Snowflake

Nowoczesna inżynieria danych w praktyce

Maja Ferle

Helion

Tytuł oryginału: Snowflake Data Engineering

Tłumaczenie: Ksawery Sosnowski

ISBN: 978-83-289-2898-5

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
helion.pl/user/opinie/snowfl
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

1 Inżynieria danych w środowisku Snowflake

Zawartość rozdziału:

- wykorzystanie Snowflake'a w inżynierii danych;
- analiza obowiązków inżyniera danych w Snowflake'u;
- konstruowanie potoków danych w Snowflake'u;
- inżynieria danych w aplikacjach tworzonych z użyciem Snowflake'a.

Organizacje w niemal każdej branży gromadzą dane, często w ogromnych ilościach. Analitycy danych, danetycy i inni specjaliści biznesowi wykorzystują te informacje do uzyskiwania spostrzeżeń wspierających podejmowanie decyzji. Surowe dane rzadko nadają się do bezpośredniego wykorzystania. Użytkownicy potrzebują danych, które zostały przetworzone zgodnie z regułami biznesowymi, przekształcone tak, aby odpowiadały konkretnym potrzebom oraz ewentualnie wzbogacone o dane zewnętrzne.

Inżynieria danych to praktyka budowania rozwiązań, które wydobywają dane z systemów źródłowych, przekształcają je w użyteczne informacje i prezentują ujednolicone dane użytkownikom do dalszego wykorzystania. Akronimy **ETL** (ang. *extract*, *transform*, *load* — wyodrębnij, przekształć, załaduj) lub **ELT** (ang. *extract*, *load*, *transform* — wyodrębnij, załaduj, przekształć) są również powszechnie używane do określenia tych rozwiązań.

Inżynierowie danych są odpowiedzialni za tworzenie potoków danych, które umożliwiają analitykom, danetykom i innym użytkownikom dostęp do potrzebnych im informacji. Dostarczanie wysokiej jakości danych na czas jest kluczowe dla skutecznej analizy, dlatego inżynierowie danych odgrywają krytyczną rolę w dziedzinie analityki danych.

1.1. Snowflake a inżynieria danych

Organizacje zazwyczaj pozyskują dane z systemów źródłowych, przechowują je w hurtowni danych i udostępniają użytkownikom w postaci wymiarowych składnic danych. Dodatkowe scenariusze obejmują tworzenie jezior danych, składnic danych lub domenowych produktów danych. W każdym z tych przypadków dane są wydobywane z systemów źródłowych, wprowadzane do docelowej platformy przechowywania i przekształcane na potrzeby odbiorców końcowych.

Snowflake to nowoczesna chmurowa platforma danych służąca do przechowywania, przetwarzania i analizy danych oraz tworzenia aplikacji. Jest popularnym wyborem dla rozwiązań obejmujących hurtownie danych, jeziora danych, analizę danych czy danetykę. W poniższych punktach szczegółowo wyjaśniam możliwości środowiska Snowflake w zakresie inżynierii danych.

> **WSKAZÓWKA** Platforma Snowflake jest nieustannie rozwijana i są do niej dodawane nowe funkcje. Najbardziej aktualne informacje są dostępne na oficjalnej stronie Snowflake'a pod adresem https://www.snowflake.com/.

1.1.1. Architektura Snowflake'a

Środowisko Snowflake to chmurowe rozwiązanie typu SaaS (ang. *Software as a Service* — oprogramowanie jako usługa). Obecnie wspieranymi dostawcami usług chmurowych są Amazon Web Services, Microsoft Azure i Google Cloud Platform. Gdy użytkownicy wybierają Snowflake'a jako swoją platformę danych w chmurze, nie muszą kupować, instalować, konfigurować, zarządzać ani serwisować sprzętu i oprogramowania.

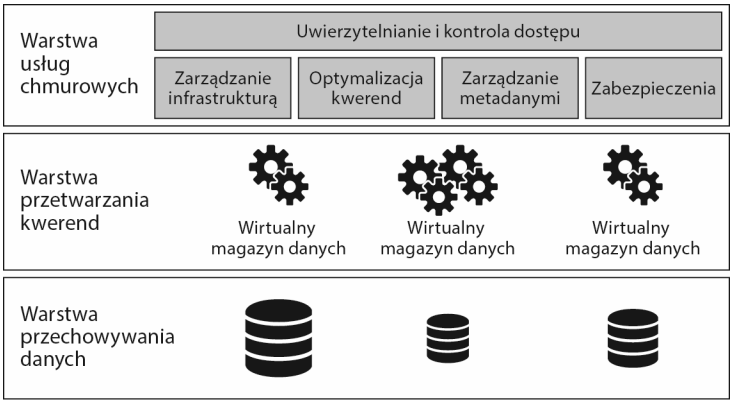
W architekturze platformy Snowflake przechowywanie i przetwarzanie danych są zarządzane oddzielnie. Dane są zapisywane jeden raz, a następnie wykorzystywane przez wiele różnych procesów, takich jak potoki inżynierii danych, silniki kwerend, rozwiązania danetyczne, udostępnianie informacji i narzędzia analityczne.

Snowflake przechowuje dane w chmurze w skompresowanym formacie kolumnowym. Zarządza on organizacją, strukturą, kompresją, metadanymi, statystykami i bezpieczeństwem danych. Użytkownicy mają dostęp do obiektów danych za pomocą kwerend SQL.

Użytkownicy wykonują kwerendy za pomocą wirtualnych magazynów danych. Każdy wirtualny magazyn wykorzystuje własne zasoby obliczeniowe, nie wpływając

na inne magazyny. Ilość zasobów dla poszczególnych magazynów można konfigurować. Snowflake może również automatycznie zwiększać lub zmniejszać przydzielone zasoby w zależności od oczekiwanej wydajności i obciążenia.

Snowflake koordynuje różne składniki za pośrednictwem warstwy usług chmurowych. Usługi te obejmują uwierzytelnianie, kontrolę dostępu, zarządzanie infrastrukturą, optymalizację kwerend, zarządzanie metadanymi oraz zabezpieczenia. Rysunek 1.1 przedstawia trzy warstwy architektury Snowflake'a: przechowywania danych, przetwarzania kwerend oraz usług chmurowych.



