

Avant-propos

1. But de l'ouvrage	11
2. De multiples concepts et quelques outils	12
2.1 Une modélisation indispensable et des outils utiles	12
2.2 Un concept, de multiples modèles	13
3. Rapports versus décisionnel	14
4. Le quiproquo du Big Data	15
5. Utilisateurs visés	16
6. À propos du contenu	16
6.1 Technologies utilisées	16
6.2 Code couleur pour les schémas	17

Aborder le décisionnel

1. Quid des termes décisionnels	19
1.1 De nombreux termes pour de nombreux concepts	19
1.2 Détails par catégorie	

Modélisation décisionnelle

Concevoir la base de données pour les traitements OLAP

1.2.1 Le projet : décisionnel, aide à la décision ou la Business Intelligence	19
1.2.2 Le système d'information : SID ou système d'information décisionnel	19
1.2.3 Les données : base de données décisionnelle relationnelle, entrepôt de données et magasin de données	20
1.2.4 La modélisation spécifique : base de données multidimensionnelle, OLAP, cubes	20
1.2.5 Le chargement des données : ETL, intégration	20
1.2.6 La restitution de la donnée : reporting, restitution, rapports, états, tableaux de bord, analyse, reporting ad hoc	21
1.3 Synthèse	21
2. La justification du décisionnel	22
2.1 Des données et un besoin déjà présents	23
2.1.1 Un besoin analytique	23
2.1.2 Bien qualifier ce besoin	23
2.2 La réponse à un besoin métier	24
2.3 Une nécessité technique	24
2.4 Une nécessaire historisation	26
2.5 Nécessaire mais dispensable	29
3. Mise en œuvre du projet	31
3.1 Les étapes	31
3.2 Découper le projet	31
	32

4. Identification du besoin	33
4.1 Identifier les utilisateurs cibles	33
4.2 Reprise de l'existant	34
4.2.1 Reprise en l'état	34
4.2.2 Identification des éléments clés	34
4.3 Demandes récurrentes	35
4.4 Identification des nouveaux besoins	35
4.5 Formaliser le besoin	36
4.5.1 Matrice des besoins	36
4.5.2 Détails complémentaires	36
4.6 La réalité	38

Repenser la donnée

1. De la technique au métier	39
2. Des "applications" aux "objets"	39
3. De l'événement à la période	41
4. Du transactionnel à l'analyse	42
5. De l'exhaustivité à la simplicité	

Comprendre les spécificités du décisionnel

1. Le multidimensionnel : hiérarchiser la donnée

1.1 Les dimensions	45
1.2 Les cubes	48
1.3 Densité d'un cube	52
1.4 Les technologies de cubes	53
1.5 L'interaction avec le datawarehouse	54
1.6 Un langage spécifique : le MDX	54
1.7 Les modes de stockage	57
1.7.1 Différents modes de stockage	57
1.7.2 MOLAP	58
1.7.3 ROLAP	59
1.7.4 HOLAP	60
1.7.5 Synthèse et choix	63
1.7.6 Autres technologies	64

2. Les fondamentaux de la base de données relationnelle

2.1 Le rôle de la base de données relationnelle	65
2.2 La dénormalisation	

2.2.1 Principe	66
2.2.2 Dénormaliser les référentiels	66
2.2.3 Dénormaliser les transactions	67
2.3 Pourquoi ne pas tout dénormaliser ?	70
2.4 Les modèles de données	73
2.4.1 Dimensions versus référentiels	75
2.4.2 Modèle en étoile	75
2.4.3 Modèle en flocon	77
2.4.4 Choix du modèle	80
2.4.5 Modèle en galaxie ou constellation	81
	86

