Avg /+)*							✓					
Prix facturé par le transporteur				✓				✓					
Coûts suppl. facturés par le transporteur				✓									
/ Revenus totaux du transporteur	*Mesure dérivée : prix reçu + coûts facturés*			✓	✓	✓							
Commission du broker			✓	✓		✓							

3.2 L'IMPORTANCE DU TRAITEMENT DE DONNÉES

Pour passer des données et exigences de départ aux visualisations, les principales étapes à entreprendre sont celles de la construction du data warehouse et de son approvisionnement en données. Rappelons, à ce stade, que l'objectif de ce mémoire n'est pas d'étudier en détail les lignes de pensée qui animent la littérature et le monde professionnel en ce qui concerne la conception du DW : il est plutôt question de se focaliser sur un paradigme d'architecture BI spécifique (à savoir l'architecture double couche) et de développer le DW et le processus ETL de BrokerRoad en fonction.

Si l'on veut construire des outils d'analyse qui soient suffisamment pertinents, les étapes de la création et de l'approvisionnement du DW sont cruciales et inévitables en BI, à l'exception du cas où l'on utiliserait une architecture à simple couche. Dans ce cas de figure, le DW serait remplacé par un middleware permettant de faire une transition virtuelle entre les données opérationnelles et les outils d'analyse décisionnelle. Bien que plus « simple » à mettre en œuvre, cette approche n'est pas adoptée dans ce cas-ci car elle ne permet pas réellement de dresser distinctement la frontière entre le traitement transactionnel et le traitement analytique des données (Vaisman et Zimanyi, 2014).

Dans le cas qui nous occupe, **il est capital de prendre le temps d'expliquer le processus ETL et le DW** mis en place car il est nécessaire, pour comprendre le *front-end* lié aux outils d'analyse (partie visible de l'iceberg), de comprendre le *back-end* (partie immergée de l'iceberg). Il est ainsi possible de se pencher sur la qualité des données brutes à disposition, sur les choix en matière de transformation de ces dernières et sur la conception du DW permettant de les intégrer.

Aborder la distinction entre data warehouse et data marts n'est par ailleurs pas vraiment approprié au vu des caractéristiques des données à disposition. Parler de data marts ne se justifie pas vraiment dans ce cadre parce qu'on n'observe pas véritablement de sous-ensembles de données qui soient propres à des départements ou à des groupes d'utilisateurs spécifiques. On fait dès lors aussi abstraction de la discussion opposant l'approche *top-down* de Inmon à l'approche *bottom-up* de Kimball (à savoir : faut-il dériver les data marts du DW ou l'inverse ?) (Vaisman et Zimanyi, 2014).

La conception du DW de BrokerRoad doit idéalement avoir pour objectif d'incorporer les 4 caractéristiques fondamentales des DW : être orienté sujet, intégré, évolutif dans le temps et non volatile (Vaisman et Zimanyi, 2014). L'idée est d'intégrer des données provenant de différentes sources afin de pouvoir se pencher sur les informations métier qui en émergent. Ces données sont accumulées de manière incrémentale au cours du temps, et une historisation se met ainsi en place.

Sachant cela, notons que les sources de données qui sont utilisées par la suite sont le fichier Excel fourni par BrokerRoad et d'autres sources extérieures visant à étoffer l'analyse (cf. section 3.3). Le fichier Excel étant statique, il n'est pas réellement possible d'intégrer de nouvelles données opérationnelles dans le DW à intervalles réguliers, autrement dit de répéter le processus ETL dans le temps. La problématique des dimensions à variation lente n'est d'ailleurs que brièvement abordée par la suite.

La **suite de ce chapitre** se penche sur la **modélisation du DW** et est déclinée en deux sections : d'une part, la conception du schéma conceptuel (appelé *MultiDim*) et, d'autre part, la conception du schéma logique qui en découle.

3.3 LE SCHÉMA CONCEPTUEL DU DATA WAREHOUSE

La ***figure* 3.1** présentée en fin de section modélise le **schéma conceptuel** de BrokerRoad. On y retrouve une table de faits incorporant les métriques propres à chacune des expéditions, ainsi que les dimensions non dégénérées suivantes : expéditeur, transporteur, équipement, type de trajet, cargaison, devise de la transaction, dates, temps et localisation. Le numéro d'expédition et les codes de statut (attribut de basse cardinalité) sont, eux, des dimensions dégénérées. Rappelons qu'en présence de dimensions à un attribut, on peut faire le choix d'injecter ce dernier directement dans la table de faits, sans avoir recours à une clé de substitution (cf. section 3.4).

Etant donné que BrokerRoad exerce ses activités aux USA, la langue utilisée pour développer les modèles de données est l'anglais.

La table de faits construite ici est une ***accumulated snapshot fact table***. Cette dernière cherche à capturer le processus d'affaire de BrokerRoad et est dès lors liée aux dimensions

temporelles à de nombreuses reprises, de manière à pouvoir enregistrer les différentes étapes du processus. Une entrée/ligne individuelle est insérée quand la transaction a débuté, c.-à-d. lorsque l'expéditeur a posté son offre et que l'ETL a été exécuté pour répercuter le changement dans le DW. Si le processus a évolué et que l'ETL a été exécuté en fonction, l'entrée/ligne est mise à jour. Il est par exemple question d'insérer la date à laquelle un transporteur a été trouvé. Cette mise à jour régulière des lignes existantes tranche avec les autres tables de faits de type *event, factless* et *snapshot* qui, elles, fonctionnent globalement sur base d'insertions sans mises à jour ultérieures ((Wisdomjobs.com, n.d.) ; (Kimball Group, n.d.)).

Les dimensions temporelles sont donc cruciales ici, et il est alors intéressant de justifier le choix de séparer le temps (hh:mm:ss) et les dates (aaaa-mm-jj) : il s'agit d'éviter d'avoir une seule table de taille considérable, c'est-à-dire une taille égale au produit cartésien du temps et des dates. Omettre le temps (hh:mm:ss) en ne conservant que les dates n'est, par ailleurs, aucunement recommandé étant donné les courts laps de temps qui séparent les étapes clés du processus.

Il est aussi utile de noter les 2 **hiérarchies** qui ont été construites : le calendrier décrit les dates (jours, semaines et notions de congés), les mois, les trimestres et les années, tandis que l'emplacement décrit les villes, les états et les pays.

La **nature des relations qui unissent les dimensions à la table de faits** est également à remarquer. La table de faits est en lien avec différentes dimensions, parmi lesquelles 3 sont des *role-playing dimensions,* c.-à-d. des dimensions qui sont reliées plusieurs fois à la table de faits. La dimension spatiale renseigne l'origine et la destination, tandis que les deux dimensions temporelles renseignent chacune 15 événements différents tels que l'enlèvement, la livraison et l'achèvement de la transaction.

Remarquons, par la même occasion, la cardinalité des relations qui unissent les différentes tables. On doit savoir que chaque fait est toujours relié à un élément de la dimension vers laquelle la relation pointe, et ce, en dépit du fait que des informations peuvent être manquantes, non applicables ou pas encore disponibles (ex. : pas de transporteur assigné jusqu'à présent). Le choix est fait de suivre la méthode préconisée notamment par le Kimball Group (n.d.) et de veiller à ce que les tables de dimensions disposent d'une entrée par défaut symbolisant l'absence d'information (pas d'expéditeur assigné, non applicable, etc.).

Contrairement aux autres tables, les 4 dimensions que sont le temps, la date, la devise et le type de trajet sont construites de toutes pièces par la suite. L'idée est que, lors du processus d'ETL, ces dernières sont construites et alimentées en données au travers de requêtes SQL uniquement (cf. section 4.2). Ceci permet de comprendre pourquoi les éléments de ces dimensions sont reliés à la table de faits par une cardinalité optionnelle (autre extrémité des relations) : certains éléments qui ont été créés ne sont probablement pas utilisés actuellement dans la table de faits comme, par exemple, une date ou un emplacement.

La majorité des attributs des dimensions ayant déjà été explicitée précédemment, notons simplement la nature et la provenance des **informations qui ont été ajoutées aux dimensions** du DW pour les rendre un tant soit peu plus consistantes. Les informations relatives aux expéditeurs et transporteurs, qui avaient été omises pour des raisons d'anonymat, ont été remplacées par des noms, e-mails et numéros d'entreprises nationaux légaux (*Enterprise Identification Number* aux USA) générés au travers du site web « fauxid.com ». Des informations ont également été ajoutées en ce qui concerne les types d'équipements, précisant ainsi quelles remorques étaient ou non réfrigérées et étaient ou non ouvertes. Par ailleurs, les informations liées aux différentes devises proviennent du site « investing.com », tandis que les données relatives aux codes des états et des pays, aux dates et aux heures sont inspirées de sources qui sont évoquées par la suite.

Au vu de la nature déstructurée du texte décrivant le contenu des cargaisons, il a aussi été entrepris d'en dériver une catégorie présumée au travers d'algorithmes de ***text mining*** qui sont résumés par la suite (cf. section 4.2). Le manque de rigueur dont ont fait preuve les employés en complétant le champ de description impacte néanmoins très fortement la qualité des résultats.

Le dernier point à évoquer est celui de certaines des particularités des **mesures de la table de faits**. Deux mesures ont été dérivées et sont symbolisées par le signe « / » : les dépenses totales des expéditeurs ainsi que les revenus totaux des transporteurs (cf. section 3.1). Ces mesures dérivées ont été incluses dans le modèle car on considère que l'on peut être amené à les utiliser régulièrement, et qu'il est dès lors utile de les calculer dès le départ. La commission du broker aurait également pu constituer une mesure dérivée puisqu'elle représente la différence entre le prix de base de l'expéditeur et celui du transporteur, mais elle était déjà fournie au départ (notons que les charges supplémentaires ne sont pas prises en compte dans le calcul de

la commission). D'autres mesures sont également dérivées lors de la phase des visualisations (cf. chapitre 5).

Enfin, soulignons que les mesures représentant le budget maximal des expéditeurs et le prix ayant permis de remporter d'éventuelles enchères sont considérées comme étant non additives et sont symbolisées par un « /+ » : il n'est vraisemblablement pas justifié de les résumer/agréger en utilisant l'addition, et ce, quelle que soit la dimension envisagée.

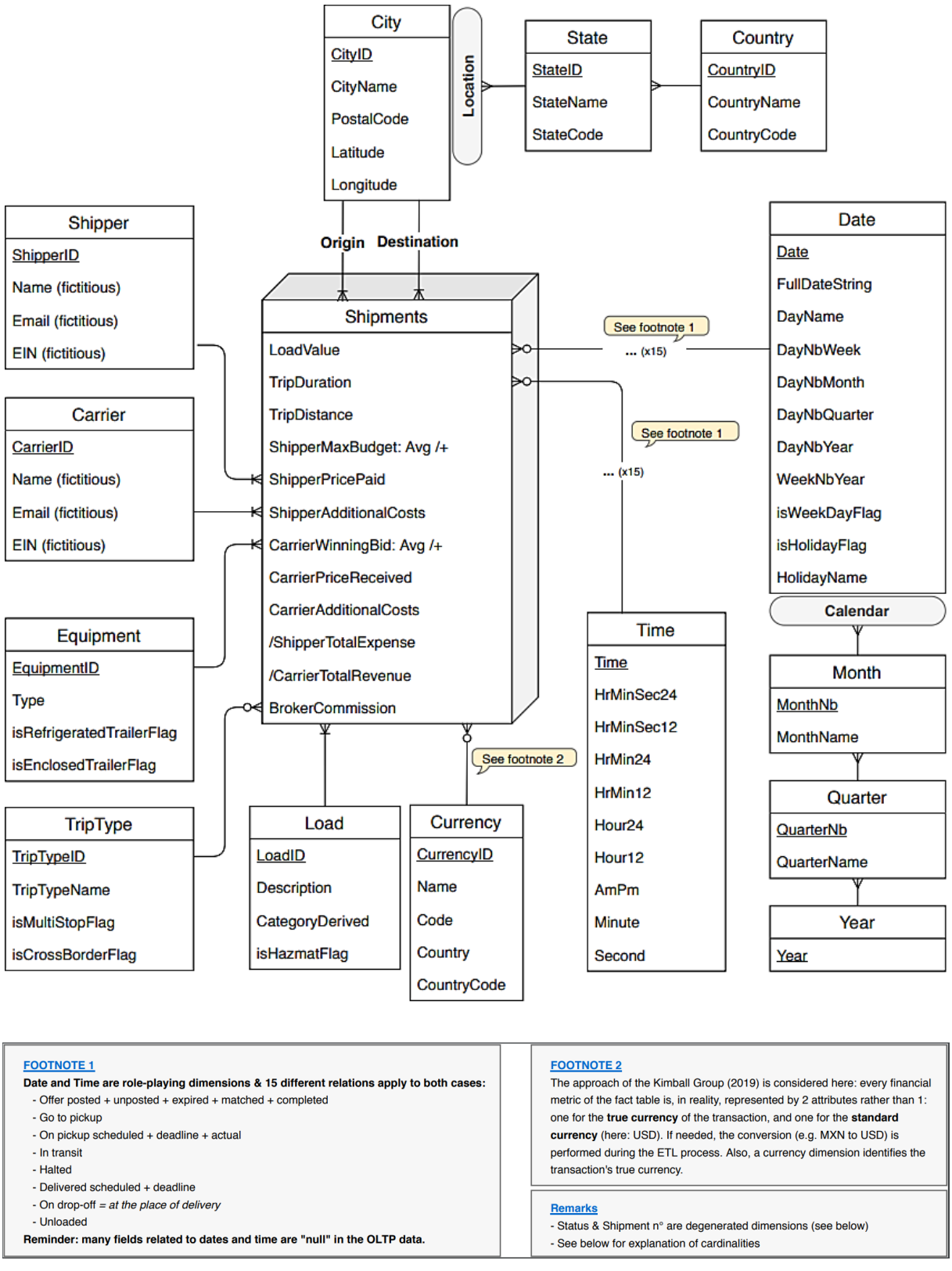


Figure 3.1 : schéma conceptuel

3.4 LE SCHÉMA LOGIQUE DU DATA WAREHOUSE

Cette dernière section a pour objectif de présenter le schéma logique de BrokerRoad. Celui-ci est dérivé du modèle MultiDim et est présenté en fin de section à la ***figure* 3.2**. Il est ici question de créer un modèle qui puisse ensuite être implémenté physiquement au travers du système de gestion de base de données (SGBD) qu'est Microsoft SQL Server. A ce stade, le but n'est pas de prendre en compte toutes les spécificités de SQL Server, mais bien de créer un modèle plus généralisable.

