Précis d'Informatique Décisionnelle

Data Warehouse, Data Mart, Data Vault, ODS, « étoile », « flocons », Business Objects, ETL Informatica ...

© Huy CHAU, 2021

Table des matières

Avant-propos

Le « Décisionnel » est l'un des domaines où de nouveaux termes sont régulièrement inventés. Par ailleurs, il n'y a pas de « normes ISO » dans ce domaine pour les préciser. Par conséquent, le jargon sans cesse enrichi du « Décisionnel » accroît la confusion même dans le milieu des professionnels.

Ce « Précis d'Informatique Décisionnelle » tente de présenter au lecteur, toujours avec le **souci de démystification**, des notions qui devraient lui servir de repères, par rapport auxquels il devrait pouvoir appréhender plus facilement et sereinement le jargon dans le sens utilisé par l'entreprise pour laquelle il travaille.

Les exemples qui y sont exposés sont très simples pour éviter que leur complexité ne disperse le lecteur des raisonnements et messages sous-jacents. Les exemples plus complexes et plus proches des situations professionnelles sont abondants dans *[A&V 98], [KBR 01], [SVL 01]*.

L'ouvrage est composé de thèmes (comme des fiches) assez courts et progressifs, avec parfois une bibliographie propre. La bibliographie à la fin du livre propose un éventail plus large de références aux lecteurs qui disposent de plus de temps. Bien entendu, la liste des ouvrages proposés dans ces bibliographies résulte d'un choix subjectif et n'est donc pas exhaustive.

Nous présenterons aussi au lecteur un outil de chargement et un outil d'exploitation du Système d'Information Décisionnel (SID) :

- l'ETL Informatica Power Center (IPC) pour le chargement
- Business Objects (BO) Designer pour l'exploitation

Cette présentation n'est pas superficielle : elle devrait fournir suffisamment d'éléments au lecteur pour pouvoir réaliser plusieurs applications (Mapping IPC, Workflow IPC, Univers BO) même dans un contexte professionnel.
Les thèmes relatifs à la présentation de ces outils sont indépendants du reste, le lecteur pourra donc les ignorer s'il n'est intéressé que par les aspects conceptuels et théoriques du « Décisionnel ».

Huy CHAU

chau.consultant@yahoo.com
BI.ETL.consultant@gmail.com

Références citées :

[A&V 98] : "Data warehouse design solutions" par Christopher Adamson & Michael Venerable (Wiley - ISBN 0-471-25195-X, 1998).

[KBR 01] : « Entrepôts de données » par Ralph Kimball (Vuibert Informatique – ISBN 2-7117-8668-7, 2001).

[SVL 01] : « The data model resource book" par Len Silverston (Wiley - ISBN 0-471-38023-7, 2001).

© Huy CHAU, 2021

Thème 1 - Données - Information – Information de synthèse

Les discussions sur « donnée » et « information » pourraient s'avérer sans fin (Barry Devlin a beaucoup publié sur ce sujet). Nous proposons ci-après notre point de vue à travers quelques exemples d'illustration.

« 11 Février 1978 », « 5 avenue du Général de Gaulle à Toulouse » .. etc … sont des « données ». Cette date du « 11 Février 1978 » est-elle celle d'un événement historique, la date de naissance ou de mariage d'un individu .. etc : elle ne nous fournit aucune « information ».

En revanche,
«Le 11 Février 1978, Dominique Durand a emménagé au 5 avenue du Général de Gaulle à Toulouse» est une information.
«Le 11 Février 1978, Camille Martin organise son anniversaire à l'adresse 5 avenue du Général de Gaulle à Toulouse » est une information.

Ainsi, les informations sont construites à partir de données.

Il est à remarquer que lorsqu'un magazine titre sa couverture par « vos données valent de l'or », il s'agit bien entendu de l'information : en effet, la subtilité se trouve dans l'utilisation de l'adjectif possessif « vos » qui permet de relier les données à un individu pour en générer des informations sur lui; ainsi, « vos données valent de l'or » est une formulation plus percutante pour annoncer « les informations sur vous valent de l'or ».

Par cet exemple, nous avons introduit un facteur, important à nos yeux, qui est la **relation** entre les données : des données disparates sans aucune relation entre elles ne permettent de générer aucune information.

