Spis treści

Część II. Zagłębiamy się w Trino

Przedmowa

Czy naprawdę minęła już ponad dekada? Wydaje się, że to było wczoraj, a jednocześnie bardzo dawno temu. Pracę nad Presto w firmie Facebook rozpoczęliśmy w 2012, a nad niektórymi komponentami i pomysłami jeszcze wcześniej. Z pewnością planowaliśmy opracować coś przydatnego. Zawsze chcieliśmy mieć udany projekt open source i społeczność. W 2013 opublikowaliśmy pierwszą wersję Presto na licencji Apache. Tak zaczęła się nasza przygoda.

Od tamtego czasu odeszliśmy z firmy Facebook, porzuciliśmy nazwę Presto i przekroczyliśmy nasze najśmielsze oczekiwania. Jesteśmy dumni z osiągnięć społeczności skupionej wokół tego projektu, a także zaszczyceni pozytywnym przyjęciem oraz pomocą, jakiej nam udzielono.

Projekt Trino rozrósł się niesamowicie i dostarczył wielkiej wartości ogromnej społeczności użytkowników. Członkowie tej społeczności pochodzą z całego świata, m.in. z Brazylii, Kanady, Chin, Niemiec, Indii, Izraela, Polski, Singapuru, Stanów Zjednoczonych, Wielkiej Brytanii i innych krajów.

Kolejnym ważnym krokiem było założenie fundacji Trino Software Foundation na początku 2019 i usamodzielnienie projektu. Ta organizacja not-for-profit dba o udoskonalanie rozproszonego silnika SQL w ramach projektu open source Trino. Fundacja ma gwarantować, że przez kolejne dziesięciolecia projekt pozostanie niezależny, otwarty i będzie rozwijany w ramach współpracy.

Zmiana nazwy projektu i fundacji na Trino pod koniec 2020 była kolejnym krokiem milowym dla społeczności. Mamy nową nazwę, nowy wygląd, nowe zainteresowanie z całego świata, a nawet wspaniałą maskotkę, naszego ukochanego Komendanta Bun Bun. Obecnie, ponad trzy lata po założeniu fundacji możemy spojrzeć wstecz i podziwiać mnóstwo ulepszeń wprowadzonych przez ogromną i stale rosnącą społeczność.

Cieszymy się, że Matt, Manfred i Martin napisali tę uaktualnioną, drugą wersję książki o Trino z pomocą wydawnictwa O'Reilly. Zawiera ona świetne wprowadzenie do Trino i uczy wszystkiego, co jest potrzebne do jego użycia.

Zachęcamy do ruszenia w podróż w głąb Trino i powiązany z nim świat analiz biznesowych, tworzenia pulpitów, hurtowni danych, eksploracji danych, uczenia maszynowego i innych zagadnień.

Oczywiście zachęcamy również do zapoznania się z dodatkowymi zasobami i pomocą oferowaną w witrynie Trino, dostępnej pod adresem *https://trino.io*, z czatem społeczności, z repozytorium źródłowym, z kanałem Trino Community Broadcast i innymi zasobami. Witamy w społeczności Trino!

— Dain Sundstrom i David Phillips
twórcy projektu Trino
i założyciele Trino Software Foundation

Wprowadzenie

Trino: Profesjonalny przewodnik jest przede wszystkim książką o rozproszonym silniku zapytań Trino. Książka jest przeznaczona zarówno dla początkujących, jak i zaawansowanych użytkowników Trino. Czytelnicy powinni już mieć pewne doświadczenie z bazami danych i SQL-em, w przeciwnym razie powinni równolegle z książką zapoznać się z tymi zagadnieniami. Niezależnie od doświadczenia, z pewnością każdy nauczy się czegoś nowego z tej książki.

W tej drugiej wersji uwzględniono najnowsze ulepszenia w Trino. Opisaliśmy nowe aspekty, takie jak pakiety Helm służące do wdrażania klastrów Trino w środowisku Kubernetes, nowe konektory Iceberg i Delta Lake dla nowoczesnych architektur typu lakehouse, odporne na awarie wykonywanie zapytań, rozszerzone funkcje języka SQL, a także ostatnie wydanie Trino, obecnie działające w środowisku Java 17.

Pierwsza część książki zawiera wprowadzenie do Trino i opisuje, jak szybko skonfigurować i uruchomić Trino. Obejmuje instalację i pierwsze użycie interfejsu wiersza poleceń, a także aplikacje klienckie i webowe, takie jak narzędzia do zarządzania bazami danych SQL, tworzenia pulpitów i raportów, z wykorzystaniem sterownika JDBC.

W drugiej części książki opisujemy architekturę Trino, wdrażanie klastra, wiele konektorów do źródeł danych, a także uwzględniamy mnóstwo informacji o najważniejszej funkcji Trino – przeszukiwaniu dowolnego źródła danych za pomocą SQL.

Trzecia część książki obejmuje inne aspekty dotyczące wdrażania Trino w środowisku produkcyjnym. Omawiamy korzystanie z interfejsu webowego Trino, konfigurację zabezpieczeń i niektóre przypadki użycia Trino w innych organizacjach.

Konwencje używane w książce

W tej książce wykorzystywane są następujące konwencje typograficzne:

Kursywa
: Jest używana do oznaczania nowych terminów, adresów URL, nazw plików i rozszerzeń plików.

`Czcionka o stałej szerokości`
: Jest używana w listingach programów, a także do oznaczenia elementów programu, takich jak nazwy zmiennych lub funkcji, baz danych, typów danych, zmiennych środowiskowych i słów kluczowych.

