Estimation de la capacité de stockage de l'entrepôt de données en fonction des éléments de la politique de la gestion des informations décisionnelles et de l'environnement du système

par

JEAN BAPTISTE LALA

Essai présenté au Centre de formation en technologie de l'information (CeFTI)

en vue de l'obtention du grade de Maître en Génie Logiciel MGL (Maîtrise en génie logiciel incluant un cheminement de type court en génie logiciel)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Longueuil, Québec, Canada, février 2017

Sommaire

Face à l'évolution des besoins en information et aux exigences technologiques liées aux besoins de changement, l'espace nécessaire pour stocker et à rassembler les données dans l'entrepôt devient de plus en plus une ressource importante. En vue d'une planification stratégique des ressources, les entreprises doivent accorder une importance capitale à l'estimation de leur capacité de stockage de l'entrepôt de données en fonction de l'environnement du système et des éléments de la politique de la gestion des informations décisionnelles qui ont une influence directe sur celle-ci.

Afin de résoudre les problèmes engendrés par l'absence d'outils d'estimation d'espace de stockage, la principale question de recherche est :

« L'estimation de la capacité de stockage de l'entrepôt de données en fonction des éléments de la politique de gestion des informations décisionnelles et de l'environnement du système par la méthode formelle permet-elle d'améliorer la précision de l'estimation de l'espace de stockage requis?»

En vue de répondre à cette question, une enquête par sondage sur les déterminants de la capacité de stockage d'entrepôt de données a été menée. Cette enquête vise à comprendre les pratiques des entreprises et à identifier à priori les variables qui sont susceptibles d'influencer la capacité de stockage selon les expériences sur le terrain pour éviter les biais dans les modèles d'estimation. Le présent essai met l'accent sur deux types de modélisation : la modélisation à travers les facteurs discriminants et la modélisation par la régression log-linéaire sur les données en panel. La modélisation à travers les facteurs discriminants permet de séparer explicitement les groupes de magasins de données homogènes et de prédire le groupe d'appartenance d'un nouvel entrepôt de données selon ses caractéristiques. La modélisation par la régression log-linéaire sur les données en panel offre non seulement la possibilité d'effectuer la prévision de l'espace de stockage dans le temps mais aussi elle aide à retrouver la structure des groupes d'entrepôts de données définie dans la modélisation par les facteurs discriminants.

Les résultats de l'enquête effectué par sondage font ressortir que les variables liées à l'environnement du système telles que la taille des tables de faits et la taille des tables de dimension, puis la variable associée à la politique de gestion des données informationnelles comme la durée de stockage des informations ou la fréquence de rétention des données sont considérées comme des facteurs qui influencent la capacité de stockage de l'entrepôt de données. Les analyses statistiques effectuées sur l'échantillon de données issue de la base de données de l'expérimentation viennent confirmer l'existence d'une forte corrélation entre la variable cible (capacité de stockage) et les variables explicatives (la taille des tables des faits, la taille de la table de dimension et la durée de stockage des données). Les résultats des modèles d'estimation à travers ces facteurs, démontrent que la méthode formelle affiche une erreur moyenne d'estimation près de 2,92 %. Quant à la méthode d'estimation par intuition, l'erreur d'estimation dépasse en moyenne 25 %. Donc, le modèle formel offre une estimation huit fois plus précise que la méthode intuitive. Ce constat conduit à la conclusion selon laquelle : «la connaissance de la capacité de stockage de l'entrepôt de données à partir des facteurs liés à la politique de gestion des informations décisionnelles et à l'environnement du système améliore la précision de l'estimation des besoins réels en espace de stockage». Donc, l'utilisation de la méthode formelle d'estimation d'espace de stockage est recommandable et aide l'entreprise à assurer la gestion rationnelle des ressources matérielles et financières à mobiliser.

Afin de conserver la force prédictive du modèle formel dans temps, la mise à jour de ses composantes (notamment les variables explicatives) est nécessaire et indispensable. La mise à jour du modèle et le processus de production des résultats d'estimation doivent être automatisés dans le but de mettre en place un système intelligent. L'automatisation pourrait se faire en exploitant les techniques utilisées dans le domaine de l'Apprentissage Machine (Machine Learning), une discipline qui a montré ses preuves dans le domaine de la science des données. La réalisation de ce système intelligent peut être considérée comme un des prolongements pertinents du présent essai.

Remerciements

Je remercie Monsieur Claude Cardinal pour ses conseils et son encouragement durant tout mon parcours à l'Université de Sherbrooke. Son appui constant m'a permis d'accomplir des grandes réalisations tant sur le plan académique que sur le plan professionnel.

Je remercie Monsieur Martin Désilets pour son précieux encadrement lors de toutes les étapes de l'élaboration du présent travail. Son professionnalisme très poussé, sa persévérance et son souci du détail ont rendu possible la pertinence des analyses effectuées dans cet essai.

Je remercie les membres de ma famille et toutes les personnes qui ont contribué de près ou de loin à la mise en œuvre de cet essai.

Table des matières

Liste des tableaux

Liste des figures

Glossaire

Analyse prédictive: Analyse statistique qui extrait l'information à partir des données historiques pour prédire les tendances futures.

Analyse statistique descriptive et exploratoire : Analyse qui a pour objectif de résumer, synthétiser l'information contenue dans la série statistique et de mettre en évidence ses propriétés.

Capacité de stockage : la capacité du disque dur nécessaire pour emmagasiner le volume de données.

Entrepôt de données : une collection de données thématiques.

Magasin de données : un sous-ensemble de l'entrepôt.

Systèmes opérationnels : systèmes dédiés aux métiers de l'entreprise pour les assister dans leurs tâches de gestion quotidiennes.

Tables de dimension : tables qui stockent les éléments des axes d'analyse à considérer dans l'entrepôt.

Tables de faits : tables qui enregistrent les indicateurs à mesurer.

Liste des sigles, des symboles et des acronymes

ACM : analyse en correspondances multiples

ACP : Analyse en Composantes Principales

CAH : Classification Ascendante Hiérarchique

ED : Entrepôt de données

ERP : Enterprise Resource Planning

ETC : Extraction, Transformation et Chargement

OLAP : On-Line Analytical Processing

OLTP : On-Line Transaction Processing

SAS : Statistical Analysis System

SGBD : Système de Gestion de Base de Données

Introduction

La transformation et le stockage des données opérationnelles dans un entrepôt de données permettent d'offrir des outils précieux d'aide à la décision aux gestionnaires. Cependant, l'une des principales contraintes qui pèsent sur l'entreprise est sa capacité limitée à rassembler et à stocker les informations dans l'entrepôt de données. D'où la nécessité de quantifier la capacité de stockage de l'entrepôt de données à partir des éléments de la politique de la gestion des informations décisionnelles et de l'environnement du système. Cette démarche permet de mettre en place les instruments ou les indicateurs d'aide à la décision pour l'entreprise.

Au regard de l'évolution des besoins en informations et des exigences technologiques liées aux besoins de changement, l'estimation de la capacité de stockage de l'entrepôt de données s'avère souvent un défi majeur. L'évaluation est soit sous-estimée, soit surestimée du fait de l'absence d'outil permettant d'effectuer la prévision en tenant compte des liens entre les éléments de la politique de gestion des informations décisionnelles, l'environnement du système et la capacité de stockage. Cette situation implique à son tour, soit la sous-évaluation, soit la surévaluation des ressources allouées, à cause de l'incohérence entre les prévisions et les objectifs.

Cet essai consiste à exploiter et à analyser les données d'une entreprise utilisant des gros volumes de données de plus d'un gigaoctet par jour. L'entreprise en question applique les techniques d'entrepôts de données séparant le système opérationnel à celui de l'informationnel.

La principale question est: l'estimation de la capacité de stockage de l'entrepôt de données en fonction des éléments de la politique de la gestion des informations décisionnelles et de l'environnement du système permet-elle de mieux gérer les ressources matérielles et financières affectées? Ainsi, l'hypothèse de l'étude est formulée de manière suivante : «la connaissance de la capacité de stockage de l'entrepôt de données à partir des éléments de

la politique de gestion des informations décisionnelles et de l'environnement du système permet la prévision des besoins réels en ressources matérielles et financières, compte tenu de l'évolution des besoins informationnels et la hausse du volume de données à exploiter.»

Afin de répondre à la question principale et de vérifier l'hypothèse, le présent essai est divisé en quatre grandes sections:

- La partie introductive met en exergue le contexte dans lequel l'étude est menée, puis la revue de littérature incluant l'inventaire de ce qui a été publié et la découverte des éléments liés au sujet;
- La méthodologie précise la description de la procédure d'analyses de données ;
- La description des résultats vise à répondre à la question de l'étude et à confirmer ou à infirmer l'hypothèse et
- La discussion mène à la conclusion et aux recommandations.

Chapitre 1

Mise en contexte

Ce chapitre met en évidence le sujet d'étude, le problème de recherche, le contexte de réalisation et l'explication de certains concepts abordés dans cet essai.

1.1 Le sujet d'étude

«Estimation de la capacité de stockage de l'entrepôt de données en fonction des éléments de la politique de gestion des informations décisionnelles et de l'environnement du système.»

Face à l'évolution des besoins en information et aux exigences technologiques liées aux besoins de changement, l'espace nécessaire pour stocker ou à rassembler les données dans l'entrepôt devient de plus en plus une ressource très importante. Donc, en vue d'une planification stratégique des ressources, les entreprises doivent accorder une importance capitale à l'estimation de leur capacité de stockage de l'entrepôt de données en fonction de l'environnement du système et des éléments de la politique de la gestion des informations décisionnelles qui ont une influence directe sur celle-ci.

1.2 Le problème de recherche et le contexte de réalisation

L'estimation de la capacité de stockage de l'entrepôt de données s'avère un défi majeur pour les organisations. Certaines entreprises effectuent leur estimation à partir de l'historique des données ou selon les expériences passées. D'autres entreprises n'utilisent pratiquement pas de méthode d'estimation et réagissent aux alertes. Cette conséquence peut entraîner des situations problématiques et coûteuses. En guise d'exemple, le manque d'espace de stockage peut priver l'entreprise de son entrepôt de données en attendant d'éventuelles

modifications au niveau des infrastructures, ou encore, l'acquisition d'espace supplémentaire non utilisé peut occasionner un coût excédentaire.

La principale question est: l'estimation de la capacité de stockage de l'entrepôt de données en fonction de l'environnement du système et des éléments de la politique de la gestion des informations décisionnelles permet-elle de mieux gérer les ressources matérielles et financières affectées? L'hypothèse de l'étude est formulée comme suit : «la connaissance de la capacité de stockage de l'entrepôt de données à partir de l'environnement du système et des éléments de la politique de la gestion des informations décisionnelles permet la prévision des besoins réels en ressources matérielles et financières, compte tenu de l'évolution des besoins informationnels et la hausse du volume de données à exploiter.
Cet essai consiste à exploiter et à analyser les données d'une entreprise utilisant des gros volumes de données avec plus d'un gigaoctet par jour. L'entreprise ciblée pour l'expérience est située dans la région de Montréal. Elle applique les techniques d'entrepôts de données séparant le système opérationnel à celui de l'informationnel (selon le modèle de Ralph Kimball). Sa méthode d'évaluation de la capacité de stockage de l'entrepôt de données est effectuée de manière aléatoire (à l'improviste) par certains techniciens de l'équipe de gestion des infrastructures dotés plus de huit années d'expérience professionnelle. La croissance annuelle de la capacité de stockage pour l'ensemble de ses entrepôts évolue au rythme de plus de 75 %. L'infrastructure à étudier est le disque dur stockant uniquement les modèles multidimensionnelles (modèles en étoile) car c'est la zone dans laquelle où se trouve la production croissante des données [1]. Les espaces occupés par les métadonnées et les zones de stockage intermédiaire seront exclus.

1.3 Explication des concepts en jeu

La compréhension des concepts clés suivants permet non seulement de connaître la signification de certains termes techniques utilisés mais aussi de faciliter le suivi du cheminement des idées développées dans le présent travail.

- **Entrepôt de données** : L'entrepôt de données (ED) est une collection de données thématiques, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision (Bill Inmon, 1992).
- **Capacité de stockage** : La capacité de stockage d'un entrepôt de données est la capacité du disque dur nécessaire pour emmagasiner le volume de données [2].
- **Systèmes opérationnels** : Les systèmes «opérationnels» ou «de gestion», également appelés systèmes OLTP (On-Line Transaction Processing), sont dédiés aux métiers de l'entreprise pour les assister dans leurs tâches de gestion quotidiennes et donc, directement opérationnels [2].
- **Systèmes décisionnels** : Également appelés OLAP (On-Line Analytical Processing), sont dédiés au management de l'entreprise pour l'aider au pilotage de l'activité, et donc indirectement opérationnels. Ils offrent au décideur une vision transversale de l'entreprise. La tendance pour réaliser un système décisionnel est la mise en place d'un entrepôt de données [2].
- **Magasin de données** : Un magasin de données («data mart» en anglais) est un sous-ensemble de l'entrepôt. Il peut servir à un groupe de décideurs intéressés par le même thème dans l'analyse stratégique de leurs activités. Son volume réduit permet un accès plus rapide aux données, qui peuvent être organisées de façon à répondre aux besoins particuliers [2].
- **Tables de dimension** : Ce sont les tables qui stockent les éléments des axes d'analyse à considérer dans l'entrepôt [2].
- **Tables de faits** : Ce sont les tables qui enregistrent les indicateurs à mesurer et les clés des tables de dimensions [2].

Quelques éléments des concepts clés cités ci-dessus seront détaillés dans le chapitre 2. Ce chapitre englobe la revue de littérature permettant de faire la recension de certains écrits pertinents pour cette étude.

Chapitre 2

Revue de la littérature

Ce chapitre présente une revue de la littérature liée aux concepts généraux d'entrepôt de données, tout en tenant compte de certains détails sur l'infrastructure du système et les différentes méthodes d'estimation de la capacité de stockage. Les outils de recherche d'informations pour cette revue de littérature sont : les livres écrits par différents auteurs, les articles scientifiques trouvés à l'aide du site Google Scholar et les outils de recherche documentaire de l'Université de Sherbrooke.

2.1 Concepts d'entrepôt de données

2.1.1 Définition d'entrepôt de données

Un entrepôt de données est une base de données consolidée. La plupart de ses sources proviennent des bases de données de production que l'on appelle système opérationnel. En général, un entrepôt de données sert à analyser la situation d'une entreprise en vue d'effectuer un suivi de toutes les activités visant à élaborer des stratégiques à travers l'exploitation des informations décisionnelles. Les données sont visualisées graphiquement ou présentées dans des tableaux de bord (dashboard) permettant aux gestionnaires de prendre les décisions au moment opportun.

Selon Bill Inmon, un entrepôt de données possède quatre caractéristiques principales [3] :

- Les données de l'entrepôt sont orientées sujets. En d'autres termes, les informations sont organisées par des thèmes conçus en fonction des besoins et des objectifs de l'entreprise. Ces thèmes sont présentés sous la forme des magasins de données (data mart), par exemple le magasin de données destiné à la gestion des stocks de produits.

- Les données sont intégrées. Elles proviennent de différentes sources hétérogènes. Elles sont stockées dans l'entrepôt après le processus d'extraction, transformation et chargement (ETC). L'intégration des données dans l'entrepôt permet aux utilisateurs d'obtenir une version globale, unique et cohérente des informations à explorer.
- Les données sont «historisées», elles sont répertoriées et conservées par date. Une dimension temporelle est associée à chaque donnée stockée dans l'entrepôt afin de suivre l'évolution des indicateurs à travers le temps.
- Les données sont non volatiles. Elles ne disparaissent pas et ne changent pas au fil des traitements et au fil du temps (Read-Only). Elles sont rarement modifiées ou supprimées par l'utilisateur.

2.1.2 Avantages de l'entrepôt de données

Dans l'entreprise, la mise en place l'entrepôt de données présente divers avantages tels que [4] :

- Meilleure compréhension du déroulement des activités à travers les tendances des résultats;
- Possibilité d'anticiper une situation par les prévisions qui permettent de réduire les risques;
- Apport des meilleurs produits sur le marché au moment opportun;
- Possibilité d'analyser les informations issues des opérations quotidiennes et de prendre immédiatement des décisions qui peuvent affecter de façon significative la performance de l'entreprise;
- Découverte des bonnes informations stratégiques et;
- Présentation des informations sous forme des rapports et des graphes.

Agarwal Bhushan B. et al ont essayé d'évaluer les impacts la mise en place de l'entrepôt de données sur quelques attributs présentés dans le tableau suivant.

Tableau 1 : Impacts de la mise en place de l'entrepôt de données

1-Meilleure qualité de données	63 %
2-Meilleure compétitivité	61 %
3-Accès direct aux données pour les utilisateurs finaux	32 %

4-Réductions de coûts/haute productivité	32 %
5-Accès plus rapide aux données	24 %
6-Meilleure disponibilité des systèmes	16 %
7-Soutien au changement organisationnel	8 %

Source: Agarwal Bhushan B., Tayal Prakash S., Data Mining and Data Warehousing, Laxmi Publications 2009, Chapitre 10, p.155

À travers ce tableau, Agarwal Bhushan B. et al soutiennent que la mise place de l'entrepôt de données dans l'entreprise a des impacts très significatifs sur la qualité de données (63 %) et la compétitivité sur le marché (61 %).

Hormis des bénéfices cités ci-haut, Agarwal Bhushan B. et al mentionnent que l'entrepôt de données procure de manière indirecte d'autres avantages tels que :

- L'amélioration des relations avec la clientèle grâce à une meilleure connaissance des exigences et des tendances individuelles, à l'amélioration des communications et des offres de produits sur mesure et
- L'élaboration des idées révolutionnaires pour la réingénierie des processus d'affaires.

2.1.3 Les défis à relever pour la mise en place d'entrepôt de données

La réussite de la mise en place de l'entrepôt de données exige l'existence d'une bonne organisation au sein de l'entreprise. Cette organisation peut toucher plusieurs axes notamment : l'organisation au niveau de la gestion des ressources humaines, l'organisation financière et l'organisation au niveau de la gestion des infrastructures matérielles. La construction d'un bon entrepôt de données d'entreprise requiert la volonté de consacrer des efforts. Pour mener à bien un projet de mise en place d'un entrepôt de données, Biere Mike [5] indique la moyenne de répartition des efforts à allouer pour l'exécution quelques principales activités comme suit :

Tableau 2 : Répartition des efforts à allouer pour la réalisation d'un projet de mise en place d'entrepôt de données

Relever les défis techniques, gérer des matériels informatiques, logiciels, personnel	42 %
Gestion des données	25 %
Convaincre la haute direction	15 %
Formation des utilisateurs	12 %
Gestion de changements et autres	6 %

Source: Biere Mike, Business Intelligence for the Enterprise, IBM Press 2003, p.171

Selon Biere Mike, les défis techniques, la gestion des matériels informatiques, des logiciels, et du personnel nécessitent la plus grande part des efforts à fournir (42 %). La gestion des matériels informatiques intègre essentiellement la planification des matériels tels que le serveur et le disque de stockage. Selon les besoins de l'entreprise et la politique de l'entreprise, l'acquisition et la gestion de l'espace du disque de stockage doivent mériter une attention particulière pour assurer la réussite du projet.

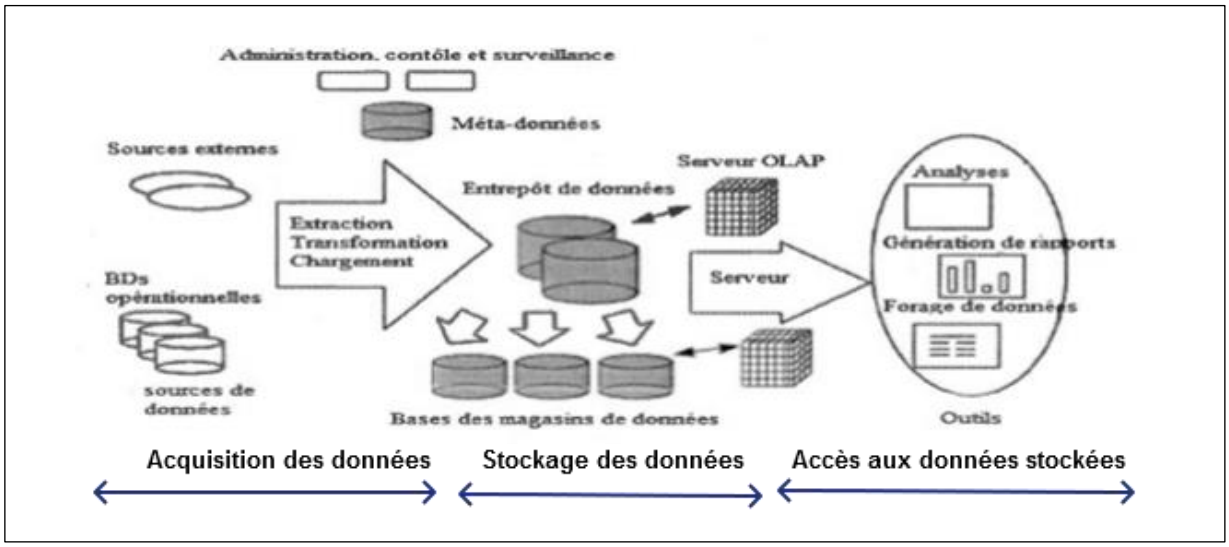
2.2 Architecture d'entrepôt de données

L'architecture adoptée joue un rôle indispensable dans mise en place du système décisionnel car elle permet d'identifier les diverses composantes à implanter. L'architecture d'entrepôt de données présente divers avantages pour l'entreprise [6] :

- Elle assure la satisfaction des besoins techniques dictés par les besoins d'affaires ;
- Elle facilite la communication car elle illustre les rôles des différents composants du système. Elle transmet de manière apparente les informations pertinentes liées à la complexité du projet aux cadres supérieurs et
- Elle facilite la planification car elle regroupe tous les détails techniques et identifie l'interdépendance entre les composants.

De manière générale, l'architecture d'un entrepôt de données est composée de trois grandes zones (voir Figure 1) : acquisition des données, stockage de données et couche de présentation [6]. Le présent essai se focalise particulièrement sur la zone de stockage de données.

Figure 1 : Architecture d'entrepôt de données.



Source: Chaudhuri et Dayal, An Overview of Data Warehousing and OLAP Technology 1997, p. 2.

2.2.1 Acquisition des données

Cette zone couvre l'extraction des données provenant des diverses sources telles que les bases de données des systèmes de production ou systèmes opérationnels, des données internes et externes, les données provenant du système «*Enterprise Resource Planning*» (ERP), des fichiers plats, etc. L'acquisition des données est effectuée à travers l'outil d'extraction, transformation et chargement (ETC) qui permet de les recevoir, nettoyer, filtrer, formater et agréger avant de les transmettre dans l'entrepôt de données. Cette zone peut contenir un espace de stockage intermédiaire de données (Staging Area) servant à rassembler les informations et à examiner chaque donnée extraite après l'application ou non de certaines règles d'affaires. L'espace de stockage occupé par cette zone dépend de la complexité des types de données et les tailles des informations à extraire.

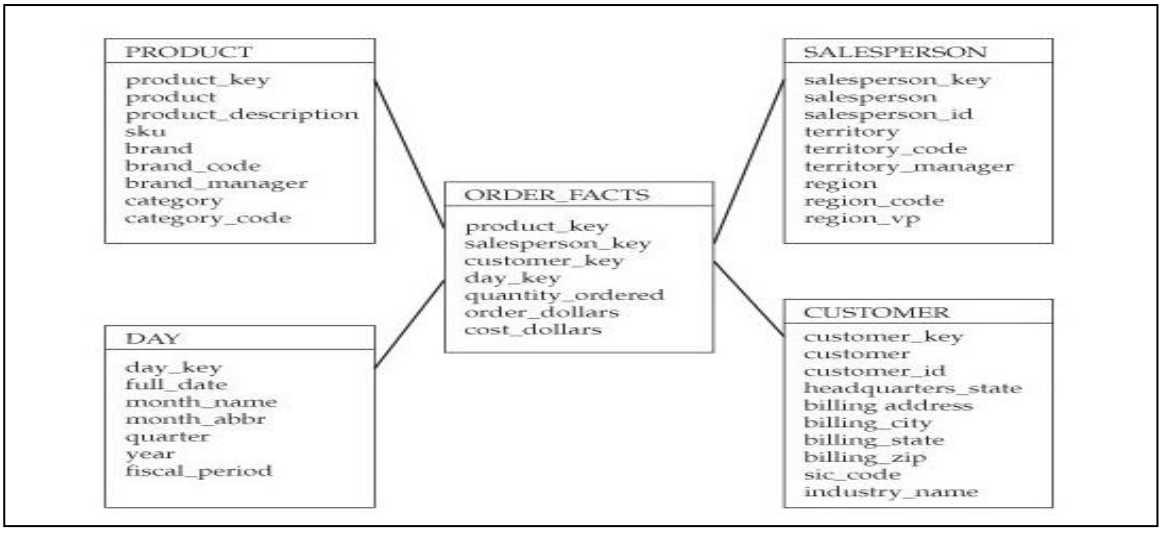
2.2.2 Zone de stockage de données

Il s'agit de l'espace dans lequel les données de l'entrepôt sont emmagasinées de façon détaillée ou agrégée. Elle contient des bases de données multidimensionnelles appelées schémas en étoile qui sont composés des tables de faits et des tables de dimensions.

L’architecture détermine au préalable la manière dont les données seront stockées. Elle peut être conçue selon différents modèles. On peut citer entre autres : l’architecture en bus de magasins de données, un modèle d'architecture centralisé, un modèle à structure concentrée et rayons (Hub and-spoke) et un modèle d'architecture fédéré (Watson et Ariyachandra 2010). Le choix d'une ou plusieurs architectures dépend des exigences et les besoins de l’entreprise.

Dans un schéma en étoile (voir la Figure 2), les tables de faits correspondent normalement à un seul processus d'affaires [7]. En d’autres termes, une table de faits représente un processus. Elles enregistrent les mesures générées par les événements du processus (par exemple : réception et envoi d'une commande). Les tables de faits contiennent typiquement un très grand nombre de lignes de données qui peuvent atteindre jusqu’à plusieurs milliards et qui couvrent environ plus de 90 % des données du modèle [8]. Les tables de dimensions contiennent les axes d’analyse (par exemple : produits, clients, employés, villes). Elles permettent de filtrer ou de restreindre les requêtes et d’étiqueter les résultats. La figure 2 montre un exemple de schéma en étoile pour un processus de commande de produits. Au centre de la figure se trouve la table de faits (ORDERS_FACTS) qui est entourée par les tables de dimensions (PRODUCT, DAY, SALESPERSON, CUSTOMER).

Figure 2 : Exemple de schéma en étoile pour un processus de commande de produits



Source: Adamson Christopher, Star Schema: The Complete Reference, Graw-Hill/Osborne

2010, chapitre 1 p.110

Afin d'accélérer la recherche d'informations et d'optimiser les temps de réponse à une requête, des index peuvent être insérés à l'intérieur des tables de faits. L'indexation des tables de faits dans un entrepôt de données est une opération délicate [9] : s'il y a peu d'index, les temps de chargement de l'entrepôt seront optimums mais les temps de réponse aux requêtes seront déplorables. Par contre, s'il y a trop d'index, les temps de chargement vont exploser mais les performances vis-à-vis des temps de réponse aux requêtes seront excellentes.

Afin de concilier un plan d'indexation optimum, certains critères doivent être tenus compte [10] :

- Le type d'entrepôt (archive en temps réel ou quasi-réel);
- La taille des tables de faits et la taille des tables de dimensions;
- Le nombre d'utilisateurs ayant accès à l'entrepôt (le nombre d'accès concurrents maximum à gérer);
- Le type d'accès aux données (ad-hoc ou via des interfaces d'applications structurées) et
- Le mode d'alimentation des données.

2.2.3 Couche de présentation

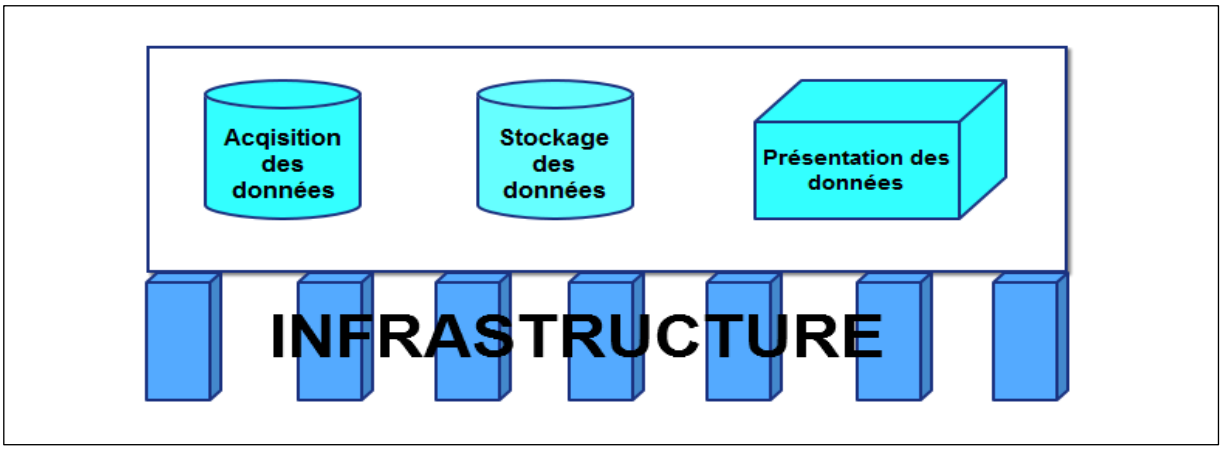
Cette zone est destinée à présenter les informations sous forme de tableaux de bord et des graphiques lisibles, interprétables et compréhensibles afin de générer des connaissances permettant aux décideurs de prendre les meilleures décisions. La couche de présentation est mise à profit par la mise en place des plusieurs logiciels permettant l'édition des états et les rapports statistiques, la fouille de données, l'exportation de données, etc.

2.3 Infrastructure d'entrepôt de données

L'infrastructure est la fondation soutenant l'architecture comme le montre la figure 3 [11]. L'infrastructure comprend plusieurs éléments tels que : le disque dur du serveur, le système

d'exploitation, le système de gestion de base de données (SGBD), le réseau Internet, les procédures et la formation, les fournisseurs d'outils de chaque composante architecturale et les LAN et WAN. Bref, l'infrastructure d'entrepôt de données intègre tous les éléments fondamentaux qui supportent l'architecture.

Figure 3 : Infrastructure d'entrepôt de données



Source: PONNIAH PAULRAJ, Data warehousing fundamentals for IT Professionals, Second Edition, p.164.

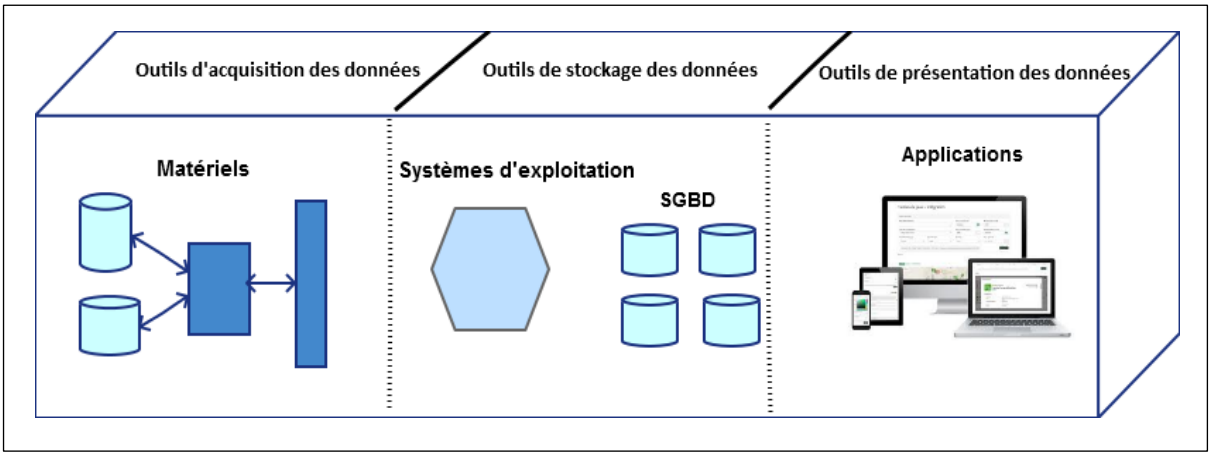
Les éléments de l'infrastructure d'entrepôt de données peuvent être classifiés en deux catégories principales : infrastructure opérationnelle et l'infrastructure physique [12]. Le présent essai s'intéresse particulièrement à l'infrastructure physique, plus précisément le disque où sont logés les schémas en étoile:

- **Infrastructure opérationnelle** : L'infrastructure opérationnelle est nécessaire pour soutenir chaque composante architecturale. Elle comprend : les individus, les procédures, la formation et la gestion des applications informatiques (l'acquisition et la maintenance). Elle permet d'assurer la fiabilité et le bon fonctionnement de l'architecture physique.

- **Infrastructure physique** : L'infrastructure physique assure les diverses fonctions et services des différents composants de l'architecture [12]. C'est une plateforme qui intègre essentiellement le disque dur, le système d'exploitation, le réseau, les

applications réseaux, le SGBD (voir la Figure 4). Dans le cadre du présent essai, on s'intéresse au disque, plus spécifiquement à sa capacité d'emmagasiner les données.

Figure 4 : Infrastructure physique d'entrepôt de données



Source: PONNIAH PAULRAJ, Data warehousing fundamentals for IT Professionals, Second Edition, p.166.

Le disque et le système d'exploitation constituent le maillon fondamental de l'environnement informatique de l'entrepôt de données. Toutes les opérations d'extraction, de transformation, d'intégration et d'agrégation sont exécutées dans le disque dur à travers le système d'exploitation choisi. Dans le contexte d'une entreprise manipulant de gros volumes de données, la taille du disque dur utilisé est de l'ordre de téraoctets. La section suivante s'intéresse aux méthodes utilisées pour estimer cet espace de stockage.

2.3.1 Méthodes d'estimation de la capacité de stockage d'un entrepôt de données

La capacité de stockage d'un entrepôt de données est la capacité du disque dur pour emmagasiner le volume de données. Dans le cadre du présent essai, il s'agit de l'espace du disque dur nécessaire pour rassembler les magasins de données (les modèles en étoile). En raison de la grosse taille de données à accumuler, les unités de mesure utilisées peuvent être le gigaoctet, le téraoctet voire le pétaoctet. En général, on peut distinguer deux

méthodes d'estimation de la capacité de l'entrepôt de données (Nabli et al. 2005) [13] : la méthode basée sur l'intuition et la méthode formelle.

2.3.2 Méthode d'estimation de la capacité de stockage basée sur l'intuition.

La méthode basée sur l'intuition est une méthode aléatoire. Elle est fondée sur les expériences passées du technicien en utilisant ou pas l'historique de données sans l'introduction d'un concept ou d'un cadre formel. Compte tenu de sa simplicité, elle peut être utilisée par des entreprises qui ne manipulent pas des gros volumes de données (en moyenne, moins d'un gigaoctet par jour). Cette méthode a ses avantages et ses inconvénients :

- **Avantages** : C'est une méthode relativement simple car elle ne nécessite pas de nombreuses informations et d'équations mathématiques. Elle n'exige pas la mobilisation de beaucoup de ressources car elle se base sur l'intuition et la connaissance des situations passées du technicien de l'infrastructure physique.
- **Inconvénients** : Les résultats de l'estimation sont imprécis. Ils sont soit surestimés, soit sous-estimés. C'est une méthode qui est très tributaire de l'expérience du technicien. Donc, les résultats obtenus ne sont pas fondés sur une base scientifique testable et répétable.

