Synthèse : BI

1. Introduction

La plupart du temps, lorsque l'on arrive dans une entreprise, le stockage des données n'a pas été pensé initialement. Il faut un système de classement comme dans les librairies par exemple. Les données, c'est le nouveau pétrole. Elles sont rarement propres, il y a en a partout, et si on a pas de stratégies correctes associées à elles, des catastrophes peuvent vite arriver.
Il ne faut pas stocker trop de données : l'important n'est pas la quantité de données, mais la quantité d'information que l'on peut tirer de ces données.

L'histoire de la Business Intelligence

La révolution est une période de grandes transformations qui a un impact incontestable sur tous les domaines techniques, économiques, sociaux et culturels. Il y a eu 4 révolutions :

- 1800 : vapeur, charbon, mécanique. On est passé d'une société agricole à une société où les gens peuvent se déplacer et commercialiser avec leurs voisins → échanges de plus en plus rapide.
- 1900 : pétrole, électricité, production de masse. Au lieu de prendre le train, on prend maintenant sa voiture. Les communications ont commencé à s'accélérer.
- 1960 : l'âge de l'information, informatique, communication de masse, internet. Plus besoin de se déplacer pour communiquer avec l'autre côté de la planète.
- 2010 : données, AI, IOT, digital. L'industrie 4.0. Le monde est encore plus ultra connecté et basé sur l'IA et les machines interconnectées.

L'impact social a été énorme. Même un clochard doit avoir un compte bancaire numérique. Les révolutions ne sont pas toujours biens partout.
Qu'est-ce qui a changé ? Ce ne sont pas de nouveaux algorithmes mais la quantité de données est montée de manière exponentielle. La vitesse de calcul également (ex : CandyCrush VS navette). Grâce au cloud et à la quantité de données, le processing de système du style IA ou décisionnel peut être exploité à sa puissance maximale. Les algorithmes IA évoluent beaucoup plus vite maintenant.

L'importance et la valeur capitale de l'entreprise, ce n'est pas d'avoir les meilleurs algorithmes, mais d'avoir les meilleurs données, claires, structurées, précises et qui ont de la valeur (Google).

Qu'est-ce que la BI ?

La Business Intelligence représente les outils et les systèmes qui jouent un rôle d'aide à la décision dans les procédés d'optimisation et de gestion d'une entreprise. Ces outils fournissent des moyens de collecter, d'enregistrer, d'accéder et d'analyser des données dans le but d'offrir une aide à la décision et afin de fournir une vue d'ensemble de l'activité traitée.
Ces systèmes sont souvent utilisés pour :

Optimiser les revenus	- Dresser le profil des clients et faire des campagnes de publicités ciblées - Corrélation de package de ventes de produits
Réduire les coûts	- Optimiser les chaînes d'approvisionnement de ressources - Compréhension des coûts actuels
Réduire les risques	- Analyser la solvabilité pour des emprunts bancaires - Automatiser la détection de fraudes en utilisant l'analyse prédictive
Régulation, respect de normes	- Respecter les normes gouvernementales afin d'éviter de payer des pénalités - Automatisation de processus administratifs

Histoire de la BI

Au début, l'aide à la décision reposait sur l'expérience individuelle, le savoir, l'avis des conseillers et des décideurs, ainsi que sur l'analyse historique. L'opinion et la subjectivité avaient une grande importance. Il était souvent difficile de prendre des décisions basées sur des faits complets et non biaisés. Petit à petit, la vision suivante rentrait dans les mœurs : Celui qui détient l'information détient le marché (McDonald).
Slide 18 pas intéressante pour ce cours.

Besoins des décideurs

~~Système~~ (application finie)→ environnement (ensemble qui permet de faire des développements concrets et modifier différentes choses, plus flexible, comme C++ ou Java). Ils ont besoin que ce soit rapide, facile à utiliser pour les non-informaticiens, indépendant du système de production (ils ne pouvaient pas avoir accès au système de production sinon ils risquaient de faire planter le programme), avec des droits d'accès restreints et fiable.

Opérationnel VS Décisionnel

Opérationnel	Décisionnel
Grand public	Nombre restreint d'utilisateurs (seulement accessible aux décideurs)
Extrêmement rapide	Rapidité suggérée (si ce n'est pas en temps réel, ce n'est pas trop grave puisque c'est juste pour faire des analyses)
Fermé (une entrée A doit exécuter une tâche, qui donne un résultat B)	Ouvert (permet de faire plus de choses)
'Petit' volume de données (juste ce qui permet de fonctionner)	Gros volume de données (pour prendre ces décisions, on va devoir connecter plusieurs systèmes ensemble, comme par exemple les ventes et les stocks)
Transactionnel (des transactions peuvent s'effectuer)	Non transactionnel
Lecture, écriture et modification des données	Données en lecture seule (la BI ne crée que des rapports à lecture seule : copie du système opérationnel, extraire les données dans un système centralisé → décisions
Décentralisé (peu importe où ça se trouve)	Centralisé (voir au-dessus)

Pourquoi aide à la décision et pas décision automatique ? Ca c'est mieux d'être dirigé par des humains et pas des machines. Il y a des risques d'anomalies non détectables par une machine. Par exemple, une boîte qui en rachète une autre fera un énorme pic. Une alliance entre pays également.

Format des données

Pour pouvoir travailler sur des données, il faut respecter ces 5 conditions :

- Propres : Une analyse correcte ne peut se faire que en se basant sur des informations justes. Il faut éviter les données erronées, les informations manquantes, les imprécisions etc. C'est pour cela qu'une des étapes les plus importante du BI est le nettoyage de données. Pas de fautes d'orthographe, ou stocker d'une certaine manière.
- Cohérentes : Les données et les valeurs doivent être les mêmes au sein de toute l'entreprise. La cohérence consiste à refléter sur la copie d'une donnée les modifications intervenues sur d'autres copies de cette donnée. Si on utilise un téléphone, ça veut dire téléphone, pas ordinateur.
- Standardisées : Les données doivent être représentées sous un format standard commun et normalisées afin de pouvoir être exploitées et étudiées uniformément. Respecter les normes écrites (ex : unité de mesure).
- Actuelles : Les données doivent être suffisamment récentes pour refléter la réalité de l'analyse voulue. Il n'est, par exemple, pas très pertinent d'étudier les données de 1830 pour voir l'évolution du smartphone en Europe. Mises à jour quotidiennes (ex : GPS).
- Compréhensives : Pour pouvoir effectuer des analyses sur des données, il est impératif de comprendre ce que ces données veulent dire, et quel est l'impact de celles-ci. Elles doivent dire quelque chose pour l'humain.

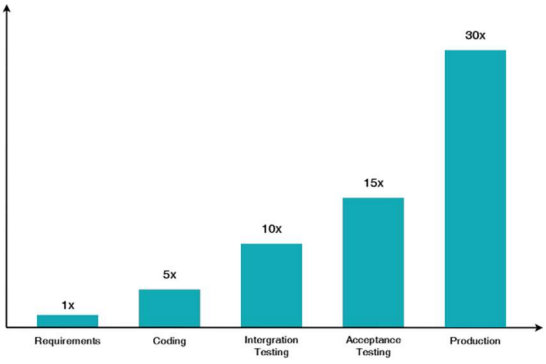
On ne peut pas faire confiance à toutes les données.

Exemples slide 30 à 37.

2. Big data

2 méthodes de production : une roue + une roue + une carrosserie + un moteur + une voiture VS un skate + un vélo + une moto + une voiture → la méthode en itération et apprentissage (2ème) rend le client beaucoup moins frustré car pas frustré durant tout le temps de la production. C'est la méthode Agile.

