

Universitatea “Alexandru Ioan Cuza”, Iași

Facultatea de informatică

Surf

Parcurgere și procesare distribuită a paginilor web în cloud
pentru extragerea informațiilor



Ovidiu Pricop

2 Octombrie 2016

Declarație privind originalitate și respectarea drepturilor de autor

Prin prezenta declar că lucrarea de licență cu titlul “Surf” este scrisă de mine și nu a mai fost prezentată niciodată la o altă facultate sau instituție de învățământ superior din țară sau străinătate. De asemenea, declar că toate sursele utilizate, inclusiv cele preluate de pe Internet, sunt indicate în lucrare, cu respectarea regulilor de evitare a plagiatului:

- toate fragmentele de text reproduse exact, chiar și în traducere proprie din altă limbă, sunt scrise între ghilimele și dețin referința precisă a sursei;
- reformularea în cuvinte proprii a textelor scrise de către alți autori deține referința precisă;
- codul sursă, imaginile etc. preluate din proiecte open-source sau alte surse sunt utilizate cu respectarea drepturilor de autor și dețin referințe precise;
- rezumarea ideilor altor autori precizează referința precisă la textul original.

Iași, 2 Octombrie 2016

Ovidiu Pricop

Declarație de consimțământ

Prin prezenta declar că sunt de acord ca Lucrarea de licență cu titlul “Surf”, codul sursă al programelor și celelalte conținuturi (grafice, multimedia, date de test etc.) care însoțesc această lucrare să fie utilizate în cadrul Facultății de Informatică. De asemenea, sunt de acord ca Facultatea de Informatică de la Universitatea Alexandru Ioan Cuza Iași să utilizeze, modifice, reproducă și să distribuie în scopuri necomerciale programele-calculator, format executabil și sursă, realizate de mine în cadrul prezentei lucrări de licență.

Iași, 2 Octombrie 2016

Ovidiu Pricop

Rezumat

La nivelul internetului, traficul global a crescut de aproximativ 850 de ori in perioada 2000 - 2015 [1]. World wide web-ul reprezinta o parte semnificativa a volumului de informatii interschimbate pe internet. In anul 2015 existau peste jumatate de miliard de situri web accesibile [2]. Fiecare pagina web raspunde anumitor nevoi (sociale, financiare, educationale etc.). O singura sursa de informatii este uneori suficienta pentru a raspunde nevoilor unui utilizator. Alteori, este necesar un ansamblu de surse informative (e.g. newsletter-e zilnice din 5 surse diferite), pentru a urmari un subiect din mai multe perspective sau a-i intregi continutul.

Prezenta lucrare urmareste elaborarea unui serviciu web distribuit in cloud pentru parcurgerea, selectarea, colectarea si indexarea informatiilor la nivelul world wide web-ului.

Cuprins

Rezumat	3
Introducere	5
1 Crawling-ul paginilor web	6
1.1 Puncte de pornire	6
1.2 Parcurgere	6
1.3 Selectia informatiilor	7
1.4 Viteza, politici de respingere si drepturi de autor	7
1.5 Paralelizare	8
1.6 Logarea actiunilor si retentia datelor	8
2 Crawling in cloud	10
2.1 Infrastructura	11

Introducere

Un serviciu web reprezinta o componenta functionala ce indeplineste anumite sarcini. Comunicarea cu un serviciu web se realizeaza independent de platforma, limbajul de programare sau sistemul de operare pe care este dezvoltat. Schimbul de informatii se realizeaza prin mesaje text ce respecta un format standardizat precum xml¹ sau json².

Aplicatia "Surf" reprezinta un serviciu web specializat in web crawling³, dezvoltat folosind tehnologii cloud din cadrul Amazon Web Services. Se urmareste crearea unui serviciu web cu disponibilitate permanenta, scalabil si de inalta putere computationala care sa orchestreze colectarea distribuita de informatii din aria definita de utilizator.

