O'REILLY®

Helion

Microsoft Fabric od podstaw

Kompleksowe projektowanie nowoczesnej analityki danych

Nikola Ilic
Ben Weissman

Tytuł oryginału: Fundamentals of Microsoft Fabric: Designing End-to-End Analytics Solutions

Tłumaczenie: Grzegorz Werner

ISBN: 978-83-289-3552-5

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
helion.pl/user/opinie/mifaod
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

ROZDZIAŁ 3.

Wszystkie drogi prowadzą do OneLake

Jedną z kluczowych cech platformy Fabric jest to, że jest ona zorientowana na jezioro danych. Wszystkie dane są przechowywane w jednym jeziorze danych o nazwie OneLake. W tym rozdziale omówimy podstawy jezior danych oraz szczegóły dotyczące OneLake.

Wprowadzenie do jezior danych

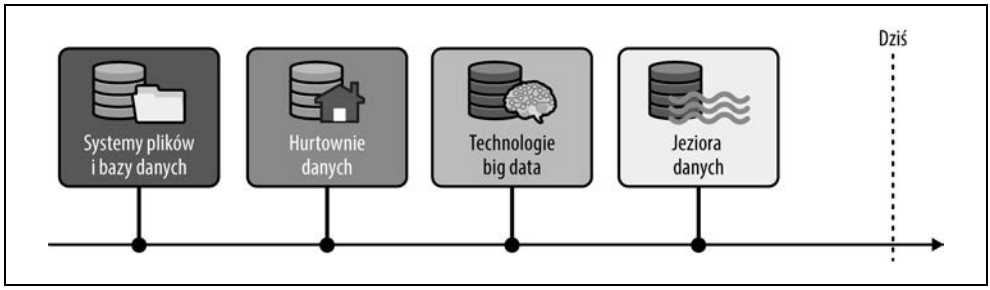
Jezioro danych to scentralizowane repozytorium umożliwiające przechowywanie danych ustrukturyzowanych, częściowo ustrukturyzowanych i niestrukturyzowanych o dowolnej skali. W przeciwieństwie do tradycyjnych hurtowni danych, które przechowują dane w predefiniowanych schematach, jeziora danych są zaprojektowane do przechowywania ogromnych ilości surowych danych w ich natywnym formacie aż do momentu, gdy będą potrzebne. Ta elastyczność pozwala na obsługę różnorodnych typów danych, w tym tekstu, obrazów, filmów i strumieni z mediów społecznościowych, co czyni jeziora danych integralną częścią nowoczesnych architektur big data.

Głównym celem jeziora danych jest zapewnienie skalowalnego i ekonomicznego rozwiązania do przechowywania dużych ilości danych. Dane te mogą być przetwarzane i analizowane w celu wydobywania cennych informacji, przeprowadzania analiz w czasie rzeczywistym oraz wspierania zastosowań z dziedziny data science i uczenia maszynowego. Struktura jeziora danych pozwala firmom przechowywać wszystkie dane w jednym miejscu, co umożliwia kompleksową analizę i integrację różnych źródeł danych.

Ewolucja rozwiązania do przechowywania danych

Rozwiązania do przechowywania danych znacząco ewoluowały na przestrzeni lat w miarę rosnącej złożoności i skali potrzeb w zakresie zarządzania danymi.

Rozwój rozwiązań do przechowywania danych przedstawiono na rysunku 3.1.



Rysunek 3.1. Ewolucja systemów przechowywania danych

Podróż zaczęła się od prostych systemów plików i baz danych, przeszła przez hurtownie danych, a obecnie dotarła do jezior danych i trwa dalej.

Systemy plików i bazy danych

Początkowo dane przechowywano głównie w systemach plików i relacyjnych bazach danych. Systemy te były odpowiednie do zarządzania małymi i średnimi ilościami ustrukturyzowanych danych, ale miały problemy z danymi nieustrukturyzowanymi i skalowalnością.

Hurtownie danych

Pojawienie się hurtowni danych stanowiło znaczący postęp w przechowywaniu informacji. Systemy te są zoptymalizowane pod kątem operacji odczytu i zaprojektowane do wspierania działań związanych z analityką biznesową. Hurtownie danych przechowują ustrukturyzowane informacje w predefiniowanym schemacie, co ułatwia efektywne zapytania i raportowanie. Są jednak mniej elastyczne w obsłudze danych nieustrukturyzowanych lub częściowo ustrukturyzowanych.

Technologie big data

Wraz z eksplozją danych z różnych źródeł, takich jak media społecznościowe, czujniki i urządzenia mobilne, tradycyjne systemy przechowywania danych okazały się ograniczone. Technologie big data, takie jak Hadoop, wprowadziły koncepcję rozproszonego przechowywania i przetwarzania, umożliwiając obsługę ogromnych zbiorów danych na klastrach standardowego sprzętu.

Jeziora danych

Jeziora danych opierają się na zasadach big data, ale są bardziej elastyczne i skalowalne. Mogą przechowywać dane w surowej formie, obsługiwać różne typy danych i stanowić podstawę do zaawansowanej analityki. Jeziora danych integrują się z nowoczesnymi platformami do przetwarzania danych, pozwalając organizacjom na efektywniejsze wyciąganie wniosków z posiadanych informacji.

Tradycyjne systemy pamięci masowej, takie jak relacyjne bazy danych, organizują informacje w ustrukturyzowanych tabelach z predefiniowanymi schematami, co sprawdza się idealnie w systemach transakcyjnych i ustrukturyzowanych kwerendach. Wymagają one oczyszczenia i sformatowania danych przed ich zapisaniem w celu zapewnienia spójności i niezawodności.

Natomiast jezioro danych jest zaprojektowane do przechowywania ogromnych ilości surowych, nieustrukturyzowanych i częściowo ustrukturyzowanych danych w otwartym formacie z zachowaniem ich pierwotnej struktury przed jakąkolwiek transformacją. Ta elastyczność pozwala na obsługę różnorodnych formatów, w tym tekstu, obrazów i plików dziennika, co czyni je idealnymi do analityki big data i uczenia maszynowego. Dzięki funkcji stosowania schematu podczas odczytu jeziora danych pozwalają naukowcom danych na eksplorowanie i przetwarzanie informacji bez sztywnych ograniczeń schematów, często z użyciem zoptymalizowanych formatów przechowywania, takich jak Parquet, do efektywnego tworzenia kwerend i analiz.

Znaczenie jezior danych

W dzisiejszym zorientowanym na dane świecie firmy wykorzystują je do podejmowania decyzji, wprowadzania innowacji i utrzymywania przewagi konkurencyjnej. Jeziora danych odgrywają kluczową rolę w tym środowisku z kilku powodów:

Skalowalność

Jeziora danych są zaprojektowane do obsługi dużych ilości danych i dostosowują się do rosnącego napływu informacji z różnych źródeł bez utraty wydajności.

Elastyczność

W przeciwieństwie do tradycyjnych hurtowni danych jeziora danych obsługują szeroki zakres typów i formatów danych. Ta elastyczność pozwala firmom przechowywać wszystko, od dzienników transakcji, po pliki multimedialne, umożliwiając kompleksową analizę danych.

Efektywność kosztowa

Jeziora danych używają ekonomicznych rozwiązań pamięci masowej, często korzystając z usług chmurowych. Dzięki temu organizacje mogą przechowywać duże zbiory danych bez ponoszenia wysokich kosztów.

Zaawansowana analityka

Zaawansowana analityka to zbiór zaawansowanych technik analizy danych, w tym uczenia maszynowego, modelowania predykcyjnego i sztucznej inteligencji, używanych do odkrywania głębszych spostrzeżeń i trendów wykraczających poza tradycyjne raportowanie. Jeziora danych, przechowujące dane w surowej formie, stanowią cenny zasób dla naukowców i analityków danych.

Integracja danych

Jeziora danych umożliwiają integrację informacji z wielu źródeł, tworząc jednolity widok, który umożliwia bardziej świadome podejmowanie decyzji. Ta integracja jest niezbędna w takich zastosowaniach jak pełne profilowanie klientów (kompleksowy, ujednolicony obraz klienta, który łączy dane z wielu źródeł, aby zapewnić całościowe zrozumienie jego interakcji, zachowań i preferencji), spersonalizowany marketing i optymalizacja operacyjna.

Przetwarzanie w czasie rzeczywistym

Nowoczesne jeziora danych obsługują przetwarzanie danych w czasie rzeczywistym, pozwalając firmom szybko reagować na zmieniające się warunki i podejmować terminowe decyzje. Ta funkcja jest kluczowa dla zastosowań takich jak wykrywanie oszustw, zarządzanie łańcuchem dostaw i dynamiczne ustalanie cen.

Nadzór i zgodność z przepisami

Jeziora danych mogą obejmować solidne rozwiązania ramowe do zarządzania danymi, zapewniając ich jakość, bezpieczeństwo i zgodność z przepisami. Staje się to coraz ważniejsze dla firm, które muszą poruszać się po złożonym środowisku regulacyjnym.