Rysunek 1.1. *Architektura Snowflake'a składa się z trzech warstw: przechowywania danych, przetwarzania kwerend oraz usług chmurowych*

Snowflake oferuje zaawansowane funkcje przechowywania danych, takie jak **klonowanie bez kopiowania** (ang. *zero-copy cloning*) i **podróże w czasie** (ang. *time travel*). Klonowanie bez kopiowania umożliwia tworzenie kopii danych bez ponoszenia dodatkowych kosztów przechowywania. Pozwala to inżynierom danych na szybkie tworzenie nowych środowisk. Na przykład mogą oni sklonować środowisko programistyczne do środowiska testowego lub skopiować środowisko produkcyjne do środowiska programistycznego w celu opracowania nowych funkcji.

Podróże w czasie mogą chronić dane przed niezamierzonymi błędami lub złośliwymi działaniami. Umożliwiają użytkownikom dostęp do danych z wybranego momentu w przeszłości i stanowią efektywny sposób tworzenia kopii zapasowych w ramach konfigurowalnego okresu.

Snowflake umożliwia replikację baz danych między kontami Snowflake'a w różnych rejonach geograficznych i platformach chmurowych. Ta funkcja zapewnia synchronizację replikowanych danych z danymi źródłowymi. Pozwala również przybliżyć dane do użytkowników w wybranym rejonie, co zmniejsza opóźnienia kwerend i służy jako kopia zapasowa.

1.1.2. Funkcje Snowflake'a używane w inżynierii danych

Snowflake oferuje różne funkcje do pozyskiwania i przekształcania danych. Funkcje te zostały opisane w poniższych podpunktach.

POBIERANIE DANYCH Z MAGAZYNÓW CHMUROWYCH

Użytkownicy mogą importować dane z plików przechowywanych w obsługiwanych dostawcach magazynów chmurowych: komorach (ang. *bucket*) AWS S3, kontenerach Azure Blob Storage oraz zasobnikach (ang. *bucket*) GCP. Pliki mogą być w formatach strukturalnych lub półstrukturalnych, takich jak CSV, JSON, XML lub Parquet. Snowflake umożliwia również pracę z plikami nieustrukturyzowanymi zawierającymi obrazy, wideo lub dźwięk. Użytkownicy mogą udostępniać adresy URL dostępu do plików oraz wczytywać te adresy URL i metadane plików do tabel Snowflake'a.

USŁUGA SNOWPIPE

Snowpipe to usługa umożliwiająca ciągłe pozyskiwanie danych. Pozwala ona wprowadzać pliki z magazynu chmurowego do tabel Snowflake'a natychmiast po pojawieniu się nowych danych w chmurze. Ciągłe pozyskiwanie danych jest możliwe dzięki integracji z usługą powiadomień dostawcy chmury, która informuje Snowpipe'a o dostępności nowych plików do wczytania.

ZADANIA W SNOWFLAKE'U

Snowflake umożliwia tworzenie zadań. Zadanie realizuje instrukcje SQL, procedury składowane lub bloki skryptów Snowflake'a zgodnie z powtarzającym się harmonogramem. Zadania mogą być zależne od innych zadań, co pozwala na implementację złożonych potoków przetwarzania danych.

INTERFEJS SNOWPARK

Snowpark udostępnia biblioteki API pozwalające inżynierom danych tworzyć potoki danych w jednym z obsługiwanych języków programowania, takim jak Python, Scala czy Java. Kod napisany w tych językach jest wykonywany natywnie w środowisku Snowflake. Dzięki temu inżynierowie danych mogą budować potoki danych, korzystając z wybranego języka programowania, na przykład Pythona, i w razie potrzeby wykorzystywać dodatkowe funkcje dostępne w bibliotekach o otwartym kodzie źródłowym.

FUNKCJE ZARZĄDZANIA DANYMI

Dane przechowywane w środowisku Snowflake mogą zawierać wrażliwe informacje, które powinny być dostępne tylko dla uprawnionych użytkowników. W architekturach wielotenantowych dane różnych organizacji lub jednostek biznesowych są przechowywane w jednej bazie danych. Mimo to dostęp do danych

jest ograniczony do użytkowników z tej samej organizacji. Aby obsłużyć takie przypadki użycia, Snowflake oferuje funkcje zarządzania danymi, takie jak reguły maskowania i reguły dostępu do wierszy, które uniemożliwiają nieupoważnionym użytkownikom dostęp do wrażliwych lub poufnych danych.

SNOWFLAKE MARKETPLACE

Snowflake Marketplace to platforma, na której użytkownicy mogą odkrywać i uzyskiwać dostęp do zestawów danych udostępnionych przez inne organizacje. Mogą wybierać interesujące ich zestawy danych i wykorzystywać je do wzbogacania własnych analiz. Mają również możliwość publikowania własnych zestawów danych, aby udostępnić je innym. Zestawy danych można udostępniać publicznie lub prywatnie, co umożliwia bezpieczne dzielenie się danymi i współpracę zarówno wewnątrz organizacji, jak i z podmiotami zewnętrznymi.

SNOWFLAKE PARTNER CONNECT

Dzięki bogatemu środowisku partnerów biznesowych Snowflake płynnie integruje się z wieloma popularnymi narzędziami i usługami zewnętrznymi z różnych kategorii, takich jak analityka biznesowa, integracja danych, uczenie maszynowe, danetyka, bezpieczeństwo, zarządzanie danymi i obserwowalność. Użytkownicy wykorzystują te narzędzia do pozyskiwania danych w procesach inżynierii danych, przeprowadzania analiz, tworzenia raportów i pulpitów oraz pracy z algorytmami uczenia maszynowego i danetycznymi

Dzięki Snowflake'owi inżynierowie danych mogą korzystać z wielu funkcji do budowania potoków danych, dostosowując je do swoich preferencji, założeń architektonicznych i potrzeb biznesowych.

1.2. Obowiązki inżyniera danych w Snowflake'u

Inżynierowie danych są odpowiedzialni za projektowanie, budowanie, orkiestrację, optymalizację, monitorowanie i utrzymywanie potoków, które pozyskują surowe dane z systemów źródłowych i przekształcają je w użyteczne informacje gotowe do dalszego wykorzystania. Dbają oni o to, by dane były łatwo dostępne, bezpieczne, wiarygodne i osiągalne dla użytkowników zgodnie z ich potrzebami.

