UNIVERSITATEA POLITEHNICA BUCUREȘTI

FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

DEPARTAMENTUL CALCULATOARE

|  |  |
| --- | --- |
| upb | cs |

PROIECT DE DIPLOMĂ

Serviciu de crawling și de indexare de date

Proiect dimplomă 2019

Barbu Iulian-Marian

**Coordonator științific:**

Prof. dr. ing. Nicolae Țăpuș

BUCUREŞTI

2019

**CUPRINS**

[Sinopsis 2](#_Toc7074723)

[Abstract 2](#_Toc7074724)

[Mulțumiri 3](#_Toc7074725)

[1 Introducere 4](#_Toc7074726)

[1.1 Context 4](#_Toc7074727)

[1.2 Problema 4](#_Toc7074728)

[1.3 Obiective 4](#_Toc7074729)

[1.4 Structura lucrării 4](#_Toc7074730)

[2 Analiza și specificarea cerințelor 5](#_Toc7074731)

[3 Studiu de piață / Abordări existente 6](#_Toc7074732)

[3.1 Indicații formatare figuri 6](#_Toc7074733)

[4 Soluția propusă 8](#_Toc7074734)

[4.1 Indicații formatare formule 8](#_Toc7074735)

[5 Detalii de implementare 9](#_Toc7074736)

[5.1 Indicații formatare tabele 9](#_Toc7074737)

[6 Studiu de caz / Evaluarea rezultatelor 11](#_Toc7074738)

[7 Concluzii 13](#_Toc7074739)

[8 Bibliografie 14](#_Toc7074740)

[9 Anexe 1](#_Toc7074741)

# Sinopsis

Acest proiect de diplomă propune o soluție pentru crawling-ul și indexarea unor resurse din internet în vederea achiziției de date text. Scopul acestui proiect este dezvoltarea unui serviciu care să ofere o modalitate simplă de a efectua procese de crawling peste anumite resurse din internet, cât și indexarea unor date preluate prin procesul de crawling. Serviciul este bazat pe o stivă de programe, implementează modelul SaaS (software as a service) și beneficiază de arhitectură distribuită, atât din punctul de vedere al procesării cât și al stocării. Interacțiunea cu serviciul se realizează printr-o interfață web, care oferă un panou de administrare al serviciului. În realizarea serviciului au fost folosite tehnologii precum Apache Nutch, Apache Hadoop, Apache Solr, Apache Wicket, Selenium, și Docker, pentru punerea în funcțiune a unui serviciu cu etape de configurare simple, care beneficiază de procesare și stocare distribuită pentru procese de crawling de mari proporții.

# Introducere

## Context

Proiectul de față urmărește dezvoltarea unei soluții în vederea achiziției de date text din internet, și ingestia acestor date text într-o platformă de căutare. Soluția prezentată în cuprinsul acestei lucrări constă într-un serviciu bazat pe o stivă de programe software (Saas) care va oferi modalități de crawling și indexare distribuită de date text, preluate din pagini web, servite prin protocoalele HTTP și HTTPS.

Motivația pentru care am ales implementarea unui astfel de serviciu izvorăște din interesul pe care îl am pentru dezvoltarea programelor software care funcționează distribuit și pentru tehnologiile folosite în dezvoltarea serviciului, foarte relevante pentru diverse sfere IT precum procesare în cloud, sisteme distribuite, big data, minare de date, infrastructură cloud, automatizare a proceselor de testare, motoare de căutare text etc. Pe lângă această motivație personală, implementarea unei astfel de soluții, care oferă în cele din urmă o interfață web pentru interacțiunea cu serviciul, pentru a efectua procese de crawling de mari proporții peste pagini web, înseamnă o simplificare enormă pentru un utilizator, care în primul rând nu va fi nevoit să folosească soluții contra cost, și în al doilea rând va avea nevoie de un nivel de pregătire tehnică minimă pentru a pune serviciul în funcțiune.

