Тема: Wayback Machine - запазване на архивирани страници в S3

Предмет: *Приложно-програмни интерфейси за работа с облачни архитектури с Амазон Уеб Услуги (AWS)* **Изготвил**: *Христо Бориславов Спасов*, **фн**: *62278*, **имейл**: *hristo.b.spasov@gmail.com* **Лектор**: *доц. д-р Милен Петров*, **година**: *2021*

Съдържание

1. Условие

Да се добави възможност Wayback Machine да запазва архивираните страници в AWS S3 bucket и от същото място да ги извлича.

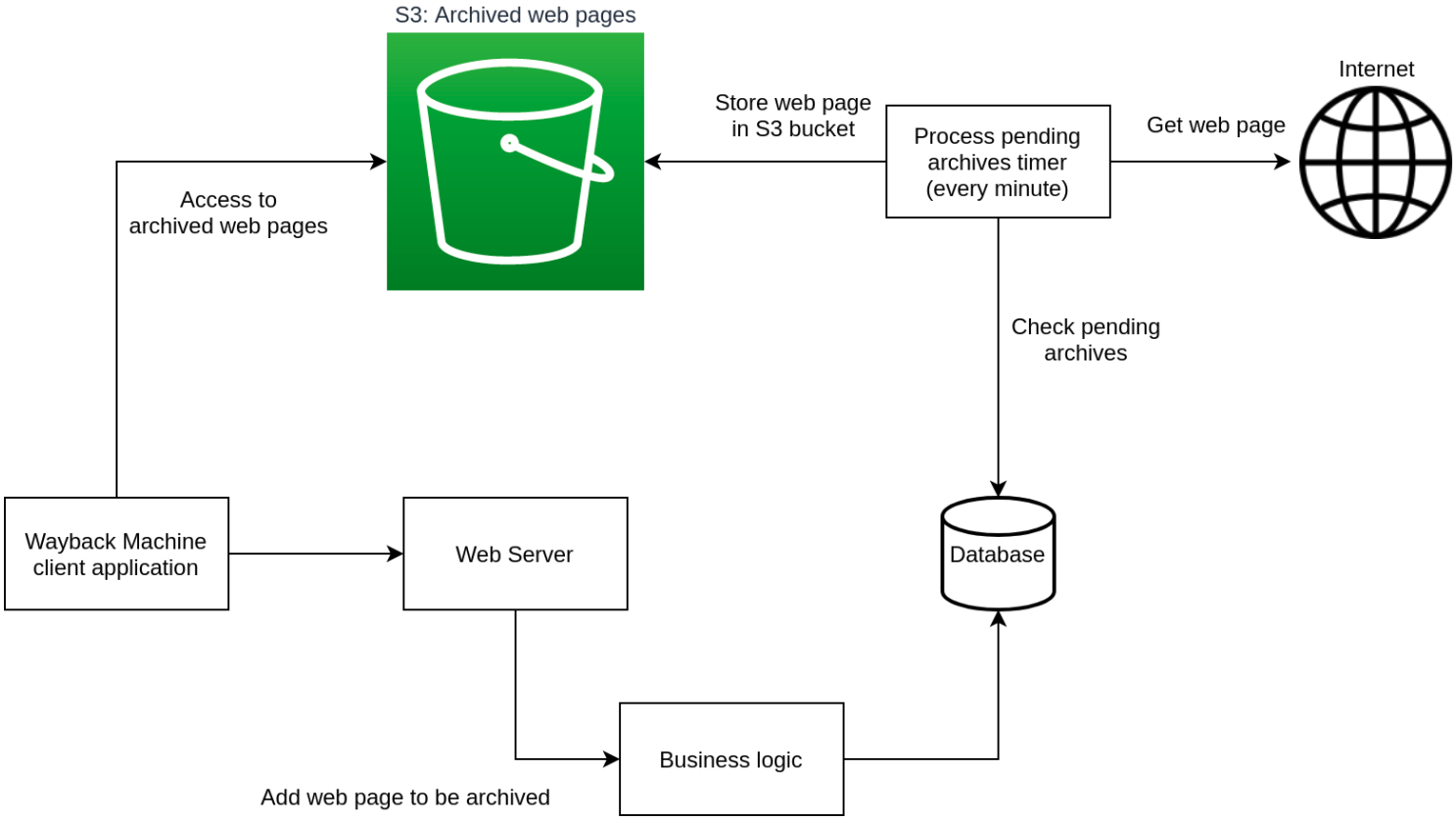
2. Въведение

Wayback Machine е система, подобна на „https://archive.org/web/“, която по зададен URL прави архив на дадена страница. След архивиране може да се влезе и да се навигира по страници и да се търси архив за страница. Ако има - всяко от архивираните се показва в календар или списък.

3. Теория

За описание на начина на работа на Wayback Machine, прочетете документацията на системата. Следва описание на начина на работа на функционалността за съхранение на архивираните сайтове в S3 bucket.

S3 услугата на AWS дава възможност на bucket да се качат статични файлове, които да се представят като уеб страница посредством опцията „Static website hosting“. Тази функционалност може да се използва от системата Wayback Machine за качване, съхранение и достъпване на архивирани сайтове в cloud. За прехвърляне на свалени ресурси за даден сайт от локалната файлова система на S3 bucket, може да се използва библиотеката AWS SDK for PHP и конкретно класовете Aws\S3\S3Client и Aws\S3\Transfer.