En général, une donnée a une nature, un type : par exemple, « 11 Février 1978 » est une date, « 5 avenue du Général de Gaulle à Toulouse » est une adresse.

Présentées ainsi, « donnée » et « information » ne devraient prêter à aucune confusion.
Cependant il est d'usage que les informations exploitées pour aboutir à une conclusion, à un résultat (ou plus généralement exploitées par des traitements qu'ils soient informatiques ou pas) sont appelées «données» *:* ainsi, on peut dire que les informations qui sont données (fournies) aux traitements sont appelées « données ».
Il est à remarquer que Barry Devlin a même introduit le terme … « données informationnelles » (« informational data »*([DLB 97],p.45*) qui, à notre avis, ne fait qu'accroître la confusion !

Information de synthèse, « Business Intelligence »
Sans être dans les services de renseignements de l'Etat ou de l'Armée, nous pouvons néanmoins imaginer que ces services collectent des informations de différentes sources mais aussi les recoupent, les « fusionnent » en informations de synthèse.

Nous pouvons déjà remarquer qu'à partir des mêmes informations collectées, deux services pourraient produire des « informations de synthèse » de pertinence inégale : cette différence résulte

de « l'intelligence » sous-jacente à l'élaboration de ces informations de synthèse, et ce n'est donc probablement pas par hasard que de tels services possèdent des noms comme « (Secret) Intelligence Service » ou « (Central) Intelligence Agency » !

De la même manière, dans les activités commerciales, une entreprise peut aussi « travailler » sur des informations collectées au cours de ses activités courantes (*les ventes par exemple*) pour en produire des informations de synthèse plus ou moins pertinentes selon « l'intelligence » intervenant dans leur élaboration: la « Business intelligence » (BI) !
Concernant ces informations de synthèse de la BI, chaque entreprise est libre de les « fabriquer » selon sa « créativité » mais il est à noter qu'il existe aussi des informations de synthèse dont les règles de production sont très réglementées comme un bilan comptable par exemple.

A propos du « Big data »

Etait-ce une coincidence : depuis les révélations d'Edward Snowden sur la collecte d'informations de la National Security Agency (NSA), la ruée vers le « Big Data » s'est affirmée.

Cependant, les entreprises donnent-elles le même contenu à leur projet « Big Data » ? en effet, à ce jour, il n'y a pas de consensus sur le « Big Data » (voir *[B&C&D 16]*) à part le fait qu'un volume « important » de données et de sources de collecte variées (dont Web, sms .. etc ..) seront à prendre en compte, leur exploitation pourra s'avérer complexe (et coûteuse) et l'entreprise pourra se trouver dans une situation semblable à celle décrite par Alain Bauer en 2016 (*[BAA 16]*) à propos des services de renseignements français : *« les services de renseignement sont extrêmement efficaces dans la collecte d'information, tellement efficaces qu'on ne sait plus quoi faire de ces informations …. »*

En bref

La donnée est un composant élémentaire des informations. Une donnée a un type (de donnée) associé.
Sans aucune relation entre elles, ces données ne peuvent produire aucune information.

Des informations peuvent être exploitées pour en produire de nouvelles qui sont des informations de synthèse. Il est à remarquer que ces informations de synthèse peuvent être exploitées à leur tour pour produire d'autres informations de synthèse.

Cependant, il est très fréquent que les informations utilisées en entrée d'un traitement ou d'une analyse sont appelées par « **données** » : par la suite, pour nous rapprocher de l'usage, nous pourrons aussi adopter cet abus de langage.

Références citées

[B&C&D 16] : « Big Data : Principles and Paradigms » par Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi (ebook – Edition Elsevier – ISBN 978-0-12-805394-2, 2016) - chapitre « Historical Interpretation of Big Data », paragraphe « 1.3.3 Summary of 7 Types Definitions of Big Data»

[BAA 16] : conférence sur le terrorisme à la Halle de la Place d'Armes à Calais publié le 25/04/2016 par « La voix du Nord»
http://www.lavoixdunord.fr/region/calais-l-expert-alain-bauer-vient-parler-de-terrorisme-ia33b48581n3468858

