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. Cel biznesowy

- zwiększenie ilości i zaangażowania użytkowników strony

W wyniku pracy należy: - zaproponować co można zmienić aby zrealizować cel główny - przygotować testowe środowisko: – przechowujące dane – przetwarajace dane – prezentujące dane - środowisko na pierwszym etapie może być testowe ale użyte technologie powinny być łatwo skalowalne, powszechnie używane w środowisku, z rozwinietą społeczności, w miare możliwości open source - modułowość - integracja - zastępowalność - jeżeli w pierwszym kroku analiza nie przyniesie efektów to ważne jest aby gotowa była koncepcja systemu przechowującego dane który może posłużyć do dalszych analiz teraz i w przyszłości - ip jako dana osobowa - anonimizacja przy zachowaniu wiedzy na temat ip

2. Dostępne dane

2.2 Struktura

2.3 Statystyki

2.2 Generator

3. Rozwiązanie koncepcyjne probelmu biznsowego

- 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

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/clickh
server/ - clickhouse docker https://github.com/enqueue/metabase-clickhouse-driver – meta-
base 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_languag
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/4
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/Dockerfi
https://stackoverflow.com/questions/37132559/add-jar-files-to-a-spark-job-spark-submit https://stackoverflow.com/qu
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 visit