
Spis treści

Wprowadzenie	9
Rozdział 1. Wprowadzenie do potoków danych	13
Czym jest potok danych?	13
Kto tworzy potok danych?	14
Podstawy pracy z SQL i hurtowniami danych	15
Python i/lub Java	15
Przetwarzanie rozproszone	16
Podstawowa administracja systemem	16
Nastawienie bazujące na celach	16
Dlaczego w ogóle są tworzone potoki danych?	17
Jak jest tworzony potok danych?	18
Rozdział 2. Nowoczesna infrastruktura danych	19
Różnorodność źródeł danych	20
Własność źródła danych	20
Interfejs pobierania danych i ich struktura	21
Wolumen danych	23
Czystość danych i ich weryfikacja	24
Opóźnienie i przepustowość systemu źródłowego	25
Jezioro danych i hurtownia danych w chmurze	26
Narzędzia pobierania danych	27
Przekształcanie danych i narzędzia modelowania	28
Platformy narzędzi koordynacji sposobu pracy	30
Skierowany graf acykliczny	30
Dostosowanie infrastruktury danych do własnych potrzeb	32

Rozdział 3. Najczęściej spotykane wzorce potoków danych33

ETL i ELT	33
Pojawienie się ELT po ETL	35
Podwzorzec EtLT	38
ELT w analizie danych	39
ELT dla naukowca	41
ELT dla produktów danych i uczenia maszynowego	41
Etapy potoku danych dla uczenia maszynowego	42
Wykorzystanie informacji zwrotnych w potoku	44
Więcej zasobów dotyczących potoków danych dla uczenia maszynowego	44

Rozdział 4. Pobieranie danych — wyodrębnianie46

Przygotowanie środowiska Pythona	47
Przygotowanie plikowego magazynu danych w chmurze	49
Wyodrębnianie danych z bazy danych MySQL	52
Pełne i przyrostowe wyodrębnianie danych z tabeli MySQL	54
Binarny dziennik zdarzeń replikacji danych MySQL	64
Wyodrębnianie danych z bazy danych PostgreSQL	74
Pełne i przyrostowe wyodrębnianie danych z tabeli PostgreSQL	76
Replikacja danych za pomocą dziennika zdarzeń WAL	78
Wyodrębnianie danych z bazy danych MongoDB	78
Wyodrębnianie danych z API REST	85
Strumieniowane pobieranie danych za pomocą Kafki i Debezium ...	89

Rozdział 5. Pobieranie danych — wczytywanie92

Konfiguracja hurtowni danych Amazon Redshift jako miejsca docelowego	92
Wczytywanie danych do hurtowni danych Redshift	94
Wczytywanie przyrostowe a pełne	99
Wczytywanie danych wyodrębnionych z dziennika zdarzeń CDC ...	102
Konfiguracja hurtowni danych Snowflake jako miejsca docelowego	103
Wczytywanie danych do hurtowni danych Snowflake	105
Używanie plikowego magazynu danych jako jeziora danych	107
Frameworki typu open source	109
Alternatywy komercyjne	110

Rozdział 6. Przekształcanie danych 113

Przekształcenia pozbawione kontekstu	114
Usunięcie powtarzających się rekordów w tabeli	114
Przetwarzanie adresów URL	119
Kiedy powinno odbywać się przekształcanie	
— podczas pobierania danych czy już po?	123
Podstawy modelowania danych	124
Najważniejsze pojęcia związane z modelowaniem danych	124
Modelowanie w pełni odświeżonych danych	125
Powolna zmiana wymiarów w celu pełnego odświeżenia danych	130
Modelowanie przyrostowo pobieranych danych	132
Modelowanie danych, które są tylko dołączane	137
Modelowanie zmiany przechwytywanych danych	147

Rozdział 7. Narzędzia instrumentacji potoków danych 153

Skierowany graf acykliczny	153
Konfiguracja Apache Airflow i ogólne omówienie	
tego frameworka	154
Instalacja i konfiguracja	155
Baza danych Apache Airflow	156
Serwer WWW i interfejs użytkownika	159
Harmonogram	163
Wykonawca	164
Operatory	165
Tworzenie skierowanego grafu acyklicznego	
za pomocą Apache Airflow	165
Prosty skierowany graf acykliczny	166
Skierowany graf acykliczny potoku danych ELT	170
Dodatkowe zadania potoku danych	175
Komunikaty i powiadomienia	175
Weryfikacja danych	176
Zaawansowane konfiguracje koordynacji	176
Połączone a niepołączone zadania potoku danych	176
Kiedy podzielić skierowany graf acykliczny?	177
Koordynacja wielu grafów za pomocą operatora Sensor	178
Zarządzane opcje Apache Airflow	181
Inne frameworki koordynacji	182

Rozdział 8. Weryfikacja danych w potoku 184

Weryfikuj wcześniej i często	184
Jakość danych w systemie źródłowym	185
Niebezpieczeństwa związane z pobieraniem danych	186
Umożliwienie analitykowi weryfikacji danych	187
Prosty framework weryfikacji	188
Kod frameworka weryfikacji	188
Struktura testu weryfikacyjnego	192
Wykonywanie testu weryfikacyjnego	194
Używanie frameworka w skierowanym grafie acyklicznym	
Apache Airflow	195
Kiedy zatrzymać wykonywanie potoku, a kiedy tylko wygenerować ostrzeżenie i kontynuować potok?	196
Rozbudowa frameworka	198
Przykłady testów weryfikacyjnych	202
Powielone rekordy po operacji pobierania danych	203
Nieoczekiwana zmiana liczby rekordów	
po operacji pobierania danych	204
Fluktuacje wartości wskaźników	207
Komercyjne i niekomercyjne frameworki do weryfikacji danych ...	212

Rozdział 9. Najlepsze praktyki podczas pracy z potokiem danych 214

Obsługa zmian w systemach źródłowych	214
Wprowadzenie abstrakcji	214
Obsługa kontraktów danych	215
Ograniczenia schematu podczas odczytu	217
Skalowanie złożoności	219
Standaryzacja pobierania danych	219
Wielokrotne używanie logiki modelu danych	222
Zapewnienie spójności zależności	225

Rozdział 10. Pomiar i monitorowanie wydajności działania potoku danych 229

Kluczowe wskaźniki potoku	229
Przygotowanie hurtowni danych	230
Schemat infrastruktury danych	231

Rejestrowanie danych i sprawdzanie wydajności działania	
operacji pobierania danych	232
Pobieranie z Apache Airflow historii wykonania	
poszczególnych skierowanych grafów acyklicznych	232
Dodawanie funkcjonalności rejestrowania danych	
do frameworka weryfikacji danych	236
Przekształcanie danych dotyczących wydajności działania	242
Wskaźnik sukcesu skierowanego grafu acyklicznego	242
Zmiana czasu wykonywania skierowanego grafu acyklicznego	
na przestrzeni czasu	244
Liczba testów weryfikacyjnych i współczynnik sukcesu	245
Koordynacja wydajności działania potoku	248
Skierowany graf acykliczny dotyczący wydajności działania	249
Przejrzystość wydajności działania	250
Skorowidz	252