الجمـــهورية الجزائريـــة الديمقراطيـــة الشعبيـــة
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur Et de la Recherche Scientifique
Universite 8 Mai 1945 Guelma
Faculté des Mathématiques, de l'Informatique et des Sciences de la Matière
Département : Informatique

وزارة التعليم العالـي والبحث العلمـي
جامعـة 8 مـــاي 1945 قالمـــة

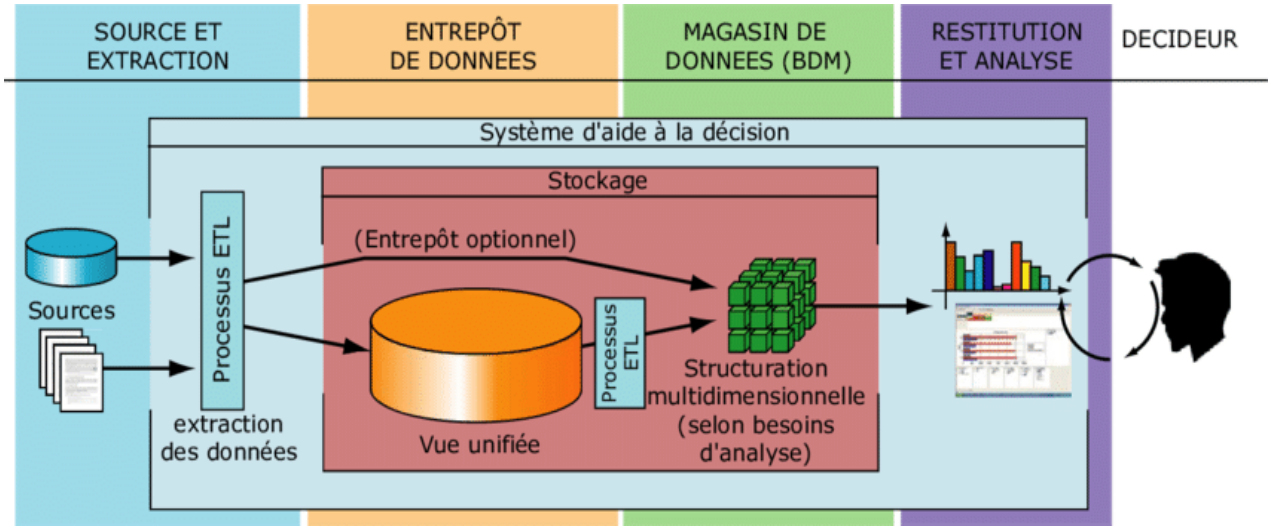
كليـة الرياضيـات، الإعـلام الآلـي وعلـوم المـادة

قسم : الإعلام الآلي

Polycopié de cours :

Business Intelligence (Informatique Décisionnelle)

Cours & Exercices



Destiné aux étudiants en 3ième année licence informatique
Spécialité : Ingénierie des Systèmes d'informations et du Logiciel (ISIL)

Année : 2022

Dr. KHEBIZI Ali

Préambule

L'accroissement spectaculaire de la connectivité, conjugué avec la démocratisation de l'utilisation des technologies de l'information et de la communication (TIC) ont complètement bouleversé les modes de fonctionnement des organisations modernes et les pratiques de consommation des personnes. En effet, à l'heure du numérique, les systèmes d'information (**S.I**) sont devenus omniprésents et les transactions commerciales sont de plus en plus automatisées. La conséquence immédiate de ce constat est l'explosion de la masse de données capturées au quotidien, aussi bien par les différents SI que par la variété des capteurs utilisés (*smartphones, serveurs web, serveur de messagerie, réseaux sociaux…etc*).

D'un point de vue stratégique, l'exploitation efficace de la mine d'informations collectées et l'analyse des données stockées peuvent être utiles pour assister les systèmes de pilotage des entreprises dans leurs prises de décisions. C'est l'objectif fondamental de l'informatique décisionnelle (*Business Intelligence ou BI*). En ce sens, la Business Intelligence a pour but de produire, à partir de la masse de données générées par les processus métiers de l'entreprise et captées par son S.I, de nouvelles informations pertinentes dont l'exploitation peut aider les cadres et les dirigeants dans le pilotage de leur entreprise.

De manière simpliste, la BI désigne l'ensemble de méthodes, de techniques et des outils informatiques pour aider à la prise de décision au niveau des organisations. Elle permet de délivrer, à chaque manager, les informations pertinentes afin qu'il puisse prendre le plus efficacement possible les meilleures décisions selon son contexte d'action, ses prérogatives et ses objectifs tactiques et stratégiques. Ces différents éléments constituent le sujet traité dans le présent polycopié de cours.

Ce polycopié pédagogique de cours intitulé « *Business Intelligence (Informatique Décisionnelle***)** Cours & Exercices » est destiné aux étudiants en informatique et il est conforme au programme de la formation proposée pour les deux parcours de la 3^{ième} année licence informatique en Systèmes Informatiques (**SI**) et Ingénierie des Systèmes d'Information et du Logiciel (**ISIL**). Par ailleurs, son exploitation peut s'étendre aux gestionnaires utilisateurs (*directeurs, cadres dirigeants, cadres supports, utilisateurs finaux,…etc.*) qui désirent s'initier et se familiariser avec les outils d'intégration des données et de prise de décision en entreprise, par l'exploitation des données opérationnelles. D'autre part, il constitue une bonne référence pour tout professionnel du traitement de l'information et du développement des systèmes décisionnels. Le polycopié est structuré de façon pédagogique qui permet, dans un premier temps, au lecteur de s'initier aux systèmes d'informations de gestion et de prendre conscience des problèmes liés à la diversité des sources de données et de leur hétérogénéité, puis de mettre en exergue le besoin et l'intérêt de centraliser les données manipulées, de les structurer et de les intégrer dans une seule structure de référence (*Entrepôt De Données (**EDD**) ou data warehouse*), en vue de leur exploitation future dans le cadre de la prise de décision. La modélisation multidimensionnelle des données et l'élaboration de l'entrepôt de données constitue le socle de l'architecture sous-jacente à la solution BI. Pour consolider la démarche de développement d'un projet BI, une présentation des outils conceptuels utiles à la phase de modélisation est présentée et illustrée par des études de cas réels. Enfin, un panorama de quelques outils commerciaux les plus utilisés est exposé en fin du polycopié.

1. Objectifs visés

Dans ce cours, le lecteur apprendra les principes et les techniques de base de la Business Intelligence et comment les appliquer pour atteindre les objectifs de performance en entreprise. Il introduit un ensemble de connaissances théoriques et pratiques permettant de

transformer les flots de données de gestion en informations exploitables qui aident les décideurs à prendre les meilleures décisions commerciales, concurrentielles et celles qui sont utiles à l'amélioration des produits et services, tout en garantissant un degré satisfaisant de satisfaction des clients.

L'enseignement de ce cours vise à définir les besoins, les enjeux et l'utilisation de la BI dans le monde de l'entreprise. Il fait découvrir aux lecteurs les principaux concepts de l'informatique décisionnelle, leur permet d'acquérir les compétences nécessaires pour pouvoir aborder l'intégration des données hétérogènes issues de différentes sources en vue de pouvoir les modéliser, de construire et d'exploiter les entrepôts de données (*EDD* ou *Data warehouse*) et enfin de savoir créer des tableaux de bord ergonomiques.

Le polycopié s'articule autour de quatre axes principaux. On commence par une présentation de la BI et les problématiques liés à l'exploitation des données massives et hétérogènes. Puis, les techniques d'intégration des données sont exposées et illustrées par des études de cas réels. Une fois, les données sont centralisées et intégrées, l'architecture et le fonctionnement des entrepôts des données sont étudiés et les différentes configurations possibles sont examinées. Enfin, la modélisation multidimensionnelle des entrepôts de données est abordée et quelques outils logiciels OLAP du marché sont exposés.

Après avoir étudié ce cours, le lecteur devrait montrer les compétences suivantes :

- Comprendre les enjeux liés à la diversité des sources de données en entreprise et prendre conscience du besoin de leur intégration dans une seule structure homogène.
- Etre capable d'analyser les sources de données existantes, d'en extraire le contenu, puis d'appliquer les transformations adéquates en vue de comprendre leurs corrélations et leurs contraintes pour les intégrer dans l'entrepôt de données.
- Comprendre l'architecture et le fonctionnement des entrepôts de données.
- Aborder la modélisation multidimensionnelle et élaborer le schéma conceptuel de futur entrepôt de données.
- Pouvoir identifier les indicateurs clés de performance (**K**ey **P**erformance **I**ndicators **KPI**) et être en mesure de proposer un tableau de bord pour plus de performances de l'organisation.
- Conduire un projet de BI et exploiter les outils logiciels existants dans le marché.

2. Structure du polycopié

Le manuscrit est composé de quatre chapitres dont chacun est consacré à un aspect particulier du domaine de la BI. A la fin de chaque chapitre, une série d'exercices est proposée aux étudiants pour leur permettre de consolider et d'approfondir les connaissances théoriques acquises dans un aspect spécifique lié aux concepts, architectures et modèles proposés dans le cours. Les exercices proposés sont souvent des scénarios réels permettant d'aborder des problématiques concrètes liées à la compréhension et à la conception des solutions BI.

Nous commençons par **le chapitre 1** intitulé *« **de l'Information à la décision*** » qui met en exergue l'intérêt de l'intégration des données générées lors de l'exécution des processus métiers de l'entreprise dans une structure commune, en vue d'en extraire des connaissances utiles à la prise de décisions. Ainsi, le contexte et les objectifs d'une solution BI seront clairement décrits. Par la suite, les différentes définitions de la BI sont présentées et discutées et les concepts afférents et nécessaires à la compréhension du reste du manuscrit y seront introduits. Le chapitre se termine par l'exposé des relations existantes entre la BI et autres domaines connexes, tels que les logiciels de gestion intégrée ou technologie ERP (*Enterprise Resource Planning*), celui de la fouille de données (*Data mining*) ou le domaine des données volumineuses (*Big data*).

Le chapitre 2 ***« Architecture et fonctionnement d'une solution BI »*** est consacré à la

description fonctionnelle d'une architecture BI. On y exposera le fonctionnement des différents composants de l'architecture BI ainsi que leurs interactions. Les fonctions assurées par le processus BI seront clairement définies et illustrées par des exemples réels et une attention particulière sera accordée à la technique **ETL** (*Extract-Transform and Load*), tout en se focalisant sur les mécanismes de migration et d'intégration des données.

Les deux premiers chapitres permettent d'introduire la matière utile à la compréhension du mécanisme de fonctionnement d'une solution BI. Il apparait, par la suite, que la centralisation des données dans une structure commune constitue l'épine dorsale du futur système décisionnel, d'où l'intérêt de se focaliser sur le futur entrepôt de données représentant le noyau de la future solution BI à concevoir. Cette préoccupation fera l'objet des deux chapitres suivants.

Dans **le chapitre 3** intitulé « ***Les entrepôts de données*** » on s'étale longuement sur l'étude des entrepôts de données, avec un examen approfondi de leurs avantages et de leurs caractéristiques. Un panorama des différentes configurations possibles est exposé et chacune des architectures est évaluée. En plus, une attention particulière sera accordée aux magasins de données.

Le chapitre 4 « ***Modélisation multidimensionnelle de l'entrepôt de données et outils OLAP*** » est dédié à la conception des entrepôts de données. On y introduira les fondements du modèle multidimensionnel et son formalisme, suivis des différentes approches de modélisation permettant d'aboutir à un schéma abstrait spécifiant le schéma du futur entrepôt de données. Divers scénarios illustratifs relatifs à des domaines d'application variés sont présentés pour consolider les concepts liés à la modélisation de l'entrepôt de données. Par ailleurs, ce chapitre examine le fonctionnement des outils OLAP utiles à l'exploitation des entrepôts de données et il est clôturé par une brève présentation de quelques logiciels OLAP les plus répandus sur le marché.

A la fin de chaque chapitre, une série d'exercices pratiques est présentée aux lecteurs pour leur permettre de consolider les notions les plus importantes et de mettre en pratique certains aspects théoriques acquis dans le chapitre. Ces exercices sont, soit des retours sur certains concepts et définitions dont la maîtrise est primordiale pour la compréhension de la BI, soit des études de cas réels qui permettent de renforcer les connaissances théoriques acquises dans le chapitre en question. Nous terminons le polycopié par un annexe contenant les solutions détaillées des séries des exercices proposés dans chaque chapitre.

Répartition du volume horaire semestriel sur les chapitres du module.

Chapitre	Intitulé du chapitre	Nombre de semaines
Chapitre I	***de l'Information à la décision***	4 semaines
Chapitre II	***Architecture et fonctionnement d'une solution BI***	4 semaines
Chapitre III	***Les Entrepôt de données (EDD)***	3 semaines
Chapitre IV	***La Modélisation multidimensionnelle des entrepôts de données et outils OLAP***	4 semaines

Enfin, quelques références bibliographiques sont données à la fin du polycopié.

Table des matières

Chapitre I : de l'Information à la décision

1. Introduction

La quantité incroyable de données produites et collectées par les différents systèmes informatiques représente aujourd'hui une mine d'or dont l'exploitation rationnelle offre une valeur ajoutée pour toute organisation. En effet, l'Internet des objets a ouvert la voie à l'apparition de nouvelles sources d'informations distribuées et hétérogènes. Ces sources varient des bases de données (**BDD**) relationnelles, les bases de données semi-structurées et celles non structurées, aux traces d'exécutions et données mobiles. Ainsi, la masse d'informations stockées et archivées sur des supports informatiques ne cesse de croitre de manière exponentielle. Elle constitue, aujourd'hui, un capital dont l'exploitation permettra d'assurer la pérennité et la survie de toute organisation évoluant dans un contexte de plus en plus concurrentiel et où l'environnement est fortement mouvant. Dans cette perspective, l'analyse et l'exploitation des données peuvent être utiles pour assister les entreprises dans leurs prises de décisions et constituent un facteur clé de leur réussite.
Dans une optique d'exploitation de la masse de données produite par les différents SI, dans un premier temps, une mise en relation des données stockées s'avère impérative. En effet, la prise en compte de la diversité des formats de données manipulées, les différentes considérations et contraintes technologiques liées aux plateformes et aux systèmes d'exploitation exigent, impérativement, une consolidation et une intégration qui aboutissent à un réservoir de données unique qui sera facilement exploitable. En ce sens, pour collecter, intégrer et analyser les données émanant de différentes sources, il est impératif d'utiliser une large variété d'outils et de technologies. C'est l'objectif de l'*informatique décisionnelle* ou *Business Intelligence* **(BI)**.

Ce premier chapitre est dédié à la présentation de la BI. Nous commençons par présenter le contexte d'utilisation d'une solution BI et les objectifs qu'elle vise à réaliser. Puis, nous exposerons les différentes définitions de la BI et les concepts qui lui sont associés. Une mise en relation de la BI avec d'autres domaines connexes, tels que les progiciels ERP, le data mining et le Big data est dressée et une comparaison de ces techniques avec la solution BI est exposée. Nous terminerons le chapitre par une conclusion.

2. Contexte de la Business Intelligence

De nos jours les systèmes d'informations de gestion sont omniprésents et il fréquent de voir la grande majorité des entreprises dotée de logiciels permettant la prise en charge et la gestion automatisée de toutes les fonctions de gestion. Néanmoins, le contexte et l'environnement actuel dans lesquels évolue toute organisation sont caractérisés par les deux aspects fondamentaux suivants.

2.1 Sur le plan économique

Les phénomènes de mondialisation et de globalisation de l'économie ont ouvert la voie à un marché mondial unique ou la concurrence est de plus en plus rude. Comme conséquence immédiate de cet environnement concurrentiel, la survie de toute entreprise est devenue un enjeu majeur. Ainsi, il est observé aujourd'hui une évolution permanente des entités organisationnelles, où les actions d'acquisition, de rachat et de fusion d'entreprises sont de plus en plus fréquentes. D'autre part, les changements qui surgissent dans le micro et macro environnements de l'entreprise, tels que les dimensions socioculturelle, écologique, politique…etc. sont devenus quasi-permanents. Ainsi, toute organisation doit, impérativement, cerner et maitriser les variables qui surgissent dans son environnement afin d'assurer sa pérennité, toute en étant flexible et en s'adaptant aux mutations et aux changements.

2.2 La dimension technologique

L'évolution technologique récente, caractérisée par la démocratisation de l'utilisation de l'Internet, l'augmentation permanente du débit ainsi que l'émergence spectaculaire des technologies de l'information et de la communication (**TIC**), a complètement bouleversé et accéléré les processus économiques, tout en intensifiant les transactions et les échanges entre les différents partenaires. En effet, aujourd'hui le secteur du numérique a impacté la grande partie de la sphère socio-économique (*e-commerce, e-gouvernance, …etc.*) et il s'étend à tous les autres secteurs d'activités. Ainsi, on constate de nos jours que les activités économiques et sociales sont soutenues et gérées par des plates-formes électroniques dédiées, telles que les sites de commerce en ligne, les réseaux sociaux, les réseaux mobiles ou les réseaux de capteurs.

2.3 Pourquoi l'informatique décisionnelle ?

Pour faire face aux enjeux économiques et aux défis technologiques cités ci-dessus, les entreprises modernes déploient des moyens colossaux et des ressources considérables pour cerner ces phénomènes, tout en étant compétitives et concurrentielles. La pierre angulaire des solutions à adopter réside dans le développement intensif des systèmes d'informations de gestion permettant de prendre en charge les différentes fonctions de l'entreprise, telles que les fonctions approvisionnement, commerciale, finances, gestion des ressources humaines …etc. Dans cette perspective, les S.I qui assurent la collecte, le stockage, le traitement et la diffusion des informations de gestion sont devenus omniprésents. Souvent, ces systèmes offrent au système de pilotage de l'entreprise des informations utiles à la prise de décision tactique et stratégique. Néanmoins, la diversité des données issues des sources variées et incompatibles, telles que les téléphones mobiles ou autres capteurs, permet de collecter des données hétérogènes dont l'exploitation devient de plus en plus complexe et problématique, voire impossible dans certaines situations. D'autre part, la surinformation induite par la multiplicité des sources de données complique d'avantage l'exploitation efficace des informations collectées par l'organisation.

L'*informatique décisionnelle* ou BI tente de répondre à ces préoccupations et à surmonter ces difficultés. C'est une discipline en plein évolution qui se situe au carrefour des S.I, des processus métiers et des stratégies de la direction générale de l'entreprise. Du point de vue technologique, elle consiste en une interface entre les bases de données et les métiers de l'entreprise qui offre aux gestionnaires et décideurs un tableau de bord utile pour l'aide à prise de décision.

En résumé, toute solution BI offre les avantages suivants :

- Faciliter la prise de décision par le traitement des données, ce qui a un impact direct sur la mise en place d'une stratégie efficiente.
- Elle permet aux différentes structures de l'entreprise d'identifier les tendances actuelles du marché et d'adopter des plans d'action plus adéquats et plus performants.
- Grâce aux outils de visualisation des données (*Data visualization*), une véritable accélération et amélioration de la prise de décisions est atteinte.
- Des modules dédiés de la solution BI prennent en charge la création de modèles de données graphiques (*nuage de points, histogramme, courbe statistique, etc.*) qui sont faciles à interpréter et à analyser. En disposant de ce genre d'applications, il n'est pas nécessaire d'engager un consultant BI.
- Puisque les données sont stockées sur le Cloud, l'analyste sera en mesure de parcourir les données n'importe où et n'importe quand, ce qui consolidera la position du décideur en tant qu'utilisateur nomade.

Néanmoins, il faut signaler que le prix à payer concerne l'investissement financier engagé pour la mise en place de toute solution BI et que, souvent, certains logiciels proposés en version gratuite sont limités en matière de fonctionnalités.

3. Objectifs de la Business Intelligence : système d'aide à la décision

L'analyse des données collectées par les S.I de l'entreprise est très utile pour assister les décideurs dans leurs prises de décisions. Les besoins en matière de gestion de l'information résident, alors, dans le processus de consolidation des données collectées et de leur analyse, dans le but d'avoir une vision globale de l'organisation, d'optimiser le patrimoine informationnel de l'entreprise et de prendre les bonnes décisions aux moments opportuns.
Ainsi, on passe d'une perspective de gestion de l'information à celle de la prise de décision. Donc, il ne s'agit pas de recueillir l'information, mais de l'exploiter en vue d'en extraire une plus-value. La réalisation de cet objectif permettra de rendre disponible l'information utile et pertinente sous la bonne forme, au bon moment et à la bonne personne afin de l'exploiter et d'en extraire de la valeur ajoutée. En définitif, l'objectif d'une solution BI est de réaliser les buts suivants :

- Offrir une vue de haut niveau sur l'efficacité opérationnelle de l'entreprise,
- Simplifier les processus de prise de décision internes,
- Permettre d'identifier les points à améliorer dans l'entreprise.

L'objectif de la BI est de créer à partir des données de l'entreprise, mais aussi externes à celle-ci, l'information et le savoir pour aider les cadres et les dirigeants dans le pilotage de l'entreprise.

La *table 1*, ci-dessous récapitule les différentes préoccupations des gestionnaires d'une entreprise commerciale et permet de situer clairement les objectifs de la BI pour le domaine commercial.

N°	Niveau de Préoccupation	Types de besoins	Exemple de questions
1	Opérationnelle	Que se passe-t-il en ce moment ?	• Quel est le total des ventes de la journée ? • Quelle est la valeur des produits périmés ?
2	Historique	Que s'est-il passé ?	• Quel sont les clients qui ont acheté le produit A et le produit B ? • Quel est l'historique des achats d'un client ?
3	Analytique	Pourquoi est-ce que cela s'est passé ?	• Quels étaient les meilleurs clients pour l'année précédente ? • Quel est le point de vente dont le montant total des ventes n'a pas atteint une valeur X pendant les deux dernières années ?
4	*Pronostique*	*Que va-t-il se passer ?*	• Le point de vente A aurait-il besoin d'un sur-stockage du produit B en période estivale ? • Est-ce que les clients forts méritent-ils une remise considérable ?
5	*Décisionnelle*	*Comment agir ?*	• Quelles sont les quantités à sur-stocker dans le point de vente A ? • Quels taux de remise accorder pour chaque catégorie de clients (*forts, faibles, moyens*) ? • Comment cibler la clientèle ? Quel est l'évolution d'un produit ?

Table 1.1 Les différentes préoccupations et les types de besoins de la BI

La lecture du tableau fait ressortir que la BI est conçue pour les décideurs et les dirigeants d'entreprises et elle leur offre la possibilité d'avoir une *vue d'ensemble complète et claire* de la totalité des données disponibles. Pour réaliser les objectifs escomptés la BI doit, impérativement, regrouper tous les moyens, outils et méthodes pour collecter, consolider et modéliser les données de l'entreprise à partir de plusieurs sources.

4. Définitions et concepts de l'informatique décisionnelle

Après avoir introduit le contexte et les objectifs de la BI, nous présentons dans cette section quelques définitions de la BI et nous exposons quelques domaines phares de son application, puis nous introduisons les concepts et les notions de base qui lui sont associés et qui seront utiles à la compréhension du reste du polycopié.

4.1 Définitions de la BI

Plusieurs définitions ont été attribuées dans la littérature à la notion de BI. Nous exposons ci-dessous quelques-unes et nous dégageons la plus pertinente et la plus significative, à notre sens, et qui sera adoptée dans le cadre de ce polycopié.

a. **Définition 1**

Le terme Business Intelligence (BI), ou informatique décisionnelle, désigne les applications, les infrastructures, les outils et les pratiques offrant l'accès à l'information, et permettant d'analyser l'information pour améliorer et optimiser les décisions et les performances d'une entreprise.

b. **Définition 2**

La Business Intelligence est le processus d'analyse de données dirigé par la technologie dans le but de déceler des informations utilisables pour aider les dirigeants d'entreprises et autres utilisateurs finaux à prendre des décisions plus informées.

c. **Définition 3**

La Business Intelligence se définit par l'ensemble des moyens, méthodes et outils qui supportent le processus de collecte, de consolidation, de modélisation, d'analyse et de restitution des informations.

Bien que les trois définitions précédentes sont exprimées et formulées de manières plus ou moins différentes, elles mettent en exergues les points communs suivants :

- *Exploitation des données collectées* (*consolidation, modélisation et analyse*).
- Par **l'***utilisation de la technologie* (*moyens, applications, méthodes, outils*).
- En vue d**'***améliorer et d'aider les dirigeants à prendre et optimiser les décisions*.

En se basant sur ces trois points communs, nous retenons la définition suivante de la BI.

d. Définition retenue :

La BI exploite la masse de données générées par les processus métiers de l'entreprise et captées par son système d'informations *et utilise l'ensemble des moyens, outils informatiques et méthodes techniques pour aider à la prise de décision et d'avoir une vue d'ensemble de l'activité traitée.*

4.2 Les domaines d'application de la BI

En se référant à la définition précédente, la BI peut être utilisée par tous les processus métiers de l'entreprise. En effet, du fait que les données relatives aux activités de l'entreprise sont partout et concernent tous les secteurs d'activités de l'entreprise comme les achats, le stock, le commerciale, la production les finances ou encore la gestion des ressources humaine et la production, alors ces domaines peuvent constituer des champs très fertiles à l'application des techniques de la BI, en vue de les analyser et les faire parler afin de prendre des décisions éclairées.

La *Table 2* suivante illustre clairement l'utilité de la BI par des scénarios réels relatifs aux différentes fonctions d'une entreprise commerciale. La dernière colonne du tableau fait ressortir les indicateurs clés de performance (Key Performance Indicators ou **KPI**) qui sont associés à chaque fonction et dont la connaissance constitue un élément primordial pour la prise de décisions. Ces indicateurs répondent à des préoccupations de gestions, telles que mentionnées dans la *colonne 2* du même tableau, et ils sont extraits à partir d'une analyse poussée (*colonne 3*) des données collectées par les différentes applications de gestion de l'entreprise.

Fonction	Préoccupations de gestion	Actions à mener	indicateurs clés de performance utiles à la prise de décision
Vente et commerciale	▪ Identifier les meilleurs points de ventes, ▪ Cibler les meilleures périodes de lancement des remises.	▪ Analyse des points de vente	▪ Montants des ventes pour toutes les dimensions (*produits, clients, équipes, promotions, emplacement*s) ; ▪ Produits les plus vendus/ périmés ; ▪ Jours avec le maximum de ventes; ▪ Produits échangés/retournés.
Marketing	▪ Etudier les impacts des promotions, ▪ Quels sont les préférences et les comportements des clients ?	▪ Analyse des comportements ▪ Analyse des compagnes marketing	▪ Paniers moyens par type de clients (*Traffic et Fréquence d'achat*). ▪ Panier moyen avec et sans action marketing ▪ Performance campagnes marketing / programme de fidélité ▪ Tendances et prévisions des ventes.
Finances	▪ Analyser le chiffre d'affaire (**CA**) réalisé. ▪ Gestion de la trésorerie, ▪ Quels sont les impacts des placements financiers ?	▪ Reporting financier ▪ Analyse budgétaire ▪ Mesure des coûts, ▪ Estimation des risques. ▪ Analyse des pertes, gaspillages	▪ CA détaillé par dimension (point de vente, marque, pays, canal de vente) ▪ CA cumulé par dimension *(semaine, point de vente, produit …*) ▪ Coût total des produits périmés. ▪ Perte par produit / point de vente.
Logistique	▪ Comment optimiser la gestion du stock et suivre les livraisons et les achats.	▪ Monitoring du stock, ▪ Analyse des coûts de stockage	▪ Valeur du stock avec précision, ▪ Produits les plus performants, ▪ Produits en situations de rupture de stock ; ▪ coûts d'entreposage.
Ressources humaines	▪ Optimiser l'allocation des ressources humaines, ▪ Estimer les performances des employés.	▪ Analyse des traces de changements de poste. ▪ Évaluation des performances des employés en fonction des ventes réalisées.	▪ Nombre d'articles vendus par employé ; ▪ Nombre d'articles différents vendus par employé.

Table 1.2 Illustration de quelques domaines d'application de la BI

4.3 Quelques concepts de base utiles dans le contexte de la BI

Pour rendre ce polycopié autonome, nous présentons dans cette section quelques

définitions et concepts de base utiles à la compréhension des chapitres suivants. Ces prérequis sont incontournables pour pouvoir suivre l'avancement du cours. A noter que la liste des concepts exposés n'est pas exhaustive, mais on a essayé de cerner ceux qui sont les plus utilisés et dont la maitrise est fondamentale dans le cadre de ce cours.

a) Base de données (BDD): désigne une collection d'informations structurées et homogènes relative à un domaine particulier et qui est facilement accessible par des logiciels adéquats.

b) Format de données: suite de caractères dotée d'une *nature* (*numérique, alphanumérique ou alphabétique*) et d'une *longueur***.** Par exemple, le format du nom du client est sur 30 caractères alphabétiques.

c) La structure ou forme des données: désigne le mode d'organisation des données en se basant sur leur format. On distingue les trois structures de base suivantes :

- **Données structurées**: correspond à des données qui sont formatées et qui obéissent à une forme bien définie. Elles se présentent sous forme de lignes/colonnes. C'est le cas des bases de données relationnelles ou des feuilles d'un tableur.
- **Données semi-structurées :** ce sont des données partiellement structurées dont les caractéristiques sont cohérentes et définies, mais elles ne se limitent pas à une structure rigide, telle que celle nécessaire aux bases de données relationnelles, car elles n'ont pas été organisées en référentiel spécialisé. On trouve dans cette catégorie les fichiers XML, RDF et les fichiers plats au format CSV.
- **Données non structurées:** ces données se présentent sous forme brute absolue et il n'y a aucune contrainte ni format qui leur sont imposés. On peut citer à titre d'exemples les documents Word, les pages web, les vidéos, les images, les commentaires Facebook, tweets, commentaires de blogs…etc. Ces données sont difficiles à traiter en raison de leur agencement et de leur formatage complexe.

d) Décision programmable : type de décisions qui portent sur des variables quantitatives aisément identifiables dans le SI et auxquelles on applique des procédures formalisées de résolution. Ainsi, un algorithme peut être élaboré facilement pour permettre d'automatiser la prise de décision. En conséquence, les décisions programmables peuvent être entièrement prises en charge par une machine. Par exemple, accorder une remise de 5% pour les clients dont le total des achats dépasse un certain seuil.

e) Les décisions non programmables : décisions difficiles à prendre car les variables qu'elles manipulent sont qualitatives et nombreuses. Ce type de décision fait appel à l'intuition et à l'intelligence. Il est difficile de les inclure dans un modèle mathématique. Par conséquent, l'automatisation du processus décisionnel est complexe, voire impossible.