Notons que le modèle logique utilisé est de **type relationnel**. Et, étant donné que le modèle relationnel n'a pas été initialement pensé pour représenter des schémas multidimensionnels, force est de constater qu'il n'est pas des plus optimaux pour traduire le modèle MultiDim et suivre la philosophie des cubes OLAP. La suite Microsoft qui est utilisée par après permet néanmoins de créer des structures de type multidimensionnel au travers du module SSAS : un **cube OLAP** est ainsi mis en place pour BrokerRoad (cf. section 4.3 pour davantage de précisions).

Le schéma logique adopté est un **schéma relationnel en étoile** qui, contrairement au schéma en flocons de neige, ne modélise pas explicitement les hiérarchies des dimensions : chacune des dimensions n'est représentée que par une seule table. On dénormalise ainsi le modèle car les informations des niveaux hiérarchiques supérieurs sont mentionnées plusieurs fois dans la table ainsi formée : la hiérarchie est aplatie (*flattened*). Malgré les inconvénients qui en découlent, cette modélisation est utilisée pour le DW de BrokerRoad car la nature des hiérarchies spatio-temporelles exploitées ici s'y prête généralement bien (Vaisman et Zimanyi, 2014).

Il est question de mettre en place des clés étrangères et des clés de substitution (***surrogate keys***). Les deuxièmes ont pour objectif de pallier certains des inconvénients principaux des clés d'affaire présentes dans les systèmes opérationnels (*business keys*). Les clés de substitution deviennent alors les clés primaires des dimensions, tandis que les clés d'affaire sont souvent appelées à devenir des clés alternatives (***alternate keys***).

Insistons sur le fait qu'à l'inverse d'une clé d'affaire, une clé de substitution ne change normalement pas et ne se répète pas dans le temps. Si on ne travaille pas avec des clés de

substitution, on prend par exemple le risque de constater que l'on a attribué l'identifiant business d'un ancien expéditeur à un nouvel arrivant, ce qui impacte l'historisation : les transactions liées à l'ancien expéditeur seraient alors attribuées au nouvel expéditeur. Les clés de substitution permettent également la mise en place de dimensions à variation lente de type 2, pour lesquelles on ajoute une nouvelle ligne quand un changement est observé dans les données : différentes clés de substitution peuvent donc être attribuées à une même entité pour assimiler les différents changements qui la caractérisent. Par ailleurs, la performance est accrue en matière d'accès aux données puisque les clés de substitution sont de type entier, contrairement à certaines clés d'affaire telles que l'ID des expéditeurs.

Les seules dimensions qui n'ont pas recours aux clés de substitution factices sont les dimensions dégénérées et les dimensions temporelles. Les deuxièmes par exemple utilisent des entiers uniques de type « aaaammjj » et « hhmmss » comme clés primaires.

Les relations qui unissent les tables sont soit de type *one mandatory to many mandatory*, soit de type *one mandatory to many optional*. Du point de vue des cardinalités maximales, les relations sont donc de type *one to many* et il n'est alors pas nécessaire de créer des *bridge tables*, puisque l'on peut placer les **clés étrangères** dans les tables qui pointent vers, au plus, un élément d'autres tables.

Du point de vue de la **table de faits**, les montants standards (en USD) sont cette fois-ci affichés aux côtés des montants originaux de la transaction (en USD ou MXN) afin de respecter l'approche du Kimball Group expliquée précédemment à la *figure* 3.1.

Le numéro d'expédition n'étant attribué à une transaction qu'une fois le transporteur trouvé, il ne permet pas d'identifier de manière unique la totalité des transactions. Rappelons à ce titre que le fichier source Excel ne dispose d'aucune clé primaire pour distinguer les transactions. Le Kimball Group préconise alors d'inclure un identifiant de transaction de substitution dans la table de faits et, plutôt que de créer une clé composite avec l'ensemble des clés étrangères de la table, on utilisera ici cet identifiant de transaction de substitution comme clé primaire (Kimball Group, n.d.).

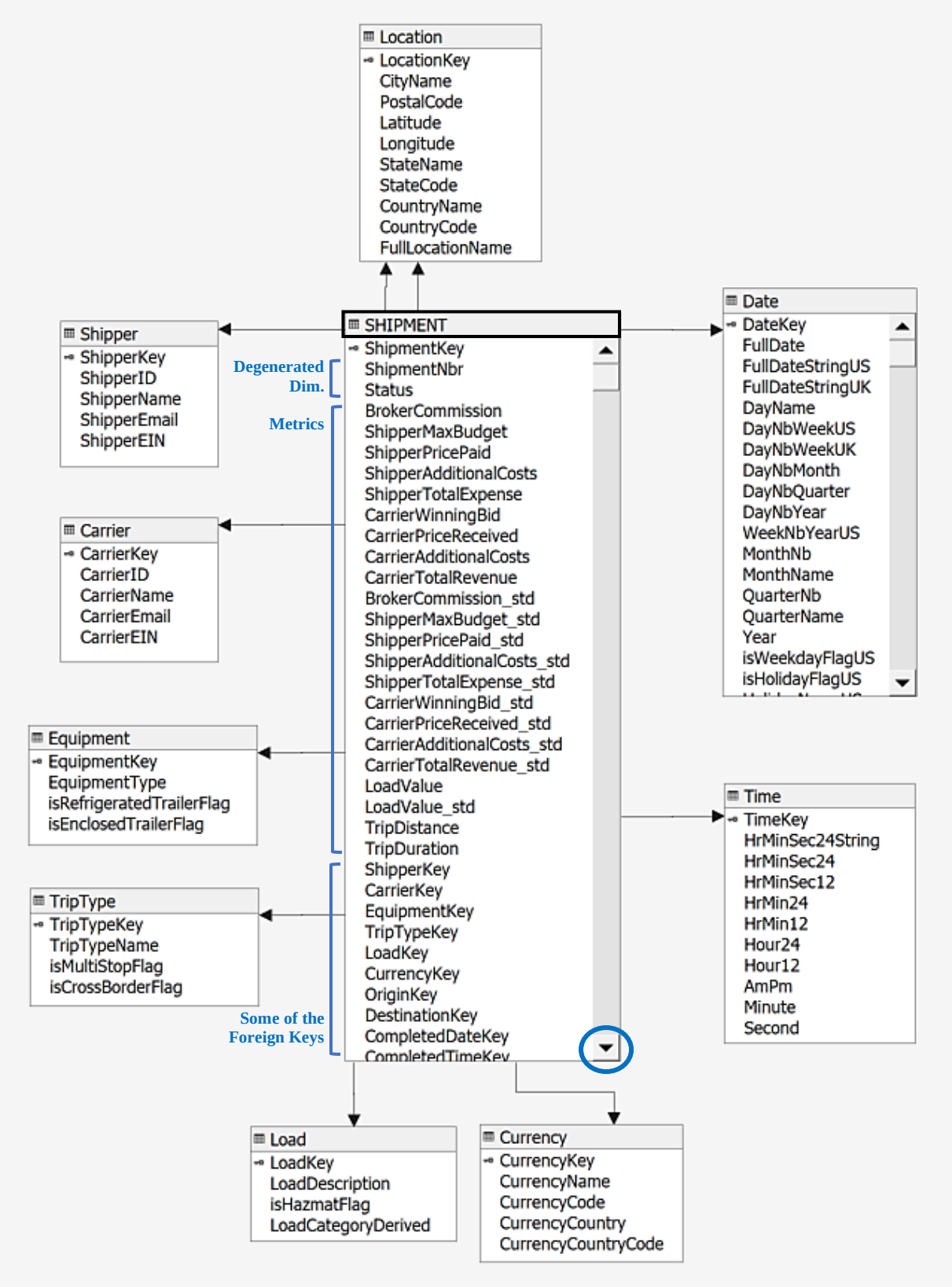


Figure 3.2 : schéma logique

Chapitre 4 : processus ETL et modélisation OLAP

4.1 L'ORCHESTRATION DU PROCESSUS

Le processus ETL a pour objectif d'extraire les données brutes, de les transformer et de les charger dans le DW. Le paradigme adopté ici pour ce faire est celui de l'architecture double couche. A titre d'information, notons qu'en architecture de DW à triple couche, il est même question de constituer un ensemble de données au niveau d'une couche que l'on appelle « réconciliée » : celle-ci matérialise les données opérationnelles une fois qu'elles ont été extraites, intégrées et nettoyées, et avant qu'elles ne soient chargées dans le DW.

Pour comprendre comment on passe des données brutes au DW, il est d'abord nécessaire d'appréhender le processus ETL d'un point de vue global. La ***figure* 4.1** est présentée dans ce but, puisqu'elle modélise l'**orchestration du processus ETL** au moyen de la notation BPMN (*Business Process Model and Notation*). Il est en fait question d'une tâche de contrôle composée d'une série de sous-tâches suivant un ordre d'exécution bien défini : les différentes sous-tâches sont reliées entre-elles par des flèches de séquence qui indiquent la précédence que certaines sous-tâches ont sur d'autres.

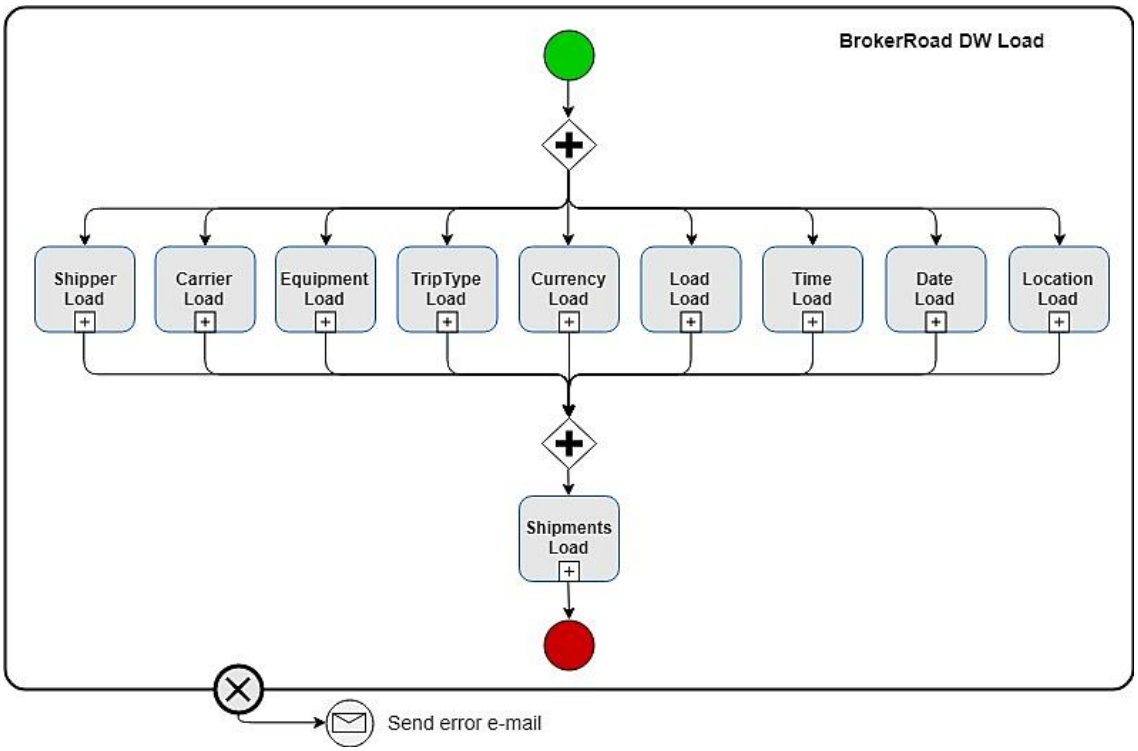


Figure 4.1 : orchestration du processus ETL

Les tables des 9 dimensions de BrokerRoad sont ainsi chargées avant la table de faits puisqu'il est nécessaire de charger les dimensions auxquelles se réfèrent les clés étrangères de la table de faits avant de charger cette dernière. La dernière *parallel gateway*, symbolisée par un « ◊ » englobant un « + », indique d'ailleurs que toutes les sous-tâches liées aux dimensions doivent être terminées avant le chargement de la table de faits. Remarquons que, puisque les 3 hiérarchies ont été condensées, il n'est pas nécessaire d'établir une relation de précédence entre les tables « parent » et les tables « enfant », étant donné qu'en l'occurrence les tables « parent » n'existent plus.

La *figure* 4.1 décrit donc un processus d'orchestration assez simple. Le flux de contrôle de l'ETL qui est mis en œuvre en pratique au travers du logiciel Microsoft SSIS est, quant à lui, plus complexe.

La ***figure* 4.2 présentée en fin de section 4.2** décrit ainsi la mise en œuvre du processus ETL avec SSIS. Il a été jugé pertinent de présenter cette figure car elle s'avère **plus détaillée que le processus BPMN théorique**. On constate en effet que certaines des sous-tâches présentées précédemment sont maintenant scindées en plusieurs sous-tâches, appelées *tâches de flux de données* ou *tâches d'exécution de script SQL*. Les choix qui sont faits ici ont été pensés de manière à créer un processus SSIS qui soit plutôt explicite et cohérent, mais d'autres configurations auraient pu être envisagées. L'objectif des quelques remarques qui suivent est donc d'insister sur les éléments clés traités ici.

En ce qui concerne les dimensions, les tâches d'exécution de script SQL ont pour objectif de vérifier que les tables soient correctement initialisées dans Microsoft SQL Server. Lors de la première exécution de l'ETL, ces tâches permettent de créer et configurer les tables des dimensions en amont des tâches de chargement des données. D'autre part, dans le cas des dimensions créées manuellement que sont la date, le temps, le type de trajet et la devise (cf. section 4.2), les scripts SQL permettent également de charger des données dans les tables.

Dans le cas des dimensions qui sont peuplées au travers de données opérationnelles internes à l'entreprise, on fait appel à des tâches de flux de données qui permettent d'incorporer certaines des données du fichier Excel dans le DW de BrokerRoad. Notons également que deux tâches de flux de données sont mises en exergue pour insister sur le fait que l'on a eu recours à des données externes en vue d'ajouter des renseignements factices aux dimensions des expéditeurs et des transporteurs (nom, e-mail et n° d'entreprise).

On a aussi veillé à respecter l'approche du Kimball Group (n.d.) en incorporant une entrée par défaut dans chaque dimension, et ce, pour s'assurer que les clés étrangères de la table de faits ne soient jamais *null* et pointent toujours vers les dimensions, même en l'absence d'information (surrogate keys = -1 ; champs = 'nom spécifié', 'n/a'…).

En ce qui concerne la table de faits, un cas de figure similaire est applicable : l'utilisation de tâches d'exécution de script SQL pour une initialisation éventuelle et l'utilisation de tâches de flux de données pour l'approvisionnement en données. L'activation des clés étrangères est, elle, permise au travers d'un script SQL.