Mais Agile ne résout pas tous les problèmes en entreprise. Chaque phase de développement nécessite différents coûts. Les phases de développement sont : analyses et besoins → développement → testing → déploiement → maintenance. Chaque suivant est plus grand que le précédent. Le déploiement prend énormément de temps. Les start-ups sont dix fois plus rapides que les grosses entreprises car elles commencent à partir de zéro et elles utilisent les nouvelles technologies.



Si il y a un problème à telle étape ça va coûter ceci :
C'est pareil pour la AI/BI.

Temps moyen d'une mise à jour : ça dépend de la taille de l'entreprise. Une start-up mettra quelques minutes car elle aura commencé à zéro et n'aura jamais été en production. Par contre, dans le système bancaire, le temps sera énorme car il faut vérifier que tout est bien sécurisé, que tout se passe bien, qu'il n'y a aucun impact sur toutes les apps qui sont connectées à cette MAJ. Il faut retester tout ce qui est connecté.

Évolutions technologiques

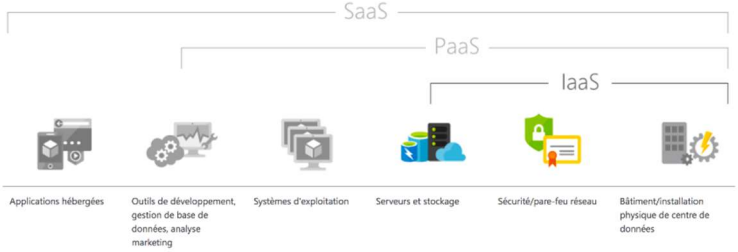
Le changement (quantité de données et vitesse de calcul) se traduit dans la croissance du Big Data. Il y a eu 2 énormes évolutions :

- Technologie de stockage : majoritairement représentée par le déploiement du Cloud Computing.
- Technologie de traitement : base de données non structurée (Hadoop) et calcul à haute performance (Mac Reduce).

Cloud

Le cloud se situe partout autour de nous. Il y a différents serveurs qui permettent de se connecter.
Il y a différents types de cloud et différentes choses possibles avec le cloud :

- Iaas : Infrastructure as a service. Au lieu d'avoir un énorme datacenter dans lequel on installe 25 serveurs câblés, on va tout mettre dans le cloud. On loue en tant que service cette infrastructure.
- Paas : Platform as a service. On rajoute 2 couches. On ne doit même pas acheter de licences de système d'exploitation et les installer sur les différentes machines qu'on a loué dans le cloud. On peut même directement louer tous les outils de développement.
- Saas : Software as a service. On rajoute la dernière couche. Au lieu d'acheter un DVD, venir dans l'entreprise et l'installer en local sur une app, on oublie. On va directement passer sur un modèle dans lequel les clients paient juste une licence d'utilisation et peuvent utiliser un service (en ligne, payant, gratuit ou autre). Amazon, GoogleDrive, Wooclap, Discord, Onedrive, Netflix, SAP, Spotify, ICloud, Facebook, Twitter, … .



Les différents services peuvent tourner.
Les dépenses dans le Cloud doublent de taille chaque année.
Voir Local VS Cloud slide 17.
Attention, le cloud ne résout pas tout tout seul. Par exemple, voir exemple Netflix slide 20.

Types de cloud

Hybrid Computing :

« Un Hybrid Cloud est un environnement de calcul qui combine un Cloud privé et un Cloud public avec des données et des applications pouvant être partagés entre eux. ». Un cloud privé, c'est lorsque l'on connecte plusieurs serveurs au sein d'une entreprise. Un cloud public, c'est le mettre sur un système ANS ou autre.

Avantages :

- Partager ce qui est nécessaire et garder certaines données privées.
- Pouvoir bénéficier sur CloudBursting (gérer les pics de demande informatique → si on fait le buzz en 5 minutes, une partie des connexions se fera sur le clou public afin d'éviter que l'app crashe.
- Toucher des localisations avec de mauvaises connexions (bateau en mer, …).

Edge Computing :

« L'Edge computing est une méthode de calcul qui permet de traiter les données au plus proche de la source. Par exemple, directement au niveau des capteurs IOT. ». Au lieu d'envoyer toutes les données dans le Cloud pour faire un méga calcul central, on utilise des parties de calcul directement plus proche de la source de données (précalcul).

Avantages :

- Diminuer la bande passante entre les capteurs et les systèmes de traitement de données.
- Ces systèmes sont utilisés pour les voitures autonomes: la puissance de calcul se situe dans la voiture, mais les mises à jours se font via le Cloud.
- Une autre utilisation est pour les drones automatiques de surveillances des pipelines de gaz: S'il y a un problème, un opérateur est directement contacté, pas besoin d'attendre le retour du drone à la base.

Pourquoi est-ce important de savoir où les données sont localisées ? Car la loi est différente pour chaque pays. Aux États-Unis, toutes les données stockées sur leur territoire peuvent être consultées par les autorités dans le cadre d'enquêtes. Maintenant, le RGPD permet de protéger ces données pour les européens. Il y a également eu une nouvelle règle aux USA : toutes les compagnies américaines seront obligées de fournir leurs données, peu importe dans quel pays se trouvent ces données.

Big Data

« Le Big Data désigne des ensembles de données numériques tellement volumineux qu'ils dépassent la capacité de calcul des outils classiques de gestion de base de données. ». Ces données proviennent des mobiles, des réseaux sociaux, des IOT, du Web, des transactions et pleins d'autres choses.
Les données sont dures à surveiller : Où est-ce qu'elles sont ? Où est-ce qu'elles vont ?

Les 3 V du Big Data

Pour être défini comme du Big Data, il faut 3 choses :

- Volume : La masse de données produites chaque seconde devient de plus en plus massive. L'ensemble des données produites depuis le Big Bang jusque 2009 est équivalent à la masse de données qui est aujourd'hui généré à chaque minute.
- Variété : Les données peuvent être structurées, semi-structurées ou non structurées. Tout devient données: Son, Texte, Vidéo, Image, Odorat, … . Seulement 20% des données sont structurées et stockées dans des tables de base de données relationnelles.
- Vélocité : Les données sont produites, récoltées et déployées de plus en plus vite. La vélocité représente la fréquence à laquelle les données sont à la fois générées, capturées, partagées et mise à jour. Par exemple, des messages de réseaux sociaux qui deviennent viraux en quelques minutes seulement. Vitesse de propagation des données.

Les 2 V bonus sont :

- Véracité : Les informations collectées doivent être aussi fiables et crédibles que possible. Ex : Il faut vérifier ses sources et éviter les attaques du type « Bombing » (noyer le web avec pleins d'informations). Comme en 2009 où la requête « trou du cul du web » menait vers le site de sarkozy.fr.
- Valeur : La valeur des données représente le profit qui peut être tiré de l'usage du Big Data. Les données et leurs analyses représentent un avantage concurrentiel non-négligeable. C'est intéressant de collecter des données d'un certain type, mais si on ne sait pas les exploitées, on ne saura rien en faire.

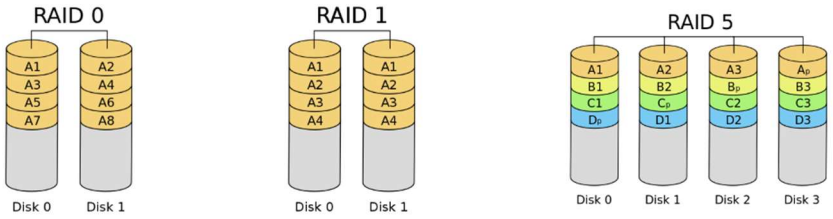
Les enjeux du Big Data

Les principaux challenges du Big Data consistent à collecter, stocker et exploiter la donnée. Les données sont toutes les informations qui ont été enregistrées numériquement.