`Pogrubiona czcionka o stałej szerokości`
: Pokazuje polecenia lub inny tekst, który użytkownik powinien wpisać.

`Pochyła czcionka o stałej szerokości`
: Pokazuje tekst, który należy zastąpić wartością podaną przez użytkownika lub zależną od kontekstu.

Ten element oznacza wskazówkę lub sugestię.

Ten element oznacza ogólną uwagę.

Ten element jest ostrzeżeniem.

Przykładowy kod, pozwolenia i atrybucja

Dodatkowe materiały dla tej książki są szczegółowo opisane w podrozdziale „Repozytorium dla książki” na stronie 15.

W przypadku pytań technicznych lub problemów z przykładowym kodem, zachęcamy do kontaktu na czacie społeczności – opisanym w podrozdziale „Czat społeczności” na stronie 13 – lub do zgłoszenia problemu w repozytorium tej książki.

Ta książka ma pomagać w pracy. Przykładowy kod można wykorzystać w swoich programach i dokumentacji. Czytelnik nie musi nas prosić o pozwolenie, o ile nie zamierza wykorzystać znacznej części kodu. Przykładowo napisanie programu zawierającego kilka fragmentów kodu z tej książki nie wymaga pozwolenia. Sprzedaż lub dystrybucja przykładów z książek wydawnictwa O’Reilly wymaga pozwolenia. Odpowiedź na pytanie zawierające cytat z tej książki oraz przykładowy kod nie wymaga pozwolenia. Umieszczenie znacznej ilości przykładowego kodu z tej książki w dokumentacji produktu wymaga pozwolenia.

Doceniamy, ale nie wymagamy atrybucji. Atrybucja zwykle obejmuje tytuł, autora, wydawnictwo i ISBN. Przykładowo: „*Trino: Profesjonalny przewodnik*, wydanie 2., Matt

Fuller, Manfred Moser i Martin Traverso (O'Reilly). Copyright 2023 Matt Fuller, Martin Traverso i simpligility technologies, Inc., 978-1-098-13723-6."

Jeśli ktoś uważa, że będzie używał przykładów kodu w sposób wymagający uzyskania pozwolenia, powinien się z nami skontaktować pisząc pod adres *permissions@oreilly.com*.

Podziękowania

Chcielibyśmy podziękować wszystkim ze społeczności Trino – począwszy od założycieli projektu, zespołu zajmującego się jego utrzymaniem oraz wszystkim współtwórcom, którzy rozwijają projekt od czasu jego przejęcia. Nasza wdzięczność sięga znacznie dalej. Jesteśmy wdzięczni wszystkim użytkownikom Trino, tym, którzy dzielą się wiedzą o tej platformie, pomagają innym użytkownikom, pracują nad powiązanymi projektami i integracjami, a także tym, którzy zadają pytania na naszym czacie. Cieszymy się, że należymy do społeczności i mamy nadzieję, że w przyszłości będziemy mogli wspólnie odnieść wiele sukcesów.

Krytyczną częścią społeczności Trino jest firma Starburst. Chcielibyśmy podziękować wszystkim w firmie Starburst za ich pomoc. Doceniamy pracę, zasoby, stabilność i wsparcie, jakie oferuje firma Starburst, jej klientów używających Trino, oraz autorów, którzy należą do zespołu tej firmy.

W odniesieniu do tej książki chcielibyśmy podziękować wszystkim, którzy podsuwali nam pomysły, dostarczali informacji i sporządzali recenzje. Ponadto, autorzy chcieliby wyrazić osobiste podziękowania:

- Matt chciałby podziękować swojej żonie Meghan oraz trójce swoich dzieci, Emily, Hannah i Liamowi, za ich cierpliwość i wsparcie podczas pracy nad książką. Podekscytowanie dzieci, wynikające z faktu, że ich tata będzie „autorem", pomagało Mattowi w ciągu długich weekendów i wieczorów.
- Manfred chciałby podziękować swojej żonie Yen i trzem synom, Lukasowi, Nikolasowi i Tobiasowi, nie tylko za tolerowanie technicznego bełkotu, ale za szczere zainteresowanie i pasję do technologii, pisania i nauki.
- Martin chciałby podziękować swojej żonie Melinie oraz czwórce dzieci, Marcosowi, Victorii, Joaquinowi i Martinie za ich wsparcie i entuzjazm przez ostatnich 10 lat pracy nad Trino.

Część I

Wprowadzenie do Trino

Trino jest silnikiem zapytań umożliwiającym dostęp do każdego zbioru danych za pomocą języka SQL. Trino można wykorzystać do przeszukiwania ogromnych zbiorów danych, dzięki horyzontalnemu skalowaniu przetwarzania zapytań.

W pierwszej części książki poznamy Trino i przypadki jego użycia. Następnie nauczymy się prostego sposobu instalacji i uruchamiania Trino. Na koniec poznamy narzędzia, za pomocą których można się połączyć z Trino i przeszukiwać dane. Skorzystamy z podstawowej konfiguracji, która pozwoli na szybkie rozpoczęcie pracy z Trino.

Rozdział 1

Wstęp do Trino

Zakładam, że Czytelnik usłyszał gdzieś o Trino i znalazł tę książkę. A może podczas przeglądania tej pierwszej części zastanawiał się, czy warto się zagłębić w tę tematykę. W tym rozdziale opisane są problemy, na jakie możemy natrafić w miarę jak dysponujemy coraz większą ilością danych, które kryją w sobie ukrytą wartość. Trino jest kluczowym silnikiem, umożliwiającym analizę wszystkich danych i dostęp do nich za pomocą skutecznych narzędzi wykorzystujących język Structured Query Language (SQL).

