

Bases de données

Concepts, utilisation
et développement

Jean-Luc Hainaut

Professeur à la faculté d'informatique,
membre du Namur Digital Institute,
université de Namur

5^e édition

DUNOD

Toutes les marques citées dans cet ouvrage
sont des marques déposées par leurs propriétaires respectifs

Illustration de couverture : © marukopum - Shutterstock

© Dunod, 2009, 2022 pour la présente édition
11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-084680-1

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.



Grand éditeur de SGBD SQL observant avec ironie, condescendance mais aussi un appétit non dissimulé les efforts des SGBD dits NoSQL¹.

1. Référence allégorique à la section 1.7.8. *Linogravure*, B. Hainaut, 2004

Table des matières

AVANT-PROPOS

CHAPITRE 1 • MOTIVATION ET INTRODUCTION	33
1.1 Des données et des hommes	33
1.2 Une première approche des données	34
1.3 Utiliser une base de données	38
1.4 Premières conclusions	39
1.5 Les systèmes de gestion de bases de données	40
1.6 Les défis des bases de données d'aujourd'hui	41
1.7 Un peu d'histoire	42
1.7.1 Le modèle collectiviste : CODASYL DBTG	43
1.7.2 La pensée unique : IMS	44
1.7.3 La naissance du modèle relationnel	44
1.7.4 Apparition d'un nouvel acteur : l'utilisateur final	45
1.7.5 La guerre des modèles	45
1.7.6 La chute d'Icare	46
1.7.7 Le modèle relationnel : la suite de l'histoire	46
1.7.8 Cronos dévore ses enfants	47

PARTIE 1

CONCEPTS DES BASES DE DONNÉES

CHAPITRE 2 • CONCEPTS DES BASES DE DONNÉES	49
2.1 Introduction	49
2.2 Tables, lignes et colonnes	50
2.3 La valeur null	51

2.4	Les identifiants	52
2.5	Les clés étrangères	53
2.6	Identifiants et clés étrangères facultatifs	54
2.7	Schéma et contenu d'une base de données	55
2.8	Un exemple représentatif de base de données	55
2.9	Représentation graphique d'un schéma	56
2.10	Modifications et contraintes d'intégrité	59
2.10.1	Les colonnes obligatoires	59
2.10.2	Les contraintes d'unicité (identifiants)	59
2.10.3	Les contraintes référentielles (clés étrangères)	59
2.11	Redondances internes dans une table	61
2.11.1	Le phénomène de redondance interne	61
2.11.2	Normalisation par décomposition	63
2.12	Les structures physiques	64
2.13	Que retenir ?	66
2.14	Pour en savoir plus	67
2.15	Pratiquer les bases de données	67
2.16	Exercices	69
CHAPITRE 3 • MODÈLE RELATIONNEL ET NORMALISATION		71
3.1	Le phénomène de redondance interne	71
3.2	Le modèle relationnel de bases de données	72
3.2.1	Concepts de base	72
3.2.2	Opérateurs d'extraction de données	73
3.3	Identifiants	76
3.4	Dépendances fonctionnelles	77
3.4.1	Le phénomène de dépendance fonctionnelle	77
3.4.2	Un exemple plus complexe	78
3.4.3	Graphe ADF d'une relation (attributs et DF)	79
3.4.4	Propriétés des dépendances fonctionnelles	79
3.5	Contraintes d'inclusion et clés étrangères	82
3.6	Calcul des identifiants d'une relation	85
3.6.1	Relation entre identifiants et dépendances fonctionnelles	85
3.6.2	La relation possède un seul identifiant	86
3.6.3	La relation possède plusieurs identifiants	87
3.6.4	Rôle des attributs racines	88
3.7	Décomposition d'une relation	89
3.8	Normalisation d'une relation	90

3.8.1	Réexamen du phénomène de redondance interne	91
3.8.2	Analyse du phénomène	91
3.8.3	Relation normalisée	92
3.8.4	Normalisation d'une relation	92
3.8.5	Normalisation d'un noyau irréductible	95
3.8.6	SGBD et dépendances fonctionnelles	97
3.9	Les relations <i>non en 1re forme normale</i>	97
3.10	Que retenir ?	100
3.11	Pour en savoir plus	102
3.12	Exercices	102
CHAPITRE 4 • ÉLÉMENTS DE TECHNOLOGIE DES BASES DE DONNÉES		105
4.1	Introduction	105
4.2	Les composants d'un ordinateur	106
4.3	Mémoire externe : le disque magnétique	107
4.3.1	Géométrie d'un disque	107
4.3.2	Lecture et écriture de données sur un disque	109
4.3.3	Optimisation : tampon d'échange et lecture anticipée	111
4.4	Les technologies alternatives	113
4.4.1	Les disques RAID	113
4.4.2	Les disques électroniques (SSD)	115
4.4.3	Les mémoires à bande	117
4.4.4	Disques optiques à très grande capacité	118
4.4.5	Les machines de bases de données	118
4.5	Organisation d'un espace de stockage	120
4.5.1	Structure d'un fichier	120
4.5.2	Allocation d'espace à un fichier	121
4.5.3	Structure des enregistrements d'un fichier	121
4.5.4	Rangement des enregistrements dans un fichier	121
4.5.5	Le phénomène de fragmentation d'un fichier	123
4.5.6	Organisation des fichiers sur le support	124
4.6	Traitement séquentiel d'un fichier	124
4.6.1	Lecture séquentielle d'un fichier	125
4.6.2	Tri d'un fichier séquentiel	128
4.7	Les index	132
4.8	Organisation séquentielle indexée	134
4.8.1	Structure d'un fichier séquentiel indexé	134
4.8.2	Lecture et modification d'un fichier séquentiel indexé	135
4.8.3	Caractéristiques et performances d'un fichier séquentiel indexé	138

4.8.4	Étude de cas	145
4.9	Organisation calculée	146
4.9.1	Principes de l'accès calculé	146
4.9.2	Fonction de calcul d'adresse	147
4.9.3	Gestion des débordements	149
4.9.4	Caractéristiques et performances d'un fichier à accès calculé	150
4.9.5	Chargement et réorganisation d'un fichier à accès calculé	153
4.9.6	Comparaison des deux techniques d'index primaires	154
4.10	Les index secondaires	154
4.10.1	Structure d'un index secondaire par listes de pointeurs	155
4.10.2	Performances d'un index secondaire par listes de pointeurs	156
4.10.3	Structure d'un index bitmap	159
4.10.4	Les index partiels	159
4.10.5	Les index sur expression	159
4.10.6	Les index secondaires augmentés	159
4.10.7	Disponibilité des techniques d'indexation primaires et secondaires	160
4.11	Les index multi-champs	161
4.12	Accès par index ou recherche séquentielle ?	162
4.13	Quand créer un index ?	162
4.14	Les techniques d'agrégation (clustering)	163
4.14.1	Clustering index	163
4.14.2	Clusters	164
4.14.3	Performances des techniques d'agrégation	165
4.15	Stabilité des adresses	166
4.16	Un exemple : SQL Server	168
4.17	Compléments sur les fonctions de hachage	168
4.17.1	Le prédicteur à hachage (filtre de Bloom)	168
4.17.2	Hachage cryptographique (Secure hashing)	169
4.17.3	Notion d'identifiant universel (UUID)	170
4.18	Futurologie : un peu de poésie dans ce monde de brutes !	171
4.18.1	Les mémoires ADN	171
4.18.2	Les mémoires atomiques	172
4.18.3	Les bases de données quantiques	173
4.19	Que retenir ?	175
4.20	Pour en savoir plus	176
4.21	Exercices	176
CHAPITRE 5 • LES SYSTÈMES DE GESTION DE BASES DE DONNÉES		179
5.1	Introduction	179

5.2	Les fonctions des systèmes de gestion de bases de données	180
5.3	Architecture des SGBD	182
5.4	Les transactions	184
5.5	Régulation de la concurrence	187
5.5.1	Mise à jour perdue	187
5.5.2	Données instables	188
5.5.3	Régulation de la concurrence par verrouillage	189
5.5.4	Sérialisabilité et verrouillage en 2 phases (2PL)	190
5.5.5	Interblocage	190
5.5.6	Les niveaux d'isolation	191
5.6	Protection contre les accidents	192
5.6.1	Sauvegardes et journaux	192
5.6.2	Reprise à froid	192
5.6.3	Reprise à chaud	193
5.6.4	Les fichiers différentiels	194
5.7	Les bases de données distribuées	194
5.7.1	Un premier exemple	194
5.7.2	Les types de distribution	195
5.7.3	Partitionnement du réseau	197
5.7.4	Le commit en deux phases (2PC)	197
5.7.5	Distribution des traitements	198
5.8	Contrôle des accès	199
5.8.1	Les modèles de contrôle d'accès	199
5.8.2	Protection des données privées - Inférence statistique	200
5.9	Les bases de données in-memory	202
5.10	Le marché des SGBD relationnels	203
5.11	Pour en savoir plus	204

