Nie zmieniaj danych

„Powiedz mi” – zaczął Wendell. „Czy marzysz czasem o pracy?” Cezar zaśmiał się. "Marzenie? No cóż, nie ostatnio, nie. „Po prostu widzę różnego rodzaju anomalie w tym zbiorze danych, który dostałem od Belindy, i po prostu nie mogę sobie z tym poradzić. Ostatnio nie dawało mi to spokoju, a ostatniej nocy nawet śniłem o tym projekcie. Cezar przesunął kursor myszy, żeby ekran nie zniknął. „To oznacza, że pracujesz zbyt ciężko”. Wendell westchnął. „Może mógłbyś mi pomóc? Może uda ci się coś wymyślić. Cezar przeciągnął się. „Nie nazywają mnie Rain Manem bez powodu”. Wendell musiał się uśmiechnąć. „Tak?” "NIE. W porządku, więc co masz?” Wendell otworzył dokument pokazujący kilka wykresów. „Mike poprosił mnie, abym lepiej zaznajomił się z naszymi modelami danych, więc koreluję informacje na lewo i prawo, ale niektóre rzeczy po prostu nie mają dla mnie sensu. Nie jestem w 100% pewien, czy moje metody są błędne lub czy coś jest nie tak z liczbami. Cezar patrzył na wykresy z kciukiem przyciśniętym do brody. "Na przykład?" Wendell wskazał na ekran laptopa. „Mike chce wiedzieć, czy istnieją jakieś powiązania między adresami domowymi naszych klientów a lokalizacjami sklepów, prawda. Ale wtedy dostaję tutaj coś takiego jak ten przykład. Mam osobę mieszkającą w Victorii w Teksasie, ale dane pokazują, że robiła zakupy w sklepie w mieście Victoria w kraju Grenada. Cezar pstryknął palcami i odchylił się do tyłu. „Och, słyszymy to cały czas. Właściwie, to Belinda naprowadziła mnie na właściwą drogę. Miałem podobne problemy i nie byłem pewien, co jest źródłem anomalii. Następnie Belinda pomogła mi z danymi klientów, przeprowadziliśmy kilka dodatkowych korelacji i w końcu to rozpracowałem. Musisz po prostu zadać inny typ zapytań.” Wendell zapisał kilka notatek. „OK, to ma sens”. „Nie czuj się źle, bracie. Te niespójności zdarzają się dość często i nie mogłeś o tym wiedzieć. Pomogłem naszemu zespołowi ds. analityki danych w pisaniu nowych raportów, które ignorują te oczywiste literówki podczas generowania tygodników. Po prostu wyklucz wszelkie dopasowania, w których kraj klienta nie odpowiada krajowi sklepu. Wendell skinął głową. „I to też nie jest żadna nowość. Mike skarżył się, że jego raporty dla kierownictwa wyglądają źle z powodu tych dziwnych błędów. Zostałem wciągnięty, żeby dowiedzieć się, jak to naprawić. Wendell potarł knykcie. "Bez problemu. To właściwie jest dobre, ponieważ pozwala mi lepiej zrozumieć, w jaki sposób analizujemy dane i jak interpretujemy uzyskiwane wyniki. Ale przeglądanie tych wszystkich rzeczy wydaje się nudne. Czy istnieje zautomatyzowany sposób sprawdzania danych pod kątem anomalii?” Cezar uśmiechnął się. "Oczywiście. Właściwie utrzymuję skrypt powłoki, który przechodzi przez bazę danych i oczyszcza te anomalie. Uruchamiam go za każdym razem, gdy muszę wygenerować jakikolwiek raport. Wyślę Ci link. Korzystaj z niego, jak chcesz, a jeśli napotkasz jakieś nowe anomalie, daj mi znać, a zaktualizuję narzędzie. "Dzięki." Wendell zapisał kolejną notatkę. „Chciałem też porozmawiać z tobą o monitorowaniu i logowaniu, ponieważ to takie twoje dziecko”. Cezar zaśmiał się. „Dumny ojciec”. – Czy możesz mi to pokrótce omówić? Cezar pochylił się do przodu. "Tak. Najpierw skonfigurowałem mnóstwo rzeczy, które ułatwią mi życie. Wiesz jak to jest z logami. Łatwo wpadasz w hałas, przez co tracisz pewne rzeczy. Mam skrypt, który wysyła mi e-mail tylko wtedy, gdy pojawi się jakikolwiek krytyczny alert. Następnie wykorzystuję dashboardy, gdy system zgłasza błąd i potrzebuję szybkiego podglądu podstawowych wskaźników jego wydajności. Wreszcie mam inny skrypt, który wysyła mi e-mail z codziennym podsumowaniem, dzięki czemu mogę szukać korelacji z problemami występującymi w całym środowisku”. „Brzmi jak całkiem solidna konfiguracja” – skomentował Wendell. „No cóż, tak musi być. Inaczej byłbym ciągle przeciążony. Muszę działać wydajnie i usuwać hałas, żeby nie karmić ciebie i Mike’a bezużytecznymi danymi. Wendell przejrzał swoje notatki z ostatnich kilku tygodni. „Korzystamy z centralnego rejestrowania w całej naszej infrastrukturze?” "Tak i nie. Wysyłamy wszystkie dzienniki serwera do centralnego węzła gromadzenia danych za pośrednictwem sieci, a następnie codziennie zmieniamy dzienniki. Następnie generujemy analizy z dzienników, dzięki czemu mamy wykresy wydajności systemu i statystyk użytkowania. Nie prowadzimy

