Badanie typów Big Data

Różnorodność jest przyprawą życia, a różnorodność jest jedną z zasad big data. W części 1 omawiamy znaczenie umiejętności zarządzania różnymi typami danych. Oczywiście duże zbiory danych obejmują wszystko, od transakcji w złotówkach, przez tweety, po obrazy i dźwięk. Dlatego korzystanie z Big Data wymaga zintegrowania wszystkich tych informacji w celu analizy i zarządzania danymi. Wykonywanie tego typu czynności jest trudniejsze niż się wydaje. Tu przeanalizujemy dwa główne typy danych, które składają się na duże zbiory danych - ustrukturyzowane i nieustrukturyzowane - oraz przedstawimy definicje i przykłady każdego z nich. Chociaż zarządzanie danymi istnieje od dawna, w świecie dużych zbiorów danych są dwa nowe czynniki:

* Niektóre źródła dużych zbiorów danych są w rzeczywistości nowe, podobnie jak generowane dane z czujników, smartfonów i tabletów.

* Wcześniej wytworzone dane nie zostały przechwycone ani zapisane i przeanalizowane w użyteczny sposób. Głównym powodem tego był brak technologii, która to umożliwiała. Innymi słowy, nie mieliśmy opłacalnego sposobu radzenia sobie z tymi wszystkimi danymi.

Istnieje wiele różnych sposobów wykorzystania Big Data do rozwiązywania problemów. Na przykład w niektórych sytuacjach chcesz zajmować się danymi w czasie rzeczywistym, na przykład podczas monitorowania danych o ruchu. W innych sytuacjach zarządzanie danymi w czasie rzeczywistym nie będzie konieczne, na przykład gdy zbierasz ogromne ilości danych, które chcesz przeanalizować w trybie wsadowym, aby określić nieoczekiwany wzorzec. Podobnie, czasami trzeba zintegrować wiele źródeł danych w ramach rozwiązania big data, więc zastanawiamy się, dlaczego warto zintegrować źródła danych. Najważniejsze jest to, że to, co chcesz zrobić ze swoimi ustrukturyzowanymi i nieustrukturyzowanymi danymi, wpływa na zakup technologii.

Definiowanie danych strukturalnych

Termin uporządkowane dane ogólnie odnosi się do danych o określonej długości i formacie. Przykłady uporządkowanych danych obejmują liczby, daty oraz grupy słów i liczb zwanych ciągami znaków (na przykład imię i nazwisko klienta, adres itp.). Większość ekspertów zgadza się, że tego rodzaju dane stanowią około 20 procent dostępnych danych. Dane strukturalne to dane, z którymi prawdopodobnie masz do czynienia. Zwykle jest przechowywany w bazie danych. Możesz zapytać go za pomocą języka, takiego jak język zapytań strukturalnych (SQL). Twoja firma może już zbierać ustrukturyzowane dane z „tradycyjnych” źródeł. Mogą to być dane zarządzania relacjami z klientami (CRM), dane planowania zasobów operacyjnych przedsiębiorstwa (ERP) oraz dane finansowe. Często te elementy danych są integrowane w hurtowni danych w celu analizy.

Odkrywanie źródeł dużych uporządkowanych danych

Chociaż może się to wydawać zwykłym biznesem, w rzeczywistości ustrukturyzowane dane przejmują nową rolę w świecie dużych zbiorów danych. Ewolucja technologii zapewnia nowe źródła danych strukturalnych, które są tworzone - często w czasie rzeczywistym i w dużych ilościach. Źródła danych są podzielone na dwie kategorie:

* Wygenerowane komputerowo lub maszynowo: dane wygenerowane maszynowo ogólnie odnoszą się do danych tworzonych przez maszynę bez udziału człowieka.

* Generowane przez człowieka: są to dane, które dostarczają ludzie w interakcji z komputerami.