2.3.3 Méthode d'estimation de la capacité de stockage par la méthode formelle.

La méthode formelle est une méthode scientifique et expérimentale. Elle utilise les techniques quantitatives basées sur des modèles d'équations statistiques et mathématiques permettant d'analyser et de comprendre le comportement d'un système à partir d'un certain nombre de facteurs. Elle tient compte d'énormes quantités d'informations à exploiter provenant de l'historique de données d'entreprise. Selon l'article publié par Cécile Favre et ses collaborateurs [14], cette méthode se base sur une équation mathématique qui met en relation deux types de variables telles que :

- **La variable expliquée comme** la capacité de stockage de l'entrepôt de données.
- **Les variables explicatives** qui peuvent être regroupées en deux grandes catégories : variables liées à l'environnement du système (le nombre de lignes ajoutées à chaque chargement dans les tables des faits. Par exemple : chaque jour ou semaine, le nombre et la taille de tables de faits et de dimensions, la taille des index dans les tables des faits) et les variables liées à la politique de gestion des informations décisionnelles telles que la fréquence de rétention de données dans l'entrepôt exprimée en nombre de lignes de données à conserver pour une période déterminée (12 mois ou 36 mois selon les besoins de l'entreprise).

D'après Cécile Favre et al. [14], la forme globale de l'équation se présenter comme suit :

$$C = F(B, S) \qquad (1)$$

Où :

- F **:** le cadre formelle;
- C : est la capacité du disque dur;
- B : représente l'ensemble des variables liées à l'environnement du système;
- S : représente l'ensemble des variables associées à la politique de gestion des informations décisionnelles.

L'objectif final de l'équation (1) est de pouvoir quantifier la capacité de l'entrepôt de données, ou la variation de l'espace occupé par le magasin de données nommée $\Delta \boldsymbol{C}$ en fonction des variations de B et S (ΔB et ΔS).

Selon la nature des données à exploiter, pour estimer l'espace de stockage, la méthode utilisée peut faire appel soit à l'analyse discriminante bayésienne qui est une méthode probabiliste, soit à l'équation de régression linéaire dénommée log-linéaire ou encore équation de mesures de sensibilité. Le modèle log-linéaire est un outil statistique adapté à l'étude de structures complexes qui permet de souligner l'influence de tous les éléments entrant en jeu [15]. Le modèle Log-linéaire s'écrit sous la forme :

$$\mathrm{Ln}(C) = a_1 \mathrm{Ln}(BX_1) + \cdots + a_n \mathrm{Ln}(BX_n) + b_1 \mathrm{Ln}(SX_1) + \cdots + b_n \mathrm{Ln}(SX_n) + \varepsilon \quad (2)$$

Où :

- Ln est le logarithme népérien;
- C est une variable dépendante ou expliquée qui peut être la capacité de l'entrepôt de données;
- BX_i : i allant de 1 à n sont des variables indépendantes qui peuvent être des variables liées à l'environnement du système de l'entrepôt de données;
- SX_i : i allant de 1 à n sont des variables indépendantes qui peuvent être des variables associées à la politique de gestion des informations décisionnelles;
- a_i et b_i pour i allant de 1 à n sont des indicateurs de sensibilité exprimés en pourcentage;
- ε : Erreur d'estimation (ou résidu).

L'équation (2) permet d'évaluer la variation en pourcentage de la variable expliquée en fonction de la variation en pourcentage d'un facteur explicatif du modèle (a_i ou b_i). La méthode formelle a également ses avantages et ses inconvénients :

- **Avantages** : Les résultats de l'estimation sont précis avec des marges d'erreurs bien déterminées. La validité et la significativité des paramètres du modèle sont testables statistiquement. Le modèle permet de tenir compte des éléments liés à l'environnement du système et aux décisions politiques prises par l'entreprise. La méthode formelle permet d'effectuer des prédictions à court, à moyen et à long terme. L'expérience est répétable selon les contextes dans lesquels l'entreprise se situe.

- **Inconvénients** : La méthode formelle exige des connaissances approfondies en mathématiques et en statistiques. Sa mise en œuvre nécessite une bonne organisation et planification des ressources comparativement à la méthode basée sur l'intuition. La mise en place de la méthode formelle oblige l'entreprise à avoir l'historique des données utilisées comme étant l'échantillon.

2.4 Conclusions

Ce chapitre a permis de définir les concepts généraux et l'infrastructure d'entrepôt de données. L'infrastructure d'entrepôt de données supporte trois principales zones telles que : l'acquisition des données, la zone de stockage et la couche de présentation. Cet essai, se focalise particulièrement sur la zone de stockage. Deux méthodes d'estimation de la capacité de stockage d'entrepôt de données ont été soulignées, à savoir : la méthode basée sur l'intuition et la méthode formelle qui utilise le modèle d'équation mathématique et statistique. Certains chercheurs comme Cécile F., Fadila B. et al. [14] ont identifié les groupes de variables qui peuvent influencer la capacité de stockage d'un entrepôt de données. Cependant, leurs travaux n'ont pas stipulé explicitement la forme mathématique modélisant la relation entre ces groupes de variables et la capacité d'accumulation des données. Selon le type et la nature des données à exploiter, la méthode d'estimation peut faire appel : soit à l'analyse discriminante bayésienne qui est une méthode probabiliste, soit au modèle de régression dit log**-linéaire.** Le modèle **log-linéaire** ou encore équation de mesures de sensibilité peut être mis à contribution afin de déterminer la croissance de cette capacité de stockage, compte tenu du fait qu'il est adapté à l'étude de structures complexes permettant de considérer l'influence de différents facteurs entrant en jeu [15].

La revue de littérature a permis de procéder à la recension des certains écrits pertinents qui sont définis dans la liste des références. Elle a permis de documenter et de délimiter le problème tout en précisant les concepts en jeu. Le chapitre suivant traitera la problématique et le détail du cadre conceptuel de l'étude.

Chapitre 3

Problématique

3.1 Introduction

Pour faire face à la concurrence sur le marché, la plupart des entreprises ont tendance à développer des stratégies d'intelligences d'affaires permettant de générer les connaissances sur leurs activités. La mise en œuvre de ces stratégies nécessite d'énormes quantités d'informations provenant de sources multiples qui doivent être emmagasinées dans un entrepôt de données. L'une des principales contraintes qui pèsent sur la compagnie est sa capacité limitée à rassembler et à accumuler ces informations. Selon l'environnement du système à mettre en place et les besoins d'affaires définies, la connaissance au préalable de l'espace du disque dur à allouer pour emmagasiner la masse d'informations revêt une importance capitale afin de suivre et contrôler les activités de façon optimale. En d'autres termes, il s'avère pertinent de quantifier de façon précise la croissance de la capacité de l'entrepôt de données à partir des éléments de la politique de gestion des informations décisionnelles et de l'environnement du système. Cette démarche permet la mise en place des instruments ou des indicateurs d'aide à la décision et facilite le travail du technicien qui doit trouver des solutions aux problèmes de manque d'espace.

Dans le chapitre 2, deux méthodes d'estimation de la capacité de stockage de l'entrepôt de données ont été soulignées : la méthode basée sur l'intuition et la méthode formelle. La méthode basée sur l'intuition est une méthode aléatoire qui n'utilise pas de concepts statistiques. Elle est fondée sur les expériences passées du technicien. La méthode formelle est une méthode scientifique et expérimentale. Elle utilise les techniques quantitatives basées sur des modèles d'équations statistiques et mathématiques permettant d'analyser et de comprendre le comportement d'un système à partir d'un certain nombre de facteurs. Le présent essai s'intéresse à l'analyse et à la mise en œuvre de la méthode formelle en partant de l'historique de données d'une entreprise.

En tenant compte du type et la nature des données à exploiter, la méthode formelle peut faire appel, soit à l'analyse discriminante bayésienne qui est une méthode probabiliste, soit à l'équation de régression linéaire dénommée log-linéaire ou encore équation de mesures de sensibilité (voir équation 2). Le modèle log-linéaire est un modèle statistique adapté à l'étude de structures complexes permettant de mettre en exergue l'influence des différents facteurs entrant en jeu [15].

Le présent chapitre vise à définir la question de recherche et à préciser l'hypothèse et le cadre conceptuel de l'étude.

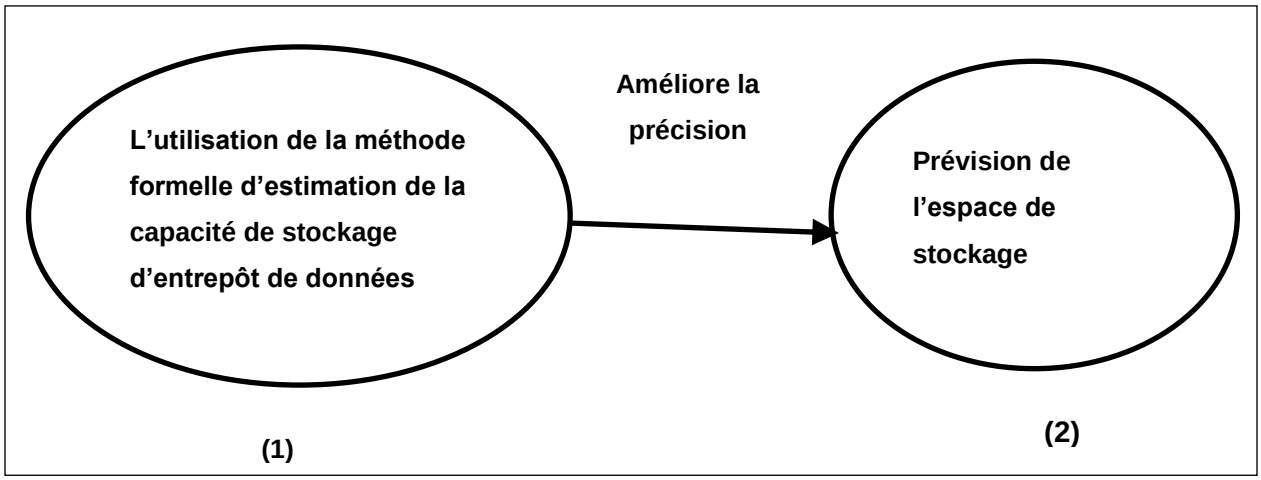
3.1.1 Question de recherche et hypothèse

Compte tenu des problèmes engendrés par l'absence d'outils d'estimation dans l'entreprise, la principale question est :

«L'estimation de la capacité de stockage de l'entrepôt de données en fonction des éléments de la politique de gestion des informations décisionnelles et de l'environnement du système par la méthode formelle permet-elle d'améliorer la précision de l'estimation de l'espace de stockage requis?»

Cette question centrale conduit à émettre l'hypothèse suivante : «la connaissance de la capacité de stockage de l'entrepôt de données à partir des facteurs liés à la politique de gestion des informations décisionnelles et à l'environnement du système à travers la méthode formelle améliore la précision de l'estimation des besoins réels en espace de stockage, compte tenu de l'évolution des besoins informationnels et la hausse du volume de données à exploiter.». L'espace mis en évidence dans le présent essai est la capacité de stockage du disque dur (exprimée en gigaoctet ou en téraoctet) de l'entrepôt de données. La question de recherche évoquée ci-haut peut être résumée par le cadre conceptuel ci-après :

Figure 5 : Cadre conceptuel de l'étude



(1) : Variable indépendante de l'étude
(2) : Variable dépendante de l'étude

Le cadre conceptuel permet de mettre en relation la variable indépendante (1) et la variable dépendante (2) de l'étude. La variable dépendante est celle mesurée, qui sert à vérifier si la précision s'améliore en fonction de l'utilisation de la méthode d'estimation.

3.1.2 Limites de l'étude

Cet essai tient compte des données d'une entreprise utilisant des gros volumes de données de plus d'un gigaoctet par jour. L'entreprise ciblée pour l'expérimentation est située dans la région de Montréal. Elle pratique les techniques d'entrepôts de données séparant le système opérationnel à celui de l'informationnel (selon le modèle de Ralph Kimball). Elle évalue la capacité de l'entrepôt de données de manière aléatoire. L'infrastructure à mettre en évidence est le disque dur enregistrant uniquement les modèles multidimensionnelles (modèles en étoile) car il s'agit de la zone où se trouve la production croissante et massive des données [1]. Les espaces occupés par les métadonnées et les zones de stockage intermédiaire seront exclus.

Le modèle formel à analyser ne considère pas de façon exhaustive tous les facteurs pouvant influencer la capacité de stockage de l'entrepôt de données. Il tient plutôt compte de ceux qui ont des corrélations significatives avec cette dernière. Le processus de sélection de ces facteurs est détaillé dans le chapitre 4.

3.2 Méthodologie proposée

3.2.1 Type de recherche

L'approche proposée dans le présent essai est l'analyse quantitative à travers laquelle il est question de :

- Déterminer s'il existe une relation entre chaque variable indépendante et la capacité de stockage de l'entrepôt de données;
- Caractériser la forme de la liaison entre chaque variable indépendante et la capacité de stockage d'entrepôt de données: positive ou négative, linéaire ou non linéaire, monotone ou non monotone;
- Vérifier si la liaison est statistiquement significative;
- Quantifier l'intensité de la liaison;
- Valider la liaison identifiée en répondant à la question : Est-ce que cette liaison n'est pas le fruit d'un simple artefact ou le produit d'autres informations sous-jacentes dans les données?
- Établir le modèle d'équation statistique décrivant la relation en vue d'effectuer la prédiction de la croissance de la capacité de stockage de l'entrepôt selon divers scénarios.

La description de cette approche est détaillée dans le chapitre 4.

Chapitre 4

Approche proposée

Ce chapitre traite l'approche utilisée pour répondre à la question de recherche évoquée dans le chapitre 3. Il permet de mettre en exergue les différentes étapes à suivre qui constituent la stratégie de recherche.

4.1 Introduction

Dans le cadre du présent essai, la majorité des variables d'analyse sont du type quantitatif. La stratégie de recherche adoptée est du type quantitatif. Une étude quantitative permet non seulement de vérifier les corrélations entre les variables, mais également d'établir un modèle d'équation de comportement liant la variable dépendante à ses facteurs explicatifs. Plus précisément cette démarche consiste à :

- Déterminer s'il existe une relation entre chaque variable indépendante et la capacité de stockage de l'entrepôt de données;
- Caractériser la forme de la relation entre chaque variable indépendante et la capacité de stockage d'entrepôt de données (positive ou négative, linéaire ou non linéaire, monotone ou non monotone);
- Vérifier si la liaison est statistiquement significative;
- Quantifier l'intensité de la liaison;
- Valider la liaison identifiée en répondant à la question : Est-ce que cette liaison est le fruit d'un simple artefact ou encore le produit d'autres informations sous-jacentes dans les données?
- Établir le modèle d'équation statistique décrivant le lien en vue d'effectuer la prédiction de la croissance de la capacité de l'entrepôt selon différents scénarios.

La stratégie de recherche à adopter respecte les étapes suivantes :

- Réalisation d'une enquête par sondage sur les déterminants de la capacité de stockage d'entrepôt de données;
- Choix de la base de données de l'entreprise ciblée pour l'expérimentation et collecte de données d'analyse afin d'établir le modèle d'estimation;
- Analyse des données;
- Validation des résultats en comparant les résultats d'estimation par la méthode formelle à ceux obtenus par la méthode intuitive vis-à-vis des besoins réels en capacité de stockage;
- Répondre à la question de recherche.

4.2 Stratégie de recherche

4.2.1 Enquête par sondage sur les déterminants de la capacité de stockage.

4.2.1.1 Objectif

L'enquête par sondage sur les déterminants de la capacité d'entrepôt de données a deux principaux objectifs : comprendre les pratiques des entreprises en matière de la gestion de la capacité de stockage, identifier à priori les variables qui influencent la capacité de stockage selon les expériences sur le terrain afin d'éviter la régression fallacieuse. La régression fallacieuse désigne une situation dans laquelle l'introduction d'une variable explicative à corrélation quasi nulle avec la variable dépendante dans le modèle d'estimation fait apparaître des résultats erronés, trop optimistes, qui font croire à une relation parfaite entre les variables alors que ce n'est pas le cas [16].

Cette enquête ne consiste pas à collecter les capacités de stockage des entrepôts de données de toutes les entreprises visées par l'échantillonnage. Elle vise plutôt l'indentification à priori des variables ou les facteurs explicatifs de la capacité de stockages selon les expériences pratiques. Elle permet d'avoir à l'avance les meilleures indications permettant de valider le modèle d'estimation. Le questionnaire d'enquête par sondage est présenté à l'annexe 1.

4.2.1.2 Échantillon

Dans le cadre de cette étude, la population cible est l'entreprise utilisant les techniques d'entrepôt et consommant des données plus d'un gigaoctet par jour. Pour des raisons financières et de simplicité, l'enquête peut se limiter au maximum aux 250 individus employés des entreprises situées dans la ville de Montréal. La méthode d'échantillonnage appliquée est le sondage aléatoire simple. Il s'agit d'une méthode probabiliste qui suppose l'existence d'une base de sondage. Pour collecter les données, des questionnaires en ligne seront adressés principalement aux techniciens responsables de la gestion des entrepôts de données à travers leurs adresses électroniques. Chaque message électronique contient :

- L'explication du contexte, l'objectif de l'enquête et la mention d'anonymat afin d'obtenir l'adhésion du répondant et
- Le lien hypertexte à travers lequel le répondant peut cliquer afin de remplir le formulaire.

Le questionnaire comprend principalement deux sections : L'identification de l'entreprise, et les questions relatives aux variables ou aux facteurs explicatifs de la capacité de l'entrepôt de données.

4.2.2 Choix de la base de données de l'entreprise ciblée pour l'expérimentation et collecte de données d'analyse

Cette étape consiste à choisir la base de données de l'entreprise pour l'expérimentation qui permet d'établir le modèle d'estimation. L'entreprise en question doit figurer dans l'échantillon de l'enquête par sondage évoquée ci-haut. Elle doit présenter le plus possible les caractéristiques de l'ensemble des entreprises échantillonnées par rapport aux réponses obtenues.

Les données d'analyse comprennent l'historique d'observations pour une période de 22 mois (janvier 2015 à d'octobre 2016) par magasin de données. Elles renseignent sur les variables selon les groupes définis dans le tableau 3 :

Tableau 3 : Liste des variables d'analyse

Groupe	Nomenclature	Désignation	Unité de mesure
Identifiant	**Code Magasin**	Identifiant du magasin de données	
Variables associées au système (Indépendantes)	**TailleTblTFact**	Taille ou espace occupé par les tables de faits	Giga-octets (GO)
	TailleTblDim	Taille ou espace occupé par les tables de dimensions	Giga-octets (GO)
	TailleIndex	Taille des index dans les tables des faits	Méga-octets (MO) ou Giga-octets (GO)
Variable liée à la politique de gestion de données associées au système (Indépendante)	**FreqRet**	Fréquence de rétention ou durée de stockage de données dans l'entrepôt exprimée en nombre de mois	Nombre entier
Variable dépendante	**Capacite**	Espace occupé par le magasin de données	Giga-octets (GO)
Variables de contrôle	**C_Intuitive**	Capacité de stockage estimée selon la méthode intuitive	Giga-octets (GO)
	C_Besoin	Besoin en croissance de la capacité de stockage après estimation par intuition	Giga-octets (GO)

Les variables indépendantes et dépendantes sont des variables d'analyse nécessaires pour bâtir le modèle d'estimation par la méthode formelle. Hormis ces variables, les indicateurs tels que C_Intuitive et C_Besoin servent de variables de contrôle. Ils permettent de juger la validité et la force prédictive du modèle à construire en comparant les résultats obtenus avec les besoins en croissance de la capacité de stockage après estimation par intuition.

Les observations collectées sont temporelles et présentées sous la forme de données en panel (données en coupe transversale). La présentation en panel permet d'étudier et de modéliser le comportement d'un système suivant le temps [17]. Le tableau 4 présente la structure du tableau de données.

Tableau 4 : Structure du tableau de données collectées

Date	Code Magasin	TailleTblTFact	TailleTblDim	TailleIndex	FreqRet	Capacite	C_Intuitive	C_Besoin
01/01/2015	1					25	x	x
01/02/2015	1					58	x	
01/03/2015	1					87		
01/04/2015	1					98		x
01/05/2015	1					100		
01/06/2015	1					101		
01/07/2015	1					102		
01/08/2015	1					193		x
01/09/2015	1					197		
01/10/2015	1					198	x	
01/01/2015	2					25	x	x
01/02/2015	2					58		
01/03/2015	2					87		x
01/04/2015	2					98	x	
01/05/2015	2					100		x
01/06/2015	2					101		
01/07/2015	2					102		
01/08/2015	2					193		x
01/09/2015	2					197		
01/10/2015	2					198	x	x

Ce tableau renseigne sur toutes les valeurs prises par les variables explicatives et la variable dépendante par magasin de données pour une période de 22 mois (janvier 2015 à octobre 2016). Il présente également les données relatives aux capacités estimées intuitivement et celles dont l'entreprise avait besoin par magasin de données.

4.2.3 Analyse des données

- Analyse statistique descriptive ou exploratoire :

Ce type d'analyse a pour objectif de résumer, synthétiser l'information contenue dans la série statistique et de mettre en évidence ses propriétés. Les données sont analysées variable par variable en faisant ressortir des indicateurs tels que la moyenne, la médiane, l'écart-type, la variance et la corrélation linéaire simple. Les outils utilisés pour cette analyse sont les tableaux, les graphiques.

Pour tenir compte des interactions entre les variables, les méthodes d'analyses multidimensionnelles telles que l'analyse en composante principale (ACP), l'analyse en correspondance multiple (ACM) et la classification ascendante hiérarchique (CAH) sont mises à contribution. Ces méthodes permettent de traiter simultanément un nombre quelconque de variables en dégageant les corrélations entre elles et les caractéristiques de chaque magasin de données vis-à-vis de des variables d'analyse. Les outils d'aide à l'interprétation utilisés sont : les valeurs propres (indicateurs de dispersion entre les variables), le cercle de corrélation, les graphiques des observations (magasins de données) et des variables explicatives.

-**Analyse prédictive :**

L'analyse prédictive, est un domaine de l'analyse statistique qui extrait l'information à partir des données historiques pour prédire les tendances futures et les motifs de comportement. Le cœur de l'analyse prédictive se fonde sur la capture des relations entre les variables explicatives et les variables expliquées, ou prédites, issues des occurrences passées, et l'exploitation de ces relations pour prédire les résultats futurs.

Dépendamment de la nature des données à exploiter, il existe plusieurs méthodes statistiques pour effectuer l'analyse prédictive. Dans le cadre du présent essai, deux méthodes seront abordées à savoir :

i) **L'analyse discriminante** : elle consiste à rechercher les quantités qui sont en combinaisons linéaires des variables explicatives permettant de séparer le mieux possible les groupes de magasins de données homogènes. Ces quantités sont appelées facteurs discriminants qui peuvent être exprimées sous la forme [18] :

$$\begin{cases} V_1 = a_{11}X_1 + \cdots + a_{n1}X_n \\ \cdots\cdots \\ V_p = a_{1p}X_1 + \cdots + a_{np}X_n \end{cases}$$

Où V_p est facteurs discriminants et a_{np} est le paramètre de la variable explicative X_n. Les facteurs discriminants peuvent être considérés comme les indicateurs d'association ou d'inclusion à un groupe.

Pour effectuer la prédiction, cette méthode se base sur des fonctions linéaires discriminantes qui sont nécessaires au calcul de probabilité d'appartenance à chaque groupe de magasins de données. Ces fonctions sont écrites comme suit :

$$\begin{cases} f_1(X) = a_{11}X_1 + \cdots + a_{n1}X_n \\ \cdots\cdots \\ f_k(X) = a_{1k}X_1 + \cdots + a_{nk}X_n \end{cases}$$

Où f_k est la fonction discriminante associée au groupe k et a_{nk} est le paramètre de la variable explicativeX_n. La probabilité d'appartenance à un groupe est alors exprimée par la quantité :

$$P_j = \frac{e^{f_j(X)}}{\sum_{j=1}^{k} e^{f_j(X)}}$$

Pour assurer la validité du modèle, différents tests statistiques doivent être concluants tels que : le test de significativité global et individuel la force discriminante des variables explicatives et le test d'égalité des matrices de covariance entre les groupes de magasin de données. L'analyse de la matrice de confusion de classement permet également d'apprécier la force prédictive du modèle.

ii) **Régression log-linéaire sur les données en panel** : Les données de panel, ou données croisées possèdent deux dimensions : la dimension individuelle et la dimension temporelle. Il s'agit des données qui rapportent les valeurs des variables prises pour un ensemble ou panel d'individus sur une suite de périodes. Il faut noter qu'en général, le mot panel désigne un échantillon fixe de consommateurs interrogés à différentes périodes, dans le cadre du présent essai, le terme données en panel est simplement synonyme de données croisées ayant une dimension individuelle et une dimension temporelle. Si on fixe l'individu observé, on obtient la série chronologique, soit la coupe longitudinale. Par contre, si l'on fixe la période d'observation, on est en présence des données en coupe transversale et instantanée pour l'ensemble d'individus. L'objectif principal est la mise en œuvre du modèle d'estimation ci- après :

$$\mathrm{Ln(C)} = \mathrm{a_{1t}Ln(X_{1t})} + \cdots + \mathrm{a_{nt}Ln(X_{nt})} + \varepsilon$$

Ce modèle permet de calculer les paramètres a_{it} (i allant de 1 à n) qui sont interprétés comme la variation en pourcentage de la capacité de l'entrepôt de données lorsque la variable explicative $\boldsymbol{X_{it}}$ (i allant de 1 à n) augmente de 1 % ($\boldsymbol{X_{it}}$ note l'observation de la variable X pour l'individu i à la période t).

Avant d'élaborer un modèle avec des données temporelles tel que les données en panel, il faut s'assurer qu'elles conservent une distribution constante dans le temps [19] : c'est le concept de stationnarité. Une série chronologique est dite stationnaire si la distribution des variables chronologiques ne varie pas dans le temps [20]. Cela veut juste dire que si les valeurs antérieures de nos variables sont semblables à celles des futures, il serait raisonnable d'utiliser le passé pour tenter de prédire le futur. Si les données ne sont pas stationnaires, alors on peut se retrouver avec trois situations inconfortables : biais de prévision; inefficacité de prévision; mauvaise inférence car les paramètres du modèle d'estimation sont biaisés. Il existe trois sources principales de non-stationnarité [21] :

- Changement structurel (break) : La fonction de régression change dans le temps, soit de façon discrète, soit de façon graduelle. Cette situation se reproduit, par exemple, dans le cas d'un changement politique ou stratégique dans l'entreprise qui tend à modifier brusquement l'espace de l'entrepôt de données;
- Tendance déterministe : les données suivent une tendance qui est exprimée en fonction du temps t;
- Tendance stochastique (racine unitaire) : Les données suivent une marche aléatoire, c'est-à-dire pour une variable explicative donnée, il existe la persistance d'une mémoire assez longue dans le temps.

Afin d'assurer sa conformité statistique, le modèle est soumis à des tests de validation tels que [22] : test de stationnarité des données d'observations, test de significativité globale (Test de Fisher), test de significativité du paramètre (a_{it}) de chaque variable explicative (test de Student)

Avant d'effectuer des simulations, l'analyse de résidus (ε) sera entreprise pour apprécier le niveau de stabilité du modèle.

4.2.4 Approche de validation des résultats

Les forces prédictives des modèles sont appréciées grâce à la comparaison des résultats d'estimation obtenus avec ceux de la méthode intuitive vis-à-vis des besoins réels en capacité de stockage. Plus les valeurs estimées sont proches des besoins réels avec une marge d'erreur de 5 %, plus le modèle a une bonne capacité à prédire les situations à venir. L'exemple de données de validation dans le cadre du modèle log-linéaire est présenté dans un tableau comparatif ci-après.

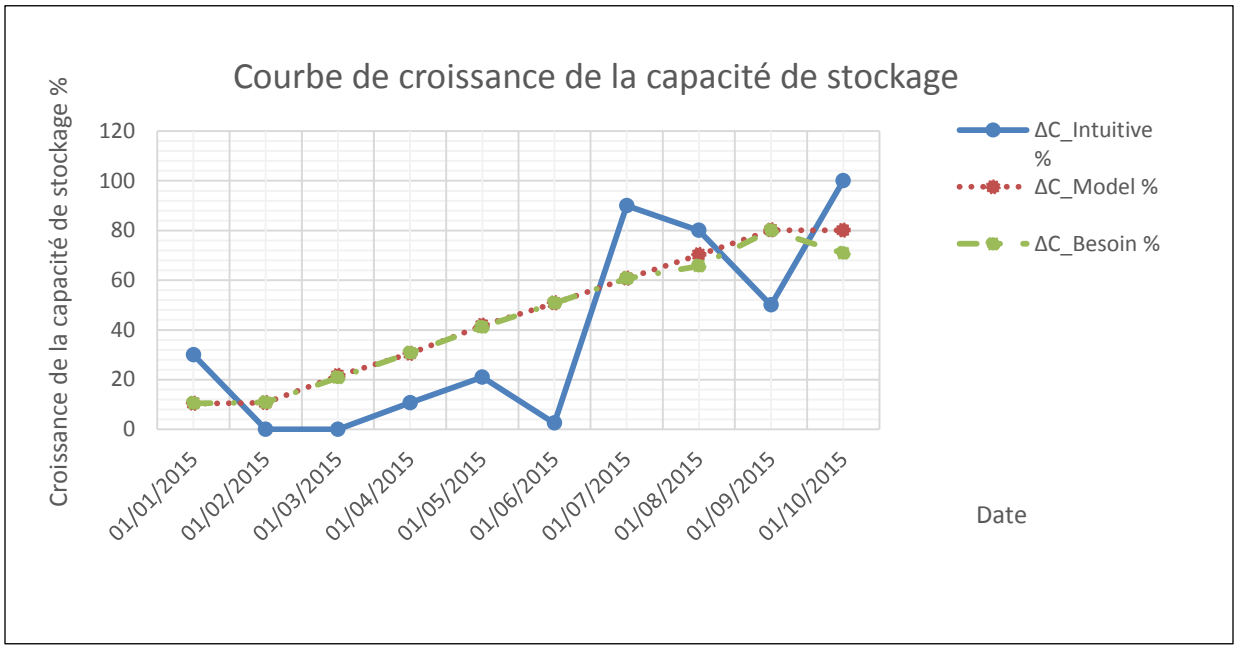
Tableau 5 : Exemple de données de validation

Date	Code Magasin	ΔC_Intuitive %	ΔC_Model %	ΔC_Besoin %
01/01/2015	1	30	10,35	10,5
01/02/2015	1	0	10,67	10,8
01/03/2015	1	0	21,55	20,8
01/04/2015	1	10,7	30,51	30,73
01/05/2015	1	21	41,75	41,2
01/06/2015	1	2,5	50,79	50,8
01/07/2015	1	90	60,7	60,7
01/08/2015	1	80	70,25	65,7
01/09/2015	1	50	80,12	80,1
01/10/2015	1	100	80,1	70,9

ΔC_Intuitive **:** Croissance de la capacité de stockage en % estimée selon la méthode intuitive.
ΔC_Model **:** Croissance de la capacité de stockage en % estimée à partir du modèle.
ΔC_Besoin **:** Besoin réel en croissance de la capacité de stockage en %.

Afin d'assurer une bonne interprétation et d'obtenir une bonne visibilité des comparaisons de résultats, les données de validation seront tracées sous la forme des graphiques indiqués ci-après :

Figure 6 : Comparaison de la capacité de stockage



Pour chaque magasin de données et pour une période précise, si la série de données générées par le modèle se rapproche davantage de la courbe de croissance des besoins réels par rapport à la courbe de la croissance de la capacité de stockage déterminée intuitivement, alors le modèle fournit les meilleurs résultats que la méthode intuitive.

4.2.5 Résultats attendus

Après avoir établi le modèle d'estimation selon la méthode formelle, la significativité statistique du paramètre de chaque variable explicative est présentée au seuil de 5 % de marge d'erreur. Cette significativité est exprimée en termes de probabilité ou de risque de se tromper dans l'interprétation du modèle. Plus la probabilité de se tromper est faible (inférieur à 5 %), plus le paramètre est significatif, donc, plus le modèle est meilleur. Le tableau de significativité des paramètres est établi comme suit :

Tableau 6 : Significativité des paramètres

Variables /Paramètres	TailleTblTFact	TailleTblDim	TailleIndex	FreqRet
Paramètres	0,90	0,56	0,86	0,89
Probabilité (significativité)	0,001(**)	0,004(**)	0,001(**)	0,003(**)

TailleTblTFact : Taille ou espace occupé par les tables de faits dans le système en giga-octets
TailleTblDim : Taille ou espace occupé par les tables de dimension en giga-octets
TailleIndex : Taille des index dans les tables des faits en méga-octets
FreqRet : Fréquence de rétention ou la durée de stockage de données dans l'entrepôt exprimée en nombre de mois
(**) : Paramètre significatif au seuil de 5 %

Pour établir le modèle d'estimation, divers outils d'analyses statistiques peuvent être utilisés, tels que : Statistical Analysis System (SAS), Excel Stat, R, Matlab, Eviews etc. Ces applications offrent des outils d'aide à l'interprétation des résultats d'analyse relativement faciles à comprendre avec des indicateurs pré-calculés.

Les signes des paramètres des variables explicatives font partie des indices d'aide à interprétation du modèle d'estimation par la méthode formelle.

En guise d'exemple, les signes escomptés peuvent être résumés dans le tableau suivant :

Tableau 7 : Signes des paramètres du modèle

Variable	Signe du paramètre (a_{it})
TailleTblTFact	+
TailleTblDim	+
TailleIndex	+
FreqRet	+

TailleTblTFact: Taille ou espace occupé par les tables de faits dans le système en giga-octets
TailleTblDim: Taille ou espace occupé par les tables de dimension en giga-octets
TailleIndex: Taille des index dans les tables des faits en méga-octets
FreqRet: Fréquence de rétention ou la durée de stockage de données dans l'entrepôt exprimée en nombre de mois

Dans le cadre du modèle de Régression log-linéaire sur les données en panel, le signe plus (+) est interprété comme suit : une augmentation de 1 % de la variable explicative (nombre de tables de faits par exemple) nécessite une hausse de x % de la capacité de stockage de l'entrepôt de données.

En conclusion, la stratégie a permis une compréhension plus raffinée de la problématique de l'étude et de développer une démarche structurée de l'approche d'analyse à suivre. Les résultats de cette démarche d'analyse sont présentés dans le chapitre 5.

Chapitre 5

Analyse des résultats

Dans le chapitre 4, il était question d'énoncer l'approche et les démarches à suivre afin de répondre à la question de recherche soulevée au chapitre 3. Le chapitre 5 consiste à présenter les résultats qui permettent de confirmer ou d'infirmer l'hypothèse de recherche qui a été stipulée comme suit : «la connaissance de la capacité de stockage de l'entrepôt de données à partir des facteurs liés à la politique de gestion des informations décisionnelles et à l'environnement du système à travers la méthode formelle améliore la précision de l'estimation des besoins réels en espace de stockage».

