Zrozumieć podstawy Big Data

Zarządzanie danymi i ich analizowanie zawsze oferowało największe korzyści i największe wyzwania dla organizacji różnej wielkości i we wszystkich branżach. Firmy od dawna zmagają się ze znalezieniem pragmatycznego podejścia do gromadzenia informacji o swoich klientach, produktach i usługach. Gdy firma miała tylko garstkę klientów, którzy kupowali ten sam produkt w ten sam sposób, sprawy były dość proste i proste. Jednak z biegiem czasu firmy i rynki, w których uczestniczą, stały się bardziej skomplikowane. Aby przetrwać lub zyskać przewagę konkurencyjną wśród klientów, firmy te dodały więcej linii produktów i zróżnicowały sposób dostarczania swoich produktów. Walka o dane nie ogranicza się do biznesu. Na przykład organizacje zajmujące się badaniami i rozwojem (B + R) walczyły o uzyskanie wystarczającej mocy obliczeniowej do obsługi zaawansowanych modeli lub przetwarzania obrazów i innych źródeł danych naukowych. Rzeczywiście, mamy do czynienia z dużą złożonością, jeśli chodzi o dane. Niektóre dane są ustrukturyzowane i przechowywane w tradycyjnej relacyjnej bazie danych, a inne dane, w tym dokumenty, zapisy dotyczące obsługi klienta, a nawet zdjęcia i filmy, są nieustrukturyzowane. Firmy muszą również wziąć pod uwagę nowe źródła danych generowanych przez maszyny, takie jak czujniki. Inne nowe źródła informacji są generowane przez człowieka, takie jak dane z mediów społecznościowych i dane o strumieniach kliknięć generowane w wyniku interakcji z witryną. Ponadto dostępność i przyjęcie nowszych, bardziej wydajnych urządzeń mobilnych w połączeniu z wszechobecnym dostępem do globalnych sieci będzie napędzać tworzenie nowych źródeł danych. Chociaż każde źródło danych może być niezależnie zarządzane i przeszukiwane, dziś wyzwaniem jest sposób, w jaki firmy mogą zrozumieć skrzyżowanie wszystkich tych różnych typów danych. Kiedy masz do czynienia z tak wieloma informacjami w tak wielu różnych formach, nie sposób myśleć o zarządzaniu danymi w tradycyjny sposób. Chociaż zawsze dysponowaliśmy dużą ilością danych, obecnie różnica polega na tym, że istnieje ich znacznie więcej i różnią się rodzajem i aktualnością. Organizacje również znajdują więcej sposobów wykorzystania tych informacji niż kiedykolwiek wcześniej. Dlatego musisz inaczej myśleć o zarządzaniu danymi. To jest szansa i wyzwanie związane z dużymi zbiorami danych.

Ewolucja zarządzania danymi

Byłoby miło pomyśleć, że każda nowa innowacja w zarządzaniu danymi jest nowym początkiem i jest oderwana od przeszłości. Jednak niezależnie od tego, czy jest to rewolucyjne, czy przyrostowe, większość nowych etapów lub fal zarządzania danymi opiera się na swoich poprzednikach. Chociaż zarządzanie danymi jest zwykle postrzegane przez pryzmat oprogramowania, w rzeczywistości należy je postrzegać z holistycznej perspektywy. Zarządzanie danymi musi obejmować postęp technologiczny w zakresie sprzętu, pamięci masowej, sieci i modeli obliczeniowych, takich jak wirtualizacja i przetwarzanie w chmurze. Konwergencja nowych technologii i redukcja kosztów wszystkiego, od pamięci masowej po cykle obliczeniowe, zmieniły krajobraz danych i stworzyły nowe możliwości. Ponieważ wszystkie te czynniki technologiczne są zbieżne, zmienia to sposób zarządzania danymi i ich wykorzystywania. Big data to najnowszy trend, który pojawił się z powodu tych czynników. Czym więc są duże zbiory danych i dlaczego są tak ważne? W dalszej części książki podajemy bardziej wyczerpującą definicję. Na początek duże zbiory danych definiuje się jako dowolne źródło danych, które ma co najmniej trzy wspólne cechy to:

* Niezwykle duże ilości danych

* Ekstremalnie duża prędkość danych

* Niezwykle szeroki Różnorodność danych

Big Data jest ważne, ponieważ umożliwia organizacjom gromadzenie, przechowywanie, zarządzanie i przetwarzanie ogromnych ilości danych z odpowiednią prędkością i we właściwym czasie, aby uzyskać właściwy wgląd. Zanim jednak zagłębimy się w szczegóły dotyczące dużych zbiorów danych, ważne jest, aby przyjrzeć się ewolucji zarządzania danymi i temu, jak doprowadziło to do powstania dużych zbiorów danych. Big data nie jest samodzielną technologią; jest to raczej połączenie ostatnich 50 lat rozwoju technologii. Dzisiejsze organizacje znajdują się w punkcie zwrotnym w zarządzaniu danymi. Przeszliśmy z epoki, w której technologia została zaprojektowana, aby wspierać określone potrzeby biznesowe, takie jak określanie, ile produktów zostało sprzedanych ilu klientów, do czasu, w którym organizacje mają więcej danych z większej liczby źródeł niż kiedykolwiek wcześniej. Wszystkie te dane wyglądają na potencjalną żyłę złota, ale podobnie jak w kopalni złota masz tylko trochę złota i dużo więcej wszystkiego innego. Wyzwania technologiczne to: „Jak nadać sens tym danym, skoro nie potrafisz łatwo rozpoznać wzorców, które mają największe znaczenie dla decyzji biznesowych? W jaki sposób Twoja organizacja radzi sobie z ogromnymi ilościami danych w znaczący sposób? ” Zanim przejdziemy do opcji, przyjrzymy się ewolucji zarządzania danymi i zobaczymy, jak te fale są połączone.

