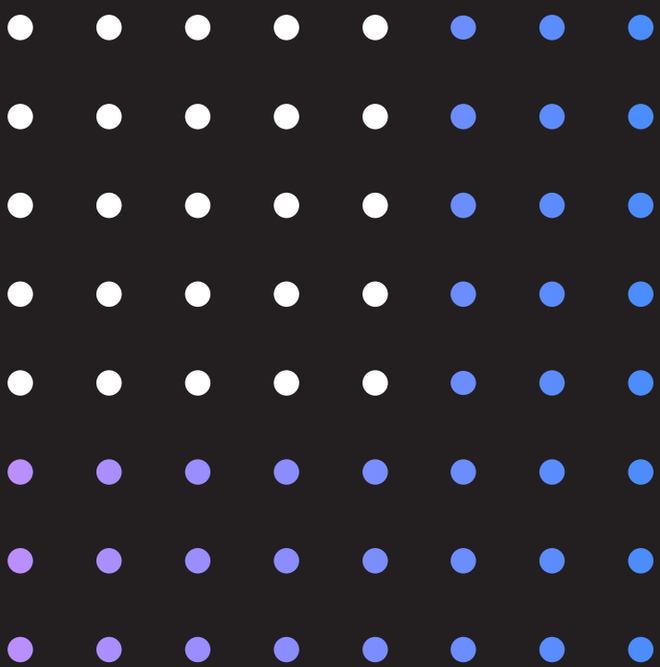
Dostarczanie gotowych do wykorzystana danych biznesowych dzięki funkcjonalności katalogowania danych i zarządzaniu repozytoriami data lake

Rozwiązanie IBM Watson Knowledge Catalog zapewnia zasilaną technologią maszynowego uczenia platformę do zarządzania danymi, która pomaga radzić sobie z wyzwaniami związanymi z obsługą repozytoriów data lake.



IBM

Spis treści

Główne wnioski

- Niewiele organizacji dostrzega wartość repozytoriów data lake utworzonych w celu przechowywania i analizowania ich danych w celu pozyskiwania na ich podstawie cennych wniosków biznesowych.
- Podejście DataOps rozwiązuje problemy organizacji z nieskutecznym dostępem, przygotowywaniem, integracją oraz udostępnianiem danych konsumentom, jednocześnie zachowując zgodność z zasadami polityk firmowych i regulacyjnych.
- Najczęściej występujące problemy z repozytoriami data lake to trudność oraz wysokie koszty importowania nowych źródeł danych do repozytoriów data lake; niemożność integracji wewnętrznych i zewnętrznych zbiorów danych; brak zaufania co do prawidłowości zarządzania danymi; brak dostępu do samoobsługowych narzędzi do przygotowywania danych; oraz niemożność odnalezienia i zrozumienia danych znajdujących się w repozytoriach data lake.
- Platforma do zarządzania danymi firmowymi z funkcją katalogowania, poprawy jakości danych oraz eksploracji danych może przeobrazić nieudany projekt repozytorium data lake w źródło danych o wysokiej wartości biznesowej.
- IBM Watson® Knowledge Catalog, obsługiwany przez system IBM Cloud Pak™ for Data, zapewnia oparty na technologii uczenia maszynowego (ML) katalog przeznaczony do przeglądania danych, katalogowania danych, poprawy jakości danych oraz zarządzania nimi. Rozwiązanie to pomaga użytkownikom danych szybko przeglądać, selekcjonować i filtrować, kategoryzować oraz udostępniać zasoby danych, zbiory danych oraz modele analityczne.
- Kiedy organizacjom brakuje dogłębnego zrozumienia danych, trudniej im zaufać oraz czynić z nich użytek za pomocą wszelkich form technologii sztucznej inteligencji (AI), w tym uczenia maszynowego oraz deep learning (głębokiego uczenia).

Rozwiązywanie problemów z repozytoriami data lake za pomocą podejścia DataOps

Dziesięć lat temu rozpoczął się proces poszukiwania elastycznego i wszechstronnego rozwiązania umożliwiającego budowę centralnego magazynu danych, który mógłby przechowywać dane całego przedsiębiorstwa. Rozwiązaniem tym jest data lake – środowisko służące do magazynowania danych o zastosowaniu ogólnym, mogące przechowywać praktycznie każdy rodzaj danych. Środowisko to pozwalało analitykom biznesowym oraz badaczom danych na zastosowanie najbardziej odpowiednich aparatów analitycznych w przypadku każdego zbioru danych, w jego oryginalnej lokalizacji.

Z reguły takie repozytoria data lake były budowane z wykorzystaniem platformy Apache Hadoop oraz systemu plików HDFS, w połączeniu z takimi systemami jak Apache Hive oraz Apache Spark. Wszystkie uzyskane w ten sposób repozytoria data lake zaczęły się powiększać, ujawniając w toku tego procesu cały szereg problemów. Chociaż ta technologia była fizycznie zdolna do skalowania w celu gromadzenia, przechowywania i analizowania ogromnych zbiorów danych ze strukturą i bez struktury, zbyt mało uwagi poświęcono kwestiom praktycznym, dotyczącym osadzenia tych funkcjonalności w firmowych przepływach pracy.

