Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Instytut Informatyki

Studia Podyplomowe
Big Data - przetwarzanie i analiza dużych zbiorów danych

PRACA KOŃCOWA

Kamil Gontarz

# Zaprojektowanie i wykonanie systemu do składowania, przetwarzania oraz analizy danych o aktywności użytkowników strony internetowej.

Opiekun pracy
mgr Patryk Pilarski

Warszawa, 2022

## Spis treści

## 1. Kontekst biznesowy

Klientem jest informacyjny portal internetowy, którego głównym źródłem dochodu są reklamy pokazywane użytkownikom. Do tej pory główny reklamodawca rozliczał się z portalem na zasadzie ryczałtu, ustalonej kwoty przelewanej na konto klienta w miesięcznych interwałach. W ostatnim czasie odstąpił od takiego modelu finansowania swoich kampanii reklamowych na rzecz ustalonej kwoty za każde 1000 odsłon stron zawierającyh materiały reklamowe. Przy takim sposobie rozliczania klient stracił znaczną część swojego głównego źródła dochodu. **Głównym zadaniem jest zdefiniowanie co wpływa na poziom ruchu na stronie oraz jakie zadania należy wykonać aby zwiększyć liczbę odsłon**. Jak się okazało klient nie dysponuje żadnym dedykowanym do tego typu zadań działem analitycznym ani działem wyspecjalizowanym działem technicznym. Techniczne kwestie związane z serwowaniem treści zostały oddelegowane do zewnętrznego podmiotu, który nie oferuje usług w zakresie doradztwa analitycznego.

Po konsultacjach z klientem oprócz głównego celu zdefiniowano również cel dodatkowy jakim jest przygotowanie testowego środowiska dla przyszłego działu analitycznego którego zadaniem będzie utrzymanie/zwiększenie efektywności dochodowej protalu. Środowisko te ma pozwalać na:

- przechowywanie surowych danych (obecnie dane ze skryptów na stronie nie są przechowywane a jedynie agregowane ilościowo w kubełkach godzinnych co nie pozwala na inną niż ilościową analizę historycznych danych)
- przetwarzanie danych w celu tworzenia cyklicznych, automatycznie powstających raportów
- prezentację danych w formie tabel, wykresów oraz tablic agregujących różne wyniki (dashboard)
- wykonywanie na surowych danych doraźnych, niestandardowych danych

Z założenia środowisko ma być testowe(aby w pierwszym okresie nie poświecać czasu na uprodukcyjnianie rozwiązania badawczego) ale użyte technologie powinny być:

- łatwo skalowalne - klient zakłada w przyszłości wzrost generowanego ruchu
- powszechnie używane w środowisku - co zapewnia dostęp do specjalistow znających daną technologię oraz potwierdza jej przydatność i możliwość zastosowania w realnych przypadkach
- w miare możliwości open source z rozwiniętą społecznością wokół technologii - podejście to nie generuje kosztów licencyjnych a jednocześnie pozwala na znalezienie rozwiązań wielu problemów w ogólnodostępnych tematycznych forach internetowych
- modułowe - system podzielony na kilka mniejszych modułów jest łatwiejszy w zrozumieniu, utrzymaniu, rozwijaniu czy znajdowaniu błędów
- integrowalne z wieloma rozwiazaniami - brak potrzeby manualnej integracji przy każdorazowej zmianie/dodaniu rozwiązania znacząco przyspiesze rozwój oraz zmniejsza ilość potencjalncyh błędów

- zastępowalne - dzięki identycznym lub podobnym interfejsom w następnych fazach projektu będzie można zastąpić wybrane rozwiązanie innym o bardziej porządanych parametrach czy cechach bez przebudowywania dużej części systemu.

Klient zdaje sobie sprawę, że w pierwszym kroku analizy mogą nie przynieść oczekiwanych efektów jednocześnie jest świadomy tego, że aby przedsiębiorstwo mogło pracować w oparciu o dane należy zaszczepić w nim kulturę pracy z tymi danymi, a do tego niezbędne jest odpowiednie środowisko techniczne.

