Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Платформа за анализу података и машинско учење h2o

(материјали за предавања)

Садржај

Платформа *h2o*

- софтверска платформа намењена анализи података и машинском учењу
  - усмерена на рад с великим количинама података
  - усмерена на ефикасно коришћење ресурса
    - паралелизоване имплементације алгоритама машинског учења
- аутори
  - компанија *H2O.ai* („Силицијумска долина”, САД)
- верзије
  - 3.46.0.7 (2025)
  - …
  - *rel-zorn* – 3.36.0.3 (2022)
  - …
  - *rel-shannon* – 3.0.0.4 (2015)
  - …
  - *rel-selberg* – 0.2.0.1 (2015)

Платформа h2o

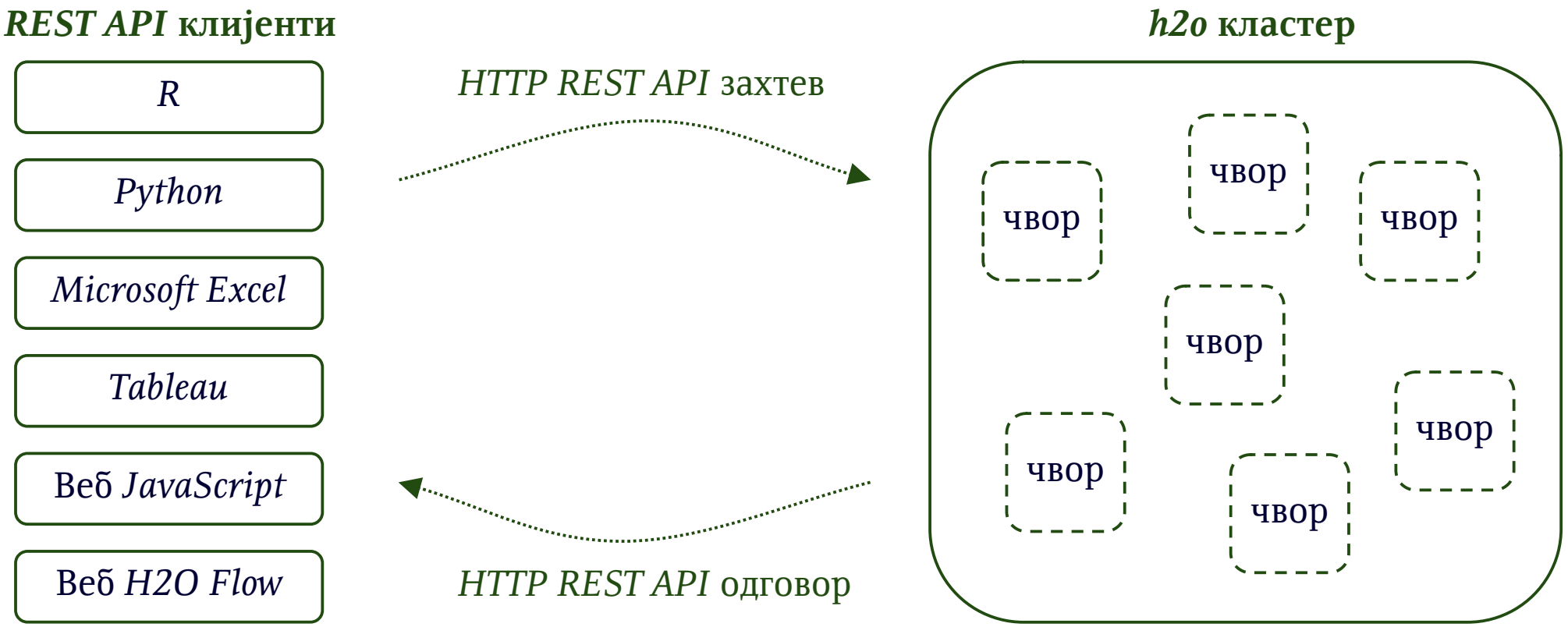
Одлике

- рад у примарној меморији
- дистрибуираност обраде
- проширивост
- доступност

Платформа h2o

Структура решења заснованог на платформи *h2o*

- *REST API* клијенти
- *h2o* кластер
- комуникација путем мреже



Платформа h2o

Структура решења заснованог на платформи *h2o*

- *REST API* клијенти
  - скриптови у језику *R*
  - скриптови у језику *Python*
  - радни листови за софтвер *Microsoft Excel*
  - визуализација кроз софтвер *Tableau*
  - интеракција кроз веб кориснички интерфејс развијен у језику *JavaScript*
  - интеракција кроз веб кориснички интерфејс *H2O Flow*