Do 2022 roku ponad 80% projektów opartych na repozytoriach data lake nie dostarczy odpowiednich wyników, ponieważ procesy wyszukiwania, inwentaryzacji oraz selekcjonowania danych okażą się największymi przeszkodami dla procesów analizy danych oraz barierą dla sukcesu nauki opartej na danych.[1] W rezultacie pytania, takie jak: „Jakie dane powinniśmy umieszczać w repozytoriach data lake?", „Kto będzie z tych danych korzystał?", „W jaki sposób ułatwić użytkownikom odnalezienie tych danych?", „Skąd pochodzą te dane?" oraz „Jak zapobiegać wykorzystywaniu tych danych do niewłaściwych celów?", często pozostawały bez odpowiedzi. Te poważne ograniczenia dotyczące ludzi, procesów oraz kwestii technologicznych zaowocowały nieudanymi wdrożeniami repozytoriów data lake.

Obecnie wiele organizacji wyciągnęło wnioski ze swoich błędów, dostosowało skład swoich zespołów kierowniczych do wdrażania repozytoriów data lake i przy drugiej, trzeciej, a nawet czwartej próbie wdrożenia wreszcie odniosło sukces – tym razem dzięki podejściu data operations DataOps.

Ten raport przedstawia ocenę częstych problemów z repozytoriami data lake oraz nowe podejścia, takie jak DataOps, które są w stanie przemienić przypominające bagno zbioru trudno dostępnych danych w centralne źródło użytecznych danych firmowych.

DataOps to oparta na współpracy praktyka zarządzania danymi skupiająca się na usprawnieniu komunikacji, integracji oraz automatyzacji przepływu danych pomiędzy zarządcami danych oraz konsumentami danych w ramach danej organizacji.

Przedstawiamy proces DataOps

DataOps powstał w wyniku połączenia najlepszych praktyk stosowanych w ramach metodyki DevOps i zarządzania danymi w jedną strukturę, która opiera się na współpracy w zakresie rozwoju oraz utrzymywania przepływu danych pomiędzy wieloma uczestnikami. Proces DataOps rozwiązuje problemy przedsiębiorstw z nieskutecznym dostępem, przygotowywaniem, integracją oraz udostępnianiem danych konsumentom, jednocześnie pozwalając im zachować zgodność z zasadami polityk firmowych i regulacyjnych. Te korzyści może odnieść zarówno jednostka biznesowa, zespół analityków, jak i nawet proces operacyjny.

Stosowanie tej metody wymaga sięgania po osoby, procesy oraz kwestie technologiczne, dzięki którym nieskuteczne wdrożenie repozytoriów data lake może stać się skuteczne. Od strony technologicznej DataOps podkreśla konieczność korzystania z w pełni zintegrowanej, kompleksowej platformy do pozyskiwania oraz integracji danych, poprawy jakości danych, zarządzania danymi oraz zużycia danych umożliwiającej utworzenie zarządzanego repozytorium data lake. Zasady poprawy jakości danych powinny być stosowane w sposób automatyczny jako element procesu pozyskiwania danych w celu podtrzymania stałego tempa przepływu danych biznesowych w ramach strumieni danych w całym przedsiębiorstwie. Proces pozyskiwania danych powinien być w pełni zintegrowany z katalogiem danych, który stanowi serce strumienia danych. Konsumenci danych powinni być w stanie uzyskać dostęp do szczegółów oceny danych oraz wyników profilowania danych zawartych w katalogu danych oraz mieć pewność, że ich organizacja pracuje z tymi samymi danymi.

Ilość danych wzrasta szybciej niż możliwości organizacji w zakresie czerpania korzyści z odpowiedniego ich wykorzystywania. Kiedy zapytaliśmy różne organizacje o ich największe wyzwania związane z korzystaniem z systemów wglądu, dowiedzieliśmy się, że: 1) 40% organizacji łączy istniejące procesy biznesowe w zakresie pozyskiwania danych z procesami ich analizy oraz 2) 39% pozyskuje, gromadzi, administruje dane oraz zarządza nimi w miarę wzrostu ich liczby.[2] Obecnie nie chodzi tylko o kwestię oszczędności czasu oraz o inwestycje w zasoby, które już zostały wprowadzone w ramach technologii data lake – chodzi o to, że nie ma już innej alternatywy. Począwszy od wdrażania technologii sztucznej inteligencji aż po przeprowadzanie skomplikowanych procesów analitycznych, istotne jest dostrzeganie jak największej liczby danych, co oznacza, że konieczna jest architektura zdolna do przechowywania oraz analizy wszystkich tych danych w jednym miejscu. W wielu przypadkach zarządzane repozytoria data lake to jedyna realistyczna możliwość spełnienia tych wymagań.

Dzisiejsze firmy potrafią – i są zmuszone – szukać sposobów na wydobywanie wartości ze swoich repozytoriów danych data lake, jednocześnie czyniąc je wsparciem dla strumieni gotowych danych biznesowych dla procesów DataOps.

Wyzwania związane z korzystaniem z firmowych repozytoriów data lake

Udostępnianie danych
Kiedy zespół pracowników firmy przejmie lub utworzy nowy zbiór danych, ma on z reguły silne poczucie wysokiego stopnia ważności zawartych w nim danych oraz dotyczących ich kwestii związanych z ich wrażliwością. Jeżeli zbiór danych zawiera na przykład poufne informacje handlowe, dane osobowe lub dane klientów, zespół będzie wiedział, w jaki sposób informacje te powinny i nie powinny być wykorzystywane, oraz podejmie odpowiednie środki ostrożności, aby żaden członek nie użył ich w nieprawidłowy sposób.

