? ... !

Párhuzamos R programok végrehajtása sparklyr segítségével az ELKH Cloud infrastruktúrán

Kacsuk Zoltán (HdM IAAI)

Understanding the role of Apache Spark in the big data ecosystem

Big Data

The theoretical approach

a) Volume
b) Variety
c) Velocity

Big Data

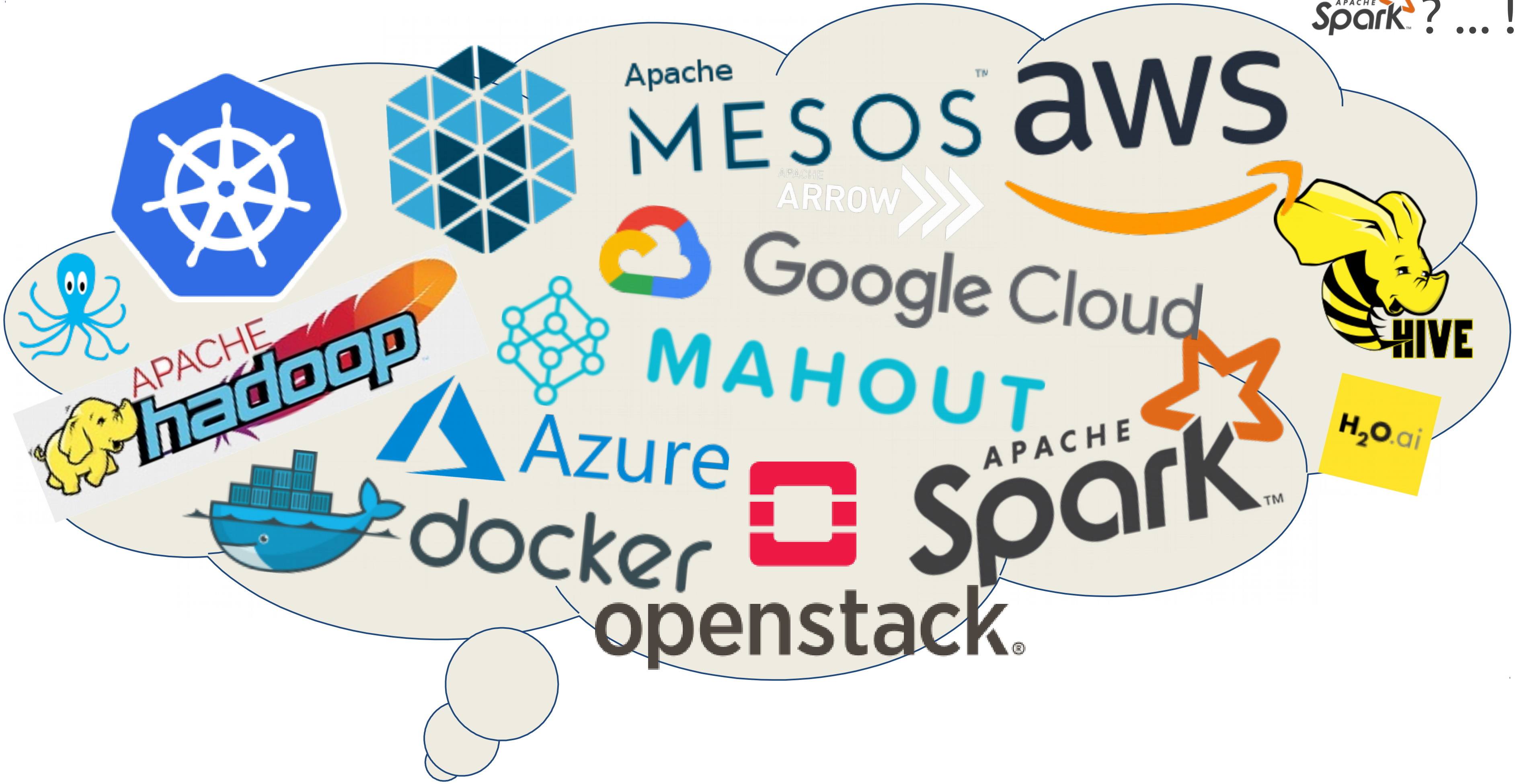
The theoretical approach

a) Volume
b) Variety
c) Velocity

The practical approach

d) Doesn’t fit
e) Too slow

APACHE Spark ? ... !



It can feel a bit overwhelming. (And this is just the tip of the iceberg.)

