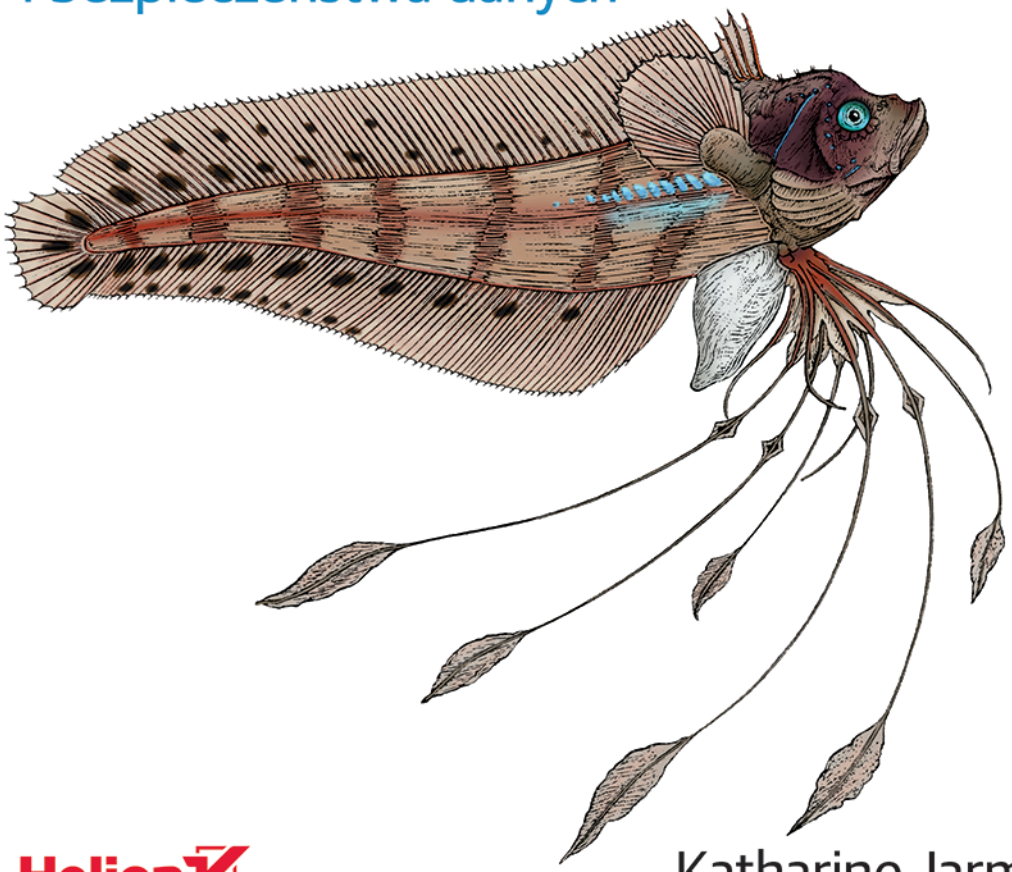
O'REILLY®

Prywatność danych w praktyce

Skuteczna ochrona prywatności i bezpieczeństwa danych



Helion

Katharine Jarmul

Tytuł oryginału: Practical Data Privacy: Enhancing Privacy and Security in Data

Tłumaczenie: Piotr Fabijańczyk, Witold Sikorski

ISBN: 978-83-289-0922-9

Authorized Polish translation of the English edition of *Practical Data Privacy*
ISBN 9781098129460 © 2023 Kjamistan, Inc.

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/prydan
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
https://ftp.helion.pl/przyklady/prydan.zip

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

ROZDZIAŁ 3.

Uwzględnianie prywatności w potokach danych

Po dokonaniu oceny różnych podejść do pseudonimizacji i anonimizacji przyjrzyjmy się, jak bezpośrednio zintegrować te podejścia z normalnymi przepływami danych. Potoki i inne rodzaje wielkoskalowej infrastruktury danych to zrównoważone i rozszerzalne podejście do projektowania prywatności w architekturze danych. Skalowanie metod ochrony prywatności, zdefiniowanych przez multidyscyplinarny zespół ekspertów i wdrożonych przez grupę rozumiejącą nie tylko technologie ochrony prywatności, ale także infrastrukturę firmy, powoduje przejście od fragmentarycznej i jednorazowej operacji do możliwej do utrzymania ochrony prywatności w fazie projektowania.

W tym rozdziale dowiesz się, jak włączyć technologie ochrony prywatności do infrastruktury i oprogramowania inżynierii danych.[1] Poznasz również wskazówki dotyczące pracy z zespołami zajmującymi się inżynierią danych (na wypadek, gdyby ktoś jeszcze nie wchodził w skład takiego zespołu!). Na koniec dowiesz się, jak uwzględnić prywatność w metodach zbierania danych, a także jak wygląda prywatność różnicowa w ramach potoku zbierania danych.

Jak wbudować prywatność w potoki danych?

W rozdziale 1. przyjrzeliśmy się podstawom zarządzania danymi i sposobom stosowania podstawowych zabezpieczeń prywatności. W rozdziale 2. omówiliśmy metody anonimizacji i prywatności różnicowej. Teraz, gdy rozumiesz już podstawowe elementy składowe prywatności, nadszedł czas, by z nimi poeksperymentować, a następnie zautomatyzować je i przeskalować do prawdziwej infrastruktury danych.

Zanim zaczniesz wbudowywać prywatność w przepływy pracy, musisz odpowiednio nakreślić ryzyko, udokumentować dane (najlepiej jak potrafisz), zrozumieć przypadek użycia i poufność danych oraz nakreślić plan przetwarzania i prywatności dla innych osób. Po zapoznaniu się z procesem określania sposobu zarządzania i wrażliwości z rozdziału 1. znasz już wymaganą metodę zbierania,

[1] Jeśli zwykle nie zajmujesz się zagadnieniami dotyczącymi inżynierii danych lub infrastruktury, to zalecam zapoznanie się z podstawami poprzez lekturę książek, filmy lub wpisy na blogu, które dotyczą potoków i architektury danych.

używania lub przekształcania danych. Niezależnie od tego, czy używasz maskowania, pseudonimizacji, czy prywatności różnicowej, możesz wykonać następujące kroki, by wbudować swoje działania związane z ochroną prywatności w potoki.

Zaprojektuj odpowiednie środki ochrony prywatności

Codziennie mogą być uruchamiane setki lub tysiące zadań inżynierii danych mające na celu przetwarzanie danych. Niezależnie od zakresu rozbudowanej konfiguracji przetwarzania danych jasno określ wszystkim zaangażowanym, jakie środki ochrony prywatności podejmiesz w odniesieniu do różnych typów danych, które wykorzystujesz. Nie wiesz, jaka powinna być odpowiednia technologia ochrony prywatności? Zacznij od prostych technik, na przykład maskowania lub redagowania wrażliwych pól, i eksperymentuj. Twoim celem jest określenie złotego środka między uzyskaną informacją a prywatnością, który umożliwia wystarczającą użyteczność danych dla celu, w jakim zostały zebrane, przy jednoczesnym zachowaniu prywatności osób. Zbieranie dostępnych osobowych lub wrażliwych danych tylko dlatego, że nikt ich nie używa, jest zawsze złym pomysłem!

Przyjrzyjmy się kilku krytycznym pytaniom i zaleceniom, które pomogą dokonać oceny, czy przyjęte miary pasują do przypadku użycia.

Zdefiniuj swój cel i przypadek użycia

Jaki jest cel zbierania tych danych? Kto i do czego będzie z nich korzystał? W jasny sposób zdefiniuj swoje przypadki użycia z użytkownikami wewnętrznymi lub końcowymi. Jeśli nie masz pewności, dlaczego dane są gromadzone lub wykorzystywane, to musisz porozmawiać z innymi o tym, czy warto ryzykować gromadzenie danych, których możesz nigdy nie użyć. Śledź stronę *https://oreil.ly/SK8Xq* i dowiedz się więcej o minimalizacji danych, która jest najlepszym dostępnym mechanizmem ochrony prywatności. Spośród wszystkich przypadków użycia nadaj najwyższy priorytet implementacji temu, który wiąże się z wyższym ryzykiem naruszenia prywatności.

Maksymalizuj prywatność

W oparciu o początkowy przypadek użycia, jaka jest najwyższa prywatność, jaką możesz zaoferować, jednocześnie wykonując swoją pracę? Czasami oznacza to eksperymentowanie z różnymi technologiami prywatności i obserwowanie wyników. Aby to zrobić bez utraty danych, utwórz tymczasową bezpieczną pamięć wewnętrzną lub próbkę, w której nie ma ochronny prywatności. Przetestuj różne metody i poproś zespół lub konsumenta danych o przyjrzenie się i sprawdzenie, czy może odpowiedzieć na swoje pytania. Jeśli możesz, zacznij od wyższej prywatności (tj. usunięcia pól, maskowania i/lub prywatności różnicowej). Na przykład, gdy definiujesz nowy produkt, dowiedz się, czego potrzebują użytkownicy. Nie dawaj im wszystkiego, o co proszą. Po pierwsze, jako osoba specjalizującą się w zagadnieniach prywatności możesz znaleźć sposoby na zaspokojenie ich potrzeb w bardziej przyjazny dla prywatności sposób. Po drugie, mogą koncentrować się na własnej analizie, a nie na równoważeniu swoich potrzeb z prawami osób, których dane dotyczą. Jeśli to tylko możliwe, rozmawiajcie! Po kilku rundach będziesz budować doświadczenie i wiedzę swoją i zespołu na temat tego, jakie środki są odpowiednie dla poszczególnych rodzajów zadań. Poznasz także nowe biblioteki i podejścia, gdy staną się one dostępne w ramach eksperymentów.

Rozwiń przypadki użycia

Po osiągnięciu odpowiedniej równowagi między prywatnością a użytecznością danych dla początkowego projektu, jakie inne zadania, przypadki użycia lub konsumenci danych będą odpowiadać tym samym lub podobnym wymaganiom? Jeśli organizacja ma silną klasyfikację danych powiązaną ze zgodą lub użyciem, to zacznij od przypadków użycia, które mają te same klasyfikacje i wymagania dotyczące użycia. Wypróbuj zwinne podejście, w którym wdrażasz małe zmiany oraz poznajesz i dostosowujesz jeden przypadek użycia lub jedno zadanie naraz. Użytkownicy danych lub użytkownicy końcowi również uczą się na bieżąco i mogą mieć różne potrzeby, które będą wymagały dostosowania. Gdy podejście działa w wielu zespołach i podobnych przypadkach użycia, wówczas może stać się standardową metodą lub podejściem. Można to udokumentować i nauczyć tego nowe osoby zajmujące się inżynierią danych lub prywatnością danych. Jeśli z jakiegoś powodu po pewnym czasie to podejście zawiedzie, upewnij się, że istnieją sposoby na to, by konsumenci danych mogli zakomunikować, co się zmieniło!

