

Stručný obsah

Poděkování	15
O autorovi	15
O odborných korektorech	15
Úvod	16
ČÁST I	
Příprava	21
1 Přehled datových skladů a systémů business intelligence	23
2 Data v organizaci	55
3 Argumenty pro implementaci	81
4 Strategie datového skladu a business intelligence	97
5 Projektové zdroje: role a přehledy	123
6 Přehled dokumentů	135
ČÁST II	
Součásti	141
7 Business intelligence: datové trhy a použití	143
8 Podnikové datové modely	185
9 Architektura datového skladu: Komponenty	217
10 ETL a kvalita dat	241
11 Plánování projektu a metodika	271
ČÁST III	
Implementace	285
12 Pracovní scénáře	287
13 Správa dat	323
14 Analýza po dokončení projektu	333

Obsah

Poděkování	15
O autorovi	15
O odborných korektorech	15
Úvod	16
O této knize	17
Zpětná vazba od čtenářů	19
Zdrojové kódy ke knize	19
Errata	19

ČÁST I

Příprava

KAPITOLA 1

Přehled datových skladů a systémů business intelligence	23
Přehled systémů business intelligence	23
Definice	24
Hodnota systému business intelligence	25
Prvky systému business intelligence	26
Faktory zajišťující úspěch systému business intelligence	27
Funkce systému BI	28
Uživatelská prezentace business intelligence	29
Nástroje a architektura business intelligence	33
Pokroky díky globalizaci	35
Přehled datových skladů	35
Definice	35
Systém datového skladu	37
Architektura datového skladu	38
Terminologie toku dat	39
Účel datového skladu	41
Strategie datových struktur	43
Datový sklad jako firma	44
Nejčastější dotazy	45
Jsou aktuální systémy dost dobré?	46
Jakou hodnotu má datový sklad?	47
Kolik to bude stát?	48
Jak dlouho to bude trvat?	49
Co zajistí úspěch projektu?	50

KAPITOLA 2

Data v organizaci	55
Podniková aktiva	55
Data v kontextu	55
Kvalita dat	57
Datová terminologie	59
Datové komponenty	60
Organizace dat	64
Vytvoření struktury dat	64
Datové modely	65
Datová architektura	70
Konkurenční výhoda	75
Vlastní vývoj nebo nákup datového modelu	75
Podnikové poradenství	78

KAPITOLA 3

Argumenty pro implementaci	81
Migrace platformy	82
Zajištění nepřerušného provozu	83
Zpětný překlad	83
Kvalita dat	84
Paralelní prostředí	85
Přidaná hodnota	85
Centralizace datového skladu	86
Fúze podniků	86
Slučování v rámci organizace	86
Centrální návrh a místní použití	86
Konsolidace datových trhů	87
Nová iniciativa	89
Nová iniciativa: dynamické vykazování	90
Důraz na budování	91
Datový zmatek	93
Důvody, proč datový sklad NEBUDOVAT	94
Nedostatečná kvalita dat	94
Nedostatek zájmu v podniku	95
Chybějící sponzor	95
Nejasný cíl	95
Aktuální systémy postačují	95
Nedostatek prostředků	95
Nestabilní prostředí	96

Příliš vysoké náklady	96
Špatné vedení	96

KAPITOLA 4

Strategie datového skladu a business intelligence	97
Strategie business intelligence	97
Podnikový cíl	97
Podnikové použití	98
Přehled architektury	99
Strategie datového skladu	101
Použití	101
Architektura datového skladu	102
Cíl a úspěch	105
Podnik nebo obor?	105
Zaměření na cíl	105
Úspěch: kdy bude projekt dokončen?	106
Kde začít?	107
Pro business intelligence	107
Pro datový sklad	108
Jak začít?	109
Pro business intelligence	109
Pro datový sklad	111
Fázování projektu	113
Vrací se otázka: Jak dlouho to bude trvat?	114
Důležitá hlediska	116
Typické příčiny selhání	116
Základní hodnoty	120