Le présent chapitre traite deux points principaux :

i) Analyse des résultats d'enquête par sondage effectuée auprès des diverses entreprises sélectionnées.
ii) Mise en œuvre du modèle d'estimation d'espace de stockage d'entrepôt par la méthode formelle et analyse des résultats.

5.1 Analyse des résultats d'enquête par sondage

Cette section vise à dépouiller et à analyser les réponses des individus enquêtés. L'enquête par sondage a été réalisée dans le but d'obtenir les opinions des techniciens par rapport aux déterminants de la capacité de stockage d'entrepôt de données. L'objectif principal est de comprendre les pratiques des entreprises au sujet de la gestion de la capacité en vue d'identifier à priori les variables qui influencent celle-ci en tenant compte des expériences sur le terrain ou des réalités vécues.

Cette partie met en évidence non seulement les caractéristiques et les profils des individus enquêtés mais également l'analyse des réponses aux questions liées aux facteurs explicatifs de la capacité de stockage de l'entrepôt de données de leurs compagnies.

5.1.1 Caractéristiques et profils des individus enquêtés

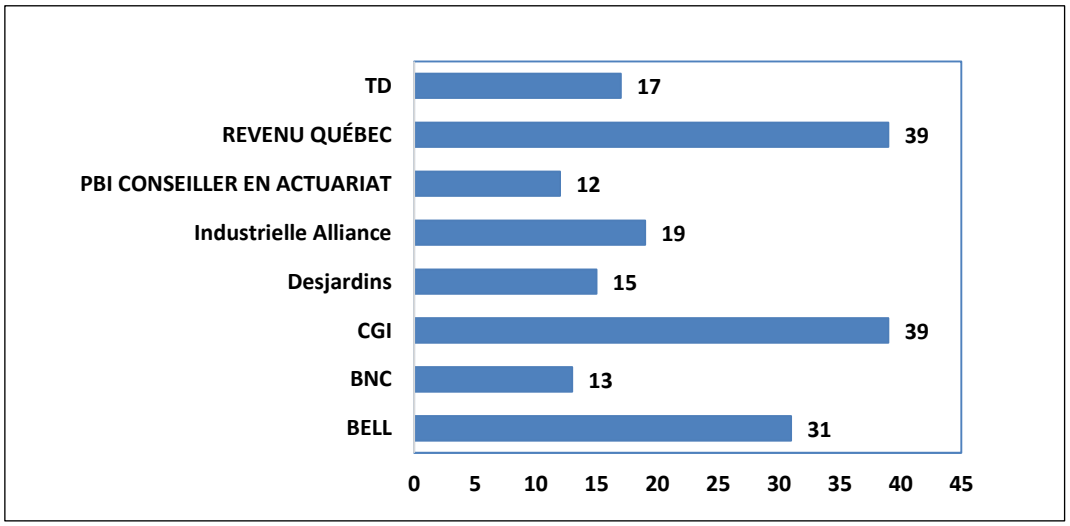
Les individus enquêtés ont été sélectionnés selon la méthode aléatoire simple à partir d'une base composée par 12 entreprises employant environ 250 personnes qui travaillent sur des projets liés directement au domaine de la gestion des entrepôts de données ou au domaine de l'informatique décisionnelle. Parmi les 12 entreprises, huit ont été sélectionnées aléatoirement. Elles emploient environ 200 employés dans divers secteurs d'activités.

Parmi les 200 employés ciblés dans le sondage, 185 ont répondu aux questionnaires en ligne qui ont été leurs adressés. Ce qui correspond au taux de non réponse égal 7,5 %. Un taux de non-réponse plus faible indique un risque peu élevé de biais dû à la non-réponse et par conséquent, un risque moins élevé d'imprécisions [23]. Le seuil utilisé est de l'ordre de 10 % à 15 % pour les enquêtes auprès des entreprises [24].

5.1.1.1 Effectif des enquêtés par entreprise

Les 185 individus enquêtés sont des employés de huit entreprises dont la répartition est indiquée dans le graphique suivant.

Figure 7 : Effectif des répondants par entreprise



Source : Enquête effectuée auprès des entreprises

Les employés de CGI et du REVENU QUÉBEC ont les nombres de répondants les plus élevés. Ces deux entreprises couvrent plus de 40 % du nombre de répondants.

5.1.1.2 Répartition des répondants selon le secteur d'activités

Les individus enquêtés travaillent principalement dans cinq secteurs d'activités différents dont la répartition est présentée dans le tableau suivant.

Tableau 8 : Effectif des répondants par secteur d'activité

Secteur d'activité	Effectif	Proportion %
FINANCES/ACTUARIAT	40	21,62
ASSURANCE	36	19,46
FONCTION PUBLIQUE	39	21,08
IT (secteur Privé)	39	21,08
TELECOM	31	16,76
TOTAL	185	100,00

Source : Enquête effectuée auprès des entreprises

La proportion du nombre de répondants est presque similaire pour chaque secteur d'activités, sauf pour le secteur de TELECOM qui affiche une proportion moins de 17 %.

5.1.1.3 Répartition des répondants par titre ou par fonction

Tableau 9 : Effectif des répondants selon leur titre ou leur fonction

Titre	Effectif	Proportion %
Analyste BI (Business Intelligence)	29	15,68
Analyste Programmeur	21	11,35
Analyste Fonctionnel	15	8,11
Gestionnaire de Base de données(DBA)	97	52,43
Gestionnaire de Projets	23	12,43
Total	185	100,00

Source : Enquête effectuée auprès des entreprises

Plus de la moitié des répondants (52,43 %) sont des gestionnaires de bases de données qui ont la responsabilité de suivre l'évolution de l'espace occupé par les entrepôts de données ainsi que celui de ses composantes (tables des faits, tables de dimensions, taille des index)

5.1.2 Analyse des réponses aux questions liées aux facteurs explicatifs de la capacité de stockage de l'entrepôt de données

L'analyse des réponses aux questions liées aux facteurs explicatifs de la capacité de stockage de l'entrepôt de données peut être menée selon deux approches :

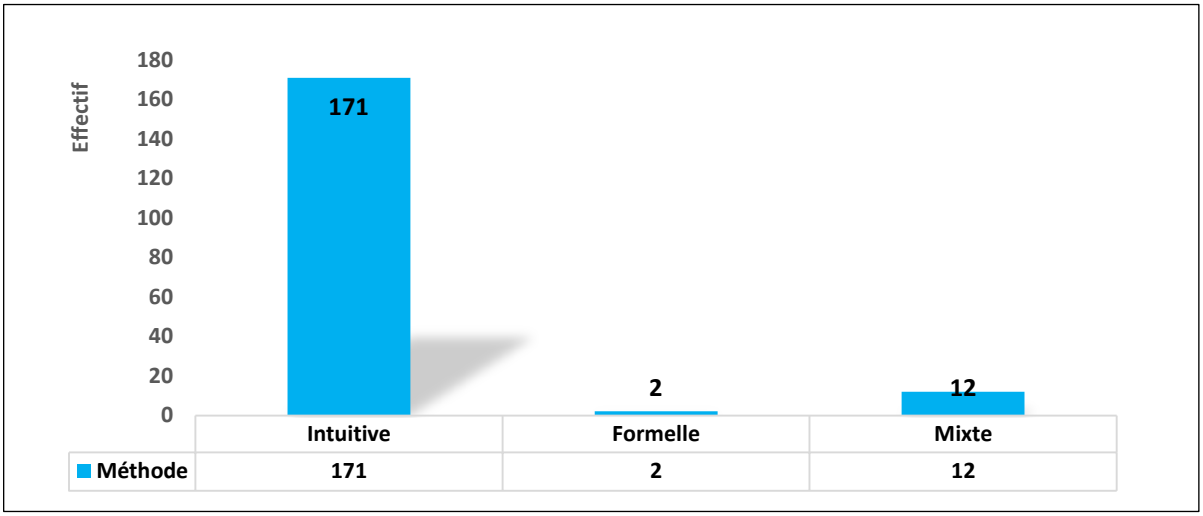
i) Approche unidimensionnelle : Elle consiste à analyser les données par variable. Cette approche met en évidence uniquement l'influence d'une variable explicative ou d'un facteur explicatif sur la capacité de stockage sans tenir compte des interactions des autres facteurs. Dans cette technique, on produit les tableaux de fréquence de chaque variable au regard de l'espace occupé par l'entrepôt. Ensuite, on effectue le test d'indépendance ou de liaison entre chaque la variable explicative et l'espace occupé à travers la statistique de khi-2.

ii) Approche multidimensionnelle : Elle tient compte de l'interaction de l'ensemble de toutes les variables explicatives. Elle permet d'apprécier globalement le degré d'association entre toutes les variables explicatives et la capacité de stockage de l'entrepôt de données. Étant donné que les variables collectées sont toutes nominales, l'analyse en correspondances multiples (ACM) sera mise en contribution pour décrire les associations.

5.1.2.1 Approche unidimensionnelle

Avant d'analyser l'influence ou la liaison de chaque variable explicative sur la capacité de l'entrepôt de données tout en s'appuyant sur les réponses des enquêtés, il s'avère pertinent de faire un bref aperçu sur les réponses liées aux méthodes utilisées pour estimer cette capacité de stockage. Le graphique ci-après illustre les réponses obtenues selon les méthodes utilisées.

Figure 8 : Effectif des répondants selon la méthode d'estimation de la capacité de stockage d'entrepôt de données adoptée

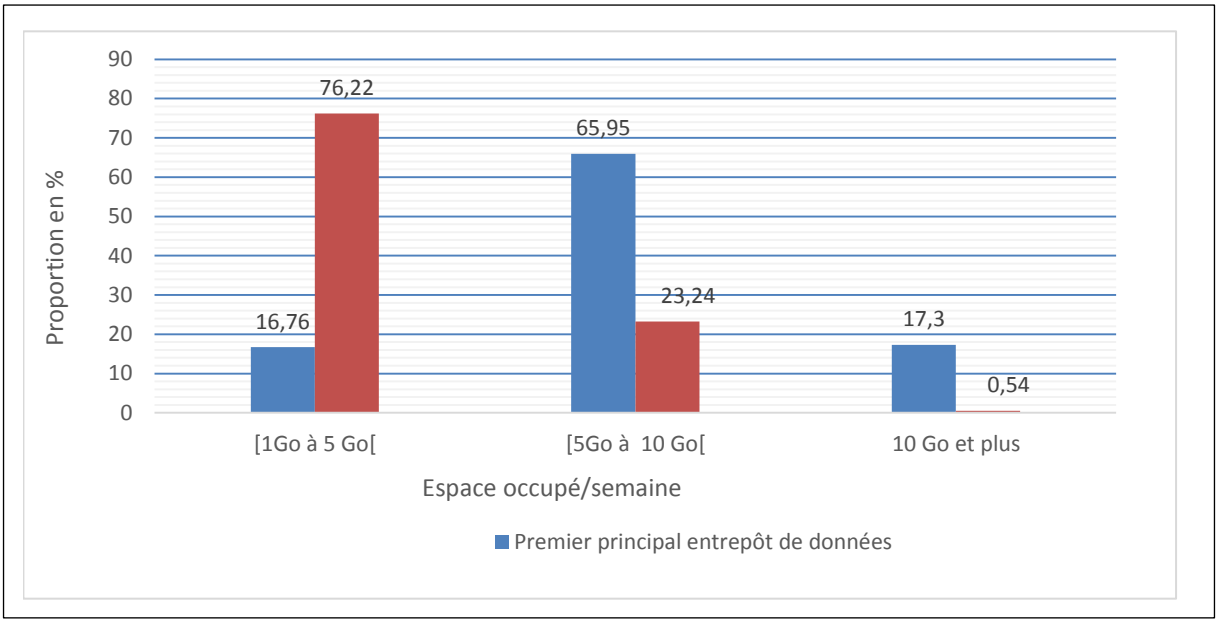


Source : Enquête effectuée auprès des entreprises

Selon ce graphique, plus de 92 % des enquêtés ont confirmé que leurs entreprises ou leurs départements procèdent de façon intuitive pour estimer la capacité de stockage de leurs entrepôts de données. Plus de 1 % seulement ont déclaré avoir utilisé la méthode formelle et moins de 7 % ont affirmé avoir adopté une méthode mixte.

Selon les réponses fournies par les enquêtés (question 5 du formulaire), la répartition de l'espace hebdomadaire occupé par les deux principaux entrepôts de données utilisés par leurs compagnies peut être illustrée par la figure suivante :

Figure 9 : Proportion des réponses en fonction d'espaces occupés par l'entrepôt de données



Source : Enquête effectuée auprès des entreprises

D'après ce graphique, 65,95 % des employés enquêtés ont souligné que leur premier principal entrepôt de données occupe un espace qui varie de 5 à moins de 10 giga-octets par semaine. Ce graphique met en évidence que 76,22 % ont considéré que leur deuxième entrepôt de données prend moins de 5 giga-octets par semaine.

Pour effectuer les analyses unidimensionnelle et multidimensionnelle, on tient compte uniquement du premier principal entrepôt de données.

Espaces occupés par l'entrepôt de données selon leurs usages (nécessités)

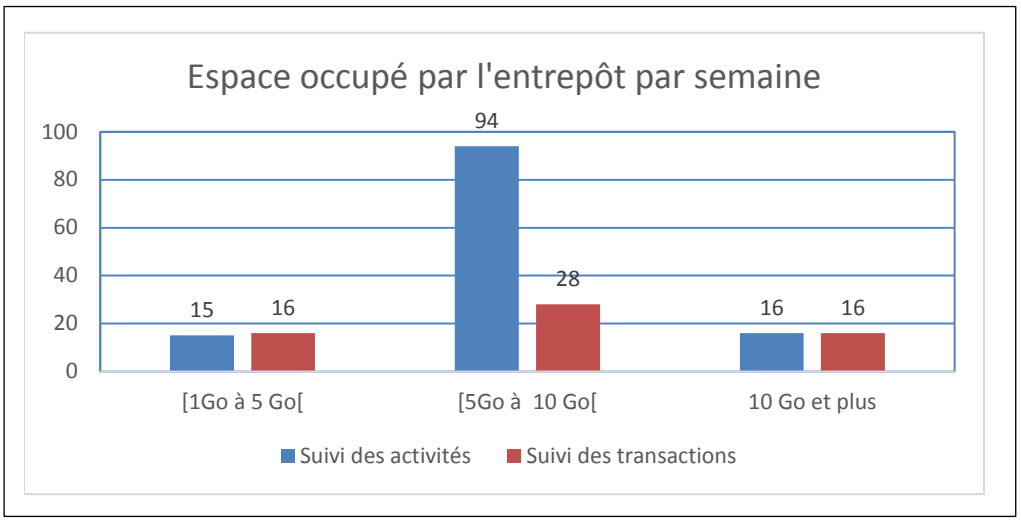
Le résultat de l'enquête par sondage distingue principalement deux catégories d'usage d'entrepôt de données telles :

i) Suivi de activités : Il s'agit d'un usage selon lequel on enregistre les indicateurs clefs de performance qui mettent en exergue les mesures telles que l'efficacité, l'efficience, la productivité et la vélocité.

ii) Suivi des transactions : Il s'agit d'un usage destiné à retracer les clients et les produits, à quantifier et à estimer les ventes et à évaluer les achats.

Le graphique suivant présente la répartition de l'espace occupé par l'entrepôt de données selon leurs usages.

Figure 10 : Espaces occupés par l'entrepôt de données selon l'usage.



Source : Enquête effectuée auprès des entreprises

On note une forte proportion de réponse des enquêtés (94) par rapport aux entrepôts destinés au suivi des activités. Ces entrepôts de données occupent un espace oscillant entre cinq et moins 10 giga-octets par semaine.

Le tableau ci-après précise les statistiques du test d'indépendance entre l'usage auquel l'entrepôt est destiné et son espace dans le système.

Tableau 10: Statistiques du test d'indépendance entre l'usage de l'entrepôt et l'espace occupé

Statistique	DDL	Valeur	Prob
Khi-2	2	14,7160	0,0006
Test du rapport de vraisemblance	2	14,3893	0,0008
V de Cramer		0,2820	

Source : Enquête effectuée auprès des entreprises (sortie du logiciel SAS)

Pour interpréter les informations fournies dans ce tableau, il faut fixer le seuil au-delà duquel on peut se tromper pour accepter ou rejeter le test d'indépendance. En général, ce seuil est de l'ordre de 0,05 soit 5 %. Il s'agit d'un seuil qui correspond à un niveau de confiance estimé à 95 %.

Si la valeur de P-value ou la probabilité (colonne Prob du tableau) correspondant à la statistique Khi-2 est inférieure 0,05, alors on peut affirmer qu'il existe un lien entre la variable explicative et l'espace occupé par l'entrepôt au risque de 5 % de se tromper. Dans le cas contraire, on peut accepter la situation d'indépendance.

Pour notre cas, on note la probabilité évaluée à 0.06 % qui est largement inférieure à 5 %. Donc, en se basant sur les réponses fournies par nos enquêtés, il y a une bonne raison de croire que l'usage auquel est destiné l'entrepôt influence significativement son espace dans le système.

Espace occupé par l'entrepôt en fonction de la taille des tables de faits

Les tables des faits sont les éléments fondamentaux dans l'analyse informationnelle car elles contiennent les mesures et indicateurs avec lesquelles les décideurs peuvent prendre les décisions stratégiques et pertinentes pour leurs compagnies. Le tableau suivant renseigne sur les tailles ou les poids des tables des faits sur l'espace occupé par l'entrepôt.

Tableau 11 : Répartition de l'espace occupé par l'entrepôt en fonction de la taille des tables de faits

	Espace Occupé par l'entrepôt par semaine			
Taille des tables des faits en %	[1Go à 5 Go [	[5Go à 10 Go [	10 Go et plus	Total
] 0-30]	8	17	5	30
] 30-60]	17	88	16	121
plus de 60%	6	17	11	34
Total	31	122	32	185

Source : Enquête effectuée auprès des entreprises (sortie du logiciel SAS)

On souligne plus de 65 % des enquêtés ont précisé que les tables des faits de leurs entrepôts contribuent environ à plus de 30 à 60 % de l'espace occupé.

Les statistiques du test d'indépendance entre les Taille des tables des faits et l'espace occupé par l'entrepôt sont indiquées dans le tableau suivant.

Tableau 12 : Statistiques du test d'indépendance entre les Taille des tables des faits et l'espace occupé par l'entrepôt

Statistique	DDL	Valeur	Prob
Khi-2	4	10,4735	0,0332
Test du rapport de vraisemblance	4	9,5327	0,0491
V de Cramer		0,1682	

Source : Enquête effectuée auprès des entreprises (sortie du logiciel SAS)

En faisant la même lecture effectuée précédemment, on peut affirmer que la taille des tables de faits influence significativement l'espace occupé par l'entrepôt de données à 95 % de niveau de confiance (Probabilité=0,0332<0,05).

Espace occupé par l'entrepôt en fonction de la taille des tables de dimensions

Les tables de dimensions contiennent les informations liées aux axes d'analyse informationnelle. En général, dans le cadre pratique, chaque table de dimensions contient peu de nombre d'enregistrements car les informations enregistrées sont presque statiques au cours des années (par exemple, les régions, les villes, les catégories de produits, le sexe).

Le tableau suivant montre la répartition de la taille des tables de dimensions par rapport à l'espace occupé par l'entrepôt de données.

Tableau 13 : Répartition de la taille des tables de dimensions par rapport à l'espace occupé

Taille des tables de dimensions en %	Espace Occupé par l'entrepôt par semaine			Total
	1Go à 5 Go	5Go à 10 Go	10 Go et plus	
0-30	25	116	26	167
30-60	6	6	6	18
Total	31	122	32	185

Source : Enquête effectuée auprès des entreprises

Selon ce tableau, on peut conclure que plus de 90 % des enquêtés ont dévoilé que les tables de dimensions contribuent à moins de 30 % de l'espace occupé par leurs entrepôts de données.

Les statistiques du test d'indépendance entre les taille des tables de dimension et l'espace occupé par l'entrepôt sont fournies par le tableau suivant.

Tableau 14 : Statistiques du test d'indépendance entre les Taille des tables de dimension et l'espace occupé par l'entrepôt

Statistique	DDL	Valeur	Prob
Khi-2	2	9,4503	0,0089
Test du rapport de vraisemblance	2	8,8740	0,0118
V de Cramer		0,2260	

Source : Enquête effectuée auprès des entreprises (sortie du logiciel SAS)

En faisant la même lecture que précédemment, on peut affirmer que la taille des tables de dimension influence l'espace occupé par l'entrepôt de données à 95 % de niveau de confiance (Probabilité=0.0089<0.05).

Espace occupé par l'entrepôt en fonction de la taille des index

Les index sont très importants pour accélérer la recherche des informations dans une table des bases de données relationnelles à travers les requêtes SQL. Le tableau suivant renseigne sur les réponses des enquêtés vis-à-vis du poids des index sur l'espace occupé par les entrepôts.

Tableau 15 : Répartition de la taille des index par rapport à l'espace occupé

Taille des index en %	Espace Occupé par l'entrepôt par semaine			Total
	1Go à 5 Go	5Go à 10 Go	10 Go et plus	
0-30	30	120	32	182
30-60	1	2	.	3
Total	31	122	32	185

Source : Enquête effectuée auprès des entreprises.

La lecture de ce tableau nous révèle que presque la totalité des enquêtés (98,38 %) ont considéré que le poids des index sur l'espace occupé par les entrepôts ne dépasse pas 30 %. Cette situation reflète les réalités dans la pratique des entreprises.

Les statistiques du test d'indépendance entre la taille des index et l'espace occupé par l'entrepôt sont présentées comme suit :

Tableau 16 : Statistiques du test d'indépendance entre les Taille des index et l'espace occupé par l'entrepôt

Statistique	DDL	Valeur	Prob
Khi-2	2	1,0278	0,5982
Test du rapport de vraisemblance	2	1,4357	0,4878
V de Cramer		0,0745	

Source : Enquête effectuée auprès des entreprises (sortie du logiciel SAS)

On note que la taille des index n'influence pas significativement l'espace occupé par l'entrepôt de données (Probabilité=0,5982>0,05). Cette situation peut être expliquée par le fait que, dans la pratique, la plupart les gestionnaires des bases de données ou les développeurs en intelligence d'affaires utilisent rarement les index particulièrement sur les tables des faits.

Espace occupé par l'entrepôt en fonction en fonction la durée de stockage des données

Le choix de la durée de rétention des données informationnelles, notamment dans la table des faits relève de la décision politique ou stratégique de l'entreprise. Le choix de la période pour emmagasiner les informations dépend de la disponibilité des ressources matérielles et financières. Plus l'entreprise veut garder d'avantage l'historique de données, plus elle est obligée de réserver plus d'espace nécessaire.

Le tableau suivant décrit les réponses des enquêtés liées à la durée de stockage.

Tableau 17: Répartition de l'espace occupé par l'entrepôt en fonction de la durée de stockage

Durée de stockage	Espace Occupé par l'entrepôt par semaine			Total
	1Go à 5 Go	5Go à 10 Go	10 Go et plus	
0-12 mois	7	24	11	42
12-24 mois	5	56	6	67
24-36 mois	6	28	6	40
Plus de 36 mois	13	14	9	36
Total	31	122	32	185

Source : Enquête effectuée auprès des entreprises

Ce tableau montre que plus de 36 % (soit 67 sur 185) des employés d'entreprises enquêtés ont affirmé avoir un entrepôt de données destiné à stocker des données pour une période allant de moins de 12 mois à 24 mois. Dans cet intervalle de temps, on retrouve la plus grande (83,58 % soit 56 sur 67) part des entrepôts qui occupent les espaces qui varient de cinq à moins de 10 giga-octets.

Les statistiques du test d'indépendance entre la durée de stockage et l'espace occupé par l'entrepôt sont indiquées dans le tableau suivant :

Tableau 18 : Statistiques du test d'indépendance entre la durée de stockage et l'espace occupé par l'entrepôt

Statistique	DDL	Valeur	Prob
Khi-2	6	25,2979	0,0003
Test du rapport de vraisemblance	6	24,8354	0,0004
V de Cramer		0,2615	

Source : Enquête effectuée auprès des entreprises (sortie du logiciel SAS)

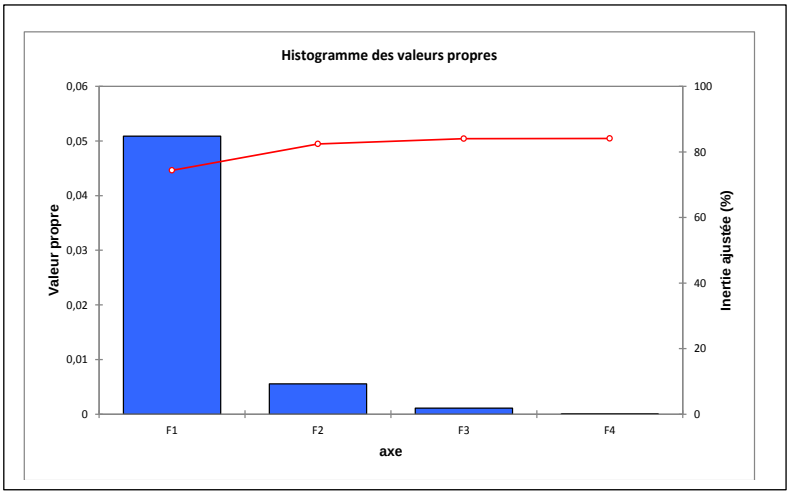
En adoptant la même démarche que précédemment, on peut conclure que la durée de stockage des informations influence significativement l'espace occupé par l'entrepôt de données à 95 % de niveau de confiance (Probabilité=0,0003<0,05).

5.1.2.1 Approche multidimensionnelle

L'approche multidimensionnelle permet d'apprécier l'interaction ou l'association entre les variables d'analyse. Étant donné que les variables utilisées dans le cadre de l'enquête par sondage sont du type nominal, la méthode appropriée pour détecter la présence des corrélations est celle de l'analyse des correspondances Multiples (ACM). L'analyse des correspondances multiples est une méthode descriptive permettant de résumer l'information contenu dans un grand nombre de variables en vue de faciliter l'interprétation des corrélations pouvant exister entre elles. En se basant sur les réponses des enquêtés, la principale question est celle de savoir quelles sont les modalités des variables corrélées à celles de l'espace occupé par l'entrepôt de données. Cette technique s'appuie sur des cartes de représentation ou cartes factorielles sur lesquelles on peut visuellement observer les associations ou les proximités entre les modalités des variables et les observations.

L'idée générale est la suivante : l'ensemble des individus enquêtés peut être représenté dans un espace à plusieurs dimensions sous la forme des nuages de points où chaque axe représente les différentes variables utilisées pour décrire chaque individu. Les caractéristiques de dispersion des individus sur les axes factoriels vis-à-vis des variables sont mesurées par ce que l'on appelle valeurs propres. La valeur propre (ou inertie) représente la proportion des quantités d'informations projetées sur un axe factoriel. Le graphique d'histogramme des valeurs propres présenté ci-après permet de choisir le nombre d'axes factoriels nécessaires pour présenter le maximum d'informations.

Figure 11 : Histogramme des valeurs propres



Source : Enquête effectuée auprès des entreprises (sortie du logiciel Excel Stat)

Ce graphique affiche que deux axes factoriels (F1 et F2) représentent plus de 80 % des informations projetées. Ces deux axes suffisent pour interpréter les corrélations entre la capacité (espace occupé par l'entrepôt de données) et les variables explicatives. Les associations entre l'espace occupé par l'entrepôt de données et les variables explicatives sont fournies par les tableaux ci-après.

Tableau 19 : Positions des variables par rapport au premier axe factoriel F1

Variable	Position des modalités des variables sur le premier axe factoriel (F1)	
	Côté positif	Côté négatif
Usage de l'entrepôt	Suivi des transactions	Suivi des activités
Espace occupé par l'entrepôt	10 Go et plus	5Go à 10 Go
Taille des tables de faits	Plus de 60 %	30-60
Taille des tables de dimensions	30-60	0-30
Durée de stockage des données	0-12 mois	12-24 mois

Source : Enquête effectuée auprès des entreprises (sortie du logiciel Excel Stat)

Tableau 20 : Positions des variables par rapport au deuxième axe factoriel F2

variable	Position des modalités des variables sur le deuxième axe factoriel	
	Côté positif	Côté négatif
Espace occupé par l'entrepôt		1Go à 5 Go
Taille des tables de faits		0-30
Durée de stockage des données		Plus de 36 mois

Source : Enquête effectuée auprès des entreprises (sortie du logiciel Excel Stat)

Selon le tableau 19, au regard du premier axe factoriel F1, deux associations entre les modalités des variables sont observées :

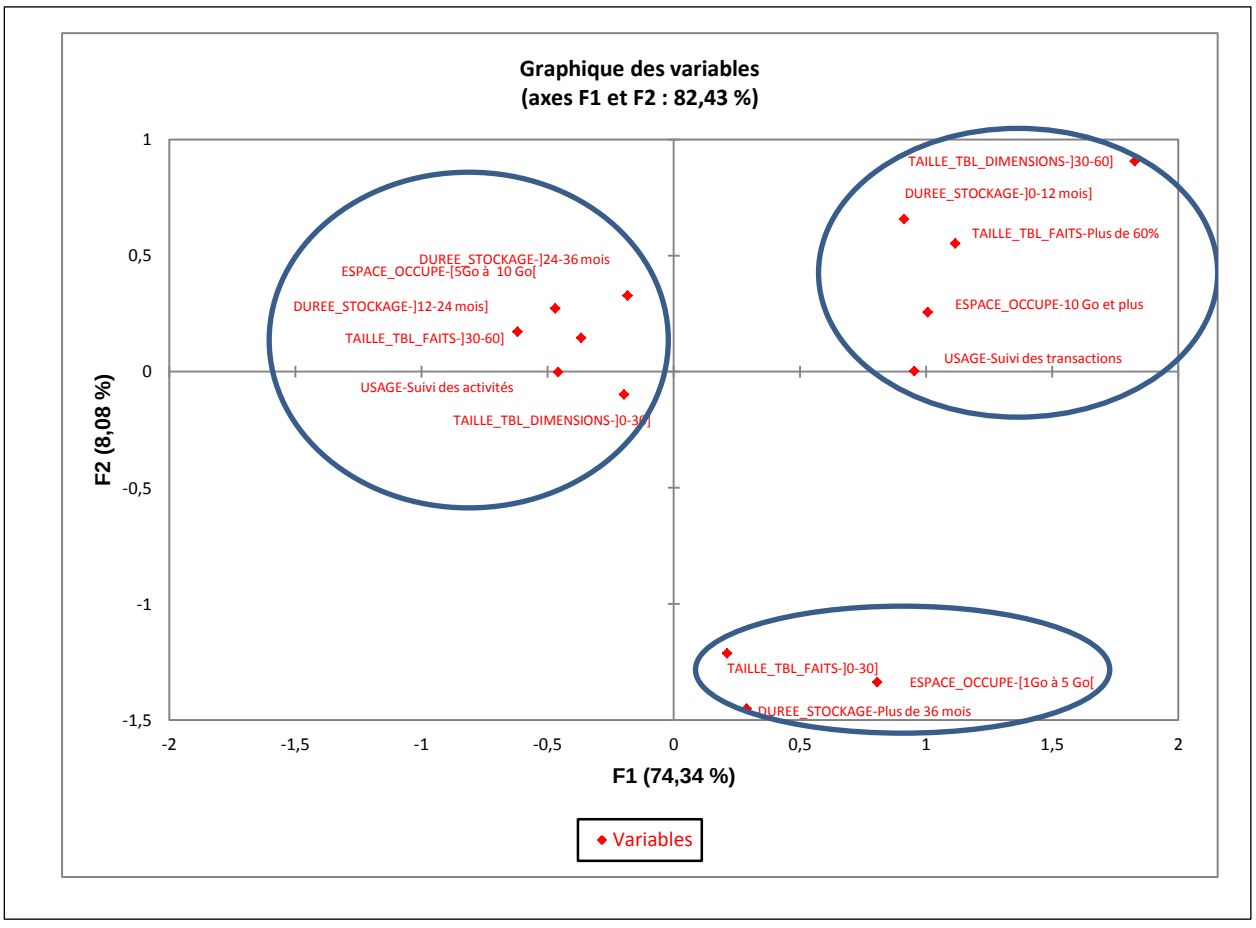
i) Du côté positif, on note la présence des entrepôts de données destinés au suivi des transactions. Ces entrepôts enregistrent des gros volumes de données de transaction plus de 10 GO par semaine dont la taille des tables de faits contribue à 60 % et plus et les poids des tables de dimension sont évalués entre 30 % et 60 %. La fréquence de rétention des données dans ces entrepôts dure moins d'un an.

ii) Du côté négatif, on remarque l'existence des entrepôts de données consacrés au suivi des activités. Leurs espaces dans le système varient de 5 à moins de 10 GO par semaine. Les tables de faits pèsent entre moins de 30 à 60 % et les tables de dimensions occupent moins de 30% de l'espace. La durée de rétention de données varie d'un à deux ans.

D’après le tableau 20, par rapport au deuxième axe factoriel F2, on observe du côté négatif les entrepôts de données occupant les espaces qui se situent entre un à moins de cinq GO. Les tables de faits pèsent moins de 30 % dans le système et la fréquence de rétention des informations peut atteindre plus de trois ans.

Les informations fournies par les tableaux 19 et 20 peuvent être résumées dans la carte factorielle présentée par le graphique suivant.

Figure 12 : Carte factorielle (F1, F2)



Source : Enquête effectuée auprès des entreprises (sortie du logiciel Excel Stat)

La carte factorielle montre de manière palpable les associations entre l’espace occupé par l’entrepôt de données dans le système et les variables explicatives.

En conclusion, les réponses fournies par les employés lors de l'enquête par sondage ont confirmé l'existence de corrélations entre la capacité d'entrepôts de données et les variables explicatives telles que :

i) Les variables liées à l'environnement du système : la taille des tables de faits et la taille des tables de dimension et
ii) La variable associée à la politique de gestion des données informationnelles : la durée de stockage des informations ou la fréquence de rétention des données.

La section suivante concerne la mise en œuvre du modèle d'estimation de la capacité de stockage d'entrepôts de données et l'analyse des résultats de prédiction.

5.2 Mise en œuvre des modèles d'estimation de la capacité de stockage d'entrepôts de données et l'analyse des résultats de prédiction

Dans cette section, il est question de décrire la source de données, d'effectuer les analyses descriptives afin de ressortir les caractéristiques de chaque variable et d'élaborer les modèles d'estimation de capacité de stockage d'entrepôt de données.

5.2.1 Source des données

Les données collectées pour l'élaboration des modèles d'estimation sont issues des bases de données de l'entreprise BELL Canada dans le Département intelligence d'affaires services extérieurs. Elle fait partie des entreprises intégrées dans l'échantillon pour l'enquête par sondage dont les résultats ont été analysés à la section 5.1. Les informations collectées auprès de cette entreprise répondent aux critères de sélection tels que : l'accessibilité des données, la disponibilité (les données sont repérées par des dates) et la qualité de données (absence des données manquantes). Le respect de ces critères est indispensable pour éviter les biais dans l'interprétation des résultats. Ces données sont principalement stockées dans les systèmes tels SQL SERVER, ORACLE et TERADATA.

