Rozpoznawanie różnych typów danych

Kiedy jesteś w zespole zajmującym się analizą danych, często masz do czynienia z wieloma różnymi typami danych. Te różne typy będą kluczowym czynnikiem przy określaniu sposobu przechowywania danych. Technologie takie jak NoSQL zapewniają dużą elastyczność w przechowywaniu różnych typów danych. Relacyjne bazy danych zapewniają mniejszą elastyczność, ale czasami są łatwiejsze w obsłudze i generalnie łatwiej jest generować raporty w relacyjnych bazach danych. Kiedy myślisz o tym, jak chcesz przechowywać swoje dane, musisz zrozumieć różne typy danych. To samo dotyczy każdego magazynu. Niektóre bazy danych są zoptymalizowane pod kątem określonych typów danych. Tak jak nie chciałbyś przechowywać kanapki w dzbanku na wodę, nie chciałbyś konfigurować relacyjnej bazy danych do przechowywania niewłaściwego typu danych. Istnieją trzy rodzaje danych, które Twój zespół powinien wziąć pod uwagę:

• Strukturalne: Dane, które mają określony format w określonej kolejności.

• Częściowo ustrukturyzowane: dane z pewną strukturą, ale także z dodatkową elastycznością zmiany nazw pól i tworzenia wartości.

• Nieustrukturyzowane: dane, które nie są zgodne ze schematem i nie mają modelu danych.

W kolejnych sekcjach szczegółowo omawiamy każdy z tych typów danych, a następnie omawiamy, czym są duże śmieci i przedstawiamy kilka wskazówek, jak je przeszukiwać.

Uproszczenie danych dzięki uporządkowanym danym

Pierwszy typ danych jest pod wieloma względami najprostszy. Jest to powszechnie określane jako dane strukturalne. Dane strukturalne to dane, które mają określony format, w określonej kolejności. To jak cegły i zaprawa w świecie baz danych - jest tani, nieelastyczny i wymaga dużo wcześniejszego projektu. Dobrym przykładem danych strukturalnych jest typowy arkusz kalkulacyjny dla biura. Kiedy wypełniasz swoje wiersze danymi, musisz trzymać się dość sztywnego formatu i struktury. Na przykład możesz mieć kolumnę o nazwie „Data zakupu”. Każde pole musi przestrzegać ścisłych wytycznych. Nie możesz umieścić „wtorku” w jednym wierszu, a „marzec” w następnym. Musisz przestrzegać określonego formatu; na przykład miesiąc numeryczny, po którym następuje ukośnik, dzień i rok (coś w formacie MM/DD/RRRR). Ten format i struktura nazywa się modelem danych. Dane strukturalne opierają się na tym modelu danych. Model danych jest podobny do schematu danych, z tą różnicą, że schemat służy do definiowania całej struktury bazy danych. Model danych definiuje strukturę poszczególnych pól. W ten sposób definiujesz, co trafia do każdego pola danych. Ty decydujesz, czy pole będzie zawierało tekst, liczby, daty lub cokolwiek innego. Pomyśl o przykładzie arkusza kalkulacyjnego i co może się stać, jeśli zignorujesz model danych. Jeśli wpiszesz wtorek w polu Data zakupu w jednym wierszu, a marzec w innym, co się stanie, jeśli zechcesz utworzyć raport, który wyświetli wszystkie zakupy w marcu? Jak byś to zrobił? Czy użyłbyś numeru trzy? Czy użyłbyś słowa marzec? Na pewno nie użyłbyś słowa wtorek. Jeśli zrobisz tego typu wprowadzanie danych, Twój arkusz kalkulacyjny zostanie wypełniony śmieciami danych. Za każdym razem, gdy próbujesz posortować dane lub utworzyć raport, pojawi się kilka wierszy z nieprawidłowymi danymi. Następnie musiałbyś wrócić i wyczyścić go lub po prostu usunąć je z raportu. Dlatego wiele aplikacji do obsługi arkuszy kalkulacyjnych ma reguły formatowania, które zmuszają Cię do przestrzegania określonego modelu podczas wprowadzania danych. To samo dotyczy baz danych. Wiele baz danych odrzuca dane niezgodne z modelem. Często witryna internetowa (lub oprogramowanie pośredniczące) używane do zbierania danych jest ustawione na określony typ i format dla różnych pól. Relacyjne bazy danych doskonale radzą sobie z gromadzeniem uporządkowanych danych, co oznacza, że istnieje wiele uporządkowanych danych. Wiele danych, do których uzyskujesz dostęp w witrynach internetowych lub aplikacjach mobilnych, pochodzi z danych