Zrozumienie fal zarządzania danymi

Każda fala zarządzania danymi rodzi się z konieczności podjęcia próby rozwiązania określonego typu problemu związanego z zarządzaniem danymi. Każda z tych fal lub faz wyewoluowała z powodu przyczyny i skutku. Kiedy nowe rozwiązanie technologiczne pojawiło się na rynku, wymagało odkrycia nowych podejść. Kiedy relacyjna baza danych pojawiła się na rynku, potrzebowała zestawu narzędzi, aby umożliwić menedżerom badanie relacji między elementami danych. Kiedy firmy zaczęły przechowywać nieustrukturyzowane dane, analitycy potrzebowali nowych możliwości, takich jak narzędzia analityczne oparte na języku naturalnym, aby uzyskać wgląd, który byłby przydatny dla biznesu. Jeśli byłeś liderem firmy zajmującej się wyszukiwarkami, zacząłeś zdawać sobie sprawę, że masz dostęp do ogromnych ilości danych, na których można było zarabiać. Aby uzyskać wartość z tych danych, potrzebne były nowe, innowacyjne narzędzia i podejścia. Fale zarządzania danymi w ciągu ostatnich pięciu dekad osiągnęły punkt kulminacyjny w miejscu, w którym jesteśmy dzisiaj: zapoczątkowaniu ery dużych zbiorów danych. Aby więc zrozumieć duże zbiory danych, musisz zrozumieć podstawy tych poprzednich fal. Musisz także zrozumieć, że przechodząc od jednej fali do drugiej, nie wyrzucamy narzędzi, technologii i praktyk, których używaliśmy do rozwiązywania innych problemów.

Fala 1: Tworzenie zarządzalnych struktur danych

Ponieważ komputeryzacja weszła na rynek komercyjny pod koniec lat 60., dane były przechowywane w płaskich plikach, które nie narzucały żadnej struktury. Gdy firmy musiały osiągnąć poziom szczegółowego zrozumienia klientów, musiały stosować metody brutalnej siły, w tym bardzo szczegółowe modele programowania, aby stworzyć jakąś wartość. Później w latach siedemdziesiątych sytuacja uległa zmianie wraz z wynalezieniem relacyjnego modelu danych i systemu zarządzania relacyjnymi bazami danych (RDBMS), które narzuciły strukturę i metodę poprawy wydajności. Co najważniejsze, model relacyjny dodał poziom abstrakcji (ustrukturyzowany język zapytań [SQL], generatory raportów i narzędzia do zarządzania danymi), tak aby programiści mogli łatwiej zaspokoić rosnące wymagania biznesowe w zakresie wydobywania wartości z danych. Model relacyjny oferował ekosystem narzędzi pochodzących z wielu powstających firm programistycznych. Spełniło to rosnącą potrzebę pomocy firmom w lepszym organizowaniu ich danych i możliwości porównywania transakcji z jednego regionu do drugiego. Ponadto pomogło menedżerom biznesowym, którzy chcieli mieć możliwość przeanalizowania informacji, takich jak zapasy i porównania ich z informacjami o zamówieniach klientów w celu podjęcia decyzji. Ale pojawił się problem z gwałtownym zapotrzebowaniem na odpowiedzi: przechowywanie tej rosnącej ilości danych było kosztowne, a

dostęp do nich był powolny. Co gorsza, istniało wiele duplikatów danych, a rzeczywista wartość biznesowa tych danych była trudna do zmierzenia. Na tym etapie istniała pilna potrzeba znalezienia nowego zestawu technologii wspierających model relacyjny. Pojawił się model Entity-Relationship (ER), który dodał dodatkową abstrakcję w celu zwiększenia użyteczności danych. W tym modelu każdy element został zdefiniowany niezależnie od jego zastosowania. Dlatego programiści mogli tworzyć nowe relacje między źródłami danych bez skomplikowanego programowania. W tamtym czasie był to ogromny postęp, który umożliwił programistom przesunięcie granic technologii i tworzenie bardziej złożonych modeli wymagających złożonych technik łączenia jednostek. Rynek relacyjnych baz danych eksplodował i nadal żyje. Jest to szczególnie ważne w przypadku transakcyjnego zarządzania danymi o dużej strukturze. Kiedy ilość danych, którymi organizacje musiały zarządzać, wymknęła się spod kontroli, hurtownia danych zapewniła rozwiązanie. Hurtownia danych umożliwiła organizacji IT wybranie podzbioru przechowywanych danych, aby firma mogła łatwiej uzyskać wgląd. Hurtownia danych miała pomóc firmom radzić sobie z coraz większymi ilościami ustrukturyzowanych danych, które musiały być w stanie przeanalizować, zmniejszając ilość danych do czegoś mniejszego i bardziej skoncentrowanego na określonym obszarze działalności. Wypełniło to potrzebę rozdzielenia wsparcia przetwarzania decyzji operacyjnych i wspomagania decyzji - ze względów wydajnościowych. Ponadto hurtownie często przechowują dane z poprzednich lat w celu zrozumienia wydajności organizacji, identyfikowania trendów i pomocy w ujawnianiu wzorców zachowań. Zapewnił również zintegrowane źródło informacji z różnych źródeł danych, które można wykorzystać do analizy. Hurtownie danych zostały skomercjalizowane w latach 90. XX wieku, a dziś zarówno systemy zarządzania treścią, jak i hurtownie danych są w stanie skorzystać z ulepszeń w zakresie skalowalności sprzętu, technologii wirtualizacji oraz możliwości tworzenia zintegrowanych systemów sprzętowych i programowych, zwanych także urządzeniami.