Eksperymentuj, ucz się i dostosowuj

Gdy pojawią się nowe technologie ochrony prywatności lub gdy rozwiniesz umiejętności zespołu w zakresie prywatności, wówczas określ, co możesz zoptymalizować i dostosować. Technologia stale się rozwija i zmienia. Zawsze oddeleguj kilka osób z zespołu do przewidywania kolejnych ruchów. Czy istnieje nowa wersja biblioteki, której używasz? Jeśli tak, to jakie ma nowe funkcjonalności? Czy w zespole pojawiła się nowa osoba, która wnosi konkretną wiedzę specjalistyczną, lub czy zatrudniono nową osobę zajmującą się inżynierią prywatności? Jak możesz zintegrować wskazówki tych osób ze swoimi przepływami pracy? Oceń planowane zmiany w polityce zapewniania prywatności i innych środkach zarządzania danymi w organizacji oraz współpracuj nad tworzeniem nowych zasad, standardów i wdrożeń. Patrzenie w przyszłość zapewni, że stosowane środki ochrony prywatności będą odpowiednie w chwili obecnej i w przyszłości.

Jak już wiesz z rozdziału 2., prywatność różnicowa to rygorystyczna definicja prywatności, która zapewnia możliwe do udowodnienia gwarancje. Wszystkie inne metody są zatem mniej bezpieczne i mniej zalecane. Jeśli chodzi o praktyczną prywatność danych, to prawdopodobnie konieczne będzie stosowanie metod, które nie są tak bezpieczne i rygorystyczne jak prywatność różnicowa. Wiedzę zawartą w tej książce można wykorzystać do określenia najbezpieczniejszej działającej metody. Jeśli oznacza to usunięcie pól zamiast stosowania prywatności różnicowej, to Twoim zadaniem jest również upewnienie się, że Twój zespół i inne osoby korzystające z danych rozumieją to ryzyko.

Po znalezieniu środków, które sprawdzają się w Twoich przypadkach użycia, nadszedł czas ich utrwalenia w każdym przepływie pracy. Jednak zanim to zrobimy, omówmy tworzenie przepływów danych funkcjonujących w stosunku do wszystkich zaangażowanych.

Spotykaj się z użytkownikami tam, gdzie się znajdują

Kto otrzymuje dane na końcu potoku? Jakie są wymagania z punktu widzenia prywatności i jakości danych? Ustalenie, kto jest Twoim „klientem danych", jest ważną częścią tworzenia lepszych i bardziej użytecznych potoków.

Poznając ich potrzeby oraz przypadki użycia, lepiej zrozumiesz najlepsze podejście do prywatności oraz to, jak wprowadzić kontrole zapewniające, że wszystko funkcjonuje zgodnie z planem.

Jak omawialiśmy w rozdziale 2., tworzenie prywatności różnicowej wymaga dobrego zrozumienia danych, przypadku ich użycia i związanej z tym wrażliwości. Każdy z tych czynników jest zmienny w czasie. Przeprowadzanie regularnych odpraw z osobami korzystającymi z danych może zapewnić, że dane z prywatnością różnicową nadal spełniają oczekiwania i potrzeby zespołu. Jeśli to się zmieni lub jeśli podstawowa dystrybucja ulegnie zasadniczej zmianie (np. w wyniku zmiany w oprogramowaniu lub aplikacji, która generuje dane, lub w wyniku zmiany w populacji), wówczas mechanizm prywatności różnicowej powinien zostać odpowiednio przeprojektowany. Może to być tak proste, jak dostosowanie granic zaciskania, czułości, jednostki prywatności lub wartości epsilon, lub tak złożone, jak opracowanie nowego algorytmu.

Równie ważne jest informowanie o zmianach, jakie mogą napotkać konsumenci, a które mogą wynikać z obowiązujących zabezpieczeń prywatności. Idealnie byłoby, gdyby była to ciągła dyskusja, ponieważ zespół wdrażający mechanizmy kontroli prywatności dostosowuje je do potrzeb organizacji, a także do wymagań regulacyjnych, kwestii prawnych i innych interesariuszy. Uczenie ludzi, jak korzystać z danych, które zostały poddane środkom ochrony prywatności, będzie nową umiejętnością wymaganą w zespołach zajmujących się kwestiami prywatności i danymi. Upewnienie się, że dyskusja przebiega bezproblemowo, a każdy czuje się upoważniony do zadawania pytań i eksperymentowania, będzie kluczowym czynnikiem sukcesu lub porażki inicjatyw dotyczących prywatności.

Osobiście jestem wielką fanką spotykania się z ludźmi tam, gdzie funkcjonują, jeśli chodzi o doświadczenie użytkownika. Jeśli natkniesz się na osobę zajmującą się analizą lub inżynierią, która przez ostatnie 20 lat wykonywała swoją pracę w określony sposób, a Ty masz zamiar to zmienić, to zastanów się, jak zminimalizować towarzyszące temu zakłócenia. Być może osoby te stosują takie same metody od wielu lat i ważne jest, by upewnić się, że ich praca pozostanie niezakłócona, przy jednoczesnym korzystaniu z technologii prywatności. Spraw, by drobne zmiany były prawie niezauważalne, przeprowadź użytkowników krok po kroku przez zmiany, wysłuchaj ich obaw i uwzględnij je w swojej architekturze i planowaniu. Kontynuuj dyskusję w trakcie tego procesu, by ustalić, czy przynosi to pożądany dla nich skutek. Odrobina bezpieczeństwa psychicznego może być kluczem do sukcesu!

Podejmując ten wysiłek, dowiesz się również, jak włączyć małe wycinki pracy do własnych potoków.

Implementowanie prywatności

Tworząc automatyzację prywatności danych, spróbuj osadzić ją bezpośrednio w swoich systemach, zamiast tworzyć jednorazowe i definiowane przez użytkownika funkcjonalności. Nawet jeśli od tego zaczniesz, cofnij się o krok w celu ustalenia, czy po uzyskaniu pozytywnych wyników eksperymentu istnieją holistyczne sposoby na zaimplementowanie prywatności w systemie.

Jednym ze sposobów jest tworzenie małych pakietów oprogramowania w dowolnym języku lub na dowolnej platformie, które będą odpowiadać za wykonanie poszczególnych etapów. Jeśli wykorzystujesz już oprogramowanie, które obsługuje niektóre funkcjonalności (takie jak Apache Beam

z wbudowaną obsługą prywatności różnicowej (*https://oreil.ly/24cOE*), to korzystaj z tych narzędzi, gdy tylko jest to możliwe! Jeśli używasz platformy Spark, to możesz również użyć opisanej w dalszej części tego rozdziału biblioteki Tumult Analytics (*https://oreil.ly/fNrKw*) lub PipelineDP (*https://oreil.ly/jA1Xs*). Wykonuj ciężką pracę tylko wtedy, gdy wymaga tego konkretny przypadek użycia, na przykład bardzo szczególny typ szyfrowania zachowującego format, mający na celu spełnienie wymagań określonego zespołu.

Oprócz implementowania technologii prywatności bezpośrednio w potoku warto również przetestować i zweryfikować, czy wszystko działa zgodnie z oczekiwaniami.

Testowanie i weryfikowanie

Już testujesz swoje potoki, prawda? Od wielu lat jest to najlepsza praktyka, pozwalająca lepiej zrozumieć i zweryfikować ich prawidłowe funkcjonowanie. Oprócz testowania jednostkowego lub uruchamiania testów integracyjnych do określenia kondycji systemu niezbędne jest testowanie rzeczywistych przesyłanych danych.

Od dawna jestem fanką i użytkowniczką platformy Great Expectations (*https://greatexpectations.io*), która ma dość intuicyjny interfejs i pozwala zarówno korzystać z wbudowanych oczekiwań, jak i budować nowe. Można to traktować jako pewnego rodzaju analizę statyczną danych. Czy dane przechodzą test dobrego nosa (*smell test*)?[2] Jeśli nie, to możesz oznaczyć takie dane lub zatrzymać cały proces. Pozwala to zminimalizować koszty obliczeniowe oraz przyspiesza identyfikację i usuwanie błędów systemu i potoków.

Kiedy myślisz o wykorzystaniu narzędzi takich jak Great Expectations w kwestiach zapewnienia prywatności, to co powinno zostać przetestowane? Cóż, możesz sprawdzić, czy określone pola są obecne lub ich brakuje, jeśli po przetworzeniu prywatności dodane zostały pola lub jeśli spodziewasz się, że określone wrażliwe pola zostaną usunięte lub nie zostaną uwzględnione. Możesz również sprawdzić, czy granice zaciskania z prywatnością różnicową działają zgodnie z oczekiwaniami. Możesz nawet opracować specjalne testy w celu upewnienia się, że określone pola są haszowane lub szyfrowane (testując entropię ciągów) lub że pewne tokeny maskujące występują w polach, w których się ich spodziewasz.

W celu wykorzystania tych zasad przewodnich, zapewnienia standardów i uczynienia ich bardziej praktycznymi przeanalizujemy na przykładzie sposób ich zaprojektowania.

Inżynieria prywatności i zarządzania danymi w potokach

W tym momencie możemy już opracować początkowy przypadek lub kilka przypadków użycia. W celu uzyskania przetestowanego i zatwierdzonego przekształcenia prywatności w środowisku produkcyjnym należy upewnić się, że są one zgodne z stosowanymi przepływami danych oraz zostały prawidłowo zaimplementowane i odpowiednio przetestowane.

[2] Test dobrego nosa (*https://oreil.ly/rnO2I*) w informatyce oznacza sprawdzenie, czy jakość kodu jest wysoka. Celem jest określenie, czy standardy prywatności są właściwie egzekwowane, i zachowanie odpowiedniego rozmieszczenia standardów prywatności w całej organizacji i wielu przepływach danych.

W dalszej części przeanalizujemy tego typu rozwiązanie na konkretnym przykładzie udostępniania danych pomiędzy różnymi działami w organizacji.

Przykładowy przepływ pracy w udostępnianiu danych

Udostępnianie danych w organizacji, a nawet pomiędzy organizacjami, jest bardzo powszechne. Jak to zrobić w sposób bezpieczny i z zachowaniem prywatności? Mamy tu dwa cele: zachowanie prywatności użytkowników lub pracowników oraz spełnienie potrzeb konsumentów danych. Przyjmujmy, że skonsultowaliśmy się z ekspertami wewnętrznymi w celu przeanalizowania ryzyka związanego z prywatnością i bezpieczeństwem w celu określenia odpowiedniego wymaganego poziomu ochrony.