strukturalnych. Twoje wyciągi bankowe, informacje o lotach, rozkłady jazdy autobusów, a nawet książka adresowa są formami danych strukturalnych. Nie oznacza to, że większość danych jest ustrukturyzowana. W rzeczywistości większość danych nie ma określonego formatu i struktury. W rzeczywistości niektóre z bardziej interesujących danych w ogóle nie mają żadnej struktury. Dane takie jak wideo, audio i strony internetowe nie mają zdefiniowanej struktury. Jako członek zespołu zajmującego się badaniem danych musisz połączyć typ danych z metodą ich zbierania. Jeśli korzystasz z relacyjnej bazy danych, ograniczasz się do w większości uporządkowanych danych. Dzięki ustrukturyzowanym danym tworzenie raportów jest zwykle dość proste. Możesz użyć strukturalnego języka zapytań lub SQL, aby pobrać dane z bazy danych i wyświetlić je w standardowym formacie. Jeśli używasz klastra NoSQL, możesz pracować ze wszystkimi typami danych, ale tworzenie raportów będzie trudniejsze. To są wszystkie decyzje, o których Twój zespół musi się zastanowić.

Udostępnianie częściowo ustrukturyzowanych danych

Kiedy masz uporządkowane dane w swojej relacyjnej bazie danych, wszystko na świecie wydaje się zdefiniowane i dobrze zorganizowane. To tak, jakbyś miał wszystkie swoje przyprawy w słoikach z przyprawami – wiesz, gdzie wszystko jest i dokładnie wiesz, gdzie to znaleźć. Jednak niewiele aplikacji pozostaje tak prostych. Dane częściowo ustrukturyzowane są nieco trudniejsze do zdefiniowania niż dane strukturalne, dlatego jako przykładu posłużymy się naszą witryną internetową z butami do biegania. Wyobraź sobie, że korzystasz z relacyjnej bazy danych dla witryny internetowej poświęconej butom do biegania. Ma cztery tabele: buty, klienci, ich adres i opcje wysyłki. Wszystkie Twoje uporządkowane dane pasują do modelu danych. Daty są standardowe, a kody pocztowe są standardowe. Sprawy idą gładko. Wszystko wydaje się w porządku na świecie. Następnie otrzymałeś e-mail od swojego przewoźnika. Przewoźnik twierdzi, że możesz radykalnie obniżyć koszty, dodając informacje bezpośrednio do swojej bazy danych. Wystarczy przeszukać ich bazę danych, pobrać jeden z regionalnych kodów wysyłkowych, a następnie dodać go do zamówienia i utworzyć nowy rekord. Powinno to być łatwe, ponieważ ich baza danych jest taka sama jak Twoja. To wszystko uporządkowane dane w relacyjnej bazie danych. Problem polega na tym, że ich schemat nie jest taki sam jak twój. Nazwałeś swój kod pocztowy „Kod pocztowy". Nazwali swoje dane kodu pocztowego „kodem pocztowym". Nie obchodzi Cię, czy buty są wysyłane do firmy lub rezydencji. Robią. Nie precyzujesz, czy to dom, czy mieszkanie. Mają różne stawki dla każdego. Teraz potrzebujesz sposobu na wymianę danych strukturalnych z ich danymi strukturalnymi, nawet jeśli każdy z nich jest innym schematem. Aby rozwiązać ten problem, musisz pobrać dane przewoźnika i powiązany schemat. Gdy klient zamawia but, Twoja baza danych wyśle kod pocztowy do bazy przewoźnika. Zwróci pakiet danych, który zawiera ich wersję adresu wraz z nazwami pól i modelem danych. Pamiętaj, że używali nazwy „kod pocztowy" dla kodów pocztowych. Zostanie to uwzględnione w nowych danych. Ich dane mają pewne cechy danych strukturalnych. Jest dobrze zorganizowany i ma standardowy format. Pola tekstowe to tekst. Pola daty to daty. Ale dane zawierają ich schemat. Przewoźnik może używać dowolnych nazw. Dlatego ten rodzaj danych nazywa się danymi częściowo ustrukturyzowanymi. Dane częściowo ustrukturyzowane są jeszcze bardziej popularne niż dane ustrukturyzowane. Ma strukturę, ale ta struktura zależy od źródła. Przez cały czas będziesz pracować z częściowo ustrukturyzowanymi danymi. Twój e-mail to częściowo ustrukturyzowane dane. Ma dość spójną strukturę. Zawsze masz nadawcę i odbiorcę, ale wiadomość może się różnić. Treść wiadomości może być po prostu tekstem lub zawierać obrazy lub załączniki. Zespoły zajmujące się analizą danych zwykle pracują z większą ilością danych częściowo ustrukturyzowanych niż danych ustrukturyzowanych. Istnieje wiele ilości wiadomości e-mail, blogów i zawartości witryn sieci społecznościowych, które można przeanalizować. Istnieje kilka terminów, które są dość powszechne, gdy mówimy o pracy z danymi częściowo ustrukturyzowanymi i ich wymianie. Jednym z nich jest typ danych Extended Markup Language (XML), który jest starszym półstrukturalnym typem danych używanym do wymiany informacji. Istnieje również JavaScript Object