KAPITOLA 5

Projektové zdroje: role a přehledy	123
Hlavní postřehy	123
Projektové týmy	124
Odborné znalosti	124
Vedení	125
Projektový sponzor	126
Ředitel pro datový sklad	127
Struktura týmu	127
Podpora výkonného vedení	128
Správci dat	129
Základní zdroje	129

Pravidelné kontroly: audit postupu	133
Centrum kompetence	133

KAPITOLA 6

Přehled dokumentů	135
Projektová listina	135
Rozsah projektu	137
Pracovní deklaráce	137

ČÁST II

Součásti

KAPITOLA 7

Business intelligence: datové trhy a použití	143
Proč modelovat data?	144
Typy datových modelů	145
Návrh dat	148
Multidimenzionální analýza	148
Tabulky faktů	155
Typy faktů	156
Typy tabulek faktů	158
Zdroj metrik	160
Klíč tabulky faktů	161
Granularita tabulky faktů	161
Hustota tabulky faktů	161
Tabulka faktů bez faktů	162
Dimenze	163
Dimenze nebo metrika	163
Historie a data	164
Klíč tabulky dimenzí	167
Granularita dimenze	169
Zdroj a hodnota atributů dimenze	169
Typy dimenzí	171
Hierarchie a pomocné tabulky	179
Profilové tabulky	182
Počet dimenzí	182
Změny velikosti	184

KAPITOLA 8

Podnikové datové modely	185
Přehled datových modelů	185
Inmon a Kimball	188

Účel podnikového datového modelu	189
Výhody podnikového datového modelu	189
Datový model: kde začít	190
Úplný datový model shora dolů	190
Model předmětných oblastí	191
Koncepční model	193
Model entit a vztahů	194
Architektura sběrnice	195
Zakoupený datový model	197
Přehledy modelu	199
Datové komponenty	199
Normalizace datového modelu	199
Modely nadřazených a podřízených typů	205
Zaznamenávání historie v normalizovaném datovém modelu	208
Náhradní klíče	212
Logický a fyzický datový model	213
Referenční integrita: ano nebo ne	214
Jiné datové modely	215
Vstupní datový model	215
Realizační datový model	215
Závěrečné poznámky	216

KAPITOLA 9

Architektura datového skladu: Komponenty	217
Přehled architektury	217
Role architekta	218
Architekt řešení	218
Architekt datového skladu	218
Technický architekt	219
Datový architekt	219
Architekt ETL	219
Architekt BI	219
Souhrn	220
Vrstvy architektury	220
Architektura s jednou vrstvou	221
Klasická architektura se dvěma vrstvami	222
Pokročilá architektura se třemi vrstvami	223
Architektury datového skladu	224
Architektura samostatných datových trhů	224
Architektura sběrnice	225
Architektura centrálního úložiště	226
Federativní architektura	227

Komponenty (vrstvy)	227
Zdroje dat	228
Pořizování dat	228
Organizace dat	228
Distribuce dat	228
Výstup informací	229
Přístupy k implementaci	229
Návrh dat a tok dat	230
Logické a fyzické modely	231
Přístup shora dolů	232
Přístup zdola nahoru	234
Hybridní přístup	235
Akcelerátory	237
Vrstva získávání dat	238
Vrstva centralizovaných dat	238
Vrstva distribuce dat	239
Výkonnostní vrstva	239
Vrstva uživatelské prezentace	239
Metodologie	239
Hotové řešení	240
KAPITOLA 10	
ETL a kvalita dat	241
Architektura	242
Pořizování dat	243
Distribuce dat	244
Mapování ETL	246
Počáteční a inkrementální načítání	247
ETL vs. ELT vs. ETTL	247
Paralelní operace	249
Role ETL	249
Diagramy toku dat	251
Úložiště provozních dat	251
Zdrojové systémy	252
Zdroj chybí	252
Více zdrojů	254
Alternativní zdroje (vstupní soubory struktury)	254
Nestrukturovaná data	254
Profilování dat	255

