

Decifrando Arquiteturas de Dados

**Escolhendo entre data warehouse
moderno, data fabric, data lakehouse
e data mesh**

James Serra

O'REILLY®
Novatec

Authorized Portuguese translation of the English edition of Deciphering Data Architectures ISBN 9781098150761 © 2024 James Serra. This translation is published and sold by permission of O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

Tradução em português autorizada da edição em inglês da obra Deciphering Data Architectures ISBN 9781098150761 © 2024 James Serra. Esta tradução é publicada e vendida com a permissão da O'Reilly Media, Inc., detentora de todos os direitos para publicação e venda desta obra.

© Novatec Editora Ltda. [2024].

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998. É proibida a reprodução desta obra, mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e da Editora.

Editor: Rubens Prates

Tradução: Aldir Coelho Corrêa da Silva

Revisão gramatical: Tássia Carvalho

ISBN do impresso: 978-85-7522-921-7

ISBN do ebook: 978-85-7522-922-4

Histórico de impressões:

Agosto/2024 Primeira edição

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110

02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

Email: novatec@novatec.com.br

Site: <https://novatec.com.br>

Twitter: twitter.com/novateceditora

Facebook: facebook.com/novatec

LinkedIn: linkedin.com/in/novatec

GRA20240807

Sumário

Apresentação	15
Prefácio	17
Parte I ■ Base	21
Capítulo 1 ■ Big Data	22
O que é big data e como ele pode ajudá-lo?	23
Maturidade dos dados.....	27
Estágio 1: Reativo	28
Estágio 2: Informativo.....	28
Estágio 3: Preditivo	29
Estágio 4: Transformador.....	29
Self-service business intelligence.....	29
Resumo	30
Capítulo 2 ■ Tipos de arquiteturas de dados.....	32
Evolução das arquiteturas de dados	33
Data warehouse relacional.....	36
Data Lake	38
Data warehouse moderno.....	40
Data Fabric	41
Data Lakehouse	41
Data Mesh	42
Resumo	43
Capítulo 3 ■ Sessão de design de arquitetura	44
O que é uma ADS?	44
Por que fazer uma ADS?	44
Antes da ADS	46
Preparação.....	46
Convidando os participantes.....	49
Conduzindo a ADS.....	50
Introduções	50
Descoberta	51
Quadro branco	57

Depois da ADS	58
Dicas para a condução de uma ADS	59
Resumo	61

Parte II ■ Conceitos comuns das arquiteturas de dados62

Capítulo 4 ■ Data warehouse relacional63

O que é um data warehouse relacional?	63
O que um data warehouse não é.....	66
Abordagem top-down	68
Por que usar um data warehouse relacional?	70
Desvantagens do uso de um data warehouse relacional	74
Preenchendo um data warehouse.....	75
Com que frequência os dados serão extraídos	76
Métodos de extração.....	76
Como saber que dados mudaram desde a última extração.....	77
É um exagero afirmar que o data warehouse relacional morreu.....	79
Resumo	80

Capítulo 5 ■ Data Lake81

O que é um data lake?.....	82
Por que usar um data lake?.....	82
Abordagem bottom-up.....	84
Melhores práticas para o design de um data lake.....	86
Múltiplos data lakes.....	93
Vantagens.....	93
Desvantagens.....	97
Resumo	97

Capítulo 6 ■ Soluções e processos de armazenamento de dados.....99

Soluções de armazenamento de dados	100
Data Marts.....	100
Armazenamentos de dados operacionais	101
Data Hubs	104
Processos de dados.....	106
Gerenciamento de dados mestres	106
Virtualização de dados e federação de dados.....	108
Catálogos de dados	113
Marketplaces de dados.....	114
Resumo	116

Capítulo 7 ■ Abordagens de design.....117

Processamento de transações online versus processamento analítico online	118
Dados operacionais e analíticos	120

Multiprocessamento simétrico e processamento massivamente paralelo	121
Arquitetura Lambda	123
Arquitetura Kappa	126
Persistência poliglota e armazenamentos de dados poliglotas	127
Resumo	129

Capítulo 8 ■ Abordagens de modelagem de dados 130

Modelagem relacional	130
Chaves	131
Diagramas entidade-relacionamento.....	131
Regras e formas de normalização	132
Rastreando alterações	133
Modelagem dimensional	134
Fatos, dimensões e chaves	134
Rastreando alterações	135
Desnormalização	137
Modelo de dados comum	138
Cofre de dados.....	139
Metodologias de data warehousing de Kimball e Inmon	141
Metodologia top-down de Inmon.....	142
Metodologia bottom-up de Kimball.....	144
Escolhendo uma metodologia	145
Modelos híbridos.....	146
Mitos das metodologias.....	149
Resumo	152

Capítulo 9 ■ Abordagens da ingestão de dados 153

ETL versus ELT	153
ETL reverso	156
Processamento em lote versus processamento em tempo real.....	157
Vantagens e desvantagens do processamento em lote	159
Vantagens e desvantagens do processamento em tempo real.....	159
Governança de dados	160
Resumo	161

Parte III ■ Arquiteturas de dados..... 163

Capítulo 10 ■ Data warehouse moderno 164

Arquitetura MDW	164
Vantagens e desvantagens da arquitetura MDW.....	170
Combinando o RDW e o data lake	172
Data Lake.....	172
Data warehouse relacional	173