Podsumowując, jeziora danych stanowią znaczący postęp w przechowywaniu danych i zarządzaniu nimi. Eliminują ograniczenia wcześniejszych systemów i są skalowalnym, elastycznym i ekonomicznym rozwiązaniem dla nowoczesnych przedsiębiorstw. Umożliwiając przechowywanie różnorodnych danych i typów plików — a następnie ich przetwarzanie za pomocą różnych narzędzi — jeziora danych pozwalają organizacjom w pełni wykorzystać potencjał danych, a przez to szybciej wprowadzać innowacje i utrzymać przewagę konkurencyjną w świecie opartym na danych. W miarę wzrostu ilości i różnorodności danych jeziora danych, wspierające nową generację aplikacji i rozwiązań analitycznych, będą odgrywać coraz ważniejszą rolę w ekosystemie przedsiębiorstw.

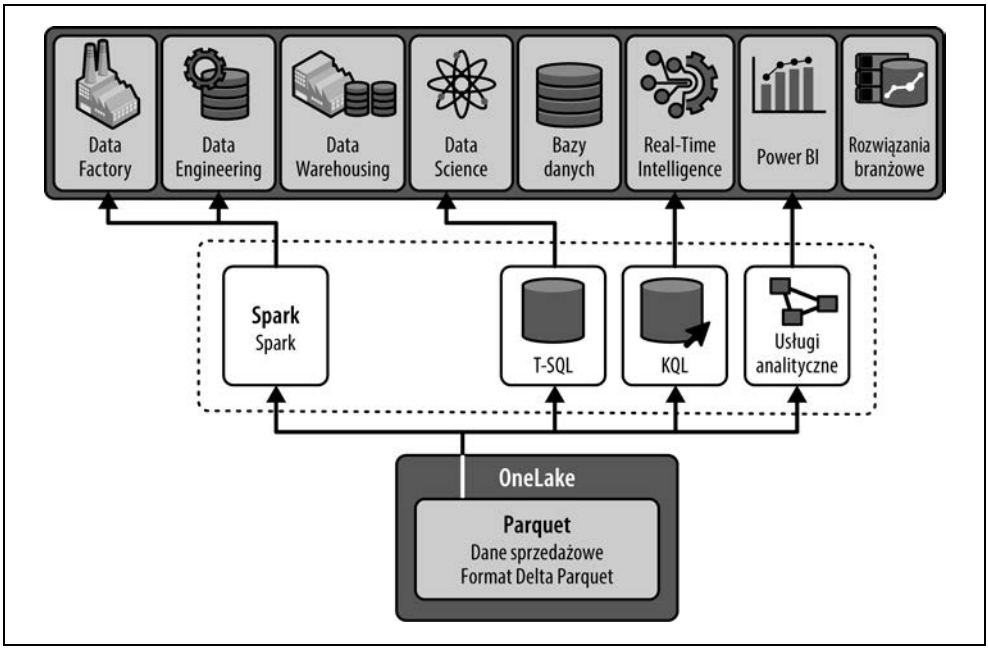
Wprowadzenie do OneLake

OneLake to nowoczesne rozwiązanie jeziora danych opracowane przez Microsoft, zaprojektowane w celu usprawnienia zarządzania danymi i analityki w chmurze. Jako część pakietu usług chmurowych Microsoftu OneLake ściśle integruje się z różnymi ofertami Azure, zapewniając ujednoliconą platformę do przechowywania, przetwarzania i analizowania danych na dużą skalę. Rozwój OneLake odzwierciedla zaangażowanie Microsoftu w dostarczanie firmom solidnych, skalowalnych i elastycznych rozwiązań do obsługi danych, odpowiadając na rosnące zapotrzebowanie na efektywne zarządzanie danymi w opartym na danych świecie.

Ogólną strukturę OneLake przedstawiono na rysunku 3.2.

Oddzielenie mocy obliczeniowej od pamięci masowej

OneLake wykorzystuje zasadę architektoniczną polegającą na rozdzieleniu przetwarzania i przechowywania danych, co stanowi znaczący postęp w projektowaniu jezior danych. Taki podział pozwala firmom na niezależne skalowanie zasobów pamięci masowej i mocy obliczeniowej, co optymalizuje koszty i wydajność. Dzięki rozdzieleniu tych komponentów OneLake umożliwia skalowanie pamięci masowej bez konieczności jednoczesnego zwiększania mocy obliczeniowej i odwrotnie. Ta elastyczność jest kluczowa dla efektywnego zarządzania zmiennymi obciążeniami i osiągnięcia optymalnej efektywności kosztowej.



Rysunek 3.2. OneLake — ogólny zarys

Eksplorator plików

OneLake zawiera przyjazny dla użytkownika interfejs eksploratora plików, który upraszcza nawigację i zarządzanie danymi. Eksplorator umożliwia przeglądanie, przesyłanie, pobieranie i organizowanie plików w jeziorze danych, zapewniając intuicyjny sposób interakcji z przechowywanymi informacjami. Ta funkcja jest szczególnie przydatna dla użytkowników, którzy nie są zaznajomieni z interfejsami wiersza poleceń lub złożonymi narzędziami do zarządzania danymi, i sprawia, że rozwiązanie OneLake staje się dostępne dla szerszego grona odbiorców.

Czym wyróżnia się OneLake?

OneLake oferuje kilka funkcji, które odróżniają go od innych jezior danych, zwiększając atrakcyjność i funkcjonalność rozwiązania.

Skróty

Jedną z najbardziej godnych uwagi funkcji OneLake jest rozbudowane wsparcie dla skrótów, przełomowego narzędzia, które znacznie usprawnia zarządzanie danymi i ułatwia dostęp do nich z poziomu Microsoft Fabric. Funkcja skrótów umożliwia użytkownikom tworzenie zarówno **skrótów wewnętrznych**, czyli odniesień do danych przechowywanych w różnych lokalizacjach OneLake, jak i **skrótów zewnętrznych**, czyli odniesień do zewnętrznych systemów pamięci masowej, takich jak Azure Data Lake Storage czy Amazon S3. Ta funkcja jest szczególnie korzystna dla dużych organizacji, w których dane są zazwyczaj rozproszone między różnymi

działami, projektami lub regionami geograficznymi. Zamiast duplikować dane w różnych obszarach roboczych lub domenach, co może prowadzić do zwiększonych kosztów przechowywania i problemów z zarządzaniem danymi, skróty umożliwiają zespołom płynne odwoływanie się do istniejących zbiorów danych.

Skróty działają na zasadzie „jednej kopii", co oznacza, że dane są fizycznie przechowywane tylko w jednym miejscu, a nie w wielu redundantnych i potencjalnie niesynchronizowanych kopiach w różnych lokalizacjach.

Wyobraźmy sobie, że zespół marketingowy firmy ma w swoim obszarze roboczym zbiór danych o zachowaniach klientów. Zespół finansowy może utworzyć skrót do tego zbioru danych w swoim własnym obszarze roboczym bez fizycznego kopiowania danych. Takie podejście nie tylko oszczędza miejsce, ale także gwarantuje, że oba zespoły pracują na najbardziej aktualnych informacjach. Co więcej, skróty te zachowują zabezpieczenia i kontrolę dostępu pierwotnych danych, więc poufne informacje pozostają chronione i dostępne tylko dla uprawnionych użytkowników, niezależnie od tego, ile obszarów roboczych się do nich odwołuje. Ponadto aktualizacja pierwotnej tabeli nie wymaga przeprowadzania żadnych procesów ETL, aby zaktualizowane dane były widoczne dla użytkowników.

Skróty upraszczają również integrację danych i współpracę między różnymi działami. Analitycy i inżynierowie danych mogą włączać dane z różnych źródeł do swoich procesów bez konieczności wielokrotnego przenoszenia lub przekształcania danych. Zmniejsza to złożoność przetwarzania danych i zapewnia spójność raportów i pulpitów nawigacyjnych. W miarę jak coraz więcej organizacji wprowadza architekturę siatki danych (ang. *data mesh*), skróty stają się jeszcze bardziej użyteczne. Umożliwiają one każdej domenie zarządzanie i udostępnianie danych jako produktu, podczas gdy inne domeny mogą szybko i łatwo integrować te dane z własnymi analizami bez zakłócania pierwotnych zbiorów danych.

Ponadto skróty można tworzyć dla danych przechowywanych w systemach zewnętrznych, co znacznie rozszerza zakres danych dostępnych w OneLake. Na przykład, jeśli organizacja ma dane historyczne zarchiwizowane w Amazon S3, można utworzyć skrót do tych danych, co sprawia, że są one natychmiast dostępne do analizy w Microsoft Fabric bez konieczności przenoszenia ich do OneLake. Funkcja ta wspiera strategie hybrydowe i wielochmurowe, pozwalając organizacjom wykorzystać istniejące inwestycje w dane, a jednocześnie korzystać ze zintegrowanych możliwości analitycznych Microsoft Fabric.

