Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Систем за обраду података Apache Spark

(материјали за предавања)

Садржај

Систем Apache Spark

Систем *Apache Spark* (*Spark*)

- систем за дистрибуирану обраду података
  - „обједињени аналитички енџин”
  - превасходно намењен раду над великим количинама података (енгл. *big data*) и захтевним рачунским обрадама (енгл. *big compute*)
  - могућност рада над подацима у примарној меморији
    - могућност и коришћења локалних дискова
  - слободно доступан

Систем *Apache Spark* (*Spark*)

- настанак
  - истраживачки пројекат на Универзитету Калифорније у Берклију (Калифорнија, САД) у оквиру лабораторије *AMPLab*, 2006
  - пројекат отвореног кôда, 2010
  - под патронатом корпорације *Apache Software Foundation*, 2013
- верзије
  - *Spark 4.0.0* (мај 2025)
  - *Spark 3.5.1* (фебруар 2024)
  - *Spark 3.4.0* (април 2023)
  - *Spark 3.2.1* (јануар 2022)
  - *Spark 3.1.1* (март 2021)
  - *Spark 3.0.2* (фебруар 2021)
  - *Spark 2.4.7* (септембар 2020)
  - …

Систем Apache Spark

Систем *Apache Spark* (*Spark*)

- уграђене библиотеке
  - *Spark SQL*
    - рад над структурираним подацима
  - *Spark Structured Streaming*
    - рад над токовима података
  - *MLlib*
    - машинско учење
  - *GraphX*
    - рад над графовима

Систем Apache Spark

Систем *Apache Spark* (*Spark*)

- подршка за језике
  - *Java*
  - *Python*
  - *R*
  - *Scala*
  - *SQL*

Успостављање сопственог кластера са системом *Spark*

- учесници
  - управљач кластером
    - иницијализује систем *Spark* по сваком чвору и региструје чворове
    - могуће опције
      - *Standalone* (доступан уз систем *Spark*)
      - *Hadoop YARN*
      - *Kubernetes*
  - радни чворови (извршиоци)
    - извршавају обраду над партицијама
    - прослеђују међурезултате обраде
  - водећи чвор
    - прослеђује задатке радним чворовима
      - оркестрација преко управљача кластером
    - агрегира резултате

Систем Apache Spark

Успостављање сопственог кластера са системом *Spark*

- отварање конекције према истуреном (ивичном) чвору кластера
  - путем *SSH*
  - путем веб прегледача

Систем Apache Spark

Обрада унутар кластера са системом *Spark*

- конфигурисање
  - подешавање потребних ресурса
    - број језгара, количина меморије, број извршилаца
- партиционисање
  - подела података по чворовима
    - партиција као један подскуп података
- извршавање
  - обрада над партицијама
- мешање
  - прераспоређивање података
- кеширање
  - очување података у меморији
- серијализација
  - трансформација података ради њиховог слања

Званичне препоруке у погледу хардвера 1/2 (*Spark 4.0.0*)

- складишни систем
  - у случају *HDFS*
    - постављање система *Spark* на чворовима где и складишни систем
    - евентуално на другим чворовима али у оквиру исте локалне мреже где и складишни систем
  - у случају *HBase*
    - пожељно извршање на другим чворовима у односу на складишни систем
- локални дискови
  - од 4 до 8 дискова по чвору
    - без коришћења повезивања типа *RAID*
- радна меморија
  - од 8 GiB меморије па навише по чвору
    - највише 75% меморије одвојити за систем *Spark*

Званичне препоруке у погледу хардвера 2/2 (*Spark 4.0.0*)

- микропроцесори
  - барем од 8 до 16 језгара по чвору
  - потенцијално уско грло у пракси при раду над подацима у примарној меморији
- мрежа
  - 10-гигабитна мрежа или бржа
  - потенцијално уско грло у пракси при раду над подацима у примарној меморији

Садржај

Коришћење система Apache Spark кроз језик R

Повезивање са системом *Apache Spark* из језика *R*

- главни пакети за рад са системом *Apache Spark* у језику *R*
  - пакет *sparkr*
    - https://spark.apache.org/docs/latest/sparkr.html
      превазиђена опција
  - пакет *sparklyr*
    - https://cran.r-project.org/web/packages/sparklyr/index.html

