Universitatea
POLITEHNICA
din București

Facultatea
de Automatică și
Calculatoare

Universitatea POLITEHNICA din București
Facultatea de Automatica și Calculatoare
Departamentul Calculatoare

**Teza de abilitare**
**în domeniul**
**Calculatoare și Tehnologia Informației**

# Metode de extragere a informațiilor din volume mari de date

Prezentată de Alexandru Boicea
Septembrie 2021, București, România

## Rezumat

Odata cu dezvoltarea Internetului volumul de date care poate fi accesat de publicul larg a crescut exponențial. Pe lânga dezvoltarea masivă a infrastructurii, in ultimii ani s-a pus accent și pe cercetare în domeniul prelucrării datelor pentru identificarea unor metode moderne de prelucrare a volumelor mari de date. Apariția unor servere din ce in ce mai performante nu a rezolvat în totalitate procesarea volumelor masive de date care necesită resurse hardware importante. În acest context, s-au dezvoltat in paralel noi tehnici de extragere a informațiilor din seturi foarte mari de date prin intermediul unor algoritmi specifici.

Această teză prezintă rezultatele cercetării mele în cadrul colectivului de baze date din facultate privind implementarea, evaluarea și îmbunătățirea metodelor de extragere a informațiilor din seturi masive de date. Algoritmii au fost testați pe diverse sisteme de gestiune a bazelor de date relaționale și NoSQL.

O altă direcție de cercetare a fost evaluarea performanțelor bazelor de date pentru a veni în sprijinul dezvoltatorilor de aplicații software în vederea alegerii sistemului de baze de date cel mai potrivit pentru tipul de aplicație ce urmează a fi implementat.

Eșantionarea este o metodă folosită des in procesări masive de date deoarece prelucrarea întregului set de date necesită costuri mari de timp și resurse hardware. Un capitol din teză este dedicat metodelor de eșantionare, algoritmilor de eșantionare si evaluarea performanțelor acestor algoritmi testați pe volume mari de date.

Interogarea după cuvinte cheie a documentelor stocate în pagini web nu este totdeauna mulțumitoare cand folosim motoare de căutare sau rețele de socializare. În cadrul tezei se prezintă rezultatele experimentale care aduc o îmbunătățire a tehnicii *Top-k* folosită pentru interogarea seturilor mari de documente text.

Pentru creșterea performanțelor de interogare a fost creat și implementat un *Inverted Index*, o structură de index de tip cheie-valoare care stochează locația unui cuvânt într-un fișier, document sau înregistrare din baza de date.

O altă metodă de interogare se bazează pe tehnica *autocomplete* utilizată in domeniul *Information Retrieval* pentru regăsirea cuvintelor, sau secvențelor de cuvinte, în fișiere text de dimensiuni foarte mari. Este prezentată o arhitectură bazată pe această tehnică ce conține module pentru crearea bazei de date a cuvintelor(*tokenization*), identificarea

cuvintelor pe baza prefixului (*prefix tree*) și identificarea secvențelor de cuvinte(*collocation*). Analiza complexității a fost realizată pentru algoritmii *Mean-variance* si *Chi-squared*. Pentru identificarea subiectului tratat intr-un document text, sau colecție de documente, este prezentat un model *text-mining* folosit in *Topic Modeling*.

Extragerea informațiilor din imagini se face cu o tehnica *data-mining* folosind algoritmul de clusterizare *K-Means*. Pentru extragerea informațiilor din pagini web s-a folosit un *crowler* pentru descărcarea documentelor pe care se face analiza de text.

Ultimele capitole prezintă metode de comparare a bazelor de date relaționale și NoSQL pe baza criteriilor de performanță. Rezultatele cercetării prezentate în această teză au fost publicate la conferințe și in jurnale internaționale de top.