Historique des stockages

Raid 0 : pair sur un disque et impaire sur l'autre. Si 1 crashe, c'est la merde.

Raid 1 : on les met tous sur les 2 disques. Plus de problèmes lors de crashes mais les coûts de stockages sont super élevés.



Raid 5 : séparer chaque disque (orange) en sous-disques. Stocker une bande de parité dans chaque disque = une sorte de back up des 3 autres disques (oranges). Si A_3 crashe, on peut utiliser A_1 , A_2 et A_p pour recalculer A_3.

Par contre, ça ne fonctionne pas pour les données de type continue.

Le Big Data a été créé par 2 technologies : Hadoop et MapReduce. Plus d'infos slides 37, 38, 39, 40.

Est-ce du Big Data :

- Churn chez American Express (permet de savoir quand un client va quitter l'entreprise) : Oui car ils utilisent tout ce qui est possible et inimaginables.
- Résultat des élections américaines : Non, beaucoup de données mais pas de variété, une seule source.
- Utilisation du service Uber : Oui, ils regardent les déplacements, l'âge, les informations des profils, le nombre de likes, le trafic routier, … .
- Bourse de New York : Oui, on regarde toutes les informations qui arrivent sur une société pour déterminer si cette société va monter ou descendre.
- Données d'enquête de satisfaction de la Playstation 5 : Non, il y a beaucoup de données mais une information d'une seule source et ce n'est pas mis à jour fréquemment.
- Données de capteurs IOT d'une récolte de lait : Non, le nombre de données est élevé, le transfert très rapide mais qu'un seul capteur avec une seule source sauf si on commence à connecter pleins de types de capteurs et qu'on lie les informations.
- Données de Palantir : Oui

3. Collecte de données

Si on nous propose de recevoir gratuitement une maison en Espagne ou un PC, on va choisir la 1ère. Pourtant, au niveau des chiffres de vente, c'est l'inverse → **on ne peut pas faire confiance aux données d'une manière aveugle.**

83% des gens ne votent pas Trump et 86% des gens aiment les jeux vidéos → **on ne peut pas faire correler plusieurs études qui n'ont rien à voir.**

Les valeurs seront aussi différentes selon la population interrogée et la localisation.

Quelles sont généralement les données les plus importantes pour le Business dans une entreprise avec au moins 50 employés ?

Les ressources par rapport au monde extérieur et les ressources par rapport à l'entreprise interne.

Il y a 2 familles d'outils : CRM et ERP. Les gros leaders se trouvent slide 4.

CRM :

Frontend. Un Customer Relationship Management (CRM) est en ensemble d'outils permettant la gestion de la relation client afin d'optimiser les processus de vente, de recueillir les informations sur les clients et permettre un suivi commercial cohérent. Lien avec les clients. Ex : fidélisation client, campagnes marketing ciblées, recherche de nouvelles tendances.

ERP :

Backend. Un Enterprise Resource Planning (ERP) est un ensemble d'outils qui permet de gérer les processus d'une entreprise en intégrant l'ensemble de ses fonctions. Gérer les processus internes d'une entreprise. Ex : gestion des ressources humaines, gestion comptable et financière, gestion des stocks et des ventes.

Quelles sont les données les plus fiables en entreprise ?

Majoritairement, ce sont les **contrats** et les **ventes**.

Quel est le plus grand concurrent interne des systèmes BI ?

Spreadmart : « un Spreadmart est un système d'analyse de données business qui se base sur des tableurs Excell ou autres bases de données bureau qui sont créées et maintenues par des groupes ou des individus pour réaliser des tâches qui pourraient l'être d'une manière plus structurées. ». Au lieu d'attendre les stratégies internes développées en général, ils cherchent à créer leur propre solution spécifique. C'est lorsqu'une personne trouve l'IT trop lent, qu'elle ne veut pas partager ses données et/ou qu'elle a 'toujours travaillé comme ça'. Sauf que ça ne peut pas fonctionner. Avec les filtres, les règles business, les manipulations de données, les macros, etc., c'est presque certain que les données deviendront inconsistantes.

Faut-il se débarrasser de tous ces Spreadmarts ?

+	-
• Contient de la connaissance business spécifique • Rapide et Flexible • Accessible • Pas cher à utiliser • Remplit les trous de fonctionnalités des grands systèmes • Efficace pour des tâches simples • Facile à utiliser pour des utilisateurs business	• Données inconsistantes dans l'entreprise • Perte de productivité • Erreurs dans l'import, l'export et le changement de sources de données • Augmente le risque • Aucun audit possible • Pas de documentation • Pas extensible • Pas de traçabilité • Single Point of Failure

Pas forcément. Si on développe un système et qu'on automatise une partie, il faut que ça couvre 90 à 95% des cas. Les 5 derniers pourcents, on essaie même pas de les développer. Ça sera la tâche du spreadmart.

Le risque : si la personne qui gère ces 5% se barre, tout s'écroule. Alors, que faire avec eux ?

➔ Il faut réussir à séparer les données, les trier. Toutes les données n'ont pas la même importance.

Il y a les **données stratégiques** et les **données tactiques**. Les données stratégiques sont importantes du point de vue de l'entreprise d'une manière générale. Les données tactiques sont un extrait des données stratégiques. Ils font leur petite analyse, une conclusion et jettent leur ancien système.

Système de tri

Si le système n'a que peu de valeur business → s'en débarrasser.

Si le système a peu d'utilisateurs business mais que c'est encore utile → pas forcément utile de le remplacer, le laisser vivre et mourir petit à petit.

Si le système fournit une analyse one shot et des rapports spécifiques et uniques → le laisser.

Si le système a une grande valeur business et/ou de nombreux utilisateurs business → l'intégrer dans une solution centralisée.

Collecte de données

« La collecte de données est une approche qui consiste à regrouper et analyser des informations spécifiques dans le but d'apporter des éléments de réponses à des questions pertinentes. ». Ces données peuvent être quantitatives ou qualitatives. Il est plus important de collecter moins de données mais en se focalisant à chaque fois sur une question.

Comment collecter des données externes à l'entreprise ?

Achat de données, questionnaires et enquêtes, interview de clients (donne moins de données mais de plus haute qualité), utiliser des outils de type web scrapper (scanner le net pour tenter de récupérer les données sur d'autres sites), créer des plateformes et de communautés, prendre des données de recherches existantes, expérimenter et mesurer les impacts, profiter du packaging d'un produit (QR code menant à une enquête de satisfaction). Exemple Puratos voir notes p.5.

Mécanismes pour collecter des données :

Acheter des données :

- Recherches marché : Analyse de l'évolution du marché global: Est-ce qu'il grandit? Quelles sont les tendances? Quelles sont les prévisions ? Quel est le potentiel du marché? Quelles sont les différenciations à envisager?
- Product intelligence : En se basant sur des données, on peut collecter et analyser les performances d'un produit. Le but est de pouvoir déduire les nouvelles fonctionnalités intéressantes, dans quels pays se focaliser et surtout comment se différencier de la concurrence.
- Market intelligence : Ce sont les informations relatives à l'environnement d'une société : Les clients, les partenaires, les compétiteurs et les Thought Leaders.

Web Scrapper : « Le Web Scrapping est une technique utilisée pour extraire des données de différents sites web. Elle peut être utilisée en utilisant un bot ou un script afin d'extraire les différentes informations se trouvant sur le net pour les transformer en données internes structurées. ». Ces données seront structurées en SQL, JSON, CSV, XML et autres.

- Différents types de critiques.
- Contenu : nouvelles tendances.
- News : être au courant de l'actualité et de ce qui pourrait changer.
- Forum : avoir l'avis des gens.
- Team.
- Produits.
- Social : ce que les gens disent sur nous, et la concurrence.