Comunicarea cu aplicatia "Surf" se realizeaza prin intermediul unui API Restful⁴ construit pe platforma AWS⁵ API Gateway. Autentificarea utilizatorilor se va realiza printr-un serviciu OpenID Connect (e.g. Google, Facebook etc.). Autorizarea va avea ca principala componenta AWS IAM. Utilizatorilor le vor fi repartizate, in functie de privilegiile asociate cu cheia de autentificare, o serie de roluri (i.e. drepturi de access asupra resurselor din cadrul serviciului "Surf"). Executia codului propriu-zis, gazduit de functii AWS Lambda, va interactiona cu serviciul pentru baze de date no-sql AWS DynamoDB pentru a permite accesul la informatii cheie pentru functionalitatea aplicatiei (metadata crawling, date acces utilizatori etc.). Mediul de procesare distribuita va fi sustinut de AWS Simple Workflow Service si configurat dupa preferintele utilizatorului. Datele extrase din procesul de web-crawling vor fi salvate in mediul persistent de stocare AWS S3. Evenimentele legate de parcurgerea siturilor vor fi expuse, ca metadata, intr-o coada AWS SQS si vor fi accesibile utilizatorilor prin procesul de long-polling asupra acestei cozi.

¹Extensible Markup Language - <https://www.w3.org/XML/>

²JavaScript Object Notation - <http://www.json.org/>

³https://en.wikipedia.org/wiki/Web_crawler

⁴https://en.wikipedia.org/wiki/Representational_state_transfer

⁵Amazon Web Services

1 Crawling-ul paginilor web

Ca sens general, un crawler web reprezinta un program care parcurge, prin cereri succesive, situri web. Programatic, vom considera urmatoarele aspecte cheie in proiectarea unui crawler distribuit:

1. Punctele de pornire in parcurgerea siturilor web;
2. Adancimea maxima a parcurgerii recursive a siturilor (i.e. cea mai indepartata pagina la care se poate ajunge, de la punctul de pornire, prin accesarea succesiva a legaturilor de tipul hyperlink;
3. Politica de selectie a informatiilor din paginile parcurse[3];
4. Viteza de crawling, fisierul *robots.txt*⁶ si respectarea drepturilor de autor;
5. Politica de paralelizare a procesului de web crawling;
6. Politica de retentie temporara a rezultatelor parcurgerii siturilor web si logare a actiunilor serviciului de crawling;

In cele ce urmeaza, se va descrie particularizarea aspectelor generale enumerate mai sus in cadrul serviciului web "Surf". Se va crea, astfel, contextul dezvoltarii aplicatiei si se vor puncta principalele componente functionale implicate.

1.1 Puncte de pornire

Pentru a putea parcurge siturile web, crawler-ul "Surf" necesita unul sau mai multe puncte de pornire. Un punct de pornire este definit printr-un URL⁷ si este furnizat, ca input din partea utilizatorului, la initializarea crawler-ului.

1.2 Parcurgere

Crawler-ul web "Surf" are ca scop extragerea informatiilor cerute de catre utilizator. Intrucat utilizatorul are posibilitatea de a furniza, ca punct de pornire, un domeniu web relevant pentru informatia cautata, parcurgerea recursiva se va executa in maniera breadth-first. Asadar, crawler-ul va vizita toate legaturile de tip hyperlink din pagina curenta a parcurgerii inainte de a accesa legaturile din pagina urmatoare din punct de vedere ierarhic. Adancimea maxima a arborelui de legaturi realizat prin parcurgerea URL-urilor va fi definita de catre utilizator, la initializarea sesiunii de crawling.

⁶<http://www.robotstxt.org/robotstxt.html>

⁷https://en.wikipedia.org/wiki/Uniform_Resource_Locator

1.3 Selectia informatiilor

Selectia informatiilor necesita parcurgerea siturilor web aflate la adresele URL pe care crawler-ul le are in vedere, parsarea informatiilor in functie de tipul lor (e.g. html, xml, json, text) si extragerea blocurilor de continut aferente.

