SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4573
Obrada podataka tehnologijom Apache Spark

Martin Matak

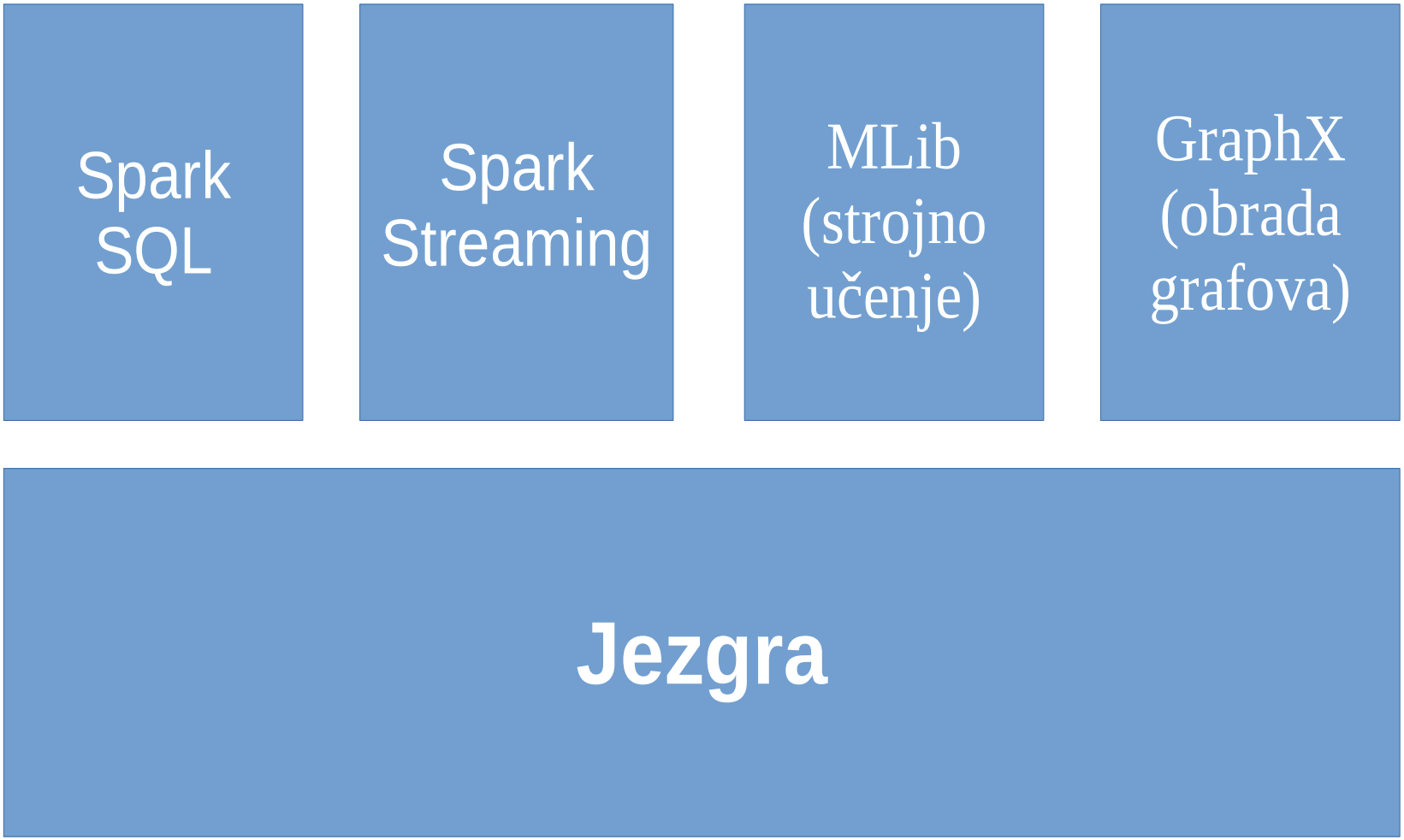
Zagreb, srpanj 2016.