żadnego zdalnego logowania w środowiskach testowych i deweloperskich. Pokażę ci raport. Spędzili kilka minut na przeglądaniu danych i przeglądaniu wykresów. Wendell zauważył coś dziwnego na jednej z działek. Nastąpił ogromny wzrost użycia, który trwał tylko około minuty. Wyglądało na to, że aktywność na stronie wzrosła o dobre 800%. „Co tam robiliśmy? Czy w tym czasie miało miejsce wprowadzenie jakiegoś produktu na rynek lub coś interesującego?" Cezar machnął ręką. "O nie. Przepraszam. To problem z logowaniem. Czasami wpisy w dzienniku są zapisywane wielokrotnie, po czym pojawiają się dziwne skoki wartości. Ale to w porządku. Po prostu odfiltrowuję duplikaty i zachowuję jeden oryginalny wpis dla tego znacznika czasu i to wszystko. Dostajemy normalne dane i możemy mieć czysty raport. Oszczędza to wielu bólów głowy, gdy programiści i marketingowcy wpadają w panikę bez powodu, ponieważ nasz ruch nagle rośnie i nie mają pojęcia dlaczego. Wendell skinął głową, ale nie był pewien, czy to najlepsze podejście do problemu. „Czy wiemy, dlaczego wpisy w dzienniku są zapisywane wielokrotnie?" Cezar zrobił smutną minę. – Nie ma czasu na rozwiązywanie tego problemu, kolego. Dotrzemy tam. Właściwie Daniel obiecał mi w tym pomóc. Ma duże doświadczenie w pracy z narzędziami do gromadzenia i raportowania danych i pomoże mi uzyskać dostęp do generatora logów. Zwykle nie można tego uzyskać w wersji narzędzia, której używamy, ale Daniel zna się na tym i myśli, że wie, jak uzyskać dostęp do dzienników oprogramowania i dowiedzieć się, czy nie występuje tam problem. Wendell ponownie skinął głową. Cóż, warto byłoby usłyszeć od Daniela, co miał do powiedzenia na temat problemu z duplikatami rejestrowania.

Jeśli kontrolujesz dane, kontrolujesz narrację

Jak wspomnieliśmy , istnieje kilka możliwych permutacji w procesie decyzyjnym. Można podejmować dobre decyzje w oparciu o dobre dane, złe decyzje w oparciu o złe dane, a także złe decyzje wynikające z niezrozumienia dostępnych informacji. Szczęście oczywiście istnieje, ale w świecie biznesu jest to towar deficytowy, zwłaszcza gdy obsługujesz klientów. Trudno jest kontrolować wszystkie czynniki wpływające na Twoją pracę. Rzadko kiedy będziesz miał luksus dyktowania każdego etapu procesu. Niezależnie od tego, czy chodzi o ludzi, strony zewnętrzne zaangażowane w pracę, czas połączenia różnych elementów, czy nawet podstawowe zasady fizyki, będziesz mieć do czynienia z całą masą prawdopodobieństw. Jest jednak jedna rzecz, którą możesz kontrolować – a są to dane. W Rozdziale 2 (Szanuj prywatność) mówiliśmy o eksplozji informacji. Rzeczywiście, świat IT to jedna gigantyczna baza danych, zawierająca miliardy miliardów rekordów danych wszelkiego rodzaju, stanowiących sumę naszego doświadczenia komputerowego, głównie z ostatnich 30 lat. Dane te stanowią podstawę procesu decyzyjnego w firmach i napędzają strategię, inwestycje i rozwój produktów. Wykorzystanie wszystkich tych danych może zdziałać cuda. W ciągu ostatnich kilku dekad poczyniliśmy ogromne postępy w postępie technologicznym i naukowym. Udało nam się zmapować ludzkie DNA, odkryliśmy bozon Higgsa i umożliwiliśmy milionom ludzi rozmowę w czasie rzeczywistym ze swoimi przyjaciółmi na drugim końcu planety. Wykorzystujemy dane do przewidywania trendów demograficznych i gospodarczych, a giełdy działają w oparciu o sprytne algorytmy, które sprawiają, że od czasu do czasu bije nam serce. Każdy aspekt naszego życia jest podyktowany masowym wykorzystaniem danych, a wykorzystanie to podwaja się co dwa lata. Te same dane mogą również siać spustoszenie, jeśli zostaną niewłaściwie wykorzystane. Administratorzy systemów, programiści i technicy IT znajdują się w dość niepewnej sytuacji. Często sami nie są opiekunami danych, ale ich dostęp do uprzywilejowanych systemów i zasobów umożliwia im dostęp do wrażliwych informacji mających znaczący wpływ na działalność biznesową. Ich obowiązki często wymagają interakcji z tymi informacjami. A to prowadzi nas do aspektu kontroli. Administratorzy systemów często siedzą na węzłach przesyłu danych i mają możliwość przekształcania surowych strumieni informacji wpływających do centrów danych w wyższe formy logiki i porządku. W rzeczywistości czasami będą do tego zobowiązani, zwłaszcza podczas zbierania takich danych, jak wskaźniki wydajności systemu, dzienniki bezpieczeństwa, a czasami nawet dane klientów. Rzeczywiście, administratorzy systemu niekoniecznie zawsze będą mieli możliwość

dyktowania, które dane będą gromadzone oraz dlaczego i w jaki sposób będą one przechowywane w długoterminowych archiwach na całym świecie. Mają jednak pełną kontrolę nad sposobem uzyskiwania dostępu do tych danych i ich wykorzystywania. Ale to nie dotyczy tylko administratorów systemów. Dotyczy to każdego, kto pracuje z danymi (i nie muszą to być dane ściśle techniczne). Ma to zastosowanie w akademii, wśród badaczy, a także menedżerów i twórców oprogramowania obsługujących logi aplikacji i klientów. Dane same w sobie nie mają wartości, dopóki nie zostaną przetworzone i przeanalizowane. To właśnie ten krok przekształca surowe liczby i litery w znaczącą, potężną logikę, której możemy następnie użyć, aby uczynić nasze życie lepszym, bezpieczniejszym i mądrzejszym. To właśnie ten krok reguluje proces podejmowania decyzji. Złe dane sprawią, że nawet osoby posiadające dużą wiedzę w danej dziedzinie będą podejmować błędne decyzje. Manipulacja danymi jest nieodłączną częścią wykorzystania danych od zarania dziejów ludzkości. Jednak słowo „manipulacja" nie jest już używane do wskazania wysokich umiejętności, poziomu lub użycia. Obecnie ma to głównie negatywne konotacje i jest kojarzone ze zmianami danych mającymi na celu celowe wypaczenie wyników w kierunku preferowanego wniosku. Powodem jest to, że firmy, przedsiębiorstwa, organizacje, a także osoby fizyczne wykorzystują – i zmieniają – dane, aby osiągnąć swoje cele; zmienić dane, zmienić narrację. Jednak niemal zbyt łatwo jest odrzucić celowe manipulowanie danymi. Prawdziwy problem polega na tym, że ludzie często nieumyślnie, nawet niewinnie, dokonują zmian w danych, ponieważ nie w pełni rozumieją lub nie doceniają ryzyka związanego z takimi działaniami, ponieważ czują się pod presją przedstawienia „różowego" lub „zawyżonego" obrazu oczekiwanych rezultatów , a także dlatego, że uważają, że zmiany danych są uzasadnione, jeśli istnieje dobry powód, aby je wprowadzić. Jako osoba mająca dostęp do danych, niezależnie od tego, czy jesteś administratorem systemu, badaczem czy menedżerem, najprawdopodobniej znajdziesz się w drugim scenariuszu: zostaniesz poproszony o zebranie i przeanalizowanie niektórych danych. Możesz odkryć, że masz za dużo informacji i że Twoja baza danych nie jest w stanie obsłużyć woluminów wejściowych, więc możesz poczuć pokusę odrzucenia części danych. Możesz też odfiltrować określone rekordy, ponieważ uważasz, że są one nieprzydatne. Możesz też otrzymać wyniki testu, które nie dają jednoznacznych wniosków, jakich się spodziewałeś, i będziesz musiał powtórzyć wszystko od nowa. Wykonując którąkolwiek z tych czynności, możesz postawić się w nieetycznej sytuacji.