Czasami same hurtownie danych były zbyt złożone i duże oraz nie zapewniały szybkości i elastyczności wymaganej przez firmę. Odpowiedzią było dalsze udoskonalenie danych zarządzanych za pośrednictwem baz danych. Te zbiorniki danych były skoncentrowane na konkretnych problemach biznesowych i były znacznie bardziej usprawnione i wspierały biznesową potrzebę szybkich zapytań niż bardziej masowe hurtownie danych. Jak każda fala zarządzania danymi, hurtownia ewoluowała, aby wspierać nowe technologie, takie jak systemy zintegrowane i urządzenia danych. Hurtownie danych i składnice danych rozwiązały wiele problemów firm potrzebujących spójnego sposobu zarządzania ogromnymi danymi transakcyjnymi. Ale jeśli chodzi o zarządzanie ogromnymi ilościami nieustrukturyzowanych lub częściowo ustrukturyzowanych danych, hurtownia nie była w stanie ewoluować na tyle, aby sprostać zmieniającym się wymaganiom. Komplikując sprawy, hurtownie danych są zwykle zasilane w okresach wsadowych co tydzień lub codziennie. Jest to dobre w przypadku planowania, sprawozdawczości finansowej i tradycyjnych kampanii marketingowych, ale jest zbyt wolne dla środowisk biznesowych i konsumenckich w czasie rzeczywistym. W jaki sposób firmy byłyby w stanie przekształcić swoje tradycyjne podejście do zarządzania danymi, aby poradzić sobie z rosnącą ilością nieustrukturyzowanych elementów danych? Rozwiązanie nie pojawiło się z dnia na dzień. Gdy firmy zaczęły przechowywać nieustrukturyzowane dane, dostawcy zaczęli dodawać funkcje, takie jak BLOB (duże obiekty binarne). Zasadniczo niestrukturalny element danych byłby przechowywany w relacyjnej bazie danych jako jedna ciągła porcja danych. Ten obiekt można oznaczyć (to znaczy zapytać klienta), ale nie można było zobaczyć, co było w środku. Najwyraźniej nie rozwiąże to zmieniających się potrzeb klientów ani biznesowych. Wejdź do systemu zarządzania bazą danych obiektów (ODBMS). Baza danych obiektów przechowała BLOBa jako adresowalny zestaw elementów, abyśmy mogli zobaczyć, co tam jest. W przeciwieństwie do BLOBa, który był niezależną jednostką dołączoną do tradycyjnej relacyjnej bazy danych, obiektowa baza danych zapewniała ujednolicone podejście do obsługi danych nieustrukturyzowanych. Bazy danych obiektów zawierają język programowania i

strukturę elementów danych, dzięki czemu łatwiej jest manipulować różnymi obiektami danych bez programowania i skomplikowanych połączeń. Obiektowe bazy danych wprowadziły nowy poziom innowacji, który pomógł doprowadzić do drugiej fali zarządzania danymi.

Fala 2: Zarządzanie siecią i treścią

Nie jest tajemnicą, że większość danych dostępnych obecnie na świecie jest nieustrukturyzowana. Paradoksalnie, firmy skupiły swoje inwestycje w systemach opartych na ustrukturyzowanych danych, które były najbardziej związane z przychodami: branżowe systemy transakcyjne. Systemy zarządzania treścią przedsiębiorstwa ewoluowały w latach 80. XX wieku, aby zapewnić firmom możliwość lepszego zarządzania danymi nieustrukturyzowanymi, głównie dokumentami. W latach dziewięćdziesiątych, wraz z rozwojem internetu, organizacje chciały wyjść poza dokumenty i przechowywać treści internetowe, obrazy, audio i wideo oraz zarządzać nimi. Rynek ewoluował od zestawu odłączonych rozwiązań do bardziej ujednoliconego modelu, który połączył te elementy w platformę obejmującą zarządzanie procesami biznesowymi, kontrolę wersji, rozpoznawanie informacji, zarządzanie tekstem i współpracę. Ta nowa generacja systemów dodała metadane (informacje o organizacji i cechach przechowywanych informacji). Te rozwiązania pozostają niezwykle ważne dla firm, które muszą logicznie zarządzać wszystkimi tymi danymi. Ale jednocześnie zaczęła się pojawiać nowa generacja wymagań, które prowadzą nas do następnej fali. Te nowe wymagania wynikały w dużej mierze z konwergencji czynników, w tym internetu, wirtualizacji i przetwarzania w chmurze. W tej nowej fali organizacje zaczynają rozumieć, że muszą zarządzać nową generacją źródeł danych z bezprecedensową ilością i różnorodnością danych, które muszą być przetwarzane z niespotykaną szybkością.