Платформа h2o

Структура решења заснованог на платформи *h2o*

- *h2o* кластер
  - кластер састављен од чворова
  - *JVM* процес по чвору
  - слојеви *JVM* процеса
    - језички слој
      - евалуација израза за језик *R*
      - слој *Shalala* за језик *Scala*
    - алгоритамски слој
      - алгоритми за учитавање података, математичко израчунавање, машинско учење и евалуацију модела
    - основни слој
      - управљање ресурсима (процесором и меморијом)

Структура решења заснованог на платформи *h2o*

- *h2o* кластер – управљање ресурсима
  - управљање меморијом – структуре података
    - *h2o* скуп података (енгл. *h2o data frame*)
      - основна јединица складиштења података намењена корисницима
      - велика дистрибуирана табела
        - именоване колоне и нумерисани редови
      - табела (енгл. *frame*) као скуп именованих вектора (енгл. *vector*)
        - може бити чувана у складишту типа кључ–вредност
        - постоји на једном чвору
      - вектор као скуп нумерисаних делова (енгл. *chunk*)
        - мора бити чуван у складишту типа кључ–вредност
        - делови у вектору су истог типа
        - вектор је дистрибуиран
      - део као скуп елемената (енгл. *element*)
        - може да садржи примитивне вредности, временске ознаке, универзално јединствене идентификаторе или стрингове
        - за већину типова садржаја примењује се компресовање

Структура решења заснованог на платформи *h2o*

- *h2o* кластер – управљање ресурсима
  - управљање меморијом – структуре података
    - дистрибуирано складиште типа кључ–вредност
      - складиште дистрибуирано у оквиру кластера
    - неблокирајућа хеш мапа
      - примењује се у имплементацији складишта типа кључ–вредност

Структура решења заснованог на платформи *h2o*

- *h2o* кластер – управљање ресурсима
  - управљање процесором – извршавање задатака
    - посао (енгл. *job*)
      - обимнија активност коју корисник може надзирати
    - задатак типа мапирање–редукција (енгл. *map/reduce task*, *MRTask*)
      - извршавање у примарној меморији
    - задатак типа подела–спој (енгл. *fork/join task*)
      - радни оквир заснован на спецификацији *JSR166* посвећеној конкурентном програмирању
        - модификација пакета *JSR166y*
      - подразумева паралелизацију и синхронизацију извршавања
        - метода **`fork()`** – асинхроно извршавање задатка
        - метода **`join()`** – прихват резултата по његовом израчунавању

Платформа h2o

Подаци

- главни извори података
  - локални датотечки систем
  - удаљени датотечки систем
  - *HDFS*
  - *Hive*
  - *JDBC*
  - *S3*

Платформа h2o

Подаци

- подржани датотечки формати
  - *ARFF*
  - *Avro*
  - *CSV*
  - *ORC*
  - *Parquet*
  - *SVMLight*
  - *XLS*
  - *XLSX*

Платформа h2o

Подршка за анализу података и машинско учење

- надгледани поступци
  - наивна Бајесова класификација
  - уопштени линеарни модели
  - машине потпорних вектора
  - неуронске мреже (дубоко учење)
  - метод дистрибуиране случајне шуме (енгл. *distributed random forest)*
  - поступак екстремног градијентног појачања (енгл. *XGBoost*)
  - ...

Платформа h2o

Подршка за анализу података и машинско учење

- ненадгледани поступци
  - метод средина̂
  - анализа главних компоненти (енгл. *principal component analysis*)
  - метод изолационе шуме (енгл. *isolation forest*)
  - ...

