Introduction à l'informatique décisionnelle

Le développement de l'informatique décisionnelle est lié au fait que les masses de données traitées par les organisations augmentent considérablement. Cela vient de plusieurs facteurs. Sans prétendre à l'exhaustivité, on peut notamment citer :

- le développement des technologies de l'information et de la communication qui permettent de collecter des données de manière automatique
- Le développement des réseaux qui permet d'accroître le champ d'action des organisations (entreprises, associations etc...)
- La taille des organisations

Si on prend par exemple le cas d'un site Web commercial, les transactions, ainsi que tout type d'informations, sont enregistrées automatiquement dans des bases de données en continu. Par ailleurs, le site Web recouvre une zone de chalandise bien plus importante que celle d'un commerçant "traditionnel". Sous l'effet de ces deux facteurs, la masse des données stockées augmente.

Les masses de données à disposition des organisations deviennet donc considérables et il est impossible pour un être humain d'en faire la synthèse. Apparaissent donc des **S**ystème **I**nteractifs d'**A**ide à la **D**écision ou SIAD qui permettent de visualiser les données, d'obtenir des analyses automatiques sur ces données, de manière rapide, pour ensuite fonder la prise de décision. En anglais, on parle de DSS (**D**ecision **S**upport **S**ystem).

Dans les chapitres précédents, on a vu comment modéliser des systèmes de traitement et de stockage de l'information, qui sont la première pierre vers la construction de systèmes de stockage de l'information. Dans cette partie, on étudie la manière dont des masses de données disponibles pour une entreprise sont traitées pour être analysées et exploitées ensuite.