Podsumowując, wykorzystanie skrótów w OneLake znacznie zmniejsza redundancję, optymalizuje wykorzystanie pamięci masowej i upraszcza nadzór nad danymi poprzez centralizację dostępu przy jednoczesnym zachowaniu autonomii poszczególnych zespołów i obszarów roboczych. Funkcja ta jest zgodna z nowoczesnymi zasadami zarządzania danymi, które kładą nacisk na zwinność, skalowalność i bezpieczeństwo. Umożliwia organizacjom szybsze i efektywniejsze uzyskiwanie wniosków, poprawiając ich ogólną strategię zarządzania danymi.

Integracja z systemem Microsoftu

Ścisła integracja OneLake z pakietem narzędzi i usług Microsoftu oferuje spójną i kompleksową platformę do zarządzania danymi i analityki. Dzięki tej integracji użytkownicy mogą korzystać ze znajomych narzędzi i procesów, co zwiększa produktywność i ułatwia naukę.

Skalowalność i elastyczność

Rozdzielenie zasobów obliczeniowych i pamięci masowej OneLake zapewnia niespotykaną skalowalność i elastyczność. Firmy mogą skalować pamięć masową, aby pomieścić rosnące ilości danych, a jednocześnie niezależnie zarządzać zasobami obliczeniowymi w zależności od potrzeb i w ten sposób optymalizować zarówno wydajność, jak i koszty.

Zaawansowane zabezpieczenia i zgodność z przepisami

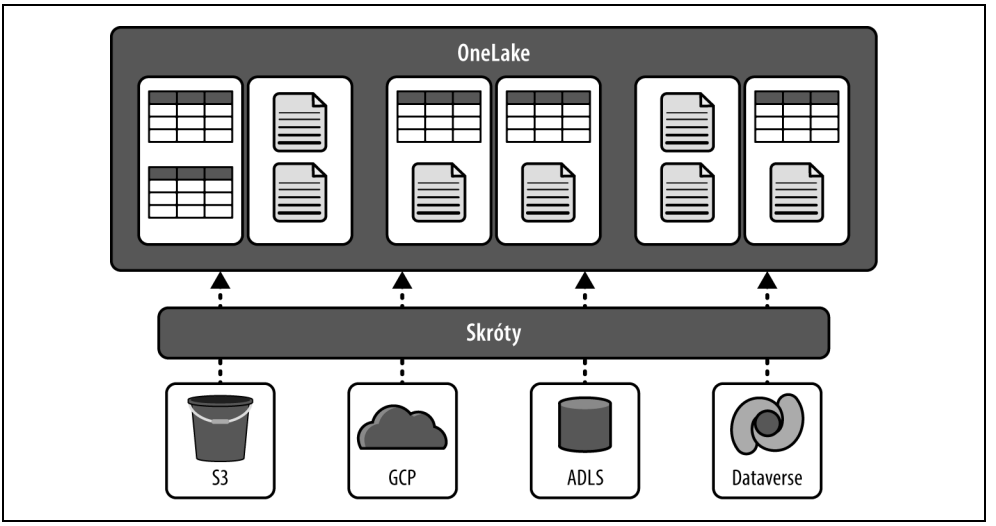
OneLake zawiera solidne funkcje bezpieczeństwa i zgodności, które zapewniają ochronę i umożliwiają zarządzanie danymi zgodnie ze standardami branżowymi i przepisami. Ten nacisk na bezpieczeństwo i zgodność jest kluczowy dla firm działających w regulowanych branżach lub przetwarzających wrażliwe dane.

Efektywność kosztowa

Architektura OneLake, wraz z możliwościami integracji i obsługą skrótów, zapewnia znaczne oszczędności. Firmy mogą zminimalizować koszty przechowywania, uniknąć niepotrzebnego powielania danych i zoptymalizować wykorzystanie zasobów, a przez to stworzyć bardziej ekonomiczne rozwiązanie do zarządzania danymi. Część tej efektywności kosztowej bierze się stąd, że skróty nie zajmują rzeczywistego miejsca w OneLake. Ponadto *dublowanie*, metoda pozyskiwania danych, którą omówimy później w tym rozdziale i szczegółowo w rozdziale 11., wiąże się z darmową przestrzenią dyskową.

Jak pokazano na rysunku 3.3, OneLake wyróżnia się na tle tradycyjnych jezior danych, zachowując jednocześnie ich kluczowe cechy.

Mówiąc w skrócie, OneLake stanowi znaczący postęp w technologii jezior danych i oferuje potężne, skalowalne, elastyczne rozwiązanie do nowoczesnego zarządzania danymi i analityki. OneLake wykorzystuje rozległą wiedzę Microsoftu w zakresie chmury i usług zarzadzania danymi, zapewniając solidną platformę, która bezproblemowo integruje się z szeroką gamą narzędzi i usług. Jej unikatowe cechy, takie jak oddzielenie mocy obliczeniowej od pamięci masowej, przyjazny dla użytkownika eksplorator plików i innowacyjne skróty, wyróżniają OneLake spośród innych rozwiązań jeziora danych, czyniąc go atrakcyjnym wyborem dla firm chcących w pełni wykorzystać potencjał swoich danych. W świecie, w którym danych jest coraz więcej, a analizy stają się coraz bardziej złożone, OneLake może stanowić kluczowe narzędzie do podejmowania decyzji opartych na danych i wspierania innowacji.



Rysunek 3.3. OneLake ze skrótami jako wyróżnikiem

Fundamenty OneLake

OneLake opiera się na ujednoliconej pamięci masowej, zaawansowanych formatach danych, zgodności operacyjnej oraz koncentracji na skalowalności i wydajności. Elementy te łączą się, tworząc solidne, elastyczne i wydajne rozwiązanie jeziora danych, które spełnia różnorodne potrzeby nowoczesnych przedsiębiorstw.

U podstaw architektury OneLake leży koncepcja **ujednoliconej pamięci masowej**. Wszystkie rodzaje danych — ustrukturyzowane, częściowo ustrukturyzowane i nieustrukturyzowane — są przechowywane w jednym, spójnym repozytorium. Ujednolicona pamięć masowa eliminuje silosy danych, ułatwiając dostęp i umożliwiając analizę różnych typów danych. Centralizacja upraszcza zarządzanie danymi, zmniejsza redundancję i zapewnia spójność danych w skali całego przedsiębiorstwa. Ujednolicona pamięć masowa wspiera również zróżnicowane potrzeby nowoczesnych firm. Niezależnie od tego, czy mamy do czynienia z danymi transakcyjnymi, plikami dzienników, treściami multimedialnymi, czy danymi strumieniowymi czasu rzeczywistego, OneLake zapewnia elastyczne i skalowalne środowisko, które spełnia pełne spektrum wymagań dotyczących danych. Umożliwia to kompleksowe analizy i wyciąganie wartościowych wniosków.

Wykorzystanie formatów Delta i Iceberg zwiększa niezawodność i wydajność przetwarzania danych, a zgodność z interfejsami API ADLS Gen2 umożliwia ścisłą integrację z istniejącymi narzędziami i systemami.

Wreszcie skalowalność i wydajność OneLake pozwala organizacjom efektywnie obsługiwać ogromne zbiory danych, przyspieszając uzyskiwanie wniosków i ułatwiając wprowadzanie innowacji opartych na danych. W rezultacie OneLake staje się fundamentalnym elementem kompleksowej strategii zarządzania danymi i analityki.

Formaty Delta i Iceberg

OneLake obsługuje zaawansowane formaty danych Delta i Apache Iceberg, które mają kluczowe znaczenie dla efektywnego zarządzania dużymi zbiorami danych.

Format Delta Lake

Delta Lake to otwarta warstwa pamięci masowej, która dodaje transakcje ACID do zadań przetwarzania dużych zbiorów danych. Format Delta zwiększa niezawodność i wydajność, umożliwiając tworzenie skalowalnych i niezawodnych potoków inżynierii danych. Obsługuje takie funkcje jak podróże w czasie (wersjonowanie danych), egzekwowanie schematów oraz możliwość efektywnego przetwarzania zarówno danych wsadowych, jak i strumieniowych. Dzięki wykorzystaniu formatu Delta OneLake zapewnia integralność danych i ułatwia wykonywanie złożonych operacji przetwarzania danych.

Format Iceberg

Apache Iceberg to inny otwarty format tabel przeznaczony do obsługi dużych zbiorów danych analitycznych. Został zaprojektowany do zarządzania tabelami o rozmiarach rzędu petabajtów i zapewnia szybkie operacje odczytu i zapisu. Obsługa ewolucji schematów, ukrytego partycjonowania i zaawansowanego zarządzania metadanymi sprawia, że Iceberg idealnie nadaje się do nowoczesnych architektur jezior danych. Obsługa formatu Iceberg w OneLake zapewnia kompatybilność z różnymi mechanizmami analitycznymi oraz poprawia wydajność zapytań poprzez optymalizację układu danych i minimalizację skanowania danych.