Sadržaj

Uvod

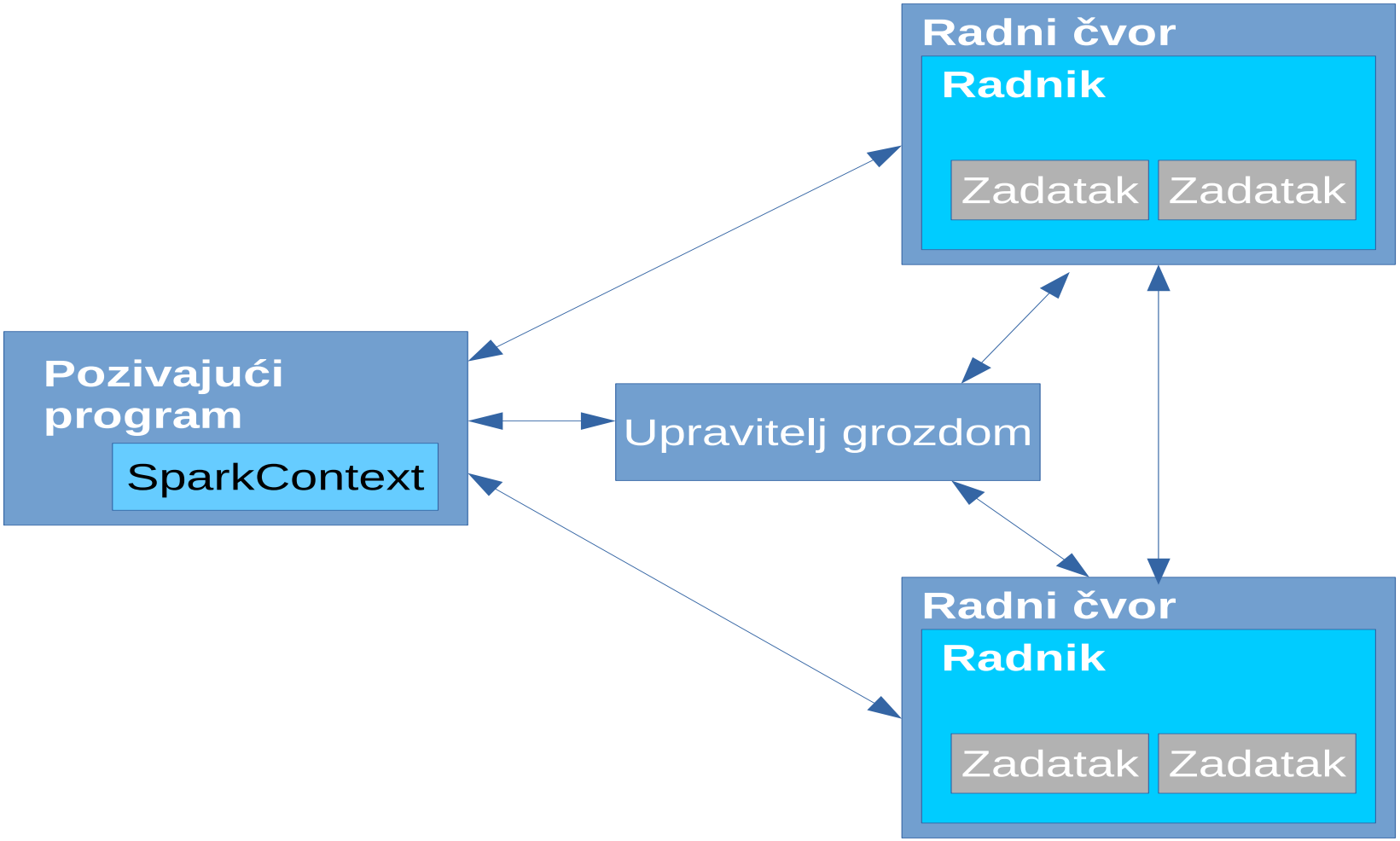

1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW LinkedIn ACCOUNTS
IN 60 SECONDS...
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WordPress DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com

Osnovni gradivni elementi



Sadržaj

Osnovni elementi aplikacije



Sadržaj

Otporni raspodijeljeni skup podataka

Resilient distributed dataset - RDD

- Nepromjenjiva kolekcija podataka

Otporni raspodijeljeni skup podataka

Resilient distributed dataset - RDD

- Nepromjenjiva kolekcija podataka

Transformacije

Operacije koje iz jednog skupa kreiraju drugi, novi skup podataka.
Lijena evaluacija - pokreću ih akcije.