Садржај

Коришћење платформе h2o кроз језик R

Платформа *h2o* кроз језик *R*

- две библиотеке за језик *R* у дистрибуцији *h2o*
  - потпуна
    - дозвољава самостални режим
  - клијентска
    - само клијентске могућности
- пакет *h2o* за језик *R*
  - верзија доступна кроз репозиторијум *CRAN*
  - верзија доступна преко посебног репозиторијума

Покретање инстанце *h2o*

- директно покретање помоћу датотеке *h2o.jar* из дистрибуције
- покретање преко клијента уз помоћ пакета *h2o*
  - функција **`h2o.init()`**
    - дозвољава покретање на локалном рачунару
      - подразумевана адреса *localhost:54321*
        - коришћење протокола *HTTP* или *HTTPS* (зависно од параметра **`https`**)
      - подешавање броја нити
        - параметар **`nthreads`**
      - подешавање количине меморије за алокацију
        - параметри **`min_mem_size`** и **`max_mem_size`**
    - дозвољава и повезивање на постојећу инстанцу
  - функција **`h2o.shutdown()`**
    - заустављање инстанце уз губитак несачуваних података и резултата

Коришћење платформе h2o кроз језик R

Коришћење инстанце *h2o* кроз језик *R*

- наредбе дате на клијентској страни се шаљу и извршавају у оквиру кластера
  - наредбе се шаљу кроз *REST API*
  - одговор стиже на клијентску страну
    - одговор као *JSON* датотека
- подаци се не чувају на клијентској страни
  - постоје именовани објекти на клијентској страни преко којих се идентификују објекти на серверској страни

Коришћење инстанце *h2o* кроз језик *R*

- основни рад над објектима инстанце
  - функција **`h2o.ls()`**
    - преглед кључева објеката из инстанце
  - функција **`h2o.rm()`**
    - уклањање објеката на основу кључа
      - могућност каскадног уклањања
  - функција **`h2o.removeAll()`**
    - уклањање података из кластера
      - могућност задржавања посебно назначених објеката

Коришћење инстанце *h2o* кроз језик *R*

учитавање података на кластер

- функција **`as.h2o()`**
  - учитавање *R* објекта
  - могуће задати назив за *h2o* податке (параметар **`destination_frame`**)
- функција **`h2o.importFile()`**
  - паралелизовано читање на серверској страни
- функција **`h2o.importFolder()`**
  - учитавање целокупног директоријума датотека
- функција **`h2o.uploadFile()`**
  - преношење података с клијента на сервер
    - није предвиђено за веће количине података

Коришћење инстанце *h2o* кроз језик *R*

- преузимање података с кластера
  - функција **`as.data.frame()`**
    - преузимање података у облику скупа података за клијентску страну
  - функција **`h2o.exportFile()`**
    - преузимање података у облику датотеке за локални датотечки систем инстанце, *HDFS* или *S3N*

Коришћење платформе h2o кроз језик R

Коришћење инстанце *h2o* кроз језик *R*

очитавања и обраде над *h2o* подацима

класа **`H2OFrame`**

- користи се за представљање *h2o* података
- подржани оператори за приступање и доделу
  - **`[]`** **`[[]]`** **`$`** **`<-`**

Коришћење платформе h2o кроз језик R

Коришћење инстанце *h2o* кроз језик *R*

- очитавања и обраде над *h2o* подацима
  - функција **`h2o.cbind()`**
  - функција **`h2o.merge()`**
  - функција **`h2o.na_omit()`**
  - функција **`h2o.ncol()`**
  - функција **`h2o.nrow()`**
  - функција **`h2o.print()`**
  - функција **`h2o.rbind()`**
  - функција **`h2o.unique()`**
  - функција **`h2o.which()`**
  - ...