Záznam dat	255
Více velkých souborů	256
Přepínací soubory	256
Strategie odolnosti proti selhání	257
Transformace a realizace	257
Příprava	258
Náhradní klíče	260
Referenční integrita	262
Agregace, profilování a souhrny	263
Tabulky kódů	263
Načítání	263
Ukládání a neukládání historie	264
Vkládání, aktualizace, jejich kombinace a odstranění	264
Informace o pořizování dat	264
Plánování načítání	265
Realizace u podnikového datového skladu a realizace u architektury sběrnice	265
Distribuce dat	265
Změna z 3NF na hvězdici	267
Kvalita dat	269
Nástroje ETL	270
 KAPITOLA 11	
Plánování projektu a metodika	271
Základy	273
Riziko: Vývoj ve fázích	274
Riziko: Kvalita dat	274
Riziko: Zdroje	275
Riziko: Náklady	276
Správa změn	276
Optimální postupy	276
Chyby	277
Metodika projektového plánu	278
Podnikové požadavky	279
Strategie a plán	280
Osnova řešení	282
Návrh	282
Sestavení	282
Nasazení	283
Použití	283

ČÁST III

Implementace

KAPITOLA 12

Pracovní scénáře	287
Šéfkuchař: Puště se do vaření!	289
Shora dolů (podnikové úložiště)	292
Terminologie	293
Centralizovaný datový model	294
Datová architektura	296
Zdroje	296
Datový model	297
Databáze	298
Získávání	299
Přehled řešení	300
Zdola nahoru (vykazování OLAP)	300
Konečný výsledek	302
Terminologie	308
Datová architektura	308
Správa potvrzených dimenzí	308
Zdroje	309
Přehled řešení	309
Hybridní přístup (normalizovaný návrh a OLAP)	311
První iniciativy	312
Datové modely	312
Datová architektura	313
Přehled řešení	315
Slučování	316
Plán činnosti	318
Chybí vstup: Vstupní soubory struktury	318
Integrační fáze 2	320
Správa změn	321
Celkový obrázek: Podniková informační architektura (EIA)	321

KAPITOLA 13

Správa dat	323
Co to je správa dat?	323
Definice	323
Důvody pro správu dat	324
Organizační struktura	325
Vedení a iniciativy	326

Správa dat: Hlavní hlediska	327
Zabezpečení a citlivost	327
Kvalita dat	328
Vlastnictví	329
Kontrola změn	329
Připravenost na správu dat	330
 KAPITOLA 14	
Analýza po dokončení projektu	333
Přehled	333
Hodnocení projektu	334
Další fáze	339
 Rejstřík	 341

*Tuto knihu věnuje své krásné a nadané ženě a drahému synovi.
Oba jste celý můj život a každý den děkuji Bohu za naši společnou
cestu. Nezáleží přitom na čísle stránky ani na počtu znaků.*

Poděkování

Zvláštní díky patří mým odborným korektorům Davidu Marcottemu a Kenovi Yu za jejich pomoc při analýze mých poznámek a za to, že mi poskytli své pohledy a komentáře. Díky vaší cenné zpětné vazbě jsem mohl celý projekt úspěšně dokončit v termínu. Mnohokrát děkuji!

David Marcotte je expert na globální maloobchod a analytiku, který shromáždil tolik znalostí jako málokterý člověk v oboru.

Ken Yu je technický odborník v oblasti datového návrhu a toku dat, který pracoval v mnoha sektorech a vyznačuje se praktickým a realistickým přístupem k datovému modelování a datové architektuře.

Speciální díky patří mé ženě Rakhee Laberge (MBA, B.E) za všechna její hodnocení, komentáře a názory.