Nie zmieniaj danych

Jeśli zmienisz dane, skutecznie zmienisz migawkę rzeczywistości w momencie gromadzenia punktów danych. Stwarza to wypaczony obraz sytuacji i prowadzi do nieprawidłowych wyników, nawet jeśli metody analizy są doskonałe. Zmiana danych ma wiele długoterminowych konsekwencji. Może to być wręcz nielegalne. Może to podważyć zaufanie, jakim obdarzyli Cię inni, co może w przyszłości doprowadzić do zlekceważenia lub zdyskredytowania Twoich wyników. Spowoduje to, że Ty lub osoby, których wyniki dotyczą, podejmiecie kosztowne, a może nawet niebezpieczne decyzje w oparciu o wyniki uzyskane na podstawie zmanipulowanych, wypaczonych lub częściowych informacji. W niektórych sytuacjach może to być zawstydzające. W innych, może to kosztować życie ludzi. Jednym z przykładów zmiany danych w firmie, która spowodowała znaczne szkody, było wycofanie poduszki powietrznej Takata. Niedokładne dane dotyczące wyników testów przyczyniły się do zgonów, obrażeń i wycofania dziesiątek milionów pojazdów.

Wyniki są wynikami, dobrymi lub złymi

Nie da się przecenić tego punktu. Bez względu na to, jak zły obraz przedstawiają dane, nie należy ich zmieniać. Firmy często tworzą niesprawiedliwe oczekiwania wobec swoich pracowników, żądając pozytywnych wyników projektów roboczych lub prosząc o pompatyczne liczby. Z biegiem czasu powoduje to, że ludzie próbują narzucać swoją pracę i wyniki zgodnie z oczekiwaniami. To z kolei

prowadzi do ignorowania lub ukrywania złych wyników, co na dłuższą metę może spowodować znaczne szkody. Idealnie byłoby, gdyby każdy projekt dający wyniki ilościowe miał z góry określone kryteria sukcesu. Powinien także opierać się na hipotezie, którą można testować w sposób powtarzalny. Jeśli wyniki testu wykazują dane, które nie odpowiadają kryteriom sukcesu, oznacza to, że albo pierwotna hipoteza była błędna, albo kryteria sukcesu zostały ustawione nieprawidłowo. Nie należy zmieniać danych w celu dopasowania któregokolwiek z nich. Złe wyniki mają zalety; mogą pomóc Ci zdecydować, czego nie robić, co jest równie ważne, jak uzyskanie dobrych wyników. Być może za pierwszym razem nie uda Ci się stworzyć udanego modelu swojego produktu. Iteracja po kilku nieudanych modelach zwykle prowadzi do bardziej udanej wersji końcowej produktu. Nie różni się to od procesu ewolucyjnego w przyrodzie. Istnieją oczywiście metody naukowe, które mogą pomóc zarówno ulepszyć zaproponowany model pracy, jak i dokładniej przeanalizować dostępne dane.

Użyj automatyzacji i filtrowania, aby zrozumieć dane

Rozmowa Wendella i Caesara dotyka wielu delikatnych punktów, głównie dlatego, że nie ma prostego, wyraźnego podziału na obszary etyczne i nieetyczne. Trochę jak dane Wendella. Nie jestem w 100% pewien, czy moje metody są błędne lub czy coś jest nie tak z liczbami. Wendell nie jest pewien, czy jego analiza danych jest prawidłowa, dlatego zdecydował się zasięgnąć porady kolegi, który ma większe doświadczenie w tej dziedzinie. To zawsze rozsądne podejście. Praca z członkami zespołu może pomóc odkryć problemy i niespójności w metodach i procedurach pracy, a także uwypuklić dodatkowe aspekty, które mogą zwiększyć niezawodność analizy danych. Co więcej, Wendell próbuje także zrozumieć, czy i gdzie w jego pracy może tkwić błąd. Systematyczne podejście jest zawsze dobrą rzeczą, ponieważ pozwala izolować problemy. Czasami problemy mogą wynikać z wielu wzajemnie powiązanych przyczyn, a ich rozwiązywanie jest dość trudne i czasochłonne. Zawsze wskazane jest, aby spróbować zredukować problem do minimalnego zestawu istotnych czynników. Oszczędza to czas i wysiłek oraz ułatwia analizę danych.

Czy istnieje zautomatyzowany sposób sprawdzania danych pod kątem anomalii?

Wendell próbuje zastosować niektóre lekcje, których nauczył się od Alexa i Belindy. Poszukuje sposobów na automatyzację filtrowania danych (niezależnie od tego, czy samo takie działanie jest właściwym wyborem). Ogólnie rzecz biorąc, automatyzacja może pomóc w usprawnieniu procesów, ograniczeniu błędów i ułatwieniu pracy.