Członkowie zespołu będą również dbać o to, aby inni potencjalni użytkownicy danych, spoza ich grona, nie uzyskali podobnej do nich wiedzy na temat wartości danych ani na temat zagrożeń związanych z ich nieprawidłowym wykorzystaniem. Zagrożenia te w naturalny sposób sprawią, że członkowie zespołu będą wykazywać niezwykłą ostrożność w sytuacji udostępniania danych lub przechowywania ich w miejscach, w których nie będą one objęte ich kontrolą.

Takie podejście wpływa niekorzystnie na popularyzację repozytoriów danych. Jeżeli firmy będą postrzegać repozytoria data lake jako pozbawione kontroli składowiska danych, nie będą skłonne umieszczać w nich swoich cennych danych. W rezultacie różne działy tych firm nie będą w stanie czerpać korzyści z tych danych, co sprawi, że cała koncepcja wykorzystywania repozytoriów data lake jako samoobsługowych narzędzi do udostępniania danych firmowych rozpadnie się na drobne kawałki.

Integracja danych
Nawet jeżeli zespół zgodzi się na integrację danych z repozytorium data lake, proces ten może być niezwykle uciążliwy. Według oryginalnej koncepcji repozytorium data lake ma za zadanie importować dane w formacie bezpośrednim, nie wymagając skomplikowanych procesów ekstrakcji, transformacji i ładowania (ETL) danych przebiegających w przypadku tradycyjnych magazynów danych. Niemniej jednak, aby źródła danych nadawały się do jakiejkolwiek wnikliwej analizy, tak naprawdę prawie wszystkie z nich wymagają jakiegoś stopnia wstępnego przetwarzania.

W rezultacie integracja nowego źródła danych z repozytorium data lake często trwa kilka miesięcy. A ponieważ wiele tych danych było do tej pory przechowywanych w niewielkich silosach operacyjnych, a nie w systemach firmowych, integracja może objąć nawet dziesiątki lub setki źródeł.

Oznacza to, że w wielu przypadkach informacje potrzebne analitykom biznesowym lub badaczom danych nie zostaną dodane do repozytoriów data lake przez wiele miesięcy, a nawet lat. Może to poważnie zniechęcać firmy do zastosowania tego rozwiązania.

Przechowywanie danych
Podczas gdy koszt magazynowania towarów oraz zasobów obliczeniowych zdecydowanie zmalał w ciągu ostatnich kilku lat, klastry Hadoop nie są bezczynne. Przechowywanie dużych ilości danych w repozytoriach data lake jest o wiele tańsze niż magazynowanie ich w magazynach danych o wysokiej wydajności. Co nie zmienia faktu, że jego koszty mogą być nadal znaczne.

Co więcej w przeciwieństwie do danych, które są w sposób tradycyjny przechowywane w magazynach danych, stosunek wartości do objętości zbiorów big data przechowywanych w repozytoriach data lake jest porównywalnie niższy. Do ukrycia kilku wyjątkowo cennych igieł może być nieraz potrzebny wyjątkowo duży stóg siana.

Jeśli nie wiemy, które zbiory danych będą naprawdę użyteczne oraz cenne dla naszych badaczy, możemy zainwestować znaczne sumy pieniędzy w integrację i przechowywanie danych, które spadną na dno repozytorium data lake i nigdy nie zostaną użyte.

Wyszukiwanie danych
Załóżmy, że udało Ci się określić zbiory najcenniejszych danych do przechowywania, przekonać swoich interesariuszy do ich udostępnienia oraz przeprowadzić udany proces integracji z repozytorium data lake. Oprócz tego musisz

Wyzwania związane z korzystaniem z firmowych repozytoriów data lake

Rys. 1. Przedsiębiorstwa, które korzystają z technologii repozytoriów data lake, mogą napotkać jeden lub więcej z tych powszechnych problemów.

jeszcze doprowadzić do tego, by inni użytkownicy byli w stanie je wyszukać, zrozumieć oraz odpowiednio wykorzystać. Jakość danych w repozytorium data lake to kolejne wyzwanie. Nie masz pewności, czy dane są wysokiej czy niskiej jakości, ale mimo to chcesz je włączyć do repozytorium.

Niestety w przypadku większości repozytoriów data lake nie jest to łatwe. Dane są często przechowywane bez żadnego kontekstu, co utrudnia albo całkowicie uniemożliwia ich odkodowanie przez użytkownika bez konsultacji z ich pierwotnym właścicielem. Terminologia jest często tak specjalistyczna, że to samo słowo może oznaczać zupełnie coś innego w zależności od firmy, w której jest używane – albo być nieco inaczej definiowane. Prawdopodobieństwo pomyłki i błędnej interpretacji może być tak duże, że wiele zbiorów danych staje się całkowicie bezwartościowych lub nawet niebezpiecznych dla analityków, którzy nie są zaznajomieni z ich zawartością.