PARTIE 2

UTILISATION DES BASES DE DONNÉES

CHAPITRE 6 • LE LANGAGE SQL DDL	205
6.1 Introduction	205
6.2 Le langage SQL DDL	206
6.3 Création d'un schéma	206
6.4 Les noms SQL	207
6.5 Création d'une table	207
6.6 Les colonnes et leurs types	207

6.7	Caractère obligatoire/facultatif d'une colonne	209
6.8	Les identifiants	210
6.9	Les clés étrangères	210
6.10	Forme synthétique des contraintes	212
6.11	Un exemple complet	213
6.12	Suppression d'une table	213
6.13	Ajout, retrait et modification d'une colonne	214
6.14	Ajout et retrait d'une contrainte	214
6.15	Les structures physiques	215
6.16	Que retenir ?	216
6.17	Pour en savoir plus	216
CHAPITRE 7 • LE LANGAGE SQL DML (1)		217
7.1	Introduction	217
7.2	Requêtes élémentaires	219
7.2.1	Extraction simple	219
7.2.2	Extraction de lignes sélectionnées	220
7.2.3	Lignes dupliquées dans le résultat	221
7.2.4	Des conditions de sélection plus complexes	222
7.2.5	Expressions composées	223
7.3	Données extraites et données dérivées	224
7.3.1	Les fonctions SQL	225
7.3.2	Les registres du système	227
7.4	Ordre des lignes du résultat	227
7.5	Les fonctions agrégatives (ou statistiques)	229
7.6	Les sous-requêtes	231
7.6.1	Condition d'association par sous-requête	231
7.6.2	Sous-requête et clé étrangère composite	232
7.6.3	Attention aux conditions d'association négatives	233
7.6.4	Références multiples à une même table	235
7.7	Les quantificateurs ensemblistes	237
7.8	Exercices	239
CHAPITRE 8 • LE LANGAGE SQL DML (2)		241
8.1	Introduction	241
8.2	Extraction de données de plusieurs tables (jointure)	242
8.2.1	La jointure de plusieurs tables	242
8.2.2	Conditions de jointure et conditions de sélection	243

8.2.3 Lignes célibataires et jointures externes	244
8.3 Les opérateurs ensemblistes	244
8.4 Le produit relationnel	246
8.5 Variantes syntaxiques de la jointure	247
8.6 Les tables statiques	249
8.7 Requêtes sur des structures cycliques	249
8.8 Compléments sur les jointures	253
8.8.1 Sous-requête ou jointure ?	254
8.8.2 Valeurs dérivées dans une jointure	256
8.8.3 Les jointures généralisées	256
8.8.4 Interprétation du résultat d'une jointure	257
8.9 Extraction de données groupées	259
8.9.1 Notion de groupe de lignes	259
8.9.2 Sélection de groupes et sélection de lignes	261
8.9.3 Groupes et jointures	262
8.9.4 Composition de la liste select	262
8.9.5 Composition du critère de groupement	263
8.9.6 Attention aux groupements multiniveaux	264
8.9.7 Peut-on éviter l'utilisation de données groupées ?	265
8.10 Extension de la structure des requêtes SFW	265
8.10.1 Extension de la clause select	265
8.10.2 Extension de la clause from	266
8.11 Limitation de la taille d'un résultat	267
8.12 Interprétation d'une requête	268
8.13 Modification des données	269
8.13.1 Ajout de lignes	269
8.13.2 Suppression de lignes	270
8.13.3 Modification de lignes	271
8.13.4 À propos de la gestion des verrous	272
8.13.5 Mise à jour et contraintes d'unicité	272
8.13.6 Mise à jour et contraintes référentielles	273
8.13.7 Déterminisme des requêtes SQL	273
8.14 Modification des structures de données	273
8.15 Que retenir ?	274
8.16 Pour en savoir plus	274
8.17 Exercices	275
 CHAPITRE 9 • LE LANGAGE SQL AVANCÉ	 279
9.1 Introduction	279

9.2	Le contrôle d'accès	279
9.3	Les vues SQL	282
9.3.1	Principes et objectifs des vues	282
9.3.2	Définition et utilisation d'une vue	282
9.3.3	Une vue sur colonne : les colonnes calculées	283
9.3.4	Les vues comme interface pour des besoins particuliers	283
9.3.5	Les vues comme mécanisme de contrôle d'accès	284
9.3.6	Les vues comme mécanisme d'évolution de la base de données	284
9.3.7	Les vues comme aide à l'expression de requêtes complexes	285
9.3.8	Mise à jour des données via une vue	285
9.4	Les Common Table Expressions (CTE)	286
9.5	Les requêtes récursives	286
9.6	Les extensions objet de SQL3	289
9.6.1	Types de données complexes (<code>row</code> et <code>array</code>)	290
9.6.2	Type défini par l'utilisateur (TUD)	292
9.6.3	Table typée	292
9.6.4	Hiérarchie de types	293
9.6.5	Hiérarchie de tables typées	294
9.6.6	Références entre tables	296
9.6.7	Remarques finales	296
9.7	Les prédicats (<code>check</code>)	296
9.8	Les procédures SQL (<code>stored procedures</code>)	298
9.9	Les déclencheurs (<code>triggers</code>)	298
9.10	Le catalogue	303
9.11	Les interfaces SQL des programmes d'application	307
9.11.1	L'interface SQL statique classique	308
9.11.2	SQLJ : une interface statique pour Java	312
9.11.3	L'interface SQL dynamique classique	313
9.11.4	JDBC : une interface CLI-SQL dynamique pour Java	316
9.11.5	Comparaison des modèles d'interaction	319
9.11.6	Un problème de sécurité : l'injection de code SQL	320
9.12	SQL et l'information incomplète	321
9.12.1	Introduction	321
9.12.2	La valeur <code>null</code> de SQL	322
9.12.3	La logique ternaire de SQL	322
9.12.4	La propagation de <code>null</code> en SQL	322
9.12.5	La propagation de <code>unknown</code> en SQL	323
9.12.6	Les problèmes de l'information incomplète en SQL	324
9.12.7	Deux recommandations	329
9.13	Que retenir ?	330

9.14 Pour en savoir plus	331
CHAPITRE 10 • SQL ET L'ANALYSE DE DONNÉES	333
10.1 Introduction	333
10.2 Rappel sur les fonctions agrégatives classiques	334
10.3 Extension des fonctions de groupement	336
10.4 La fonction window de SQL - Introduction	337
10.5 Anatomie d'une fonction window	340
10.6 La fenêtre	342
10.7 Les partitions	342
10.8 L'ordre	343
10.9 Le frame	343
10.10 Les fonctions de calcul	345
10.10.1 Les fonctions agrégatives classiques	346
10.10.2 Les fonctions de ranking	346
10.10.3 Les fonctions positionnelles	347
10.10.4 Les frames par défaut	348
10.10.5 Le filtre	349
10.11 Que retenir	349
10.12 Pour en savoir plus	350
10.13 Exercices	351
CHAPITRE 11 • LES BASES DE DONNÉES NON RELATIONNELLES	353
11.1 Introduction	353
11.2 Les bases de données NoSQL	355
11.2.1 Modèles clé-valeur	356
11.2.2 Modèles orientés documents	357
11.2.3 Modèles orientés colonnes	357
11.3 Les autres modèles non relationnels	358
11.4 Les architectures de distribution	359
11.4.1 La cohérence finale	359
11.4.2 Les protocoles à quorum	360
11.4.3 Le théorème CAP	361
11.5 Un SGBD orienté colonnes : Cassandra	361
11.6 Un SGBD orienté documents : MongoDB	365
11.7 Les SGBD NewSQL : l'exemple de VoltDB	368
11.8 Conclusions sur les SGBD NoSQL	369
11.9 Les chaînes de blocs	371

11.9.1 Principes généraux	372
11.9.2 Le protocole BitCoin	376
11.9.3 Le protocole Ethereum	380
11.10 Pour en savoir plus	383

