Krok 5: Analiza danych

Ta część obejmuje następujące tematy:

* Rzeczy do rozważenia podczas analizy danych źródłowych dla aplikacji BI

* Różnica między fazą analizy systemów tradycyjnej metodologii a skoncentrowaną na biznesie analizą danych wykonywaną na tym etapie

* Odgórne logiczne modelowanie danych, w tym logiczne modele danych specyficzne dla projektu, zintegrowane logiczne modele danych przedsiębiorstwa i specyficzne dla danych komponenty metadanych biznesowych zebrane podczas logicznego procesu modelowania danych

* Analiza danych źródłowych oddolna, w tym sposób zastosowania trzech zestawów reguł transformacji do danych źródłowych: reguły konwersji danych technicznych, reguły domeny danych biznesowych i reguły integralności danych biznesowych

* Odpowiedzialność za archeologię danych, czyszczenie danych i egzekwowanie jakości danych, a także konieczność segregowania (priorytetowania) działań związanych z czyszczeniem danych

* Krótkie opisy działań związanych z analizą danych, rezultatów wynikających z tych działań oraz zaangażowanych ról

* Ryzyko niewykonania Kroku 5

Rzeczy do rozważenia

Dane źródłowe

* Czy wiemy, gdzie znajdują się dane źródłowe? W jakich systemach? W jakich plikiach? W jakich bazach danych?

* Czy istnieje wiele potencjalnych źródeł tych samych danych?

* Czy żądane dane źródłowe zostały już zamodelowane?

* Jak aktualne są metadane biznesowe w tych modelach?

* Czy właściciele danych ratyfikowali metadane biznesowe?

* Czy wiemy, kim są właściciele danych? Kto ma władzę nad danymi źródłowymi?

* Czy istnieje inny rodzaj dokumentacji dla żądanych danych źródłowych? Czy jest aktualny i kompletny?

* Gdzie jest ta dokumentacja? W repozytorium metadanych? W programach? W instrukcjach?

Jakość danych

* Czy wiemy, jak czyste są dane źródłowe?

* Jak czyste muszą być dane zgodnie z naszą działalnością przedstawiciela?

* Czy będzie to wystarczająco czyste dla innych pracowników wiedzy, analityków biznesowych i menedżerów biznesowych, którzy będą korzystać z tych samych danych?

* Czy wiemy, kim oni są?

* Skąd bierzemy reguły biznesowe dla danych? Od właścicieli danych? Od przedstawiciela biznesowego w projekcie?

Oczyszczanie danych

* Czy błędy danych zostały już udokumentowane przez inne zespoły projektowe?

* Jeśli tak, gdzie jest ta dokumentacja?

* Kto mógłby wiedzieć, jakie są znane błędy danych?

* Czy kody są tłumaczone w programach operacyjnych? Jeśli tak, w jakich programach?

* Czy istnieje książka tłumaczeń kodów dla pól zakodowanych?

* Czy wiemy już, które dane są krytyczne, a które ważne i co jest nieistotne (dla celów segregacji danych)?

Systemy operacyjne są opracowywane jako automatyka kominowa rozwiązania dla poszczególnych jednostek biznesowych, a nie jako wsparcie procesu podejmowania decyzji wykonawczych. Dlatego systemy operacyjne nie są zaprojektowane tak, aby integrować lub uzgadniać ze sobą w celu zapewnienia spójnego obrazu między organizacjami. Z drugiej strony aplikacje BI są zaprojektowane właśnie do tego celu - do dostarczania społeczności biznesowej zintegrowanych i uzgodnionych danych biznesowych.

Analiza danych skoncentrowana na biznesie

Dla wielu organizacji inicjatywa wspomagania decyzji BI jest pierwszą próbą połączenia danych biznesowych z wielu źródeł w celu udostępnienia ich w różnych działach. Organizacje, które używają tradycyjnej metodologii rozwoju systemów w swoich projektach BI, zwykle napotykają poważne problemy z danymi źródłowymi, gdy próbują wdrożyć procesy ekstrakcji/transformacji/ładowania (ETL), ponieważ tradycyjne metodologie rozwoju nie mają kroków do analizy domen danych na wczesnym etapie rozwoju proces. Mają w najlepszym razie fazę analizy systemów dla funkcji aplikacji, ale nie mają fazy analizy danych skoncentrowanej na biznesie dla danych bazowych.

Krok analizy danych zorientowanej na biznes jest najbardziej krytycznym krokiem między organizacjami.