Łączenie danych wewnętrznych i zewnętrznych
Nawet największe repozytoria data lake nie powinny dążyć do przechowywania wszystkich zbiorów danych, jakie tylko mogłyby przydać się firmowym analitykom. Na przykład nie ma sensu importować do repozytorium kompletnej repliki danych z Google Maps, Weather.com® lub serwisu Bloomberg, tylko dlatego że firmowy badacz wyrazi chęć przeprowadzenia analizy geoprzestrzennej lub utworzenia algorytmu na podstawie danych pogodowych bądź cen akcji.

Twoje repozytoria data lake nie są w stanie pomieścić wszystkich danych, których analitycy potrzebują, aby wykonywać swoje obowiązki, dlatego są oni zmuszeni sięgać również po inne źródła. Fakt, że wiele przydatnych analiz to wynik pracy z wewnętrznymi i zewnętrznymi zbiorami danych, stanowi kolejne utrudnienie

i z perspektywy użytkownika jest to również czynnik obniżający oczekiwaną wartość repozytoriów data lake.

Przygotowywanie danych
Istnieje wiele czynników, które sprawiają, że przygotowywanie danych to trudny proces – począwszy od zrozumienia, gdzie szukać odpowiednich danych, aż po sposób ich formatowania. Przygotowywanie danych do wykorzystania przez analityków to najmniej efektywne i czasochłonne zadanie użytkowników danych. Użytkownicy danych przeznaczają większość swojego czasu na ich wyszukiwanie, czyszczenie oraz formatowanie zamiast skupiać się na analizie danych, modelowaniu oraz wyciąganiu wniosków mających pozytywny wpływ na rozwój ich firm.

Ograniczona dostępność zarządzanych zbiorów danych doprowadziła do zbyt dużego polegania na procesach IT w trakcie fazy przygotowawczej. Ograniczony dostęp sygnalizuje potrzebę usprawnienia funkcji samodzielnej obsługi oraz umiejętności korzystania z danych w ramach przedsiębiorstw, dzięki którym uda się pokonać te utrudnienia.

Jakość danych
Umieszczenie danych w repozytorium data lake może sprawić, że staną się one niezdatne do użycia. Przed umieszczeniem danych w repozytoriach data lake nie mają zastosowania żadne zasady dotyczące jakości ani zatwierdzania danych, stąd nie zawierają one danych, którym można zaufać i z których można korzystać. Wysokiej jakości dane to kluczowy element, który określa ich wiarygodność w procesie podejmowania istotnych decyzji. Dane to cenne zasoby, którymi należy zarządzać na każdym etapie procesu ich przemieszczania się w ramach organizacji. W miarę jak pojawia się coraz więcej źródeł informacji i stają się one coraz bardziej zróżnicowane, a inicjatywy w zakresie monitorowania zgodności przepisami coraz bardziej skupione na detalach, kluczowa staje się potrzeba integracji oraz dostępu do informacji pochodzących z różnych źródeł z zachowaniem ich spójności, wiarygodności i możliwości wielokrotnego wykorzystania.

Holistyczne podejście do budowania zarządzanych repozytoriów data lake

Warstwy magazynowania danych i aparaty analityczne w przypadku większości repozytoriów data lake opierają się na platformie Apache Hadoop i jej bogatym ekosystemie projektów open source. Nie budzi więc zdziwienia fakt, że społeczność open source związana z platformą Hadoop dostrzega problemy, które występują podczas aktualnych wdrożeń repozytoriów data lake, a ostatnimi czasy wiele projektów obrało sobie za cel rozwiązywanie tych problemów pojedynczo. Na rynku funkcjonuje wiele autorskich narzędzi, które mają na celu rozwiązywanie tych samych problemów.

Kuszące może być więc naprawianie problemów z danymi w sposób fragmentaryczny, w miarę jak się pojawiają. Kiedy liczba zbiorów danych staje się zbyt duża, aby zarządzanie nią było możliwe, stosuje się narzędzie katalogujące. Kiedy użytkownicy narzekają, że nie są w stanie odnaleźć danych, których potrzebują, wprowadza się funkcję wyszukiwania. A kiedy zarządca danych nie jest już w stanie ustalić skąd pochodzą dane lub kto ich używa, uruchamia się narzędzia śledzenia przepływu danych oraz zasady zarządzania danymi.

Brzmi to nieskomplikowanie, ale w praktyce to fragmentaryczne podejście niesie za sobą koszty w postaci znacznego zwiększenia stopnia skomplikowania danych oraz utrudnienia procesów ich konserwacji, a skala i zakres repozytoriów danych data lake wciąż wzrasta. W taki sam sposób, w jaki dodawanie nowych źródeł danych do repozytoriów data lake zwiększa stopień skomplikowania wymagań w zakresie procesów ETL, dodawanie nowych narzędzi skutkuje zwiększeniem stopnia skomplikowania wymagań niefunkcjonalnych dotyczących data lake.

W przypadku korzystania z osobnych narzędzi, a nie z kompleksowej platformy, która jest w stanie przeprowadzić procesy integracji, poprawy jakości oraz katalogowania danych w celu ich efektywnego wykorzystania przez firmowych analityków, okazuje się, że każde z tych narzędzi w inny sposób radzi sobie z błędami oraz w inny sposób prowadzi procesy rejestrowania. W rezultacie procesy, takie jak diagnostyka oraz rozwiązywanie błędów, mogą być bardzo czasochłonne.

