



Wydział Matematyki i Nauk Informatycznych

POLITECHNIKA WARSZAWSKA

Raport projektu

Składowanie danych w systemach Big Data

Autorzy:

Konrad Komisarczyk, Mikołaj Malec, Patryk Wrona

Prowadzący laboratorium:

mgr inż. Michał Możdżonek

Warszawa
24.01.2022

Spis treści

1. Cel projektu	2
2. Opis wykorzystanych zbiorów danych	3
2.1. Wayback Machine	3
2.2. Dodatkowe dane	3
2.2.1. Ranking Tranco	3
2.2.2. Lista znaczników HTML	3
3. Architektura systemu i opis narzędzi	4
4. Pozyskiwanie, przetwarzanie i składowanie danych źródłowych	10
4.1. Pozyskiwanie danych	10
4.2. Przetwarzanie danych	10
4.3. Przechowywane dane	10
5. Analiza danych i widoki wsadowe	12
5.1. Analiza danych	12
5.1.1. Strony z rzadkim rodzajem taga	12
5.1.2. Średnia długość strony internetowej	13
5.1.3. Średnie liczby występowania tagów	14
5.1.4. Najdłuższe strony internetowe	20
5.2. Widoki wsadowe	21
6. Testy funkcjonalne komponentów	22
6.1. Przepływ danych w Nifi	22
6.1.1. Pliki przed zapytaniem do serwera	22
6.1.2. Wynik zapytania od serwera	24
6.1.3. Niepoprawne wczytane linki	24
6.1.4. Wynik analizy	24
6.1.5. Plik zawierający dane wrzucane do HDFS	30
6.1.6. Plik z zgrupowanymi rekordami do HDFS	30
6.2. Analiza stron	30
6.3. Zawartość baz danych	30
6.3.1. HBase	30
6.3.2. Hadoop	34
7. Podsumowanie	35
8. Podział pracy w grupie	36

1. Cel projektu

Celem projektu była analiza zmian stron internetowych w czasie na podstawie danych z Wayback Machine.

Wayback Machine jest archiwum internetowym należącym do amerykańskiej organizacji non-profit Internet Archive. Celem archiwum jest zapobieganie bezpowrotnej utracie danych w sytuacji, gdy treść stron internetowych się zmienia, czy strony są zamykane. Projekt pozwala użytkownikom na przeglądanie zarchiwizowanych wersji witryn. Dane z Wayback Machine były już w przeszłości obiektem wielu badań naukowych.

Inspiracją do powstania projektu była udostępniana przez Wayback Machine funkcjonalność "Summary" pozwalająca na wyświetlenie krótkiego podsumowania dla danej strony internetowej i wybranego okresu w czasie zliczającego poszczególne typy MIME treści udostępnianych przez witrynę.

Celem projektu było przygotować podobne podsumowania, ale szczegółowo skupiające się na kodzie HTML stron internetowych - w szczególności liczbie wystąpień poszczególnych znaczników i długości kodu.

W celu prezentacji wyników działania systemu przeanalizowano 1500 najpopularniejszych stron internetowych według rankingu Tranco w okresie ostatnich 20 lat (od 2002 do 2022 roku). Rozwiązanie bardzo łatwo zmodyfikować tak, aby analizować inny okres czasu, czy inny zbiór stron internetowych.

2. Opis wykorzystanych zbiorów danych

2.1. Wayback Machine

Dane pochodzące z Wayback Machine to pliki HTML zawierające kody ustalonych stron internetowych w ustalonych momentach czasu. Poza kodem oryginalnej strony, w pliku znajdują się wstawki dodane przez archiwum - w początkowej części pliku kod HTML dodający do strony toolbar Wayback Machine wyświetlany na górze strony przy przeglądaniu strony w przeglądarce, a pod koniec pliku komentarze zawierające dodatkowe informacje o zrzucie strony, m.in. kiedy zrzut był wykonany, czy informacje o prawie autorskim.

Zrzuty stron wykonywane są w różnych odstępach czasowych. Odstępy mogą się zmieniać i zależeć od archiwizowanej strony. Wayback Machine nie udostępnia funkcji API pozwalających otrzymać wygodne do przetworzenia informacje o datach zrzutów danej strony.

Zrzuty stron opisane wcześniej pobieramy z Wayback Machine wykonując zapytanie HTTP GET o odpowiedni adres URL zawierający adres strony, której zrzut chcemy pobrać oraz odpowiednio zakodowaną datę, z kiedy chcemy pobrać zrzut - zwracany jest zrzut o dacie najbliższej podanej spośród dat wykonanych zrzutów.

Projekt ogranicza się do pobierania jednego zrzutu strony na rok. Dla każdego roku zostało wysłane zapytanie o zrzut z 1 lipca o godzinie 1:01. Niekiedy strona archiwizowana była rzadziej niż raz na rok i zapytania o różne lata zwracają ten sam wynik, co jest dalej odpowiednio obsługiwane - duplikaty są pomijane.

Na potrzeby prezentacji wyników dokonano $1500 \cdot 20 = 30000$ zapytań do Wayback Machine, pobierając ostatecznie 13676 unikalnych stron internetowych.

2.2. Dodatkowe dane

2.2.1. Ranking Tranco

Tranco jest projektem, którego celem jest udostępnianie aktualnego, odpornego na manipulacje rankingu najpopularniejszych stron internetowych do wykorzystania w badaniach. Projekt łączy różne źródła danych, m. in. ranking Alexa Internet Top 1 Million, czy ranking oparty na danych z serwerów DNS firmy Cisco. Oficjalna strona projektu znajduje się pod adresem tranco-list.eu.

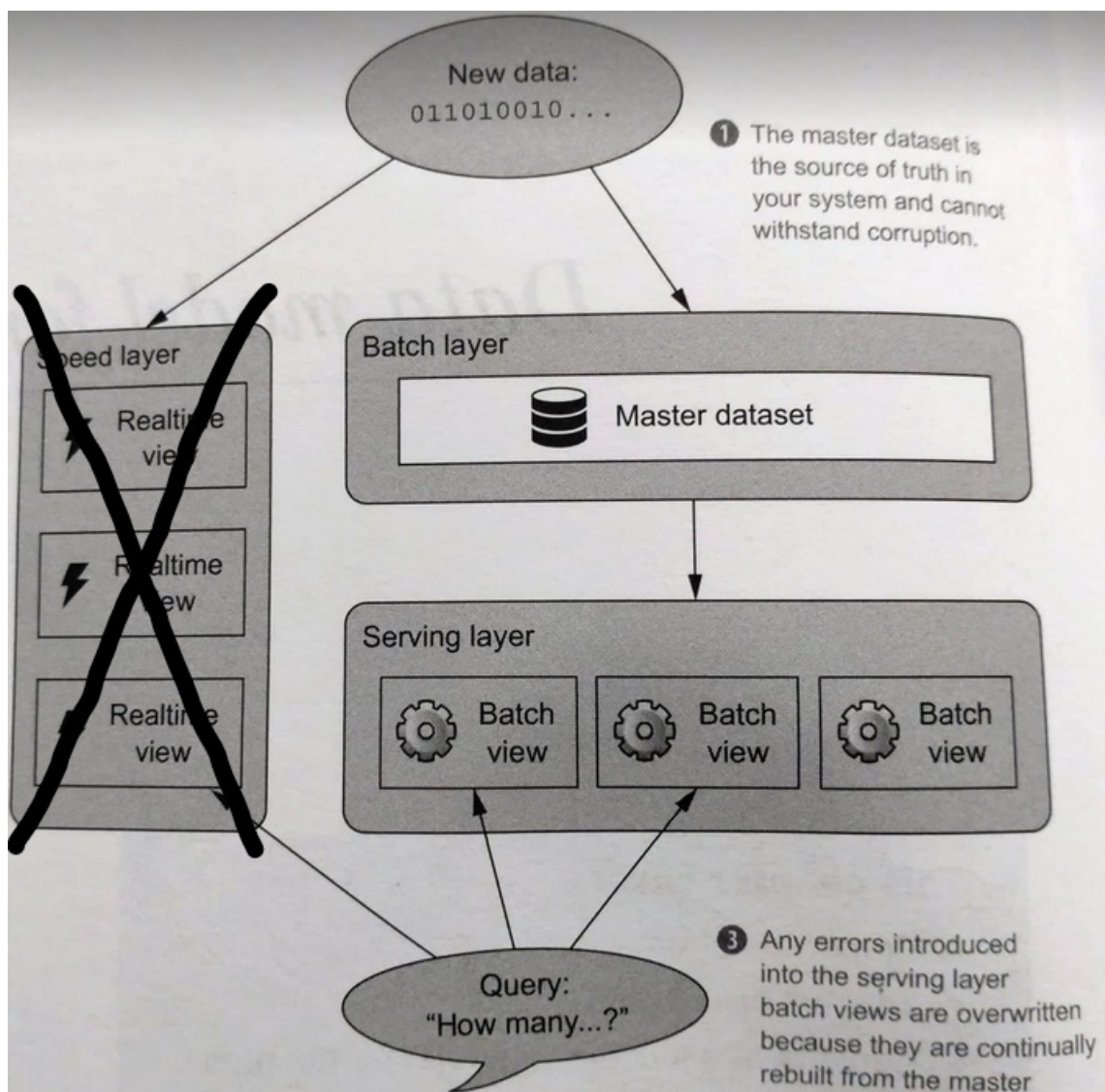
Wykorzystujemy listę 1 miliona najpopularniejszych witryn pobraną ze strony tranco w dniu 20 stycznia 2022. Lista jest plikiem w formacie CSV, którego druga kolumna to adresy WWW stron internetowych. Lista zapisana jest w pliku `webpages_list.csv` znajdującym się w podkatalogu `scripts/data` rozwiązania.

2.2.2. Lista znaczników HTML

Lista zapisana jest w pliku `html_tags.txt` znajdującym się w podkatalogu `scripts/data` rozwiązania. Dane pochodzą ze strony <https://www.w3schools.com/TAGS/default.ASP>, ręcznie zostały przetworzone do formatu `.txt` na potrzeby niniejszego projektu.

3. Architektura systemu i opis narzędzi

Architektura zastosowana w naszym rozwiązaniu jest podobna do architektury lambda, przedstawionej na rysunku 3.1:

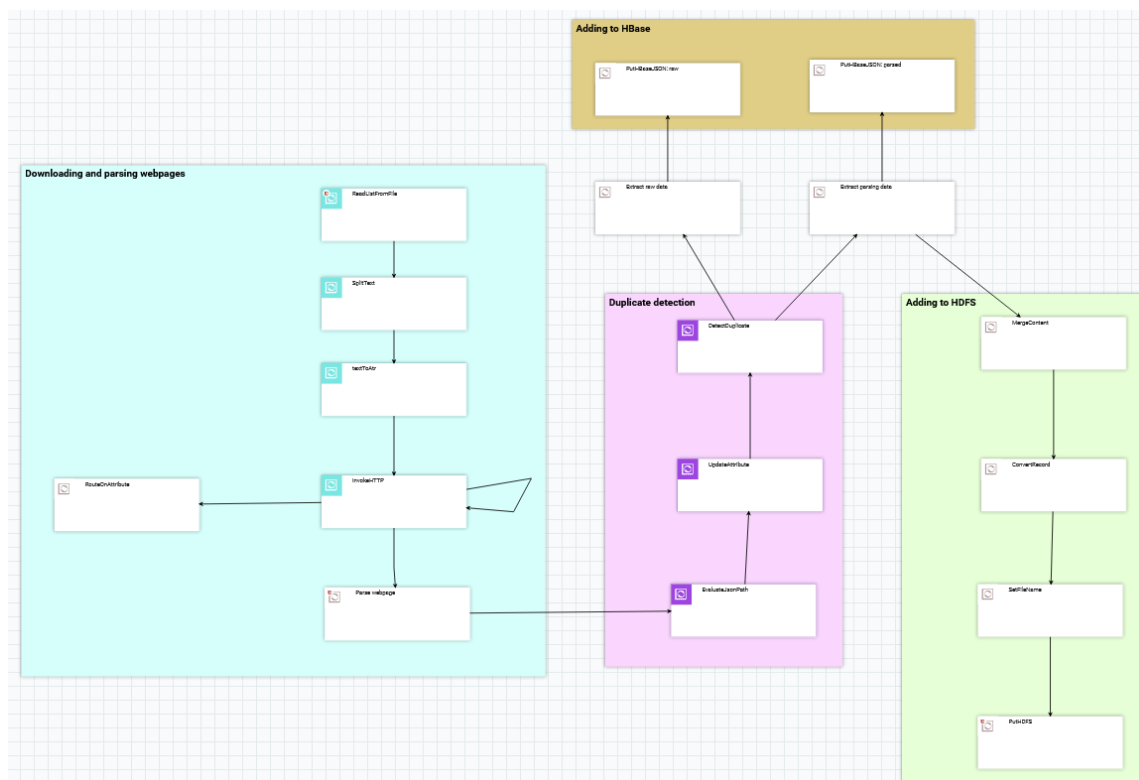


Rys. 3.1. Architektura lambda będąca naszym rozwiązaniem - pominięto speed layer - zdjęcie z Nathan Marz 'Principles and best practices of scalable real-time data systems.

Na początku, dane pobierane są ze źródeł opisanych w rozdziale 2. Za pomocą Apache NiFi dane są odpowiednio przekształcane oraz gwarantowana jest ich jakość. Ich surowa wersja zapisywana jest jako master data set do tabeli *webpages*, jako rodzina kolumn *raw* bazy HBase będącej de facto **Batch layer** - dzięki temu strony są udostępniane w szybki sposób użytkownikowi (random read) i nie musi on czekać na odpowiedź ze strony Wayback Machine. Dodatkowo, poprzez Apache NiFi, dla każdej strony zliczana jest liczba wystąpień danych tagów html, dłu-

gość strony, oraz wyciągany jest rok archiwizacji oraz link do strony. Tego typu agregowane dane są przechowywane w **Serving layer** po pierwsze w tabeli *webpages*, jako rodzina kolumn *parsed* bazy HBase, po drugie w rozproszonym systemie plików HDFS w katalogu */user/bigdata/data* jako pliki w formacie Parquet. Następnie wykonywane analizy danych zostały wykonane dzięki Apache Spark z wykorzystaniem składowanych plików z HDFS.