PARTIE 3

DÉVELOPPEMENT DES BASES DE DONNÉES

CHAPITRE 12 • CONSTRUCTION D'UNE BASE DE DONNÉES	385
12.1 De l'artisanat à la méthode	385
12.2 Les acteurs et les métiers	386
12.3 Deux parcours	387
12.4 Une démarche simplifiée	387
12.5 Les outils	388
CHAPITRE 13 • LE MODÈLE ENTITÉ-ASSOCIATION DE BASE	389
13.1 Introduction	389
13.2 Les types d'entités	390
13.3 Les attributs	391
13.4 Les types d'associations	392
13.4.1 Rôles d'un type d'associations	392
13.4.2 Propriétés d'un type d'associations	393
13.4.3 Classe fonctionnelle d'un type d'associations	393
13.4.4 Type d'associations (ou rôle) obligatoire ou facultatif	395
13.4.5 Cardinalités d'un type d'associations	396
13.4.6 Type d'associations cyclique	396
13.4.7 Représentation graphique des populations	397
13.5 Une question de vocabulaire	397
13.6 Les identifiants	398
13.6.1 Identifiants hybrides	399
13.6.2 Composition des identifiants	400
13.6.3 Identifiants minimaux et identifiants implicites	401
13.6.4 Importance du concept d'identifiant	402
13.7 Les contraintes d'intégrité	403
13.7.1 Contraintes de base	403
13.7.2 Contraintes additionnelles	403
13.8 Contenu informationnel d'un schéma	403
13.9 Exemples	405

13.9.1 Une structure administrative	405
13.9.2 Gestion d'une bibliothèque	406
13.9.3 Voyages en train	407
13.10 Que retenir ?	408
13.11 Pour en savoir plus	409
13.12 Dessiner un schéma conceptuel à l'aide d'un outil	409
13.13 Exercices	410
 CHAPITRE 14 • ÉLABORATION D'UN SCHÉMA CONCEPTUEL	 413
14.1 Introduction	413
14.2 Un premier exemple	414
14.3 Décomposition de l'énoncé	416
14.3.1 Notion de proposition élémentaire	416
14.3.2 Cardinalités	417
14.3.3 Propositions générales et propositions particulières	418
14.3.4 Attention aux propositions complexes irréductibles !	418
14.3.5 Les propositions non binaires réductibles	421
14.4 Pertinence d'une proposition	421
14.5 Représentation d'une proposition	422
14.5.1 Nouveau type d'entités et son attribut	422
14.5.2 Type d'associations entre types d'entités existants	423
14.5.3 Attribut d'un type d'entités existant	423
14.5.4 Nouveau type d'entités et ses deux attributs	424
14.5.5 Restructuration pour intégration	424
14.5.6 Propriétés multivaluées	427
14.5.7 Recommandations pratiques	430
14.6 Non-redondance des propositions	431
14.6.1 Redondance explicite	431
14.6.2 Variantes d'expression	431
14.6.3 Redondance indirecte	431
14.6.4 Redondance apparente	432
14.6.5 Synonymes et homonymes	432
14.7 Non-contradiction des propositions	433
14.8 Les contraintes d'intégrité	434
14.9 Normalisation du schéma	435
14.9.1 Simplification du schéma	435
14.9.2 Élimination des redondances internes	435
14.10 Documentation du schéma	437
14.11 Complétude du schéma	438

14.12	Que retenir ?	438
14.13	Pour en savoir plus	439
14.14	Exercices	439
CHAPITRE 15 • PRODUCTION DU SCHÉMA DE LA BASE DE DONNÉES		443
15.1	Introduction	443
15.2	Représentation des types d'entités	444
15.3	Représentation des attributs	444
15.4	Représentation des types d'associations	444
15.4.1	Types d'associations un-à-plusieurs (1:N)	444
15.4.2	Types d'associations un-à-un (1:1)	446
15.4.3	Types d'associations plusieurs-à-plusieurs (N:N)	448
15.4.4	Types d'associations cycliques	449
15.5	Représentation des identifiants	449
15.6	Traduction des noms	450
15.7	Synthèse des règles de traduction	451
15.8	Les structures physiques	452
15.9	Traduction des structures en SQL	454
15.10	Que retenir ?	455
15.11	Pour en savoir plus	456
15.12	Produire un schéma de base de données à l'aide d'un outil	456
15.13	Exercices	457
CHAPITRE 16 • MÉTHODOLOGIE DES BASES DE DONNÉES		459
16.1	Introduction	459
16.2	Méthodes, processus et produits	460
16.3	Méthodes de développement d'une base de données	461
16.4	Les modèles de développement	466
16.5	Techniques transformationnelles	468
16.6	Rôle des outils	468
CHAPITRE 17 • LE MODÈLE ENTITÉ-ASSOCIATION ÉTENDU		469
17.1	Introduction	469
17.2	Le modèle de base - rappel	470
17.3	Sémantiques	471
17.4	Modèle, schéma et population	472
17.5	Les stéréotypes	474

17.6	Schéma de base de données	475
17.7	Attributs	475
17.7.1	Attributs composés	475
17.7.2	Attributs multivalués	476
17.7.3	Attributs complexes	476
17.7.4	Domaine et type d'un attribut	476
17.7.5	Cardinalité d'un attribut	477
17.8	Types d'associations	478
17.8.1	Types d'associations fonctionnels	478
17.8.2	Types d'associations n-aires	478
17.8.3	Types d'associations à rôles polymorphiques	479
17.8.4	Attributs de type d'associations	480
17.8.5	Généralisation des cardinalités des rôles	480
17.9	Types d'entités	481
17.9.1	Relation de généralisation-spécialisation (relation is-a)	481
17.9.2	Propriétés et contraintes ensemblistes des sous-types	482
17.9.3	L'héritage	484
17.9.4	Les sous-types non disjoints sont-ils vraiment utiles ?	485
17.9.5	Surtypes et héritage multiples	486
17.9.6	Contraintes portant sur des composants hérités	488
17.9.7	Répartitions multiples	488
17.9.8	Règle de répartition en sous-types	489
17.10	Identifiants	490
17.10.1	Identifiants d'un type d'entités	491
17.10.2	Identifiants d'un type d'associations	493
17.10.3	Identifiants d'un attribut complexe	495
17.11	Contraintes d'intégrité	495
17.11.1	Contraintes de domaine	496
17.11.2	Contraintes entre valeurs	497
17.11.3	Contraintes d'existence	497
17.11.4	Dépendances fonctionnelles	499
17.12	Contraintes d'intégrité additionnelles	499
17.12.1	Expression des contraintes	499
17.12.2	Remarques sur les langages d'expression de contraintes	500
17.12.3	Cycles et contraintes cycliques	501
17.12.4	Propriétés des graphes d'instances	502
17.12.5	Contraintes de redondance	503
17.13	Les noms	504
17.13.1	Règles de dénomination	504
17.13.2	Dénomination des attributs	505

17.13.3 Dénomination des types d'associations	505
17.14 Classes d'objets génériques	506
17.14.1 Type d'associations de composition	506
17.14.2 Type d'associations de matérialisation	507
17.15 Exemple de schéma Entité-association	509
17.16 Métadonnées et métaschémas	510
17.16.1 Une première approche : le catalogue d'une base de données	510
17.16.2 Métaschéma du modèle Entité-association	511
17.17 Sémantique interne du modèle Entité-association	512
17.17.1 Types d'entités et relations is-a	512
17.17.2 Types d'associations	513
17.17.3 Attributs	513
17.17.4 Attributs et rôles obligatoires	513
17.17.5 Identifiants	513
17.17.6 Autres contraintes	514
17.17.7 Application de la théorie relationnelle au modèle Entité-association	514
17.18 Que retenir ?	515
17.19 Créer un schéma conceptuel à l'aide d'un outil (compléments)	515
17.20 Exercices	516
 CHAPITRE 18 • LES DIAGRAMMES DE CLASSES UML	 519
18.1 Introduction	519
18.2 UML et les bases de données	520
18.3 Les mécanismes d'extension	521
18.4 Les classes d'objets	522
18.5 Les associations binaires	522
18.6 Les associations n-aires	523
18.6.1 Multiplicité maximale (max)	524
18.6.2 Multiplicité minimale (min)	524
18.7 Les attributs	525
18.8 Les opérations	525
18.9 Les associations qualifiées	526
18.10 Les classes associations	528
18.11 Les associations d'agrégation et de composition	528
18.12 Les contraintes	529
18.13 Le langage de contraintes OCL	530
18.14 Une question de vocabulaire	533
18.15 Exemple de schéma de classes en UML	533