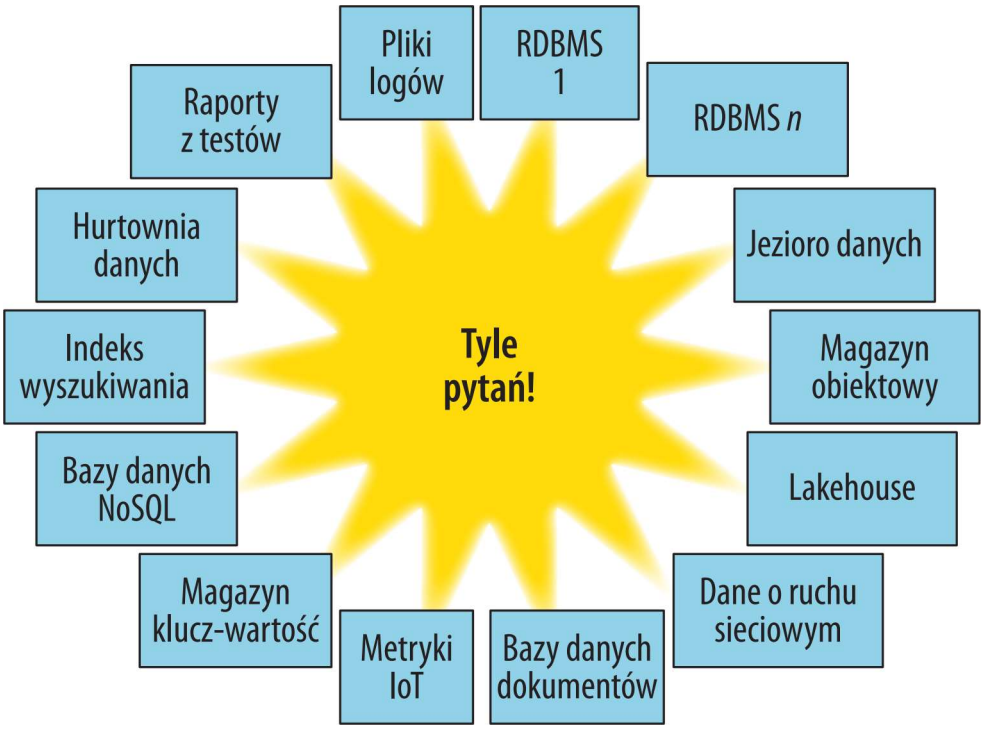
Sposób, w jaki został zaprojektowany Trino oraz dostępne w nim funkcje, umożliwiają uzyskanie lepszego wglądu w dane, znacznie wykraczającego poza dotychczasowe możliwości. Odbywa się to szybciej, a ponadto można uzyskać informacje niedostępne w przeszłości ze względu na wysoki koszt lub zbyt długi czas ich pozyskania. Wszystko to można osiągnąć za pomocą mniejszej ilości zasobów, a zatem znacznie taniej, dzięki czemu można uzyskać jeszcze więcej informacji!

Opiszemy także inne zasoby, wykraczające poza zakres tej książki, ale mamy nadzieję, że Czytelnik najpierw zapozna się z jej treścią.

Problemy związane z wielkimi zbiorami danych

Wszyscy zbierają coraz więcej danych z urządzeń, ze śledzenia zachowania użytkowników, transakcji biznesowych, danych o lokalizacji, procedur testowania oprogramowania i systemów, i z wielu innych źródeł. Informacje uzyskane na podstawie analizy danych i ich wykorzystanie mogą przyczynić się do powodzenia dowolnej inicjatywy, a nawet firmy.

Jednocześnie znacznie wzrosła różnorodność mechanizmów magazynowania danych. Mamy do dyspozycji relacyjne bazy danych, bazy danych NoSQL, bazy danych opartych na dokumentach, magazyny typu klucz-wartość, obiektowe systemy magazynowe, itd. Wiele z nich odgrywa kluczową rolę we współczesnych organizacjach i nie można się już opierać tylko na jednej, wybranej technologii. Jak widać na rysunku 1-1, przetwarzanie danych może być trudnym, przytłaczającym zadaniem.



Rysunek 1-1 *Wielkie zbiory danych bywają przytłaczające*

Ponadto, danych zgromadzonych w tych wszystkich systemach nie da się przeszukiwać i analizować za pomocą standardowych narzędzi. Wszędzie mamy do czynienia z różnymi językami zapytań i narzędziami analitycznymi dla niszowych systemów. Natomiast analitycy biznesowi są zwykle przyzwyczajeni do korzystania ze standardowego języka SQL. Wiele potężnych narzędzi wykorzystuje SQL do analizy, tworzenia pulpitów, atrakcyjnych raportów i innych zadań związanych z analizą biznesową.

Dane znajdują się w różnych silosach, a niektórych z nich nie da się analizować z wymaganą wydajnością. Inne systemy, w przeciwieństwie do nowoczesnych aplikacji w chmurze, zapisują dane w monolitycznych systemach, których nie da się horyzontalnie skalować. Rezygnując z tych możliwości, ograniczamy liczbę potencjalnych przypadków użycia i użytkowników, a zatem przydatność danych.