Les variables utilisées sont celles qui ont été définies dans le Tableau1 de la section 4.2.2 du chapitre 4. Ces variables, principalement quantitatives, sont observées pour une période de 22 mois (janvier 2015 à 'octobre 2016) par rapport aux 10 magasins de données.

5.2.2 Analyses descriptives des données

L'analyse descriptive permet de mettre en évidence les caractéristiques des variables telles que le minimum, le maximum, la moyenne, la médiane, la variance et les quartiles. Pour dévoiler les informations qui se cachent derrière les données, certaines méthodes statistiques sont utilisées, à savoir : l'analyse des corrélations entre les variables suivies de l'analyse en composantes principales (ACP), et la classification ascendante hiérarchique (CAH).

5.2.2.1 Corrélations linéaires entres les variables

Avant d'entamer l'analyse de corrélation, le tableau ci-après donne l'aperçu global des caractéristiques de chaque variable d'analyse.

Tableau 21 : Caractéristiques des variables d'analyse

Variable	Minimum	Maximum	Moyenne	Écart-type
Capacite	320,463	569,001	418,103	83,473
TailleTblDim	0,031	2,761	0,982	0,675
TailleIndex	0,026	3,913	1,627	0,982
FreqRet	10	35	18,586	8,110
TailleTblFact	96,702	315,942	182,565	73,153

Source **:** Sortie du logiciel SAS

Capacite : Espace occupé par l'entrepôt de données
TailleTblFact : Taille des tables de faits
FreqRet : Durée de stockage des données
TailleTblDim: Taille des tables de dimension
TailleIndex : Taille des index

Pour tout entrepôt de données confondu, on note que la taille moyenne des tables des faits représente 43,67 % de la capacité de stockage pour une durée moyenne de rétention de données avoisinante d'un an et demi. Pour 22 mois d'observation, la durée maximale de rétention de données est 35 mois, soit deux ans et 11 mois. Les corrélations linéaires entre les variables sont présentées dans le tableau suivant.

Tableau 22 : Matrice de corrélations

Variables	Capacite	TailleTblDim	TailleIndex	FreqRet	TailleTblFact
Capacite	1	0,866(*)	-0,107	0,927(*)	0,981(*)
TailleTblDim	0,866(*)	1	-0,113	0,831(*)	0,854(*)
TailleIndex	-0,107	-0,113	1	-0,097	-0,103
FreqRet	0,927(*)	0,831(*)	-0,097	1	0,931(*)
TailleTblFact	0,981(*)	0,854(*)	-0,103	0,931(*)	1

Source : Sortie du logiciel SAS
Capacite : Espace occupé par l'entrepôt de données
TailleTblFact : Taille des tables de faits
FreqRet : Durée de stockage des données
TailleTblDim: Taille des tables de dimension
TailleIndex : Taille des index
Les valeurs marquées en (*) sont différentes de 0 à un niveau de signification alpha=0,05

Au regard de ce tableau, la taille des tables de faits, la durée de rétention et la taille des tables de dimension sont corrélées positivement avec l'espace occupé par l'entrepôt de données. En d'autres termes, l'augmentation de la taille des tables de faits, de la durée de stockage et de la taille des tables de dimension engendre significativement la hausse de l'espace occupé par l'entrepôt de données (variable Capacite). Il est à noter que la durée de rétention est corrélée positivement avec la taille des tables de faits. Cette situation paraît logique car plus la durée de rétention des informations dans les tables des faits s'étale, plus leur poids dans le système augmente.

uant à la taille des index dans les tables des faits, sa corrélation avec la variable cible (Capacite) n'est pas significative au seuil de 5 %. Cette situation pourrait signifier la quasi-absence des index dans les tables, en général dans les tables de faits afin d'accélérer la recherche des informations à travers les requêtes d'interrogation des données. La taille des tables de faits et celle des tables de dimension sont également corrélées positivement. Dans la pratique, l'influence de la taille des tables de dimension sur celle de la table de faits est plausible mais le sens contraire paraît absurde car les clefs primaires des tables de dimension sont des clés secondaires dans les tables de faits pour matérialiser les axes d'analyse. Donc, il est tout à fait logique que l'ajout d'autres axes d'analyse entraine la hausse du poids des tables de faits. Les corrélations entre les variables sont résumées dans

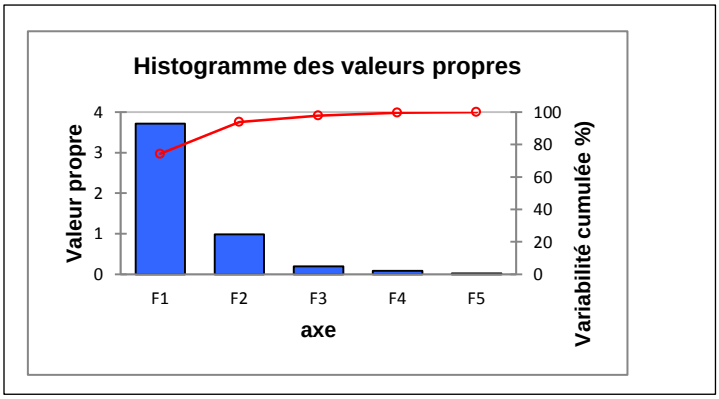
le graphique présenté à l'Annexe II. Ce graphique montre de façon apparente les corrélations entre les variables d'analyse.

Afin de vérifier l'existence d'une structure et d'identifier la présence des entrepôts de données qui se ressemblent entre eux, on applique la méthode d'analyse en composantes principales (ACP). L'ACP fait partie du groupe des méthodes descriptives multidimensionnelles. Elle permet d'étudier les structures de liaisons linéaires sur l'ensemble des variables, c'est-à-dire les corrélations entre les variables et de détecter les individus ayant les comportements similaires par rapport à ces variables. L'ACP s'appuie sur les axes factoriels ou axes de projection pour assurer une représentation fidèle (minimum de perte d'informations) à la structure des variables. Les logiciels statistiques comme ExcelStat fournissent plusieurs outils d'interprétation des résultats, mais on va se baser sur quelques-uns pour affiner notre analyse, à savoir :

i) L'histogramme des valeurs propres permettant de choisir le nombre optimal d'axes factoriels;
ii) Tableau d'association des variables avec les axes : il aide à identifier les variables qui sont corrélées entre elles;
iii) Cercle de corrélation : il s'agit d'un graphique montrant de façon visuelle les corrélations entre les variables et
iv) Graphiques de projections des individus et variables permettant de détecter les entrepôts ayant un comportement similaire vis-à-vis des variables.

À la suite du traitement de données avec le logiciel ExcelStat, on obtient les outils d'interprétation des résultats cités ci-haut.

Figure 13 : Histogramme des valeurs propres



Source : Sortie du logiciel ExcelStat

Ce graphique montre que les axes factoriels (F1 et f2) restituent presque de 94% des informations projetées. Donc, ces axes suffisent pour interpréter les corrélations entre les variables d'analyse.

Tableau 23 : Association des variables au regard des axes factoriels

Axe1 (F1) côté positif	Axe2 (F2) côté positif
Espace occupé par l'entrepôt (Capacite)	Taille des index(TailleIndex)
Taille des tables de faits (TailleTblFact)	
Durée de stockage des données (FreqRet)	
Taille des tables de dimension (TailleTblDim)	

Source : Sortie du logiciel ExcelStat

Ce tableau présente, au regard du premier axe factoriel (F1) l'association entre la variable dépendante (espace occupé par l'entrepôt de données) et les variables explicatives (taille des tables de faits, durée de stockage des données, taille des tables de dimension). La variable cible (Capacite) est corrélée positivement avec la taille des tables de faits, la durée de stockage des données et la taille des tables de dimension. Cette corrélation est nettement visible dans le graphique du cercle de corrélations présenté dans l'Annexe III

La séparation entre des groupes d'entrepôts de données présentée dans le graphique à l'Annexe IV n'est pas tellement apparente mais la présence de trois groupes distincts peut

être appréhendée au regard du premier axe factoriel (F1). Le premier groupe (situé à l'extrême gauche du premier axe factoriel) est formé par les entrepôts de données caractérisés par les faibles capacités de stockage, les durées de rétention de données assez courtes, les faibles tailles des tables de faits et des tables de dimension. Le second groupe (positionné relativement au milieu du premier axe factoriel) englobe les entrepôts de données identifiés par les valeurs moyennes de la capacité de stockage, des durées de rétention de données, des tailles des tables de faits et de dimension. Le troisième groupe fait référence aux entrepôts de données disposant des fortes valeurs de la capacité de stockage, des tailles des tables de faits et dimension, puis des durées de rétention des données relativement longues. Afin de mettre en exergue la frontière entre ces trois groupes, une analyse multidimensionnelle dénommée classification ascendante hiérarchique (CAH) sera abordée dans la sous-section suivante.

5.2.2.2 Classifications des entrepôts de données selon leur profil par la méthode de classification ascendante hiérarchique (CAH).

La classification ascendante hiérarchique est une méthode d'analyse statistique multidimensionnelle capable de classifier un ensemble d'individus. Elle permet de répartir en classes un ensemble d'entrepôts de données décrits par différentes variables ou caractéristiques. L'objectif est d'obtenir :

i) Des classes homogènes, c'est à dire les entrepôts d'une même classe partageant de nombreuses caractéristiques (les entrepôts qui se ressemblent vis-à-vis des variables d'analyse);
ii) Des classes séparées, c'est à dire les entrepôts de classes différentes qui ont peu de caractéristiques en commun.

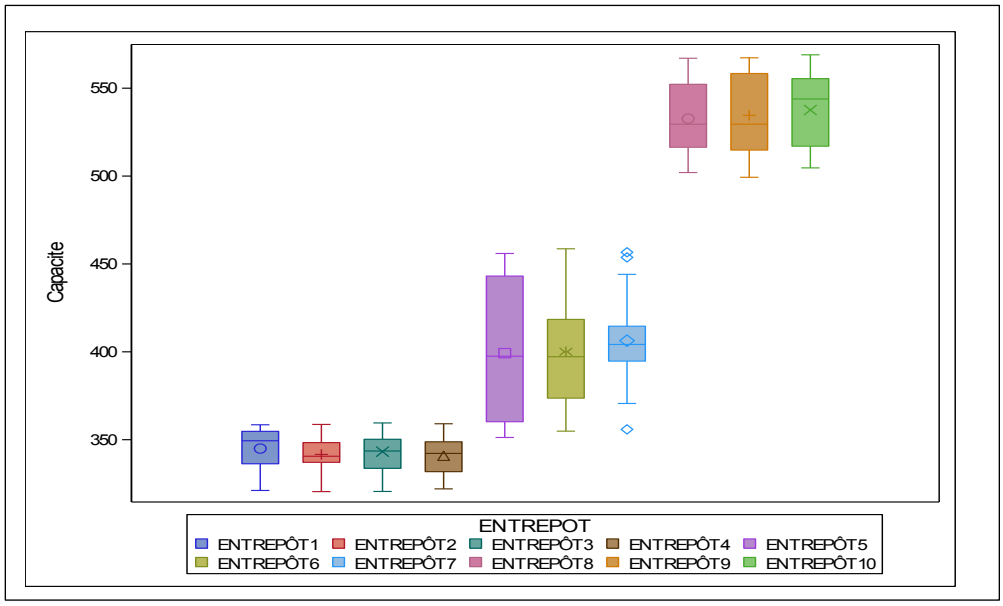
Les outils d'aide à l'interprétation utilisés dans cette méthode d'analyse sont multiples. Dans le cadre du présent travail, trois outils essentiels seront présentés:

i) L'histogramme ou diagramme des indices de niveaux permettant de trancher sur le choix du nombre optimal de partitions ou de classes d'entrepôts à tenir compte;
ii) Le dendrogramme indiquant l'arborescence des entrepôts formant les différentes classes;

iii) Le tableau dressant la liste des entrepôts avec leurs groupes d'appartenance.

Avant d'appliquer la CAH, il s'avère pertinent de présenter la capacité de stockage de chaque entrepôt de données à travers un graphique en boite à moustaches (box plot). La boite à moustaches est un moyen rapide permettant d'identifier le profil essentiel d'une série statistique quantitative. Elle résume à travers un diagramme quelques caractéristiques de position d'une variable étudiée telles : la médiane, les quartiles, le minimum, le maximum ou les déciles. Ce diagramme est utilisé principalement pour détecter la présence éventuelle des groupes d'individus par rapport à une même variable. Le graphique suivant montre la boite à moustaches de chaque entrepôt de données associée à la capacité de stockage.

Figure 14 : Boîtes à moustaches des capacités de stockage des entrepôts

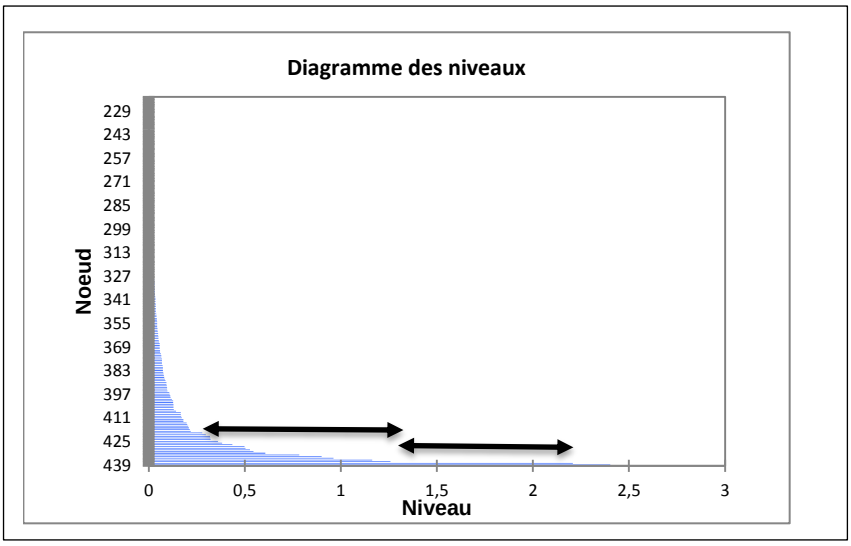


Source : Sortie du logiciel SAS

L'observation de ce graphique permet d'identifier l'existence assez remarquable de trois groupes d'entrepôts de données. Le premier groupe est formé par les entrepôts de données affichant des capacités de stockage variant de 320 à 360 GO, le second groupe concerne les entrepôts de données ayant des capacités de stockage oscillant autour de 320 à 360 GO et le troisième est caractérisé par les entrepôts occupant des espaces allant de près de 500 à 560 GO. Afin de confirmer l'existence de ces trois groupes, la classification ascendante hiérarchique est appliquée.

À la suite des traitements de données à travers le logiciel d'analyse ExcelStat, on obtient à la sortie : le diagramme des indices de niveaux présenté par la figure 15 et le dendrogramme montré par la figure 16

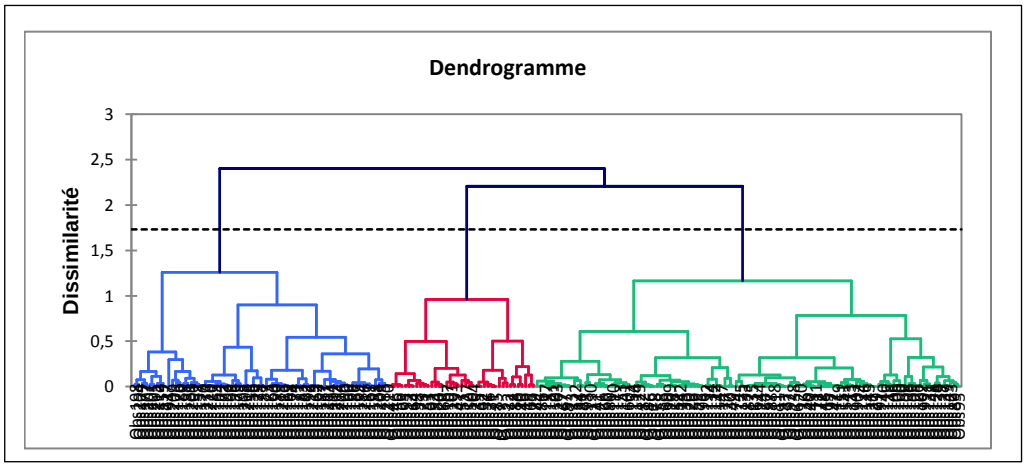
Figure 15 : Diagramme des indices de niveaux



Source : Sortie du logiciel Excel Stat

Ce diagramme montre l'apparition de deux importants sauts de niveau. Le nombre optimal de classes ou de groupes à considérer est trois. Le dendrogramme ci-dessous permet de confirmer ce choix.

Figure 16 : Dendrogramme



Source : Sortie du logiciel Excel Stat

La hauteur du dendrogramme traduit le niveau de dissemblance entre les entrepôts regroupés. Plus le palier est haut, moins les entrepôts réunis se ressemblent. La partition optimale consiste à mettre en évidence la présence de trois groupes ou de classes d'entrepôts. Ce qui justifie l'apparition du phénomène observé précédemment. Donc, nous avons une bonne raison de retenir le regroupement en 3 classes dans la suite de l'analyse.

En exploitant le tableau dressant la liste des entrepôts avec leurs groupes d'appartenance montré à l'Annexe V, les caractéristiques de chaque groupe d'entrepôts selon nos variables d'analyse peuvent être présentées comme suit :

Tableau 24 : Caractéristiques des groupes d'entrepôts selon les variables d'analyse

	VARIABLE	NOBS	MEAN	STD	MIN	MAX	Q1	MEDIAN	Q3
GROUPE1: -Entrepôt1 -Entrepôt2 -Entrepôt3 -Entrepôt4	Capacite	88	342,65	11,02	320,46	359,53	333,57	343,00	351,76
	TailleTblFact		119,14	10,60	96,70	140,10	109,70	119,18	128,00
	FreqRet		11,07	0,78	10	12	10	11	12
	TailleTblDim		0,40	0,17	0,03	0,82	0,27	0,40	0,53
GROUPE2: -Entrepôt5 -Entrepôt6 -Entrepôt7	Capacite	66	401,87	31,67	351,26	458,63	375,10	401,72	418,73
	TailleTblFact		162,70	27,40	112,40	220,14	141,54	158,88	181,72
	FreqRet		17,71	3,72	10,00	24,00	15,00	18,00	21,00
	TailleTblDim		0,94	0,43	0,25	1,74	0,54	1,02	1,24
GROUPE3 : -Entrepôt8 -Entrepôt9 -Entrepôt10	Capacite	66	534,94	21,79	499,31	569,00	516,38	532,92	555,43
	TailleTblFact		287,00	14,43	263,35	315,94	275,79	285,22	298,53
	FreqRet		29,48	3,01	24,00	35,00	27,00	29,50	32,00
	TailleTblDim		1,80	0,42	0,98	2,76	1,49	1,79	2,15

Source : Sortie du logiciel SAS

TailleTblFact : Taille des tables de faits
FreqRet : Durée de stockage des données
TailleTblDim : Taille des tables de dimension
TailleIndex : Taille des index

À la suite de l'analyse descriptive, il a été constaté une forte corrélation entre la variable cible (capacité de stockage) et les variables explicatives (la taille des tables des faits, la taille de la table de dimension et la durée de stockage des données). La méthode de classification a permis de confirmer la présence de trois groupes distincts d'entrepôts de données. Chaque groupe est constitué par les entrepôts caractérisés par un comportement homogène au regard des variables d'analyse. Les résultats de l'analyse descriptive faciliteront la mise en œuvre des modèles d'estimation de la capacité de stockage qui est développée dans la section suivante.

5.2.3 Analyse Prédictive : Mise en œuvre des modèles d'estimation de la capacité de stockage.

Cette section a pour objet d'élaborer les modèles d'estimation de la capacité d'entrepôts de données. Pour cela, deux modèles de prédiction sont abordés :

i) Modélisation à travers les facteurs discriminants : Il s'agit de l'estimation obtenue en appliquant la méthode d'analyse discriminante linéaire et celle de l'analyse discriminante bayésienne;
ii) Modélisation à travers la régression log-linéaire sur les données en panel.

5.2.3.1 Modélisation à travers les facteurs discriminants de la capacité de stockage

On distingue deux aspects en analyse discriminante :

i) Un aspect descriptif (géométrique) : En tenant compte des trois groupes d'entrepôts observés lors de la classification ascendante hiérarchique, on va rechercher les combinaisons linéaires des variables explicatives (la taille des tables des faits, la taille de la table de dimension et la durée de stockage des données) qui permettent de séparer le «mieux possible» les trois classes. Les variables discriminantes ainsi construites sont donc non corrélées entre elles. La première variable discriminante la plus significative peut être considérée comme un score attribué à une observation connaissant les valeurs prise par les variables explicatives.
ii) Un aspect aide à la décision (probabiliste) : une nouvelle observation se présente pour laquelle on connaît les valeurs de la taille des tables des faits, de la taille de la table de dimension et de la durée de rétention des données. Dans quelle classe a-t-elle le plus de chance d'appartenir? C'est l'analyse discriminante bayésienne qui permet de répondre à cette question.

Avant de calculer les valeurs des facteurs discriminants, il faut s'assurer que chaque variable explicative ait un pouvoir discriminant de manière globale et individuelle afin que le modèle soit interprétable. Les résultats des tests présentés à l'Annexe VI montrent que le rapport de corrélation associé à chaque variable explicative (R-carré) sont tous significatifs au seuil de de 5 %. Le test du pouvoir discriminant global confirme le même résultat. Donc, la taille des

tables des faits, la taille de la table de dimension et la durée de stockage des données ont un pouvoir discriminant significatif sur la capacité de stockage que ce soit de façon globale ou individuelle.

Après les traitements de données effectués avec le logiciel statistique SAS, on obtient le tableau qui fournit le coefficient de chaque variable explicative associé à chaque facteur discriminant.

Tableau 25 : Coefficients des variables explicatives associés aux facteurs discriminants.

Variable	Facteur discriminant	
	Can1	**Can2**
TailleTblFact	2,937142257	-3,046531961
FreqRet	1,350309226	2,145617581
TailleTblDim	0,319976806	1,082442297

Source : Sortie du logiciel SAS

TailleTblFact : Taille des tables de faits
FreqRet : Durée de stockage des données
TailleTblDim : Taille des tables de dimension

De ce tableau, on peut en déduire les équations qui déterminent les scores ou les valeurs des facteurs discriminants. Comme nous avons trois classes, il n'y aura que deux facteurs (Can1 et Can2).

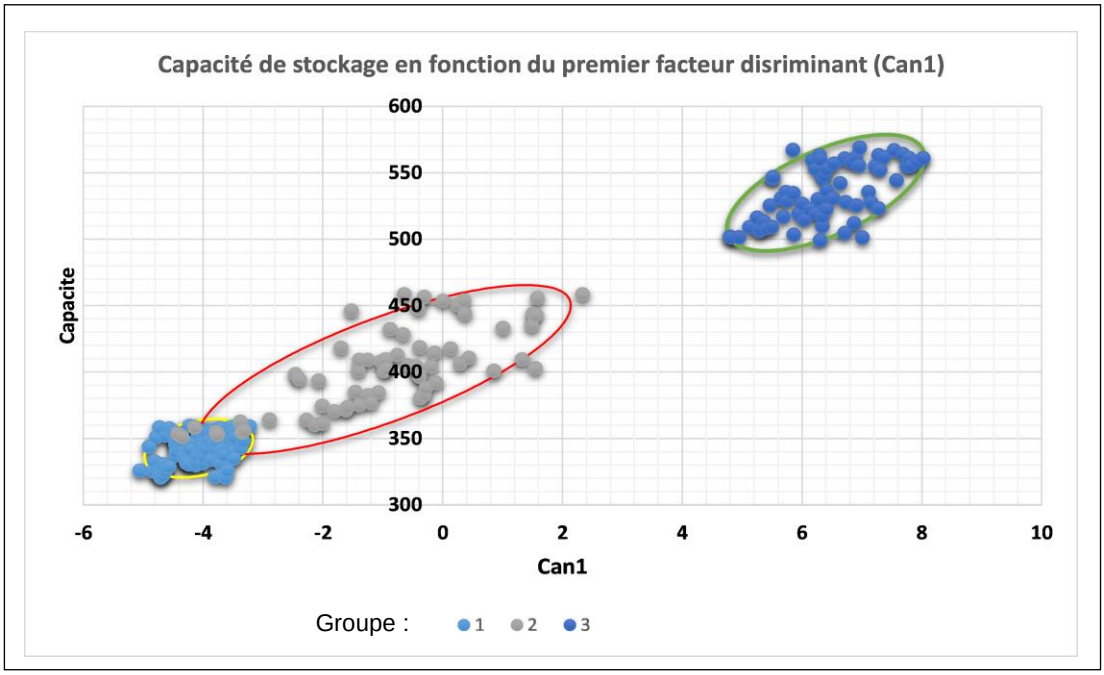
$$\begin{cases} \text{Can1} = 2{,}937 * \text{T}\mathbf{ailleTblFact} + 1{,}350 * \mathbf{FreqRet} + 0{,}320 * \mathbf{TailleTblDim} \\ \text{Can2} = -3{,}046 * \text{T}\mathbf{ailleTblFact} + 2{,}146 * \mathbf{FreqRet} + 1{,}082 * \mathbf{TailleTblDim} \end{cases}$$

À partir de ces deux équations, on obtient les valeurs de ces deux facteurs pour chaque entrepôt de données (voir l'Annexe VII). Il faut noter que les valeurs des variables explicatives utilisées dans les deux équations sont centrées par rapport à leur écart type et réduite par rapport à leur moyenne. Parmi les deux facteurs discriminants, lequel admet un pouvoir discriminant le plus élevé? La réponse à cette question est fournie par les tests développés à l'annexe VIII.

Ces tests montrent que le facteur discriminant Can1 affiche le carré d'une corrélation canonique proche de un (0,95). Donc, le pouvoir discriminant de Can1 est meilleur que celui de Can2. En effet, nous ne devons pas aller au-delà de Can1 dans notre analyse. C'est-à-dire, nous ne retiendrons donc qu'une variable discriminante (Can1) dans notre interprétation. En croisant les valeurs de la capacité de stockage avec celle du facteur

discriminante (ou score) retenu, on peut tracer le graphique permettant de distinguer de façon plus claire les trois classes d'entrepôts.

Figure 17 : Capacité de stockage en fonction de pointage



Source : Sortie du logiciel SAS

Ce graphique fait ressortir les caractéristiques de chaque groupe d'entrepôts vis-à-vis de leur score. Le premier groupe a un pointage qui varie de -5,2 à -3,5 (Capacité qui varie de 300 à 360 GO). Le second groupe a un score variant de -3,5 à 2,5 (Capacité oscillant autour de 360 à 460 GO). Le troisième fait référence au score allant de 4,79 à 8 (Capacité allant de 500 à 569 GO). Pour affecter de nouveaux individus dans des classes, il est préférable d'utiliser l'analyse discriminante bayésienne qui offre des résultats beaucoup plus précis [25].

À la suite des traitements de données effectués avec le logiciel statistique SAS, on obtient le tableau décrivant le coefficient de chaque variable explicative associé à chaque fonction linéaire discriminante liée au groupe d'appartenance correspondant.

Tableau 26 : Fonction linéaire discriminante par groupe

Fonction discriminante linéaire par GROUPE			
Variable	**1**	**2**	**3**
Constant	-24,68201	-49,92785	-150,66489
TailleTblFact	0,32663	0,41789	0,74186
FreqRet	0,99530	1,72685	2,77031
TailleTblDim	-1,41104	1,35926	3,74753

Source : Sortie du logiciel SAS

TailleTblFact : Taille des tables de faits

FreqRet : Durée de stockage des données

TailleTblDim: Taille des tables de dimension

Les fonctions linéaires discriminantes affichées dans ce tableau peuvent être présentées sous la forme suivante :

$$\begin{cases} f_1(X) = -24{,}682 + 0{,}327 * \mathbf{TailleTblFact} + 0{,}995 * \mathbf{FreqRet} - 1{,}411 * \mathbf{TailleTblDim} \\ f_2(X) = -49{,}928 + 0{,}418 * \mathbf{TailleTblFact} + 1{,}727 * \mathbf{FreqRet} + 1{,}359 * \mathbf{TailleTblDim} \\ f_3(X) = -150{,}665 + 0{,}742 * \mathbf{TailleTblFact} + 2{,}770 * \mathbf{FreqRet} + 3{,}747 * \mathbf{TailleTblDim} \end{cases}$$

Les probabilités d'appartenance à chaque groupe sont exprimées de la manière suivante :

$$\begin{cases} p_1 = \dfrac{\exp(f_1(X))}{\exp(f_1(X)) + \exp(f_2(X)) + \exp(f_3(X))} \\ p_2 = \dfrac{\exp(f_2(X))}{\exp(f_1(X)) + \exp(f_2(X)) + \exp(f_3(X))} \\ p_3 = \dfrac{\exp(f_3(X))}{\exp(f_1(X)) + \exp(f_2(X)) + \exp(f_3(X))} \end{cases}$$

Les probabilités d'appartenance de chaque observation obtenue à partir de ce modèle sont calculées dans le tableau affiché à l'Annexe IX. Connaissant la probabilité d'appartenance au groupe à partir des valeurs prises par les variables explicatives (la taille des tables des faits, la taille de la table de dimension et la durée de rétention des données), que peut-on faire pour prédire l'espace nécessaire? La réponse à cette question nous renvoie à choisir les meilleurs indicateurs de tendance de la capacité par groupe d'affectation présenté dans le tableau 20 (moyenne, médiane, minimum et maximum). En fonction de l'expérience du technicien responsable du système, il pourrait choisir la médiane de la capacité du groupe car c'est un indicateur qui n'est pas influencé par les valeurs extrêmes. Cependant, il est beaucoup plus prudent d'opter pour la valeur maximale de la capacité de stockage du groupe afin de ne pas subir les conséquences causées par le manque d'espace.

Validation de la force prédictive du modèle :

La force prédictive du modèle peut être validée par un certain nombre d'outils, tels que :

i) Le test d'égalité de matrices de variance et covariance entre les groupes : Si ce test n'est pas validé alors le modèle de prédiction par la méthode d'analyse discriminante bayésienne est biaisé, donc, non interprétable. Selon les résultats obtenus à l'Annexe X, le test est significatif à 1 % de risque de se tromper.

ii) La matrice de confusion : Il s'agit d'un tableau croisant les effectifs des classes obtenues par la méthode de classification ascendante hiérarchique aux effectifs des groupes prédits par le modèle de d'analyse discriminante bayésienne.

Tableau 27 : Matrice de confusion de classement

de \ Vers	1	2	3	Total	% correct
1	88	0	0	88	100,00%
2	10	56	0	66	84,85%
3	0	0	66	66	100,00%
Total	99	55	66	220	95,45%

Source : Sortie du logiciel SAS

Cette matrice de confusion de classement montre un taux d'erreur de prédiction moins de 5 %, ce qui correspond à un taux de bon classement évalué à 95,45 %. Donc, le modèle ainsi établit, a un pouvoir prédictif très élevé.

Simulation des quelques prévisions de capacité de stockage

Pour simuler quelques prévisions, on va se servir des données qui n'ont pas été utilisées pour construire le modèle. Elles sont issues des bases de données de l'entreprise BELL Canada, dans le Département intelligence d'affaires service extérieurs. Les résultats de prévision sont affichés dans le tableau suivant :

Tableau 28 : Résultats de prévision en utilisant la classification bayésienne

Date	Besoin en Capacité (GO)	Taille des tables de faits (GO)	Taille des tables de dimension (GO)	Durée de rétention des données	Estimation Probabilité d'appartenance			Estimation Groupe	Estimation espace de stockage
					Groupe1	Groupe2	Groupe3		
30-nov-16	343,12	119	0.4	11	0,9947	0,0053	0,0000	1	343,00
31-Dec-2016	401,02	159	1,02	18	0,0052	0,9948	0,0000	2	401,72
31-janv-17	532,95	285	2,00	30	0,0000	0,0000	1,0000	3	532,92

Source : Bases de données de l'entreprise BELL Canada, dans le Département intelligence d'affaires service extérieurs

Ce tableau présente les prévisions qui sont très proches des besoins en espaces.

5.2.3.2 Modélisation par la régression sur les données en panel

Dans cette section, il est question d'analyser la stationnarité des séries de données à travers le temps, d'élaborer les différents types de modèles de régression en panel tels que : Modèle à effets fixes, modèle à effets aléatoires, afin de présenter les résultats d'analyse et de valider les modèles ainsi établis. Les modèles d'estimation sont bâtis à partir des programmes (scripts) écrits en Python car ce langage possède de nombreuses bibliothèques spécialisées en analyses statistiques facilitant l'écriture des codes.

Analyse de stationnarité :

Tel qu'il est stipulé dans la sous-section 4.2.3 du chapitre 4, l'établissement d'un modèle de régression en panel exige le test de stationnarité des séries de données. La création d'un modèle statistique à partir données temporelles non stationnaires entraîne les biais au niveau des paramètres de l'estimation, l'absence de la force prédictive, les erreurs de prévisions, par conséquent, la non validité du modèle.

L'analyse des graphiques à l'annexe XI permet d'observer, de manière générale, qu'il n'y a pas de changement brusque de structure ou de changement brusque de valeurs de séries données. Donc, les données comportent des structures presque stables au cours des périodes d'observation, ce qui laisse présager la présence de la stationnarité au niveau de chaque variable. Pour confirmer cette situation, le test de stationnarité (ou test de racine

unitaire) est effectué au niveau de chaque variable en utilisant le test ADF (Augmented Dickey-Fuller). Le test ADF utilise un modèle autorégressif et optimise un critère d'information sur plusieurs valeurs de retard différentes [26]. L'hypothèse nulle (H0) du test suppose que la série temporelle présente une racine unitaire, c'est-à-dire qu'elle n'est pas stationnaire (présence de changement brusque de structure temporelle). L'hypothèse alternative (H1) confirme la stationnarité, ce qui correspond au rejet de H0. L'interprétation du résultat se fait à travers la valeur p-value du test. Une valeur de p-value inférieure à un seuil (5 % ou 1 %) confirme le rejet de l'hypothèse nulle (Acceptation de H1 : série stationnaire). Sinon, le test confirme la présence d'une série non stationnaire.

Les résultats du test de stationnarité sont présentés à l'Annexe XII. Les résultats du test affichent toutes les valeurs de p-value inférieures à 0,01. Donc, les séries de données (capacité de stockage, la taille des tables de faits, la taille des tables de dimensions et la fréquence de rétention de données) sont stationnaires au risque de 1 % de se tromper.

Modèles de régression de données en panel

Après avoir validé tous les tests de stationnarité, on peut procéder à l'élaboration des modèles de régression des données en panel à partir de l'équation log-linéaire formulée à la sous-section 4.2.3 : $\mathrm{Ln(C)} = \mathrm{a_{1t}Ln(X_{it})} + \cdots + \mathrm{a_{nt}Ln(X_{it})} + \varepsilon$

La composante résiduelle du modèle peut être décomposée comme suit [27] :

$$\varepsilon = u_i + v_t + w_{it}$$

Où u_i désigne un terme constant au cours du temps ne dépendant que de l'individu i (c'est-à-dire chaque magasin de données), v_t est un terme ne dépendant que de la période t (chaque temps d'observation), et w_{it} est un terme aléatoire croisé.

