

ENTREPÔTS DE DONNÉES

Guide pratique de modélisation dimensionnelle

2^e édition

Ralph Kimball et Margy Ross

Traduction de Claude Raimond



L'édition originale de ce livre a été publiée aux États-Unis par John Wiley & Sons, Inc., 605 Third Avenue, New York, 10158, sous le titre :

The Data Warehouse Toolkit – Second Edition

© Ralph Kimball et Margy Ross – 2002.

© Vuibert – 2003, 2008 – 5, allée de la 2^e DB, 75015 Paris

ISBN 978-2-7117-4811-2

Contact : sciences@vuibert.fr

Web : www.vuibert.fr

Conception de la couverture : Jean Widmer

Les programmes et exemples figurant dans ce livre ont pour but d'illustrer les sujets traités. Il n'est donné aucune garantie quant à leur utilisation dans le cadre d'une activité professionnelle ou commerciale.

Toute représentation ou reproduction intégrale ou partielle, faite sans le consentement de l'auteur, ou de ses ayants droit, ou ayants cause, est illicite (loi du 11 mars 1957, alinéa 1^{er} de l'article 40). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait une contrefaçon sanctionnée par les articles 425 et suivants du Code pénal. La loi du 11 mars 1957 n'autorise, aux termes des alinéas 2 et 3 de l'article 41, que les copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective d'une part et d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration.

Table des matières

Remerciements XVII

Introduction XIX

Chapitre 1

Initiation à la modélisation dimensionnelle I

- 1.1 Des univers d'information différents 2
- 1.2 Objectifs d'un entrepôt de données 2
 - Métaphore de la publication 5
- 1.3 Composants d'un entrepôt de données 7
 - Les applications opérationnelles sources 7
 - Préparation des données 8
 - Présentation des données 10
 - Outils d'accès aux données 14
 - Observations complémentaires 15
 - Métadonnées 15
 - Magasin de données opérationnelles 16
- 1.4 Vocabulaire de la modélisation dimensionnelle 17
 - Table de faits 18
 - Tables de dimension 20
 - Relier les faits et les dimensions 23
- 1.5 Mythes de la modélisation dimensionnelle 26
 - Dix erreurs fréquentes 28
- Résumé 29

Chapitre 2

Grande distribution 31

- 2.1 Processus de modélisation dimensionnelle en quatre étapes 32
- 2.2 Étude de cas de la distribution 34
 - Étape 1. Sélection du processus d'entreprise à modéliser 35

- Étape 2. Déclaration du grain 36
- Étape 3. Choix des dimensions 37
- Étape 4. Identification des faits 38
- 2.3 Attributs de table de dimension 40
 - Dimension date 40
 - Dimension produit 44
 - Dimension magasin 47
 - Dimension promotion 49
 - Table de faits sans fait relative aux promotions 52
 - Dimension numéro de transaction dégénérée 52
- 2.4 Mise en œuvre du schéma de grande distribution 54
- 2.5 Extensibilité du schéma vente au détail 55
- 2.6 Ne pas céder à la facilité 57
 - Normalisation des dimensions (flocons de neige) 58
 - Trop de dimensions 60
- 2.7 Clés artificielles 61
- 2.8 Analyse de panier de marché 65
- Résumé 68

Chapitre 3

Stocks 69

- 3.1 Présentation de la chaîne de valeur 70
- 3.2 Modèles de stock 71
 - Instantané périodique de stock 71
 - Faits semi-additifs 73
 - Faits de stock améliorés 74
 - Transactions de stock 76
 - Instantané récapitulatif de stock 77
- 3.3 Intégration de la chaîne de valeur 78
- 3.4 Architecture de bus de l'entrepôt de données 79
 - Matrice de bus de l'entrepôt de données 81
 - Dimensions conformes 84
 - Faits conformes 89
- Résumé 90

Chapitre 4

Achats 91

- 4.1 Étude de cas des achats 92
- 4.2 Transactions d'achat 93
 - Tables de faits mélangeant ou non les types de transaction 93

- Instantané complémentaire d'achat 97
- 4.3 Dimensions à évolution lente 97
 - Type 1 : écrasement de la valeur précédente 98
 - Type 2 : ajout d'une ligne de dimension 99
 - Type 3 : ajout d'une colonne de dimension 103
- 4.4 Techniques hybrides de traitement des dimensions à évolution lente 104
 - Changements prévisibles et application aux données de multiples versions des attributs modifiés 104
 - Changements imprévisibles avec application aux données antérieures de la version actuelle de l'attribut modifié 106
- 4.5 Dimensions à évolution plus rapide 107
- Résumé 107