4.4 Notions élémentaires du domaine de la BI

En plus des concepts de base précédents, nous anticipons sur la présentation de quelques définitions relatives au domaine de la BI et dont la compréhension rendra la lecture du manuscrit plus pratique.

a) Sources de données : c'est l'endroit d'où proviennent les données utilisées. C'est à dire, l'emplacement physique où les données ont été créées et numérisées. Concrètement, une source de données peut être une base de données, un fichier plat, des mesures provenant directement d'appareils physiques, des données obtenues par web scraping ou des données en streaming provenant d'Internet.

b) Indicateurs clés de performance (Key performance Indicators : KPI)

Un indicateur clé de performance, ou **KPI** (*Key Performance Indicator*), est un élément de mesure métier utilisé pour évaluer différents facteurs essentiels à la réussite d'une entreprise ou d'un projet. Les KPI varient selon les entités; les KPI d'une entreprise pourront correspondre au bénéfice net ou à une mesure de la fidélité des clients, tandis qu'un gouvernement s'intéressera aux taux de chômage ou au revenu des ménages.

Les KPI sont utilisés en BI pour évaluer les tendances métier et pour conseiller des orientations tactiques de l'entreprise. Ils prennent en charge des quantités mesurables et sont reliés à des objectifs d'affaires.

Exemples de KPI : dans le domaine marketing et ventes, on peut identifier les KPI suivants : le nombre de nouveaux clients acquis, le nombre de clients perdus ou le taux de réponse à une campagne marketing.

c) Entrepôt de données (EDD) ou *Data warehouse* : structure de données destinée à centraliser, nettoyer, et uniformiser les données de l'entreprise à des fins de reporting et d'analyse. Elle stocke l'historique des données avec la granularité la plus fine.

d) Magasin de données ou *Data-Mart* **: e**ntrepôt de données dédié à un métier particulier. Il est situé en aval de l'entrepôt de données.

5. Liens entre BI et autres domaines connexes

La BI est une discipline émergente à la croisée de plusieurs domaines informatiques qui traitent, aussi bien du stockage et de l'exploitation des données que de leur intégration et de leur analyse. Dans ce qui suit, nous essayons de mettre en relief les liens existants entre la BI et d'autres domaines connexes qui ont des relations plus ou étroites avec cette discipline.

5.1 BI et Progiciel de gestion intégrée ou Enterprise Resources Planning (ERP)

La constitution *d'une base de données unique* depuis de multiples départements et fonctions permet à toute entreprise d'automatiser et de mettre à jour les informations opérationnelles nécessaires à la prise de décisions. Par ailleurs, l'analyse en temps réel est idéale pour procéder à des pistes d'audit afin de définir l'origine de données spécifiques. C'est le rôle que doit assurer une solution ERP.

Définition: *Un ERP est un logiciel de gestion intégré permettant de gérer l'ensemble des processus d'une entreprise en se basant sur l'intégration de toutes les fonctions, dont la gestion des ressources humaines, la gestion comptable et financière,* ***l'aide à la décision****, mais aussi la vente, la distribution, l'approvisionnement et le commerce électronique.*

La définition précédente stipule que la prise de décision est un axe fonctionnel qui caractérise une solution ERP. *Se pose alors la question des liens existants entre la BI et ERP ?*
L'utilisation d'un ERP permet de centraliser chaque donnée transactionnelle et opérationnelle afin de présenter une vision globale et détaillée de la situation d'une entreprise. Néanmoins, *ces analyses ne considèrent aucune tendance*. Par contre, une solution BI est dédiée à l'analyse de haut niveau, où la prise de chaque décision stratégique peut être déterminante. Donc, la BI considère l'ensemble des informations d'une entreprise, aussi bien les données opérationnelles (*historiques des ventes comptabilisées au quotidien*) et les données relatives aux stratégies (*revenus, croissance et recettes, tendances…*)
Ainsi, la déférence majeure entre BI et ERP réside dans le niveau de prise de décisions. Les analyses fournies par les ERP sont des analyses des données opérationnelles, alors que la BI favorise des analyses de stratégies.

5.2 BI et Data mining

La fouille de données ou *Data mining* est une branche de la science des données (*data science*) qui recherche dans de grands ensembles de données des informations d'intérêt par l'application d'algorithmes efficaces permettant d'identifier des motifs spécifiques. Elle manipule et expose différents modèles ou patterns associés à des ensembles de données massifs qui peuvent fournir des informations utiles dans le cadre de la BI. Ainsi, le data mining peut être considéré comme le précurseur de la BI.

Il existe plusieurs méthodes de data mining. On distingue, notamment :

- **La classification :** divise des grands ensembles de données en catégories spécifiques. Cette technique peut être utilisée, par exemple, dans le cadre des activités de marketing pour permettre aux entreprises de publier différentes annonces dans différents domaines, garantissant que les bonnes annonces ciblent les clients qui répondraient le plus favorablement.
- **Le clustering :** porte la classification à un nouveau niveau, en détectant de petites anomalies ou similitudes que les humains ne peuvent pas observer. En tant que tel, le clustering peut découvrir des moyens de rendre le marketing ciblé, l'efficacité opérationnelle et l'innovation de produit encore plus puissants.
- **Les règles d'association :** permettent de découvrir les relations entre les variables au fil du temps. En suivant et en analysant l'activité des clients, les entreprises peuvent commencer à prédire le comportement futur. Par exemple, les clients qui achètent le produit « A » achètent dans 90% des cas le produit « B », donc changer en conséquence la mise en rayon des produits.

Le data mining et la BI diffèrent dans leur objectif. Mais, ils sont complémentaires dans le sens où l'étude des modèles « patterns » via le data mining aide les entreprises à développer de nouveaux KPI pour la BI. La BI vise donc à montrer les besoins et les évolutions vers de nouveaux KPI, tels que spécifiés par le data mining.

5.3 BI et big data

Inventé par les géants du Web, (*Google, Apple, Facebook, Amazon et Microsoft*), le concept Big data ou données massives est très en vogue durant cette dernière décennie.

Avant de mettre en exergue les liens existants entre BI et Big data, nous commençons par présenter la définition et les concepts de base afférents à la technologie big data pour de stockage et le traitement des données.

Définition : *Le big data désigne un ensemble très volumineux de données de différents formats (structurées, semi-structurées et non structurées) qui offre à tout le monde une solution pour accéder en temps réel à des bases de données géantes, hétérogènes et versatiles.*

Selon cette définition, les données massives doivent vérifier le principe des trois V : *Volume***,** *Vitesse* et *Variété.*

Du point de vue utilité, les données gérées par les technologies big data peuvent potentiellement être exploitées à des fins de gestion, d'analyse de données et de *prise de décisions*. En outre, le big data manipule plusieurs sources de données simultanées qui ne pourraient pas être intégrées avec les solutions classiques. Par exemple, un projet d'analyse de données volumineuses peut tenter d'évaluer le succès d'une compagne marketing sur la base des ventes réalisées d'un produit en mettant en corrélation les données de ventes antérieures, les données de retour marchandise et celles provenant du site de commerce électronique ainsi que les commandes courantes.

Donc, le premier point commun entre la BI et le big data est qu'ils permettent tous les deux de récupérer et de traiter des données pour aider les entreprises dans leur prise de décision. Le deuxième point commun est que le big data consolide la BI par l'apport de l'un des V (*le volume***).** Cependant, les différences essentielles entre BI et big data, se résument aux points suivants :

- **Source de données :** elles sont uniquement opérationnelles pour la BI, mais elles sont opérationnelles et aussi externes pour le big data.
- **Types des données manipulées:** le big Data traite des données brutes (*structurées et non structurées*) issues de différentes sources, notamment celles externes à l'entreprise**,** tels que les réseaux sociaux), ce qui n'est pas le cas de la BI qui analyse des données structurées ou semi-structurées, centralisées… et pour la plupart internes à l'entreprise. Les formats sont donc systématiquement moins variés.

- **Temporalité des données:** la BI utilise des données historiques pour prendre des décisions futures, alors que les solutions big data peuvent, non seulement, aller chercher l'information dans des données historiques, mais aussi des *sources de données en temps réel*.
- **Support de stockage des données:** dans le cadre de la BI, l'information est stockée sur un serveur central **(Entrepôt e données)**, alors que le big data implique un *système de fichiers distribués* **(D**istributed **F**ile **S**ystem **: DFS)**, ce qui rend les opérations plus souples mais aussi la préservation des données plus sûre.
- **Processus de traitement des données:** le big Data utilise une approche de traitement massivement parallèle qui, entre autres, *accélère le traitement et l'analyse des données*.

En résumé, on peut plutôt voir la BI et le big data comme des *approches à forte valeur ajoutée complémentaires*, en particulier en intégrant les apports du big data aux architectures BI déjà puissantes des entreprises actuelles.

6. Conclusion

Dans ce premier chapitre nous avons présenté l'informatique décisionnelle et nous avons montré que pour les entreprises contemporaines, la prise de décision puis l'action stratégique à suivre sont basées, non pas sur les données brutes, mais sur le résultat des transformations, puis l'intégration des données extraites et traitées. La représentation de ces données sous forme de rapports, diagrammes, tableaux de bords…etc., facilitera l'interprétation et le choix de l'action à mener en vue d'optimiser les ressources et les performances de l'entreprise.
Dans cette perspective, le processus BI est un ensemble d'outils, d'application et de méthodologies permettant de récupérer les données brutes (*contenues dans les outils ERP, Customer Relationship Management (CRM), sources externes….)*, à les transformer en informations utiles et à les diffuser sous formes directement exploitables par les managers et les preneurs de décisions.
Le prochain chapitre sera dédié la description de l'architecture et du fonctionnement des solutions BI.

Ce qu'il faut retenir

L'explosion de la masse de données générée par les systèmes d'information des organisations exige de nouvelles technologies et des approches plus performantes pour l'exploitation rationnelle de ces données. La BI vise à offrir aux décideurs des outils efficaces pour l'aide à la prise de décisions stratégiques.

Série de TD N° 1 : de l'Information à la décision

Exercice 1 : Choisissez la ou les bonnes réponses parmi celles proposées

1. Les valeurs ajoutées de l'informatique décisionnelle ou BI sont :
   a. Le contrôle permanent de la gestion interne des activités de l'entreprise.
   b. Anticiper et prévoir les tendances et habitudes des consommateurs.
   c. Offrir une vue de haut niveau sur l'efficacité opérationnelle.
   d. Recueillir l'information, la traiter puis diffuser les résultats aux acteurs concernés.
   e. Organiser et consolider un grand volume de données de façon homogène.
   f. Toutes les réponses précédentes.
2. Les problèmes posés par la diversité des sources de données résident dans:
   a. L'augmentation conséquente da la masse de données à manipuler.
   b. Des difficultés pour centraliser les données sur un support commun et unique.
   c. Le coût d'acquisition des données externes utiles à la prise de décision.
   d. Les données issues des différentes sources sont incompatibles et hétérogènes.
3. La BI est une discipline au carrefour des domaines suivants :
   a. Les S.I opérationnels, les métiers de l'entreprise et la stratégie de l'entreprise.
   b. Les S.I géographiques, les réseaux d'entreprise et le cloud.
   c. Les ERP, le domaine du big data et le data mining.
   d. La gestion des connaissances, la data science et le business analytics.
   e. Toutes les réponses précédentes.
4. L'informatique décisionnelle (BI) fournit comme résultats les éléments suivants
   a. Propose de nouveaux KPI.
   b. Des tableaux de bord qui aident à la prise de décision.
   c. Des réponses aux requêtes d'interrogations sur que s'est-il passé.
   d. Des rapports de synthèses et de visualisation des données.
   e. Toutes les réponses précédentes
5. Répondez aux affirmations suivantes par *VRAI* ou *FAUX* **(***Corrigez dans le cas FAUX***)**
   a. La BI exploite les données structurées et semi-structurées.
   b. Les données utilisées par une solution BI sont seulement les données internes.
   c. La BI aide à la prise des décisions opérationnelles, tactiques et stratégiques.
   d. Le point commun entre les 3 V du big data et la BI se limite au **V** de la variété.

Exercice 2

La discipline fouille de données (data mining) est considérée comme le précurseur de la BI.

Discuter cette assertion, tout en mettant en évidence les principales différences entre les deux domaines ?

Exercice 3

Les S.I opérationnels diffèrent en plusieurs points des SI décisionnels.

En se basant sur vos connaissances préalables sur les SI :

a. Enumérer un ensemble de critères pertinents à la comparaison des deux types de systèmes d'informations ?
b. Dressez une comparaison des deux systèmes d'informations?
c. A partir des deux questions précédentes, dégager une définition d'un SI décisionnel qui met en exergue ses avantages par rapport aux S.I opérationnels ?

Chapitre II : Architecture et fonctionnement d'une solution BI

1. Introduction

Le premier chapitre a été consacré à la présentation de la BI et de son importance pour l'entreprise. Ce chapitre est dédié à l'examen de l'architecture classique d'une solution BI et à l'étude de son fonctionnement. On y abordera la description des différents composants d'une solution BI et leurs interactions, puis on expliquera le fonctionnement détaillé de cette

architecture. Par la suite, nous nous focaliserons sur le processus d'élaboration et de mise en place d'un projet BI. Une attention particulière sera accordée à l'aspect gestion des données, tout en traitant le fonctionnement de la technique ETL (**E**xtract-**T**ransform and **L**oad) pour l'intégration des données. On termine le chapitre par une conclusion.

2. Architecture classique d'une solution BI

D'une manière générale, une architecture décrit les éléments constituants d'un système quelconque, ainsi que les relations entre ces éléments pour atteindre un objectif particulier. Dans notre contexte, une architecture BI peut être perçue comme un système composé d'un ensemble d'éléments (*outils, méthodes, et technologies*) logiciels et matérielles qui, une fois mis en relation, permettent de créer de la connaissance et de répondre aux besoins des décideurs de l'entreprise en matière de prise de décisions stratégiques.
Il est d'une importance capitale de bien cerner et de comprendre chaque élément qui entre dans la composition de la solution BI avant de passer à l'étude des interactions possibles entre les composants et leur assemblage pour réaliser une solution architecturale globale. Toute confusion entre ces différents éléments peut engendrer un échec certain de la perception de l'ensemble.
Les composants d'une solution BI peuvent être classés selon trois dimensions complémentaires qui sont les données, les logiciels et les résultats attendus. Ces trois dimensions seront abordées dans ce qui suit.

2.1. L'aspect statique (*les données*)

L'étape fondamentale du processus d'élaboration et de mise en œuvre de toute solution BI, consiste à identifier les sources de données, à cerner celles qui sont pertinentes à la prise de décision, puis de les intégrer dans une structure commune. En ce sens, il s'agira de recenser les sources de données internes et externes potentielles, de récupérer les données, de les intégrer dans des structures plus élaborées, puis de les organiser et de les stocker de manière à pouvoir en extraire une plus-value exprimant des connaissances utiles à la prise de décision. Ces différentes activités sont décrites et examinées ci-dessous.

2.1.1. Identification des sources et collecte des données

Plusieurs sources de données en entreprise peuvent contenir les données utiles à la prise de décision. Ces sources regroupent, souvent, les données opérationnelles issues directement des activités de l'entreprise et manipulées par son système d'information, comme elles peuvent provenir d'autres sources externes à l'entreprise. Quel que soit leur origine (*interne ou externe*) et indépendamment de leur format initial (*base de données structurées, données semi-structuré ou données plates*), il s'agira d'extraire ces données et de les structurer de façon sémantiquement cohérente pour permettre la création d'un ensemble d'informations homogènes et compatibles dont l'exploitation future servira à produire de la valeur ajoutée et de faciliter la prise de décisions. Pour réaliser cet objectif, il convient d'aller chercher les données où elles se trouvent. Les données applicatives métiers peuvent être stockées dans une ou plusieurs bases de données sous-jacentes à chaque application utilisée *(marketing, ventes, finance, GRH,…),* ou bien dans les applications en contact avec les clients, ou encore dans les systèmes d'informations externes possédés par les partenaires de l'entreprise.
L'opération de collecte des données des différentes sources est déclenchée à intervalles réguliers. Elle consiste à explorer les sources déjà identifiées, d'accéder aux supports de stockage et d'extraire les données considérées. Ces données subissent ensuite des opérations de formatage et d'importation pour être alignées sur les données déjà présentes dans le système (*dans l'entrepôt de données*). La fréquence à laquelle les données sont collectées et les opérations de formatage varient en fonction des besoins de l'entreprise.

La figure 2.1 ci-dessous met en exergue les différentes sources de données possibles de l'entreprise ainsi que leurs formats.

Sources de données	Formats des données
CRM / Ventes ERP GRH Finances Sources de données diverses (Internet, Web services, réseaux sociaux, blogs…	**1. Données structurées (BDD SQL et Tableur)** • ORACLE • MySQL • DB2 • MS-SQL • Fichiers EXCEL (.XLS) **2. Données semi-structurées** • Fichiers XML, RDF, CSV **3. Données non structurées** • Fichiers textes • Pages Web, images, messages…etc

Figure 2.1 Variété des sources de données en entreprise

2.1.2. Le stockage des données dans l'entrepôt

Les volumes considérables de données provenant des différentes sources sont traités et stockés dans une base de données relationnelle unique, appelée *Entrepôt De Données* **(EDD)** ou **(***Data warehouse*). Cette structure de stockage est pensée et conçue pour prendre en charge des informations relatives à l'activité historique de l'entreprise, tout en offrant une vue d'ensemble des différentes transactions qui ont eu lieu au fil du temps dans la perspective de les exploiter à des fins de prise de décisions.

La figure 2.2, ci-dessous met en relief le mécanisme de centralisation et du stockage des données dans une structure unique, appelée l'**EDD**.

(*Le fonctionnement e l'EDD sera abordé dans le prochain chapitre et sa modélisation sera étudiée dans le chapitre IV****)***

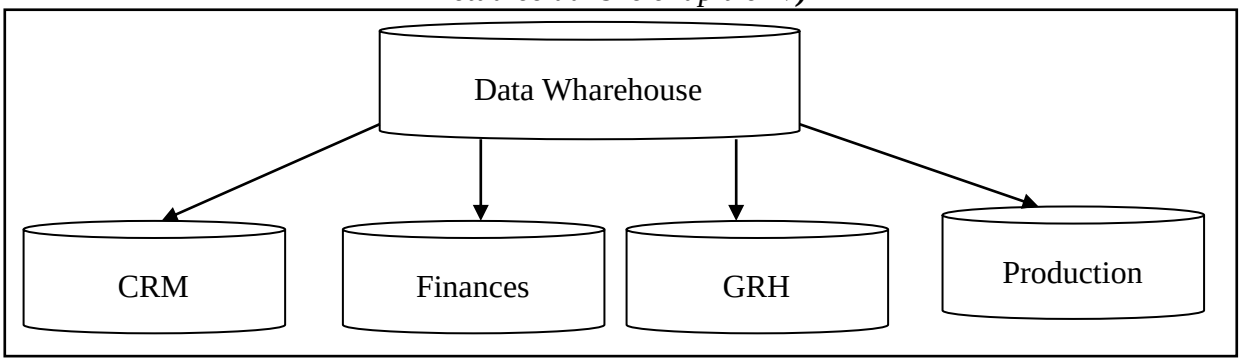


Figure 2.2 Stockage des données dans l'entrepôt et les magasins de données

L'avantage du stockage des données dans l'entrepôt de données est de permettre de :

- Répondre à des requêtes et à des analyses de données,
- Faciliter la prise de décisions et les activités de type Business Intelligence.

L'entrepôt de données constitue, ainsi, un réservoir des données reflétant l'activité de l'entreprise et qui sont utiles à la prise de décision. Ce réservoir peut être éclaté en plusieurs briques élémentaires spécifiques à des activités ciblées (*vente, finance, GRH, production…etc.*). Ces ilots spécialisés sont appelées magasin de données ou (*Data mart*).

2.2. L'aspect traitement (*les logiciels*)

En plus des données, une batterie d'outils logiciels dédiés est déployée pour faire fonctionner les différents composants de l'architecture BI. Ces logiciels permettent, d'une part, d'assurer la jonction entre les divers composants assurant la collecte, la transformation et l'exploitation des données. Et d'autre part, de produire et de réaliser les fonctions spécifiques relatives à l'exploitation des données, telles que les outils d'analyse et de reporting.

Dans ce qui suit, nous identifions les types de logiciels nécessaires à toute solution BI, puis nous abordons leur description, tout en illustrant leur utilisation avec des exemples concrets.

2.2.1 Les outils ETL (*Extract-Transform and Load*)

Les outils ETL sont des logiciels qui répondent à des besoins d'intégration des données diverses. Les premiers ETL ont fait leur apparition dans les années 1970 et étaient utilisés pour agréger et stocker des données de différents types en provenance de multiples sources.

Définition : *un logiciel ETL est un intergiciel (middleware) qui permet de collecter les données en provenance de sources multiples pour ensuite les convertir dans un format adapté à un entrepôt de données, et enfin de les y charger.*

D'après cette définition, on constate que le logiciel ETL permet la consolidation de grandes quantités de données à l'aide des trois opérations d'extraction, de transformation et de chargement dont l'explication est donnée ci-après.

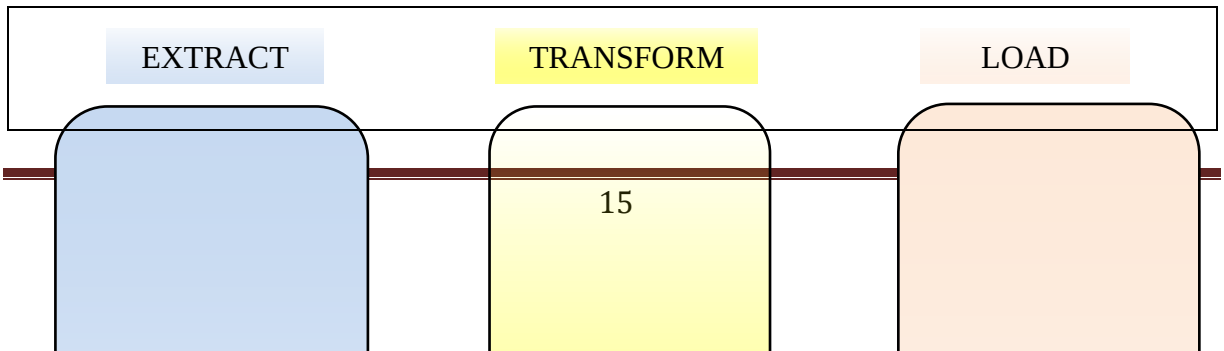
a) **Extraction**: cette opération permet d'identifier et d'extraire les données des sources qui ont subi des modifications depuis la dernière exécution du chargement.

b) **Transformation**: consiste à appliquer diverses transformations aux données pour les nettoyer, les intégrer et les agréger;

c) **Chargement**: action qui consiste à insérer les données transformées dans l'entrepôt et de gérer les changements des données existantes.

La figure 2.3 suivante illustre le séquencement des trois opérations assurées par un outil ETL.

(*La section 5 de ce chapitre sera dédiée exclusivement à l'étude du fonctionnement des ETL*)



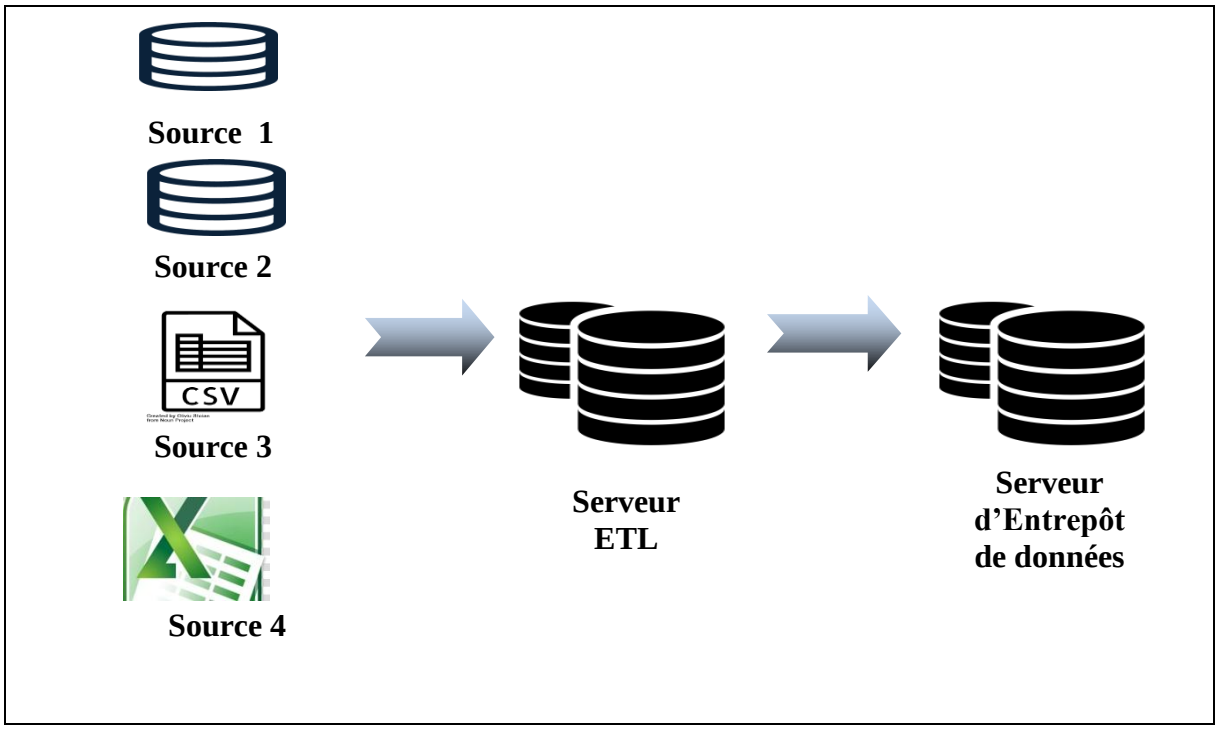


Figure 2.3 Les étapes du processus ETL

Quelques exemples d'outils ETL commerciaux

Il existe une large gamme d'outils logiciels ETL sur le marché. Certains d'entre eux sont des outils commerciaux sous licence et peu sont des outils gratuits en open source. Nous citons, à titre d'exemples, quelques outils ETL les plus populaires sur le marché.

- Oracle Warehouse Builder;
- IBM Infosphere Information Server;
- Microsoft SQL Server Integration Services (SSIS);
- SAS Data Integration Studio.
- SAP - Intégrateur de données BusinessObjects
- Intégrateur de données Oracle

2.2.2 Les logiciels d'exploitation des données

La couche logicielle de toute architecture BI intègre un large éventail d'applications et de logiciels qui sont dédiés à l'exploitation des données collectées dans l'entrepôt et permettant d'assurer différentes fonctions d'exploration et d'analyse. Les résultats fournis en sortie sont des rendus qui sont généralement des graphiques ergonomiques dont l'exploitation facilitera la prise de décision par le système de pilotage de l'entreprise.

Parmi ces logiciels, on distingue les catégories suivantes.

a) Les logiciels de reporting (*logiciels de communication de données)*

Ce type de logiciels permet de matérialiser l'illustration des données pour les rendre compréhensibles par tous les utilisateurs, afin qu'elles deviennent de bons indicateurs de performance. Le principal avantage des outils de reporting est la *visualisation des données*. Dans cette perspective, ils mettent en valeur les données récupérées sur une période souhaitée et les présentent de manière claire afin qu'elles puissent être analysées et exploitées par une tierce personne. Ainsi, ils permettent de rendre compte périodiquement des indicateurs de performance et améliorent considérablement le processus de prise de décision. A titre

d'exemple, les graphiques Microsoft Excel (*histogrammes, graphiques en secteur, courbes, nuages de points, ...etc.*) sont très expressifs pour faire les premiers pas pour la visualisation des données.

b) Les logiciels de traitement analytique en ligne (OLAP)

Un outil OLAP est un logiciel qui permet aux utilisateurs d'analyser des données issues de plusieurs sources en même temps. Ces données doivent être intégrées préalablement dans un entrepôt de données. Les données exploitées par les logiciels OLAP sont multidimensionnelles et représentées sous forme de cubes, ce qui signifie que l'information peut être comparée de nombreuses façons différentes, contrairement aux bases de données relationnelles qui sont considérées comme bidimensionnelles.

(*Pour plus de détails sur les outils OLAP, consultez la section 5 du chapitre IV****).***

c) Les logiciels de visualisation de données (logiciels end–users)

Pour donner vie aux données de l'entreprise, l'architecture BI inclue également des applications *de visualisation de données* **(***data visualization***)** qui fournissent des conceptions de graphiques et des outils pour la création de tableaux de bord de performances BI permettant d'afficher des métriques métier de l'entreprise et autres indicateurs clés de performance. Ces logiciels facilitent la visualisation des données pour que les utilisateurs professionnels réduisent et découpent les données de leur propre chef ou effectuent des rapports ad hoc.

d) Logiciels d'analyse avancée

Les programmes BI peuvent également incorporer des formes d'analyses avancées comme :

- Les outils de data mining, d'analyses prédictives, de fouille de texte (*texte mining*) et les outils d'analyses statistiques.
- Les outils d'analyses big data.
- Les logiciels de mobile BI et de BI en temps réel.
- Des logiciels en tant que service (Software as a Service : **SaaS**).

2.3. Les sorties de l'architecture BI

Les outils logiciels décrits précédemment produisent en sortie différents types de résultats facilement exploitables par les preneurs de décisions. Afin d'offrir des représentations des données sous des formes qui seront directement exploitables, l'architecture BI produit en sortie les types de résultats suivants.

2.3.1. Les analyses multidimensionnelles

Ce type d'analyse permet de visualiser les données manipulées sous forme de tableaux croisés appelés *pivots*. Le principe de fonctionnement se base sur la représentation des données sous forme d'un cube multidimensionnel (*Hypercube*), où chaque côté est une dimension d'analyse et chaque case est une métrique de mesure des performances.

Exemple de cube exprimant les données commerciales (ventes)

Table des ventes				
Code Client	**Numéro Article**	**Date de Vente**	**Montant**	

10	100	15/06/2021	4000,00
20	300	15/06/2021	3000,00
20	200	16/06/2021	70000,00
30	100	10/10/2021	400,00
10	300	10/10/2021	500,00
10	400	20/10/2021	8000,00
20	300	27/11/2021	5000,00
30	500	27/11/2021	8500,00
10	100	27/11/2021	300,00
10	300	28/11/2021	1000,00
…..	…..	……..	……

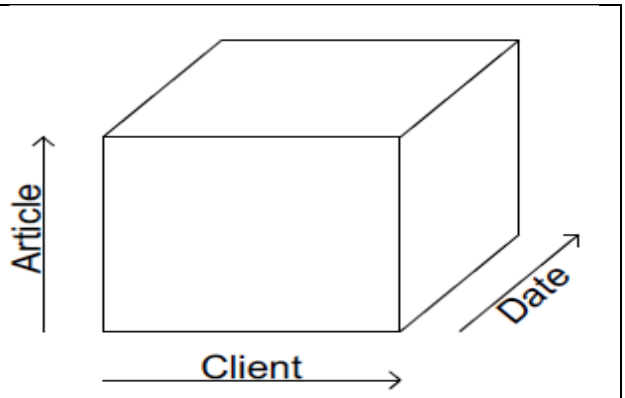


Figure 2.4 Exemple de données commerciales et leur analyse multidimensionnelle

2.3.2. Les requêtes ad-hoc

Cette sortie du système BI vise à traduire les besoins d'affaires en des requêtes interprétables par des systèmes de gestion de données adéquats. Les requêtes ad-hoc sont des requêtes SQL de sélection, combinées avec des requêtes d'agrégation, telles que *SUM, AVG, MOY, MAX, COUNT*…etc. Ces requêtes composées permettent d'effectuer des opérations statistiques sur un ensemble d'enregistrements.