Le Web Scrapper est légal lorsque le site ne mentionne nulle part que c'est interdit.

Comment réaliser un bon questionnaire ?

1) Définir l'objectif : Définir les informations qui doivent être collectées. Quelle est l'information, voir le problème spécifique auquel on voudrait répondre? Quels sont les critères pertinents qui devraient être mesurés ? Comment distribuer le questionnaire (si on publie juste à l'IPL, on aura pas le même résultat que dans le monde) ? Quel public viser ? Age, Sexe, Activité, Géographie, …
2) Rédiger le questionnaire : Introduction (on aime bien savoir dans quel but ce questionnaire à été rédigé) → Questions pour savoir le profil de la personne interrogée → Listes de questions → Remerciements (accroche le client).
3) L'envoyer : Tester le questionnaire sur un échantillon (pas toute la population, attendre un feedback) → Le diffuser progressivement.

Voici les éléments à prendre en compte :

- Il faut faire attention à la compréhension des questions: elles doivent être claires, compréhensibles, et il ne peut y avoir qu'une seule interprétation possible de la question.
- Dans la structure d'un questionnaire, les questions globales doivent être posées avant les questions générales : dans le cas contraire, les réponses détaillées pourraient influencer la perception globale.

- Le questionnaire doit être aussi court que possible: il faut éviter les phénomènes de lassitude.
- Il y a 2 manières de définir l'ordre des questions :
- Le classement par ordre décroissant d'importance : On maximise les chances d'avoir des réponses de meilleures qualités pour les questions les plus importantes.
- Le classement chronologique: suit le déroulé d'une expérience client : accueil - > commande -> livraison. Cette organisation permet de réactiver les conditions de l'expérience tout au long du processus ce qui permet plus de fiabilité dans les réponses à la clé.

Le type de questions ainsi que la façon de rédiger une interview se trouvent aux slides 28, 29 et 30, et ne sont pas à connaitre.

Méthode d'expérimentation : A/B testing

« L'A/B testing est une méthode permettant de comparer deux versions d'un système – A et B – afin d'identifier celle qui convient le plus aux besoins analysés. La version A est la version actuelle(version de contrôle), tandis que la version B est la version modifiée (la page de traitement). Les performances de ces deux systèmes peuvent être facilement comparées si elles sont testées simultanément. ». Lorsque l'on change quelque chose sur notre système et un principe, on doit pouvoir expérimenter ce qu'on a fait. On envoie alors à un échantillon la nouvelle variation et on analyse.

Exercice : « Après avoir terminé vos études à l'IPL, vous décidez de monter avec des amis une start up de livraison de matériel par drones robotisés. Le fonctionnement est le suivant : Les clients installent une application gratuite qu'ils utilisent pour faire appel au service. Quand un utilisateur veut envoyer une commande par drone robotisé, il fait appel à l'application qui lui envoie un drone. Ce dernier arrive et se pose sur le point relais le plus proche. L'utilisateur peut ensuite placer un paquet de moins de 5 kilos dans le drone robotisé. Après avoir spécifié une destination via l'application, le drone robotisé prendra son envol et livrera le colis au destinataire. Etant des spécialistes Data, vous voudriez utiliser des données afin d'optimiser les services proposés. Décrivez 3 techniques que vous utiliseriez pour collecter des données dans ce contexte, et comment les utiliseriez-vous dans votre système d'aide à la décision ? ».

Le but est d'optimiser le service proposé. On peut l'optimiser de 2 manières différentes : améliorer le service existant ou réduire le prix ou autre. Avec un questionnaire, le taux de réponse est faible donc bof.

1) Question rapide avant l'envoi du colis : voulez-vous le livrer urgemment ? Pour ensuite faire des statistiques sur la fréquence des colis urgents. Comment vous nous avez trouvé ? Est-ce que la limite du poids courante est suffisante ? → frustration ou non.
2) Prendre les données de l'app pour voir les points relais les plus souvent utilisés.
3) Voir l'utilisation qui a été faite et voir combien de drones sont en bon état des différents endroits → si tous les drones de Charleroi sont cassés, revoir la manière de communiquer ou créer un système particulier de livraison.
4) Questions du formulaires qualitatives et quantitatives : le temps de livraison est-il satisfaisant pour le client ? Sur 5 étoiles → permet de prendre des décisions. Quelle est la partie de frustration → utiliser WorldCloud pour savoir les mots les plus utilisés pour se baser sur les recommandations des clients.
5) Habitudes de paiement clients.
6) Synchroniser avec les réseaux sociaux : inscription en un clic (mais attention à l'éthique), invitation à liker une page, … .
7) Analyser le type de colis envoyés (rajouter donc la fonctionnalité de mettre un type lors de l'envoi.

Il faut au moins une question liée à l'amélioration interne (login et autre), une question liée à une collecte extérieure et une question liée à une étude de marché ou autre.

Exercice : « Vous acceptez un stage de Data Engineer pour le gouvernement belge. Pour votre premier jour au travail, Sophie Wilmes vient vous voir et vous demande de mettre en place une stratégie de collecte de données dans le cadre de la crise du COVID-19. Le nombre de cas de contaminés fluctuent plus rapidement que le cours du Bitcoin et il est temps de mettre un peu d'ordre dans cette histoire. Pour prendre des décisions

précises aussi bien sur la taille de la « bulle sociale » que sur les évènements qui pourront avoir lieu, il est important d'avoir accès à des informations précises et pertinentes. La première ministre belge vous donne une tape sur l'épaule en vous donnant carte blanche. a) Décrivez 3 techniques que vous utiliseriez pour collecter des données dans ce contexte. b) A quoi ces données serviront elles au gouvernement pour prendre des décisions sur les prochaines mesures COVID ?

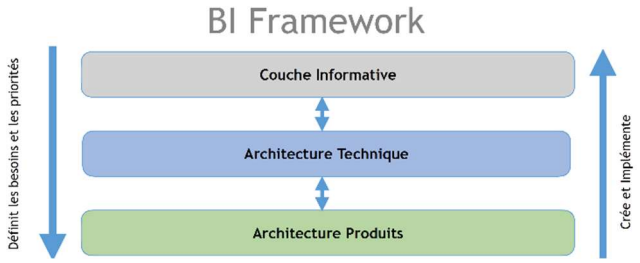
- collecter le plus d'informations possibles sur le Covid : statut des différents vaccins/médicaments, facteurs qui peuvent aggraver, études en cours.
- Envoi aléatoire de visites de police dans les foyers en quarantaine : respect ou non ?
- Récupération et mise en place d'une DB centralisée reprenant les données de tous les hôpitaux publics : meilleure analyse des données.
- Questionnaire dans la rue pour savoir si les personnes comprennent les mesures → qualitatif.
- Analyser les heures des transaction bancaires pour déterminer les heures où le peuple est le plus actif.
- Comparer les données/la situation par rapport à d'autres pays (dans une situation de libre échange).
- Augmenter le nombre de tests pour améliorer la qualité des statistiques.
- Mécanisme où la personne répond à une question en échange de quelque chose.
- Utiliser les données de machines Stib pour constater où se situent les forts déplacements.

Il faut faire attention aux sondages car ils peuvent être biaisés : respect de la bulle → les gens ont peur de la condamnation donc risques de réponses mensongères. Dans ce cas, il faut préciser que c'est anonymisé.