Коришћење система Apache Spark кроз језик R

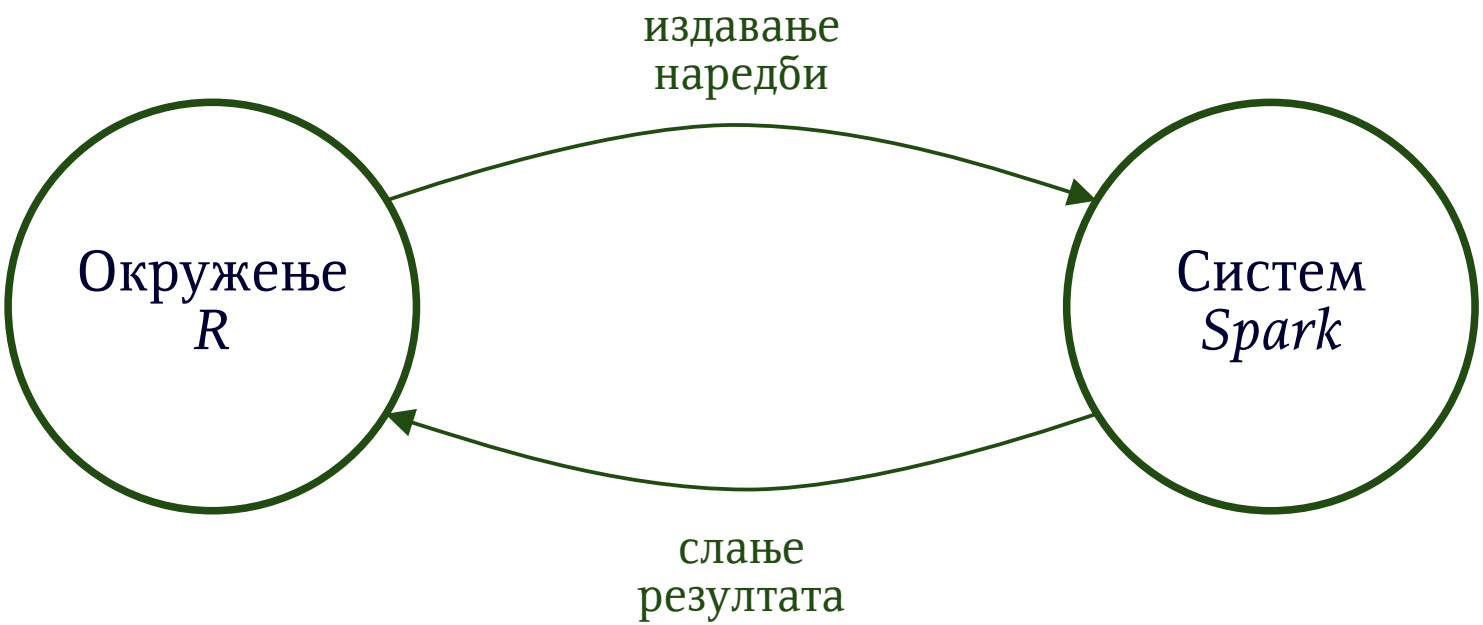
Пакет *sparklyr*

- аутори
  - Хавијер Лураски и сарадници
- верзије
  - верзија 1.9.0 (2025)
  - ...
  - верзија 1.8.6 (2024)
  - ...
  - верзија 1.8.1 (2023)
  - ...
  - верзија 1.7.5 (2022)
  - ...
  - верзија 1.6.2 (2021)
  - ...
  - верзија 0.4 (2016)

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- већина функција користи позиве *Spark API*
- основни приступ коришћењу система *Spark*
  - „*push compute, collect results*”
    - у *R* окружењу се издаје наредба за обраду у оквиру система *Spark*
      резултати обраде се прикупљају за *R* окружење, где се даље могу користити



Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- повезивање са системом *Spark*
  - функција **`spark_connect()`**
    - успостављање конекције према систему *Spark*
  - функција **`spark_disconnect()`**
    - прекидање конекције према систему *Spark*
  - класа **`spark_connection-class`**
    - опис конекције према систему *Spark*
    - успостављена конекција се користи приликом издавања наредби за извршавање у оквиру система *Spark*
  - …

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- надзор система *Spark*
  - функција **`spark_web()`**
    - приступање веб корисничком интерфејсу за систем *Spark*
  - функција **`spark_log()`**
    - очитавање најновијих уписа у лог система *Spark*

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- основни рад с подацима у систему *Spark*
  - често подржани протоколи при коришћењу датотека
    - **`file://`**
    - **`hdfs://`**
    - **`s3a://`**

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

основни рад са скуповима података у систему *Spark*

функција **`copy_to()`**

копирање скупа података из *R* окружења у скуп података у систему *Spark*

функција **`spark_read_csv()`**

учитавање података из *CSV* датотеке у скуп података у систему *Spark*

функција **`spark_read_jdbc()`**

учитавање података помоћу *JDBC* конекције у скуп података у систему *Spark*

функција **`spark_read_json()`**

учитавање података из *JSON* датотеке у скуп података у систему *Spark*

…

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- основни рад са скуповима података у систему *Spark*
  - функција **`spark_write_csv()`**
    - снимање података у *CSV* датотеку из скупа података у систему *Spark*
  - функција **`spark_write_jdbc()`**
    - снимање података помоћу *JDBC* конекције из скупа података у систему *Spark*
  - функција **`spark_write_json()`**
    - снимање података у *JSON* датотеку из скупа података у систему *Spark*
  - …