Chapitre 5

Gestion des commandes client 109

- 5.1 Présentation de la gestion des commandes 110
- 5.2 Transactions de commande 110
 - Normalisation des faits 111
 - Jeux de rôles d'une dimension 112
 - Retour sur la dimension produit 113
 - Dimension adresse de livraison 115
 - Dimension affaire 118
 - Dimension dégénérée pour le numéro de commande 119
 - Dimensions fourre-tout 119
 - Devises multiples 122
 - Faits d'en-tête et de ligne à des grains différents 123
- 5.3 Transactions de facturation 124
 - Faits de profits et pertes 126
 - Rentabilité — le plus puissant des marchés d'infos 129
 - Mise en garde à propos de la rentabilité 129
 - Faits de satisfaction des clients 130
- 5.4 Instantané récapitulatif du pipeline de traitement des commandes 130
 - Calculs de délais 132
 - Unités de mesure multiples 133
 - Au-delà du rétroviseur 134
- 5.5 Comparaison des tables de faits 135
 - Tables de faits de transaction 135
 - Tables de faits instantané périodique 136
 - Tables de faits instantané récapitulatif 137

- 5.6 Conception de partitions temps réel 138
 - Cahier des charges de la partition temps réel 138
 - Partition temps réel au grain de la transaction 139
 - Partition temps réel d'instantané périodique 140
 - Partition temps réel d'instantané récapitulatif 140
- Résumé 141

Chapitre 6

Gestion des relations client 143

- 6.1 Vue d'ensemble de la GRC 144
 - GRC opérationnelle et GRC analytique 145
 - GRC sur étagère 147
- 6.2 Dimension client 148
 - Analyse syntaxique des noms et adresses 149
 - Considérations sur les noms et adresses dans un contexte international 151
 - Autres attributs client courants 152
 - Dates 152
 - Attributs de segmentation des clients et scores 153
 - Attributs représentant des faits agrégés 153
 - Table de dimension déportée pour un ensemble d'attributs de faible cardinalité 154
 - Grandes dimensions client changeantes 155
 - Ensembles d'attributs de largeur variable 160
 - Incidence des changements de la solution de type 2 sur les dimensions client 161
 - Groupes d'étude de comportement des clients 162
 - Hierarchies des entreprises clientes 163
 - Hierarchies à profondeur fixe 164
 - Hierarchies à profondeur variable 164
 - Combinaison de sources multiples de données client 170
- 6.3 Analyse de données client provenant de multiples processus d'entreprise 171
- Résumé 172

Chapitre 7

Comptabilité 175

- 7.1 Étude de cas de comptabilité 176
- 7.2 Données de comptabilité générale 177
 - Instantané périodique de la comptabilité générale 177
 - Plan comptable 177
 - Clôture d'une période 178

- Faits sur un an à la date du jour 179
- Retour sur les devises multiples 179
- Transactions de journaux de comptabilité générale 180
 - Types de faits 181
 - Multiples calendriers comptables 181
- Documents comptables 182
- 7.3 Budget 183
 - Tables de faits consolidées 186
- 7.4 OLAP et packages d'applications analytiques 188
- Résumé 189

Chapitre 8

Gestion des ressources humaines 191

- 8.1 Suivi de transactions horodatées dans une dimension 192
- 8.2 Dimension horodatée avec des faits d'instantané périodique 195
- 8.3 Dimension audit 197
- 8.4 Table de dimension déportée mot-clé 198
 - Le dilemme AND/OR 199
 - Recherche de sous-chaînes 200
- 8.5 Données de questionnaires d'enquête 201
- Résumé 202

Chapitre 9

Services financiers 203

- 9.1 Étude de cas bancaire 204
- 9.2 Ségrégation des dimensions 204
 - Dimension foyer 207
 - Dimensions à valeurs multiples 208
 - Retour sur les minidimensions 209
- 9.3 Plages de valeurs de faits à limites arbitrairement définies 211
- 9.4 Soldes à un point dans le temps 212
- 9.5 Schémas de produits hétérogènes 214
 - Produits hétérogènes avec faits de transaction 219
- Résumé 219