Inny, również istotny mankament podejścia fragmentarycznego uwidacznia się, kiedy spojrzymy na problemy często napotykane przez repozytoria data lake nie z punktu widzenia technicznego, ale koncepcyjnego. Kluczowy wniosek jest następujący: problemy dotyczące skalowalności, łatwości wyszukiwania, integracji, jakości i zarządzania danymi są ze sobą nierozerwalnie powiązane. Rozwiązanie ich wymaga bardziej holistycznego podejścia.

Problemy dotyczące skalowalności, łatwości wyszukiwania, integracji, jakości i zarządzania danymi są ze sobą nierozerwalnie powiązane. Ich rozwiązanie wymaga bardziej holistycznego podejścia do kwestii zarządzania informacjami.

Opracowanie rozwiązanie IBM Watson Knowledge Catalog Data, katalogowanie i jakość danych

Rozwiązanie IBM Watson Knowledge Catalog obsługiwane przez IBM Cloud Pak for Data pomaga użytkownikom danych szybko przeglądać, selekcjonować, kategoryzować oraz udostępniać zasoby danych, zbiory danych, a także modele analityczne i relacje pomiędzy nimi innym członkom organizacji. Ponadto pomaga zespołom ds. zarządzania danymi opracowywać glosariusze biznesowe, polityki i zasady oraz zapewnia zaawansowane przepływy danych służących do zarządzania. Katalog pełni rolę pojedynczego punktu dostępu do prawidłowych danych dla inżynierów danych, zarządców danych, badaczy danych oraz analityków biznesowych umożliwiającego uzyskanie przez nich samodzielnego dostępu do danych, którym mogą zaufać i które mogą wykorzystać bez obaw.

Rozwiązania takie jak IBM Watson Knowledge Catalog obsługiwane przez by IBM Cloud Pak for Data mogą dostarczyć wszystkich funkcjonalności wymaganych do rozwiązywania głównych problemów dotyczących repozytoriów data lake za pomocą jednej kompleksowej platformy. Katalog pomaga docierać do źródła tych wzajemnie powiązanych ze sobą problemów powodujących częste awarie repozytoriów data lake i zapewnia skuteczne narzędzia umożliwiające gromadzenie i przechowywanie metadanych, a także zarządzanie nimi oraz śledzenie przepływu danych.

Pod wieloma względami wartość repozytoriów data lake zależy od zawartych w nich metadanych w takim samym stopniu, w jakim polegają one na samych danych. Bez metadanych wyjaśniających pochodzenie zbiorów danych, ich twórców, ich zawartość, zakres osób, które mogą z nich korzystać, oraz sposób, w jaki można z nich korzystać, same dane są praktycznie bezużyteczne. Użytkownicy danych nie byliby w stanie ich odnaleźć, a nawet gdyby im się to udało, nie będą w stanie zrozumieć, co znaczą, wykorzystać ich bez obaw, ani dowiedzieć się, w jaki sposób mogą je wykorzystać.

Watson Knowledge Catalog
Dostarczanie sprawdzonych i istotnych danych

Organizuj swoje dane

Wiedza

Dane muszą być kompletne, odpowiednie oraz dostępne z każdego punktu. Przeglądanie, klasyfikacja i zrozumienie wszystkich typów danych.

Zarządzaj swoimi danymi

Pewność

Dane muszą być bezpieczne, czyste oraz łatwe do odnalezienia, tak aby zachęcały użytkowników do zaufanego samodzielnego dostępu. Zrozumienie źródła pochodzenia danych oraz ich jakości.

Demokratyzuj swoje dane

Konsumowanie

Umiejętność napędzania procesów samodzielnego przeglądania danych oraz automatyzacji procesów podejmowania decyzji w celu rozwoju firmy. Przedstawianie wszystkich informacji użytkownikom, którzy ich potrzebują, i wyrażenie pozwolenia na ich dostęp do tych informacji.

Rys. 2. Rozwiązanie IBM Watson Knowledge Catalog zapewnia szeroki zakres funkcjonalności służących do przeglądania, katalogowania i zarządzania danymi.

Pojedyncze źródło sprawdzonych danych oraz pojedynczy punkt dostępu

IBM Watson Knowledge Catalog obsługiwany przez IBM Cloud Pak for Data odnosi się do tych kwestii, nadając kluczowy priorytet metadanym. Centralnym elementem rozwiązania jest potężny aparat katalogujący, który indeksuje wszystkie zbiory danych oraz zbiory analityczne, do których przedsiębiorstwo ma dostęp, bez względu na to, gdzie się one znajdują, może to być na przykład repozytorium data lake, magazyn danych lub system transakcyjny, a nawet zbiór arkuszy kalkulacyjnych. Bez względu na to, czy są to zbiory danych ze strukturą, czy bez struktury, lub zbiory przechowywane stacjonarnie czy w chmurze. Co więcej katalog może również obejmować zewnętrzne zbiory danych i źródeł, takie jak zastrzeżone usługi dotyczące przetwarzania danych, które subskrybuje Twoja firma, lub otwarte API.