Exemple de requêtes ad-hoc relatives à la gestion commerciale

En se référant aux données commerciales de la *figure 2.4* précédente, on peut formuler la requête ad-hoc suivante.

SELECT *Code-client, Numéro-Article, SUM (Montant)*
FROM *Vente* ***GROUP BY*** *Code-client, Numéro-Article*

Code Client	Numéro Article	Montant
10	100	4300,00
10	300	1500,00
10	400	8000,00
20	200	70000,00
20	300	8000,00
30	100	400,00
30	500	8500,00

Table 2. Exemple de résultats d'une requête ad-hoc pour les données commerciales

2.3.3. Le reporting

Le reporting est probablement l'application la plus utilisée de l'informatique décisionnelle, il permet aux gestionnaires de :

- Sélectionner des données vérifiant des contraintes particulières, tels que une période, un produit, un secteur de clientèle, une zone géographique… etc.
- Trier, regrouper ou répartir ces données selon les critères de leur choix.
- Réaliser divers calculs (*totaux, moyennes, écarts, comparatif d'une période à l'autre…*).
- Présenter les résultats d'une manière synthétique ou détaillée, le plus souvent graphique selon les besoins ou les attentes des dirigeants de l'entreprise.

Cette représentation des données offre des visualisations ergonomiques sous forme de graphiques facilement exploitables. Les logiciels commerciaux offrent une large gamme de modèles en sortie, comme les histogrammes, les graphiques en courbes, courbes en surfaces, graphiques en radar ou en secteur ou encore en anneau…etc.

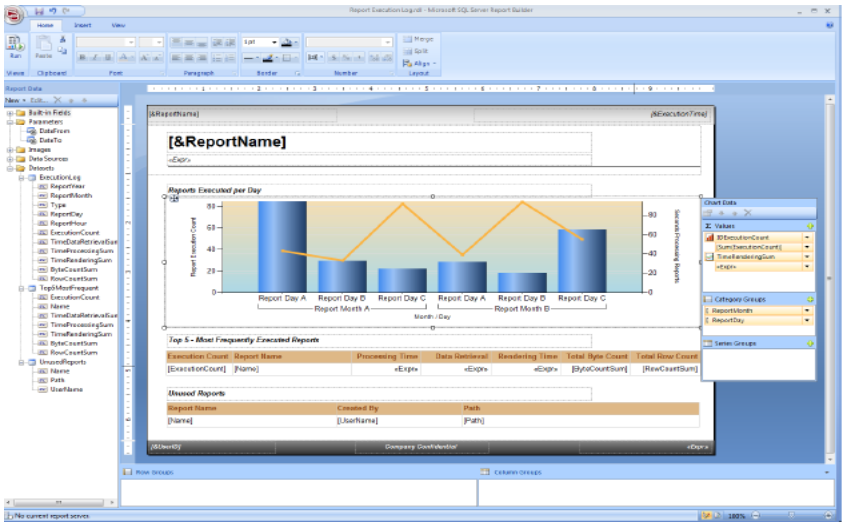


Figure 2.5 Création de rapports graphiques avec MicrosoftSQLServer Reporting service

2.3.4. Les tableaux de bord (*dashboard*)

Par la combinaison des données issues de divers systèmes, les tableaux de bord permettent de mettre en valeur les indicateurs de performance et les éventuels problèmes qui leur sont associés à l'aide des éléments visuels suivants :

- **Graphique** : par exemple des courbes, des tartes,…etc.
- **Jauges** (*pour illustrer par exemple les profits par utilisateur*).
- **Des feux de circulation** : par exemple le rouge signifie qu'il y a un problème.

Les tableaux de bord offrent une vision globale et unifiée de haut niveau de l'entreprise. Ils tiennent compte des changements ponctuels des données en mettant en œuvre un mécanisme de rafraichissement périodique à intervalles réguliers (*toutes les heures, par exemple*).

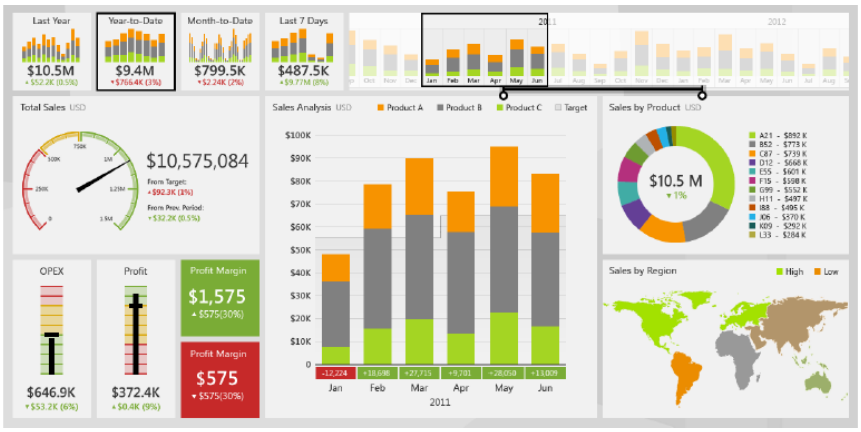


Figure 2.6 Exemple de tableau de bord avec Microsoft PowerPivot

2.3.5. Les Scorecards

Ces graphiques fournissent des métriques de performance de l'entreprise et offrent une représentation graphique des KPI caractérisant les objectifs stratégiques. Ces métriques sont comparées avec les valeurs cibles de l'entreprise pour une certaine période, par exemple durant les cinq dernières années.

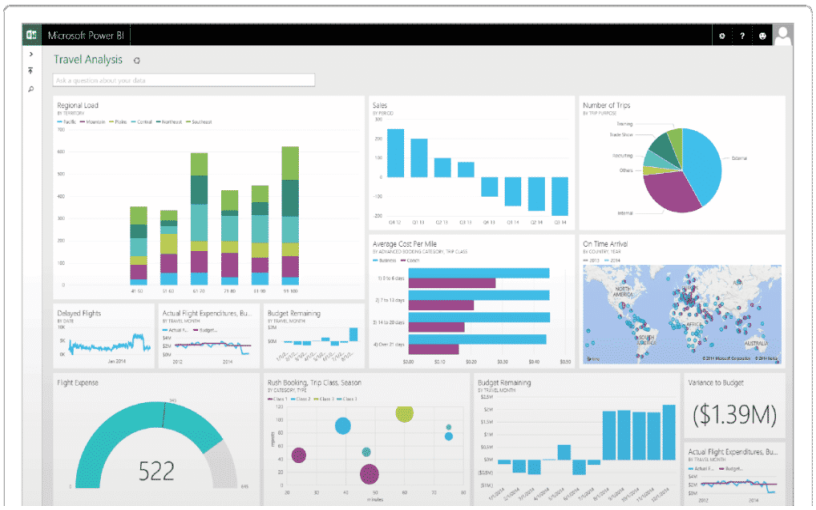


Figure 2.7 Exemple de Scorecard avec l'outil Microsoft Power BI

2.3.6. Les Cockpits

Un Cockpit d'administration BI est un outil graphique de représentation des données qui permet de surveiller l'évolution des performances des systèmes BI. Il fournit un point d'entrée central et il met à la disposition des décideurs des graphiques qui fournissent des moniteurs en temps réel et des statistiques d'exécution. En plus, il offre un accès contextuel à des rapports et applications complètes qui aident à identifier et à analyser les problèmes.

Les cockpits visent à :

- Optimiser les performances des activités BI au niveau de l'organisation.
- Gérez les systèmes BI à partir d'un emplacement unique, réduisant le coût total de possession (*Total Cost of Ownership : **TCO***).
- Suivre le statut des objets BI, des opérations BI, …etc.

Ainsi, le cockpit d'administration BI permet de naviguer à travers les différents composants du système BI, tout en effectuant des transactions et des requêtes pertinentes pour analyser les performances du système et résoudre les problèmes sans avoir à se connecter explicitement à un système.

3. Fonctionnement de l'architecture BI

Comme illustré dans la *figure 2.8* suivante, les données (*colonnes 1 et 3*) et les logiciels (*colonnes 2 et 4*), sont mis en interaction pour aboutir aux sorties attendues de l'architecture BI (*colonne 5*). Le fonctionnement de l'architecture BI est un processus progressif qui se déroule comme suit :

a) Identification des sources, extraction et stockage des données : cette opération consiste à récupérer les données émanant du système d'information de l'entreprise et autres type de données de production (*colonne 1*), en utilisant un logiciel ETL (*colonne 2*). Ce dernier assurera, par la suite, la transformation et le chargement des données collectées dans une structure dédiée. Ces données seront traitées afin qu'elles soient disponibles pour un usage décisionnel, puis chargées et stockées dans l'entrepôt de données. Souvent l'entrepôt de données et scindé en plusieurs silos spécialisés, appelés magasins de données ou Data mart (*colonne 3*).

b) Traitement et production des résultats *:* l'accès et l'exploitation des informations stockées dans l'entrepôt de données (*ou les magasins de données*) sont opérés par des outils logiciels dédiés qui assurent différentes fonctions de filtrage, d'analyse et de visualisation

des données. Ce type de logiciels réalise des fonctions et des types d'utilisation spécifiques qui répondent aux besoins des décideurs (*colonne 4*).

c) Diffusion des résultats : le traitement des données produit différents types de résultats facilement exploitable (rapports, *tableaux de bords, analyse multidimensionnelle, cockpit,…etc.*) qui seront utilisés à des fins de prise de décision (*colonne 5*).

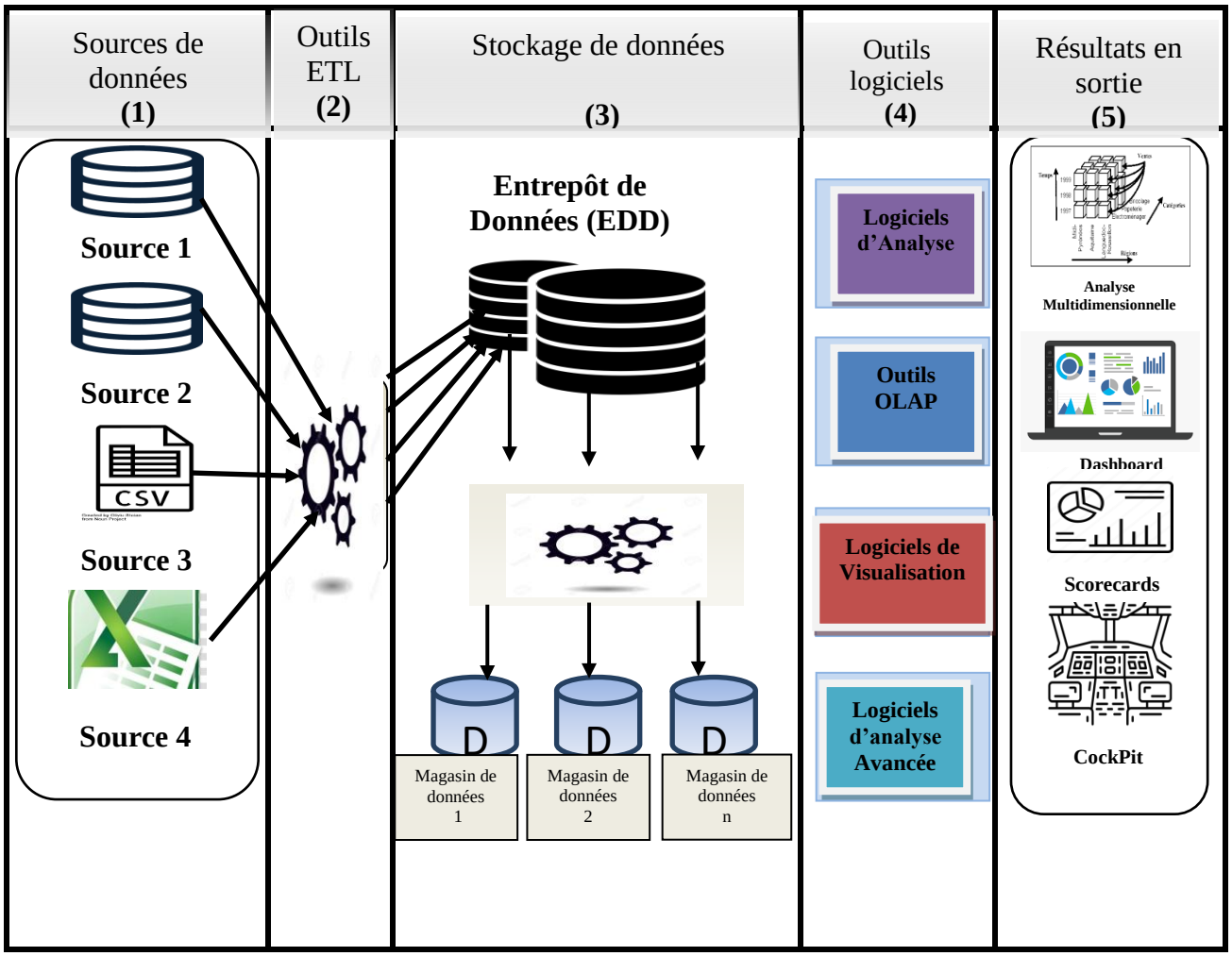


Figure 2.8 Architecture et fonctionnement d'un système BI

4. Les étapes du processus BI ou la chaine décisionnelle

Un système d'information décisionnel orienté BI assure quatre fonctions qui sont la collecte, l'intégration, la diffusion et la restitution des données.
Comme illustré dans la *figure 2.9* suivante, la chaine décisionnelle est un processus incrémental qui débute par la collecte des données et qui se termine par la restitution des résultats aux acteurs concernés.

Les quatre fonctions de la chaine décisionnelle sont explicitées de manière claire et détaillée dans ce qui suit.

4.1 La collecte (*ou alimentation*)

Cette fonction est la plus délicate à mettre en place dans un système BI. Elle consiste à identifier, à sélectionner, à extraire et à filtrer les données brutes issues du micro et/ou du macro-environnements de l'entreprise. Ainsi, il s'agira dans un premier temps de cerner les données pertinentes dans le cadre de l'aide à la prise de décision. Pour cela il convient d'aller chercher les données internes et/ou externes où elles se trouvent. Généralement, les données applicatives

métier sont stockées dans une ou plusieurs bases de données correspondantes à chaque application utilisée. Néanmoins, il faut surmonter les problèmes induits par la variété des sources de données et relatifs à d'hétérogénéité tant sur le plan technique que sur le plan sémantique. Dans ce contexte, l'utilisation d'un outil ETL dédié est incontournable. Elle permettra, à la fois, d'extraire les données, de les transformer, puis de les charger dans un entrepôt de données. Ces trois actions peuvent être effectuées de manière périodique via l'utilisation de batchs configurés préalablement. Par exemple, les actions d'extraction, de transformation et de chargement peuvent être lancées tous les soirs à minuit lorsque plus aucune application n'est active.

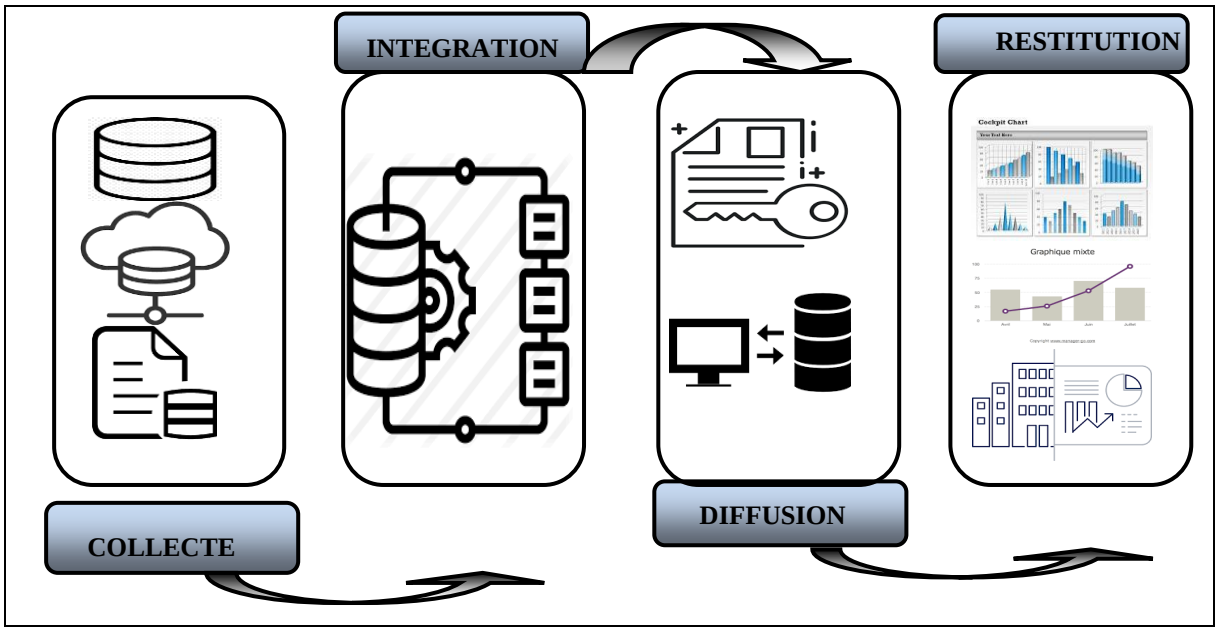


Figure 2.9 Les étapes d'un processus décisionnel

4.2 L'intégration

Afin de s'abstraire de la diversité des sources de données, l'intégration assure la concentration des données collectées dans une structure homogène et unifiée qui est l'entrepôt de données. Ce dernier est le noyau central de toute l'architecture BI, dans le sens où il permet aux applications d'aide à la décision de bénéficier d'une source d'information commune, normalisée et fiable. Durant cette étape les données sont transformées et filtrées en vue de maintenir leur cohérence globale. Enfin, c'est aussi durant cette étape que sont effectués les éventuels calculs et agrégations communs à l'ensemble du système BI.

Une fois les données centralisées par un outil ETL, celles-ci doivent être structurées au sein de l'entrepôt de données. Cette opération est toujours faite par un outil ETL grâce à un connecteur permettant l'écriture dans l'entrepôt de données. En définitif, l'intégration consiste en un prétraitement dont le but est de faciliter l'accès aux données centralisées par les outils d'analyse et de visualisation.

4.3 La diffusion (*ou organisation*)

Cette étape permet de mettre les données à la disposition des différents utilisateurs de la solution BI. C'est durant cette étape que la gestion de droits d'accès correspondant au profil ou au métier de chacun est configurée et déployée. De cette façon l'accès direct à l'entrepôt de données ne sera pas autorisé à tout utilisateur. En effet, un mécanisme de personnalisation doit être mis en place pour répondre aux besoins spécifiques de chacun, tout en lui offrant les variables ou les indicateurs qui l'intéressent.

L'objectif principal de l'étape de diffusion est de segmenter les données collectées en *contextes de diffusion* qui soient cohérents, simples à utiliser et qui correspondent à une activité décisionnelle particulière (*par exemple aux besoins d'un service particulier*). Chaque contexte de diffusion peut correspondre à *un magasin de données* qui est généralement multidimensionnel et modélisable sous la forme d'un hyper-cube pouvant être mis à disposition des utilisateurs via un outil OLAP.
A signaler que les différents contextes d'un même système BI n'ont pas forcement tous besoin du même niveau de détails selon la cible visée. En effet, de nombreux agrégats n'intéressent que certaines applications et ne sont, donc, pas considérés comme des agrégats communs. Ces cumuls ne doivent pas être gérés par la fonction d'intégration, mais par celle de la diffusion. Pour des raisons de performances, ces agrégats peuvent être stockés de manière persistante dans les magasins de données spécifiques, au lieu de les recalculer dynamiquement chaque fois.

4.4 La restitution

C'est la fonction la plus visible pour l'utilisateur final d'un système décisionnel de BI. Elle assure le fonctionnement du poste de travail, le contrôle d'accès aux rapports, la prise en charge des requêtes et la visualisation des résultats sous différentes formes pour qu'ils soient directement exploitables. Donc, cette dernière étape offre une présentation des informations à valeur ajoutée et des KPI, de telle sorte qu'elles apparaissent de la façon la plus lisible possible. La représentation ergonomique des résultats consolide fortement l'aide à la prise décision. Dans cette perspective, les données sont principalement modélisées par des représentations à base de requêtes afin de constituer des tableaux de bord ou des rapports via des outils d'analyse décisionnelle.

5. Gestion des données de la solution BI et outils ETL

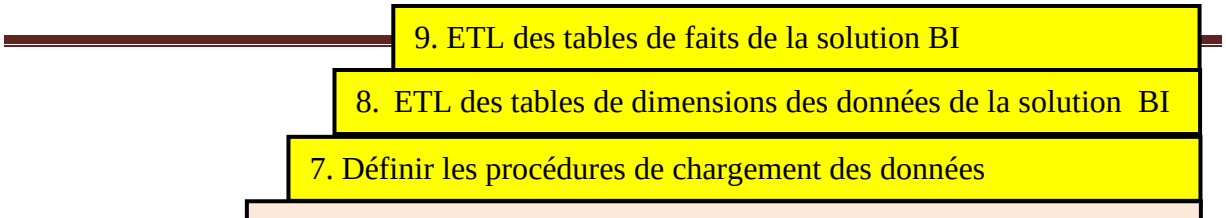
Les données constituent la colonne vertébrale de toute la gestion opérationnelle de l'entreprise et elles sont au cœur de l'action décisionnelle. Le terme « *gestion des données* » désigne l'ensemble des opérations techniques et organisationnelles nécessaires à la construction et à la maintenance d'un Framework pour l'importation, la transformation, le stockage et l'archivage des données qui sont nécessaires aux activités de l'entreprise.
Vue l'importance de l'aspect gestion des données dans le système décisionnelle global, dans cette section nous nous focalisons sur les questions inhérentes à l'importation, à la transformation et au chargement des données. Les problématiques de leur modélisation et de leur exploitation seront abordées dans les prochains chapitres du polycopié.

5.1 Fonctionnement général d'un outil ETL

Comme il a été déjà expliqué dans la sous-section **2.2.1**, les outils ETL sont des logiciels inévitables à toute solution BI. En effet, ils assurent l'extraction des données des différentes sources, puis opèrent leurs transformations en des formats plus adéquats et enfin, ils les stockent dans l'entrepôt de données. Partant de ce rôle primordial des ETL, cette section est réservée exclusivement à l'analyse de leur fonctionnement.
La *figure 2.10***,** ci-dessous illustre le principe général de fonctionnement d'un processus ETL. Comme il est observé dans la figure, le mécanisme ETL est un processus incrémental qui passe par plusieurs opérations complémentaires, dont l'explication détaillée est donnée ci-dessous. Les trois premières opérations constituent l'étape d'extraction, les trois suivantes l'étape de transformation et les trois dernières forment l'étape de chargement.

5.2 Etape d'extraction des données (*Extract*)

Avant toute action d'extraction, il faut tout d'abord identifier les sources de données. En effet, la grande diversité des sources de données impose une recherche exhaustive des données pertinentes pour la solution BI cible.



5.2.1 Identification des sources

La procédure suivante cerne les actions à suivre pour l'identification des sources et la conduite à tenir pour surmonter les contraintes rencontrées durant cette phase.

- ***Recenser les métriques et attributs de dimension*:** consiste à énumérer les attributs cibles nécessaires à l'entrepôt de données;
- ***Trouver les correspondances source-cible:*** pour chaque attribut cible, il faut trouver la source et l'attribut correspondant de cette source;
- ***Sélection des sources pertinentes:*** si plusieurs sources sont trouvées lors de l'opération précédente, alors il faut choisir la plus pertinente;
- ***Consolidation des attributs:*** Dans le cas où l'attribut cible exige des données de plusieurs sources, alors il faut formaliser les règles de consolidation;
- ***Expression des règles de découpage:*** si l'attribut source renferme plusieurs attributs cibles, alors il faut spécifier les règles de découpage. Par exemple, si l'attribut cible *Nom-client* contient, à la fois le nom et le prénom du client, il faut opérer son découpage en deux attributs distincts (*Nom-client, Prénom-client*).
- **Elimination des données manquantes:** dans le cas où plusieurs attributs cibles contiennent encore des valeurs manquantes, alors il faut inspecter toutes les sources possibles afin de localiser ces valeurs.

5.2.2 Extraction des données

Une fois les sources de données identifiées, l'extraction proprement dite peut être lancée. Cette opération peut être activée de deux manières différentes suivant le contexte de déploiement de la solution BI.

a) Extraction complète: ce type d'extraction est employé lors d'un chargement initial des données dans l'entrepôt ou bien lors d'un rafraichissement complet des données (*dans le cas de changement d'une source, par exemple*). L'extraction complète permet de capturer l'ensemble des données à un certain instant (*snapshot de l'état opérationnel*). Néanmoins, le chargement complet peut être très coûteux en temps, du fait que toutes les données seront chargées (*de plusieurs heures à plusieurs jours en fonction du volume des données manipulé*).

b) Extraction incrémentale: cette extraction capture uniquement les données qui ont changé ou ont été ajoutées depuis la dernière extraction. Elle peut être faite en temps-réel, c'est-à-dire

au moment où les transactions surviennent dans les systèmes sources (*par des triggers ou par les journaux des transactions*), ou bien en différé, en analysant tous les changements effectués pendant une certaine période grâce à des programmes de comparaison des états des sources pour des périodes différentes (*heure, jours, mois…*).

5.3 Etape de transformation des données (*Transform*)

Avant de charger les données émanant des différentes sources dans l'entrepôt, plusieurs catégories de transformations doivent être opérées sur ces données. Les transformations peuvent porter aussi bien sur le format des données que sur le contenu lui-même.
*La table 3***,** ci-dessous met en exergue les différents types de transformations qu'un outil ETL standard doit garantir. Pour chaque type de transformation, un exemple illustratif est montré dans la dernière colonne du tableau.

N°	Type de Transformation	Description	Exemples
1	Redressement de format	Changer le type ou la longueur d'un attribut.	*Adresse-Client* sur 30 caractères α-numériques au lieu de 40 α-bétiques.
2	Unification du codage de champs.	Consolider les données de sources multiples.	[*Homme, Femme*], [*H,F*], [*1,2*]
3	Transcription des valeurs en codes.	Faire correspondre des codes à des valeurs.	*G :Gros client, M : moyen* et *F :Faible*
4	Pré-calcul des valeurs dérivées.	Appliquer les règles de calcul.	*Profit=Prix vente- coûts* *Prix TTC= Prix HT + TVA*
5	Découpage de champs complexes.	Extraire des informations atomiques à partir d'autres articulées.	*Prénom, nomFamille* à partir d'une chaine de caractère *Nomcomplet*.
6	Fusion de plusieurs champs.	Regrouper les informations d'une même entité.	Source 1 : *Code* et *libellé* produit Source 2 : Type de *forfaits* et *remises* Source 3 : *Coût de fabrication produit* et *Conditions de stockage*
7	Conversion de jeu de caractères.	Unifier les divers jeux de caractères.	*EBCDIC (IBM)* vers *ACSII* *UNICODE* vers *UTF8*
8	Conversion des dates.	Harmoniser les formats des dates.	*24FEB 2021* vers *24/02/2021* *02/24/2021* vers *24/02/2021*
9	Conversion des unités de mesure.	Utiliser les mêmes unités de mesures. (*système international*)	Changer les unités impériales à métrique. *Exemple : inch en cm*
10	Pré-calcul des agrégats.	Calculer les sommes, produits, moyennes.	*Total des ventes* par semaine et par mois de chaque produit
11	Déduplication ou redondances des tuples.	Plusieurs enregistrements pour la même entité.	*Client* au magasin, *client* en ligne, *client* potentiel.

Table 3. Les types de transformations assurées par les outils ETL

5.4 Etape de chargement des données (*Load*)

Une fois les données sont extraites et transformées dans des formats adéquats, la dernière étape du processus ETL standard consiste à les charger dans leur nouvel emplacement qui est l'entrepôt de données. En général, les entrepôts de données supportent trois modes pour le chargement des données: le chargement *initial***,** le chargement *incrémentiel* et le chargement *complet***.**

5.4.1 Chargement initial

Ce type de chargement n'est opéré qu'une seule fois, lors de l'activation de l'entrepôt de données. A cause de la longue durée que peut prendre le processus de chargement initial et afin d'éviter la génération d'incohérences au niveau de l'entrepôt, il est impératif de désactiver temporairement les indexes et les contraintes d'intégrité référentielles relatives aux clés étrangères.

5.4.2 Chargement incrémentiel

Ce type de chargement peut être fait soit en temps réel, soit en batch (*traitement par lots*), mais une fois le chargement initial terminé. Il doit tenir compte de la nature des changements survenus dans les sources de données. A cet effet, une stratégie de gestion des changements doit être adoptée pour chaque situation. On parle de dimension de changement lent (*Slowly Changing Dimension* **:** *SCD*) qui peut être de différents types. Les stratégies d'historisation possibles pour les différents CSD sont les suivantes :

- **SCD Type 1**: consiste à écraser l'ancienne valeur avec la nouvelle valeur. Par exemple, le client a changé son adresse de livraison.
- **SCD Type 2**: consiste à ajouter une ligne dans la table de dimension pour la nouvelle valeur. Par exemple, si le client a changé son adresse de livraison de A à B, alors préserver les deux valeurs A et B. Donc, on aura deux enregistrements du même client avec deux valeurs distinctes pour l'attribut adresse.
- **SCD Type 3**: permet d'avoir deux colonnes dans la table de dimension correspondantes à l'ancienne et la nouvelle valeur dans la colonne courante. Pour l'exemple de changement d'adresse, il faut créer une nouvelle colonne dont le libellé sera *NOUVELLE-ADRESSE*, tout en gardant l'ancienne colonne *ADRESSE*.
- **Stratégie Hybride: o**n combine les stratégies de gestion des types de changements 2 et 3.

5.4.3 Chargement complet

Ce type de chargement est employé lorsque le nombre de changements rend le chargement incrémental trop complexe. Par exemple, lorsque plus de 20% des enregistrements ont changé depuis le dernier chargement.

A signaler que pour les différents types de chargement précédents, certaines considérations supplémentaires sont à prendre en compte, à savoir :

- Opérer le chargement des données en périodes creuses (*entrepôts de données non utilisé*);
- Considérer la bande passante requise pour le chargement ;
- Prévoir un plan pour la vérification et l'évaluation de la qualité des données chargées.
- Commencer par le chargement des données des tables de dimension avant celles des faits.