W organizacjach na całym świecie zauważono, że tworzenie i utrzymywanie ogromnych hurtowni danych jest niezwykle kosztowne. Okazało się również, że dla wielu użytkowników i przypadków użycia to podejście było zbyt powolne i niewygodne. Często wybierane rozwiązanie, polegające na utworzeniu jeziora danych, prowadziło do powstania swego rodzaju wysypiska danych, których nikt nie przeglądał lub musiał podjąć ogromny wysiłek, aby je przeanalizować. Nawet nowe podejście, polegające na wykorzystaniu magazynu typu lakehouse, rozwiązania łączącego zalety hurtowni i jeziora danych, nie będzie ostatecznym remedium. Dane nadal będą rozproszone, zapisywane w wielu miejscach i będą się pojawiać w coraz większej liczbie systemów.

Widać wyraźnie, że pojawiają się ogromne możliwości dla systemu, który pozwoli wykorzystać całą wartość zgromadzoną w danych.

Trino na ratunek

Trino może rozwiązać wszystkie wspomniane problemy, a nawet stworzyć nowe perspektywy dzięki zapytaniom federacyjnym do odmiennych systemów, zapytaniom równoległym, horyzontalnemu skalowaniu klastrów i wielu innym funkcjom. Na rysunku 1-2 pokazano logo projektu Trino.

Rysunek 1-2 *Logo Trino z maskotką Komendantem Bun Bun*

Trino jest rozproszonym silnikiem zapytań SQL, dystrybuowanym w modelu open source. Został zaprojektowany i napisany z myślą o skutecznym przeszukiwaniu danych zgromadzonych w źródłach wszystkich rozmiarów, od gigabajtów po petabajty. Trino burzy rzekomy wybór między możliwością szybkiej analizy z użyciem drogich komercyjnych narzędzi a użyciem powolnych „darmowych" rozwiązań wymagających dużych zasobów sprzętowych.

Stworzony z myślą o wydajności i skalowalności

Trino jest narzędziem, które zostało zaprojektowane z myślą o wydajnym przeszukiwaniu ogromnej ilości danych z wykorzystaniem procesów rozproszonych. Jeśli trzeba przeszukać wiele terabajtów, a nawet petabajtów danych, często używa się narzędzi, takich jak Apache Hive, które współpracuje z platformą Hadoop i z magazynem Hadoop Distributed File System (HDFS), lub podobnych systemów, takich jak Amazon Simple Storage Service (S3) i innych udostępnianych przez dostawców rozwiązań chmurowych, lub używanych w samodzielnych instalacjach. Trino jest alternatywą dla tych narzędzi, umożliwiającą skuteczniejsze przeszukiwanie danych.

Z Trino powinni korzystać analitycy, którzy oczekują odpowiedzi na zapytanie SQL w ciągu kilku milisekund podczas analiz w czasie rzeczywistym, a w innych przypadkach w ciągu kilku sekund lub minut. Trino obsługuje SQL, język zwykle używany w hurtowniach danych oraz analizie danych, agregując ogromne ilości danych i tworząc raporty. Te zadania zwykle określa się mianem *online analytical processing* (OLAP).

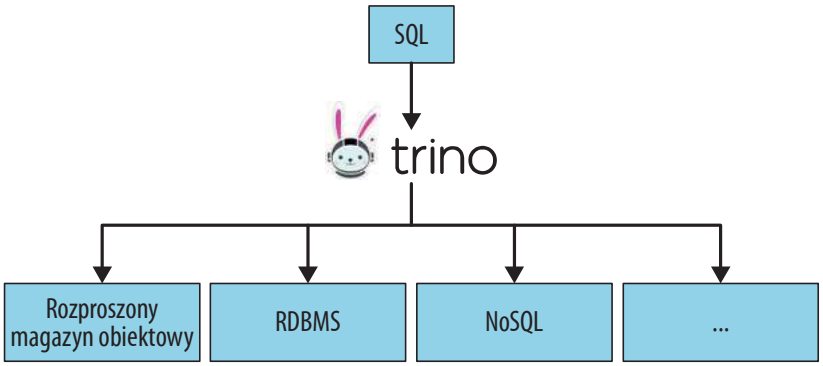
Chociaż Trino rozumie i wydajnie przetwarza zapytania SQL, *nie* jest bazą danych, ponieważ nie zawiera własnego systemu magazynowania. Nie jest to relacyjna baza danych ogólnego przeznaczenia, która mogłaby zastąpić Microsoft SQL Server, Oracle Database, MySQL lub PostgreSQL. Ponadto, Trino nie obsługuje procesów *online transaction processing* (OLTP). To samo można napisać o innych bazach danych, zaprojektowanych i zoptymalizowanych pod kątem hurtowni danych lub analizy, takich jak Teradata, Netezza, Vertica i Amazon Redshift.

Trino wykorzystuje zarówno dobrze znane, jak i nowe techniki rozproszonego przetwarzania danych. Obejmują one równoległe przetwarzanie w pamięci, przetwarzanie w potokach w węzłach klastra, model wielowątkowego przetwarzania wykorzystujący wszystkie rdzenie CPU, wydajne płaskie struktury danych w pamięci minimalizujące odśmiecanie pamięci w języku Java oraz generowanie kodu bajtowego języka Java. Szczegółowy opis tych skomplikowanych wewnętrznych mechanizmów Trino wykracza poza zakres tej książki. Użytkownicy Trino zyskują dzięki nim szybszy wgląd w dane, za ułamek kosztów innych rozwiązań.

SQL do wszystkiego

Trino początkowo miał przeszukiwać dane z magazynu HDFS. Robi to bardzo skutecznie, o czym przekonamy się później. Ale to nie wszystko. Wręcz przeciwnie, Trino jest silnikiem zapytań, który może przeszukiwać dane z magazynów obiektowych, systemów zarządzania relacyjnymi bazami danych (Relational Database Management System – RDBMS), baz danych NoSQL i innych systemów, zgodnie z rysunkiem 1-3.