Krok 5, Analiza Danych, różni się od fazy analizy systemów w tradycyjnej metodologii. Czynności wykonywane tradycyjnie podczas analizy systemów mają na celu podjęcie decyzji projektowej dla systemu, który ma zostać zbudowany. Czynności wykonywane podczas analizy danych mają na celu zrozumienie i skorygowanie istniejące rozbieżności w danych biznesowych, niezależnie od sposobu projektowania i implementacji systemu. Analiza danych jest zatem czynnością skoncentrowaną na biznesie, a nie czynnością skoncentrowaną na systemie. Do przeprowadzenia rygorystycznej analizy danych wymagane są dwie uzupełniające się metody:

1. Odgórne logiczne modelowanie danych w celu integracji i spójności

2. Oddolna analiza danych źródłowych pod kątem standaryzacji i jakości

Modelowanie danych logicznych od góry do dołu

Najskuteczniejszą techniką wykrywania i dokumentowania pojedynczego zintegrowanego i uzgodnionego między organizacjami widoku danych biznesowych jest modelowanie relacji między podmiotami (E-R), znane również jako modelowanie danych logicznych. Popularnym podejściem do

modelowania E-R we wczesnych latach 80. było modelowanie wszystkich danych dla całej organizacji jednocześnie. Chociaż takie podejście było wartościowym przedsięwzięciem architektonicznym, nie przyniosło lepszych systemów, ponieważ proces ten nie był zintegrowany z cyklem rozwoju systemu. Bardziej efektywnym podejściem jest włączenie modelowania E-R do każdego projektu, a następnie łączenie logicznych modeli danych specyficznych dla projektu w jeden skonsolidowany model danych przedsiębiorstwa w miarę upływu czasu.

Logiczny model danych specyficznych dla projektu

Modelowanie E-R opiera się na regułach normalizacyjnych, które są stosowane zarówno podczas modelowania danych odgórnych, jak i oddolnej analizy danych źródłowych. Stosowanie reguł normalizacyjnych wraz z innymi zasadami administrowania danymi zapewnia, że każdy element danych w ramach projektu BI jest jednoznacznie zidentyfikowany, poprawnie nazwany i właściwie zdefiniowany, a jego domena jest walidowana dla wszystkich ludzi biznesowych, którzy będą mieli dostęp do danych. W ten sposób znormalizowany logiczny model danych specyficznych dla projektu zapewnia formalną reprezentację danych dokładnie taką, jaka istnieje w świecie rzeczywistym, bez nadmiarowości i niejednoznaczności. Ta formalna reprezentacja danych jest zgodna z inną zasadą normalizacji: niezależność procesu. Dlatego z definicji logiczny model danych, który jest oparty na regułach normalizacyjnych, jest również niezależny od procesu. Niezależność procesu oznacza, że na strukturę i zawartość logicznego modelu danych nie ma wpływu żaden typ bazy danych, ścieżka dostępu, projekt, program, narzędzie ani sprzęt . Ze względu na niezależność procesu, logiczny model danych jest widokiem biznesowym, a nie widokiem bazy danych czy widokiem aplikacji. W związku z tym unikalna część danych, która istnieje tylko raz w rzeczywistym świecie biznesowym, również istnieje tylko raz w logicznym modelu danych, mimo że może być fizycznie przechowywana w wielu plikach źródłowych lub wielu docelowych bazach danych BI.

Logiczny model danych przedsiębiorstwa

Obowiązkiem grupy architektury korporacyjnej lub administracji danymi, jeśli organizacja nie ma grupy architektury korporacyjnej, jest połączenie modeli danych logicznych specyficznych dla projektu w zintegrowany i ustandaryzowany model danych logicznych przedsiębiorstwa. Ten logiczny model danych przedsiębiorstwa, znany również jako architektura informacji przedsiębiorstwa, nie jest tworzony od razu, ani nie jest warunkiem wstępnym ukończenia projektów BI. Zamiast tego logiczny model danych przedsiębiorstwa ewoluuje w czasie i może nigdy nie zostać ukończony. Nie trzeba go wypełniać, ponieważ celem tego procesu nie jest stworzenie gotowego modelu, ale wykrycie i rozwiązanie rozbieżności danych między różnymi widokami i implementacjami tych samych danych. Te rozbieżności danych występują masowo w systemach operacyjnych typu „Stoppipe” i są głównymi przyczynami niezdolności organizacji do dostarczania swoim pracownikom biznesowym zintegrowanych i spójnych informacji międzyorganizacyjnych. Odkrycie tych rozbieżności powinno zostać przyjęte i uczczone przez zespół projektu BI, a zwłaszcza przez ludzi biznesu, ponieważ dane o niskiej jakości są w końcu rozwiązywane i rozwiązywane. Przejęcie kontroli nad istniejącym chaosem danych jest przecież jedną z głównych funkcji każdej inicjatywy wspierającej podejmowanie decyzji BI.

Gdyby organizacje postępowały zgodnie z najlepszymi praktykami analizy biznesowej, opracowując logiczne modele danych dla wszystkich swoich aplikacji operacyjnych i łącząc je (z czasem) w logiczny model danych przedsiębiorstwa, wysiłek związany z rozwojem wspomagania decyzji BI mógłby być znacząco zredukowany. Umożliwiłoby to zespołom projektowym BI zwiększenie szybkości dostarczania ludziom biznesu wiarygodnych informacji wspierających podejmowanie decyzji. Innymi słowy, zespoły

projektowe BI mogą dostarczać „szybkie trafienia”, których każdy chce - i dostarczać je z wysoką jakością.