6. Conclusion

Dans ce chapitre nous avons examiné l'architecture et le fonctionnement d'une solution BI. Les différents composants de l'architecture, à savoir les données, les logiciels et les sorties attendues ont été cernés et leurs interactions ont été profondément exposées et étudiées. Aussi, le fonctionnement global de l'architecture BI et le processus d'élaboration de la chaine décisionnelle ont été largement expliqués. D'autre part, une attention particulière a été attachée à la gestion des données alimentant l'entrepôt de données de la solution BI. Dans cette perspective, nous nous sommes focalisés, particulièrement sur le mécanisme de fonctionnement des outils ETL et les opérations d'extraction, de transformation et de chargement des données ont été largement examinées et illustrées par des exemples.

Le prochain chapitre est consacré à l'étude des entrepôts de données.

Ce qu'il faut retenir

Une solution BI s'articule autour de deux dimensions complémentaires. Un aspect statique matérialisé par l'entrepôt de données et un aspect dynamique exprimé par une batterie d'outils logiciels. Les outils ETL constituent le point nodal de la solution BI, car ils permettent l'extraction des données des différentes sources, puis de les convertir en des formats hétérogènes avant de les charger dans l'entrepôt de données.

Série de TD N° 2 : Architecture et fonctionnement d'une solution BI

Exercice 1 : Choisissez la ou les bonnes réponses parmi celles proposées

1. Dans une architecture Business Intelligence, l'objectif principal de la phase d'intégration de données est qu'à terme :

a. Les données soient utilisables de façon homogène comme si elles constituaient une seule base de données permettant ainsi leur analyse.
b. Les données soient duplicables à travers le Cloud pour pouvoir les partager.
c. Les données soient contrôlées seulement par l'administrateur pour des raisons de sécurité.
d. Offrir une vision transversale de l'entreprise pour répondre aux besoins décisionnels.

2. L'entrepôt de données est conçu pour :
   a. Répondre à des requêtes et à des analyses de données.
   b. Permettre l'exploitation des données par des outils OLTP.
   c. Organiser et stocker les données de manière à pouvoir en extraire une plus-value.
   d. Faciliter la prise de décisions et les activités de type Business Intelligence.
3. Le chargement incrémentiel des données par un outil ETL:
   a. Doit tenir compte de la nature des changements survenus dans les sources de données.
   b. Tient compte de la stratégie de gestion des changements adéquate à chaque situation.
   c. Doit être lancer en batch (*traitement par lots*).
   d. Est opéré une fois le chargement initial terminé.
4. L'opération de collecte de données à partir des différentes sources consiste à :
   a. Exploiter les bases de données internes pour récupérer les données utiles.
   b. Explorer, accéder aux supports de stockage et extraire les données considérées.
   c. Identifier, sélectionner, extraire et filtrer les données brutes .
   d. Cerner les données externes pertinentes pour la prise de décision.
5. Répondez par **VRAI** ou **FAUX** aux affirmations suivantes et corrigez si **FAUX**
   a. L'entrepôt de données s'appuie sur le principe de la traçabilité des informations.
   b. L'extraction complète des données par un outil ETL est employée uniquement lors d'un chargement initial des données dans l'entrepôt.
   c. Un outil ETL sert à collecter, convertir et charger les données dans l'entrepôt.
   d. Les transformations effectuées par un outil ETL portent sur le contenu des données uniquement.

Exercice 2

La chaine décisionnelle ou le processus BI s'articule autour des quatre fonctions suivantes: la collecte, l'intégration, la diffusion et la restitution.

a. Rappeler les objectifs visés par la fonction de diffusion ?
b. Quels seront les problèmes potentiels engendrés par la suppression et/ou l'omission de cette fonction ?
c. Illustrez votre réponse par un exemple réel de votre choix ?

Exercice 3

La phase de transformation est l'étape la plus complexe de tout outil ETL, car souvent plusieurs types de problèmes peuvent se poser à ce niveau. En se basant sur les types de transformations à effectuer déjà vues dans le cours :

a. Identifier les types de problèmes potentiels rencontrés lors de cette étape ?
b. Pour chaque type de problèmes, proposez et décrivez des techniques pour les surmonter ?

Exercice 4 : Gestion des flux de données en temps réel

Une des limitations principales des outils ETL est leur incapacité à traiter et à intégrer des flux de données en temps réel (*real-time data streaming*).

a. Discuter en détails ce constat, tout en mettant en évidence le besoin de toute solution BI à intégrer les données en temps-réel.

b. Enumérer les inconvénients des outils ETL conventionnels en matière de traitement des données en temps réels.

c. Les nouvelles approches d'intégration proposent un changement de paradigme et utilise, plutôt la technique ELT (Extract-Load and Transform).

   Dressez une comparaison entre les deux approches ETL et ELT.

d. En se basant sur la comparaison précédente, proposez une technique pour améliorer les architectures BI afin de prendre en compte la nature des données en flux continu**.**

Chapitre III : Les Entrepôts de données (EDD)

1. Introduction

Les sources de données d'une entreprise proviennent généralement des bases de production qui sont souvent réparties dans les systèmes et applications multiples, comme la gestion commerciale, les finances, la production ou la GRH. Les données gérées par ces applications ne sont pas nécessairement compatibles entre elles, car elles sont conçues pour prendre en charge et être efficaces pour les fonctions critiques qui leur sont dédiées. Ainsi, ces données

sont peu ou mal adaptées pour une vision stratégique et une analyse de prise de décision. Il est alors, impératif de les agréger, de les restructurer et de les valoriser en les intégrant dans une nouvelle structure qui sera dotée d'un nouveau format plus adéquat permettant de prendre en charge les préoccupations des preneurs de décisions. Cette nouvelle structure n'est autre que l'entrepôt de données (EDD) ou le Data warehouse.

Ce chapitre traite des différents aspects relatifs aux entrepôts de données. On y décrira leurs caractéristiques architecturales et fonctionnelles, puis on abordera les différentes configurations possibles avec leur évaluation en termes de performances. Par ailleurs, l'utilisation des magasins de données ou data-Mart sera également abordée dans ce chapitre.

2. Définitions, caractéristiques et objectifs d'un entrepôt de données

De manière très simpliste, nous pouvons considérer un entrepôt de données comme un vaste gisement de données qui facilite la prise de décision dans l'entreprise. Il est l'élément principal du système d'information décisionnel.

2.1 Définitions d'un entrepôt de données

Un entrepôt de données représente une agrégation des bases de données opérationnelles (*commerciale, comptabilité, production, GRH,…etc.*) en une nouvelle structure qui permettra à l'utilisateur d'y accéder de manière simple et ergonomique à des fins de prise de décisions. Le processus de collecte, de transformation et de chargement des données sources dans l'entrepôt de données est assuré par un outil ETL, comme le montre la **figure 3.1** suivante.

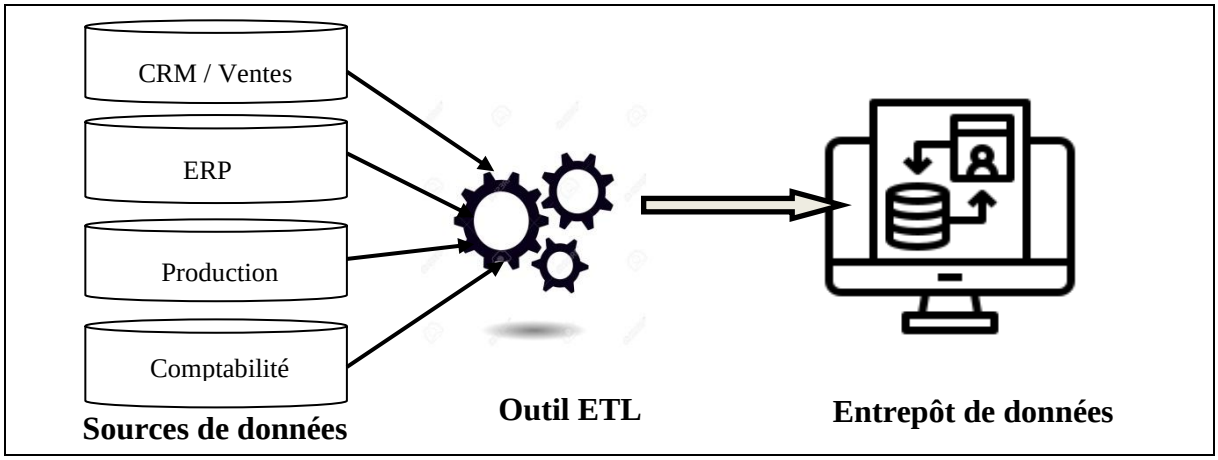


Figure 3.1 Agrégation des données sources dans l'entrepôt de données

Nous donnons, ci-dessous, trois définitions de l'entrepôt de données parmi lesquelles nous allons retenir la plus expressive et la plus pertinente.

Définition 1

Un entrepôt de données regroupe des données structurées provenant de sources variées et sert de référentiel pour l'ensemble de l'entreprise.

Définition 2

Un entrepôt de données est une base de données relationnelle pensée et conçue pour satisfaire les actions suivantes:

- *La prise de décision et les activités de type BI ;*
- *Les requêtes et les analyses de données ;*
- *Les informations stockées dans l'entrepôt de données sont historiques, et offrent une vue d'ensemble des différentes transactions qui ont eu lieu au fil du temps.*

Bill Inmon, théoricien fondateur des entrepôts de données, défini l'entrepôt de données comme suit :

Définition 3

« *L'entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision* ».

Nous allons retenir cette dernière définition, car elle met en exergue les quatre critères que doit satisfaire un entrepôt de données. Ces critères sont explicités dans la sous-section 2.3.

2.2 Les objectifs d'un entrepôt de données

L'entrepôt de données est utilisé pour stocker les données provenant d'autres bases de données dans le cadre de prise de décisions. En tant que structure de données centralisée il permet d'assurer les fonctions suivantes :

- Offrir une vision transversale de l'entreprise pour répondre aux besoins décisionnels;
- Garantir une intégration cohérente des différentes données dans une seule base, ce qui facilitera leur analyse et d'anticiper sur les évolutions qui peuvent surgir dans l'environnement de l'entreprise;
- Améliorer la prise de décision et le déploiement de stratégies plus efficaces.
- Faire des prédictions utiles à la vision stratégique de l'entreprise,
- Offrir un avantage concurrentiel important pour une entreprise.

En se basant sur les fonctionnalités précédentes, les entrepôts de données sont conçus et mis en œuvre afin de réaliser les objectifs suivants:

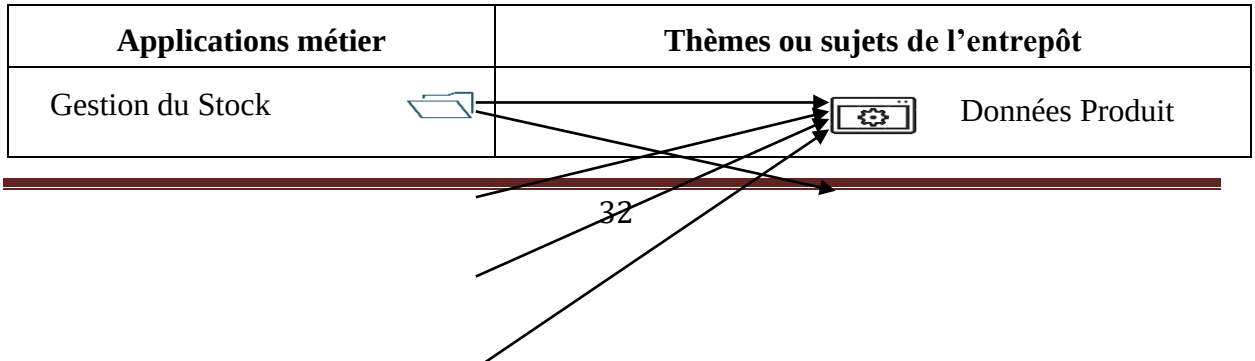
- Offrir une vue unique et consolidée des données de l'entreprise ;
- Procurer de l'information de qualité, le plus rapidement possible;
- Simplifier l'accès aux données;
- Permettre de mener des analyses poussées sur les données par rapport à différents sujets d'affaires;
- Libérer les ressources (par exemples des serveurs) dédiées au traitement des transactions (OLTP) de celui réservé aux tâches d'analyse (OLAP).

2.3 Caractéristiques d'un entrepôt de données

Pour atteindre les objectifs attendus dans le cadre d'une solution BI, les données d'un entrepôt doivent satisfaire les quatre propriétés spécifiées dans la définition 3 et que nous allons examiner ci-après.

a) **Orientées sujet :** contrairement aux bases de données de production qui sont organisées par type de processus fonctionnel, l'entrepôt de données est lui organisé autour des sujets majeurs de l'entreprise. Les données sont donc structurées par *thèmes*, par exemple les ventes ou les produits. Ces thèmes étant souvent transverses par rapport aux structures fonctionnelles et organisationnelles de l'entreprise (et donc *transverses par rapport aux systèmes de production*). Ainsi, un entrepôt de données regroupe les informations des différents métiers et ne tient pas compte de l'organisation fonctionnelle des données.

La *figure 3.2* suivante montre quelques applications relatives aux processus fonctionnels d'une entreprise et les thèmes possibles pris en charge par l'entrepôt de données.



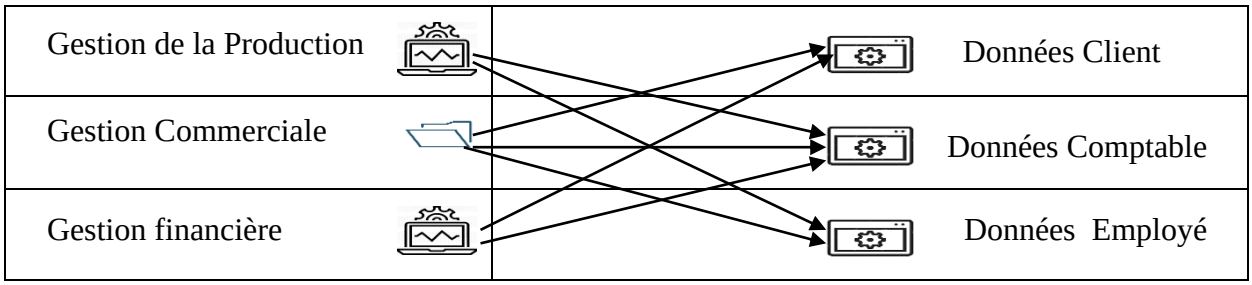


Figure 3.2 Orientation sujet des données de l'entrepôt

b) Intégrées : avant d'être chargées dans l'entrepôt, les différentes données doivent être mises en forme et unifiées dans un format unique afin d'assurer leur cohérence globale, ce qui facilitera leur exploitation et leur analyse. Comme il a été expliqué dans la section *5.3* du chapitre précédent, l'utilisation des outils ETL permettra d'assurer la cohérence et l'intégrité des données par la conduite des actions suivantes : l'adaptation des formats des données, la prise en compte des contraintes référentielles et la maitrise de la sémantique des données.

c) Non volatiles : les données stockées dans l'entrepôt de données sont conservées de manière permanente afin de garder la traçabilité des décisions prises. Par conséquent, elles ne sont ni modifiées ni supprimées. Ainsi, la même requête appliquée sur les mêmes données à des intervalles de temps différents doit donner toujours les mêmes résultats.

d) Historisées (ou chronologiques): à la différence des bases de données de production, les données de l'entrepôt ne sont jamais mises à jour et chaque nouvelle donnée doit être insérée séparément. Cela garantira le suivi dans le temps de l'évolution des différentes valeurs des indicateurs. Par conséquent, un référentiel de temps doit être mis en place afin de pouvoir identifier chaque donnée dans le temps.

3. Les magasins de données ou Data-marts

Le concept de magasin de données est souvent associé à celui d'entrepôt de données et il est souvent utilisé dans le contexte des systèmes BI. Nous consacrons cette section à clarifier la nuance qui peut exister entre ces deux notions voisines.

3.1. Définition d'un magasin de données

Un magasin de données (ou *data-mart*) est un sous ensemble d'un entrepôt de données. Il est structuré et formaté pour répondre à un besoin spécifique ou à une fonction particulière de l'entreprise. Donc, il représente un point de vue spécifique selon des critères métiers bien précis ou un usage particulier.

Comme il est montré dans la *figure 3.3* suivante, on peut avoir plusieurs magasins de données dans une même entreprise. Par exemple, le premier enregistre les données du service marketing, le deuxième est dédié au service comptabilité et le dernier est destiné au service des achats.

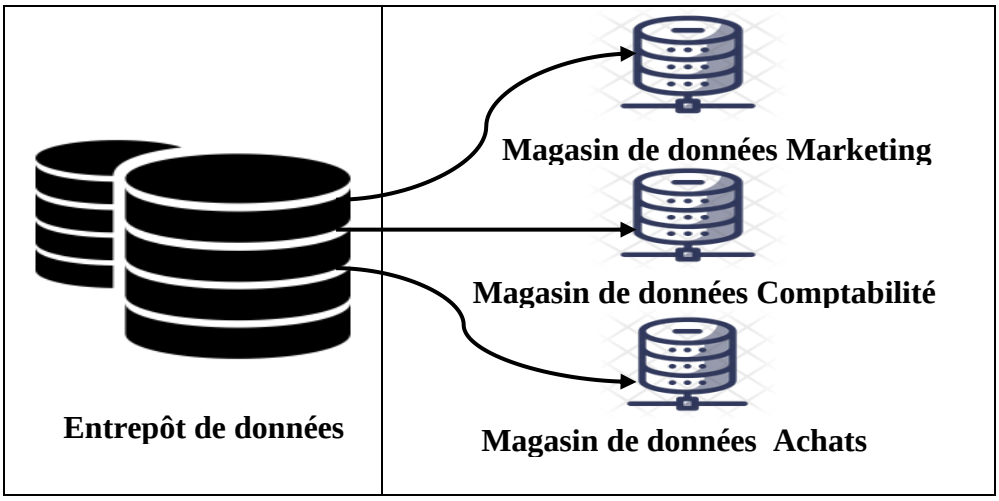


Figure 3. 3 Eclatement d'un entrepôt de données en magasins de données

3.2. Comparaison entrepôt et magasin de données (*Data warehouse vs Data-mart*)

L'intérêt d'utiliser plusieurs magasins de données au lieu d'un seul entrepôt de données réside dans la simplicité de leur compréhension et de leur manipulation. En plus, les magasins de données sont destinés à des utilisateurs qui sont plus ciblés.

En sus de ces différences de base relatives à l'exploitation des entrepôts et des magasins de données, d'autres différences fondamentales concernent le volume des données manipulées et stockées par chacune des structures. En effet, l'entrepôt de données enregistre toutes les données et offre une vision globale de l'entreprise, alors que le magasin de données ne traite qu'une portion spécifique des données qui sont relatives à seul sujet d'analyse (*un domaine de gestion particulier : vente, comptabilité, GRH, …etc.*).

La différence précédente concerne le volume des données et elle est causée par le nombre de sources exploitées par chacune des deux structures. Ainsi, pour l'entrepôt de données, les sources de données sont diversifiées et regroupent toutes les sources possibles, aussi bien internes qu'externes, alors que les sources de données d'un magasin de données sont restreintes et se limitent, la plus part du temps, à un seul département (*une seule fonction de l'entreprise*).

Evidemment, le nombre des sources de données impacte directement le niveau de complexité du processus ETL qui sera à déployé. En effet, ce processus est très simple dans le cas d'un magasin de données et devient de plus en plus complexe avec l'accroissement du nombre de sources de données utilisées dans le cas d'un entrepôt de données.

Néanmoins, il faut signaler que bien que les processus de conception d'un entrepôt de données et d'un magasin de données demeurent les mêmes, les ressources déployées pour chacune des structures varient en fonction du volume et du nombre de sources de données gérées.

4. Architecture d'un entrepôt de données

L'intérêt d'étudier l'architecture de l'entrepôt est de spécifier et de décrire les composants du système ainsi que leurs interactions.

A rappeler qu'en amont de l'entrepôt de données, la phase d'intégration des données à partir des différentes sources permet de récupérer les données, de les transformer et enfin de les charger dans l'entrepôt de données lui-même. Ces fonctions sont, généralement, réalisées par des outils ETL dédiés. Le chargement des données, proprement dit, se fera au niveau d'une structure de stockage qui n'est autre que l'entrepôt de données. D'autre part, en aval de l'entrepôt, les outils logiciels d'exploitation et d'analyse assureront l'exploitation des données contenues dans l'entrepôt en vue de leurs analyses futures.

Cette section est consacrée à la description de l'architecture fonctionnelle de l'entrepôt de données ainsi qu'à l'exposé des différentes configurations possibles répondant aux besoins spécifique de chaque entreprise.

4.1. Description de l'architecture de l'entrepôt de données

La description de l'architecture de l'entrepôt de données consiste à déterminer les différents rôles de chaque composant et à décrire les dépendances et interactions existantes.

Comme illustré dans la *figure 3.4* ci-après, une architecture fonctionnelle standard d'un entrepôt de données est constituée des trois niveaux suivants :

- **Le niveau inférieur de l'architecture**: représente le serveur de base de données, soit l'endroit où les données sont chargées et stockées. Il s'agit de l'entrepôt de données proprement dit (*colonne (a) dans la figure 3.4).*
- **Le niveau intermédiaire** : comprend le moteur d'analyse utilisé pour accéder et analyser les données (Serveur OLAP et outils d'analyse : *colonne (b)*)
- **Le niveau supérieur** : c'est le niveau client frontal qui affiche les résultats aux utilisateurs finals via des outils de création de rapports, d'analyse et d'exploration de données (*voir colonne (c)*)

Le premier niveau exprime la phase d'intégration des données, alors que les deux derniers correpsondent à la pahse d'exploitation et d'analyse des données de l'entrepôt.

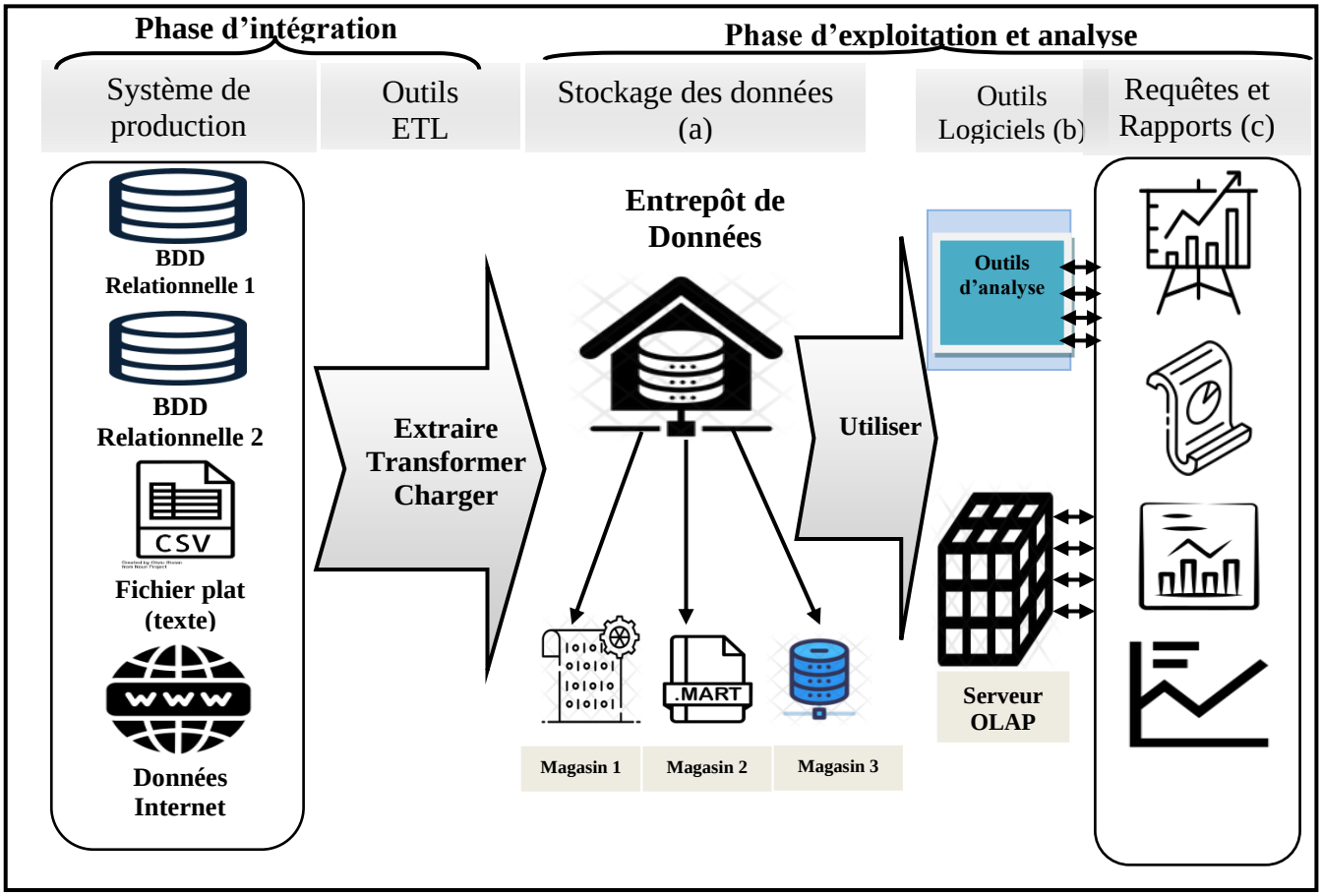


Figure 3.4 Schéma d'une architecture standard d'un entrepôt de données

4.2. Les différentes architectures des entrepôts de données

En fonction de l'utilisation ou non des magasins de données et des interactions possibles entre magasins et entrepôts de données, différentes configurations sont envisageables

engendrant, ainsi, différents schémas architecturaux. D'autres facteurs, tels que le volume des données et leurs sources respectives, sont à considérer lors de la spécification de l'organisation des composants de l'entrepôt de données.
Nous exposons, ci-après, les architectures d'entrepôt de données les plus fréquemment utilisées dans les entreprises. Pour chacune d'elles, une description est présentée, puis une évaluation contenant ses avantages et ses inconvénients est dressée.

4.2.1. Architecture en entrepôt de données centralisé

C'est l'architecture standard et la plus répandue d'un entrepôt de données (*voir figure 3.5*). Elle consiste en un gigantesque réservoir unique contenant toutes les données servant l'entreprise entière, et où les magasins de données spécialisées sont complètement absents.

a) Avantages

- Les utilisateurs peuvent accéder à toutes les données de l'organisation.
- A cause de la centralisation des données en un seul endroit, le processus d'intégration ETL et la maintenance des données sont considérablement facilités.
- Les performances du système sont optimales (le *temps de réponses aux requêtes est réduit*).

b) Inconvénients

- Le processus de développement de l'entrepôt de données est lent est couteux.
- L'extensibilité de l'entrepôt est limitée et la gestion de son évolution est coûteuse.

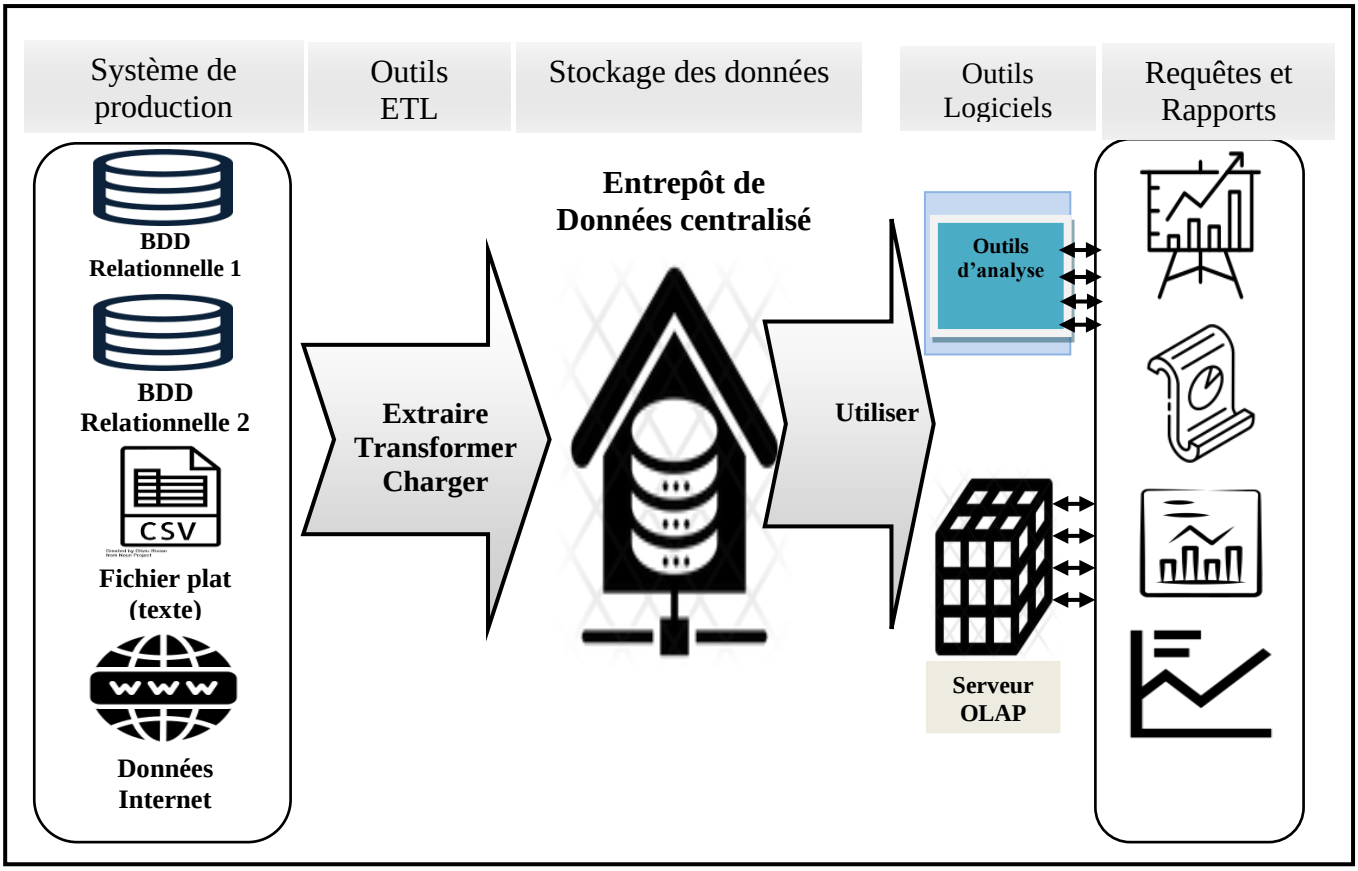


Figure 3.5 Schéma d'un entrepôt centralisé de données

4.2.2. Architecture en magasins de données indépendants

Contrairement à l'architecture en entrepôt de données centralisé, décrite précédemment, cette configuration est basée sur l'utilisation de plusieurs magasins de données qui sont disposés

en silos fonctionnels et qui opèrent de manière indépendante. En effet, les magasins de données ont été développés individuellement, car ne disposant pas de dimensions communes.

c) Avantages

- Cette architecture est très simple à développer et elle est moins coûteuse.

d) Inconvénients

- L'analyse inter-fonctionnelle est difficile à réaliser, voire impossible.
- Possibilité d'incohérence et de redondance des données contenues dans les différents magasins de données.
- Elle n'offre pas une vision globale de l'entreprise et la perception des activités est parcellaire et limitée.