Trino przeszukuje dane tam, gdzie się znajdują i nie wymaga ich przenoszenia do jednego miejsca. A zatem Trino umożliwia przeszukiwanie danych w magazynach HDFS i innych rozproszonych obiektowych systemach magazynowych. Umożliwia przeszukiwanie systemów RDBM i innych źródeł danych. Dzięki możliwości przeszukiwania danych tam, gdzie się znajdują, Trino może zastąpić tradycyjne, drogie i obciążające procesy ekstrakcji, transformacji i ładowania (extract, transform, load – ETL). Jako minimum, może ułatwić nam pracę udostępniając silnik zapytań, pozwalający na uruchomienie wysoce wydajnych procesów ETL za pomocą nowoczesnych narzędzi, takich jak dbt, we wszystkich źródłach danych oraz zmniejszyć, a nawet uniknąć konieczności przeprowadzania procesów ETL. A zatem Trino zdecydowanie nie jest kolejnym rozwiązaniem typu SQL dla platformy Hadoop.



Rysunek 1-3 *Obsługa SQL dla zróżnicowanych źródeł danych za pomocą Trino*

Obiektowe systemy magazynowe obejmują Amazon S3, Microsoft Azure Blob Storage, Google Cloud Storage i magazyny zgodne z S3, takie jak MinIO i Ceph. Te systemy mogą wykorzystywać tradycyjny format tabel Hive, a także nowsze rozwiązania, takie jak Delta Lake i Iceberg. Trino może przeszukiwać tradycyjne systemy RDBMS, takie jak Microsoft

SQL Server, PostgreSQL, MySQL, Oracle, Teradata i Amazon Redshift. Trino może też przeszukiwać systemy NoSQL, takie jak Apache Cassandra, Apache Kafka, MongoDB lub Elasticsearch. Trino może przeszukiwać właściwie wszystko i jest prawdziwym systemem typu SQL do wszystkiego.

Dla użytkowników oznacza to, że nie muszą już korzystać ze specyficznych języków zapytań lub narzędzi do interakcji z danymi w tych systemach. Mogą po prostu użyć Trino oraz znanego języka SQL, a także znanych już narzędzi do analizy, sporządzania pulpitów i raportów. Narzędzia te, oparte na języku SQL umożliwiają analizę dodatkowych zbiorów danych z oddzielnych systemów, które w przeciwnym razie pozostają niedostępne. Użytkownicy mogą nawet za pomocą Trino przeszukiwać dane z różnych systemów korzystając ze znanego już języka SQL.

Oddzielenie magazynu danych od zasobów służących do przetwarzania zapytań

Trino nie jest bazą danych z magazynem; jest to narzędzie do przeszukiwania danych tam, gdzie się znajdują. W Trino magazynowanie oraz przetwarzanie danych są od siebie oddzielone i można je niezależnie skalować. Trino stanowi warstwę obliczeniową, a przetwarzane źródła danych są warstwą magazynową.

Dzięki temu zasoby obliczeniowe Trino można skalować w górę i w dół podczas przetwarzania zapytań, w zależności od potrzeb analitycznych. Nie trzeba przenosić danych, zapewniać mocy obliczeniowych i magazynowych pasujących dokładnie do potrzeb bieżących zapytań, ani regularnie ich zmieniać ze względu na zmienne potrzeby zapytań.

Trino może skalować możliwości obliczeniowe zapytań poprzez dynamiczne skalowanie klastra obliczeniowego. Dane można przeszukiwać w miejscu, w którym znajduje się ich źródło. Dzięki temu można doskonale optymalizować potrzebne zasoby sprzętowe, a co za tym idzie, zmniejszać koszty analiz.

Przypadki użycia Trino

Elastyczność oraz możliwości Trino pozwalają na samodzielny wybór sposobu użycia Trino oraz korzyści, jakie chcemy osiągnąć. Można zacząć od jednego niewielkiego projektu, pozwalającego rozwiązać określony problem. Tak zaczyna większość użytkowników Trino.

Gdy wraz z innymi użytkownikami Trino w swojej organizacji zapoznamy się z korzyściami i funkcjonalnością tego narzędzia, odkryjemy nowe scenariusze. Wieści będą się rozchodzić i szybko okaże się, że istnieje wiele potrzeb, które można zaspokoić za pomocą Trino, analizując dane z różnorodnych źródeł.

W tym podrozdziale omówiono kilka przypadków użycia. Pamiętajmy, że można je rozszerzyć na własne scenariusze. Z drugiej strony, można użyć Trino tylko do rozwiązania jednego problemu. Przygotujmy się jednak na to, że ostatecznie będziemy częściej korzystać z Trino.

Jeden analityczny punkt dostępu za pomocą SQL

Systemy RDBMS oraz język SQL są dostępne od dawna i okazały się bardzo przydatne. Żadna organizacja nie może się bez nich obejść. W rzeczywistości większość firm korzysta z wielu systemów. Prawdopodobnie oprogramowanie w naszej firmie opiera się na wielkich komercyjnych bazach danych, takich jak Oracle Database lub IBM DB2. Mogą być używane również otwarte systemy, takie jak MariaDB lub PostgreSQL, a także własne aplikacje firmy. Klienci oraz analitycy prawdopodobnie zmagają się z wieloma następującymi problemami:

- Czasem nie wiadomo, gdzie znajdują się dane, nawet te gotowe do użycia. Tylko wiedza plemienna w firmie lub lata doświadczenia z wewnętrznymi systemami umożliwiają znalezienie właściwych danych.
- Przeszukiwanie różnych baz danych wymaga używania różnych połączeń, a także innych zapytań wykorzystujących różne dialekty SQL. Są one na tyle podobne, że wyglądają tak samo, ale zachowują się wystarczająco odmiennie, by wprowadzić zamieszanie i wymagają poznania szczegółów.
- Danych z różnych systemów nie można połączyć w jednym zapytaniu bez użycia hurtowni danych.