Uczestnicy modelowania danych logicznych

Sesje logicznego modelowania danych są zazwyczaj ułatwiane i prowadzone przez administratora danych, który ma solidne doświadczenie biznesowe. Jeżeli administrator danych nie rozumie dobrze biznesu, w tym zadaniu musi mu pomóc ekspert w danej dziedzinie. Przedstawiciel biznesowy oraz ekspert merytoryczny przydzielony do projektu BI są aktywnymi uczestnikami sesji modelowania. Jeśli dane są wyodrębniane z kilku różnych systemów operacyjnych, wielu właścicieli danych może być zmuszonych do udziału w projekcie BI, ponieważ każdy system operacyjny może podlegać zarządzaniu przez innego właściciela. Właściciele danych to osoby biznesowe, które mają uprawnienia do ustanawiania reguł biznesowych i ustalania zasad biznesowych dla tych fragmentów danych, które pochodzą z ich działów. W przypadku wykrycia rozbieżności w danych, obowiązkiem właścicieli danych jest uporządkowanie różnych poglądów biznesowych i zatwierdzenie zgodnego z prawem wykorzystania ich danych. Ten proces uzgadniania danych jest i powinien być funkcją biznesową, a nie funkcją informatyczną (IT), chociaż administratorzy danych, którzy zwykle pracują dla IT, ułatwiają proces wyszukiwania. Analitycy systemowi, programiści i administratorzy baz danych powinni być również dostępni do udziału w niektórych sesjach modelowania w miarę potrzeb. Ci technicy IT utrzymują aplikacje i struktury danych organizacji i często wiedzą więcej niż ktokolwiek inny o danych – jak i gdzie są przechowywane, jak są przetwarzane i ostatecznie jak są wykorzystywane przez ludzi biznesu. Ponadto technicy ci często mają dogłębną wiedzę na temat dokładności danych, ich związku z innymi danymi, historii ich wykorzystania oraz tego, jak treść i znaczenie danych zmieniały się w czasie. Ważne jest, aby uzyskać zaangażowanie w projekt BI od tych zasobów IT, ponieważ często są one zajęte „gaszeniem pożarów” i pracą nad ulepszeniami systemów operacyjnych.

Standaryzowane metadane biznesowe

Logiczny model danych, reprezentujący pojedynczy, międzyorganizacyjny widok biznesowy danych, składa się z diagramu E-R i wspierających biznesowych metadanych. Metadane biznesowe obejmują informacje o obiektach danych biznesowych, ich elementach danych oraz relacjach między nimi. Metadane biznesowe oraz metadane techniczne, które są dodawane na etapie projektowania i budowy, zapewniają spójność danych oraz poprawiają zrozumienie i interpretację danych w środowisku wspomagania decyzji BI.

* Nazwa danych, oficjalna etykieta opracowana z formalnej taksonomii nazewnictwa danych, powinna składać się ze słowa pierwszego, słowa klasy i kwalifikatorów. Każda nazwa danych jednoznacznie identyfikuje jedną część danych w logicznym modelu danych. Nie powinny istnieć żadne synonimy ani homonimy.

* Definicja danych to jedno- lub dwuzdaniowy opis obiektu danych lub elementu danych, podobny do definicji w słowniku językowym. Jeśli obiekt danych ma wiele podtypów, każdy podtyp powinien mieć własną unikalną definicję danych. Definicja danych wyjaśnia znaczenie obiektu danych lub elementu danych. Nie obejmuje tego, kto stworzył obiekt, kiedy był ostatnio aktualizowany, jaki system go stworzył, jakie wartości zawiera i tak dalej. Informacje te są przechowywane w innych komponentach metadanych (np. własność danych, zawartość danych).

* Relacja danych to powiązanie biznesowe między wystąpieniami danych w działalności biznesowej. Każda relacja dotycząca danych opiera się na regułach biznesowych i politykach biznesowych dotyczących powiązanych wystąpień danych w ramach każdej działalności biznesowej.

* Identyfikator danych jednoznacznie identyfikuje wystąpienie obiektu danych. Identyfikator danych powinien być znany przedsiębiorcom. Powinna być również „minimalna", co oznacza, że powinna być jak najkrótsza (składająca się z wystarczającej liczby elementów danych, aby była unikalna). Ponadto identyfikator danych powinien być nieinteligentny, bez wbudowanej logiki. Na przykład numery kont 0765587654 i 0765563927, gdzie 0765 jest wbudowanym numerem oddziału, byłyby słabymi identyfikatorami danych.

Logiczny identyfikator danych to nie to samo, co klucz podstawowy w bazie danych. Chociaż identyfikator danych może być używany jako klucz podstawowy, często jest zastępowany kluczem zastępczym („wymyślonym") podczas projektowania bazy danych.

* Typ danych opisuje strukturę elementu danych, kategoryzując typ wartości (znak, liczba, dziesiętna, data), które mogą być w nim przechowywane.

* Długość danych określa rozmiar elementu danych dla określonego typu danych. Na przykład dziesiętny element danych może być polem kwoty z dwiema cyframi po przecinku lub polem stawki z trzema cyframi po przecinku.

* Zawartość danych (domena) identyfikuje rzeczywiste dopuszczalne wartości elementu danych specyficzne dla jego typu danych i długości danych. Domena może być wyrażona jako zakres wartości, lista dopuszczalnych wartości, ogólna reguła biznesowa lub reguła zależności między dwoma lub większą liczbą elementów danych.