Zgodność operacyjna

Projekt OneLake kładzie nacisk na interoperacyjność z szeroką gamą narzędzi i systemów. Ma to kluczowe znaczenie dla organizacji korzystających z różnorodnych technologii, które muszą zintegrować swoje jezioro danych z istniejącymi procesami i aplikacjami.

OneLake jest w pełni kompatybilne z interfejsami API Azure Data Lake Storage Gen2, które zapewniają obsługę hierarchicznej przestrzeni nazw, zaawansowane funkcje bezpieczeństwa i zoptymalizowaną wydajność. Zgodność ta umożliwia integrację OneLake z aplikacjami i usługami, które już korzystają z ADLS Gen2, zapewniając płynne przejście i spójne wrażenia użytkownika. Wykorzystanie tych interfejsów API pozwala na solidną kontrolę dostępu, uproszczone zarządzanie i wydajne przetwarzanie danych.

Skalowalność i wydajność

Skalowalność i wydajność to fundamentalne aspekty rozwiązania OneLake, które zaprojektowano z myślą o sprostaniu wymaganiom dużych środowisk danych.

Skalowalność

OneLake zbudowano z myślą o skalowalności poziomej, aby umożliwić organizacjom zwiększanie pojemności pamięci masowej w miarę wzrostu ilości danych. Ta skalowalność zapewnia, że OneLake może obsługiwać ogromne zbiory danych, od terabajtów do petabajtów, bez

utraty wydajności. Wykorzystując infrastrukturę chmurową, OneLake zapewnia praktycznie nieograniczoną pojemność pamięci masowej, umożliwiając firmom przechowywanie i analizowanie stale rosnących ilości danych.

Wydajność

OneLake jest zoptymalizowane pod kątem wysokiej wydajności i wspiera szybkie pozyskiwanie, przetwarzanie i badanie danych. Wykorzystanie zaawansowanych formatów danych Delta i Iceberg przyczynia się do poprawy szybkości odczytu i zapisu, efektywnej kompresji danych oraz minimalizacji opóźnień. Ponadto rozdzielenie zasobów obliczeniowych i pamięci masowej zapewnia, że wydajność może być dostosowywana i skalowana niezależnie w celu zoptymalizowania użycia zasobów dla konkretnego obciążenia. Takie podejście umożliwia szybki dostęp do danych i analizę w czasie rzeczywistym, dając organizacjom możliwość szybkiego uzyskiwania wglądu w dane i podejmowania decyzji opartych na danych.

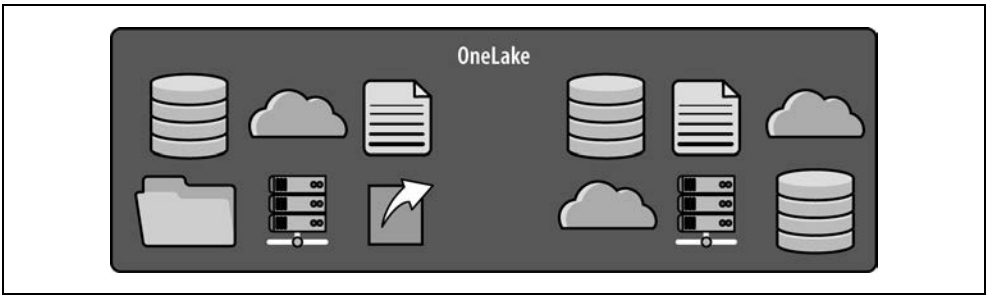
Dane przechowywane w OneLake

OneLake zaprojektowano do przechowywania szerokiej gamy typów danych, aby spełnić różnorodne potrzeby nowoczesnych przedsiębiorstw. Większość danych przechowywanych w OneLake to oczywiście dane podstawowe — można je nawet nazwać danymi rzeczywistymi — a w OneLake i ogólnie w systemach informatycznych istnieje wiele typów danych operacyjnych.

Elastyczna i skalowalna architektura OneLake pozwala na przechowywanie różnorodnych typów danych w ramach kompleksowej strategii zarządzania danymi. Dzięki obsłudze danych ustrukturyzowanych, częściowo ustrukturyzowanych i nieustrukturyzowanych, a także obsłudze danych czasu rzeczywistego i big data OneLake zapewnia ujednoliconą platformę do przechowywania i analizowania różnorodnych zbiorów danych. Ta funkcjonalność gwarantuje, że organizacje mogą wykorzystać wszystkie swoje zasoby danych do generowania wniosków, innowacji i przewagi konkurencyjnej.

Integralną częścią OneLake są metadane, w tym informacje techniczne, operacyjne, biznesowe i dotyczące bezpieczeństwa. Platforma obsługuje przechowywanie różnych typów metadanych oraz elementów analitycznych, takich jak notatniki i skrypty. To kompleksowe zarządzanie metadanymi usprawnia odkrywanie danych, zarządzanie nimi, poprawia ich jakość i ułatwia współpracę, czyniąc OneLake potężnym i efektywnym jeziorem danych.

Różne dane i formaty przechowywane w OneLake pokazano na rysunku 3.4.



Rysunek 3.4. Różnorodne rodzaje danych przechowywanych w OneLake

Dane ustrukturyzowane

Dane ustrukturyzowane są wysoce zorganizowane i sformatowane tak, aby nadawały się do przechowywania w tradycyjnych bazach danych lub arkuszach kalkulacyjnych. Mają spójny schemat i zazwyczaj są przechowywane w wierszach i kolumnach.

Dane ustrukturyzowane pochodzą z relacyjnych systemów zarządzania bazami danych (RDBMS), takich jak SQL Server, Oracle, MySQL i PostgreSQL, i mogą być importowane do OneLake. Obejmują tabele, indeksy i widoki. Jeśli chodzi o zawartość, często określa się je jako *dane transakcyjne* i *operacyjne*. Dane transakcyjne są generowane podczas transakcji biznesowych, takich jak sprzedaż, transakcje finansowe i przetwarzanie zamówień. Dane operacyjne powstają w wyniku codziennych działań biznesowych i obejmują poziomy zapasów, dokumentację kadrową oraz dane z systemów zarządzania relacjami z klientami (CRM).

Oprócz surowych danych pochodzących z innych systemów i źródeł typowym przykładem danych ustrukturyzowanych są *dane pochodne*. Są to surowe dane, które zostały przetworzone i przekształcone w celu analizy, takie jak *dane zagregowane* (podsumowane dane z różnych źródeł, np. miesięczne sumy sprzedaży, średnie oceny klientów i statystyki zbiorcze) oraz *dane analityczne* (przetworzone dane gotowe do analizy, w tym kostki danych, pulpity nawigacyjne i raporty).

Dane częściowo ustrukturyzowane

Dane częściowo ustrukturyzowane nie mają sztywnego schematu, ale zawierają znaczniki lub markery oddzielające elementy danych, co oznacza, że są bardziej elastyczne niż dane ustrukturyzowane. Typowe przykłady danych częściowo ustrukturyzowanych to:

Pliki JSON i XML

Dane w formacie JSON (JavaScript Object Notation) i XML, często używane w aplikacjach internetowych i interfejsach API.

Pliki dziennika

Dzienniki systemowe i aplikacyjne generowane przez serwery, aplikacje i urządzenia sieciowe. Te pliki często zawierają cenne informacje przydatne do monitorowania i rozwiązywania problemów.

Dane z czujników

Dane z urządzeń internetu rzeczy (IoT), w tym odczyty temperatury, wilgotności i inne dane z czujników środowiskowych.

Poczta elektroniczna

Treść i metadane z komunikacji e-mail.

Dane nieustrukturyzowane

Dane nieustrukturyzowane nie mają zdefiniowanego formatu, przez co najtrudniej się je przechowuje i analizuje. OneLake potrafi efektywnie obsługiwać ogromne ilości danych nieustrukturyzowanych. Dane nieustrukturyzowane pochodzą najczęściej z następujących źródeł:

Pliki multimedialne

Obrazy, filmy i pliki audio używane w produkcji medialnej, marketingu i komunikacji.

Dokumenty

Dokumenty tekstowe, pliki PDF, prezentacje i inne typy plików używane w działalności biznesowej i komunikacji.

Dane z mediów społecznościowych

Informacje pochodzące z platform społecznościowych, takie jak posty, komentarze, polubienia i udostępnienia.

Dane internetowe

Treści pobrane ze stron internetowych, w tym pliki HTML, CSS i JavaScript.