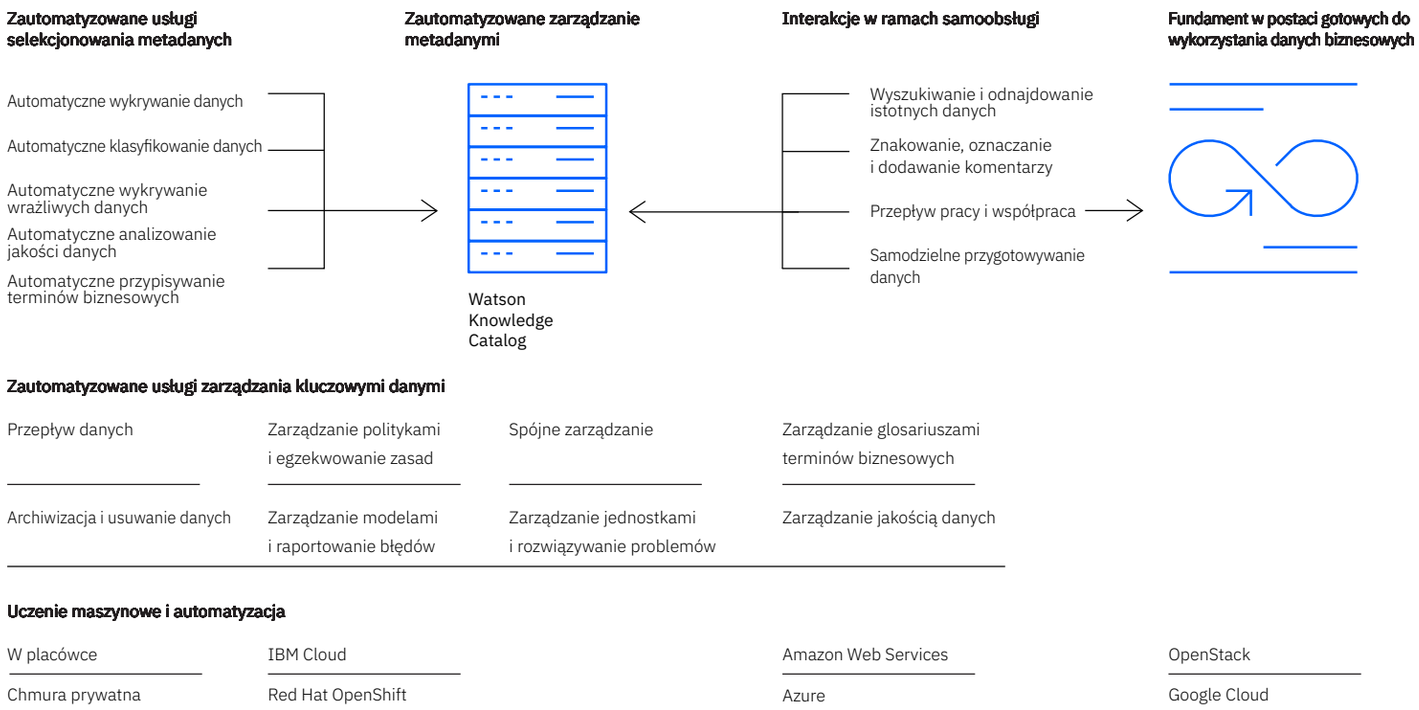
Oprócz zapewniania pojedynczego źródła sprawdzonych zbiorów danych, katalog danych stanowi również pojedynczy punkt dostępu do danych. Oparte na technologii sztucznej inteligencji funkcjonalności wyszukiwania oraz podpowiedzi pomagają analitykom biznesowym, badaczom, inżynierom danych oraz zespołom ds. zarządzania danymi w łatwy sposób odnajdować zasoby i prezentować dostępne metadane w celu ułatwienia użytkownikom zrozumienia, co znaleźli i do czego uzyskali dostęp, o ile taka wiedza jest dla nich przydatna.

Wbudowane samoobsługowe funkcjonalności dotyczące przygotowywania danych skracają czas potrzebny na transformację danych, aby skutecznie wykorzystać je w celach analitycznych oraz do zastosowań z wykorzystaniem technologii opartej na AI. Analitycy biznesowi oraz badacze danych nie muszą już tracić czasu na przygotowywanie i analizowanie danych. Integracja z dostępnym w ramach całego przedsiębiorstwa rozwiązaniem do przygotowywania danych, takim jak IBM® InfoSphere® Advanced Data Preparation, pomaga wyłowić utworzone za pomocą katalogu zarządzane zbiory danych, które mają największy wpływ na uzyskanie najbardziej trafnych wniosków biznesowych oraz działań dla ich użytkowników. Taka integracja pozwala pogłębić współpracę w ramach strumienia danych.

Problemy dotyczące skalowalności, łatwości wyszukiwania, integracji, jakości i zarządzania danymi są ze sobą nierozerwalnie powiązane. Ich rozwiązanie wymaga bardziej holistycznego podejścia do kwestii zarządzania informacjami.

Katalog ułatwia również zarządzanie danymi przez inspektorów danych osobowych dzięki funkcji znakowania i klasyfikacji zbiorów danych oraz automatycznego śledzenia ich przepływu i wykorzystania, a także dzięki wbudowanym glosariuszom biznesowym umożliwiającym standaryzację terminologii biznesowej. W rezultacie zarządcom danych łatwiej jest zrozumieć, co zawiera każdy zbiór danych, gdzie znajdują się dane wrażliwe lub osobowe i kto powinien mieć do nich dostęp.