Un "bloc de continut" asociat unui URL reprezinta o parte a intregului continut aflat la URL-ul respectiv, filtrata dupa anumite caracteristici stabilite de utilizatorul serviciului web. Aceste caracteristici pot fi:

- Pentru fisiere HTML/XML:
 - selectori CSS/jQuery
 - cuvinte cheie
 - o combinatie logica a punctelor de mai sus (e.g. toate tagurile <p> in care se afla cuvintele cheie "om" si "luna")
- Pentru fisiere text:
 - cuvinte cheie

Pentru tipurile de continut enumerate mai sus sau pentru alte tipuri de continut aflate la adresele URL vizitate de crawler, utilizatorul poate defini filtre bazate pe continutul textual al URL-ului. Spre exemplu, se pot evita toate URL-urile care nu satisfac o anumita expresie regulata sau un anumit tip de continut MIME⁸ (e.g. "image/png").

Evitarea selectarii informatiilor duplicate se amelioreaza prin normalizarea⁹ URL-urilor parcurse de catre crawler.

1.4 Viteza, politici de respingere si drepturi de autor

Siturile web pot implementa variate modalitati de contracarare a incercarilor de crawling. Aplicatia "Surf" incearca sa minimizeze riscul de respingere a cererilor de accesare a anumitor resurse web printr-o implementare neintruziva a procesului de crawling. Cateva aspecte esentiale care sunt luate in considerare in ceea ce priveste o astfel de implementare sunt urmatoarele:

- Minimizarea volumului de date preluat de pe un anumit domeniu prin diferite metode de filtrare a linkurilor urmarite (e.g. o anumita structura a URL-ului, un anumit tip de date care se gaseste la URL-ul respectiv);

⁸https://en.wikipedia.org/wiki/Media_type

⁹https://en.wikipedia.org/wiki/URL_normalization

- Introducerea pauzelor temporale aleatoare intre accesari succesive a datelor apartinand aceluiasi domeniu.

Crawler-ul web "Surf" poate satisface numeroase cerinte ale utilizatorilor. O parte dintre aceste cerinte poate veni din partea sistemelor anti-malware. In acest caz, nu se doreste respectarea fisierului *robots.txt* (utilizat drept referinta, pentru crawleri, asupra URL-urilor accesibile ale domeniului vizitat), deoarece exista pericolul ca un sit malitios sa blocheze o eventuala scanare. De aceea, aplicatia "Surf" va putea fi configurata in ceea ce priveste ignorarea fisierului *robots.txt* in procesul de parcurgere a unui domeniu.

Crawler-ul "Surf" implementeaza un mecanism de *blacklisting*¹⁰. Siturile web sau domeniile care interzic procesul de crawling (e.g. prin ToS¹¹) vor fi adaugate unei liste de excludiune din procesul de parcurgere executat de crawler.

1.5 Paralelizare

"Surf" implementeaza un mecanism de crawling distribuit. Mai exact, exista posibilitatea de a imparti sarcinile de parcurgere a paginilor web prin rularea concurenta a mai multor instante de functii Lambda¹².

Exista doua moduri bine cunoscute de alocare a sarcinilor de web crawling: statica si dinamica[4]. Aplicatia "Surf" implementeaza un mecanism de distribuire statica a sarcinilor:

Fie n , *gradul de paralelizare*¹³ configurat pentru executia curenta a web crawler-ului. Alocarea statica va asocia fiecarui URL candidat pentru crawling un numar intreg in intervalul $[1, n]$ reprezentand identificatorul uneia dintre cele n instante paralele ale web crawler-ului. Sarcina va fi distribuita pe acea instanta si va fi executata.

1.6 Logarea actiunilor si retentia datelor

In timpul executiei, crawler-ul genereaza evenimente ce sunt adaugate la istoricul executiei curente. Aceste evenimente sunt utile, in primul rand, in cazul in care executia instantei curente a crawler-ului esueaza. Sarcinile de crawling web pot dura o perioada considerabila si nu este dezirabila refacerea intregului

¹⁰<https://en.wikipedia.org/wiki/Blacklisting>

¹¹https://en.wikipedia.org/wiki/Terms_of_service

¹²Amazon Web Services Lambda Functions

¹³Numarul maxim de executii simultane de sarcini de crawling

proces de parcurgere a siturilor, atat timp cat exista rezultate parțiale; executia se poate relua pornind de la ultimul eveniment valid inregistrat in istoric. In al doilea rand, istoricul este util pentru utilizator, deoarece il poate avertiza in legatura cu statutul executiei curente a web crawler-ului.