Pomogłem naszemu zespołowi ds. analityki danych w pisaniu nowych raportów, które ignorują te oczywiste literówki podczas generowania tygodników.

Godne pochwały są także metody pracy Cezara. Bardzo interesuje go również automatyzacja i angażuje się w pomaganie innym zespołom w usprawnianiu ich pracy. Współpraca między zespołami jest niezbędna, szczególnie w obszarach bogatych w dane w branży IT, ponieważ podział na segmenty prowadzi do niepotrzebnego powielania wysiłków i rozwiązań. Angażując się, Cezar ma również możliwość przedstawienia bezstronnego spojrzenia z zewnątrz na problem, przed którym mogą stanąć inne zespoły, bez możliwości jego skutecznego rozwiązania, ponieważ mogłyby być zbyt zaangażowane emocjonalnie, aby uzyskać bardziej filozoficzne zrozumienie problemu . Cezar stara się działać metodycznie i śledzić swoje działania. Przejrzystość jego pracy jest zgodna z wnioskami, jakie wyciągnęliśmy z rozdziału 2 (Szanuj prywatność) i może pomóc w rozwiązaniu problemów, jeśli wystąpią podczas analizy danych. Co więcej, Cezar zasugerował, aby Wendell wracał do niego w celu uzyskania aktualizacji narzędzia. Scentralizowana funkcja raportowania danych jest pomocna w unikaniu powielania, niespójności i nieetycznego dostępu do danych.

Mam skrypt, który wysyła mi e-mail tylko wtedy, gdy pojawi się jakikolwiek krytyczny alert.

Oprócz automatyzacji naleganie Cezara na używanie skryptów do sortowania i filtrowania danych ma inne zalety. Mianowicie, filtrując informacje do segmentów na podstawie ich ważności, priorytetu i ważności, może skupić się najpierw na analizie ważnych danych. W ten sposób Caesar może uniknąć przeciążenia danymi, co jest powszechne w wielu systemach monitorowania i ostrzegania. Wysyłamy wszystkie dzienniki serwera do centralnego węzła gromadzenia danych za pośrednictwem sieci, a następnie codziennie zmieniamy dzienniki. To kolejny przykład dobrych praktyk w zakresie przechowywania danych i rotacji dzienników. Omówiliśmy je już w powiązaniu z kontami użytkowników w Rozdziale 1 (Oddzielne role) i mają one pełne zastosowanie we wszystkich aspektach technologii informatycznych i administrowania systemami. Dobrze udokumentowany proces gromadzenia, przetwarzania i przechowywania danych minimalizuje ryzyko nieetycznego dostępu, nawet przypadkowego. Co więcej, wysyłanie logów do centralnego węzła gromadzenia odbywa się bez filtrowania. Oznacza to, że kopiowane są wszystkie wpisy dziennika, w tym potencjalnie nieprawidłowe dane, które mogą być przydatne do szeregu różnych analiz, w tym do celów bezpieczeństwa, kryminalistyki, wzorców aktywności klientów, wykorzystania i obciążenia zasobów i nie tylko.

Usunięcie złych danych usuwa dobre informacje

Cezar wprawdzie wprowadził kilka pożytecznych, etycznych praktyk, ale dopuścił się także szeregu nieetycznych naruszeń. Rola Wendella w tej sytuacji jest delikatna. Sam nie dokonał żadnych zmian, ale jest wtajemniczony w sytuację, w której uzyskano dostęp do danych (być może bez odpowiedniego upoważnienia) i je zmieniono, tworząc wypaczony obraz operacyjnego środowiska biznesowego. Czasami trudno jest znaleźć równowagę pomiędzy donosem na kolegów a próbą ich skorygowania, ponieważ takie działania mogą zostać źle zinterpretowane przez zainteresowane osoby. Istnieją jednak pewne jasne wytyczne etyczne mające zastosowanie do historii Wendella i Cezara.

Po prostu wyklucz wszelkie dopasowania, w których kraj klienta nie odpowiada krajowi sklepu. Po prostu odfiltrowuję duplikaty i zachowuję jeden oryginalny wpis dla tego znacznika czasu i to wszystko.

Widzimy tu wspólny motyw. Cezara nie zadowala zło

punktów danych i podjął kroki w celu usunięcia takich danych ze swoich raportów. Chociaż Cezar jest semantycznie poprawny, usunięcie „złych” punktów danych jedynie maskuje podstawowy problem. Nie wiemy, dlaczego rekordy bazy danych zawierają te fałszywe wpisy danych. Usuwając je, Cezar eliminuje także możliwość rozwiązania problemu i zrozumienia problemu. Długoterminowy efekt jego zmian może zamaskować dodatkowe problemy – usuwając wpisy danych, Cezar usuwa również widoczność anomalii, usuwając w ten sposób szansę, aby ktoś inny dostrzegł problem i ewentualnie naprawił jego źródło, zapewnił zasoby rozwiązać problem lub odpowiednio ustalić priorytety wysiłków mających na celu zrozumienie anomalii. Na przykład, jeśli istnieje duża liczba „złych” wpisów danych, usunięcie ich może skutkować zaniżeniem wartości raportu, co można interpretować na wiele sposobów. Leady firmy mogą mieć fałszywe poczucie samozadowolenia, nie podejrzewając, że pewna część danych klientów została błędnie sklasyfikowana w bazie danych lub że mogą to być w rzeczywistości problemy związane z doświadczeniami klientów, co może bezpośrednio wpłynąć na wyniki finansowe.

Jeśli zmienisz dane, utracisz kluczowe wskaźniki.

Niedostateczne zgłaszanie może również zaszkodzić przyszłemu planowaniu wydajności. Możliwe jest, że zespoły IT planują aktualizacje swojej infrastruktury sprzętowej w oparciu o bieżące i prognozowane zapotrzebowanie na wykorzystanie zasobów, takich jak serwery i baza danych. Korzystanie z wypaczonych wskaźników (po oczyszczeniu przez zespół) może wprowadzić w błąd i wyprowadzić błędne prognozy; wykorzystanie powyżej optymalnych parametrów, ponieważ rzeczywiste zużycie

będzie wyższe niż zużycie raportowane; i pogorszenie doświadczenia użytkownika końcowego, co ponownie może mieć wpływ na działalność biznesową.