Chapitre 10

Télécommunications, distribution d'eau et d'électricité 221

- 10.1 Étude de cas télécommunications 222

- 10.2 Considérations générales sur la révision d'une conception 224
 - Granularité 224
 - Granularité des faits 224
 - Granularité des dimensions 225
 - Dimension date 225
 - Données réparties par période à la place des tables de dimension 225
 - Dimensions dégénérées 226
 - Décodages et descriptions dans les dimensions 226
 - Clés artificielles 227
 - Trop (ou trop peu) de dimensions 227
- 10.3 Discussion de l'ébauche de conception 227
- 10.4 Dimension emplacement géographique 229
 - Table d'emplacement déportée 230
 - Exploitation des possibilités des systèmes d'information géographique 231
- Résumé 231

Chapitre 11

Transports 233

- 11.1 Cas des passagers réguliers d'une compagnie aérienne 234
 - Diverses granularités de table de faits 234
 - Regroupement des segments en voyages 237
- 11.2 Extension à d'autres industries 238
 - Transport maritime 238
 - Agences de voyages 239
- 11.3 Combinaison de petites dimensions en une superdimension 240
 - Classe 240
 - Origine et destination 241
- 11.4 Considérations supplémentaires sur la date et l'heure 243
 - Calendriers spécifiques par pays 243
 - L'heure comme dimension ou comme fait 244
 - Date et heure dans le cas de multiples fuseaux horaires 245
- Résumé 246

Chapitre 12

Enseignement 247

- 12.1 Étude de cas de l'université 248
- 12.2 Instantané récapitulatif pour le suivi des admissions 248
- 12.3 Tables de faits sans fait 250
 - Événements d'inscription des étudiants 251
 - Suivi de l'utilisation des ressources 253

- Événements de présence des étudiants aux cours 254
 - Lignes explicites pour ce qui ne s'est pas produit 255
 - Autres options de traitement de ce qui ne s'est pas produit 255
 - Traitement multidimensionnel de ce qui ne s'est pas produit 257

12.4 Autres domaines susceptibles d'analyse 257

Résumé 258

Chapitre 13

Santé 259

- 13.1 Cercle de valeur de la santé 260
- 13.2 Facture de soins 262
 - Rôles joués par la dimension date 265
 - Dimension diagnostic à valeurs multiples 266
 - Extension d'une table de faits de facturation pour faire apparaître la rentabilité 269
 - Dimensions pour la facturation des soins aux patients hospitalisés 270
- 13.3 Événements de soin complexes 271
- 13.4 Dossiers médicaux 273
 - Dimension fait pour les faits épars 273
- 13.5 Remonter dans le temps 275
 - Lignes de faits en retard 275
 - Lignes de dimension en retard 277
- Résumé 278

Chapitre 14

Commerce électronique 281

- 14.1 Initiation aux interactions client-serveur sur le Web 282
- 14.2 Le flux interactif n'est pas une source de données comme les autres 285
 - Obstacles à l'identification des données de flux interactif 286
 - Identification de l'origine du visiteur 286
 - Identification de la session 287
 - Identification du visiteur 288
 - Serveurs proxy 289
 - Caches de navigateur 290
 - Dimensions spécifiques du flux interactif 291
 - Dimension page 292
 - Dimension événement 293
 - Dimension session 293
 - Dimension prescription 295

- 14.3 Table de faits de flux interactif pour sessions complètes 296
- 14.4 Table de faits flux interactif pour événements de page individuels 298
- 14.5 Tables de faits de flux interactif agrégées 301
- 14.6 Intégration du marché d'infos flux interactif dans l'entrepôt de données d'entreprise 303
- 14.7 Marché d'infos de rentabilité du commerce électronique 305
- Résumé 308