Rolę inżyniera danych zwykle pełnią specjaliści z wcześniejszym doświadczeniem, ponieważ wymaga ona szerokiego zakresu umiejętności. Wielu inżynierów danych ma za sobą pracę w obszarze tworzenia oprogramowania lub analizy biznesowej i stopniowo rozszerza swoje kompetencje. Pracując w Snowflake'u, zdobywają oni wiedzę i doświadczenie związane z funkcjami tej platformy.

Znajomość chmury obliczeniowej jest kluczowa podczas korzystania ze Snowflake'a. Na przykład inżynierowie danych często pracują z chmurowymi magazynami obiektów i wiedzą, jak bezpiecznie łączyć się ze środowiskami chmurowymi. Wiele organizacji przechodzi na rozwiązania chmurowe i przenosi swoją

infrastrukturę do dostawców usług w chmurze, co oznacza, że inżynierowie danych coraz swobodniej poruszają się w środowisku chmurowym.

Inżynierowie danych znają organizację przechowywania danych w środowisku Snowflake, co pozwala im pisać wydajne kwerendy SQL pobierające i przekształcające dane. Rozumieją, w jaki sposób dane są przechowywane w mikropartycjach Snowflake'a oraz kiedy i jak stosować klasteryzację, optymalizację wyszukiwania, przyspieszanie kwerend lub inne funkcje w celu poprawy wydajności kwerend.

Znają oni również sposoby efektywnego kosztowo wykorzystania wirtualnych magazynów Snowflake'a i rozumieją, jak ich rozwiązania wpływają na zużycie mocy obliczeniowej. Wirtualne magazyny mogą także poprawić wydajność kwerend, ponieważ utrzymują pamięć podręczną danych, którą inne kwerendy wykonywane w tym samym magazynie mogą ponownie wykorzystać.

Podczas tworzenia rozwiązań w środowiskach korporacyjnych inżynierowie danych stosują, tam gdzie jest to możliwe, zasady inżynierii oprogramowania, takie jak najlepsze praktyki pisania kodu, zautomatyzowane testy, kontrola wersji kodu, ciągła integracja oraz ciągłe wdrażanie. Jeśli korzystają z interfejsu Snowpark, posiadają również znajomość języka programowania takiego jak Python.

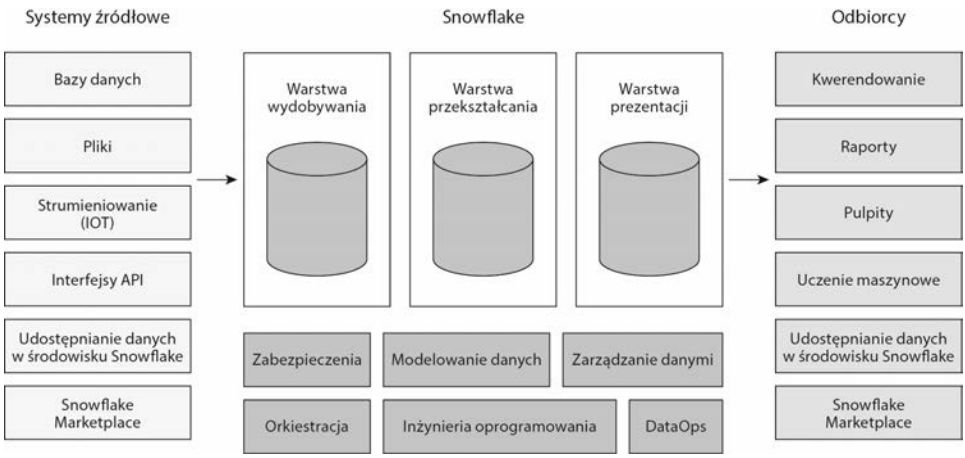
Umiejętność bezpiecznego łączenia się z platformą Snowflake ma kluczowe znaczenie dla uniknięcia naruszeń bezpieczeństwa i przypadkowego udostępnienia poufnych danych osobom nieupoważnionym. Każdy inżynier danych musi koniecznie i szczegółowo znać różne metody nawiązywania połączenia ze środowiskiem Snowflake.

Rysunek 1.2 przedstawia ogólne elementy składowe inżynierii danych służące do budowy potoków danych w ramach platformy chmurowej Snowflake. Nie wszystkie składniki są istotne w każdym projekcie. W zależności od rodzaju tworzonego rozwiązania inżynierowie danych wykorzystują tylko te elementy, które są odpowiednie i niezbędne w danym środowisku. Głównymi komponentami inżynierii danych są:

- pozyskiwanie danych z systemów źródłowych i wprowadzanie ich do Snowflake'a,
- wykonywanie przekształceń danych istotnych dla rozwiązania;
- prezentowanie danych odbiorcom końcowym do celów analityki, tworzenia raportów, danetyki lub innych zastosowań.

Oprócz głównych składników inżynierowie danych wykorzystują również komponenty bazowe podczas budowania potoków przetwarzania danych. Te bazowe elementy obejmują:

- zabezpieczenia,
- modelowanie danych,
- zarządzanie danymi,



Rysunek 1.2. *Elementy składowe inżynierii danych w Snowflake'u obejmują wydobywanie danych z systemów źródłowych, wprowadzanie ich na Snowflake'a, wykonywanie przekształceń oraz udostępnianie danych odbiorcom końcowym. Proces ten uwzględnia podstawowe elementy i najlepsze praktyki związane z bezpieczeństwem, modelowaniem danych, zarządzaniem danymi, orkiestracją, inżynierią oprogramowania oraz metodologią DataOps*

- inżynierię oprogramowania,
- orkiestrację,
- metodologię DataOps.

Każdy z tych elementów został szczegółowo opisany w kolejnych podrozdziałach (zobacz rysunek 1.2).

1.2.1. Pozyskiwanie danych z systemów źródłowych

Pierwszym krokiem w procesie inżynierii danych jest pozyskiwanie danych z systemów źródłowych. Inżynierowie danych zdobywają podstawową wiedzę na temat biznesowego znaczenia danych i ich wykorzystania w rozwiązaniach końcowych. Często kilka systemów źródłowych dostarcza dane do rozwiązania, a inżynierowie danych tworzą potoki, które płynnie pobierają te dane na jedną platformę, tworząc ujednolicony model dla użytkowników.