* Reguła danych to ograniczenie obiektu danych lub elementu danych. Ograniczenie danych może również dotyczyć relacji danych. Ograniczenie danych może mieć postać reguły biznesowej lub reguły zależności między obiektami danych lub elementami danych, na przykład „Gimnalna stopa procentowa musi być wyższa niż dolna stopa procentowa".

* Zasady dotyczące danych określają zawartość i zachowanie obiektu danych lub elementu danych. Zwykle wyraża się to jako politykę organizacyjną lub rozporządzenie rządowe. Na przykład „Pacjenci korzystający z Medicare muszą mieć co najmniej 65 lat".

* Własność danych identyfikuje osoby, które mają uprawnienia do ustanawiania i zatwierdzania metadanych biznesowych dla obiektów danych i elementów danych będących pod ich kontrolą.

Oddolna analiza danych źródłowych

Analiza danych nie może zostać zatrzymana po logicznym modelowaniu danych od góry do dołu, ponieważ dane źródłowe często nie są zgodne z regułami i zasadami biznesowymi uchwyconymi podczas sesji modelowania. Gdyby nie przeprowadzono analizy danych źródłowych typu bottom-up, problemy z danymi i naruszenia reguł biznesowych nie zostałyby wykryte do momentu wdrożenia procesu ETL. Pewne problemy z jakością danych nie zostałyby wykryte dopiero po wdrożeniu, i to tylko wtedy, gdyby ktoś na nie poskarżył.

Zasady konwersji danych technicznych

Za każdym razem, gdy dane są mapowane z jednego systemu do drugiego, czy to w przypadku tradycyjnej konwersji systemów, czy też mapowania źródło-cel w aplikacjach BI, należy przestrzegać następujących zasad technicznych. v

1. Typy danych źródłowych elementów danych muszą odpowiadać typom danych docelowych elementów danych.

2. Długości danych muszą być odpowiednie, aby umożliwić przenoszenie, rozszerzanie lub skracanie źródłowych elementów danych do docelowych elementów danych.

3. Logika programów manipulujących elementami danych źródłowych musi być zgodna z zawartością elementów danych źródłowych i mieć do niej zastosowanie. W przeciwnym razie wyniki będą nieprzewidywalne.

Zasady domeny danych biznesowych

Znacznie większy wysiłek związany z analizą danych źródłowych kręci się wokół reguł domeny danych biznesowych. Te zasady są ważniejsze dla ludzi biznesu niż techniczne zasady konwersji danych. Element danych źródłowych może spełniać wszystkie trzy reguły konwersji danych technicznych, ale nadal ma nieprawidłowe wartości. Reguły domeny danych biznesowych to reguły dotyczące semantyki (znaczenia i interpretacji) zawartości danych. Służą do identyfikacji i korygowania naruszeń danych

Zasady integralności danych biznesowych

Podobnie jak w przypadku reguł domeny danych biznesowych, reguły integralności danych biznesowych są znacznie ważniejsze dla poprawy jakości informacji niż reguły konwersji danych technicznych. Reguły integralności danych biznesowych regulują zawartość semantyczną pomiędzy zależnymi lub powiązanymi elementami danych, a także ograniczenia nałożone przez reguły biznesowe i politykę biznesową.

Każdy krytyczny i ważny element danych musi zostać zbadany pod kątem tych defektów i należy podjąć decyzję, czy i jak je poprawić. Konsumenci informacji (osoby biznesowe, które będą wykorzystywać te elementy danych do podejmowania decyzji biznesowych) oraz właściciele danych powinni podjąć tę decyzję po omówieniu wpływu działań czyszczących ze sponsorem biznesowym, kierownikiem projektu i głównym zespołem.

Oczyszczanie danych

Jednym z najczęściej wymienianych celów dla aplikacji BI jest dostarczanie społeczności biznesowej czystych, zintegrowanych i uzgodnionych danych. Jeśli nie zostaną uwzględnione wszystkie trzy zestawy reguł mapowania danych, ten cel nie może zostać osiągnięty. Wiele organizacji znajdzie w swoich systemach źródłowych znacznie wyższy procent brudnych danych niż oczekiwano, a ich wyzwaniem będzie podjęcie decyzji, ile z nich wyczyścić.

Odpowiedzialność za jakość danych

Archeologia danych (proces wyszukiwania złych danych), czyszczenie danych (proces korygowania złych danych) i egzekwowanie jakości danych (proces zapobiegania defektom danych u źródła) to wszystkie obowiązki biznesowe, a nie obowiązki IT. Oznacza to, że ludzie biznesu (odbiorcy informacji, a także właściciele danych) muszą być zaangażowani w działania związane z analizą danych i znać reguły mapowania danych źródłowych. Ponieważ właściciele danych tworzą dane i ustanawiają reguły biznesowe i zasady dotyczące danych, są bezpośrednio odpowiedzialni przed dalszymi konsumentami informacji (pracownicy wiedzy, analitycy biznesowi, menedżerowie biznesu), którzy muszą korzystać z tych danych. Jeżeli dalsi odbiorcy informacji opierają swoje decyzje biznesowe na danych o niskiej jakości i ponoszą z tego powodu straty finansowe, właściciele danych muszą zostać pociągnięci do odpowiedzialności. W przeszłości tej odpowiedzialności nie było w systemach rurowych. Odpowiedzialność za jakość danych nie jest ani tymczasowa, ani specyficzna dla BI, a ludzie biznesu

muszą zobowiązać się do trwałego przyjęcia tych obowiązków. Jest to część wymaganej zmiany kulturowej, której omówienie wykracza poza zakres tej książki. Wyzwaniem dla IT i dla sponsora biznesowego w projekcie BI jest wymuszenie nieuniknionych zadań archeologii danych i czyszczenia danych w celu spełnienia celów jakości środowiska wspomagania decyzji BI.