Implementing a machine learning process: the **how**

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations is achieved

Data distribution and **I/O operations**

? ... !

Implementing a machine learning process: the **how**

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations is achieved

Data distribution and **I/O operations**

Implementing a machine learning process: the **how**

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations is achieved

Data distribution and **I/O operations**

Implementing a machine learning process: the **how**

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations is achieved

Data distribution and **I/O operations**

Implementing a machine learning process: the **how**

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations

Data distribution and **I/O operations**

Implementing a machine learning process: the **how**

The type of **statistical model** MAHOUT

The **exact formula** being implemented MAHOUT

The way the **algorithm** works MAHOUT

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations

Data distribution and **I/O operations**

Spark ? ... !

Implementing a machine learning process: the **how**

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations

Data distribution and **I/O operations**

Implementing a machine learning process: the **how**

The type of **statistical model**

This is the Spark Machine Learning Library

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved

The way the **control of the distribution** of the operations

Data distribution and **I/O operations**

Apache Spark and Hadoop architectures in a nutshell

Spark architecture basics

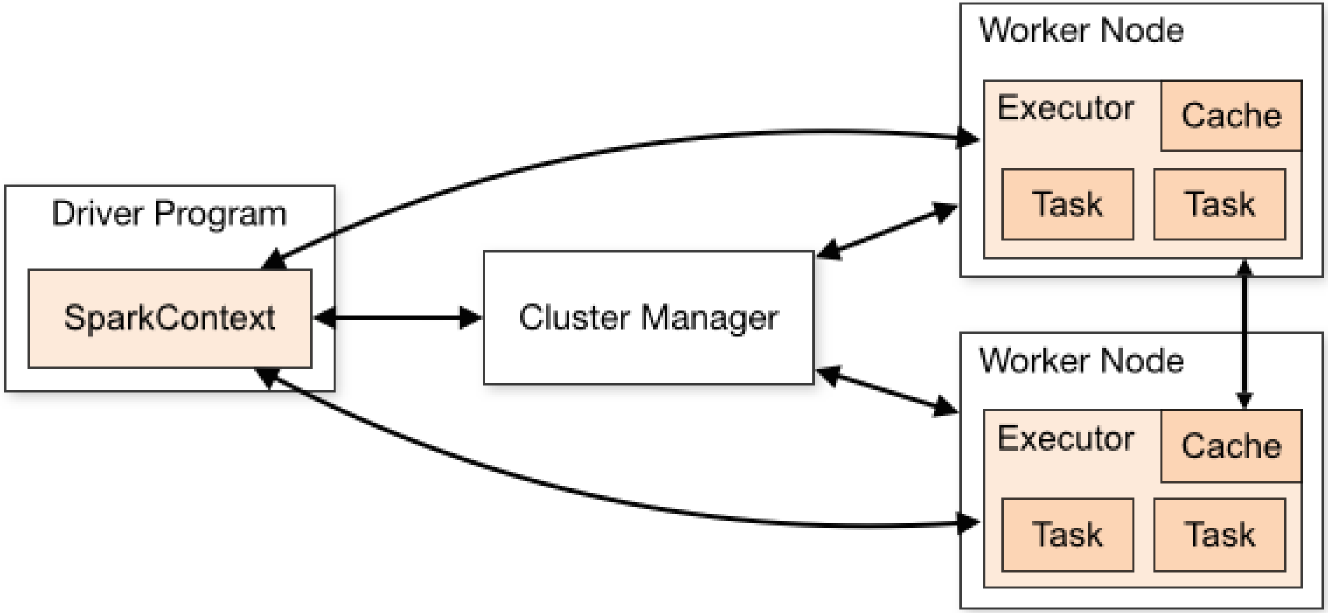


Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Spark architecture basics

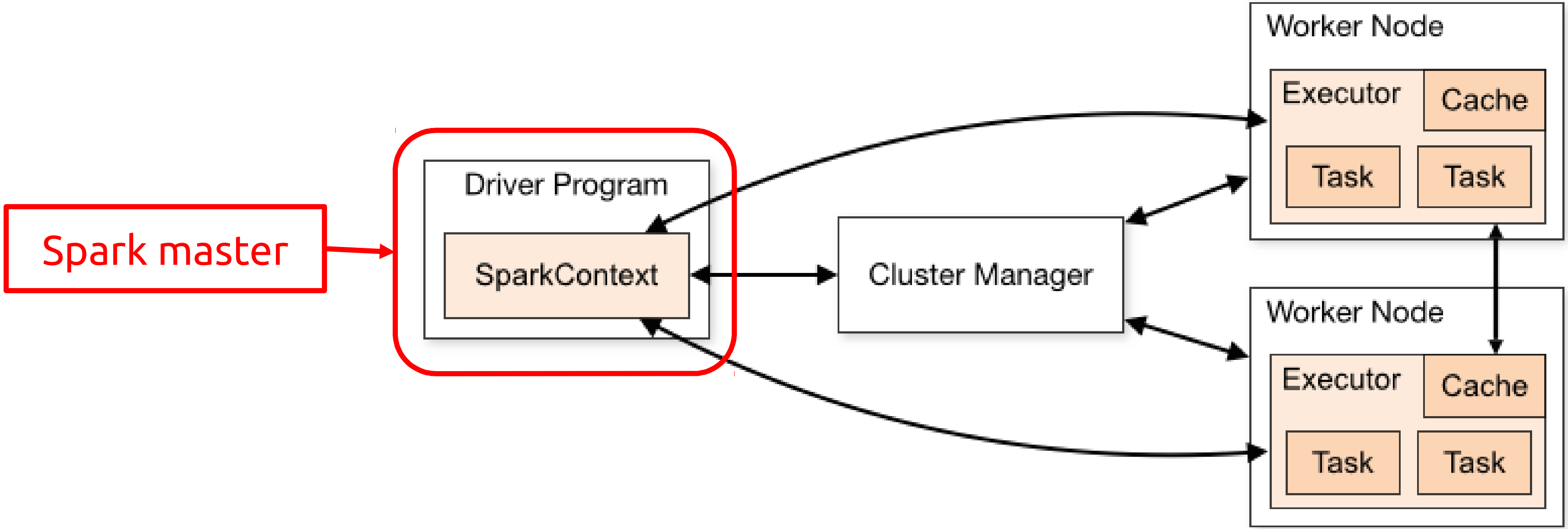


Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Spark architecture basics

Spark ? ... !