Inżynierowie danych zapoznają się z technologiami używanymi w systemach źródłowych — na przykład sprawdzają, czy zestawy danych istnieją jako tabele w bazach danych, czy pliki w chmurze, czy są generowane jako wynik zapytań API z aplikacji, czy pochodzą ze strumieni danych z urządzeń IoT, czy też z innych źródeł. Identyfikują najlepsze sposoby dostępu do danych i przygotowania ich do umieszczenia w Snowflake'u. Dane mogą być dostępne nie tylko z zewnętrznych źródeł, lecz również poprzez bezpieczne udostępnianie danych Snowflake'a lub Snowflake Marketplace.

PRZYROSTOWE POBIERANIE DANYCH

Potoki danych zazwyczaj pobierają dane z systemów źródłowych przyrostowo, przetwarzając tylko nowe i zmienione informacje. Inżynierowie danych współpracują z właścicielami systemów źródłowych, aby ustalić najlepszą metodę wykrywania zmian. Wykorzystują filtrowanie według znaczników czasu, mechanizmy śledzenia zmian w bazach danych lub inne techniki. W przypadku pobierania danych z przestarzałych systemów, których właściciele są niedostępni, inżynierowie danych czasami muszą sami określić najlepsze podejście do wykrywania zmian, opierając się na własnym doświadczeniu i wiedzy specjalistycznej.

Podczas pracy z systemami, które nie umożliwiają łatwej identyfikacji zmian w danych, inżynierowie mogą wykorzystać *strumienie* Snowflake'a. Strumień śledzi operacje takie jak wstawianie, aktualizowanie czy usuwanie, wykonywane na danych źródłowych między dwoma punktami w czasie.

NARZĘDZIA A KORZYSTANIE Z WŁASNEGO KODU

Rozpoczynając nowy projekt z zakresu inżynierii danych, inżynierowie danych współpracują z architektami rozwiązań i innymi zainteresowanymi stronami, aby wybrać najlepsze narzędzia i technologie do wydobywania danych z systemów źródłowych do środowiska Snowflake. Mogą to być narzędzia o otwartym kodzie źródłowym lub komercyjne, a także własny kod napisany z wykorzystaniem bibliotek i technologii otwartych lub własnościowych.

Podczas tworzenia własnego kodu inżynierowie danych rozumieją formaty danych źródłowych oraz znają odpowiednie polecenia i opcje do wprowadzania tych danych do Snowflake'a. Są również zaznajomieni z podstawowymi komendami i opcjami Snowflake'a używanymi przez zewnętrzne narzędzia do ekstrakcji danych. Wszystkie narzędzia korzystają z tych poleceń w tle, a inżynierowie danych powinni być w stanie rozwiązywać ewentualne problemy.

1.2.2. Przekształcanie danych

Dane pobrane z systemów źródłowych są przekształcane z ich surowego formatu w użyteczne i przydatne informacje zgodnie z wymaganiami użytkowników. Inżynierowie danych zazwyczaj otrzymują reguły transformacji od analityków danych, architektów systemów lub czasami bezpośrednio od użytkowników. W mniejszych zespołach inżynierii danych inżynierowie mogą współpracować z użytkownikami biznesowymi w celu zdefiniowania zasad przekształcania danych.

NARZĘDZIA DO PRZEKSZTAŁCANIA DANYCH

Narzędzia używane przez inżynierów danych do przekształcania danych są rozmaite. Mogą one przeprowadzać transformacje za pomocą kwerend SQL, procedur składowanych lub funkcji. Niektóre narzędzia do pobierania danych z systemów źródłowych również obsługują przekształcenia danych i mogą być wykorzystywane w tym celu.

Inżynierowie danych mogą też pisać kod w usłudze Snowpark, co umożliwia im tworzenie przekształceń w jednym z obsługiwanych języków programowania, takim jak Python, Java czy Scala. Mogą także korzystać z popularnych otwartoźródłowych bibliotek do przekształcania danych.

SPRAWDZANIE POPRAWNOŚCI DANYCH

Oprócz transformacji danych inżynierowie danych dodają do potoków etapy walidacji, dzięki czemu odbiorcy mogą mieć pewność, że otrzymywane dane są wiarygodne i spójne. Potoki danych często zawierają techniczne reguły walidacji, takie jak sprawdzanie naruszeń kluczy głównych lub obcych, ponieważ Snowflake nie wymusza ograniczeń kluczy głównych ani obcych w standardowych tabelach. Odbiorcy mogą definiować dodatkowe reguły walidacji danych. Dane, które nie spełniają tych reguł, mogą zostać oznaczone lub wykluczone z dalszych analiz.

1.2.3. Udostępnianie danych odbiorcom końcowym

Po pobraniu danych z systemów źródłowych i odpowiednim ich przekształceniu są one prezentowane odbiorcom w postaci dostosowanej do narzędzi i rozwiązań używanych w dalszych etapach przetwarzania. Inżynierowie danych zaspokajają potrzeby biznesowe odbiorców, dostarczając dane zgodnie z oczekiwaniami dotyczącymi terminowości i wydajności.

Inżynierowie danych tworzą odpowiednie interfejsy użytkownika tam, gdzie są one potrzebne, takie jak API, jeśli ma to zastosowanie, lub prezentują dane w modelu odpowiednim dla narzędzia używanego do analizy. Dostarczają również metadane, aby ułatwić odnajdywanie danych, na przykład przez udostępnianie komentarzy do tabel i atrybutów w bazie danych oraz innych istotnych informacji, takich jak między innymi czas ostatniej aktualizacji danych, liczba zaimportowanych rekordów i wskaźniki jakości danych (jeśli są stosowane).

W przeciwieństwie do wielu tradycyjnych baz danych kwerendy SQL w środowisku Snowflake nie wymagają ręcznej optymalizacji, która zwykle pochłania sporo czasu administratorów. Tutaj większość kwerend działa wydajnie bez dodatkowych zabiegów. Niemniej zdarzają się sytuacje, w których administratorzy i inżynierowie danych mogą poprawić wydajność kwerend o wysokiej złożoności, operujących na dużych zestawach danych lub zużywających zbyt wiele zasobów. W takich przypadkach inżynierowie mogą zwiększyć wydajność, wykorzystując funkcje Snowflake'a takie jak klasteryzacja, optymalizacja wyszukiwania, widoki zmaterializowane czy tabele dynamiczne.