Krok 5, Analiza Danych, może być czasochłonny, ponieważ wśród ludzi biznesu może toczyć się wiele bitew o słuszne znaczenie i dziedzinę danych.

Chociaż narzędzia do czyszczenia danych mogą pomóc w procesie archeologii danych, opracowywanie specyfikacji czyszczenia danych jest głównie procesem ręcznym. Menedżerowie IT, menedżerowie biznesowi i właściciele danych, którzy nigdy nie przeszli oceny jakości danych i inicjatywy czyszczenia danych, często nie doceniają czasu i wysiłku wymaganego od swoich pracowników czterokrotnie lub więcej.

Proces wyboru danych źródłowych

Dane o niskiej jakości są tak przytłaczającym problemem, że większość organizacji nie będzie w stanie skorygować wszystkich rozbieżności.

1. Zidentyfikuj wymagane dane. : Zidentyfikuj interesujące Cię dane i znaczenie tych danych. Czyszczenie danych to wspólny wysiłek analityków biznesowych zaznajomionych z semantyką danych i analityków jakości danych, którzy znają specyficzne dla programu znaczenie danych (np. użycie i znaczenie wartości „flagi” lub przedefiniowanych układów rekordów) .

2. Przeanalizuj zawartość danych. : Przeanalizuj dane pod kątem treści, znaczenia i ważności. Wiele organizacji zgromadziło ogromne ilości danych w plikach i bazach danych. Dane te stanowią potencjalną kopalnię cennej wiedzy biznesowej i są potencjalnie dobrym źródłem do eksploracji danych. Jednak najpierw należy ocenić jakość zawartości danych, ponieważ eksploracja brudnych danych ma niewielką wartość.

3. Wybierz dane do BI. : Określ, które dane należy uwzględnić w aplikacji BI. Wybierz tylko te dane, które spełnią podstawowe wymagania biznesowe. Nawet w przypadku zautomatyzowanych narzędzi, koszt zapewnienia jakości danych dla kompleksowego środowiska wspomagającego podejmowanie decyzji BI staje się dla większości organizacji niedopuszczalny. Poniżej znajdują się pytania, które należy wziąć pod uwagę przy wyborze danych.

- Czy te dane są wystarczająco czyste do wykorzystania przy podejmowaniu decyzji?

- Jeśli nie, czy te dane mogą zostać oczyszczone, przynajmniej częściowo? Czy wiemy jak?

- Czy brudne dane są powodem zbudowania tej aplikacji BI? Czy czyszczenie tych danych jest zatem obowiązkowe?

- Ile wysiłku zajmie wymyślenie sposobu oczyszczenia danych?

- Ile będzie kosztować czyszczenie danych?

- Jaka jest korzyść z czyszczenia danych w przeciwieństwie do przenoszenia ich do aplikacji BI przy obecnym poziomie zabrudzenia?

- Jakie są oczekiwania co do jakości danych od konsumentów informacji i ogólnie od kierownictwa firmy?

4. Przygotuj specyfikacje czyszczenia danych. : Personel IT współpracujący z przedstawicielem biznesowym pozna zasady biznesowe potrzebne do napisania specyfikacji czyszczenia danych. W istocie jest to proces reengineeringu danych źródłowych.

5. Wybierz narzędzia. : Wybierz ETL i narzędzia do oczyszczania. Określ, czy nabycie narzędzia ETL, narzędzia do oczyszczania lub obu tych elementów jest odpowiednie i opłacalne. Sprawdź przydatność i skuteczność tych narzędzi. Niektóre specyfikacje czyszczenia danych mogą być bardzo skomplikowane. Upewnij się, że narzędzia są w stanie je obsłużyć.

Kluczowe punkty wyboru danych

Identyfikując i wybierając dane operacyjne, które mają być wykorzystane do zapełnienia docelowych baz danych BI, należy wziąć pod uwagę kilka kluczowych punktów

* Integralność danych: Jak wewnętrznie spójne są dane? To jest najważniejsze kryterium.

- Im większy odsetek ręcznie wprowadzonych danych (dane wprowadzone z niewielką liczbą lub bez kontroli danych, edycji i walidacji), tym niższa integralność.

- Błędy programistyczne również zanieczyszczają ogromne ilości danych - i robią to automatycznie.

- Im niższa integralność, tym większe zapotrzebowanie na oczyszczenie.

* Dokładność danych: jak dokładne są dane? To kolejne ważne kryterium.