Etapas importantes para a criação de um MDW	173
Ampliação do EDW	174
Data lake temporário mais EDW	175
Multifuncional.....	176
Estudo de caso: Mudança estratégica da Wilson & Gunkerk para um MDW	177
Desafio.....	178
Solução.....	178
Resultado	178
Resumo	179
Capítulo 11 ■ Data Fabric.....	181
Arquitetura data fabric	182
Políticas de acesso a dados	183
Catálogo de metadados.....	184
Gerenciamento de dados mestres	185
Virtualização de dados.....	185
Processamento em tempo real	185
APIs	186
Serviços.....	186
Produtos.....	186
Por que fazer a transição de um MDW para uma arquitetura data fabric?	186
Possíveis desvantagens	187
Resumo	188
Capítulo 12 ■ Data Lakehouse.....	189
Recursos do Delta Lake.....	190
Melhorias no desempenho.....	192
Arquitetura data lakehouse	194
E se você evitar usar o data warehouse relacional?.....	196
Camada de atendimento relacional (relational serving layer).....	199
Resumo	200
Capítulo 13 ■ Aspectos básicos do data mesh	201
Uma arquitetura de dados descentralizada.....	202
Badalação em cima do data mesh	203
Os quatro princípios de Dehghani para o Data Mesh	205
Princípio #1: Propriedade de domínio	205
Princípio #2: Dados como produto.....	206
Princípio #3: Plataforma de autoatendimento de dados.....	208
Princípio #4: Governança computacional federada	209
O data mesh “puro”	211
Domínios de dados	213
Arquitetura lógica do data mesh.....	214

Diferentes topologias.....	215
Data mesh versus data fabric	218
Casos de uso.....	218
Resumo	220

Capítulo 14 ■ Você deve adotar o data mesh? Mitos, problemas e o futuro 222

Mitos.....	222
Mito: Usar o data mesh é uma bala de prata que resolve todos os desafios de dados rapidamente	223
Mito: Um data mesh substituirá seu data lake e seu data warehouse	223
Mito: Todos os projetos de data warehouse estão falhando, e um data mesh resolverá esse problema.....	223
Mito: Criar um data mesh significa descentralizar absolutamente tudo.....	224
Mito: Você pode usar a virtualização de dados para criar um data mesh.....	224
Problemas.....	225
Questões filosóficas e conceituais	226
Combinando dados em um ambiente descentralizado	227
Outros problemas da descentralização.....	228
Complexidade	229
Duplicação	230
Viabilidade	231
Pessoas	233
Obstáculos no nível de domínios	234
Avaliação organizacional: você deve adotar um data mesh?	236
Recomendações para a implementação de um data mesh bem-sucedido	238
Futuro do data mesh	239
Olhando de forma geral: entendendo as arquiteturas de dados e suas aplicações	240
Resumo	242

Parte IV ■ Pessoas, processos e tecnologia..... 243

Capítulo 15 ■ Pessoas e processos 244

Organização de equipes: papéis e responsabilidades.....	245
Papéis para o MDW, Data Fabric ou Data Lakehouse.....	245
Papéis para o data mesh	247
Por que os projetos falham: armadilhas e prevenção.....	251
Armadilha: Permitir que os executivos achem que inteligência empresarial é “fácil”..	251
Armadilha: Usar as tecnologias erradas	252
Armadilha: Coletar muitos requisitos empresariais	252
Armadilha: Coletar poucos requisitos empresariais.....	252
Armadilha: Apresentar relatórios sem antes validar seu conteúdo.....	253
Armadilha: Contratar uma empresa de consultoria inexperiente	253

Armadilha: Contratar uma empresa de consultoria que terceirize o desenvolvimento para profissionais de outras regiões	253
Armadilha: Passar a responsabilidade do projeto para os consultores	254
Armadilha: Negligenciar a necessidade de transferir conhecimento para a empresa	254
Armadilha: Reduzir o orçamento no meio do projeto	254
Armadilha: Começar com uma data final e trabalhar baseando-se nela	254
Armadilha: Estruturar o data warehouse para refletir os dados de origem, e não as necessidades da empresa	255
Armadilha: Apresentar para os usuários finais uma solução com tempos de resposta lentos ou outros problemas de desempenho	255
Armadilha: Criar um design excessivo (ou insatisfatório) para sua arquitetura de dados..	256
Armadilha: Comunicação deficiente entre a área de TI e os domínios de negócios .	256
Dicas para o sucesso	257
Não economize ao investir	257
Envolve os usuários, mostre-lhes os resultados e empolgue-os.....	258
Adicione valor aos novos relatórios e dashboards	258
Peça aos usuários finais que criem um protótipo	259
Encontre um defensor/patrocinador do projeto	259
Crie um plano para o projeto que vise a 80% de eficiência.....	260
Resumo	260
Capítulo 16 ■ Tecnologias.....	262
Escolhendo uma plataforma	262
Soluções open source	262
Soluções locais.....	265
Soluções de provedores de nuvem.....	267
Modelos de serviços de nuvem.....	270
Principais provedores de nuvem	272
Soluções multicloud	273
Frameworks de software.....	276
Hadoop	276
Databricks	280
Snowflake.....	282
Resumo	284
Índice remissivo	287