Metadane

Metadane odgrywają kluczową rolę w zarządzaniu danymi przechowywanymi w jeziorze danych. Dostarczają kontekstu, ułatwiają odkrywanie danych i zapewniają ich odpowiednie wykorzystanie. Na platformie Fabric metadane są również przechowywane w OneLake.

Metadane można podzielić na kilka typów, z których każdy zawiera różne rodzaje informacji:

Metadane techniczne

Metadane techniczne obejmują:

Informacje o schemacie

Informacje o strukturze danych, takie jak definicje tabel, typy kolumn i formaty danych.

Pochodzenie danych

Informacje o źródle danych, ich przepływie przez różne etapy przetwarzania i zastosowanych transformacjach.

Szczegóły przechowywania

Ścieżki do plików, lokalizacje w jeziorze danych i formaty przechowywania.

Metadane operacyjne

Metadane operacyjne obejmują:

Miary jakości danych

Informacje o dokładności, kompletności, spójności i poprawności danych.

Statystyki użycia

Dane dotyczące częstotliwości dostępu do zbiorów danych, modyfikacji lub kwerend.

Dzienniki przetwarzania

Zapisy procesów pozyskiwania danych, zadań ETL i kroków transformacji danych.

Metadane biznesowe

Metadane biznesowe obejmują:

Katalogi danych

Opisy zbiorów danych, terminologia biznesowa i klasyfikacje danych.

Tagi i adnotacje

Słowa kluczowe, etykiety i notatki dodawane do zbiorów danych w celu ułatwienia wyszukiwania i odkrywania

Metadane bezpieczeństwa

Metadane bezpieczeństwa obejmują:

Informacje o kontroli dostępu

Uprawnienia i role przypisane do różnych zbiorów danych i użytkowników.

Dzienniki audytu

Zapisy dotyczące tego, kto i kiedy uzyskał dostęp do danych lub je zmodyfikował.

Notatniki i elementy analityczne

Oprócz tradycyjnych metadanych OneLake może również przechowywać elementy analityczne, takie jak:

Notatniki

Interaktywne dokumenty zawierające aktywny kod, równania, wizualizacje i opisowy tekst. Są one powszechnie używane do analizy i eksploracji danych oraz zadań związanych z uczeniem maszynowym.

Skrypty i kod

Skrypty i kod napisane w Pythonie, R, Scali i innych językach programowania używanych do przetwarzania i analizy danych.

Modele uczenia maszynowego

Wytrenowane modele wraz z ich konfiguracjami i miarami trafności.

Pulpity nawigacyjne i raporty

Wizualizacje i raporty utworzone przy użyciu Power BI.

Organizowanie danych w OneLake

Aby porządkować i grupować lub dzielić dane w OneLake, trzeba rozumieć dwa współistniejące pojęcia: *domeny* i *obszary robocze*.

Domeny

Domeny to logiczne partycje OneLake zaprojektowane w celu organizowania danych i zarządzania nimi według konkretnych jednostek biznesowych, działów lub innych kryteriów organizacyjnych. Każdą domenę można postrzegać jako oddzielną, bezpieczną przestrzeń danych z własnym zestawem uprawnień, zasad zarządzania i struktur danych.

Domeny są dostosowane do struktury organizacyjnej firmy, co ułatwia zarządzanie danymi istotnymi dla konkretnych jednostek biznesowych lub działów. Umożliwiają one precyzyjną kontrolę nad dostępem do danych i nad zarządzaniem nimi, gwarantując, że tylko upoważnieni użytkownicy mają dostęp do wrażliwych informacji. Określają ramy zarządzania cyklem życia danych, ich jakością i spójnością w określonych kontekstach organizacyjnych.

Domeny w OneLake to podstawowe elementy, które poprawiają nadzór nad danymi, bezpieczeństwo, organizację, współpracę, skalowalność i integrację. Są one niezbędne do efektywnego zarządzania danymi w dużych organizacjach i umożliwiają organizowanie danych pod kątem wspierania celów biznesowych.

Domena może rozciągać się na wiele zasobów obliczeniowych!

Obszary robocze

Obszary robocze to środowiska współpracy, w których użytkownicy mogą tworzyć i udostępniać modele semantyczne, raporty, pulpity nawigacyjne oraz inne elementy analityczne, takie jak jeziora danych czy potoki. Każdy obszar roboczy działa jak kontener na te zasoby, ułatwiając pracę zespołową i zarządzanie projektami.

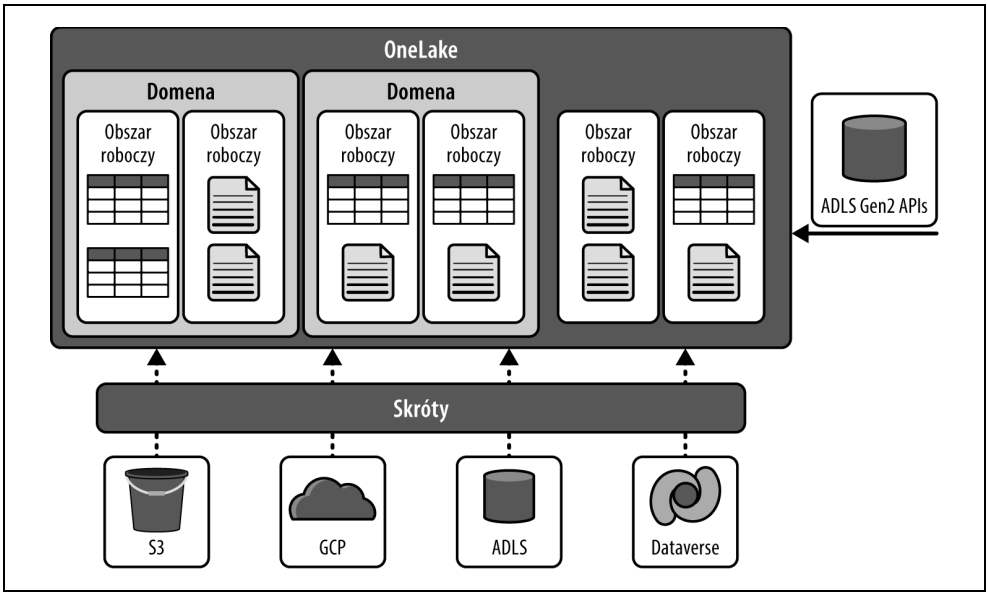
Obszary robocze zaprojektowano po to, aby umożliwić zespołom współpracę nad projektami związanymi z danymi oraz efektywne dzielenie się spostrzeżeniami i zasobami. Pomagają one w logicznym organizowaniu projektów i zasobów, ułatwiając nawigację i zarządzanie. Obszary robocze pozwalają na ustawianie uprawnień i kontroli dostępu, żeby tylko upoważnieni użytkownicy mogli przeglądać lub edytować ich zawartość.

Główne różnice między domenami a obszarami roboczymi

Domeny i obszary robocze odgrywają kluczowe, ale różne role w OneLake. Domeny zapewniają ustrukturyzowane i bezpieczne środowisko zgodne z hierarchią organizacyjną, podczas gdy obszary robocze oferują elastyczne, wspólne przestrzenie do zarządzania projektami i pracą zespołową. Razem gwarantują dobrą organizację danych, ich bezpieczeństwo oraz dostępność dla tych, którzy ich potrzebują.

Każdy element Fabric musi być przypisany do obszaru roboczego, ale nie każdy obszar roboczy wymaga domeny. Domeny stanowią dodatkową, opcjonalną warstwę grupowania, więc domena bez obszaru roboczego nie miałaby sensu, ponieważ nie mogłaby zawierać żadnych elementów.

Relacje między domenami, obszarami roboczymi i skrótami w OneLake przedstawiono na rysunku 3.5.



Rysunek 3.5. Różnice między domenami a obszarami roboczymi w OneLake

Przyjrzyjmy się kluczowym różnicom między domenami a obszarami roboczymi.

Zakres i cel

Domeny skupiają się na organizacji danych w całej firmie według jednostek biznesowych, działów lub innych struktur, zapewniając nadzór nad danymi, bezpieczeństwo i spójność organizacyjną.

Obszary robocze natomiast koncentrują się na ułatwianiu współpracy i zarządzaniu projektami w ramach konkretnych zespołów lub projektów, centralizując zasoby i umożliwiając efektywną pracę zespołową.

Wyobraźmy sobie dużą firmę technologiczną, która decyduje się wdrożyć podejście siatki danych, aby lepiej zarządzać swoim zróżnicowanym i rozproszonym środowiskiem danych. W tym kontekście domeny w Microsoft Fabric reprezentują całą firmę, służąc jako nadrzędna struktura zawierająca wszystkie dane, zasoby i użytkowników w organizacji. Domenami mogą być sprzedaż, HR, IT, marketing lub inne działy. W ramach każdej domeny obszary robocze funkcjonują jako konkretny produkt danych lub jednostka biznesowa, jak samodzielny zespół projektowy odpowiedzialny za własne dane, raporty i analizy. Na przykład międzynarodowy zespół sprzedaży i krajowy zespół sprzedaży mogą mieć własne obszary robocze, przy czym oba należą do domeny sprzedaży. Takie podejście jest zgodne z zasadami siatki danych — daje każdemu zespołowi możliwość zarządzania swoimi danymi przy jednoczesnym zapewnieniu nadzoru i łączności w całej organizacji.

