

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I
RAČUNARSTVA

ZAVRŠNI RAD br. 4573
Obrada podataka tehnologijom Apache Spark

Martin Matak

Zagreb, srpanj 2016.

Sadržaj

Uvod

Osnovni gradivni elementi

Prvi programi

- Postavljanje temelja

- Otporni rasopdijeljeni skup podataka

Napredno programiranje

- Algoritam PageRank

- Skupovi podataka kao uređeni parovi

- Dijeljene varijable

U radu implementirano

Zaključak



Osnovni gradivni elementi

Spark
SQL

Spark
Streaming

MLib
(strojno
učenje)

GraphX
(obrada
grafova)

Jezgra

Sadržaj

Uvod

Osnovni gradivni elementi

Prvi programi

Postavljanje temelja

Otporni rasopdijeljeni skup podataka

Napredno programiranje

Algoritam PageRank

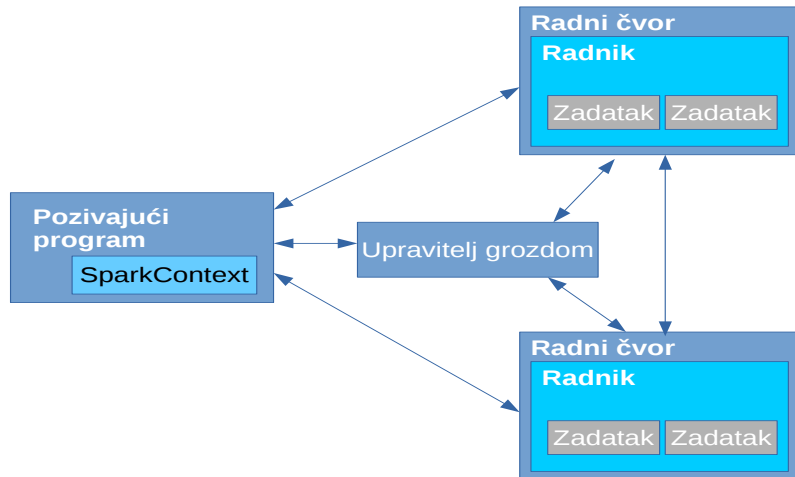
Skupovi podataka kao uređeni parovi

Dijeljene varijable

U radu implementirano

Zaključak

Osnovni elementi aplikacije



Sadržaj

Uvod

Osnovni gradivni elementi

Prvi programi

Postavljanje temelja

Otporni rasopdijeljeni skup podataka

Napredno programiranje

Algoritam PageRank

Skupovi podataka kao uređeni parovi

Dijeljene varijable

U radu implementirano

Zaključak

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

Uvod

Osnovni gradivni elementi

Prvi programi

Postavljanje temelja

Otporni rasopdijeljeni skup podataka

Napredno programiranje

Algoritam PageRank

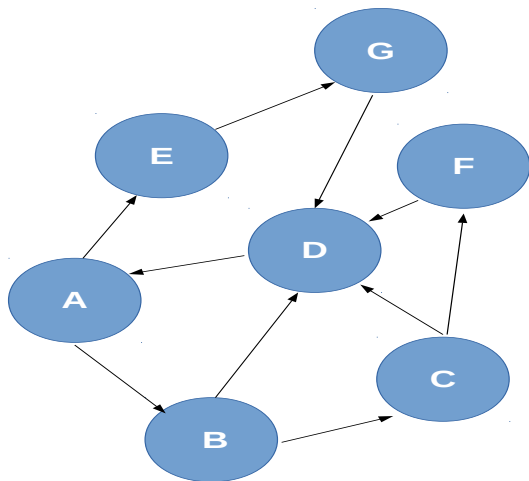
Skupovi podataka kao uređeni parovi

Dijeljene varijable

U radu implementirano

Zaključak

Koja stranica je najvažnija?

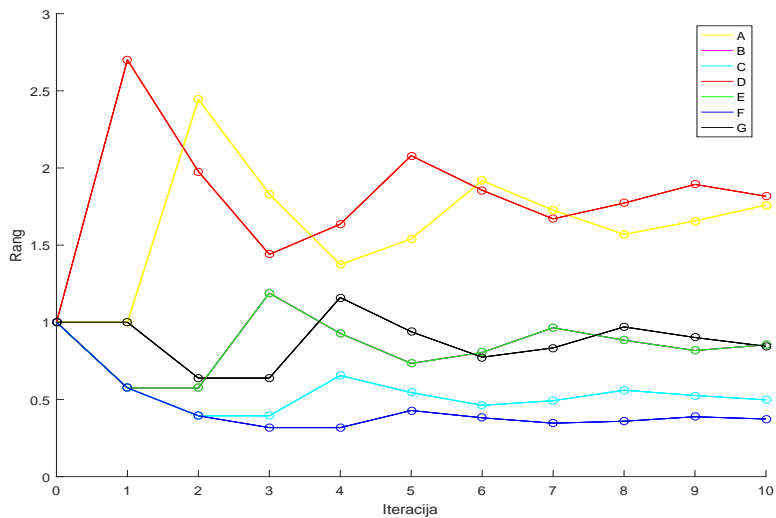


Algoritam PangeRank

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
 $\text{rang}(n) / \text{brojSusjeda}(n)$
 - 4: Postavi ukupni rang stranice prema formuli:
 $0.15 + 0.85 * \text{ukupan primljeni doprinos}$
 - 5: **end for**
-

Rezultati



Sadržaj

Uvod

Osnovni gradivni elementi

Prvi programi

Postavljanje temelja

Otporni rasopdijeljeni skup podataka

Napredno programiranje

Algoritam PageRank

Skupovi podataka kao uređeni parovi

Dijeljene varijable

U radu implementirano

Zaključak

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

Uvod

Osnovni gradivni elementi

Prvi programi

Postavljanje temelja

Otporni rasopdijeljeni skup podataka

Napredno programiranje

Algoritam PageRank

Skupovi podataka kao uređeni parovi

Dijeljene varijable

U radu implementirano

Zaključak

Dijeljene varijable: odašiljačelji 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

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.