Notation (JSON), który jest zaktualizowanym sposobem wymiany danych częściowo ustrukturyzowanych. Często jest to preferowany typ danych dla usług internetowych.

Uwzględnienie danych częściowo ustrukturyzowanych to dobry sposób na zadawanie ciekawszych pytań. Wróćmy do przykładu butów do biegania. Załóżmy, że chciałeś uzyskać opinie klientów na temat zamówień na buty do biegania. Możesz pobrać częściowo ustrukturyzowane dane z niektórych największych serwisów społecznościowych, a następnie połączyć te dane z ustrukturyzowanymi danymi, które posiadasz o swoim kliencie. Jeśli klient jest niezadowolony ze swoich butów, możesz wysłać mu kupon z przeprosinami. Są to problemy, które można wykryć, korzystając z danych ustrukturyzowanych i częściowo ustrukturyzowanych. Twój zespół może zacząć badać zadowolenie klienta z zakupu.

Zbieranie nieustrukturyzowanych danych

Najpopularniejszym typem danych jest wszystko, co nie jest ustrukturyzowane lub częściowo ustrukturyzowane: dane nieustrukturyzowane. Niektórzy analitycy szacują, że 80% danych jest nieustrukturyzowanych. Kiedy się nad tym zastanowisz, ma to sens. Pomyśl o danych, które tworzysz każdego dnia: za każdym razem, gdy zostawiasz wiadomość głosową, każde zdjęcie przesłane do serwisu Facebook, notatkę OneNote lub prezentację PowerPoint utworzoną w pracy, a nawet dane generowane podczas wyszukiwania w sieci Web. To wszystko jest nieustrukturyzowane. Co zatem mają ze sobą wspólnego te wszystkie dane? To największe wyzwanie. Odpowiedź to niewiele. Jest pozbawiony schematów. Pamiętaj, że schemat to mapa, która pokazuje pola danych, tabele i relacje. Nie masz tej mapy z nieuporządkowanymi danymi. Ponadto format danych nieustrukturyzowanych zależy od pliku. Dokument Microsoft Word może mieć ustawiony format, ale ten format jest używany tylko przez tę aplikację. To nie jest format dla całego tekstu. Dlatego zazwyczaj nie możesz edytować dokumentów Microsoft Word w innym programie. Oznacza to również, że nie ma ustalonego modelu danych. Nie ma spójnego miejsca, w którym można szukać nazw pól i danych. Gdybyś miał tuzin dokumentów, jak mógłbyś ustalić ich tytuł i zawartość? Co by było, gdyby niektóre z nich były plikami PDF, niektóre były dokumentami Microsoft Word, a niektóre były prezentacjami PowerPoint? Każdy z nich ma zastrzeżony format. Nie ma pola do wyszukania z etykietą „tytuł dokumentu”. To wyzwanie, nad którym od lat pracują firmy zajmujące się wyszukiwaniem, takie jak Google. Jak pracujesz z danymi bez ustawionego formatu i bez spójnego modelu danych? Za każdym razem, gdy przeszukujesz te wyszukiwarki, zobaczysz owoce ich pracy. Jeśli wyszukasz termin taki jak „kot”, zobaczysz tekst, filmy, zdjęcia, a nawet pliki audio. Praca z danymi nieustrukturyzowanymi to jeden z najciekawszych obszarów nauki o danych. Nowe bazy danych, takie jak NoSQL, umożliwiają przechwytywanie i przechowywanie dużych plików. Dużo łatwiej jest przechowywać to wszystko w jednym miejscu. Wszystkie pliki audio, wideo, obrazy lub pliki tekstowe mogą trafić do klastra NoSQL. Jeśli chcesz wszystko uchwycić, są też do tego nowe narzędzia. Możesz użyć technologii Big Data, takiej jak Hadoop, do przetwarzania danych w partiach lub w czasie rzeczywistym. Wróćmy więc do Twojej witryny internetowej z butami do biegania. Firma trochę się rozwinęła, a teraz jesteś częścią nowego zespołu ds. analityki danych. Współpracujesz z marketingiem i kierownictwem, aby zadać pierwsze interesujące pytanie: kto jest najlepszym klientem butów do biegania? Zbierasz podstawowe informacje biograficzne, które dość łatwo było znaleźć w Twojej bazie klientów. Masz ich adres e-mail oraz miasto i województwo, w którym mieszkają. Bierzesz te informacje i zaczynasz przeszukiwać posty w sieciach społecznościowych klienta. Zaczynasz zbierać wszystkie nieustrukturyzowane dane. Może Twój klient opublikował film z ukończenia maratonu. Możesz wysłać gratulacyjny tweet. Możesz również zdecydować się na przeszukiwanie postów znajomych klienta. Może Twój klient opublikuje zdjęcie, na którym biegają z grupą ludzi. Możesz użyć nieustrukturyzowanych danych do identyfikacji tych osób i wysyłania im specjalnych promocji. Ten typ projektu danych jest zwykle nazywany widokiem 360°