Notons également que le *SQL Server Agent* peut être utilisé au sein de SQL Server Management Studio pour programmer les exécutions successives du processus ETL et, ainsi, automatiser l'approvisionnement en données du DW. On parle ici de programmer des *jobs*.

4.2 LES SOUS-TÂCHES DU PROCESSUS

L'orchestration du processus ETL ayant été explicitée, il est question à présent de se pencher sur certaines de ses sous-tâches. L'objectif n'est pas de s'étendre outre mesure sur le sujet, mais plutôt d'étudier d'un peu plus près certains des aspects les plus caractéristiques du processus établi avec SSIS. On explicite ainsi certains des choix de modélisation, sans pour autant s'attarder sur les avantages et inconvénients de ces derniers, sans quoi on risquerait de sortir du cadre du sujet de ce mémoire. Il est donc question, dans un premier temps, de se pencher sur les tâches d'exécution de script SQL et, dans un deuxième temps, sur les tâches de flux de données.

Du point de vue **des tâches d'exécution de script SQL**, le choix qui a été fait en matière d'initialisation des dimensions implique de suivre deux étapes : il faut d'abord vérifier que la table n'a pas encore été construite et, ensuite, la créer si nécessaire, en veillant à déclarer une clé de substitution jouant le rôle de clé primaire et incrémentée par pas de 1 en partant de 1. L'éventuelle clé business/alternative et les attributs de la dimension sont ensuite renseignés avec leur type et leurs contraintes éventuelles (not null, unique, etc.). Voici un extrait de code générique :

```
---------------------------------------
IF NOT EXISTS (SELECT * FROM sysobjects WHERE name='MyDimTable' and xtype='U')
BEGIN
CREATE TABLE MyDimTable (
MyDimTableKey int PRIMARY KEY IDENTITY (1,1),
MyDimTableBusinessID int NOT NULL,
Attribute1 …,
Attribute2 …,
…)
END
---------------------------------------
```

Les tâches SQL permettent également d'approvisionner en données des dimensions standards telles que la date, le temps et la devise. Ces types de scripts étant abondants dans la littérature et d'une longueur significative, il n'a pas été jugé utile d'en présenter la structure ici. D'autre part, le script permettant d'insérer l'entrée symbolisant l'absence d'information au sujet d'une dimension suit la structure de l'exemple fourni ci-dessous. L'idée est, comme on le constate, de vérifier à chaque reprise si l'entrée est déjà présente et, si ce n'est pas le cas, de l'insérer. Notons que l'on pourrait envisager d'exécuter l'insertion au moment de la création de la table, sans devoir effectuer de vérification à chaque exécution ultérieure de l'ETL.

```
---------------------------------------
IF NOT EXISTS (select * from Carrier where CarrierKey=-1)
BEGIN
SET IDENTITY_INSERT Carrier  ON
INSERT INTO Carrier (CarrierKey, CarrierID, CarrierName, CarrierEmail, CarrierEIN) VALUES (-1,'No carrier assigned','No carrier assigned','No carrier assigned','No carrier assigned')
SET IDENTITY_INSERT Carrier  OFF
END
---------------------------------------
```

Il faut dresser également quelques constats en ce qui concerne les **tâches de flux de données**. Notons qu'il n'a pas été jugé opportun de présenter ici la totalité de ces tâches de flux, sous-jacentes au processus d'orchestration présenté à la *figure* 4.2 : la plupart de celles-ci n'apportent en effet pas une information particulièrement enrichissante au lecteur (à la lectrice). Certains exemples caractéristiques de ces dernières sont néanmoins mis à disposition du lecteur (de la lectrice) en ***annexe* A-3**. D'autres exemples comportant certaines particularités sont présentés ci-dessous.

Un constat important doit être dressé en matière d'insertion et de mise à jour des données des dimensions. Le fichier Excel de BrokerRoad étant statique, il n'est pas possible de constater

une évolution des données dans le temps et, donc, de répéter fidèlement le processus ETL. Dans un DW disposant d'une *accumulated snapshot fact table*, cet aspect est d'autant plus impactant : on ne peut en effet pas mettre à jour les faits au fur et à mesure de l'avancement du processus en ajoutant, par exemple, des informations relatives aux dates.

Le choix a par ailleurs été fait de ne pas recourir explicitement à des techniques telles que celle des dimensions à variation lente (***Slowly Changing Dimensions*** ou **SCD**). Il est néanmoins utile d'illustrer, au travers du simple exemple de la dimension des équipements/remorques, les techniques de modélisation de SCD qui auraient pu être mises en place pour que l'ETL soit plus complet (cf. *figure* 4.3). On a ici recours à une SCD de type 2.

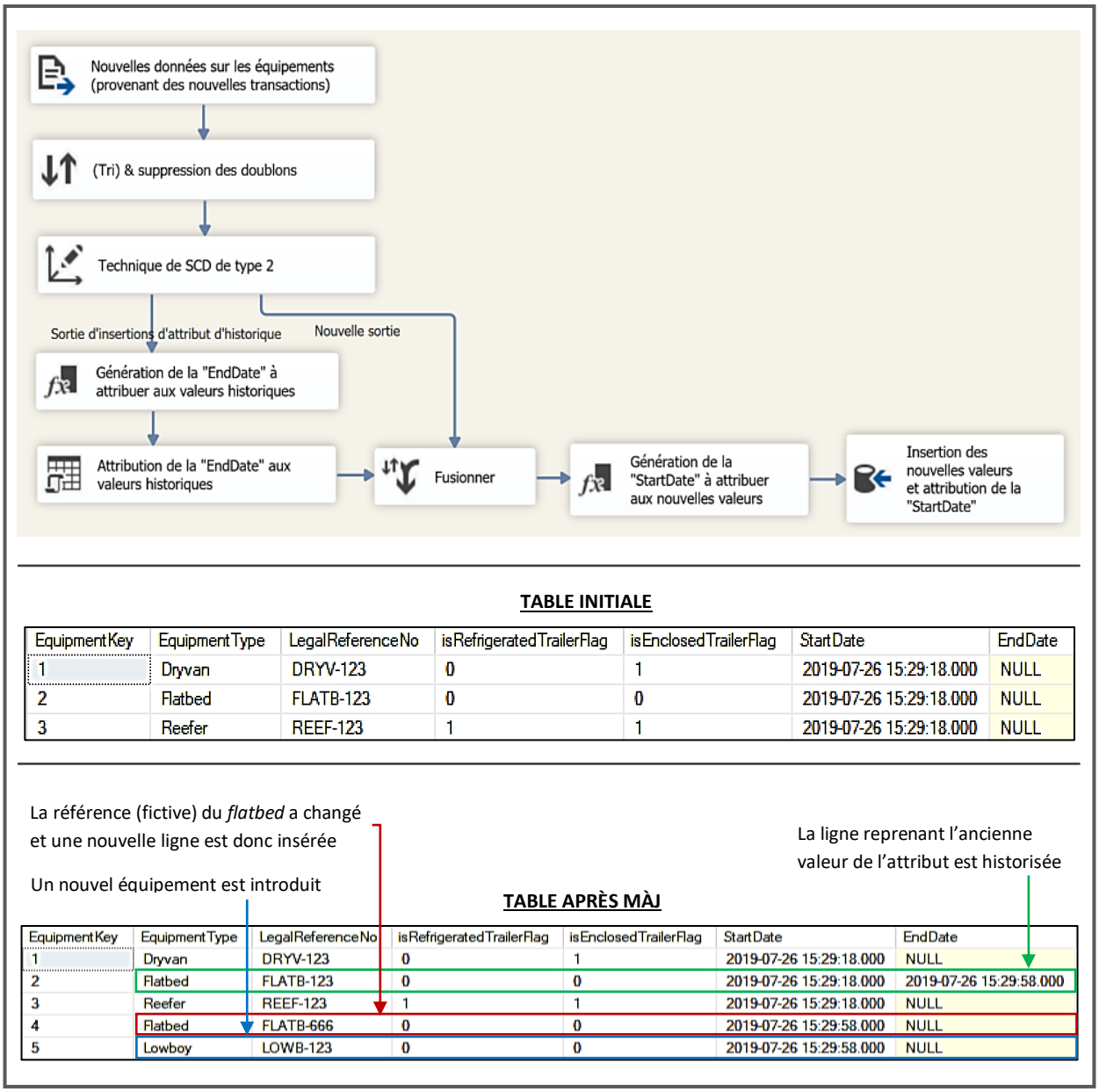


TABLE INITIALE

EquipmentKey	EquipmentType	LegalReferenceNo	isRefrigeratedTrailerFlag	isEnclosedTrailerFlag	StartDate	EndDate
1	Dryvan	DRYV-123	0	1	2019-07-26 15:29:18.000	NULL
2	Flatbed	FLATB-123	0	0	2019-07-26 15:29:18.000	NULL
3	Reefer	REEF-123	1	1	2019-07-26 15:29:18.000	NULL

La référence (fictive) du *flatbed* a changé et une nouvelle ligne est donc insérée

Un nouvel équipement est introduit

La ligne reprenant l'ancienne valeur de l'attribut est historisée

TABLE APRÈS MÀJ

EquipmentKey	EquipmentType	LegalReferenceNo	isRefrigeratedTrailerFlag	isEnclosedTrailerFlag	StartDate	EndDate
1	Dryvan	DRYV-123	0	1	2019-07-26 15:29:18.000	NULL
2	Flatbed	FLATB-123	0	0	2019-07-26 15:29:18.000	2019-07-26 15:29:58.000
3	Reefer	REEF-123	1	1	2019-07-26 15:29:18.000	NULL
4	Flatbed	FLATB-666	0	0	2019-07-26 15:29:58.000	NULL
5	Lowboy	LOWB-123	0	0	2019-07-26 15:29:58.000	NULL

Figure 4.3 : illustration du fonctionnement des SCD de type 2

La tâche de flux liée au chargement de la dimension des cargaisons requiert également une attention toute particulière. Il a été mentionné précédemment qu'au vu de la nature déstructurée du texte décrivant le contenu des cargaisons, il est utile d'en dériver une catégorie présumée au travers d'un algorithme de **text mining**. Le schéma de la tâche de flux dont il est question ici est fourni en *annexe* A-3.1. Mais, sans avoir réellement besoin de consulter ce schéma, concentrons-nous sur le fonctionnement global de l'algorithme.

Une *tâche d'extraction de terme* détecte les termes qui apparaissent régulièrement dans les champs de description des cargaisons et met ainsi en évidence des catégories potentielles : on choisit un seuil de fréquence minimale et une longueur maximale pour les chaînes de caractères à mettre en évidence. Cette fouille de texte fait appel à un dictionnaire de *stop words* : ce dernier recense l'ensemble des mots les plus communs en anglais (USA) et en espagnol (MX) pour éviter de prendre en compte des termes tels que « the » et « and ». Une fois ces catégories déterminées, une *tâche de recherche floue* cherche à assigner chaque description à sa catégorie potentielle. Le manque de rigueur dont ont fait preuve les employés en complétant les champs de description, la diversité des langues utilisées et le fait que certaines catégories ne soient que peu présentes impactent néanmoins fortement la qualité des résultats.

Rappelons que d'autres schémas sont également mis à la disposition du lecteur (de la lectrice) en *annexe* A-3.

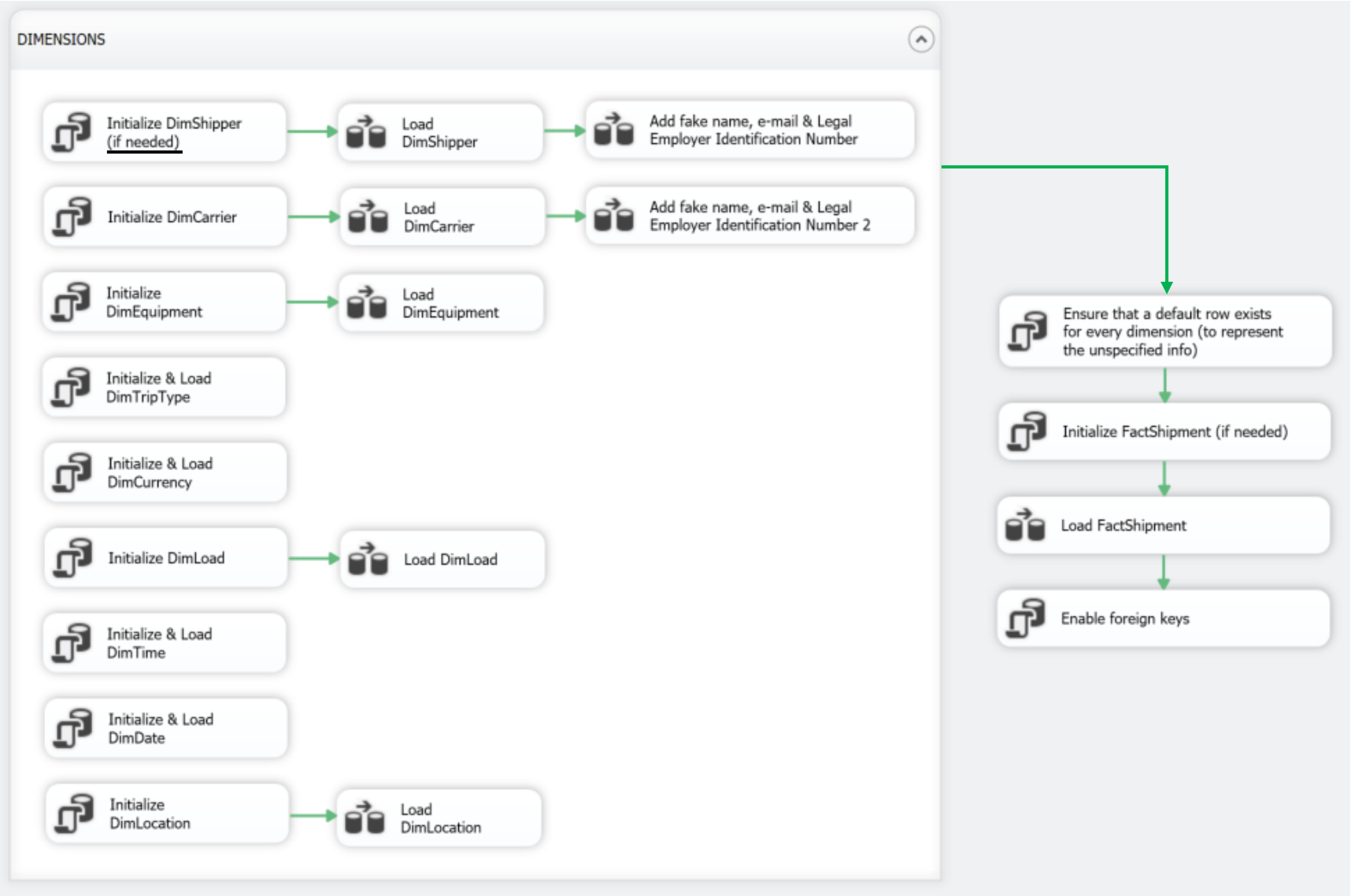


Figure 4.2 : mise en œuvre de l'ETL avec SSIS

4.3 LA MODÉLISATION DU CUBE OLAP

Le fait de stocker des données aux caractéristiques multidimensionnelles dans une base de données relationnelle implique forcément des inconvénients, puisque cette dernière n'est, au départ, pas conçue pour le stockage de ce type de données. Avant de se plonger dans le dernier et très important chapitre des visualisations, il est donc nécessaire de **comprendre les choix qui s'offrent concrètement à nous en matière de structures de données physiques** et de les avoir bien en tête lors de la mise en place de l'outil de visualisation (Power BI).