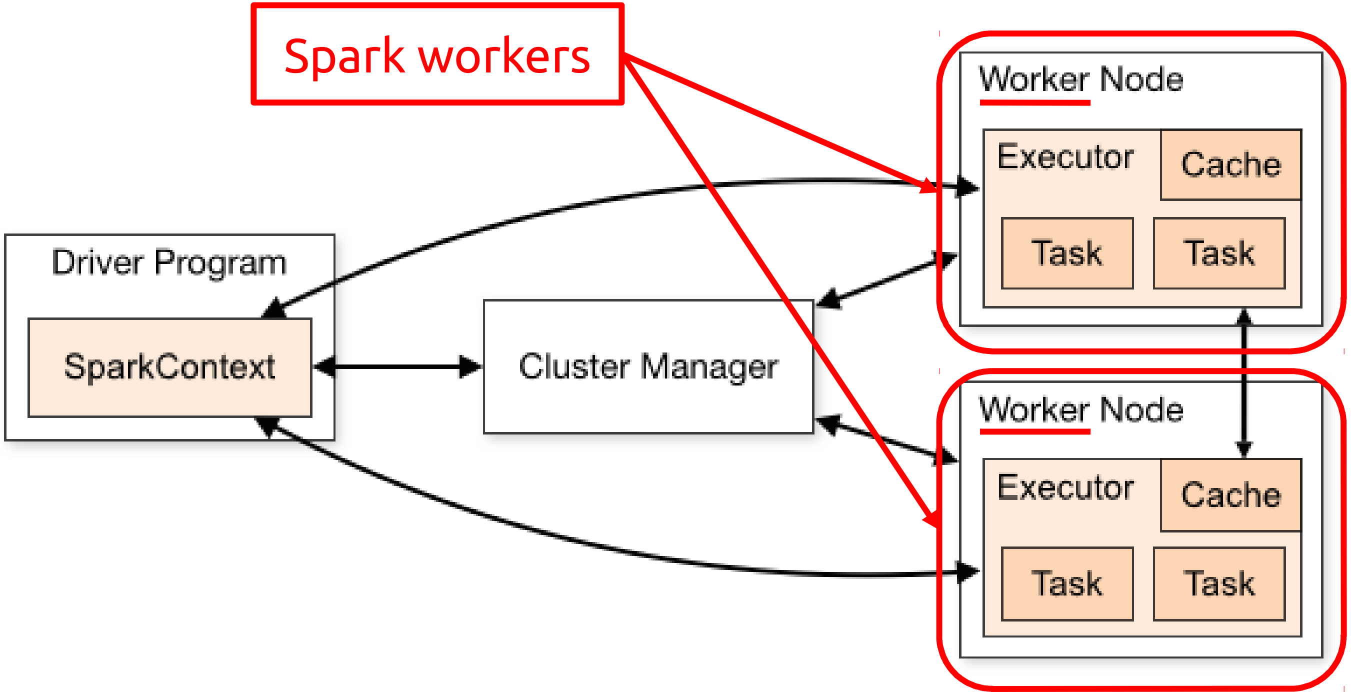


Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Spark architecture basics

APIs:
Scala
Java
Python
R

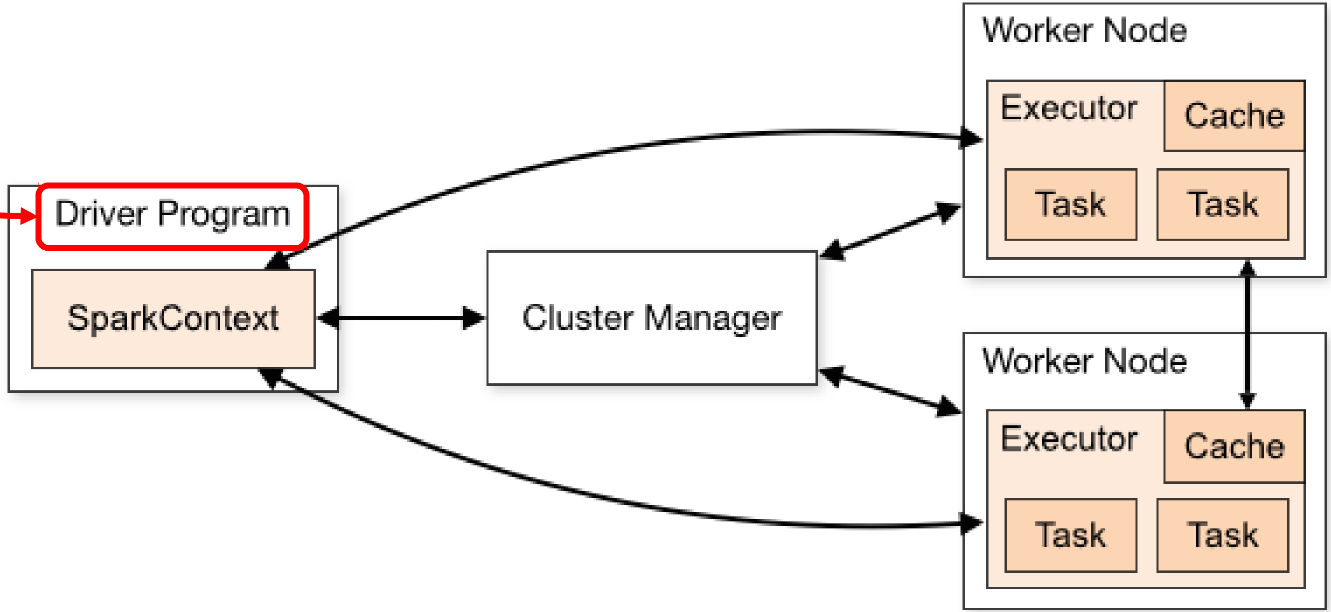


Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Spark architecture basics

Think of this as a single shell, session or job

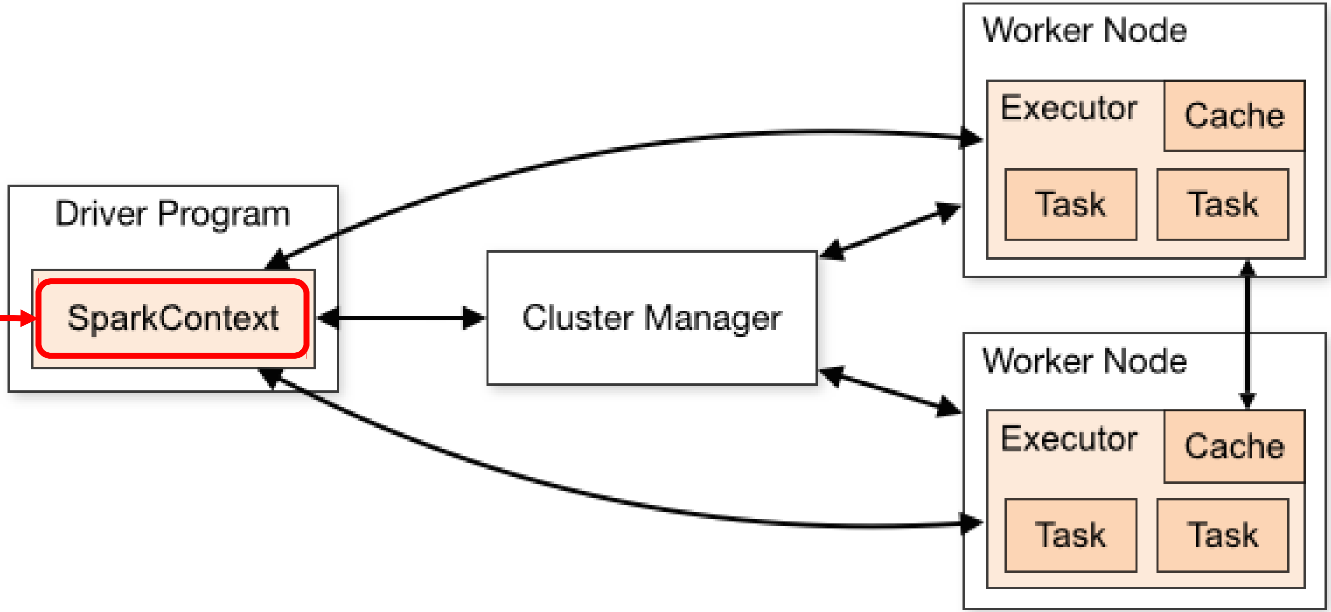


Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Spark architecture basics

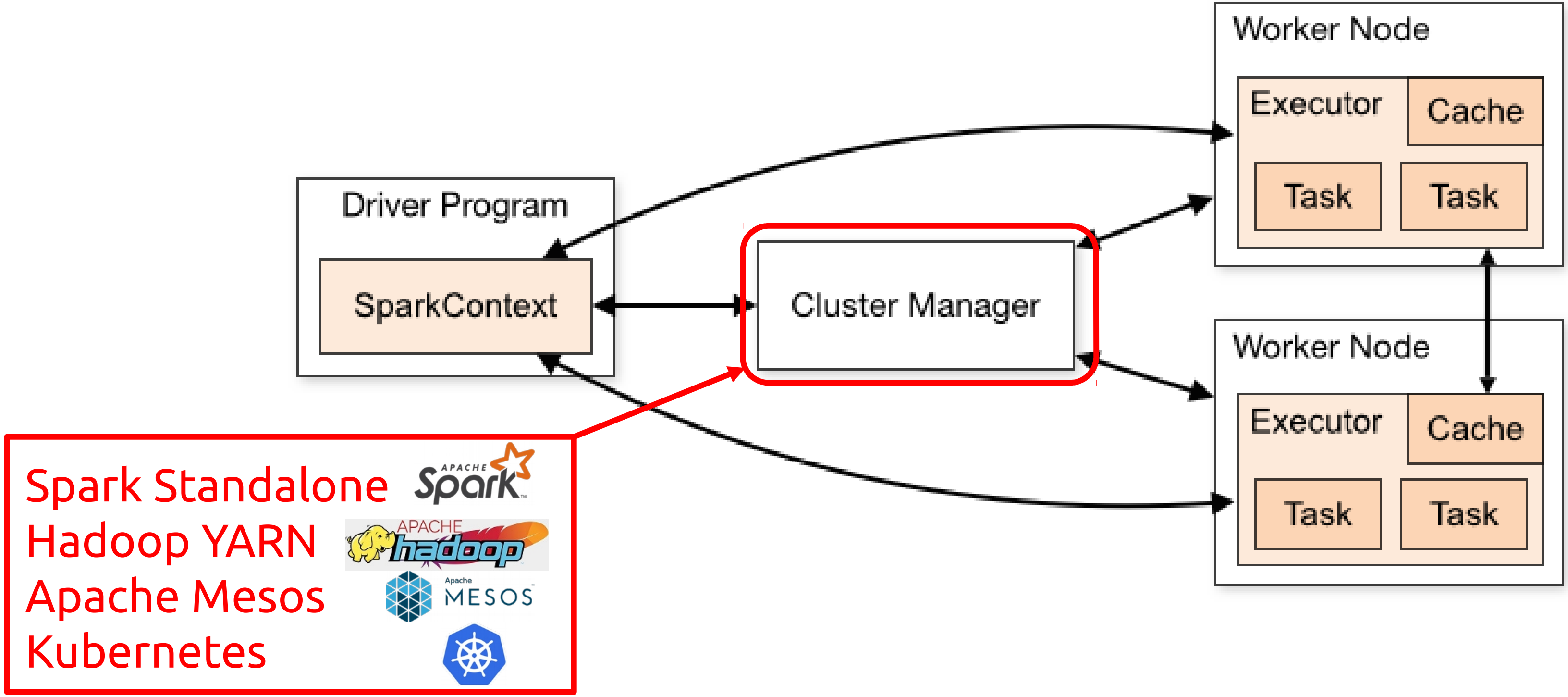
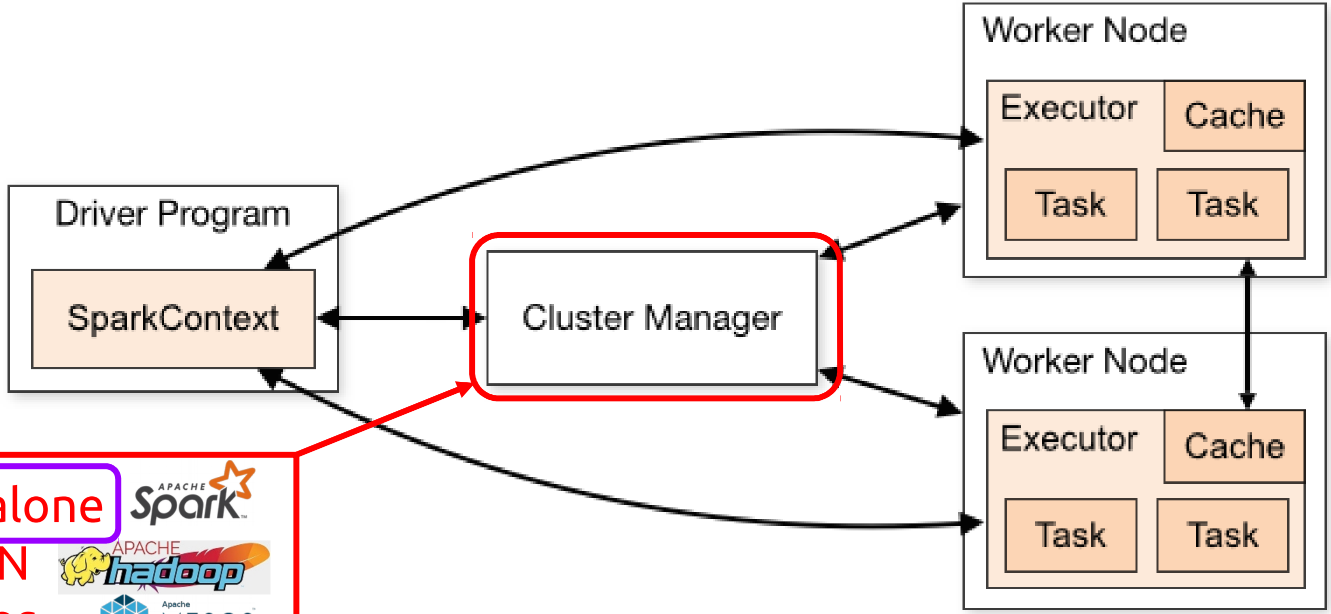


Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Spark architecture basics

IMPORTANT:
Standalone
does not equal
local mode!



Spark Standalone
Hadoop YARN
Apache Mesos
Kubernetes

Image source and further details: https://spark.apache.org/docs/latest/cluster-overview.html

Understanding Hadoop for Spark

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved: **Hadoop MapReduce**

The way the **control of the distribution** of the operations: **Hadoop YARN**

Data distribution and **I/O operations**: **Hadoop Distributed File System (HDFS)**

Understanding Hadoop for Spark

The type of **statistical model**

The **exact formula** being implemented

The way the **algorithm** works

The way **individual calculations** are computed

The way the **distribution of the operations** is achieved: **Hadoop MapReduce**

The way the **control of the distribution** of the operations: **Hadoop YARN**

Data distribution and **I/O operations**: **Hadoop Distributed File System (HDFS)**

The Hadoop Distributed File System (HDFS)

Spark ? … !