Fala 3: Zarządzanie dużymi zbiorami danych

Czy duże zbiory danych są naprawdę nowe, czy jest to ewolucja w zarządzaniu danymi? Odpowiedź brzmi: tak - w rzeczywistości jest to jedno i drugie. Podobnie jak w przypadku innych fal w zarządzaniu danymi, duże zbiory danych są oparte na ewolucji praktyk zarządzania danymi w ciągu ostatnich pięciu dekad. Nowością jest to, że po raz pierwszy koszt cykli obliczeniowych i pamięci masowej osiągnął punkt krytyczny. Dlaczego to jest ważne? Zaledwie kilka lat temu organizacje zazwyczaj godziły się na przechowanie migawek lub podzbiorów ważnych informacji, ponieważ koszt przechowywania i ograniczenia przetwarzania uniemożliwiały im przechowywanie wszystkiego, co chcieli przeanalizować. W wielu sytuacjach ten kompromis działał dobrze. Na przykład firma produkcyjna mogła zbierać dane maszynowe co dwie minuty, aby określić kondycję systemów. Mogą jednak zaistnieć sytuacje, w których migawka nie będzie zawierała informacji o nowym typie usterki i może pozostać niezauważona przez wiele miesięcy. Dzięki Big Data możliwa jest teraz wirtualizacja danych, aby można było je wydajnie przechowywać, a przy wykorzystaniu pamięci masowej w chmurze również bardziej efektywnie kosztowo. Ponadto poprawa szybkości i niezawodności sieci usunęła inne fizyczne ograniczenia związane z możliwością zarządzania ogromnymi ilościami danych w akceptowalnym tempie. Dodajmy do tego wpływ zmian ceny i wyrafinowania pamięci komputera. Przy tych wszystkich zmianach technologicznych można teraz wyobrazić sobie, w jaki sposób firmy mogą wykorzystać dane, które byłyby nie do pomyślenia jeszcze pięć lat temu. Ale żadne przejście technologiczne nie odbywa się w izolacji; dzieje się tak, gdy istnieje ważna potrzeba, którą można zaspokoić dzięki dostępności i dojrzewaniu technologii. Wiele technologii stanowiących podstawę dużych zbiorów danych, takich jak wirtualizacja, przetwarzanie równoległe, rozproszone systemy plików i bazy danych w pamięci, istnieje od dziesięcioleci. Zaawansowane analizy istnieją również od dziesięcioleci, chociaż nie zawsze były praktyczne. Inne technologie, takie jak Hadoop i MapReduce, istnieją na rynku zaledwie od kilku lat. To połączenie postępów technologicznych może teraz rozwiązać istotne problemy biznesowe. Firmy chcą mieć możliwość uzyskiwania wglądu i przydatnych wyników z wielu różnych rodzajów danych z odpowiednią szybkością - niezależnie od ilości danych. Jeśli firmy mogą analizować petabajty danych

(co odpowiada 20 milionom czteroszufladowych szafek na dokumenty wypełnionych plikami tekstowymi lub 13,3 lat zawartości HDTV) z akceptowalną wydajnością, aby dostrzec wzorce i anomalie, firmy mogą zacząć nadawać sens danym w nowy sposób. Przejście na duże zbiory danych to nie tylko biznes. Nauka, badania i działania rządowe również przyczyniły się do tego. Wystarczy pomyśleć o analizie ludzkiego genomu lub zajęciu się wszystkimi danymi astronomicznymi zebranymi w obserwatoriach, aby lepiej zrozumieć otaczający nas świat. Weź pod uwagę ilość danych, które rząd zbiera również w swoich działaniach antyterrorystycznych, a zrozumiesz, że duże zbiory danych to nie tylko biznes. Istnieją różne podejścia do obsługi danych w zależności od tego, czy są to dane w ruchu, czy w spoczynku. Oto krótki przykład każdego z nich. Dane w ruchu byłyby wykorzystywane, gdyby firma była w stanie przeanalizować jakość swoich produktów podczas procesu produkcyjnego, aby uniknąć kosztownych błędów. Dane w stanie spoczynku byłyby wykorzystywane przez analityka biznesowego do lepszego zrozumienia aktualnych wzorców zakupów klientów w oparciu o wszystkie aspekty relacji z klientami, w tym sprzedaż, dane z mediów społecznościowych i interakcje z obsługą klienta. Pamiętaj, że wciąż jesteśmy na wczesnym etapie wykorzystywania ogromnych ilości danych, aby uzyskać 360-stopniowy wgląd w biznes i przewidywać zmiany i zmiany oczekiwań klientów. Technologie wymagane do uzyskania odpowiedzi, których potrzebuje biznes, są nadal odizolowane od siebie. Aby osiągnąć pożądany stan końcowy, technologie ze wszystkich trzech fal będą musiały się połączyć. Jak zobaczysz czytając tę książkę, duże zbiory danych to nie tylko jedno narzędzie czy jedna technologia. Chodzi o to, w jaki sposób wszystkie te technologie łączą się, aby zapewnić właściwy wgląd we właściwym czasie na podstawie odpowiednich danych - niezależnie od tego, czy są one generowane przez ludzi, maszyny czy sieć.

Definicja Big Data

Big data to nie pojedyncza technologia, ale połączenie starych i nowych technologii, które pomagają firmom uzyskać praktyczny wgląd. Dlatego duże zbiory danych to możliwość zarządzania ogromną ilością różnych danych, z odpowiednią prędkością i we właściwych ramach czasowych, aby umożliwić analizę i reakcję w czasie rzeczywistym. Jak zauważyliśmy wcześniej w tym rozdziale, duże zbiory danych są zazwyczaj podzielone według trzech cech:

* Wolumen : ilość danych

* Prędkość : jak szybko te dane są przetwarzane

* Różnorodność : różne typy danych

Chociaż wygodnie jest uprościć duże zbiory danych do tego, może to być mylące i zbyt uproszczone. Na przykład możesz zarządzać stosunkowo niewielką ilością bardzo różnych, złożonych danych lub przetwarzać ogromną ilość bardzo prostych danych. Te proste dane mogą być w całości ustrukturyzowane lub niestrukturalne. Jeszcze ważniejszy jest czwarty element: prawdomówność. Jak dokładne są te dane w przewidywaniu wartości biznesowej? Czy wyniki analizy dużych zbiorów danych rzeczywiście mają sens? Bardzo ważne jest, abyś nie lekceważył wykonywanego zadania. Dane muszą być możliwe do zweryfikowania zarówno na podstawie dokładności, jak i kontekstu. Innowacyjna firma może chcieć mieć możliwość analizowania ogromnych ilości danych w czasie rzeczywistym, aby szybko ocenić wartość tego klienta i potencjał dostarczenia mu dodatkowych ofert. Konieczne jest określenie odpowiedniej ilości i rodzajów danych, które można przeanalizować, aby wpłynąć na wyniki biznesowe. Big data obejmuje wszystkie dane, w tym dane ustrukturyzowane i nieustrukturyzowane z poczty e-mail, mediów społecznościowych, strumieni tekstowych i nie tylko. Ten rodzaj zarządzania danymi wymaga, aby firmy wykorzystywały zarówno dane ustrukturyzowane, jak i nieustrukturyzowane.

Budowanie udanej architektury zarządzania Big Dat

Przeszliśmy z epoki, w której organizacja mogła wdrożyć bazę danych, aby spełnić określone potrzeby projektowe i zostać zakończone. Ponieważ jednak dane stały się paliwem wzrostu i innowacji, ważniejsze niż kiedykolwiek jest posiadanie podstawowej architektury obsługującej rosnące wymagania.

Zaczynając od przechwytywania, organizowania, integracji, analizy i działania

Zanim zagłębimy się w architekturę, ważne jest, aby wziąć pod uwagę wymagania funkcjonalne dotyczące dużych zbiorów danych. Po pomyślnym wdrożeniu tej fazy dane mogą być analizowane w oparciu o rozwiązywany problem. Wreszcie, kierownictwo podejmuje działania na podstawie wyników tej analizy. Na przykład Amazon.com może polecić książkę na podstawie wcześniejszego zakupu lub klient może otrzymać kupon rabatowy na przyszły zakup produktu pokrewnego do właśnie zakupionego. Chociaż brzmi to prosto, pewne niuanse tych funkcji są skomplikowane. Walidacja to szczególnie ważna kwestia. Jeśli Twoja organizacja łączy źródła danych, bardzo ważne jest, aby mieć możliwość sprawdzenia, czy te źródła mają sens w połączeniu. Ponadto niektóre źródła danych mogą zawierać poufne informacje, dlatego należy wdrożyć wystarczające poziomy zabezpieczeń i nadzoru. Oczywiście każde zagłębienie się w duże zbiory danych musi najpierw zacząć się od problemu, który próbujesz rozwiązać. To będzie dyktować rodzaj danych, których potrzebujesz i jak architektura może wyglądać.

Ustawienie fundamentów architektonicznych

Oprócz obsługi wymagań funkcjonalnych ważne jest, aby wspierać wymaganą wydajność. Twoje potrzeby będą zależeć od charakteru analizy, którą wspierasz. Będziesz potrzebował odpowiedniej ilości mocy obliczeniowej i szybkości. Chociaż część analiz, które wykonasz, będzie wykonywana w czasie rzeczywistym, nieuchronnie będziesz również przechowywać pewną ilość danych. Twoja architektura musi mieć również odpowiednią ilość nadmiarowości, aby być chronionym przed nieoczekiwanymi opóźnieniami i przestojami. Twoja organizacja i jej potrzeby określą, ile uwagi musisz poświęcić tym problemom z wydajnością. Zacznij więc od zadania sobie następujących pytań:

* Jak dużą ilością danych będzie potrzebna moja organizacja do zarządzania dzisiaj i w przyszłości?

* Jak często moja organizacja będzie musiała zarządzać danymi w czasie rzeczywistym lub prawie w czasie rzeczywistym?

* Na jakie ryzyko może sobie pozwolić moja organizacja? Czy moja branża podlega surowym wymaganiom w zakresie bezpieczeństwa, zgodności i nadzoru?

* Jak ważna jest szybkość dla mojej potrzeby zarządzania danymi?

* Jak pewne lub dokładne muszą być dane?

Aby zrozumieć duże zbiory danych, pomocne jest rozplanowanie elementów architektury. Architektura zarządzania dużymi zbiorami danych musi obejmować różnorodne usługi, które umożliwiają firmom korzystanie z niezliczonych źródeł danych w szybki i efektywny sposób.

Interfejsy i kanały

Zanim przejdziemy do sedna samego stosu technologii big data, chcielibyśmy, abyście zrozumieli, jak duże zbiory danych działają w prawdziwym świecie, należy zacząć od zrozumienia tej konieczności. W rzeczywistości to, co sprawia, że duże zbiory danych są duże, to fakt, że opiera się na zbieraniu dużej ilości danych z wielu źródeł. Dlatego będą otwarte interfejsy programowania aplikacji (API) ,rdzeń do

dowolnej architektury Big Data. Ponadto należy pamiętać, że interfejsy istnieją na każdym poziomie i między każdą warstwą stosu. Bez usług integracji duże zbiory danych nie mogą się wydarzyć.