Selon la forme du résidu, on peut distinguer trois types de modèles :

i) Modèle à effets fixes : ce modèle, également appelé modèle de la covariance, suppose que u_i et v_t sont des effets constants, non aléatoires, qui viennent donc simplement modifier la valeur de la constante la de l'équation de régression selon les valeurs de i et de t.

ii) Modèle à effets aléatoires : ce modèle, encore appelé modèle à erreur composée, suppose que u_i et v_t sont véritablement aléatoires.

iii) Modèle à effets croisés : ce modèle tient compte de l'existence des effets temporels et des effets individuels matérialisés par la composante w_{it}.

Dans le cadre du présent essai, le modèle le modèle à effets fixes est mis à contribution car il permet de contrôler l'hétérogénéité non observée lorsque celle-ci est constante dans le temps et corrélée avec des variables indépendantes [28]. Donc, ce modèle permet de capter l'influence ou le poids de chaque magasin de données vis-à-vis de leur espace de stockage et l'influence du temps dans le modèle d'estimation.

Mise en œuvre du modèle à effets fixes :

Le modèle à effets fixes se présente sous la forme :

$$Ln(Capacite) = a_1 Ln(TailleTblFact_{it}) + a_2 Ln(TailleTblDim_{it}) + a_3 Ln(FreqRet_{it}) + u_i + v_t$$

Tels que $Capacite$ est la capacité de stockage à estimer, $TailleTblFact_{it}$ est la taille des tables de faits pour un entrepôt de données i à la date t, $FreqRet_{it}$ est la fréquence de rétention des données dans l'entrepôt i à la date t, u_i est l'effet apporté par l'entrepôt de données i qui est constant dans le temps, v_t est l'effet temporel qui ne varie pas selon l'entrepôt, et a_i $(i\ allant\ de\ 1 \text{ à } 3)$ sont des coefficients des variables explicatives.

Si $u_i = 0$ et $v_t \neq 0$, alors on est en présence d'une modèle à effet temporel. Par contre, si $u_i \neq 0$ et $v_t = 0$, alors le modèle spécifié capte l'effet apporté par chaque entrepôt de données indicé i. Après avoir exécuté les scripts écrits en Python présentés à l'Annexe XIII, on obtient dans le tableau de spécifications du modèle captant l'effet temporel suivant :

Tableau 29 : Spécifications du modèle à effet temporel

----------Summary of Regression Analysis----------

Formula: Y ~ <TailleTblFact> + <TailleTblDim> + <FreqRet>+ Intercept

Source	DDL	Somme des carrés	Carré moyen	Valeur F	p-value (Pr > F)
Model	3	7,78053619	2,59351206	1617,71	0,0001
Error	216	0,34629151	0,00160320		
Corrected Total	219	8,12682770			

R^2	Coef de Var	Racine MSE	Ln(Capacite) Moyenne
0,957389	0,665470	0,040040	6,016804

----------Summary of Estimated Coefficients----------

Variable	Coefficient	Std Err	t-stat	p-value	Interval de conf. à 95%	
Intercept	3,788390440	0,06872301	55,13	0,0001	3,652936870	3,923844009
TailleTblFact	0,403694187	0,01945682	20,75	0,0001	0,365344646	0,442043728
TailleTblDim	0,015367544	0,00561700	2,74	0,0067	0,004296391	0,026438697
FreqRet	0,057118435	0,01720564	3,32	0,0011	0,023205988	0,091030883

Source : Sortie obtenue à la suite de l'exécution du programme en Python à l'Annexe XIII

Ce tableau affiche que le modèle à effet temporel est globalement significatif au seuil de 5 % (p-value ou Pr>F=0,0001 inférieure 0,05). En d'autres termes, la taille des tables de faits, la taille des tables de dimensions et la durée ou la fréquence de rétention sont des facteurs explicatifs de l'espace de l'entrepôt de données. Ces variables expliquent près de 96 % ($R^2 = 0{,}957389$) de la variabilité de la capacité de stockage. Les 4 % restants sont proviennent de l'influence d'autres variables qui sont non observées dans le cadre du présent travail. Il faut noter que la taille des index (TailleIndex) n'est pas introduite dans le modèle car sa corrélation avec la capacité de stockage n'est pas significative (voir Tableau 22).

La lecture du Tableau 29 permet de déduire que tous les coefficients associés à chaque variable explicative sont significatifs avec un niveau de confiance évalué à 95 % (p-value inférieure à 0,05). La taille des tables de faits, la taille des tables de dimensions et la durée de rétention sont statistiquement suffisantes pour estimer de l'espace de l'entrepôt de données. Par conséquent, les valeurs des coefficients des variables explicatives sont interprétables et leurs significations sont définies telles qu'elles sont mentionnées à la sous-section 4.2.5 du chapitre 4 :

i) L'augmentation de 1 % de la taille de la table de faits entraîne une augmentation de plus de 0,4 % ($a_1 = 0{,}403694187$) d'espace. Soit, une hausse de 10 % la taille

de la table de faits occasionne une hausse de plus de 4 % de la capacité de stockage;

ii) L'accroissement de 1 % de la taille de la table de dimension engendre une augmentation de plus de 0,01 % ($a_2 = 0{,}015367544$) de l'espace de stockage. Autrement dit, une hausse de 10 % de la taille de la table de dimensions correspond à une hausse de 1,5 % de l'espace de stockage;

iii) La hausse de 1% de la durée de rétention provoque un accroissement près de 0,05 % de l'espace de stockage. Soit l'accroissement de 10 % de la durée de rétention des données correspond à une hausse de 5 % de l'espace de stockage;

iv) La constante ($Intercept = v_t = 3{,}788390440$) correspond à l'effet dû à la variation du temps (effet temporel) qui est indépendante des variables explicatives.

Les résultats obtenus à l'issue de l'établissement du modèle à effet temporel coïncident avec ce que l'on a attendu et corroborent de façon générale les réponses obtenues auprès des techniciens et experts lors de l'enquête par sondage. Avant de procéder à une simulation de prévisions avec ce modèle, il est pertinent valider son pouvoir prédictif.

Validation de la force prédictive du modèle à effets temporels

Les valeurs des capacités de stockage estimées à travers le modèle ainsi que les erreurs de prédiction associées sont présentées dans l'Annexe XIV. Ces valeurs permettent de construire les courbes (voir l'Annexe XV) permettant d'apprécier le pouvoir prédictif du modèle à effet temporel. La lecture de ces courbes fait ressortir que les valeurs estimées à partir du modèle sont très proches des besoins en espace de stockage, comparées à celles qui ont été calculées par la méthode intuitive. Le tableau suivant montre les répartitions des valeurs des erreurs d'estimation pour la méthode formelle d'estimation et pour la méthode intuitive.

Tableau 30 : Erreurs d'estimation

Erreur d'estimation pour le modèle à effet temporel en %				Erreur d'estimation pour la méthode intuitive en %		
Niveau	**Valeur**	**Intervalle de confiance 95 %**		**Valeur**	**Intervalle de confiance 95 %**	
100 Max 100 %	11,1752005			27,0660		
99 %	11,0320133	9,6126542	11,175200	26,9478	26,8651	27,0660
95 %	7,0098281	6,1643189	9,612654	26,7168	26,6087	26,8651
90 %	5,8335461	5,3510718	6,898795	26,5527	26,4245	26,6774
75 % Q3	3,9765670	3,6963903	4,741222	26.0607	25,7521	26,1982
50 % Médiane	2,3921238	2,0660570	2,851324	25,0892	24,8468	25,2996
25 % Q1	1,2742556	0,9814843	1,512175	24,0704	23,8479	24,2474
10 %	0,4399286	0,2356174	0,644924	23,5819	23,2286	23,6649
5 %	0,2182753	0,0906770	0,347386	23,2199	23,0508	23,3456
1 %	0,0859381	0,0247932	0,090677	23,0264	22,8780	23,0508
0 % Min	0,0247932			22,8780		
Moyenne	2,91852	2,61806	3,21899	25,0424	24,8935	25,1913

Source : Annexe XIII

Ce tableau présente le pourcentage d'erreurs absolues d'estimation d'espace par rapport aux besoins selon les deux méthodes (modèle à effet temporel et la méthode intuitive). La méthode formelle (le modèle à effet temporel) affiche une erreur moyenne d'estimation près de 2,92 %. Pour la méthode d'estimation par intuition, la moyenne atteint plus de 25 %. Pour le modèle à effet temporel, on note que 50 % (c'est-à-dire la médiane) des erreurs absolues d'estimation observées ont des valeurs comprises entre 0,02 % et 2,39 %. Par contre, pour la méthode d'estimation intuitive, 50 % des erreurs associées se situent entre 22,90 % et 25,09 %. Donc, modèle formel offre une estimation plus précise (8 fois plus en moyenne) que la méthode intuitive.

Simulation des quelques prévisions d'espace de stockage.

Pour simuler quelques prévisions, on utilise les mêmes données que le Tableau 28.

Tableau 31 : Résultats de prévision en utilisant le modèle à effet temporel

Date	Besoin en Capacité(GO)	Taille des tables de faits (GO)	Taille des tables de dimension (GO)	Durée de rétention des données	Estimation espace de stockage
30-nov-16	343,12	119	0.4	11	345,222
31-Dec-2016	401,02	159	1,02	18	395,652
31-janv-17	532,95	285	2,00	30	535,884

Source : Bases de données de l'entreprise BELL Canada, dans le Département intelligence d'affaires service extérieurs

Ce tableau affiche les prévisions qui restent très proches des besoins en espaces.

À la suite de l'exécution du programme à l'Annexe XVI, les spécifications du modèle à effets individuels c'est-à-dire le modèle qui tient compte de l'effet de chaque entrepôt de données, sont présentées comme suit :

Tableau 32 : Spécifications du modèle à effets individuels

----------Summary of Regression Analysis----------

Formula: Y ~ <TailleTblFact> + <TailleTblDim> + <FreqRet> + <FE_ENTREPOT1>
+ <FE_ENTREPOT2> + <FE_ENTREPOT3> + <FE_ENTREPOT4>
+ <FE_ENTREPOT5> + <FE_ENTREPOT6> + <FE_ENTREPOT7> + <FE_ENTREPOT8>
+ <FE_ENTREPOT9>+ <ENTREPOT10>

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	12	7,84033289	0,65336107	472,07	<.0001
Error	207	0,28649481	0,00138403		
Corrected Total	219	8,12682770			

R^2	Coef de Var	Racine MSE	Ln(Capacite) Moyenne
0.964747	0.618312	0.037203	6.016804

----------Summary of Estimated Coefficients----------

Paramètre	Valeur estimée	Erreur type	Valeur du test t	p-value	Intervalle de conf. à 95%	
TailleTblFact	0,3104	0,0247	12,58	0,0000	0,2621	0,3588
TailleTblDim	0,0148	0,0055	2,71	0.0074	0,0041	0,0254
FreqRet	0,0112	0,0194	0,58	0,5655	-0,0269	0,0492
FE_ENTREPOT1	4,3512	0,1153	37,74	0,0000	4,1253	4,5772
FE_ENTREPOT2	4,3279	0,1157	37,40	0,0000	4,1011	4,5547
FE_ENTREPOT3	0,3104	0,0247	12,58	0,0000	0,2621	0,3588
FE_ENTREPOT4	0,0148	0,0055	2,71	0,0074	0,0041	0,0254
FE_ENTREPOT5	0,0112	0,0194	0,58	0,5655	-0,0269	0,0492
FE_ENTREPOT6	4,3512	0,1153	37,74	0,0000	4,1253	4,5772
FE_ENTREPOT7	0,3104	0,0247	12,58	0,0000	0,2621	0,3588
FE_ENTREPOT8	0,0148	0,0055	2,71	0,0074	0,0041	0,0254
FE_ENTREPOT9	4,4772	0,1352	33,12	0,0000	4,2122	4,7422
FE_ENTREPOT10	4,4839	0,1348	33,26	0,0000	4,2196	4,7481

Source : Sortie obtenue à la suite de l'exécution du programme en Python à l'Annexe XVI

Ce tableau montre que le modèle à effet individuel est globalement significatif au seuil de 5 % (p-value ou Pr>F=0,0001 inférieure 0,05). La taille des tables de faits, la taille des tables de

dimensions et la durée ou la fréquence de rétention sont des déterminants de l'espace de stockage de l'entrepôt de données. On constate également la significativité (p-value<0.05) de l'effet apporté par chaque entrepôt de données ($FE\_ENTREPOT_i$ (i $variant$ de 1 à 10). On observe l'existence des groupes d'entrepôts admettant des effets semblables, tels que :

- Groupe 1 : entrepôts 1, 2, 3, 4;
- Groupe 2 : entrepôts 5, 6, 7 et
- Groupe 3 : entrepôts 8, 9 ,10.

Cette situation démontre la validité des résultats de regroupement des entrepôts de données obtenus lors de l'analyse à travers la Classification Ascendante Hiérarchique à la sous-section 5.2.2.2.

5.3 Conclusions et recommandations

Le chapitre 5 a permis de mettre en exergue les résultats des analyses qui ont été développées dans le présent travail. L'enquête par sondage a permis de recueillir les opinions des experts vis-à-vis de l'espace de l'entrepôt de données. Leurs réponses à cette enquête stipulent que les variables liées à l'environnement du système telles que la taille des tables de faits et la taille des tables de dimension, puis la variable associée à la politique de gestion des données informationnelles comme la durée de stockage des informations ou la fréquence de rétention des données sont des facteurs qui influencent la capacité de l'entrepôt de données. L'analyse descriptive effectuée sur l'échantillon de données issue de la base de données de l'entreprise BELL Canada montre l'existence d'une forte corrélation entre la variable cible (capacité de stockage) et les variables explicatives (la taille des tables des faits, la taille de la table de dimension et la durée de stockage des données). Les résultats de l'analyse descriptive des données viennent corroborer les opinions des techniciens experts lors de l'enquête par sondage.

Les résultats de l'analyse descriptive ont facilité la formalisation et la mise en œuvre des modèles d'estimation de la capacité de stockage. Deux types de modèles formels de prédiction ont été développés : modélisation à travers les facteurs discriminants de la capacité de stockage et la modélisation par la régression sur les données en panel. La

modélisation à travers les facteurs discriminants permet de regrouper les entrepôts de données et de prédire la probabilité du groupe d'appartenance d'un nouvel entrepôt de données selon les variables d'analyse. La Modélisation par la régression sur les données en panel permet de capter l'influence et le poids de chaque entrepôt de données vis-à-vis de leur espace de stockage et l'influence du temps. Ce modèle offre non seulement la possibilité d'effectuer la prévision de l'espace dans le temps mais il permet également de retrouver la structure de groupes d'entrepôts définie dans la modélisation par les facteurs discriminants. La comparaison des résultats issus de ces deux modèles d'estimation à ceux obtenus par la méthode intuitive fait ressortir que le modèle formel offre en moyenne une estimation huit fois plus précise (valeur plus proche du besoin en espace de stockage) que la méthode intuitive. L'hypothèse de l'étude annoncée au chapitre 3 est confirmée, c'est-à-dire la connaissance de la capacité de stockage de l'entrepôt de données à partir des facteurs liés à la politique de gestion des informations décisionnelles et à l'environnement du système à travers la méthode formelle améliore la précision de l'estimation des besoins réels en espace de stockage. D'où l'importance capitale d'opter pour la méthode formelle pour la mise des projets d'entreprise de grande envergure.

Malgré la meilleure précision d'estimation offerte par le modèle formel, celui-ci nécessite un processus de mise à jour afin de conserver sa force prédictive. En guise d'exemple, la variable taille des index sur les tables de faits n'a pas été introduite dans le modèle car sa corrélation avec la capacité de stockage n'est statistiquement pas significative. Selon l'évolution de la structure des données d'analyse dans le temps, la prise en compte de cette variable ou d'autres pourrait s'avérer importante car leur poids ou leur corrélation par rapport à l'espace de stockage pourrait être significatif. La mise à jour et le processus de production des résultats fournis par le modèle doivent être automatisés. L'automatisation pourrait se faire en utilisant par exemple les techniques utilisées dans le domaine de l'apprentissage machine (Machine Learning). L'apprentissage machine, regroupent les technologies telles que l'apprentissage en profondeur, les réseaux neuronaux et le Data Mining. Ces technologies peuvent aussi englober des systèmes plus avancés qui sont en mesure de comprendre, d'apprendre, de s'adapter, de prédire les besoins en espace de stockage, et potentiellement de fonctionner de manière autonome. Ces systèmes peuvent apprendre et modifier le comportement futur du phénomène étudié, ce qui donne lieu à la création de systèmes et de programmes plus intelligents.

Conclusion

Le présent essai a eu pour objectif de répondre à la question de recherche : «L'estimation de la capacité de stockage de l'entrepôt de données en fonction des éléments de la politique de gestion des informations décisionnelles et de l'environnement du système par la méthode formelle permet-elle d'améliorer la précision de l'estimation de l'espace requis?»

Pour répondre à la question de l'étude, une revue de littérature permettant la compréhension et l'inventaire de ce qui a été publié a été procédée. Ce volet a permis la découverte des éléments liés au sujet, notamment l'architecture et l'infrastructure d'entrepôt de données, les différentes méthodes d'estimation de l'espace et les facteurs qui peuvent l'influencer. Deux méthodes d'estimation de l'espace d'entrepôt de données ont été stipulées, à savoir : la méthode basée sur l'intuition et la méthode formelle qui utilise les modèles d'équations mathématiques et statistiques.

Afin de mieux organiser le cheminement de la pensée lors de l'élaboration du présent travail, quelques approches méthodologiques ont été adoptées. Celles-ci a permis de préciser la stratégie de recherche et la procédure d'analyses de données. Pour cela, une enquête par sondage sur les déterminants de l'espace d'entrepôt de données a été menée. Cette enquête visait deux principaux objectifs : comprendre les pratiques des entreprises en matière de la gestion de de l'espace, identifier à priori les variables qui l'influencent selon les expériences sur le terrain afin d'éviter l'introduction des biais dans les modèles d'estimation. Cette enquête ne consistait pas à collecter les capacités de stockage des entrepôts de données de toutes les entreprises visées par l'échantillonnage. Elle se focalisait plutôt sur l'identification des facteurs explicatifs de l'espace selon les expériences pratiques. L'enquête a permis non seulement de recueillir les opinions des 185 enquêtés tirés des 200 techniciens issus de la base de sondage selon la méthode aléatoire simple, mais également de détecter les éléments à tenir compte dans la mise en œuvre des modèles d'estimation. Afin de bâtir ces modèles, les données issues des bases de données de l'entreprise BELL Canada (Département intelligence d'affaires services extérieurs) ont été choisies. Elles concernent les observations pour une période de 22 mois (janvier 2015 à octobre 2016) par rapport aux 10

magasins de données. Le choix des données destinées à l'expérimentation a été guidé par les critères de sélection pertinents tels que : l'accessibilité des données, la disponibilité (les données sont repérées par des dates précises), la qualité de données (absence des données manquantes). Le respect de ces critères est nécessaire et indispensable en vue d'éviter les biais dans l'interprétation des résultats.

Deux méthodes d'analyse de données ont été principalement abordées dans le présent travail : l'analyse statistique descriptive ou exploratoire et l'analyse prédictive. L'analyse descriptive unidimensionnelle a rendu possible le résumé et la synthèse des informations contenues dans la base de données. L'analyse descriptive multidimensionnelle, quant à elle, permettait de tenir compte de toutes les variables dans son ensemble en vue de dégager les corrélations entre elles et de définir les groupes d'entrepôts de données ayant un comportement similaire. L'analyse prédictive a permis d'établir les modèles d'estimation de l'espace à travers les traitements des informations provenant de l'historique de données afin de prédire les tendances futures et les motifs de comportement. Le présent essai a mis l'accent sur deux types de modélisation : la modélisation à travers les facteurs discriminants et la modélisation par la régression log-linéaire sur les données en panel. La modélisation à travers les facteurs discriminants a permis de séparer le mieux possible les groupes de magasins de données homogènes et de prédire le groupe d'appartenance d'un nouvel entrepôt de données dont ses caractéristiques n'ont pas été introduites lors de l'établissement du modèle. La modélisation par la régression log-linéaire sur les données en panel a offert non seulement la possibilité d'effectuer la prévision de l'espace dans le temps mais aussi elle nous a aidé à retrouver la structure des groupes d'entrepôts de données définies dans la modélisation par les facteurs discriminants.

Après avoir dépouillé les réponses à l'enquête par sondage sur les déterminants de la capacité de stockage, il a été constaté que les variables liées à l'environnement du système telles que la taille des tables de faits et la taille des tables de dimension, puis la variable associée à la politique de gestion des données informationnelles comme la durée de stockage des informations ou la fréquence de rétention des données sont considérées comme des facteurs qui influencent la capacité de stockage de l'entrepôt de données. Les analyses statistiques effectuées sur l'échantillon de données issue de la base de données de de l'expérimentation viennent confirmer l'existence d'une forte corrélation entre la variable

cible (capacité de stockage) et les variables explicatives (la taille des tables des faits, la taille de la table de dimension et la durée de rétention des données). À la suite de l'établissement des modèles d'estimation à travers ces facteurs, il a été démontré que la méthode formelle affiche une erreur moyenne d'estimation près de 2,92 %. Tandis que pour la méthode d'estimation par intuition, cette erreur dépasse 25 % en moyenne. Le modèle formel offre une estimation huit fois plus précise que la méthode intuitive. Cette situation a permis de valider l'hypothèse de l'étude qui est celle de savoir : «la connaissance de la capacité de stockage de l'entrepôt de données à partir des facteurs liés à la politique de gestion des informations décisionnelles et à l'environnement du système à travers la méthode formelle améliore la précision de l'estimation des besoins réels en espace de stockage». L'utilisation de la méthode formelle d'estimation d'espace de stockage d'entrepôt de données pour la mise des projets d'entreprise de grande envergure est recommandable. Et ce, en vue d'obtenir plus de précisions, afin d'assurer par la suite la gestion rationnelle des ressources matérielles et financières affectées.

Malgré la meilleure précision d'estimation offerte par le modèle formel, celui-ci exige la mise en place d'un processus de mise à jour pour pouvoir conserver sa force prédictive. Selon l'évolution la structure des données d'analyse dans le temps, la prise en compte d'une autre variable explicative pourrait s'avérer importante car son influence sur l'espace de stockage pourrait être significative. La mise à jour du modèle et le processus de production des résultats d'estimation doivent être automatisés dans le but de mettre en place un système intelligent. L'automatisation pourrait se faire en exploitant les techniques utilisées dans le domaine de l'apprentissage machine (Machine Learning), une discipline qui a montré ses preuves dans le domaine de la science des données. La réalisation de ce système intelligent peut être considérée comme un des prolongements pertinents du présent essai.

Liste des références

[1] Cécile F., Fadila B., Omar B. (2012). « Les entrepôts de données pour les nuls . . . ou pas !». Laboratoire ERIC - Lyon 2, Université de Lyon, p.7

[2] Elisabeth METAIS, «SYSTÈMES INFORMATIQUES - Systèmes d'aide à la décision», Encyclopædia Universalis [en ligne], consulté le 14 janvier 2017. URL: http://www.universalis.fr/encyclopedie/systemes-informatiques-systemes-d-aide-a-la-decision/

[3] Bill Immon. Building the Data Warehouse. John Wiley and Son 1996

[4] Agarwal Bhushan B., Tayal Prakash S., Data Mining and Data Warehousing, Laxmi Publications 2009, Chapitre 10, p.155

[5] Biere Mike, Business Intelligence for the Enterprise, IBM Press 2003, p.171

[6] Chaudhuri et Dayal, An Overview of Data Warehousing and OLAP Technology 1997, p. 2.

[7] Adamson Christopher, Star Schema: The Complete Reference, Graw-Hill/Osborne 2010, chapter 1 p.110

[8] Williams Steve, Business Intelligence Strategy and Big Data Analytics: A General Management Perspective, Morgan Kaufmann Publishers 2016, p.121

[9] Silvers Fon, Building and Maintaining a Data Warehouse, Auerbach Publications 2008, p.211

[10] Kantardzic Mehmed, «Data Mining: Concepts, Models, Methods, and Algorithms», John Wiley & Sons 2003, p.214

[11] PONNIAH PAULRAJ, Data warehousing fundamentals for IT Professionals, Second Edition, p.164.

[12] PONNIAH PAULRAJ, Data warehousing fundamentals for IT Professionals, Second Edition, p.166.

[13] Nabli, A., A. Soussi, J. Feki, H. Ben-Abdallah, et F. Gargouri (2005). Towards an Automatic Data Mart Design. In VIIth International Conference on Enterprise Information Systems (ICEIS 05), Miami, Florida, USA, pp. 226–231.

[14] Cécile F., Fadila B., Omar B. (2007). Évolution de modèle dans les entrepôts de données : existant et perspectives. Laboratoire ERIC - Lyon 2, Université de Lyon

[15] Jacques Ledent. Une analyse log-linéaire des courants migratoires interprovinciaux : Canada, 1961-1983. Cahiers québécois de démographie, Volume 12, numéro 2, octobre 1983, p. 224

[16] Jean-Bernard CHATELAIN, Kirsten RALF, Les liaisons fallacieuses : quasi-colinéarité et « suppresseur classique », Documents de Travail du Centre d'Economie de la Sorbonne, nov. 2012.

[17] DUGUET Emmanuel, Économétrie des panels avec applications, Mars 2010, p.7

[18] Saporta Gilbert, Probabilités Analyse de Données et Statistique, Technip, 201, P.468

[19] Ahn, S.C., et Schmidt, P., (1995), Efficient Estimation of Models for Dynamic Panel Data, Journal of Econometrics, 68, p.27.

[20] Gaulier G, Hurlin C et Jean–Pierre P. (1999), Testing Convergence : A Panel Data Approach, Annales d'Economie et de Statistiques, p.56.

[21] Levin, A., et Lin, C-F., (1992), Unit Root Test in Panel Data: Asymptotic and Finite-Sample Properties", Discussion Paper 92-23, Department of Economics, University of California, San Diego.

[22] C. Girard, T.B.M.J. Ouarda et B. Bobée. Étude du biais dans le modèle log-linéaire d'estimation régionale. Presses scientifiques du CNRC, avril 2004

[23] Saporta Gilbert. (2002) Data fusion and data grafting. Computational Statistics and Data Analysis, p.465-473

[24] Rousseau Sylvie, Saporta G. (Décembre 2011) Non-réponse et données manquantes, p.2

[25] T. W. Anderson: An Introduction to Multivariate Statistical Analysis, Wiley-Interscience, 3rd edition, 2003 p.137.

[26] Elliot Graham Elliot, Timmermann A.: Handbook of Economic Forecasting, Vol 2A Elsevier Science and Technology Books, 2013 p.525

[27] HSIAO CHENG, Analysis of Panel Data, Cambridge University Press, second edition, 2003, p.30

[28] Badi H. Baltagi, A Companion to Econometric Analysis of Panel Data, Wiley, 2009, p.145

Bibliographie

Adamson Christopher, Star Schema: The Complete Reference, McGraw-Hill/Osborne, McGraw-Hill/Osborne, 2010

Alan Julian Izenman: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, 2008.

David J. Hand, Heikki Mannila, and Padhraic Smyth: Principles of Data Mining, MIT Press, 2001.

Hughes Ralph, Agile Data Warehousing for the Enterprise: A Guide for Solution Architects and Project Leaders, Morgan Kaufmann Publishers, 2016

Ian H. Witten and Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2nd edition, 2005.

Kanti V. Mardia, J. T. Kent, and J. M. Bibby: Multivariate Analysis, Academic Press, 1980.

Ludovic Lebart, Alain Morineau, and Marie Piron : Statistique Exploratoire Multidimensionnelle : Visualisations et Inférences en Fouille de Données, Dunod, 4th edition, 2006

Olivia Parr Rud: Data Mining Cookbook, Wiley, 2000.

Robert Johansson: Numerical Python: A Practical Techniques Approach for Industry, Apress, 2015

Ron Cody: Learning SAS by Example: A Programmer's Guide, SAS Publishing, 2007.

Annexe I : Questionnaire d'enquête par sondage

UNIVERSITÉ DE SHERBROOKE

Estimation de la Capacité de Stockage de l'entrepôt de données

CONTEXTE ET OBJECTIFS DU SONDAGE

Cette enquête par sondage est réalisée dans le cadre l'obtention du grade de maître en Technologies de l'Information (Maîtrise en Génie Logiciel à l'Université de Sherbrooke). Elle a pour objet d'identifier les déterminants de la capacité de stockage d'un entrepôt de données selon les expériences pratiques de l'entreprise.
Dans cette enquête, l'anonymat du répondant ou de l'entreprise est scrupuleusement gardé et les informations liées à l'identification de l'entreprise seront exploitées de manière confidentielle.

Le questionnaire comprend 2 sections :
1- Identification de l'entreprise ou du répondant
2- Facteurs explicatifs de la capacité de stockage de l'entrepôt de données

Note: Les questions marquées (*) sont obligatoires

SECTION 1: IDENTIFICATION DE L'ENTREPRISE

1. Nom du répondant

* 2. Rôle ou responsabilité du répondant

* 3. Nom de l'entreprise

* 4. Secteur d'activité

1

UNIVERSITÉ DE SHERBROOKE

Estimation de la Capacité de Stockage de l'entrepôt de données

SECTION 2: LES DÉTERMINANTS DE LA CAPACITÉ DE STOCKAGE D'UN ENTREPÔT DE DONNÉES

*** 5. Quelle méthode utilisez-vous afin d'estimer l'espace requis pour le stockage des informations dans un entrepôt de données?**

- Intuitive (sans utilisation des équations mathématiques ou statistiques. Exemple: estimation selon juste l'impression du technicien)
- Formelle (utilisation des équations mathématiques ou statistiques à travers l'historique des données. Exemple: équation de régression linéaire)
- Mixte (formelle et intuitive)

*** 6. Pour deux (2) importants entrepôts de données (2 entrepôts de données qui occupent le plus d'espace en Gigaoctets), en moyenne, combien de quantités de données en (gigaoctets) par semaine y sont stockées?**

	[1Go à 5 Go[/Semaine	[5Go à 10 Go[/Semaine	10 Go et plus /Semaine
Entrepôts de données 1			
Entrepôts de données 2			

*** 7. Parmi les éléments cités ci-dessous, lesquels occupent le plus d'espace de stockage pour le premier important entrepôt de données mentionnés à la question 5 (Choisissez la tranche du taux d'occupation en % proposée dans la liste)**

	]0-30]	]30-60]	plus de 60%
Tables de faits (Facts tables)			
Tables de dimensions (Dimensions tables)			
Taille des Index dans la table de faits (Index size of facts tables)			

*** 8. Pendant quelle période (en nombre de mois) les historiques des données sont-ils stockés pour les 2 principaux entrepôts de données mentionnés à la question 5**

	]0-12 mois]	]12-24 mois]	]24-36 mois]	Plus de 36 mois
Entrepôt de données 1				
Entrepôt de données 2				

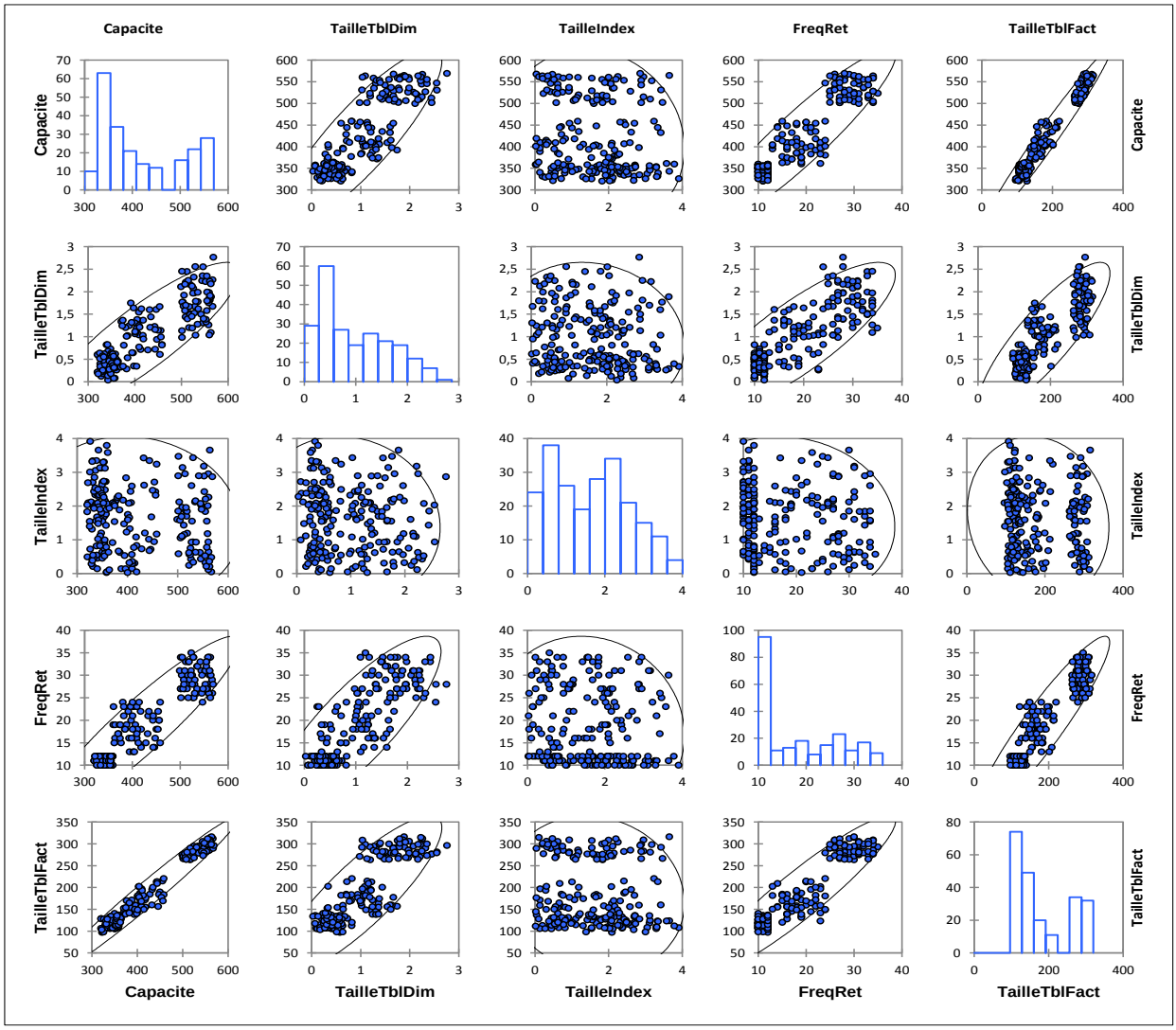
2

* 9. Selon les pratiques de votre entreprise ou selon les vos expériences professionnelles, y a t-il d'autres facteurs qui peuvent occuper de façon considérable l'espace de stockage de vos entrepôts de données? (si OUI préciser au maximum 3 facteurs à l'étape 10)

- OUI
- NON

10. Veuillez préciser ici les autres facteurs mentionnés à l'étape 9

Annexe II : Matrice de corrélations entre les variables



Source : Bases de données de l'entreprise BELL Canada, BELL Canada dans le Département Intelligence d'Affaires Service Extérieurs