klienta. Próbujesz dowiedzieć się wszystkiego, co możesz o tym, co ich motywuje. Następnie możesz wykorzystać te informacje, aby znaleźć najlepszych klientów i wysyłać promocje. Może się również okazać, że masz kilku klientów, którzy polecają wielu swoich znajomych. Możesz zaoferować im specjalne zachęty i rabaty. W miarę upływu czasu możesz przechwycić coraz więcej nieustrukturyzowanych danych swoich klientów, co pozwoli ci zadawać bardziej wyrafinowane pytania dotyczące klientów. Na przykład: czy częściej podróżują? Czy są bardziej konkurencyjne? Jak często chodzą do restauracji? Każde z tych pytań może pomóc Ci nawiązać kontakt z klientem i sprzedawać więcej produktów. Gdy zbierasz te dane, możesz chcieć wyświetlić je na wykresie. Dane nieustrukturyzowane to zasób, który rośnie z każdym dniem. Pomyśl o rzeczach, które zrobiłeś dzisiaj, a które mogą być interesujące dla firmy. Czy wysłałeś tweeta o swoim długim spacerze do pracy? Może potrzebujesz lepszych butów. Narzekałeś na deszczowy dzień? Powinieneś kupić parasol. Dane nieustrukturyzowane pozwalają firmom oferować taki poziom interakcji.