18.16 Le modèle de classes d'UML revisité (DB-UML)	533
18.17 Sémantique du modèle de classes DB-UML	535
18.18 Que retenir ?	537
18.19 Pour en savoir plus	537
18.20 Créer un schéma conceptuel UML à l'aide d'un outil	538
 CHAPITRE 19 • ANALYSE CONCEPTUELLE DU DOMAINE D'APPLICATION	 539
19.1 Introduction	539
19.2 Analyse des énoncés textuels	542
19.2.1 Proposition existentielle	542
19.2.2 Proposition relationnelle binaire	543
19.2.3 Proposition relationnelle n-aire	546
19.2.4 Intégration d'une proposition dans un schéma	550
19.3 Les patrons de conception	551
19.4 Analyse des documents structurés	553
19.4.1 Analyse des documents de l'organisation	553
19.4.2 Analyse d'un écran de saisie de données	555
19.4.3 Analyse d'un document papier	556
19.5 Dénomination des objets d'un schéma	557
19.6 Processus de correction d'un schéma	558
19.6.1 Constructions syntaxiquement incorrectes	559
19.6.2 Constructions non satisfiables et incohérentes	560
19.7 Processus de normalisation du schéma conceptuel	564
19.7.1 Simplification des constructions non minimales	566
19.7.2 Explicitation des constructions insuffisamment expressives	570
19.7.3 Élimination des redondances structurelles	573
19.7.4 Élimination des redondances internes	575
19.7.5 Autres critères de qualité	578
19.8 Évaluation d'un schéma	578
19.8.1 Paraphrase du schéma	579
19.8.2 Prototypage de la base de données	580
19.8.3 Inférence structurelle	580
19.8.4 Les heuristiques d'erreurs	581
19.8.5 Remarque	582
19.9 Intégration de schémas	582
19.9.1 Scénarios d'intégration	583
19.9.2 Principes de l'intégration binaire de schémas	584
19.9.3 Intégration basée sur les correspondances	584
19.9.4 Repérage des objets en correspondance	589

19.9.5 Unification : le cas des correspondances hétérogènes	590
19.9.6 Normalisation et résolution de conflits sémantiques	592
19.9.7 Compléments	593
19.9.8 Étude de cas	594
19.10 Analyse conceptuelle à l'aide d'un atelier	594
19.10.1 Transformation de schémas	595
19.10.2 Analyse de schémas	596
19.10.3 Intégration de schémas	597
19.11 Que retenir ?	597
19.12 Pour en savoir plus	599
19.13 Exercices	600
 CHAPITRE 20 • CONCEPTION LOGIQUE D'UNE BASE DE DONNÉES RELATIONNELLE	 605
20.1 Introduction	605
20.2 Le modèle logique relationnel	607
20.2.1 Le modèle relationnel de base	607
20.2.2 Le modèle relationnel enrichi	608
20.3 Représentation des attributs	610
20.3.1 Représentation d'un attribut simple	610
20.3.2 Représentation d'un attribut composé	610
20.3.3 Représentation d'un attribut multivalué	613
20.3.4 Représentation d'un attribut composé multivalué	614
20.3.5 Traduction des types et domaines de valeurs	615
20.3.6 Compléments	616
20.4 Retour aux types d'associations fonctionnels	617
20.5 Représentation des types d'associations complexes	618
20.5.1 Types d'associations n-aires, avec attributs et/ou contraintes	618
20.5.2 Représentation d'un type d'associations à rôle polymorphique	618
20.5.3 Clé étrangère et identifiant absent	620
20.6 Représentation des relations is-a	620
20.6.1 Les trois techniques de base	621
20.6.2 Représentation de relations is-a par matérialisation	622
20.6.3 Représentation de relations is-a par héritage descendant	626
20.6.4 Représentation de relations is-a par héritage ascendant	627
20.6.5 Évaluation comparative	628
20.6.6 Hiérarchies multiples et clés étrangères liées	629
20.7 Compléments	630
20.7.1 Les identifiants facultatifs	630
20.7.2 Les identifiants primaires complexes	631

20.7.3 Traduction des contraintes conceptuelles	632
20.7.4 Noms des objets	632
20.7.5 Delete et update modes des clés étrangères	632
20.7.6 Au sujet des rôles de cardinalité [1-N]	633
20.8 Traduction d'un schéma conceptuel	634
20.8.1 Choix des représentations privilégiées	634
20.8.2 Construction du plan de transformation	636
20.8.3 Limitations d'un plan de transformation	636
20.8.4 Validation d'un plan de transformation	637
20.9 Définition des vues	638
20.10 Conception logique : un exemple	638
20.11 Exploitation des constructions SQL3	640
20.11.1 Le modèle relationnel objet	642
20.11.2 Représentation des attributs complexes	642
20.11.3 Représentation des hiérarchies is-a de types d'entités	642
20.11.4 Conception logique relationnelle objet : un exemple	643
20.11.5 Extension de la méthode	643
20.12 Produire un schéma logique à l'aide d'un outil	644
20.12.1 Transformation de schémas	644
20.12.2 Analyse de schémas	646
20.13 Que retenir ?	646
20.14 Pour en savoir plus	647
20.15 Exercices	648
 CHAPITRE 21 • CONCEPTION PHYSIQUE D'UNE BASE DE DONNÉES RELATIONNELLE	 653
21.1 Introduction	653
21.2 Notion d'efficacité d'une base de données	655
21.3 Conception physique d'une base de données en 16 règles	656
21.4 Détermination des index	656
21.4.1 Index induits par les structures logiques	656
21.4.2 Index induits par l'utilisation des données	657
21.4.3 Index à écarter	658
21.4.4 Maintenance du jeu d'index	660
21.5 Technologie des index	660
21.6 Sélection des espaces de stockage	661
21.7 Optimisation avancée	661
21.7.1 Restructurations sans redondance	662
21.7.2 Redondances structurelles	664
21.7.3 Dénormalisation	665

21.8	Estimation des volumes et des temps de traitement	666
21.9	Complément : modèle de choix d'un index	666
21.10	Que retenir ?	667
21.11	Pour en savoir plus	668
21.12	Exercices	668
CHAPITRE 22 • PRODUCTION DU CODE D'UNE BASE DE DONNÉES		671
22.1	Introduction	671
22.2	Expression déclarative des structures	672
22.3	Les concepts SQL d'expression de contraintes	673
22.4	Les comportements non standard	675
22.5	Les contraintes d'intégrité de base	677
22.6	Les contraintes d'intégrité additionnelles	678
22.7	Les contraintes dérivées du modèle Entité-association	681
22.8	Les contraintes d'intégrité dynamiques	682
22.9	Traduction des relations is-a	683
22.9.1	Gestion à partir des tables de base	683
22.9.2	Gestion via des vues modifiables	685
22.10	Traduction des vues	688
22.11	Que retenir ?	689
22.12	Génération du code à l'aide d'un outil	689
22.13	Exercices	690
CHAPITRE 23 • RÉTRO-INGÉNIERIE D'UNE BASE DE DONNÉES		693
23.1	Introduction	693
23.2	Méthode générale de rétro-ingénierie	697
23.3	Extraction physique	698
23.4	Reconstruction du schéma logique	699
23.4.1	Les sources et leurs techniques d'analyse	699
23.4.2	Recherche des constructions implicites	703
23.5	Conceptualisation du schéma logique	705
23.6	Que retenir ?	709
23.7	Pour en savoir plus	710
23.8	Rétro-ingénierie à l'aide d'un outil	710
CHAPITRE 24 • ÉTUDE DE CAS		711
24.1	Les animaux du zoo	711

24.1.1 Énoncé	711
24.1.2 Construction du schéma conceptuel	712
24.1.3 Production du schéma de tables et du code SQL	713
24.1.4 Exploitation	715
24.2 Autres études de cas	718
 BIBLIOGRAPHIE	 719
 INDEX	 727

ANNEXES

Les annexes sont disponibles en ligne sur le site de l'ouvrage, à l'adresse
<http://www.info.fundp.ac.be/libd> > Ouvrages
ou sur le site www.dunod.com sur la page dédiée à cet ouvrage

ANNEXE 1 • MOTIVATION ET INTRODUCTION

ANNEXE 2 • CONCEPTS DES BASES DE DONNÉES

ANNEXE 3 • MODÈLE RELATIONNEL ET NORMALISATION

ANNEXE 4 • ÉLÉMENTS DE TECHNOLOGIE DES BASES DE DONNÉES

ANNEXE 5 • LES SYSTÈMES DE GESTION DE BASES DE DONNÉES

ANNEXE 6 • LE LANGAGE SQL DDL

ANNEXE 7 • LE LANGAGE SQL DML (1)