1.2.4. Stosowanie podstawowych składników

Choć inżynieria danych zasadniczo obejmuje pozyskiwanie danych z systemów źródłowych, wprowadzanie ich do środowiska Snowflake i przekształcanie na potrzeby dalszego wykorzystania, to inżynierowie danych stosują również dodatkowe komponenty w zależności od potrzeb.

ZABEZPIECZENIA

Niezależnie od tego, czy rozwiązania IT są przechowywane w chmurze, czy lokalnie, mogą być narażone na zagrożenia i luki w zabezpieczeniach. Ponieważ Snowflake działa w chmurze, konieczne są dodatkowe środki ostrożności, aby zapobiec nieautoryzowanemu dostępowi. Administratorzy Snowflake'a zazwyczaj zabezpieczają dostęp na różne sposoby, takie jak ograniczenie adresów IP, z których użytkownicy mogą łączyć się z platformą, lub wymaganie silnych metod uwierzytelniania, na przykład za pomocą uwierzytelniania wieloskładnikowego.

Podczas tworzenia potoków danych inżynierowie nieuchronnie w pewnym momencie łączą się z platformą Snowflake. Wszelkie dane uwierzytelniające używane do połączenia z nią są bezpiecznie przechowywane w skrytkach dostawcy chmury lub lokalnie w bezpiecznym miejscu, które nigdy nie jest udostępniane innym użytkownikom ani zapisywane w repozytorium, gdzie inni mogliby je zobaczyć. Inżynierowie danych muszą działać odpowiedzialnie, aby zagwarantować, że dane dostępowe do konta Snowflake'a nie zostaną celowo ani przypadkowo udostępnione komukolwiek innemu.

Użytkownicy powinni mieć minimalne uprawnienia niezbędne do wykonania zadania. Nawet gdy inżynierom danych przyznaje się bardziej zaawansowane role administracyjne w Snowflake'u do wykonywania określonych zadań, nie powinni oni korzystać z tych ról domyślnie. Powinni używać swoich wyznaczonych ról programistycznych i świadomie przełączać się na bardziej zaawansowane role administracyjne tylko wtedy, gdy jest to konieczne do wykonania konkretnego zadania administracyjnego.

Inżynierowie danych rozumieją zasady kontroli dostępu opartej na rolach w Snowflake'u i używają odpowiednich ról podczas tworzenia obiektów w tym systemie. Choć niekoniecznie muszą być odpowiedzialni za konfigurację mechanizmu kontroli dostępu, to powinni znać jego strukturę, aby móc prawidłowo z niego korzystać.

ZARZĄDZANIE DANYMI

Większość organizacji posiada zasady zarządzania danymi, które ograniczają nieupoważnionym użytkownikom dostęp do poufnych lub wrażliwych informacji. Czasami, szczególnie w środowiskach o wysokim poziomie bezpieczeństwa i ścisłej regulacji, inżynierowie danych nie otrzymują dostępu do rzeczywistych danych produkcyjnych, lecz pracują na próbkach danych, na danych z zamaskowanymi informacjami wrażliwymi lub na podzbiorach, z których usunięto poufne informacje.

Inżynierowie danych współpracują z architektami danych i właścicielami domen, aby zrozumieć wymagania dotyczące zarządzania danymi. Przestrzegają tych wymagań podczas budowania potoków danych. Na przykład mogą przechowywać wrażliwe atrybuty w oddzielnych tabelach z bardziej restrykcyjnymi regułami dostępu.

MODELOWANIE DANYCH

W zależności od wielkości organizacji, w której pracują inżynierowie danych, może przypaść im obowiązek modelowania danych. W dużych organizacjach często istnieją wyspecjalizowane zespoły lub eksperci odpowiedzialni za tworzenie modeli danych. W mniejszych inżynierowie danych mogą modelować dane w zakresie swoich obowiązków.

Niezależnie od tego, kto tworzy model danych, inżynier danych musi rozumieć koncepcje modelowania, aby poprawnie wczytywać dane do docelowych modeli. Na przykład modele mogą wymagać, aby potoki danych generowały klucze zastępcze podczas procesu wczytywania danych.

Oto przykłady standardowych modeli danych stosowanych w rozwiązaniach hurtowni danych i analityce:

- modele relacyjne dla korporacyjnych hurtowni danych;
- modele zespołów danych, takie jak skarbce danych (ang. *data vault*) czy modelowanie kotwicowe (ang. *anchor modeling*), które są szczególnie przydatne jako korporacyjne hurtownie danych przy pobieraniu informacji z wielu źródeł;
- modele danych wymiarowych dla składnic danych i narzędzi do tworzenia raportów.

Niektóre modele danych umożliwiają stosowanie powtarzalnych wzorców ładowania, które automatyzują proces wprowadzania danych. Inżynierowie danych implementują takie wzorce, co pomaga im tworzyć bardziej ustandaryzowane i wydajne potoki danych.

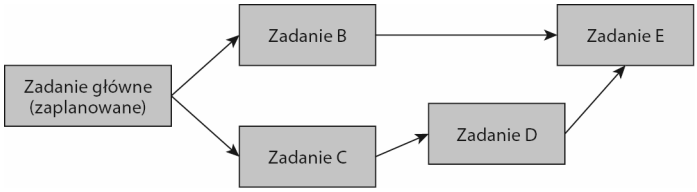
ORKIESTRACJA

Większość potoków danych działa według harmonogramu ustalonego z właścicielami systemów źródłowych i odbiorcami końcowymi. W zależności od potrzeb użytkowników potoki mogą być uruchamiane masowo albo w mikropartiach, na przykład codziennie, co godzinę lub co kilka minut.

Gdy system zawiera wiele potoków danych, inżynierowie danych projektują zależności między potokami i ich składnikami. Mogą oni wykorzystywać zewnętrzne narzędzia do koordynowania przepływu danych, polecenia systemu operacyjnego lub zadania Snowflake'a do planowania potoków w uporządkowany sposób.