Capacite : Espace occupé par l'entrepôt de données

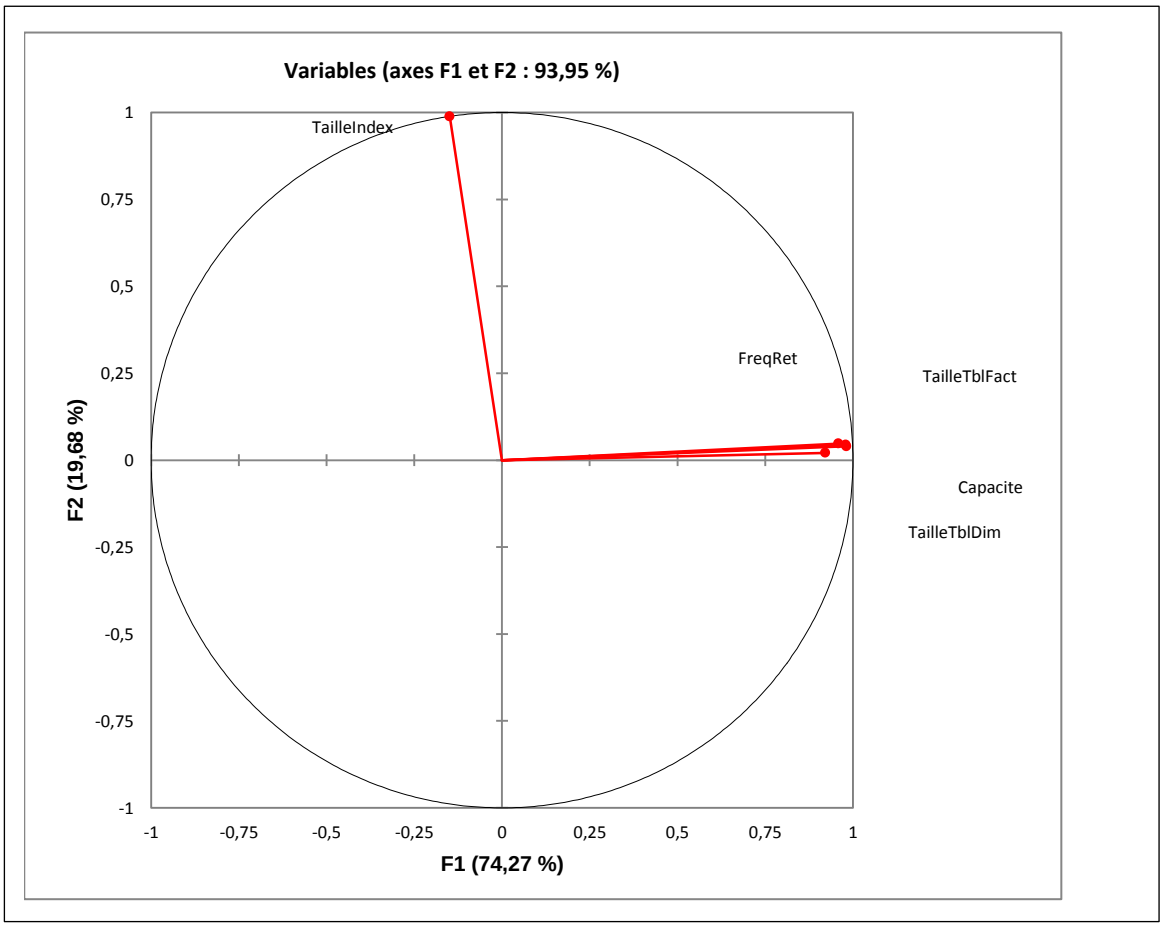
TailleTblFact : Taille des tables de faits

FreqRet : Durée de stockage des données

TailleTblDim: Taille des tables de dimension

TailleIndex : Taille des index

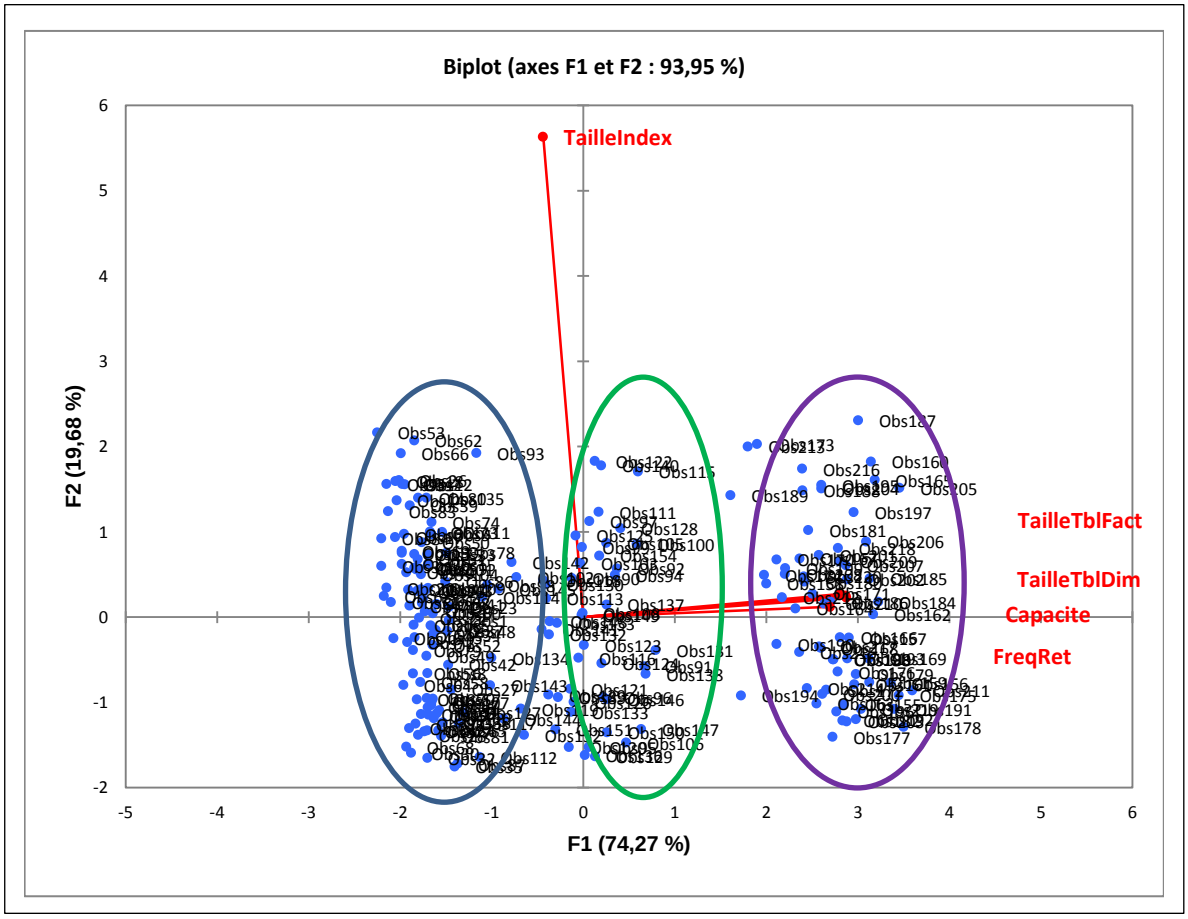
Annexe III : Cercle de corrélations entre les variables



Source : Sortie du logiciel ExcelStat

Capacite : Espace occupé par l'entrepôt de données
TailleTblFact : Taille des tables de faits
FreqRet : Durée de stockage des données
TailleTblDim: Taille des tables de dimension
TailleIndex : Taille des index

Annexe IV : Graphique de projection des observations et des variables



Source : Sortie du logiciel ExcelStat

Capacite : Espace occupé par l'entrepôt de données
TailleTblFact : Taille des tables de faits
FreqRet : Durée de stockage des données
TailleTblDim: Taille des tables de dimension
TailleIndex : Taille des index

Annexe V : Tableau de données avec le résultat de la classification (CAH)

OBS	Entrepôt	GROUPE	Capacite	TailleTblDim	FreqRet	TailleTblFact
1	ENTREPÔT1	1	355,28	0,57	12	127,90
2	ENTREPÔT1	1	336,30	0,38	12	117,70
3	ENTREPÔT1	1	343,85	0,09	10	106,59
4	ENTREPÔT1	1	355,39	0,42	12	117,28
5	ENTREPÔT1	1	352,73	0,53	10	123,46
6	ENTREPÔT1	1	357,28	0,08	10	135,76
7	ENTREPÔT1	1	353,97	0,31	11	109,73
8	ENTREPÔT1	1	346,23	0,34	11	128,10
9	ENTREPÔT1	1	350,81	0,49	11	133,31
10	ENTREPÔT1	1	358,49	0,37	10	107,55
11	ENTREPÔT1	1	354,75	0,61	11	134,80
12	ENTREPÔT1	1	352,24	0,40	11	105,67
13	ENTREPÔT1	1	351,28	0,26	10	115,92
14	ENTREPÔT1	1	347,99	0,26	10	118,32
15	ENTREPÔT1	1	339,82	0,50	11	132,53
16	ENTREPÔT1	1	340,00	0,25	12	119,00
17	ENTREPÔT1	1	331,57	0,32	10	119,37
18	ENTREPÔT1	1	331,47	0,28	11	122,64
19	ENTREPÔT1	1	322,34	0,58	12	96,70
20	ENTREPÔT1	1	328,28	0,12	12	105,05
21	ENTREPÔT1	1	321,11	0,21	11	128,45
22	ENTREPÔT1	1	357,69	0,44	10	110,88
23	ENTREPÔT2	1	348,57	0,73	11	128,97
24	ENTREPÔT2	1	340,79	0,29	11	129,50
25	ENTREPÔT2	1	337,70	0,46	11	118,20
26	ENTREPÔT2	1	338,89	0,27	11	122,00
27	ENTREPÔT2	1	354,17	0,58	11	131,04
28	ENTREPÔT2	1	330,44	0,38	11	115,65
29	ENTREPÔT2	1	320,46	0,60	11	128,19
30	ENTREPÔT2	1	331,86	0,17	12	112,83
31	ENTREPÔT2	1	330,64	0,36	10	109,11
32	ENTREPÔT2	1	340,38	0,42	12	105,52
33	ENTREPÔT2	1	355,95	0,44	11	124,58
34	ENTREPÔT2	1	337,63	0,49	12	124,92
35	ENTREPÔT2	1	358,65	0,61	12	111,18
36	ENTREPÔT2	1	326,19	0,40	12	127,21
37	ENTREPÔT2	1	347,88	0,47	11	132,19
38	ENTREPÔT2	1	339,57	0,82	10	112,06
39	ENTREPÔT2	1	337,11	0,56	11	107,88
40	ENTREPÔT2	1	343,01	0,03	12	133,77
41	ENTREPÔT2	1	358,69	0,36	12	129,13
42	ENTREPÔT2	1	348,34	0,57	11	135,85
43	ENTREPÔT2	1	347,18	0,24	12	135,40
44	ENTREPÔT2	1	341,60	0,49	10	126,39
45	ENTREPÔT3	1	350,84	0,46	10	105,25
46	ENTREPÔT3	1	325,64	0,24	11	107,46
47	ENTREPÔT3	1	342,74	0,58	11	109,68
48	ENTREPÔT3	1	342,95	0,67	11	133,75
49	ENTREPÔT3	1	329,84	0,56	12	112,15
50	ENTREPÔT3	1	349,72	0,63	10	111,91
51	ENTREPÔT3	1	332,77	0,29	10	106,49
52	ENTREPÔT3	1	357,80	0,30	12	125,23
53	ENTREPÔT3	1	326,20	0,34	10	107,64
54	ENTREPÔT3	1	350,25	0,24	10	140,10
55	ENTREPÔT3	1	359,53	0,49	10	118,64
56	ENTREPÔT3	1	346,34	0,17	12	114,29
57	ENTREPÔT3	1	338,73	0,45	11	128,72
58	ENTREPÔT3	1	342,00	0,43	10	119,70
59	ENTREPÔT3	1	344,17	0,45	11	120,46
60	ENTREPÔT3	1	348,61	0,59	10	125,50
61	ENTREPÔT3	1	333,73	0,46	11	126,82
62	ENTREPÔT3	1	359,52	0,40	12	111,45
63	ENTREPÔT3	1	320,55	0,21	11	105,78
64	ENTREPÔT3	1	355,12	0,07	10	117,19
65	ENTREPÔT3	1	350,18	0,54	12	105,05
66	ENTREPÔT3	1	342,99	0,29	11	123,47
67	ENTREPÔT4	1	354,45	0,26	11	131,15
68	ENTREPÔT4	1	328,82	0,21	12	105,22
69	ENTREPÔT4	1	336,23	0,38	10	117,68
70	ENTREPÔT4	1	354,90	0,26	11	124,22
71	ENTREPÔT4	1	345,66	0,28	12	107,15
72	ENTREPÔT4	1	342,58	0,13	10	133,61
73	ENTREPÔT4	1	353,36	0,61	11	116,61
74	ENTREPÔT4	1	334,90	0,62	12	127,26
75	ENTREPÔT4	1	347,47	0,34	11	121,62
76	ENTREPÔT4	1	348,82	0,18	12	108,13
77	ENTREPÔT4	1	333,42	0,47	11	113,36
78	ENTREPÔT4	1	341,81	0,82	12	119,63
79	ENTREPÔT4	1	325,26	0,63	11	97,58
80	ENTREPÔT4	1	345,76	0,46	11	124,47
81	ENTREPÔT4	1	331,03	0,57	12	122,48
82	ENTREPÔT4	1	331,81	0,25	12	116,13
83	ENTREPÔT4	1	325,73	0,51	10	97,72
84	ENTREPÔT4	1	322,01	0,17	12	103,04
85	ENTREPÔT4	1	345,66	0,27	12	107,16
86	ENTREPÔT4	1	355,75	0,53	12	131,63
87	ENTREPÔT4	1	359,07	0,41	12	136,45

OBS	Entrepôt	GROUPE	Capacite	TailleTblDim	FreqRet	TailleTblFact
88	ENTREPÔT4	1	333,92	0,55	10	126,89
89	ENTREPÔT5	2	428,39	0,70	15	184,21
90	ENTREPÔT5	2	407,78	1,36	16	163,11
91	ENTREPÔT5	2	443,11	1,22	20	212,69
92	ENTREPÔT5	2	391,91	1,74	22	156,76
93	ENTREPÔT5	2	361,16	0,25	23	122,79
94	ENTREPÔT5	2	403,08	1,08	23	201,54
95	ENTREPÔT5	2	388,18	1,24	19	170,80
96	ENTREPÔT5	2	446,44	1,60	14	156,25
97	ENTREPÔT5	2	454,14	0,60	22	181,66
98	ENTREPÔT5	2	360,23	0,50	19	133,28
99	ENTREPÔT5	2	406,56	1,07	19	187,02
100	ENTREPÔT5	2	455,96	1,13	20	214,30
101	ENTREPÔT5	2	351,26	0,41	11	112,40
102	ENTREPÔT5	2	376,81	0,81	15	169,57
103	ENTREPÔT5	2	380,36	1,20	24	148,34
104	ENTREPÔT5	2	353,55	0,54	10	113,13
105	ENTREPÔT5	2	450,40	1,14	18	189,17
106	ENTREPÔT5	2	409,70	1,03	23	196,65
107	ENTREPÔT5	2	358,54	0,31	12	114,73
108	ENTREPÔT5	2	448,00	0,94	16	183,68
109	ENTREPÔT5	2	353,02	0,36	12	123,56
110	ENTREPÔT5	2	354,77	0,31	11	127,72
111	ENTREPÔT6	2	397,89	1,63	20	159,16
112	ENTREPÔT6	2	364,38	0,45	16	127,53
113	ENTREPÔT6	2	384,81	1,02	18	157,77
114	ENTREPÔT6	2	364,17	0,40	19	131,10
115	ENTREPÔT6	2	458,63	0,82	24	220,14
116	ENTREPÔT6	2	458,63	0,81	15	183,45
117	ENTREPÔT6	2	354,86	0,57	12	131,30
118	ENTREPÔT6	2	371,39	1,33	17	144,84
119	ENTREPÔT6	2	385,12	0,53	18	154,05
120	ENTREPÔT6	2	396,53	0,64	23	158,61
121	ENTREPÔT6	2	418,44	1,36	14	154,82
122	ENTREPÔT6	2	432,30	1,60	17	159,95
123	ENTREPÔT6	2	405,21	1,38	16	174,24
124	ENTREPÔT6	2	417,70	1,02	22	171,26
125	ENTREPÔT6	2	409,20	1,54	16	155,50
126	ENTREPÔT6	2	382,72	1,00	21	164,57
127	ENTREPÔT6	2	362,78	0,43	15	119,72
128	ENTREPÔT6	2	433,33	1,20	21	195,00
129	ENTREPÔT6	2	400,71	1,30	21	144,26
130	ENTREPÔT6	2	373,63	1,24	19	138,24
131	ENTREPÔT6	2	435,00	1,44	22	200,10
132	ENTREPÔT6	2	393,64	1,31	16	137,77
133	ENTREPÔT7	2	409,24	1,24	16	155,51
134	ENTREPÔT7	2	398,60	0,34	16	139,51
135	ENTREPÔT7	2	355,87	0,31	12	135,23
136	ENTREPÔT7	2	418,73	0,94	18	175,87
137	ENTREPÔT7	2	400,98	1,06	24	180,44
138	ENTREPÔT7	2	453,72	1,67	18	176,95
139	ENTREPÔT7	2	413,00	0,71	15	181,72
140	ENTREPÔT7	2	444,07	0,72	21	213,15
141	ENTREPÔT7	2	374,66	1,36	16	138,62
142	ENTREPÔT7	2	382,54	0,28	23	141,54
143	ENTREPÔT7	2	370,61	0,32	17	151,95
144	ENTREPÔT7	2	394,71	0,64	13	149,99
145	ENTREPÔT7	2	375,10	0,34	19	153,79

OBS	Entrepôt	GROUPE	Capacite	TailleTblDi	FreqRet	TailleTblFact
146	ENTREPÔ	2	444,04	1,16	14	208,70
147	ENTREPÔ	2	414,58	1,62	24	149,25
148	ENTREPÔ	2	402,37	1,10	13	181,07
149	ENTREPÔ	2	404,11	1,18	18	177,81
150	ENTREPÔ	2	410,85	1,21	18	193,10
151	ENTREPÔ	2	409,54	0,77	17	167,91
152	ENTREPÔ	2	404,29	0,34	13	202,15
153	ENTREPÔ	2	401,07	1,11	18	148,40
154	ENTREPÔ	2	456,64	0,99	20	168,96
155	ENTREPÔ	3	526,67	2,32	25	289,67
156	ENTREPÔ	3	544,42	2,24	31	304,88
157	ENTREPÔ	3	559,62	1,98	28	285,41
158	ENTREPÔ	3	516,38	1,20	30	263,35
159	ENTREPÔ	3	561,56	1,89	31	292,01
160	ENTREPÔ	3	552,22	2,22	31	298,20
161	ENTREPÔ	3	535,55	1,77	32	294,55
162	ENTREPÔ	3	512,16	2,44	33	276,56
163	ENTREPÔ	3	545,96	1,31	30	289,36
164	ENTREPÔ	3	528,79	1,67	25	290,83
165	ENTREPÔ	3	561,16	1,81	34	308,64
166	ENTREPÔ	3	519,30	2,15	29	280,42
167	ENTREPÔ	3	503,51	1,77	30	271,90
168	ENTREPÔ	3	509,45	2,21	27	270,01
169	ENTREPÔ	3	530,23	2,55	28	281,02
170	ENTREPÔ	3	534,77	1,27	26	294,12
171	ENTREPÔ	3	517,60	1,52	33	274,33
172	ENTREPÔ	3	523,37	1,19	35	293,09
173	ENTREPÔ	3	502,01	1,77	25	266,07
174	ENTREPÔ	3	509,08	1,44	27	269,81
175	ENTREPÔ	3	567,05	2,26	29	311,88
176	ENTREPÔ	3	558,32	1,26	34	284,75
177	ENTREPÔ	3	567,29	1,67	26	289,32
178	ENTREPÔ	3	561,33	2,23	30	314,35
179	ENTREPÔ	3	525,59	2,07	30	294,33
180	ENTREPÔ	3	519,00	1,49	33	264,69
181	ENTREPÔ	3	553,04	1,32	31	282,05
182	ENTREPÔ	3	563,40	1,09	34	298,60
183	ENTREPÔ	3	527,93	1,52	34	279,80
184	ENTREPÔ	3	558,36	1,75	34	307,10
185	ENTREPÔ	3	501,44	2,45	34	275,79
186	ENTREPÔ	3	499,31	2,11	31	274,62
187	ENTREPÔ	3	564,18	1,88	30	315,94
188	ENTREPÔ	3	552,84	1,39	28	298,53
189	ENTREPÔ	3	501,40	1,13	27	265,74
190	ENTREPÔ	3	513,43	1,39	28	272,12
191	ENTREPÔ	3	555,31	1,99	33	305,42
192	ENTREPÔ	3	522,84	2,12	31	277,10
193	ENTREPÔ	3	560,88	1,86	26	308,49
194	ENTREPÔ	3	501,62	0,98	27	270,87
195	ENTREPÔ	3	514,79	2,25	27	283,13
196	ENTREPÔ	3	531,07	1,79	29	292,09
197	ENTREPÔ	3	527,83	1,97	34	285,03
198	ENTREPÔ	3	536,69	1,62	32	279,08
199	ENTREPÔ	3	560,22	1,04	26	308,12
200	ENTREPÔ	3	525,45	1,97	28	267,98
201	ENTREPÔ	3	535,58	1,99	27	278,50
202	ENTREPÔ	3	554,91	1,58	32	299,65
203	ENTREPÔ	3	557,05	1,80	27	300,80
204	ENTREPÔ	3	542,30	1,50	33	281,99
205	ENTREPÔ	3	569,00	2,76	28	295,88
206	ENTREPÔ	3	561,85	1,91	31	303,40
207	ENTREPÔ	3	546,84	2,55	24	278,89
208	ENTREPÔ	3	516,99	1,66	25	289,52
209	ENTREPÔ	3	545,45	2,35	26	272,73
210	ENTREPÔ	3	504,66	1,68	34	277,56
211	ENTREPÔ	3	554,87	2,35	33	299,63
212	ENTREPÔ	3	507,78	1,58	29	264,04
213	ENTREPÔ	3	530,86	1,03	26	291,97
214	ENTREPÔ	3	522,71	1,45	30	282,26
215	ENTREPÔ	3	505,98	1,92	27	268,17
216	ENTREPÔ	3	509,85	1,70	31	280,42
217	ENTREPÔ	3	508,15	2,16	28	264,24
218	ENTREPÔ	3	555,43	1,80	27	311,04
219	ENTREPÔ	3	563,27	2,17	25	298,53
220	ENTREPÔ	3	547,90	1,40	27	301,35

Source : Bases de données de l'entreprise BELL Canada, BELL Canada dans le Département Intelligence d'Affaires Service Extérieurs

Annexe VI : Test de significativité du pouvoir discriminant global et du pouvoir discriminant individuel des variables explicatives

Test du pouvoir discriminant individuel des variables explicatives

Statistiques de tests à une variable							
F Statistics, Num DF=2, Den DF=217							
Variable	Total Ecart-type	Ecart-type groupé	Ecart-type intra	R-carré	R-carré / (1-RSq)	Valeur F	Pr > F
TailleTblFact	73.1525	18.2280	86.5959	0.9385	15.2541	1655.07	<.0001
FreqRet	8.1104	2.6654	9.3652	0.8930	8.3444	905.37	<.0001
TailleTblDim	0.6748	0.3462	0.7090	0.7392	2.8337	307.46	<.0001

Source : Sortie du logiciel SAS

TailleTblFact : Taille des tables de faits

FreqRet : Durée de stockage des données

TailleTblDim: Taille des tables de dimension

Test de pouvoir discriminant global

Statistiques multivariées et Approximations F					
S=2 M=0 N=106.5					
Statistique	Valeur	Valeur F	DDL Num.	DDL Res.	Pr > F
Wilks' Lambda	0.04350774	271.92	6	430	<.0001
Pillai's Trace	1.06146779	81.43	6	432	<.0001
Hotelling-Lawley Trace	19.57161337	699.70	6	284.9	<.0001
Roy's Greatest Root	19.44754624	1400.22	3	216	<.0001
NOTE: la statistique F pour la plus grande racine de Roy est une borne supérieure.					
NOTE: la statistique F pour Lambda de Wilks est exacte.					

Annexe VII : Capacité de stockage en fonction du premier facteur discriminant

Obs	ENTREPOT	Capacite	GROUPE	Can1	Can2
1	ENTREPOT4	325,73	1	-5,06	0,50
2	ENTREPOT1	343,85	1	-4,90	-0,54
3	ENTREPOT4	325,26	1	-4,84	0,97
4	ENTREPOT3	332,77	1	-4,81	-0,21
5	ENTREPOT3	350,84	1	-4,78	0,11
6	ENTREPOT3	326,20	1	-4,74	-0,19
7	ENTREPOT1	322,34	1	-4,73	1,19
8	ENTREPOT1	358,49	1	-4,73	-0,13
9	ENTREPOT3	320,55	1	-4,71	-0,04
10	ENTREPOT2	330,64	1	-4,67	-0,21
11	ENTREPOT4	322,01	1	-4,67	0,27
12	ENTREPOT3	325,64	1	-4,63	-0,06
13	ENTREPOT1	352,24	1	-4,63	0,26
14	ENTREPOT1	328,28	1	-4,62	0,10
15	ENTREPOT4	328,82	1	-4,57	0,24
16	ENTREPOT1	357,69	1	-4,56	-0,15
17	ENTREPOT1	353,97	1	-4,51	-0,05
18	ENTREPOT3	355,12	1	-4,49	-1,01
19	ENTREPOT4	348,82	1	-4,47	0,06
20	ENTREPOT4	345,66	1	-4,46	0,25
21	ENTREPOT2	337,11	1	-4,46	0,43
22	ENTREPOT4	345,66	1	-4,46	0,27
23	ENTREPOT2	340,38	1	-4,46	0,56
24	ENTREPOT1	351,28	1	-4,45	-0,65
25	ENTREPOT3	349,72	1	-4,43	0,11
26	ENTREPOT5	353,55	2	-4,43	-0,09
27	ENTREPOT3	350,18	1	-4,42	0,78
28	ENTREPOT3	342,74	1	-4,38	0,38
29	ENTREPOT1	347,99	1	-4,35	-0,76
30	ENTREPOT5	351,26	2	-4,35	0,00
31	ENTREPOT2	339,57	1	-4,34	0,40
32	ENTREPOT4	336,23	1	-4,32	-0,54
33	ENTREPOT4	333,42	1	-4,28	0,06
34	ENTREPOT1	331,57	1	-4,28	-0,70
35	ENTREPOT2	331,86	1	-4,28	-0,13
36	ENTREPOT2	330,44	1	-4,24	-0,19
37	ENTREPOT3	359,52	1	-4,23	0,28
38	ENTREPOT3	359,53	1	-4,23	-0,39
39	ENTREPOT3	346,34	1	-4,22	-0,20
40	ENTREPOT3	342,00	1	-4,22	-0,54
41	ENTREPOT5	358,54	2	-4,14	0,00
42	ENTREPOT2	358,65	1	-4,14	0,64
43	ENTREPOT3	329,84	1	-4,12	0,51
44	ENTREPOT4	331,81	1	-4,11	-0,15
45	ENTREPOT2	337,70	1	-4,09	-0,16
46	ENTREPOT4	353,36	1	-4,09	0,15
47	ENTREPOT2	338,89	1	-4,03	-0,62
48	ENTREPOT1	352,73	1	-4,02	-0,54
49	ENTREPOT4	347,47	1	-4,01	-0,49
50	ENTREPOT3	344,17	1	-4,01	-0,28
51	ENTREPOT1	331,47	1	-4,00	-0,64
52	ENTREPOT1	340,00	1	-4,00	-0,28
53	ENTREPOT1	336,30	1	-3,99	-0,01
54	ENTREPOT1	355,39	1	-3,99	0,07
55	ENTREPOT3	342,99	1	-3,97	-0,66
56	ENTREPOT4	354,90	1	-3,95	-0,73
57	ENTREPOT2	341,60	1	-3,92	-0,71
58	ENTREPOT3	348,61	1	-3,91	-0,52
59	ENTREPOT4	333,92	1	-3,87	-0,65
60	ENTREPOT2	355,95	1	-3,85	-0,46
61	ENTREPOT4	345,76	1	-3,84	-0,42
62	ENTREPOT1	321,11	1	-3,80	-1,00
63	ENTREPOT4	342,58	1	-3,80	-1,59
64	ENTREPOT5	354,77	2	-3,78	-0,80
65	ENTREPOT5	353,02	2	-3,76	-0,29
66	ENTREPOT1	346,23	1	-3,75	-0,77
67	ENTREPOT3	333,73	1	-3,75	-0,52
68	ENTREPOT1	357,28	1	-3,74	-1,78
69	ENTREPOT3	357,80	1	-3,72	-0,45
70	ENTREPOT2	340,79	1	-3,72	-0,90
71	ENTREPOT4	331,03	1	-3,70	0,10
72	ENTREPOT4	341,81	1	-3,70	0,62
73	ENTREPOT3	338,73	1	-3,68	-0,62
74	ENTREPOT4	354,45	1	-3,67	-1,03
75	ENTREPOT2	337,63	1	-3,64	-0,13
76	ENTREPOT2	320,46	1	-3,63	-0,35
77	ENTREPOT2	326,19	1	-3,60	-0,38
78	ENTREPOT2	358,69	1	-3,54	-0,51
79	ENTREPOT2	348,57	1	-3,53	-0,18
80	ENTREPOT2	347,88	1	-3,53	-0,74
81	ENTREPOT2	354,17	1	-3,52	-0,50
82	ENTREPOT2	343,01	1	-3,51	-1,24
83	ENTREPOT1	339,82	1	-3,50	-0,70
84	ENTREPOT4	334,90	1	-3,49	-0,02
85	ENTREPOT3	350,25	1	-3,49	-1,70

Obs	ENTREPOT	Capacite	GROUPE	Can1	Can2
86	ENTREPOT1	355,28	1	-3,48	-0,12
87	ENTREPOT1	350,81	1	-3,47	-0,74
88	ENTREPOT6	362,78	2	-3,38	0,79
89	ENTREPOT3	342,95	1	-3,37	-0,47
90	ENTREPOT1	354,75	1	-3,36	-0,62
91	ENTREPOT4	355,75	1	-3,36	-0,35
92	ENTREPOT6	354,86	2	-3,35	-0,27
93	ENTREPOT2	347,18	1	-3,34	-0,97
94	ENTREPOT2	348,34	1	-3,33	-0,72
95	ENTREPOT7	355,87	2	-3,32	-0,85
96	ENTREPOT4	359,07	1	-3,22	-0,75
97	ENTREPOT6	364,38	2	-2,89	0,75
98	ENTREPOT7	398,60	2	-2,46	0,09
99	ENTREPOT7	394,71	2	-2,40	-0,67
100	ENTREPOT6	364,17	2	-2,27	1,32
101	ENTREPOT5	360,23	2	-2,14	1,40
102	ENTREPOT6	393,64	2	-2,07	1,71
103	ENTREPOT7	374,66	2	-2,01	1,76
104	ENTREPOT5	361,16	2	-2,01	2,49
105	ENTREPOT7	370,61	2	-1,81	-0,21
106	ENTREPOT6	418,44	2	-1,70	0,55
107	ENTREPOT6	371,39	2	-1,61	1,71
108	ENTREPOT6	373,63	2	-1,59	2,37
109	ENTREPOT5	446,44	2	-1,53	0,88
110	ENTREPOT6	385,12	2	-1,46	0,30
111	ENTREPOT7	401,07	2	-1,41	1,47
112	ENTREPOT7	409,24	2	-1,39	0,86
113	ENTREPOT7	375,10	2	-1,39	0,28
114	ENTREPOT6	409,20	2	-1,25	1,34
115	ENTREPOT7	382,54	2	-1,24	1,75
116	ENTREPOT5	376,81	2	-1,20	-0,68
117	ENTREPOT6	384,81	2	-1,07	0,94
118	ENTREPOT5	407,78	2	-1,03	0,74
119	ENTREPOT6	400,71	2	-0,98	2,75
120	ENTREPOT7	409,54	2	-0,95	-0,15
121	ENTREPOT7	402,37	2	-0,93	-1,23
122	ENTREPOT6	432,30	2	-0,88	1,51
123	ENTREPOT7	413,00	2	-0,76	-1,34
124	ENTREPOT5	428,39	2	-0,66	-1,46
125	ENTREPOT6	458,63	2	-0,64	-1,26
126	ENTREPOT6	405,21	2	-0,58	0,30
127	ENTREPOT7	404,29	2	-0,45	-3,33
128	ENTREPOT5	448,00	2	-0,41	-0,80
129	ENTREPOT6	397,89	2	-0,39	2,40
130	ENTREPOT6	396,53	2	-0,39	1,61
131	ENTREPOT7	418,73	2	-0,38	0,06
132	ENTREPOT5	380,36	2	-0,37	3,20
133	ENTREPOT6	382,72	2	-0,31	1,42
134	ENTREPOT7	456,64	2	-0,31	0,95
135	ENTREPOT5	388,18	2	-0,28	1,02
136	ENTREPOT7	404,11	2	-0,19	0,37
137	ENTREPOT7	414,58	2	-0,14	3,84
138	ENTREPOT5	391,91	2	-0,11	3,20
139	ENTREPOT7	453,72	2	0,00	1,18
140	ENTREPOT6	417,70	2	0,13	1,44
141	ENTREPOT5	450,40	2	0,24	-0,18
142	ENTREPOT5	406,56	2	0,29	0,07
143	ENTREPOT5	454,14	2	0,35	0,33
144	ENTREPOT7	444,04	2	0,37	-2,02
145	ENTREPOT7	410,85	2	0,43	-0,23
146	ENTREPOT7	400,98	2	0,85	1,64
147	ENTREPOT6	433,33	2	1,01	0,48
148	ENTREPOT5	409,70	2	1,33	0,67
149	ENTREPOT6	435,00	2	1,49	0,91
150	ENTREPOT7	444,07	2	1,50	-1,06
151	ENTREPOT5	403,08	2	1,54	0,54
152	ENTREPOT5	443,11	2	1,56	-0,49
153	ENTREPOT5	455,96	2	1,58	-0,70
154	ENTREPOT6	458,63	2	2,33	-0,39