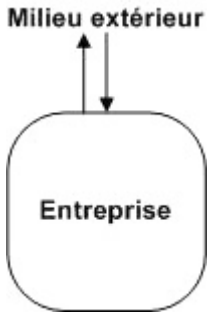
[DLB 97] : « Data Warehouse : from architecture to implementation» par Barry Devlin (Addison Wesley - ISBN : 0-201-96425-2, 1997)

© Huy CHAU, 2021

Thème 2 - Donnée opérationnelle - Système d'information

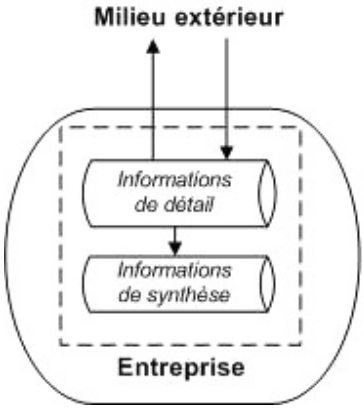
Pour faciliter l'illustration, nous définissons deux acteurs : l'entreprise et le « milieu extérieur » à l'entreprise.

L'entreprise vit et croît en développant ses échanges avec le milieu extérieur (par exemple Clients, Fournisseurs … etc …).



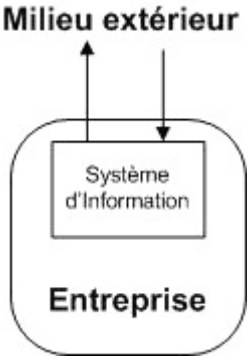
Ces flux d'échange concrétisent les activités *opérationnelles* de l'entreprise.

Ces flux d'échange peuvent être de nature matérielle (par exemple retrait d'argent à la banque, paiement des marchandises … etc ...) ou immatérielle (par exemple prises de contact, présentation de produits .. etc ...)



Les informations relatives à ces flux d'échange sont généralement conservées par l'entreprise (ne serait-ce que par obligation légale pour certaines comme les écritures comptables) : ce sont des informations opérationnelles *(communément appelées aussi par « données opérationnelles »)* de l'entreprise.

L'entreprise peut utiliser ces informations opérationnelles pour produire les « informations de synthèse » dont elle a besoin , ainsi, par opposition à ces dernières, nous pouvons aussi appeler par « informations de détail » ces informations opérationnelles.



En général l'entreprise stocke et range ses différentes informations sur divers supports : elles constituent son " Système d'Information " (SI).

Il est à remarquer que jusqu'à maintenant, nous avons fait abstraction de l'informatique : ces concepts sont indépendants de la technologie, on pourrait imaginer que ce SI est composé d'étagères et d'armoires contenant des classeurs rangés selon une certaine classification de manière à ce que les documents, les informations soient facilement retrouvés.

© Huy CHAU, 2021

Le lecteur familier avec la Comptabilité Générale pourrait constater que

- *le « Journal» contient les « informations de détail »,*
- *le « Grand Livre » contient ces « informations de détail » mais aussi les premières « informations de synthèse » construites à partir de ces « informations de détail »*
- *la « Balance » et le « Bilan » contiennent des « informations de synthèse » construites à partir de celles du « Grand Livre » (même si, dans la pratique, il est plus commode d'établir le « Bilan » à partir de la « Balance », pas directement du « Grand Livre »).*

En bref

Les opérations que l'entreprise effectue avec son « milieu extérieur » (fournisseurs, clients … etc …) permettent de générer des « informations de détail » appelées aussi par « informations détaillées » *[TBY 86] (p.42)* ou par «données opérationnelles ».

Ces « données opérationnelles » pourront être exploitées pour générer des « informations de synthèse ».

Les « informations de détail» et les « informations de synthèse » font partie du « Système d'Information » (SI) de l'entreprise.

C'est une présentation simple et pratique d'un SI que nous proposons ainsi au lecteur, une vision plus élaborée est proposée dans, par exemple, *[T&R&C 94] et [TBY 86]*.

Références citées

[T&R&C 94] : « La méthode Merise – Tome 1 – Principes et outils » par Hubert Tardieu, Arnold Rochfeld, René Coletti (Les éditions d'organisation – ISBN 2-7081-1699-1, 1994).

[TBY 86] : « De l'autre côté de Merise – Systèmes d'information et modèles d'entreprise » par Yves Tabourier (Les éditions d'organisation – ISBN 2-7081-0762-3 1986).