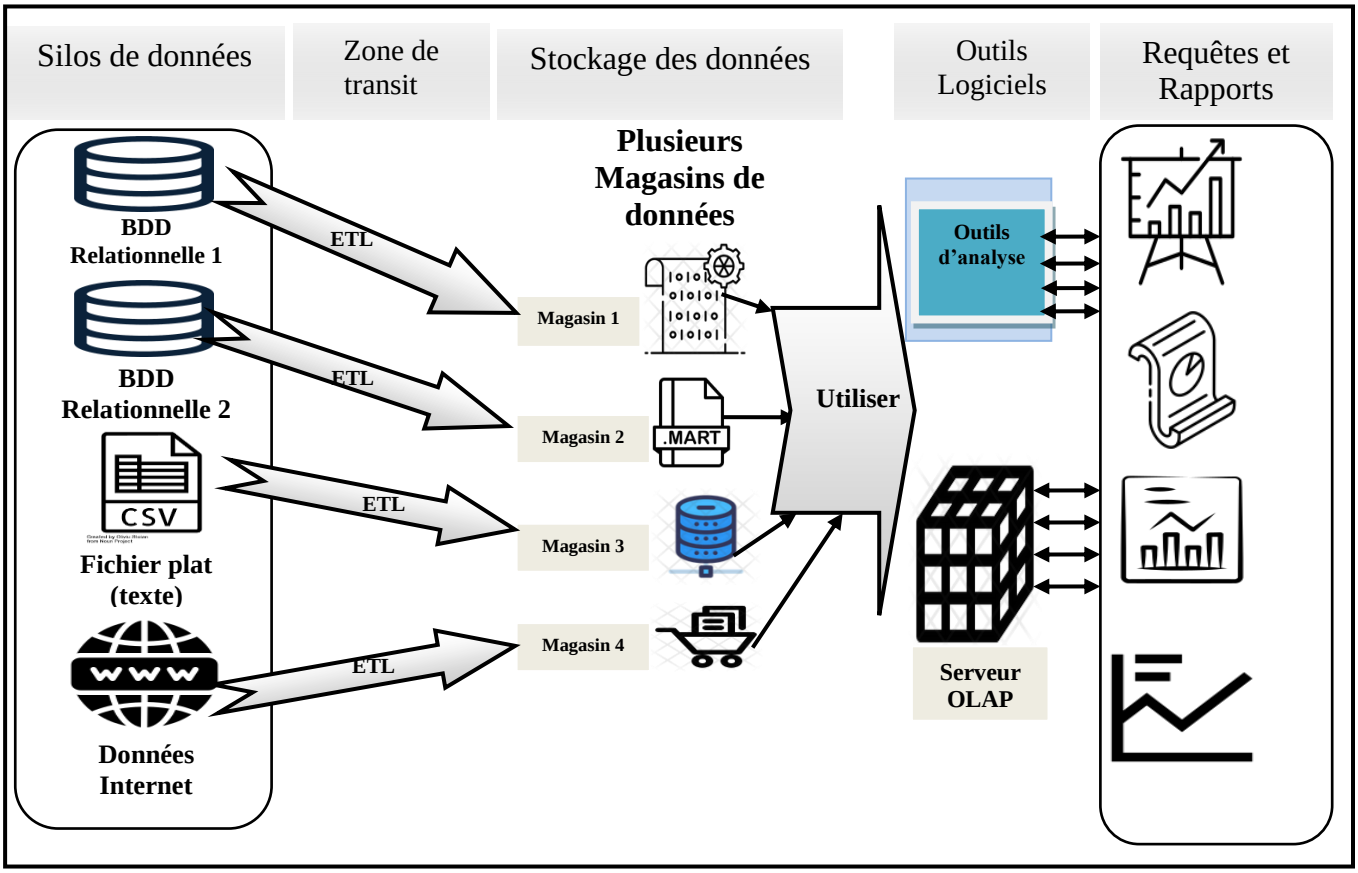


Figure 3.6 Schéma d'une architecture en magasins de données indépendants

4.2.3. Architecture en bus de magasin de données

Cette architecture vise à tirer profit des avantages respectifs des deux configurations précédentes, en opérant leur hybridation. Elle est articulée autour de plusieurs magasins de données qui sont développés par sujet (*ou processus d'affaires*), en se basant sur des dimensions conformes. En plus, un entrepôt de données conceptuel est formé à partir des magasins de données inter-reliés grâce à une couche d'intergiciels spécifiques.

e) Avantages

- Cette approche incrémentale de conception de l'entrepôt de données donne rapidement des résultats, car elle commence par prendre en charge le processus fonctionnel le plus important dans l'organisation.
- L'intégration des données est assurée par les dimensions conformes (*données communes*).

f) Inconvénients

- Les itérations futures pour implémenter les magasins de données à venir sont plus difficiles à planifier et à réaliser.
- Les performances globales de l'architecture ne sont pas optimales, car plusieurs magasins de données sont exploités simultanément pour répondre à une requête particulière d'un utilisateur.

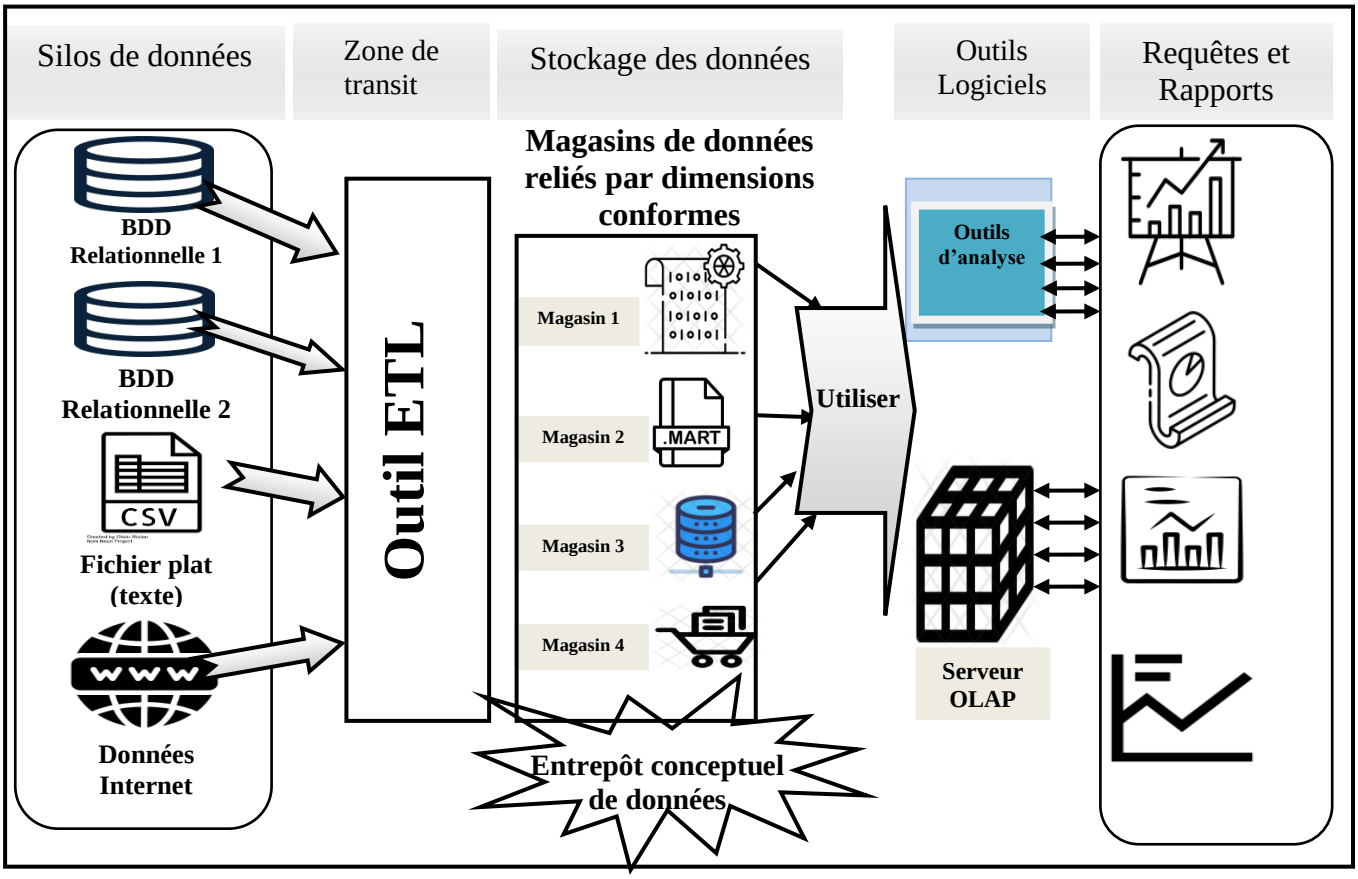


Figure 3.7 Schéma d'une architecture hybride d'un entrepôt de données

5. Fonctionnement de l'entrepôt de données

Abstractions faite aux différents types d'architecture décrites précédemment, le fonctionnement de l'entrepôt de données est basé sur l'accès et l'exploitation des données qui y sont stockées. L'intérêt est de donner la possibilité aux différents utilisateurs d'analyser et de visualiser les données stockées selon différentes perspectives.

Dans le cas ou l'entreprise dispose d'administrateurs de base de données (*ou d'entrepôt de données*), une utilisation directe et brute des données est possible grâce à la formulation des requêtes d'interrogation exprimées dans un langage dédié, tel que SQL. Néanmoins, dans la plus part des situations, le recours à l'utilisation et à l'assistance de plusieurs outils logiciels d'exploration des données est inévitable. Ces outils peuvent répondre à différentes préoccupations en matière d'accès aux données de l'entrepôt de données. On distingue, notamment :

- **Les outils pour les requêtes :** ces logiciels spécialisés dans l'interrogation des données de l'entrepôt permettent de filtrer les données qui vérifient certaines contraintes spécifiées par l'utilisateur (requêtes SQL d'interrogation de type *SELECT… FROM …WHERE*).

- **Les outils de reporting :** aident les utilisateurs à produire des rapports d'analyse pour l'entreprise. Les rapports peuvent être sous la forme de feuilles de calcul ou de graphiques visuels (*diagramme en secteur, barres, nuages…etc.)*
- **Outils de développement d'applications :** destinés à la création des rapports personnalisés et de les présenter dans formats facilement interprétables.
- **Outils d'exploration de données pour l'entreposage de données :** ces outils sont destinés à systématiser la procédure d'identification des tableaux et des liens dans d'énormes quantités de données en utilisant des méthodes de modélisation statistique de pointe.
- **Outils OLAP** : assurent l'analyse des données d'entreprise à partir de nombreux points de vue.

A signaler que d'autres types d'outils d'analyses avancées peuvent être utilisés pour accéder et interroger les données de l'entrepôt, tels que les logiciels de fouille et d'analyse de données.

NB : Pour plus de détails sur les fonctionnalités des logiciels, voir la sous-section 2.2 du premier chapitre. Concernant les outils OLAP, consultez la section 5 du prochain chapitre.

6. Conclusion

Dans ce chapitre, nous nous sommes focalisés sur l'élément fondamental de toute solution BI qui est l'entrepôt de données. Ses caractéristiques, ses objectifs ainsi que les différentes architectures possibles ont été exposés et son fonctionnement décrit. D'autre part, nous avons mis en relief la notion de magasin de données (*ou data-mart*), tout en le comparant avec l'entrepôt de données.

Le prochain chapitre sera consacré exclusivement à l'aspect modélisation des entrepôts de données.

Ce qu'il faut retenir

L'entrepôt de données est un vaste gisement de données issues des différentes sources internes et externes à l'organisation. Il est l'élément principal du système d'information décisionnel.

L'entrepôt de données peut être scindé en plusieurs magasins de données qui sont des sous- ensembles de données exprimant des points de vue spécifiques selon des critères métiers bien précis ou un usage particulier.

Série de TD N° 3 : Les entrepôts de données

Exercice 1 : Maitrise des concepts

a. Expliquez pourquoi dit-on qu'un entrepôt de données donne une vision transversale d'une organisation ?

b. Donnez trois motivations principales pour l'utilisation d'un entrepôt de données au lieu d'une base de données?

c. Quel est l'intérêt d'utiliser plusieurs magasins de données à la place d'un entrepôt de données unique ?

Exercice 2 : Propriétés d'une base de données vs propriétés d'un entrepôt de données

Souvent on assimile le fonctionnement d'un entrepôt de données à celui d'une base de données. Néanmoins, plusieurs différences existent entre ces deux types de structures de donnes.

a. Dressez une comparaison entre une base de données et un entrepôt de données en termes de fonctions assurées par chaque structure, la nature des données manipulées et le modèle de données utilisé ?

b. Du point de vue technique, les bases de données et les entrepôts de données abordent les questions de normalisation, de gestion de la concurrence et de la sécurité des données avec des niveaux d'importance différents.
Mettez en relief ces aspects pour chacune des structures de données ?

Exercice 3 : Une solution BI sans entrepôt de données

Une entreprise désire exploiter ses données à des fins de prise de décision de manière directe sans passer par leur centralisation et sans faire recours aux entrepôts de données. Cette manière de faire lui permettra d'analyser directement ses données stockées dans les différentes applications en utilisant des middlewares conventionnels dédiés.

a. Spécifiez un schéma illustratif de ce scénario d'exploitation des données à des fins décisionnelles sans entrepôts de données?

b. Quels seront les difficultés rencontrées par ce choix architectural ?

c. Quels seront les inconvénients d'une telle approche ?

Chapitre IV : Modélisation multidimensionnelle et outils OLAP

1. Introduction

Dans le chapitre précédent, nous avons mis en exergue le rôle et les caractéristiques de l'entrepôt de données et nous avons précisé qu'il constitue une mine d'informations consolidées et intégrées de l'entreprise. Par ailleurs, les différentes architectures fonctionnelles ont été abordées. Vue son importance, la modélisation de l'entrepôt est traitée dans le chapitre courant de manière approfondie.

Ce chapitre est consacré à la modélisation multidimensionnelle des entrepôts de données. On y présente le formalisme utilisé et les différents modèles d'entrepôt de données. Le chapitre se termine par un exposé des outils OLAP utiles à l'exploitation des données de l'EDD.

2. Fondement de la modélisation multidimensionnelle de l'entrepôt

Les objectifs de cohérence et de centralisation des données de l'entrepôt, conjugués avec les besoins de leur analyse, exigent une perception différente de la façon dont ces données doivent être modélisées et organisées. En effet, on passe d'une modélisation orientée processus fonctionnel à celle orientée *sujet* ou *thème*. Par conséquent, le principe de la modélisation de l'entrepôt de données sera basé sur une prise en compte de plusieurs axes d'analyse ou dimensions. Ainsi, on parle de modélisation *multidimensionnelle*.

2.1. Intérêt de la modélisation multidimensionnelle

Les EDD sont destinés à mettre en place des systèmes décisionnels qui sont censés répondre à des objectifs différents de ceux des systèmes transactionnels. En effet, dans un système opérationnel les données sont essentiellement destinées à satisfaire un processus fonctionnel et obéissent à des règles de gestion bien définies, alors que celles d'un entrepôt de données sont destinées à un processus analytique, en premier lieu.

Nous motivons, ci-dessous, le recours à une nouvelle approche de modélisation des données des EDD, comparativement à celle des BDD traditionnelles.

- Les objectifs de performances dans les BDD ne sont pas les mêmes que ceux des EDD. Effectivement, dans les BDD les requêtes sont simples à cause des méthodes d'accès et d'indexation, alors que dans les EDD les requêtes d'analyse de type OLAP sont plus complexes.
- Le besoin de combiner les données provenant des diverses sources de données et la nécessité d'effectuer des agrégations dans un seul EDD afin d'offrir des vues multidimensionnelles.
- Les données d'un EDD sont souvent non volatiles et ont donc une durée de vie plus longue que celle des BDD.

Partant des considérations précédentes, il est impératif d'aborder la modélisation de l'EDD d'une manière fondamentalement différente de celle des BDD et de recourir à un modèle de données qui sera simple et compréhensible, tout en permettant la prise en charge des besoins stratégiques des décideurs au lieu de se focaliser sur les détails inhérents à la gestion des données associées aux processus fonctionnels.

2.2. Principe de la modélisation dimensionnelle

Lors de la modélisation de l'EDD, on s'intéresse à des sujets (*ou thèmes*) au lieu d'applications métiers. Cette vision affecte forcément la perception, la conception et l'organisation des données contenues dans l'EDD. Donc, le processus de modélisation doit être spécifique pour répondre aux besoins de centralisation et d'analyse. Plus concrètement, la *modélisation dimensionnelle* consiste à considérer un sujet d'analyse comme un cube à plusieurs dimensions (*3, 4 et même 5 dimensions*). D'où le concept de modélisation multidimensionnelle. Cette modélisation offre des vues en tranches qui correspondent à des analyses selon différents axes.

3. Formalisme de modélisation multidimensionnelle

Alors que la modélisation des bases de données relationnelles utilise les concepts d'*entités* et de *relations* afin de construire des tables, la modélisation dimensionnelle d'un EDD utilise les concepts de *tables de faits* et *tables de dimension*, que nous allons détailler ci-après.

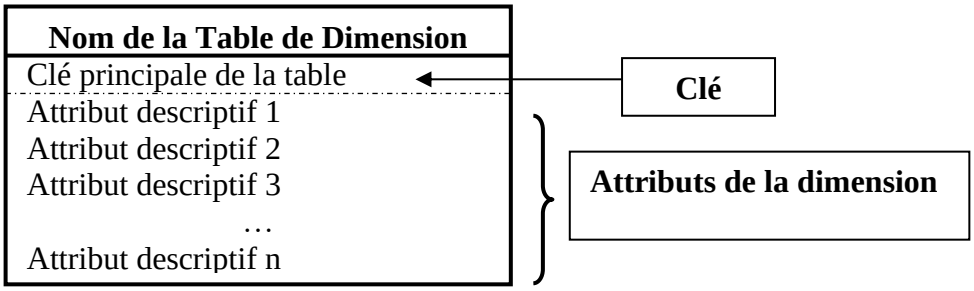
Pour chaque type de table, on expose son formalisme de représentation suivi d'exemples illustratifs

3.1. Les tables de dimensions

Les tables de dimensions représentent le point de vue selon lequel on veut voir les données décrites par un ensemble d'attributs. Donc, elles servent à enregistrer les descriptions textuelles des dimensions de l'activité de l'entreprise. Chaque réalisation de ces descriptions textuelles concerne une occurrence de la dimension concernée. Par exemple, chaque enregistrement de la dimension produit représente un produit spécifique. Puisque les attributs de dimensions sont utilisés pour décrire des propriétés d'une dimension, alors ils sont plus utiles sous forme de texte que sous toute autre forme.

3.1.1 Formalisme de représentation des tables de dimension

Une table de dimension est schématisée par un rectangle contenant dans sa partie haute le nom de la dimension, et dans la partie inférieure les noms des attributs qui la décrivent, avec l'identifiant en premier. (*Ce formalisme est inspiré de celui du modèle entité/association*)



3.1.2 Exemples de tables de dimension

Nous illustrons ci-dessous, le concept de tables de dimensions par trois exemples relatifs à une entreprise commerciale : table *Produit*, table *Magasin* et table *Temps*.

Dimension Produit
Code-Prod
Libellé
Marque
Catégorie
Unité de mesure
Type emballage
Prix unitaire

Dimension Magasin
Code Magasin
Nom-magasin
Adresse
Téléphone
Superficie

Dimension Temps
#ID-TEMPS
Jour de semaine
Mois
Trimestre
Semestre
Année

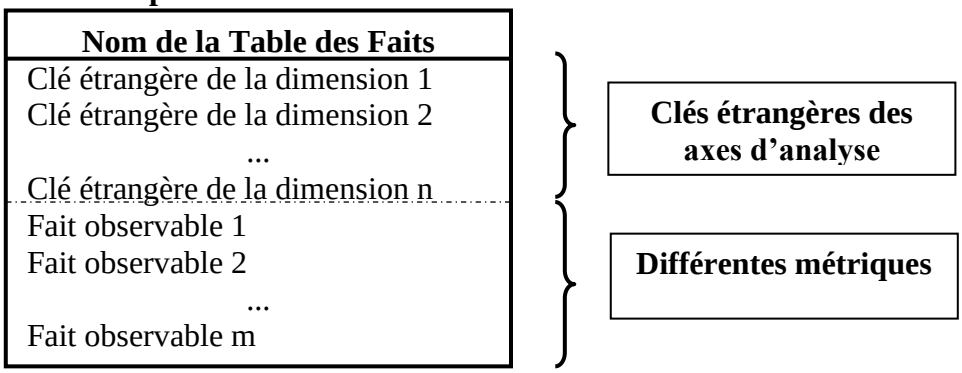
Figure 4.1 Illustration des tables de dimension d'un entrepôt de données

L'un des rôles essentiels des attributs figurant dans une table de dimension, par exemples les propriétés *Libellé, marque et catégorie* de la table *Produit*, est de servir comme source de contraintes d'une requête ou de fournir les en-têtes de lignes pour les états de sortie finaux.

3.2. Les tables de faits

Une table de fait sert à stocker les mesures de l'activité et contient les informations mesurables (*métriques*) sur ce qu'on veut analyser. Pour assurer la jonction des faits avec les dimensions d'analyse, la table de faits doit contenir, également, les clés étrangères des tables de dimensions. Ainsi, elle représente la table centrale du modèle dimensionnel à laquelle seront reliées les tables de dimensions.

3.2.1. Formalisme de représentation des tables de faits



3.2.2. Exemples de tables de faits

Par exemple, les tables de faits : *Ventes* et *Mouvements compte* suivantes expriment différentes métriques de l'activité commerciale. Dans la table *Ventes*, en plus des clés étrangères *date, N° produit et code magasin*, trois mesures utiles à l'évaluation de l'activité commerciale sont ajoutées : la *quantité vendue, le montant des ventes* et *l'unité monétaire*.

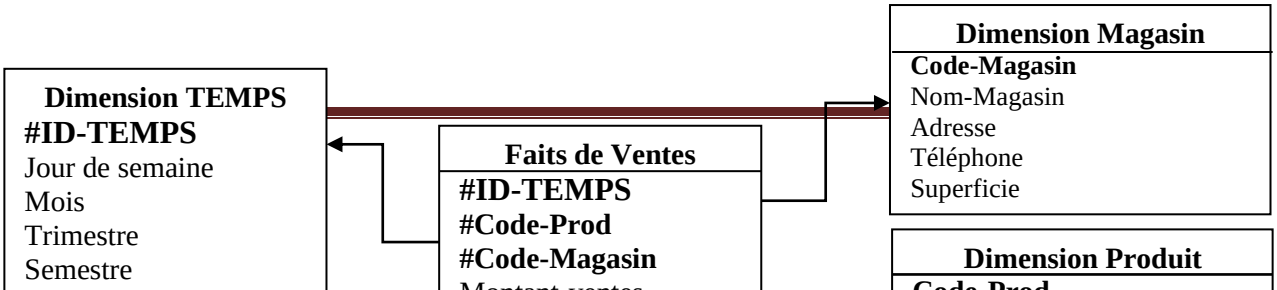
Fait de Ventes
#ID-TEMPS
Code-Prod
Code Magasin
Quantité vendue
Montant des ventes
Unité monétaire

Fait de Mouvement compte
#ID-TEMPS
NUM-COMPTE
CODE-CLIENT
Type-Mouvement
Montant mouvement

Figure 4.2 Deux table de faits du domaine commercial

Les faits les plus importants et les plus utiles dans un EDD sont *numériques* et *valorisés de façon continue*. En plus ils sont *additifs*. Ces trois caractéristiques sont cruciales pour la conception de l'EDD. L'intérêt de ce point de vue basé sur le trio (*numériques, valorisés d'une manière continue et additifs*) est que pratiquement toute requête visant cette table de faits nécessitera l'utilisation par le SGBD de centaines, de milliers, voire de millions d'enregistrements pour construire le jeu de réponse à fournir à l'utilisateur. Après exploration et filtrage, ce grand nombre d'enregistrements sera comprimé en quelques lignes sur le jeu de réponse de l'utilisateur. Il est évident que la seule façon de condenser ces enregistrements est de les additionner, du fait qu'ils sont additifs.

3.3. Exemple complet de modèle multidimensionnel



En termes d'occurrences ou de réalisations, chaque enregistrement de la table de faits *Ventes* exprime le total des quantités vendues, le montant total des ventes ainsi que l'unité monétaire utilisée, et ce pour le produit en question dans un magasin et pour une référence temps donnée. A titre d'illustration, nous montrons ci-dessous un sous-ensemble d'occurrences possibles de cette table de fait.

Date	Code-Prod	Code Magasin	Quantité vendue	Montant ventes	Unité monétaire
10/10/2021	A100	10	500	50.000,00	Da
10/10/2021	A200	10	400	80.000,00	Da
10/10/2021	B400	20	200	16.000,00	Euro
10/10/2021	B600	30	20	20.000,00	Euro
20/12/2021	A200	10	50	500,00	Da
20/12/2021	C300	10	100	10.000,00	Euro
20/12/2021	C500	30	300	45.000,00	Euro
20/12/2021	B600	40	400	8.0000,00	Euro
20/12/2021	D800	10	1000	100.000,00	Dollars
20/12/2021	D900	10	1200	60.0000	Dollars

Table 4.1 Exemples d'occurrences de la table de faits Ventes

A noter que la table de fait précédente peut contenir des milliers, voire des millions d'enregistrements qui reflètent les historiques des ventes de l'entreprise. Cette table stockée dans l'entrepôt de données est exploitée par des outils OLAP adéquats, en formulant des requêtes SQL.

La requête analytique suivante montre un exemple permettant de calculer le total des ventes par année de chaque produit.

```
SELECT Produit.code_Prod,Produit.libellé, Temps.annee, SUM(Ventes.montant_ventes)
       as total
FROM Produit, Temps, Ventes
WHERE Produit. Code_Prod = Ventes. Code_Prod AND Temps.année = Ventes.année
GROUP BY Produit.code_Prod, Temps.annee
```

On peut améliorer la requête précédente pour restreindre le montant total des ventes à une certaine catégorie de produits. Par exemple, la requête suivante limite le total des ventes aux seuls produits dont le libellé est « *Scanner* »

```
SELECT Produit.code_Prod,Produit.libellé, Temps.annee, SUM(Ventes.montant_ventes)
       as total
```

```
FROM Produit, Temps, Ventes
WHERE Produit. Code_Prod = Ventes. Code_Prod AND Temps.année = Ventes.année
      AND Produit.libellé = 'Scanner'
GROUP BY Produit.code_Prod, Temps.annee
```

4. Les approches de modélisation des entrepôts de données

Comme l'entrepôt de données est orienté vers le sujet afin de faciliter l'analyse des données en ligne, alors son schéma (*ou modèle conceptuel*) doit être unique, tout en intégrant les tables de faits et de dimensions, leurs contraintes et leurs relations. Ce schéma multidimensionnel servira à décrire de manière logique la base de données entière et aussi à assurer sa maintenance et son évolution. Chaque type de structuration et de mise en association des tables de dimensions et de faits, ainsi que leur contenu permettront de déterminer des modèles différents des EDD.

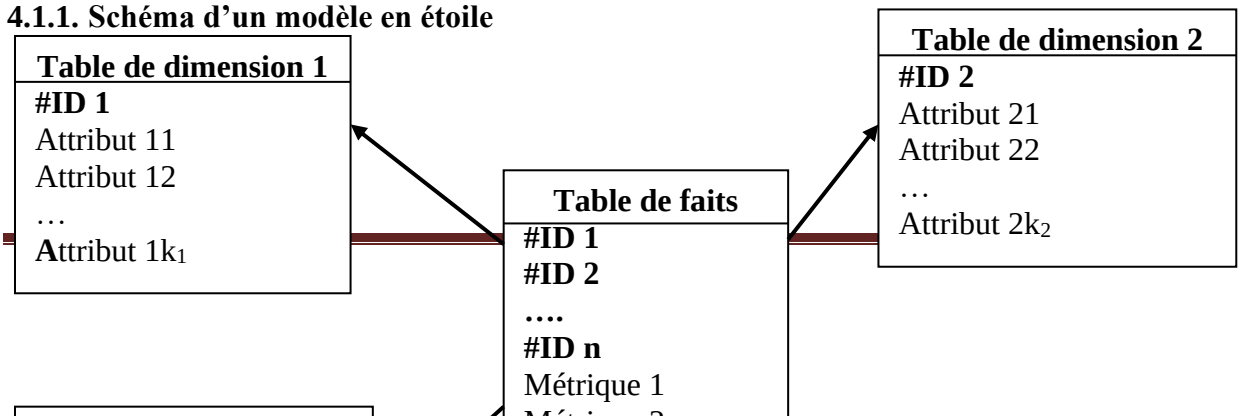
En termes de typologie, les schémas en étoile et en flocon de neige sont les modèles de données multidimensionnels qui sont généralement les plus utilisés pour la modélisation des EDD. Un troisième type de modèle possible est basé sur la combinaison des deux modèles précédents. Ces trois types de modèles sont présentés et illustrés ci-dessous. Pour chaque type, nous présentons sa spécification et nos l'illustrons avec un exemple réel.

4.1. Le modèle en étoile

Dans ce modèle simple, l'entrepôt de données comprend une table de faits unique et autant de tables de dimension qu'il existe de dimension d'analyse. La table de fait centrale est entourée par les différentes tables de dimension qui lui sont connectées via la clé primaire et les clés étrangères. Ainsi, la table de faits contient les clés de chaque dimension et les attributs des mesures à analyser.

A signaler que cette représentation est fortement *dénormalisée***,** car elle ne respecte pas les trois premières formes normales. Par conséquent, le problème de la redondance de données et les problèmes sous-jacents (*perte d'espace, possibilité d'incohérence des données*) demeurent résiduels. Néanmoins, ce modèle assure un haut niveau de performance des requêtes même sur de gros volumes de données.