Jusqu'à présent, on a eu recours au paradigme du *Relational-OLAP* (**ROLAP**) car les données sont stockées dans une base de données relationnelle. Dans le cas du *Multidimensional-OLAP* (**MOLAP**), on stocke plutôt ces données dans une forme d'hypercube à n-dimensions. L'avantage du MOLAP est qu'il permet de créer des requêtes plus performantes, puisqu'il est pensé pour respecter la philosophie multidimensionnelle et ne nécessite donc pas de faire des jointures de manière intempestive entre les tables. Mais il consomme cependant assez bien d'espace (pensons à toutes les cases du cube qui sont vides, c.-à-d. toutes les combinaisons de valeurs de dimensions pour lesquelles il n'existe pas de fait). C'est pour cette raison que l'on travaille régulièrement en *Hybrid-OLAP* (**HOLAP**), qui est une forme de compromis entre performance de calcul et optimisation de l'espace de stockage. Les agrégations qui sont fréquemment calculées sont stockées sous forme multidimensionnelle, tandis que le reste est stocké sous forme relationnelle (Vaisman et Zimanyi, 2014).

SSAS nous permet de mettre en place un hypercube OLAP pour BrokerRoad et nous offre ainsi des capacités analytiques avancées, principalement au travers du langage MDX (*MultiDimensional eXpressions*). En ayant recours à une base de données hébergée sur un serveur analytique, on a en effet la possibilité d'interroger les données directement au travers d'un langage intrinsèquement multidimensionnel. Ce qui n'est pas le cas lorsque l'on interroge une base de données relationnelle car, quel que soit le langage initialement utilisé pour interroger les données, un moteur multidimensionnel devra toujours traduire les requêtes en langage SQL ou SQL-OLAP. Ayons bien à l'esprit que le langage SQL-OLAP est une forme d'adaptation du langage SQL qui permet de remédier à quelques-unes des lacunes du langage SQL en matière d'analyse multidimensionnelle (Vaisman et Zimanyi, 2014).

Ces aspects ayant maintenant été remis dans leur contexte, il est utile d'illustrer le cube *Shipments* de SSAS à l'aide d'un exemple. La *figure* 4.4 présente donc une requête en MDX.

La requête présentée ci-après permet enfin d'obtenir un aperçu du potentiel stratégique des nouvelles structures de données en répondant aux questions suivantes : en considérant les expéditions en partance du Texas et complétées en 2018, quels expéditeurs génèrent les plus grandes commissions pour BrokerRoad ? Quels sont les revenus qu'ils génèrent et la marge de profit (commissions/revenus) en % ?

En considérant les expéditions en partance du Texas et complétées en 2018, quels expéditeurs génèrent les plus grandes commissions pour BrokerRoad ? Quels sont les revenus qu'ils génèrent et la marge de profit (commissions/revenus) en % ?

```
WITH
Member [Total Commission ($)]
as [Measures].[Broker Commission],
format_string = '### ### ###'

Member [Total Revenue ($)]
as [Measures].[Shipper Total Expense],
format_string = '### ### ###'

Member [Profit margin (%)]
as IIF([Measures].[Shipper Total Expense]<>0,
       Round(([Measures].[Broker Commission]
       /[Measures].[Shipper Total Expense])*100,0),
       NULL)

SELECT
NON EMPTY { [Total Revenue ($)],
            [Profit margin (%)],
            [Total Commission ($)] } ON COLUMNS,
NON EMPTY ( ORDER ( [Shipper].[Shipper Name].[Shipper Name].ALLMEMBERS,
                    [Total Commission ($)], DESC ) ) ON ROWS
FROM [Shipments DW]
WHERE ( [Origin].[State Name].&[Texas], [Completed Date].[Year].&[2018])
```

Création des alias des métriques + définition du format (espaces entre les milliers)

Création de l'alias + définition de la métrique temporaire

Choix des colonnes (métriques) + choix des lignes (expéditeurs) + tri décroissant sur base des commissions + filtres

Arrondi

	Total Revenue ($)	Profit margin (%)	Total Commission ($)
Cummerata Inc	556 674	12	64 945
Bailey-Bechtelar	446 947	9	39 980
Goldner-Botsford	463 839	8	37 416
Larson-Pacocha	87 452	9	8 092
Mayer, VonRueden and Streich	254 224	3	7 597
Wunsch, Osinski and O'Hara	54 300	13	7 250
Upton, Tromp and Littel	100 303	6	6 228

Figure 4.4 : illustration du potentiel stratégique du client front-end OLAP

Chapitre 5 : visualisations

Bon nombre de constats ont été dressés au sujet des visualisations dans la revue de la littérature et tout au long de l'analyse. L'objectif est donc ici de les prendre en considération en modélisant les visualisations de BrokerRoad, tout en ajoutant un certain nombre d'éléments qu'il semblait utile d'analyser de manière directement appliquée.

Les éléments qui viennent compléter les concepts étudiés précédemment sont les suivants : les exigences fonctionnelles et non-fonctionnelles auxquelles l'outil de visualisation doit répondre, la méthodologie qui guide le choix des types de visualisations (histogramme, *treemap*, flow map, etc.), la définition des KPIs et l'utilisation d'une « balanced scorecard ». Le but étant de compléter les dashboards présentés en section 5.5.

5.1 LES EXIGENCES FONCTIONNELLES ET NON-FONCTIONNELLES DE L'OUTIL

L'outil qui a été choisi ici pour développer les visualisations est Microsoft Power BI. Le marché de la BI dispose de nombreux outils de visualisation et le choix n'est pas toujours évident. Il est vrai que l'on a utilisé la suite Microsoft jusqu'à présent et il semble donc logique de continuer sur cette voie. On peut néanmoins tout à fait faire le choix de recourir à d'autres outils de visualisation si la nature des données et des exigences le justifie.

Le *Magic Quadrant* de Gartner dans sa version 2019 est ici utilisé pour se pencher sur les exigences auxquelles doit répondre un outil tel que Power BI. Notons qu'au-delà du fait que Power BI dispose de nombreux modules permettant de développer des visualisations fidèles aux exigences fonctionnelles de BrokerRoad, il répond également à de nombreuses exigences non-fonctionnelles. On a donc pris en compte non seulement les **exigences fonctionnelles**, mais aussi les **exigences non-fonctionnelles**, **ayant trait à la qualité et à l'ergonomie** de l'outil (performance, maintenabilité, facilité d'utilisation, interface *user-friendly*, etc.).

Froese et Tory (2016) sont de ceux qui rappellent que, puisque les utilisateurs des visualisations BI sont principalement des cadres moyens ou supérieurs, ils ont très souvent une connaissance avancée des indicateurs liés à leur business, ce qui facilite largement leur compréhension des figures. Ces auteurs insistent également sur le fait que, bien que les cadres

aient relativement peu de temps à consacrer à l'analyse des visualisations, ils ont souvent déjà une expérience significative en matière de dashboards.

Le choix des outils de visualisation est, quoi qu'il en soit, capital puisque leur qualité et leur bonne ergonomie doivent être assurées pour que les employés de BrokerRoad soient en mesure, par exemple, d'explorer et de personnaliser les interfaces sans avoir réellement besoin de suivre des formations en lien avec l'outil (Weiner, Balijepally et Tanniru, 2015).

Gartner (2019) considère que Microsoft est un des leaders du marché dans le domaine des plateformes de BI, comme en témoigne son **Magic Quadrant** consultable en *annexe* A-4. Pour Gartner, les leaders sont ceux qui font preuve d'une capacité d'exécution élevée et d'une vision globale : ils doivent parvenir à mettre en œuvre leur vision actuelle au travers de l'outil de visualisation et être bien positionnés pour l'avenir.

L'entreprise américaine de conseil et de recherche met en garde contre les différences significatives qui peuvent exister entre les différentes versions de Power BI mais souligne également de **nombreux avantages** : les faibles prix en vigueur (2ème raison la plus citée par les entreprises ayant choisi Power BI), la bonne expérience d'achat (dans le top 3 des meilleurs vendeurs), la maintenabilité, la robustesse, la facilité d'utilisation pour des analyses plus complexes et des fonctionnalités complètes, avancées et visionnaires (Gartner, 2019).

Bien que Tableau, Qlik et ThoughtSpot soient également des leaders du marché, et malgré certaines faiblesses de Power BI, les aspects énoncés ci-dessus et le fait que Power BI soit le leader le mieux positionné sur le Magic Quadrant en 2019 confortent le choix effectué.

5.2 LES CHOIX EN MATIÈRE DE MODÉLISATION DES VISUALISATIONS

Dans la revue de littérature (cf. sous-section 1.3.2), une checklist avait été construite. Elle reprenait, rappelons-le, les éléments clés permettant d'établir une synthèse des constats dressés alors : pictogrammes familiers, densité visuelle, mélange entre visualisations courantes et moins courantes, aspects naturels, thèmes de couleurs pour les daltoniens, absence de 3D, possibilité de personnaliser les interfaces… Il est donc question ici de veiller à ce que **les dashboards présentés en section 5.5 incorporent ces éléments clés**.

Par ailleurs, c'est l'approche de Froese et Tory (2016) qui est adoptée ici. **Les employés de BrokerRoad doivent donc pouvoir consulter différents sous-ensembles d'informations et de KPIs répondant chacun à différentes questions bien spécifiques, tout en ayant la possibilité d'adapter quelque peu les interfaces en fonction des besoins du moment** (cf. section 1.3.1).

Le **choix des figures** de visualisation n'est bien évidemment pas anodin. Chaque sous-ensemble de données a une nature et une taille qui lui sont propres, et c'est précisément ce qui doit guider les décisions en matière de visualisations. Pour choisir une figure, il est nécessaire d'analyser le quoi, le comment, le pourquoi et le combien de ce choix (Dumas, 2017).

Le « quoi » fait référence au type de données à traiter : variable quantitative, variable catégorique, variable dérivée… Le « comment » reflète les caractéristiques principales de la figure envisagée et la manière dont elle projette les données dans l'espace. Le « pourquoi » détaille la tâche principale à laquelle la visualisation peut répondre : étudier la corrélation entre les données, les classifier… Enfin, le « combien » est lié à l'échelle : il s'agit de déterminer la quantité de données qui peut être projetée sans surcharger la figure et en conservant son potentiel explicatif (Dumas, 2017).

Il n'a pas été jugé pertinent de développer ici chacune des raisons ayant motivé les choix des différentes figures présentées dans les dashboards. Cela alourdirait inutilement la lecture. Un exemple est toutefois fourni afin d'illustrer les concepts énoncés.

Une des figures qui est utilisée dans les dashboards de BrokerRoad est la carte proportionnelle (*treemap*). Elle représente des données hiérarchiques au travers de rectangles imbriqués ayant une aire proportionnelle à la valeur numérique de l'élément auquel ils sont liés. Les données sont de type hiérarchique (le quoi). Elles sont encodées au travers d'un *layout* linéaire plutôt que sous forme d'arbre (le comment) et sont disposées de manière à étudier les différents niveaux de la hiérarchie en insistant sur les éléments prédominants (le pourquoi). Le nombre d'éléments que l'on peut afficher par niveau hiérarchique est significativement élevé, mais l'inconvénient majeur est que l'on occulte les nœuds dont la valeur numérique est faible (le combien). La *treemap* est recommandée pour les hiérarchies peu profondes, c'est-à-dire celles de 3 ou 4 niveaux au maximum. Pour des hiérarchies plus profondes, l'aspect plus naturel des structures en arbre (nœuds liés par des arêtes) est privilégié (Dumas, 2017).

Dans le cas de BrokerRoad, on comprend pourquoi des dimensions intrinsèquement hiérarchiques à 2-3 niveaux telles que celle de l'emplacement (pays → état → ville) se prêtent bien à ce type de figure pour étudier des métriques telles que les commissions.

5.3 LES INDICATEURS CLÉS DE PERFORMANCE

Les **Key Performance Indicators** sont essentiels. Ils ont été abordés précédemment, mais il est important à présent de préciser certains aspects. Parmenter (2010, p.2) explique qu'ils représentent « un ensemble de mesures qui se concentrent sur les aspects de la performance organisationnelle qui sont les plus critiques pour le succès actuel et futur de l'organisation » : un KPI est une mesure porteuse de la stratégie de l'entreprise. Il est donc question de **poser certains des concepts pris en compte par la suite pour construire les KPIs des visualisations**, et de noter que **certains critères importants sont difficilement applicables au cas de BrokerRoad**.

Les KPIs sont « clés » car ils apportent une contribution majeure au succès/échec d'un projet, et sont « indicateurs de performance » car ils cherchent à maximiser le ratio résultats/moyens mis en place au regard d'un objectif spécifique, et ce, dans l'état actuel et futur des choses (Parmenter, 2010).

Parmenter (2010) explique que leurs **caractéristiques principales** sont les suivantes : ils doivent être régulièrement mesurables, intuitifs, contextualisés et actionnables par les dirigeants et les cadres supérieurs, voire moyens, de l'entreprise (indication des actions à entreprendre), ils doivent identifier clairement les personnes qui sont tenues de réagir à la situation et, enfin, ils doivent avoir un impact positif et considérable sur l'organisation.

Eckerson (2009) ajoute qu'ils doivent être peu nombreux (aux alentours de 10), *drillable* (un indicateur peut en cacher un autre), alignés les uns par rapport aux autres et validés sur le terrain (L'indicateur est-il effectif ? Dispose-t-on des informations nécessaires ? Les employés peuvent-ils falsifier l'indicateur pour qu'il corresponde à leurs désirs ? …).