Niektórzy eksperci twierdzą, że istnieje trzecia kategoria, która jest hybrydą między maszyną a człowiekiem. Tutaj jednak zajmujemy się dwiema pierwszymi kategoriami. Strukturyzowane dane generowane maszynowo mogą obejmować:

* Dane czujnika: Przykłady obejmują tagi ID częstotliwości radiowej (RFID), inteligentne liczniki, urządzenia medyczne i dane Globalnego Systemu Pozycjonowania (GPS). Na przykład RFID szybko staje się popularną technologią. Wykorzystuje małe chipy komputerowe do śledzenia przedmiotów na odległość. Przykładem tego jest śledzenie pojemników z produktami z jednego miejsca do drugiego. Gdy informacje są przesyłane z odbiornika, mogą trafić na serwer, a następnie zostać przeanalizowane. Firmy są tym zainteresowane w zakresie zarządzania łańcuchem dostaw i kontroli zapasów. Innym przykładem danych z czujników są smartfony zawierające czujniki, takie jak GPS, które można wykorzystać do zrozumienia zachowań klientów w nowy sposób.

* Dane dzienników internetowych: gdy działają serwery, aplikacje, sieci itp., Przechwytują wszelkiego rodzaju dane dotyczące ich aktywności. Może to oznaczać ogromne ilości danych, które mogą być przydatne, na przykład, do radzenia sobie z umowami o poziomie usług lub do przewidywania naruszeń bezpieczeństwa.

* Dane punktu sprzedaży: gdy kasjer przeciąga kod kreskowy dowolnego kupowanego produktu, generowane są wszystkie dane powiązane z produktem. Wystarczy pomyśleć o wszystkich produktach wszystkich osób, które je kupiły, i zrozumiesz, jak duży może być ten zestaw danych.

* Dane finansowe: Wiele systemów finansowych jest teraz zautomatyzowanych; działają w oparciu o predefiniowane reguły automatyzujące procesy. Dobrym tego przykładem są dane giełdowe. Zawiera uporządkowane dane, takie jak symbol firmy i wartość w dolarach. Niektóre z tych danych są generowane maszynowo, a inne przez ludzi.

Przykłady ustrukturyzowanych danych generowanych przez ludzi mogą obejmować:

* Dane wejściowe: są to dowolne dane, które człowiek może wprowadzić do komputera, takie jak imię i nazwisko, wiek, dochód, odpowiedzi w ankietach, które nie są dowolne i tak dalej. Te dane mogą być przydatne do zrozumienia podstawowych zachowań klientów.

* Dane strumienia kliknięć: dane są generowane za każdym razem, gdy klikasz łącze w witrynie internetowej. Dane te można analizować w celu określenia zachowań klientów i wzorców zakupowych.

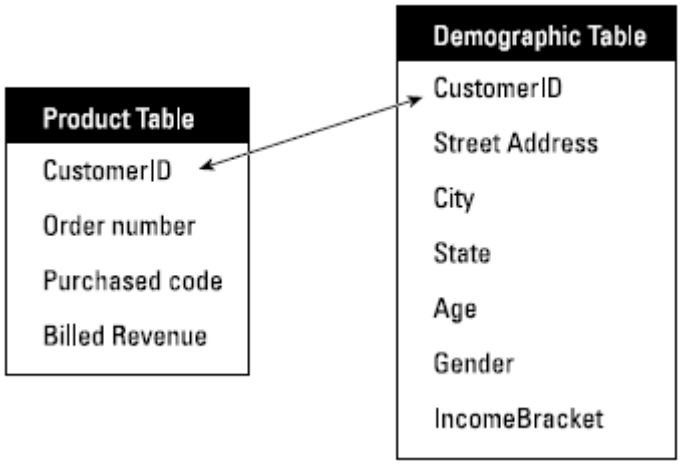
* Dane związane z grami: każdy Twój ruch w grze może zostać nagrany. Może to być przydatne do zrozumienia, w jaki sposób użytkownicy końcowi poruszają się po portfolio gier.