Datele rezultate in urma procesului de crawling al paginilor web sunt salvate in serviciul persistent de stocare S3, din cadrul Amazon Web Services. Deoarece aceste date pot deveni voluminoase (in functie de politica de selectie configurata), aplicatia "Surf" salveaza metadate asociate acestor rezultate intr-o tabela DynamoDB. Simultan, o parte dintre metadate sunt plasate intr-o coada SQS¹⁴. Utilizatorul poate sa efectueze long-polling asupra acestei cozi pentru a extrage informatiile legate de procesul de crawling. Pe baza acestor informatii (metadate), utilizatorul are optiunea de a accesa rezultatele complete corespunzatoare rularii crawler-ului web, aflate in S3.

¹⁴Amazon Simple Queue Service

2 Crawling in cloud

Aplicatia "Surf" este construita si ruleaza in cloud. Dezvoltarea unei aplicatii pe suportul unei infrastructuri cloud este substantial diferita fata de abordarea clasica, pe o singura masina de calcul. Cloud computing-ul pune la dispozitie, la cerere, un set de resurse (e.g. putere de procesare, spatiu de stocare) pe care un utilizator (e.g. aplicatia "Surf") le poate folosi. Faptul ca toate aceste resurse sunt administrate de fiecare serviciu cloud in parte usureaza sarcinile consumatorului: se pot solicita mai multe sau mai putine dupa gradul de incarcare al aplicatiei dezvoltate in cloud. Acest lucru se realizeaza intr-o maniera transparenta pentru utilizator, fara a necesita efort suplimentar din partea acestuia.

Dezvoltarea aplicatiei "Surf" folosind servicii cloud mai prezinta un avantaj considerabil, deoarece comunicarea prin internet intre componentele aplicatiei si, implicit, utilizarea unui format agnostic de arhitectura sau limbaj de programare, conduce la un grad mare de decuplare. Fiind decuplate, componentele aplicatiei pot fi construite, testate si depanate¹⁵ in izolare si pot functiona ca module (plugin-uri) in alcatuirea arhitecturii sistemului.

A dezvolta un serviciu web folosind exclusiv tehnologii cloud permite realizarea unui sistem de subscriptii granular. Un utilizator al aplicatiei "Surf" poate solicita una sau mai multe instante ale crawler-ului care sa ruleze in paralel, in functie de necesitatile proprii. Serviciile cloud permit aplicatiei sa scaleze¹⁶ orizontal ca timp, permitand utilizatorului realizarea mai multor sarcini de parcurugere automata web (deci, in consecinta, acoperirea unui volum mai mare de informatii) in acelasi interval de timp.

Un alt aspect relevant in realizarea serviciului web "Surf" este caracterul sau *serverless*. Infrastructura *serverless* implica executia codului sursa al programului in cadrul serviciilor cloud folosite pentru dezvoltarea aplicatiei. Cererile de executie a codului sunt monetizate conform unei masuri abstracte corespunzatoare resurselor utilizate pentru satisfacerea solicitarii, ceea ce difera fata de modelul clasic, in care ar fi trebuit achizitionat hardware in acest scop[5].

¹⁵*engl.* debugged

¹⁶<https://en.wikipedia.org/wiki/Scalability>

2.1 Infrastructura

Esenta aplicatiei "Surf" se afla in codul executat de serviciul web AWS Lambda, care preia componentele dezvoltate pentru crawling web si le executa drept programe autonome folosind infrastructura AWS. AWS Lambda executa acest cod doar cand este nevoie (e.g. la cererea utilizatorului care doreste sa demareze operatiunea de crawling). Se minimizeaza, astfel, atat costurile utilizatorului in ceea ce priveste rularea crawler-ului web (nu necesita hardware dedicat pentru procesarea paginilor web), cat si costurile dezvoltatorilor crawler-ului, care nu trebuie sa rezerve hardware in cloud (e.g. masini virtuale Amazon EC2, inchiriere de hosturi), ca mai apoi, cand nu exista cereri suficiente, aceste masini de calcul sa ramana nefolosite. De asemenea, se renunta la necesitatea de a avea un load-balancer care sa distribuie sarcinile de crawling pe capacitatea hardware disponibila, deoarece acest lucru este administrat, in backend, de catre AWS Lambda.