Ces différents éléments sont donc pris en compte ici pour choisir les KPIs de BrokerRoad. On comprend néanmoins que, dans le cadre de ce mémoire, la vérification du caractère actionnable des KPIs ainsi que leur validation sur le terrain ne sont pas envisageables.

On prête également attention au fait que les KPIs doivent être équilibrés (Eckerson, 2009). Il s'agit premièrement de mettre en place des indicateurs stratégiques qui sont focalisés sur du moyen/long terme, mais également de créer quelques KPIs d'ordre plus tactique voire opérationnel : c'est pourquoi un dashboard d'analyse *on the fly* a été développé pour donner un aperçu des expéditions en cours ou complétées dernièrement chez BrokerRoad (cf. section 5.5). Deuxièmement, il faut être en mesure de combiner les indicateurs historiques (*lagging KPIs*) aux indicateurs avancés (*leading KPIs*) : certains indicateurs des dashboards de BrokerRoad sont en effet historiques, tandis que d'autres ont davantage trait aux performances actuelles et futures (cf. sections 5.4 et 5.5). Enfin, notons que les indicateurs peuvent être basés sur des mesures qualitatives et/ou quantitatives.

La complexité de la mise en place des KPIs qui nous occupe ici réside principalement dans le fait qu'il est difficile de trouver des standards qui soient facilement généralisables à toute situation : les KPIs sont spécifiques aux compagnies et doivent être le reflet des stratégies actuelles et futures. Une partie significative de l'analyse s'est d'ailleurs déjà focalisée sur ces problématiques et leur impact sur le cas d'étude de BrokerRoad.

Quoi qu'il en soit, la **définition d'un KPI** consiste normalement à préciser son nom, sa valeur actuelle, sa valeur cible, son statut (ex. : « dans le vert »), sa tendance et, au besoin, sa variance et son pourcentage de variance.

Ayant ces constats en tête, on pourra appréhender plus facilement les visualisations présentées à la section 5.5.

5.4 LES DASHBOARDS ET LES SCORECARDS

Les outils d'analyse, de reporting et de monitoring sont si variés qu'il n'est ni possible ni pertinent de tous les passer en revue ici. Dans cette dernière partie du travail, il est question de combiner les deux outils de monitoring que sont les *dashboards* et les *balanced scorecards* puisqu'ils sont utiles dans le cas qui nous occupe. Rappelons aussi qu'en matière d'outils d'analyse, un cube OLAP avait été déployé en section 4.3 et un exemple de requête MDX avait alors été proposé.

Les **dashboards** ayant déjà été étudiés, mentionnons simplement ici qu'il s'agit d'outils de visualisation qui permettent, au travers d'une interface dynamique, d'afficher des KPIs, des métriques et d'autres éléments importants pour l'organisation.

Les **balanced scorecards** nous intéressent également car, comme le disent Kaplan et Norton (1996, p. 18), elles « transcrivent la mission et la stratégie d'une organisation en un ensemble complet de mesures de performance […] ». Une balanced scorecard permet de monitorer les KPIs au travers des **4 (ou 5) perspectives** suivantes : financière, orientée client, orientée processus internes, orientée apprentissage et croissance (Kaplan et Norton, 1996) et, éventuellement, orientée durabilité. Cette dernière est de plus en plus mentionnée au travers des *sustainability balanced scorecards* (Naro et Noguéra, 2008).

La perspective financière est liée à la relation avec les actionnaires et à la manière dont on parvient à transcrire leurs objectifs financiers en stratégie. Les informations qui sont à notre disposition, en ce qui concerne BrokerRoad, nous indiquent que des KPIs historiques tels que les revenus, les commissions et les marges de profit ainsi que leur croissance sont cruciaux.

Du point de vue de la perspective client, il s'agit globalement pour BrokerRoad d'analyser la mesure dans laquelle elle parvient à satisfaire ses clients, à les garder et à en acquérir de nouveaux. On peut étudier des KPIs historiques tels que la fréquence des expéditions par client, le nombre de clients et les retards ayant impacté les clients (métrique qui a également trait aux processus internes). On pourrait également envisager des indicateurs avancés (*lead indicators*) qui seraient en lien avec des études de satisfaction, mais ces derniers ne sont malheureusement pas disponibles (cf. section 2.3).

La perspective des processus internes est principalement liée à l'excellence des processus d'affaire établis par BrokerRoad pour satisfaire ses actionnaires et ses clients : il s'agit de KPIs historiques liés à des aspects tels que le nombre de transactions annulées, mais aussi d'indicateurs avancés ayant trait à l'analyse de l'efficacité du cycle de livraison.

La perspective de l'apprentissage et de la croissance compare les objectifs stratégiques de BrokerRoad, formulés au travers des 3 perspectives susmentionnées, aux capacités qu'a BrokerRoad pour les mettre en place (employés, systèmes et procédures). L'entreprise doit alors apprendre à innover et à croître en fonction de ces constats.

Résumons maintenant les idées sous forme de tableau :

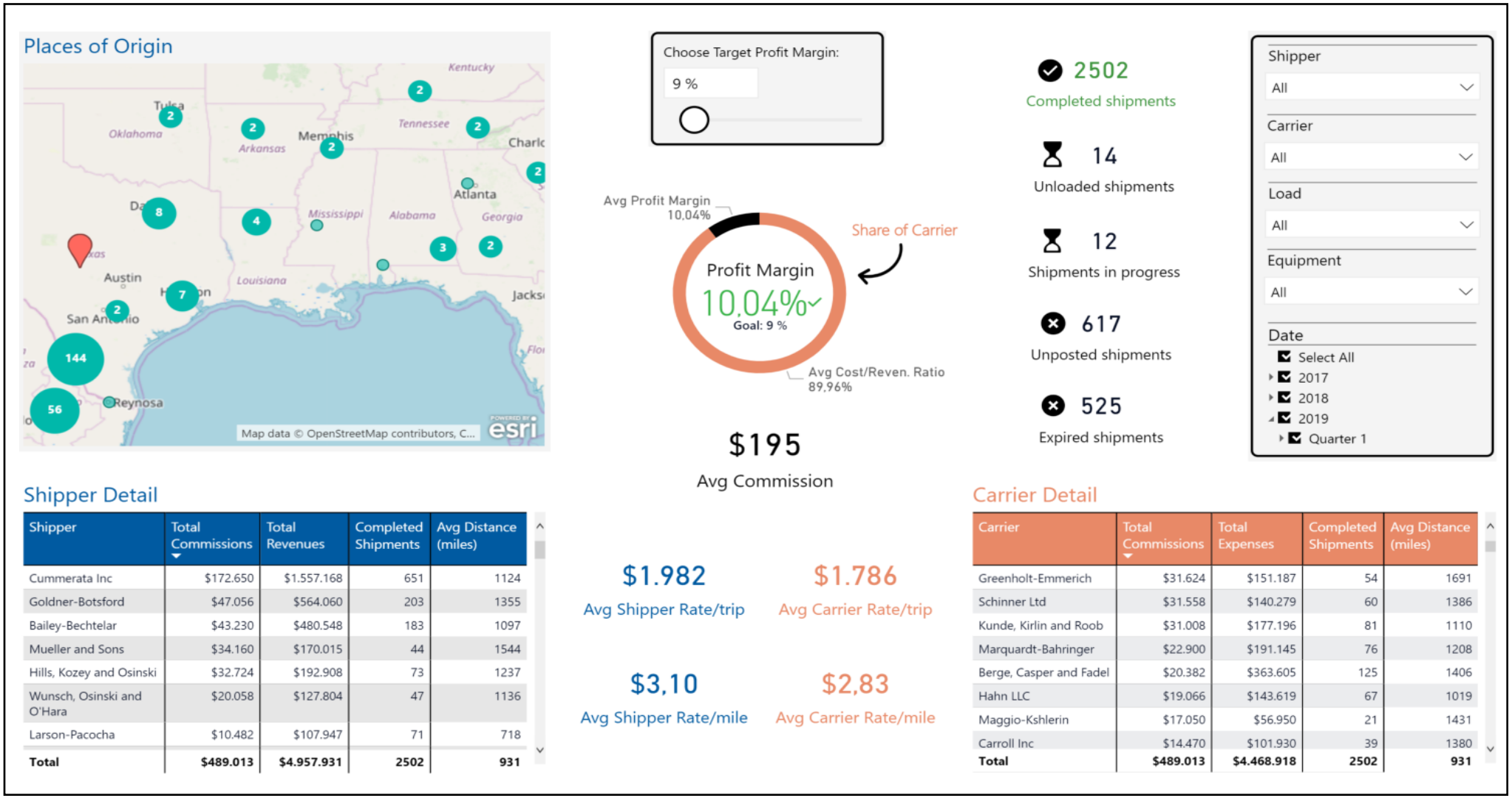
Tableau 5.1 : balanced scorecard de BrokerRoad (exemples d'indicateurs)

Perspective	Objectifs stratégiques	Indicateurs historiques (*lagging indicators*)	Indicateurs avancés (*lead indicators*)
Financière	**Entreprise à but lucratif → profit et croissance financière**	• Marges de profit • Commissions • Revenus • Croissance dans le temps …	
Client	**Satisfaction, rétention et acquisition de clients**	• Nombre de clients • Fréquence des expéditions par client • Evolution du nombre de clients …	• Enquêtes de satisfaction (pas disponibles pour BrokerRoad)
Interne	**Excellence des processus internes**	• Nombre de transactions annulées • Retards …	• Cycles/étapes de la mise en œuvre des expéditions …
Apprentissage et croissance	**Analyse de l'écart entre les objectifs et la capacité à les appliquer**	Difficile à mettre en place pour BrokerRoad (absence d'informations au sujet d'aspects tels que les performances et la satisfaction des employés)	
(Durabilité)	***Sustainability scorecard***	Absence d'informations à ce sujet concernant BrokerRoad	

Les balanced scorecards ont donc permis d'établir des KPIs en prenant notamment en compte l'approche « équilibrée » d'Eckerson mentionnée en section 5.3. **Le choix est fait ici d'incorporer les KPIs mis en évidence par les scorecards dans les dashboards, afin que ces indicateurs soient étudiés de manière interactive**.

5.5 LES DASHBOARDS CONÇUS

Sur base des constats dressés tout au long de l'analyse et particulièrement de ce chapitre, le lecteur (la lectrice) a maintenant bon nombre de cartes en main pour appréhender les visualisations présentées ci-dessous. Ces dernières marquent la fin de ce travail.



Shipper	Total Commissions	Total Revenues	Completed Shipments	Avg Distance (miles)
Cummerata Inc	$172.650	$1.557.168	651	1124
Goldner-Botsford	$47.056	$564.060	203	1355
Bailey-Bechtelar	$43.230	$480.548	183	1097
Mueller and Sons	$34.160	$170.015	44	1544
Hills, Kozey and Osinski	$32.724	$192.908	73	1237
Wunsch, Osinski and O'Hara	$20.058	$127.804	47	1136
Larson-Pacocha	$10.482	$107.947	71	718
Total	**$489.013**	**$4.957.931**	**2502**	**931**

Carrier	Total Commissions	Total Expenses	Completed Shipments	Avg Distance (miles)
Greenholt-Emmerich	$31.624	$151.187	54	1691
Schinner Ltd	$31.558	$140.279	60	1386
Kunde, Kirlin and Roob	$31.008	$177.196	81	1110
Marquardt-Bahringer	$22.900	$191.145	76	1208
Berge, Casper and Fadel	$20.382	$363.605	125	1406
Hahn LLC	$19.066	$143.619	67	1019
Maggio-Kshlerin	$17.050	$56.950	21	1431
Carroll Inc	$14.470	$101.930	39	1380
Total	**$489.013**	**$4.468.918**	**2502**	**931**

Figure 5.1 : analyse stratégique des profils des expéditeurs et des transporteurs (expéditions complétées)

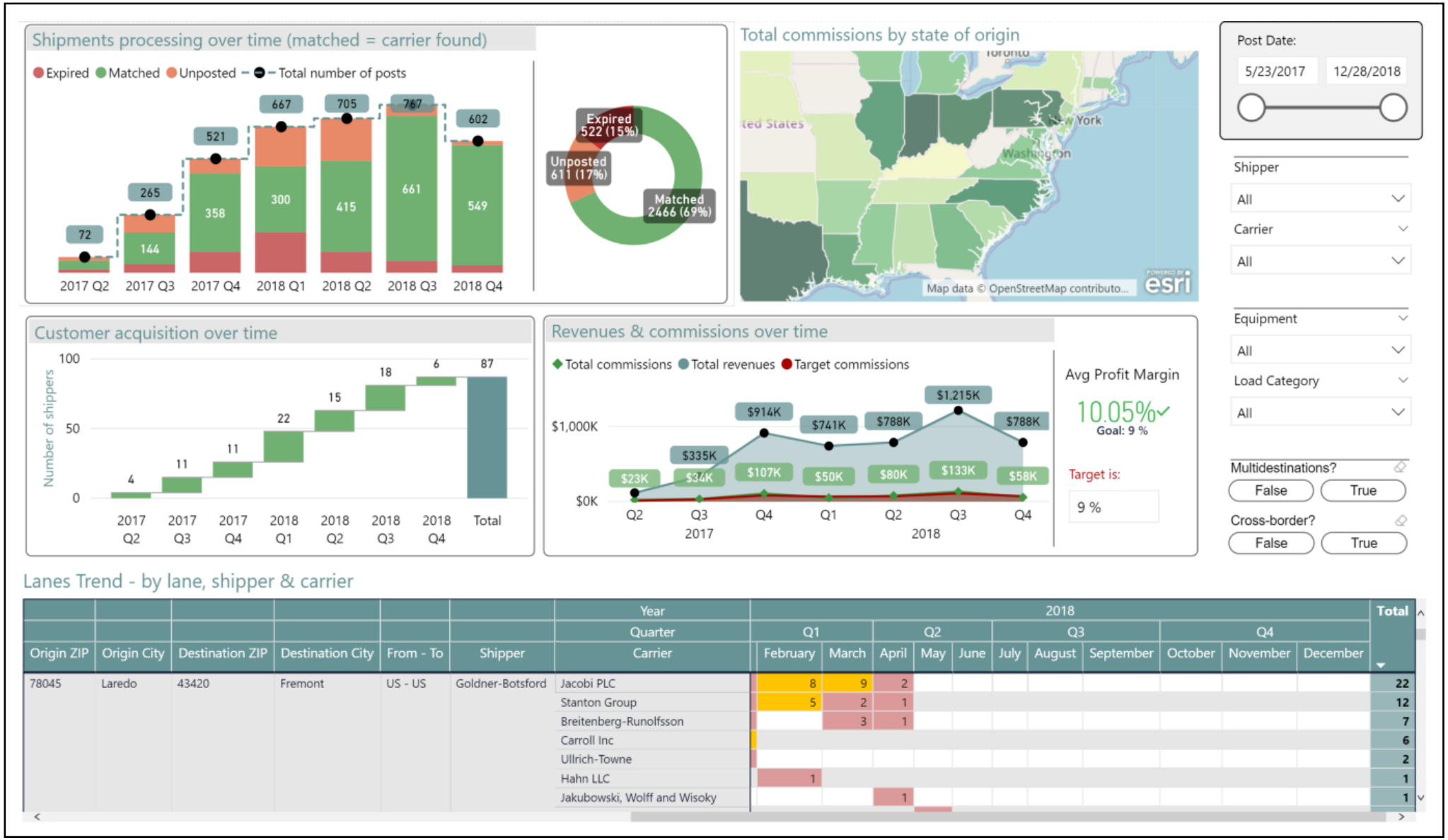


Figure 5.2 : analyse stratégique financière et logistique (processus)