Masz pomysł. Niektóre z tych danych mogą same w sobie nie być tak duże, na przykład dane profilu. Jednak biorąc pod uwagę miliony innych użytkowników przesyłających te same informacje, rozmiar jest astronomiczny. Ponadto wiele z tych danych zawiera komponent czasu rzeczywistego, który może być przydatny do zrozumienia wzorców, które mają potencjał do przewidywania wyników. Najważniejsze jest to, że tego rodzaju informacje mogą być potężne i mogą być wykorzystywane do wielu celów.

Zrozumienie roli relacyjnych baz danych w Big Data

Trwałość danych odnosi się do sposobu, w jaki baza danych zachowuje swoje wersje po zmodyfikowaniu. Pradziadkiem trwałych magazynów danych jest system zarządzania relacyjnymi bazami danych (RDBMS). W początkach swojego istnienia przemysł komputerowy stosował obecnie uważane za prymitywne techniki trwałości danych. Około 1980 roku. Chociaż mechanizmy te były przydatne, były bardzo trudne do opanowania i zawsze wymagały od programistów systemowych

pisania niestandardowych programów do manipulowania danymi. Model relacyjny został wymyślony przez Edgara Codda, naukowca z IBM, w latach 70. XX wieku i był używany przez IBM, Oracle, Microsoft i innych. Jest nadal w powszechnym użyciu i odgrywa ważną rolę w ewolucji dużych zbiorów danych. Zrozumienie relacyjnej bazy danych jest ważne, ponieważ w przypadku dużych zbiorów danych używane są inne typy baz danych. Porównujemy różne rodzaje baz danych używane do dużych zbiorów danych w całej książce. W modelu relacyjnym dane są przechowywane w tabeli. Ta baza danych zawierałaby schemat - czyli strukturalną reprezentację tego, co znajduje się w bazie danych. Na przykład w relacyjnej bazie danych schemat definiuje tabele, pola w tabelach i relacje między nimi. Dane są przechowywane w kolumnach, po jednej dla każdego określonego atrybutu. Dane są również przechowywane w wierszach. Na przykład dwie tabele pokazane na rysunku przedstawiają schemat prostej bazy danych.



Pierwsza tabela przechowuje informacje o produkcie; druga przechowuje informacje demograficzne. Każdy ma różne atrybuty (identyfikator klienta, numer zamówienia, kod zakupu produktu itd.). Każdą tabelę można aktualizować nowymi danymi, a dane można usuwać, odczytywać i aktualizować. Jest to często realizowane w modelu relacyjnym przy użyciu strukturalnego języka zapytań (SQL). Innym aspektem modelu relacyjnego używającego języka SQL jest to, że można przesyłać zapytania do tabel przy użyciu wspólnego klucza (czyli relacji). Na rysunku wspólnym kluczem w tabelach jest IDKlienta. Możesz na przykład przesłać zapytanie, aby określić płeć klientów, którzy kupili określony produkt. Może to wyglądać mniej więcej tak:

Select CustomerID, State, Gender, Product from

“demographic table”, “product table” where

Product= XXYY

Chociaż relacyjne bazy danych dominowały w ciągu ostatnich kilku dziesięcioleci, mogą być trudne w użyciu, gdy masz do czynienia z ogromnymi strumieniami różnych typów danych. Producenci relacyjnych baz danych nie stoją jednak w miejscu i zaczynają wprowadzać relacyjne bazy danych przeznaczone dla dużych zbiorów danych. Ponadto ewoluowały nowe modele baz danych, aby pomóc ludziom zarządzać dużymi zbiorami danych. PostgresSQL (www.postgressql.org) to najczęściej używana relacyjna baza danych typu open source. Jego rozszerzalność i fakt, że jest dostępny na wielu odmianach komputerów typu mainframe, sprawiają, że jest to technologia podstawowa dla niektórych relacyjnych baz danych Big Data.