ANNEXE 8 • LE LANGAGE SQL DML (2)

ANNEXE 9 • LE LANGAGE SQL AVANCÉ

ANNEXE 10 • SQL ET L'ANALYSE DE DONNÉES

ANNEXE 11 • LES BASES DE DONNÉES NON RELATIONNELLES

ANNEXE 13 • LE MODÈLE ENTITÉ-ASSOCIATION DE BASE

ANNEXE 14 • ÉLABORATION D'UN SCHÉMA CONCEPTUEL

ANNEXE 15 • PRODUCTION DU SCHÉMA DE LA BASE DE DONNÉES

ANNEXE 16 • MÉTHODOLOGIE DES BASES DE DONNÉES

ANNEXE 17 • LE MODÈLE ENTITÉ-ASSOCIATION ÉTENDU

ANNEXE 18 • LES DIAGRAMMES DE CLASSES UML

ANNEXE 19 • ANALYSE CONCEPTUELLE DU DOMAINE D'APPLICATION

ANNEXE 20 • CONCEPTION LOGIQUE D'UNE BASE DE DONNÉES RELATIONNELLE

ANNEXE 21 • CONCEPTION PHYSIQUE D'UNE BASE DE DONNÉES RELATIONNELLE

ANNEXE 22 • PRODUCTION DU CODE D'UNE BASE DE DONNÉES

ANNEXE 23 • RÉTRO-INGÉNIERIE D'UNE BASE DE DONNÉES

ANNEXE 24 • ÉTUDES DE CAS

ANNEXE 25 • SQL, LES ENSEMBLES ET LA LOGIQUE

ANNEXE 26 • TRANSFORMATIONS DE SCHÉMAS

ANNEXE 27 • CONCEPTION LOGIQUE D'UNE BASE DE DONNÉES RELATIONNELLE OBJET

ANNEXE 28 • APPLICATIONS AVANCÉES DES BASES DE DONNÉES

ANNEXE 29 • INTRODUCTION À SQLfast

ANNEXE 30 • CONVENTIONS GRAPHIQUES ER

Avant-propos

Cet ouvrage est le fruit de plus de trente ans d'enseignement, de recherche et de pratique, tant à l'université que dans les entreprises et les administrations. J'y ai rassemblé et structuré les concepts, les techniques et les méthodes du domaine des bases de données qui me sont apparus les plus *utiles* à l'informaticien, depuis l'étudiant jusqu'au praticien en passant par l'utilisateur motivé.

Le lecteur ne trouvera ici que ce qu'il est susceptible de rencontrer dans sa pratique.¹ Même les sections un peu plus théoriques ont leur utilité, soit pour résoudre un problème concret de manière rigoureuse, soit pour mieux comprendre l'un ou l'autre concept délicat ou controversé. En revanche, le lecteur intéressé par les détails techniques propres à un SGBD particulier trouvera ceux-ci dans l'abondante littérature qui leur est consacrée, tant sur le web que dans les librairies.

J'ai donc résisté à la tentation de l'*encyclopédisme*, qui voudrait que je traite ici, comme il est souvent d'usage dans un ouvrage généraliste, de thèmes tels que les SGBD historiques (tels CODASYL DBTG et IMS), les dépendances multivaluées, la quatrième forme normale, le calcul relationnel, Datalog, l'optimisation des requêtes, les bases de données orientées objet natives, déductives, XML ou multimédias. Malgré leur incontestable intérêt je les ai écartés de la table des matières dans la mesure où ils restent marginaux par rapport au cœur de métier de l'informaticien. De même, les chapitres consacrés à SQL sacrifient la complétude syntaxique au profit des formes les plus utiles orientées vers les problèmes concrets.

Avant de décrire plus en détail les objectifs et le mode d'emploi de l'ouvrage, je voudrais remercier tous ceux qui de loin ou de près, peu ou prou, ont contribué à son élaboration. Certains ont été, ou sont encore, des collaborateurs de recherche. D'autres sont des collègues avec lesquels j'ai à l'occasion des discussions sur des points particuliers techniques ou théoriques.

1. À une exception près, que le lecteur est invité à identifier !

Je pense en particulier, dans le désordre, à Véronique Goemans, Yves Delvaux, Lysiane Goffaux, Jean-Pierre Thiry, Alain Delcourt, Benoît Vanhoute, Carine Charlot, Bertrand Mignon, Catherine Tonneau, Muriel Chandelon, Michel Joris, Mario Cadelli, Bernard Decuyper, Olivier Marchand, Jean Henrard, Jean-Marc Hick, Vincent Englebert, Didier Roland, Alain Gofflot, Pierre Delvaux, Anne-France Brogneaux, Arnaud Deflorenne, Stéphane Bodart, Majid Chougrani, Philippe Thiran, Virginie Detienne, Denis Zampunieris, Christine Delcroix, Aurore François, Fabrice Estiévenart, Frédéric Vesentini, Jean-Roch Meurisse, Julien Vilz, Ravi Ramdoyal, Anthony Cleve, Jonathan Lemaitre, Karin Becker, Yannis Tzitzikas, Rokia Bendaoud, Cédric Libert, Gilles Perrouin, François Bodart, Patrick Heymans, Wim Vanhoof et Yves Pigneur. Une pensée toute particulière à Anne-France, Vincent et Anthony, pour leur aide dans le décryptage de SQL lorsqu'on tente de le pousser dans ses derniers retranchements.

Je remercie tous les étudiants auxquels j'ai enseigné dans les facultés d'informatique, d'économie, de gestion et de droit, cette discipline particulièrement riche et passionnante qu'est celle des bases de données. Leurs questions, leurs suggestions et leurs difficultés sont à l'origine d'une partie significative de cet ouvrage. Celui-ci leur est tout naturellement dédié.

Merci enfin aux maisons d'édition Masson, InterEditions et Dunod pour leur aide et la confiance qu'elles m'ont accordée au fil des années.

Objectifs et structure de l'ouvrage

L'ouvrage a pour objectif d'offrir au lecteur un texte d'introduction aux bases de données à la fois pratique, moderne et détaillé. Les concepts et les problèmes qu'il aborde sont ceux que le lecteur, quel que soit son niveau de départ et quelles que soient ses attentes (il s'adresse aussi bien à l'*étudiant* qu'au *praticien*) rencontrera sur le terrain. Sa structure générale suit une approche rationnelle : pour apprendre à *utiliser* les bases de données il faut au préalable en *comprendre* les concepts et pour apprendre à *construire* des bases de données il faut savoir les *utiliser*. L'ouvrage est par conséquent structuré en trois parties :

- La première, intitulée **Concepts des bases de données**, est constituée des quatre chapitres qui décrivent les concepts fondamentaux des bases de données : les structures de données, leur implémentation, les systèmes de gestion de bases de données et, pour les lecteurs plus curieux, des éléments du modèle relationnel et de la normalisation. Cette partie est complétée d'un chapitre sur les bases de données non relationnelles, y compris les SGBD dits NoSQL, avec une extension vers les *chaînes de blocs*.
- La deuxième partie, **Utilisation des bases de données**, décrit les différents aspects du langage SQL, depuis les formes et fonctions élémentaires jusqu'aux fonctions avancées du modèle relationnel objet. Elle inclut un nouveau chapitre consacré au domaine de l'analyse des données. Cette partie comprend cinq chapitres.
- La troisième partie, **Développement des bases de données**, décrit de manière détaillée les techniques et les méthodes de construction de bases de données relationnelles et relationnelles objet. On y trouve également une introduction à la

rétro-ingénierie des bases de données. Cette partie constituée de douze chapitres est suivie d'une étude de cas complètement résolue et implémentée.

Chaque chapitre est prolongé par une annexe qui propose des exercices additionnels ainsi que des matériaux complémentaires. D'autres annexes transversales contiennent des exposés de mise à niveau, des applications et des manuels pour les outils utilisés dans cet ouvrage. Ces annexes et ces outils, ainsi que des matériaux complémentaires, destinés à l'étudiant, à l'enseignant et au praticien, sont disponibles en ligne sur le site de l'ouvrage.

Les options de cet ouvrage

L'ouvrage a été structuré et rédigé sur la base de quelques principes suggérés par l'expérience et l'étude des besoins actuels en matière de bases de données.