Dodatkowe wymaganie pochodzi z działu prawnego. IP jest jedną z informacji przysyłanych przez skrypty umieszczone na stronie portalu. Adres IP w pewnych przypadkach może być uznawany za daną osobową[1]

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa ip jako dana osobowa - anonimizacja przy zachowaniu wiedzy na temat ip

[1] https://archiwum.giodo.gov.pl/pl/319/2258

## 2. Dostępne dane

## 2.2 Struktura

## 2.3 Statystyki

## 2.2 Generator

## 3. Rozwiązanie koncepcyjne probelmu biznsowego

- zaproponować co można zmienić aby zrealizować cel główny
- zwiększenie ilości i zaangażowania użytkowników strony
- odsłony
- sesje
- czas
- podział na segmenty technicze (os, deviceType, Browser)
- próba znalezienia odpowiedzi na pytanie jak zwiększyć liczbę użytkowników oraz ich zaangażowanie

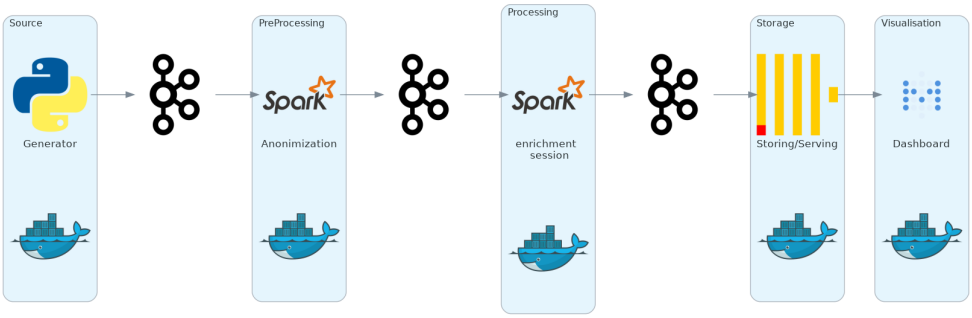
## 4. Model danych

- tabela eventów
- tabela sesji

# 5. Opis modułow

## 5.1 Ogólny schemat rozwiązania

opis




**Rysunek 1:** Przepływ danych między modułami.

dalszy opis

## 5.2 Preprocessing

## 5.3 Processing

## 5.4 Storage

## 5.5 Wizualizacja

## 6. Napotkane problemy oraz ich rozwiązanie

## 7. Wnioski

## 8. Bibliografia


- https://altinity.com/blog/2020/5/21/clickhouse-kafka-engine-tutorial – virtual columns
- https://hub.docker.com/r/clickhouse/clickhouse-server/ - clickhouse docker
- https://github.com/enqueue/metabase-clickhouse-driver – metabase clickhouse community driver
- https://clickhouse.com/docs/en/integrations/kafka/kafka-table-engine/
- https://www.metabase.com/docs/latest/operations-guide/running-metabase-on-docker.html
- https://www.metabase.com/docs/latest/administration-guide/01-managing-databases.html#database-sync-and-analysis
- https://vincent.doba.fr/posts/20211004_spark_data_description_language_for_defining_spark_schema/
- https://towardsdatascience.com/spark-3-2-session-windowing-feature-for-streaming-data-e404d92e267
- https://pypi.org/project/user-agents/
- https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/96810098
- https://stackoverflow.com/questions/70991571/stream-data-from-one-kafka-topic-to-another-using-pyspark
- https://stackoverflow.com/questions/2013124/regex-matching-up-to-the-first-occurrence-of-a-character
- https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#window-operations-on-event-time
- https://github.com/ykursadkaya/pyspark-Docker/blob/master/Dockerfile
- https://stackoverflow.com/questions/37132559/add-jar-files-to-a-spark-job-spark-submit
- https://stackoverflow.com/questions/50217212/how-do-i-write-to-kafka-using-pyspark
- https://faker.readthedocs.io/en/master/
- https://github.com/joke2k/faker
- https://docs.docker.com/develop/develop-images/dockerfile_best-practices/ - multiline env
- https://support.google.com/analytics/answer/2731565?hl=en#overview&zippy=%2Cin-this-article – definition of visits