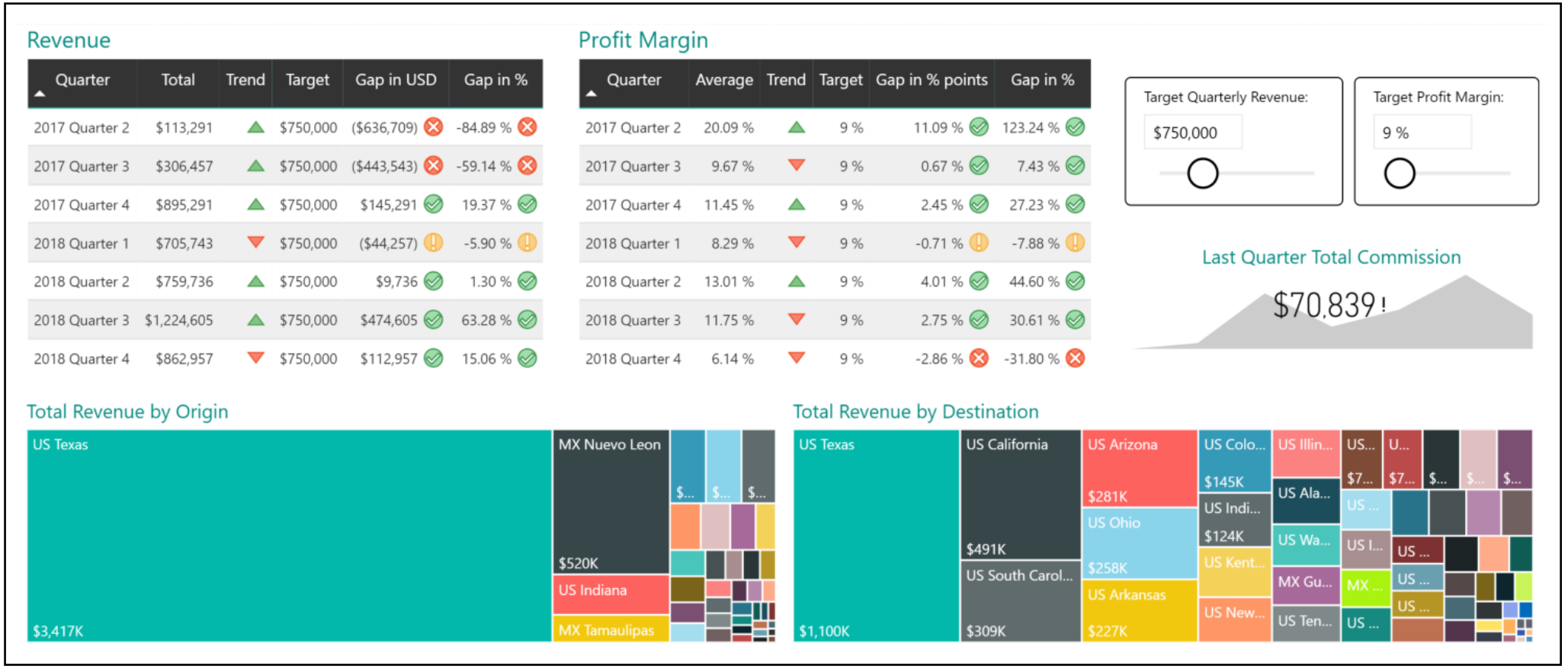
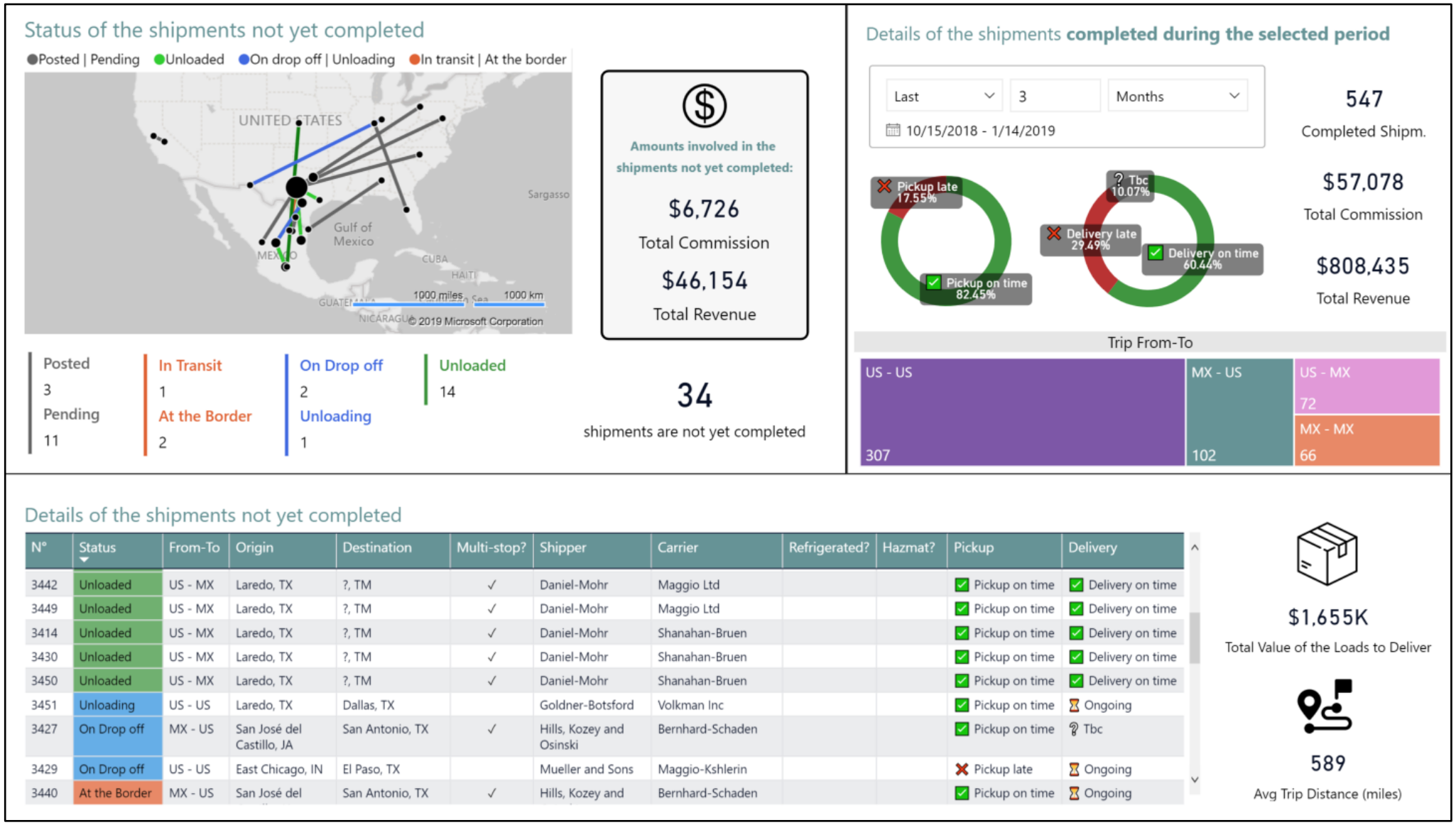


Figure 5.3 : focus sur certains des KPIs financiers principaux



N°	Status	From-To	Origin	Destination	Multi-stop?	Shipper	Carrier	Refrigerated?	Hazmat?	Pickup	Delivery
3442	Unloaded	US - MX	Laredo, TX	?, TM	✓	Daniel-Mohr	Maggio Ltd			Pickup on time	Delivery on time
3449	Unloaded	US - MX	Laredo, TX	?, TM	✓	Daniel-Mohr	Maggio Ltd			Pickup on time	Delivery on time
3414	Unloaded	US - MX	Laredo, TX	?, TM	✓	Daniel-Mohr	Shanahan-Bruen			Pickup on time	Delivery on time
3430	Unloaded	US - MX	Laredo, TX	?, TM	✓	Daniel-Mohr	Shanahan-Bruen			Pickup on time	Delivery on time
3450	Unloaded	US - MX	Laredo, TX	?, TM	✓	Daniel-Mohr	Shanahan-Bruen			Pickup on time	Delivery on time
3451	Unloading	US - US	Laredo, TX	Dallas, TX		Goldner-Botsford	Volkman Inc			Pickup on time	Ongoing
3427	On Drop off	MX - US	San José del Castillo, JA	San Antonio, TX	✓	Hills, Kozey and Osinski	Bernhard-Schaden			Pickup on time	Tbc
3429	On Drop off	US - US	East Chicago, IN	El Paso, TX		Mueller and Sons	Maggio-Kshlerin			Pickup late	Ongoing
3440	At the Border	MX - US	San José del	San Antonio, TX	✓	Hills, Kozey and	Bernhard-Schaden			Pickup on time	Ongoing

Figure 5.4 : analyse tactique/stratégique des expéditions en cours ou dernièrement complétées

Conclusion

Ce mémoire a permis de prêter une attention particulière à certains aspects de la Business Intelligence. Il va sans dire que, pour se pencher sur des questions telles que celle de la conception des visualisations, il était nécessaire de mettre en place une démarche rigoureuse en amont. Le cas réel de l'entreprise fictivement appelée BrokerRoad a été particulièrement utile pour illustrer ces aspects.

Il a ainsi été question de procéder en 5 grandes étapes : appréhender les exigences business et la nature des données opérationnelles mises à disposition, développer des structures de données (data warehouse relationnel et cube OLAP), intégrer les données opérationnelles dans les structures de données ainsi formées, concevoir des visualisations et, enfin, étudier la problématique liée à l'interaction entre visualisations et utilisateurs.

L'analyse n'a évidemment pas la prétention d'apporter une contribution majeure au domaine de la BI, mais elle a néanmoins permis d'étudier de plus près certains aspects qui étaient parfois relativement peu couverts dans la littérature. Les écrits liés aux visualisations BI sont par exemple significativement moins abondants que ceux liés à l'ingénierie des exigences et au traitement de données. Cette carence relative en informations et en marches à suivre dans le cadre de projets BI est souvent liée au fait que ce domaine est avant tout une « science de terrain » : les possibilités sont beaucoup plus nombreuses que l'on ne pourrait le croire quand il s'agit de déployer un processus BI (diversité des DWs, multitude d'outils et de techniques de visualisation...) et beaucoup de techniques avancées sont développées au sein des entreprises et font parfois même l'objet de brevets.

Les visualisations stratégiques conçues pour BrokerRoad sont donc le fruit de ce cheminement, et se basent sur des principes émergeant de la confrontation de divers points de vue recensés dans la littérature ainsi que sur l'analyse à proprement parler : débat de la *junk chart*, débat de la *task-oriented chart*, critères visuels, modèles de formalisation des exigences, définition des KPIs, etc.

Pour conclure, il semble important de mettre en exergue certaines des faiblesses de l'analyse. Cette liste est bien entendu non exhaustive…

Les choix faits ici en matière de revue de la littérature sont bien évidemment partiels et subjectifs. Ils ont néanmoins déjà été justifiés à diverses reprises.

L'analyse qui a suivi la revue de la littérature a été construite sur base du cas d'étude de BrokerRoad, c'est-à-dire au regard d'un domaine d'activité bien déterminé, avec des particularités qui lui sont propres et qui ne sont pas toujours généralisables à d'autres cas. C'est précisément à cet égard que l'on a cherché à souligner les avantages et les limites propres au cas d'étude, notamment pour que le lecteur (la lectrice) ait une idée précise des aspects qui n'ont pas pu être mis en pratique, en dépit de leur importance.

Les principaux désavantages du cas d'étude ont trait à l'absence de contact avec les décideurs et à certaines caractéristiques des données. L'absence de contact avec les décideurs a laissé quelques incertitudes au niveau des exigences initiales et a empêché de développer le projet BI de manière agile : pour qu'un projet BI soit efficace, tant pour le client que pour le fournisseur de solutions, il faut que la solution réponde aux besoins changeants du client et que le client paie en fonction, « ni plus, ni moins ». Il n'a pas non plus été possible de valider empiriquement les modèles sur le terrain. Le caractère statique de l'ensemble des données, sa taille relativement faible et le fait que de nombreux champs ne soient pas complétés ont également complexifié l'analyse.

Ces différentes remarques et limites ouvrent dès lors la voie à d'autres pistes de réflexion qui mériteraient d'être étudiées…

Bibliographie

Articles, ouvrages et supports de cours

- Bar, M. et Neta, M. (2006), "Humans prefer curved visual objects", *Journal of psychological science*, 17(8), pp. 645-648.

- Basili, V.R., Heidrich, J., Lindvall, M., Müch, J., Regardie, M., Rombach, D., Seaman, C. et Trendowicz, A. (2010), "Linking Software Development and Business Strategy Through Measurement", *Computer Journal*, 43(4), pp. 57-65.

- Basili, V.R. et Rombach, H.D. (1994), *Goal Question Metric Paradigm*, 1ère édition, John Wiley & Sons, Hoboken.

- Bateman, S., Mandryk, R.L., Gutwin, C., Genest, A., McDine, D. et Brooks, C. (2010), "Useful junk? The effects of visual embellishment on comprehension and memorability of charts", *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, 10, pp. 2573-2582.

- Bendoly, E. (2016), "Fit, Bias, and Enacted Sensemaking in Data Visualization: Frameworks for Continuous Development in Operations and Supply Chain Management Analytics", *Journal of Business Logistics*, 37(1), pp. 6-17.

- Bertin, J. (1983), *Semiology of Graphics*, University of Wisconsin Press, Madison.

- Bleggi, C.C. et Zhou, F. (2017), "A Study of Freight Performance and Carrier Strategy", *Massachusetts Institute of Technology Libraries*, pp. 1-66.