Historia z okopów IT: Wydaje się, że jest więcej rzeczywistych przykładów nadmiernego zgłaszania wykorzystania niż zaniżania. W świecie administracji systemami skutkuje to zwykle nadmiernym zakupem zasobów obliczeniowych. W jednym przykładzie kierownictwo i finanse przestały ufać danym, co spowodowało brak zasobów w późniejszych latach. W innym przykładzie dane dotyczące wykorzystania zostały sprawdzone po miesiącach bezczynności serwerów; osoba odpowiedzialna za dane została przeniesiona na nową „szansę".

Działalność Cezara narusza zarówno drugie (Szanuj prywatność), jak i trzecie przykazanie (Nie zmieniaj danych) omawiane tutaj. Potrzebowałby pozwolenia na dostęp do danych i ich analizę, a następnie zwróciłby się konkretnie do podmiotu przechowującego dane lub właściciela o pozwolenie na dokonanie zmian. Wendell powinien wskazać je Cezarowi. Powinien zaproponować alternatywne podejście, które pozwala zarówno na zachowanie oryginalnych danych bez żadnych zmian, jak i na osobne wykonanie dodatkowych analiz.

Właściwie utrzymuję skrypt powłoki, który przegląda dane i oczyszcza te anomalie. Uruchamiam go za każdym razem, gdy muszę wygenerować jakikolwiek raport. Wyślę Ci link. Korzystaj z niego, jak chcesz, a jeśli napotkasz jakieś nowe anomalie, daj mi znać, a zaktualizuję narzędzie.

Widzieliśmy ten cytat już w poprzednim tekście i powód

jak tu napisano, jest to, że ma to również kilka nieetycznych konotacji. Skrypt Cezara nie jest częścią udokumentowanego procesu dostępu do danych. Jeśli w skrypcie występują problemy, mają one wpływ na raporty, a osoby je czytające nie są świadome, że dane zostały zmienione (tzn. integralność danych została naruszona podczas przesyłania). Cezar wplątuje również Wendella, sugerując, aby zrobił to samo, co utrudnia zrozumienie problemów w logice danych i późniejsze rozwiązanie.

Jeśli podejrzewasz błąd podczas wprowadzania danych, przejdź do źródła

Czasami będziesz przekonany, że wyniki są błędne – a nie tylko złe (lub wyjątkowo dobre). Jeśli uważasz, że Twoja hipoteza jest poprawna, a w Twoich modelach lub analizie nie ma błędów, możesz podejrzewać dane. Dane powinny być nienaruszalne – ale to nie znaczy, że powinieneś mieć dostęp do nich lub ich interpretacji za dobrą monetę. Rzeczywiście, jeśli uważasz, że oryginalne, surowe dane są z jakiegoś powodu nieprawidłowe, powinieneś wrócić do źródła. W niektórych sytuacjach mogą to być odczyty czujników z urządzenia IoT, wartości wydajności z serwera renderującego grafikę, statystyki witryny internetowej lub dowolne inne dane. Jeśli uważasz, że odczyty są nieprawidłowe, musisz zrozumieć proces pobierania i gromadzenia danych oraz wyeliminować wszelkie błędy. Na przykład czujnik może zostać błędnie skalibrowany lub statystyki witryny internetowej mogą opierać się na wektorze danych, który tworzy fałszywe wartości. Jeśli zidentyfikujesz błąd (i nie dlatego, że chcesz poprawić wyniki), możesz wprowadzić zmiany w sposobie gromadzenia danych, ale po ich zebraniu nie należy ich zmieniać.

Poproś o pomoc w znalezieniu źródła

Wendell i Cezar rzeczywiście starali się być sumienni w swojej pracy. Nie pracowali ślepo ze swoimi zbiorami danych, przeprowadzali analizy i starali się skorelować informacje przed podjęciem kolejnych kroków.

Mam osobę mieszkającą w Victorii w Teksasie, ale dane pokazują, że robiła zakupy w sklepie w mieście Victoria w kraju Grenada.

Wendell miał rozsądne podejście do swojej pracy. Postanowił nie wprowadzać żadnych zmian i zamiast tego skonsultować się z kolegą. Przeglądał także dane dotyczące zakupów i adres klienta, aby zrozumieć rozbieżność w swojej analizie. Co więcej, Wendell powinien nadal trzymać się tego podejścia, jeśli chodzi o dane Caesara.

Następnie Belinda pomogła mi z danymi klientów i pobiegliśmy kilka dodatkowych korelacji i w końcu to rozgryzłem.

Podobnie jak Wendell, Cezar również włożył wiele pracy, próbując zrozumieć, dlaczego otrzymywane przez niego wyniki wydawały się nieprawidłowe. Poprosił kolegę o pomoc, a on pracował z dodatkowymi zbiorami danych, aby zrozumieć i skorelować wyniki. Zawsze przydatne jest porównanie (i zbadanie) wielu zestawów wyników, aby zrozumieć, czy występują jakieś anomalie, zwłaszcza jeśli wiadomo, że konkretny zestaw jest prawidłowy i może być stosowany jako standard lub punkt odniesienia. Rzeczywiście, jeśli występują anomalie, możliwe jest zrozumienie pierwotnej przyczyny problemu lub przynajmniej wyizolowanie etapu procesu analizy danych komponentu w łańcuchu dostępu do danych, w którym występuje problem.