Te problemy można rozwiązać za pomocą Trino. Wszystkie bazy danych można udostępnić w jednym miejscu: w Trino.

Wszystkie systemy można przeszukiwać za pomocą jednego standardu SQL – wystandaryzowanego języka SQL, funkcji i operatorów obsługiwanych przez Trino.

Wszystkie narzędzia do tworzenia pulpitów i analiz, a także inne systemy wykorzystywane do przeprowadzania analiz biznesowych mogą korzystać z jednego systemu Trino i mieć dostęp do wszystkich danych z organizacji.

Punkt dostępu do hurtowni danych i systemów źródłowych

Gdy w organizacjach pojawia się potrzeba lepszego zrozumienia i analizy danych z wielu systemów RDBMS, zwykle pojawia się pomysł utworzenia i utrzymywania systemów hurtowni danych. Pobieranie danych z różnych systemów odbywa się następnie poprzez złożone procesy ETL i ostatecznie, za pośrednictwem długotrwałych zadań wsadowych, dane są magazynowane w ściśle kontrolowanej, olbrzymiej hurtowni danych.

Chociaż w wielu przypadkach jest to pomocne rozwiązanie, to jednak analitycy danych mogą się natknąć na nowe problemy:

- Obok istniejących już baz danych pojawia się kolejny punkt wejścia dla narzędzi i zapytań.
- Dane, które są potrzebne w określonym momencie, nie są dostępne w hurtowni danych. Umieszczenie w niej danych jest trudnym, pełnym przeszkód i kosztownym procesem.

Trino pozwala na dodanie dowolnej hurtowni danych jako źródła danych, podobnie jak każdej innej relacyjnej bazy danych.

Jeśli ktoś chce się zagłębić w zapytanie do hurtowni danych, może to zrobić w tym samym systemie. W jednym miejscu można uzyskać dostęp do hurtowni danych *oraz* do źródłowego systemu bazy danych, a nawet pisać zapytania obejmujące te wszystkie systemy. Za pomocą Trino można przeszukiwać dowolną bazę danych w tym samym systemie, hurtownię danych, źródłową i dowolną inną bazę danych.

Dostęp do wszystkiego za pomocą SQL

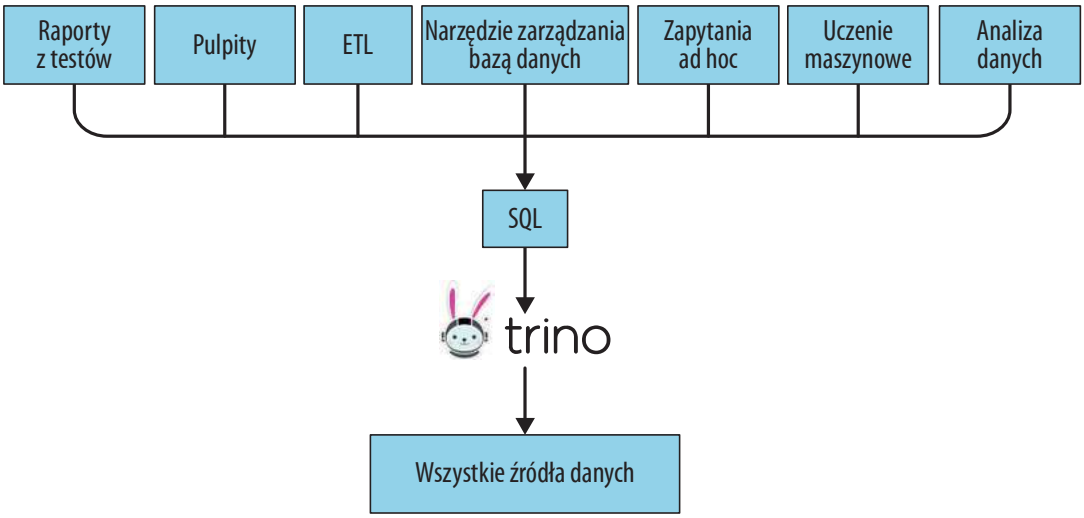
Minęły już czasy, gdy używano tylko systemów RDBMS. Dane są obecnie magazynowane w wielu różnorodnych systemach, zoptymalizowanych pod kątem określonych przypadków użycia. Magazyny obiektowe, magazyny typu klucz-wartość, bazy danych oparte na dokumentach, grafowe bazy danych, systemy generujące strumienie zdarzeń i inne systemy NoSQL zapewniają unikalne funkcje i korzyści.

W wielu firmach używa się co najmniej kilku z nich, magazynując dane, które są niezbędne do zrozumienia i doskonalenia działalności biznesowej.

Oczywiście wszystkie te systemy wymagają użycia różnych narzędzi i technologii do przeszukiwania i wykorzystywania danych.

Jest to skomplikowane zadanie, wymagające długiej nauki. Jednak najczęściej okazuje się, że analizujemy dane tylko powierzchownie, bez ich głębszego zrozumienia. Nie dysponujemy dobrym sposobem analizy danych. Trudno znaleźć narzędzia do wizualizacji i głębszej analizy danych, a często ich po prostu brakuje.