Zadania w Snowflake'u umożliwiają tworzenie skierowanego grafu acyklicznego (ang. *Directed Acyclic Graph* — DAG), który zawiera serię powiązanych ze sobą zadań. Każdy DAG zaczyna się od jednego zadania głównego, a zadania zależne są wykonywane w tym samym kierunku bez tworzenia pętli. Rysunek 1.3 przedstawia przykładowy DAG zadań w Snowflake'u.

Inżynierowie danych monitorują zaplanowane potoki, aby upewnić się, że działają one bez błędów i poprawnie ładują dane. Podczas projektowania potoków danych uwzględniają mechanizmy rejestrowania zdarzeń, które umożliwiają



Rysunek 1.3. *Skierowany graf acykliczny (DAG) zadań. Graf składa się z zadania głównego oraz zadań zależnych, które rozpoczynają się po zakończeniu poprzednich zadań*

skuteczne rozwiązywanie problemów w przypadku wystąpienia błędów. Projektują również potoki danych w taki sposób, aby możliwe było ponowne uruchomienie całego procesu od początku lub wznowienie go od miejsca, w którym wystąpił błąd.

INŻYNIERIA OPROGRAMOWANIA

Inżynieria oprogramowania stanowi znaczną część pracy inżynierów danych. Tworzą oni, testują i utrzymują potoki przetwarzania danych, które często są rozbudowanymi systemami składającymi się z kodu programistycznego. Nawet jeśli większość tego kodu to kwerendy SQL, a nie tradycyjne konstrukcje języków programowania, nadal stosuje się zasady inżynierii oprogramowania, takie jak konwencje nazewnictwa, modularyzacja kodu, czytelność kodu, wersjonowanie i łatwość utrzymania systemu.

Inżynierowie danych tworzą również wydajne, solidne i niezawodne systemy. Weźmy na przykład potok danych, który zostałby przypadkowo uruchomiony dwukrotnie. W takiej sytuacji nie powinno to prowadzić do duplikowania danych, a potok nie powinien zostać uruchomiony, gdy poprzednia instancja tego samego potoku jest wciąż w trakcie przetwarzania.

Istotnym elementem inżynierii danych jest zrozumienie i wykorzystanie różnych środowisk w cyklu życia oprogramowania. Organizacje zazwyczaj korzystają ze środowiska programistycznego, w którym programiści pracują nad swoim kodem. Następnie kod jest wdrażany do środowiska testowego, gdzie sprawdza się, czy utworzone potoki danych działają zgodnie z oczekiwaniami. Niektóre duże firmy mogą również posiadać środowiska do testów akceptacyjnych. Po przetestowaniu i zatwierdzeniu kodu jest on wdrażany do środowiska produkcyjnego.

DATAOPS

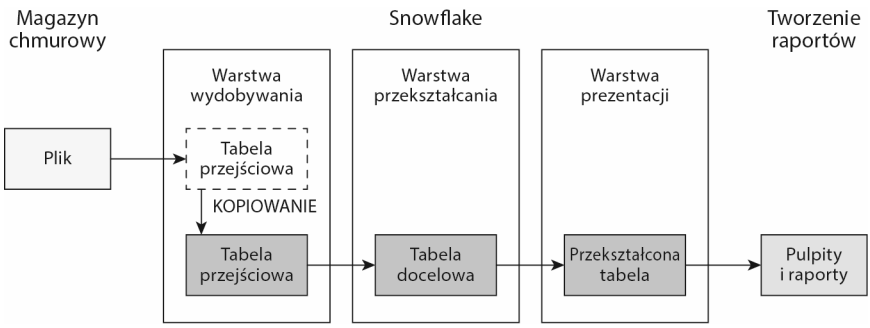
DataOps to ogólne pojęcie wywodzące się z metodyki zwinnego tworzenia oprogramowania, które ma zastosowanie w całym cyklu życia danych, obejmującym wydobywanie, pozyskiwanie, przekształcanie i prezentowanie. Łączy ona procesy i technologie, które poprawiają jakość oraz umożliwiają ciągły rozwój i integrację potoków inżynierii danych.

DataOps, podobnie jak DevOps w inżynierii oprogramowania, zapewnia ciągłe dostarczanie działającego oprogramowania poprzez wykorzystanie narzędzi i technologii umożliwiających pracę zespołową. Na przykład zespoły inżynierii danych korzystają z repozytoriów kodu, takich jak Git, zintegrowanych z procesami CI/CD. Mogą one również automatyzować testy kodu potoków danych i walidację samych danych. Nawet w małych organizacjach, gdzie pracuje tylko jeden inżynier danych, praktyki DataOps pomagają lepiej zorganizować i ustrukturyzować pracę, zapewniając spójne i ciągłe wdrażanie do środowiska produkcyjnego.

1.3. Budowanie potoków danych

Potok danych (ang. *data pipeline*) to zestaw składników przetwarzających dane, który przenosi je ze źródła do miejsca docelowego. W zależności od wymagań potok może również wykonywać przekształcanie danych. Każdy potok danych jest inny. Na przykład inżynier danych pobiera pliki z chmurowego magazynu obiektów do środowiska Snowflake i udostępnia dane z tych plików odbiorcom końcowym, którzy tworzą na ich podstawie raporty.

Rysunek 1.4 przedstawia przykładowy potok danych. Pliki z chmurowego magazynu obiektów są udostępniane Snowflake'owi poprzez zewnętrzne miejsce przechowywania i wczytywane do **tabeli przejściowej** (ang. *staging table*). Następnie potok ładuje dane do tabeli docelowej, przeprowadza niezbędne transformacje i umieszcza przetworzone dane w odpowiedniej tabeli. Na koniec udostępnia te dane do dalszego raportowania, kwerendowania lub analizowania przez użytkowników końcowych.



Rysunek 1.4. *Potok danych, który pobiera plik danych z chmurowego dostawcy pamięci masowej, wczytuje dane do tabeli Snowflake'a i przekształca je na potrzeby tworzenia raportów*

Przykład przedstawiony na rysunku 1.4 to jeden z najczęściej spotykanych potoków inżynierii danych w środowisku Snowflake. Wykorzystuje on miejsce zewnętrzne wskazujące na lokalizację w chmurowym magazynie obiektów. Następnie używa polecenia `COPY` do wczytania danych z magazynu chmurowego do

tabeli przejściowej Snowflake'a. Potem dane te są ładowane do tabeli docelowej poprzez dołączanie ich za pomocą polecenia `INSERT` lub scalanie za pomocą polecenia `MERGE`. Tabela docelowa jest zwykle projektowana, jeszcze zanim inżynier danych zbuduje potok danych jako część modelu danych warstwy przekształcania.