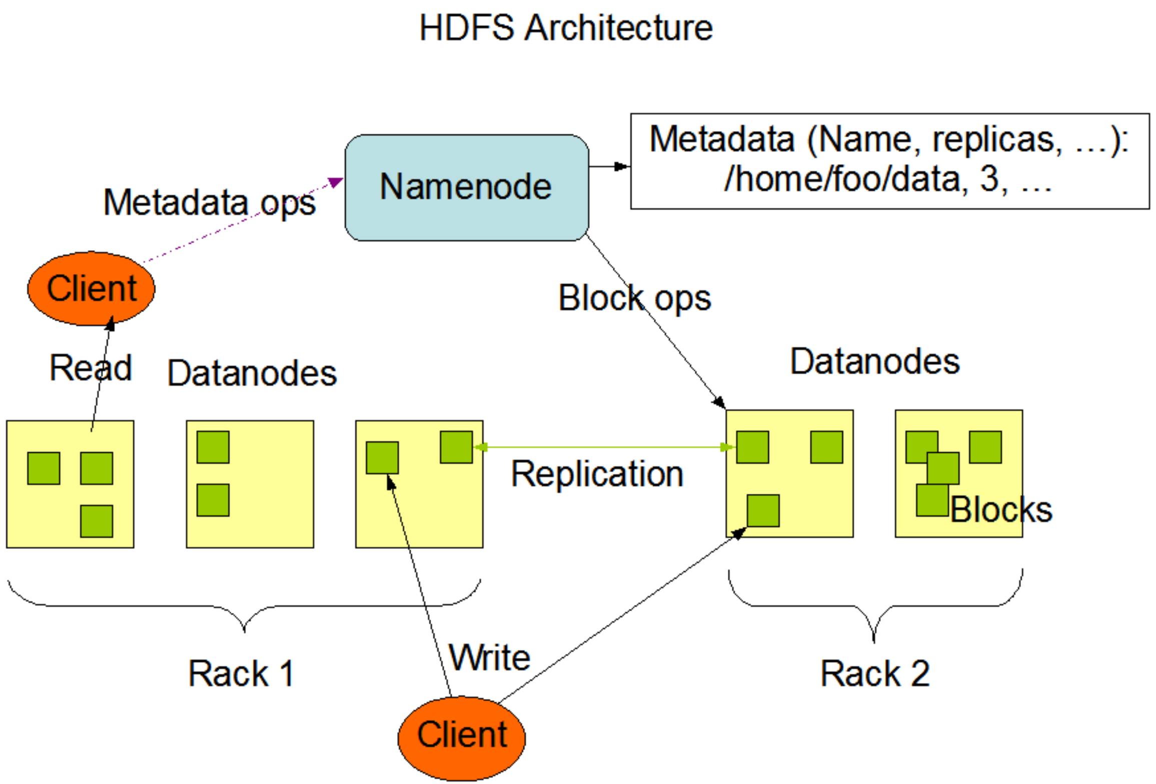
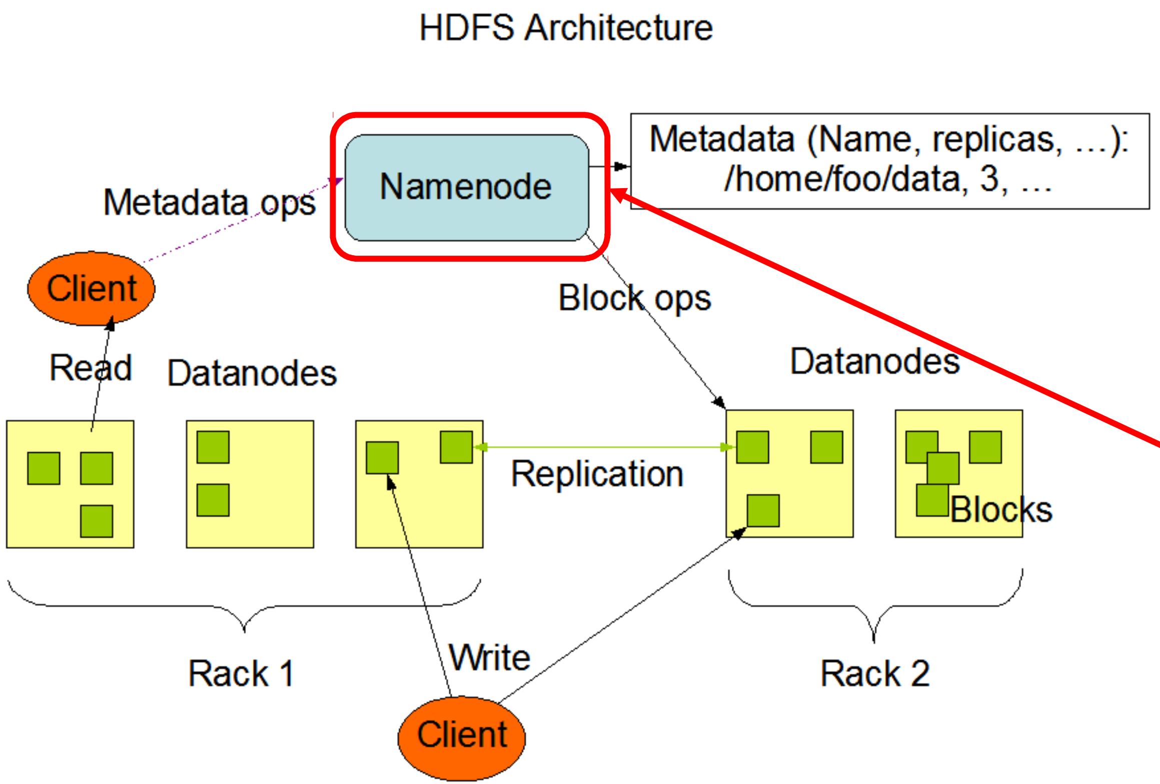