O autorovi

Robert (Bob) Laberge je zakladatel několika internetových podniků a hlavní konzultant týmu IBM Industry Models and Assets Lab, který se zaměřuje na řešení datových skladů a business intelligence.

Svou kariéru začal koncem 70. let, kdy byl Bill Gates pouhým milionářem. Od té doby pracoval jako vývojář, správce databáze, datový modelář, vedoucí projektu, datový architekt, podnikový informační architekt, auditor datových skladů a business intelligence a stratég. Navíc se projevil jako inovativní podnikatel. Od té doby často cestuje po celém světě, vyučuje, školí a demonstruje praktické aspekty a řešení datových skladů a business intelligence z hlediska návrhu, optimalizací, optimálních postupů a zdravého rozumu na koncepční, logické i fyzické úrovni. Úspěšně pomohl více než 50 velkým organizacím, které působí v různých oborech – např. v maloobchodu, pojišťovnictví, zdravotní péči, železniční dopravě, telekomunikacích, státní správě, elektronickém obchodování a bankovníctví.

O odborných korektorech

Bob Laberge je držitelem titulu MBA (Masters of Business Administration) z britské University of Durham. Můžete jej kontaktovat na adrese datawarehousementor@gmail.com.

David Marcotte je viceprezidentem společnosti Kantar Retail pro divizi Retail Insights Americas, který má zkušenosti s mezinárodním obchodem, podporou prodeje, fungováním maloobchodu, dodavatelským řetězcem a technologiemi business intelligence. Profesionálně přednáší na mnohá témata a pracuje také jako školitel na pokročilých seminářích. Spolupracoval přitom s mnoha společnostmi ze seznamu Fortune 500 z celého světa. V nedávné době byl hlavním výzkumníkem a autorem studie 2010 Retailing Emergent Markets pro radu Coca-Cola Retailer Research Council, která shrnula zjištěná data z Číny, Peru, Turecka, Polska, Jižní Afriky a Brazílie. Před svým nástupem do společnosti Kantar Retail zastával vedoucí pozici ve skupině Global Business Intelligence ve společnosti IBM. Přitom se účastnil práce týmů, které vyvíjely různá řešení pro maloobchod, včetně ochrany před ztrátami, tvorby cen, zpracování správy kategorií, plánování podpory prodeje, dodavatelského řetězce, podpory prodeje a optimalizace prodejního personálu. Pracoval také na pozici podnikového a projektového vedoucího v divizi společnosti IBM Retail Data Warehouse Solution, kdy poprvé navázal spolupráci s autorem

knihy Bobem Labergem. Tvořil také řešení pro mnohé maloobchodní řetězce a výrobce, včetně společností Sony, Sobeys, D&S, Falabella, Shoppers Drug Mart a Kroger. Zasedal v mnoha oborových radách a panelech, z nichž lze jmenovat například komisi IRI Product Advisory Board a skupinu Efficient Consumer Response (ECR). Můžete jej kontaktovat na adrese *David.Marcotte@KantarRetail.com*.

Ken Yu pracuje v oboru správy informací a datové architektury více než 25 let. Má zkušenosti s prací jako podnikový informační architekt, vedoucí projektu, datový modelář, správce databáze a vývojář v mnoha různých oborech a aplikacích. V roce 1995 nastoupil do společnosti IBM, kde měl za úkol podporovat klienty velkých databází a poskytovat provozní vedení. Jako konzultant správy informací vyučuje a školí v různých částech podniku na témata správy dat, správy hlavních dat a metadat a hodnoty datového modelování a referenčních modelů. Pracoval jako vedoucí pracovník při návrhu a vývoji technické architektury a provozních procesů u mnoha aplikací kritických pro činnost podniku a řešení datových skladů a business intelligence. V současnosti zastává pozici hlavního konzultanta společnosti Cypselurus Inc. se sídlem v Torontu v kanadské provincii Ontario.