Załóżmy, że w naszym przykładzie pracujesz w organizacji, która produkuje czekoladę. Przechowujesz dane dotyczące zakupów pochodzące od użytkowników za pośrednictwem strony internetowej. Dział marketingu chce wykorzystać te dane do mierzenia skuteczności swoich kampanii, analizując poszczególnych użytkowników.[3]

Na podstawie wstępnej analizy opracowujesz plan, w jaki sposób dane powinny zostać przekształcone, tak by zastosować odpowiednią metodę zapewnienia prywatności, przy jednoczesnym zachowaniu użyteczności danych dla tego konkretnego przypadku użycia. Twój plan jest następujący:

- Usunięcie identyfikatorów osobistych, z wyjątkiem identyfikatora użytkownika, który jest haszowany, tak by można było odpowiedzieć na konkretne pytania działu marketingu, znajdując pasujący identyfikator użytkownika.
- Jeśli sesja użytkownika została dołączona do otagowanej kampanii, to zachowywane są informacje o tej kampanii. W przeciwnym razie pole danych pozostaje puste.
- Zachowywane są informacje dotyczące adresu rozliczeniowego (miasto i województwo), ponieważ było to specjalne żądanie działu marketingu, który wymaga tych danych do skutecznego odpowiedzenia na pytania dotyczące prowadzonej kampanii.
- Dane dotyczące zamówienia są łączone i agregowane względem ilości (liczba zamówień) i wartości zamówienia (suma) dla każdego użytkownika.
- Powiązanie wartości odstających, które mają wyjątkowo dużą lub małą wartość (w razie potrzeby możliwe jest utworzenie pisemnego podsumowania dla przeglądu marketingowego).

Wyjaśnijmy, jak powyższy plan może wpisywać się w przepływ pracy. Kod Pythona może wyglądać mniej więcej tak:

```
order_dataframe.drop(['street_address', 'first_name', 'last_name',...])
browser_dataframe.drop(['ip_address', 'browser_user_agent',...])
order_campaign_df = order_dataframe.merge(browser_dataframe, how='inner', on=['order_id'])
# w tym miejscu używany jest klucz, bezpiecznie wygenerowany i przechowywany
```

[3] Dane na poziomie użytkownika nigdy nie są dobrym pomysłem na zachowanie prywatności, ale są powszechną praktyką w działach marketingu. Kiedy masz do czynienia z takimi przypadkami w swojej pracy, musisz określić najlepszy sposób postępowania dla danego przypadku. Moim zdaniem na pytania marketingowe często można odpowiedzieć na podstawie danych zagregowanych (które mogą nawet skorzystać na dodaniu prywatności różnicowej, w zależności od odbiorców i tego, jak szeroko będą udostępniane).

```
# dla tego konkretnego zadania klucz jest utrzymywany tylko na czas trwania badania danych, a następnie kasowany
order_campaign_df.user_id.map(lambda x: encrypt(x, key))
order_campaign_df = order_campaign_df.groupby('order_id').agg(
    {'campaign_uri': 'first',
     'user_id': 'first',
     'city': 'first',
     'state': 'first'
     'total': 'sum',
     'num_items': 'count',
    })
order_campaign_df = order_campaign_df.total.map(remove_outliers)
# Następnie wyeksportuj i udostępnij ze szczegółami dotyczącymi przetwarzania!
```

Działający przykład tego procesu zamieszczony został w notatniku Jupyter Notebook w repozytorium książki (*https://ftp.helion.pl/przyklady/prydan.zip*). W razie potrzeby możesz dostosować te przykłady do struktury i języków programowania używanych w organizacji.

Skąd możesz wiedzieć, czy potok działa prawidłowo? Oczywiste jest, że musisz to przetestować!

Aby to zrobić, skorzystajmy z kilku oczekiwań z biblioteki Great Expectations. Najpierw możesz zaimportować Great Expectations bezpośrednio do notatnika i skonfigurować go tak, by można było próbkować dane bezpośrednio z ramki danych biblioteki Pandas:

```
Import great_expectations jako ge
context = ge.get_context()
```

Następnie możesz sprawdzić, jakie oczekiwania są dostępne, tworząc rozszerzoną ramkę danych, która ma wbudowane oczekiwania. Zrób to przy użyciu funkcji autouzupełniania w programie Jupyter, korzystając z polecenia `expect`, a następnie naciskając *Tab*. Wynikiem jest długa lista, którą możesz przeglądać:

```
ge_df = ge.from_pandas(summary_by_order)
ge_df.expect # Teraz naciśnij Tab!
```

Poniższy kod testuje, czy z kolumny `total_price` usunięto wartości odstające. Kod można również zdefiniować tak, by używane były wartości percentyli, maksimum i minimum, a także odchylenia standardowego.

```
ge_df.expect_column_values_to_be_between('total_price', 1500, 27000)
```

Zwykle oczekiwania można dodawać ręcznie, aczkolwiek możesz również skorzystać z Great Expectations i automatycznie tworzyć zestawy oczekiwań na podstawie danych. Na końcu eksploracji należy zapisać oczekiwania i skonfigurować potok tak, by uruchamiał je jako etap przetwarzania. Dostosowanie do konkretnej konfiguracji może być bardziej szczegółowe, a wiele przewodników jest dostępnych w dokumentacji Great Expectations (*https://docs.greatexpectations.io*):

```
ge_df.get_expectation_suite(discard_failed_expectations=False)

# Tutaj sprawdzisz, czy oczekiwania odpowiadają Twoim potrzebom!
# Po wykonaniu zapisz wszystko w pliku JSON i skonfiguruj system,
# aby użyć tego pliku i GE do testowania danych podczas ich przetwarzania!

import json
with open("order_summary_for_sharing_expecation_file.json", "w") as my_file:
```

```
my_file.write(
    json.dumps(ge_df.get_expectation_suite().to_json_dict())
)
```

Teraz można mieć już pewność, że jeśli coś będzie nie tak, to uzyskane zostanie odpowiednie powiadomienie. Po pierwszych kilku wykonaniach tego procesu skontaktuj się z odbiorcami danych i zapytaj ich, czy analiza danych przebiega prawidłowo. Dowiedz się, czy zauważyli zmiany w swojej analizie lub przetwarzaniu danych. Czy rozmawiali ze swoimi konsumentami danych lub osobami, które korzystają z wyników ich analizy? Polecam codzienne przeprowadzenie odprawy w pierwszym tygodniu implementacji, a następnie zmniejszenie częstotliwości do jednego spotkania w tygodniu przez kolejny miesiąc. Poproś kogoś z zespołu o przeanalizowanie ogólnych statystyk Great Expectations w ramach cotygodniowego raportu efektywności wszystkich potoków i zwróć uwagę na wszystkie nienaturalne odchylenia, które mogą wymagać dalszego zbadania. Skonfiguruj alerty o błędach, które nie powinny się zdarzyć, i upewnij się, że osoby wchodzące w skład zespołu otrzymują te alerty i rozumieją, jak rozwiązywać problemy.

Przyjrzyjmy się teraz innemu przypadkowi użycia, tak by zbadać zarządzanie danymi w potokach. Jak w ogóle uzyskać informacje o zamówieniu i sesji?

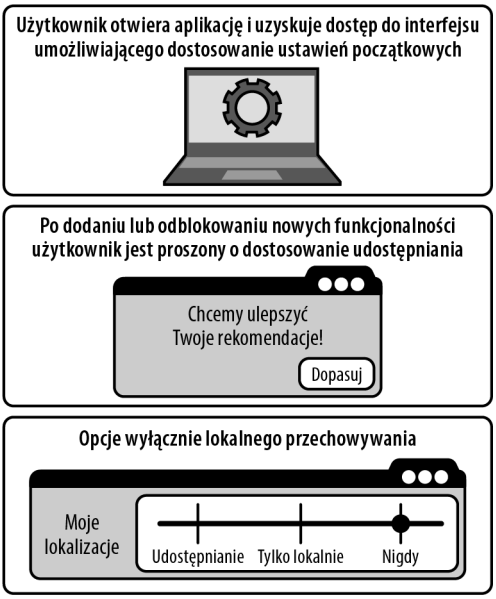
Dodawanie do gromadzonych danych informacji o pochodzeniu i zgodzie

W rozdziale 1. podkreśliliśmy, jak ważne jest zbieranie i dołączanie zgód podczas zbierania danych. Automatyzacja tego typu gromadzenia danych i ustrukturyzowanie ich w celu łatwego użycia są niezbędne do wbudowania prywatności w proces gromadzenia danych.

Idealnie byłoby, gdyby została wyrażona zgoda nie tylko na gromadzenie danych, ale także na ich wykorzystywanie. Nie jest to sposób, w jaki projektuje się obecnie większość systemów i interfejsów. Zazwyczaj użytkownik widzi listę danych, które będą zbierane, a następnie może zapoznać się z długim regulaminem lub polityką prywatności. Zazwyczaj jest to bardzo ogólny dokument, nienakreślający konkretnych przypadków użycia, na które użytkownik może następnie wyrazić zgodę i z których może się wycofać.

Jeśli pracujesz w firmie, która chce inaczej podejść do kwestii prywatności, to możesz inaczej zaprojektować swoje zasady i sposób gromadzenia danych. Może to wyglądać mniej więcej tak, jak na rysunku 3.1. Gdy użytkownik po raz pierwszy otworzy aplikację lub witrynę, wówczas załadują się ustawienia początkowe i zostanie wyświetlony interfejs. Ustawienia domyślne powinny być przede wszystkim ukierunkowane na zachowanie prywatności, zapewniając, że stosowane jest jedynie absolutnie wymagane zbieranie danych, a wszystkie inne ustawienia są wyłączone. Wszelkie nowe funkcjonalności lub przetwarzania muszą zostać zaakceptowane przez użytkownika poprzez dostosowanie ustawień. Ponadto powinna być dostępna opcja przechowywania danych tylko lokalnie, kiedy dane są przechowywane tylko na urządzeniu i nie są udostępniane centralnemu serwerowi lub aplikacji, chyba że jest to absolutnie wymagane do jej poprawnego działania (więcej na ten temat w rozdziale 6.).

Jeśli chcesz zmienić przypadki użycia lub masz nowy sposób wykorzystania danych, taki jak nowy model uczenia maszynowego, to przedstaw go użytkownikom i jawnie poproś o zgodę na ten nowy przypadek użycia danych.