4. Architecture BI

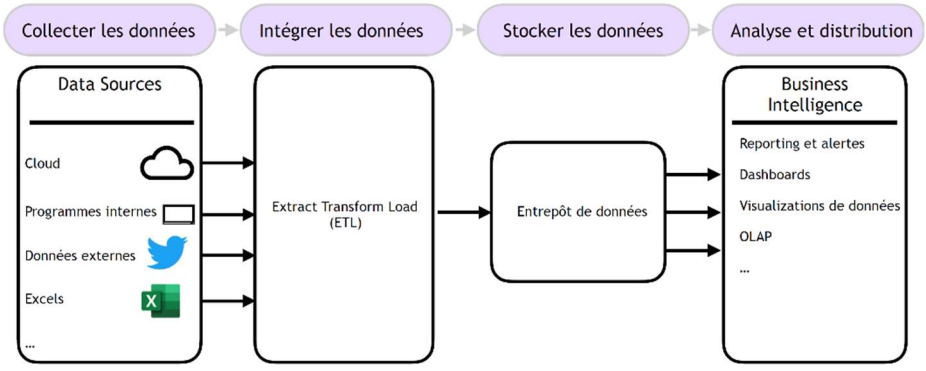
Quelle est la fréquence de chargement de sources de données en entreprise ?

Tout le temps ! **Il y a changements, ajouts, suppressions NON STOP !**



Il va alors falloir créer une architecture résistante aux changements constants. On va alors définir plusieurs couches :

- Couche informative : elle définit le contexte business nécessaire pour déployer une solution BI. Quoi ? Quel type d'analyse sera nécessaire? Quel type de décision sera prise? Quelles fonctions doivent être supportées? Qui ? Qui aura accès à la plateforme ? Où ? Où se situe la donnée. Où sera-t-elle intégrée? Où sera-t-elle consommée? Pourquoi ? Quels sont les besoins business et techniques ?
- Architecture technique : c'est la partie qui compose l'architecture des données et les techniques utilisées. Il y a 4 phases : Collecte de données (en connectant les différentes sources). Ensuite, en prenant ces données, on doit les intégrer (mettre différents connecteurs pour prendre les données aux différents endroits où on les a stockés), les intégrer. Ensuite, les stocker sous un format particulier. Pour ensuite faire de l'analyse et de la distribution.



Qu'est-ce que la BI et à quoi ressemble une architecture générale ? Dire qu'il y a ces 4 blocs qui existent + préciser avec nos mots 1 des blocs (qu'est-ce qu'il veut dire en particulier).

- Architecture produits : slide 7.

RGPD

Le Règlement Général de Protection des Données (RGDP) est le texte de référence en matière de protection des données au niveau européen. Pour utiliser les données personnelles des Européens, les entreprises devront obtenir leur consentement et être clair sur leur utilisation. L'utilisateur a un droit à l'oubli, et le consentement de l'utilisateur a une durée dans le temps. Mis en place pour protéger la vie des citoyens. Tous les développeurs ont dû rajouter dans leur DB un lien avec une date d'expiration (consentement valable 1 an). On a aussi le droit de demander à n'importe quel moment qu'une entreprise nous oublie. Ce n'est pas lié à la localisation de l'entreprise mais à l'origine de l'utilisateur (Europe). Il y a vraiment des entreprises qui ont été condamnées pour ne pas avoir respecter le RGPD. Les amendes sont ultras lourdes. **On peut utiliser les données de manière illimitée si c'est anonymisé, mais ça requiert le consentement.**

L'Europe a été le premier continent à mettre ces règles de protection en place. En Asie et en Afrique, pas encore. En Amérique, c'est plus selon les états.

Pour suivre les règles du RGPD, il est important de savoir répondre à ces questions :

- Quoi ? Quelles données sont utilisées pour créer les applications BI ?
- Qui ? Qui aura accès à la plateforme ?
- Pourquoi ? Pourquoi les utilisateurs ont accès aux données et ce qu'ils auront le droit de faire avec ?
- Quel ? Quelle est la procédure à mettre en place en cas de crise ?

Combien de sources de données y a-t-il en général dans une entreprise ?

Ça dépend d e l'entreprise. Dans une entreprise, chaque type de données est mis dans une boîte regroupant toutes les données du même type. L'importance du BI est de **savoir dans ce bordel, lesquelles de ces données ont de la valeur**.

Intégrer les données

On va prendre pleins de sources de données et on va les faire passer par l'ELT. Prendre toutes les données → les extraire → les transformer pour arriver sur une méthodologie commune → les charger dans l'entrepôt de données.

Le but de l'ELT est de nettoyer les données. Les données devront être stockée le plus proprement possible.

ELT

« Un ETL permet ainsi l' extraction, la transformation et le chargement de données depuis des sources diverses vers des cibles préalablement définies. ». Donne une logique aux données.

- Connectivité : il faut pouvoir se connecter à de nombreuses sources de données: bases de données, xml, textes, … .
- Performance : les ETL travaillent sur de grands volumes de données.
- Flexibilité : Il faut pouvoir merger des données, matcher des données, splitter des données, etc… .
- Validation : vérifier la qualité des données.
- Standardisation : mettre les données dans un format standard et commun. La définition sera différente si c'est une BD européenne ou américaine (football). Si on ne standardise pas, ça ne sera pas du tout la même chose. Pour les noms des pays, il vaut mieux avoir des codes (BEL).

Des problèmes d'intégrations peuvent être rencontrés (voir slide 24).

Structurer les données

Les noms des colonnes doivent être bons. Adresses séparées. L'importance de la splitter est de permettre plus facilement la détection d'erreur pour le nom d'une ville par rapport à un code postal ou autre.

Nettoyage

1) Définir les métadonnées voulues (noms de colonnes, etc.).
2) Définir l'encodage voulut (Homme/Femme en H et M, null en 0).
3) Standardiser les unités de mesures (format de date, valeur monétaire, etc.).
4) Corriger les erreurs d'encodage (Smif, Smith, Smiss, etc.). Comment (3 méthodes):
   - Soundex basé sur la phonétique et la prononciation, mais bof (notes p.8 et slide 33). Pas ouf pour les mots longs.
   - Fuzzy Matching (slide 35). Pas fonctionnel pour les mots courts mais bien pour les longs.
   - Named Entity Resolution (slide 36). Permet de trouver la structure des mots.
5) Corriger les noms qui peuvent s'écrire différemment (Belgique, Belgie, etc.).
6) Corriger les données en double. Comment (choix):
   - Regarder la meilleure source de données (la plus officielle)
   - Regarder la date la plus récente
   - Alerter et validation manuelle
7) Prendre une décision sur les données vides.
8) Splitter des colonnes (avenue pagodes 25 => rue: pagodes numéro: 25).
9) Valider les données (l'âge ne peut pas dépasser 150 ans, etc.).

Comment charger les données ?

Il faut faire attention a ne pas charger deux fois la même information.
Il faut ajouter des informations sur la provenance des données: si elles viennent des archives ou des données courantes. Si les données proviennent d'un chargement qui a échoué et qui a été rechargé correctement.
Chargement initial : une fois. Si la BD est vide, prendre toutes les DB et faire un chargement complet de toutes les données.
Chargement incrémental : se répète dans le temps. Si la DB n'est pas vide, regarder tous les X temps (par un scheduler) quelles sont les nouvelles données par rapport aux anciennes (par rapport aux dates : delta).
Les systèmes BI sont des systèmes de lecture seule : on prend les données, on les copie (en les modifiant peut-être) sur un système central nettoyé pour pouvoir faire des analyses. On ne modifie pas les vraies données des vrais systèmes.

5. Architecture et Data Warehouse

Quelles sont les grandes difficultés en entreprises pour implémenter des projets BI?