Úvod

Tato kniha představuje souhrn různých témat z oboru datových skladů. Kniha je uspořádána tak, aby reprezentovala výstavbu systému datového skladu z obchodního i technického hlediska s důrazem na budování konkrétního řešení. Používá přitom jednoduchou a srozumitelnou terminologii. Uvedené přehledy jsou založeny na třech desetiletích osobních zkušenostech z více než 50 organizací ve více než 20 zemích, kde jsem pracoval jako nezávislý konzultant, zaměstnanec a jako partner týmu Industry Models and Assets Lab společnosti IBM.

V této knize vysvětlíme komponenty a různé alternativy budování datového skladu spolu s kombinací výhod a nevýhod při volbě jednoho z možných postupů. Budování datového skladu je v každé organizaci jedinečné, ale může využít zkušeností získaných při práci v různých prostředích datových skladů a business intelligence v organizacích po celém světě. Kniha předkládá témata datových skladů nejprve formou obecného přehledu, který zajišťuje seznámení s terminologií a kontextem, a pokračuje podrobným vysvětlením jednotlivých témat. Tato témata se vesměs týkají datových skladů, business intelligence a správy výkonu.

Při budování datového skladu neexistují žádná pevná pravidla, ale lze využít mnoho doporučení. Hlavní motiv této knihy je takový, že byste měli budovat tak, aby systém vyhovoval požadavkům vaší organizace s ohledem na specifické a zjištěné podnikové požadavky. Zároveň byste měli vytvořit otevřenou a pružnou architekturu, která umožní budoucí rozšiřování projektu. Hodně organizací vyplývá značné prostředky na vytvoření centralizovaného datového skladu s funkcemi vykazování business intelligence a poté zjišťuje, že řešení je pro jednu nebo dvě aplikace příliš specifické a nedokáže uspokojit budoucí požadavky. Budoucnost samozřejmě nelze předvídat, ale můžeme odhadovat budoucí datové požadavky a použití do té míry, že je možné zaručit, že se návrh a budované prostředí vyznačují dostatečnou pružností a otevřeností ke změnám a že je nebude nutné pokaždé znovu navrhovat a budovat od začátku.

Management mnoha organizací si uvědomuje, že podniková data představují základní aktiva a vyžadují organizaci, strukturování a správu. Tím lze zajistit, že bude podniková informační hodnota sdílena v rámci podniku a bude zajištěna dostatečná kvalita, správa a vlastnictví dat. Podniky nemohou fungovat bez informačních systémů a informační systémy nemohou exis-

tovat bez podnikového účelu. Díky spolupráci, která zajistí vzájemnou podporu obou a která je založena na znalostech informační architektury a použití, mohou podniky zdokonalit své fungování.

O této knize

Část I: Příprava

Část I předkládá základní koncepce a přehledy problematiky datových skladů a business intelligence. Cílem této části je poskytnout přehled základů správy.

Kapitola 1: Přehled datových skladů a systémů business intelligence – kapitola 1 poskytuje přehled datových skladů a technologie business intelligence a končí často kladenými obecnými otázkami, které se týkají implementace systému datového skladu.

Kapitola 2: Data v organizaci – kapitola 2 probírá data jako podniková aktiva a přehledně popisuje organizaci dat.

Kapitola 3: Argumenty pro implementaci – kapitola 3 rozebírá důvody pro a proti budování systému datového skladu. Důvody „pro“ jsou typické scénáře budování datového skladu a důvody „proti“ jsou založeny na možnostech postupu vývojového projektu v rámci organizace s ohledem na její kulturu a omezení.

Kapitola 4: Strategie datového skladu a business intelligence – kapitola 4 poskytuje přehled plánování činností spolu s rozбором toho, kde a jak začít projekt v závislosti na tom, zda je iniciativa zaměřena na řešení podnikového vykazování, nebo se jedná o iniciativu organizace a strukturování dat.