Siatka danych (ang. *data mesh*) to nowoczesne podejście do zarządzania danymi, które decentralizuje kontrolę, przekazując odpowiedzialność za dane zespołom dziedzinowym traktującym swoje zbiory danych jako niezależne produkty. Wykorzystuje ona infrastrukturę samoobsługową, umożliwiając zespołom efektywny dostęp, przetwarzanie i udostępnianie danych przy zachowaniu standardów zarządzania i jakości. Eliminując scentralizowane wąskie gardła i promując zgodność operacyjną, siatka danych wspiera skalowalne, zorientowane na biznes zarządzanie danymi w skali całej organizacji.

Bezpieczeństwo i nadzór

Domeny implementują zasady bezpieczeństwa i zarządzania w skali całej organizacji, kontrolując dostęp do danych w zależności od potrzeb jednostek biznesowych lub działów. Obszary robocze implementują kontrolę dostępu specyficzną dla projektu lub zespołu, zarządzając uprawnieniami do wspólnej pracy nad modelami semantycznymi, raportami i innymi zasobami.

Organizacja

Podczas gdy domeny organizują dane na wyższym poziomie, często zgodnie ze strukturą firmy, obszary robocze organizują zasoby na poziomie projektu lub zespołu, koncentrując się na potrzebach konkretnych działań zespołowych.

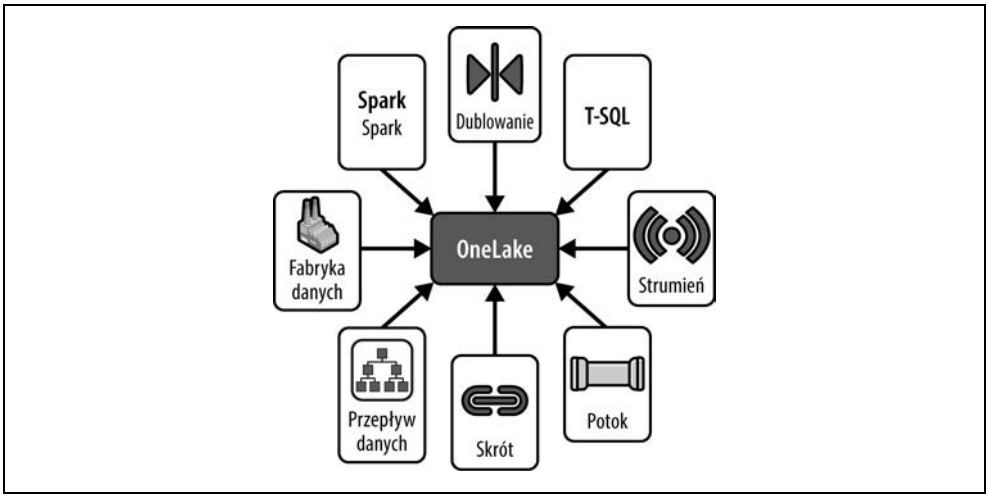
Pozyskiwanie danych i integrowanie ich z OneLake

Pozyskiwanie i integrowanie danych to kluczowe procesy zapewniające efektywne i skuteczne wprowadzanie danych z różnych źródeł do OneLake. OneLake obsługuje różnorodne metody pozyskiwania danych i mechanizmy integracji, dzięki czemu jest wszechstronną platformą do zarządzania złożonymi środowiskami danych.

Metody pozyskiwania danych

Istnieje wiele sposobów wprowadzania danych do OneLake, jak pokazano na rysunku 3.6. Omówimy je bardziej szczegółowo w drugiej części tej książki, ale oto krótki przegląd.

Możesz użyć skrótu, który tak naprawdę nie wprowadza danych, a jedynie udostępnia je w Twoim jeziorze danych, lub możesz skorzystać z narzędzi takich jak Spark, przepływy danych, działania fabryki danych, potoki, strumienie, T-SQL lub dublowanie. Prawdopodobnie słyszałeś o większości z nich, ale powiedzmy nieco więcej o dublowaniu. **Dublowanie** (ang. *mirroring*) w Microsoft Fabric to funkcja zaprojektowana w celu uproszczenia integracji danych i usprawnienia procesów analitycznych poprzez tworzenie repliki danych w czasie rzeczywistym, dostępnej tylko do odczytu, w otwartym formacie Delta, bez konieczności stosowania procesów ETL. Jest to w zasadzie inteligentny sposób replikowania danych z innych źródeł do OneLake w czasie niemal rzeczywistym, bez ręcznego wczytywania danych. Dublowanie omówimy dokładniej w rozdziale 11.



Rysunek 3.6. Wprowadzanie danych do OneLake

Mechanizmy integracji

OneLake ściśle integruje się z szeroką gamą usług, tworząc spójny ekosystem zarządzania danymi i analityki.

Usługi Microsoftu

Fabric i OneLake oczywiście bardzo dobrze integrują się z innymi usługami Microsoftu, takimi jak:

Power BI

Power BI może łączyć się bezpośrednio z OneLake, umożliwiając użytkownikom tworzenie interaktywnych raportów i pulpitów nawigacyjnych na podstawie danych przechowywanych w jeziorze. Integracja ta umożliwia zaawansowaną wizualizację danych i analitykę biznesową. Dzięki pełnej, natywnej integracji Power BI z OneLake używanie OneLake jako źródła danych pozwala na pełne wykorzystanie potencjału obu usług. Więcej na ten temat powiemy w rozdziale 9.

Azure Machine Learning

Specjaliści data science mogą wykorzystać Azure Machine Learning do budowania, trenowania i wdrażania modeli uczenia maszynowego z użyciem danych przechowywanych w OneLake. Ta integracja ułatwia tworzenie rozwiązań do analityki predykcyjnej i systemów AI.

Microsoft Dynamics 365

Integracja z Dynamics 365 umożliwia firmom łączenie danych operacyjnych pochodzących z systemów CRM i ERP z innymi źródłami danych w OneLake w celu kompleksowej analizy. Zazwyczaj odbywa się to za pomocą funkcji o nazwie Fabric Link, która automatycznie generuje skróty.

Azure Databricks

Microsoft Fabric integruje się z Databricks, zapewniając ujednolicone środowisko, w którym użytkownicy mogą korzystać z zaawansowanych funkcji analityki i uczenia maszynowego oferowanych przez Databricks wraz z kompleksowymi narzędziami do zarządzania danymi dostępnymi w Fabric. Integracja ta umożliwia bezproblemowe udostępnianie danych i współpracę, pozwalając użytkownikom na wykonywanie złożonych operacji przetwarzania i analizy danych w Databricks oraz wykorzystanie funkcji wizualizacji i raportowania Fabric. Usprawnia to procesy oraz wyciąganie wniosków z danych dzięki spójnej i wydajnej platformie.

Narzędzia firm trzecich

Otwarta architektura OneLake pozwala również na integrację z różnorodnymi zewnętrznymi narzędziami do przetwarzania i analizy danych. Choć nie sposób stworzyć wyczerpującej listy, oto kilka popularnych rozwiązań:

Apache Spark

Microsoft Fabric integruje się z Apache Spark, umożliwiając wykorzystanie zaawansowanych funkcji przetwarzania rozproszonego Sparka do analizy dużych zbiorów danych bezpośrednio w środowisku Fabric. Integracja ta pozwala na wydajne wykonywanie złożonych zadań analitycznych, uczenie maszynowe i przetwarzanie danych przy użyciu Sparka przy jednoczesnym wykorzystaniu łączności z danymi i narzędzi wizualizacyjnych platformy Fabric. Połączenie to ułatwia tworzenie zaawansowanych przepływów pracy i przyspiesza analizy dzięki skalowalnemu i wydajnemu przetwarzaniu danych.

Tableau

Microsoft Fabric integruje się z Tableau, umożliwiając użytkownikom wizualizowanie i analizowanie danych przechowywanych w Microsoft Fabric bezpośrednio w potężnej platformie wizualizacyjnej Tableau. Integracja ta zapewnia płynne przekazywanie i spójność danych, pozwalając użytkownikom tworzyć interaktywne pulpity nawigacyjne, które korzystają z kompleksowych możliwości zarządzania danymi i analityki Fabric. Połączenie intuicyjnych wizualizacji Tableau z solidną infrastrukturą danych Fabric wspomaga podejmowanie decyzji w oparciu o dane.

