Link do logów:

<https://raw.githubusercontent.com/elastic/examples/master/Common%20Data%20Formats/apache_logs/apache_logs>

**1. Dokonaj zgrubnej ceny danych zawartych w pliku access.log Jakiego rodzaju informacje możesz odczytać z zapisu poszczególnych zapytań do serwera?**

Logi z pliku access.log zawierają informacje o żądaniach, jakie otrzymał serwer WWW, wraz z dodatkowymi danymi takimi jak adres IP klienta, datą i czasem żądania, metodą HTTP (np. GET), żądanym zasobem (ścieżką URL), kodem odpowiedzi serwera, rozmiarem odpowiedzi oraz informacjami o refererze (stronie, z której pochodził link do żądanego zasobu) i user-agent (informacje o przeglądarce i systemie operacyjnym klienta).

Na przykład, pierwsza linijka:

83.149.9.216 - - [17/May/2015:10:05:03 +0000] "GET /presentations/logstash-monitorama-2013/images/kibana-search.png HTTP/1.1" 200 203023 "http://semicomplete.com/presentations/logstash-monitorama-2013/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_9\_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.77 Safari/537.36"

oznacza:

Adres IP klienta: 83.149.9.216

Data i czas: 17/May/2015:10:05:03 +0000

Metoda HTTP: GET

Żądany zasób: /presentations/logstash-monitorama-2013/images/kibana-search.png

Kod odpowiedzi serwera: 200 (OK)

Rozmiar odpowiedzi: 203023 bajty

Referer: http://semicomplete.com/presentations/logstash-monitorama-2013/

User-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_9\_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.77 Safari/537.36

**2. Sposób identyfikacji sesji:**

Sesje zostały zidentyfikowane na podstawie interwałów czasowych między żądaniami użytkownika.

**Procedura/algorytm:**

* Podział logów na adresy IP i znaczniki czasowe.
* Sortowanie czasów dla każdego adresu IP.
* Podział czasów na sesje z uwzględnieniem 15-minutowego interwału.

**3. Czas identyfikacji sesji:**

Czas identyfikacji sesji opiera się na 15-minutowym interwale między kolejnymi żądaniami, co rozdziela sesje.

**4. Długość sesji:**

Sesje zostały ograniczone do maksymalnie 15 minut, co pozwala na lepsze śledzenie aktywności użytkownika. Krótsze sesje są preferowane.

**5. Moment zakończenia sesji:**

Sesje kończą się, gdy czas między żądaniami przekracza 15 minut.

**5. Traktowanie pojedynczych żądań:**

Pojedyncze żądanie bez aktywności przez ponad 15 minut jest traktowane jako osobna sesja.

**6. Uzyskano dwa częste zbiory:**

Zbiór ("Mozilla/5.0) otrzymał wsparcie (support) wynoszące około 65%.

Zbiór ("msnbot/2.0b) otrzymał wsparcie (support) wynoszące około 11%.

Nie zostały znalezione żadne reguły asocjacyjne, co wskazuje na brak istotnych zależności między danymi elementami.

**7. Co możesz powiedzieć o tych regułach w kontekście budowy serwisu oraz użytkowania serwisu?**

Brak znalezionych reguł asocjacyjnych sugeruje brak silnych powiązań między analizowanymi danymi (user-agent string w tym przypadku) w logach. Może to oznaczać, że analizowane dane mogą być losowe lub nieposortowane, co z kolei może wynikać z różnorodności urządzeń, przeglądarek lub sposobów dostępu do serwisu.

W przypadku budowy serwisu może to wskazywać na brak specyficznych trendów lub zachowań użytkowników, które można byłoby zidentyfikować na podstawie analizy danych logów.

**8. Co możesz powiedzieć na temat topologii witryny na podstawie otrzymanej analizy asocjacyjnej?**

Z braku wykrytych reguł asocjacyjnych nie można wyciągnąć konkretnych wniosków na temat topologii witryny.

Brak znalezionych zależności sugeruje, że elementy analizowane (np. user-agent strings) nie mają silnych współzależności ani powiązań między sobą, co uniemożliwia wywnioskowanie o strukturze czy topologii witryny na podstawie tych danych.

9. Roboty internetowe

W logach rejestrowane były wpisy charakteryzujące się obecnością agentów internetowych (botów). Proces identyfikacji opierał się na analizie nagłówka User-Agent w logach dostępu do serwera WWW. Spostrzeżono wpisy, w których nagłówek User-Agent zawierał informacje identyfikujące typowe roboty internetowe, takie jak Googlebot, Baiduspider, Sogou web spider, archive.org\_bot, MJ12bot, YandexBot, Mail.RU\_Bot, Twitterbot, msnbot, Exabot, MojeekBot oraz Majestic

10. 1498 roznych żądań

11. Najczęstsze adresy IP:

66.249.73.135: 480 żądań

46.105.14.53: 364 żądań

130.237.218.86: 356 żądań

75.97.9.59: 269 żądań

50.16.19.13: 113 żądań

12. Najczęściej odwiedzane strony:

- [18/May/2015:04:05:55 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 4 odwiedzeń**

- [18/May/2015:03:05:11 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "http://www.semicomplete.com/blog/tags/puppet?flav=rss20" "Tiny Tiny RSS/1.11 (http://tt-rss.org/)"

**: 2 odwiedzeń**

- [17/May/2015:13:05:11 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 2 odwiedzeń**

- [18/May/2015:03:05:23 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 2 odwiedzeń**

- [18/May/2015:04:05:27 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

: **2 odwiedzeń**

**13.**