Definiowanie danych nieustrukturyzowanych

Dane nieustrukturyzowane to dane, które nie mają określonego formatu. Jeśli 20 procent danych dostępnych dla przedsiębiorstw to dane ustrukturyzowane, pozostałe 80 procent to dane nieustrukturyzowane. Dane nieustrukturyzowane to tak naprawdę większość danych, które napotkasz. Jednak do niedawna technologia ta nie pozwalała robić z nią zbyt wiele, poza przechowywaniem lub ręczną analizą.

Odkrywanie źródeł nieustrukturyzowanych danych

Dane nieustrukturyzowane są wszędzie. W rzeczywistości większość osób i organizacji żyje w oparciu o nieustrukturyzowane dane. Podobnie jak w przypadku danych ustrukturyzowanych, dane nieustrukturyzowane są generowane maszynowo lub przez człowieka. Oto kilka przykładów nieustrukturyzowanych danych generowanych maszynowo:

* Zdjęcia satelitarne: obejmują dane pogodowe lub dane, które rząd rejestruje w swoich zdjęciach satelitarnych. Wystarczy pomyśleć o Google Earth, a otrzymasz obraz (gra słów przeznaczona).

* Dane naukowe: obejmuje zdjęcia sejsmiczne, dane atmosferyczne i fizykę wysokich energii.

* Zdjęcia i wideo: obejmuje to bezpieczeństwo, nadzór i wideo o ruchu drogowym.

* Dane radarowe lub sonarowe: obejmują profile sejsmiczne pojazdów, meteorologiczne i oceanograficzne.

Poniższa lista przedstawia kilka przykładów nieustrukturyzowanych danych generowanych przez człowieka:

* Tekst wewnętrzny dla Twojej firmy: Pomyśl o całym tekście w dokumentach, dziennikach, wynikach ankiety i wiadomościach e-mail. Informacje przedsiębiorstwa stanowią obecnie duży procent informacji tekstowych na świecie.

* Dane z mediów społecznościowych: dane te są generowane z platform mediów społecznościowych, takich jak YouTube, Facebook, Twitter, LinkedIn i Flickr.

* Dane mobilne: obejmuje dane, takie jak wiadomości tekstowe i informacje o lokalizacji.

* Treść witryny: pochodzi z dowolnej witryny dostarczającej treści nieustrukturyzowane, takiej jak YouTube, Flickr lub Instagram.

Lista jest długa.

Niektórzy uważają, że termin nieustrukturyzowane dane jest mylący, ponieważ każdy dokument może zawierać własną, specyficzną strukturę lub formatowanie oparte na oprogramowaniu, które go stworzyło. Jednak to, co jest wewnętrzne w dokumencie, jest naprawdę nieustrukturyzowane. Jak dotąd dane nieustrukturyzowane są największym elementem równania danych, a przypadki użycia danych nieustrukturyzowanych szybko się rozszerzają. Po samej stronie tekstu analiza tekstu może być wykorzystywana do analizowania nieustrukturyzowanego tekstu oraz do wyodrębniania odpowiednich danych i przekształcania tych danych w ustrukturyzowane informacje, które można wykorzystać na różne sposoby. Na przykład popularnym przypadkiem użycia dużych zbiorów danych jest analiza mediów społecznościowych do wykorzystania w rozmowach z klientami o dużym natężeniu. Ponadto nieustrukturyzowane dane z notatek call center, e-maili, pisemnych komentarzy w ankiecie i innych dokumentów są analizowane w celu zrozumienia zachowań klientów. Można to połączyć z mediami społecznościowymi z dziesiątek milionów źródeł, aby zrozumieć doświadczenia klientów.

Przeglądanie danych częściowo ustrukturyzowanych

Dane częściowo ustrukturyzowane to rodzaj danych, które mieszczą się między danymi ustrukturyzowanymi i nieustrukturyzowanymi. Dane częściowo ustrukturyzowane niekoniecznie są zgodne z ustalonym schematem (czyli strukturą), ale mogą samo się opisywać i mogą mieć proste pary etykieta / wartość. Na przykład pary etykieta / wartość mogą obejmować: < family > = Jones, < mother > = Jane i < daughter > = Sarah. Przykłady danych częściowo ustrukturyzowanych obejmują EDI, SWIFT i XML. Można je traktować jako rodzaj ładunków służących do przetwarzania złożonych zdarzeń.