W celu lepszego wyobrażenia działania architektury, na rysunku 3.2 przedstawiono cały przepływ danych z użyciem Apache NiFi.



Rys. 3.2. NiFi flow - cały system

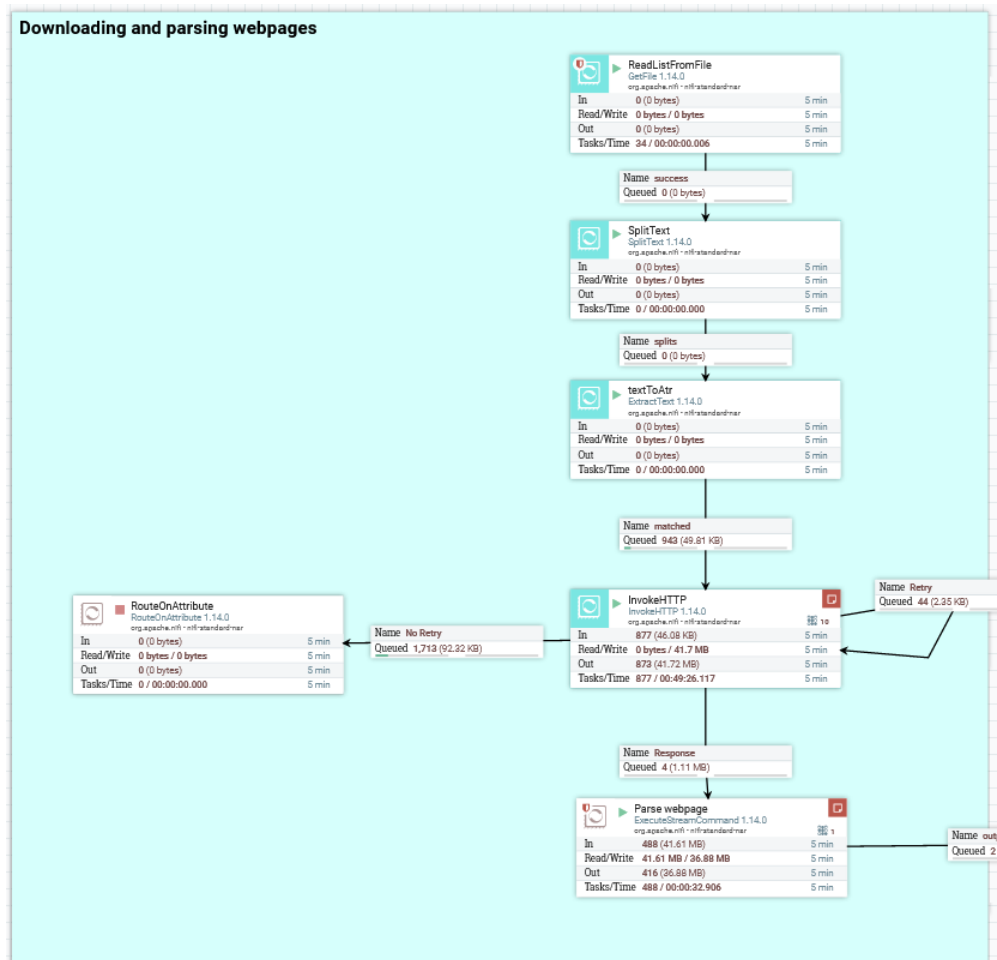
Dodatkowo na następnych rysunkach przedstawiono przybliżenia odpowiednich części przepływu.

Część 1 - na początku pobierana jest lista stron ze statycznego pliku. Proces *InvokeHTTP* wysyła równolegle 32 zapytań (32 wątków) do Wayback Machine i otrzymuje stronę html. W przypadku zwrócenia błędu przez stronę, albo zapytanie jest powtarzane (w przypadku części odpowiedzi), albo informacja o błędnej odpowiedzi są dodawane do kolejki i przechowywane w pamięci NiFi. Następnie w procesie *Parse webpage* strony są przetwarzane (zliczana jest liczba wystąpień tagów) za pomocą skryptu python. Dalej przekazywany jest plik w postaci json zawierający 2 elementy - surową stronę i przetworzoną.

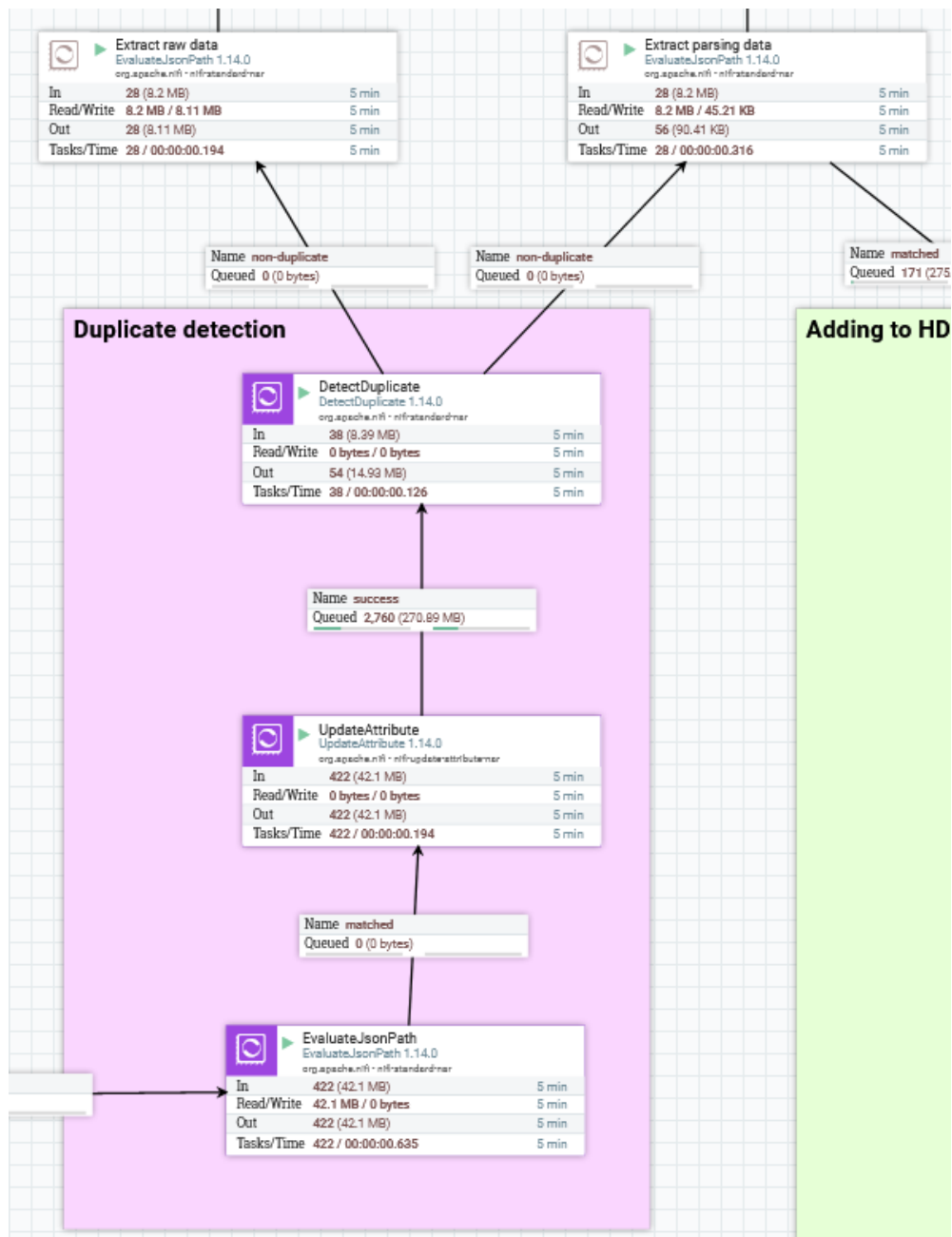
Część 2 - odpowiednio w ramach kolejnych procesów rozdzielany jest plik json oraz rekordy sprawdzane są z pomocniczą tabelą HBase *duplicates_cache* w celu eliminacji powtarzających się rekordów.

Część 3 - dane surowe (raw) oraz przetworzone (parsed) umieszczane są w bazie HBase.

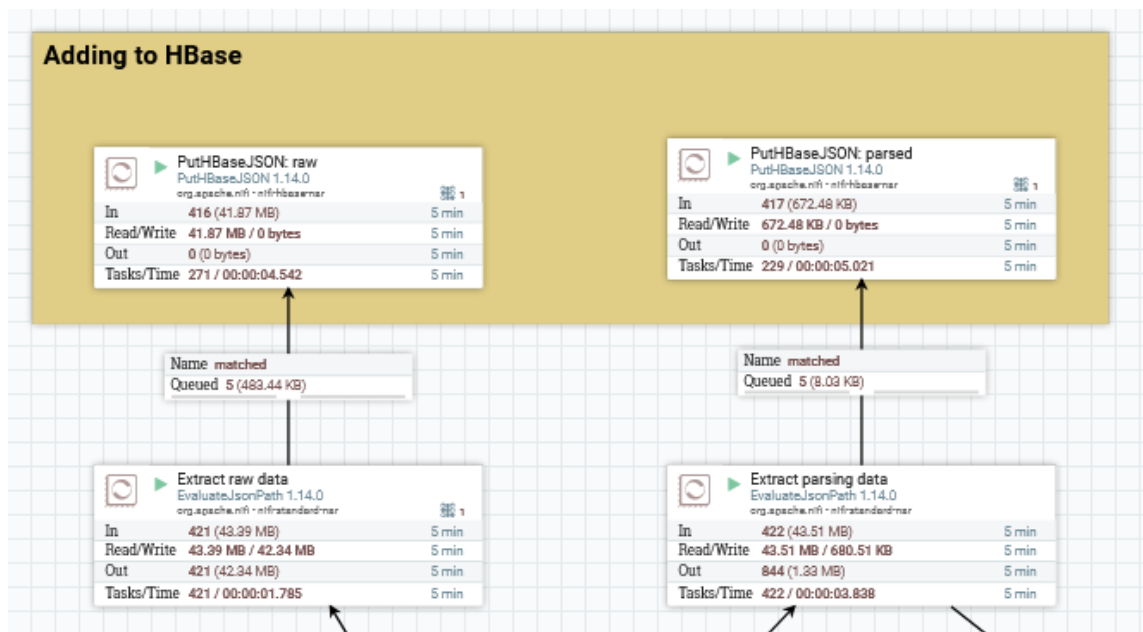
Część 4 - dane przetworzone kolejkowane są do momentu aż zbierze się 1000 stron. Następnie są one łączone w 1 plik w *MergeContent*, którego format zmieniany jest na format Parquet w ramach *ConvertRecord*. Dodatkowo zmieniana jest nazwa pliku na umożliwiający zapis w HDFS (bez niedozwolonych znaków) - nowa nazwa jest generowana w sposób losowy w procesie *SetFileName*. Ostatecznie proces *PutHDFS* dodaje plik w formacie Parquet do HDFS do katalogu */user/bigdata/data/*



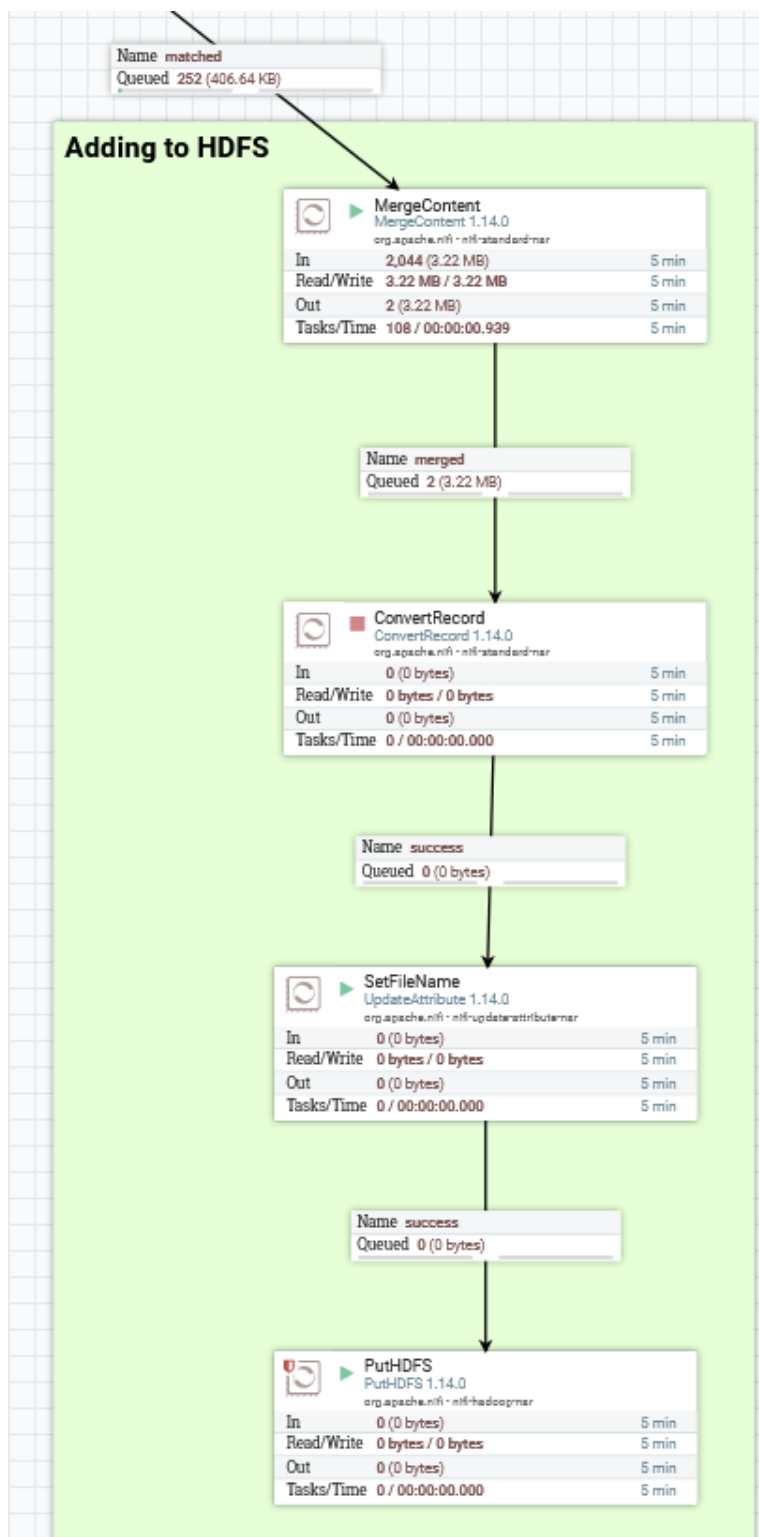
Rys. 3.3. NiFi flow 1 - pobieranie i parsowanie stron internetowych



Rys. 3.4. NiFi flow 2 - detekcja duplikatów



Rys. 3.5. NiFi flow 3 - dodawanie surowych i przetworzonych danych do HBase



Rys. 3.6. NiFi flow 4 - akumulowanie 1000 rekordów przetworzonych stron i zapis plików w formacie Parquet do HDFS

4. Pozyskiwanie, przetwarzanie i składowanie danych źródłowych

4.1. Pozyskiwanie danych

Zrzuty stron z Wayback Machine pobierane są przez procesor `InvokeHTTP` w Apache NiFi. Procesor wykonuje zapytania HTTP GET pod dostarczony mu w przepływie danych odpowiedni adres URL zawierający pożądaną datę zrzutu oraz adres archiwizowanej strony.