Pojedynczy katalog do zarządzania wieloma źródłami danych wewnątrz i na zewnątrz organizacji



Rys. 3. Dzięki rozwiązaniu IBM Watson Knowledge Catalog inteligentny indeks metadanych i dane – ze strukturą i bez – mogą znajdować się w oryginalnych systemach, ale użytkownicy mogą przeglądać je szybko, co pozwala na szybsze przeprowadzanie procesów analitycznych.

Rozwiązanie IBM Watson Knowledge Catalog nadaje kluczowy priorytet metadanym, zapewniając tym samym spójne źródło danych oraz pojedynczy punkt dostępu do wszystkich zbiorów danych, które są dostępne dla Twojego przedsiębiorstwa.

Wbudowane funkcje inteligentnego odnajdowania danych

W celu jeszcze bardziej skutecznego odnajdowania danych katalog umożliwia użytkownikom znakowanie i umieszczanie komentarzy na temat zbiorów danych oraz zasobów analitycznych, co wzbogaca metadane oraz stanowi dodatkowy kontekst, dzięki któremu innym użytkownikom łatwiej jest znaleźć to, czego szukają. Rozwiązanie obejmuje również wbudowane algorytmy służące do inteligentnego odnajdowania danych, wykorzystujące technologię uczenia maszynowego w celu automatycznej klasyfikacji treści każdego zbioru danych. Dzięki identyfikacji wspólnych typów pól, takich jak imiona i nazwiska, adresy, kody pocztowe oraz numery ubezpieczenia społecznego, rozwiązanie redukuje potrzebę autorów do ręcznego oznaczania danych. Wykorzystuje procesy automatyzacji oraz uczenia maszynowego do automatyzacji w celu selekcjonowania danych oraz zarządzania metadanymi. Dzięki wbudowanym funkcjom poprawy jakości danych, rozwiązanie umożliwia wprowadzenie głębokiego profilowania danych oraz stosowanie zasad dotyczących jakości i zatwierdzania danych.

Zautomatyzowane operacje dotyczące danych zapewniają strumień wyselekcjonowanych danych o odpowiedniej jakości i poziomie zarządzania oraz pomaga zapewnić stały przepływ wysokiej jakości zarządzanych danych do repozytorium data lake.

W podobny sposób dodanie inteligentnego modelu metadanych zasobów firmowych stanowi wyjątkowy sposób pozwalający na automatyczne wdrażanie wymagań rozporządzenia RODO oraz CCPA.

Rozwiązanie IBM Watson Knowledge Catalog oparte na IBM Cloud Pak for Data pomaga dostarczać pewnych, wysokiej jakości, gotowych do wykorzystania w celach biznesowych danych praktycznie wszystkim użytkownikom danych.

Wszystkie komponenty rozwiązania zostały zaprojektowane jako mikrousługi za pomocą pojedynczego zestawu zasad projektowania oraz wspólnego podejścia do wymagań niefunkcjonalnych, takich jak skalowalność, zarządzanie błędami, bezpieczeństwo oraz rejestrowanie.

Rozwiązanie IBM Watson Knowledge Catalog zapewnia opartą na technologii uczenia maszynowego firmową platformę do zarządzania danymi, co oznacza, że jest ono gotowe na wykorzystanie technologii sztucznej inteligencji na jeszcze szerszą skalę.

Zamiast dezorientujących błędów oraz wąskich gardeł wydajnościowych, które są z reguły konsekwencją fragmentarycznego podejścia „zrób to sam”, IBM Watson Knowledge Catalog zapewnia opartą na uczeniu maszynowym platformę zarządzania, a więc jest gotowy na wdrażanie technologii AI.

Rozwiązanie IBM Watson Knowledge Catalog jest dostępne w trzech wersjach:

- Jako oprogramowanie, jako usługa i jako rozwiązanie SaaS oparte na IBM Cloud™
- W ramach IBM Cloud Pak for Data
- Wersja zintegrowana z IBM Watson Studio

Rozwiązania takie jak IBM Watson Knowledge Catalog mogą uwolnić pierwotnie zapowiadany potencjał inicjatyw opartych na repozytoriach data lake. Rozwiązanie Watson Knowledge Catalog wyposażone w inteligentne funkcjonalności dotyczące katalogowania oraz zarządzania pomaga budować zaufane i zarządzane repozytoria data lake dla technologii sztucznej inteligencji

Cztery korzyści związane z budowaniem zarządzanych repozytoriów data lake dla technologii sztucznej inteligencji

1. Budowanie zaufania i pewności co do prawidłowości danych dzięki wysokiej jakości oraz zarządzaniu

- Funkcjonalności poprawy jakości danych umożliwiają poprawę jakości danych Twojego przedsiębiorstwa oraz udostępnienie w ramach repozytoriów data lake danych o najwyższej jakości.
- Polityki w zakresie zarządzania są automatycznie ustalane oraz egzekwowane, a więc w przypadku znalezienia zbioru danych użytkownik wie, czy i jak może z niego korzystać.
- Istnieje możliwość selekcjonowania danych według ocen oraz komentarzy użytkowników, a także innych informacji, które pomagają określić, czy dany zbiór danych jest użyteczny.