Zrozumienie roli CMS w zarządzaniu dużymi zbiorami danych

Organizacje przechowują niektóre nieustrukturyzowane dane w bazach danych. Wykorzystują jednak również systemy zarządzania treścią przedsiębiorstwa (CMS), które mogą zarządzać całym cyklem życia treści. Może to obejmować zawartość sieci Web, treść dokumentów i inne nośniki formularzy. Według Association for Information and Image Management (AIIM; www.aiim.org;), organizacji non-profit, która zapewnia edukację, badania i najlepsze praktyki, zarządzanie treścią w przedsiębiorstwie (ECM) obejmuje „strategie, metody i narzędzia używane do przechwytywanie, przechowywanie, przechowywanie i dostarczanie treści oraz dokumentów związanych z procesami organizacyjnymi, zarządzanie nimi. ” Technologie zawarte w ECM obejmują zarządzanie dokumentami, zarządzanie rekordami, obrazowanie, zarządzanie przepływem pracy, zarządzanie treścią WWW i współpracę. Wokół zarządzania treścią rozwinęła się cała branża, a wielu dostawców usług zarządzania treścią skaluje swoje rozwiązania, aby obsługiwać duże ilości nieustrukturyzowanych danych. Jednak nowe technologie również ewoluują, aby pomóc w obsłudze danych nieustrukturyzowanych i analizie danych nieustrukturyzowanych. Niektóre z nich obsługują zarówno dane ustrukturyzowane, jak i nieustrukturyzowane. Niektóre obsługują strumienie w czasie rzeczywistym. Należą do nich technologie takie jak Hadoop, MapReduce i przesyłanie strumieniowe. Systemy przeznaczone do przechowywania treści w postaci systemów zarządzania treścią nie są już samodzielnymi rozwiązaniami. Są raczej częścią ogólnego rozwiązania do zarządzania danymi. Na przykład Twoja organizacja może monitorować kanały Twittera, które mogą następnie programowo uruchamiać wyszukiwanie CMS. Teraz osoba, która uruchomiła tweet (być może szukając rozwiązania problemu), otrzymuje odpowiedź, która wskazuje lokalizację, w której dana osoba może znaleźć produkt, którego może szukać. Największą korzyścią jest to, że tego typu interakcja może mieć miejsce w czasie rzeczywistym. Ilustruje również wartość wykorzystania nieustrukturyzowanych, ustrukturyzowanych (dane klientów dotyczące osoby, która tweetowała) i częściowo ustrukturyzowanych (rzeczywista treść w systemie CMS) w czasie rzeczywistym. W rzeczywistości prawdopodobnie użyjesz podejścia hybrydowego, aby rozwiązać swoje problemy związane z dużymi zbiorami danych. Na przykład nie ma sensu przenoszenie całej treści wiadomości, na przykład, do Hadoop w Twojej siedzibie, ponieważ ma to pomóc w zarządzaniu nieustrukturyzowanymi danymi.