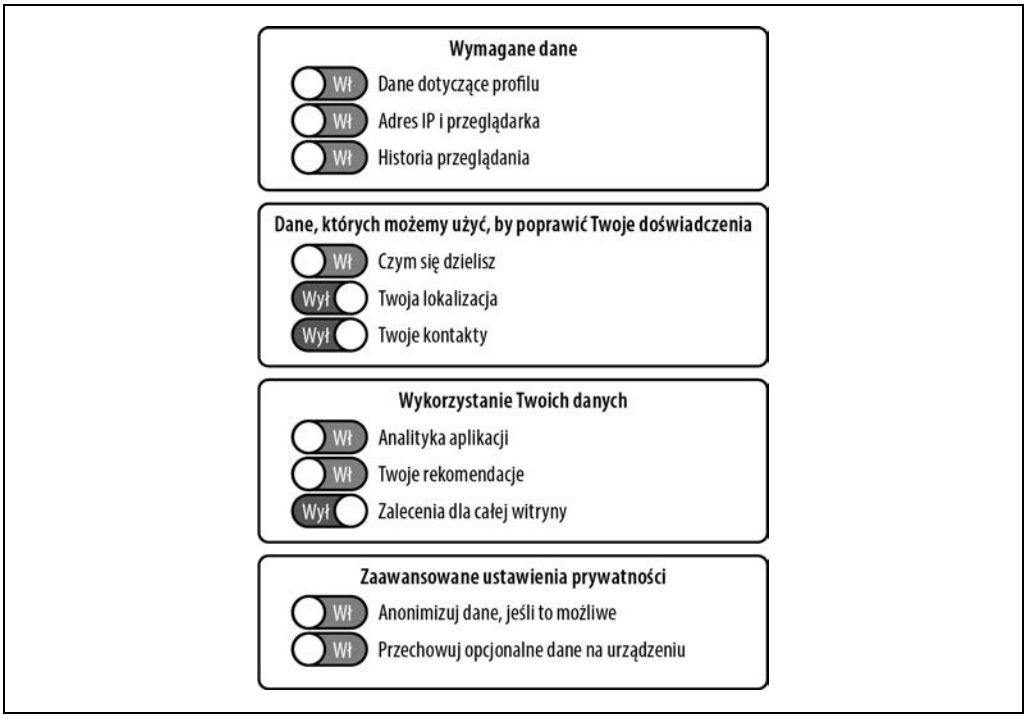
Rysunek 3.1. Zgoda i gromadzenie danych uwzględniające prywatność

Inną przyjazną dla użytkowników opcją jest udostępnienie interfejsu, który umożliwia przeglądanie wszystkich opcji zgody dla każdego przypadku użycia i ich indywidualne dostosowanie. Mogłoby to wyglądać mniej więcej tak, jak na rysunku 3.2. W tym przypadku interfejs określa, jakie dane są wymagane do każdego przypadku użycia danych, a używane słownictwo jest łatwe do zrozumienia. Użytkownicy mają do wyboru różne opcje, w tym możliwość anonimizacji danych lub przechowywania ich na urządzeniu.

Do tej pory prowadzono wiele badań dotyczących projektowania interfejsów i doświadczeń użytkownika w zakresie prywatności. Przykładem może być ind.ie Ethical Design Framework (*https://oreil.ly/Oqzox*) i Privacy UX (*https://oreil.ly/o0cO5*). Prowadzono także prace nad złymi wzorcami projektowania prywatności, czego przykładem jest praca uniwersytetu w Ulm (*https://oreil.ly/qrmP1*).

W ramach organizacji warto przeprowadzić dyskusję z działem projektowania obsługi klienta i ustalić sposób dostosowania przepływu pracy, tak by ulepszyć zbieranie zgód oraz tworzyć bardziej przejrzyste komunikaty.[4] Znając interfejsy zgody na przetwarzanie danych, wiesz, że o wiele łatwiej jest wyrazić zgodę, gdy jest jasne, do czego i dlaczego są wykorzystywane dane, a także gdy jest miejsce, w którym można sprawdzić te ustawienia. Jest to również podstawowy wymóg przepisów RODO (więcej na ten temat w rozdziale 8.).

[4] Polecam przeczytanie artykułów Vitaly'ego Friedmana z serii *Privacy UX Experience* na portalu Smashing Magazine (*https://oreil.ly/HE4NK*) oraz skonsultowanie się z zespołem UX w celu określenia potencjalnych zmian dla swojej organizacji.



Rysunek 3.2. Szczegółowy interfejs zarządzania zgodami na zbieranie danych

Nawet jeśli nie ma możliwości przeprojektowania sposobu zbierania zgód w znaczący sposób, to należy zwrócić uwagę na to, jakie dane dotyczące zgody są dostępne w momencie zbierania. Oczywiste jest, że regulacje prawne dynamicznie się zmieniają. Wiedza o tym, które zasady polityki prywatności zostały użyte podczas gromadzenia danych, może pomóc w późniejszym podejmowaniu lepszych decyzji dotyczących wykorzystania, przechowywania i minimalizowania danych.

Jak w praktyce miałoby działać udzielanie zgody i gromadzenie danych z uwzględnieniem prywatności? Przedstawmy wymagania dotyczące zbierania danych, a następnie zaprojektujmy schemat, który będzie przydatny na potrzeby późniejszej ochrony prywatności. Dane, które należałoby zebrać, to:

- Wersja polityki prywatności.
- Okres przechowywania.
- Data zaakceptowania zasad przez użytkownika.
- Cel przetwarzania danych (tj. z jakiego powodu te dane są gromadzone?).
- Wersja regulaminu.
- Lokalizacje, których dotyczą dane.
- Preferencje użytkowania (np. proszę używać moich danych do tej usługi, ale nie do powiązanych usług partnerskich, lub proszę używać moich danych do uczenia maszynowego, ale nie udostępniać ich innym organizacjom).
- Szczegóły dotyczące pochodzenia danych.

Pamiętaj, że tak naprawdę schemat nie będzie w pełni opracowany, dopóki nie zostanie przetestowany i zatwierdzony. Jak można się upewnić, że dane dotyczące udzielania zgody są odpowiednio śledzone?

Oto przykładowa walidacja schematu mająca na celu upewnienie się, że dane są prawidłowo zbierane i przechowywane. W tym przykładzie została wykorzystana składnia Apache Avro (*https://avro.apache.org*) jako kolejny ze sposobów testowania schematu danych i weryfikowania struktur danych w potokach.

```
{"namespace": "example.avro",
 "type": "record",
 "name": "User Consent Data",
 "fields": [
     {"name": "username", "type": "string"},
     {"name": "policy_version", "type": "float"},
     {"name": "retention_months", "type": "int"},
     {"name": "agreement_date", "type": "datetime"},
     {"name": "processing_purposes", "type": "array",
                                          "items" : "string",
                                          "default": []},
     {"name": "terms_version", "type": "float"},
     {"name": "data_localities", "type": "array", "items" : "string",
                                         "default": ["us-aws-east"]},
     {"name": "usage_detail__location_on", "type": "bool"},
     {"name": "usage_detail__location_ml", "type": "bool"},
     {"name": "usage_detail__location_analytics", "type": "bool"},
     {"name": "usage_detail__location_sharing", "type": "bool"},
     {"name": "usage_detail__actions_on", "type": "bool"},
     {"name": "usage_detail__actions_ml", "type": "bool"},
     {"name": "usage_detail__actions_analytics", "type": "bool"},
     {"name": "usage_detail__actions_sharing", "type": "bool"},
     {"name": "provenance_location", "type": "string"},
                          /* aplikacja lub strona internetowa itp. */
 ]
}
```

W powyższym przykładzie pola szczegółów użycia zostały zapisane z wykorzystaniem znaku _, co ułatwia filtrowanie danych na podstawie typu szczegółowej zgody. Dzięki temu potoki ciągłej integracji wykorzystywane w uczeniu maszynowym mogą łatwo automatycznie wybierać użytkowników, którzy wyrazili zgodę, i natychmiast kierować odpowiednie dane do wnioskowania lub procesu uczenia.

W tym momencie masz również dane dotyczące okresu przechowywania danych oraz daty zawarcia umowy, co ułatwia grupową anonimizację lub usuwanie rekordów, które ze względu na osiągnięcie okresu przechowywania wkrótce wygasną. Korzystne jest tworzenie potoków dla zadań, których daty wygaśnięcia będą zbyt bliskie. Jeśli chcesz zachować dane, w stosunku do których jest używany zgodny z prawem mechanizm anonimizacji bazujący na prywatności różnicowej (więcej w rozdziale 8.), to oznacza to konieczność zintegrowania prywatności różnicowej jako normalnej części przepływów danych. Przyjrzyjmy się temu bliżej!

Wykorzystywanie bibliotek prywatności różnicowej w potokach

Podczas dodawania prywatności różnicowej do bieżących lub nowych potoków najlepiej jest użyć dobrze obsługiwanej biblioteki, która zintegruje się z obecnie używaną technologią. Jeśli korzystasz już z Apache Beam na Google Cloud Platform, to polecam skorzystanie z notatników Codelab (*https://oreil.ly/IKHm4*), by poznać sposoby implementacji prywatności różnicowej.

Jeśli używasz Sparka, to polecam bibliotekę Tumult Analytics (*https://oreil.ly/idW3D*), z wbudowanym rozliczaniem prywatności i wieloma innymi funkcjonalnościami.

W momencie pisania tej książki biblioteka Tumult Analytics była stosunkowo nowa, więc polecam zapoznać się z dokumentacją (*https://oreil.ly/gI2ia*) na wypadek, gdyby coś uległo zmianie. Notatniki z repozytorium książek są również regularnie aktualizowane, a więc zapoznaj się z nimi w celu wypróbowania biblioteki.

Przyjrzyjmy się bibliotece prywatności różnicowej Tumult Analytics i przeanalizujmy, jak działałaby w ramach normalnego przepływu pracy. Możesz użyć nowego zestawu danych, do którego odwołuje się dokumentacja Tumult (*https://oreil.ly/gI2ia*), lub po prostu skorzystać bezpośrednio z repozytorium kodu książki.