Wszelkie niezbędne transformacje są przeprowadzane za pomocą instrukcji SQL lub procedur składowanych, a przetworzone dane są zapisywane w przeznaczonej do tego tabeli. Na koniec dane są prezentowane w formacie odpowiednim do sporządzania raportów. Zamiast budować etapy transformacji danych w potokach, inżynierowie danych mogą tworzyć widoki SQL do przekształcania danych, ponieważ widoki są łatwe w utrzymaniu i nie wymagają ponownego ładowania danych. Jednak widoki mogą nie zapewniać wydajności wymaganej przez użytkowników. W takich sytuacjach inżynier danych decyduje o najlepszym podejściu do poprawy wydajności, często poprzez materializację danych.

W praktyce stosuje się wiele rodzajów potoków danych wykorzystujących różne składniki. Omówię je dokładniej w kolejnych rozdziałach.

1.4. Inżynieria danych w zastosowaniach Snowflake'a

Najczęstsze zastosowania inżynierii danych z wykorzystaniem platformy Snowflake obejmują jeziora danych, hurtownie danych, składnice danych, siatki danych, analizę biznesową, danetykę oraz dogenerowywanie danych przy użyciu dużych modeli językowych.

JEZIORO DANYCH

Jezioro danych (ang. *data lake*) to repozytorium przechowujące duże ilości danych o pełnej, częściowej lub zerowej strukturze. Snowflake umożliwia przechowywanie danych w różnych typach tabel, w tym w tabelach standardowych, zewnętrznych, hybrydowych i innych. Wiele równoczesnych zadań może uzyskiwać dostęp do danych bez rywalizacji o zasoby. Użytkownicy mogą odpytywać dane relacyjne lub częściowo ustrukturyzowane w jeziorze danych za pomocą składni SQL lub innych języków obsługiwanych w środowisku programistycznym Snowpark.

HURTOWNIA DANYCH

Hurtownia danych (ang. *data warehouse*) to korporacyjne repozytorium danych gromadzonych z jednego bądź wielu systemów źródłowych lub zewnętrznych źródeł, takich jak Snowflake Marketplace. Głównym celem hurtowni danych jest dostarczanie ujednoliconych danych historycznych do tworzenia raportów, pulpitów i analiz dla użytkowników biznesowych. Umożliwia to użytkownikom

biznesowym wyciąganie z tych danych wniosków, które stanowią podstawę do podejmowania decyzji biznesowych.

SKŁADNICA DANYCH

Składnice danych (ang. *data mart*) przechowują informacje dla pojedynczych działów biznesowych, takich jak marketing czy finanse. Działy te mogą szybko i łatwo uzyskać dostęp do danych w składnicach, zamiast przeszukiwać całą hurtownię danych. Ponieważ budowa pełnej hurtowni danych może pochłonąć znaczne zasoby i zająć dużo czasu, niektóre organizacje decydują się na tworzenie wyłącznie składnic bez budowania centralnej hurtowni danych.

SIATKA DANYCH

Siatka danych (ang. *data mesh*) to stosunkowo nowe podejście do analizy danych, w którym zamiast monolitycznej hurtowni danych poszczególne zespoły dziedzinowe tworzą własne produkty danych. Często przypominają one składnice danych. Dzięki lepszym narzędziom oraz łatwiejszym w obsłudze platformom i technologiom użytkownicy biznesowi mogą samodzielnie przygotowywać i eksplorować dane. Na przykład mogą oceniać jakość danych w swoich jednostkach biznesowych. Mogą również wykonywać proste przekształcenia. W rezultacie część zadań związanych z inżynierią danych przechodzi od inżynierów danych na użytkowników biznesowych.

Nawet jeśli zadania związane z inżynierią danych zostaną przekazane użytkownikom biznesowym, zapotrzebowanie na wykwalifikowanych inżynierów danych nie zniknie. Wiele technicznych aspektów potoków danych wymaga solidnej wiedzy z zakresu inżynierii oprogramowania, której użytkownicy biznesowi mogą nie posiadać ze względu na inny obszar specjalizacji. Inżynierowie danych nadal odgrywają istotną rolę w podejściu opartym na siatce danych, jednak ich rola może się przenieść z działu IT do odpowiednich działów biznesowych.

Snowflake może przechowywać produkty danych, a funkcja udostępniania danych w niej umożliwia ekspozycję produktów danych w całej organizacji.

ANALITYKA BIZNESOWA

Analityka biznesowa (ang. *business intelligence*) to proces analizowania danych i uzyskiwania na ich podstawie informacji umożliwiających podejmowanie decyzji. Rozwiązania z zakresu analityki biznesowej zazwyczaj czerpią dane z firmowych hurtowni danych, składnic danych, produktów danych lub kombinacji tych źródeł.

DANETYKA

Danetyka (ang. *data science*) to szeroka dziedzina obejmująca zastosowania takie jak zaawansowana analityka, sztuczna inteligencja i uczenie maszynowe. Zasadniczo danetyka ma na celu odkrywanie użytecznych informacji ze źródeł

danych, które są wykorzystywane do podejmowania decyzji i planowania strategicznego w organizacjach. Usługa Snowpark umożliwia danetykom budowanie modeli przy użyciu wybranego języka programowania, szczególnie Pythona, który ma bogaty wybór bibliotek do uczenia maszynowego. Na przykład danetycy mogą tworzyć, trenować i wdrażać modele uczenia maszynowego w środowisku Snowpark, korzystając z popularnych bibliotek takich jak scikit-learn, TensorFlow czy PyTorch. Wraz z wprowadzeniem środowiska Snowflake Cortex niektóre funkcje uczenia maszynowego, takie jak prognozowanie szeregów czasowych, wykrywanie anomalii czy klasyfikacja, są dostępne bezpośrednio z poziomu poleceń SQL w usłudze Snowflake.