- Borgo, R., Abdul-Rahman, A., Mohamed, F., Grant, P.W., Reppa, I., Floridi, L. et Chen, M. (2012), "An empirical study on using visual embellishments in visualizations", *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2759-2768.

- Borkin, M.A., Vo, A.A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A. et Pfister, H. (2013), "What Makes a Visualization Memorable?", *IEEE Transactions on Visualizations and Computer Graphics*, 19(12), pp. 2306-2315.

- Brady, T.F., Konkle, T., Alvarez, G.A. et Oliva, A. (2008), "Visual long-term memory has a massive storage capacity for object details", *Proceedings of the National Academy of Sciences*, 105(38), pp. 14325-14329.

- Casner, S.M. (1991), "A task-analytic approach to the automated design of graphic presentation", *ACM Transactions on Graphics*, 10, pp. 111-151.

- Cho, J. (2008), “Issues and Challenges of Agile Software Development with SCRUM”, *Issues in Information Systems*, 9(2), pp. 188-195.

- Cleveland, W.S. et McGill, R. (1984), “Graphical perception: Theory, experimentation and application to the development of graphical methods”, *Journal of the American Statistical Association*, 79(387), pp. 531-554.

- Crainic, T.G., Damay, J. et Gendreau, M. (2007), “An integrated freight transportation modelling framework”, *Proceedings of the International Network Optimization Conference*, pp. 1-6.

- Dumas, B. (2017), *Information Visualisation (4): Tables, Spatial data, Networks and Trees*, notes de cours, Visualisation de l’information [IDASM103], Université de Namur, donné le 11 septembre 2017.

- Eckerson, W.W. (2009), “Performance Management Strategies. How to Create and Deploy Effective Metrics”, *TDWI Best Practices Report*.

- Few, S. (2006), *Information Dashboard Design*, O’Reilly Media, Sebastopol.

- Few, S. (2011), “The chartjunk debate: A close examination of recent findings”, *Journal of Visual Business Intelligence*, pp. 1-10.

- Froese, M.-E. et Tory, M. (2016), “Lessons Learned from Designing Visualization Dashboards”, *IEEE Computer Graphics and Applications*, pp. 83-89.

- Golfarelli, M., Rizzi, S. et Cella, I. (2004), “Beyond Data Warehousing: What’s Next in Business Intelligence?”, *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, pp. 1-6.

- Highsmith, J. (2002), *Agile Software Development Ecosystems*, Addison-Wesley, Boston.

- Huang, M., Homem-de-Mello, T., Smilowitz, K. et Driegert, B. (2011), “Supply Chain Broker Operations – Network Perspective”, *Journal of the Transportation Research Bord*, 2224, pp. 1-2.

- Hullman, J., Adar, E. et Shah, P. (2011), “Benefiting infovis with visual difficulties”, *IEEE Transactions on Visualization and Computer Graphics*, 17(12), pp. 2213-2222.

- Isola, P., Xiao, J., Torralba, A. et Oliva, A. (2011), “What makes an image memorable?”, *Conference on Computer Vision and Pattern Recognition*, pp. 145-152.

- Janes, A. et Succi, G. (2009), “To Pull or Not to Pull”, *Proceedings of OOPSLA*, pp. 1-10.

- Janes, A., Succi, G. et Silliti, A. (2013), “Effective dashboard design”, *Cutter IT Journal*, pp. 17-23.

- Johnson, J.C. et Schneider, K.C. (1995), “Outsourcing in Distribution: The Growing Importance of Transportation Brokers”, *Business Horizons*, pp. 40-48.

- Kaplan, R. et Norton, D. (1996), “Strategic learning & the balanced scorecard”, *Strategy & Leadership*, 24(5), pp. 18-24. [traduction]

- Lindvall, M., Basili, V.R., Boehm, B., Costa, P., Dangle, K.C., Shull, F., Tesoriero Tvedt, R., Williams, L. et Zelkowitz, M.V. (2002), “Empirical Findings in Agile Methods”, *Extreme Programming and Agile Methods - Agile Universe 2002*, pp. 197-207.

- Kosslyn, S.M. (1989), “Understanding charts and graphs”, *Applied cognitive psychology*, 3(3), pp. 185-225.

- Mackenzie, P.D., Jesson, J.E., Salvo, J.J., Mangino, K.M., Graziano, R.A., Theurer, C.B. et Ratsimor, O. (2009), “Freight Commerce System and Method”, *US Patent Application Publication No.US 2010/0250446 A1*, pp. 5-6.

- Mirza, M. et Datta, S. (2019), “Strengths and Weaknesses of Traditional and Agile Processes - A Systematic Review”, *Journal of Software*, 14(5), pp. 209-219.

- Naro, G. et Noguéra, F. (2008), “L’intégration du développement durable dans le pilotage stratégique de l’entreprise : enjeux et perspectives des sustainability balanced scorecards”, *Revue de l’organisation responsable*, 1, pp. 24-38. [traduction]

- Nowell, L., Schulman, R. et Hix, D. (2002), “Graphical Encoding for Information Visualization: An Empirical Study”, *Proceedings of the IEEE Symposium on Information Visualization*.

- Parmenter, D. (2010), *Key Performance Indicators: Developing, Implementing and Using Winning KPIs*, Wiley, New York. [traduction]

- Rust, R.T., Thompson, D.V. et Hamilton, R. (2006), “Defeating feature fatigue”, *Harvard Business Review*, 84(2), pp. 98-107.

- Shukla, A. et Dhir, S. (2016), “Tools for Data Visualization in Business Intelligence: Case Study Using the Tool Qlikview”, *Information Systems Design and Intelligent Applications*, 434, pp. 319-326.

- Tufte, E.R. (2001), *The Visual Display of Quantitative Information*, 2ème édition, Graphics Press, Cheshire.

- Tweedie, L. (1997), “Characterizing Interactive Externalizations”, *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

- Vaisman, A. et Zimanyi, E. (2014), *Data Warehouse Systems: Design and Implementation*, Springer, New York.

- Ware, C. (2012), *Information Visualization: Perception for Design*, 3ème édition, Morgan Kaufmann Publishers, Burlington.

- Wattenberg, M. et Fisher, D. (2004), “Analysing perceptual organization in information graphics”, *Information Visualization*, 3, pp. 123-133.

- Weiner, J., Balijepally, V. et Tanniru, M. (2015), “Integrating Strategic and Operational Decision Making Using Data-Driven Dashboards: The Case of St. Joseph Mercy Oakland Hospital”, *Journal of Healthcare Management*, 60(5), pp. 319-330.

- Williams, L. et Cockburn, A. (2003), “Agile Software Development: It’s about Feedback and Change”, *IEEE Computing*, 36(6), pp. 39-43.

- Zhu, Y. (2007), “Measuring Effective Data Visualization”, *International Symposium on Visual Computing*, 2, pp. 652-661.

Sites Internet

- Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J. et Thomas, D. (2001). *Manifeste pour le Développement Agile de Solutions*. [En ligne] Disponible sur : https://manifesteagile.fr [Consulté le 28/07/19].

- Brandbucket.com. (2019). *Mobile App Development Company Names.* [En ligne] Disponible sur : https://www.brandbucket.com/industries/mobile-application-development-company-names [Consulté le 01/07/19].

- Flaticon.com. (2019). *Flaticon.* [En ligne] Disponible sur : https://www.flaticon.com [Consulté le 01/07/19].

- Fmcsa.dot.gov. (2017). *What are the definitions of motor carrier, broker and freight forwarder authorities?* [En ligne] Disponible sur : https://ask.fmcsa.dot.gov/app/answers/detail/a_id/248/~/what-are-the-definitions-of-motor-carrier%2C-broker-and-freight-forwarder [Consulté le 01/07/19].

- Freightquote.com. (2016). *What is the difference between LTL and FTL?* [En ligne] Disponible sur : https://www.freightquote.com/blog/what-is-the-difference-between-ltl-ftl [Consulté le 01/07/19].

- Freightquote.com. (2017). *Dry van trucking vs. refrigerated shipping and flatbed trucks.* [En ligne] Disponible sur : https://www.freightquote.com/blog/dry-van-trucking-vs-refrigerated-shipping-and-flatbed-trucks [Consulté le 01/07/19].

- Gartner.com. (2019). *Magic Quadrant for Analytics and Business Intelligence Paltforms.* [En ligne] Disponible sur : https://www.gartner.com/doc/reprints?id=1-3TXXSLV&ct=170221&st=sb [Consulté le 26/07/19].

- Kimballgroup.com. (n.d.). *Data Warehouse and Business Intelligence Resources.* [En ligne] Disponible sur : https://www.kimballgroup.com/data-warehouse-business-intelligence-resources [Consulté le 20/07/19].

- Masterslogistical.co.uk. (n.d.). *3PL vs. Freight Broker: who do you decide to use?* [En ligne] Disponible sur : https://www.masterslogistical.co.uk/3pl-vs-freight-broker-decide-use [Consulté le 01/07/19].

- Oecd.org. (2017). *Transport de marchandises*. [En ligne] Disponible sur : https://data.oecd.org/fr/transport/transport-de-marchandises.htm [Consulté le 02/07/19].

- Project44.com. (2019). *Visibility Your Customers Expect.* [En ligne] Disponible sur : https://project44.com [Consulté le 02/07/19].

- Welna, P. (n.d.). *Transportation: asset-based vs. brokerage*. [En ligne] Disponible sur : http://www.murphywarehouse.com/blog/transportation-asset-based-vs-brokerage [Consulté le 02/07/19].

- Wisdomjobs.com. (n.d.). *Fundamental grains – data warehouse ETL toolkit.* [En ligne] Disponible sur : https://www.wisdomjobs.com/e-university/data-warehouse-etl-toolkit-tutorial-201/fundamental-grains-8209.html [Consulté le 10/07/19].

Documents annexes

Annexe A-1. Données à disposition

N° d'expédition et parties prenantes	
shipper_id	ID de l'expéditeur
shipment_no	Numéro d'expédition (est assigné uniquement si un transporteur a été trouvé et n'est donc pas une clé primaire)
carrier_id	ID du transporteur assigné pour l'expédition

Cargaison	
load_description	Description de la cargaison
load_value	Valeur de la cargaison, telle que spécifiée par l'expéditeur
load_equipment	Type d'équipement (remorque) exigé par l'expéditeur
is_hazmat	La cargaison est-elle ou non de nature dangereuse ?

Prix et coûts	
currency_code	Devise (USD ou MXN) utilisée pour les montants renseignés pour l'expédition
fh_commission	Commission du broker *(shipper_closed_price - carrier_closed_price)*
shipper_aux_total_charges	Coûts supplémentaires que l'expéditeur doit payer au broker
carrier_aux_total_charges	Coûts supplémentaires que le broker doit payer au transporteur
shipper_ask_price	Budget max. de l'expéditeur
shipper_closed_price	Prix effectivement payé par l'expéditeur au broker
carrier_winning_bid	Prix qui a permis au transporteur de remporter les enchères (si enchères il y a eu)
carrier_closed_price	Prix effectivement payé par le broker au transporteur

Trajet	
distance	Distance min. estimée
duration	Durée estimée sur base de la distance
is_cross_border	La frontière US-MX est-elle ou non franchie ?
is_multistop	Le trajet fait-il ou non l'objet d'arrêts à des destinations intermédiaires ?

Origine	
origin_point	Coordonnées de l'origine
origin_postal_code	Code postal de l'origine
origin_city	Ville d'origine
origin_state	Etat d'origine
origin_country	Pays d'origine

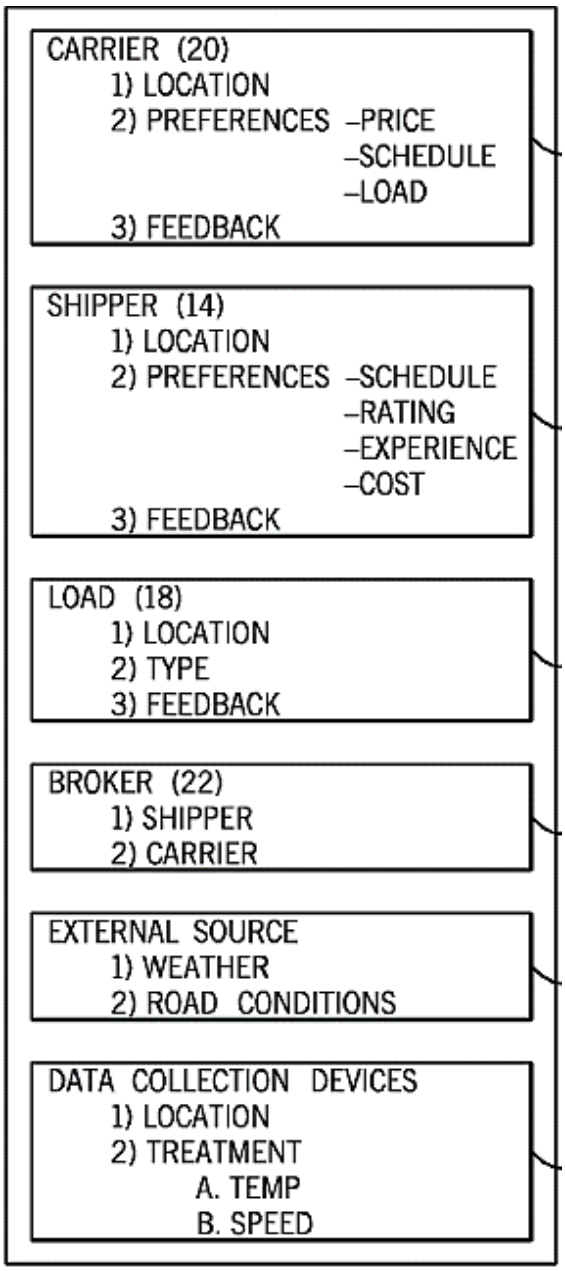
Destination	
destination_point	Coordonnées de la destination
destination_postal_code	Code postal de la destination
destination_city	Ville de destination
destination_state	Etat de destination
destination_country	Pays de destination