4.2. Przetwarzanie danych

Kody HTML pobranych zrzutów przekazywane są na wejście skryptu `parse.py`. Skrypt przetwarza stronę wyznaczając z niej informacje używane dalej do analiz. Wynikiem skryptu jest plik w formacie JSON zawierający informacje o zrzucie strony:

- surowy kod strony `raw` - słownik składający się z par klucz, wartość
 - `link` adres URL wykorzystany do pobrania zrzutu z Wayback Machine
 - `content` surowy kod HTML strony
- informacje przeznaczone do analiz `parsed` - słownik składający się z par klucz, wartość
 - `link` adres URL wykorzystany do pobrania zrzutu z Wayback Machine
 - `year`: rok, w którym została zarchiwizowana strona
 - `webpageLink`: adres archiwizowanej strony internetowej
 - `length`: długość kodu HTML strony, z pominięciem wstawek dodanych przez Wayback Machine
 - dla każdego tagu z listy `html_tags.txt` para tag: liczba wystąpień tagu w kodzie strony, z pominięciem wstawek dodanych przez Wayback Machine

Dalej dane surowe i przeznaczone do analiz są rozdzielane, a następnie dane przeznaczone do analiz łączone są w pliki zawierające co najmniej 1000 rekordów oraz konwertowane do formatu Parquet. Ten format plików został wybrany ze względu na liczne agregacje danych po kolumnach w przypadku analizy w Apache Spark, oraz relatywnie rzadkie dodawanie nowych wierszy - jako pliki wsadowe z co najmniej 1000 obserwacjami.

4.3. Przechowywane dane

System przechowuje dane w dwóch platformach: Apache HBase oraz Hadoop HDFS.

Apache HBase

W bazie danych HBase system wykorzystuje dwie tabele: `'webpages'` oraz `'duplicates_cache'`.

Tabela `'duplicates_cache'` służy do szybkiego sprawdzania czy dana strona została już zarchiwizowana przez system. Zawiera jedną rodzinę kolumn `'f'` z tylko jedną kolumną `'q'`. Klucz stanowi złączona para (rok, adres strony internetowej).

Tabela `'webpages'` służy do przechowywania danych wszystkich analizowanych stron. Składają się na nią dwie rodziny kolumn `'raw'` oraz `'parsed'`. W rodzinie `'raw'` mamy jedną kolumnę

zawierającą cały kod html strony. W rodzinie 'parsed' mamy kolumnę odpowiadającą długości kodu strony, jej adresowi, temu, z którego roku pochodzi kod strony oraz, dla każdego znacznika HTML, kolumnę odpowiadającą ilości jego wystąpień w kodzie strony. Jako klucz wykorzystujemy adres używany jako zapytanie do Wayback Machine.

Apache Hadoop

System Hadoop wykorzystujemy do przechowywania danych przygotowanych do przeprowadzenia analiz. W systemie plików dane przechowujemy jako pliki w formacie Parquet zebrane w jednym folderze. Pliki to ramki danych zawierające między 1000, a 5000 rekordów. Pliki, podobnie jak w rodzinie 'parsed' tabeli 'webpages' w bazie HBase, zawierają kolumny odpowiadające długości kodu strony, jej adresowi, temu, z którego roku pochodzi kod strony oraz, dla każdego znacznika HTML, kolumnę odpowiadającą ilości jego wystąpień w kodzie strony, a także adres używany jako zapytanie do Wayback Machine.

5. Analiza danych i widoki wsadowe

Celem analizy danych jest pozyskanie z danych wartościowych z perspektywy biznesowej informacji. Analiza danych i tworzenie widoków wsadowych zostały wykonane z użyciem Apache Spark, używając do tego języka Python (wykorzystując de facto interfejs Pythona PySpark). Plik z analizą jest dostępny jako notatnik Jupyterowy `/bigdata/pyspark-analysis/analysis-ia.ipynb` na repozytorium github.

5.1. Analiza danych

Do analiz wykorzystano wszystkie pliki w formacie Parquet składowane w katalogu HDFS `user/bigdata/data/`.

5.1.1. Strony z rzadkim rodzajem taga

Na początku odfiltrowano te strony w których występują rzadkie tagi - analizę wykonano dla tagu `var`. Wynikowane strony zawierające dany tag:

```
webpages.filter(col("__var__") > 0).select("year", "webpageLink").show()
```

```
+-----+-----+
|      year| webpageLink|
+-----+-----+
|{2015, null}|express.co.uk|
|{2016, null}|express.co.uk|
|{2013, null}|express.co.uk|
|{2014, null}|express.co.uk|
|{2017, null}|express.co.uk|
|{2018, null}|express.co.uk|
|{2019, null}|express.co.uk|
|{2021, null}|express.co.uk|
|{2020, null}|express.co.uk|
|{2011, null}| mashable.com|
|{2012, null}| mashable.com|
+-----+-----+
```

Rys. 5.1. DataFrame zawierający przykładowe strony mające wystąpienia taga `var` w swoim kodzie

Następnie sprawdziliśmy stronę `express.co.uk` i rzeczywiście posiada ona wystąpienia taga `var`:



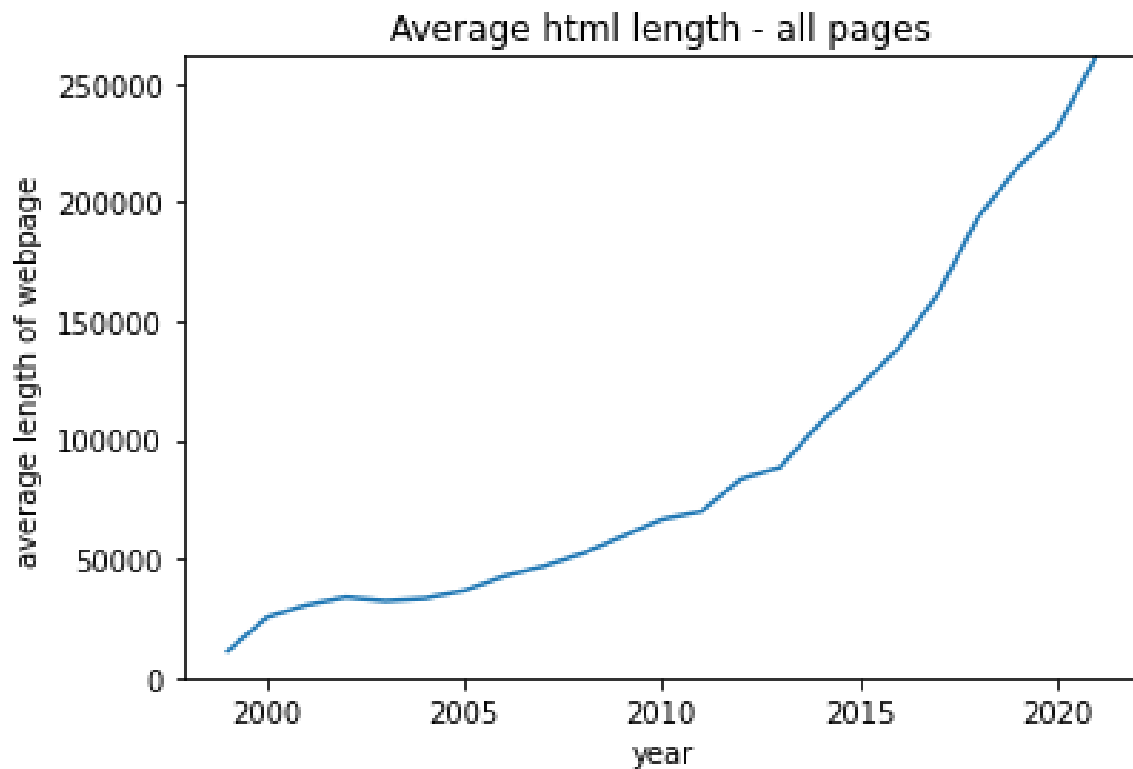
Rys. 5.2. Html strony express.co.uk - wystąpienie taga var - 1



Rys. 5.3. Html strony express.co.uk - wystapienie taga var - 2

5.1.2. Średnia długość strony internetowej

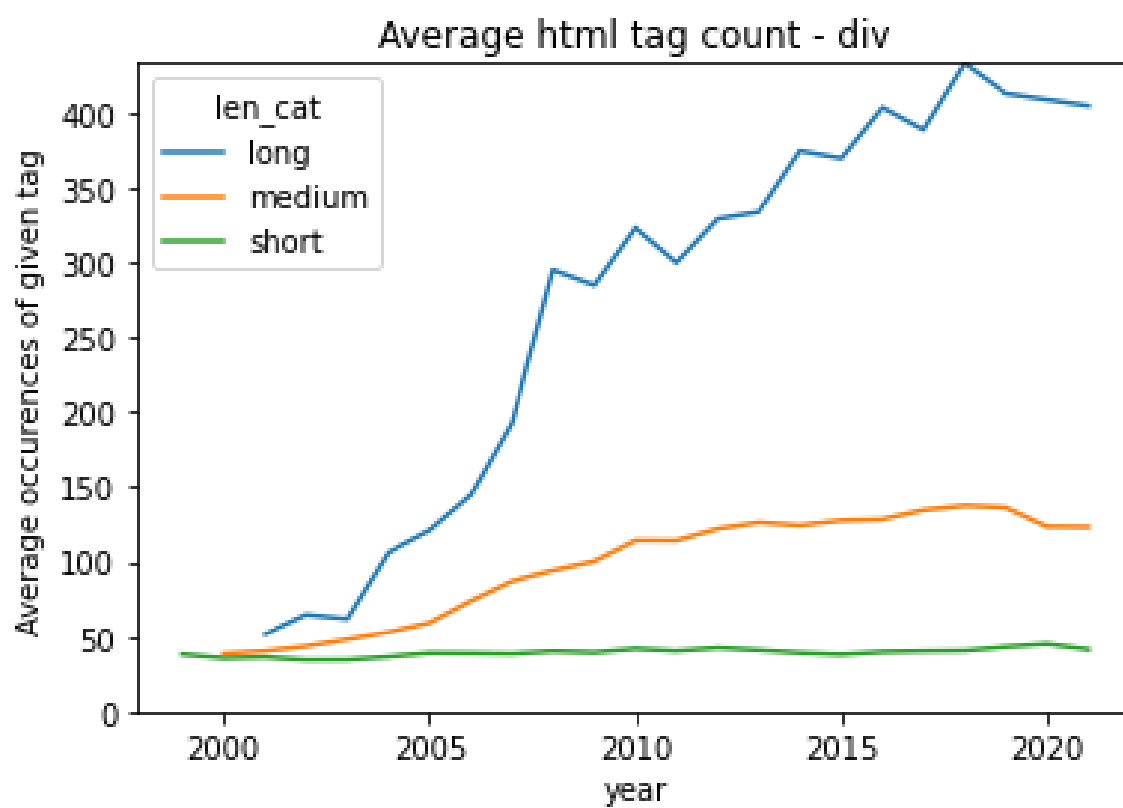
Obliczono i przedstawiono na następnym wykresie średnią długość strony internetowej w zależności od roku:



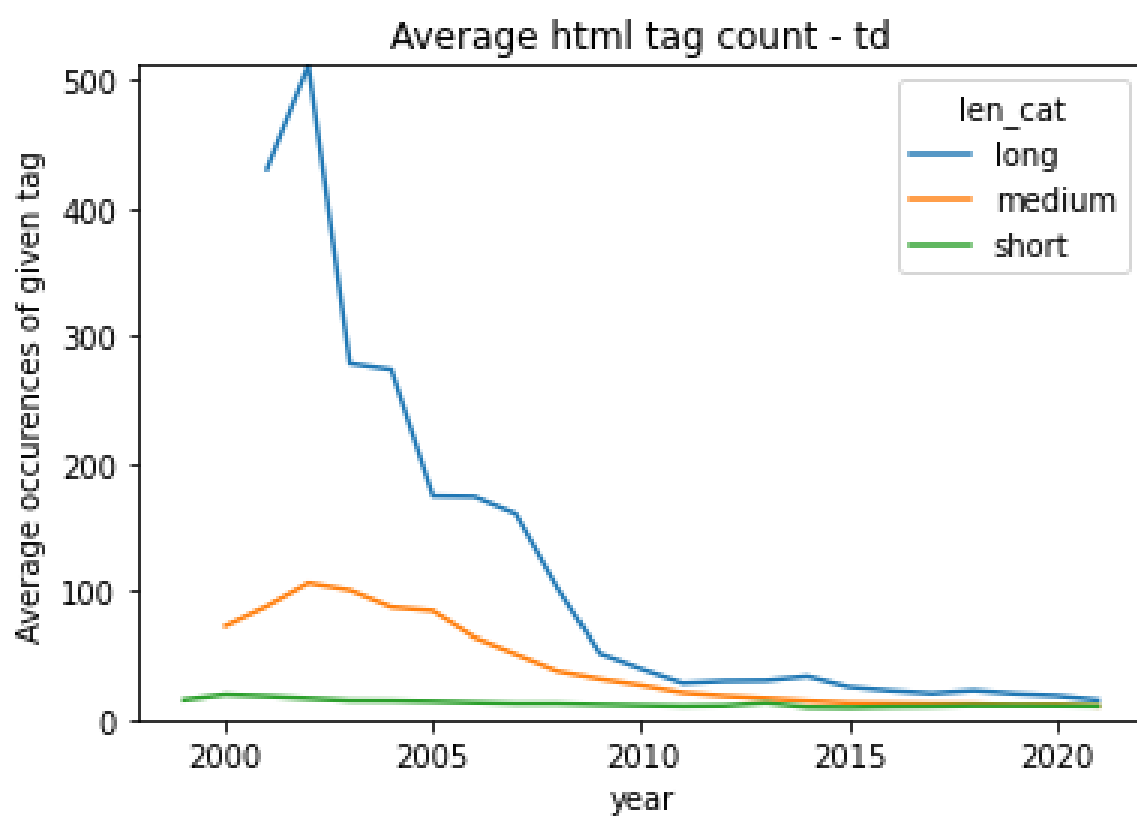
Rys. 5.4. Średnia długość strony internetowej w funkcji roku

5.1.3. Średnie liczby występowania tagów

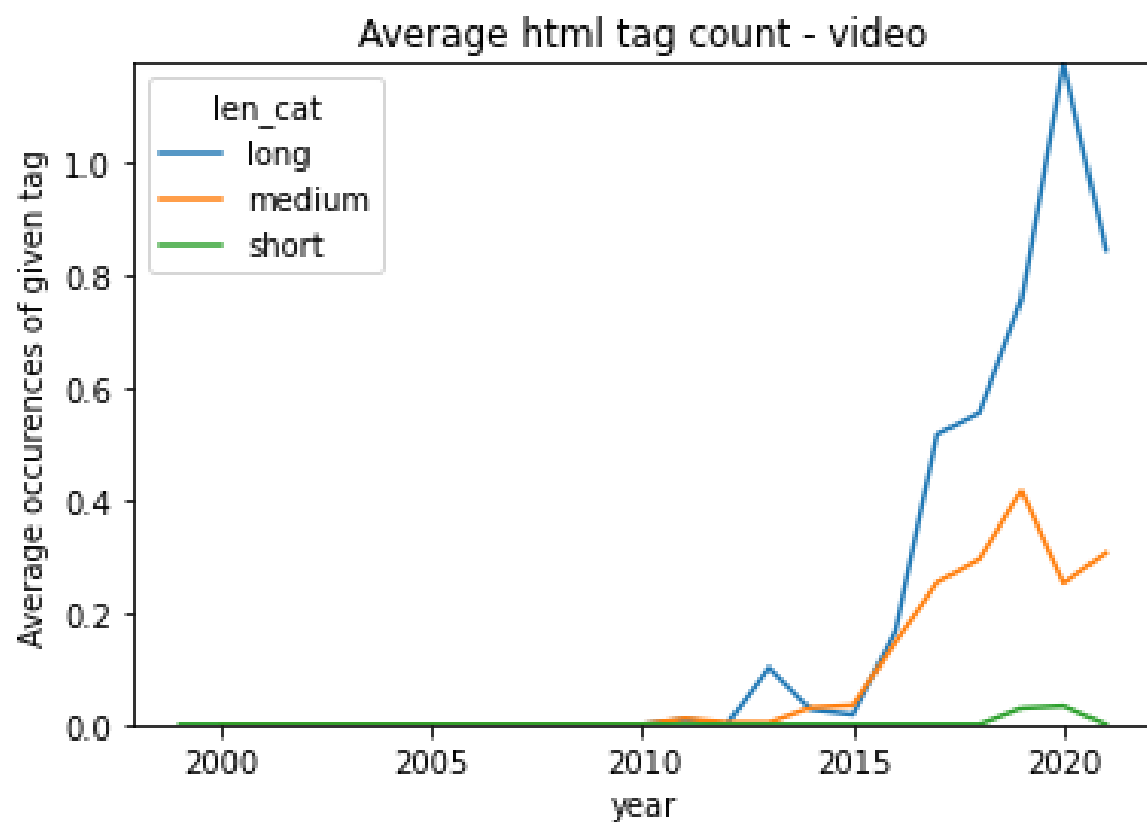
Bardzo ważną częścią analizy było wyznaczenie trendów średniej liczby wystąpień danego taga html w czasie. Podzielono strony na kategorie *len_cat* określające czy dana strona jest krótka (mniej niż 20 tysięcy znaków w pliku .html), długa (więcej niż 80 tysięcy znaków), czy też średniej długości. Przykładowe analizy przedstawiamy dla tagów *div*, *td*, *video*, *a*, *br*, *small*:



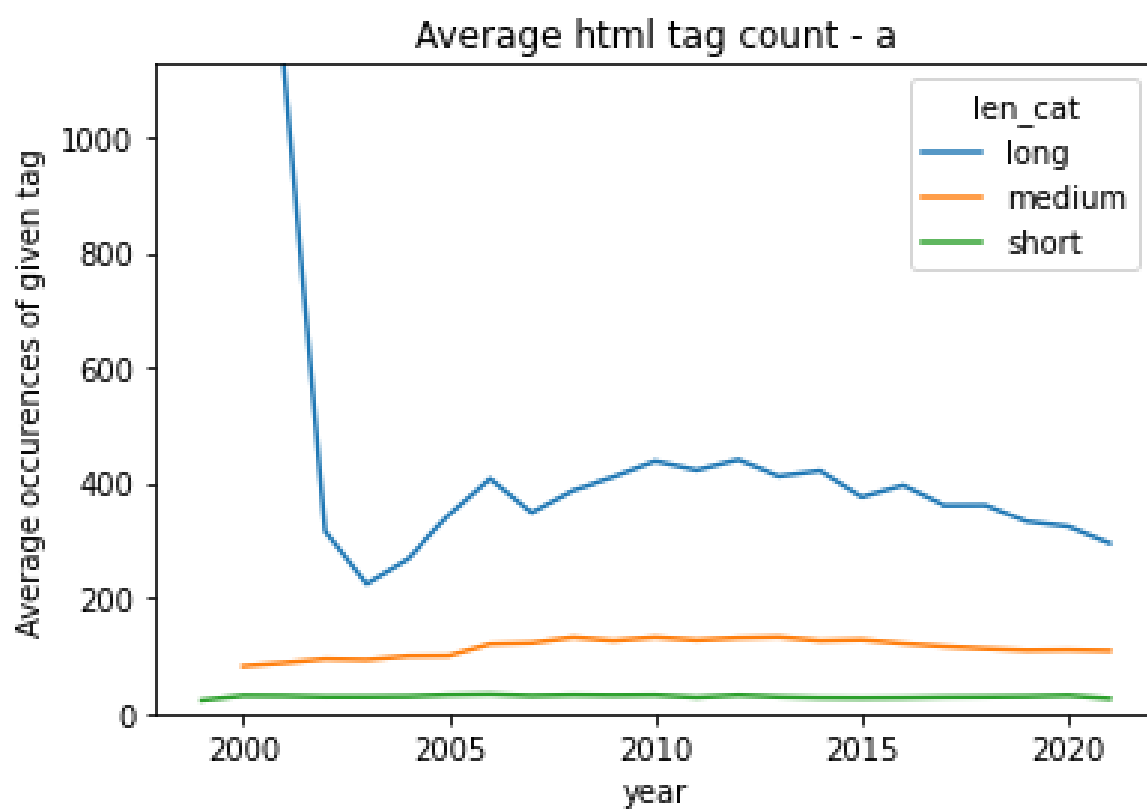
Rys. 5.5. Średnia liczba wystąpień taga div



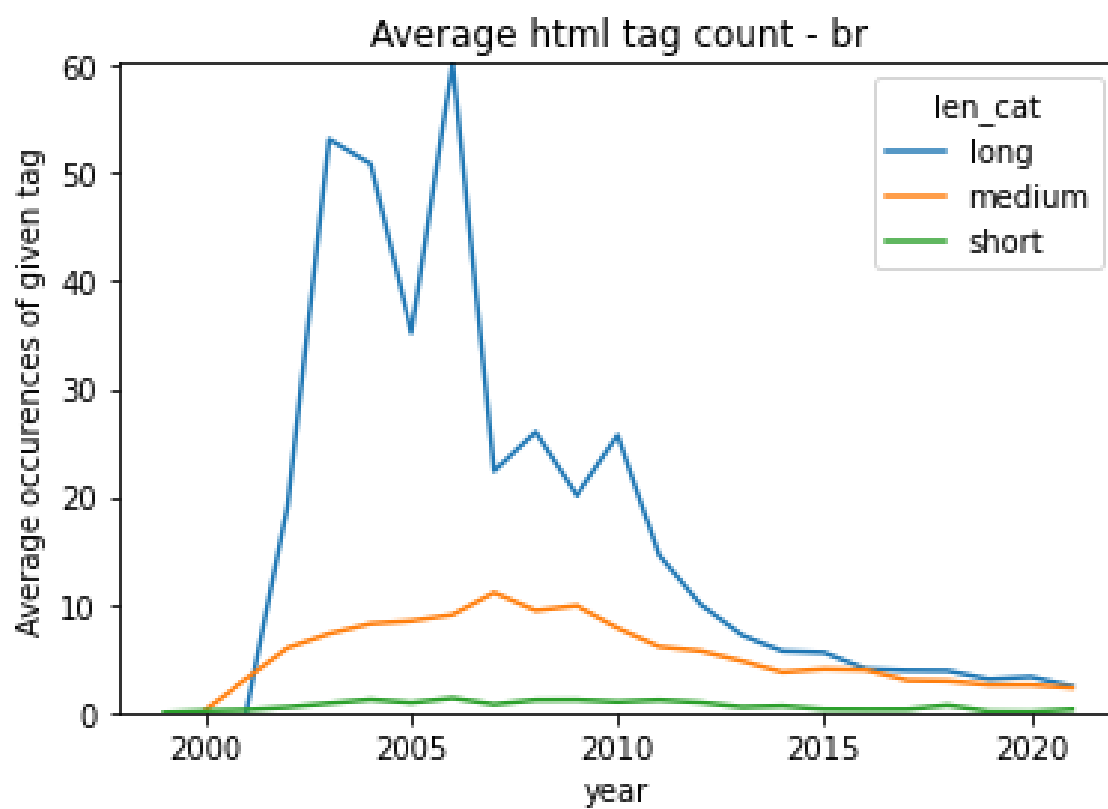
Rys. 5.6. Średnia liczba wystąpień taga td - table data cell



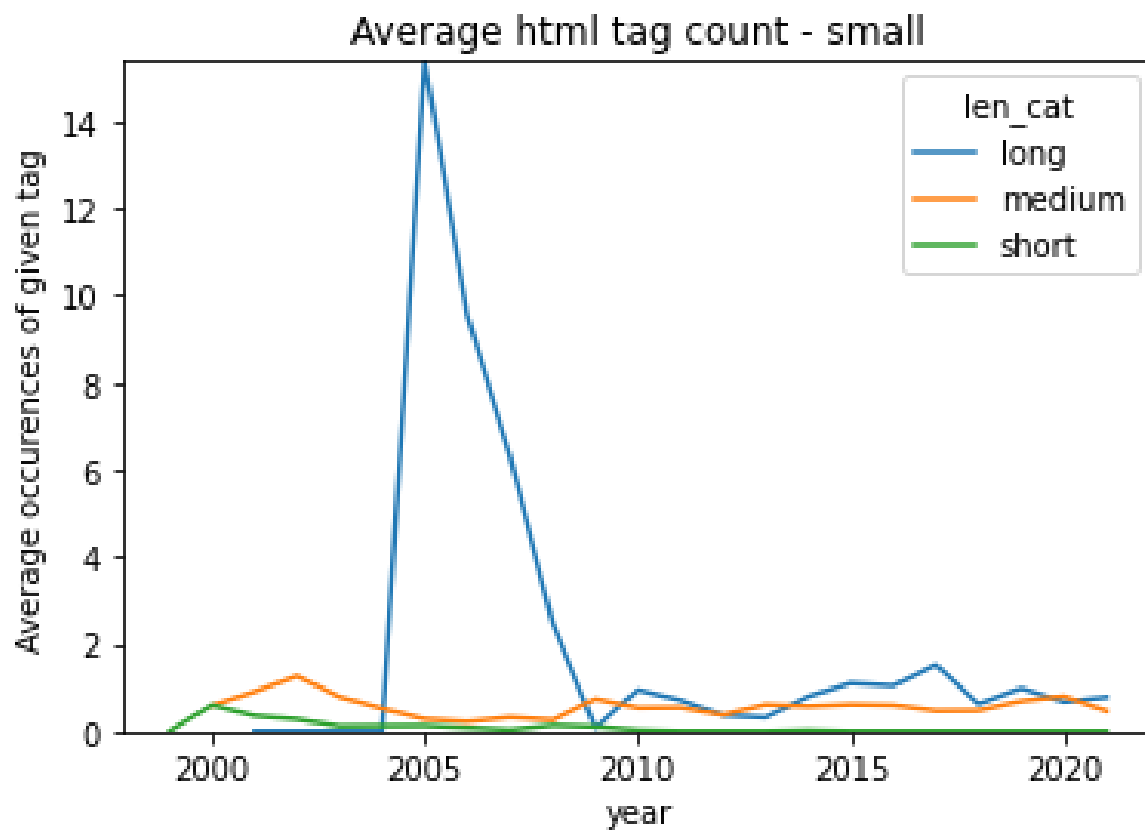
Rys. 5.7. Średnia liczba wystąpień taga video