Besoins :	- Pas assez détaillés - Trop techniques ou trop business (utilisateur final) - N'implique pas les bonnes personnes initialement (trop de gens marketing qui n'ont aucune technique en informatique) - Pas mis à jour quand nécessaire - Cadre pas bien défini - Personne ne sait ce qui doit être fait
Support externe :	- Pas les bonnes personnes impliquées dans le projet - Mauvaises priorités - Coûts - Pas de documentation
Accès aux données :	- Problèmes de droit d'accès et de sécurité - Personne ne sait où sont les données - Problèmes légaux (RGPD)
Architecturaux :	- Trop de technologies dans l'entreprise - Différentes instances créées dans différents endroits, époques, équipes, etc… - Legacy - Besoins très spécifiques qui ne sont pas standards

User experience :	- Utilisateurs habitués à leurs anciens systèmes -Les utilisateurs n'aiment pas toujours le changement
Changements :	- Les changements technologiques sont constants - Les besoins changent d'une manière incessante

A chaque fois que l'on cherche a créé un nouveau système BI, c'est important d'avoir au moins l'un de ces éléments (une valeur) :

Augmenter les revenus | Nouveautés pour les produits | S'adapter aux réglementations légales | Réduire les coûts | Réduire les risques

Stocker les données : entrepôt de données

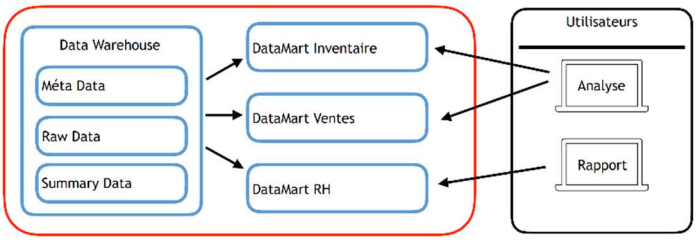
Data Warehouse : « Un entrepôt de données, ou data Warehouse, est un regroupement d'informations structurées, historiées et provenant de différentes sources de données. Il sert à avoir une vision centralisée et universelle des informations de l'entreprise afin de pouvoir effectuer des analyses et créer des systèmes d'aide à la décision. ». Il permet d'être la vue de toutes les différentes sources de données qu'il y a au sein d'une entreprise.

Un Data Warehouse est important car il permet de n'avoir qu'un seul point d'accès. Au lieu d'avoir 20 000 imports dans chaque, on fait 1 grand import.

Pour éviter un dépassement du nombre maximal de lignes, il faut splitter en plus de fichiers chaque fichier initial. « Sur son site, le gouvernement britannique indique qu'une solution de contournement est d'ores et déjà en place. Elle consiste à fractionner les fichiers volumineux en unités que la version d'Excel utilisée est capable de prendre en charge. De plus, un examen complet de bout en bout de tous les systèmes a fait l'objet d'instauration pour éviter que des erreurs similaires ne se reproduisent. Les observateurs restent néanmoins d'avis que l'utilisation d'un véritable système de gestion de base de données (SGBD) reste plus indiquée pour des cas de figure de ce type. ».

Décomposition d'un Data Warehouse

Les Data Warehouses contiennent généralement de gros volumes de données et sont très complexes à concevoir. Pour faciliter la création et la gestion de ces derniers, ces derniers sont divisés en de plus petits modules appelés Data Marts. Ces modules peuvent être regroupé par fonctions (un data mart pour les ventes, pour les commandes, pour les ressources humaines) ou par une structure organisationnel de la société.



Meta data : signalement de la structure (de quelle manière les données sont structurées).
Raw data = les vraies données chargées complètement.
Summary data : résumé des données pour avoir certains précalculs faits pour ne pas surcharger le serveur.

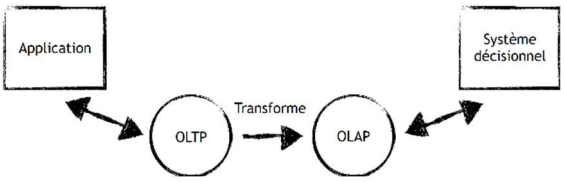
DataMart : petits marchés sur lesquels l'utilisateur va faire ses analyses. DataMart ventes contient les données de ventes.

Base de données

Elles sont différentes en BI et celles qu'on voit en SQL. Qu'est-ce que c'est ?

« OLTP - Online Transaction Processing » =

- Système transactionnel rapide, performant, de production
- Données normalisées
- Les transactions OLTP sont spécifiques : elles impliquent généralement un seul enregistrement ou une petite sélection d'enregistrements. Ex : un client envoie une transaction à un autre: ça ne concerne que ces 2 personnes.
- Pas adapté pour les systèmes d'analyses



OLAP : traitement analytique en ligne.
Si on veut faire des analyses sur des OLTP, c'est compliqué car il y a énormément de select à faire.

Modèle dimensionnel (en étoile slide 20) : utilisé pour partir de nos BD pour une nouvelle représentation sur un autre type de BD. Le centre de l'étoile est la table de fait. Les autres sont de dimensions.

OLAP

Tables de faits et dimensions. « OLAP(On-Line Analytical Processing) représente les technologies qui permettent une prise de décision stratégique rapide et fiable sur des données modélisées de manière multi dimensionnelles. Les données sont regroupées selon des catégories qui ont un sens pour les utilisateurs business.

Une table de fait est une table qui contient les données observables (les faits) que l'on a sur un sujet que l'on veut étudier, selon divers axes d'analyse (les **dimensions**). Une table de fait contient 2 éléments : les clés primaires de toutes les dimensions d'analyse qu'on a créé autour et les mesures (ce que ça va contenir).

Les « faits » sont majoritairement numériques, puisque d'ordre quantitatif. Il peut s'agir des ventes, des dépenses, du niveau d'un inventaire,… . Ils représentent en moyenne 90/100 des données. Les dimensions fournissent le contexte: qui, quoi, quand, où, pourquoi et comment des faits. Une table de faits suit généralement un schéma en étoile et est entourée de plusieurs tables de dimensions.

Table de faits : types de mesures

Mesures additives : Ce sont des mesures qui peuvent être additionnées à travers toutes les dimensions. Ex: le nombre d'articles acheté en ligne, … . Ce type de donnée peuvent être agrégée par date, clients, produits, vendeurs, … . Nombre de jeux vendus, bénéfice d'un produit = montent vente – coût, nombre de vues de la vidéo Baby Shark, chiffre d'affaires d'un stand de vente de jus d'orange.

Mesures semi-additives : Ce sont des mesures qui peuvent être additionnées à travers certaines dimensions mais pas toutes. Ex: Solde de compte en banque, nombre d'étudiants en classe, niveau d'inventaire, … . On ne peut pas additionner 12 mois de valeur d'un compte bancaire pour avoir le total sur le compte. Mais on peut à un moment donné dans le temps T voir le total d'argent sur le compte en additionnant les valeurs à cet instant T. Stock de blés dans un hangar (~~additif~~ : ça ne signifie pas que j'avais en tout la somme des 2, ~~non additif~~ : je peux voir les sommes de T à chaque instant T), stock de papier toilettes à l'IPL.

Mesures non additives : Ce sont des mesures de tables de faits qui ne peuvent être additionnées à travers aucune dimension. Ex: Les prix unitaires, les températures, les ratios, … . Pourcentage de profit (on ne sait pas additionner les %), prix de ventes d'un produit, pourcentage de promotion durant les soldes, marge de vente d'un stand de jus d'orange.

Une table dimension est une entité qui définit le contexte business pour les faits utilisés dans une entreprise. Elle permet à la base de donnée de ne pas être surchargée de données redondante. D'un point de vue business, le but principal d'une dimension est d'utilisé ses attributs pour filtrer et analyser les données.