Obs	ENTREPOT	Capacite	GROUPE	Can1	Can2
155	ENTREPOT8	502,01	3	4,79	-0,52
156	ENTREPOT9	501,40	3	4,81	-1,00
157	ENTREPOT9	501,62	3	4,95	-1,45
158	ENTREPOT8	509,08	3	5,12	-0,68
159	ENTREPOT8	516,38	3	5,25	0,00
160	ENTREPOT10	505,98	3	5,28	0,16
161	ENTREPOT10	507,78	3	5,29	0,32
162	ENTREPOT9	513,43	3	5,36	-0,58
163	ENTREPOT10	508,15	3	5,41	0,99
164	ENTREPOT10	525,45	3	5,46	0,51
165	ENTREPOT8	509,45	3	5,49	0,55
166	ENTREPOT10	545,45	3	5,50	0,40
167	ENTREPOT10	546,84	3	5,51	-0,06
168	ENTREPOT10	530,86	3	5,65	-2,52
169	ENTREPOT10	516,99	3	5,69	-1,66
170	ENTREPOT10	535,58	3	5,73	-0,16
171	ENTREPOT8	528,79	3	5,74	-1,70
172	ENTREPOT9	567,29	3	5,85	-1,38
173	ENTREPOT8	534,77	3	5,85	-2,22
174	ENTREPOT8	503,51	3	5,86	0,56
175	ENTREPOT9	519,00	3	5,94	1,21
176	ENTREPOT8	526,67	3	6,00	-0,62
177	ENTREPOT9	514,79	3	6,04	0,07
178	ENTREPOT10	522,71	3	6,12	-0,39
179	ENTREPOT8	559,62	3	6,17	-0,19
180	ENTREPOT8	519,30	3	6,22	0,56
181	ENTREPOT9	553,04	3	6,22	-0,31
182	ENTREPOT8	530,23	3	6,27	0,91
183	ENTREPOT10	563,27	3	6,29	-1,22
184	ENTREPOT9	499,31	3	6,30	1,26
185	ENTREPOT10	560,22	3	6,30	-3,18
186	ENTREPOT10	509,85	3	6,33	0,35
187	ENTREPOT8	517,60	3	6,34	0,85
188	ENTREPOT8	545,96	3	6,34	-0,91
189	ENTREPOT10	547,90	3	6,37	-2,05
190	ENTREPOT9	522,84	3	6,40	1,18
191	ENTREPOT9	536,69	3	6,41	0,55
192	ENTREPOT9	552,84	3	6,42	-1,69
193	ENTREPOT9	531,07	3	6,51	-0,51
194	ENTREPOT10	557,05	3	6,53	-1,39
195	ENTREPOT10	542,30	3	6,64	0,50
196	ENTREPOT9	560,88	3	6,71	-1,87
197	ENTREPOT10	504,66	3	6,71	1,24
198	ENTREPOT9	527,93	3	6,73	0,90
199	ENTREPOT8	558,32	3	6,80	0,27
200	ENTREPOT8	512,16	3	6,87	2,24
201	ENTREPOT8	561,56	3	6,89	0,19
202	ENTREPOT9	525,59	3	6,91	0,12
203	ENTREPOT10	555,43	3	6,95	-1,82
204	ENTREPOT10	569,00	3	6,96	0,63
205	ENTREPOT9	501,44	3	7,01	2,55
206	ENTREPOT8	535,55	3	7,11	0,16
207	ENTREPOT9	527,83	3	7,15	1,40
208	ENTREPOT10	554,91	3	7,22	-0,37
209	ENTREPOT8	523,37	3	7,27	0,08
210	ENTREPOT9	563,40	3	7,28	-0,58
211	ENTREPOT8	552,22	3	7,30	0,46
212	ENTREPOT10	561,85	3	7,36	-0,26
213	ENTREPOT8	567,05	3	7,53	-0,58
214	ENTREPOT8	544,42	3	7,57	0,21
215	ENTREPOT9	564,18	3	7,68	-1,09
216	ENTREPOT10	554,87	3	7,75	1,13
217	ENTREPOT9	561,33	3	7,78	-0,47
218	ENTREPOT9	555,31	3	7,81	0,31
219	ENTREPOT9	558,36	3	7,93	0,13
220	ENTREPOT8	561,16	3	8,02	0,16

Annexe VIII : Test de validation du choix de nombre de facteurs discriminants linéaires

	Corrélation canonique	Corrélation canonique ajustée	Erreur type approchée	Corrélation canonique au carré	Valeurs propres de Inv(E)*H = CanRsq/(1-CanRsq)			
					Valeur propre	Différence	Proportion	Cumulé
1	0.975241	0.975005	0.003305	0.951094	19.4475	19.3235	0.9937	0.9937
2	0.332225	0.326865	0.060115	0.110373	0.1241		0.0063	1.0000

	Test de H0 : les corrélations canoniques de la ligne en cours et suivantes sont égales à zéro				
	Rapport de vraisemblance	Valeur de F approchée	DDL Num.	DDL Res.	Pr > F
1	0.04350774	271.92	6	430	<.0001
2	0.88962658	13.40	2	216	<.0001

Coefficients de corrélation de Pearson, N = 220 Proba > |r| sous H0: Rho=0		
	Can1	Can2
TailleTblFact	0.99253 <.0001	-0.11815 0.0804
FreqRet	0.96663 <.0001	0.19756 0.0033
TailleTblDim	0.87742 <.0001	0.25084 0.0002

Source **:** Sortie du logiciel SAS

TailleTblFact : Taille des tables de faits

FreqRet : Durée de stockage des données

TailleTblDim: Taille des tables de dimension

Annexe IX : Classification des entrepôts et probabilités d'appartenance

Obs	De GROUPE	Classé dans GROUPE		Probabilité a posteriori d'un membre de GROUPE 1	2	3
1	1	1		0,961	0,039	0,000
2	1	1		0,991	0,009	0,000
3	1	1		1,000	0,000	0,000
4	1	1		0,990	0,010	0,000
5	1	1		0,995	0,006	0,000
6	1	1		0,995	0,005	0,000
7	1	1		0,998	0,002	0,000
8	1	1		0,990	0,010	0,000
9	1	1		0,975	0,025	0,000
10	1	1		0,999	0,001	0,000
11	1	1		0,962	0,039	0,000
12	1	1		0,998	0,002	0,000
13	1	1		0,999	0,001	0,000
14	1	1		0,998	0,002	0,000
15	1	1		0,976	0,024	0,000
16	1	1		0,993	0,007	0,000
17	1	1		0,998	0,002	0,000
18	1	1		0,995	0,005	0,000
19	1	1		0,998	0,002	0,000
20	1	1		0,999	0,001	0,000
21	1	1		0,993	0,007	0,000
22	1	1		0,999	0,001	0,000
23	1	1		0,968	0,032	0,000
24	1	1		0,990	0,010	0,000
25	1	1		0,994	0,006	0,000
26	1	1		0,995	0,005	0,000
27	1	1		0,974	0,026	0,000
28	1	1		0,996	0,004	0,000
29	1	1		0,979	0,021	0,000
30	1	1		0,997	0,003	0,000
31	1	1		0,999	0,001	0,000
32	1	1		0,997	0,003	0,000
33	1	1		0,990	0,010	0,000
34	1	1		0,976	0,024	0,000
35	1	1		0,990	0,010	0,000
36	1	1		0,977	0,023	0,000
37	1	1		0,979	0,021	0,000
38	1	1		0,996	0,004	0,000
39	1	1		0,997	0,003	0,000
40	1	1		0,985	0,015	0,000
41	1	1		0,975	0,025	0,000
42	1	1		0,961	0,039	0,000
43	1	1		0,969	0,031	0,000
44	1	1		0,994	0,007	0,000
45	1	1		0,999	0,001	0,000
46	1	1		0,999	0,001	0,000
47	1	1		0,996	0,004	0,000
48	1	1		0,958	0,042	0,000
49	1	1		0,991	0,009	0,000
50	1	1		0,997	0,003	0,000
51	1	1		0,999	0,001	0,000
52	1	1		0,985	0,015	0,000
53	1	1		0,999	0,001	0,000
54	1	1		0,989	0,011	0,000
55	1	1		0,997	0,003	0,000
56	1	1		0,996	0,004	0,000
57	1	1		0,985	0,015	0,000
58	1	1		0,997	0,003	0,000
59	1	1		0,993	0,007	0,000
60	1	1		0,992	0,008	0,000
61	1	1		0,987	0,013	0,000
62	1	1		0,995	0,006	0,000
63	1	1		0,999	0,001	0,000
64	1	1		0,999	0,001	0,000
65	1	1		0,995	0,005	0,000
66	1	1		0,994	0,006	0,000
67	1	1		0,989	0,011	0,000
68	1	1		0,998	0,002	0,000
69	1	1		0,998	0,002	0,000
70	1	1		0,994	0,006	0,000

Obs	De GROUPE	Classé dans		Probabilité a posteriori d'un membre de GROUPE 1	2	3
71	1	1		0,997	0,003	0,000
72	1	1		0,995	0,005	0,000
73	1	1		0,992	0,008	0,000
74	1	1		0,959	0,042	0,000
75	1	1		0,994	0,006	0,000
76	1	1		0,998	0,002	0,000
77	1	1		0,996	0,004	0,000
78	1	1		0,963	0,037	0,000
79	1	1		0,999	0,001	0,000
80	1	1		0,990	0,010	0,000
81	1	1		0,976	0,024	0,000
82	1	1		0,994	0,006	0,000
83	1	1		1,000	0,001	0,000
84	1	1		0,999	0,001	0,000
85	1	1		0,997	0,003	0,000
86	1	1		0,952	0,048	0,000
87	1	1		0,948	0,053	0,000
88	1	1		0,992	0,008	0,000
89	2	2		0,011	0,989	0,000
90	2	2		0,006	0,994	0,000
91	2	2		0,000	1,000	0,000
92	2	2		0,000	1,000	0,000
93	2	2		0,030	0,970	0,000
94	2	2		0,000	1,000	0,000
95	2	2		0,001	1,000	0,000
96	2	2		0,024	0,976	0,000
97	2	2		0,000	1,000	0,000
98	2	2		0,099	0,902	0,000
99	2	2		0,000	1,000	0,000
100	2	2		0,000	1,000	0,000
101	2	1	*	0,997	0,003	0,000
102	2	2		0,031	0,969	0,000
103	2	2		0,000	1,000	0,000
104	2	1	*	0,998	0,002	0,000
105	2	2		0,000	1,000	0,000
106	2	2		0,000	1,000	0,000
107	2	1	*	0,994	0,006	0,000
108	2	2		0,003	0,997	0,000
109	2	1	*	0,985	0,015	0,000
110	2	1	*	0,991	0,009	0,000
111	2	2		0,000	1,000	0,000
112	2	1	*	0,659	0,341	0,000
113	2	2		0,006	0,994	0,000
114	2	2		0,152	0,848	0,000
115	2	2		0,000	0,918	0,082
116	2	2		0,009	0,991	0,000
117	2	1	*	0,949	0,051	0,000
118	2	2		0,016	0,984	0,000
119	2	2		0,031	0,969	0,000
120	2	2		0,000	1,000	0,000
121	2	2		0,052	0,948	0,000
122	2	2		0,002	0,998	0,000
123	2	2		0,002	0,998	0,000
124	2	2		0,000	1,000	0,000
125	2	2		0,007	0,993	0,000
126	2	2		0,000	1,000	0,000
127	2	1	*	0,895	0,105	0,000
128	2	2		0,000	1,000	0,000
129	2	2		0,001	0,999	0,000
130	2	2		0,009	0,991	0,000
131	2	2		0,000	1,000	0,000
132	2	2		0,065	0,935	0,000
133	2	2		0,016	0,984	0,000
134	2	2		0,464	0,536	0,000
135	2	1	*	0,963	0,037	0,000

Obs	De	Classé dans		Probabilité a posteriori d'un membre de GROUPE 1	2	3
136	2	2		0,001	0,999	0,000
137	2	2		0,000	1,000	0,000
138	2	2		0,000	1,000	0,000
139	2	2		0,014	0,987	0,000
140	2	2		0,000	1,000	0,000
141	2	2		0,053	0,947	0,000
142	2	2		0,005	0,995	0,000
143	2	2		0,125	0,875	0,000
144	2	1	+	0,567	0,433	0,000
145	2	2		0,026	0,974	0,000
146	2	2		0,001	0,999	0,000
147	2	2		0,000	1,000	0,000
148	2	2		0,021	0,979	0,000
149	2	2		0,001	0,999	0,000
150	2	2		0,000	1,000	0,000
151	2	2		0,010	0,991	0,000
152	2	2		0,025	0,975	0,000
153	2	2		0,011	0,990	0,000
154	2	2		0,001	1,000	0,000
155	3	3		0,000	0,000	1,000
156	3	3		0,000	0,000	1,000
157	3	3		0,000	0,000	1,000
158	3	3		0,000	0,000	1,000
159	3	3		0,000	0,000	1,000
160	3	3		0,000	0,000	1,000
161	3	3		0,000	0,000	1,000
162	3	3		0,000	0,000	1,000
163	3	3		0,000	0,000	1,000
164	3	3		0,000	0,000	1,000
165	3	3		0,000	0,000	1,000
166	3	3		0,000	0,000	1,000
167	3	3		0,000	0,000	1,000
168	3	3		0,000	0,000	1,000
169	3	3		0,000	0,000	1,000
170	3	3		0,000	0,000	1,000
171	3	3		0,000	0,000	1,000
172	3	3		0,000	0,000	1,000
173	3	3		0,000	0,000	1,000
174	3	3		0,000	0,000	1,000
175	3	3		0,000	0,000	1,000
176	3	3		0,000	0,000	1,000
177	3	3		0,000	0,000	1,000
178	3	3		0,000	0,000	1,000
179	3	3		0,000	0,000	1,000
180	3	3		0,000	0,000	1,000
181	3	3		0,000	0,000	1,000
182	3	3		0,000	0,000	1,000
183	3	3		0,000	0,000	1,000
184	3	3		0,000	0,000	1,000
185	3	3		0,000	0,000	1,000
186	3	3		0,000	0,000	1,000
187	3	3		0,000	0,000	1,000
188	3	3		0,000	0,000	1,000
189	3	3		0,000	0,000	1,000
190	3	3		0,000	0,000	1,000
191	3	3		0,000	0,000	1,000
192	3	3		0,000	0,000	1,000
193	3	3		0,000	0,000	1,000
194	3	3		0,000	0,000	1,000
195	3	3		0,000	0,000	1,000
196	3	3		0,000	0,000	1,000
197	3	3		0,000	0,000	1,000
198	3	3		0,000	0,000	1,000
199	3	3		0,000	0,000	1,000
200	3	3		0,000	0,000	1,000
201	3	3		0,000	0,000	1,000
202	3	3		0,000	0,000	1,000
203	3	3		0,000	0,000	1,000
204	3	3		0,000	0,000	1,000
205	3	3		0,000	0,000	1,000
206	3	3		0,000	0,000	1,000
207	3	3		0,000	0,000	1,000
208	3	3		0,000	0,000	1,000
209	3	3		0,000	0,000	1,000
210	3	3		0,000	0,000	1,000
211	3	3		0,000	0,000	1,000
212	3	3		0,000	0,000	1,000
213	3	3		0,000	0,000	1,000
214	3	3		0,000	0,000	1,000
215	3	3		0,000	0,000	1,000
216	3	3		0,000	0,000	1,000
217	3	3		0,000	0,000	1,000
218	3	3		0,000	0,000	1,000
219	3	3		0,000	0,000	1,000
220	3	3		0,000	0,000	1,000

Annexe X : Tests d'égalité des variances covariances entre les groupes d'entrepôts

I-1 Test de Box (Approximation asymptotique du khi²)

-2Log(M)	289,775
Khi² (Valeur observée)	283,865
Khi² (Valeur critique)	21,026
DDL	12
p-value	< 0,0001
alpha	0,05

Interprétation du test :
H0 : Les matrices de covariance intra-classe sont égales.
Ha : Les matrices de covariance intra-classe sont différentes.
Étant donné que la p-value calculée est inférieure au niveau de signification alpha=0,05, on doit rejeter l'hypothèse nulle H0, et retenir l'hypothèse alternative Ha.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inférieur à 0,01%

I-2 Test de Box (Approximation asymptotique du F de Fisher) :

-2Log(M)	289,775
F (Valeur observée)	23,654
F (Valeur critique)	1,752
DDL1	12
DDL2	197576
p-value	< 0,0001
alpha	0,05

Interprétation du test :
H0 : Les matrices de covariance intra-classe sont égales.
Ha : Les matrices de covariance intra-classe sont différentes.
Étant donné que la p-value calculée est inférieure au niveau de signification alpha=0,05, on doit rejeter l'hypothèse nulle H0, et retenir l'hypothèse alternative Ha.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inférieur à 0,01%.

I-3 Test de Kullback

K (Valeur observée)	144,888
K (Valeur critique)	21,026
DDL	12
p-value	< 0,0001
alpha	0,05

Interprétation du test :
H0 : Les matrices de covariance intra-classe sont égales.
Ha : Les matrices de covariance intra-classe sont différentes.
Étant donné que la p-value calculée est inférieure au niveau de signification alpha=0,05, on doit rejeter l'hypothèse nulle H0, et retenir l'hypothèse alternative Ha.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inférieur à 0,01%.

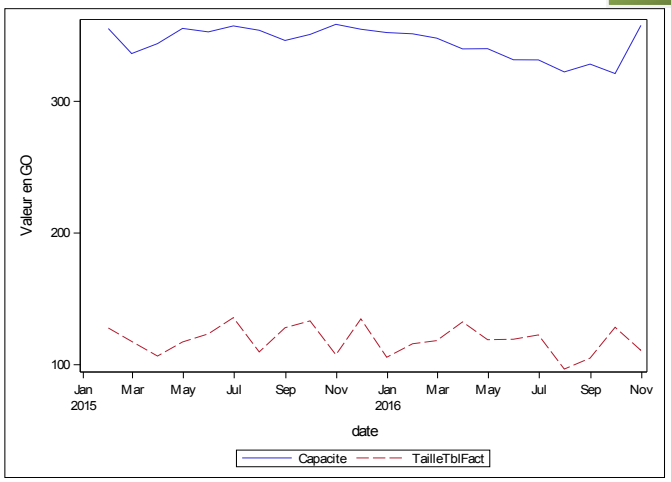
I-4 Test du Lambda de Wilks (approximation de Rao) :

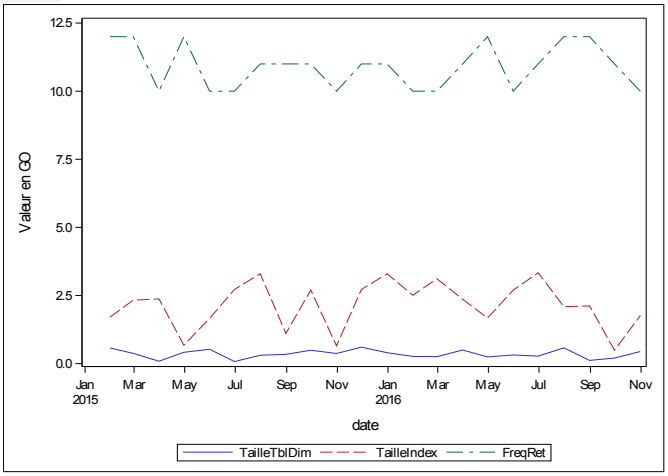
Lambda	0,044
F (Valeur observée)	271,918
F (Valeur critique)	2,120
DDL1	6
DDL2	430
p-value	< 0,0001
alpha	0,05

Interprétation du test :
H0 : Les vecteurs moyens des 3 classes sont égaux.
Ha : Au moins l'un des vecteurs moyens est différent d'un autre.
Étant donné que la p-value calculée est inférieure au niveau de signification alpha=0,05, on doit rejeter l'hypothèse nulle H0, et retenir l'hypothèse alternative Ha.
Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inférieur à 0,01%.

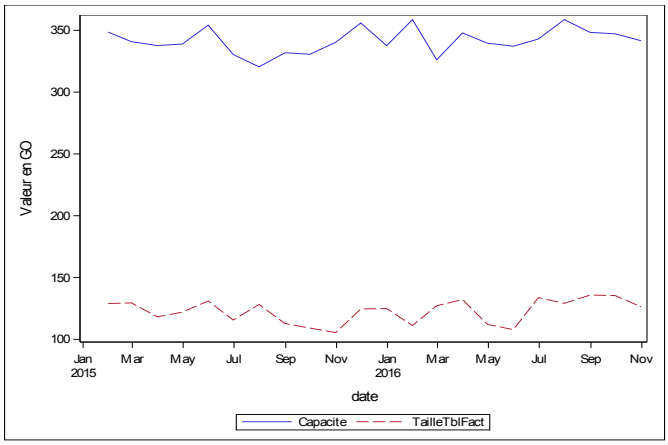
Annexe XI : Évolution des variables d'analyse (capacité de stockage, taille de la table des faits, taille des tables de dimension, taille des index, fréquence de rétention) par magasin de données selon les périodes d'observation

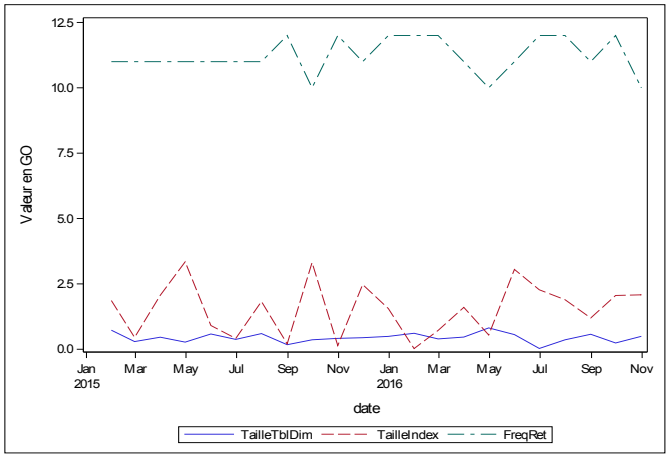
Magasin 1



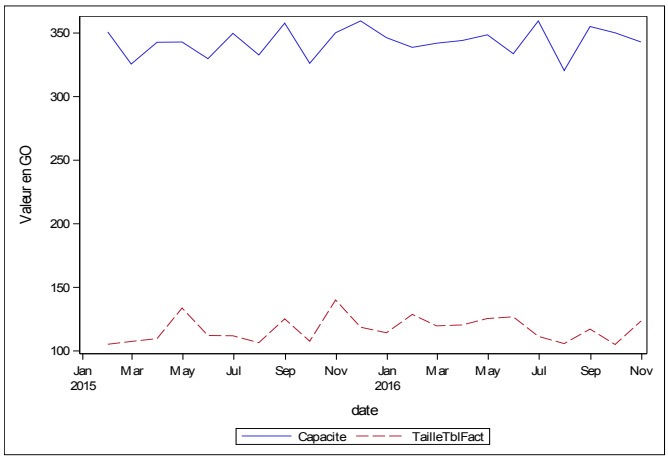


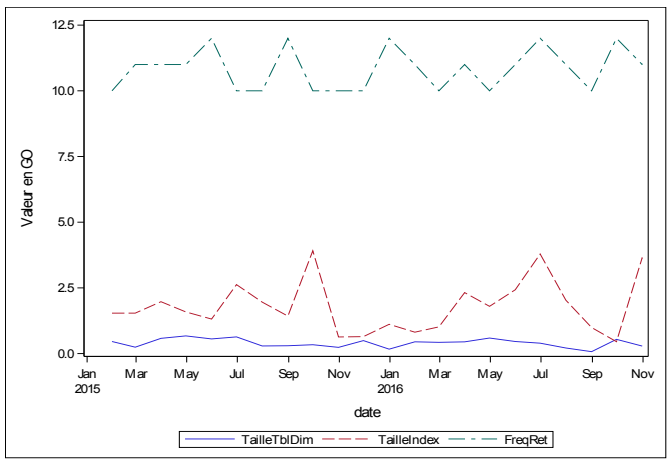
Magasin 2



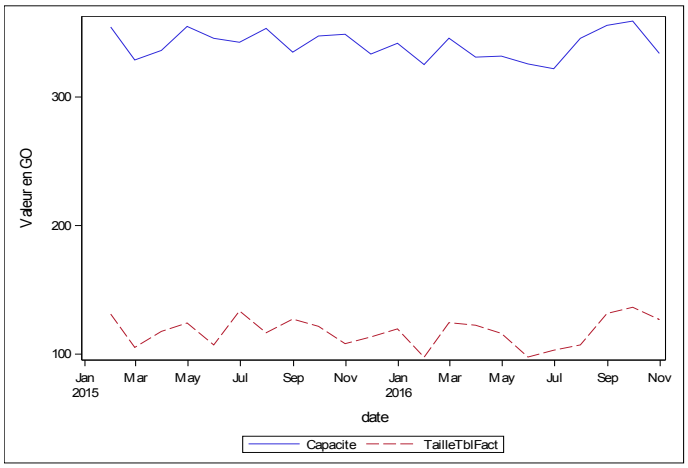


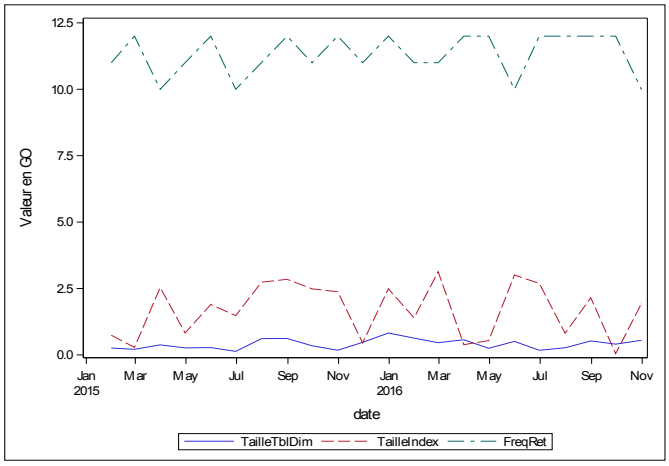
Magasin 3



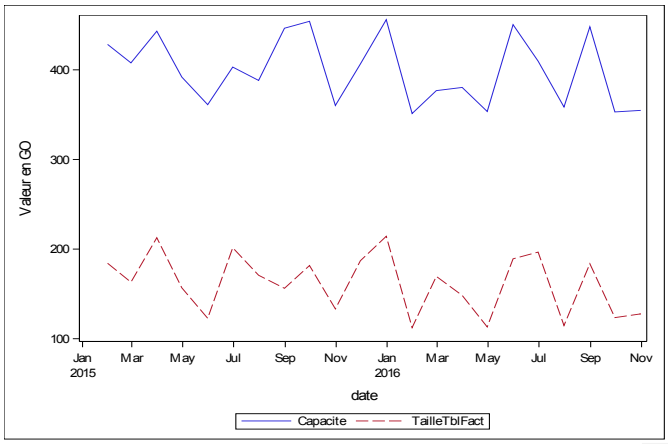


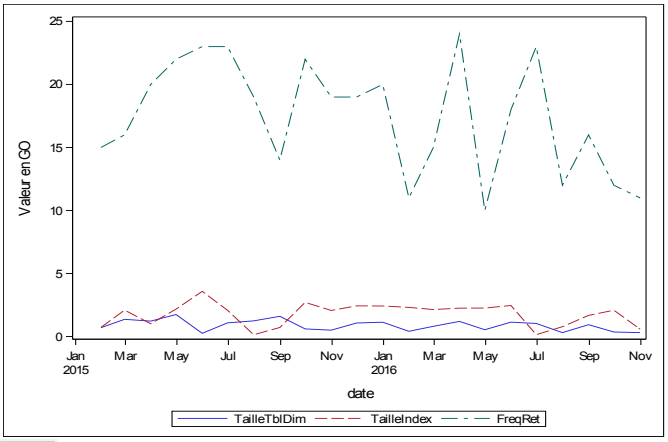
Magasin 4



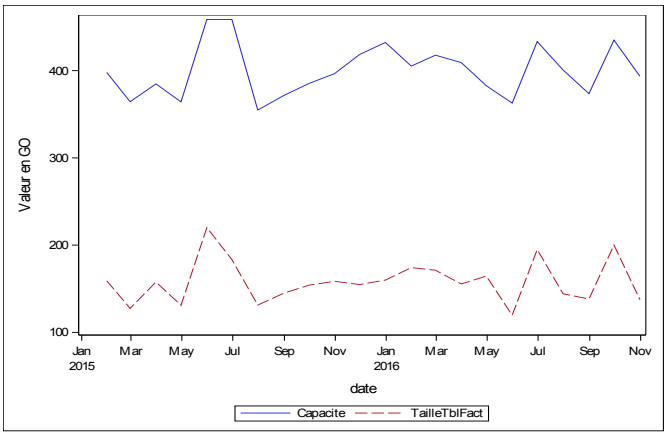


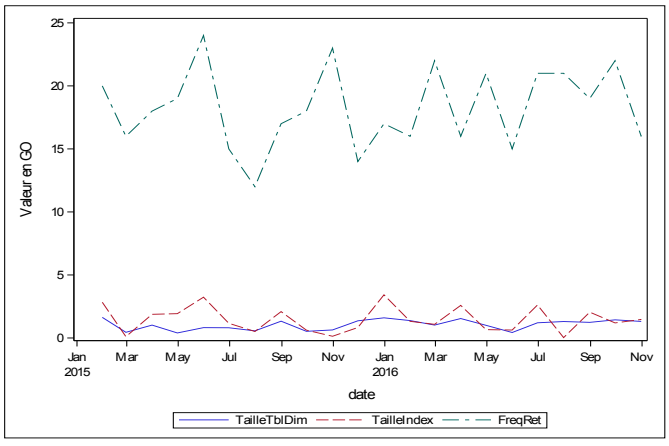
Magasin 5



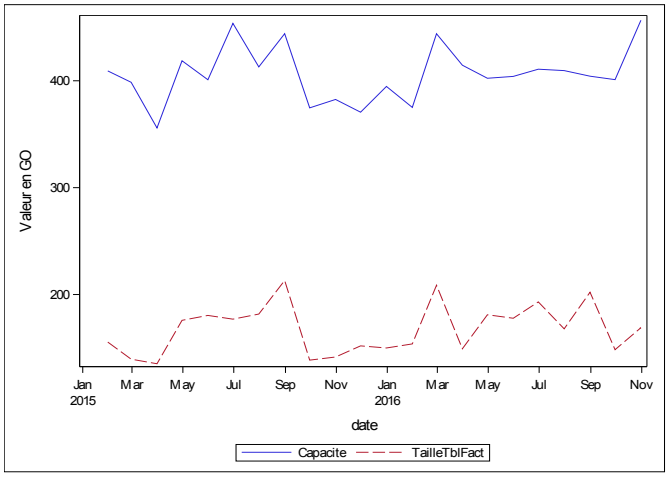


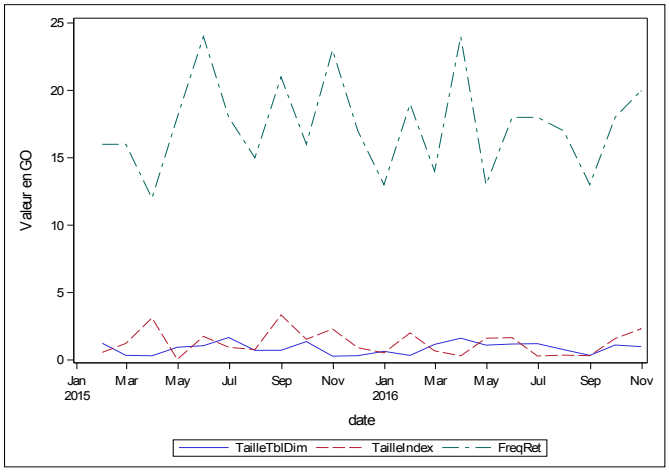
Magasin 6



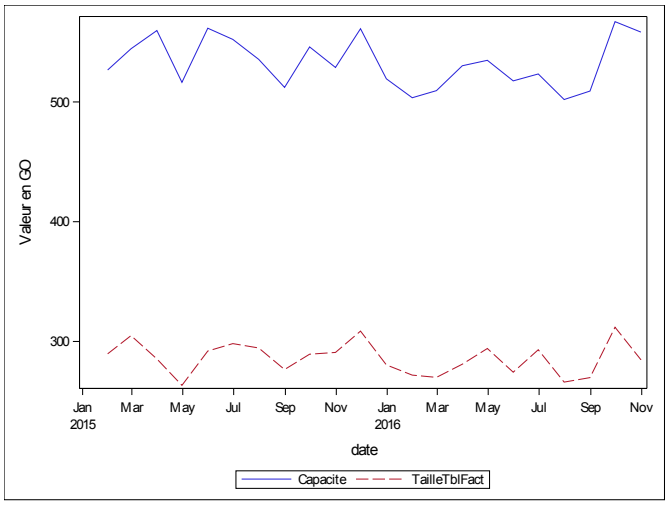


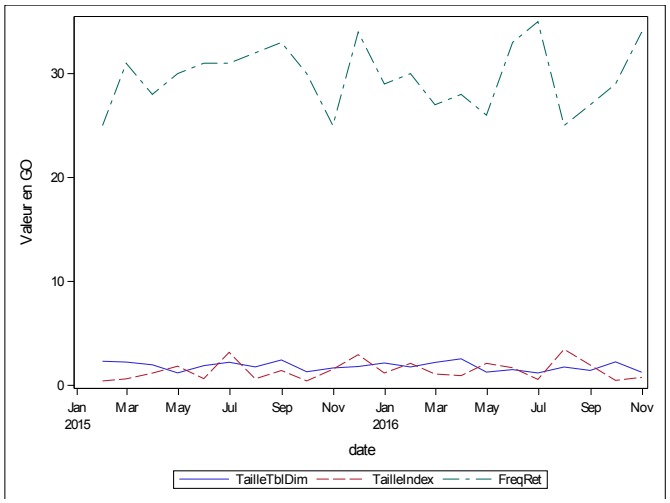
Magasin 7



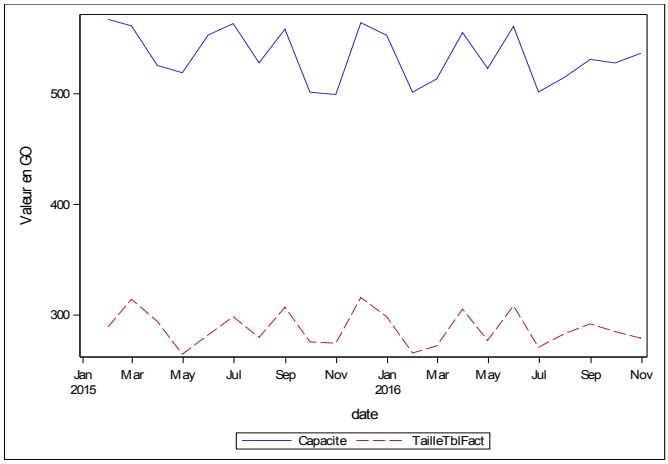


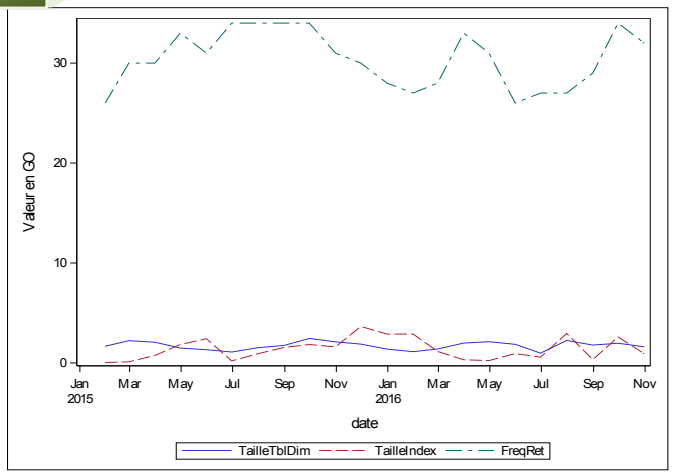
Magasin 8



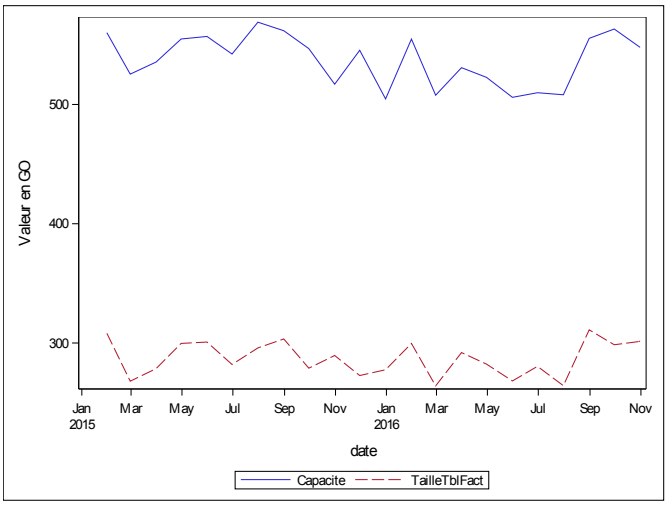


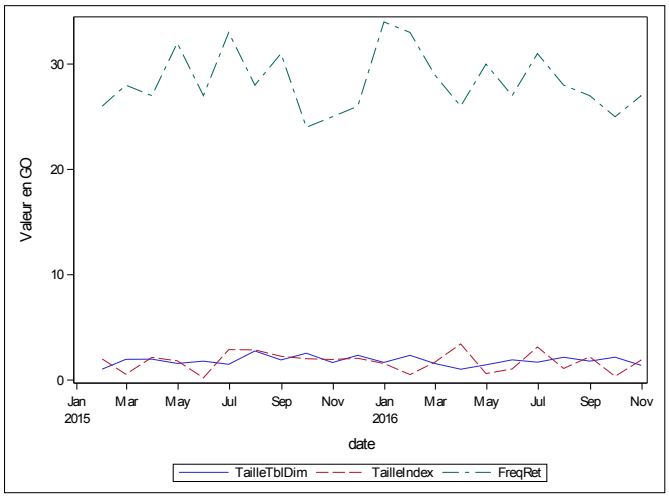
Magasin 9





Magasin 10


Valeur en GO
500
400
300
Jan 2015
Mar
May
Jul
Sep
Nov
Jan 2016
Mar
May
Jul
Sep
Nov
date
Capacite
TailleTblFact