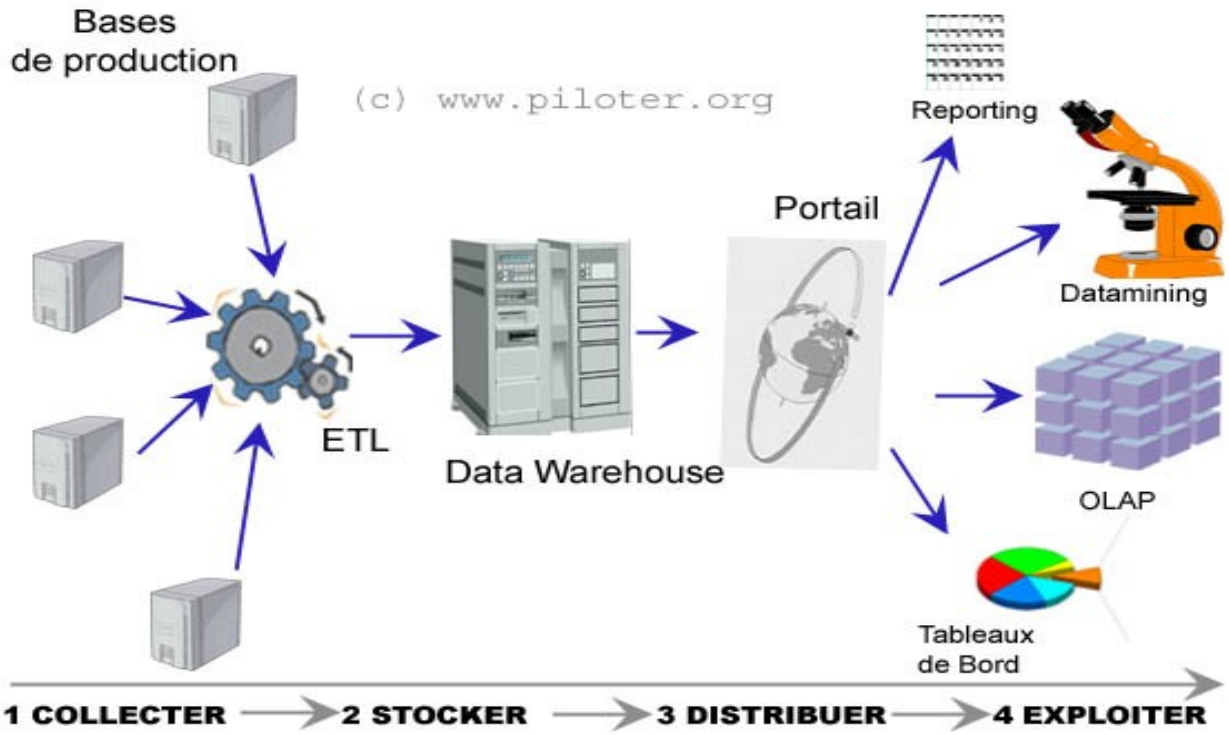
La notion de SIAD

La décision est un moment clé de la vie des organisations. Les systèmes d'aide à la décision sont là pour aider les utilisateurs à prendre des décisions. Le SIAD est la partie du système d'information qui étudie l'information stockée par le système d'information pour en donner des représentations, pour en tirer des informations permettant de fonder la décision. On parle de système interactif d'aide à la décision, parce que la décision reste un processus complexe qu'il n'est pas possible d'automatiser : un système d'aide à la décision doit être conçu comme un système qui produit des résultats, sachant que le décideur tentera généralement d'obtenir d'autres résultats du système d'aide à la décision pour

affiner sa compréhension, son analyse.

En réalité, il existe nombre de SIAD différents au sein desquels on retrouve toujours au moins deux fonctions :

- **Un SIAD fournit des représentations des données stockées. Ces informations sont censés être pertinentes.**
- **Un SIAD traite l'information (exécute des algorithmes de traitement sur les données)**



La cycle de vie d'une information est le suivant : on la collecte, on la stocke, on la distribue et on l'exploite. Considérons ces 4 phases. Une notion centrale est celle de **Data Warehouse** (entrepôt de données) : il s'agit d'un ensemble de données. Souvent, les données sont stockées dans une base de données relationnelle, mais la notion de data Warehouse s'abstrait du type de stockage : on pourrait avoir d'autres solutions qu'une base de donnée relationnelle.

La collecte des données se fait en suivant les étapes de l'acronyme ETL : extract, transfert, load. Cette étape consiste à récupérer l'information depuis les bases de production (par exemple, depuis les commerciaux qui remontent les informations clients, depuis le site Web, en récupérant les informations sur les produits etc). Ces informations sont ensuite chargées dans le data warehouse : comme elles peuvent provenir de sources hétérogènes, utilisant des formats hétérogènes, il est souvent nécessaire de les transformer pour les mettre au format des données du data warehouse. Tout ce travail est bien sûr généralement automatisé.

Le stockage des données se fait donc dans le data warehouse. Il permet de rassembler l'ensemble des données de l'entreprise, ou l'ensemble des données d'un sous-système de l'entreprise (le domaine des ventes par exemple). Il est une **collection de données intégrées** (c'est à dire formant un tout homogène, c'est à dire des données qui

sont bien du domaine attendu et des données mises en forme pour être intégrée au data warehouse) **et non volatiles**. Pour cette dernière propriété, elle signifie que les données sont historicisées : on peut toujours retrouver telle information sur les ventes, même si il s'agit d'une information sur une période passée.

La distribution des données correspond au fait que les machines qui vont exécuter l'exploitation peuvent être distinctes et/ou distantes, de sorte qu'il peut y avoir une phase de distribution.

La phase d'exploitation est la phase d'analyse décisionnelle proprement dite. Cette exploitation peut être réalisée de différentes manières que l'on développe dans la suite. Sur le schéma, on présente différents types d'exploitation :

