SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1337

Obrada podataka tehnologijom Apache Spark

Martin Matak

Zagreb, svibanj 2016.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`. Na ovoj stranici se

nalazi izvornik.

Zahvala - TODO . . . :)

CONTENTS

1. Uvod

1.1. Motivacija

Pametni mobilni uređaji postaju neizostavan dodatak svakog modernog čovjeka. Većina pametnih mobilnih uređaja u sebi sadrži sljedeće senzore:

akcelerometar - elektromehanička komponenta koja mjeri sile ubrzanja;

barometar - mehanički senzor za mjerenje atmosferskog pritiska (na trenutnoj lokaciji uređaja);

senzor svjetlosti - mjeri intenzitet, tj. jačinu svjetlosti i uglavnom se nalazi s prednje strane uređaja, iznad ekrana;

senzor blizine - u stanju prepoznati situacije kada mu neki objekt stoji u blizini - ovo omogućava automatske pozive prilikom primicanja telefona licu uz zaključavanje telefona da bi onemogućili slučajno prekidanje poziva uhom ili slično;

senzor gestikulacije - prepoznaje kretnje ruke tako što detektira infracrvene zrake koje se reflektiraju - omogućuje nam djelomično upravljanje telefonom bez dodirivanja ekrana;

žiroskop - uređaj koji se koristi za navigaciju i merenje kutne brzine;

geomagnetski senzor - mjeri okolno geomagnetsko polje za sve tri fizičke osi i u biti služi kao kompas na mobilnim uređajima i

Hall Sensor - magnetski senzor zadužen za prepoznavanje je li maska telefona zatvorena ili otvorena.

Pretpostavimo da je mobilni uređaj spojen na internet i da svakih nekoliko sekundi pošalje vrijednost koju u tom trenutku mjeri pojedini senzor. U samo jednom danu može se skupiti dosta podataka. A što kada to ne bi radili za jedan uređaj nego za sve izdane uređaje nekog modela? Količina podataka bi jako brzo narasla.

Kako količina podataka postaje sve veća, dolazimo do pojma *Velika količina podataka* (engl. *Big Data*). U današnje vrijeme imamo više podataka u digitalnom obliku nego što smo ikada imali. Jedan od zanimljivijih izazova je kako ih efektivno obraditi i zaključiti nešto iz toga. Kako od te velike količine podataka doći do nekih pametnih zaključaka iz kojih ćemo nešto novo naučiti.

Apache Spark je otvorena (engl. *open source*) tehnologija koja omogućava pisanje programa za obradu podataka u tri programskih jezika: *Java*, *Python* i *Scala*; a nudi i mogućnost interaktivnog rada.
U okviru ovog rada biti će proučene mogućnosti ove tehnologije, razrađeno nekoliko konkretnih primjera obrade podataka te ostvarena programska rješenja koja obavljaju tu obradu koristeći *Apache Spark*.

Svi primjeri će biti napisani u programskom jeziku *Java*.

1.2. Instalacija i izvorni kodovi

1.2.1. Instalacija

Apache Spark je pisan u programskom jeziku *Scala* i izvršava se na *Javinom virtualnom stroju* (engl. *Java Virtual Machine*) (JVM). Instalacija na osobno računalo je prilično jednostavna, a u ovdje će biti prikazana instalacija na operacijskom sustavu *Ubuntu 15.10*.
Za početak je potrebno imati instaliranu Javu, a je li Java instalirana na računalu se može provjeriti tako što se u naredbenom retku unese sljedeća naredba:
`java -version`. Kao rezultat bi trebali dobiti trenutno instaliranu verziju Jave. Ukoliko Java nije instalirana, potrebno ju je najprije instalirati. Taj dio neće biti objašnjen ovdje.

Jednom kada imamo instaliranu Javu, sve što treba napraviti je otići na službene stranice: `https://spark.apache.org/downloads.html`, odabrati najnoviju verziju (Za vrijeme pisanja ovog rada to je verzija *1.6.1 (Mar 09 2016*), izabrati odgovarajući paket te pokrenuti dohvaćanje odgovarajuće *.tgz* arhive. Najjednostavnije je odabrati neki *pre-built* paket, primjerice *Pre-built for Hadoop 2.6 and later* te će daljnji koraci instalacije biti napisani pod pretpostavkom da je korisnik dohvatio tu verziiju paketa. Ukoliko korisnik želi, moguće je instalirati i *Source code* varijantu paketa, ali taj postupak instalacije ovdje nije opisan.