Une approche pratique et pédagogique. Destiné à un public diversifié et évolutif, l'ouvrage adopte une démarche progressive et propose plusieurs parcours de lecture. Le contenu de l'ouvrage est articulé autour de problèmes et de questions effectivement rencontrés en pratique. Les aspects théoriques sont proposés dans la mesure où ils permettent de résoudre aisément et efficacement ces problèmes. Le texte inclut de nombreux exemples, exercices et études de cas, accompagnés de leurs solutions. Les concepts sont illustrés par des outils disponibles gratuitement (**SQLfast** et **DB-Main**) que le lecteur est invité à utiliser. La plupart des chapitres se clôturent par une synthèse (*Que retenir ?*), par des pistes d'approfondissement et un état de l'art (*Pour en savoir plus*) et par un jeu d'exercices corrigés.

Un contenu et une approche modernes. L'ouvrage est destiné aux lecteurs confrontés aux problèmes que posent les systèmes d'information et les technologies d'aujourd'hui. Il inclut des résultats récents issus de la recherche et de l'industrie. Ainsi, des méthodes pratiques relatives au *calcul des identifiants*, à la *normalisation de schémas conceptuels*, à la construction de *bases de données relationnelles objet*, aux *bases de données NoSQL*, aux *chaînes de blocs*, à l'*analyse des données* et à la *rétro-ingénierie* ont-elles été incluses dans cet ouvrage. Les méthodes d'ingénierie qu'il décrit sont basées sur des approches transformationnelles et peuvent être mises en œuvre à l'aide d'outils (les AGL).

Une approche rigoureuse. L'ouvrage ne se contente pas de décrire les concepts et les méthodes à la manière d'une encyclopédie. Il construit progressivement un édifice cohérent et raisonné. Les techniques et les méthodes de résolution de problèmes sont présentées, non comme des recettes prêtes à l'emploi, mais comme le résultat naturel de raisonnements objectifs et rigoureux. Le lecteur peut ainsi comprendre la rationalité de ces techniques et méthodes et en appréhender les limites. Il peut dès lors se les approprier et les adapter à ses propres besoins. Certaines parties délicates de l'exposé sont soutenues par des éléments plus théoriques qui en garantissent la validité. Ainsi en est-il de la théorie relationnelle et de la normalisation (chapitre 3).

Une approche critique. Les concepts, les technologies et les méthodes ne sont pas livrées comme allant de soi. Leur histoire, leurs points forts et leurs faiblesses sont largement discutés. Pour citer quelques exemples, les normes SQL2 et SQL3, les valeurs null, le modèle relationnel objet, les modèles NoSQL, le modèle Entité-association, les diagrammes de classes d'UML et les options choisies par les implémenteurs de SGBD sont analysés et critiqués, souvent de manière détaillée.

Publics cibles et parcours personnalisés

Cet ouvrage est destiné à la fois aux *étudiants et enseignants*, et aux *professionnels* de l'informatique, auxquels il propose deux parcours différents :

- **Parcours introductif.** Le lecteur ou l'étudiant débutant se contentera dans un premier temps de la suite des neuf chapitres suivants :
 - *Concepts* : chapitres 1 et 2.
 - *Utilisation* : chapitres 6, 7 et 8 consacrés au langage SQL.
 - *Développement* : chapitres 12, 13, 14 et 15. Étude de cas : chapitre 24.

Ces chapitres constituent un parcours d'initiation autonome couvrant les aspects fondamentaux du domaine (figure 1). Il correspond à un module d'introduction aux bases de données de 30 heures, travaux pratiques non compris.

- **Parcours complet.** Il sera proposé au lecteur ou à l'étudiant à la recherche d'un exposé plus approfondi et plus technique du domaine. Les trois parties de l'ouvrage sont constituées comme suit :
 - *Concepts* : chapitres 2, 3, 4, 5
 - *Utilisation* : chapitres 6, 7, 8, 9, 10 et 11
 - *Développement* : chapitres 16, 17, 18, 19, 20, 21 et 22. Étude de cas : chapitre 24. Le chapitre 23 sur la *rétro-ingénierie* intéressera surtout le praticien.

Ces chapitres constituent un parcours de lecture complet couvrant en détail l'ensemble du domaine des bases de données (figure 2). Il correspond à un module approfondi de 60 heures, travaux pratiques non compris, ou encore à un module de 45 heures pour l'étudiant qui a suivi le parcours introductif.

En outre, il est possible de définir des parcours plus spécifiques et d'exploiter des matériaux additionnels.

- L'annexe 25 (*SQL, les ensembles et la logique*) constitue une mise à niveau utile pour aborder les chapitres 3, 17 et 19. De même, l'annexe 26 détaille l'approche transformationnelle utilisée informellement dans les chapitres 14, 15, 17, 19 et 20. L'annexe 29 est une introduction à SQLfast, outil dont nous parlerons plus loin.
- Les exercices et les applications sont essentiels pour maîtriser les concepts et les techniques étudiés. On les trouvera à la fin de chaque chapitre ainsi que dans leurs annexes respectives. En outre, les annexes 24 et 28 présentent des applications et des études de cas de complexité variable.
- Sur le plan de la méthodologie (troisième partie), les chapitres 16, 17, 19 et 23 (*rétro-ingénierie*) apportent aux *concepteurs, consultants et administrateurs de*

données une spécialisation en modélisation des données. Celle-ci pourra être complétée par l'étude critique d'UML (chapitre 18).

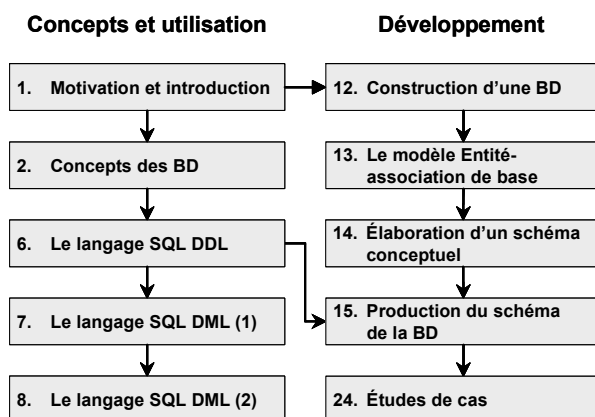


Figure 1 - Parcours de lecture introductif

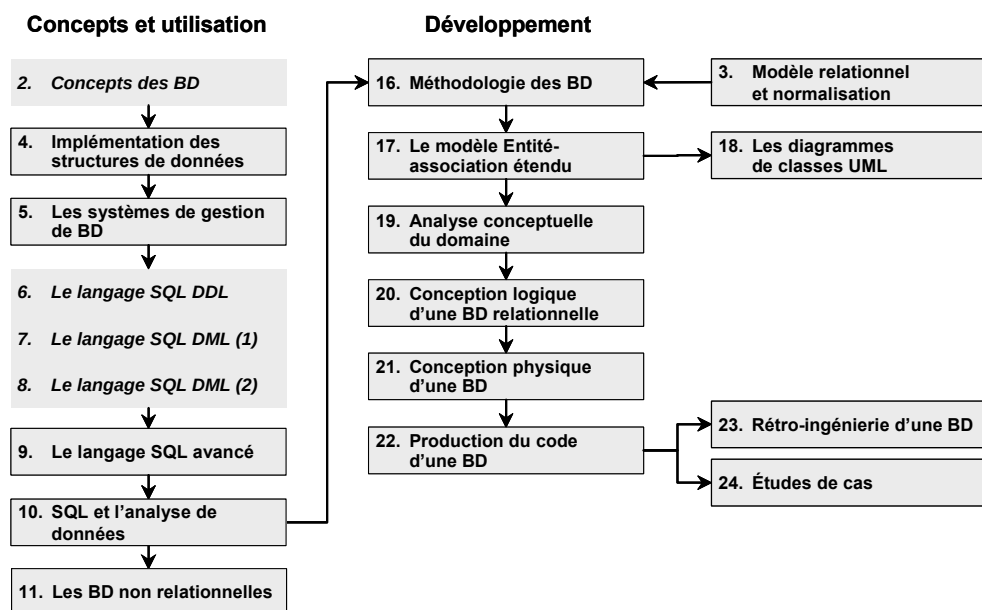


Figure 2 - Parcours de lecture complet

Cet ouvrage est accompagné et complété d'un site web dont le contenu est en libre accès. Celui-ci met en ligne les annexes 1 à 30 et des matériaux complémentaires : tutoriels, logiciels recommandés dans l'ouvrage, matériaux didactiques, documents d'approfondissement et errata (que le lecteur attentif consultera régulièrement). Ces matériaux sont disponibles à l'adresse <http://www.info.fundp.ac.be/libd> (onglet *Books/Ouvrages*) ou sur le site de Dunod sur la page dédiée à l'ouvrage.