- Jak dane są reprezentowane wewnętrznie?

- Jaka jest skala i precyzja danych liczbowych?

- Jak są sformatowane dane dat?

* Dokładność danych: Jak poprawne są dane?

- Czy w programie do wprowadzania danych są kontrole edycji?

- Czy wartości zależne są sprawdzane krzyżowo? Na przykład, czy program do wprowadzania danych zabrania, aby data wygaśnięcia poprzedzała datę wejścia w życie?

- Czy istnieje proces operacyjny służący do poprawiania danych?

- Czy są przechowywane obliczone wartości? Jakie mechanizmy, jeśli w ogóle, są wdrożone, aby te wartości były dokładne?

* Wiarygodność danych: ile lat mają dane?

- Jakie jest generowanie danych (na koniec miesiąca, co tydzień, codziennie)?

- Czy dane zostały uzyskane z bezpośrednich źródeł, czy z pobrań?

- Czy znane jest źródło danych?

- Czy dane są duplikatem danych w innym magazynie danych? Jeśli tak, czy dane są aktualne?

* Format danych: im dane są bliższe docelowemu formatowi danych, tym mniej będzie wymagań dotyczących konwersji. Od najwyższego do najniższego priorytety formatu to:

- Dane z relacyjnej bazy danych (np. DB2, Oracle)

- Dane z nierelacyjnej bazy danych (np. IMS, CA-IDMS)

- Pliki płaskie (np. VSAM, ISAM)

Jakość danych źródłowych będzie tak dobra, jak egzekwowanie procesów jakościowych w systemach operacyjnych. Obowiązkowe procesy jakości powinny obejmować zasady wprowadzania danych i kontroli edycji w programach. Jeśli te procesy nie są wymuszane lub nie istnieją, dane zwykle ulegają uszkodzeniu, niezależnie od tego, czy dane znajdują się w relacyjnej bazie danych, czy w starym pliku VSAM.

Oczyszczać czy nie oczyszczać

Wiele organizacji zmaga się z tym pytaniem. Badania nad czyszczeniem danych wskazują, że niektóre organizacje bagatelizują czyszczenie danych, aby osiągnąć cele krótkoterminowe. Konsekwencje nieuwzględnienia danych o niskiej jakości zwykle uderzają w sedno, gdy ich przedsięwzięcia biznesowe kończą się niepowodzeniem lub napotykają na negatywne skutki z powodu niedokładnych danych. Należy pamiętać, że czyszczenie danych jest procesem pracochłonnym, czasochłonnym i kosztownym. Czyszczenie wszystkich danych zwykle nie jest ani uzasadnione kosztowo, ani praktyczne, ale czyszczenie żadnych danych nie jest równie nie do przyjęcia. Dlatego ważne jest, aby dokładnie przeanalizować dane źródłowe i sklasyfikować elementy danych jako krytyczne, ważne lub nieistotne dla firmy. Skoncentruj się na oczyszczeniu wszystkich krytycznych elementów danych, pamiętając, że nie wszystkie dane są równie ważne dla wszystkich ludzi biznesu. Następnie wyczyść tyle ważnych elementów danych, ile pozwala na to czas, i przenieś nieistotne elementy danych do docelowych baz danych BI bez ich czyszczenia. Innymi słowy, nie musisz czyścić wszystkich danych i nie musisz robić tego od razu.

Oczyszczanie systemów operacyjnych

Gdy wybrane dane są czyszczone, standaryzowane i przenoszone do docelowych baz danych BI, należy rozważyć, czy należy również wyczyścić pliki źródłowe i źródłowe bazy danych. Kierownictwo może zapytać, dlaczego nie wydać trochę dodatkowych pieniędzy i czasu na wyczyszczenie plików źródłowych i baz danych, tak aby dane były spójne zarówno w źródle, jak iw miejscu docelowym? Jest to ważne pytanie i zdecydowanie należy skorzystać z tej opcji, jeśli działanie naprawcze w systemie źródłowym jest tak proste, jak dodanie kontroli edycji do programu do wprowadzania danych. Jeśli akcja korygująca wymaga zmiany struktury pliku, co oznacza modyfikację (jeśli nie przepisanie) większości programów, które uzyskują dostęp do tego pliku, koszt tak inwazyjnej akcji korygującej w systemie operacyjnym jest prawdopodobnie nieuzasadniony - zwłaszcza jeśli złe dane są nie ingerowanie w potrzeby operacyjne tego systemu. Pamiętaj, że wiele firm nie chciało nawet wprowadzać tak drastycznych zmian w niesławnym teraz problemie Y2K; dokonali tych zmian tylko wtedy, gdy stało się jasne, że ich przetrwanie jest zagrożone. Z pewnością niewłaściwie użyte pole kodu nie zagraża przetrwaniu organizacji. Stąd szanse na naprawę systemów operacyjnych są nikłe.

Czynności związane z analizą danych

Czynności do analizy danych nie muszą być wykonywane liniowo. Poniższa lista krótko opisuje czynności związane z Krokiem 5, Analiza Danych.