Rys. 5.8. Średnia liczba wystąpień taga a



Rys. 5.9. Średnia liczba wystąpień taga br



Rys. 5.10. Średnia liczba wystąpień taga small

5.1.4. Najdłuższe strony internetowe

Wyznaczono również tabelę z najdłuższymi stronami, jakie udało nam się do tej pory pobrać i składować w bazie danych oraz HDFS:

webpageLink	length	year
jmw.com.cn	3147822	{2021, null}
allegro.pl	2639957	{2020, null}
opensea.io	2560538	{2019, null}
aljazeera.com	2425943	{2021, null}
flipkart.com	2220556	{2021, null}
scmp.com	2210676	{2021, null}
cnn.com	2061486	{2018, null}
nytimes.com	2053475	{2020, null}
walgreens.com	1885306	{2020, null}
trendyol.com	1874076	{2018, null}
tripadvisor.com	1734808	{2019, null}
nymag.com	1713391	{2021, null}
audible.com	1649510	{2021, null}
houzz.com	1647372	{2021, null}
coursera.org	1576667	{2016, null}
fastcompany.com	1540185	{2021, null}
autodesk.com	1395225	{2018, null}
go.com	1316388	{2017, null}
alexa.com	1308152	{2021, null}
euronews.com	1254791	{2018, null}

Rys. 5.11. Ranking najdłuższych stron wraz z podaną ich długością oraz rekordowym rokiem dla określonej strony

5.2. Widoki wsadowe

Widoki wsadowe stworzone za pomocą PySparka są następujące:

- **pagesLengthView** - widok, w którym dla każdego roku przechowywana jest średnia długość każdej strony internetowej - zagregowana po wszystkich stronach dla określonego roku
- **average_tag_count_per_year_longLengthPages** - dla stron o długości większej lub równej 80 tysięcy znaków, widok ze zliczeniami średniej liczby wystąpień danego taga dla danego roku (agregacja po wszystkich stronach). Rok posortowano rosnąco.
- **average_tag_count_per_year_mediumLengthPages** - analogicznie jak powyżej, dla stron o długości większej niż 20 tysięcy znaków i mniejszej niż 80 tysięcy znaków
- **average_tag_count_per_year_shortLengthPages** - analogicznie jak powyżej, dla stron o długości mniejszej równej niż 20 tysięcy znaków

6. Testy funkcjonalne komponentów

Celem testowania jest pokazanie, że:

1. przepływ danych w Nifi przebiega poprawnie,
2. analiza stron przebiega poprawnie,
3. w bazach HBase oraz Hadoop są składowane dane.

6.1. Przepływ danych w Nifi

Pierwszy punkt testowany będzie poprzez sprawdzanie zawartości plików Flow File w Nifi przed lub po danym procesie. Zostanie opisane, co jest oczekiwane przed danym procesem, wybranie kilku przypadków i sprawdzenie, czy w pliku znajduje się to, co było oczekiwane na danym etapie procesu przetwarzania. W raporcie zostały zamieszczone po jednym lub dwa przykłady dla zachowania zwięzłości raportu.

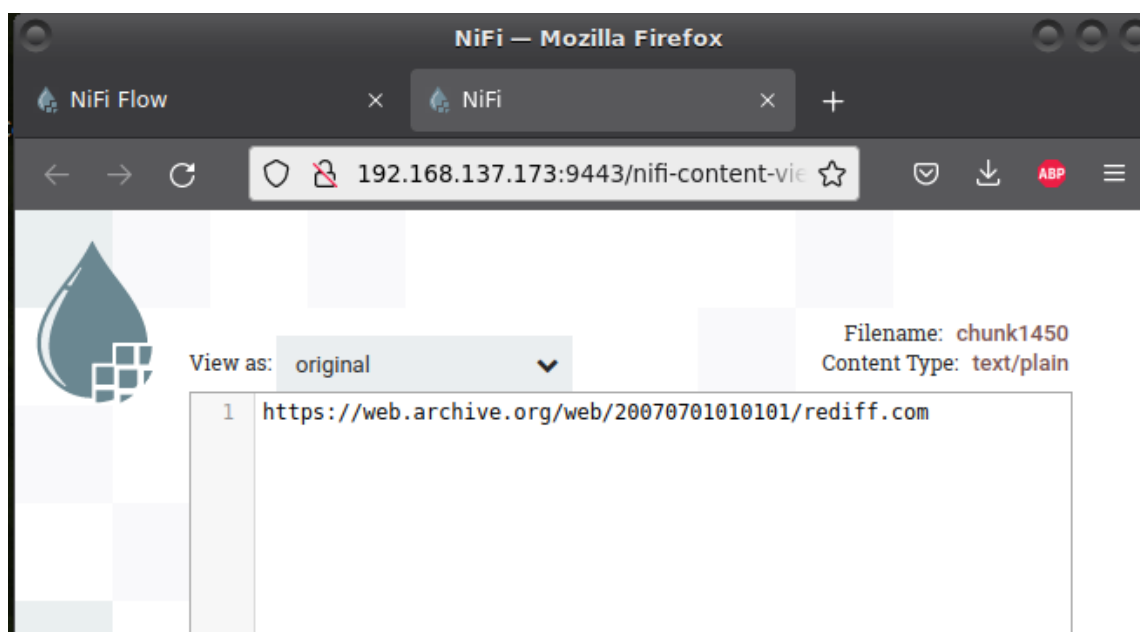
6.1.1. Pliki przed zapytaniem do serwera

Plik z zapytaniem powinien zawierać jeden link w postaci:

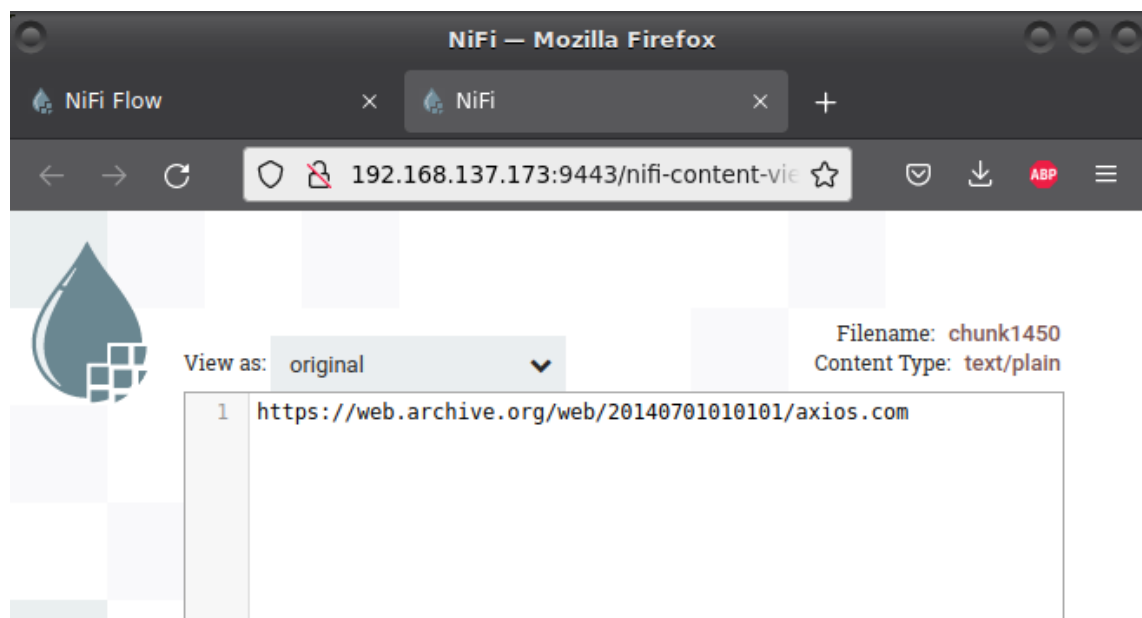
`https://web.archive.org/web/[czas yyyyymmddhhmmss]/[adres strony internetowej]`

Czas oznacza rok wykonania i zawsze sprawdzamy 1 lipca o godzinie 1:01. Zmieniany jest tylko rok, co oznacza że cyfra powinna być w postaci yyy0701010101.

Poniższe zrzuty ekranu potwierdzają oczekiwany wynik. 6.1 6.2



Rys. 6.1. Przykładowy plik z linkiem do zapytania Wayback Machine

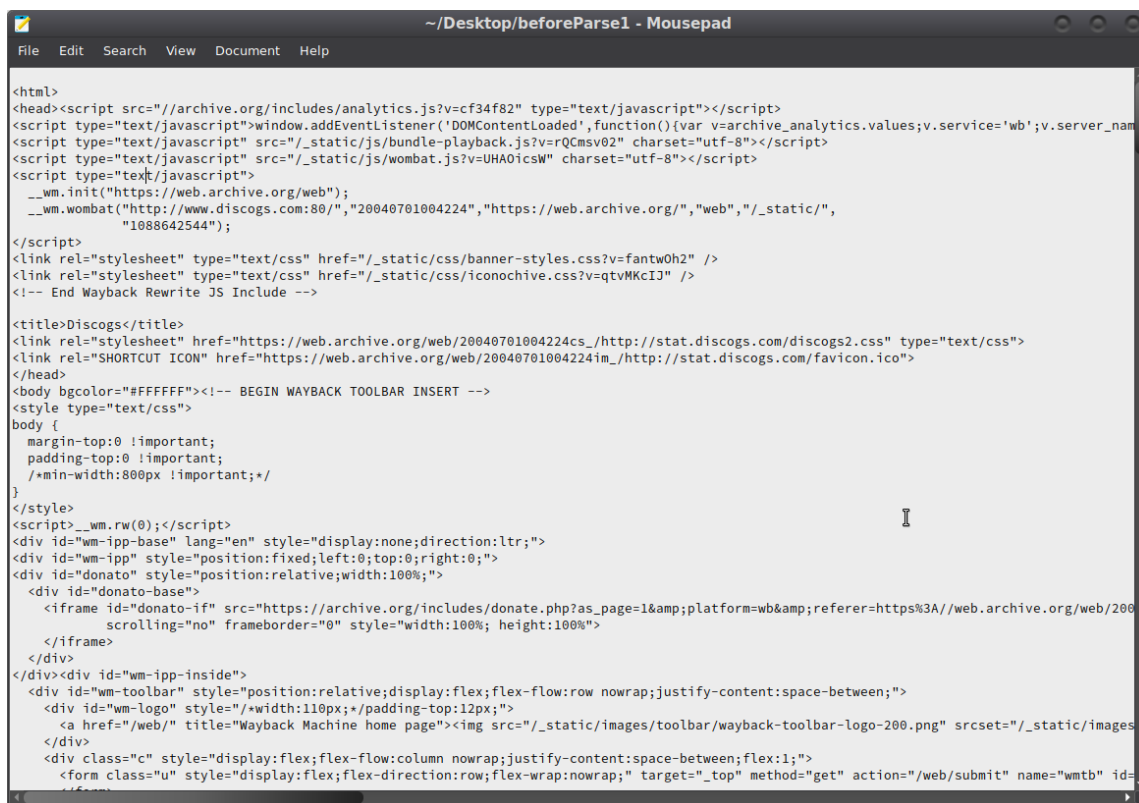


Rys. 6.2. Przykładowy plik z linkiem do zapytania Wayback Machine

6.1.2. Wynik zapytania od serwera

Od razu po zapytaniu powinno się otrzymać plik zawierający stronę internetową od Wayback Machine.

Sam plik jest w postaci binarnej, ale można go odczytać ściągając kopię pliku z kolejki i wyświetlając go w notatniku. Poniższe zrzuty ekranu pokazują część zawartości takiego pliku, który jest poprawną odpowiedzią ze strony Wayback Machine. 6.3 6.4



```
<html>
<head><script src="//archive.org/includes/analytics.js?v=cf34f82" type="text/javascript"></script>
<script type="text/javascript">window.addEventListener('DOMContentLoaded',function(){var v=archive_analytics.values;v.service='wb';v.server_nam
<script type="text/javascript" src="/_static/js/bundle-playback.js?v=RQcmsv02" charset="utf-8"></script>
<script type="text/javascript" src="/_static/js/wombat.js?v=UHA01csW" charset="utf-8"></script>
<script type="text/javascript">
  __wm.init("https://web.archive.org/web/");
  __wm.wombat("http://www.discogs.com:80/", "20040701004224", "https://web.archive.org/", "web", "/_static/",
    "1088642544");
</script>
<link rel="stylesheet" type="text/css" href="/_static/css/banner-styles.css?v=fantW0h2" />
<link rel="stylesheet" type="text/css" href="/_static/css/iconochive.css?v=qtvMKcI3" />
<!-- End Wayback Rewrite JS Include -->

<title>Discogs</title>
<link rel="stylesheet" href="https://web.archive.org/web/20040701004224cs_/http://stat.discogs.com/discogs2.css" type="text/css">
<link rel="SHORTCUT ICON" href="https://web.archive.org/web/20040701004224im_/http://stat.discogs.com/favicon.ico">
</head>
<body bgcolor="#FFFFFF"><!-- BEGIN WAYBACK TOOLBAR INSERT -->
<style type="text/css">
body {
  margin-top:0 !important;
  padding-top:0 !important;
  /*min-width:800px !important;*/
}
</style>
<script>__wm.rw(0);</script>
<div id="wm-ipp-base" lang="en" style="display:none;direction:ltr;">
<div id="wm-ipp" style="position:fixed;left:0;top:0;right:0;">
<div id="donato" style="position:relative;width:100%;">
  <div id="donato-base">
    <iframe id="donato-if" src="https://archive.org/includes/donate.php?as_page=1&platform=wb&referrer=https%3A//web.archive.org/web/20040701004224cs_/http://stat.discogs.com/discogs2.css" scrolling="no" frameborder="0" style="width:100%; height:100%">
    </iframe>
  </div>
</div><div id="wm-ipp-inside">
  <div id="wm-toolbar" style="position:relative;display:flex;flex-flow:row nowrap;justify-content:space-between;">
    <div id="wm-logo" style="/*width:110px;*/padding-top:12px;">
      <a href="/web/" title="Wayback Machine home page">
    <div class="c" style="display:flex;flex-flow:column nowrap;justify-content:space-between;flex:1;">
      <form class="u" style="display:flex;flex-direction:row;flex-wrap:nowrap;" target="_top" method="get" action="/web/submit" name="wmtb" id=
```