Au sujet de la 5^e édition

Par rapport à l'édition précédente, cette cinquième édition inclut un nouveau chapitre consacré aux fonctions du langage SQL dédiées à l'analyse de données (chapitre 10). Elle comprend également de nombreuses mises à jour, corrections et améliorations de présentation ainsi que de nouveaux matériaux conformes à l'état de l'art en 2022. Ces ajouts et modifications sont répartis entre les chapitres de cet ouvrage et leurs annexes.

Les outils

L'ouvrage est accompagné de deux outils que le lecteur voudra bien considérer comme composants essentiels de la démarche pédagogique proposée.

- **SQLfast** est un environnement graphique convivial à large spectre destiné à la fois à l'apprentissage des bases de données et du langage SQL et au développement rapide de petites applications interactives de bases de données. Rédigé en Python et construit sur le SGBD SQLite, SQLfast ne nécessite aucune installation. La version 5, recommandée comme complément à cet ouvrage, comprend :
 - Une interface graphique permettant de visualiser, sélectionner et modifier le contenu d'une base de données sans connaissances en programmation
 - Un interpréteur du langage SQLfast, une extension de SQL, permettant de rédiger et exécuter des requêtes SQL et des petits scripts de manière interactive.
 - Des tutoriels qui reprennent toutes les requêtes et les exercices de chaque chapitre tout en les augmentant d'exercices supplémentaires. En particulier, le tutoriel du chapitre 10, consacré à l'analyse de données, comporte, outre une synthèse des concepts de base, plus de 50 exemples et exercices d'application, tous exécutables *in situ*.
- **DB-MAIN** est un atelier logiciel d'ingénierie de bases de données. Il est recommandé en support des chapitres de la troisième partie de l'ouvrage (*Développement des bases de données*).

Ces deux outils sont disponibles gratuitement sur le site de l'ouvrage.

Jean-Luc Hainaut
Namur, le 31 janvier 2022

Chapitre 1

Motivation et introduction

Ce chapitre introduit de manière intuitive le lecteur au concept de *donnée*. Il montre que le contenu des documents que nous produisons et utilisons quotidiennement sont naturellement structurés d'une manière qui suggère leur introduction dans un ordinateur. Ce processus d'analyse conduit à la notion de *base de données*. On identifie les principales questions qui se posent en pratique dans le monde des bases de données et on rappelle quelques étapes de leur histoire.

1.1 DES DONNÉES ET DES HOMMES

Il n'est pas d'interactions avec nos collègues, avec des commerçants ou des administrations lors desquelles nous n'échangeons pas des **informations** sous forme de **données**. La consultation du site d'un fournisseur d'équipement multimédia, le suivi de notre commande de matériel informatique, la déclaration de naissance du petit dernier, la lecture inquiète des notes obtenues en fin d'année scolaire (les nôtres ou celles de notre fainéant de fiston), nos résultats dans Fortnite, la consultation des films proposés ce samedi dans notre quartier, une visite chez notre médecin, nos communications téléphoniques, le téléchargement de l'épisode 3 de la saison 2 ou l'usage d'un GPS n'en sont que quelques exemples parmi les plus évidents.

Où vont les données que nous fournissons et d'où viennent celles qui nous sont livrées ? Elles sont évidemment stockées dans un ordinateur. Dès que ces données présentent un certain degré de complexité, qu'elles sont soumises à des contraintes de qualité, qu'elles doivent résister à des incidents ou à des attaques, qu'elles sont partagées par plusieurs usagers ou qu'elles sont utilisées dans plusieurs départe-

ments pour des fonctions différentes, il est hautement probable qu'elles sont stockées dans une **base de données**.

C'est désormais une évidence, nous sommes quotidiennement, en tant que simples usagers ou dans nos activités professionnelles, et souvent sans en être conscients, en relation étroite avec des milliers de bases de données que nous alimentons et que nous consultons.

Dans les entreprises, ces données, dites permanentes, constituent le matériau de base à partir duquel vont s'élaborer la plupart des applications informatiques quelle qu'en soit la taille. À tel point d'ailleurs que certaines de celles-ci se réduisent pratiquement à la gestion et la consultation de données : répertoire d'adresses, références bibliographiques ou catalogue de pièces de rechange. Ces données sont rangées dans ce qu'on appelle un **fichier** et sont structurées en **enregistrements**. Chaque correspondant du répertoire d'adresses y est décrit par un enregistrement qui reprend son nom et son prénom, ainsi que ses coordonnées postales et électroniques.

Les données nécessaires aux applications informatiques sont le plus généralement réparties dans plusieurs fichiers en fonction des objets qu'elles décrivent : *fichier des clients*, *fichier des produits*, *fichier des commandes*, *fichier des factures*, etc. Il existe entre les fichiers des liens qui sont à l'image des relations entre les objets décrits. C'est via ces liens entre les données qu'on indique les *commandes qui ont été émises par tel client* ou les *produits qui sont référencés par telle commande*.

Les données nécessaires à une application pourraient être utiles à d'autres applications, voire même à d'autres utilisateurs. Ces données constituent alors une ressource commune qu'on appelle une **base de données**. Gérer de telles données n'est pas une mince affaire. Garantir leur qualité (retrouve-t-on ce qu'on y a enregistré ?), leur cohérence (le client de chaque commande est-il répertorié ?), les protéger en cas d'incident, permettre à plusieurs utilisateurs d'y accéder simultanément sans conflits, tout en contrôlant strictement l'accès aux données confidentielles, offrir de bonnes performances d'accès à toutes les applications, en particulier celles qui sont interactives, sont des fonctions qui réclament des logiciels puissants et complexes, les **systèmes de gestion de bases de données**, ou **SGBD**.

Ces logiciels constituent un des outils fondamentaux de développement des grosses applications informatiques, mais aussi d'applications plus légères ou de sites web. Ils existent sur toutes les plates-formes, depuis les gros ordinateurs (*mainframes* ou serveurs de grande puissance) jusqu'au PC portable et au *smartphone*, à la carte à puce.

1.2 UNE PREMIÈRE APPROCHE DES DONNÉES

Sur votre bureau, là à gauche, la pile des bons de commande de la semaine que vos clients vous ont envoyés par courrier postal (figure 1.1). En face de vous l'ordinateur que vous venez d'acquérir grâce aux premiers bénéfices de votre petite entreprise de matériaux de construction. Votre intention est d'introduire celle-là (la pile) dans celui-ci (l'ordinateur).

Commande N° : 30188		Date : 2/1/2009		
Numéro client	B512			
Nom	GILLET			
Adresse	14, r. de l'Eté			
Localité	Toulouse			
N° PRODUIT	LIBELLE PRODUIT	PRIX	QUANTITE	SOUS-TOTAL
CS464	CHEV. SAPIN 400x6x4	220	180	39600
PA45	POINTE ACIER 45 (20K)	105	22	2310
PA60	POINTE ACIER 60 (10K)	95	70	6650
PH222	PL. HETRE 200x20x2	230	92	21160
TOTAL COMMANDE				69720

Figure 1.1 - La pile de bons de commande de la semaine

Vous pensez dans un premier temps à **scanner** ces documents, mais vous réalisez bien vite que ce procédé ne vous apporterait pas grand-chose d'utile, le résultat ne se prêtant pas à des manipulations ultérieures, telles que l'édition de factures, la gestion des stocks ou votre comptabilité.

L'idée de les copier dans un **document Word** ne vous semble guère plus sérieuse : mise en page laborieuse et exploitation des données très limitée.

Vous pourriez introduire leur contenu dans une **feuille Excel**, mais la structure des données vous paraît trop complexe à traduire dans un simple tableau. Et puis comment cette solution supportera-t-elle la croissance rapide que vous prévoyez pour votre entreprise ?

Vous n'y échapperez pas : il faudra vous résoudre à introduire ces données dans une **base de données** ! À ce stade de la lecture de ce chapitre, vous savez déjà que les données s'organisent en *fichiers*, constitués chacun d'*enregistrements*.