Nadmiarowa infrastruktura fizyczna

Wspierająca infrastruktura fizyczna ma fundamentalne znaczenie dla działania i skalowalności architektury Big Data. W rzeczywistości bez dostępności solidnej infrastruktury fizycznej duże zbiory danych prawdopodobnie nie pojawiłyby się jako tak ważny trend. Aby obsłużyć nieoczekiwaną lub nieprzewidywalną ilość danych, fizyczna infrastruktura dla dużych zbiorów danych musi być inna niż dla tradycyjnych danych. Infrastruktura fizyczna oparta jest na modelu przetwarzania rozproszonego. Oznacza to, że dane mogą być fizycznie przechowywane w wielu różnych lokalizacjach i mogą być łączone ze sobą poprzez sieci, wykorzystanie rozproszonego systemu plików oraz różnych narzędzi i aplikacji do analizy dużych zbiorów danych. Nadmiarowość jest ważna, ponieważ mamy do czynienia z dużą ilością danych z wielu różnych źródeł. Nadmiarowość przybiera różne formy. Jeśli Twoja firma utworzyła chmurę prywatną, będziesz chciał mieć wbudowaną redundancję w środowisku prywatnym, aby można ją było skalować w celu obsługi zmieniających się obciążeń. Jeśli Twoja firma chce ograniczyć wewnętrzny rozwój IT, może skorzystać z zewnętrznych usług w chmurze, aby zwiększyć swoje wewnętrzne zasoby. W niektórych przypadkach to nadmiarowość może przybrać formę oferty oprogramowania jako usługi (SaaS), która umożliwia firmom przeprowadzanie zaawansowanej analizy danych jako usługi. Podejście SaaS oferuje niższe koszty, szybsze uruchamianie i płynną ewolucję podstawowej technologii.

Infrastruktura bezpieczeństwa

Im ważniejsza będzie analiza dużych zbiorów danych dla firm, tym ważniejsze będzie zabezpieczenie tych danych. Na przykład, jeśli jesteś firmą z branży opieki zdrowotnej, prawdopodobnie będziesz chciał użyć aplikacji Big Data do określenia zmian demograficznych lub zmian w potrzebach pacjentów. Te dane o Twoich wyborcach muszą być chronione zarówno w celu spełnienia wymogów zgodności, jak i ochrony prywatności pacjentów. Będziesz musiał wziąć pod uwagę, kto ma prawo wglądu do danych iw jakich okolicznościach ma to prawo. Będziesz musiał mieć możliwość weryfikacji tożsamości użytkowników, a także ochrony tożsamości pacjentów. Tego typu wymagania dotyczące bezpieczeństwa muszą być częścią struktury dużych zbiorów danych od samego początku, a nie po namyśle.

Operacyjne źródła danych

Kiedy myślisz o dużych zbiorach danych, ważne jest, aby zrozumieć, że musisz uwzględnić wszystkie źródła danych, które dadzą Ci pełny obraz Twojej firmy i zobacz, jak dane wpływają na sposób prowadzenia firmy. Tradycyjnie operacyjne źródło danych składało się z wysoce ustrukturyzowanych danych zarządzanych przez linię biznesową w relacyjnej bazie danych. Jednak w miarę jak zmienia się świat, ważne jest, aby zrozumieć, że dane operacyjne muszą teraz obejmować szerszy zestaw źródeł danych, w tym źródła nieustrukturyzowane, takie jak dane klientów i media społecznościowe we wszystkich ich formach. Znajdziesz nowe, pojawiające się podejścia do zarządzania danymi w świecie dużych zbiorów danych, w tym architektury dokumentów, wykresów, kolumnowych i geoprzestrzennych baz danych. Łącznie są one nazywane bazami danych NoSQL lub nie tylko SQL.

Zasadniczo musisz odwzorować architektury danych na typy transakcji. Pomoże to zapewnić dostępność odpowiednich danych wtedy, gdy ich potrzebujesz. Potrzebujesz także architektur danych obsługujących złożoną, nieustrukturyzowaną zawartość. W podejściu do wykorzystywania dużych zbiorów danych należy uwzględnić zarówno relacyjne, jak i nierelacyjne bazy danych. Konieczne jest również uwzględnienie nieustrukturyzowanych źródeł danych, takich jak systemy zarządzania treścią,

aby można było zbliżyć się do tego 360-stopniowego widoku biznesowego. Wszystkie te operacyjne źródła danych mają kilka wspólnych cech:

* Reprezentują systemy ewidencji, które śledzą krytyczne dane wymagane do codziennego funkcjonowania firmy w czasie rzeczywistym. Są one stale aktualizowane w oparciu o transakcje zachodzące w jednostkach biznesowych oraz w Internecie.

* Aby te źródła zapewniały dokładną reprezentację firmy, muszą łączyć dane ustrukturyzowane i nieustrukturyzowane.

* Te systemy muszą być również w stanie skalować się, aby obsługiwać tysiące użytkowników w sposób spójny. Mogą to być transakcyjne systemy e-commerce, systemy zarządzania relacjami z klientami lub aplikacje call center.