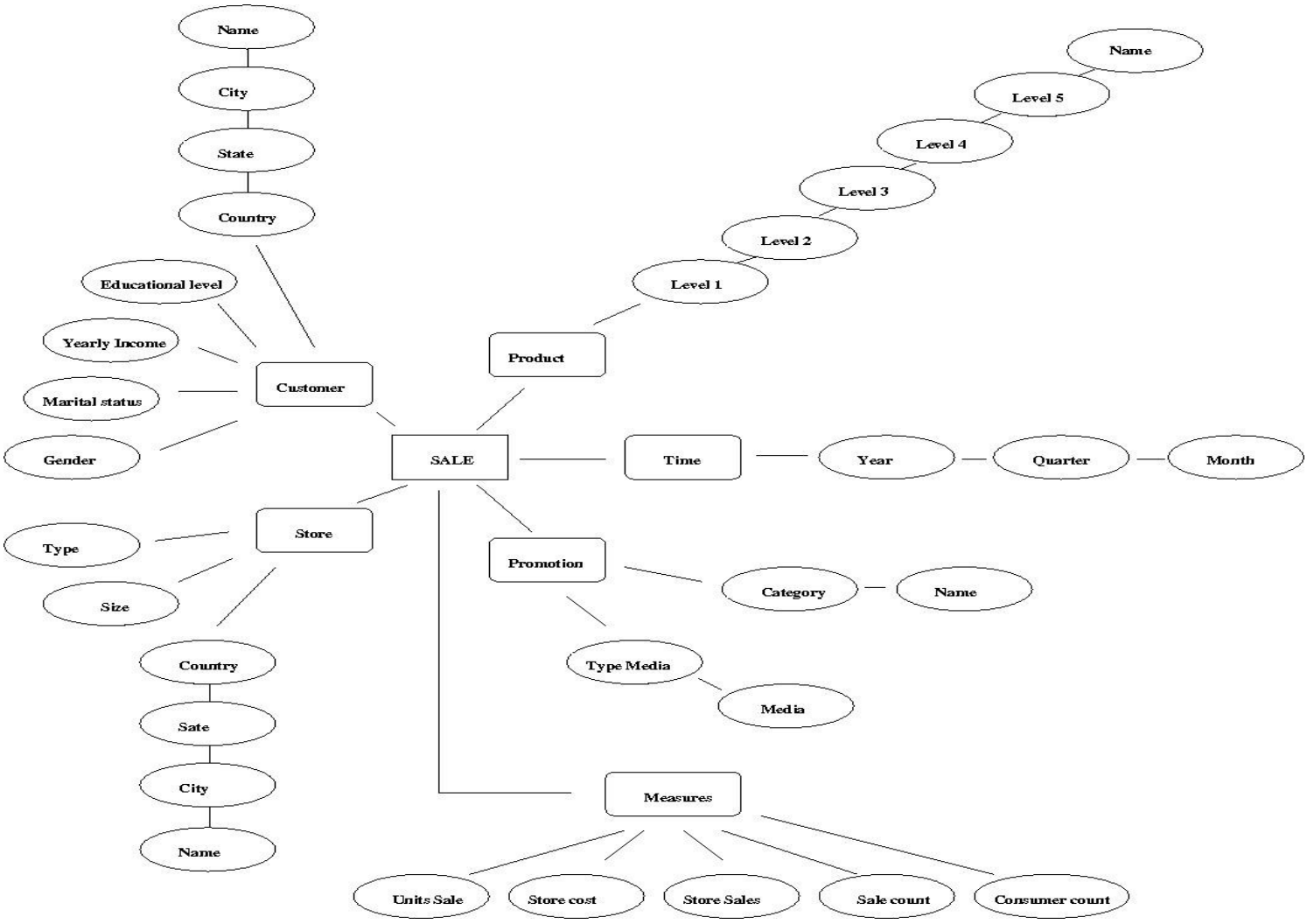
- le reporting, qui consiste à établir un compte rendu de l'activité.
- le datamining qui consiste à utiliser des algorithmes permettant de rechercher des régularités, des règles, en explorant automatiquement par des algorithmes puissants l'ensemble des données,
- l'OLAP est un outil permettant de visualer les données, en permettant d'obtenir une ou des vues synthétiques sur ces données pour les décideurs.
- Les tableaux de bord qui sont une autre méthode classique en gestion pour évaluer l'évolution de l'activité de manière synthétique.

Bien sûr, ce schéma général n'est pas toujours entièrement implémenté : dans beaucoup d'organisations, les bases de production des données sont déjà intégrées au data warehouse : il n'est pas nécessaire de passer par une phase d'extraction et de transfert. De même pour la phase de distribution, si les systèmes d'exploitation sont intégrés au data warehouse, il n'y a pas réellement de phase de distribution.

La mise en place de SIAD n'est pas propre à la démarche informatique. La notion de tableau de bord a par exemple été formalisée par Norton et Kaplan hors du champ de l'informatique. Un tableau de bord peut être considéré comme un SIAD. Il reste cependant un outil limité, notamment par rapport à un outil comme OLAP. Dans la suite, on insiste particulièrement sur deux formes d'exploitation des données : OLAP et le datamining.