Łatwe prowadzi do problemów

W całej historii widzieliśmy kilka przypadków, w których wydawało się, że wystąpiły anomalie w gromadzeniu i analizie danych. Najrozsądniejszym podejściem jest rozwiązanie problemu u źródła – czyli na etapie generowania i zapisywania danych. Może to ujawnić problemy z logiką biznesową aplikacji lub nieprawidłową konfigurację systemu. Mogą istnieć również inne przyczyny nieprawidłowych punktów danych, z których żadnej nie można zidentyfikować ani rozwiązać na późniejszych etapach procesu. Nie ma czasu na rozwiązywanie tego problemu, kolego. To odwieczna wymówka w branży IT. Dość często ludzie mają duże obciążenie pracą i zmieniają się priorytety, a są pewne rzeczy, które trudno będzie uwzględnić w napiętych, napiętych harmonogramach pracy. Wendell ma właściwe podejście, próbując przeanalizować problem u źródła, i powinien nalegać, aby robiono to w zorganizowany i zaplanowany sposób. Podobnie jak fizyczne prawa zachowania energii, spinu czy ładunku elektrycznego, istnieje również uniwersalny odpowiednik IT – zasada zachowania problemów. Jeśli nie zostaną rozwiązane, problemy nie znikną. Ogólnie rzecz biorąc, jest to zdecydowanie trudny scenariusz. Dość często istnieje kompromis między rozwiązywaniem problemów a otwieraniem puszki Pandory. Czasami ludzie mogą po prostu szukać „łatwiejszego" wyjścia. Ale najważniejsze jest, aby niczego nie ukrywać. Skoki danych mogą wywołać panikę w dziale marketingu, co zwróci uwagę kierownictwa na zakup oprogramowania, które pomoże rozwiązać problem i określić pierwotną przyczynę duplikatów. W końcu niezmienianie danych jest rzeczą słuszną. Ponadto te skoki mogą wynikać z włamania się do systemu w celu kradzieży danych.

Zwykle nie jest to możliwe w wersji narzędzia, której używamy, ale Daniel zna się na tym i myśli, że wie, jak uzyskać dostęp do dzienników oprogramowania i dowiedzieć się, czy nie występuje w nich problem.

Jest tu wiele kwestii. Podejście Daniela omija znane, ustalone procesy, podobnie jak to widzieliśmy w przypadku Alexa. Jest to naruszenie szóstego przykazania (Nie będziesz chodzić tam, gdzie cię nie chcą). Daniel i Cezar potrzebują autoryzacji, aby uzyskać dostęp do plików w generatorze logów. Potrzebują pozwolenia właściciela danych, którym w tym przypadku może być dostawca urządzenia lub oprogramowania do generowania logów. Mogą również występować kwestie związane z prywatnością, ponieważ mogą dotyczyć danych wrażliwych i wymagają one osobnej uwagi.

Metoda naukowa

Analiza danych jest czynnością bardzo złożoną i często czasochłonną. Co więcej, wymaga stosunkowo głębokiego zrozumienia metod i narzędzi statystycznych, co nie jest ani trywialne, ani powszechne w szerszej branży IT. Chociaż większość administratorów systemów ma dobrą wiedzę o systemach, zazwyczaj nie są oni dobrze zaznajomieni z analizą matematyczną. Co więcej, często wychodzi się z założenia, że statystyka jest domeną nauki i badań i że w przyziemnym świecie operacji IT nie ma miejsca na takie podejście. Czasami analizą danych zajmują się dedykowane zespoły analityki biznesowej. A mimo to obserwujemy eksplozję danych na wszystkich poziomach świata IT. Chociaż pozornie przyziemne wpisy w dziennikach systemowych nie wyglądają ekscytująco ani odkrywczo, gdy są sprawdzane sporadycznie, często mogą ujawnić ważne prawidłowości, gdy są analizowane na dużą skalę. Jednak nawet w przypadku mniejszych, izolowanych projektów posiadanie odpowiedniej metodologii i rygorystycznego rygoru analitycznego jest dość ważne.

Integralność danych

Jeśli musisz analizować dane, niezależnie od ich rozmiaru i postrzeganej ważności, jest dziesięć rzeczy, które musisz zrobić przed, w trakcie i po analizie.

Źródła danych

• Źródło danych musi być wyraźnie zidentyfikowane – musisz być w stanie określić ostateczne pochodzenie nieedytowanych, surowych wartości danych. W niektórych przypadkach dane wyjściowe mogą być już przetwarzane przez systemy „czarnych skrzynek" (takich jak czujniki), a dostęp do kolejnych wyników będziesz mieć jedynie jako swoje dane.

• Źródło danych musi być chronione – Twoje środowisko powinno posiadać mechanizmy zapobiegające samowolnym zmianom danych. Mogą one mieć formę uprawnień do dzienników i baz danych, szyfrowania lub alertów dotyczących dostępu do nieprzetworzonych danych i ich zmian. Urządzenia typu Write Once, Read Many (WORM) mogą być używane do wrażliwych i ważnych danych. W takich urządzeniach lub systemach danych nie można zmienić po ich zapisaniu.

• Informacje i przypuszczenia pochodne muszą być przechowywane oddzielnie od danych źródłowych – jeśli zachodzi potrzeba przetwarzania i analizowania danych, należy to zrobić bez modyfikowania w jakikolwiek sposób oryginalnego zbioru danych, ewentualnie pracując z tymczasową kopią oryginalnych wartości danych . W ten sposób inne osoby również mogą pracować z nieedytowanymi danymi. Co więcej, nie będą oni narażeni na Twoją analizę, co mogłoby wypaczać lub wypaczać ich interpretację i zrozumienie wyników.