Otporni raspodijeljeni skup podataka

Resilient distributed dataset - RDD

- Nepromjenjiva kolekcija podataka

Transformacije

Operacije koje iz jednog skupa kreiraju drugi, novi skup podataka.
Lijena evaluacija - pokreću ih akcije.

Akcije

Dohvat jednog ili više elemenata iz nekog skupa.

Otporni raspodijeljeni skup podataka

Resilient distributed dataset - RDD

- Nepromjenjiva kolekcija podataka

Transformacije

Operacije koje iz jednog skupa kreiraju drugi, novi skup podataka.
Lijena evaluacija - pokreću ih akcije.

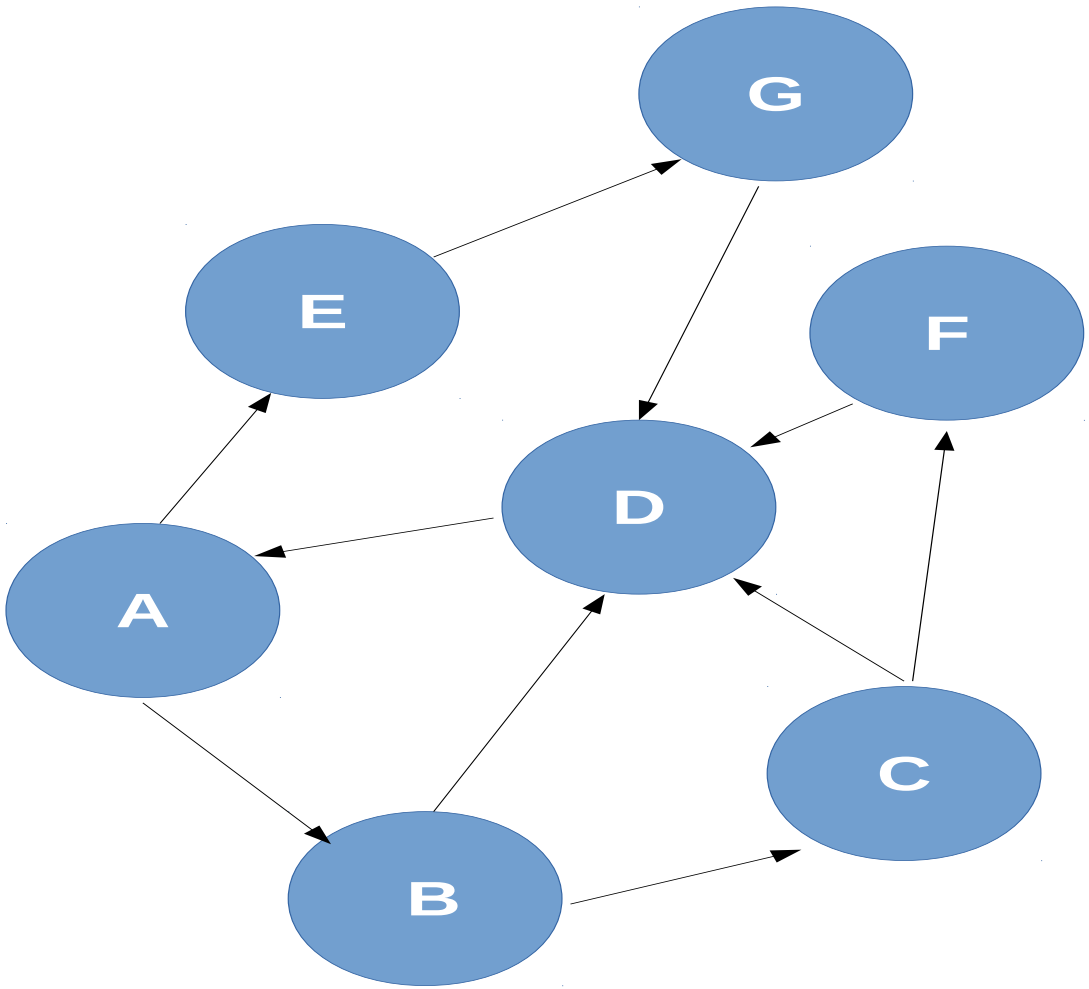
Akcije

Dohvat jednog ili više elemenata iz nekog skupa.

- Stanje u memoriji?

Sadržaj

Koja stranica je najvažnija?