OLAP

La démarche OLAP est très utilisée. OLAP représente l'information sous la forme d'un **schéma en étoile**. A partie du moment où le décideur veut étudier les ventes d'une organisation, la démarche OLAP va être soutenue par ce type de schéma :



Dans ce schéma, les ventes sont représentées avec différentes **dimensions** : le temps, le consommateur, le magasin, le produit etc... Par exemple, pour une vente, elle a été réalisée le 12/07/2005, le consommateur est M. Durant, le magasin est le 2 rue Montmartre, le produit est un lave-vaisselle : une vente est bien définie sur ces différentes dimensions. Pour l'unité considérée, ici la vente, on donne aussi un ensemble de **mesures** ("Measures" sur le schéma) : le nombre de ventes, le montant des ventes, le nombre des consommateurs qui sont autant de mesures possibles pour les ventes. On pourrait multiplier les exemples de mesures : le montant de la vente moyenne, la vente la plus importante, etc...Enfin, on définit des **opérateurs d'agrégation** : la moyenne, le min, le max, la variance etc... On peut ensuite faire des **requêtes** à partir des schéma en étoile. Une requête c'est le fait de fixer quelques valeurs sur certaines dimensions :

- Dimensions : pour le pays France, pour les ventes de l'année 2008, pour les produits de type livre

le fait de choisir une ou des mesures :

- Mesures : le nombre de client,

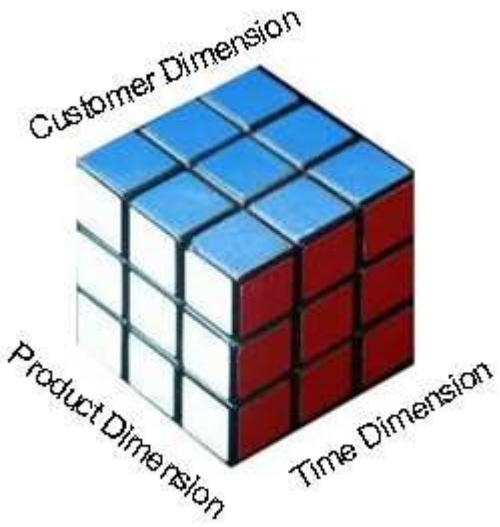
un opérateur d'agrégation :

- La somme.

On vient de construire une requête qui est la somme du nombre de clients pour la France pour l'année 2008.

D'autres requêtes sont possibles que l'on pourrait aussi décomposer en opérateurs d'agrégation, mesures et dimensions. Combien de vente pour le magasin du 7 avenue Charles de Gaulle ? Quel est le produit le plus vendu dans les magasins du Mans ? Quel age et quel genre ont les consommateurs qui achètent les lave-vaisselles ? Ces informations sont très importantes pour l'analyse.

A noter que la représentation en schéma étoile est équivalente à une représentation de cette forme :



On parle de cube OLAP : chaque petit cube est un ensemble de ventes pour lequel on peut appliquer un opérateur d'agrégation et choisir une mesure. Le cube est cependant une représentation plus limitée que celle du schéma étoile : on peut y représenter au plus 3 dimensions. Cependant, l'aisance visuelle qui lui est associée fait que la notion de cube OLAP reste très employée.

OLAP est un outil qui permet de visualiser l'information. Au niveau global d'une entreprise, on constate un résultat, un chiffre d'affaires qui sont des agrégats. Ces agrégats peuvent être décomposés : la comptabilité propose le calcul de SIG pour comprendre la formation du résultat depuis les comptes. OLAP est une démarche permettant de regarder l'information sous toutes ses formes et permet également de décomposer l'information : on peut commencer à regarder le résultat global puis voir le résultat par pays, puis voir le résultat global par magasin. Si dans un pays le résultat est plus mauvais que dans d'autres, on peut essayer de voir si cela vient de la fréquentation du site, du nombre de produits visités par visites, du nombre d'achat par visite de produits etc... Ces analyses permettent de fortement affiner l'analyse sur des résultats de vente.