1. Przeanalizuj zewnętrzne źródła danych. Oprócz wymagania wewnętrznych operacyjnych danych źródłowych wiele aplikacji BI wymaga danych ze źródeł zewnętrznych. Scalanie danych zewnętrznych z danymi wewnętrznymi stwarza własny zestaw wyzwań. Dane zewnętrzne są często brudne i niekompletne i zwykle nie mają tego samego formatu lub struktury klucza, co dane wewnętrzne. Zidentyfikuj i rozwiąż te różnice na tym etapie.

2. Udoskonal logiczny model danych. Logiczny model danych wysokiego poziomu, specyficzny dla projektu, powinien zostać utworzony w jednym z poprzednich kroków. Ponadto niektóre lub wszystkie dane wewnętrzne i zewnętrzne mogły być modelowane na podstawie innych projektów i mogą już stanowić część logicznego modelu danych przedsiębiorstwa. W takim przypadku wyodrębnij reprezentatywną część logicznego modelu danych przedsiębiorstwa i rozszerz ją o nowe obiekty danych, nowe relacje danych i nowe elementy danych. Jeśli wymagane dane nie były wcześniej modelowane, utwórz nowy logiczny model danych dla zakresu tego projektu BI. Powinien zawierać wszystkie wewnętrzne i zewnętrzne elementy danych.

3.Analizuj jakość danych źródłowych. Równolegle do tworzenia lub rozbudowy logicznego modelu danych należy szczegółowo przeanalizować jakość wewnętrznych i zewnętrznych plików źródłowych oraz źródłowych baz danych. Często zdarza się, że istniejące dane operacyjne nie są zgodne z określonymi regułami biznesowymi i politykami biznesowymi. Wiele elementów danych jest używanych do różnych celów lub po prostu pozostawia się je puste. Zidentyfikuj wszystkie te rozbieżności i uwzględnij je w logicznym modelu danych.

4. Rozwiń logiczny model danych przedsiębiorstwa. Gdy model danych logicznych specyficznych dla projektu jest względnie stabilny, należy go ponownie połączyć z modelem danych logicznych przedsiębiorstwa. Podczas tego procesu scalania mogą zostać zidentyfikowane dodatkowe rozbieżności lub niespójności danych. Zostaną one odesłane do projektu BI w celu rozwiązania.

5. Rozwiąż rozbieżności danych. Czasami rozbieżności danych wykryte podczas analizy danych dotyczą innych przedstawicieli biznesowych z innych projektów. W takim przypadku wezwij innych przedstawicieli biznesu, a także właścicieli danych, aby ustalili różnice. Albo odkryją nowy legalny podtyp obiektu danych lub nowy element danych, który musi być modelowany jako taki, albo będą musieli rozwiązać i ujednolicić niespójności.

6. Napisz specyfikacje czyszczenia danych. Po zidentyfikowaniu i zamodelowaniu wszystkich problemów z danymi napisz specyfikacje dotyczące czyszczenia danych. Specyfikacje te powinny być w prostym języku angielskim, aby mogły zostać zweryfikowane przez właściciela danych i ludzi biznesu, którzy będą z nich korzystać.

Rezultaty wynikające z tych działań

1.Znormalizowany iw pełni przypisany logiczny model danych. Ten specyficzny dla projektu logiczny model danych jest w pełni znormalizowanym diagramem relacji encji, pokazującym encje jądra, encje asocjacyjne, encje charakterystyczne, liczność, opcjonalność, unikalne identyfikatory i wszystkie atrybuty.

2.Metadane biznesowe. Podmioty biznesowe i atrybuty z logicznego modelu danych muszą być opisane za pomocą metadanych. Komponenty metadanych biznesowych specyficzne dla danych obejmują nazwy danych, definicje danych, relacje danych, unikalne identyfikatory, typy danych, długości danych, domeny, reguły biznesowe, zasady i własność danych. Są one zwykle rejestrowane w repozytorium narzędzi narzędzia do wspomagania komputerowego inżynierii oprogramowania (CASE).

3. Specyfikacje czyszczenia danych. W tym dokumencie opisano logikę czyszczenia, którą należy zastosować do danych źródłowych, aby zapewnić ich zgodność z regułami konwersji danych technicznych, regułami domeny danych biznesowych oraz regułami integralności danych biznesowych. Ten dokument będzie używany do tworzenia specyfikacji transformacji w dokumencie mapowania źródło-cel w kroku 9, projektowanie ETL.

4. Rozszerzony logiczny model danych przedsiębiorstwa. Ten element dostarczany jest zakulisowo przez administrację danych lub grupę zajmującą się architekturą przedsiębiorstwa, która łączy logiczny model danych konkretnego projektu z logicznym modelem danych przedsiębiorstwa. Wszelkie odrzucone encje lub atrybuty oraz wszelkie rozbieżności między modelami zostaną przedstawione zespołowi projektu BI do rozwiązania.

Role zaangażowane w te działania

* Przedstawiciel firmy. Przedstawiciel biznesowy przypisany do projektu BI jest głównym uczestnikiem działań związanych z modelowaniem danych logicznych odgórnych, a także oddolnymi działaniami analizy danych źródłowych. Dostarcza metadane biznesowe administratorowi danych i pomaga analitykowi jakości danych w analizie plików źródłowych.