Identifier le projet

1. Entrepôts et magasins de données

1.1 Deux éléments distincts	87
1.2 Identification des caractéristiques	87
	90

2. Méthodologies de conception

2.1 Inmon : de l'entrepôt aux magasins	92
2.2 Kimball : des magasins à l'entrepôt	92
2.3 Une même vision cible	93
	94

2.4 Choisir une méthode et Middle-Out	95
2.5 Le piège de la technique	97
3. Formaliser l'entrepôt et les magasins de données	104
3.1 Les différents rôles de l'entrepôt de données	104
3.2 Les multiples concepts du magasin de données	106
3.2.1 Les limites de la définition de base	106
3.2.2 Le magasin de données comme nécessité technique	107
3.2.3 Ordonancement avec l'entrepôt	111
3.2.4 Cardinalités entre datamarts et restitutions	115
4. La base de données opérationnelle	124
4.1 L'ODS	124
4.2 La Staging Area ou Landing Zone	125
5. Identifier les contraintes	126
5.1 Avant-propos	126
5.2 Contraintes technologiques	127
5.2.1 Processus de choix des technologies	127
5.2.2 Choix d'une solution pour la modélisation de la base de données relationnelle	129
5.2.3 Choix d'une solution pour la modélisation des cubes	131
5.3 Contraintes techniques	

5.3.1 Accès aux données	133
5.3.2 Espace disque	133
5.3.3 Exploitation	133
5.4 Contraintes projet	134
5.4.1 Attentes utilisateurs	134
5.4.2 Planning	134
5.4.3 Normes et modélisation	135
5.4.4 Sécurité	135

Concevoir le modèle

1. Identifier les étoiles	137
1.1 L'importance des étoiles	137
1.2 Structuration ascendante	138
1.3 Identifier un fait	142
2. Formaliser une dimension	145
2.1 Dimension référentielle et dimension de faits	145
2.2 Les hiérarchies	148
2.2.1 Identifier une hiérarchie	148
2.2.2 Construire une hiérarchie	150

2.2.3 Ascendance unique	152
2.2.4 Homogénéité des niveaux	158
2.2.5 Profondeur limitée et prédéfinie	161
2.2.6 Les éléments communs entre les branches	164
2.3 Dimension basée sur les faits	166
2.3.1 Dans la base de données relationnelle	166
2.3.2 Dans la base de données multidimensionnelle	170
2.3.3 Alternatives à l'absence de référentiel	172
2.4 Structure d'une table de dimension	173
2.4.1 Identifiants techniques	173
2.4.2 Codes, libellés et attributs fonctionnels	174
2.4.3 Structure des tables	176
2.5 Le cas particulier de la dimension temps	178
3. Formaliser les indicateurs	181
3.1 Différents termes et différents concepts	181
3.2 Indicateur intensif et indicateur extensif	183
3.3 Indicateurs calculés simples	186
3.4 Indicateurs calculés complexes	187
4. Identifier les données	194
4.1 Dans la base de données décisionnelle	

4.2 Dans la base de données opérationnelle	194
	198
5. Formaliser la table de faits	201
5.1 La structure type d'une table de faits	201
5.2 Les sources des tables de faits	202
5.3 Consolidation de tables de faits	203
5.4 Contraintes techniques	207
5.5 La gestion des valeurs nulles	208
5.6 Granularité de la table de faits	209
5.6.1 Finesse des données	209
5.6.2 Impact sur la volumétrie	210
	210
6. Formaliser la table d'agrégats	212
6.1 Rôle et structure d'une table d'agrégats	212
6.2 Utilisation de filtres	213
6.3 Utilisation d'agrégats de dimension	214
6.3.1 Méthodes et limites	214
6.3.2 Impacts sur la granularité	218
6.4 Utilisation d'agrégats de faits	218
6.5 Suppression de dimension	220
6.6 Contraintes techniques	222

7. Nommer les objets	223
8. Schéma synthétique des flux et types d'objets	224
8.1 Entrepôt en flocon et magasins en étoile	224
8.2 Entrepôt en étoile et magasins en étoile	225
8.3 Entrepôt en flocon et magasins en flocon	225
Historiser les données	
1. Les raisons de l'historisation	227
2. Historisation des dimensions	229
2.1 Une évolution de la réalité métier	229
2.2 Dimension à évolution lente	230
2.2.1 Les méthodes courantes	230
2.2.2 Ajout d'une nouvelle ligne	230
2.2.3 Ajout d'une nouvelle colonne	241
2.3 Dimension à évolution rapide	245
2.4 Choix de la période	246
3. Historisation des faits	248
3.1 Identification des faits	