Wydajność ma znaczenie

Twoja architektura danych musi również działać w zgodzie z infrastrukturą pomocniczą organizacji. Na przykład możesz być zainteresowany uruchomieniem modeli w celu ustalenia, czy wiercenie ropy naftowej na obszarze morskim jest bezpieczne, biorąc pod uwagę dane w czasie rzeczywistym dotyczące temperatury, zasolenia, resuspensji osadów i wielu innych biologicznych, chemicznych i fizycznych właściwości słup wody. Uruchomienie tego modelu przy użyciu tradycyjnej konfiguracji serwera może zająć kilka dni. Jednak przy zastosowaniu modelu przetwarzania rozproszonego zajęło to kilka dni. Wydajność może również określać rodzaj używanej bazy danych. Na przykład w niektórych sytuacjach możesz chcieć zrozumieć, w jaki sposób dwa bardzo różne elementy danych są powiązane. Jaki jest związek między szumem w sieci społecznościowej a wzrostem sprzedaży? Nie jest to typowe zapytanie, które można zadać w przypadku ustrukturyzowanej, relacyjnej bazy danych. Graficzna baza danych może być lepszym wyborem, ponieważ została specjalnie zaprojektowana w celu oddzielenia „węzłów" lub jednostek od ich „właściwości" lub informacji definiujących tę jednostkę oraz „krawędzi" lub relacji między węzłami i właściwościami. Korzystanie z odpowiedniej bazy danych również poprawi wydajność. Zwykle baza danych wykresów będzie używana w zastosowaniach naukowych i technicznych. Inne ważne podejścia do operacyjnej bazy danych obejmują kolumnowe bazy danych, które wydajnie przechowują informacje w kolumnach, a nie wierszach. Takie podejście prowadzi do szybszej wydajności, ponieważ wejście / wyjście jest niezwykle szybkie. Gdy przechowywanie danych geograficznych jest częścią równania, przestrzenna baza danych jest zoptymalizowana pod kątem przechowywania i wyszukiwania danych na podstawie tego, jak obiekty są powiązane w przestrzeni.

Organizowanie usług i narzędzi związanych z danymi

Nie wszystkie dane używane przez organizacje są operacyjne. Rosnąca ilość danych pochodzi z różnych źródeł, które nie są tak zorganizowane lub proste, w tym z danych pochodzących z maszyn lub czujników oraz ogromnych publicznych i prywatnych źródeł danych. W przeszłości większość firm nie była w stanie przechwycić ani przechowywać tak ogromnej ilości danych. To było po prostu zbyt drogie lub zbyt przytłaczające. Nawet jeśli firmy były w stanie przechwycić dane, nie miały narzędzi, aby cokolwiek z tym zrobić. Bardzo niewiele narzędzi mogłoby nadać sens tym ogromnym ilościom danych. Narzędzia, które istniały, były skomplikowane w użyciu i nie dawały rezultatów w rozsądnych ramach czasowych. W końcu ci, którzy naprawdę chcieli podjąć ogromny wysiłek analizy tych danych, zostali zmuszeni do pracy z migawkami danych. Ma to niepożądany efekt polegający na pomijaniu ważnych wydarzeń, ponieważ nie wystąpiły one w określonej migawce.

MapReduce, Hadoop i Big Table

Wraz z ewolucją technologii komputerowej możliwe jest teraz zarządzanie ogromnymi ilościami danych, które wcześniej mogły być obsługiwane tylko przez superkomputery, kosztem. Ceny systemów spadły, w wyniku czego nowe techniki przetwarzania rozproszonego stały się głównym nurtem. Prawdziwy przełom w Big Data nastąpił, gdy firmy takie jak Yahoo!, Google i Facebook zdały sobie sprawę, że potrzebują pomocy w zarabianiu na ogromnych ilościach danych, które tworzyły ich oferty. Te firmy musiały znaleźć nowe technologie, które pozwoliłyby im przechowywać, uzyskiwać dostęp i analizować ogromne ilości danych w czasie zbliżonym do rzeczywistego, aby mogły zarabiać na korzyściach wynikających z posiadania tak dużej ilości danych o uczestnikach ich sieci. Powstałe rozwiązania zmieniają rynek zarządzania danymi. W szczególności innowacje MapReduce, Hadoop i Big Table okazały się iskrą, która doprowadziła do nowej generacji zarządzania danymi. Technologie te rozwiązują jeden z najbardziej podstawowych problemów - zdolność do wydajnego, ekonomicznego i terminowego przetwarzania ogromnych ilości danych.

MapReduce

MapReduce zostało zaprojektowane przez Google jako sposób na wydajne wykonywanie zestawu funkcji na dużej ilości danych w trybie wsadowym. Komponent „mapa" rozdziela problem lub zadania programistyczne na wiele systemów i obsługuje rozmieszczenie zadań w sposób równoważący obciążenie i zarządzający odtwarzaniem po awarii. Po zakończeniu obliczeń rozproszonych inna funkcja o nazwie „redukuj" agreguje wszystkie elementy z powrotem w celu uzyskania wyniku. Przykładem użycia MapReduce może być określenie, ile stron książki jest napisanych w każdym z 50 różnych języków.

Big Table

Big Table został opracowany przez Google jako rozproszony system pamięci masowej przeznaczony do zarządzania wysoce skalowalnymi danymi strukturalnymi. Dane są zorganizowane w tabele z wierszami i kolumnami. W przeciwieństwie do tradycyjnego modelu relacyjnej bazy danych, Big Table to rzadka, rozproszona, trwała, wielowymiarowa, posortowana mapa. Jest przeznaczony do przechowywania ogromnych ilości danych na serwerach towarowych.

Hadoop

Hadoop to platforma programowa zarządzana przez Apache, wywodząca się z MapReduce i Big Table. Hadoop pozwala aplikacjom opartym na MapReduce działać na dużych klastrach standardowego sprzętu. Projekt jest podstawą architektury obliczeniowej wspierającej biznes Yahoo!. Platforma Hadoop została zaprojektowana w celu zrównoleglenia przetwarzania danych w węzłach obliczeniowych w celu przyspieszenia obliczeń i ukrycia opóźnień. Istnieją dwa główne komponenty Hadoop: masowo skalowalny rozproszony system plików, który może obsługiwać petabajty danych oraz masowo skalowalny silnik MapReduce, który oblicza wyniki wsadowo.