Serviciul prezentat în cuprinsul lucrării prezente, în contextul actual al dezvoltării unor servicii asemănătoare, este relevant din mai multe puncte de vedere, și anume:

- nu sunt percepute costuri bănești pentru utilizare, ceea ce îl face corespunzător pentru acei utilizatori care nu sunt interesați să investească într-o astfel de soluție, din diverse motive: din lipsa unui buget, din intenția de a experimenta, în vederea investirii ulterioare într-un serviciu contra cost, cu beneficii suplimentare, sau dintr-un motiv pe care eu îl consider cel mai important, și anume utilizarea personală, cu interesul de a exploata serviciul, îmbunătățind-ul și contribuind la dezvoltarea lui.

- proiectul principal pe baza căruia serviciul a fost dezvoltat, Apache Nutch, este un proiect open source matur, apărut în 2003, care oferă un web crawler distribuit. Nutch are o bază de utilizatori întinsă, iar unii dintre ei încearcă să ofere tutoriale și documentație pentru proiect, însă de-a lungul timpului, acestea au rămas învechite. Pe lângă acest aspect, există găuri importante în documentație. Documentația este constituită din liste de articole publicate de utilizatori Apache Nutch. În opinia mea, aceste liste de articole și tutoriale nu reprezintă o documentație oficială și suficientă, pentru utilizatorii care vor să înceapă să folosească web crawler-ul, sau care vor să contribuie, și de aceea poate descuraja pe cei mai mulți dintre noii veniți. Relevanța serviciului dezvoltat în cuprinsul acestei lucrări, în contextul tocmai introdus, apare prin faptul că oferind un serviciu de web crawling, bazat pe Apache Nutch, pentru utilizatorii fără pregătire tehnică și nu numai, un punct important al lucrării prezente este încercarea atașării unei documentații oficiale actualizate, alături de componentele principale care formează serviciul, pentru Apache Nutch versiunea 1.15.

## Problema

Problema pe care acest proiect încearcă să o rezolve este lipsa unui program software, ori a unui serviciu, gratuite, care să ofere posibilitatea efectuării unui proces de crawling de mari proporții peste pagini web din internet. Această problemă este rezolvată, în cuprinsul acestei lucrări, prin dezvoltarea unui serviciu (SaaS), care necesită etape simple de configurare, pentru utilizatori care nu dispun de timpul necesar înțelegerii tehnice a stivei de programe software pe care trebuie să o creeze pentru a pune în funcțiune un crawler distribuit, și pentru a avea disponibile datele achiziționate într-o platformă de căutare.

## Obiective

Prin dezvoltarea serviciului de crawling și indexare de date text intenționez să ating trei obiective principale, și anume:

- serviciul să suporte mijloacele pentru configurarea infrastructurii de rulare și a procesului de crawling, după specificațiile utilizatorilor. Aceștia vor fi ghidați pe parcursul configurării, și li se va expune mici exemple de configurări uzuale peste care ar putea să adauge preferințele personale. Spre exemplu, utilizatorii vor putea să aleagă ce plugin-uri **(*REFERINȚĂ PLUGIN*)** pot folosi în diversele etape ale procesului de crawling. Dacă utilizatorii nu sunt experimentați, pot folosi opțiunile implicite. Acest scop este unul la care se va lucra o perioadă nedeterminată, întrucât modurile de configurare ale serviciului vor fi multiple și vor fi descoperite în timp.

- serviciul să fie controlat printr-un panou de control, expus printr-o interfață web, care să abstractizeze accesul la internele stivei de programe. Panoul de control trebuie să ofere posibilitatea pornirii unui proces de crawling, monitorizarea acestuia, și oprirea, în ultimul caz.

## Structura lucrării

Un paragraf în care fiecare dintre secțiunile următoare este prezentată în 1-2 fraze, punând accentul pe elementele cele mai semnificative din fiecare secțiune.