L'information servant à la démarche OLAP est souvent stockée dans des bases de données relationnelles avec le but de minimiser la taille du stockage (où l'information est stockée selon un schéma en 3ème forme normale). La démarche OLAP peut utiliser les données stockées dans les bases de données relationnelles pour fabriquer les cubes

OLAP en fonction des besoins. Il est à noter que les algorithmes de réalisation de cette opération sont assez lourds : OLAP est un outil de visualisation et d'exploitation de l'information dont les besoins sont différents de ceux satisfaits par une base de données relationnelles.

A noter qu' OLAP existe sous différentes formes : ROLAP correspond à la pratique d'OLAP depuis une base de données relationnelles : dans ce cas, il est nécessaire de faire beaucoup de retraitements pour parvenir à obtenir le cube à partir de la base de données. MOLAP correspond à la pratique d'OLAP à partir d'une base de données multidimensionnelles : dans ce cas, la base de données est plus lourde à stocker, mais le temps de calcul du cube est plus rapide.

Datamining

Les techniques de datamining sont nombreuses et évoluent constamment du fait du travail en IA, en bio-informatique. Par exemple, pour étudier le génome, il faut des algorithmes de recherche de régularité puissants. De fait, on dispose d'outils pour l'exploitation des données et la recherche de règles de manière automatique. Alors qu'en OLAP, on cherche des règles à la machine de chercher automatiqueme sions, les réseaux de neurones, des classif

Des application dans l'étude des données automatiques is

On peut propos **du domaine des statistiques, des mathé ortant volume de données brutes, de co à découvrir de l'information cachée que les données renferr e relations ou de régularités"**.

Pour affiner la on selon l'un des algorithmes les plus utilis

n,2K€<<5K€,1K€<<2K€,<10K€,Fonctionnaire,Chomage,10%<<20%
o,5K€<<10K€,<1K€,>1M€,Agriculteur,auFoyer,>50%
n,1K€<<2K€,<1K€,10K€<<30K€,Cadre et ass.,Cadre et ass.,<5%
n,1K€<<2K€,2K€<<5K€,<10K€,Fonctionnaire,Prof intermediaire,<5%
n,1K€<<2K€,10K€<,>1M€,auFoyer,Agriculteur,10%<<20%
n,2K€<<5K€,1K€<<2K€,10K€<<30K€,Prof intermediaire,Chomage,5%<<10%
n,2K€<<5K€,10K€<,<10K€,Chomage,Chomage,10%<<20%
o,10K€<,2K€<<5K€,100K€<<500K€,Cadre et ass.,Cadre et ass.,30%<<50%
o,10K€<,5K€<<10K€,10K€<<30K€,Chomage,Fonctionnaire,>50%
n,1K€<<2K€,<1K€,30K€<<100K€,Cadre et ass.,auFoyer,>50%
n,<1K€,<1K€,>1M€,Agriculteur,Prof intermediaire,>50%
n,<1K€,<1K€,10K€<<30K€, Ouvrier et ass.,Cadre et ass.,5%<<10%
o,5K€<<10K€,10K€<,30K€<<100K€, Ouvrier et ass., Ouvrier et ass.,>50%
n,10K€<,<1K€,<10K€,Prof intermediaire,Agriculteur,<5%
n,1K€<<2K€,<1K€,10K€<<30K€,Cadre et ass.,Agriculteur,10%<<20%
n,1K€<<2K€,<1K€,<10K€,Cadre et ass.,Cadre et ass.,10%<<20%
n,5K€<<10K€,5K€<<10K€,>1M€,Agriculteur,Prof intermediaire,>50%
n,1K€<<2K€,2K€<<5K€,500K€<<1M€, Ouvrier et ass., Ouvrier et ass.,30%<<50%
n,<1K€,1K€<<2K€,30K€<<100K€, Ouvrier et ass.,Chomage,30%<<50%
n,<1K€,<1K€,100K€<<500K€,Cadre et ass., Ouvrier et ass.,>50%
n,5K€<<10K€,<1K€,30K€<<100K€,Agriculteur, Ouvrier et ass.,<5%
o,5K€<<10K€,10K€<,500K€<<1M€,Prof intermediaire, Ouvrier et ass.,10%<<20%
o,1K€<<2K€,1K€<<2K€,500K€<<1M€,Agriculteur, Ouvrier et ass.,20%<<30%
n,5K€<<10K€,<1K€,>1M€,auFoyer,Prof intermediaire,30%<<50%
o,5K€<<10K€,5K€<<10K€,30K€<<100K€,Prof intermediaire,auFoyer,>50%
n,5K€<<10K€,<1K€,10K€<<30K€,Fonctionnaire,Cadre et ass.,20%<<30%
n,1K€<<2K€,1K€<<2K€,10K€<<30K€,Cadre et ass.,Agriculteur,30%<<50%
n,10K€<,<1K€,100K€<<500K€,auFoyer,Prof intermediaire,20%<<30%
n,<1K€,10K€<,10K€<<30K€,Cadre et ass., Ouvrier et ass.,10%<<20%
o,<1K€,5K€<<10K€,>1M€,auFoyer,Agriculteur,>50%
n,2K€<<5K€,5K€<<10K€,500K€<<1M€,Agriculteur,Agriculteur,<5%
n,<1K€,2K€<<5K€,500K€<<1M€,Prof intermediaire,Prof intermediaire,20%<<30%
n,10K€<,10K€<,30K€<<100K€,Fonctionnaire,Fonctionnaire,<5%
n,5K€<<10K€,<1K€,500K€<<1M€,auFoyer,Agriculteur,<5%
n,1K€<<2K€,5K€<<10K€,<10K€, Ouvrier et ass., Ouvrier et ass.,<5%
n,5K€<<10K€,<1K€,>1M€,Prof intermediaire,Chomage,20%<<30%
n,2K€<<5K€,<1K€,100K€<<500K€,Prof intermediaire,Prof intermediaire,10%<<20%