Rys. 6.3. Przykładowy plik ze stroną otrzymaną od Wayback Machine

6.1.3. Niepoprawne wczytane linki

Niekiedy dany link zwraca błąd przy próbie zapytania serwera. W tym punkcie, pokazane zostanie, że jest to wynik błędu zwracanego przez samą stronę. Poniżej znajdują się przykładowe 3 strony, które zwróciły błąd i nie zostały zapisane w bazie danych. 6.5 6.6 6.7

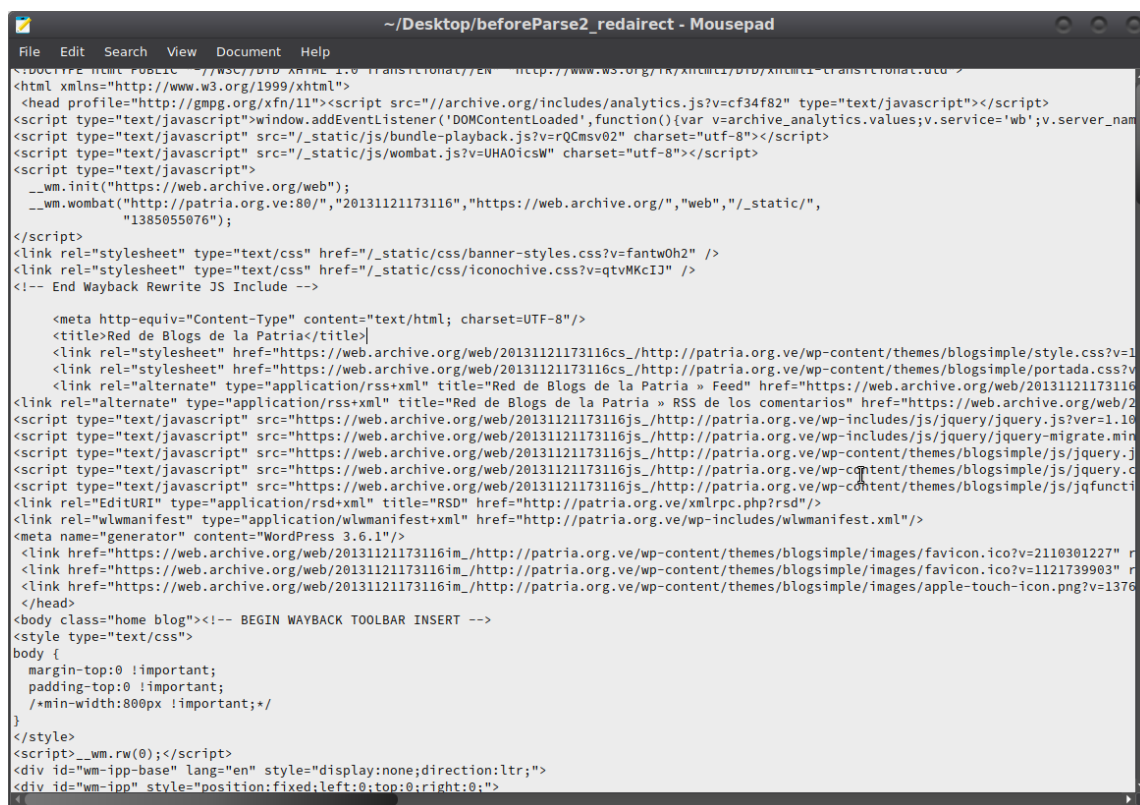
6.1.4. Wynik analizy

Po wykonaniu skryptu analizującego zawartość strony internetowej powinno otrzymać się plik z atrybutami postaci:

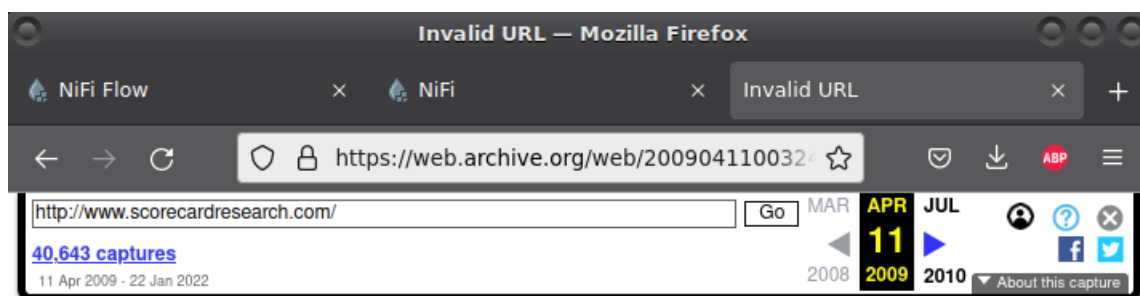
```
{ "raw": { "link": [link], "content": [zawartość strony internetowej] }, "parsed": { "link": [link], "year": ["yyyy"], "webpageLink": [nazwa strony internetowej], [lista atrybutów w postaci [tag zamykający]: [liczba wystąpień]] }
```

Poniżej znajdują się części pliku zawierającego analizę, potwierdzające, że przetworzone pliki mają taką postać. 6.3 6.4

Może zaistnieć przypadek, kiedy nie uda się otrzymać z przetwarzania strony roku jej zarzyciwizowania na Wayback Machine. Wtedy zamiast liczby roku pojawia się cyfra 0, co zostanie potraktowane jako brak danych w analizach. Taki przypadek jest pokazany poniżej. 6.11



Rys. 6.4. Przykładowy plik ze stroną otrzymaną od Wayback Machine



Invalid URL

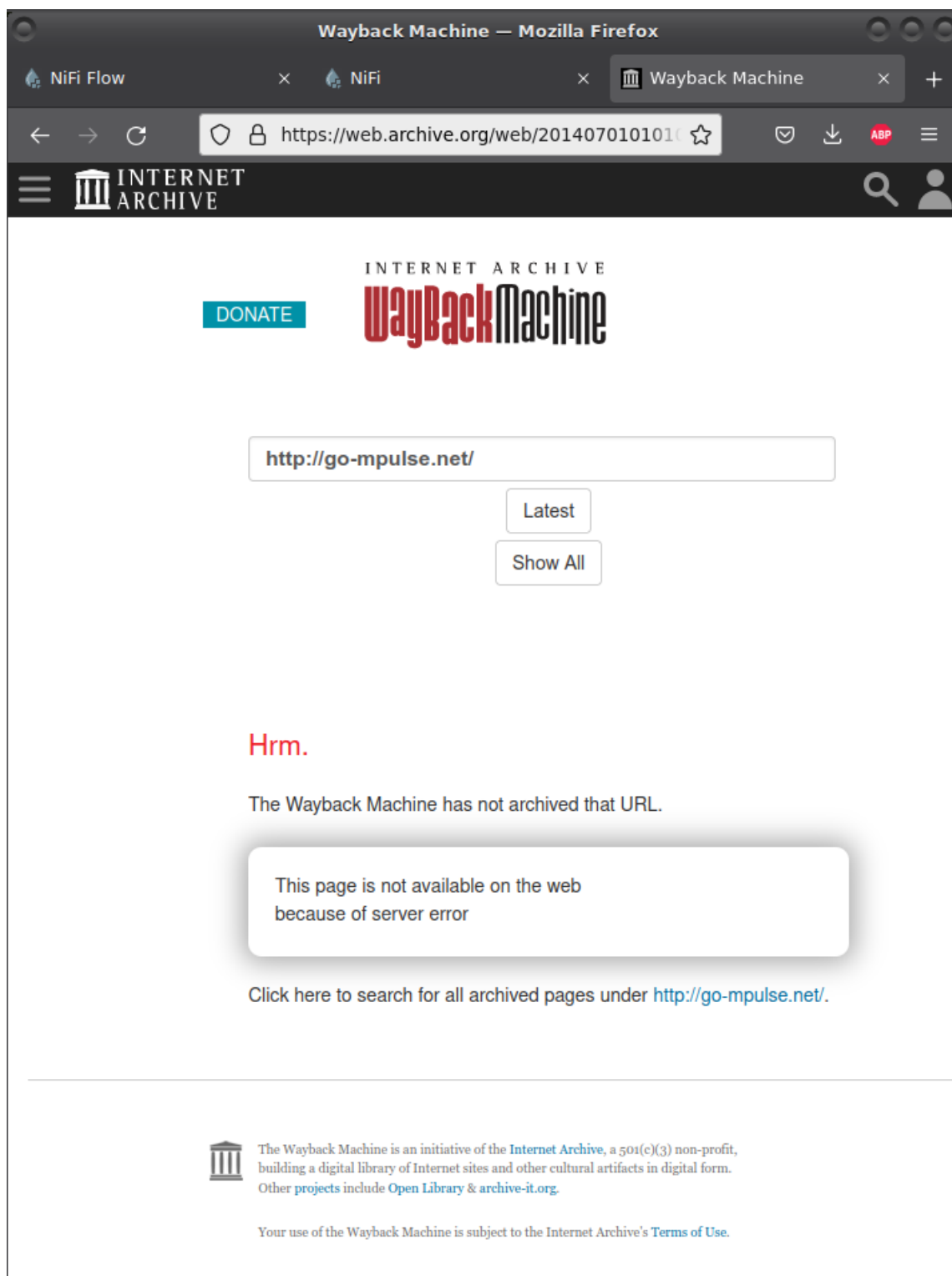
The requested URL "/", is invalid.

Reference #9.c3a81160.1239409964.0

Rys. 6.5. Przykładowy link zwracający błąd



Rys. 6.6. Przykładowy link zwracający błąd



Rys. 6.7. Przykładowy link zwracający błąd

[illegible]

Rys. 6.8. Przykładowy plik zawierający analizę; początek

```
INTERNET ARCHIVE ON 15:59:12 Jan 23, 2022.\n\n JAVASCRIPT APPENDED BY WAYBACK MACHINE,\nCOPYRIGHT INTERNET ARCHIVE.\n\n ALL OTHER CONTENT MAY ALSO BE PROTECTED BY COPYRIGHT\n(17 U.S.C.\n SECTION 108(a)(3)).\n\n-->\n\n!-\n\n\nplayback timings (ms):\n captures_list: 962.867\n exclusion robots: 0.214\n exclusion robots.policy: 0.206\n RedisCDXSource: 0.618\n esindex: 0.008\n LoadShardBlock: 98.681 (3)\n PetaboxLoader3.datanode: 75.837 (4)\n\n CDXLines.iter: 85.194 (3)\n\n load_resource:\n133.79\n PetaboxLoader3.resolve: 114.684\n\n-->}, "parsed": {"link":\n"https://web.archive.org/web/20090701010101/redcross.org", "year": "2009",\n"webpageLink": "redcross.org", "length": 38111, "<a>": 83, "</abbr>": 0, "</acronym>":\n0, "</address>": 0, "</applet>": 0, "</area>": 0, "</article>": 0, "</aside>": 0,\n"</audio>": 0, "</b>": 0, "</base>": 0, "</basefont>": 0, "</bb>": 0, "</bdo>": 0,\n"</big>": 0, "</blockquote>": 0, "</body>": 1, "<br/>": 0, "</button>": 0, "</canvas>":\n1, "</caption>": 0, "</center>": 0, "</cite>": 0, "</code>": 0, "</col>": 0,\n"</colgroup>": 0, "</command>": 0, "</datagrid>": 0, "</datalist>": 0, "</dd>": 0,\n"</del>": 0, "</details>": 0, "</dfn>": 0, "</dialog>": 0, "</dir>": 0, "</div>": 99,\n"</dl>": 0, "</dt>": 0, "</em>": 0, "</embed>": 0, "</eventsourcing>": 0, "</fieldset>":\n0, "</figure>": 0, "</font>": 0, "</footer>": 0, "</form>": 3, "</frameset>": 0,\n"</frameset>": 0, "</hl>": 2, "</h2>": 1, "</h3>": 2, "</h4>": 1, "</h5>": 0, "</h6>":\n0, "</head>": 1, "</header>": 0, "</hgroup>": 0, "</hr>": 0, "</html>": 1, "</i>": 0,\n"</iframe>": 1, "</img>": 0, "</input>": 0, "</ins>": 0, "</isindex>": 0, "</kbd>": 0,\n"</keygen>": 0, "</label>": 0, "</legend>": 0, "</li>": 57, "</link>": 1, "</map>": 0,\n"</mark>": 0, "</menu>": 0, "</meta>": 0, "</meter>": 0, "</nav>": 0, "</noframes>": 0,\n"</noscript>": 0, "</object>": 0, "</ol>": 0, "</optgroup>": 0, "</option>": 0,\n"</output>": 0, "</p>": 5, "</param>": 0, "</pre>": 0, "</progress>": 0, "</q>": 0,\n"</rp>": 0, "</rt>": 0, "</ruby>": 0, "</s>": 0, "</samp>": 0, "</script>": 10,\n"</section>": 0, "</select>": 0, "</small>": 0, "</source>": 0, "</span>": 20,\n"</strike>": 0, "</strong>": 8, "</style>": 1, "</sub>": 0, "</sup>": 0, "</table>": 2,\n"</tbody>": 1, "</td>": 12, "</textarea>": 0, "</tfoot>": 0, "</th>": 3, "</thead>": 0,\n"</time>": 0, "</title>": 1, "</tr>": 5, "</track>": 0, "</tt>": 0, "</u>": 0, "</ul>":\n13, "</var>": 0, "</video>": 0, "</wbr>": 0}}}
```