# Analiza și specificarea cerințelor

[Dezvoltare de produs] Acest capitol va analiza cerințele produsului din prisma potențialilor clienți și a scenariilor de utilizare preconizate, urmând a fi generată o lista de funcționalități.

[Cercetare] Acest capitol va introduce motivația realizării proiectului propus.

Dacă proiectul de licență face parte dintr-un proiect mai amplu (de exemplu un proiect complex, la care lucrează 2 studenți (ex: 1 student la front-end-ul aplicației, 1 student la back-end-ul aplicației), în acest capitol va fi explicat pe scurt ansamblul proiectului și ce parte din proiect este adresată de lucrarea propusă.

Criterii pentru calificativul Nesatisfăcător:

* [Dezvoltare de produs] Cerințele sunt imaginate de student pe baza unei analize a pieței;
* [Cercetare] Nu se oferă o motivație valida.

Criterii pentru calificativul *Satisfăcător*:

* [Dezvoltare de produs] Există un interviu, un client, analiza cerințelor este elaborată pe baza interviului;
* [Cercetare] Motivația este doar personala.

Criterii pentru calificativul *Bine*:

* [Dezvoltare de produs] Proces iterativ pe baza unor interviuri cu mai mulți clienți, dezvoltare MVP, reevaluare cerințe;
* [Cercetare] Motivația este legata de o necesitate științifica / tehnica explicită.

# Studiu de piață / Abordări existente

[Dezvoltare de produs] Ce soluții similare există pe piață? Care sunt limitările lor / pentru ce cazuri de utilizare sau pentru ce tip de clienți produsele existente pe piață nu răspund cerințelor? Care sunt indicatorii pe baza cărora sunt evaluate aceste produse, de către potențiali clienți, și unde sunt lipsurile/ care este oportunitatea generată de lipsurile acestea?

[Cercetare] Metode existente (sau „State of the Art“) se referă, de regulă, la nivelul curent de dezvoltare: care este starea curentă a domeniului, unde ne găsim, care este contextul. Care sunt soluțiile actuale prezente în literatura de specialitate și care sunt limitările lor? Ce direcții de explorare sunt recomandate în literatura de specialitate? Literatura de specialitate se refera la articole științifice recente, publicate în reviste cu factor de impact mare, sau în volumele unor conferințe de top, sau în cărți.

[Ambele] În încheierea acestui capitol se dorește descrierea tehnologiilor folosite în lucrare, cu alternative și cu argumente convingătoare calitative și cantitative.

Criterii pentru calificativul *Nesatisfăcător*:

* [Dezvoltare de produs] Sunt analizate superficial câteva produse de pe piață;
* [Cercetare] analiza literaturii limitata la grupuri de cercetare din România;
* [Ambele] Sunt descrise tehnologiile folosite în lucrare.

Criterii pentru calificativul *Satisfăcător*:

* [Dezvoltare de produs] Există un interviu, un client, analiza cerințelor este elaborată pe baza interviului.
* [Cercetare] analiza literaturii de specialitate din lume, fără poziționarea precisă a lucrării în peisajului domeniului studiat;
* [Ambele] Sunt descrise câteva tehnologii alternative pentru fiecare din tehnologiile folosite în lucrare. Există o argumentare referitoare la alegere.

Criterii pentru calificativul *Bine*:

* [Dezvoltare de produs] Proces iterativ pe baza unor interviuri cu mai mulți clienți, dezvoltare MVP, reevaluare cerințe;
* [Cercetare] analiza literaturii de specialitate din lume, cu poziționarea precisă a lucrării în peisajul actual al domeniului studiat;
* [Ambele] Sunt descrise tehnologii alternative. Sunt analizate cantitativ și calitativ, folosite benchmarkuri și teste efectuate de student. Analiza este rezumată prin tabele și grafice.

## Indicații formatare figuri

Figurile utilizate în document vor fi centrate și numerotate (de exemplu Figura 1).