Les attributs d'une dimension doivent être:

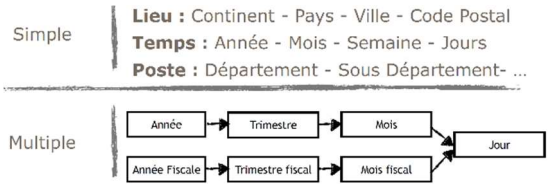
- Descriptif : pour que les personnes business puissent les comprendre
- De qualité : pas de valeur manquante, obsolète, erronée
- Unique : les valeurs doivent être identifiable uniquement
- Valide : les données doivent être utile au business

Les attributs servent à:

- Filtrer/Restreindre les requêtes
- Décrire les résultats

Hiérarchie dimensionnelle

Les dimensions sont souvent hiérarchisées: Elles regroupent les informations d'une manière compréhensible et utilisable par le business. Normalement tout est enregistré dans une seule table de dimension. Ces hiérarchies sont utilisées pour définir des chemins d'accès aux données.



Règle sur la clef primaire d'une table de dimension

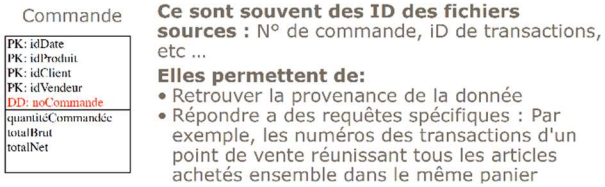
Il ne faut pas avoir de logique dans un identifiant ou une PK car parfois, il faut les changer.
Une règle de bonne pratique est d'utiliser une SurrogateKey comme clé primaire : c'est une clé unique, dans un seul champ et ayant une valeur sans aucune logique. Avantages :

- Performance : accès par index et jointures accélérés
- Robustesse : ne change jamais contrairement à une clé naturelle
- Cohérence : quand on regroupe des dimensions de différentes sources de données, il y a souvent des inconsistances et des incompatibilités dans les clés primaires utilisées dans les différents systèmes.
- Évolution : Les clés primaires peuvent évoluer à travers le temps et avec d'autres conventions utilisées à différents moments dans le temps.

Au lieu de faire ça : 

, il vaut mieux faire ça : 

Dimensions dégénérées



Une dimension dégénérée agit comme une clé de dimension dans la table de faits, cependant elle n'est liée à aucune dimension parce que tous ces attributs dignes d'intérêts sont déjà dans d'autres dimensions. C'est donc une clé de la table de faits n'ayant que elle-même en attribut. Dimension qui n'a pas de table. Ça parait comme une PK mais ça ne crée pas d'axe d'analyse.

Mise à jour d'une table de dimensions

Type 1 : Si on ne veut pas conserver l'historique d'un champ. Par exemple: il ne faut pas forcément conserver les changements de mail d'un client? Dans ce cas, un simple update de la table fera l'affaire. Attention: il devient impossible de faire des analyses sur l'ancienne valeur.

ID Produit	Description	Code
123456	Article révolutionnaire et vendu beaucoup trop cher	AZBCDG

↓

ID Produit	Description	Code
123456	Article révolutionnaire et vendu beaucoup trop cher	1 - AZBCDG

Type 2 : Ce sont les données où on voudrait garder l'historique d'un champ. Par exemple: le prix d'un produit. Pour ce faire il faut dupliquer la ligne existante, modifier la nouvelle entrée et ensuite rendre la nouvelle entrée active soit via un flag d'activité, soit via des champs dates.

ID Produit	Description	Code	Prix	DateFrom	DateTo
123456	Rocket League	AZBCDG	10,99 Euros	14/02/2018	14/02/2019

↓

ID Produit	Description	Code	Prix	DateFrom	DateTo
123456	Rocket League	AZBCDG	10,99 Euros	14/02/2018	14/02/2019
123456	Rocket League	AZBCDG	6,99 Euros	14/02/2019	31/12/9999

Type 3 : Ce sont les données où on voudrait garder un historique limité. Par exemple le dernier changement. Il faut ajouter autant de colonnes que de changements désirés, avec les dates associées.

ID Produit	Description	Code	Ancien Code	Date du Changement
123456	Rocket League	ABCDE	ACDC	14/02/2018

↓

ID Produit	Description	Code	Ancien Code	Date du Changement
123456	Rocket League	ACDC	3615	14/03/2019

Comment assembler les schémas en étoiles pour créer un Data Warehouse ?

1) Mettre en œuvre un Data Warehouse :

		+	-
Top-Down	Définir toutes les 'étoiles' possible et les implémenter d'un coup. Il faut connaitre à l'avance toutes les dimensions et faits de l'entreprise.	Ça donne une vision très claire des données et du travail à faire.	C'est très dur à mettre en place.
Bottom-Up	Créer les étoiles une par une. Ensuite les regrouper par des niveaux intermédiaires jusqu'à avoir un schéma pyramidal.	Simple à réaliser.	L'intégration des données est lourde et il peut y avoir de la redondance entre les étoiles (car elles sont créées indépendamment les unes des autres).
Middle-Out	Approche hybride qui consiste à totalement définir l'entrepôt de données (toutes les dimensions, tous les faits, toutes les relations), pour ensuite créer des divisions plus petites et plus gérables et les mettre en œuvre. Il faut découper la conception par éléments en commun et réaliser les découpages un par un	Tire le meilleurs des méthodes Top-Down et Bottom-Up.	Il faut parfois faire des compromis par exemple: dupliquer des dimensions identiques pour des besoins pratiques

Possibilités de navigation dans les cubes

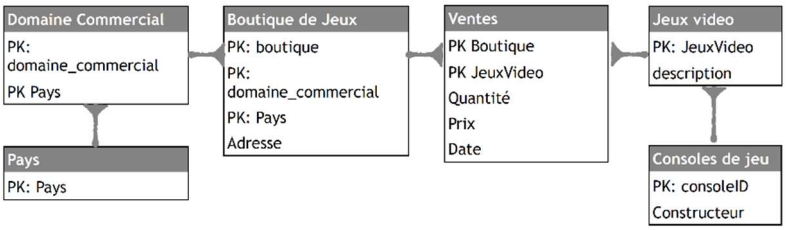
Le drill down : c'est la possibilité de « zoomer » sur une dimension, par exemple d'éclater les années en 4 trimestres pour avoir une vision plus fine, ou de passer du pays aux différentes régions. Zoomer.

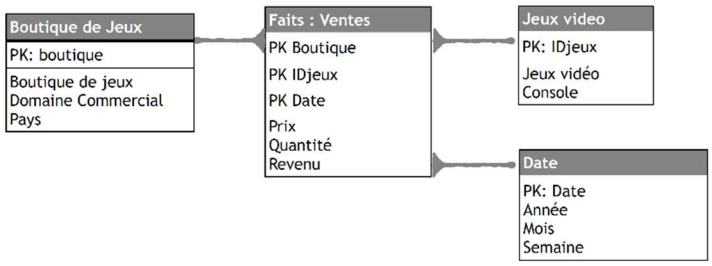
Le drill up : c'est l'opération inverse qui permet d'« agréger » les composantes de l'un des axes, par exemple de regrouper les mois en trimestre, ou de totaliser les différentes régions pour avoir le total par pays. Dézoomer.

Le slice : c'est une opération qui entraîne une permutation des axes d'analyse, par exemple, on peut vouloir remplacer une vue par pays/régions par une nouvelle vue par familles et gammes de produits. Permuter une dimension d'analyse.

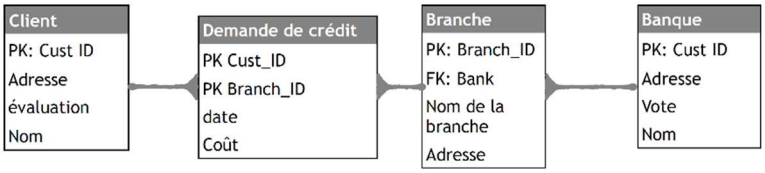
Le drill through : lorsqu'on ne dispose que de données agrégées, le drill through permet d'accéder au détail élémentaire des informations (chaque vente de chaque produit à chaque client dans chaque magasin).