4.1.1. Schéma d'un modèle en étoile



4.1.2. Exemple de schéma en étoile

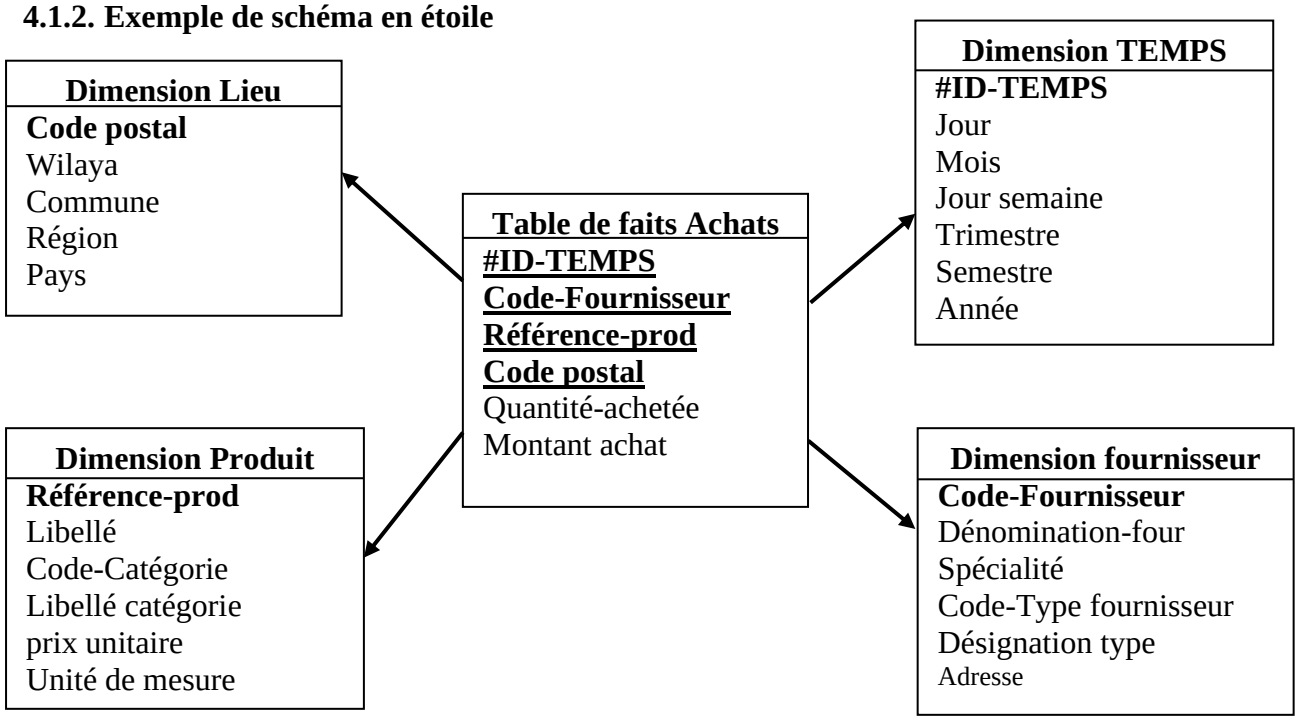


Figure 4.5 Modèle dimensionnel en étoile de l'activité Approvisionnement

4.2. Le modèle en flocon de neige

Le principe de la modélisation en flocon est de créer des hiérarchies de dimensions, de telle manière à avoir moins de lignes par dimensions. Contrairement, au schéma précédent, ce modèle utilise la normalisation qui divise les données des tables de dimensions en tables supplémentaires, afin d'éliminer les dépendances transitives qui peuvent exister entre les clés primaires et les autres données des tables de dimension. Ce fractionnement des tables de

dimension permet de réduire la redondance et de prévenir les pertes d'espace mémoire. Ainsi, on obtient un type de schéma en étoile qui inclut la forme hiérarchique des tables dimensionnelles, d'où le nom de flocon de neige car sa structure ressemble à un flocon de neige. Un schéma en flocon de neige est plus facile à gérer mais complexe à concevoir et à comprendre. Cela peut également réduire l'efficacité de la navigation car davantage de jointures seront nécessaires pour exécuter une requête.

Concernant sa structure, ce schéma est articulé autour d'une table de faits centrale reliée à différentes tables de dimension qui sont à leur tour reliées à des tables de sous-dimension. En jouant le rôle de clés étrangères, les clés primaires des tables de dimensions et de sous dimensions assurent la jonction avec la table de faits centrale.

En résumé, ces modèles ne sont pas performants dans le contexte des solutions BI à cause des nombreuses jointures à effectuer et qui vont engendrer une complexité des requêtes d'interrogation.

4.2.1. Schéma du modèle en flocon de neige

Nous exposons, ci-après le schéma général d'un modèle en flocon de neige, dont les dimensions principales sont désignées par *dimension i* et leur identifiant *IDi* (i :1…n). Par ailleurs, les tables de sous dimensions de chaque *dimension i* sont notées *dimension ij* et leur clés *IDij* (j :1…m). Pour les attributs d'une table de *dimension i*, ils sont désignés respectivement par *attribut i1, attribut i2,… attribut il*, alors que pour une table de *sous-dimension ij*, les attributs sont notés *attribut ij1, attribut ij2,… attribut ijk.*

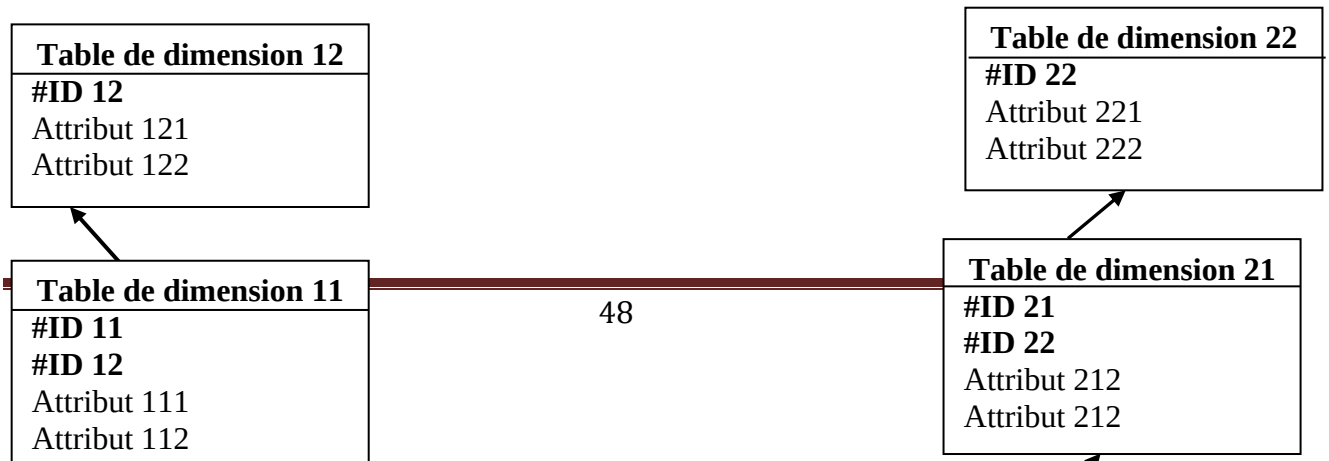
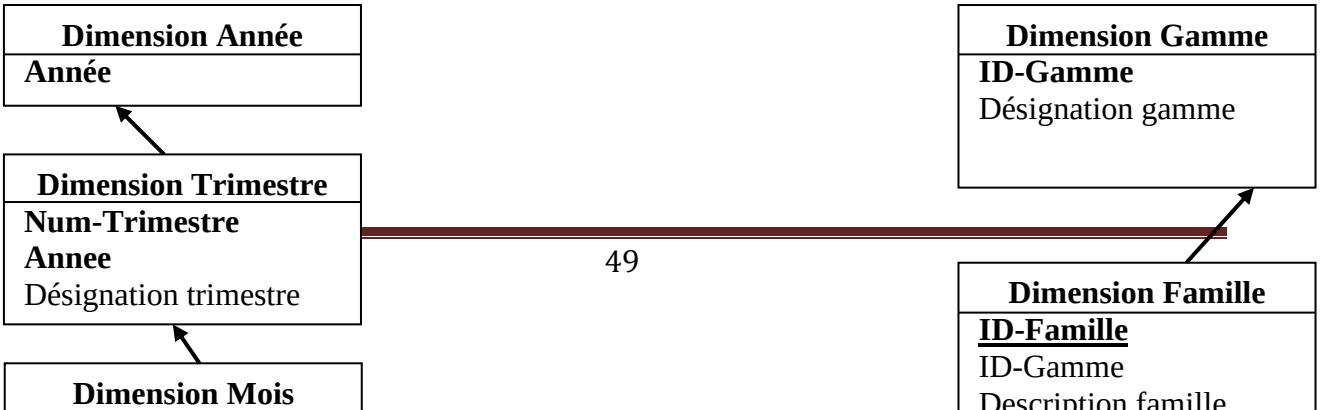


Table de dimension 3
#ID 3 **#ID 31** Attribut 31 Attribut 32 … Attribut $3k_3$

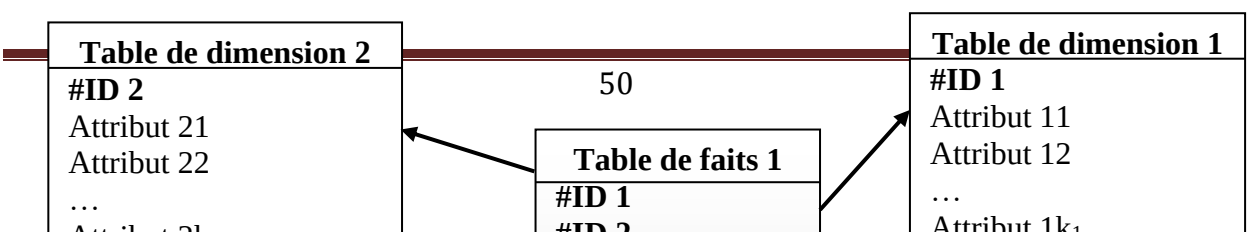
4.2.2. Exemple de schéma en flocon de neige



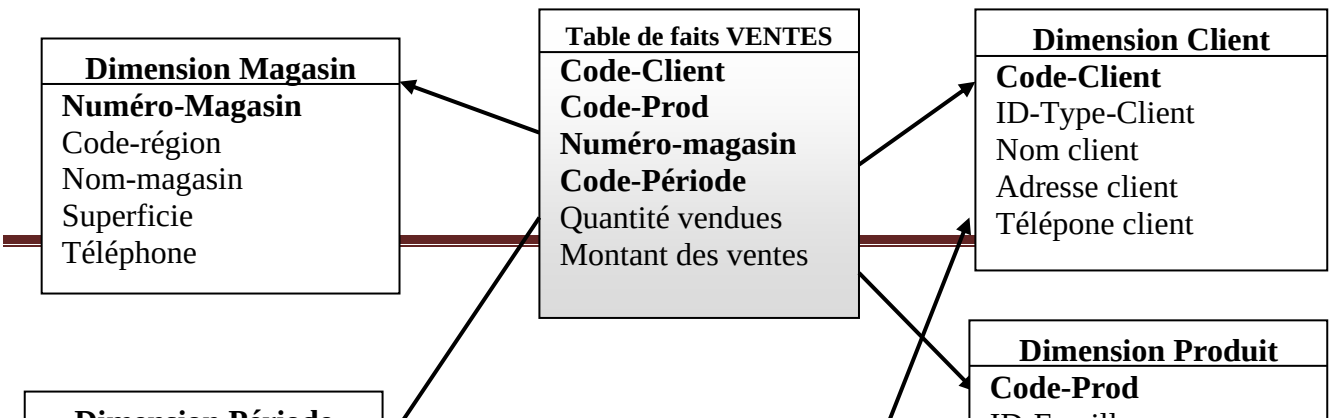
4.3. Modèle multidimensionnel en constellation

Ce modèle est basé sur la combinaison de plusieurs modèles en étoile qui utilisent des dimensions communes. Il contient plusieurs tables de dimensions et aussi plusieurs tables de faits. Les tables de dimensions peuvent êtres communes aux tables de faits ou non.

4.3.1. Schéma du modèle en constellation



4.3.2. Exemple de schéma en constellation



5. Les outils OLAP

Les bases de données multidimensionnelles s'opposent aux bases de données relationnelles à deux dimensions. Cela est dû, fondamentalement, au fait que les applications de traitement de transactions en ligne (*Online Transaction Processing* OLTP) s'inscrivent dans une optique d'un *système opérationnel*, c'est-à-dire destiné aux métiers de l'entreprise pour assister les utilisateurs dans leurs tâches de gestion, alors qu'un logiciel OLAP (*Online Analytical Processing* OLTP) est un type d'application informatique orienté vers *l'analyse sur-le-champ d'informations selon plusieurs axes (ou dimensions)***.** Le but d'un outil OLAP est, donc, de produire des rapports de synthèse qui aident la direction à avoir une vue transversale de l'activité de l'entreprise.

5.1. Définition d'un outil OLAP

Pour atteindre les objectifs d'analyse de données, les applications OLAP exploitent les données recueillies à partir de multiples sources de données et qui sont stockées dans des EDD, puis ils les nettoient et les organisent en cubes de données. Donc, les outils OLAP constituent une catégorie d'applications destinées à traiter et à restituer les données stockées dans l'entrepôt de données (*ou dans les magasins*) à des fins d'analyse des données.

Définition : *« Un logiciel OLAP est une technologie de traitement informatique qui permet à un utilisateur d'analyser, de consulter et d'extraire facilement des données de l'entrepôt pour les comparer de différentes façons à des fins d'analyse* **».**

Pour faciliter ce type d'analyses, les données OLAP sont stockées dans *une base multidimensionnelle* aussi appelées *Cubes OLAP* dont l'exploitation est effectuée en utilisant un *serveur OLAP***.**

5.2. Le cube OLAP et ses caractéristiques

Un cube OLAP est une représentation abstraite d'un ensemble d'informations multidimensionnelles exclusivement numériques utilisées par l'approche OLAP. Cette structure est prévue à des fins d'analyses interactives par une ou plusieurs personnes du métier qui désirent exploiter les données.

Les cubes OLAP ont les caractéristiques suivantes :

- Simplicité et rapidité d'accès ;
- Capacité à manipuler les données agrégées selon différentes dimensions ;
- Servent à obtenir des informations déjà agrégées selon les besoins de l'utilisateur ;
- Un cube utilise les fonctions classiques d'agrégation : *min, max, count, sum, avg,* mais peut utiliser d'autres fonctions d'agrégation spécifiques.

Du fait que les cubes OLAP sont des bases de données multidimensionnelles destinées à des analyses complexes sur des données, alors les systèmes OLAP doivent:

- Supporter les exigences complexes des décideurs en termes d'analyse,
- Analyser les données à partir de différentes perspectives ou dimensions métiers,
- Supporter les analyses complexes impliquant des ensembles de données volumineux.

La figure 4.10, ci-après montre un cube de données avec les trois dimensions d'analyse suivantes: *Produit, Magasin et Temps.*

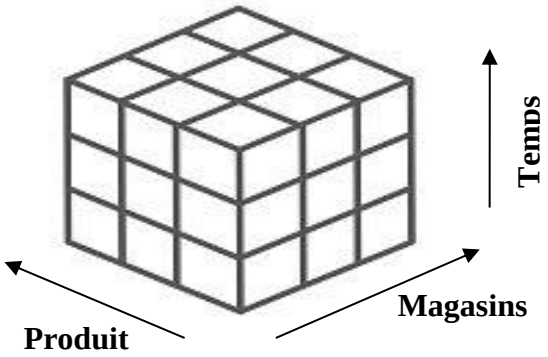


Figure 4.10 Exemple de cube de donnée avec les dimensions d'analyse

5.3. Principe de fonctionnement d'un outil OLAP

Le fonctionnement d'un logiciel OLAP est articulé autour de la notion de *cube***,** telle que explicitée précédemment. En effet, un système OLAP agit sur un cube à N dimension où toutes les intersections sont calculées. Ainsi, il offre un accès rapide à l'information située au niveau de l'intersection souhaitée. Plus concrètement, en réponse à une requête d'analyse, le système OLAP peut localiser l'intersection des dimensions et afficher les données qui y sont stockées. Il est ainsi possible d'analyser et de comparer les données de différentes façons. Les attributs ou dimensions peuvent aussi être séparés en plusieurs sous-attributs, on parle alors d'hiérarchie de la dimension.

L'analyse multidimensionnelle utilise les cinq structures suivantes :

a) La dimension : elle exprime les données utilisées comme contraintes dans les requêtes d'analyse ou comme en têtes dans les rapports de sortie.

Comme illustré dans la figure 4.10 précédente, *Produit, Magasin* et *temps* sont trois dimensions distinctes pour la gestion des ventes.

b) Hiérarchie : une dimension peut être élémentaire ou bien hiérarchisée dans le cas où elle est composée d'un ensemble de niveaux.
Par exemples : pour la dimension *Produit*, on a la hiérarchie suivante :
Branche → Famille → Sous-famille → Gamme. Alors que pour la dimension *Magasin*, on aura la hiérarchie : *Pays →Région →Ville*.

c) Niveaux : les niveaux expriment le degré de granularité dans une hiérarchie.
Par exemple, *branche* et *gamme* sont deux niveaux de la dimension *Produit*, alors que ***Région*** et ***Ville*** sont des niveaux de la dimension *Magasin*.

d) Membres : expriment une instanciation des attributs des niveaux d'une dimension. C'est-à-dire les valeurs prises par le niveau d'une hiérarchie relative à une dimension.
Par exemple : Les branches de produit sont notés par : *Produit:: branche* et prennent pour membres les valeurs suivantes: *Produit:: branche : informatique, alimentation, mercerie.*
De même pour le membre *produit ::branche.catégorie* qui peut prendre comme valeurs produit ::branche.catégorie : informatique.imprimante ou mercerie.bouton

e) Les cellules : une cellule d'un cube correspond à l'intersection des membres des différentes dimensions du même cube.

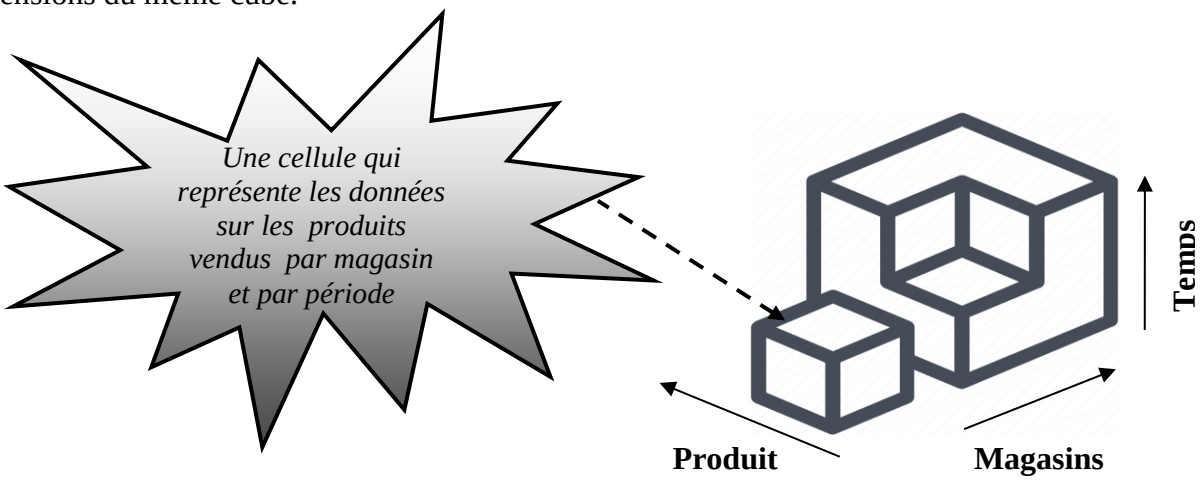


Figure 4.11 Illustration du contenu des cubes de données

Les cubes OLAP sont souvent pré-résumés dans toutes les dimensions afin d'améliorer considérablement le temps de réponse aux requêtes par rapport aux bases de données relationnelles. Les systèmes OLAP sont conçus pour repérer les intersections entre ces multiples dimensions. Ainsi, un cube OLAP est vu comme une base de *données multidimensionnelle* optimisée pour les entrepôts de données et les applications OLAP. Il s'agit d'une méthode permettant de stocker les données sous forme multidimensionnelle, notamment pour le reporting.

Exemple de cube de données

À titre d'illustration, en se basant sur le modèle en flocon de neige de la *figure 4.7*, il est possible d'analyser le chiffre d'affaires de l'entreprise commerciale selon les quatre dimensions suivantes :

- **Géographie** : Région > Ville
- **Temps** : Année > Trimestre > Mois > Période
- **Gamme de produits** : Gamme > Famille
- **Organisation** : Type-client > Client

Dans cet exemple, chaque cube OLAP contient des données classées par dimensions (*telles que la région géographique, la période de temps, la gamme de produit, ou l'entité organisationnelle)* et qui sont dérivées par tables dimensionnelles dans les EDD. Les dimensions sont ensuite complétées par les membres (*tels que les noms de clients, les pays et les mois*) qui sont organisés de manière hiérarchique.

5.4. Déploiement du langage d'analyse de données

Pour effectuer des requêtes au sein des cubes OLAP, on utilise le *langage MDX (multimensional expressions)* destiné aux développeurs décisionnels qui pourront s'en aider pour créer des membres calculés ou concevoir des rapports plus complexes, par exemple.

Ce langage fut développé par Microsoft à la fin des années 1990 avant d'être adopté par les autres vendeurs de bases de données multidimensionnelles.
Les cubes sont conçus pour pouvoir être *utilisés par tous les employés de l'entreprise*, et non uniquement par les responsables de la division informatique. Ils sont capables de rapporter des millions d'enregistrements en une seule fois.
Le langage MDX est semblable à SQL et la base d'une requête MDX est une instruction **SELECT**, mais dont la syntaxe est plus complexe. Le format général d'une instruction MDX est le suivant.

```
SELECT
{<Liste de mesures>} on 0
<dimension1> * <dimension2> * <dimension-n> on 1
FROM [Cube]
WHERE {<filtre>}
```

Les analystes peuvent ensuite effectuer différents types d'opérations d'analyse OLAP à partir de ces bases de données multidimensionnelles. Les opérations les plus courantes sont les suivantes.

- **La rotation** (**Pivot**) : permet de présenter une autre face du cube.
- **Le découpage (Slice) :** consiste à ne travailler que sur une tranche du cube. Dans ce mode, une des dimensions est réduite alors à une seule valeur.
- **Sous cube** (**Dice**) : permet l'extraction d'un bloc de données. Cette opération ne travaille que sur un sous-cube.
- **Forage vers le haut (Roll-up) :** permet d'obtenir un niveau de granularité supérieur, en utilisant les fonctions d'agrégation (*sum, average,…etc*).
- **Forage vers le bas (Drill-down**) : permet un forage vers le bas, pour obtenir un niveau de granularité inférieur, c'est-à-dire des données plus détaillées.

5.5. Quelques outils OLAP

Le marché des logiciels manifeste une large variété de suites applicatives garantissant les fonctions de BI ou bien des solutions dédiées spécialement aux analyses OLAP. Nous exposons ci-dessous, quelques exemples de logiciels prioritaires et open source qui sont les plus connus dans le marché.

5.5.1. Logiciels propriétaires

Voici quelques outils propriétaires d'informatique décisionnelle.

- **IBM Cognos :** c'est une suite BI qui assure l'élaboration des prévisions, des plans et des budgets, permettant de piloter la performance de l'entreprise. L'outil OLAP utilisé dans la suite cagnos est PowerPlay.

- **Power-OLAP :** est un outil complet qui permet de manipuler des données relationnelles à l'aide d'un browser et une feuille Excel. Il intègre un moteur OLAP qui organise les données en temps réel.
- **Oracle :** Oracle propose, depuis la version 9i, un moteur OLAP directement intégré à sa base de données relationnelle.
- **SQL Server (Analysis Services) :** Microsoft SQL Server 7.0 contient un composant OLAP qui fournit des fonctions d'exploration de données pour les applications décisionnelles. Ce composant permettant de concevoir, de créer et de gérer des structures multidimensionnelles qui contiennent des données agrégées provenant d'autres sources de données, telles que des bases de données relationnelles.

5.5.2. Logiciels Open source

Ci-dessous, une liste non exhaustive de quelques logiciels open source d'informatique décisionnelle.

- **Mondrian :** Mondrian est un serveur OLAP écrit en langage Java. Il utilise le langage d'interrogation MDX. Mondrian, précurseur du décisionnel Open source est désormais intégré au projet Pentaho.
- **Palo :** commercialisée par l'éditeur Jedox, Palo OLAP Server est un serveur de bases de données multidimensionnelles dédié à la gestion décisionnelle pour le planning, l'analyse, le reporting et la consolidation des données.
- **Spago BI :** est une solution de Business Intelligence open Source italienne. Spago BI permet de faire des analyses multidimensionnelles comme Mondrain.

6. Conclusion

La modélisation multidimensionnelle est la pierre angulaire de l'informatique décisionnelle. Elle permet de concevoir le schéma conceptuel du futur entrepôt de données qui supporte les données utiles à toute solution BI.

Dans ce chapitre, nous avons présenté l'utilité de cette phase dans le processus global d'élaboration d'une solution BI, puis nous avons exposé le formalisme utilisé pour aborder cette modélisation. Enfin, nous nous sommes largement étalés sur les différentes approches de modélisation. Le chapitre a été clôturé par la présentation des outils OLAP incontournables pour l'exploitation des données de l'entrepôt et quelques exemples d'outils commerciaux ont été exposés aux lecteurs désirant se familiariser avec la pratique des entrepôts de données.

Ce qu'il faut retenir

Les besoins de performance d'une solution BI et la nature des données manipulées au niveau de l'entrepôt de données exigent une orientation sujet (ou thème), lors de la perception et la conception de ces entrepôts de données.

Les tables de dimension et les tables de faits sont les deux catégories de structures gérées au niveau de l'entrepôt de données et les outils OLAP offrent une exploitation efficace de cette structure de données à des fins de prise de décision en entreprise.

Série de TD N° 4 : Modélisation multidimensionnelle et outils OLAP

Exercice 1 :

a. Quels sont les facteurs à prendre en considération lors de la conception de l'entrepôt de données ?

b. Quelles différences existent entre une table de faits et une table de dimensions ?

Exercice 2 : Mettez en évidence les différences qui peuvent exister entre un modèle en étoile et un modèle en flocon de neige ?

Exercice 3 : Une entreprise de fabrication de matériel informatique souhaite mettre en place un système d'information décisionnel sous forme d'un magasin de données (*data mart*) pour observer et analyser son activité de vente au niveau des différents lieux de distribution de ses articles et cela dans plusieurs villes *(magasins spécialisés, grandes surfaces*). Ces lieux de distribution sont renseignés par leur enseigne, leur type (en fonction de leur surface), leur adresse (code postal et ville), leur département et leur région. Les ventes sont comptabilisées mensuellement, par trimestre et par année en fonction du nombre d'articles vendus et le chiffre d'affaire réalisé.

1. Quel est le fait observable pour cette entreprise ?
2. Identifiez les axes d'analyse et la (les) mesure (s) associée (s) à l'activité de vente ?
3. Construire le modèle conceptuel en étoile de ce magasin de données ?
4. Améliorez le modèle précédent en un modèle en flocon de neige qui prend en charge différents types d'hiérarchies qu'il faut identifier ?

Exercice 4 : Une école spécialisée en TIC assure des formations trimestrielles pour des étudiants et cadres d'entreprises. L'école est organisée en plusieurs départements et pour des raisons de performances, elle désire développer un magasin de données afin d'améliorer la gestion de ses programmes d'enseignement. La principale source d'informations serait la base de données de gestion des historiques des formations déjà assurées, dont le schéma relationnel est le suivant.

*Etudiant (**Code-Etudiant**, nom-étudiant, courriel, #ID-adresse, #N° Groupe)*
*Adresse (**ID-Adresse**, N° Porte, Rue, Ville, Code-postal)*
*Assurer-cours (**Matricule-ens, #Sigle-cours, #N° trimestre, #N° Groupe**, Heures réalisées)*
*Département (**Code-département**, nom département)*
*Programme (**Code-programme**, nom programme, Nombre heures prévues, #Code-département)*
*Cours (**Sigle-cours**, titre cours, description cours, #Code-programme)*
*Diplôme (**ID-Diplôme**, Libellé-dip, Année-dip, Institution-dip, pays-dip, #Code-Etudiant)*
*Enseignant (**Matricule-ens**, nom-enseignant, grade)*
*Trimestre (**N° trimestre**, Description trimestre)*
*Groupe (**N° Groupe**, Nombre étudiants)*
*Inscription-Formation (**Code-Etudiant, Code-programme, N° trimestre,** Date-inscription**,** Montant facturé, montant payé**)***
*Assister-Cours (**Code-Etudiant, sigle cours, N° trimestre,*** Total-heures-cours***)***

Le magasin de données à concevoir devrait permettre de répondre, entre autres, aux questions analytiques utiles à la prise de décision (cours ayant le plus (le moins) d'inscriptions, nombre moyen d'étudiants par groupe, enseignants qui ont eu les plus gros (plus petits) groupes d'étudiants, chiffre d'affaire…)

1. Indiquez quelles sont les tables de faits et les tables de dimension du magasin de données ?
2. Donnez la multi-hiérarchie de deux dimensions de votre choix et précisez leurs niveaux ?
3. Elaborer le schéma conceptuel du magasin de données et spécifiez son type ?

Annexe : Solutions des exercices

Solutions des exercices de la série de TD N°1

Exercice 1 : Choisissez la ou les bonnes réponses parmi celles proposées

1. Les valeurs ajoutées de l'informatique décisionnelle ou BI sont :
   a. Le contrôle permanent de la gestion interne des activités de l'entreprise

b. Anticiper et prévoir les tendances et habitudes des consommateurs.
c. Offrir une vue de haut niveau sur l'efficacité opérationnelle.
d. Recueillir l'information, la traiter puis diffuser les résultats aux acteurs concernés.
e. Organiser et consolider un grand volume de données de façon homogène.
f. Toutes les réponses précédentes.

2. Les problèmes posés par la diversité des sources de données résident dans:
   a. L'augmentation conséquente da la masse de données à manipuler.
   b. Des difficultés pour centraliser les données sur un support commun et unique.
   c. Le coût d'acquisition des données externes utiles à la prise de décision.
   d. Les données issues des différentes sources sont incompatibles et hétérogènes.
3. La BI est une discipline au carrefour des domaines suivants
   a. Les S.I opérationnels, les métiers de l'entreprise et la stratégie de l'entreprise.
   b. Les S.I géographiques, les réseaux d'entreprise et le cloud.
   c. Les ERP, le domaine du big data et le data mining.
   d. La gestion des connaissances, la data science et le business analytics.
   e. Toutes les réponses précédentes.
4. L'informatique décisionnelle (BI) fournit comme résultats les éléments suivants :
   a. Propose de nouveaux KPI.
   b. Les rapports de synthèses et de visualisation des données.
   c. Les tableaux de bord qui aident à la prise de décision.
   d. Des réponses aux requêtes d'interrogations sur que s'est t-il passé.
   e. Toutes les réponses précédentes.
5. Répondez par *VRAI* ou *FAUX* aux affirmations suivantes et corrigez si *FAUX*
6. La BI exploite les données structurées et semi-structurées. *FAUX*

   Correction. La BI exploite les données structurées, semi-structurées et aussi les données non structurées.

   a. Les données utilisées par une solution BI sont seulement internes. FAUX

   Correction. Les données utilisées par une solution BI peuvent être internes et/ou externes mais aussi issues des réseaux sociaux, de l'Internet ou le cloud.

   b. La BI aide à la prise des décisions opérationnelles, tactiques et stratégiques. FAUX

   Correction. Les décisions stratégiques seulement. Les décisions opérationnelles c'est le système opérationnel qui s'en occupe. Pour les décisions tactiques c'est la responsabilité des cadres moyens et personnels de soutien agissant sur le S.I de l'entreprise, alors que les décisions stratégiques sont prises par le système de pilotage qui déploie la solution BI.

   c. Le point commun entre les 3 V du big data et la BI se limite au **V** de la variété. FAUX

   Correction. Le point commun entre les 3 V du big data et la BI et le V du Volume.