Коришћење инстанце *h2o* кроз језик *R*

- очитавања и обраде над *h2o* подацима
  - функција **`h2o.describe()`**
    - добијање прегледних информација о *h2o* подацима
  - функција **`h2o.splitFrame()`**
    - приближна подела *h2o* података према задатој размери

Коришћење платформе h2o кроз језик R

Коришћење инстанце *h2o* кроз језик *R*

математичка и статистичка израчунавања над *h2o* подацима

- функција **`h2o.abs()`**
- функција **`h2o.cos()`**
- функција **`h2o.exp()`**
- функција **`h2o.kurtosis()`**
- функција **`h2o.log10()`**
- функција **`h2o.max()`**
- функција **`h2o.mean()`**
- функција **`h2o.median()`**
- функција **`h2o.min()`**
- функција **`h2o.scale()`**
- функција **`h2o.sd()`**
- функција **`h2o.sin()`**
- функција **`h2o.skewness()`**
- ...

Коришћење инстанце *h2o* кроз језик *R*

формирање модела машинског учења

- посебне функције за разне моделе машинског учења
- функција **`h2o.glm()`**
  - употреба уопштених линеарних модела
- функција **`h2o.naiveBayes()`**
  - употреба наивне Бајесове класификације
- функција **`h2o.randomForest()`**
  - употреба метода случајне шуме (енгл. *random forest*)
- функција **`h2o.psvm()`**
  - употреба паралелизоване варијанте машине потпорних вектора
    - ограниченост на бинарну класификацију
- функција **`h2o.xgboost()`**
  - употреба поступка екстремног градијентног појачања (енгл. *XGBoost*)
- ...

Коришћење инстанце *h2o* кроз језик *R*

- формирање модела машинског учења
  - посебне функције за разне моделе машинског учења
    - одређени заједнички параметри
      - параметар **`x`**
        - предикторска обележја
      - параметар **`y`**
        - циљно обележје
      - параметар **`training_frame`**
        - ознака скупа података за обучавање
      - параметар **`validation_frame`**
        - ознака скупа података за валидацију
      - параметар **`model_id`**
        - ознака за модел
      - параметар **`seed`**
        - подешавање генератора бројева (у случају потребе)
    - параметри у вези с унакрсном валидацијом
      - углавном доступни
        - параметар **`nfolds`**
        - ...

Коришћење инстанце *h2o* кроз језик *R*

- формирање модела машинског учења
  - функција **`h2o.grid()`**
    - примена поступне (свеобухватне) претраге (енгл. *grid search*) у подешавању вредности параметара приликом обучавања
  - функција **`h2o.automl()`**
    - аутоматизација обучавања и евалуације већег броја модела

Коришћење инстанце *h2o* кроз језик *R*

коришћење и испитивање модела машинског учења

- функција **`h2o.explain()`**
  - подршка тумачењу модела
    - углавном помоћу визуализација
- функција **`h2o.performance()`**
  - анализа перформанси модела
- функција **`h2o.predict()`**
  - давање предвиђања помоћу модела

Коришћење инстанце *h2o* кроз језик *R*

- додатне анализе
  - функција **`h2o.kmeans()`**
    - употреба метода средина
  - функција **`h2o.prcomp()`**
    - извођење анализе главних компоненти
  - ...

Садржај

Извори и литература

Основни извори и литература

- Cook D. Practical machine learning with H2O: Powerful, scalable techniques for AI and deep learning. O'Reilly; 2017.
- H2O.ai. Convergence of the world's best predictive and generative AI for private, protected data. Internet: https://h2o.ai/
- H2O.ai. Overview – H2O 3.46.0.7 documentation. Internet: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html
- H2O.ai. H2O architecture – H2O 3.46.0.7 documentation. Internet: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/architecture.html
- H2O.ai. H2O architecture. Internet: https://h2o.ai/blog/2014/h2o-architecture/
- H2O.ai. Overview (h2o-core 3.46.0 API). Internet: https://docs.h2o.ai/h2o/latest-stable/h2o-core/javadoc/index.html
- R Project. CRAN: Package h2o. Internet: https://cran.r-project.org/web/packages/h2o/index.html

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Платформа за анализу података и машинско учење h2o

(материјали за предавања)