Kapitola 5: Projektové zdroje: role a přehledy – kapitola 5 analyzuje klíčové role iniciativy projektu datového skladu a struktury týmu s optimálními postupy.

Kapitola 6: Přehled dokumentů – kapitola 6 představuje stručný přehled obsahuje běžné projektové listiny, rozsahu projektu a dokumentů pracovních výkazů.

Část II: Komponenty

Část II prezentuje základní komponenty systému datového skladu. Cílem této sekce je přiblížit technické aspekty databázového systému datového skladu a business intelligence. Tato část se zabývá složkami systému datového skladu, které umožňují realizaci iniciativy správy podnikových aktiv a business intelligence.

Kapitola 7: Business intelligence: datové trhy a použití – kapitola 7 se týká podrobností datových trhů a použití od datových modelů po výkonnostní aspekty.

Kapitola 8: Podnikové datové modely – kapitola 8 popisuje podnikové datové modelování spolu s příklady postupů a obecnými hledisky.

Kapitola 9: Architektura datového skladu: Komponenty – kapitola 9 se zaměřuje na rozbor různých typů architektury datových skladů z perspektivy modelování a toku dat.

Kapitola 10: ETL a kvalita dat – kapitola 10 se zaměřuje na obecné vlastnosti vrstvy získávání dat a distribuční vrstvy v systému datového skladu spolu z hledisky kvality dat.

Kapitola 11: Plánování projektu a metodika – kapitola 11 vysvětluje metodiku plánování projektu datového skladu a business intelligence.

Část III: Implementace

Část III předkládá praktický pohled na budování systému datového skladu. V této části se zaměříme na konkrétní implementační scénáře a iniciativy spolu s popisem problematiky správy dat a hodnocení po dokončení projektu.

Kapitola 12: Pracovní scénáře – kapitola 12 se týká budování systému datového skladu a business intelligence na základě metodiky shora dolů a zdola nahoru a také hybridním přístupem. Analyzuje i několik dalších témat, včetně stručného pohledu na podnikovou informační architekturu.

Kapitola 13: Správa dat – kapitola 13 se věnuje správě podnikových dat včetně organizační struktury, kvality dat, vlastnictví a kontroly změn.

Kapitola 14: Analýza po dokončení projektu – kapitola 14 rozebírá aspekty projektu datového skladu a business intelligence po dokončení vývojových aktivit.

Tuto knihu jsem napsal s úmyslem, aby posloužila jako obecný průvodce při budování systému datového skladu. Snažil jsem se podělit se se znalostmi mnoha témat, která se uplatňují v současných projektech datových skladů, a představit celou problematiku z více úhlů. Doufám, že vám tato kniha pomůže ve vašich projektech datových skladů.

Hodně štěstí!

Robert (Bob) Laberge

Kontakt: datawarehousementor@gmail.com

Zpětná vazba od čtenářů

Nakladatelství a vydavatelství Computer Press, které pro vás tuto knihu přeložilo, stojí o zpětnou vazbu a bude na vaše podněty a dotazy reagovat. Můžete se obrátit na následující adresy:

Computer Press

Albatros Media a.s., pobočka Brno

IBC

Příkop 4

602 00 Brno

nebo

sefredaktor.pc@albatrosmedia.cz

Computer Press neposkytuje rady ani jakýkoli servis pro aplikace třetích stran. Pokud budete mít dotaz k programu, obraťte se prosím na jeho tvůrce.

Errata

Přestože jsme udělali maximum pro to, abychom zajistili přesnost a správnost obsahu, chybám se úplně vyhnout nelze. Pokud v některé z našich knih najdete chybu, budeme rádi, pokud nám ji oznámíte. Ostatní uživatelé tak můžete ušetřit frustrace a pomoci nám zlepšit následující vydání této knihy.

Veškerá existující errata zobrazíte na adrese <http://knihy.cpress.cz/K1994> po klepnutí na odkaz Soubory ke stažení.