2. Moc sprawcza dla użytkowników danych

- Firmowe zespoły LOB chętnie udostępniają swoje dane, ponieważ mają pewność, że będą one odpowiednio zarządzane i chronione przed niewłaściwym wykorzystaniem.
- Możliwość generowania ściślejszej współpracy i transformacji danych w zaufane zasoby biznesowe za pośrednictwem dynamicznych polityk ochrony danych oraz egzekwowania ich zasad.
- Dane są łatwiejsze do odnalezienia oraz ponownego wykorzystania w dłuższej perspektywie czasu, w miarę jak użytkownicy dodają do nich odpowiednie znaczniki oraz metadane, które pomagają innym użytkownikom wykorzystać je w najbardziej wartościowy sposób.
- Pojedynczy interfejs zapewnia dostęp do wszystkich zbiorów danych, będących własnością całej organizacji, bez względu na to, gdzie są one przechowywane.

3. Odzyskaj stracony czas

- Zautomatyzowane algorytmy służące do odnajdowania danych znacznie redukują czas i nakłady, które do tej pory przeznaczano na dodawanie metadanych do nowych zbiorów danych.
- Automatyczne procesy selekcjonowania danych oraz zarządzania metadanymi skracają czas odnajdowania metadanych oraz przypisywania terminów, a także skracają czas tworzenia glosariuszy terminów biznesowych.
- Dzięki prostym i intuicyjnym narzędziom do samodzielnego przygotowywania danych ich użytkownicy mogą poświęcać mniej czasu na przygotowywanie danych oraz więcej czasu na wyciąganie wniosków.
- Badacze danych oraz analitycy biznesowi osiągają lepsze wyniki w krótszym czasie.
- Inteligentne, oparte na technologii AI funkcje wyszukiwania ułatwiają odnajdowanie pożądanych danych w kilka sekund zamiast wielotygodniowego oczekiwania na dostarczenie ich przez inny zespół.

4. Zarządzanie rosnącą liczbą danych oraz kosztami

- Możliwość optymalizacji kosztów przechowywania poprzez unikanie wydatków związanych z umieszczaniem w repozytoriach data lake mało istotnych zbiorów danych.
- Możliwość wyświetlania wszystkich zewnętrznych zbiorów danych, które subskrybuje Twoja organizacja, co zmniejsza ryzyko opłacania większej liczby subskrypcji, niż jest to konieczne.
- Możliwość nadawania priorytetu nowym źródłom danych umieszczanych w repozytoriach data lake na podstawie zapotrzebowania użytkowników na dane, co pomaga w pierwszej kolejności zintegrować najważniejsze źródła.

Uwolnij potencjał swoich danych

Bez względu na to, czy pracujesz w biurze inspektora danych osobowych, w dziale IT czy jako badacz danych lub analityk zespołu LOB, Ty i Twoi koledzy dążycie do tego samego celu. Możesz utworzyć repozytorium data lake, które naprawdę spełnia obietnice. Dzięki niemu Twoja praca będzie łatwiejsza, a jednocześnie bardziej skuteczna. Dodatkowo masz możliwość odegrania kluczowej roli w zapewnieniu swojej firmie przewagi nad kilkoma konkurencyjnymi przedsiębiorstwami.

Jeżeli będziesz w stanie uporządkować zawartość swoich data lake, otworzysz przed swoją firmą możliwości, o których Twoi konkurenci mogą jedynie śnić. Prawdziwa korzyść czeka na tych, którzy jako pierwsi uwolnią potencjał dotychczas niewykorzystanych danych.

Zakończenie

Wiedza na temat lokalizacji wszystkich danych, tego, kto ich używa, a także ich przydatności analitycznej dla Twojej firmy.

Kluczowym elementem procesów DataOps są katalogi danych. To one ułatwiają zautomatyzowane zarządzanie otwartymi metadanymi poprzez integrację procesów zarządzania danymi, poprawę jakości danych oraz aktywne zarządzanie zasadami polityk ochrony danych.

Rozwiązanie Watson Knowledge Catalog wyposażone w inteligentne funkcjonalności dotyczące katalogowania oraz zarządzania pomaga budować zaufane i zarządzane repozytoria data lake dla technologii sztucznej inteligencji Katalog osadza w Twoim repozytorium data lake procesy integracji danych, poprawy jakości danych oraz zarządzania danymi, co ułatwia dostarczanie biznesowych danych gotowych do wykorzystania w procesach DataOps i sprawia że repozytorium stanowi spójne źródło sprawdzonych danych.

Więcej informacji

Więcej informacji można znaleźć na stronie:
ibm.com/cloud/watson-knowledge-catalog

IBM Polska
ul. 1 sierpnia
02-134 Warszawa

Dokument przygotowano w Stanach Zjednoczonych w październiku 2019 r. IBM, logo IBM, **ibm.com**, IBM Cloud, IBM Cloud Pak, IBM Watson oraz InfoSphere są znakami towarowymi International Business Machines Corp. zastrzeżonymi w jurysdykcjach wielu krajów.

1.Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders–Gartner, Sept 2019
2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018

ASW12449-PLPL-03