Orice figură ce nu este realizată de către autorul lucrării va fi în mod obligatoriu citată fie la final (de exemplu Figura 2 este preluată din documentul [1]), fie cel puțin într-o notă de subsol (a se vedea Figura 2). Orice figură ce depășește ca dimensiune 50% dintr-o pagină, va fi mutată la anexe. Toate figurile din cadrul tezei vor fi referite în text. Exemplu: Figura 1 prezintă o schemă de principiu pentru un amplificator inversor cu AO.

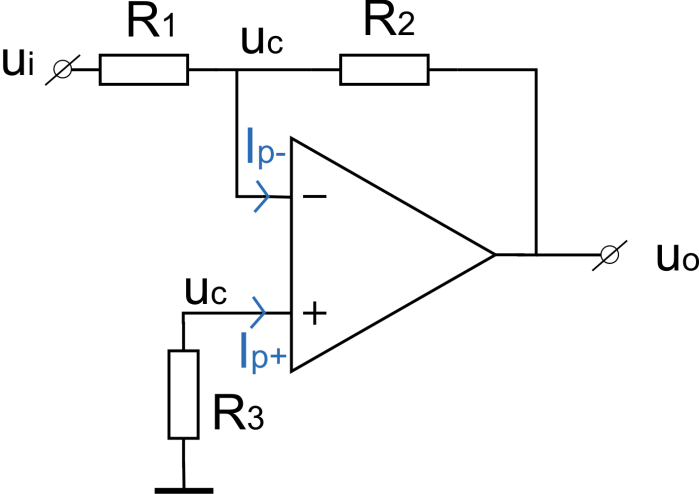


Figura 1 Amplificator inversor

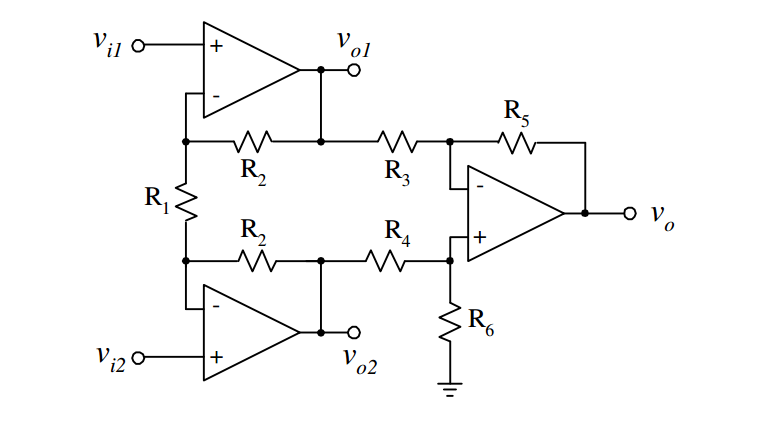


Figura 2 Amplificator de instrumentație cu 3 AO-uri [[1]](#footnote-2)

# Soluția propusă

Capitolul conține o privire de ansamblu a soluției ce rezolvă problema, prin prezentarea structurii / arhitecturii acesteia. În funcție de tipul lucrării acest capitol poate conține diagrame (clase, distribuție, workflow, entitate-relație), demonstrații de corectitudine pentru algoritmii propuși de autor, abordări teoretice (modelare matematică), structura hardware, arhitectura aplicației.

Criterii pentru calificativul *Nesatisfăcător*:

* Descriere în limbaj natural.

Criterii pentru calificativul *Satisfăcător*:

* Descriere + diagrame de baze de date, workflow, clase, algoritmi.

Criterii pentru calificativul *Bine*:

* Descriere + diagrame de baze de date, workflow, clase, algoritmi + descrierea unui proces prin care s-a realizat arhitectura/structura soluției.