Prenons le premier document pour en étudier la structure (figure 1.2). Nous y repérons immédiatement un bloc de données relatives au client : *Numéro client*, *Nom*, *Adresse* et *Localité*. La commande elle-même est caractérisée par un *Numéro*, une *Date* (en haut) et un *Total Commande* (en bas). Le tableau de la moitié inférieure regroupe les détails de la commande : pour chacun on y trouve le *Numéro du produit*, son *Libellé*, son *Prix* (à l'unité), la *Quantité* (commandée) et le *Sous-total*. Nous avons identifié dans le bon de commande trois catégories de données, relatives respectivement aux *clients*, aux *commandes* et aux *détails* de commande.

Nous décidons logiquement de ranger ces données dans trois fichiers qui accueilleront respectivement les données sur le *client*, les données sur la *commande* et les données sur les *détails* (figure 1.3). Nous procéderons à l'encodage des autres bons de commande selon cette même distribution des données.

Commande N° : 30188		Date : 2/1/2009		
Numéro client : B512 Nom : GILLET Adresse : 14, r. de l'Eté Localité : Toulouse				
N° PRODUIT	LIBELLE PRODUIT	PRIX	QUANTITE	SOUS-TOTAL
CS464	CHEV. SAPIN 400x6x4	220	180	39600
PA45	POINTE ACIER 45 (20K)	105	22	2310
PA60	POINTE ACIER 60 (10K)	95	70	6650
PH222	PL. HETRE 200x20x2	230	92	21160
TOTAL COMMANDE				69720

données du client

données de la commande

données d'un détail

Figure 1.2 - Analyse des données du premier bon de commande

données de la commande			données du client			
NCOM	DATECOM	TOTAL-COMMANDE	NCLI	NOM	ADRESSE	LOCALITE
30188	2/1/2009	69720	B512	GILLET	14, r. de l'Eté	Toulouse

données des détails				
NPRO	LIBELLE	PRIX	QCOM	SOUS-TOTAL
CS464	CHEV. SAPIN 400x6x4	220	180	39600
PA45	POINTE ACIER 45 (20K)	105	22	2310
PA60	POINTE ACIER 60 (10K)	95	70	6650
PH222	PL. HETRE 200x20x2	230	92	21160

Figure 1.3 - Extraction des données du premier bon de commande

On avance bien sûr, mais à y regarder de plus près, le résultat n'est pas tout à fait satisfaisant :

1. D'une part, certaines données sont *calculées*, de sorte qu'il n'est pas absolument nécessaire de les stocker. Nous pouvons ainsi supprimer les champs SOUS-TOTAL (= PRIX x QCOM) et TOTAL-COMMANDE (= somme des valeurs de SOUS-TOTAL).
2. D'autre part, nous observons qu'il est impossible de reconstituer les documents initiaux. Comment en effet repérer le client d'une commande, puisque nous avons extrait et rangé ailleurs le fragment décrivant ce client ? De même, comment identifier la commande de laquelle nous avons extrait un détail ? Il est évident qu'il faut prévoir dans les fragments relatifs aux commandes et aux détails des *données de référence* nécessaires à ces liaisons.

Nous complétons donc la structure des données en introduisant dans le fichier des commandes un champ NCLI indiquant le numéro du client et dans le fichier des détails un champ NCOM spécifiant la commande de chaque détail (figure 1.4). Nous en profitons pour introduire également les données de la commande 30179.

Remarquons en passant que nous avons fait l'hypothèse qu'on pouvait *identifier* un client par son numéro NCLI et une commande par son numéro NCOM, ce qui est bien naturel. Pour les détails, nous réexaminerons cette question plus tard.

données des commandes			données des clients			
NCOM	NCLI	DATECOM	NCLI	NOM	ADRESSE	LOCALITE
30188	B512	2/1/2009	B512	GILLET	14,r. de l'Eté	Toulouse
30179	C400	22/12/2008	C400	FERARD	63,r. du Tertre	Poitiers

données des détails				
NCOM	NPRO	QCOM	LIBELLE	PRIX
30188	CS464	180	CHEV. SAPIN 400x6x4	220
30188	PA45	22	POINTE ACIER 45 (20K)	105
30188	PA60	70	POINTE ACIER 60 (10K)	95
30188	PH222	92	PL. HETRE 200x20x2	230
30179	CS262	60	CHEV. SAPIN 200x6x2	75
30179	PA60	20	POINTE ACIER 60 (10K)	95

Figure 1.4 - Fichiers de données améliorés

Le résultat semble prometteur, mais il reste un malaise : pour tous les détails qui référencent le même produit (ceux qui ont la même valeur du champ NPRO), on reprend dans le fichier des détails le LIBELLE et le PRIX de ce produit, créant ainsi d'importantes redondances non seulement inutiles mais dangereuses. Quelle garantie en effet avons-vous que tous les détails d'un même produit indiqueront exactement le même *Libellé* et le même *Prix* ? Il serait plus intéressant de construire un quatrième fichier destiné à recueillir les données spécifiques aux produits. Nous retirons donc celles-ci du fichier des détails (en y laissant cependant une copie du champ NPRO pour ne pas retomber dans le piège précédent) et nous les rangeons dans ce nouveau fichier. C'est ce qui est proposé à la figure 1.5.

données des commandes			données des clients			
NCOM	NCLI	DATECOM	NCLI	NOM	ADRESSE	LOCALITE
30188	B512	2/1/2009	B512	GILLET	14,r. de l'Eté	Toulouse
30179	C400	22/12/2008	C400	FERARD	63,r. du Tertre	Poitiers

données des détails			données des produits		
NCOM	NPRO	QCOM	NPRO	LIBELLE	PRIX
30188	CS464	180	CS464	CHEV. SAPIN 400x6x4	220
30188	PA45	22	PA45	POINTE ACIER 45 (20K)	105
30188	PA60	70	PA60	POINTE ACIER 60 (10K)	95
30188	PH222	92	PH222	PL. HETRE 200x20x2	230
30179	CS262	60	CS262	CHEV. SAPIN 200x6x2	75
30179	PA60	20			

Figure 1.5 - Les données de bons de commande distribuées dans quatre fichiers

1.3 UTILISER UNE BASE DE DONNÉES

Nous venons de construire une base de données en bonne et due forme, respectant les bonnes pratiques en la matière. La figure 2.4, que le lecteur est invité à consulter, montre ce qu'on obtiendrait après avoir enregistré les données de quelques commandes supplémentaires. Ainsi mises en forme, ces données permettent de répondre simplement à une infinie variété de questions. C'est à l'aide du langage SQL (*Structured Query Language*) que nous étudierons dans cet ouvrage, que nous pourrons interagir avec cette base de données, c'est-à-dire créer, consulter, modifier et supprimer les données. Par exemple, l'extraction des données sur les *clients de Toulouse* s'obtiendra par l'instruction suivante, dont la compréhension à ce stade ne pose pas de problèmes :

```
select NCLI, NOM, ADRESSE
from   CLIENT
where  LOCALITE = 'Toulouse'
```

On y spécifie les données désirées (NCLI, NOM, ADRESSE), la table dont elles sont extraites (CLIENT) et la condition que doivent vérifier les lignes qui nous intéressent (LOCALITE = 'Toulouse'). La recherche les *commandes des clients de Toulouse* est un peu plus délicate mais reste simple (elle sera étudiée dans la section 7.6)¹ :

```
select NCOM
from   COMMANDE
where  NCLI in (select NCLI
                from   CLIENT where LOCALITE = 'Toulouse')
```

Ces données vont pouvoir servir à d'innombrables applications, y compris certaines auxquelles nous ne pensons pas aujourd'hui. Citons quelques exemples.

- Vérifier les commandes du jour : le client est-il déjà enregistré ? Son adresse n'a-t-elle pas changé ? Les produits référencés existent-ils ?
- Produire les factures.
- Calculer votre chiffre d'affaires hebdomadaire, mensuel, annuel.
- Préparer votre déclaration de revenu.
- Étudier la répartition géographique de votre chiffre d'affaires.
- Identifier les produits qui se vendent le mieux.
- Produire le chiffre d'affaires de chaque produit par localité et par quadrimestre.

Avec un peu d'imagination et en acceptant d'ajouter quelques données indispensables (la quantité en stock de chaque produit par exemple), voire une ou deux nouvelles tables, vous pourriez ajouter à cette liste quelques dizaines d'applications supplémentaires toutes utiles au fonctionnement et à la gestion de votre entreprise.

1. Un mot d'explication cependant. Cette requête demande d'extraire (**select**) les valeurs de NCOM des lignes de la table COMMANDE dont la valeur de NCLI apparaît dans un ensemble défini comme suit : l'ensemble des valeurs de NCLI extraites des lignes de CLIENT dont la colonne LOCALITE a la valeur 'Toulouse'.