Na początek rozpocznij sesję platformy Spark i zainicjuj budżet prywatności różnicowej. W naszym przykładzie została użyta wartość epsilon równa 1,1:

```
session = Session.from_dataframe(
    privacy_budget=PureDPBudget(epsilon=1.1),
    source_id="members",
    dataframe=members_df,
    protected_change=AddOneRow(),
)
```

Wykorzystywany zestaw danych obejmuje listę osób używających biblioteki oraz ich aktywność. Sprawdźmy najpierw kolumny:

```
members_df.columns
```

Spowoduje to wyświetlenie następujących danych:

```
['id',
 'name',
 'age',
 'gender',
 'education_level',
 'zip_code',
 'books_borrowed',
 'favorite_genres',
 'date_joined']
```

Załóżmy, że interesuje nas istnienie korelacji między liczbą wypożyczanych książek a poziomem wykształcenia. Aby przyjrzeć się kolumnom kategorii przy użyciu Tumult, musimy utworzyć zestaw KeySet (*https://oreil.ly/4vHVJ*). Umożliwia on zdefiniowanie wartości zmiennych jakościowych, których chcemy używać do grupowania.

```
edu_levels = KeySet.from_dict({
    "education_level": [
        "up-to-high-school",
        "high-school-diploma",
        "bachelors-associate",
        "masters-degree",
        "doctorate-professional",
    ]
})
edu_average_books_query = (
    QueryBuilder("members")
    .groupby(edu_levels)
    .average("books_borrowed", low=0, high=100)
)
edu_average_books = session.evaluate(
    edu_average_books_query,
    privacy_budget=PureDPBudget(0.6),
    # Oszczędzam trochę budżetu prywatności na później, więc zużywam teraz tylko 0,6 epsilonu (łącznie 1,1)
)
edu_average_books.sort("books_borrowed_average").show(truncate=False)
```

W tym miejscu klasy Tumult KeySet i QueryBuilder (*https://oreil.ly/-DcfZ*) tworzą zapytanie z prywatnością różnicową. Przykładowy kod ogranicza liczbę wypożyczonych książek do zakresu od 0 do 100, o czym wspominaliśmy w rozdziale 2. W celu wykonania zapytania musimy poinformować bibliotekę, jaką część budżetu prywatności chcemy wydać. Po uruchomieniu kodu otrzymałam następujące dane wyjściowe. Uwaga: Twoje wyniki będą inne ze względu na użycie prywatności różnicowej:

```
+----------------------+----------------------+
|education_level       |books_borrowed_average|
+----------------------+----------------------+
|doctorate-professional|18.929587482219063    |
|masters-degree        |19.1402224030377      |
|bachelors-associate   |19.173858890761228    |
|up-to-high-school     |19.361286812215194    |
|high-school-diploma   |19.57674149725407     |
+----------------------+----------------------+
```

Widać, że w tym konkretnym zbiorze danych nie ma zauważalnej korelacji między poziomem wykształcenia a liczbą wypożyczonych książek. Osoby korzystające z biblioteki nie zmieniają znacząco swoich zachowań związanych z wypożyczaniem w zależności od ich poziomu wykształcenia. Zakres zaciskania liczby wypożyczonych książek może być również wysoki, więc możemy go zmienić w kolejnych zapytaniach.

Zobaczmy, czy osoby pożyczają książki w różny sposób w zależności od wykształcenia i wieku. Najpierw musimy utworzyć przedziały wiekowe (*https://oreil.ly/ihlxx*), by można było pogrupować dane ze względu na ten parametr:

```
age_binspec = BinningSpec([10*i for i in range(0, 11)])

age_bin_keys = KeySet.from_dict({
    "age_binned": age_binspec.bins()
})
```

```
binned_age_with_filter_query = QueryBuilder("members")\
      .filter("education_level='masters-degree'" or\
              "education_level='doctorate-professional'")\
      .bin_column("age", age_binspec)\
      .groupby(age_bin_keys)\
      .average("books_borrowed", low=0, high=22)
session.evaluate(binned_age_with_filter_query,
                 privacy_budget=PureDPBudget(0.4)).show(truncate=False)
```

Po uruchomieniu tego kodu zostały wyświetlone następujące wyniki:

```
+----------+----------------------+
|binned_age|books_borrowed_average|
+----------+----------------------+
|100-109   |-2.0                  |
|80-89     |11.476923076923077    |
|40-49     |11.034418604651163    |
|30-39     |11.501822600243013    |
|70-79     |11.256830601092895    |
|20-29     |11.08816705336427     |
|50-59     |11.599250936329588    |
|10-19     |14.0                  |
|90-99     |-24.0                 |
|60-69     |10.970472440944881    |
|0-9       |19.0                  |
+----------+----------------------+
```

Tym razem możemy zaobserwować, że do niektórych kolumn dodano wiele szumu prywatności różnicowej, ponieważ niektóre wartości są nawet ujemne. Co poszło nie tak? W tym przypadku kod filtrował dane według wieku i nie uwzględniał tego, że niektóre z grup wiekowych będą prawdopodobnie niedostatecznie reprezentowane. Prawdopodobieństwo, że osoba w wieku 8 lat ma tytuł magistra, jest raczej małe.

Ten przykład doskonale zwraca uwagę na istotną sprawę. Myślenie o danych oraz precyzowanie zapytań i hipotez w celu zaplanowania analizy przed uruchomieniem jakiegokolwiek kodu jest niezbędne podczas pracy z prywatnością różnicową. W takim przypadku można było uruchomić zapytanie w celu przeanalizowania kategorii wiekowych lub sporządzić wykres obrazujący dostępne zakresy wiekowe (z liczbami lub bez), co dałoby nam lepsze zrozumienie przed tworzeniem zapytań dotyczących celu analizy. Jeśli korzystasz z Tumult Analytics, to możesz zacząć od nieograniczonego budżetu prywatności, by zorientować się, jak małe zmiany w epsilon wpłyną na otrzymywane wyniki. W rozdziale 5. dowiesz się również więcej o tym, jak podejść do prywatności różnicowej i innych mechanizmów w całym procesie nauki o danych.

Więcej przykładowych zapytań odnoszących się do bibliotecznego zestawu danych zostało zamieszczonych w repozytorium książki, możesz zatem poeksperymentować z biblioteką Tumult Analytics na własnym zestawie danych. Jeśli dopiero eksperymentujesz, to możesz ustawić swój budżet prywatności na nieskończoność! Jednak w rzeczywistości konieczne jest użycie pewnej wartości epsilon oraz określenie sposobu podziału budżetu prywatności. Możliwe, że otrzymasz błąd podobny do tego poniżej:

```
RuntimeError: Cannot answer query without exceeding privacy budget: it needs approximately
0.100, but the remaining budget is approximately 0.000 (difference: 1.000e-01)
```

Dostęp do danych i prywatność

W scenariuszu korzystania z prywatności różnicowej na dużą skalę może dojść do utworzenia różnych ścieżek dostępu z różnymi gwarancjami prywatności. Oto kilka scenariuszy do przemyślenia:

Ograniczenia dostępu z prywatnością różnicową

Dostęp do danych w całej firmie odbywa się tylko za pośrednictwem zapytań z prywatnością różnicową, aczkolwiek umożliwia niewielkiej liczbie osób z zespołu zajmującego się danymi dostęp z bardzo dużą wartością epsilon lub w ogóle bez prywatności różnicowej.

Udostępnianie danych przez osoby trzecie z zachowaniem prywatności różnicowej

Mechanizm prywatności różnicowej jest wykorzystywany wobec wszystkich danych udostępnianych stronom trzecim.

Prywatność różnicowa w przepływach pracy w nauce o danych

Po wykonaniu wstępnej eksploracyjnej analizy danych i lepszym zrozumieniu danych prywatność różnicowa jest stosowana jako część przepływu pracy w nauce o danych. Pamiętaj, że celem jest określenie sposobu najlepszej ochrony danych i jednocześnie umożliwienie osobom korzystającym z danych wykonywania analizy i odpowiadania na ważne zapytania!

Eksperymentuj i określ, co sprawdza się w Twojej organizacji, upewniając się, że oceniasz prywatność w porównaniu z kontinuum informacji oraz jednocześnie bierzesz pod uwagę to, jak chronić dane i zapewniać dostęp.

Oznacza to, że przydzielony został zbyt mały budżet prywatności, uniemożliwiający udzielenie odpowiedzi na dane zapytanie.

Gorąco polecam skorzystanie z przykładowego zestawu danych i zapoznanie się z możliwościami biblioteki Tumult lub innej dowolnej i dobrze zrecenzowanej biblioteki prywatności różnicowej o otwartym kodzie źródłowym! Tumult nie jest jedyną biblioteką, która współpracuje ze Sparkiem. Możesz użyć innych rozwiązań, takich jak PipelineDP, który jest utrzymywany przez niektórych członków zespołu Google wraz ze społecznością OpenMined (*https://www.openmined.org*). Dostępnych jest również kilka przykładów w repozytorium GitHub (*https://oreil.ly/Xjk4z*).

Nie każda biblioteka oferuje tak zwane kompleksowe gwarancje prywatności różnicowej, co oznacza, że śledzenie i alokacja budżetu prywatności staje się zadaniem trudnym i podatnym na błędy. Jeśli zaczynasz korzystać z prywatności różnicowej, to polecam trzymać się bibliotek, które będą za Ciebie zarządzać budżetem prywatności.

Jak powiedzieliśmy w rozdziale 2., zbudowanie własnego mechanizmu prywatności różnicowej i śledzenie budżetu prywatności może być bardzo pracochłonne i obarczone złożonymi przypadkami brzegowymi i błędami. To wspaniałe, że nauka o danych wkracza w erę, w której biblioteki *open source* mogą wspierać tego typu pracę. Teraz, gdy znasz już podstawy tworzenia własnego prostego mechanizmu, zrozumiałe powinno być także to, jak odpowiednio rozdysponować budżet prywatności i zapewnić, że gwarancje prywatności są dobrze zrozumiałe dla osób korzystających z danych. Możesz również poinformować dalszych odbiorców o procesie przetwarzania danych, by mogli go lepiej zrozumieć i zająć się wprowadzonym błędem.

Być może zastanawiasz się, czy możesz korzystać z prywatności różnicowej podczas zbierania danych. W rzeczywistości istnieje kilka różnych sposobów na zrobienie tego. W dalszej części przyjrzymy się anonimizacji w trakcie gromadzenia danych!

Anonimowe gromadzenie danych

W momencie, gdy możesz już korzystać z bibliotek prywatności różnicowej typu *open source* w swoich potokach danych, możesz pomyśleć o zbieraniu danych w sposób anonimowy. Wiąże się to z kilkoma nowymi zagadnieniami. W jaki sposób możesz identyfikować wrażliwość danych w sytuacji, gdy nie masz do nich dostępu? Jakie gwarancje prywatności możesz zaoferować, jeśli anonimizacja zostanie przeniesiona do etapu gromadzenia danych?

W celu przeanalizowania dostępnych opcji przyjrzyjmy się, w jaki sposób firma Apple zaimplementowała swój próbnik prywatności oparty na emotikonach, by zobaczyć, jak działa lokalna prywatność różnicowa (w porównaniu z centralną prywatnością różnicową, o której dyskutowaliśmy w rozdziale 2.).