Statuts & dates	
status	Statut actuel de l'expédition (cf. table « Statuts »)
is_completed	Statut="complété" ou non (la livraison a été effectuée et les démarches administratives sont terminées)
is_expired	Statut="expiré" ou non (aucun transporteur n'a été trouvé à temps)
is_dropped	Statut="unposted" ou non (l'expéditeur a retiré son offre)
completed_at	Date-heure à laquelle l'expédition a été renseignée comme complétée
expires_at	Date-heure à laquelle l'offre expire
expires_at_week	Semaine au cours de laquelle l'offre expire
dropped_at	Date-heure à laquelle l'offre a été retirée
dropped_at_week	Semaine au cours de laquelle l'offre a été retirée
posted_at	Date-heure à laquelle l'offre a été soumise par l'expéditeur sur la plateforme

matched_at	Date-heure à laquelle un transporteur a été trouvé
matched_at_week	Semaine au cours de laquelle un transporteur a été trouvé
matched_at_month	Mois au cours duquel un transporteur a été trouvé
pickup_scheduled_at	Date-heure à partir de laquelle le pickup (enlèvement) est prévu
pickup_scheduled_at_month	Mois au cours duquel le pickup est prévu
pickup_scheduled_until	Date-heure avant laquelle le pickup doit être effectué
on_pickup_at	Date-heure à laquelle le pickup a été effectué
on_pickup_at_week	Semaine au cours de laquelle le pickup a été effectué
go_to_pu_at	Date-heure à laquelle on est passé en mode "go to pickup"
in_transit_at	Date-heure à laquelle le transporteur s'est mis en mouvement avec la cargaison
halted_at	Date-heure à laquelle le transporteur s'est arrêté pour des raisons anormales (ex. : panne moteur)
is_halted	Le transporteur est-il ou non à l'arrêt pour des raisons anormales ?
last_seen_at	Date-heure à laquelle le véhicule a été géolocalisé pour la dernière fois
delivery_scheduled_at	Date-heure à partir de laquelle la livraison est prévue
delivery_scheduled_at_month	Mois au cours duquel la livraison est prévue
delivery_scheduled_until	Date-heure avant laquelle la livraison doit être effectuée
dropped_off_at_week	Semaine au cours de laquelle le transporteur est arrivé sur le lieu de livraison
dropped_off_at_month	Mois au cours duquel le transporteur est arrivé sur le lieu de livraison
unloaded_at	Date-heure à laquelle la cargaison a été renseignée comme étant déchargée
unloaded_at_week	Semaine au cours de laquelle la cargaison a été renseignée comme étant déchargée
unloaded_at_month	Mois au cours duquel la cargaison a été renseignée comme étant déchargée

Statuts	
Posted	L'expéditeur a posté l'offre sur la plateforme
Unposted	L'expéditeur a enlevé son offre
Expired	L'offre est arrivée à expiration (le broker n'a pas trouvé de transporteur dans le temps imparti par l'expéditeur)
Pending	Un transporteur a été trouvé mais certaines informations doivent encore être fournies (ex. : équipement utilisé)
In transit	La cargaison est en transit
At stop	La cargaison est à l'arrêt à la frontière US-MX
On d.o.	Le transporteur est arrivé au lieu de livraison *(on drop off)*
Unloading	Le transporteur est en train de décharger la cargaison
Unloaded	La cargaison a été déchargée
Completed	L'expédition est complétée (démarches administratives y compris)

Annexe A-2. Liste de données utiles pour établir les KPIs



(Mackenzie et al., 2009)

Annexe A-3. Schémas des sous-tâches de l'ETL

Figure A-3.1 : tâche de flux de chargement de la dimension DimLoad

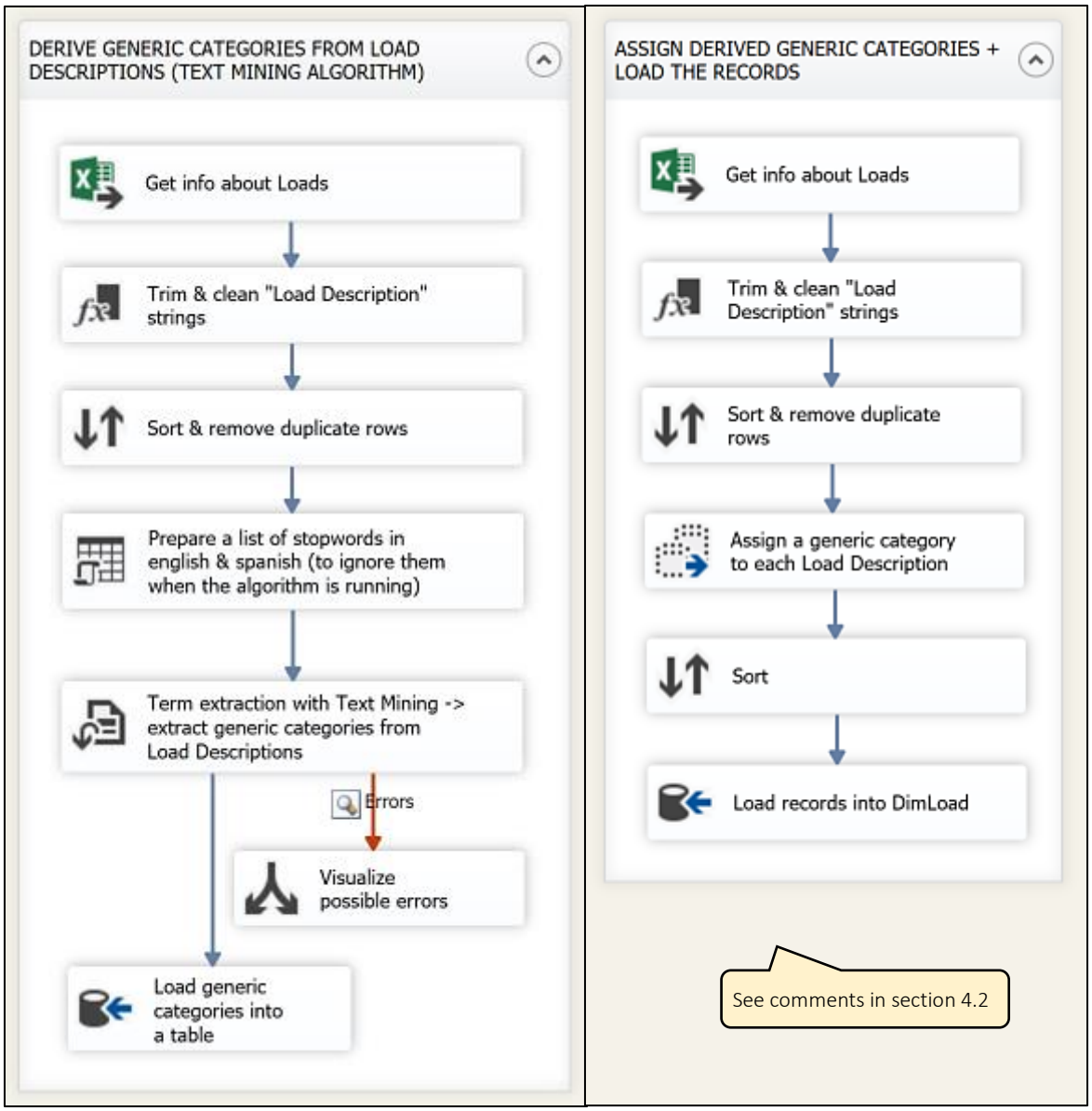


Figure A-3.2 : 2ème exemple de tâche de flux de chargement d'une dimension (DimLocation)

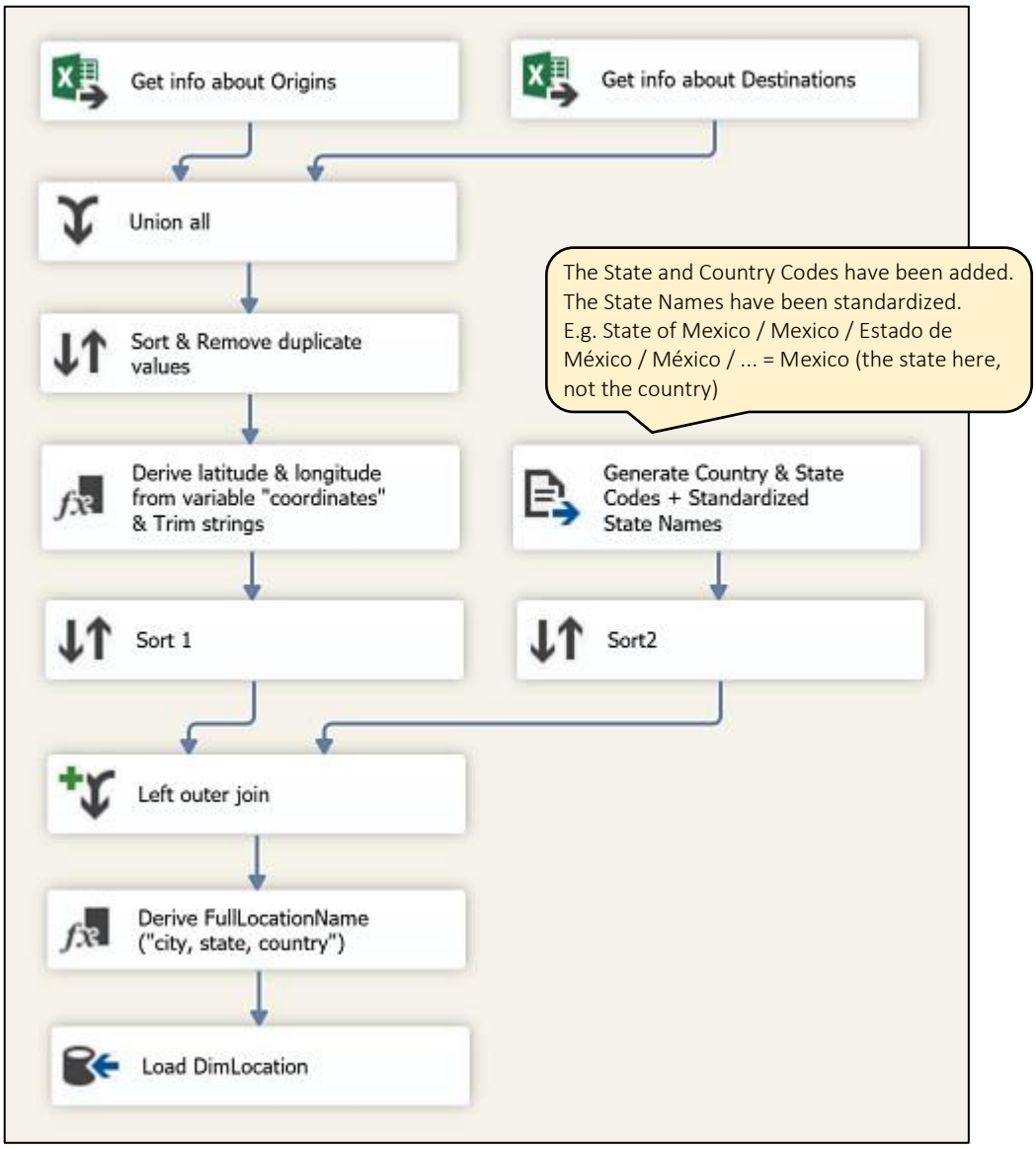
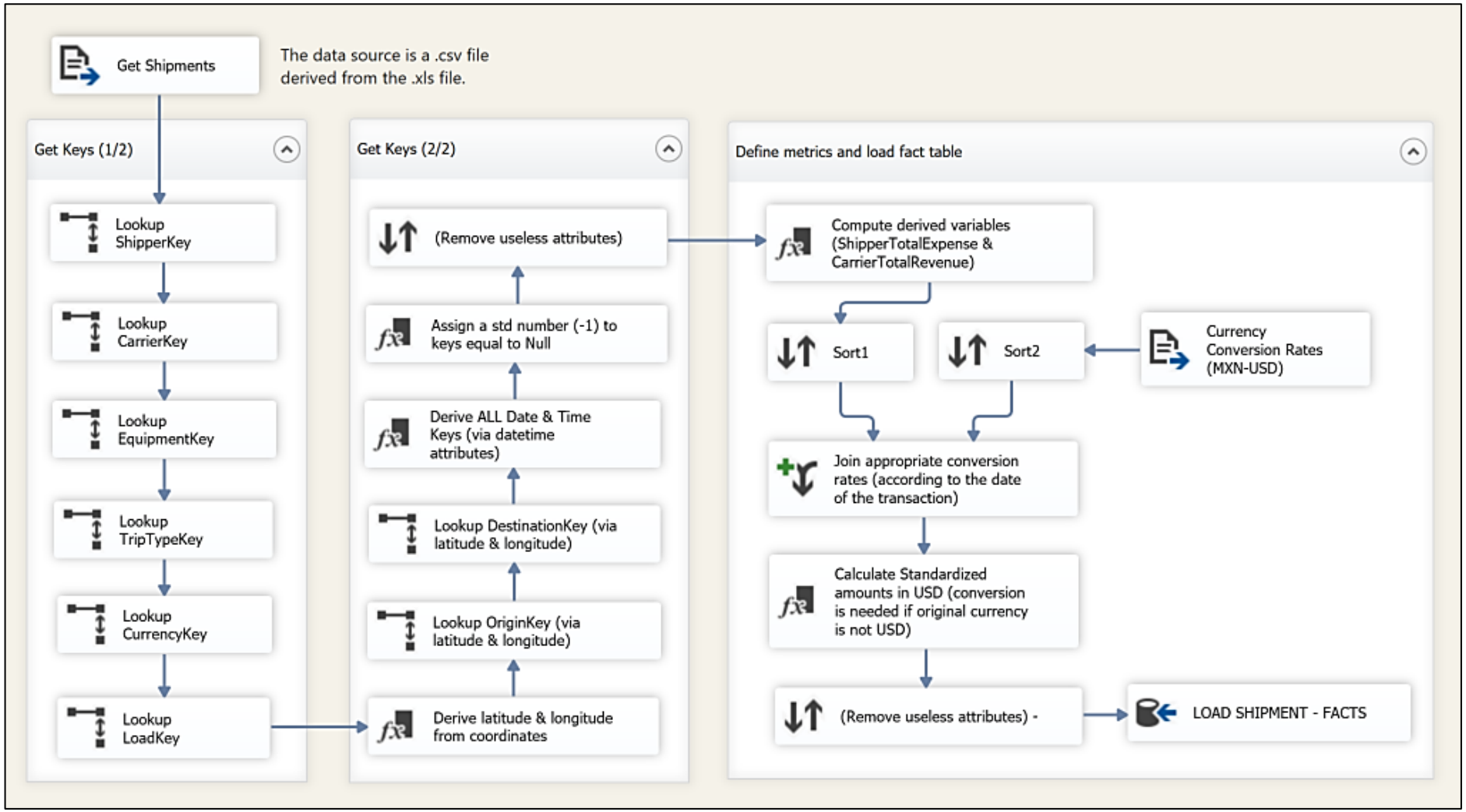


Figure A-3.3 : tâche de flux de chargement de la table de faits



Annexe A-4. Gartner Magic Quadrant lié aux plateformes de BI