© Huy CHAU, 2021

Thème 3 - Notions diverses

Domaine « opérationnel » - Domaine « décisionnel »

Le terme « Décisionnel » ne nous paraît pas des plus heureux mais c'est un terme qui a réussi à s'imposer.

Dans une entreprise, le personnel qui travaille dans le domaine « Décisionnel » exploite (par analyses) les informations que possède déjà l'entreprise tandis que celui qui travaille dans le domaine « Opérationnel » effectue des actions, des « opérations » indispensables au fonctionnement de l'entreprise au quotidien : ainsi, c'est le domaine « Opérationnel » qui permet à l'entreprise de vivre alors que le « Décisionnel » devrait favoriser la pérennité et l'expansion de l'entreprise.
Le lecteur familier avec la « Comptabilité » pourrait y voir un parallèle avec le couple (« Comptabilité générale », « Comptabilité analytique »).

Les critères de distinction entre « domaine opérationnel » et « domaine décisionnel » ne font pas forcément toujours l'unanimité.

Il y a des cas simples : par exemple, lorsqu'un employé, face à un client, effectuait une opération pour le servir (caisse dans un magasin, guichet à la banque ou à la poste), tous s'accorderaient qu'il s'agissait du domaine « opérationnel ».

En revanche, lorsqu'un directeur d'agence d'une banque étudiait le portefeuille d'un client pour décider de l'accord d'un crédit ou lorsqu'une campagne publicitaire, découlant d'une application CRM/GRC (*Customer Relationship Management/Gestion de la Relation Client*) générait du mailing ciblé … etc …, pour les uns il s'agissait déjà du « domaine décisionnel », pour les autres il s'agissait encore du « domaine opérationnel ».

Pour notre part, nous considérons que

- les informations de synthèse peuvent être exploitées tout autant par le « domaine décisionnel » que par le « domaine opérationnel »
- les actions qui « ciblent» un individu, une occurrence particulière, sont du « domaine opérationnel »
- les actions qui analysent un ensemble d'occurrences sont du « domaine décisionnel »

Prenons l'exemple d'un réseau d'autobus dans une ville.
Chaque voyageur doit valider son titre de transport magnétique lorsqu'il rentre dans l'autobus.
La société de transports peut donc collecter ces validations et produire différentes informations de synthèse.
Elle peut donc savoir sur quels tronçons de trajet, vers quelles heures, pour quelles lignes de bus il y a de fortes affluences; le service d'Exploitation pourrait alors étudier et proposer des mesures pour mieux réguler le trafic.
Par ailleurs, à partir de ces mêmes validations, la société de transports peut aussi savoir sur quelles lignes de bus, vers quelles périodes de l'année, quels types de titres de transports (abonnement « étudiant », carte « visite touristique », … etc …) sont les plus utilisés et cette fois ce serait le service Commercial qui pourrait exploiter ces informations de synthèse pour réfléchir à des nouvelles offres commerciales à proposer aux voyageurs occasionnels ou habitués.

Pour nous, ces deux cas d'utilisation sont tous du domaine Décisionnel (*même si, notamment dans le premier cas, c'est le service Exploitation (un service typiquement opérationnel) qui est utilisateur*) : en effet, ces informations de synthèse utilisées ont bien été générées à partir d'un ensemble de validations effectuées par un ensemble de voyageurs, pas en suivant l'utilisation de son titre de transport par un voyageur particulier *(comme avec la vision lagrangienne dans la « Mécanique des milieux continus »).*

Prenons un autre exemple d'illustration :
« Identifier les 3 produits les plus achetés par le client Martin » est du domaine opérationnel
alors que
« Identifier les 3 produits les plus achetés par les (clients qui sont) agriculteurs » est du domaine décisionnel.

Ces exemples nous permettent de constater que le terme « Décisionnel » peut paraître contestable lorsque son champ d'application n'est pas réservé au pilotage stratégique des décideurs.