Chapitre 15

Assurances 309

- 15.1 Étude de cas assurances 310
 - Chaîne de valeurs des assurances 312
 - Première version de la matrice de bus des assurances 313
- 15.2 Transactions de police 314
 - Précisions techniques relatives aux dimensions 314
 - Jeux de rôles des dimensions 315
 - Dimensions à évolution lente 315
 - Minidimensions pour les dimensions très grandes ou à changement rapide 316
 - Attributs de dimension à valeurs multiples 317
 - Dimension dégénérée 318
 - Dimension audit 318
 - Produits hétérogènes 319
 - Instantané récapitulatif de police complétant ou remplaçant la table des transactions police 320
- 15.3 Instantané périodique des polices 320
 - Dimensions conformes 321
 - Faits conformes 321
 - Mesures d'entités payées d'avance 321
 - D'autres produits hétérogènes 323
 - D'autres dimensions à valeurs multiples 323
- 15.4 Autres informations générales sur le métier d'assureur 323
 - Mise à jour de la matrice de bus des assurances 325
- 15.5 Transactions de sinistre 327
- 15.6 Instantané récapitulatif des sinistres 328
- 15.7 Instantané consolidé police/sinistre 329
- 15.8 Événements d'accident sans fait 330
- 15.9 Erreurs courantes de modélisation dimensionnelle 331
- Résumé 335

Chapitre 16

Construction de l'entrepôt de données 337

- 16.1 Cycle de vie d'un entrepôt dimensionnel d'entreprise 338
 - Grandes lignes du cycle de vie 339
- 16.2 Planification et gestion du projet 340
 - Détermination du degré d'engagement et de préparation 340
 - Définition de l'ampleur du projet 342
 - Justification 343
 - Constitution de l'équipe 343
 - Développement et maintenance du plan du projet 346
- 16.3 Définition des besoins de l'entreprise 347
 - Préparatifs en vue de la collecte des besoins 347
 - Choix du type de rencontre 347
 - Constitution et préparation de l'équipe de recueil des besoins 348
 - Sélection, programmation et préparation des représentants des utilisateurs 349
 - Recueil des besoins de l'entreprise 350
 - Démarrage 350
 - Déroulement de l'interview 350
 - Conclusion 351
 - Interviews consacrées aux données 352
 - Documentation et suivi postérieur au recueil des besoins 352
 - Établissement des priorités et recherche d'un consensus 353
- 16.4 Les tâches liées à la technologie 354
- 16.5 Conception de l'architecture technique 354
 - Les huit étapes d'élaboration de l'architecture technique 355
 - Constituer un comité pour l'architecture 356
 - Rassembler les besoins ayant un rapport avec l'architecture 356
 - Documenter les exigences et les contraintes architecturales 356
 - Développer un modèle architectural de haut niveau 357
 - Concevoir et spécifier les sous-systèmes 357
 - Définir les phases d'implémentation de l'architecture 358
 - Documenter l'architecture technique 358
 - Réviser et finaliser l'architecture technique 358
 - Sélection et installation des produits 358
- 16.6 Tâches se rapportant aux données 360
- 16.7 Modélisation dimensionnelle 360
 - Conception physique 362
 - Stratégie d'agrégation 363
 - Stratégie initiale d'indexation 364
- 16.8 Conception et développement de la préparation des données 365

- Préparation des tables de dimension 366
- Préparation des tables de faits 368
- 16.9 Développement des applications analytiques 370
 - Spécification des applications analytiques 370
 - Développement des applications analytiques 371
- 16.10 Mise en service 372
- 16.11 Maintenance et expansion 373
- 16.12 Erreurs à ne pas commettre dans la construction
 - d'un entrepôt de données 375
- Résumé 377

Chapitre 17

Impératifs actuels et perspectives d'avenir 379

- 17.1 Avancées constantes de la technologie 380
- 17.2 Forces politiques exigeant la sécurité
 - et affectant la confidentialité 383
 - Conflit entre les utilisations bénéfiques et les abus insidieux 383
 - Qui est propriétaire de vos informations personnelles ? 385
 - Que pourrait-il arriver ? Surveillance des surveillants... 385
 - Effets de la surveillance des surveillants
 - sur l'architecture de l'entrepôt de données 386
- 17.3 Conception visant à survivre aux catastrophes 388
 - Arrêts catastrophiques 388
 - Parades aux arrêts catastrophiques 389
- 17.4 Propriété intellectuelle et usage équitable 391
- 17.5 Tendances culturelles influençant
 - les entrepôts de données 392
 - Management par les chiffres dans toute l'entreprise 392
 - L'utilisation croissante d'indicateurs
 - de performance sophistiqués 393
 - Le comportement est la nouvelle application phare 393
 - Les packages d'application ont atteint leur apogée 394
 - L'intégration des applications doit être faite par quelqu'un 395
 - La sous-traitance de l'entrepôt de données
 - mérite un examen objectif des risques associés 395
- 17.6 Le mot de la fin 396

Glossaire 397

Index 425