Valeur en GO
30
20
10
0
Jan 2015
Mar
May
Jul
Sep
Nov
Jan 2016
Mar
May
Jul
Sep
Nov
date
TailleTblDim
TailleIndex
FreqRet

Annexe XII : Test de stationnarité (Augmented Dickey-Fuller test : ADF) des variables d'analyse

```
#---------------SCRIPT PYTHON-------------------------------------------------------------
#   AUTHOR:       JEAN BAPTISTE LALA
#   DESCRIPTION:  Test de stationnarité de chaque série (ADF test)
#-----------------------------------------------------------------------------------------


import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller
for columName in cols:
    result= adfuller(Panel_Datas[columName])
    print('----------Test ADF for : ' +columName+'----------------------' )
    print('ADF Statistic: %f' % result[0])
    print('p-value: %f' % result[1])
    print('Critical Values:')
    for key, value in result[4].items():
        print('\t%s: %.3f' % (key, value))
    #Tracer les graphiques d'autocorrélation
    acf = pd.DataFrame(sm.tsa.stattools.acf(Panel_Datas[columName]), columns=['ACF'])
    fig = acf[1:].plot(kind='bar', title='Autocorrelations')

#-----------------------------------------------------------------------------------------|
```

Résultat du test

```
----------Test ADF for : Capacite-------------------------
ADF Statistic: -3.473488
p-value: 0.00869
Critical Values:
        1%: -3.462
        10%: -2.574
        5%: -2.876

----------Test ADF for : TailleTblFact--------------------
ADF Statistic: -4.666015
p-value: 0.000097
Critical Values:
        1%: -3.462
        10%: -2.574
        5%: -2.876

----------Test ADF for : TailleTblDim---------------------
ADF Statistic: -4.106493
p-value: 0.000945
Critical Values:
        1%: -3.462
        10%: -2.574
        5%: -2.875

----------Test ADF for : FreqRet--------------------------
ADF Statistic: -4.056466
p-value: 0.001142
Critical Values:
        1%: -3.462
        10%: -2.574
        5%: -2.876
-----------------------------------------------------------
```

Annexe XIII : Spécification du modèle à effet temporel en langage Python

```
#----------------SCRIPT PYTHON-------------------------------------------------------------
#   AUTHOR:       JEAN BAPTISTE LALA
#   DESCRIPTION:  Time effect panel data analysis model
#-------------------------------------------------------------------------------------------


import pandas as pd
from numpy import log

#Read Data
df=pd.read_excel('Data.xlsx')
df.tail()

#Extract the analysis variables
DATAS = log(df[[3,4,5,7]].values)
print(DATAS)

#Get colum names for analysis variables
cols=df[[3,4,5,7]].columns
for c in cols:
    print(c)

#define time dimension and entity dimension
Temps_individu=[df['date'],df['ENTREPOT']]

#Input Panel data analisis using multindex
Panel_Datas = pd.DataFrame(DATAS ,index=pd.MultiIndex.from_tuples(list(zip(*Temps_individu))),
                           columns=cols)

#--Fit panel data analysis model

from pandas.stats.plm import PanelOLS
#1- Model with time fixed effects
Time_Fixed_Effects_Model  = PanelOLS(y=Panel_Datas['Capacite'],
                                     x=Panel_Datas[['TailleTblFact','TailleTblDim','FreqRet']],
                                     intercept = True,time_effects=True)
#Display result
print(Time_Fixed_Effects_Model)
```

Annexe XIV : Résultats d'estimations fournies par le modèle à effet temporel

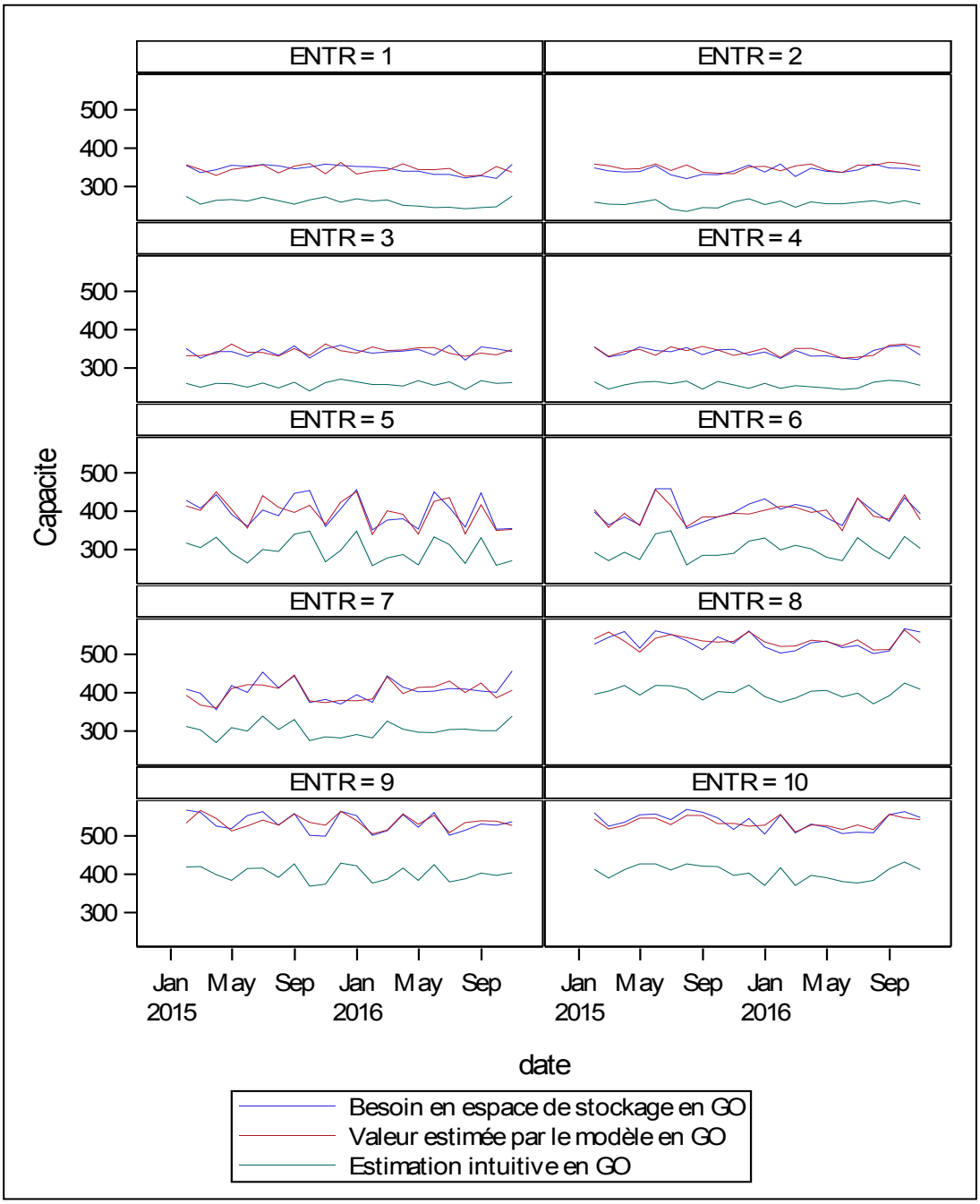
date	ENTREPOT	Capacite	Precicted Value	Confident Interval 95%		Error_model	Prev_Intuitive	PCT_error_Model	PCT_error_Intuitive
31-janv-15	ENTREPOT1	355,28	356,45	326,18	386,73	-1,17	274	0,330	22,878
31-janv-15	ENTREPOT2	348,57	358,51	328,19	388,83	-9,94	259	2,851	25,697
31-janv-15	ENTREPOT3	350,84	332,06	301,75	362,38	18,77	260	5,351	25,892
31-janv-15	ENTREPOT4	354,45	355,16	324,82	385,51	-0,72	264	0,202	25,518
31-janv-15	ENTREPOT5	428,39	413,83	383,46	444,19	14,57	317	3,400	26,003
31-janv-15	ENTREPOT6	397,89	404,36	373,68	435,03	-6,46	293	1,624	26,362
31-janv-15	ENTREPOT7	409,24	393,41	363,05	423,78	15,82	312	3,866	23,760
31-janv-15	ENTREPOT8	526,67	540,32	509,57	571,07	-13,65	396	2,592	24,810
31-janv-15	ENTREPOT9	567,29	533,25	502,77	563,73	34,05	419	6,001	26,141
31-janv-15	ENTREPOT10	560,22	543,88	512,85	574,92	16,33	413	2,916	26,279
28-févr-15	ENTREPOT1	336,30	344,45	314,16	374,74	-8,15	254	2,424	24,471
28-févr-15	ENTREPOT2	340,79	354,01	323,68	384,34	-13,22	254	3,880	25,467
28-févr-15	ENTREPOT3	325,64	332,42	302,11	362,74	-6,78	250	2,083	23,229
28-févr-15	ENTREPOT4	328,82	330,65	300,31	360,99	-1,83	245	0,558	25,490
28-févr-15	ENTREPOT5	407,78	402,03	371,63	432,44	5,74	305	1,408	25,204
28-févr-15	ENTREPOT6	364,38	357,61	327,26	387,97	6,77	271	1,857	25,627
28-févr-15	ENTREPOT7	398,60	367,81	337,47	398,16	30,78	303	7,723	23,983
28-févr-15	ENTREPOT8	544,42	558,34	527,87	588,81	-13,91	404	2,555	25,793
28-févr-15	ENTREPOT9	561,33	566,49	535,96	597,01	-5,15	420	0,918	25,178
28-févr-15	ENTREPOT10	525,45	517,80	487,42	548,17	7,65	390	1,456	25,777
31-mars-15	ENTREPOT1	343,85	329,08	298,73	359,43	14,77	264	4,295	23,222
31-mars-15	ENTREPOT2	337,70	345,16	314,88	375,45	-7,46	253	2,209	25,082
31-mars-15	ENTREPOT3	342,74	338,41	308,09	368,72	4,34	260	1,265	24,141
31-mars-15	ENTREPOT4	336,23	342,96	312,65	373,28	-6,73	256	2,002	23,863
31-mars-15	ENTREPOT5	443,11	450,63	420,36	480,89	-7,52	332	1,698	25,074
31-mars-15	ENTREPOT6	384,81	394,49	364,21	424,78	-9,69	293	2,517	23,858
31-mars-15	ENTREPOT7	355,87	360,40	330,08	390,71	-4,53	270	1,273	24,129
31-mars-15	ENTREPOT8	559,62	534,54	504,14	564,94	25,08	419	4,482	25,128
31-mars-15	ENTREPOT9	525,59	545,62	515,20	576,03	-20,03	399	3,810	24,085
31-mars-15	ENTREPOT10	535,58	527,31	496,90	557,72	8,27	412	1,544	23,074
30-avr-15	ENTREPOT1	355,39	344,50	314,21	374,78	10,89	266	3,064	25,152
30-avr-15	ENTREPOT2	338,89	346,62	316,31	376,93	-7,73	259	2,282	23,573
30-avr-15	ENTREPOT3	342,95	362,42	332,10	392,74	-19,47	259	5,677	24,479
30-avr-15	ENTREPOT4	354,90	348,63	318,31	378,95	6,27	263	1,768	25,896
30-avr-15	ENTREPOT5	391,91	404,80	373,89	435,72	-12,89	291	3,290	25,748
30-avr-15	ENTREPOT6	364,17	362,64	332,07	393,22	1,53	274	0,419	24,760
30-avr-15	ENTREPOT7	418,73	410,85	380,62	441,08	7,88	309	1,881	26,205
30-avr-15	ENTREPOT8	516,38	506,01	475,49	536,54	10,36	394	2,007	23,699
30-avr-15	ENTREPOT9	519,00	512,86	482,23	543,49	6,14	384	1,183	26,011
30-avr-15	ENTREPOT10	554,91	546,46	515,97	576,95	8,45	427	1,523	23,051
31-mai-15	ENTREPOT1	352,73	350,19	319,87	380,51	2,54	262	0,720	25,722
31-mai-15	ENTREPOT2	354,17	358,82	328,52	389,12	-4,65	266	1,312	24,896
31-mai-15	ENTREPOT3	329,84	341,24	310,94	371,55	-11,40	250	3,457	24,206
31-mai-15	ENTREPOT4	345,66	333,24	302,92	363,56	12,42	265	3,594	23,335
31-mai-15	ENTREPOT5	361,16	356,02	324,74	387,30	5,14	265	1,423	26,625
31-mai-15	ENTREPOT6	458,63	456,06	425,65	486,46	2,57	341	0,561	25,647
31-mai-15	ENTREPOT7	400,98	420,96	390,48	451,44	-19,98	300	4,983	25,183
31-mai-15	ENTREPOT8	561,56	542,09	511,70	572,48	19,47	419	3,468	25,387
31-mai-15	ENTREPOT9	553,04	526,03	495,51	556,56	27,01	415	4,884	24,961
31-mai-15	ENTREPOT10	557,05	546,39	515,87	576,91	10,66	427	1,913	23,346
30-juin-15	ENTREPOT1	357,28	356,73	326,25	387,22	0,54	272	0,152	23,868
30-juin-15	ENTREPOT2	330,44	341,77	311,48	372,06	-11,33	241	3,430	27,066
30-juin-15	ENTREPOT3	349,72	340,43	310,09	370,77	9,29	261	2,657	25,369
30-juin-15	ENTREPOT4	342,58	355,33	324,89	385,77	-12,74	259	3,720	24,398
30-juin-15	ENTREPOT5	403,08	440,61	410,32	470,89	-37,53	300	9,311	25,573
30-juin-15	ENTREPOT6	458,63	414,33	383,99	444,67	44,30	349	9,660	23,904
30-juin-15	ENTREPOT7	453,72	420,21	389,67	450,74	33,51	339	7,386	25,284
30-juin-15	ENTREPOT8	552,22	551,75	521,29	582,21	0,47	418	0,086	24,306
30-juin-15	ENTREPOT9	563,40	541,34	510,41	572,27	22,06	416	3,915	26,162
30-juin-15	ENTREPOT10	542,30	529,44	498,88	560,00	12,86	411	2,371	24,211
31-juil-15	ENTREPOT1	353,97	335,32	305,02	365,63	18,65	263	5,268	25,700
31-juil-15	ENTREPOT2	320,46	356,28	325,98	386,58	-35,81	235	11,175	26,669
31-juil-15	ENTREPOT3	332,77	331,29	300,98	361,60	1,48	248	0,445	25,474
31-juil-15	ENTREPOT4	353,36	345,39	315,08	375,69	7,97	266	2,256	24,722
31-juil-15	ENTREPOT5	388,18	410,18	379,88	440,49	-22,00	295	5,667	24,005
31-juil-15	ENTREPOT6	354,86	359,59	329,32	389,86	-4,73	260	1,334	26,731
31-juil-15	ENTREPOT7	413,00	411,57	381,23	441,91	1,43	304	0,346	26,392
31-iuil-15	ENTREPOT8	535.55	543.87	513.45	574.30	-8.32	409	1.554	23.630

Source : bases de données de l'entreprise BELL Canada dans le Département intelligence d'affaires services extérieurs

date	ENTREPOT	Capacite	Precicted Value	Confident Interval 95%		Error model	Prev_Intuitive	PCT_error_Model	PCT_error_Intuitive
31-août-15	ENTREPOT1	346,23	353,21	322,90	383,52	-6,98	254	2,017	26,638
31-août-15	ENTREPOT2	331,86	337,48	307,15	367,82	-5,62	245	1,695	26,174
31-août-15	ENTREPOT3	357,80	350,73	320,43	381,03	7,08	263	1,978	26,496
31-août-15	ENTREPOT4	334,90	356,35	326,07	386,62	-21,44	245	6,403	26,844
31-août-15	ENTREPOT5	446,44	396,80	366,10	427,49	49,64	340	11,119	23,841
31-août-15	ENTREPOT6	371,39	385,01	354,52	415,50	-13,62	285	3,668	23,261
31-août-15	ENTREPOT7	444,07	446,00	415,62	476,38	-1,93	330	0,435	25,687
31-août-15	ENTREPOT8	512,16	535,17	504,44	565,90	-23,01	381	4,492	25,609
31-août-15	ENTREPOT9	558,36	557,05	526,53	587,57	1,32	427	0,236	23,527
31-août-15	ENTREPOT10	561,85	553,12	522,71	583,54	8,73	421	1,553	25,069
30-sept-15	ENTREPOT1	350,81	359,91	329,60	390,22	-9,10	265	2,594	24,461
30-sept-15	ENTREPOT2	330,64	334,62	304,31	364,92	-3,98	244	1,203	26,204
30-sept-15	ENTREPOT3	326,20	332,93	302,63	363,24	-6,74	240	2,066	26,425
30-sept-15	ENTREPOT4	347,47	347,07	316,77	377,37	0,40	265	0,116	23,735
30-sept-15	ENTREPOT5	454,14	415,41	384,97	445,85	38,74	348	8,529	23,372
30-sept-15	ENTREPOT6	385,12	385,26	354,93	415,59	-0,14	285	0,037	25,997
30-sept-15	ENTREPOT7	374,66	378,72	348,17	409,27	-4,06	275	1,085	26,600
30-sept-15	ENTREPOT8	545,96	532,05	501,51	562,58	13,91	403	2,548	26,184
30-sept-15	ENTREPOT9	501,44	535,24	504,43	566,04	-33,79	369	6,739	26,412
30-sept-15	ENTREPOT10	546,84	531,95	500,97	562,93	14,89	420	2,722	23,195
31-oct-15	ENTREPOT1	358,49	333,24	302,93	363,54	25,25	273	7,043	23,847
31-oct-15	ENTREPOT2	340,38	333,30	302,98	363,62	7,08	260	2,079	23,614
31-oct-15	ENTREPOT3	350,25	362,71	332,27	393,16	-12,46	262	3,559	25,196
31-oct-15	ENTREPOT4	348,82	333,02	302,67	363,36	15,80	256	4,529	26,609
31-oct-15	ENTREPOT5	360,23	365,94	335,42	396,46	-5,71	268	1,585	25,602
31-oct-15	ENTREPOT6	396,53	394,57	363,91	425,23	1,95	290	0,493	26,865
31-oct-15	ENTREPOT7	382,54	374,21	343,19	405,22	8,33	285	2,177	25,497
31-oct-15	ENTREPOT8	528,79	533,98	503,42	564,54	-5,19	400	0,982	24,355
31-oct-15	ENTREPOT9	499,31	527,99	497,53	558,46	-28,69	374	5,746	25,096
31-oct-15	ENTREPOT10	516,99	532,64	502,10	563,18	-15,64	397	3,026	23,210
30-nov-15	ENTREPOT1	354,75	362,65	332,34	392,97	-7,90	259	2,227	26,991
30-nov-15	ENTREPOT2	355,95	351,02	320,73	381,32	4,92	268	1,382	24,708
30-nov-15	ENTREPOT3	359,53	345,22	314,91	375,54	14,31	271	3,979	24,624
30-nov-15	ENTREPOT4	333,42	340,69	310,40	370,98	-7,27	247	2,181	25,920
30-nov-15	ENTREPOT5	406,56	423,70	393,47	453,93	-17,14	298	4,215	26,702
30-nov-15	ENTREPOT6	418,44	392,65	362,15	423,14	25,79	322	6,164	23,048
30-nov-15	ENTREPOT7	370,61	380,13	349,75	410,51	-9,51	282	2,567	23,910
30-nov-15	ENTREPOT8	561,16	559,20	528,70	589,71	1,95	420	0,348	25,154
30-nov-15	ENTREPOT9	564,18	564,04	533,54	594,55	0,14	429	0,025	23,961
30-nov-15	ENTREPOT10	545,45	525,29	494,64	555,93	20,17	403	3,697	26,116
31-déc-15	ENTREPOT1	352,24	332,52	302,21	362,82	19,72	268	5,599	23,915
31-déc-15	ENTREPOT2	337,63	352,67	322,40	382,94	-15,04	253	4,456	25,065
31-déc-15	ENTREPOT3	346,34	338,81	308,47	369,15	7,53	264	2,175	23,775
31-déc-15	ENTREPOT4	341,81	351,44	321,09	381,78	-9,63	260	2,817	23,934
31-déc-15	ENTREPOT5	455,96	451,15	420,88	481,42	4,81	348	1,054	23,677
31-déc-15	ENTREPOT6	432,30	402,48	371,89	433,07	29,82	330	6,899	23,665
31-déc-15	ENTREPOT7	394,71	379,02	348,74	409,30	15,69	291	3,974	26,274
31-déc-15	ENTREPOT8	519,30	532,54	502,10	562,98	-13,23	390	2,548	24,900
31-déc-15	ENTREPOT9	552,84	540,23	509,67	570,80	12,61	422	2,281	23,667
31-déc-15	ENTREPOT10	504,66	528,03	497,42	558,63	-23,37	371	4,630	26,485
31-janv-16	ENTREPOT1	351,28	339,99	309,67	370,31	11,29	262	3,214	25,416
31-janv-16	ENTREPOT2	358,65	340,94	310,63	371,26	17,71	262	4,937	26,948
31-janv-16	ENTREPOT3	338,73	355,03	324,73	385,33	-16,30	257	4,813	24,128
31-janv-16	ENTREPOT4	325,26	327,48	297,10	357,86	-2,21	247	0,680	24,062
31-janv-16	ENTREPOT5	351,26	339,08	308,79	369,37	12,18	258	3,468	26,551
31-janv-16	ENTREPOT6	405,21	412,84	382,44	443,24	-7,63	299	1,883	26,211
31-janv-16	ENTREPOT7	375,10	383,58	353,11	414,05	-8,48	282	2,261	24,820
31-janv-16	ENTREPOT8	503,51	520,72	490,35	551,09	-17,21	375	3,418	25,523
31-janv-16	ENTREPOT9	501,40	505,27	474,81	535,74	-3,87	377	0,773	24,810
31-janv-16	ENTREPOT10	554,87	556,05	525,52	586,59	-1,19	417	0,214	24,847
29-févr-16	ENTREPOT1	347,99	342,18	311,85	372,50	5,81	265	1,670	23,848
29-févr-16	ENTREPOT2	326,19	353,74	323,45	384,02	-27,55	246	8,446	24,583
29-févr-16	ENTREPOT3	342,00	345,47	315,15	375,78	-3,47	257	1,014	24,854
29-févr-16	ENTREPOT4	345,76	351,13	320,84	381,42	-5,37	254	1,554	26,538
29-févr-16	ENTREPOT5	376,81	401,08	370,81	431,35	-24,26	278	6,440	26,223
29-févr-16	ENTREPOT6	417,70	410,33	379,93	440,72	7,37	311	1,764	25,544

29-févr-16	ENTREPOT7	444,04	441,64	410,98	472,30	2,40	326	0,540	26,583
29-févr-16	ENTREPOT8	509,45	521,74	491,24	552,24	-12,30	386	2,414	24,232
29-févr-16	ENTREPOT9	513,43	515,15	484,77	545,53	-1,72	387	0,335	24,625
29-févr-16	ENTREPOT10	507,78	510,32	479,97	540,67	-2,54	371	0,501	26,936
31-mars-16	ENTREPOT1	339,82	359,26	328,95	389,57	-19,44	251	5,721	26,137
31-mars-16	ENTREPOT2	347,88	358,56	328,26	388,87	-10,68	260	3,071	25,262
31-mars-16	ENTREPOT3	344,17	347,15	316,86	377,44	-2,98	253	0,866	26,490
31-mars-16	ENTREPOT4	331,03	351,23	320,96	381,51	-20,20	251	6,104	24,176
31-mars-16	ENTREPOT5	380,36	391,97	361,06	422,88	-11,61	287	3,053	24,545
31-mars-16	ENTREPOT6	409,20	396,84	366,26	427,42	12,36	302	3,021	26,198
31-mars-16	ENTREPOT7	414,58	397,65	366,53	428,78	16,93	305	4,083	26,432
31-mars-16	ENTREPOT8	530,23	536,99	506,26	567,73	-6,77	404	1,276	23,807
31-mars-16	ENTREPOT9	555,31	557,44	527,00	587,88	-2,13	416	0,384	25,087
31-mars-16	ENTREPOT10	530,86	528,40	497,60	559,20	2,46	397	0,464	25,216
30-avr-16	ENTREPOT1	340,00	344,18	313,87	374,49	-4,18	249	1,230	26,764
30-avr-16	ENTREPOT2	339,57	342,66	312,26	373,07	-3,10	255	0,912	24,904
30-avr-16	ENTREPOT3	348,61	352,88	322,55	383,21	-4,28	267	1,226	23,409
30-avr-16	ENTREPOT4	331,81	341,50	311,19	371,81	-9,69	248	2,920	25,259
30-avr-16	ENTREPOT5	353,55	340,48	310,16	370,79	13,07	260	3,696	26,459
30-avr-16	ENTREPOT6	382,72	402,99	372,62	433,37	-20,27	280	5,297	26,840
30-avr-16	ENTREPOT7	402,37	413,90	383,41	444,40	-11,53	297	2,867	26,187
30-avr-16	ENTREPOT8	534,77	533,27	502,61	563,93	1,50	406	0,280	24,079
30-avr-16	ENTREPOT9	522,84	530,51	500,05	560,97	-7,67	384	1,467	26,555
30-avr-16	ENTREPOT10	522,71	526,89	496,46	557,32	-4,18	391	0,800	25,198
31-mai-16	ENTREPOT1	331,57	343,89	313,57	374,21	-12,32	245	3,716	26,109
31-mai-16	ENTREPOT2	337,11	336,47	306,16	366,79	0,64	255	0,191	24,358
31-mai-16	ENTREPOT3	333,73	353,36	323,06	383,65	-19,63	255	5,882	23,590
31-mai-16	ENTREPOT4	325,73	325,45	295,11	355,79	0,28	244	0,087	25,092
31-mai-16	ENTREPOT5	450,40	425,78	395,53	456,03	24,62	333	5,465	26,065
31-mai-16	ENTREPOT6	362,78	349,25	318,90	379,61	13,53	271	3,729	25,300
31-mai-16	ENTREPOT7	404,11	415,47	385,21	445,73	-11,36	296	2,812	26,752
31-mai-16	ENTREPOT8	517,60	522,34	491,77	552,92	-4,75	389	0,917	24,845
31-mai-16	ENTREPOT9	560,88	553,73	523,05	584,40	7,16	425	1,276	24,226
31-mai-16	ENTREPOT10	505,98	516,69	486,33	547,06	-10,71	381	2,117	24,701
30-juin-16	ENTREPOT1	331,47	347,26	316,94	377,57	-15,79	246	4,764	25,784
30-juin-16	ENTREPOT2	343,01	355,79	325,35	386,23	-12,78	259	3,727	24,491
30-juin-16	ENTREPOT3	359,52	338,72	308,42	369,02	20,80	264	5,785	26,568
30-juin-16	ENTREPOT4	322,01	328,15	297,79	358,51	-6,14	247	1,906	23,295
30-juin-16	ENTREPOT5	409,70	435,40	405,09	465,71	-25,70	313	6,274	23,602
30-juin-16	ENTREPOT6	433,33	434,30	404,05	464,54	-0,96	331	0,223	23,615
30-juin-16	ENTREPOT7	410,85	430,33	400,06	460,60	-19,48	304	4,741	26,007
30-juin-16	ENTREPOT8	523,37	537,97	507,06	568,89	-14,61	399	2,791	23,763
30-juin-16	ENTREPOT9	501,62	508,46	477,87	539,06	-6,85	380	1,365	24,245
30-juin-16	ENTREPOT10	509,85	528,75	498,36	559,15	-18,91	377	3,708	26,056
31-juil-16	ENTREPOT1	322,34	326,77	296,38	357,16	-4,43	242	1,375	24,924
31-juil-16	ENTREPOT2	358,69	355,17	324,88	385,46	3,52	263	0,981	26,677
31-juil-16	ENTREPOT3	320,55	330,46	300,14	360,78	-9,92	244	3,094	23,880
31-juil-16	ENTREPOT4	345,66	333,16	302,84	363,49	12,50	263	3,617	23,915
31-juil-16	ENTREPOT5	358,54	340,83	310,53	371,14	17,71	264	4,938	26,368
31-juil-16	ENTREPOT6	400,71	387,09	356,40	417,78	13,62	300	3,399	25,133
31-juil-16	ENTREPOT7	409,54	400,52	370,28	430,75	9,02	305	2,202	25,526
31-juil-16	ENTREPOT8	502,01	511,45	481,07	541,82	-9,44	371	1,880	26,097
31-juil-16	ENTREPOT9	514,79	534,72	504,18	565,26	-19,94	388	3,873	24,629
31-juil-16	ENTREPOT10	508,15	516,50	486,03	546,98	-8,36	384	1,645	24,431
31-août-16	ENTREPOT1	328,28	329,42	299,05	359,78	-1,14	245	0,347	25,368
31-août-16	ENTREPOT2	348,34	363,28	332,97	393,60	-14,94	256	4,290	26,508
31-août-16	ENTREPOT3	355,12	338,99	308,61	369,37	16,12	267	4,541	24,814
31-août-16	ENTREPOT4	355,75	359,45	329,17	389,72	-3,70	268	1,040	24,665
31-août-16	ENTREPOT5	448,00	416,75	386,46	447,03	31,26	331	6,977	26,116
31-août-16	ENTREPOT6	373,63	379,13	348,55	409,72	-5,50	276	1,472	26,131
31-août-16	ENTREPOT7	404,29	425,24	394,33	456,14	-20,94	301	5,180	25,549
31-août-16	ENTREPOT8	509,08	512,70	482,34	543,05	-3,62	392	0,710	22,999
31-août-16	ENTREPOT9	531,07	539,45	509,06	569,84	-8,38	403	1,579	24,115
31-août-16	ENTREPOT10	555,43	556,13	525,51	586,76	-0,70	414	0,126	25,464
30-sept-16	ENTREPOT1	321,11	351,98	321,63	382,33	-30,87	247	9,613	23,080
30-sept-16	ENTREPOT2	347,18	359,76	329,41	390,10	-12,57	263	3,621	24,247
30-sept-16	ENTREPOT3	350,18	334,30	303,97	364,63	15,88	260	4,535	25,752
30-sept-16	ENTREPOT4	359,07	362,65	332,35	392,94	-3,58	265	0,996	26,198
30-sept-16	ENTREPOT5	353,02	349,81	319,53	380,10	3,21	259	0,909	26,634
30-sept-16	ENTREPOT6	435,00	442,59	412,30	472,89	-7,60	334	1,747	23,218
30-sept-16	ENTREPOT7	401,07	386,59	356,22	416,96	14,49	301	3,612	24,952
30-sept-16	ENTREPOT8	567,05	563,76	533,19	594,32	3,29	425	0,580	25,050
30-sept-16	ENTREPOT9	527,83	538,52	507,97	569,06	-10,69	397	2,025	24,786
30-sept-16	ENTREPOT10	563,27	547,07	516,34	577,80	16,20	432	2,876	23,305
31-oct-16	ENTREPOT1	357,69	337,24	306,93	367,55	20,45	275	5,717	23,118
31-oct-16	ENTREPOT2	341,60	352,62	322,29	382,95	-11,02	254	3,227	25,643
31-oct-16	ENTREPOT3	342,99	348,17	317,86	378,48	-5,19	262	1,512	23,612
31-oct-16	ENTREPOT4	333,92	353,72	323,39	384,05	-19,80	255	5,931	23,634
31-oct-16	ENTREPOT5	354,77	352,48	322,17	382,80	2,29	271	0,645	23,613
31-oct-16	ENTREPOT6	393,64	377,31	346,79	407,83	16,33	303	4,148	23,026
31-oct-16	ENTREPOT7	456,64	406,26	375,97	436,55	50,38	339	11,032	25,762
31-oct-16	ENTREPOT8	558,32	530,08	499,30	560,86	28,24	409	5,059	26,745
31-oct-16	ENTREPOT9	536,69	527,32	496,86	557,78	9,36	404	1,745	24,723
31-oct-16	ENTREPOT10	547,90	542,32	511,69	572,95	5,58	412	1,018	24,804

Annexe XV : Comparaison des valeurs d'estimation par les méthodes formelles et intuitives par rapport aux besoins en espace de stockage



Annexe XVI : Modèle à effet individuel en langage Python

```
#--------SCRIP PYTHON-------------------------------------

#AUTHOR: JEAN BAPTISTE LALA
#DESCRIPTION: Model with entity fixed effects
#-----------------------------------------------------------

import pandas as pd
from numpy import log

#Read Data
df=pd.read_excel('Data_essai_definitive_trie.xlsx')
df.tail()

#Extract the analysis variables
DATAS = log(df[[3,4,5,7]].values)
print(DATAS)

#Get colum names for analysis variables
cols=df[[3,4,5,7]].columns
for c in cols:
    print(c)

#define time dimension and entity dimension
Temps_individu=[df['date'],df['ENTREPOT']]
#print(Temps_individu)

#Input Panel data analisis using multindex
Panel_Datas = pd.DataFrame(DATAS ,index=pd.MultiIndex.from_tuples(list(zip(*Temps_individu))),columns=cols)
#print(Panel_Datas)

#--Fit panel data analysis data model

from pandas.stats.plm import PanelOLS
#1- Model with entity fixed effects
Entity_Fixed_Effects_Model  = PanelOLS(y=Panel_Datas['Capacite'],x=Panel_Datas[['TailleTblFact','TailleTblDim','FreqRet']],intercept = False,entity_effects=True)
print(Entity_Fixed_Effects_Model)
```