« Drill down » vs. « Drill up »

Il est fréquent que, partant de l'analyse d'un ensemble d'occurrences, l'analyste, selon ses intérêts de recherche, pourrait éprouver le besoin de poursuivre son étude vers des niveaux plus détaillés jusqu'à l'occurrence (« l'individu ») elle-même : par exemple, l'analyste, intrigué par les résultats exceptionnels des ventes du dernier mois (par rapport aux mois précédents), pourrait chercher à savoir à quels jours précis («occurrences») du mois et/ou pour quels produits précis («occurrences») les résultats de ventes ont été exceptionnels.

Cet exemple permet d'illustrer une analyse courante du monde « décisionnel » : l'analyste ne partait ni d'un jour précis, ni d'un produit précis pour chercher le montant des ventes pour la simple raison qu'il ne savait pas quels seraient le produit ou le jour pertinents à étudier; il est parti d'un phénomène considéré comme « anormal » (un montant de ventes exceptionnel par rapport aux mois précédents) pour découvrir les origines (le jour et/ou produit qui sont concernés par ces résultats de ventes exceptionnels).
Cette démarche « d'approfondissement » partant de « l'ensemble » vers « l'individu » est appelée « Zoom avant » ou « Forage vers le bas » ou encore plus connue sous le nom de « drill down ».
La démarche inverse est « Zoom arrière » ou « Forage vers le haut » ou « drill up ».

Il est à remarquer qu'ainsi, ces démarches permettent de « naviguer » des informations de synthèse vers les informations de détail et *vice-versa*.

« OLAP»

Les outils « OLAP » (« On Line Analytical Processing») sont des applications informatiques facilitant cette navigation par interface Homme-Machine entre informations de niveaux différents.

Stratégie -Tactique

Certains auteurs proposent l'introduction de la notion de « stratégie » et par opposition, celle de « tactique » dans la classification des actions : les informations de synthèse peuvent être utilisées dans des analyses et actions à objectifs stratégiques mais aussi tactiques.
Ainsi, le directeur d'une agence de banque qui étudie l'octroi d'un crédit à un client aurait à prendre une décision, en général, tactique (*sauf des cas exceptionnels où, par exemple, le client (personne physique ou morale) est très influent*).

Il est à remarquer que cette distinction (stratégie vs. tactique) ne concerne pas les informations de détail.

En bref

Nous considérons que

- les informations de synthèse peuvent être exploitées par le « domaine décisionnel » mais aussi par le « domaine opérationnel »
- les actions qui « ciblent» une occurrence particulière sont du « domaine opérationnel »
- les actions qui analysent un ensemble d'occurrences sont du « domaine décisionnel »

Les démarches de « drill down » et « drill up » permettent de « naviguer » des informations de synthèse vers les informations de détail et *vice-versa.*

Les outils OLAP offrent un moyen technique facilitant cette « navigation » par interface Homme-Machine.

« Bibliographie »

[DLB 97] : « Data Warehouse : from architecture to implementation» par Barry Devlin (Addison Wesley - ISBN : 0-201-96425-2, 1997)

Nigel Pendse, « The origins of today's OLAP products »
http://dssresources.com/papers/features/pendse10062002.html

Thème 4 - SI, SIO, SID

Nous avons vu que les informations d'une entreprise sont « rangées » dans son Système d'Information (SI).
Pendant longtemps, au sein de ce SI, ont été développées de manière peu différenciée à la fois des applications du domaine opérationnel et celles du domaine décisionnel.

Depuis, probablement favorisées aussi par la baisse des coûts des matériels, les constructions des Data Warehouses se sont développées offrant un espace dédié au domaine Décisionnel séparé explicitement de celui dédié au domaine Opérationnel et permettant ainsi le clivage physique du SI en Système d'Information Décisionel (SID) et Système d'Information Opérationnel (SIO).
Physiquement, il est fréquent que le SID et le SIO sont implantés sur des machines différentes (*par exemple IBM z/OS pour le SIO et IBM AIX ou notamment Teradata, Netezza pour le SID*).

Le SIO est alimenté par des applications opérationnelles qui collectent les données en provenance du « milieu extérieur » : par exemple, lorsqu'un client achète un produit, son paiement sera enregistré dans le SIO.
(Il est à remarquer que le « SI Opérationnel » défini dans [TBY 86] (p.39) n'était pas exactement celui qu'on vient d'introduire et qu'on utilisera tout au long de cet ouvrage).