Фиг. 1 Архитектура на системата

4. Използвани технологии

- Системата е проектирана да работи на операционна система Ubuntu 20.04
- За обработка на заявки се използва уеб сървър Apache2
- Периодичното изпълнение на скриптове се осъществява от systemd timer
- Базата от данни е MySQL
- Сървърната логика е изградена с езика PHP
- Използвана е библиотеката AWS SDK for PHP за изпращане на команди към AWS S3
- За потребителския интерфейс са използвани технологиите HTML, CSS и JavaScript

5. Инсталация и настройки

5.1. Инсталация

1. Създайте bucket в S3
2. Задайте bucket policy по подразбиране достъпът до ресурси да е публичен. Примерно policy:

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "PublicReadGetObject",
            "Effect": "Allow",
            "Principal": "*",
            "Action": "s3:GetObject",
            "Resource": "arn:aws:s3:::wbmachine/*"
        }
    ]
}
```

3. Задайте опцията Static website hosting да бъде Enabled.
4. Свалете изходния код на Wayback Machine.
5. В config/wbmachine.conf задайте стойности на environment variable S3_ENABLED=1 и на S3_URL - URL-а на S3 bucket-а (например https://wbmachine.s3.amazonaws.com/)
6. В config/wbmachine-process-pending-archives.service задайте стойности на environment variable S3_ENABLED=1, на S3_REGION, и на S3_LOCATION. Примерни стойности: S3_LOCATION=s3://wbmachine S3_REGION=us-east-1
7. Изпълнете със супер потребител инсталационния скрипт

