Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Conseils relatifs à la conception et à la réalisation de bases de données

Depuis la création de la SCM en 1995, nous avons rencontré très peu de bases de données qui soient correctement réalisées et dont l'exploitation ne requière pas un traitement préalable. Voici des exemples :

- Base de données "incidents trains" de la SNCF, un million de lignes par an : forte proportion de données inexploitables ou incohérentes ;

- Bases de données "post Tchernobyl", expertisées par la SCM pour le compte de l'IRSN : 20 années d'observations (taux de radioactivité, taux de cancers, etc.), mais données mal enregistrées, formats incohérents. Exploitation finale presque impossible.

- Base de données clients, pour une institution de prévoyance : très nombreuses erreurs de saisie, informations incomplètes (on ne sait pas les dates de début ou de fin des contrats). Exploitation finale presque impossible.

Bien que ces BD aient coûté fort cher à constituer, leur exploitation finale est décevante.

Seule exception notable : la base de données d'intervention de la Brigade de Sapeurs Pompiers de Paris, sur laquelle nous avons travaillé en 2010, était d'excellente qualité et contenait moins de 1% de données aberrantes.

Pour la constitution d'une base de données, les règles suivantes devraient être observées :

I. Réfléchir d'abord à ce que l'on souhaite enregistrer

Beaucoup d'organismes enregistrent n'importe quoi, avec n'importe quel pas de temps (toutes les minutes, tous les jours, etc.). Ceci n'a pas de sens : la nature de l'information collectée dépend de ce que l'on veut en faire.

Par exemple, pour Veolia Transport, pour la conception d'un réseau dans une ville donnée, nous avions fait la suggestion de deux BD, une grossière et une précise :

- La BD grossière contient les densités de population et les densités d'emplois sur chaque carré de 400 m de côté ;

- La BD précise contient les "points of interest" (lycées, hôpitaux, etc.) avec leurs coordonnées et leur fréquentation.

De manière générale, il faut commencer par réfléchir à l'exploitation que l'on veut faire de la base de données. Par exemple :

- Exploitation annuelle, dans un but de tableau de bord d'entreprise (très grossier) ;

- Exploitation instantanée, pour commander des pièces détachées (très fin).

Entre les deux, de nombreuses variantes sont possibles. Par exemple, pour Veolia Environnement Région Ouest, nous avons constitué un "panel" de consommateurs, pour anticiper la consommation d'eau chaque trimestre : il suffit de connaître les relevés tous les trois mois.

Les bases de données "environnement" (taux de pollution, etc.) se contentent en général d'une moyenne annuelle, si l'exploitation est politique.

Il ne faudrait pas croire que la BD fine soit toujours préférable à la BD grossière ; c'est tout l'inverse. Si la BD est trop fine, elle comporte une énorme quantité de références diverses, dont on ne sait pas quoi faire. Nous avons eu un contrat avec une fédération d'hôpitaux : il s'agissait d'étudier l'impact de changements tarifaires. Mais si on descend au niveau de la nomenclature des cotons-tiges, on ne peut savoir quelle information est pertinente.

II. Mettre en place des systèmes d'aides et d'alerte lors de la constitution de la BD

Notre expérience est catégorique : une fois qu'une fiche est mal remplie, c'est irrattrapable, car elle est noyée au milieu de toutes les autres. Il faut donc aider la personne qui réalise les fiches :

1. Par des guides

- Le format de date doit être imposé ;

- La date de fin d'une opération doit être postérieure à celle de début ;

- Les grandeurs sont positives, etc.

Tout ceci peut être vérifié de manière informatique au cours de la saisie, et un menu d'aide apparaît : "fiche mal remplie, vérifier xxx".

2. Par des vérifications

Des vérifications simples peuvent permettre de détecter les anomalies : cohérence entre l'amont et l'aval, entre deux capteurs voisins, entre une journée et la veille, etc.

Notre expérience est ici que ces mesures simples permettraient d'éviter 90 % des erreurs commises.

Nous avons développé des méthodes probabilistes robustes pour la détection de données aberrantes et la reconstruction des données manquantes, qui sont développées dans deux livres :

[RDM] Bernard Beauzamy et Olga Zeydina : Méthodes probabilistes pour la reconstruction de données manquantes. ISBN 2-9521458-2-2, ISSN 1767-1175. SCM SA, avril 2007.

[PIT] Olga Zeydina et Bernard Beauzamy : Probabilistic Information Transfer (en anglais), ISBN 978-2-9521458-6-2, ISSN 1767-1175, SCM SA, avril 2013.

La SCM a assuré une formation sur ces questions en octobre 2013.

En conclusion, nous dirons que pour disposer d'une base de données exploitable, il faut :

- Un minimum de réflexion préalable : que veut-on faire des données ?

- Des vérifications permanentes : lors de la saisie, au moyen de comparaisons, etc.

Les méthodes utilisées pour détecter les données aberrantes et reconstituer les données manquantes sont décrites dans notre fiche "Qualité de l'Information" :
http://scmsa.eu/fiches/SCM_Qualite_Information.pdf