Funcțiile programatice disponibile prin serviciul AWS Lambda pot fi accesate de catre utilizatori printr-un API Restful gazduit de AWS Api Gateway. Aceasta platforma permite integrarea cererilor HTTP externe cu mecanismul de autentificare si autorizare a utilizatorilor si codul executat de instantele functiilor Lambda. De asemenea, aplicatia utilizeaza functionalitati de management, monitorizare si analiza a traficului din cadrul API-ului "Surf" care permit, printre altele, inregistrarea si monetizarea functionalitatilor oferite de serviciul web pentru fiecare utilizator in parte.

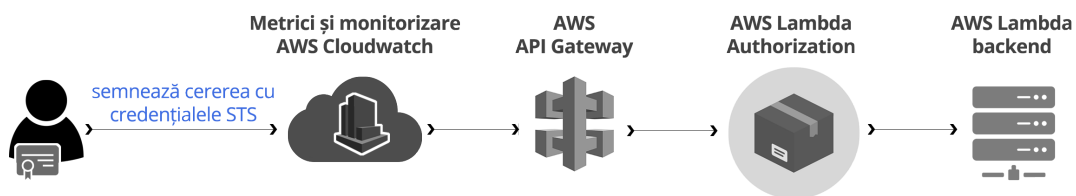
Pentru a putea accesa functionalitatile oferite de API-ul "Surf", utilizatorii trebuie sa se autentifice cu un furnizor tert de identitati ce suporta OpenID¹⁷. Credentialele obtinute sunt impachetate intr-o cerere catre AWS Security Token Service¹⁸, unde sunt validate. In cazul in care validarea este indeplinita cu succes, Security Token Service returneaza utilizatorului credentiale temporare pentru a accesa servicii din cadrul AWS. Utilizatorul trebuie sa includa aceste credentiale in fiecare cerere catre API-ul "Surf" (autentificare). Odata autentificat, utilizatorul poate interactiona doar cu acele resurse ale API-ului pentru care are drepturi de acces.

¹⁷<https://en.wikipedia.org/wiki/OpenID>

¹⁸Serviciu web ce permite unui utilizator sa solicite credentiale temporare pentru autentificarea in cadrul platformei Amazon Web Services



Cuprins 1: Procesul de autentificare [6]



Cuprins 2: Procesul de autorizare [6]

Prin AWS Identity Access Management (IAM) se vor stabili restricțiile de acces asupra API-ului. Mecanismul de asignare a permisiunilor va fi susținut prin definirea unor roluri. Utilizatorii vor fi grupați, din punct de vedere al drepturilor de acces, în mai multe categorii. Spre exemplu, un utilizator poate avea rolul "administrator" (prescurtat *Ad*), iar un alt utilizator poate avea atât rolul "administrator" (*Ad*) cât și "utilizator de baza" (prescurtat *Ub*). Atât *Ad*, cât și *Ub* vor avea asignate politici de acces asupra datelor AWS. O astfel de politică este următoarea:

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "apigateway:*"
      ],
      "Resource": [
        "*"
      ]
    }
  ]
}

```

Politica de acces de mai sus ii permite utilizatorului ce ii este asignata accesul la toate resursele API-ului "Surf". Aceasta politica este potrivita pentru un administrator dar foarte periculoasa pentru un utilizator ce acceseaza pentru prima data serviciul "Surf". De aceea, se vor defini drepturi de acces granulare care sa aiba in vedere restrictionarea la nivel "need-to-know"¹⁹ pentru fiecare entitate ce trimite cereri catrei API (autorizare).

Procesul de asignare a rolurilor pentru utilizatori este coordonat atat manual (pe baza de ierarhie: administrator - utilizator premium - utilizator standard) cat si automat (i.e. printre altele, toti utilizatorii abia inregistrati sunt utilizatori standard). Asignarea manuala are prioritate asupra asignarii automate. Datele despre rolul fiecarui utilizator vor fi pastrate in baza de date.

Crawler-ul "Surf" utilizeaza AWS DynamoDB ca suport pentru baza de date. DynamoDB reprezinta un serviciu cloud scalabil pentru baze de date no-sql. Datele dintr-o baza de date non-relationala (no-sql) pot fi modelate fara constrangerile unei baze de date relationale (e.g. tabularitate). Acest lucru atrage, in cadrul aplicatiei "Surf", elemente operationale cheie pentru beneficiul carora s-a renuntat la o abordare sql (e.g. Oracle), printre care:

- *data sharding*²⁰ - este necesar un sistem distribuit de baze de date care sa poata scala orizontal odata cu dimensiunea datelor din baza de date si odata cu cresterea numarului de utilizatori;

¹⁹https://en.wikipedia.org/wiki/Need_to_know

²⁰[https://en.wikipedia.org/wiki/Shard_\(database_architecture\)](https://en.wikipedia.org/wiki/Shard_(database_architecture))

- *scheme de date dinamice*²¹ - se dorește posibilitatea schimbării structurii datelor din baza de date fără a recrea tabelele corespunzătoare; acest lucru trebuie avut în vedere deoarece crawler-ul stochează în baza de date metadate obținute prin parcurgerea paginilor web; poate apărea oricând necesitatea introducerii unor noi metadate sau schimbării structurii celor existente, cu scopul satisfacerii nevoilor utilizatorilor.

Câteva dintre cele mai importante roluri ale bazei de date sunt gazduirea metadatelor asociate rezultatelor procesului de crawling, persistarea asignarilor între identitățile utilizatorilor și rolurile lor, menținerea istoricului evenimentelor de crawling cu scopul reconstruirii stării procesului de parcurgere a paginilor web în cazul unei erori și stocarea datelor asociate volumului de trafic în cadrul infrastructurii AWS generat de fiecare utilizator.

Pentru a obține un serviciu web distribuit pentru crawling este necesară coordonarea activităților independente din cadrul sistemului, sincronizarea pasilor necesari procesului de crawling și, în final, integrarea rezultatelor. Pentru acest lucru aplicația ”Surf” folosește serviciul web AWS Simple Workflow Service (SWF). SWF are capacități de coordonare a activităților ce se doresc îndeplinite (i.e. executia funcțiilor Lambda, sau *Lambda Tasks*) și management al stării unei aplicații (i.e. se porneste un proces de agregare a datelor de la crawleri care au rulat în paralel doar după ce aceștia și-au terminat executia).

Datele obținute în urma procesului de crawling sunt stocate utilizând serviciul web Amazon S3. Fiecare rulare a unei instanțe a crawler-ului distribuit generează un *bucket*²² S3. Datele din bucket-uri sunt disponibile, pentru utilizatori, prin plasarea de metadate precum numele bucket-ului într-o coadă Amazon SQS, asupra căreia se execută un mecanism de long-polling pentru extragerea informațiilor. Datele din bucket-uri au un timp limitat de viață, configurabil relativ la preferințele utilizatorului. O funcție Lambda, programată să se execute, periodic, prin serviciul AWS CloudWatch, elimină bucket-urile a căror durată de viață a depășit termenul limită stabilit.

²¹<https://www.mongodb.com/scale/dynamic-schema-design>

²²Unitate în cloud (AWS S3) ce stochează date și careia îi pot fi atribuite permisiuni de acces și metadate corespunzătoare

Bibliografie

- [1] "Visual Networking Index", Cisco Systems
- [2] <http://www.internetlivestats.com/total-number-of-websites/>
- [3] https://en.wikipedia.org/wiki/Web_crawler#Selection_policy
- [4] https://en.wikipedia.org/wiki/Distributed_web_crawling
- [5] Miller, Ron (24 Nov 2015). "AWS Lambda Makes Serverless Applications A Reality". TechCrunch. Retrieved 10 July 2016.
- [6] Sursele pictogramelor includ *iconfinder.com* si *simpleicon.com*.