DOGENEROWYWANIE DANYCH PRZY UŻYCIU DUŻYCH MODELI JĘZYKOWYCH

Generatywna sztuczna inteligencja to technologia, która tworzy tekst, obrazy, kod programistyczny i inne rodzaje treści. **Duże modele językowe** (ang. *Large Language Models* — LLM) to podkategoria generatywnej sztucznej inteligencji specjalizująca się w tworzeniu tekstu. Wiele zastosowań biznesowych, takich jak klasyfikacja tekstu, analiza dokumentów, ocena nastroju czy tłumaczenie maszynowe, opiera się na LLM-ach. Wyniki generowane przez LLM-y mogą wzbogacać dane w różnych rozwiązaniach platformy Snowflake.

Programiści mogą korzystać z funkcji zewnętrznych modeli językowych w usłudze Snowpark, wywołując odpowiednie API i zapisując wyniki w tabelach Snowflake'a do dalszej analizy. Wraz z wprowadzeniem środowiska Snowflake Cortex wybrane funkcje LLM-ów, takie jak ocena nastroju, uzupełnianie podpowiedzi, streszczanie, tłumaczenie, wydobywanie odpowiedzi z nieustrukturyzowanego tekstu czy osadzanie wektorowe stały się dostępne bezpośrednio z poziomu poleceń SQL w usłudze Snowflake.

Podsumowanie

- Snowflake to nowoczesna platforma danych w chmurze, która doskonale nadaje się do rozwiązań wymagających intensywnego przetwarzania danych, takich jak jeziora danych, hurtownie danych, składnice danych, siatki danych, analityka biznesowa, danetyka oraz dogenerowywanie danych przy użyciu modeli językowych.
- Inżynierowie danych w środowisku Snowflake są odpowiedzialni za projektowanie, tworzenie, koordynowanie, optymalizowanie, monitorowanie i utrzymywanie potoków danych. Potoki te przetwarzają surowe dane z systemów źródłowych, przekształcając je w użyteczne informacje gotowe do dalszego wykorzystania.

- Pierwszym krokiem w procesie przetwarzania danych jest ich pozyskiwanie z systemów źródłowych. Inżynierowie danych uzyskują dostęp do tych systemów i wprowadzają dane do Snowflake'a, korzystając z własnoręcznie napisanych poleceń lub używając specjalistycznych narzędzi.
- Dane pobrane z systemów źródłowych na Snowflake'a są przekształcane ze stanu surowego do postaci i formatu odpowiednich dla odbiorców końcowych.
- Transformacje danych można przeprowadzać za pomocą narzędzi, instrukcji SQL lub procedur składowanych, a także kodować w usłudze Snowpark przy użyciu jednego z obsługiwanych języków programowania.
- Dane dla odbiorców końcowych są prezentowane w formie, która ułatwia dostęp do nich i efektywne ich wykorzystanie w narzędziach raportowych lub innych rozwiązaniach, takich jak modele uczenia maszynowego w środowisku Snowpark.
- Potok danych to zestaw etapów przetwarzania, który przenosi dane ze źródła do miejsca docelowego, wykonując niezbędne transformacje po drodze.
- Podczas tworzenia potoków danych dla Snowflake'a inżynierowie danych wykorzystują podstawowe komponenty, takie jak zabezpieczenia, zarządzanie danymi, modelowanie danych, orkiestracja, inżynieria oprogramowania oraz metodyka DataOps.
- Najczęstsze zastosowania inżynierii danych z wykorzystaniem Snowflake'a to jeziora danych, hurtownie danych, składnice danych, siatki danych, analityka biznesowa, danetyka oraz dogenerowywanie danych przy użyciu modeli językowych.

Snowflake

Pozycja obowiązkowa, jeżeli chcesz osiągnąć wyżyny w dziedzinie inżynierii danych!
— **Isabella Renzetti**, ekspertka z dziedziny analizy danych

Snowflake to kompleksowa platforma chmurowa do przechowywania i analizy danych, oferująca niemal nieograniczoną skalowalność i szybkie, elastyczne usługi obliczeniowe. Umożliwia tworzenie i rozwijanie potoków danych, a jej nowe funkcjonalności, takie jak wyszukiwanie wektorowe, automatyczne konwersje tekstu do SQL czy generowanie kodu, korzystają z technik AI. Jeśli pracujesz z danymi, Snowflake otwiera przed Tobą zupełnie nowe możliwości.

Niesamowita książka! Przejdziesz od zera do środowiska produkcyjnego Snowflake!
— **Doyle Turner**, Microsoft

Z tą książką krok po kroku będziesz rozwijać umiejętności potrzebne do wykonywania codziennych zadań z zakresu inżynierii danych w Snowflake. Utworzysz swój pierwszy potok, a potem będziesz go rozbudowywać o coraz bardziej zaawansowane funkcje: zarządzanie danymi i bezpieczeństwem, integrację CI/CD czy wzbogacanie danych przy użyciu generatywnej AI. To praktyczny poradnik pełen kodu, przykładów i wskazówek, które pozwolą Ci w krótkim czasie wejść na zupełnie nowy poziom pracy z danymi.

Wyczerpująca, aktualna i wypełniona praktycznymi fragmentami kodu!
— **Albert Nogués**, Danone

W książce:
- **wprowadzanie danych z usług chmurowych za pomocą API lub ze Snowflake Marketplace**
- **orkiestracja potokami danych za pomocą strumieni i zadań**
- **optymalizacja wydajności i kosztów**
- **projektowanie mechanizmów kontroli dostępu i ochrony danych**
- **wdrażanie ciągłe oraz ciągła integracja obiektów i kodu Snowflake**
- **wzbogacanie danych za pomocą generatywnej AI**

Mistrzostwo! Odkryjesz rzeczywisty potencjał platformy Snowflake!
— **Shankar Narayanan**, Microsoft

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-2898-5
9 788328 928985
Cena: 99,00 zł

Maja Ferle jest architektką danych z ponad 30-letnim stażem w dziedzinie analizy danych, hurtowni danych, analityki biznesowej, inżynierii danych, modelowania danych i administracji bazami danych. Posiada certyfikaty SnowPro Advanced Data Engineer i SnowPro Advanced Data Analyst. Jest również ekspertką Snowflake (uzyskała certyfikaty SnowPro Subject Matter Expert i Snowflake Data Superhero).