Rys. 6.9. Przykładowy plik zawierający analizę: koniec

```
href="//archive.org/about/terms.php">terms of Use</a>.\n      </p>\n    </div>\n  </footer>\n</body>\n</html>"}, "parsed": {"link": "https://web.archive.org/web/20130701010101/msdn.com", "year": 0, "webpageLink": "msdn.com", "length": 93263, "</a>": 33, "</abbr>": 0, "</acronym>": 0, "</address>": 0, "</applet>": 0, "</area>": 0, "</article>": 0, "</aside>": 0, "</audio>": 0, "</b>": 0, "</base>": 0, "</basefont>": 0, "</bb>": 0, "</bdo>": 0, "</big>": 0, "</blockquote>": 0, "</body>": 1, "<br/>": 2, "</button>": 4, "</canvas>": 0, "</caption>": 0, "</center>": 0, "</cite>": 0, "</code>": 0, "</col>": 0, "</colgroup>": 0, "</command>": 0, "</datagrid>": 0, "</datalist>": 0, "</dd>": 0, "</del>": 0, "</details>": 0, "</dfn>": 0, "</dialog>": 0, "</dir>": 0, "</div>": 18, "</dl>": 0, "</dt>": 0, "</em>": 0, "</embed>": 0, "</eventsource>": 0, "</fieldset>": 1, "</figure>": 0, "</font>": 0, "</footer>": 1, "</form>": 2, "</frame>": 0, "</frameset>": 0, "</h1>": 0, "</h2>": 0, "</h3>": 0, "</h4>": 0, "</h5>": 0, "</h6>": 0, "</head>": 1, "</header>": 0, "</hgroup>": 0, "</hr>": 0, "</html>": 1, "</i>": 0, "</iframe>": 0, "</img>": 0, "</input>": 0, "</ins>": 0, "</isindex>": 0, "</kbd>": 0, "</keygen>": 0, "</label>": 4, "</legend>": 0, "</li>": 11, "</link>": 0, "</map>": 0, "</mark>": 0, "</menu>": 0, "</meta>": 0, "</meter>": 0, "</nav>": 3, "</noframes>": 0, "</noscript>": 2, "</object>": 0, "</ol>": 0, "</optgroup>": 0, "</option>": 0, "</output>": 0, "</p>": 9, "</param>": 0, "</pre>": 0, "</progress>": 0, "</q>": 0, "</rp>": 0, "</rt>": 0, "</ruby>": 0, "</s>": 0, "</samp>": 0, "</script>": 19, "</section>": 2, "</select>": 0, "</small>": 0, "</source>": 0, "</span>": 18, "</strike>": 0, "</strong>": 0, "</style>": 13, "</sub>": 0, "</sup>": 0, "</table>": 0, "</tbody>": 0, "</td>": 0, "</textarea>": 0, "</tfoot>": 0, "</th>": 0, "</thead>": 0, "</time>": 0, "</title>": 16, "</tr>": 0, "</track>": 0, "</tt>": 0, "</u>": 0, "</ul>": 2, "</var>": 0, "</video>": 0, "</wbr>": 0}}}
```

Rys. 6.10. Plik zawierający analizę z nieodczytanym rokiem zarchiwizowania; koniec

6.1.5. Plik zawierający dane wrzucane do HDFS

Do HDFS jest zamieszczana tylko część zawierającą analizę strony internetowej. Jest ona postaci:

```
{ "link": [link], "year": ["yyyy"], "webpageLink": [nazwa strony internetowej], [lista atrybu-  
tów w postaci [tag zamykający]: [liczba wystąpień]] }
```

Poniżej znajduje się przykład takiego pliku.

```
{ "link": "https://web.archive.org/web/20210701010101/redcross.org", "year": "2021", "webpageLink": "redcross.org", "len  
gth": 123957, "<a>": 204, "</abbr>": 0, "</acronym>": 0, "</address>": 0, "</applet>": 0, "</area>": 0, "</article>": 0, "  
</aside>": 0, "</audio>": 0, "</b>": 7, "</base>": 0, "</basefont>": 0, "</bb>": 0, "</bdo>": 0, "</big>": 0, "</blockquote>": 0, "  
</body>": 1, "</br>": 8, "</button>": 14, "</canvas>": 1, "</caption>": 0, "</center>": 0, "</cite>": 0, "</code>": 0, "  
</col>": 0, "</colgroup>": 0, "</command>": 0, "</datagrid>": 0, "</datalist>": 0, "</dd>": 0, "</del>": 0, "</details>": 0, "  
</dfn>": 0, "</dialog>": 0, "</div>": 0, "</div>": 536, "</dl>": 0, "</dt>": 0, "</em>": 0, "</embed>": 0, "</eventsource>": 0, "  
</fieldset>": 1, "</figure>": 0, "</font>": 0, "</footer>": 1, "</form>": 3, "</frame>": 0, "</frameset>": 0, "</h1>": 0, "  
</h2>": 11, "</h3>": 0, "</h4>": 15, "</h5>": 1, "</h6>": 0, "</head>": 1, "</header>": 1, "</hgroup>": 0, "</hr>": 0, "  
</html>": 1, "</i>": 9, "</iframe>": 1, "</img>": 0, "</input>": 0, "</ins>": 0, "</isindex>": 0, "</kbd>": 0, "</keygen>": 0, "  
</label>": 0, "</legend>": 0, "</li>": 96, "</link>": 0, "</map>": 0, "</mark>": 0, "</menu>": 0, "</meta>": 0, "</meter>": 0, "  
</nav>": 31, "</noframes>": 0, "</noscript>": 0, "</object>": 0, "</ol>": 0, "</optgroup>": 0, "</option>": 13, "</output>": 0, "  
</p>": 18, "</param>": 0, "</pre>": 0, "</progress>": 0, "</q>": 0, "</rp>": 0, "</rt>": 0, "</ruby>": 0, "</s>": 0, "</samp>": 0, "  
</script>": 13, "</section>": 0, "</select>": 1, "</small>": 0, "</source>": 0, "</span>": 54, "</strike>": 0, "</strong>": 3, "  
</style>": 9, "</sub>": 1, "</sup>": 2, "</table>": 1, "</tbody>": 1, "</td>": 9, "</textarea>": 0, "</tfoot>": 0, "</th>": 0, "  
</thead>": 0, "</time>": 0, "</title>": 1, "</tr>": 3, "</track>": 0, "</tt>": 0, "</u>": 0, "</ul>": 20, "</var>": 0, "  
</video>": 0, "</wbr>": 0 }
```

Rys. 6.11. Plik zawierający rekord do bazy HDFS

6.1.6. Plik z zgrupowanymi rekordami do HDFS

Przed wrzuceniem rekordów do HDFS rekordy są grupowane w jeden plik zawierający przy-
najmniej 1000 rekordów. Poniższy przykład pokazuje zawartość takiego pliku i potwierdzenie,
że zawiera on przyjemniej 1000 przeanalizowanych stron internetowych. 6.12

6.2. Analiza stron

W tym rozdziale testowana jest poprawność działania skryptu analizującego stronę interneto-
wą. Porównane zostanie zliczenie wystąpień wykonane przez skrypt i poprzez "ręczne" zliczenie.

Na przykładzie wyniku analizy strony na podstawie zapytania

<https://web.archive.org/web/20090701010101/http://redcross.org>

otrzymano następujące wystąpienia tagów: < /a >:83 razy, < /b >:0 razy, < /canvas >:1
raz. 6.13

Te same tagi zliczono "ręczne" analizując zawartość strony w ten sam sposób, jaki to robi
skrypt. We wszystkich 3 przypadkach zliczenie pokazało tyle samo wystąpień co w analizie
(odpowiednio 83, 0, 1 wystąpień). 6.14 6.15 6.16

6.3. Zawartość baz danych

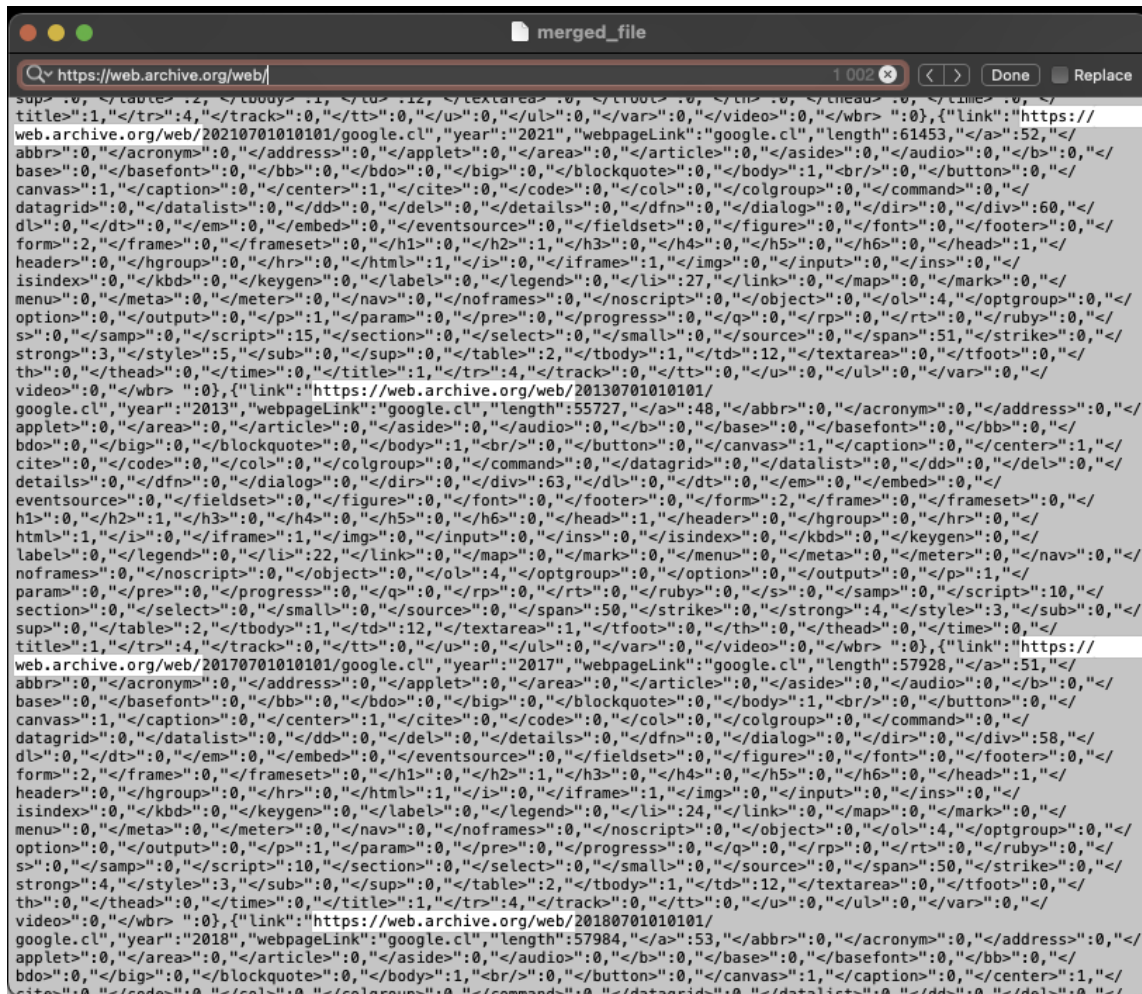
Pokazane zostanie, że dane są składowane w bazach danych Hbase i HDFS.

6.3.1. HBase

W bazie danych HBase powinny znajdować się dwie tabele: 'duplicates.cache' i 'webpages'.
Tabela 'duplicates.cache' powinna zawierać klucz [yyyy:website], natomiast tabela 'webpages'
powinna zawierać klucz będący zapytaniem (linkiem), na podstawie którego utworzono dany
rekord.

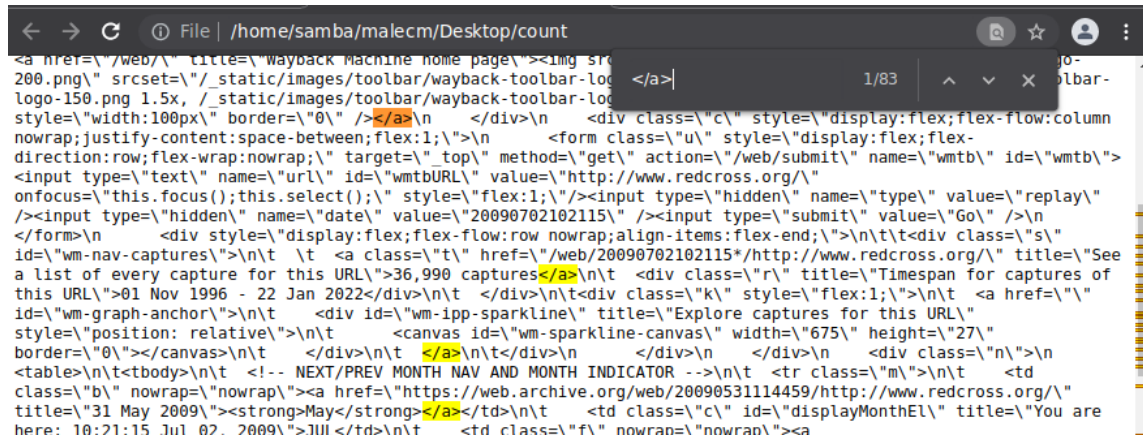
Test polegał na wywołaniu w "hbase shell" komendy "count", która wypisuje liczbę rekordów,
ale także pokazuje wybrane klucze.