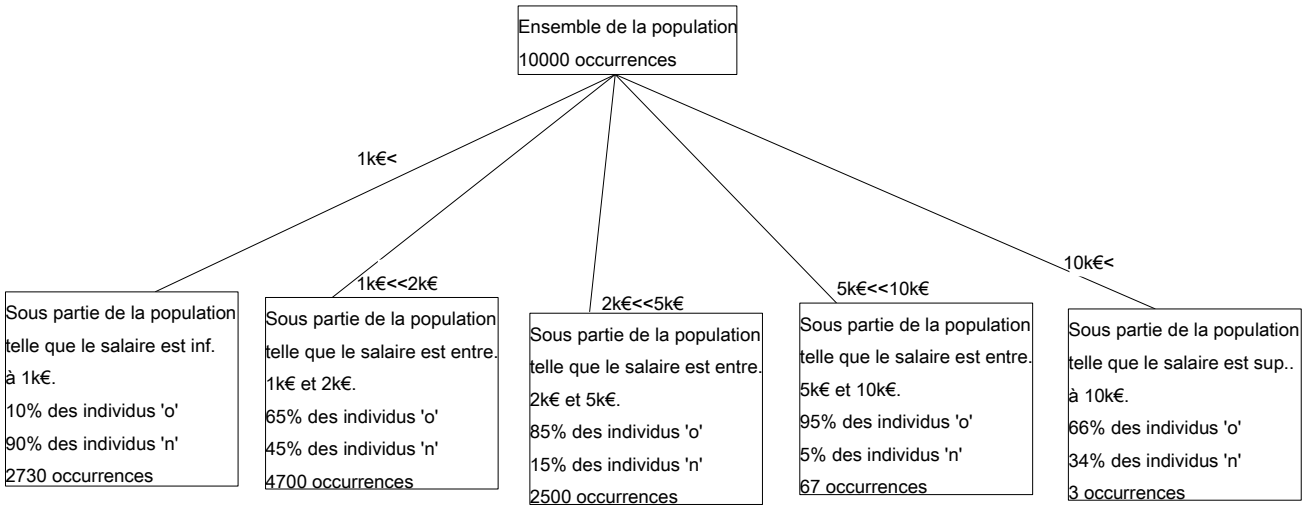
Chaque ligne correspond au fait qu'un couple rembourse ou non un prêt. Dans la première colonne : 'n' si la personne rembourse, ensuite, le niveau de salaire du premier membre du couple, ensuite le salaire du deuxième membre du couple, le patrimoine du couple, la profession du premier membre du couple, la profession du second et le niveau d'endettement du couple après l'intégration du prêt. On veut **apprendre** les critères les plus discriminant pour savoir si le prêt va être ou non remboursé. Une fois des **règles** apprises sur la base des prêts anciens, on rentra les données d'un nouveau couple (salaire, profession, endettement...) et la machine nous donnera une probabilité de remboursement en fonction de laquelle on pourra décider de prendre le prêt.