Exercice 2

Au même titre que la BI, la fouille de données vise à extraire, à partir d'une grande masse de données, des connaissances utiles et d'intérêt. Ces informations sont spécifiées par le concept de motifs (*patterns*). Dans le contexte où ces motifs d'intérêt reflètent des données décisionnelles, il y aura une convergence du domaine du data mining avec celui de la BI. Ce qui confirme l'assertion précédente du fait que la discipline data mining a vu le jour avant celle de la BI. Néanmoins, les différences majeures à distinguer entre la discipline data mining et celle de la BI se répartissent sur les quatre aspects suivants.

• **En termes d'objectifs :** le data mining vise à explorer les données pour trouver des réponses aux problèmes de l'entreprise, alors que la BI interprète et présente les données aux décideurs pour éclairer leurs prises de décisions.

• **Les données manipulées :** le data mining traite de petits ensembles de données spécifiques pour une analyse ciblée, alors que la BI exploite de volumineux entrepôts de données pour suivre l'évolution de certaines métriques reflétant le niveau macro de l'entreprise.

• **En termes de résultats rendus :** les techniques du data mining retournent comme résultats des ensembles de données utilisables dans un format unique, alors que la BI offre toute une panoplie de formats en sortie, tels que les tableaux de bord, graphiques et rapports.

• **Utilisation des indicateurs clés de performance (KPI) :** le data mining se focalise sur l'identification des nouveaux KPIs utiles à l'entreprise, alors que la BI traite et analyse l'évolution de ces KPIs.

Exercice 3

a. Dans une entreprise, les critères pertinents pour aborder une comparaison d'un SI opérationnel et un S.I décisionnel peuvent se résumer aux aspects suivants :
   1. Les objectifs visés par chaque système ;
   2. Les utilisateurs potentiels de chaque système ;
   3. Le volume des données manipulées par chaque système ;
   4. Le type d'accès aux données ;
   5. La vitesse de réponse aux requêtes des utilisateurs ;
   6. Le nature du processus d'exploitation du système ;
   7. Les structures de données;
   8. Le niveau de granularité des données.

b. En se basant sur la liste précédente de critères, la comparaison d'un S.I opérationnel avec un S.I décisionnel est exposée dans le tableau suivant.

N°	Cirière de comparaison	S.I Opérationnel	S.I Décisionnel
1	Les objectifs visés	Suivi permanent des données de gestion	Extraction des données utiles à la prise de décision
2	Les utilisateurs du système	Tout le personnel de l'entreprise	Les décideurs et analystes seulement
3	Le volume de données manipulées	Réduit (*informations du système de production*)	Enorme (*données de production, de synthèse et les historiques*)
4	Le type d'accès aux données	En lecture-écriture	En lecture seule
5	La vitesse de réponse aux requêtes	Très rapide	Rapidité moyenne
6	Structures de données	Décentralisées sur différentes bases de données	Centralisées (*vision unique : entrepôt de données*)
7	Niveau de granularité des données	Très fin (pour cibler les détails)	Très élevé (données de synthèse et résumés)

c. Définition d'un S.I décisionnel qui met en exergue les avantages des SI décisionnels par rapport aux S.I opérationnels. *Un S.I décisionnel est un système qui centralise un grand volume de données à granularité élevée et qui est destiné aux décideurs de l'entreprise pour les aider dans leurs prises de décisions***.**

Solutions des exercices de la série de TD N°2

Exercice 1 Choisissez la ou les bonnes réponses parmi celles proposées

1. Dans une architecture Business Intelligence, l'objectif principal de la phase d'intégration de données est qu'à terme :
   a. Les données soient utilisables de façon homogène comme si elles constituaient une seule base de données permettant ainsi leur analyse.
   b. Les données soient duplicables à travers le Cloud pour pouvoir les partager.
   c. Les données soient contrôlées seulement par l'administrateur pour des raisons de sécurité.
   d. Offrir une vision transversale de l'entreprise pour répondre aux besoins décisionnels.
2. L'entrepôt de données est conçu pour :
   a. Répondre à des requêtes et à des analyses de données.
   b. Permettre l'exploitation des données par des outils OLTP. (*Corrigé : des outils OLAP*)
   c. Organiser et stocker les données de manière à pouvoir en extraire une plus-value.
   d. Faciliter la prise de décisions et les activités de type Business Intelligence.
3. Le chargement incrémentiel des données par un outil ETL:
   a. Doit tenir compte de la nature des changements survenus dans les sources de données.
   b. Tient compte de la stratégie de gestion des changements adéquate à chaque situation.
   c. Doit être lancer en batch (*traitement par lots*). (*Corrigé : batch ou temps réel*)
   d. Est opéré une fois le chargement initial terminé.
4. L'opération de collecte de données à partir des différentes sources consiste à :
   a. Exploiter les bases de données internes pour récupérer les données utiles.
   Corrigé (BDD internes et externes + autres formats de données fichier, documents…)
   b. Explorer, accéder aux supports de stockage et extraire les données considérées.
   c. Identifier, sélectionner, extraire et filtrer les données brutes .
   d. Cerner les données externes pertinentes pour la prise de décision.
   Corrigé (externes et internes)
5. Répondez par *VRAI* ou *FAUX* aux affirmations suivantes et corrigez si *FAUX*
6. Le Data warehouse s'appuie sur le principe de la traçabilité des informations. *VRAI*
7. L'extraction complète des données par un outil ETL est employée uniquement lors d'un chargement initial des données dans l'entrepôt. *FAUX*
   Corrigé (ou lors d'un rafraichissement complet des données)
8. Un outil ETL sert à collecter, convertir et charger les données dans l'entrepôt. *VRAI*
9. Les transformations effectuées par un outil ETL portent sur le contenu des données uniquement. *FAUX*. *Corrigé (ou sur leur format)*

Exercice 2

a. La phase de diffusion est la 3ième étape du processus décisionnel qui vise à :

- Mettre les données à la disposition des différents utilisateurs de la solution BI.
- Répartir les données collectées et intégrées suivant différents contextes d'utilisation.
- Définir et mettre en œuvre les droits d'accès aux données de l'entrepôt au profil de chaque futur utilisateur de la solution BI.
- Identifier et calculer les agrégats (cumuls) associés à chaque contexte d'utilisation.

b. Deux problèmes majeurs sont induits par la suppression ou l'omission de la phase de diffusion :

- **Le problème de sécurité :** les utilisateurs de la solution BI n'ont pas forcément tous besoin du même niveau de détails, ni le même niveau d'intérêt pour utiliser les données mises à leur disposition. En effet, les niveaux de préoccupation différent d'un utilisateur à l'autre, suivant leur position hiérarchique dans l'organisation. Donc, une politique de droits d'accès aux données doit être instaurée afin de préserver la confidentialité et l'intégrité des données.

Cette politique est mise en œuvre via le mécanisme de gestion des privilèges de chaque utilisateur et qui est assuré par la quasi-totalité des SGBD actuels. Le cas échéant, des violations des droits d'accès peuvent survenir, engendrant ainsi des accès illicites aux données commerciales et informations personnelles des employés de l'entreprise.

- **La dégradation des performances de la solution BI :** pour des raisons de performances, les agrégats spécifiques à une catégorie d'utilisateurs peuvent être stockés de manière persistante dans les entrepôts de données, mais leur accès doit être sécurisé. Une autre solution consiste à stocker les agrégats pré-calculés dans des magasins de données spécifiques.

Le principe du calcul préalable et unique des agrégats, suivi de leur stockage de manière persistante, permettra de gagner en temps de calcul au lieu de les recalculer dynamiquement chaque fois où ils sont sollicités.

c. Exemple réel sur l'intérêt de la phase de diffusion.

Considérons l'activité GRH dont l'objectif est de faire le suivi de la carrière du personnel d'une entreprise commerciale. L'activité de cette entreprise s'étend sur plusieurs régions et plusieurs points de ventes mais aussi sur le Web (*des activités de commerce électronique*). La solution BI pour cette entreprise et relative à la fonction GRH doit intégrer les données provenant des sources suivantes :

- Les données du personnel affecté aux différents points de vente et magasins.
- Les données du personnel de la direction général et des succursales.
- Les données sur les associés et les revendeurs temporaires.
- Les données sur le personnel temporaire rattaché aux activités du numérique et vente en ligne.

Cette entreprise désire réaliser une solution BI qui aidera les décideurs à la prise de décisions relatives à la GRH. Lors de l'élaboration de la solution BI, pour cette entreprise il faut exploiter certains agrégats (*cumuls et moyennes*), tels que le taux de présence des employés, le montant du chiffre d'affaire réalisé par chacun, le niveau scolaire moyen de chaque structure, la moyenne des écarts de caisse de chaque point de vente, l'assiduité des salariés de la direction générale…etc.

Pour cette entreprise, la phase de diffusion de la solution BI, consistera à pré-calculer les cumuls précédents et les stocker de manière permanente dans l'entrepôt de données à concevoir afin d'optimiser les performances globales de la solution BI. D'autre part, il est impératif que l'accès aux informations relatives à la carrière du personnel soit restreint aux seuls utilisateurs concernés (*les chefs de services respectifs et les directeurs centraux uniquement*). A titre d'illustration, le chef service marketing ne pourra pas consulter la moyenne des écarts de caisse de chaque point de vente. Cette information ne sera accessible qu'aux responsables des ventes et le chef service comptabilité.

Exercice 3

Les outils ETL assurent l'extraction, la transformation et le chargement des données dans l'entrepôt. Plusieurs problèmes sont rencontrés lors des opérations de transformation permettant d'intégrer les données des différentes sources dans l'entrepôt cible.

a. Parmi les problèmes les plus fréquent, on peut distinguer notamment :

1. **Problème de sources multiples :** ce problème survient lorsqu'une entité quelconque possède des représentations différentes dans des sources multiples. Cela là peut être dû, soit à des entités de gestion sémantiquement divergentes, soit à des dénominations variées, ou encore à cause des identifiants des entités qui sont distincts. Par exemple, l'entité **Etudiant** peut avoir différents modèles de représentation qui changent d'un système de gestion à un autre (*des systèmes opérationnels de gestion distincts*). A titre d'illustration, cette entité peut être considérée comme *Lecteur* au niveau de la gestion de la bibliothèque, **Résidant** au

niveau de la cité universitaire et *Adhérent* au niveau du domaine des activités sportives et culturelles.

2. **Problème de résolution d'entités :** ce problème survient lorsqu'une entité se trouve dans plusieurs sources, sans quand ait la correspondance entre ces sources.
   Par exemple, suite à une opération de fusion d'entreprises, les clients ont différents identifiants et différents statuts, tels que : Client, Touriste, abonné, locataire. Chaque entité est décrite par un ensemble d'attributs, plus ou moins communs avec les autres entités. Se pose, alors, la question de leur concordance et de leur jumelage.
3. **Problème de la gestion des changements dimensionnels :** bien que les tables de dimension sont plus statiques que les tables de faits, elles peuvent à leur tour subir des changements, mais avec une fréquence faible. Le problème de la gestion des changements dimensionnels est associé au type de changements affectant les attributs de dimension et à leur prise en charge (Slowly Changing dimension ou *SCD*).
   Par exemple, l'adresse d'un client, son numéro de téléphone, sa catégorie et la dénomination d'un produit sont des attributs de dimension mais qui peuvent changer avec le temps.

b. Proposition de techniques pour résoudre les problèmes des transformations.

1. **Résolution du problème des sources multiples**. Plusieurs stratégies sont possibles,
   - Choisir la source la plus prioritaire (*Etudiant dans notre exemple précédent*)
   - Choisir la source qui contient l'information la plus récente.
   - Choisir la source qui contient le maximum d'informations.
2. **Solution pour le problème de résolution d'entités**. La solution la plus connue consiste à faire un couplage d'enregistrements qui se base sur la comparaison de leur paires d'attributs. Donc, il s'agira de trouver les correspondances entre entités. Plusieurs approches existent dans la littérature. Certaines se basent sur des règles de résolution et d'autres sur les modèles probabilistes. Un exemple de règle de résolution est que les entités doivent avoir au moins N champs identiques (fuzzy lookup / matching).
3. **Solution pour la gestion des changements dimensionnels :** en fonction du type du SCD (1,2 ou 3), il faut déterminer la stratégie de gestion des changements adéquate.
   Les stratégies d'historisation possibles pour les différents SCD sont les suivantes :
   - **SCD Type 1**: dans ce cas, l'ancienne valeur est écrasée avec la nouvelle. Par exemple, si le client a changé son adresse de livraison, on ne retient que la nouvelle adresse.
   - **SCD Type 2:** ajouter une ligne dans la table de dimension pour la nouvelle valeur.

   Par exemple, si le client a changé son adresse de livraison de A à B, alors préserver les deux valeurs A et B. Donc, on aura deux enregistrements du même client avec deux adresses différentes.
   - **SCD Type 3:** avoir deux colonnes dans la table de dimension correspondant à l'ancienne et la nouvelle valeur dans la colonne courante. Pour l'exemple de changement d'adresse, il faut créer une nouvelle colonne dont le libellé sera *NOUVELLE-ADRESSE*, tout en gardant l'ancienne colonne (*ADRESSE*).
   - **Stratégie Hybride:** on combine les types de stratégies de gestion des changements SCD 2 et SCD 3 pour préserver les anciennes valeurs, au même titre que les nouvelles.

<u>Exercice</u> 4 : Data streaming

Une des limitations principales des outils ETL est leur incapacité à traiter et à intégrer des flux de données en temps réel (real-time data streaming).

a. Intérêt de la prise en charge des données temps-réel pour les solutions BI

Le traitement par lot était suffisant pour répondre aux exigences de gestion des données à manipuler par les solutions BI. Mais, avec l'augmentation du débit, les évolutions récentes des TIC et leur démocratisation, les données internes et externes à toute organisation sont devenues de plus en plus variées, instantanées et volumineuses. Et comme les outils ETL traditionnels ont été perçus pour prendre en charge des données locales, généralement relationnelles, et ils n'ont pas été conçus pour faire face à des flux de données provenant du Cloud, le problème de la gestion des flux de données est posé avec acuité dans les environnements temps-réel. En effet, de nombreux environnements d'entreprise modernes ne peuvent pas attendre des heures ou des jours pour que les applications gèrent des lots de données. Ils doivent répondre aux nouvelles données en temps réel au fur et à mesure que les données sont générées. En effet, les organisations contemporaines génèrent et traitent des données sous forme de flux continus en temps réel qui sont caractérisées par :

- Une nature éphémère,
- Elles proviennent d'utilisateurs nomades,
- Elles sont non structurées,
- Leurs volumes sont très importants.

De ce qui précède, les outils ETL conventionnels, demeurent limités pour le traitement des données en temps-réel. Cela est dû fondamentalement au fait que les volumes de données exponentiellement importants brisent les pipelines ETL au niveau des passerelles. Par ailleurs, plus il faut du temps et des ressources pour transformer ces données, plus la file d'attente des données sources est sauvegardée et les données deviennent obsolètes.

b. Les inconvénients majeurs des outils ETL pour le traitement des flux de données temps-réel se résument aux deux points suivants :

- Pour pouvoir traiter des flux données temps-réel, toutes les exigences de la phase de transformation d'ETL, telles que le nettoyage, l'enrichissement et le traitement des données, doivent être effectuées *plus fréquemment* à mesure que le nombre de sources de données et le volume montent en flèche. Ce qui n'est pas assuré par les outils ETL conventionnels.
- Les ETL sont incapables de gérer, *instantanément*, les données importantes qui pourraient générer de meilleures informations à valeur ajoutées (*informations commerciales, par exemple*) pouvant être intégrées aux systèmes d'analyse de données avancés, tels que les systèmes l'apprentissage automatique et les algorithmes d'intelligence artificielle, comme les systèmes de recommandations et les systèmes de prédiction...

c. Comparaison entre les deux approches ETL et ELT

Le tableau ci-dessous met en relief une comparaison détaillée entre ETL et ELT qui est basée sur un ensemble de critères.

Critère de comparaison	ETL	ELT
Principe de fonctionnement	Extraction, transformation puis chargement des données.	Extraction, chargement puis transformation des données.
Volume et variété des données gérées	Plus adéquat pour les données moins volumineuses et qui	Plus adéquat pour des grands volumes de données structurés et non-structurées.

	exigent des transformations complexes.	
Emplacement où sont effectuées les transformations.	Dans une zone de transit (*staging area*)	au sein de la base de données cible (*ou l'entrepôt cible*).
Temps d'activation des transformations.	Transformations faites après l'extraction (*au moment de la conception*).	Transformations faites au moment de la requête d'exploitation.
Vitesse de chargement des données	Le chargement se fait par lots (*batch*), donc plus long (*initial, incrémentiel, complet*)	Le chargement est rapide car il se fait en temps-réel (*instantané*).
Ressources utilisées	utilise les ressources locales avec le parallélisme.	utilise la technologie Big-data (*Hadoop, Spark…etc*) et profite des ressources du cloud.
Avancée technologique	Technologie mature et disponibilité des experts spécialisés.	Technologie moins mature et implémentation plus complexe (*pas d'outils dédiés qui offrent un support complet*)

d. Proposition d'une technique pour améliorer l'architecture BI afin de prendre en charge la nature des données en flux continu.

Avec le volume et la variété croissante des données, la complexité du processus de leur prise en charge et l'émergence de flux de données temps-réel, la problématique de la gestion moderne des données reste un défi majeur pour toute architecture BI. La technique ETL a évolué de plusieurs manières, où l'extraction, la transformation et le chargement sont des processus simultanés fonctionnant sur des données locales sans prise en compte de l'aspect temps réel des données manipulées.

Pour surmonter les limites des technologies ETL traditionnelles, les solutions BI modernes doivent se tourner vers le streaming ETL, avec traitement de flux en temps réel à l'aide d'outils dédiés.

Le Streaming ETL (S-ETL) vise à exploiter des données en temps réel. Son mécanisme de fonctionnement est basé sur l'extraction des données, leur transformation automatiquement, puis leur chargement instantané vers n'importe quelle destination. Dans ce type d'architecture, les sources de données sont toujours en entrée de l'architecture (*à gauche du système BI*). Ces sources fournissent des données à une plate-forme de traitement de flux qui sert de colonne vertébrale aux applications ETL de streaming, mais également à de nombreux autres types d'applications et de processus de streaming. L'application ETL de diffusion en continu peut extraire des données de la source, ou la source peut publier des données directement dans l'application ETL de diffusion en continu. Lorsqu'un processus ETL en continu est terminé, il peut transmettre des données vers la sortie de destination (*située à droite de l'architecture*) qui est potentiellement un entrepôt de données, ou bien il peut renvoyer un résultat à la source d'origine sur la gauche. De plus, il peut fournir simultanément des données à d'autres applications et référentiels. Du point de vue technologique, l'API Kafka d'apache offre une solution dédiée pour le S-ETL.

Solutions des exercices de la série de TD N° 3

Exercice 1 :

a. L'entrepôt donne une vision transversale de l'organisation parce que sa structure ainsi que l'organisation des données qu'il stocke permettent de disposer de l'ensemble des

informations utiles sur un sujet, le plus souvent *transversal* aux structures fonctionnelles et organisationnelles de l'entreprise. En utilisant un entrepôt de données on peut ainsi passer d'une *vision* verticale de l'entreprise à une *vision transversale* beaucoup plus riche en informations et qui considère, en même temps, les données de plusieurs entités structurelles. A titre d'exemple, un entrepôt de données offre une vision transversale sur les données relatives à la gestion financière, à la gestion commerciale et à la gestion des ressources humaines simultanément.

b. On utilise un entrepôt de données à la place d'une base de données classique pour les raisons suivantes :

• **Pour des objectifs d'intégration :** la nécessité de combiner les données provenant de diverses sources dans l'entrepôt, d'effectuer des agrégations et d'offrir des vues multidimensionnelles motivent fortement le recours aux entrepôts de données.

• **Pour des besoins de traçabilité :** dans l'entrepôt, les données sont non volatiles et elles ont, donc, une plus longue durée de vie que celles des BDD. En effet, contrairement aux BDD, dans les entrepôts de données les opérations de suppressions ne sont pas tolérables.

• **Pour des objectifs de performances** : dans les BDD, les requêtes sont simples alors que dans l'entrepôt de données elles sont souvent complexes (*requêtes OLAP*). Ce constat motive le besoin d'utiliser la technologie afférente aux entrepôts de données au lieu de celle des bases de données afin de réduire le temps de réponse aux requêtes complexes.

c. L'intérêt d'utiliser plusieurs magasins de données au lieu d'un seul entrepôt de données réside dans les points suivants :

• Lorsqu'on s'intéresse à un seul sujet d'analyse, le magasin de données s'avère mieux indiqué qu'un entrepôt de données. En effet, se concentrer sur un seul sujet d'analyse engendre une simplicité dans la conception et la compréhension du magasin de données. En plus, la manipulation des données qu'il contient sera aisée, contrairement aux entrepôts des données qui sont plus complexes à implémenter et à gérer.

• Lorsque le nombre de futurs utilisateurs est *restreint* et si leurs besoins sont bien ciblés, les magasins de données seront plus adéquats que les entrepôts de données qui sont, plutôt, destinés à une large gamme de décideurs.

• Si les décideurs désirent s'intéresser à un nombre limité de sources de données qui proviennent la plus part du temps d'un même département (*une même activité : ventre, GRH, finances…*), dans ce cas les magasins de données sont plus avantageux que les entrepôts de données.

<u>Exercice</u> 2 : Propriétés d'une base de données vs propriétés d'un entrepôt de données

a. Effectivement, un entrepôt de données peut être vu comme une base de données multidimensionnelle. Cependant, plusieurs différences fonctionnelles et structurelles sont décelées.

Le tableau suivant met en exergue une comparaison des deux structures sur la base d'un ensemble de critères.

N°	Critère de comparaison	Base de données	Entrepôt de données
1	Nature des données	Quotidiennes et récentes.	Historiques et agrégées.
2	Volume des données	Moyen (giga-octets).	Très élevé (téraoctets).

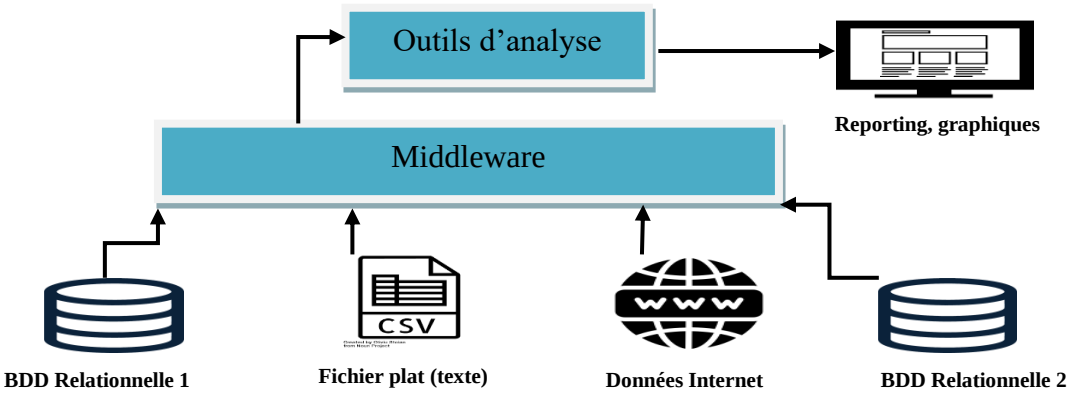
3	Durée de vie des données	Peuvent être supprimées.	Non volatiles (*permanentes*)
4	Opérations sur les données	Lecture et écriture.	Lecture et rafraichissement.
5	Contexte d'exploitation	Gestion opérationnelle	Support d'aide à la décision
6	Objectifs visés	Traitement des requêtes transactionnelles (OLTP)	Analyse des données (requêtes OLAP)
7	Perception	Bidimensionnelle	Multidimensionnelle
8	Modèle conceptuel utilisé	Entité-Association	Etoile, Flocon de neige, Hybride

b. La gestion de la concurrence, de la normalisation et de la sécurité par les systèmes de gestion des bases de données diffère de celle des entrepôts de données. Le tableau ci-dessous met en relief le degré d'importance accordé par chacun des systèmes à chaque aspect.

N°	Aspects	Base de données	Entrepôt de données
1	**Gestion de la concurrence**	Gestion très fréquente pour garantir les propriétés ACID (*Transaction Atomique, Cohérente, Isolée et Durable*).	Gestion rare de la concurrence à cause du nombre d'utilisateurs réduit et du fait que les données sont accédées en lecture seulement.
2	**Normalisation des données**	Gestion rigoureuse, formelle et avancée afin d'assurer au moins les trois première formes normales (1FN, 2FN et 3FN).	Normalisation rudimentaire (*certains types de schémas tolèrent même certaines redondances, par exemple le schéma en étoile*).
3	**Sécurité des données**	La sécurité est une fonction intrinsèque qui est prise en charge par le SGBD (*Confidentialité et droits d'accès*).	La sécurité des données est spécifiée lors de la phase de diffusion.

Exercice 3 :

a. Schéma illustratif de l'exploitation directe des données sans l'utilisation d'un entrepôt



b. Les difficultés rencontrées par le déploiement de ce choix résident dans les aspects suivants :
   - Des difficultés relatives aux accès aux différentes sources.
   - Impossibilité de faire le suivi de l'évolution historique des KPI et des métriques d'analyse.
   - Complexité de développement des middlewares adéquats et augmentation de leur cout avec l'accroissement du nombre de sources de données.
   - Risque d'incohérences sémantiques et d'anomalies des données manipulées à cause de la grande diversité des données.

c. Les inconvénients majeurs d'une telle approche sont :
   - Cette manière d'exploiter les données à partir des différentes sources n'offre *pas une réelle intégration des données*. Ce qui engendre différentes vues qui ne seront pas consolidées pour tous les utilisateurs.
   - Les vues offertes aux utilisateurs expriment des indicateurs et des métriques qui reflètent un état courant du système à un instant précis. Cependant, *aucune analyse des évolutions historiques de ces paramètres* n'est permise.
   - A cause des requêtes d'analyse portant sur un volume important de données et qui exploitent les données directement de leurs sources respectives, une *dégradation considérables des performances globales du système* est observée (*temps de réponse*), voire même des situations *de blocage des transactions*.

Solutions des exercices de la série de TD N° 4

Exercice 1 :

a. Plusieurs facteurs sont à prendre en compte lors de la conception d'un entrepôt de données. Les facteurs les plus prépondérants sont énumérés ci-dessous et classés suivant les trois dimensions : la nature des données, leur exploitation et autres contraintes.

- **La nature des données :** les caractéristiques des données à prendre en charge par l'entrepôt de données sont un facteur déterminant pour sa conception. Ces propriétés sont :
  - **La quantité de données à gérer :** la volumétrie des données à gérer par l'entrepôt influe de manière forte sur l'élaboration de son schéma conceptuel. En effet, quand le volume des données à gérer croit, le nombre d'attributs dimensionnels croit proportionnellement. D'autre part, potentiellement le nombre de métriques, d'agrégats et d'indicateurs clés de performance augmente en conséquence. Cela induit un impact sur les propriétés et faits à spécifier et à stocker dans les tables de faits correspondantes. Ce qui influence sur le schéma conceptuel de l'entrepôt de données.
  - **Les sources de données :** le nombre de sources de données impacte directement la conception du futur entrepôt de données. En effet, le schéma conceptuel dépendra étroitement du nombre de sources et il se complique davantage avec l'augmentation du nombre de ces sources. Ainsi, le schéma sera simple dans le cas où le nombre de sources est réduit et il est plus complexe si le nombre de sources est élevé (*schéma en flocon de neige*).
  - **L'interdépendance informationnelle entre les unités de l'entreprise:** si les informations provenant des entités organisationnelles de l'entreprise sont déjà relativement intégrées, alors le schéma conceptuel du futur entrepôt de données sera facile à établir. Cette simplicité croit avec une bonne intégration préalable des données. Néanmoins, si les données sont organisées en silos indépendants spécifiques à chaque entité (*service ou activité*) sans aucune intégration, même partielle, n'est existante dans ce cas l'effort de conception sera considérable.
  - **La latence des données :** cette considération prend en compte la fréquence de mise à jour des données et de leur chargement dans l'entrepôt de données. Dans le cas où les mises à jour sont opérées quotidiennement ou de manière hebdomadaire, le chargement peut se faire de manière incrémentielle. Par contre si les mises à jour sont faites en temps réel, le rafraichissement de l'entrepôt doit être instantané. Cette latence des données aura des conséquences conceptuelles sur le schéma à élaborer.
- **L'exploitation de l'entrepôt de données :** concerne les critères relatifs à la future exploitation de l'entrepôt par les différents utilisateurs, à savoir:
  - **L'urgence d'obtenir une solution fonctionnelle** : si la solution BI est exigée dans l'urgence, alors de simples magasins de données sont à prévoir. Ces magasins vont évoluer avec le temps, puis intégrés pour donner naissance à un seul entrepôt de données. Néanmoins, si la solution n'est pas impérative dans l'immédiat, il est préférable d'engager une réflexion globale qui prend en compte la modélisation de toutes les activités de l'entreprise, ce qui conduira à un entrepôt de données intégré et unique, même si la solution tarde à être opérationnelle.
  - **Le nombre d'utilisateurs de l'entrepôt** : Comme chaque futur utilisateur de l'entrepôt aura une vision qui lui est spécifique, le nombre d'utilisateurs de l'entrepôt est donc un facteur déterminant lors de la phase de conception. Cet aspect est justifié par la phase de diffusion qui exige de spécifier les normes de sécurité et de partage des données de l'entrepôt pour chaque classe d'utilisateurs.
  - **La nature des tâches des utilisateurs** : le schéma conceptuel de l'entrepôt de données dépend des tâches que les futurs utilisateurs vont effectuer lors de son l'exploitation. Ainsi, les sorties limitées à des rapports simples engendrent un schéma conceptuel simple, alors que les besoins d'effectuer des analyses détaillées et des fouilles de données

avec des paramètres statistiques compliqués induisent, impérativement, des schémas plus complexes (*augmentation du nombre d'agrégats, de KPI, …etc.*)

- **Autres contraintes :** le troisième aspect concerne les ressources à déployer et les contraintes de l'environnement. On peut citer, essentiellement :
  - **Les contraintes sur les ressources :** le budget alloué au projet d'élaboration d'une solution BI (*avec son entrepôt de données*), la maitrise technologique par le personnel de l'entreprise et si la solution est interne ou sous-traitée sont autant de facteurs qui peuvent impacter la modélisation de l'entrepôt de données.
  - **Les facteurs environnementaux:** tels que la disponibilité de solutions clés en mains, disponibilité de coopération avec les partenaires pour l'accès aux données, facteurs économiques (*coûts des outils logiciels*).

b. Différences entre table de faits et table de dimension

La différence fondamentale entre ces deux tables est que la table *de dimensions* contient des attributs le long desquels des mesures sont prises dans la *table de faits***.**

Le tableau suivant récapitule les différences majeures entre ces deux structures de données.