## Indicații formatare formule

Formulele matematice utilizate în document vor fi centrate în pagină și numerotate. Se vor utiliza fontul Cambria Math, de dimensiune 11. Pentru a insera o nouă ecuație, utilizați Insert > Quick Parts > AutoText > Ecuație.

Toate formulele prezente în lucrare vor fi referite în text. Exemplu: *Utilizând sistemul de Insert > Bookmark*, respectiv *Reference > Cross-reference* putem cita ecuația (1 respectiv ecuația (2, citările fiind actualizate și în urma unor adăugări/ ștergeri de ecuații, cu *Select All – Update Field*. Pentru mai multe detalii despre utilizarea acestui sistem de referire și formatare puteți consulta:

<https://www.youtube.com/watch?v=9YGTH4WrY_8>.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

# Detalii de implementare

În plus fata de capitolul precedent acesta conține elemente specifice ale rezolvării problemei care au presupus dificultăți deosebite din punct de vedere tehnic. Pot fi incluse configurații, secvențe de cod, pseudo-cod, implementări ale unor algoritmi, analize ale unor date, scripturi de testare. De asemenea, poate fi detaliat modul în care au fost utilizate tehnologiile introduse in capitolul 3.

Criterii pentru calificativul Nesatisfăcător:

* Sunt prezentate pe scurt scheme și pseudo-cod.

Criterii pentru calificativul Satisfăcător:

* Descriere sumara a implementării, prezentarea unor secvențe nerelevante de cod, scheme, etc.

Criterii pentru calificativul Bine:

* Descrierea detaliată a algoritmilor/structurilor utilizați; Prezentarea etapizată a dezvoltării, inclusiv cu dificultăți de implementare întâmpinate, soluții descoperite; (dacă este cazul) demonstrarea corectitudinii algoritmilor utilizați.

## Indicații formatare tabele

Se recomandă utilizarea tabelelor de forma celui de mai jos. Font: Calibri, 9.

Orice tabel prezent în teză va fi referit în text; exemplu: a se vedea Tabel 1.

Tabel 1 – Sumarizare criterii

|  |  |  |
| --- | --- | --- |
| **Calificativ** | **Criteriu** | **Observații** |
| **Nesatisfăcător** | Sunt prezentate pe scurt scheme și pseudo-cod |  |
| **Satisfăcător** | Descriere sumara a implementării, prezentarea unor secvențe nerelevante de cod, scheme, etc. |  |
| **Bine** | Descrierea detaliată a algoritmilor/structurilor utilizați; Prezentarea etapizată a dezvoltării, inclusiv cu dificultăți de implementare întâmpinate, soluții descoperite; (dacă este cazul) demonstrarea corectitudinii algoritmilor utilizați. | Pot fi incluse configurații, secvente de cod, pseudo-cod, implementări ale unor algoritmi, analize ale unor date, scripturi de testare. |

# Studiu de caz / Evaluarea rezultatelor

Acest capitol trebuie să răspundă, în principiu, la **2 întrebări** și să se încheie cu **o discuție** a rezultatelor obținute. Cele doua întrebări la care trebuie sa se răspundă sunt:

1) **Merge corect**? (Conform specificațiilor extrase în capitolul 2);

Evaluarea dacă merge corect se face pe baza cerințelor identificate în capitolele anterioare.

2) Cât de bine merge / cum se compară cu soluțiile existente? (pe baza unor metrici clare).

Evaluarea cât de bine merge trebuie să fie bazată pe procente, timpi, cantitate, numere, **comparativ cu soluțiile prezentate în capitolul 3**. Poate fi vorba de performanță, overhead, resurse consumate, scalabilitate etc.