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- основни рад с токовима података у систему *Spark*
  - функција **`stream_read_csv()`**
    - учитавање података из *CSV* тока у ток података у систему *Spark*
  - функција **`stream_read_json()`**
    - учитавање података из *JSON* тока у ток података у систему *Spark*
  - функција **`stream_read_kafka()`**
    - учитавање података из *Kafka* тока у ток података у систему *Spark*
  - …

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- основни рад с токовима података у систему *Spark*
  - функција **`stream_write_csv()`**
    - снимање у *CSV* ток података из тока података у систему *Spark*
  - функција **`stream_write_json()`**
    - снимање у *JSON* ток података из тока података у систему *Spark*
  - функција **`stream_write_kafka()`**
    - снимање у *Kafka* ток података из тока података у систему *Spark*
  - …

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

основни рад са скуповима података у систему *Spark*

функција **`sdf_bind_rows()`**

спајање скупова података по редовима (вертикално)

функција **`sdf_bind_cols()`**

спајање скупова података по колонама (хоризонтално)

функција **`sdf_dim()`**

утврђивање димензија скупа података

функција **`sdf_nrow()`**

утврђивање броја редова у скупу података

функција **`sdf_ncol()`**

утврђивање броја колона у скупу података

…

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- основни рад са скуповима података у систему *Spark*
  - функција **`sdf_collect()`**
    - прикупљање података из система *Spark* за *R* окружење
  - функција **`sdf_random_split()`**
    - подела скупа података
  - функција **`sdf_sample()`**
    - формирање случајног узорка за скуп података
  - функција **`sdf_sort()`**
    - сортирање скупа података у систему *Spark*
  - …

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- основни рад са скуповима података у систему *Spark*
  - функција **`ft_dct()`**
    - примена дискретне косинусне трансформације
  - функција **`ft_imputer()`**
    - попуњавање недостајућих вредности
  - функција **`ft_sql_transformer()`**
    - трансформисање на основу *SQL* наредбе типа *Select*
  - функција **`ft_standard_scaler()`**
    - примена стандардизације
  - ...

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- могућности за напредну обраду података
  - примена *SQL*
    - пакет *DBI*
  - примена граматике манипулације подацима
    - пакет *dplyr*
      - наредбе бивају трансформисане у *Spark SQL* исказе
  - визуализација података
    - пакет *dbplot*
      - потребна припрема података се извршава у оквиру система *Spark*
      - исцртавање се изводи помоћу пакета *ggplot2*

Пакет *sparklyr*

- могућности за машинско учење
  - класификација
    - функција **`ml_naive_bayes()`**
      наивни Бајесов класификатор
    - функција **`ml_logistic_regression()`**
      логистичка регресија
    - функција **`ml_linear_svc()`**
      линеарна машина потпорних вектора
    - функција **`ml_multilayer_perceptron_classifier`**
      вишеслојна неуронска мрежа
    - функција **`ml_decision_tree_classifier()`**
      стабло одлучивања
    - функција **`ml_gbt_classifier()`**
      метод градијентно појачаних стабала
    - функција **`ml_random_forest_classifier()`**
      метод случајне шуме

Коришћење система Apache Spark кроз језик R

Пакет *sparklyr*

- могућности за машинско учење
  - кластеризација
    - метод средина̂
      - функција **`ml_kmeans()`**
      - функција **`ml_bisecting_kmeans()`**

Пакет *sparklyr*

- могућности за машинско учење
  - опште радње
    - функција **`ml_predict()`**
      генерисање предикција
    - функција **`ml_evaluate()`**
      испитивање перформанси за модел машинског учења
    - функција **`ml_summary()`**
      издвајање метрике из описа модела машинског учења
    - ...

Садржај

Извори и литература

Основни извори и литература

- Luraschi J, Kuo K, Ruiz E. Mastering Spark with R: The complete guide to large-scale analysis and modeling. O'Reilly; 2019. Internet: https://therinspark.com/
- Damji JS, Wenig B, Das T, Lee D. Learning Spark: Lightning-fast data analytics. 2nd edition. O'Reilly; 2020.
- Apache Spark. Apache Spark™ - Unified engine for large-scale data analytics. Internet: https://spark.apache.org/
- Apache Spark. Downloads | Apache Spark. Internet: https://spark.apache.org/downloads.html
- Apache Spark. News | Apache Spark. Internet: https://spark.apache.org/news/index.html
- Apache Spark. Overview - Spark 4.0.0 documentation. Internet: https://spark.apache.org/docs/latest/
- Apache Spark. Hardware provisioning - Spark 4.0.0 documentation. Internet: https://spark.apache.org/docs/latest/hardware-provisioning.html

Извори и литература

Основни извори и литература

- GitHub. spark/R at master · apache/spark · GitHub. Internet: https://github.com/apache/spark/tree/master/R
- R Project. CRAN: Package sparklyr. Internet: https://cran.r-project.org/web/packages/sparklyr/index.html

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Систем за обраду података Apache Spark

(материјали за предавања)