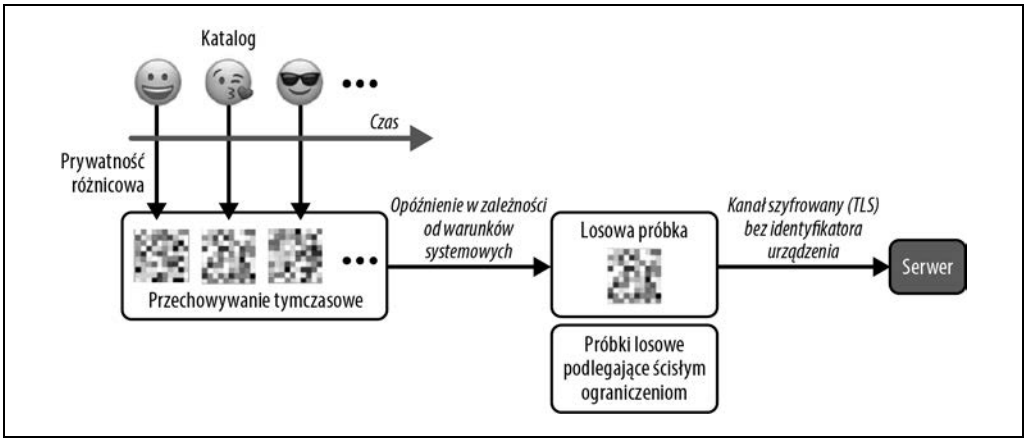
Gromadzenie danych z prywatnością różnicową przez Apple

Firma Apple była jedną z pierwszych firm, które wdrożyły lokalną prywatność różnicową na skalę całego internetu, dodając mechanizm prywatności różnicowej przed etapem zebrania i scentralizowania danych. Apple, jako marka i firma dysponująca odpowiednimi zespołami naukowymi i inżynieryjnymi, które pracują nad systemami prywatności, zdecydowała, że zapewnienie prywatności będzie jej wyróżnikiem.

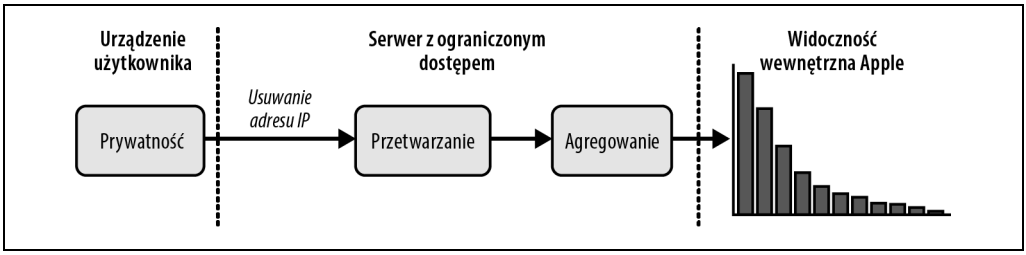
Firma Apple chciała ulepszyć swoje sugestie emotikonów, ale nie chciała wysyłać prywatnych danych tekstowych na swoje serwery, by dowiedzieć się, jakie emotikony pojawiają się w pobliżu danego tekstu. Ustalono, że częstotliwość występowania emotikonów i sąsiadujące tokeny rozwiążą ten problem, i przystąpiono do wdrożenia tego rozwiązania na urządzeniach z systemem iOS z uwzględnieniem prywatności różnicowej.

W opublikowanym wstępnym dokumencie na temat mechanizmów prywatności różnicowej (*https://oreil.ly/AtH_I*) firma Apple zawarła kilka przydatnych grafik, które pokazują prywatność różnicową na poziomie urządzenia przed etapem zebrania danych. Dane są wysyłane na serwery Apple dopiero po zastosowaniu mechanizmu prywatności różnicowej.

Jak można zauważyć na rysunku 3.3, określony szum jest dodawany do emotikonów na rzeczywistym urządzeniu. Jeśli masz urządzenie z systemem iOS, to możesz sprawdzić wartości epsilon w telefonie, przechodząc do pozycji *Ustawienia/Prywatność i ochrona/Analizy i udoskonalenia* i przeglądając dane analityczne we wpisach zaczynających się od *DifferentialPrivacy*. Po dodaniu szumu reprezentacja emotikonów jest wysyłana na serwery Apple, gdzie umieszczane są adres IP i informacje o pochodzeniu. Wyniki są następnie agregowane w rozkładzie statystycznym, który może zostać odszumiony ze względu na skalę i znajomość zastosowanych parametrów prywatności różnicowej. Przebieg tego procesu został przedstawiony na rysunkach 3.3 i 3.4.



Rysunek 3.3. Emotikony Apple wykorzystujące zbieranie danych z lokalną prywatnością różnicową



Rysunek 3.4. Potok zbierania danych dotyczących emotikonów firmy Apple

Na rysunku 3.4 można zaobserwować, że początkowy mechanizm prywatności różnicowej jest stosowany na urządzeniu (widoczne również na rysunku 3.3). W momencie, gdy zanonimizowane dane są wysyłane na serwer, usuwany jest adres IP, tak by nikt nie mógł wyśledzić danych bezpośrednio do użytkownika, ponieważ natychmiast zniweczyłoby to wszelkie gwarancje prywatności uzyskane dzięki prywatności różnicowej. Wewnętrznie firma Apple używa przetwarzania i agregowania, które przekształcają dane przychodzące i agregują je na potrzeby użytku analitycznego. Końcowym elementem są wewnętrzne pulpity nawigacyjne używane przez zespoły zajmujące się analizą danych i programowaniem aplikacji. Na rysunku każda z linii kropkowanych reprezentuje granicę zaufania. Jeśli przetwarzanie narusza tę granicę lub jeśli osoba zatrudniona w Apple może śledzić dane, unieważnia to gwarancje prywatności.

Po zakończeniu potoku można przystąpić do interpretowania wyników analiz. Oznacza to, że wobec każdej osoby istnieje wiarygodne zaprzeczenie co do najczęściej używanych emotikonów, a jednocześnie zespół prowadzący analizę ma informacje o tym, jak lepiej dostarczać emotikony dla różnych języków klawiatury. Wyniki analiz przedstawione na rysunku 3.5 sugerują istnienie różnicy w popularności emotikonów między francusko- i angielskojęzycznymi osobami korzystającymi z klawiatury.

Czym są granice zaufania?

Granica zaufania jest jak granica, na której bezpieczeństwo zmienia się w zasadniczy sposób. Zazwyczaj granice te oznaczają, że użytkownicy i operatorzy jednego obszaru nie powinni mieć takiego samego dostępu do drugiego obszaru. Poniżej przedstawiono kilka przykładów wyjaśniających tę definicję.

Od zewnętrznych do wewnętrznych

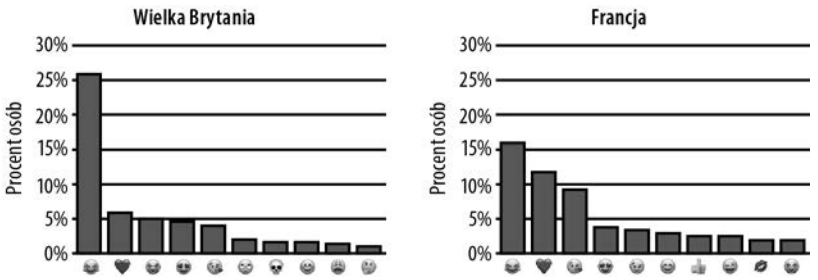
W momencie, gdy użytkownik wprowadza dane do formularza aplikacji internetowej i wysyła je do serwera aplikacji, przekracza granicę zaufania. Konieczne jest potwierdzenie, że żądanie zostało prawidłowo sformułowane, zweryfikowane i nie zawiera złośliwych poleceń (takich jak atak polegający na wstrzyknięciu kodu SQL), zanim serwer będzie mógł udzielić prawidłowej odpowiedzi. Dodatkowo musi się to odbywać pod nadzorem administratorów serwera i aplikacji.

Od środowiska bezpiecznego do mniej bezpiecznego

Przy przenoszeniu danych z bezpiecznej architektury kontrolowanej przez określone standardy do mniej bezpiecznej przekraczana jest granica zaufania. Może to oznaczać przejście z systemu lokalnego do chmury lub z restrykcyjnie ograniczonego obszaru architektury do obszaru o słabszych zabezpieczeniach.

Podrzędne przepływy danych

Podczas tworzenia przepływów danych może istnieć wielu konsumentów danych. Często będą mieli oni różne prawa dostępu i odbiorców, co oznacza, że takie przepływy danych działają również ponad granicami zaufania. Odpowiedni dobór granic zaufania do przepływów pracy danych jest ważnym krokiem w określaniu, gdzie należy zastosować mechanizmy i technologie kontroli prywatności.



Rysunek 3.5. Analiza z prywatnością różnicową dotycząca wykorzystania emotikonek na klawiaturze

Takie podejście może być przydatne w przypadku dysponowania kilkoma prostymi szkicami lub rozkładami, które mają zostać przeanalizowane, oraz w przypadku zbierania danych na urządzeniach. Jedną z istotnych rzeczy, o których dowiesz się w następnej części, jest to, że jest to naprawdę wykonalne tylko na dużą skalę, ze względu na ilość szumu wprowadzonego na poziomie indywidualnym.

Zespół zajmujący się zagadnieniami prywatności różnicowej firmy Apple regularnie publikuje swoje najnowsze badania (*https://oreil.ly/20UxD*), które mogą być bardzo pouczające, jeśli rozważasz zaimplementowanie prywatności różnicowej na poziomie urządzenia.[5]

Co jednak z innymi sposobami gromadzenia danych z zachowaniem prywatności? Firma Google opracowała pewne podejście, którego później zaniechano, jednak jest to przypadek użycia warty przeanalizowania.

Dlaczego pierwotne zbieranie danych z prywatnością różnicową w Chrome zostało porzucone?

RAPPOR był jedną z pierwszych udanych implementacji prywatności różnicowej na dużą skalę. Była to technologia wydana przez Google w 2014 roku i nadal można zapoznać się ze szczegółami tej implementacji (*https://oreil.ly/sqy3M*) oraz z kodem (*https://oreil.ly/bqz1o*) za pośrednictwem repozytorium *open source* i opublikowanej dokumentacji.

Technologia RAPPOR została opracowana z myślą o zbieraniu wrażliwych danych z przeglądarek Chrome w sposób chroniący prywatność. W tym celu zespoły zaimplementowały odpowiedź losową w bibliotece, która została wykorzystana do zbierania podstawowych statystyk w postaci ciągów bitowych. Na komputerach użytkowników może się znajdować wiele informacji dotyczących zdarzeń lub innych danych, które są interesujące dla osób zajmujących się rozwijaniem aplikacji Chrome. Może to być na przykład fakt użycia określonego rozszerzenia, wystąpienie określonego błędu lub wybrane informacje dotyczące historii przeglądarki Chrome.