L'algorithme fonctionne de cette manière :

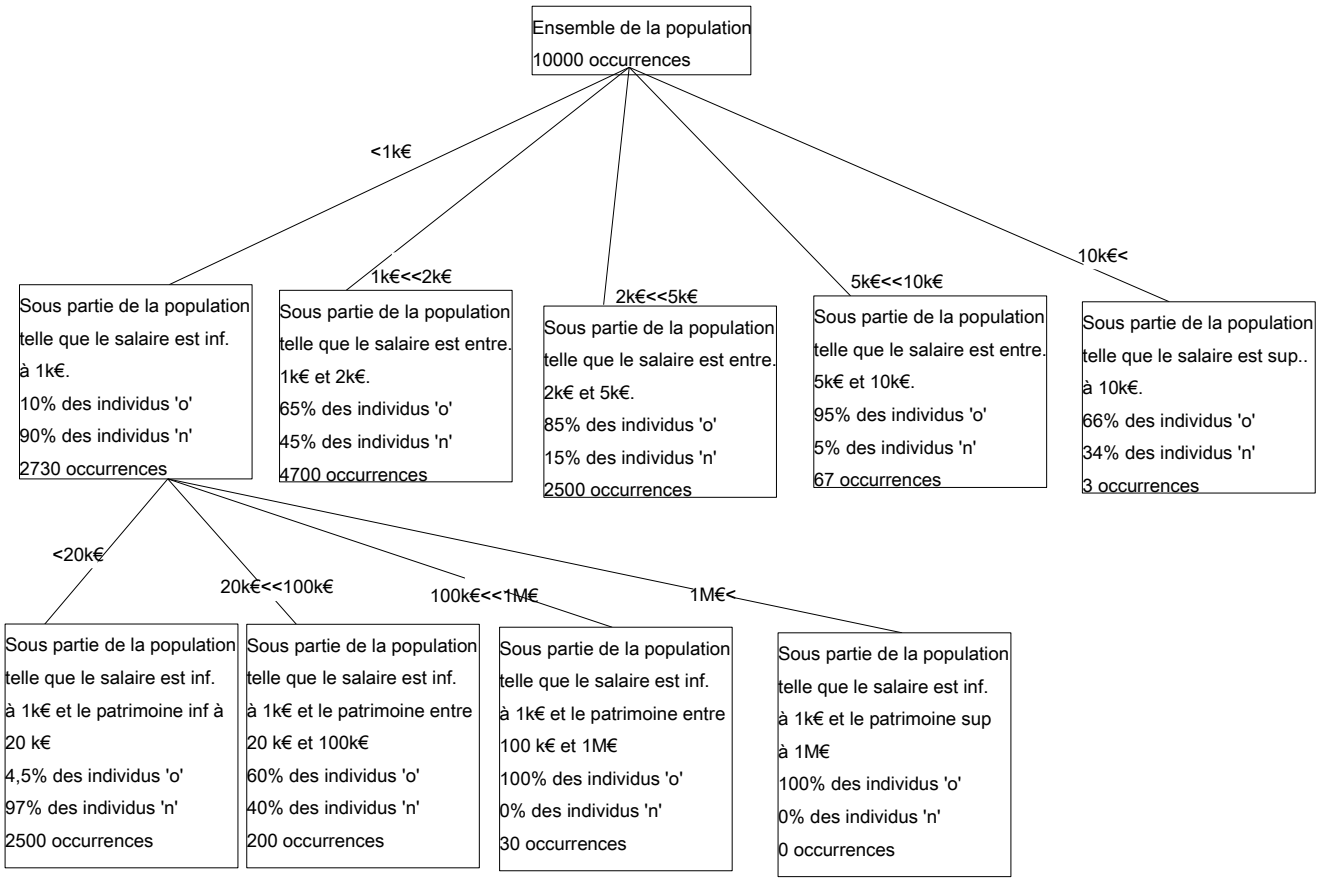
- Il cherche celui des critères qui classe le plus d'individus : ie celui qui permet le plus de discriminer entre les individus qui remboursent et ceux qui ne remboursent pas.