3.2 Problématique et historisation des faits	248
	248
4. Historiser ou corriger ?	252
4.1 Identification du contexte analytique de la donnée	252
4.2 Le temps des choix et concessions	259
	259
5. Initialisation d'un système historisé	259
5.1 Des données historiques mais non historisées	259
5.2 Utiliser le contexte initial du SID	260
5.3 Autres méthodes	262
	262
6. Durée d'historisation idéale	263
6.1 Entre contrainte technique et utilité métier	263
6.2 Historisation des indicateurs agrégés	269
	269
7. Remarques complémentaires	270
7.1 Reconsidérer le problème	270
7.2 Calcul sur une période	270
7.3 Impact sur la volumétrie	271
7.4 Dates d'effet, saisie et chargement	272
	272

Enrichir le modèle

1. Problématique de conception des objets	275
1.1 Limiter le découpage des objets	275
1.2 Les plages de dates dans les faits	276
1.2.1 Périodes basiques	276
1.2.2 Chevauchements de périodes	285
1.3 L'utilisation d'une mesure comme dimension	293
1.4 Les indicateurs distinctifs	295
1.5 Transformer un attribut en dimension	297
1.6 La relation many-to-many dans une dimension	301
1.6.1 Présentation du problème	301
1.6.2 Répartir les parents dans plusieurs hiérarchies	302
1.6.3 Répartir les parents dans plusieurs dimensions	304
1.7 Diviser une étoile en plusieurs étoiles	307
1.7.1 En enlevant une dimension	307
1.7.2 En rajoutant une dimension	310
1.7.3 Influence des données sources	311
2. Problématiques de la dimension temps	312
2.1 Les valeurs par défaut sur la dimension temps	312
2.2 Les dimensions temps métier	313

2.3 L'utilisation parcimonieuse des dates	318
2.4 L'analyse sur les heures	321
3. Problèmes de granularité des transactions reçues	324
3.1 Description du problème	324
3.2 Des données opérationnelles trop fines	325
3.3 Des données opérationnelles trop larges	326
3.4 Des données opérationnelles hétérogènes	330
4. Problématiques des référentiels	330
4.1 Stratégies de rejet	330
4.2 Utilisation d'alias	333
4.3 Les référentiels non analytiques	335
4.4 Les référentiels autoalimentés	337
4.4.1 Justification et création d'un référentiel autoalimenté	337
4.4.2 Regroupements et hiérarchies des référentiels autoalimentés	339
5. Problématiques de correction des faits	342
5.1 Création d'un flux de correction	342
5.2 Corriger une mesure	343
5.3 Corriger une ventilation	344

Charger les données

1. Le choix de la solution d'intégration	347
1.1 SQL versus ETL	347
1.2 Choix techniques	348
1.2.1 Performances	348
1.2.2 Fonctions avancées	348
1.2.3 Formats de données	348
1.3 Une interface explicite	349
1.4 Compétences	350
1.5 Bilan	350
2. Chargement des éléments du modèle	351
2.1 Les différents types de chargements utilisables dans un SID	351
2.1.1 Annule et Remplace	351
2.1.2 Insertion ou Mise à jour (alias UPSERT)	352
2.1.3 Insertion des modifications (alias INSERT)	352
2.2 Utilisation des types de chargements dans le SID	353
2.2.1 Synthèse par flux	353
2.2.2 L'effet boule de neige dans les flux transactionnels	354
2.2.3 L'évolutivité du contexte	358

3. Ordonnement des éléments	360
3.1 Dépendances des objets	360
3.2 Compartimentage des objets	363
3.3 Fréquence d'intégration	364
4. En dehors des phases d'exploitation	365
4.1 Chargement initial	365
4.2 Reprises de données	366
Optimiser le modèle	
1. Identifier les optimisations	367
2. Optimisation de la base de données relationnelle	368
2.1 Utilisation des index	368
2.2 Utilisation des partitions	371
2.3 Types de colonnes	374
2.4 Utilisation des vues et tables temporaires	375
2.5 Suppression des contraintes	378
3. Optimisation de l'utilisation du serveur ETL	379

Modélisation décisionnelle

Concevoir la base de données pour les traitements OLAP

Glossaire	381
Index	387