Zatrzymajmy się tutaj i zdefiniujmy losową odpowiedź (*https://oreil.ly/EPUBT*). Była to technika po raz pierwszy zaproponowana w 1965 roku przez zespoły zajmujące się analizowaniem ankiet, ponieważ wiązało się to z przeprowadzaniem wywiadów z ludźmi na drażliwe tematy, na które chciano uzyskać prawdziwe odpowiedzi. Opracowano zatem metodę z prywatnością różnicową, by dać ludziom wiarygodną możliwość zaprzeczenia.

W przykładowej losowej odpowiedzi opartej na rzucie monetą metoda ta jest następująca:

1. Pada wrażliwe pytanie, a osoba odpowiadająca rzuca monetą.
2. Jeśli moneta pokazuje reszkę, to odpowiada „tak” niezależnie od prawdy.
3. Jeśli na monecie pojawi się orzeł, to odpowiada zgodnie z prawdą.[6]

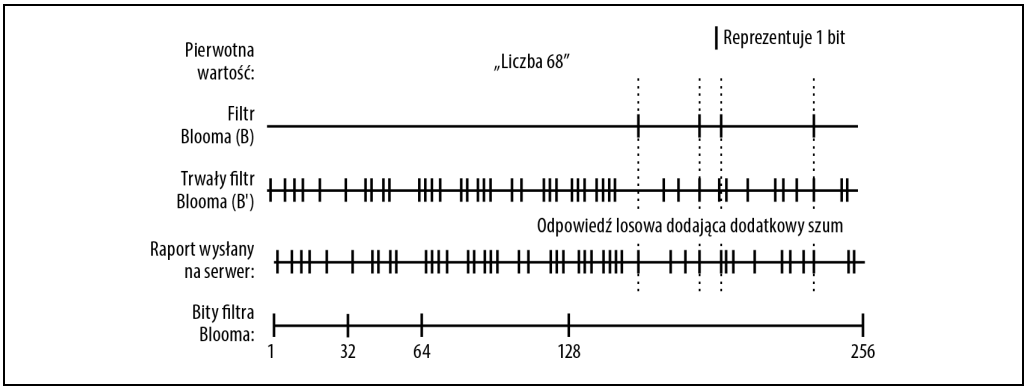
[5] Chociaż Apple jest liderem na rynku w zakresie wdrażania technologii prywatności, nie oznacza to, że nie ma przykładów na to, że ta firma podąża śladami wielu innych firm technologicznych podczas gromadzenia danych. W trakcie pisania tej książki toczył się proces sądowy, który dotyczył kwestionowania głoszonego przez Apple stawiania prywatności na pierwszym miejscu (*https://oreil.ly/gysAJ*). Podnoszono argument braku realnego sposobu na wyłączenie udostępniania wszystkich danych. W tej książce poznasz więcej konsensualnych sposobów komunikowania gromadzenia danych osobom użytkującym przez stosowanie projektowania zorientowanego na nie i ich prywatność oraz z większym upoważnieniem do kontrolowania własnych danych.

[6] Istnieje wiele wariantów udzielania odpowiedzi losowej, niektóre z nich obejmują wykorzystanie kości, kart lub rzucanie monetą co sekundę. Jest to najprostsza wersja, ale jeśli chcesz dowiedzieć się więcej na ten temat, to polecam zagłębić się w badania dotyczące stosowania odpowiedzi losowej w ankietach.

Ta metoda zapewnia mechanizm prywatności różnicowej z dość niskim epsilonem i dużą ilością szumu! W celu zmniejszenia obciążenia wyników należy założyć, że około połowa odpowiedzi „tak" jest fałszywa, i usunąć je. Jeśli wiesz więcej o badanej populacji, to możesz to zrobić bardziej inteligentnie. Jeśli chcesz dowiedzieć się więcej i poeksperymentować, to w repozytorium książki został zamieszczony odpowiedni notatnik wraz z przewodnikiem i kilkoma linkami (*https://github.com/kjam/practical-data-privacy*).

W ramach projektu RAPPOR zastosowano trzyczęściowy przepływ pracy w celu zbierania danych z komputera lokalnego w dłuższym okresie, przy jednoczesnym zapewnieniu gwarancji prywatności różnicowej. Gwarancje te są podobne do tych z metody odpowiedzi losowej. Na wysokim poziomie następowało kodowanie odpowiedzi w sekwencji bitowej. Każda z tych odpowiedzi była związana z małym punktem danych. Następnie sekwencja bitowa była haszowana do filtru Blooma. Filtr ten był poddawany procesowi losowej odpowiedzi i zapisywany jako stały magazyn danych na urządzeniu. Następnie przygotowywana była tablica bitowa z dodatkowym szumem. Takie postępowanie miało na celu wyeliminowanie możliwości, że ktoś z czasem pozna stałe odpowiedzi przechowywane w tablicy filtrów.

Analiza rysunku 3.6, bazującego na oryginalnym artykule, jest dość pouczająca. Zapisywana jest stała odpowiedź RAPPOR (*B'*), natomiast chwilowa losowa odpowiedź zawiera dodatkowy szum, zapewniający lokalne gwarancje prywatności. Możesz prześledzić przechodzenie bitów wiadomości w kierunku zbieranych wartości, korzystając z pionowych linii, i zaobserwować kombinację informacji (pierwotna odpowiedź) i szumu (stały filtr Blooma i odpowiedź losowa) w działaniu. Które bity zostały wysłane, a które nie?



Rysunek 3.6. Wizualizacja zbierania danych w metodzie RAPPOR

Google udostępniło projekt na licencji *open source* w 2014 roku, a następnie przestało go aktualizować w 2016 roku. Dlaczego?

Względny błąd odpowiedzi losowej w porównaniu z innymi metodami scentralizowanej prywatności różnicowej (jak w rozdziale 2.) jest dość wysoki. Aby uzyskać częściowo dokładne wyniki dla odpowiedzi losowej, należy nie tylko zebrać duże ilości danych, ale także uwzględnić duży błąd w danych.

W celu uzyskania gwarancji prywatności różnicowej w scentralizowanym modelu z rozkładem Laplace'a rozkład i wynikowy szum są skalowane do czułości/ε dla wszystkich punktów danych (k). Aby uzyskać te same gwarancje dla lokalnego mechanizmu prywatności różnicowej, należy dodać błąd czułości/ε dla *każdej* osoby. Oznacza to, że błąd względny dla lokalnej prywatności różnicowej skaluje się z osobami do około $\frac{1}{\sqrt{k}\cdot\varepsilon}$ i jest znacznie bardziej stromy niż błąd względny modelu scentralizowanego, który wynosiłby około $\frac{1}{k\cdot\varepsilon}$.

Jeśli chcesz zapewnić lokalną prywatność różnicową (tak jak robił to RAPPOR i jak robi to Apple podczas zbierania danych), to z czasem musisz uwzględniać błąd o naprawdę dużej wielkości. Jest prawdopodobne, że analiza danych wymaganych przez zespół Chrome nie była możliwa przy tak dużym błędzie i właśnie dlatego ostatecznie ta technologia przestała być utrzymywana i nie jest już obecnie używana ani obsługiwana.

Upewnij się zatem, że rozumiesz, co chcesz zrobić z danymi i jak chcesz je analizować w czasie, zanim zainwestujesz czas i zasoby inżynieryjne w zaprojektowanie rozwiązania ochrony z prywatnością różnicową, które w rzeczywistości nie będzie pasować do Twoich potrzeb. W chwili pisania tej książki lokalna prywatność różnicowa i towarzyszący jej duży szum są nieodpowiednim rozwiązaniem, uniemożliwiającym udaną analizę eksploracyjną, chyba że dokładnie wiesz, co chcesz przeanalizować (jak w przypadku analizy emotikonów firmy Apple).

W kolejnych latach Google opracowało metodę Prochlo (*https://oreil.ly/-z4dX*), która wykorzystuje własną infrastrukturę przetwarzania, granice zaufania, szum i szyfrowanie w celu zapewnienia gwarancji prywatności różnicowej dla procesu zbierania danych. Warto zapoznać się z tym dokumentem, by zobaczyć, w jaki sposób gwarancje prywatności różnicowej na dużą skalę mogą łączyć rozwiązania lokalnej i centralnej prywatności różnicowej w celu implementacji rozwiązania. Tego typu rozwiązanie działa tylko wtedy, gdy architektura może obsługiwać twarde granice zaufania!

Jak wynika z naszych dotychczasowych rozważań, na projektowanie odpowiedniego systemu inżynierii prywatności składa się wiele części: współpraca z interesariuszami w celu właściwego wyjaśnienia wprowadzonego błędu, wyjaśnienie pytań, na które można lub nie można odpowiedzieć, ustalenie elementów dopasowanych do strategii długoterminowej.

W tym rozdziale zostały omówione sposoby uwzględniania zarządzania i prywatności różnicowej w przepływach pracy. Należy jednak pamiętać, że nie jest to praca dla jednostki. Jak stworzyć podejście zespołowe i zdobyć wsparcie w całej organizacji?

Współpraca z zespołem inżynierii danych i kierownictwem

Jeśli Twoja organizacja jest wystarczająco duża, to prawdopodobnie w jej skład wchodzą oddzielne zespoły zajmujące się nauką o danych i inżynierią danych. Moim zdaniem jest to niefortunna sytuacja, ponieważ te dwa zespoły są ze sobą nierozerwalnie powiązane. Jeśli komunikacja między zespołami inżynierii danych, analizy danych i zarządzania danymi zostanie przerwana, to zostanie również przerwana możliwość odpowiedniego wykorzystania danych. Może to wpływać na decyzje dotyczące

danych i prywatności. Metody ochrony prywatności danych sprawdzają się we wszystkich częściach organizacji pod warunkiem, że kierownictwo oraz cały zespół inżynierii i analizy danych rozumieją standardy i metody, ich wpływ na dane, pracę z danymi oraz ich niezbędność.

Jeśli w Twojej organizacja istnieją twarde rozgraniczenia między działami, które muszą współpracować w sprawach dotyczących danych, to sprawdź, czy możesz być łącznikiem pomiędzy nimi. Odrobina wysiłku w tym zakresie znacznie przyczynia się do usprawnienia komunikacji.

Podziel się odpowiedzialnością

Automatyzacja przepływów pracy związanych z danymi to nie tylko zadanie zespołów inżynierii danych lub zarządzania danymi. Zadaniem osób naukowo zajmujących się danymi jest również właściwe zdefiniowanie istotnych problemów oraz tego, jakie dane są wymagane.