În realizarea discuției, se vor utiliza tabele cu procente, rezultate numerice și grafice. În mod obișnuit, aici se fac comparații și teste comparative cu alte proiecte similare (dacă există) și se extrag puncte tari și puncte slabe. Se ține cont de avantajele menționate și se demonstrează viabilitatea abordării / aplicației, de dorit prin comparație cu alte abordări (dacă acest lucru este posibil). Cuvântul cheie la evaluare este „metrică“: trebuie să aveți noțiuni măsurabile și cuantificabile. În cadrul procesului de evaluare, explicați datele, tabelele și graficele pe care le prezentați și insistați pe relevanța lor, în următorul stil: „este de preferat ... deoarece …“; explicați cititorului nu doar datele ci și semnificația lor și cum sunt acestea interpretate. Din această interpretare trebuie să rezulte poziționarea proiectului vostru printre alternativele existente, precum și cum poate fi acesta îmbunătățit în continuare.

Criterii pentru calificativul *Nesatisfăcător*:

* Aplicația este testată dar rulează pe calculatorul studentului, nu există posibilități de testare, nu a fost validată cu clienți / utilizatori;
* Nu au fost realizate comparații cu alte sisteme similare.

Criterii pentru calificativul *Satisfăcător*:

* [Dezvoltare de produs] Există teste unitare și de integrare, există o strategie de punere în funcțiune (*deployment*), există validare minimală cu clienții / utilizatorii.
* [Cercetare] Principalele componente și soluția în ansamblu au fost evaluate din punct de vedere al performanței, însă nu sunt folosite seturi de date standard, există unele erori de interpretare a datelor.
* [Ambele] Discuție minimală asupra relevanței rezultatelor prezentate, comparație minimală cu alte sisteme similare.

Criterii pentru calificativul *Bine*:

* [Dezvoltare de produs] Teste unitare și de integrare, instrumente de punere in funcțiune (*deployment*) utilizate și care arată lucru constant de-a lungul semestrului, lucrare validată cu clienții / utilizatorii, produs în producție.
* [Cercetare] Componentele și soluția în ansamblu au fost evaluate din punct de vedere al performanței, folosind seturi de date standard și cu o interpretare corectă a rezultatelor.
* [Ambele] Discuție cu prezentarea calitativă și cantitativă a rezultatelor, precum și a relevanței acestor rezultate printr-o comparație complexă cu alte sisteme similare.

# Concluzii

În acest capitol este sumarizat întreg proiectul, de la obiective, la implementare, si la relevanta rezultatelor obținute. În finalul capitolului poate exista o subsecțiune de „Dezvoltări ulterioare“.  
Criterii pentru calificativul *Nesatisfăcător*:

* Concluziile nu sunt corelate cu conținutul lucrării;

Criterii pentru calificativul *Satisfăcător*:

* Concluziile sunt corelate cu conținutul lucrării, însă nu se oferă o imagine asupra calității și relevantei rezultatelor obținute;

Criterii pentru calificativul *Bine*:

* Concluziile sunt corelate cu conținutul lucrării, și se oferă o imagine precisa asupra relevantei și calității rezultatelor obținute în cadrul proiectului.

# Bibliografie

* Trebuie respectat **un singur standard** de trimiteri bibliografice (citare), **dintre** următoarele alternative:
  + APA (<http://pitt.libguides.com/c.php?g=12108&p=64730>)
  + IEEE ([https://ieee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf](https://ieee-dataport.org/sites/default/files/analysis/27/IEEE Citation Guidelines.pdf))
  + Harvard (<https://libweb.anglia.ac.uk/referencing/harvard.htm>)
  + Cu numerotarea referințelor în ordine alfabetică sau în ordinea apariției în text (de exemplu, stilul cu numere folosit de unele publicații ACM - <https://www.acm.org/publications/authors/reference-formatting>)
* Toate referințele din acest capitol trebuie să fie referite în text. Exemple:
  + [Articol jurnal]: [2];
  + [Articol conferință]: [3];
  + [Carte]: [4]
  + [Weblink]: [5]

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  | * + [Application report] [6] |

1. © http://www.ece.tamu.edu/~spalermo/ecen325/Section%20III.pdf [↑](#footnote-ref-2)