Critère de comparaison	Table de dimension	Table des faits
Différence de base	Contient les attributs que la table de faits exploite pour calculer les métriques.	Contient les mesures le long des attributs des tables de dimension.
Nombre de tables dans le schéma de l'EDD	Le schéma contient plus de tables de dimension.	Le schéma conceptuel de l'EDD contient moins de tables de faits.
Taille de la table	Les tables des dimensions croient horizontalement. (ajout d'attributs descriptifs)	Les tables de faits évoluent verticalement (insertion de nouveaux enregistrements)
Attributs & enregistrements	La table de dimension contient plus d'attributs et moins d'enregistrements.	La table de faits contient moins d'attributs et plus d'enregistrements.
Nature des Clés	Chaque table de dimension contient sa propre clé primaire.	La table de faits a comme clé primaire une concaténation des clés primaires de toutes les tables de dimension.
Ordre de création des tables	Les tables de dimensions sont créées en premier lieu.	Une table de faits peut être créée uniquement lorsque les tables de dimension sont terminées.

Exercice 2 : La différence cruciale entre les schémas en étoile et celui en flocon de neige est que le schéma en étoile n'utilise pas la normalisation alors que le schéma en flocon de neige utilise la normalisation pour éliminer les redondances des données. Cette différence entraine des conséquences sur les types de tables manipulées par chaque schéma, la complexité des requêtes d'exploitation des données avec variation du temps de leur exécution ainsi que le nombre de jointures qui seront utilisées.

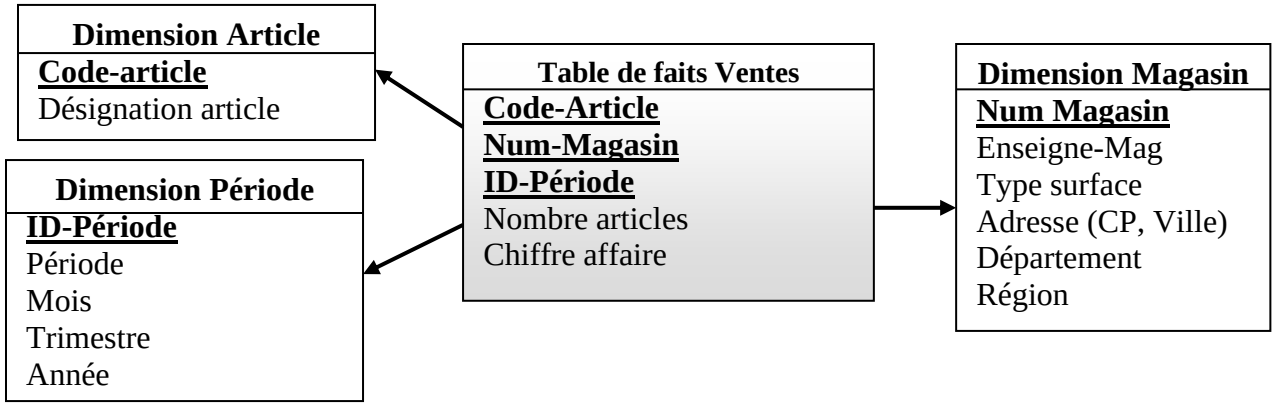
Le tableau suivant résume les différences entre ces deux types de schémas.

Critère de comparaison	Modèle en étoile	Modèle en flocon de neige

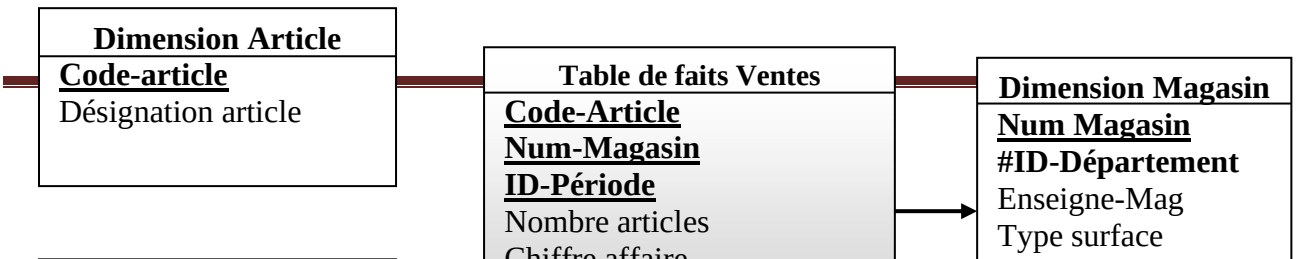
Nature du modèle	Schéma simple, facile à comprendre et implique des requêtes moins complexes.	Schéma difficile à comprendre et implique des requêtes complexes.
Structure du schéma	Contient des tables de faits et de dimension.	Tables de faits et de dimension plus les tables de sous-dimensions.
Normalisation des données	Modèle non normalisé ce qui engendre des redondances.	Modèle normalisé. Absence de redondance
Exploitation du schéma	Nombre restreint de requêtes de jointures.	Utilise des requêtes avec un nombre élevé de jointures.
Temps d'exécution des requêtes	Temps limité et réduit	Dégradation des performances en raison de l'utilisation excessive des jointures.
Espace de stockage	Légère perte d'espace à cause des informations redondantes.	Petite économie de l'espace, suite à l'élimination des redondances.
Approche de conception	Approche Top-down (*du général au particulier*)	Approche Bottom-up (*du particulier au général*)

Exercice 3

1. Le fait observable pour cette entreprise est : **Les ventes des articles**.
2. Les axes d'analyse et la (les) mesure (s) associée (s) à l'activité de vente
   - **Les axes d'analyse sont :** Type d'article, Magasin et Période.
   - **Les mesures sont :** Nombre d'articles vendus et total chiffre d'affaire.
3. Le modèle conceptuel en étoile du magasin de données



4. Amélioration du schéma en étoile obtenu : le schéma en étoile précédent est simple et compréhensible, mais il n'est pas normalisé. Donc, certaines redondances sont présentes.
   Les hiérarchies à prendre en compte pour éliminer ces redondances et rendre le modèle normalisé sont les suivantes :
   La *LOCALISATION* du magasin et la hiérarchie *TEMPS*.
   - *LOCALISATION* : Magasin → Département → Région
   - *TEMPS* : Période → Mois → Trimestre → Année
   Le schéma en flocon de neige suivant permet de modéliser le magasin de données avec des tables normalisées.



Exercice 4

1. Tables de faits et de dimensions
   - Les tables de dimension décrivent les axes d'analyse de l'activité. Elles sont les suivantes :

*a) Etudiant (**Code-Etudiant**, nom-étudiant, courriel, #ID-adresse, #N° Groupe)*
*a) Adresse (**ID-Adresse**, N° Porte, Rue, Ville, Code-postal)*
*b) Diplôme (**ID-Diplôme**, Libellé-dip, Année-dip, Institution-dip, pays-dip, #Code-Etudiant)*
*c) Département (**Code-département**, nom département)*
*d) Programme **(Code-programme**, nom programme, Nombre heures prévues, #Code-département)*
*e) Cours (**Sigle-cours**, titre cours, description cours, #Code-programme)*
*f) Enseignant (**Matricule-ens**, nom-enseignant, grade)*
*g) Trimestre (**N° trimestre**, Description trimestre)*
*h) Groupe (**N° Groupe**, Nombre étudiants)*

   - Les tables de faits servent à stocker les mesures de l'activité. Elles sont les suivantes :

a) *Assurer-cours (**Matricule-ens, #Sigle-cours**, **#N° trimestre**, **#N° Groupe**, Heures réalisées)*
b) *Inscription-Formation (**Code-Etudiant**, **Code-programme, N° trimestre,** Date-inscription**,** Montant facturé, montant payé**)***
c) *Assister-Cours (**Code-Etudiant**, **Sigle cours, N° trimestre,*** Total-heures-cours).

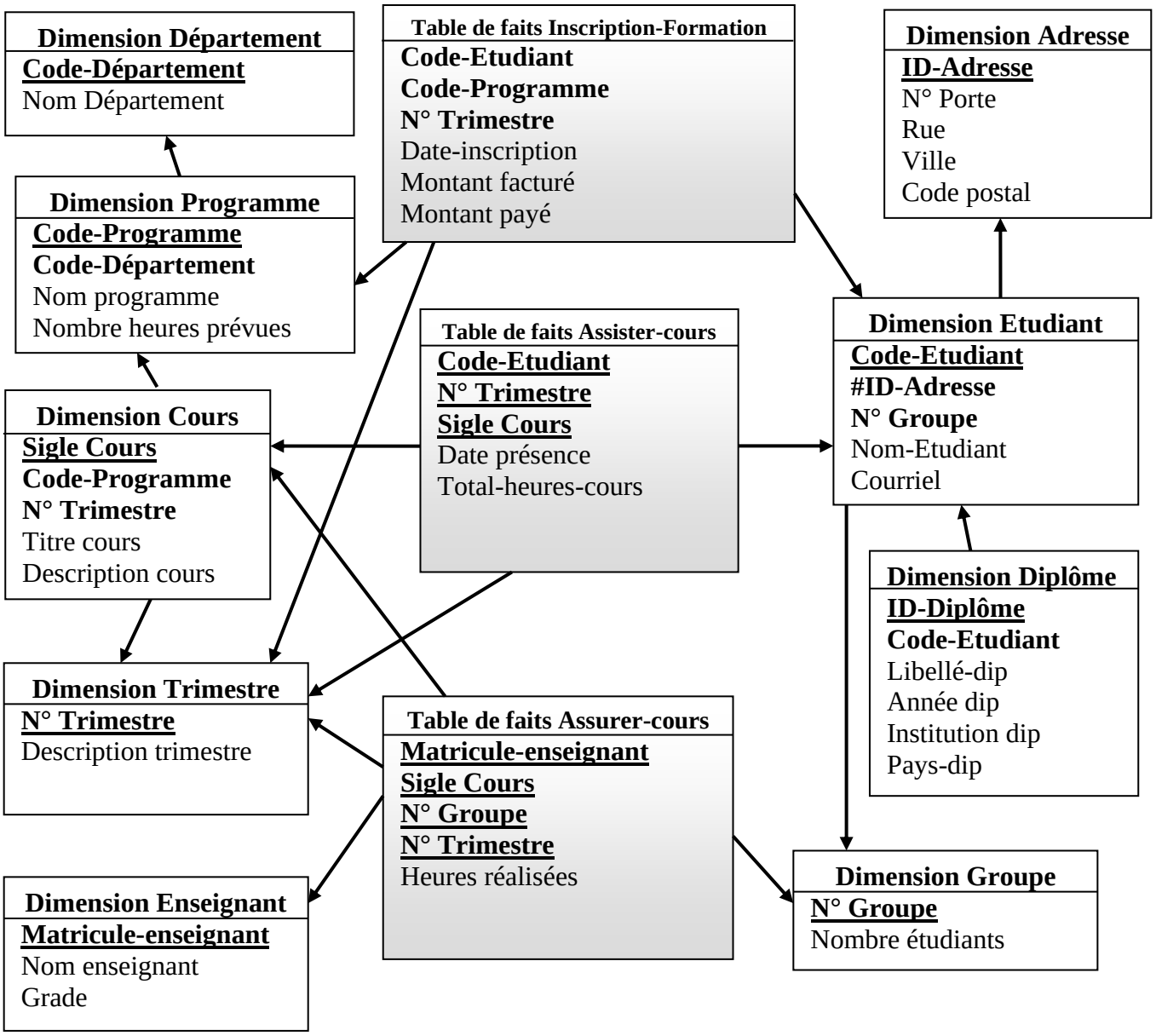
   - Nous pouvons considérer les multi-hiérarchies *Adresse* et *Cours*.
   - Pour la dimension *Adresse*, nous aurons les niveaux Ville et Rue.

   Ville → *Rue* → *Adresse*

   Et pour la dimension Cours, nous aurons les niveaux suivants :

   Département → *Programme* → *Cours.*

2. Construction du schéma conceptuel du magasin de données



Le schéma conceptuel obtenu contient plusieurs tables de faits et plusieurs tables de dimensions. En plus, certaines tables de dimensions sont communes aux tables de faits, donc le schéma élaboré est *un schéma en constellation*.

Exemple d'Examen avec corrigé type

Université 8 Mai 1945 –Guelma- / Faculté MISM / Département d'Informatique
3ième Année Licence ISIL-SI / Module : Business Intelligence

Examen Final 2021 Durée : 2H

I. Partie cours (*8 pts*)

1. Quels sont les composants d'une architecture décisionnelle ? Illustrer votre réponse par un schéma qui met en relation les différents composants ? (*2 pts*)
2. Identifiez les problèmes associés aux sources de données ? Quelle est la raison d'être des intergiciels de type ETL ? (*2 pts*)
3. Dressez une étude comparative qui met en évidence les principales différences entre une solution ERP et une solution BI ? (*2 pts*)
4. Quelles sont les transformations à effectuer sur les données sources avant de les charger dans l'entrepôt ? Enumérer, au moins, 4 types de transformations à opérer (*2 pts*)

II. Questions à choix multiples QCM (*4.5 pts*)

1. L'informatique décisionnelle ou BI vise à répondre aux préoccupations (*1.5 pts*)
   a. Opérationnelles : « Que se passe-t-il en ce moment ? »
   b. Historiques : « Que s'est-il passé ? »
   c. Analytiques : « Pourquoi est-ce que cela s'est passé ? »
   d. Pronostiques : « Que va-t-il se passer ? ».
   e. Toutes les réponses précédentes
2. Les outils utilisés par la BI permettent de (*1.5 pts*)
   a. Générer des rapports de synthèse
   b. Elaborer des tableaux de bord d'aide à la décision
   c. Offrir des visualisations ergonomiques des données (graphiques)
   d. Permettre des analyses de données (Data mining)
   e. Toutes les réponses précédentes
3. L'informatique décisionnelle (BI) est une vaste catégorie de logiciels et de technologies qui comprend (*1.5 pts*)
   a. Les outils d'aide à la décision
   b. Les applications d'exploration de données
   c. Les outils OLAP
   d. Les langages de programmation des systèmes d'informations de gestion (SIG)
   e. Toutes les réponses précédentes

III. Etude de cas : Modélisation multidimensionnelle (*7.5 pts*)

On s'intéresse au domaine de la gestion commerciale d'une entreprise et nous désirons gérer les procédures suivantes : définir le prix de vente, prévoir ses ventes, gérer ses stocks et fournir des données sur un client ou un fournisseur rapidement.

1. Identifiez les tables de dimensions avec, au moins, trois attributs pour chaque table ? (*2 pts*)
2. Quelles sont les tables de faits possibles du domaine commercial (*2 pts*)
3. Mettre en évidence les liens entre table de faits et tables de dimensions (*1.5 pts*)
4. Proposez des KPI pour le domaine commercial et spécifiez leurs règles de calcul ? (*1 pt*)
5. Soit l'indicateur K : chiffre d'affaire par point de vente et par semaine.
   Quelles sont les sources de données possibles pour cet indicateur (*1 pt*)

Bon courage

Corrigé type de l'Examen

Université 8 Mai 1945 –Guelma- / Faculté MISM / Département d'Informatique

3ième Année Licence ISIL-SI / Module : Business Intelligence

Examen Final 2021 **Durée : 1H 30**

I. Partie cours (*8 pts*)

1. Les composants d'une architecture décisionnelle se répartissent sur trois dimensions complémentaires:

a) **Les Données** : composées des sources de données et de l'entrepôt de données.

b) **Les outils logiciels** : ETL, outils d'analyse et de reporting, suites OLAP, logiciels d'analyse avancées et logiciels end-users.

c) **Les sorties attendus de l'architecture :** rapports, tableaux de bords, scoreboards, cockpit,…etc).

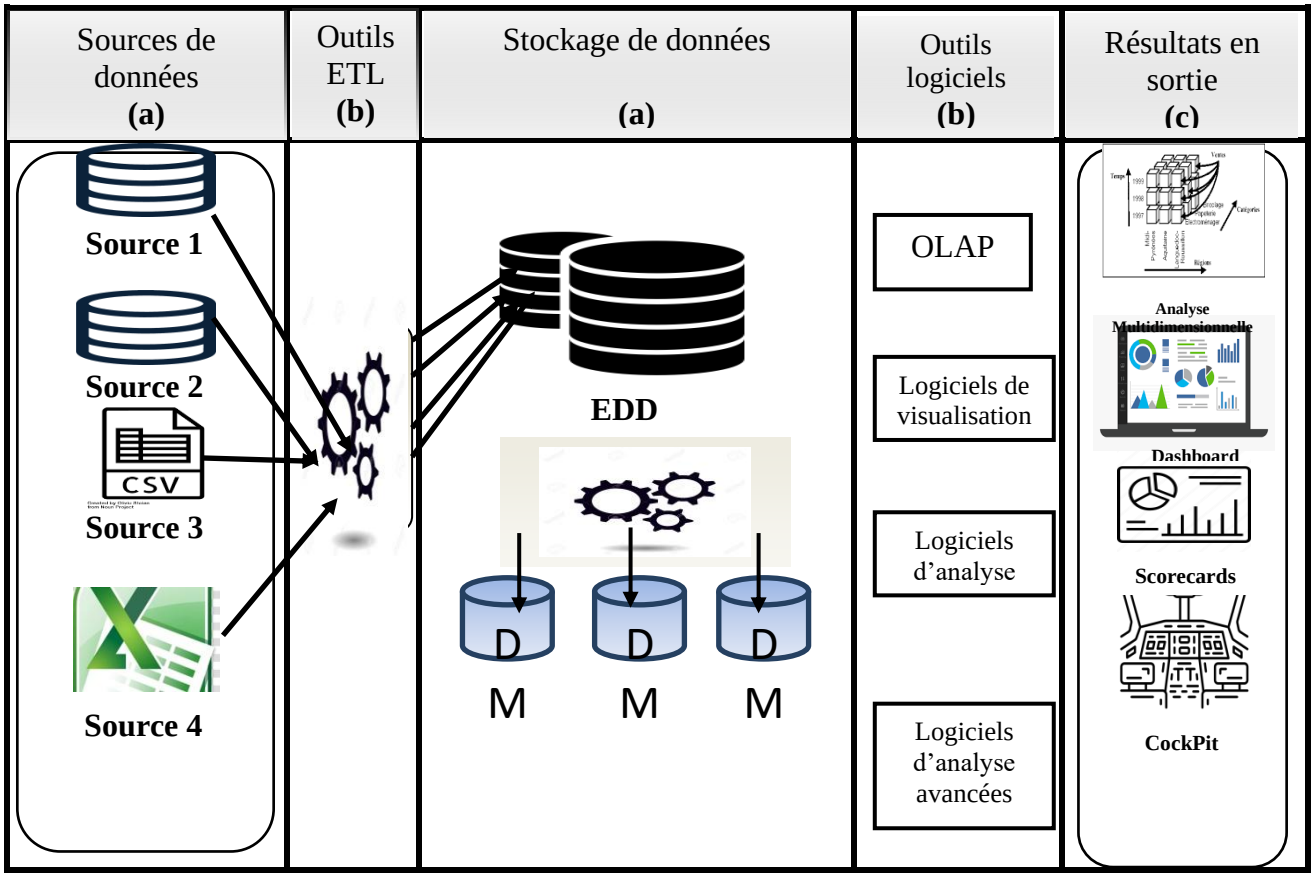


Schéma de fonctionnement d'une architecture business intelligence.

2. Les problèmes associés aux sources de données se résument aux points suivants.

- La *diversité* des sources de données et l'augmentation du nombre de sources peut engendrer des problèmes liés aux contenus de ses sources. En effet, les données peuvent êtres *incompatibles et hétérogènes* ce qui induit de difficultés dans leur *intégration*.
- La raison d'être des intergiciels de type ETL est de réaliser *l'intégration des données*. Ce processus passe par trois étapes complémentaires qui sont *l'extraction* des données des différentes sources, la réalisation des *transformations* sur ces données afin de surmonter les problèmes d'hétérogénéité et enfin le *chargement* des données dans l'entrepôt de données.

3. Comparaison une solution ERP et une solution BI

Une solution *ERP* (Entreprise Resource Planning) permet aux entreprises de centraliser de nombreuses informations opérationnelles au sein d'un seul et unique système d'information qui

exploite une BDD unique. La constitution d'une BDD depuis de multiples départements permet d'automatiser et de mettre à jour les informations manipulées par les différentes structures de manière cohérente. Le rôle d'une ERP est, donc, de gérer et d'analyser quotidiennement chaque information de l'entreprise. Cette analyse en temps réel est idéale pour procéder à des pistes d'audit afin de définir l'origine de données spécifiques. Néanmoins, pour une solution BI, il faut distinguer les constats suivants :

-Une solution BI est dédiée à l'analyse de haut niveau où la prise de chaque décision stratégique peut être déterminante. Pour réaliser ces études, l'ensemble de informations de l'entreprise est considéré, à savoir :

- Les données opérationnelles (*ventes, comptabilisées, GRH,…au quotidien*) issues de l'ERP.
- Les informations relatives aux stratégies (*revenus, croissance et recettes, tendances…*)

Donc, on définitif l'utilisation d'un ERP permet de centraliser chaque donnée transactionnelle et opérationnelle afin de présenter une vision globale et détaillée de la situation d'une entreprise. Néanmoins, ces analyses ne considèrent aucune tendance. Alors que dans le cadre d'une solution BI, on bascule d'une analyse des données opérationnelles issues de l'ERP à des analyses d'ordre stratégiques qui tiennent compte des objectifs visés par l'organisation.

4. Plusieurs types de transformations doivent être opérés sur les données avant de les charger dans l'entrepôt. Ces transformations peuvent porter, aussi bien sur la structure des attributs manipulés que sur leur contenu. On peut citer, par exemple les transformations suivantes:

a) **Unification du codage des champs** : standardiser la codification utilisée par les divers sources. Par exemple [*Homme, Femme*], [*H,F*], [*1,2*].

b) **Conversion des dates** : harmoniser les différents formats de la date. Exemple *25 JAN 2022* devient *25/01/2022*.

c) **Redressement de format :** permet de changer le type ou la longueur d'un attribut. Par exemple, le nom étudiant sur 30 caractères au lieu de 20.

d) **Fusion de plusieurs champs:** regrouper les informations (attributs) d'une même entité. Par exemple, la source 1 contient le Code et le libellé du produit et la source 2 contient son pourcentage de remises.

II. Questions à choix multiples QCM (*4.5 pts*)

1. L'informatique décisionnelle ou BI vise à répondre aux préoccupations (*1.5 pts*)
   a. Opérationnelles : « Que se passe-t-il en ce moment ? »
   b. Historiques : « Que s'est-il passé ? »
   c. Analytiques : « Pourquoi est-ce que cela s'est passé ? »
   d. Pronostiques : « Que va-t-il se passer ? ».
   e. Toutes les réponses précédentes
2. Les outils utilisés par la BI permettent de (*1.5 pts*)
   a. Générer des rapports de synthèse
   b. Elaborer des tableaux de bord d'aide à la décision
   c. Offrir des visualisations ergonomiques des données (*graphiques*)
   d. Permettre des analyses de données (Data mining)
   e. Toutes les réponses précédentes
3. L'informatique décisionnelle (BI) est une vaste catégorie de logiciels et de technologies qui comprend (*1.5 pts*)
   a. Les outils d'aide à la décision.
   b. Les applications d'exploration de données.

c. Les outils OLAP.
d. Les langages de programmation des systèmes d'informations de gestion (SIG).
e. Toutes les réponses précédentes

III. Etude de cas : Modélisation multidimensionnelle (*7.5 pts*)

1. Les tables de dimension de la gestion commerciale avec leurs attributs :
   - **Table Produit** : Code-produit, Nom produit, Prix unitaire, Unité mesure.
   - **Table Client** : Numéro-client, Nom client, Adresse client, Courriel client.
   - **Table Magasin** : Code-magasin, Libellé magasin, Adresse magasin.
   - **Table Fournisseur** : Code-Fournisseur, Nom fournisseur, Téléphone fournisseur, Courriel fournisseur.
   - **Table point de vente** : ID-point vente, enseigne point vente, adresse point vente.
   - **Table temps :** ID-temps, description période, jours, mois, année.
2. Les tables de faits du domaine commercial sont :
   - **La table Ventes :** mémorise les historiques des achats réalisés par les clients, dans les différents magasins et pour les différentes périodes de temps.
   - **La table Achats :** mémorise les historiques des acquisitions de produits opérées par l'entreprise auprès des différents fournisseurs et pour les différentes périodes.
3. Les liens entre table de faits et tables de dimensions sont schématisés dans le modèle conceptuel suivant :

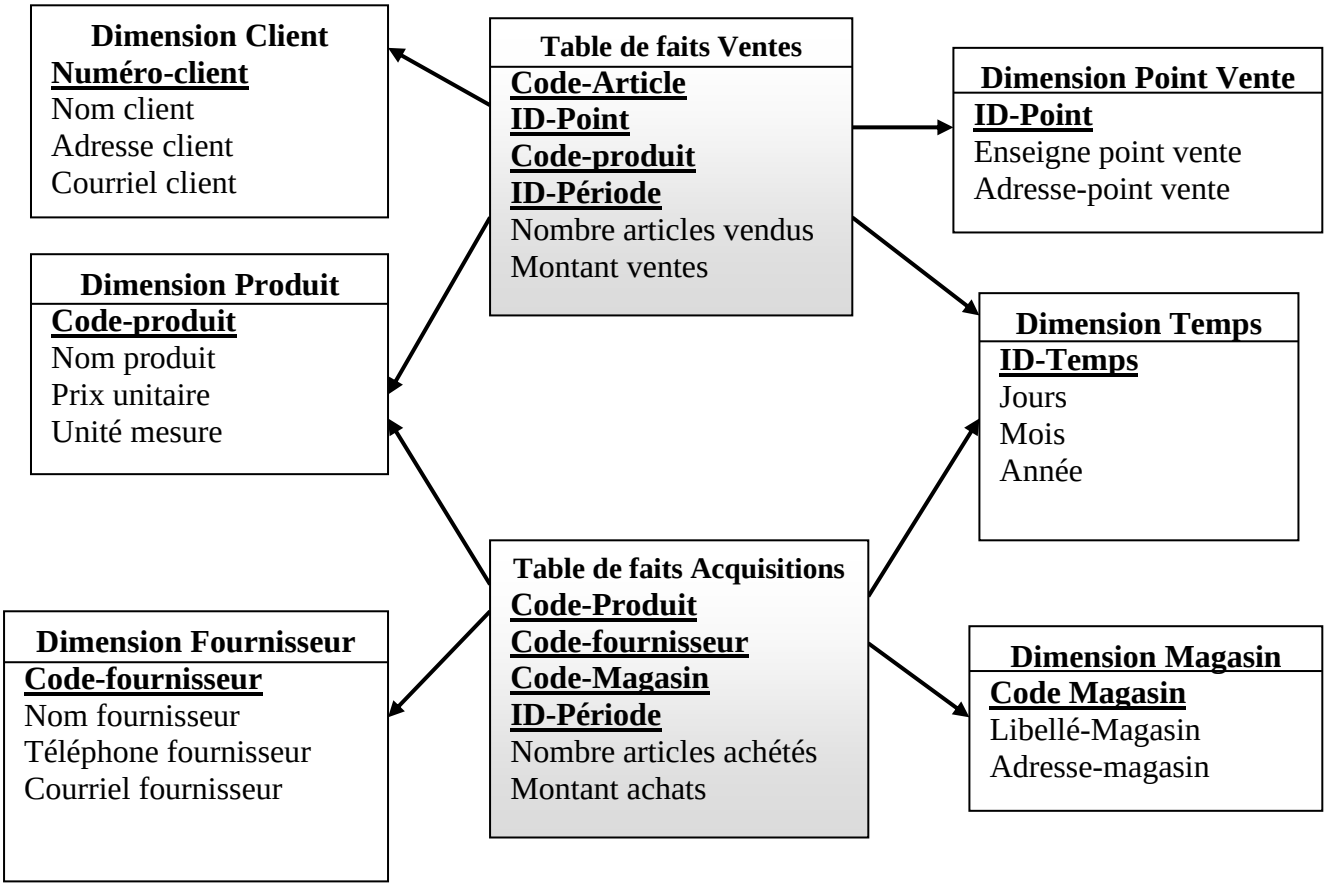


Schéma Conceptuel mettant en relation les tables de faits et les tables de dimensions

4. Proposition de quelques KPI pour la gestion commerciale.
   - Le nombre d'articles vendus par Produit /Point de Vente /Période.
   - Le chiffre d'affaire réalisé par Produit / Point de vente / Période.
   - Le chiffre d'affaire total d'une période, comparé à celui d'une autre période.

- Période de forte / faible consommation. (nombre d'articles et total des ventes).
- Le ratio montant des ventes de chaque magasin / total chiffre d'affaire d'une période.

5. Soit l'indicateur **K : chiffre d'affaire par point de vente et par semaine**.
   Les sources de données possibles pour cet indicateur sont les différentes *tables des ventes* de l'entreprise. Ces tables peuvent êtres matérialisées par un fichier unique dans le cas d'une base de données centralisée, comme elles peuvent êtres issues de plusieurs tables partielles chacune appartenant à un point de vente particulier. Les tables de ventes peuvent aussi provenir du site web de l'entreprise, dans le cas ou des ventes en ligne sont faites via le site marchand de l'entreprise. Dans le cas d'existence de plusieurs sources, l'analyse des données de vente exige une intégration préalable des différents fichiers dans une structure unique.

Bibliographie

1. Ralph Kimbal «**Entrepôt de données, guide pratique du concepteur de data warehouse »**, Wiley, 3ième édition 1997.
2. Ralph Kimball, Laura Reeves, Margy Ross, Warrren Thornthwaite, **Le data warehouse, Guide de conduite de projet**, Collection Blanche, 2005.
3. Ralph Kimball, Margy Ross. **The Data Warehouse Toolkit**: **The Complete Guide to Dimensional Modeling**, 2nd Edition, Wiley (2002).
4. Ralph Kimball, Joe Caserta. **The Data Warehouse ETL Toolkit, Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data**, Wiley (2004).
5. Efraim Turban, Ramesh Sharda, Dursun Delen, David King. ***Business Intelligence: A Managerial Approach,*** *2nd Edition*, Prentice Hall (2010).
6. Paulraj Ponniah. **Data Warehousing Fundamentals for IT Professionnals**, 2nd Edition, Wiley (2010).
7. Gloria J.Miller, Dagmar Brautigam, Stefanie V. Gerlach. ***Business Intelligence Competency Centers***, Wiley (2006).
8. https://www.coursehero.com/file/57429374/Chap19-solutionspdf/
9. https://cours.etsmtl.ca/mti820/public_docs/acetates/mti820-acetates-architecturedw_1pp.pdf/