Snowflake

Microsoft Fabric integruje się ze Snowflake, zapewniając płynne połączenie, które pozwala użytkownikom na dostęp, analizę i wizualizację danych Snowflake bezpośrednio w środowisku Microsoft Fabric. Integracja ta wykorzystuje możliwości hurtowni danych Snowflake i narzędzia analityczne Fabric, umożliwiając efektywne zarządzanie danymi i kompleksowe analizy poprzez jednolity interfejs. Usprawnia procesy decyzyjne, upraszczając przepływy pracy i zapewniając spójność danych między platformami.

Amazon S3

Microsoft Fabric integruje się z Amazon S3, umożliwiając użytkownikom zarządzanie danymi S3 bezpośrednio w środowisku Fabric. Integracja ta pozwala na płynny dostęp do danych, ich transfer i analizę z użyciem zaawansowanych narzędzi administracyjnych

i analitycznych platformy Fabric. Użytkownicy mogą efektywnie włączać dane S3 do swoich przepływów pracy, usprawniając podejmowanie decyzji w oparciu o dane dzięki kompleksowym możliwościom wizualizacji i raportowania dostępnym w Fabric.

Integracja oparta na API i punktach końcowych

OneLake obsługuje integrację opartą na API, umożliwiając programistom tworzenie niestandardowych aplikacji i usług, które współpracują z jeziorem danych. W przypadku organizacji działających w środowiskach hybrydowych OneLake może również integrować się ze źródłami lokalnymi, co zapewnia płynne zarządzanie danymi.

OneLake udostępnia interfejsy REST API do programowego dostępu i zarządzania danymi. Interfejsy te mogą być wykorzystywane do automatyzacji zadań związanych z pozyskiwaniem i pobieraniem danych oraz zarządzaniem nimi. Dzięki obsłudze API ADLS Gen2 platforma zapewnia kompatybilność z szeroką gamą istniejących aplikacji i narzędzi, które już korzystają z tych API do operacji na danych. SSIS (SQL Server Integration Services) i inne narzędzia ETL mogą być używane do tworzenia przepływów integracji danych, które przenoszą dane między systemami lokalnymi a OneLake, ułatwiając hybrydowe zarządzanie danymi, głównie przy użyciu punktów końcowych SQL Fabric.

Podsumowując, wszystkie te funkcje pokazują, że OneLake oferuje kompleksową i elastyczną platformę do pozyskiwania i integracji danych, która oferuje szeroką gamę metod i narzędzi do przenoszenia danych z różnych źródeł do jeziora. Niezależnie od tego, czy chodzi o przetwarzanie wsadowe, czy też w czasie rzeczywistym, o narzędzia migracyjne czy integrację z usługami Microsoft i firm trzecich, OneLake pomaga firmom efektywnie zarządzać danymi. Obsługa integracji opartej na API, wirtualizacji danych i hybrydowych środowisk danych dodatkowo zwiększa wszechstronność platformy i czyni ją ważnym elementem nowoczesnych architektur danych. W miarę jak organizacje coraz częściej przyjmują strategie oparte na danych, możliwości oferowane przez OneLake będą kluczowe w wykorzystaniu pełnego potencjału zasobów danych.

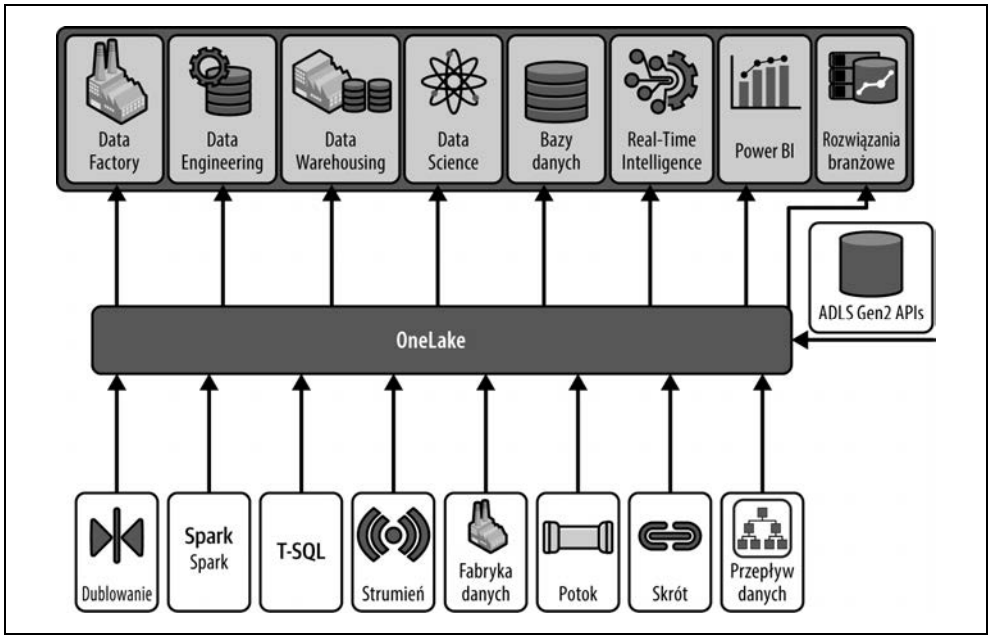
Różnorodne sposoby współdziałania Fabric z innymi narzędziami zilustrowano na rysunku 3.7.

Katalog OneLake

Katalog OneLake to centralne miejsce, w którym możesz przeglądać elementy Fabric i zarządzać nimi, a także nadzorować swoje dane. Składa się z dwóch kart: *Explore* i *Govern*.

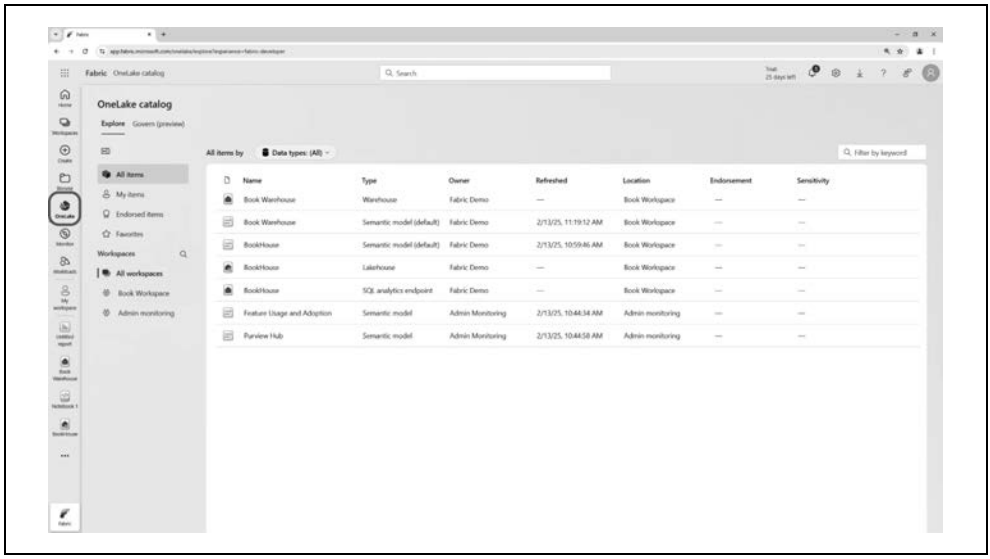
Karta *Explore* zawiera listę elementów wraz z widokiem szczegółów, co pozwala na przeglądanie i eksplorację elementów bez utraty kontekstu listy. Oferuje również filtry i selektory ułatwiające zawężanie widoku i skupianie się na konkretnych pozycjach, co ułatwia znajdowanie poszukiwanych informacji. Domyślnie katalog OneLake otwiera się na zakładce *Explore*.

Zakładka *Govern* daje wgląd w status nadzoru nad wszystkimi danymi, które masz w Fabric, oraz dostarcza praktycznych zaleceń dotyczących zwiększenia skuteczności nadzoru.



Rysunek 3.7. OneLake i jego interakcja z ekosystemem danych

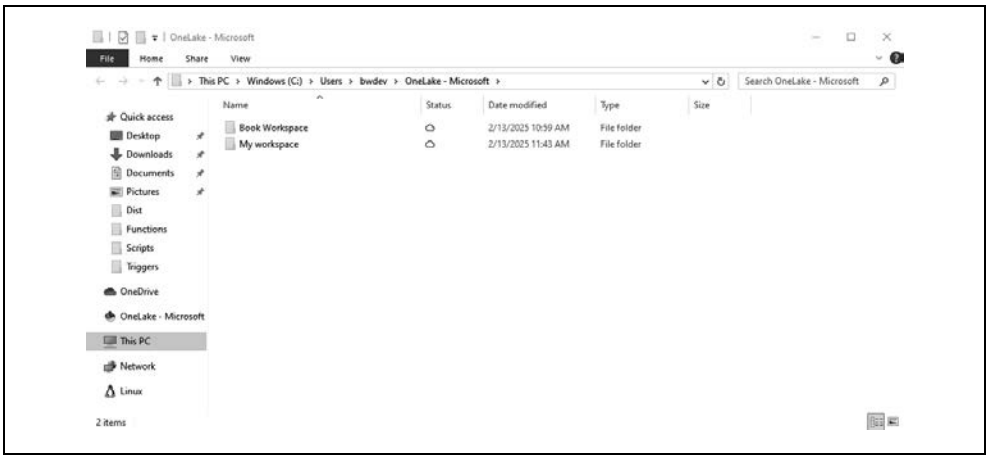
Aby uzyskać dostęp do katalogu OneLake, kliknij ikonę OneLake w panelu nawigacyjnym Fabric. Jeśli pożądana karta nie jest wyświetlana domyślnie, wybierz ją tak, jak pokazano na rysunku 3.8.