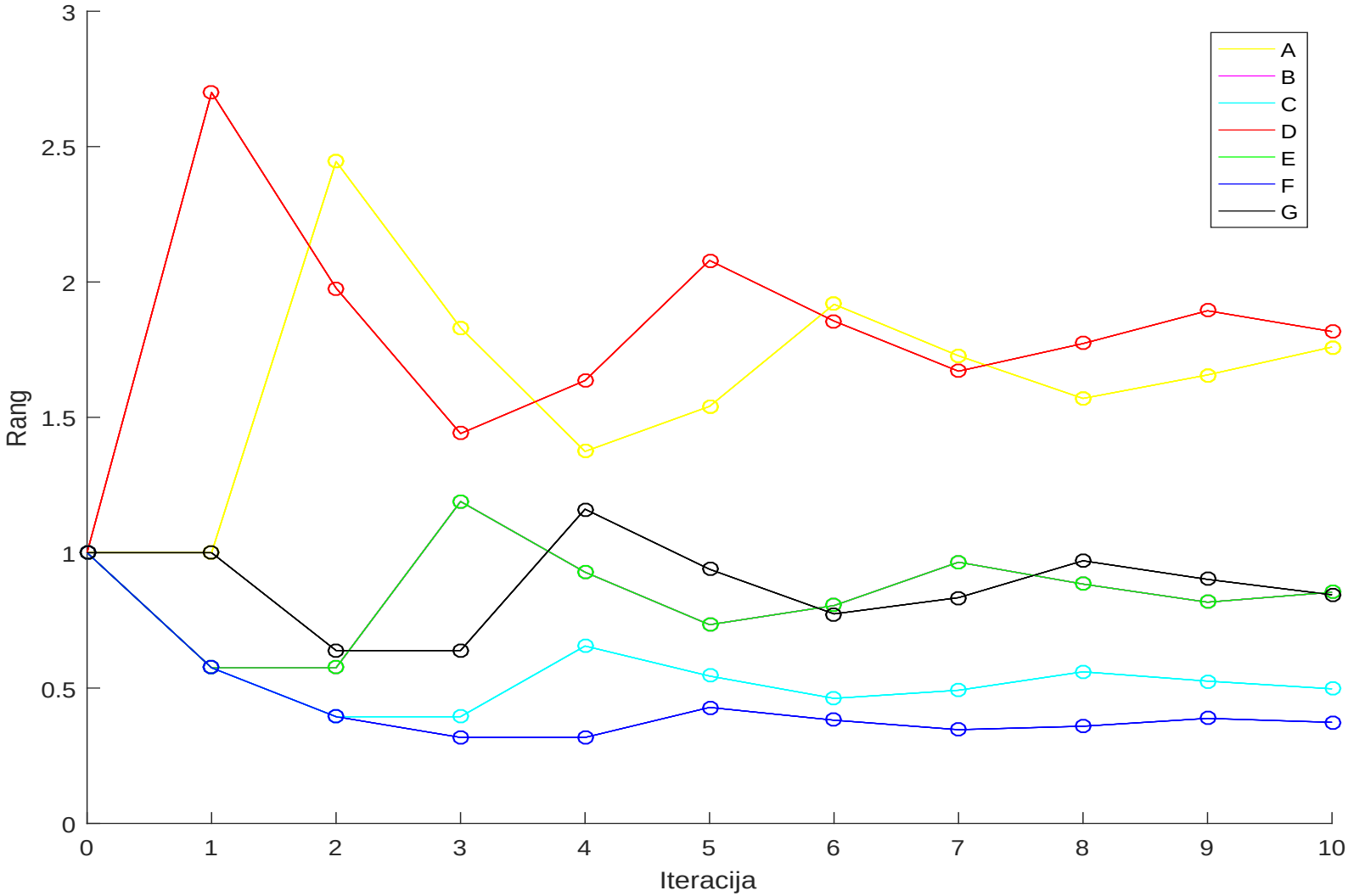


Algoritam PangeRank

Algoritam 1 Algoritam Pagerank

1: Početni rang svake stranice postavi se na 1.0
2: **for** $i = 1$ to P **do**
3: Svaka stranica n šalje svojim susjedima doprinos
`rang(`n`) / brojSusjeda(`n`)`
4: Postavi ukupni rang stranice prema formuli:
$0.15 + 0.85 *$ *ukupan primljeni doprinos*
5: **end for**

Rezultati



Sadržaj

Uređeni parovi te spremanje i čitanje podataka

- Podatci u obliku ključ - vrijednost

Uređeni parovi te spremanje i čitanje podataka

- Podatci u obliku ključ - vrijednost
- Posebne **transformacije** i **akcije**

Uređeni parovi te spremanje i čitanje podataka

- Podatci u obliku ključ - vrijednost
- Posebne **transformacije** i **akcije**
- Mogućnosti spremanja i čitanja: baze podataka, tekstualne datoteke (JSON, CSV, TSV) ...

Sadržaj

Dijeljene varijable: odašiljatelji i akumulatori

- Problem: Svaki zadatak na grozdu ima svoju kopiju varijabli

Dijeljene varijable: odašiljatelji i akumulatori

- Problem: Svaki zadatak na grozdu ima svoju kopiju varijabli

Odašiljatelj

Nepromjenjiva varijabla koja zauzima malo memorije čiju vrijednost je moguće dohvatiti na cijelom grozdu.

Dijeljene varijable: odašiljatelji i akumulatori

- Problem: Svaki zadatak na grozdu ima svoju kopiju varijabli

Odašiljatelj

Nepromjenjiva varijabla koja zauzima malo memorije čiju vrijednost je moguće dohvatiti na cijelom grozdu.

Akumulator