* Administrator danych. Administrator danych jest przeszkolony w zakresie logicznego modelowania danych, metadanych biznesowych, technik normalizacji, reguł domeny danych biznesowych, reguł integralności danych biznesowych oraz metod standaryzacji. Opis stanowiska administratora danych

odpowiada czynnościom kroku analizy danych. Administrator danych będzie osobą wiodącą na tym etapie i będzie ułatwiać wszystkie sesje modelowania danych. Do jego obowiązków należy również dokumentowanie logicznego modelu danych oraz wspierających biznes metadanych w narzędziu CASE.

* Analityk jakości danych. Analityk jakości danych jest analitykiem systemowym, przeszkolonym w posługiwaniu się regułami konwersji danych technicznych, w czytaniu i pisaniu programów oraz w wydobywaniu danych ze wszystkich typów plików źródłowych i baz danych źródłowych. Znalezienie naruszeń danych w plikach źródłowych i źródłowych bazach danych jest głównym obowiązkiem analityka jakości danych. Ściśle współpracuje z administratorem danych w celu modelowania anomalii danych i korygowania naruszeń danych z pomocą przedstawiciela biznesowego i właścicieli danych.

* Główny programista ETL. Główny programista ETL musi być zaangażowany w przeglądy modelowania i musi być świadomy skali problemów z jakością danych znalezionych w plikach źródłowych i źródłowych bazach danych. Musi zrozumieć złożoność czyszczenia danych, ponieważ algorytmy czyszczenia muszą być włączone do procesu ETL. W niektórych przypadkach narzędzie ETL nie będzie w stanie obsługiwać niektórych algorytmów czyszczących i może być konieczne napisanie niestandardowego kodu.

* Administrator metadanych. Jako opiekun metadanych i administrator repozytorium metadanych, administrator metadanych musi wiedzieć, jakie biznesowe komponenty metadanych są zbierane iw jaki sposób. Niektóre komponenty metadanych można wprowadzić do narzędzia CASE, podczas gdy inne komponenty mogą zostać przechwycone w dokumentach edytora tekstu lub arkuszach kalkulacyjnych. Administrator metadanych będzie musiał wyodrębnić komponenty metadanych z różnych plików i narzędzi i połączyć je z repozytorium metadanych.

* Interesariusze (w tym właściciele danych) . Ludzie biznesu korzystający z aplikacji BI są zwykle dalszymi konsumentami informacji, a nie właścicielami danych. Na tym etapie zarówno konsumenci informacji, jak i właściciele danych są odpowiedzialni za standaryzację danych biznesowych oraz ustalanie reguł i zasad dotyczących danych. Ciągłe spory dotyczące danych między właścicielami danych a konsumentami informacji muszą być przerzucane do kadry kierowniczej w celu rozwiązania.

* Ekspert merytoryczny. Ekspert merytoryczny pomaga administratorowi danych i analitykowi jakości danych interpretując dane biznesowe, wyjaśniając zasady i polityki biznesowe dla danych oraz określając dziedzinę (prawidłowe wartości) danych. Ponadto ekspert merytoryczny odpowiada za

wyszukiwanie problemów z danymi w źródłowych zbiorach danych i źródłowych bazach danych oraz za sugerowanie sposobów ich korygowania.

Ryzyko niewykonania kroku 5

Menedżerowie biznesowi, menedżerowie IT i technicy IT często nie chcą poświęcać czasu na rygorystyczną analizę danych, która obejmuje logiczne modelowanie danych, archeologię danych źródłowych i czyszczenie danych. Traktują te działania jako stratę czasu. Oceniają sukces projektu BI na podstawie szybkości, z jaką jest dostarczany, a nie jakości jego rezultatów. W rezultacie organizacje często tworzą hurtownie danych typu „superpipe" i wypełniają je w stylu „ssij i wrzuć" tymi samymi danymi, które mają w plikach źródłowych i źródłowych bazach danych, kopiując w ten sposób wszystkie istniejące uszkodzenia danych do nowego środowiska wspomagania decyzji BI. Zamiast eliminować istniejące problemy z danymi, po prostu je spotęgowali — teraz trzeba utrzymywać dodatkowe nadmiarowe i niespójne docelowe bazy danych i aplikacje BI. Spośród wszystkich 16 kroków przedstawionych w Business Intelligence Roadmap, Krok 5, Analiza Danych, jest najbardziej krytycznym krokiem między organizacjami. Ten krok jest głównym wyróżnikiem między tradycyjnym podejściem do tworzenia systemów a podejściem do rozwoju międzyorganizacyjnego. Działania związane z analizą danych zorientowaną na biznes zmuszają konsumentów informacji i właścicieli danych do zrekonstruowania spojrzenia międzyorganizacyjnego i oczyszczenia ich kosztownego chaosu danych, nie tylko w środowisku wspomagania decyzji BI, ale także w ich systemach operacyjnych. To wszystko są warunki wstępne dla poprawy zdolności kadry kierowniczej do podejmowania decyzji. Bez tego kroku budujesz kolejny tradycyjny system wspomagania decyzji, a nie rozwiązanie BI.