```
$ sudo ./setup/install.sh
```

8. Добавете данни за оторизация в /home/wbmachine/.aws/credentials

5.2. Настройки

Пълното описание на възможностите на настройка на системата можете да прочетете в документацията на Wayback Machine. Представените опции в този документ се отнасят само до функционалността за съхранение на архивирани сайтове в S3 bucket.

Връзката към S3 bucket се конфигурира посредством следните environment variables:

5.2.1. Конфигурация на сървърното приложение

- **S3_ENABLED** - Приема стойности 1 (архивираните сайтове да се търсят на S3 bucket) или 0 (архивираните сайтове да се търсят на локалната файлова система)
- **S3_URL** - URL на S3 bucket, в който се архивират сайтовете (напр. https://wbmachine.s3.amazonaws.com/). Задължителна опция, ако S3_ENABLED=1

Можете преди да пуснете инсталационния скрипт да промените параметрите във файл config/wbmachine.conf

5.2.2. Конфигурация на Process Pending Archives Timer

- **S3_ENABLED** - Приема стойности 1 (сайтовете да се архивират в S3 bucket) или 0 (сайтовете да се архивират на локалната файлова система).
- **S3_REGION** - AWS регион, в който S3 bucket се намира. Задължителна опция, ако S3_ENABLED=1
- **S3_LOCATION** - S3 scheme URI на bucket (напр. "s3://wbmachine"). Задължителна опция, ако S3_ENABLED=1

Можете преди да пуснете инсталационния скрипт да промените параметрите във файл wbmachine/config/wbmachine-process-pending-archives.service

6. Кратко ръководство за потребителя

Упътване как можете да използвате системата можете да прочетете в документацията на Wayback Machine.

7. Примерни данни

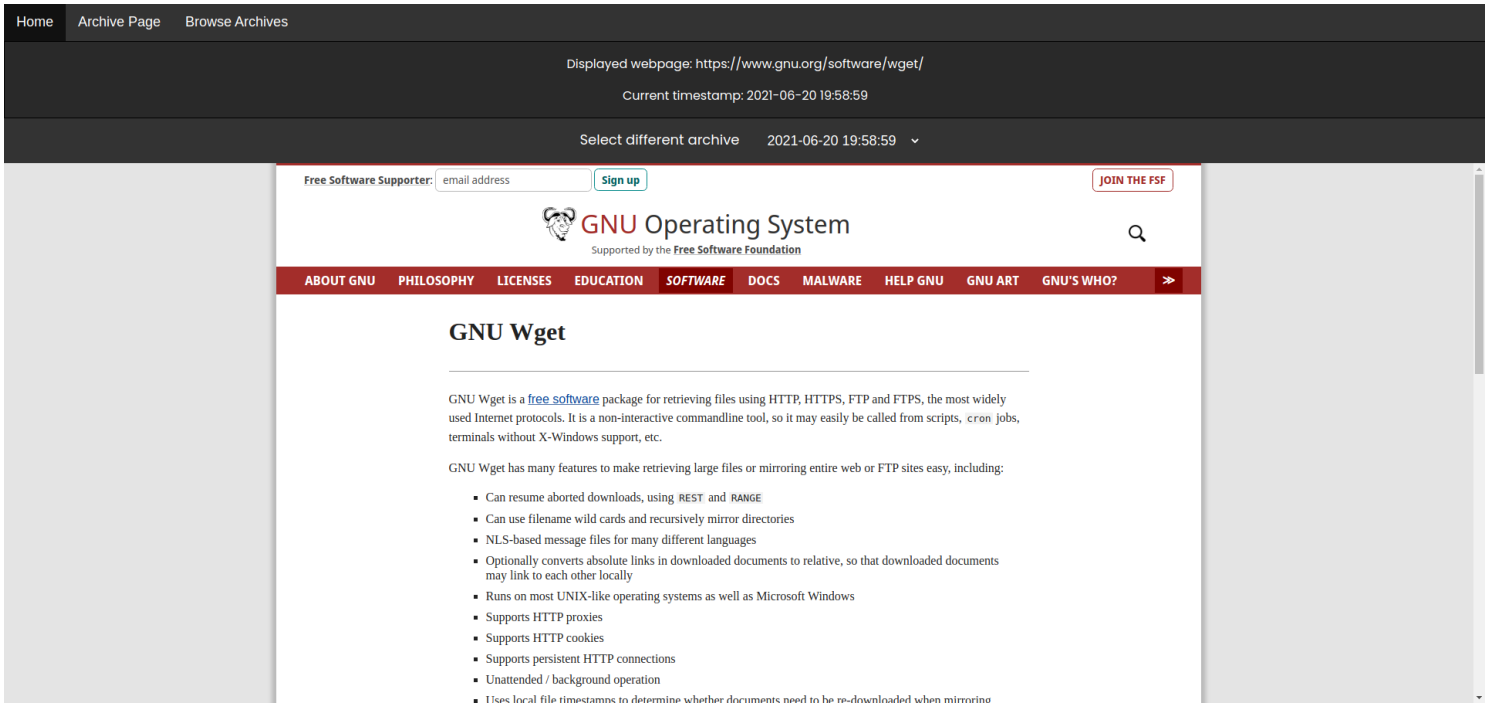
За демонстрация на работата на системата с примерни данни е използвана следната страница: https://www.gnu.org/software/wget/

Преди инсталация на системата можете да зададете следните environment variables, за да достъпвате предварително качени архивирани данни върху съществуващ S3 bucket:

S3_ENABLED=1 S3_URL=https://wbmachine.s3.amazonaws.com/ S3_LOCATION=s3://wbmachine S3_REGION=us-east-1

След инсталация на системата можете като суперпотребител да изпълните скрипта ./setup/db_fill.sh или ръчно да изпълните в базата заявките в ./setup/db_fill.sql

Отворете страницата Browse Archives, сложете в полето URL „https://www.gnu.org/software/wget/“, натиснете Submit, изберете timestamp и архивираната страница ще се визуализира с ресурси, свалени от S3 bucket-а.