Spojrzenie na wymagania w czasie rzeczywistym i inne niż w czasie rzeczywistym

Jak omówiliśmy w poprzednich sekcjach, duże zbiory danych często dotyczą robienia rzeczy, które nie były powszechnie możliwe, ponieważ technologia nie była wystarczająco zaawansowana lub koszt jej wykonania był zaporowy. Wielką zmianą, z którą mamy do czynienia w przypadku dużych zbiorów danych, jest możliwość wykorzystania ogromnych ilości danych bez całego skomplikowanego programowania, które było wymagane w przeszłości. Wiele organizacji znajduje się w punkcie zwrotnym, jeśli chodzi o zarządzanie dużymi ilościami złożonych danych. Podejścia związane z dużymi zbiorami danych pomogą zachować równowagę, abyśmy nie przekraczali granic, gdy ilość, różnorodność i szybkość zmian danych. Firmom trudno było zarządzać rosnącymi ilościami danych, którymi trzeba było zarządzać z dużą szybkością. Organizacje musiały zadowolić się analizą małych podzbiorów danych, w których często brakowało krytycznych informacji, aby uzyskać pełny obraz, który dane mogłyby ujawnić. Wraz z ewolucją i wdrażaniem technologii big data będziemy mogli łatwiej

analizować dane i je wykorzystywać do podejmowania decyzji lub działań. Aspekty Big Data w czasie rzeczywistym mogą być rewolucyjne, gdy firmy muszą rozwiązywać istotne problemy. Jaki jest wpływ organizacji na przetwarzanie danych przesyłanych strumieniowo w czasie rzeczywistym? Ogólnie rzecz biorąc, podejście w czasie rzeczywistym jest najbardziej odpowiednie, gdy odpowiedź na problem zależy od czasu i ma kluczowe znaczenie dla firmy. Może to być związane z zagrożeniem dla czegoś ważnego, takiego jak wykrywanie działania sprzętu szpitalnego lub przewidywanie potencjalnego ryzyka włamania. Poniższa lista przedstawia przykłady sytuacji, w których firma chce wykorzystać te dane w czasie rzeczywistym, aby uzyskać szybką przewagę:

* Monitorowanie wyjątku z nową informacją, taką jak oszustwa / dane wywiadowcze

* Monitorowanie kanałów informacyjnych i mediów społecznościowych w celu określenia wydarzeń, które mogą wpłynąć na rynki finansowe, takich jak reakcja klienta na ogłoszenie nowego produktu

* Zmiana pozycji reklamy podczas dużego wydarzenia sportowego na podstawie transmisji na Twitterze w czasie rzeczywistym

* Dostarczenie kuponu klientowi na podstawie tego, co kupił w punkcie sprzedaży

Czasami dane strumieniowe pojawiają się naprawdę szybko i nie obejmują wielu różnych źródeł, czasami istnieje duża różnorodność, a czasami jest to połączenie tych dwóch. Pytanie, które musisz sobie zadać, jeśli przenosisz się do czasu rzeczywistego, brzmi: Czy ten (problem) można rozwiązać za pomocą tradycyjnych funkcji zarządzania informacjami, czy też potrzebujemy nowszych funkcji? Czy sama objętość lub prędkość przytłoczy nasze systemy? Często jest to połączenie tych dwóch. Jeśli więc potrzebujesz funkcji czasu rzeczywistego, jakie są wymagania infrastruktury, aby to obsługiwać? Jednak na poniższej liście wymieniono kilka rzeczy, które należy wziąć pod uwagę, jeśli chodzi o zdolność systemu do pozyskiwania danych, przetwarzania ich i analizowania w czasie rzeczywistym:

* Niskie opóźnienie: opóźnienie to opóźnienie, które umożliwia wykonanie usługi w środowisku. Niektóre aplikacje wymagają mniejszych opóźnień, co oznacza, że muszą odpowiadać w czasie rzeczywistym. Strumień w czasie rzeczywistym będzie wymagał małego opóźnienia. Musisz więc pomyśleć o mocy obliczeniowej, a także o ograniczeniach sieci.

* Skalowalność: Skalowalność to zdolność do utrzymania określonego poziomu wydajności nawet przy rosnącym obciążeniu.

* Wszechstronność: system musi obsługiwać zarówno ustrukturyzowane, jak i nieustrukturyzowane strumienie danych.

* Format natywny: użyj danych w ich natywnej formie. Transformacja wymaga czasu i pieniędzy. Możliwość wykorzystania idei przetwarzania złożonych interakcji w danych, które wyzwalają zdarzenia, może mieć charakter transformacyjny.