Wywołanie tej komendy potwierdza oczekiwany wynik. 6.17



Rys. 6.12. Plik zawierający 10002 rekordy do bazy HDFS

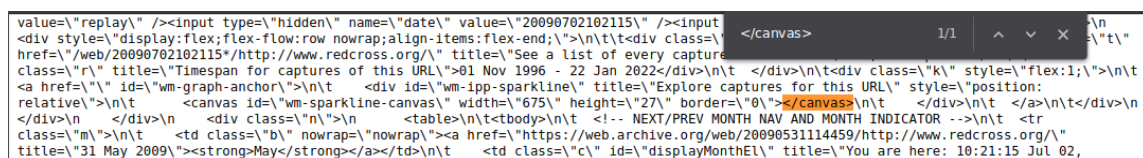
```
"parsed": { "Link": "https://web.archive.org/web/20180701010101/http://redcross.org",
"</a>": 83, "</abbr>": 0, "</address>": 0, "</area>": 0, "</article>": 0, "</aside>": 0, "</audio>": 0, "</b>": 0, "</base>": 0, "</basefont>": 0, "</bb>": 0, "</bdo>": 0, "</big>": 0, "</blockquote>": 0, "</body>": 1, "</br>": 0, "</button>": 0, "</canvas>": 1, "</caption>": 0, "</center>": 1, "</cite>": 0, "</code>": 0, "</col>": 0, "</colgroup>": 0, "</command>": 0, "</datagrid>": 0, "</datalist>": 0, "</dd>": 0, "</del>": 0, "</details>": 0, "</dfn>": 0, "</dialog>": 0, "</div>": 60, "</dl>": 0, "</dt>": 0, "</em>": 0, "</embed>": 0, "</eventsource>": 0, "</fieldset>": 0, "</figure>": 0, "</font>": 0, "</footer>": 0, "</form>": 2, "</frame>": 0, "</frameset>": 0, "</h1>": 0, "</h2>": 1, "</h3>": 0, "</h4>": 0, "</h5>": 0, "</h6>": 0, "</head>": 1, "</header>": 0, "</hgroup>": 0, "</hr>": 0, "</html>": 1, "</i>": 0, "</iframe>": 1, "</img>": 0, "</input>": 0, "</ins>": 0, "</isindex>": 0, "</kbd>": 0, "</keygen>": 0, "</label>": 0, "</legend>": 0, "</li>": 22, "</link>": 0, "</map>": 0, "</mark>": 0, "</menu>": 0, "</meta>": 0, "</meter>": 0, "</nav>": 0, "</noframes>": 0, "</noscript>": 0, "</object>": 0, "</ol>": 4, "</optgroup>": 0, "</option>": 0, "</output>": 0, "</p>": 1, "</param>": 0, "</pre>": 0, "</progress>": 0, "</q>": 0, "</rp>": 0, "</rt>": 0, "</ruby>": 0, "</s>": 0, "</samp>": 0, "</script>": 10, "</section>": 0, "</select>": 0, "</small>": 0, "</source>": 0, "</span>": 50, "</strike>": 0, "</strong>": 4, "</style>": 3, "</sub>": 0, "</sup>": 0, "</table>": 2, "</tbody>": 1, "</td>": 12, "</tfoot>": 0, "</th>": 0, "</thead>": 0, "</time>": 0, "</title>": 1, "</tr>": 4, "</track>": 0, "</tt>": 0, "</u>": 0, "</ul>": 0, "</var>": 0, "</video>": 0, "</wbr>": 0, "</link>": "https://web.archive.org/web/20180701010101/http://redcross.org", "length": 57984, "</a>": 53, "</abbr>": 0, "</acronym>": 0, "</address>": 0, "</applet>": 0, "</area>": 0, "</article>": 0, "</aside>": 0, "</audio>": 0, "</b>": 0, "</base>": 0, "</basefont>": 0, "</bb>": 0, "</bdo>": 0, "</big>": 0, "</blockquote>": 0, "</body>": 1, "</br>": 0, "</button>": 0, "</canvas>": 1, "</caption>": 0, "</center>": 1, "</cite>": 0, "</code>": 0, "</col>": 0, "</colgroup>": 0, "</command>": 0, "</datagrid>": 0, "</datalist>": 0, "</dd>": 0, "</del>": 0, "</details>": 0, "</dfn>": 0, "</dialog>": 0, "</div>": 58, "</dl>": 0, "</dt>": 0, "</em>": 0, "</embed>": 0, "</eventsource>": 0, "</fieldset>": 0, "</figure>": 0, "</font>": 0, "</footer>": 0, "</form>": 2, "</frame>": 0, "</frameset>": 0, "</h1>": 0, "</h2>": 1, "</h3>": 0, "</h4>": 0, "</h5>": 0, "</h6>": 0, "</head>": 1, "</header>": 0, "</hgroup>": 0, "</hr>": 0, "</html>": 1, "</i>": 0, "</iframe>": 1, "</img>": 0, "</input>": 0, "</ins>": 0, "</isindex>": 0, "</kbd>": 0, "</keygen>": 0, "</label>": 0, "</legend>": 0, "</li>": 24, "</link>": 0, "</map>": 0, "</mark>": 0, "</menu>": 0, "</meta>": 0, "</meter>": 0, "</nav>": 0, "</noframes>": 0, "</noscript>": 0, "</object>": 0, "</ol>": 4, "</optgroup>": 0, "</option>": 0, "</output>": 0, "</p>": 1, "</param>": 0, "</pre>": 0, "</progress>": 0, "</q>": 0, "</rp>": 0, "</rt>": 0, "</ruby>": 0, "</s>": 0, "</samp>": 0, "</script>": 10, "</section>": 0, "</select>": 0, "</small>": 0, "</source>": 0, "</span>": 50, "</strike>": 0, "</strong>": 4, "</style>": 3, "</sub>": 0, "</sup>": 0, "</table>": 2, "</tbody>": 1, "</td>": 12, "</tfoot>": 0, "</th>": 0, "</thead>": 0, "</time>": 0, "</title>": 1, "</tr>": 4, "</track>": 0, "</tt>": 0, "</u>": 0, "</ul>": 0, "</var>": 0, "</video>": 0, "</wbr>": 0, "</link>": "https://web.archive.org/web/20180701010101/http://redcross.org", "length": 57984, "</a>": 53, "</abbr>": 0, "</acronym>": 0, "</address>": 0, "</applet>": 0, "</area>": 0, "</article>": 0, "</aside>": 0, "</audio>": 0, "</b>": 0, "</base>": 0, "</basefont>": 0, "</bb>": 0, "</bdo>": 0, "</big>": 0, "</blockquote>": 0, "</body>": 1, "</br>": 0, "</button>": 0, "</canvas>": 1, "</caption>": 0, "</center>": 1, "</cite>": 0, "</code>": 0, "</col>": 0, "</colgroup>": 0, "</command>": 0, "</datagrid>": 0, "</datalist>": 0, "</dd>": 0, "</del>": 0, "</details>": 0, "</dfn>": 0, "</dialog>": 0, "</div>": 58, "</dl>": 0, "</dt>": 0, "</em>": 0, "</embed>": 0, "</eventsource>": 0, "</fieldset>": 0, "</figure>": 0, "</font>": 0, "</footer>": 0, "</form>": 2, "</frame>": 0, "</frameset>": 0, "</h1>": 0, "</h2>": 1, "</h3>": 0, "</h4>": 0, "</h5>": 0, "</h6>": 0, "</head>": 1, "</header>": 0, "</hgroup>": 0, "</hr>": 0, "</html>": 1, "</i>": 0, "</iframe>": 1, "</img>": 0, "</input>": 0, "</ins>": 0, "</isindex>":
```

Rys. 6.14. Zliczenie tagu `` na stronie
<https://web.archive.org/web/20090701010101/http://redcross.org>



Rys. 6.15. Zliczenie tagu `` na stronie
<https://web.archive.org/web/20090701010101/http://redcross.org>



Rys. 6.16. Zliczenie tagu `</canvas>` na stronie
<https://web.archive.org/web/20090701010101/http://redcross.org>

```
hbase(main):005:0> count 'duplicate_cache'
Current count: 1000, row: 2001:cornell.edu
Current count: 2000, row: 2004:hulu.com
Current count: 3000, row: 2006:xnxx.com
Current count: 4000, row: 2008:xing.com
Current count: 5000, row: 2010:searchenginejournal.com
Current count: 6000, row: 2012:edx.org
Current count: 7000, row: 2013:pnas.org
Current count: 8000, row: 2014:welt.de
Current count: 9000, row: 2016:ebay.co.uk
Current count: 10000, row: 2017:letsencrypt.org
Current count: 11000, row: 2018:salesforce.com
Current count: 12000, row: 2019:vrbo.com
Current count: 13000, row: 2021:cloudinary.com
13671 row(s)
Took 2.0605 seconds
=> 13671
hbase(main):006:0> count 'webpages'
Current count: 1000, row: https://web.archive.org/web/20020701010101/python.org
Current count: 2000, row: https://web.archive.org/web/20040701010101/yy.com
Current count: 3000, row: https://web.archive.org/web/20070701010101/fast.com
Current count: 4000, row: https://web.archive.org/web/20090701010101/daum.net
Current count: 5000, row: https://web.archive.org/web/20100701010101/usembassy.gov
Current count: 6000, row: https://web.archive.org/web/20120701010101/huffpost.com
Current count: 7000, row: https://web.archive.org/web/20130701010101/tamu.edu
Current count: 8000, row: https://web.archive.org/web/20150701010101/atlassian.com
Current count: 9000, row: https://web.archive.org/web/20160701010101/gnu.org
Current count: 10000, row: https://web.archive.org/web/20170701010101/networkadvertising.org
Current count: 11000, row: https://web.archive.org/web/20180701010101/stanford.edu
Current count: 12000, row: https://web.archive.org/web/20190701010101/xnxx.com
Current count: 13000, row: https://web.archive.org/web/20210701010101/computerworld.com
13676 row(s)
Took 6.8605 seconds
=> 13676
hbase(main):007:0> █
```

Rys. 6.17. Wynik wywołania komendy "count" w "hbase shell";
komentarz: liczba rekordów jest różna ponieważ komenda była wywoływana w trakcie zapewniania bazy

6.3.2. Hadoop

W bazie danych Hadoop powinny znajdować się pliki zawierające po około 1000 rekordów.

```
vagrant@node1:~/bigdata/scripts$ hdfs dfs -ls /user/bigdata/data
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.6/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-tez-0.9.1-bin/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 15 items
-rw-r--r-- 1 root supergroup 138261 2022-01-23 12:39 /user/bigdata/data/195768d5-e096-47c9-8e36-128c61793c82
-rw-r--r-- 1 root supergroup 140201 2022-01-23 11:57 /user/bigdata/data/211c98f8-b4ef-45d8-aec0-df891bf7bf7e
-rw-r--r-- 1 root supergroup 138990 2022-01-23 12:10 /user/bigdata/data/23ff5089-c72f-45f4-8864-43d7f8d5a671
-rw-r--r-- 1 root supergroup 139007 2022-01-23 12:38 /user/bigdata/data/30cb85d6-eb7c-4c44-afd4-4ed458f65827
-rw-r--r-- 1 root supergroup 141257 2022-01-23 16:01 /user/bigdata/data/50c30575-c54c-4fd3-9391-f53ee641bffa
-rw-r--r-- 1 root supergroup 45137 2022-01-23 10:56 /user/bigdata/data/5bb37cdc-7c01-4a67-be7d-57f52d93d3a2
-rw-r--r-- 1 root supergroup 47033 2022-01-23 10:58 /user/bigdata/data/6c343feb-74de-4498-b5e3-c424675c1bff
-rw-r--r-- 1 root supergroup 153847 2022-01-23 15:56 /user/bigdata/data/70b37f88-81f8-49f0-bb40-31649725e980
-rw-r--r-- 1 root supergroup 139405 2022-01-23 15:54 /user/bigdata/data/7ce01ec3-1e36-43c6-b85a-f4b17a4d98fe
-rw-r--r-- 1 root supergroup 141728 2022-01-23 11:39 /user/bigdata/data/8763479f-a58a-4a97-8afa-9948fb092639
-rw-r--r-- 1 root supergroup 143483 2022-01-23 14:32 /user/bigdata/data/90cf5575-3100-46d7-9e84-bd1c1a6926c3
-rw-r--r-- 1 root supergroup 137790 2022-01-23 12:25 /user/bigdata/data/d237125a-d74a-4ac8-b06d-1f7e097b04a5
-rw-r--r-- 1 root supergroup 143782 2022-01-23 11:22 /user/bigdata/data/e18a7357-8b89-42b0-a744-e5a23f9ccf7c
-rw-r--r-- 1 root supergroup 139560 2022-01-23 11:11 /user/bigdata/data/f936d626-56f9-47c3-b401-31224f51d93e
-rw-r--r-- 1 root supergroup 148999 2022-01-23 12:39 /user/bigdata/data/fc9f4984-d6b2-4a4e-9cad-0705e64f5b1d
```

Rys. 6.18. zawartość bazy danych Hadoop

7. Podsumowanie

Wykorzystane narzędzia do przetwarzania danych klasy Big Data umożliwiły nam zautomatyzowaną analizę danych pochodzących z Internet Archive. Niska przepustowość wynikająca z ograniczeń Internet Archive sprawiła, że musieliśmy zmienić naszą architekturę - w naszym przypadku nie było większego sensu zbierania napływających olbrzymich wolumenów danych, ponieważ nasze dane wejściowe pobierane z Internet Archive były relatywnie małe w jednostce czasu. Mimo trudności dotyczących przepustowości Wayback Machine, nasze rozwiązanie umożliwiło na wykonanie cennych analiz struktury internetu oraz jego rozwoju na przestrzeni ostatnich kilkunastu lat.

Zauważalny jest trend wzrostowy długości stron, a co za tym idzie zwiększanie się liczby wystąpień tagów *div*. Z analiz w PySpark można było dostrzec pojawienie się tagu *video* w około 2010 roku oraz znaczny wzrost jego występowania w ostatnich latach. Można było również dostrzec liczne wykorzystywanie tagu *td* przed rokiem 2008 - obecnie poza nadzwyczajnymi przypadkami nie jest on stosowany. Odnośnie tagu dotyczącego linków *a* - ostatnio jest on stały w czasie, lecz ze względu na stały wzrost długości stron, można stwierdzić, że procent zawartości strony zawierający linki maleje. Dzięki analizie, mogliśmy też znaleźć rok występowania tagu *small* (2005), przykład stron wykorzystujących nietypowy tag *var* albo stworzyć ranking najdłuższych stron internetowych.

8. Podział pracy w grupie

Podział pracy w grupie był następujący:

L.p.	Zakres pracy	Autor
1	Składowanie danych w HBase Automatyzacja przepływu danych w Apache Nifi Przygotowanie raportu	Konrad Komisarczyk
2	Projekt architektury rozwiązania Testy funkcjonalności Prezentacja projektu	Mikołaj Malec
3	Analiza danych w Apache Spark Skrypt parsujący strony Przygotowanie raportu	Patryk Wrona

Tab. 8.1. Podział pracy