Фиг. 2 Визуализация на архивирана страница

8. Описание на програмния код

За описание на програмния код на Wayback Machine, прочетете документацията на системата. Представеното описание се отнася за добавката на S3 възможностите към Wayback Machine.

8.1. Архивиране на сайтове в S3 bucket

Към скрипта за обработка на заявки за архивиране е добавена нова функция „fetch_site_s3“, която изпълнява обработка на заявка за архивиране чрез запазване в S3 bucket. Тя изпълнява следните действия:

1. Създава временна директория на диска на сървъра, в която да се свалят данните за архивиране
2. Сайтът, който ще се архивира, се изтегля в новосъздадената директория
3. Създава се обект за връзка към AWS API: Aws\S3\S3Client
4. Създава се обект за трансфер на файлове от локалната файлова система към S3 bucket: Aws\S3\Transfer
5. Изпълнява се трансфера
6. Изтриват се временно създадените файлове на диска на сървъра

Чрез environment variable S3_ENABLED скриптът за обработка на заявки за архивиране определя дали да се изпълнява функция „fetch_site_s3“ или оригиналната, за съхранение на локалната файлова система, „fetch_site_local“

8.2. Изтегляне на сайтове от S3 bucket

Чрез environment variable S3_ENABLED сървърът определя дали да предостави линк към ресурс от локалната файлова система, или от зададен S3 bucket.

9. Приноси на студента, ограничения и възможности за бъдещо развитие

9.1. Приноси на студента

- Добавяне на възможност за архивиране на сайтове в S3 bucket
- Добавяне на възможност за изтегляне на архивирани сайтове от S3 bucket
- Добавяне на опции за конфигурация, свързани с S3 функционалността

9.2. Ограничения и бъдещо развитие

В документацията на системата Wayback Machine са описани ограниченията, пред които системата е изправена, както и идеи за бъдещо развитие.

10. Какво научих

- Phar archives, добавяне на AWS SDK към PHP чрез phar
- Изпращане на команди към S3 bucket чрез S3Client
- Качване на директория и всички файлове в нея в S3 чрез Amazon S3 Transfer Manager
- Предоставяне на публичен достъп до ресурси в S3 bucket
- Концепцията за директории в S3 bucket

11. Списък с фигури и таблици

12. Използвани източници

```
[1] Class S3Client, AWS SDK for PHP 3.x [https://docs.aws.amazon.com/aws-sdk-php/v3/api/c
[2] Amazon S3 Transfer Manager with AWS SDK for PHP Version 3 [https://docs.aws.amazon.co
```