Potrzeba przetwarzania stale rosnących ilości odmiennych danych jest jednym z kluczowych czynników wpływających na wdrażanie usług w chmurze.

Łączenie dużych ilości danych

To, co chcesz zrobić ze swoimi ustrukturyzowanymi i nieustrukturyzowanymi danymi, wskazuje, dlaczego możesz wybrać jedną technologię zamiast innej. Określa również potrzebę zrozumienia struktur danych przychodzących, aby umieścić te dane we właściwym miejscu.

Zarządzanie różnymi typami danych

Rysunek przedstawia pomocną tabelę, która przedstawia niektóre cechy dużych zbiorów danych i typy systemów zarządzania danymi, których możesz chcieć użyć do rozwiązania każdego z nich. Nie spodziewamy się, że wiesz jeszcze, co to jest; są one opisane w kolejnych częściach.

	Batch	Streaming	Complex Query
Structured	Hadoop	Key/Value	RDBMS
Unstructured	Document	Graph Spatial	Columnar
Both	Hybrid	Hybrid	Hybrid

Integracja typów danych w środowisku dużych zbiorów danych

Innym ważnym aspektem dużych zbiorów danych jest to, że często nie musisz posiadać wszystkich danych, których będziesz używać. Wiele przykładów wskazuje na to. Możesz wykorzystywać dane z mediów społecznościowych, dane pochodzące ze statystyk branżowych stron trzecich, a nawet dane pochodzące z satelitów. Wystarczy pomyśleć o mediach społecznościowych, a zrozumiesz, o co nam chodzi. Często konieczne jest zintegrowanie różnych źródeł. Dane te mogą pochodzić ze wszystkich systemów wewnętrznych, zarówno ze źródeł wewnętrznych, jak i zewnętrznych, lub ze źródeł całkowicie zewnętrznych. Wiele z tych danych mogło zostać wcześniej wyciszonych. Dane nie muszą przychodzić do Ciebie w czasie rzeczywistym. Po prostu możesz go mieć dużo i ma to odmienny charakter. Może to nadal kwalifikować się jako problem z dużymi zbiorami danych. Oczywiście możesz również zmierzyć się ze scenariuszem, w którym widzisz ogromne ilości danych, z dużą prędkością, i ma on odmienny charakter. Chodzi o to, że nie uzyskasz wartości biznesowej, jeśli masz do czynienia z różnymi źródłami danych jako zbiorem odłączonych silosów informacji. Potrzebne komponenty obejmują łączniki i metadane, które omówimy w dalszej części.

Złącza

Chcesz mieć kilka łączników, które umożliwiają pobieranie danych z różnych źródeł dużych zbiorów danych. Może chcesz mieć złącze do Twittera lub Facebooka. Być może musisz zintegrować swoją hurtownię danych ze źródłem dużych zbiorów danych, które znajduje się poza Twoją siedzibą, aby móc analizować oba te źródła danych razem.

Metadane

Krytycznym elementem integracji wszystkich tych danych są metadane. Metadane to definicje, odwzorowania i inne cechy używane do opisania sposobu znajdowania, uzyskiwania dostępu i korzystania z komponentów danych (i oprogramowania) firmy. Przykładem metadanych są dane o numerze konta. Może to obejmować numer, opis, typ danych, imię i nazwisko, adres, numer telefonu i poziom prywatności. Metadane mogą pomóc w organizowaniu magazynów danych i radzeniu sobie z nowymi i zmieniającymi się źródłami danych. Chociaż idea metadanych nie jest nowa, zmienia się i ewoluuje w kontekście dużych zbiorów danych. W tradycyjnym świecie metadanych ważne jest, aby mieć katalog, który zapewnia jeden widok wszystkich źródeł danych. Ale ten katalog będzie musiał być inny, jeśli nie będziesz kontrolować wszystkich tych źródeł danych. Możesz potrzebować narzędzia analitycznego, które pomoże ci zrozumieć podstawowe metadane.