- On classe les individus selon ce critère.
- On recommence la procédure sur chaque branche de l'arbre.

Supposons que l'on lance l'algorithme. Lors de la première étape, on constate que c'est le salaire qui permet de discriminer le plus entre les individus qui remboursent et les autres. On crée Un premier nœud de l'arbre :



Ensuite, à chacun des nœuds, on applique le même algorithme en éliminant l'attribut du salaire. On suppose que sur le premier nœud, on trouve que c'est le patrimoine qui est le critère le plus discriminant à partir du premier noeud de second niveau :



On poursuit l'algorithme, au final, on pourra avoir jusqu'à 4 niveaux de profondeur (4 attributs) : généralement, on ne veut pas descendre jusqu'à ce niveau qui présente un important niveau de complexité et est assez illisible. Même en s'arrêtant avant, on pourra avoir classe l'ensemble une large partie de la population. En suivant les arcs de la population, on verra que pour des individus de salaire compris entre 1k€ et 2 k€, pour des patrimoines entre 20k€ et 100 k€, pour des fonctionnaires et pour un prêt ne dépassant pas 20% du budget, on a une 95% des gens qui remboursent : on pourra donc considérer que un nouvel individu qui arrive avec ces caractéristiques, il y aura une probabilité de 95% qu'il rembourse le prêt

Éléments de comparaison et autres outils

OLAP est un outil d'aide à la décision qui fournit une visualisation de la donnée : l'utilisateur cherche des règles en exploitant la donnée et en la regardant sous différents angles. Au contraire, le datamining, donne un classement automatique de la donnée, à l'instar des arbres de décision.

Le datamining existe sous d'autres formes que les arbres de décisions : les régressions linéaires, les réseaux de neurones etc...

Le datamining (qui n'a été que brièvement évoqué dans ce qui précède), n'est qu'une partie de ce qui peut être fait en analyse décisionnelle, ainsi que l'indique le schéma général plus haut : des analyses par tableaux de bord, du reporting, des analyses statistiques etc... seraient possibles.

Les outils qui viennent d'être présentés dans cette dernière partie peuvent être considérés comme des outils de BI ou Business Intelligence.

Quelques outils pour l'analyse décisionnelle

Quelques outils standards pour l'analyse décisionnelle peuvent être cités, on peut noter qu'ils peuvent également être regroupés sous le nom d'informatique décisionnelle ou de Business Intelligence (ce terme est cependant un peut plus général, il comprend également la notion de veille informationnelle) :

- SAS est une solution d'analyse décisionnelle qui permet à la fois de faire des analyses OLAP et de chercher des régularités dans les données au travers de méthodes comme les régressions linéaires, les arbres de décision et... L'outil est édité par la société SAS Institute.
- Microsoft SQL Server : à la base, le produit est un SGBD, mais cette fonctionnalité initiale s'est vue augmentée de différents module, dont un module Analysis Services, Integration Services et Reporting services
- Penthao, une solution du monde libre,
- Oracle, éditeur de SGBD, a augmenté son offre de modules permettant l'analyse décisionnelle.

- Excel permet une série de manipulations qui se rapprochent de l'analyse OLAP : tableaux croisés dynamiques, filtres, etc... Il est également possible de trouver des extensions d'Excel open source qui permettent de faire des arbres de décision.
- Etc...

Pour aller plus loin :

- http://www.cigref.fr/cigref_publications/2009/10/2009-business-intelligence-place-de-la-bi-et-pilotage-des-projets-d%C3%A9cisionnels-dans-les-grandes-entr.html
- http://www.decideo.fr/
- voir les sites sur le BI, l'analyse décisionnelle.