Globalna varijabla čiju vrijednost je moguće mijenjati iz cijelog grozda.

U radu implementirano

- Svi primjeri napisani su u Javi

U radu implementirano

- Svi primjeri napisani su u Javi
- Algoritam brojanja riječi - bez i s korištenjem lambda izraza

U radu implementirano

- Svi primjeri napisani su u Javi
- Algoritam brojanja riječi - bez i s korištenjem lambda izraza
- Primjer korištenja transformacija i akcija

U radu implementirano

- Svi primjeri napisani su u Javi
- Algoritam brojanja riječi - bez i s korištenjem lambda izraza
- Primjer korištenja transformacija i akcija
- Algoritam PageRank

U radu implementirano

- Svi primjeri napisani su u Javi
- Algoritam brojanja riječi - bez i s korištenjem lambda izraza
- Primjer korištenja transformacija i akcija
- Algoritam PageRank
- Korištenje akumulatora za brojanje grešaka u *log* datotekama

U radu implementirano

- Svi primjeri napisani su u Javi
- Algoritam brojanja riječi - bez i s korištenjem lambda izraza
- Primjer korištenja transformacija i akcija
- Algoritam PageRank
- Korištenje akumulatora za brojanje grešaka u *log* datotekama
- Korištenje odašiljatelja za dohvaćanje imena naselja pomoću identifikatora

Zaključak

- Tehnologija za obradu velike količine podataka.
- Algoritam PageRank.
- Dijeljene varijable.

- Još bi trebalo:
  - Postaviti i isprobati Apache Spark na grozdu.
  - Razraditi svaku od komponenata - SparkSQL, MLib, Spark Streaming i GraphX.