Przesiewanie wielkich śmieci

Nieuporządkowane dane niosą ze sobą nowe wyzwania. Jednym z pierwszych pytań, na które natrafisz, jest to, czy chcesz usunąć dane. Pamiętaj, że zespół data science używa metody naukowej na swoich danych. Chcesz móc zadawać ciekawe pytania. Musisz zdecydować, czy istnieje jakiś limit pytań, które chcesz zadać. Istnieją dobre argumenty za przechowywaniem i wyrzucaniem danych. Niektórzy analitycy danych twierdzą, że nigdy nie poznasz każdego pytania, które możesz zadać, więc po co wyrzucać dane? Stosunkowo tanie jest również przechowywanie ogromnych ilości danych – często tylko kilka centów za gigabajt. Równie dobrze możesz to wszystko zatrzymać, zamiast decydować, co wyrzucić. Czasami taniej jest kupić nowe dyski twarde niż spędzać czas na spotkaniach dotyczących przechowywania danych. Inni analitycy twierdzą, że powinieneś wyrzucić swoje dane. W klastrach Big Data może być dużo śmieci. Im więcej masz śmieci, tym więcej trudno jest znaleźć interesujące wyniki, ponieważ w twoich informacjach jest zbyt dużo szumu (bezsensownych danych). Podjęcie decyzji o tym, czy zachować, czy usunąć dane, to problem, nad którym wciąż pracuje wiele zespołów zajmujących się analizą danych. Pracowałem kiedyś w firmie, która mierzyła się z tym wyzwaniem. Byli właścicielami strony internetowej, która łączyła potencjalnych nabywców samochodów z dealerami samochodowymi. Stworzyli system tagowania, który rejestrował wszystko, co klienci oglądali na ich stronie internetowej. Za każdym razem, gdy klient przewinął obraz, baza danych doda nowy rekord; wszystkie kliknięte linki zostały zebrane. System rozrósł się do tysięcy tagów. Każdy z tych tagów zawierał miliony transakcji. Tylko kilka osób w firmie rozumiało, co uchwycił każdy tag, co bardzo utrudniało im tworzenie interesujących raportów. Używali tego samego systemu tagowania z nieustrukturyzowanymi danymi. Zaczęli zbieranie informacji o reklamach i filmach Flash. Chcieli połączyć tag z obrazem i transakcją, co pozwoliło im zobaczyć obraz, który kliknął klient, a także tag, który wskazywał, gdzie znajduje się na stronie. Wszystkie te informacje były przechowywane w rosnącym klastrze Hadoop. Niektórzy członkowie zespołu twierdzili, że wiele danych jest przestarzałych. Tylko kilka osób znało system tagowania i reklamy stale zmienione. Ponadto osoby znające system tagowania zaczęły zmieniać nazwy tagów. Tak wiele danych było przestarzałych. Inni członkowie zespołu argumentowali, że jest to bardzo mała ilość danych w porównaniu z tym, co można przechowywać w klastrze Hadoop. Kogo obchodzi, że masz kilka gigabajtów przestarzałych danych? Posprzątanie nie było warte wysiłku. Są szanse, że poradzisz sobie również z tego typu wyzwaniami. Kiedy to robisz, pamiętaj o tych rzeczach:

• Naprawdę nie ma właściwej odpowiedzi. Twój zespół ds. analityki danych musi tylko dowiedzieć się, co jest dla niego najlepsze.

• Jeśli zdecydujesz się zachować wszystko, prawdopodobnie będziesz musiał trochę popracować nad tworzeniem interesujących raportów. Będziesz musiał zrobić trochę więcej filtrowania i będzie trochę więcej szumu w twoich danych.

• Jeśli zdecydujesz się wyrzucić dane, uzyskasz czystszy klaster. Istnieje jednak szansa, że nieumyślnie wyrzucisz coś, czego pewnego dnia możesz żałować. To tak, jakbyś sprzątał swoją szafę. Nigdy nie wiadomo, czy ta zamszowa kurtka z kołnierzem wróci w wielkim stylu. Ale jeśli trzymasz za dużo kurtek, możesz zapomnieć, co masz.

Najważniejszą rzeczą jest upewnienie się, że Twój zespół podejmie decyzję. Nie chcesz mieć zasad dotyczących danych, które zmieniają się co kilka miesięcy. Albo zdecyduj na początku, że planujesz zachować wszystko, albo chcesz wyrzucić niektóre dane. Współpracuj z zespołem, aby upewnić się, że wszyscy zgadzają się z polityką i co można wyrzucić. Jeśli nie masz ustalonej polityki, możesz uszkodzić wszystkie dane. Jeśli nie wiesz, co wyrzuciłeś, a co zatrzymałeś, trudno jest zrozumieć sens raportów. Spróbuj się zdecydować na wczesnym etapie, co najlepiej sprawdza się w Twoim zespole.

PODSUMOWANIE

W tym rozdziale dowiedziałeś się, że dane strukturalne to dane, które mają określony format w określonej kolejności. Widziałeś również, że dane częściowo ustrukturyzowane to dane o pewnej strukturze, ale istnieje dodatkowa elastyczność zmiany nazw pól. Wreszcie są dane nieustrukturyzowane, czyli wszystko inne. To dane, które nie są zgodne ze schematem i nie mają modelu danych. Dowiedziałeś się również o dużych śmieciach i poznałeś kilka wskazówek, jak je przeszukiwać. W rozdziale 4 dowiesz się, jak zastosować analizę statystyczną do swoich danych.