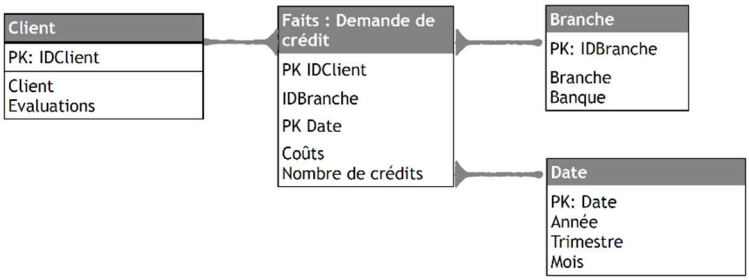
Exercice : Le schéma DB suivant représente les ventes de jeux-vidéos dans différentes boutiques de jeux . La dimension temporelle est: Année – Mois – Semaines. Identifiez la table de faits et les dimensions d'analyses possible. Modélisez ensuite le schéma en étoile.



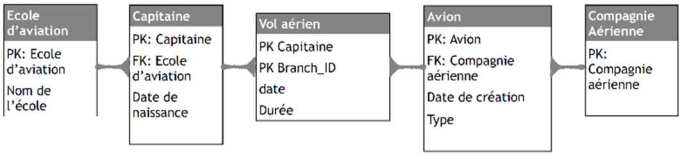


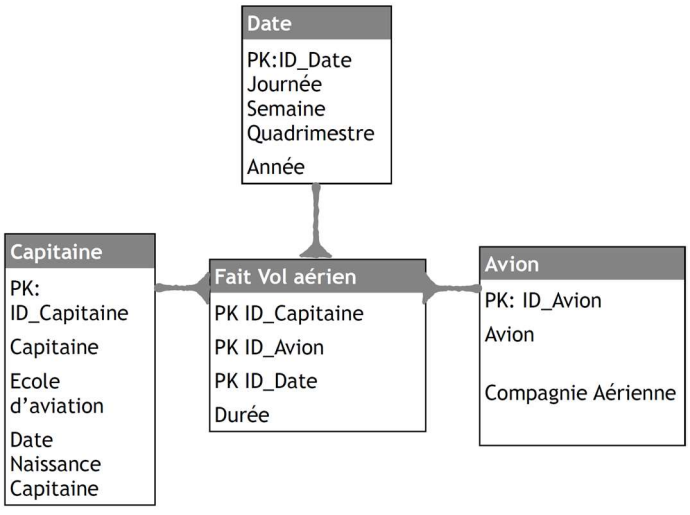
Exercice : Le schéma DB suivant représente un client demandant un crédit a une banque. La dimension temporelle est: Année – Trimestre – Mois. Identifiez la table de faits et les dimensions d'analyses possible. Ajouter un nouvel attribut de fait : nombre de crédit. Modélisez ensuite le schéma en étoile.





Exercice : Le schéma DB suivant représente un commandement de bord volant dans un avion. La dimension temporelle est: Journée, Semaine, Quadrimestre, Année. Identifiez la table de faits et les dimensions d'analyses possible. Modélisez ensuite le schéma en étoile.





6. Power BI

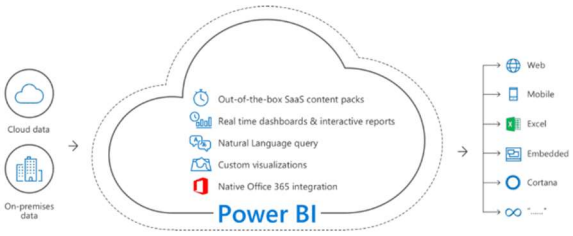
Pourquoi Power BI ?

Car c'est facilement intégrable à tout ce qui tourne sur Windows, la solution est gratuite et c'est le leader du marché BI. Concurrents slide 3.

Analyse et distribution

C'est le logiciel Power Bi qui s'en occupe. L'offre Power Bi :

- On prend des données locales ou dans le cloud.
- On permet d'utiliser Power Bi comme un service (Saas).
- On permet d'utiliser Power Bi pour fournir des dashboards et des rapports qui se mettent à jour en temps réel.
- …



Il y a différents outils cachés dans Power Bi : Power Query (solution ELT→Excel), Power Pivot (composant de modélisation et d'analyse), Power View (visualisation interactive des données), Power BI Published (dashboard en ligne).

Avantages de Power BI : intégration complète avec Office 365, analyse en temps réel, facilité d'utilisation.

7. Outil de recommandation

Spotify utilise 3 types de recommandations :

- Filtres collaboratifs : regarder nos préférences VS celles des autres.
- Natural Language Processing : regarder le type de ration de paroles, réseau neuronal, comme la reconnaissance d'images → matcher les similaires.
- Modèles audio : basé sur les fréquences → on écoute pas à chaque moment de la journée les mêmes types de fréquences.

Collaborative filtering

« Le filtrage Collaboratif est une méthodologie de prédiction automatique (filtering) des centres d'intérêt d'un utilisateur se basant sur la collection des préférences de nombreux utilisateurs (collaborative). ».

Généralement, on se base simplement sur l'avis des personnes qui ont des goûts similaires.

Un algorithme de Filtres Collaboratifs fonctionne en général en cherchant un sous-groupe de personnes ayant des goûts similaires.

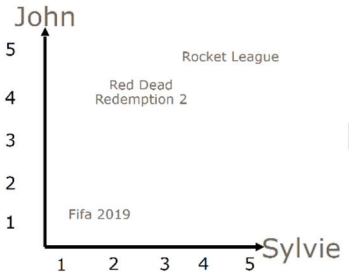
Les Filtres Collaboratifs ont souvent besoin des éléments suivants:

- Une participation active des utilisateurs (like, don't like, etc.)
- Une manière de représenter l'intérêt d'un utilisateur
- Un algorithme pour grouper les personnes avec des centres d'intérêts similaires

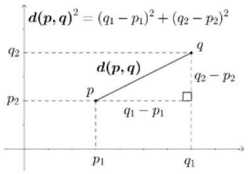
1) Un Utilisateur indique sa préférence sur le système (ex: like sur Netflix, ou nombres d'heures de visionnage d'une série).
2) Le système Match le rating de cet utilisateur avec celui des autres utilisateurs pour trouver le groupe de gens avec des goûts similaires.
3) Avec la liste des utilisateurs similaires, le système recommande les items que d'autres personnes du cluster ont appréciées et qui n'ont pas encore été notées par l'utilisateur.

Trouver des utilisateurs similaires

Pour calculer un score de similarité entre 2 utilisateurs, il existe 2 méthodes : la distance euclidienne et la corrélation de Pearson.



La distance euclidienne, se basant sur le théorème de Pythagore, est calculée de la manière suivante



Distance euclidienne : Une manière simple de calculer la similarité entre 2 personnes est de représenter les scores que les utilisateurs ont donnés en commun sur un graphe. Si la distance entre la diagonale parfaite et la leure est petite, alors ils sont similaires. Le désavantage est que si l'un est très gentil et l'autre très méchant, leur note ne se valeront pas même si ils ont les mêmes goûts.

Corrélation de Pearson : Une corrélation de Pearson est un nombre entre -1 et 1 qui indique la corrélation linéaire entre 2 variables X et Y. L'avantage de la corrélation de Pearson sur la distance euclidienne, est que cette dernière corrige les erreurs d'inflations. Par exemple: Un utilisateur qui donne toujours un score légèrement plus grand qu'un autre, mais qui a les mêmes préférences. Plus c'est proche de 0, moins on a de préférences similaires.