Nakon što smo dohvatili odgovarajuću arhivu, potrebno ju je raspakirati.
Raspakiranje arhive moguće je napraviti preko naredbe:

```
$ tar -xvf spark-1.6.1-bin-hadoop2.6.tgz
```

Nakon toga, dobra je praksa premjestiti instalaciju u neki prikladniji direktorij. Tako nešto može se napraviti na sljedeći način:

```
$ mv Downloads/spark-1.6.1-bin-hadoop2.6 faks/ZavrsniRad/spark/
```

Službeno, sada je instalacija gotova. Idemo u sljedećem odlomku malo detaljnije pogledati što smo to zapravo dobili instalacijom. Koje datoteke se sada nalaze na našem računalu, a nije ih bilo ranije. Također, biti će objašnjena i uloga nekih datoteka i direktorija.

1.2.2. Izvorni kodovi

Ako izlistamo što nam se trenutno nalazi u novonastalom direktoriju, dobiti ćemo sljedeći ispis:

```
mmatak@martins-beast:~/faks/ZavrsniRad/spark$ ls -l
total 1408
drwxr-xr-x 2 mmatak mmatak    4096 Feb 27 06:02 bin
-rw-r--r-- 1 mmatak mmatak 1343562 Feb 27 06:02 CHANGES.txt
drwxr-xr-x 2 mmatak mmatak    4096 Feb 27 06:02 conf
drwxr-xr-x 3 mmatak mmatak    4096 Feb 27 06:02 data
drwxr-xr-x 3 mmatak mmatak    4096 Feb 27 06:02 ec2
drwxr-xr-x 3 mmatak mmatak    4096 Feb 27 06:02 examples
drwxr-xr-x 2 mmatak mmatak    4096 Feb 27 06:02 lib
-rw-r--r-- 1 mmatak mmatak   17352 Feb 27 06:02 LICENSE
drwxr-xr-x 2 mmatak mmatak    4096 Feb 27 06:02 licenses
-rw-r--r-- 1 mmatak mmatak   23529 Feb 27 06:02 NOTICE
drwxr-xr-x 6 mmatak mmatak    4096 Feb 27 06:02 python
drwxr-xr-x 3 mmatak mmatak    4096 Feb 27 06:02 R
-rw-r--r-- 1 mmatak mmatak    3359 Feb 27 06:02 README.md
-rw-r--r-- 1 mmatak mmatak     120 Feb 27 06:02 RELEASE
drwxr-xr-x 2 mmatak mmatak    4096 Feb 27 06:02 sbin
```

README.md Sadrži kratke instrukcije za upoznavanje sa Spark-om.

bin Sadrži izvršive datoteke koje se koriste za interaktivni rad sa Spark-om. Zanimljivo je spomenuti da postoji interaktivna ljuska *Spark shell* za programske jezike *Python* i *Scala*. Datoteke u ovom direktorij služe upravo za to. Budući

da je ovaj rad ograničen isključivo na programski jezik *Java*, ovaj dio neće biti detaljnije obrađen.

core, streaming, python, ... Sadrži glavne komponente projekta *Apache Spark*

examples Sastoji se od nekoliko jednostavnih primjera koje pomažu korisniku da se uhoda i što bezbolnije nauči koristiti odgovarajući API.

1.3. Kratki pregled

U ovom poglavlju je bila ideja da čitatelj dobije motivaciju i želju za upoznavanjem s tehnologijom *Apache Spark*. Detaljno je opisan postupak instalacije koji bi čitatelju trebao biti sasvim dovoljan za samostalnu instalaciju. Također, dan je pregled nekih najosnovnijih datoteka i direktorija koje dolaze s instalacijom. Zainteresiranog čitatelja se ohrabruje da samostalno prouči kako se koristi *Spark shell*.

2. Zaključak

Zaključak.

Obrada podataka tehnologijom Apache Spark

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.