Zadaniem zespołu nauki o danych jest określenie potrzeb i celów biznesowych, które powinny być sprecyzowane na tyle dobrze, by przewidywać powiązane decyzje dotyczące produktów lub funkcjonalności. Wiedza z zakresu nauki o danych powinna być wykorzystywana do przewidywania i wypełniania potencjalnych luk w bieżącym procesie gromadzenia danych i kierowania innymi zespołami zarządzającymi tym procesem oraz oprogramowaniem. Jeśli należysz do zespołu zajmującego się nauką o danych, to jest to również świetny moment na zasugerowanie potencjalnych wymagań dotyczących zachowania prywatności podczas zbierania tych danych.

Jakość, interoperacyjność i standaryzacja danych są częścią zarządzania danymi i powinny być koordynowane w zespołach niezależnie od wielkości organizacji. Aby zrobić to właściwie, organizacje powinny mieć komitet zarządzający lub radę i regularnie aktualizować standardy oraz wytyczne na podstawie opinii społeczności. Twoja nowa wiedza na temat zarządzania danymi i prywatności może być pomocna w tych rozmowach i może działać jako pomost między poszczególnymi częściami organizacji.

Sprawdź w ramach multidyscyplinarnego zespołu, czy standardy i wytyczne są opracowane oraz czy dane spełniają potrzeby wszystkich zainteresowanych stron, regularnie testując i przeglądając dane. Niepożądane jest wystąpienie sytuacji, w której dowiadujesz się, że dodatkowe wymagane pole danych nie było uwzględnione w procesie zbierania danych lub że format innego pola jest nieodpowiedni i nie jest tym, czego potrzebujesz. Upewnij się, że potokom i innym procesom przetwarzania towarzyszą odpowiednie testy, takie jak te zaimplementowane w tym rozdziale, które są regularnie aktualizowane przy użyciu danych wejściowych od wszystkich użytkowników danych.

Wreszcie, w celu zapewnienia lepszej koordynacji i komunikacji w organizacji udokumentuj swoją pracę. Sposoby dokumentowania w przepływach pracy związanych z zapewnianiem prywatności jest tematem kolejnej części.

Tworzenie przepływów pracy uwzględniających dokumentowanie i prywatność

Dobrze udokumentowane systemy są łatwiejsze w użyciu, obsłudze i zarządzaniu. Oznacza to, że jeśli pracujesz już z udostępnianym repozytorium, a zadania są wyraźnie dokumentowane jako część tego repozytorium, to dodanie nowej części dotyczącej prywatności do tej dokumentacji powinno być łatwe!

W rozdziale 1. zostało przedstawionych wiele wskazówek dotyczących prowadzenia dokumentacji. Poniżej znajduje się lista kilku dodatkowych pomysłów, które mogą inspirować do integrowania dokumentacji z normalnym procesem pracy:

- Dodaj dokumentację dotyczącą prywatności do repozytoriów kodu i notatników.
- Opisz w wewnętrznej bazie wiedzy, w jaki sposób można wykorzystać techniki ochrony prywatności.
- Przeprowadź serię wewnętrznych wykładów na tematy, których nauczysz się z tej książki.
- Zachęcaj do publikowania treści i wysyłania żądań do wewnętrznych repozytoriów w celu zarządzania prywatnością.
- Dodaj krok „Uwzględnione środki ochrony prywatności" do definicji ukończenia zespołu (*https://oreil.ly/KJ1dw*).
- Utwórz grupę analityczną i przeanalizuj tę książkę, repozytorium książki i związane z nią badania. Dokumentuj wszystko na bieżąco!

Jeśli masz kilka dobrych przykładów przepływów pracy, w których dodano odpowiednie środki ochrony prywatności, upewnij się, że są one odpowiednio udokumentowane. Następnie połącz się ze współpracownikami lub przeprowadź krótką wewnętrzną rozmowę o tym, jak się to udało. Często ludzie potrzebują tylko kilku dobrych przykładów, aby zacząć, i osoby wskazującej na pojawiające się pytania. Wszystkie te prace wspierają dokumentację dotyczącą zarządzania danymi, a także tworzenie kultury prywatności w organizacji.

Prywatność jako podstawowa propozycja wartości

Twoje życie będzie o wiele łatwiejsze, jeśli prywatność będzie postrzegana przez wielu członków zespołu lub całą firmę jako podstawowa propozycja wartości. Jeśli jest to nowość, to spróbuj się zastanowić, które części produktu lub oferty są powiązane z prywatnością i jaka byłaby wartość dodana integracji prywatności dla organizacji.

Jeśli bezpośrednio zajmujesz się danymi konsumentów, to jest to dość oczywiste. Oferowanie klientom większej gwarancji prywatności to długa droga do zbudowania zaufania i stworzenia marki, którą ludzie będą postrzegać pozytywnie. Publiczne informowanie o swoich wysiłkach nie tylko buduje zaufanie, ale także może przyciągnąć nowe talenty, które chcą pracować w środowisku poważnie traktującym prywatność.

Jeśli pracujesz w organizacji *business-to-business* (B2B) lub tylko z danymi wewnętrznymi, to postawienie na prywatność jest również korzystne. Praca z danymi wewnętrznymi w sposób szanujący prywatność oznacza, że istnieje mniej naruszeń prywatności Twojej i innych pracowników organizacji. Klienci B2B, których danymi zarządzasz, z pewnością docenią dodatkowy wysiłek związany z ochroną prywatności w fazie projektowania i mogą przekazać tę informację swoim klientom. Wreszcie może się okazać, że sama organizacja ceni tematy prywatności jako normalną część codziennej działalności.

Budowanie kultury prywatności, w której pytania poruszone w tej książce stają się naturalną częścią rozmów o danych, to długotrwały proces. Podobnie jak w przypadku budowania kultury mistrzów bezpieczeństwa, znajdą się ludzie, którzy natychmiast zajmą się tematem (jak być może Ty?) i zwrócą na niego uwagę innych.

Społeczność zajmująca się bezpieczeństwem wykonała świetną pracę w celu tworzenia kultury bezpieczeństwa wewnątrz organizacji. Nawiązanie współpracy z tymi ekspertami w organizacji i poznanie, w jaki sposób wbudowują zabezpieczenia w kulturę i wspierają liderów w całej organizacji, będzie bardzo pomocne w ustaleniu, jak zrobić to samo w odniesieniu do prywatności danych. Często okazuje się, że są oni skłonni włączyć propagowanie prywatności do swojej normalnej pracy edukacyjnej.[7]

Ostatecznym celem jest upewnienie się, że wystarczająca liczba osób spojrzy na tematy prywatności z wielu perspektyw. Nie muszą to być wszyscy pracownicy, jednak powinna to być wystarczająca liczba, by zagadnienia te nie zostały potraktowane bardzo powierzchownie oraz by zostały odpowiednio wdrożone.

Podsumowanie

W tym rozdziale przedstawiono sposoby automatyzowania ochrony prywatności w potokach, najpierw analizując wymagania systemu w zakresie ochrony prywatności, a następnie znajdując miejsca do wbudowania prywatności. Gratulacje! Jest to główna część ochrony prywatności w przepływach pracy danych i produktach.

W ten sposób nauczyliśmy się, jak projektować przepływy pracy z myślą o zachowaniu prywatności. Wdrożyliśmy nowe technologie do ochrony prywatności i zarządzania danymi na kilku różnych etapach potoku. Mieliśmy okazję poeksperymentować z biblioteką prywatności różnicowej, którą można zintegrować z systemami opartymi na platformie Spark. Przedstawiono informacje o lokalnej prywatności różnicowej i mechanizmach zbierania danych z prywatnością różnicową. Oceniono również, jak współpracować między zespołami i organizacją, by zapewnić świadomość prywatności wszystkim zespołom odpowiedzialnym za zarządzanie przepływami danych.

W następnym rozdziale omówimy, czym są ataki na prywatność. Dowiemy się, przed czym dokładnie trzeba się chronić, i pogłębimy wiedzę na temat zagrożeń prywatności.

[7] Aby uzyskać dodatkowe wskazówki dotyczące budowania kultury wokół prywatności, dowiadując się więcej o technikach bezpieczeństwa, polecam przeczytać rozdział 15 książki *Agile Application Security* autorstwa Laury Bell i in. (O'Reilly, 2017) (*https://oreil.ly/YNS4v*).

Książka w przystępny sposób przedstawia głęboką perspektywę techniczną wraz z przeglądem najnowszych podejść i architektur technologicznych.

Emily F. Gorcenski, główna analityczka danych, Thoughtworks

Chyba nikogo nie trzeba przekonywać, że ochrona danych i zabezpieczenie prywatności są kwestiami absolutnie kluczowymi w cyfrowym świecie. Na szczęście zdajemy sobie coraz lepiej sprawę, że incydenty naruszeń w dziedzinie bezpieczeństwa danych mogą nas narazić na realne szkody. Z drugiej strony niedopełnienie obowiązków wynikających z RODO okazuje się dla organizacji niezwykle kosztowne, a także naraża na szwank ich wizerunek. Zapewnienie należytej ochrony danych to wymagające wyzwanie. Z tego względu inżynieria prywatności z roku na rok staje się coraz ważniejszą dziedziną.

Tę książkę docenią osoby, które w ramach codziennej pracy integrują tematy związane z prywatnością i bezpieczeństwem danych. To przewodnik dla pragmatyków, zapewniający gruntowną wiedzę o współczesnych elementach ochrony danych, takich jak prywatność różnicowa, uczenie federacyjne i obliczenia szyfrowane. Znajdziesz tu przydatne wskazówki, jak również najlepsze, wielokrotnie sprawdzone praktyki integracji przełomowych technologii, pozwalające skutecznie i na wysokim poziomie dbać o prywatność i bezpieczeństwo danych.

Wreszcie znalazłem książkę, którą mogę polecać wszystkim unikającym tematu prywatności danych!

Vincent Warmerdam, twórca Calm Code, inżynier uczenia maszynowego, Explosion

Najważniejsze zagadnienia:

- Jak przepisy (RODO i CCPA) mają się do przepływów danych i przypadków ich użycia?
- Jak właściwie anonimizować dane?
- Czy szyfrowanie homomorficzne jest właściwym rozwiązaniem?
- Jak wybierać technologie i metody ochrony prywatności?
- Jak zapewnić bezpieczeństwo danych w projektach opartych na ich analizie?
- Jak odpowiednio wdrożyć wewnętrzne zasady ochrony prywatności danych?

Katharine Jarmul jest znaną badaczką, programistką i wykładowczynią. W swojej pracy koncentruje się na zapewnianiu prywatności i bezpieczeństwa w przepływie danych. Z powodzeniem wdraża systemy przetwarzania danych zapewniające wysoki stopień ich prywatności i bezpieczeństwa.

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-0922-9
9 788328 909229
Cena: 79,00 zł