Automatyzacja i audyt

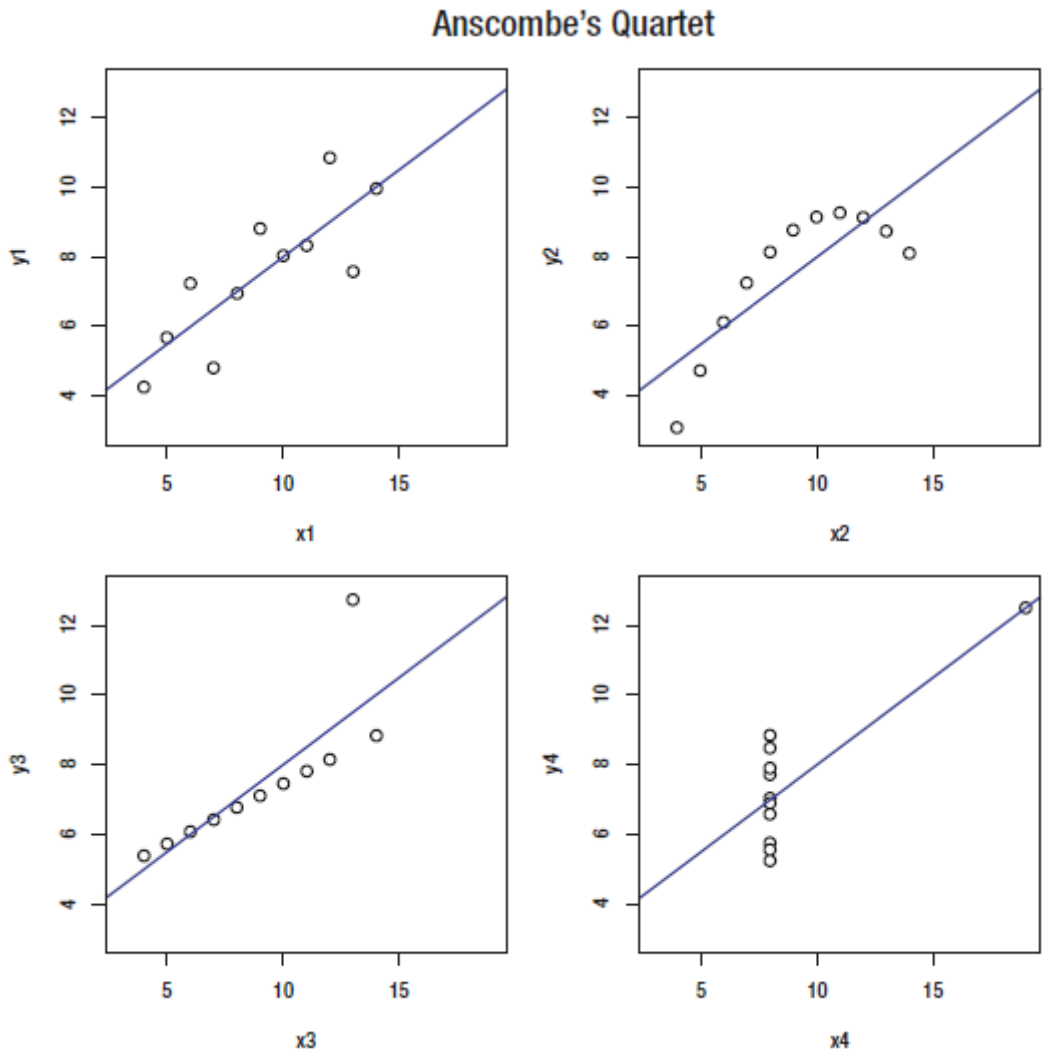
• Dane należy kontrolować pod kątem integralności, a nie dokładności – Twoje systemy powinny wykryć, czy występują problemy z gromadzeniem danych lub strukturą zbiorów danych, a nie to, czy wartości mają sens lub są zgodne z Twoimi oczekiwaniami. Analiza wyników wykaże, czy występują inne problemy z Twoimi danymi. Jeżeli dane zostaną uznane za niedokładne, należy zmienić proces gromadzenia danych na następny cykl. Pamiętaj, aby zarejestrować i udokumentować zmianę. Co więcej, automatyzacja pomaga w powtarzalnych zadaniach i ogranicza błędy ludzkie.

Hipoteza, metodologia i błędy

• Zanim zostanie przeprowadzona jakakolwiek analiza danych, musi istnieć dobrze sformułowana, możliwa do przetestowania propozycja. Nie należy po prostu ślepo szukać wzorców w danych, ponieważ zawsze jakieś będą. Korelacja nie oznacza związku przyczynowego. Co więcej, niektóre interpretacje danych najprawdopodobniej będą miały zerową wartość biznesową.

• Nie ma dobrych ani złych wyników – jeśli Twoja hipoteza okaże się błędna, jest to słuszny wniosek i nie powinieneś zmieniać danych ani metod, aby dostosować wyniki do swoich upodobań.

• Analiza danych wymaga świadomości tego, jak dane wyglądają – ślepe stosowanie metod statystycznych może prowadzić do rażących błędów. Świetnym przykładem ilustrującym to zjawisko jest Kwartet Anscombe'a3, cztery zbiory danych, które mają prawie identyczne proste statystyki, ale wyglądają zupełnie inaczej na wykresie, jak pokazano na rysunku



• Metody muszą być powtarzalne – jeśli oddasz swoją pracę komuś innemu, powinna ona być w stanie wykonać tę samą pracę i uzyskać takie same wyniki jak Ty.

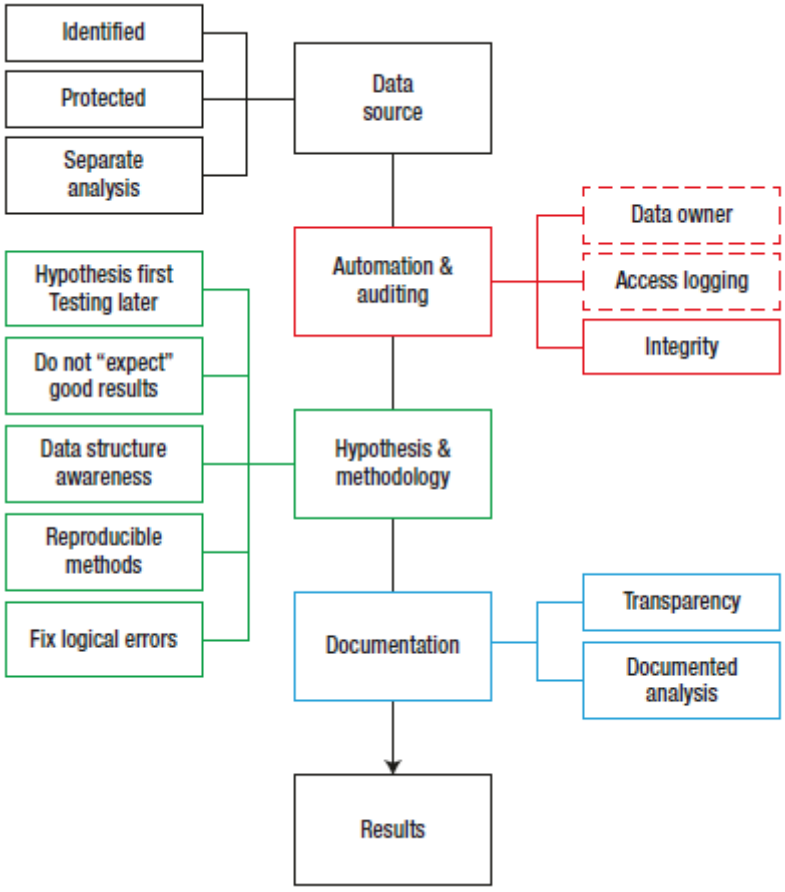
• Błędy logiczne i problemy należy naprawić przed jakąkolwiek dodatkową analizą – zawsze mogą wystąpić problemy i komplikacje podczas pracy z danymi. Możliwe, że Twoje metody są niekompletne, Twoje systemy nie przetwarzają lub nie analizują wszystkich dostępnych informacji, albo gromadzisz niekompletne lub nieprawidłowe zestawy danych. Musisz upewnić się, że cały łańcuch, od surowych danych po wyniki, jest kompletny.