Rysunek 3.8. Katalog OneLake

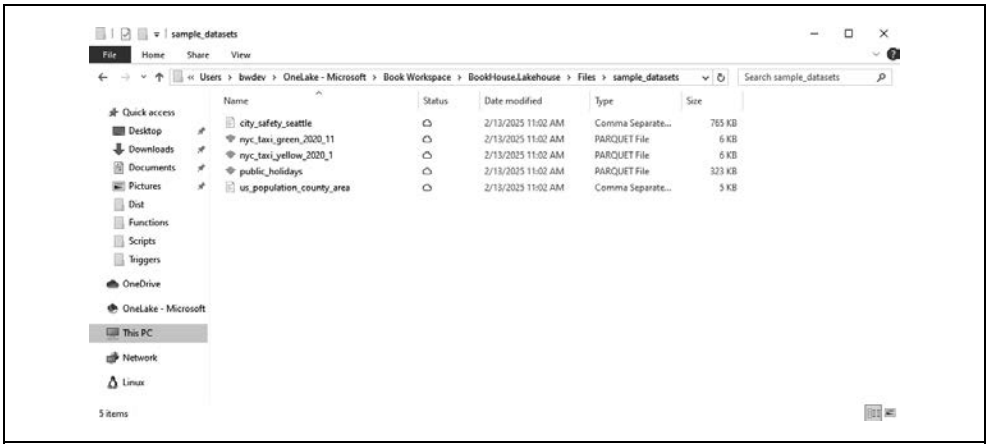
Eksplorator OneLake

Ponieważ OneLake często bywa nazywane „OneDrive'em dla danych", istnieje oczywiście również aplikacja umożliwiająca przeglądanie elementów OneLake na komputerze bez konieczności korzystania z portalu. Eksplorator plików OneLake (pokazany na rysunku 3.9) integruje OneLake bezpośrednio z Eksploratorem plików systemu Windows, pozwalając na łatwy dostęp do wszystkich elementów OneLake, do których masz uprawnienia, bezpośrednio z interfejsu Eksploratora plików Windows. Dane w eksploratorze OneLake można również synchronizować z powrotem do Fabric. Eksplorator plików OneLake można pobrać z oficjalnej strony pobierania (*https://oreil.ly/Mi8D9*).



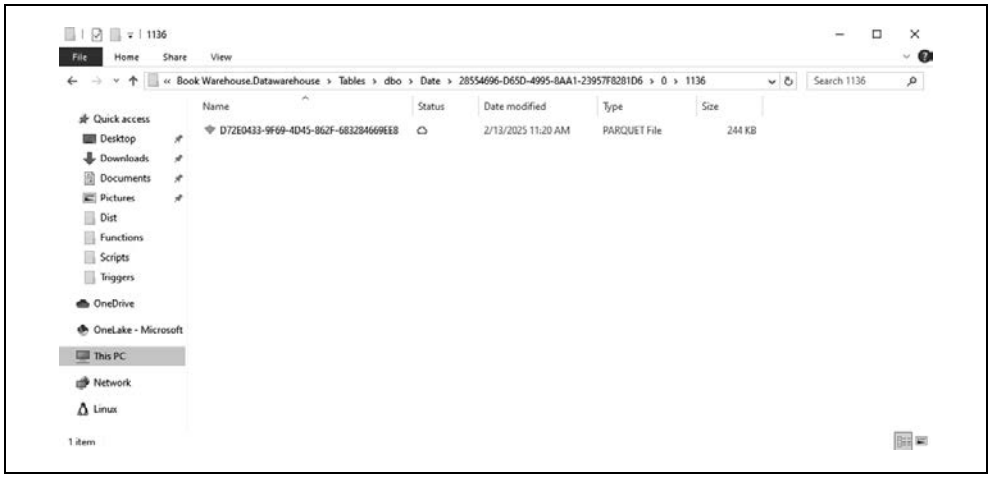
Rysunek 3.9. Przeglądarka plików OneLake

W naszym przypadku eksplorator OneLake pokazuje obszar roboczy, który utworzyliśmy w rozdziale 2. Możemy przechodzić do poszczególnych plików, podobnie jak w naszym repozytorium lakehouse (patrz rysunek 3.10).



Rysunek 3.10. Poziom elementów w eksploratorze plików OneLake

Ponadto widoczne są pliki Parquet stanowiące podstawę naszej hurtowni danych (patrz rysunek 3.11).



Rysunek 3.11. Pliki Parquet stanowiące podstawę tabeli hurtowni danych w eksploratorze plików OneLake

Podsumowanie

OneLake to nowoczesne i solidne rozwiązanie jeziora danych, stanowiące integralną część Microsoft Fabric i zaprojektowane z myślą o złożonych potrzebach współczesnych firm w zakresie zarządzania danymi i analityki. Przechowując wszystkie dane w scentralizowanym jeziorze, OneLake wykorzystuje ujednoliconą pamięć masową do wyeliminowania silosów danych i obsługi różnorodnych typów danych, od ustrukturyzowanych po nieustrukturyzowane. Wykorzystuje zaawansowane formaty danych Delta i Iceberg do efektywnego przetwarzania danych, a rozdzielenie obliczeń i magazynowania zwiększa skalowalność i efektywność kosztową platformy. Bezproblemowa integracja OneLake z pakietem narzędzi Microsoft i usługami zewnętrznymi w połączeniu z zaawansowanymi funkcjami bezpieczeństwa i zgodności czyni go wszechstronną i potężną platformą. Ta elastyczność ułatwia firmom wprowadzanie innowacji, przeprowadzanie kompleksowych analiz i utrzymywanie przewagi konkurencyjnej w świecie opartym na danych.

W kolejnej części tej książki poznasz różne „doświadczenia” — czyli funkcje — i możliwości platformy Fabric.

Budowanie efektywnych rozwiązań w obszarze analityki danych wymaga dojrzałej architektury i odpowiednich narzędzi. Microsoft Fabric oferuje ogromne możliwości, ale opanowanie tej platformy bywa trudne — mnogość dostępnych opcji może przytłaczać. Konkretne zadania można realizować na wiele sposobów, lecz nie każdy jest równie wydajny.

Tę książkę możesz potraktować jako wygodną, szczegółową mapę drogową, dzięki której zrozumiesz specyfikę Microsoft Fabric i dowiesz się, jak się poruszać po architekturze i możliwościach tej platformy. Znajdziesz tu wyjaśnienia, najlepsze praktyki i przykłady rzeczywistych zastosowań, które pomogą Ci płynnie przejść od teorii do praktyki. Zagłębisz się w szczegóły pracy z Fabric, dzięki czemu Twój zespół będzie mógł w pełni korzystać z integracji danych, inżynierii danych, magazynowania danych, data science, analiz czasu rzeczywistego i analityki biznesowej. Wszystko to w ramach wspólnej przestrzeni roboczej, z zunifikowanymi procesami i prostą obsługą.

Ta książka jest niezbędnym przewodnikiem dla specjalistów do spraw danych. Dzięki niej szybciej rozpoczniesz pracę z opracowaną przez Microsoft platformą danych nowej generacji!

Patrick LeBlanc, Microsoft

Zagadnienia:

- podstawowe komponenty Microsoft Fabric
- kluczowe koncepcje i techniki budowania solidnej platformy danych
- efektywne stosowanie Microsoft Fabric w codziennej pracy
- architektura oparta na jeziorze danych
- implementacja skalowalnych i wydajnych rozwiązań do analizy danych
- zarządzanie dzierżawą Fabric

Nikola Ilic jest niezależnym konsultantem i instruktorem. Specjalizuje się w pracy z Power BI i Microsoft Fabric. Mieszka w Salzburgu z żoną i dwojgiem dzieci.

Ben Weissman jest przedsiębiorcą i ekspertem w dziedzinie hurtowni danych i analityki biznesowej. Regularnie występuje na międzynarodowych konferencjach. Posiada tytuły: Microsoft Data Platform MVP, MCSE, MPP, a także Certified Data Vault Data Modeler.

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 978-83-289-3552-5

9 788328 935525

Cena: 99,00 zł