Analityka tradycyjna i zaawansowana

Co Twoja firma robi teraz ze wszystkimi danymi we wszystkich ich formach, aby nadać im sens? Wymaga wielu różnych podejść do analizy, w zależności od rozwiązywanego problemu. Niektóre analizy będą wykorzystywać tradycyjną hurtownię danych, podczas gdy inne analizy będą korzystać z zaawansowanej analizy predykcyjnej. Zarządzanie dużymi zbiorami danych w sposób holistyczny wymaga wielu różnych podejść, aby pomóc firmie w pomyślnym planowaniu przyszłości. Hurtownie danych analitycznych i zbiorniki danych Gdy firma posortuje ogromne ilości dostępnych danych, często pragmatyczne jest pobranie podzbioru danych ujawniających wzorce i umieszczenie ich w formie

dostępnej dla firmy. Te magazyny i sklepy zapewniają kompresję, wielopoziomowe partycjonowanie i masowo równoległą architekturę przetwarzania.

Analityka Big Data

Zdolność do zarządzania i analizowania petabajtów danych umożliwia firmom radzenie sobie z grupami informacji, które mogą mieć wpływ na biznes. Wymaga to mechanizmów analitycznych, które mogą zarządzać tymi wysoce rozproszonymi danymi i dostarczać wyników, które można zoptymalizować w celu rozwiązania problemu biznesowego. Analiza dużych zbiorów danych może być dość złożona. Na przykład niektóre organizacje używają modeli predykcyjnych, które łączą razem ustrukturyzowane i nieustrukturyzowane dane w celu przewidywania oszustw. Analityka mediów społecznościowych, analiza tekstu i nowe rodzaje analiz są wykorzystywane przez organizacje, które chcą uzyskać wgląd w duże zbiory danych.

Raportowanie i wizualizacja

Organizacje zawsze polegały na możliwości tworzenia raportów, aby zrozumieć, jakie dane mówią im o wszystkim, od miesięcznych danych o sprzedaży po prognozy wzrostu. Big data zmienia sposób zarządzania danymi i ich wykorzystywania. Jeśli firma może gromadzić, zarządzać i analizować wystarczającą ilość danych, może użyć nowej generacji narzędzi, aby pomóc kierownictwu naprawdę zrozumieć wpływ nie tylko zbioru elementów danych, ale także tego, jak te elementy danych oferują kontekst oparty na problemie biznesowym. zaadresowany. W przypadku dużych zbiorów danych raportowanie i wizualizacja danych stają się narzędziami do spojrzenia na kontekst powiązania danych i wpływu tych relacji na przyszłość.

Aplikacje Big Data

Tradycyjnie firma oczekiwała, że dane zostaną wykorzystane do odpowiedzi na pytania o to, co robić i kiedy. Dane były często integrowane jako pola w aplikacjach biznesowych ogólnego przeznaczenia. Wraz z pojawieniem się dużych zbiorów danych to się zmienia. Obecnie obserwujemy rozwój aplikacji zaprojektowanych specjalnie w celu wykorzystania unikalnych cech danych big data. Niektóre z pojawiających się aplikacji znajdują się w takich obszarach, jak opieka zdrowotna, zarządzanie produkcją, zarządzanie ruchem i tak dalej. Co mają wspólnego te wszystkie aplikacje Big Data? Opierają się na ogromnych ilościach, szybkościach i różnorodności danych, aby zmienić zachowanie rynku. W opiece zdrowotnej aplikacja do dużych zbiorów danych może być w stanie monitorować wcześniaki w celu określenia, kiedy dane wskazują, kiedy konieczna jest interwencja. W produkcji aplikacja do dużych zbiorów danych może być używana do zapobiegania wyłączaniu maszyny podczas przebiegu produkcyjnego. Aplikacja do zarządzania ruchem dużych zbiorów danych może zmniejszyć liczbę korków na ruchliwych autostradach miejskich, aby zmniejszyć liczbę wypadków, oszczędzić paliwo i zmniejszyć zanieczyszczenie.

Podróż do Big Data

Firmy zawsze miały do czynienia z dużą ilością danych w różnych formach. Zmiana, jaką przynosi duże zbiory danych, polega na tym, co możesz zrobić z tymi informacjami. Jeśli masz odpowiednią technologię, możesz wykorzystać duże zbiory danych do przewidywania i rozwiązywania problemów biznesowych oraz reagowania na pojawiające się okazje. Dzięki big data możesz. Firmy podróżujące po Big Data zawsze miały do czynienia z dużą ilością danych w różnych formach. Zmiana, jaką przynosi duże zbiory danych, polega na tym, co możesz zrobić z tymi informacjami. Jeśli masz odpowiednią technologię, możesz wykorzystać duże zbiory danych do przewidywania i rozwiązywania problemów biznesowych oraz reagowania na pojawiające się okazje. Dzięki big data możesz analizować wzorce

danych, aby zmienić wszystko, od sposobu zarządzania miastami, zapobiegania awariom, przeprowadzania eksperymentów, zarządzania ruchem, zwiększania satysfakcji klientów lub poprawy jakości produktu, żeby wymienić tylko kilka przykładów. Nowe technologie i narzędzia, które są sercem tej książki, mogą pomóc Ci zrozumieć i uwolnić ogromną moc big data, zmieniając świat, jaki znamy.