Natomiast Trino umożliwia połączenie się z tymi wszystkimi systemami i traktowanie ich jak źródła danych. Udostępnia dane do analizy za pomocą języka SQL w standardzie ANSI (American National Standards Institute), a także zawiera wszystkie narzędzia wykorzystujące SQL, co pokazano na rysunku 1-4.



Rysunek 1-4 *Jeden punkt dostępu przez SQL do wszystkich źródeł danych, dla wielu przypadków użycia*

Dzięki Trino zrozumienie danych z tych wszystkich, bardzo zróżnicowanych systemów, jest znacznie prostsze, a często możliwe po raz pierwszy.

Zapytania federacyjne

Udostępnienie wszystkich silosów danych w Trino jest olbrzymim krokiem ku zrozumieniu danych. Wszystkie można przeszukiwać za pomocą języka SQL i standardowych narzędzi. Jednak zwykle pytania, na które chcemy znaleźć odpowiedź, wymagają zajrzenia do silosu danych, znalezienia w nich pewnych informacji i połączenia w jednym miejscu.

W Trino można to osiągnąć za pomocą zapytań federacyjnych. *Zapytanie federacyjne* jest zapytaniem SQL, które w jednej instrukcji odwołuje się do baz danych i schematów z zupełnie różnych systemów. Wszystkie źródła danych z Trino można przeszukiwać jednocześnie, za pomocą jednego zapytania SQL.

Można definiować relacje między danymi ze śledzenia użytkowników z magazynu obiektowego a danymi o klientach z systemu RDBMS. Jeśli magazyn typu klucz-wartość zawiera więcej powiązanych informacji, można go również uwzględnić w zapytaniu.

Za pomocą zapytań federacyjnych w Trino można uzyskać taki wgląd w dane, jaki nie jest dostępny za pomocą innych środków.

Warstwa semantyczna dla wirtualnej hurtowni danych

Systemy hurtowni danych nie tylko przynoszą olbrzymie korzyści użytkownikom, ale są też ciężarem dla organizacji:

- Utrzymywanie hurtowni danych jest olbrzymim, kosztownym przedsięwzięciem.
- Wyznaczone zespoły utrzymują hurtownię danych oraz związane z nią procesy ETL.
- Umieszczenie danych w hurtowni wymaga walki z biurokracją i zwykle zajmuje zbyt dużo czasu.

Natomiast Trino może służyć jako wirtualna hurtownia danych. Za pomocą tego jednego narzędzia i standardu ANSI języka SQL można zdefiniować swoją warstwę semantyczną. Wystarczy skonfigurować wszystkie źródła danych w postaci baz danych w Trino i można przystąpić do analiz. Trino zapewnia niezbędne moce obliczeniowe, potrzebne do przeszukiwania magazynów baz danych. Za pomocą języka SQL oraz obsługiwanych funkcji i operatorów Trino może zwrócić potrzebne dane prosto ze źródła. Nie trzeba ich kopiować, przenosić, ani przekształcać w celu przygotowania do analizy.

Dzięki obsłudze standardowego języka SQL do przeszukiwania wszystkich połączonych źródeł danych, można utworzyć własną warstwę semantyczną, z której będą korzystać narzędzia i użytkownicy do przeszukiwania danych w prostszy sposób. Ta warstwa może obejmować wszystkie źródła danych bez konieczności migracji żadnych danych. Trino może przeszukiwać dane na poziomie źródłowym i magazynowym. Dzięki traktowaniu Trino jako „hurtowni danych tworzonej w locie" organizacje mogą udoskonalić

swoje istniejące hurtownie danych, wzbogacając je o nowe możliwości. Pozwala to nawet uniknąć tworzenia i utrzymywania hurtowni danych.

Silnik zapytań w jeziorze danych

Termin *jezioro danych* jest zwykle stosowany w odniesieniu do ogromnych systemów HDFS lub podobnych rozproszonych obiektowych systemów magazynowych, w których umieszcza się wszelkiego rodzaju dane bez przejmowania się dostępem do nich. Dzięki Trino tego typu magazyny można traktować jako przydatne hurtownie danych. W rzeczywistości, Trino początkowo rozwijano w firmie Facebook, gdzie narzędzie to służyło do obsługi szybszych i potężniejszych zapytań do ogromnych hurtowni danych Hadoop, których nie udało się obsłużyć za pomocą tabel Hive i innych narzędzi. W ten sposób powstał konektor Hive dla Trino, opisany w podrozdziale „Konektor Hive dla rozproszonych źródeł danych" na stronie 100.

Nowoczesne jeziora danych zwykle używają innych obiektowych systemów magazynowania, wykraczających poza magazyn HDFS, udostępnianych przez dostawców chmury i inne projekty open source. Trino może je obsługiwać za pomocą konektora Hive, dzięki czemu rozszerza możliwości analizy jeziora danych za pomocą SQL, niezależnie od jego lokalizacji i sposobu magazynowania danych.

Dzięki powstaniu nowych formatów tabel, takich jak Delta Lake i Iceberg, niesłychanie wzrosły możliwości użycia jezior danych, co doprowadziło do utworzenia terminu *data lakehouse*. Dzięki konektorom Delta Lake i Iceberg, Trino jest pierwszorzędnym wyborem, jeśli chodzi i przeszukiwanie magazynów typu lakehouse.