Dokumentacja

• Analiza musi być w pełni udokumentowana – istnieje wiele korzyści z utrzymywania solidnego, papierowego zapisu technik analizy danych. Przejrzysta i dokładna komunikacja jest zawsze przydatna. Pisemna dokumentacja Twojej dotychczasowej pracy może być użyteczna dla innych osób w firmie.

Umożliwi to także Tobie lub komukolwiek innemu powrót do poprzedniego problemu lub eksperymentu bez konieczności ponownego odkrywania wszystkich ustaleń.

W ten sposób analiza danych staje się ściśle kontrolowaną pętlą, od jasno zidentyfikowanego źródła, które jest chronione i kontrolowane pod kątem zmian za pomocą sformułowanych hipotez ilościowych, po analizę metodyczną z określonymi celami i otwartymi, przejrzystymi procedurami. Ponadto, opierając się na wnioskach wyciągniętych z poprzednich rozdziałów, możesz wykorzystać elementy składowe własności i prywatności, aby jeszcze bardziej zwiększyć niezawodność przetwarzania danych. Dane powinny mieć właścicieli, ich wykorzystanie powinno być mapowane, a każdy dostęp rejestrowany. W ten sposób możesz mieć pewność, że nie tylko poszanujesz prywatność swoich klientów, ale także zachowasz integralność przechowywanych informacji i zapewnisz jasne, bezstronne wyniki oparte wyłącznie na danych. Przebieg pracy przedstawiono na rysunku



Pomóż innym postępować etycznie

Jako administrator systemu poruszający się po autostradach danych znajdujesz się w wyjątkowej sytuacji, w której możesz posiadać znacznie wyższy poziom świadomości sytuacyjnej niż osoby pracujące z danymi. Możesz wykryć anomalie i problemy w systemach, metodach gromadzenia danych, przesyłaniu danych, a nawet analizie. Jeśli tak się stanie, należy poinformować właścicieli danych lub opiekunów danych i pomóc im naprawić wszelkie niespójności i luki w ich przepływach pracy. Czasami Twoimi „klientami" mogą być osoby z ograniczoną wiedzą lub dostępem do podstawowych systemów informatycznych, np. badacze lub zespoły marketingowe. Mogą nie być w stanie zajrzeć pod maskę i dowiedzieć się, dlaczego ich zestawy danych lub wyniki wyglądają źle. Będziesz mógł im pomóc, zarówno wskazując etyczne sposoby przetwarzania i analizowania danych,

jak i przeprowadzając badania techniczne i rozwiązywanie problemów z systemami, których dotyczy problem.

Historia z okopów IT: W poprzednim miejscu pracy mieliśmy administratora baz danych, który również miał dobrą praktyczną wiedzę na temat danych przechowywanych w systemie. Kiedy zauważył punkty danych, które wyglądały na literówki lub niedokładne pomiary, aktualizował wartości danych źródłowych w bazie danych, nie przestrzegając obowiązującego formalnego procesu czyszczenia danych. Nie powiadomił nikogo, że dokonał tych zmian. Wiedziałem tylko, że to zrobił, ponieważ widziałem, jak to robił, gdy mnie szkolił!

Co więcej, możesz mieć lepsze ogólne zrozumienie szerszego obrazu, biorąc pod uwagę szerszy dostęp do systemów i danych, ponieważ Twoi klienci mogą mieć kontakt tylko z niewielkim podzbiorem całkowitej puli informacji. Niekoniecznie będą w stanie dostrzec wszystkie wzorce, ale możesz im pomóc, pomagając im zrozumieć współzależności i interakcje pomiędzy różnymi komponentami i systemami w środowisku.

Wniosek

Praca z danymi jest nieuniknioną, integralną częścią administracji systemem. Czasami zostaniesz poproszony o pomoc w procesach obejmujących analizę danych. Przestrzeganie rygorystycznych zasad etycznych w swojej pracy jest niezwykle istotne. Dowiedzieliśmy się, że dane muszą być własnością i klasyfikowane, a ich wykorzystanie należy mapować i dokumentować, czasami za pomocą zautomatyzowanych narzędzi. Teraz możemy pójść o krok dalej. Źródła danych należy identyfikować, chronić i kontrolować pod kątem integralności. Analiza danych musi opierać się na powtarzalnej teorii z dobrze udokumentowanym, powtarzalnym procesem. Nie powinieneś szukać dobrych ani złych rezultatów, ponieważ mogą one i będą wypaczać postrzeganie Twojego otoczenia biznesowego i na dłuższą metę negatywnie wpłynąć na Twoją pracę. Możesz użyć automatyzacji, aby pomóc Ci w przechowywaniu i przetwarzaniu danych i wyników, aby zminimalizować błędy ludzkie. Jednak przetwarzanie danych to tylko jedna z wielu pokus Ogrodu IT. Narażenie na dane doprowadzi Cię również do informacji, które w przypadku niewłaściwego wykorzystania mogą potencjalnie nieść poważne konsekwencje dla Ciebie i Twojej firmy.