Le SID, en général, n'a pas de contact avec le « milieu extérieur » : il est alimenté par des « données » venant du SIO.

Remarque :
Il peut exister des cas marginaux où parmi les flux qui alimentent le SID, certains ne viennent pas du SIO de l'entreprise mais de « l'extérieur » (société partenaire, autre filiale du groupe, maison mère ... etc ...)

Cette alimentation a pour cible le « Data Warehouse » (entrepôt de données), pilier du SID.

En bref

Le SID est la partie du SI dédiée au domaine Décisionnel.
Le SIO est la partie du SI dédiée au domaine Opérationnel.

Le SID et le SIO sont fréquemment implantés sur des machines différentes.

Au cours des échanges avec le « monde extérieur », les applications opérationnelles enregistrent les données dans le SIO.
Et, en général, c'est le SIO qui alimente le SID, notamment son « entrepôt des données » (Data Warehouse), pilier du SID.

Référence citée

[TBY 86] : « De l'autre côté de Merise – Systèmes d'information et modèles d'entreprise » par Yves Tabourier (Les éditions d'organisation – ISBN 2-7081-0762-3, 1986).

© Huy CHAU, 2021

Thème 5 - Le Data warehouse, pilier du SID

A ce jour, il n'existe pas de norme ISO dans la définition d'un Data Warehouse (DW) ou « Entrepôt de données » (ED) en traduction française.

Nous nous référons à Bill Inmon (« *père du DW* ») pour énumérer les caractéristiques suivantes d'un DW :

- les données du DW sont historisées et non volatiles
- les données du DW sont « intégrées » et « orientées sujets »
- utilisées dans les prises de décisions stratégiques

Les données du DW sont historisées et non volatiles : les données dans le DW sont conservées sans altérations (pas de modifications des données déjà entreposées dans le DW); on peut imaginer que les données ont été prises en photos régulièrement et que ces photos sont conservées dans le DW, il en résulte qu'on peut toujours retrouver une photo du passé et en plus cette photo reflète exactement ce moment du passé sans aucune altération dû au présent, l'instant présent étant pris sur une autre photo.
(*Imaginons un album de famille qui contient les photos de chaque membre en bébé, adolescent, adulte .. etc ..)*
Cependant, malgré cet exemple, il est à éviter de penser que le DW est juste composé de successions de « photos » prises du SIO ! En effet, avant d'être rangées dans le DW, ces données venant du SIO devaient être préalablement « intégrées ».

Les données du DW sont « intégrées » : le DW « intègre » les données venant du SIO en les « homogénéisant », pour qu'elles soient compatibles et cohérentes.
Au cours des opérations d'homogénéisation, les problématiques rencontrées sont diverses.
Une problématique des plus simples est par exemple celle d'un indicateur booléen pouvant être codé par un caractère O/N dans une application opérationnelle mais par un nombre 0/1 dans une autre, le DW devra faire un choix lors de son intégration.
Une problématique plus liée à la technique est par exemple le fait que la date 0001-01-01 est valide avec le Système de Gestion de Bases de Données Relationnelles DB2 mais pas avec SQLServer (*versions d'avant 2008*).
Une problématique plus délicate et plus conceptuelle très fréquemment rencontrée est par exemple le fait qu'un « client » (une même personne qui ouvre plusieurs comptes bancaires, notamment dans des agences différentes, est-elle un seul ou plusieurs clients ? un prospect est-il un client ?), qu'un « produit » (un package de produits est-il un nouveau produit ?) peuvent être perçus différemment par les différents services (s*ervice commercial, service comptable ... etc* ...) et applications opérationnelles de l'entreprise : le DW devra résoudre ces problématiques lors de l'intégration de ces données.

Les données du DW sont « orientées sujets » : c'est, à notre avis, la caractéristique la plus délicate à expliciter.
Pour commencer, l'expression originale en anglais « subject-oriented » a pu être traduite par « orienté sujet » : certains en concluaient qu'un DW ne traitait qu'un seul sujet; nous optons donc pour « orienté sujets » afin de mieux faire ressortir le fait qu'un DW peut prendre en compte plusieurs sujets (*parmi ces sujets, si on en extrait un, on pourrait en générer un « Data Mart » qu'on verra prochainement (« Thème 6 – Data mart »*)).