Przekształcenia SQL i ETL

Dzięki obsłudze systemów RDBMS i innych podobnych magazynów danych, Trino można wykorzystać do przenoszenia danych. Za pomocą SQL i bogatego zbioru dostępnych funkcji SQL można przeszukiwać dane, przekształcać je, a następnie zapisywać w tym samym lub w dowolnym innym źródle danych.

W praktyce oznacza to, że można skopiować dane ze swojego obiektowego systemu magazynowego lub z magazynu typu klucz-wartość do systemu RDBMS i wykorzystać w dalszych analizach. Oczywiście, dane można też przekształcić i je zagregować, aby uzyskać na ich podstawie nowe informacje.

Z drugiej strony, często pobiera się dane z systemu RDBMS, lub z systemu generującego strumienie zdarzeń, takiego jak Kafka, i przenosi się je do jeziora danych, aby zmniejszyć uciążliwość korzystania z systemów RDBMS podczas przeszukiwania danych przez wielu analityków. Procesy ETL, obecnie często określane mianem *przygotowywania danych*, mogą być ważnym etapem tego procesu, poprzez ulepszanie danych i tworzenie modelu danych, lepiej nadającego się do przeszukiwania i analizy.

W tym przypadku Trino odgrywa kluczową rolę w ogólnym zarządzaniu danymi, a dzięki wykonywaniu zapytań w sposób odporny na awarie w nowoczenych magazynach typu lakehouse, Trino doskonale nadaje się do tego typu zadań.

Lepszy wgląd dzięki szybszym wynikom

Zadawanie skomplikowanych pytań i używanie ogromnych zbiorów danych podlega ograniczeniom. Kopiowanie danych i wczytywanie ich do swojej hurtowni danych w celu analizy może się okazać zbyt kosztowne. Wykonanie wszystkich zapytań może wymagać zbyt wielkich mocy obliczeniowych lub może zająć wiele dni.

Trino z założenia unika kopiowania danych. Przetwarzanie równoległe i wysoka optymalizacja prowadzi do lepszej wydajności analiz przeprowadzanych z użyciem Trino.

Jeśli zapytanie, którego wykonanie wcześniej trwało trzy dni, obecnie jest przetwarzane w ciągu 15 minut, warto rozważyć jego wykonanie. Wiedza uzyskana na podstawie jego wyników może przynieść korzyści i umożliwić wykonanie kolejnych zapytań.

Szybszy czas przetwarzania zapytań przez Trino pozwala na wykonanie lepszych analiz i uzyskanie lepszych wyników.

Wielkie zbiory danych, uczenie maszynowe i sztuczna inteligencja

Udostępnianie przez Trino większej ilości danych dla standardowych narzędzi na bazie SQL, a także skalowanie przetwarzania ogromnych zbiorów danych sprawia, że jest to podstawowe narzędzie do przetwarzania wielkich zbiorów danych. Obecnie zwykle obejmuje to analizę statystyczną i coraz bardziej złożone zadania z dziedziny uczenia maszynowego i systemów sztucznej inteligencji. Dzięki obsłudze języka R i innych narzędzi Trino zdecydowanie może odegrać znaczącą rolę w tych przypadkach użycia.

Inne przypadki użycia

W poprzednich podrozdziałach opisano ogólnie przypadki użycia Trino. Stale pojawiają się nowe przypadki oraz ich kombinacje.

W rozdziale 13. opisano szczegółowo użycie Trino w znanych firmach i organizacjach. Informacje te zostały umieszczone pod koniec książki, aby umożliwić najpierw zdobycie wiedzy potrzebnej do zrozumienia posiadanych danych.

Zasoby dotyczące Trino

Oprócz tej książki dostępnych jest wiele innych zasobów dotyczących Trino. W tym podrozdziale przedstawiono istotne źródła, od których można zacząć. Większość z nich zawiera mnóstwo informacji oraz odwołania do kolejnych zasobów.

Witryna

Fundacja *Trino Software Foundation* zarządza społecznością otwartego projektu Trino i utrzymuje poświęconą mu witrynę. Strona główna jest pokazana na rysunku 1-5. Witryna zawiera dokumentację, dane kontaktowe, blog społeczności z wpisami o nowościach

i wydarzeniach, odcinki Trino Community Broadcast w formacie dźwiękowym i wideo, a także inne informacje. Jest dostępna pod adresem *https://trino.io*.

Rysunek 1-5 *Strona główna witryny Trino dostępnej pod adresem trino.io*

Dokumentacja

Szczegółowa dokumentacja Trino wchodzi w skład kodu źródłowego i jest dostępna w witrynie projektu. Obejmuje ogólne informacje, a także szczegóły referencyjne o obsłudze SQL, funkcjach i operatorach, konektorach, konfiguracji i wielu innych. Dostępne są też informacje o poszczególnych wydaniach, zawierające szczegóły o ostatnich zmianach. Dokumentacja jest dostępna pod adresem *https://trino.io/docs*.

Czat społeczności

Społeczność początkujących, zaawansowanych i doświadczonych użytkowników, a także programistów rozwijających i utrzymujących Trino jest bardzo pomocna i aktywnie udziela się na czacie dostępnym pod adresem *https://trinodb.slack.com*.

Najpierw trzeba dołączyć do kanału *announcements*, a następnie sprawdzić inne kanały dotyczące różnorodnych zagadnień, np. pomocy dla początkujących, selekcji błędów, kolejnych wydań, rozwoju i innych, umożliwiających dyskusję na temat platformy Kubernetes lub określonych konektorów.

Matt, Manfred i Martin uczestniczą w dyskusjach na czacie społeczności niemal codziennie. Zachęcamy do dołączenia.