Image source and further details: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

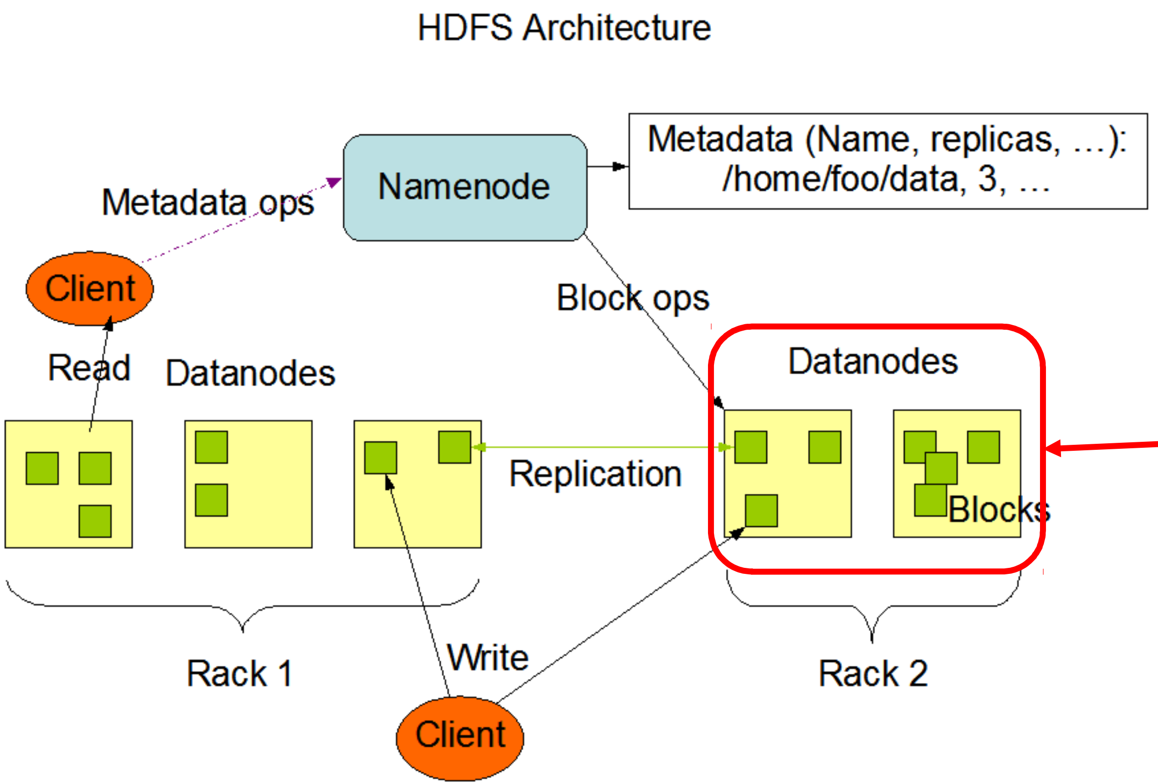
The HDFS NameNode



- This is running **on the Spark master**.
- This is what we see when we browse the filesystem.

Image source and further details: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

The HDFS DataNodes



- These are running **on the Spark workers**.
- This is where the data actually is.

Image source and further details: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Data replication and fault tolerance in the HDFS

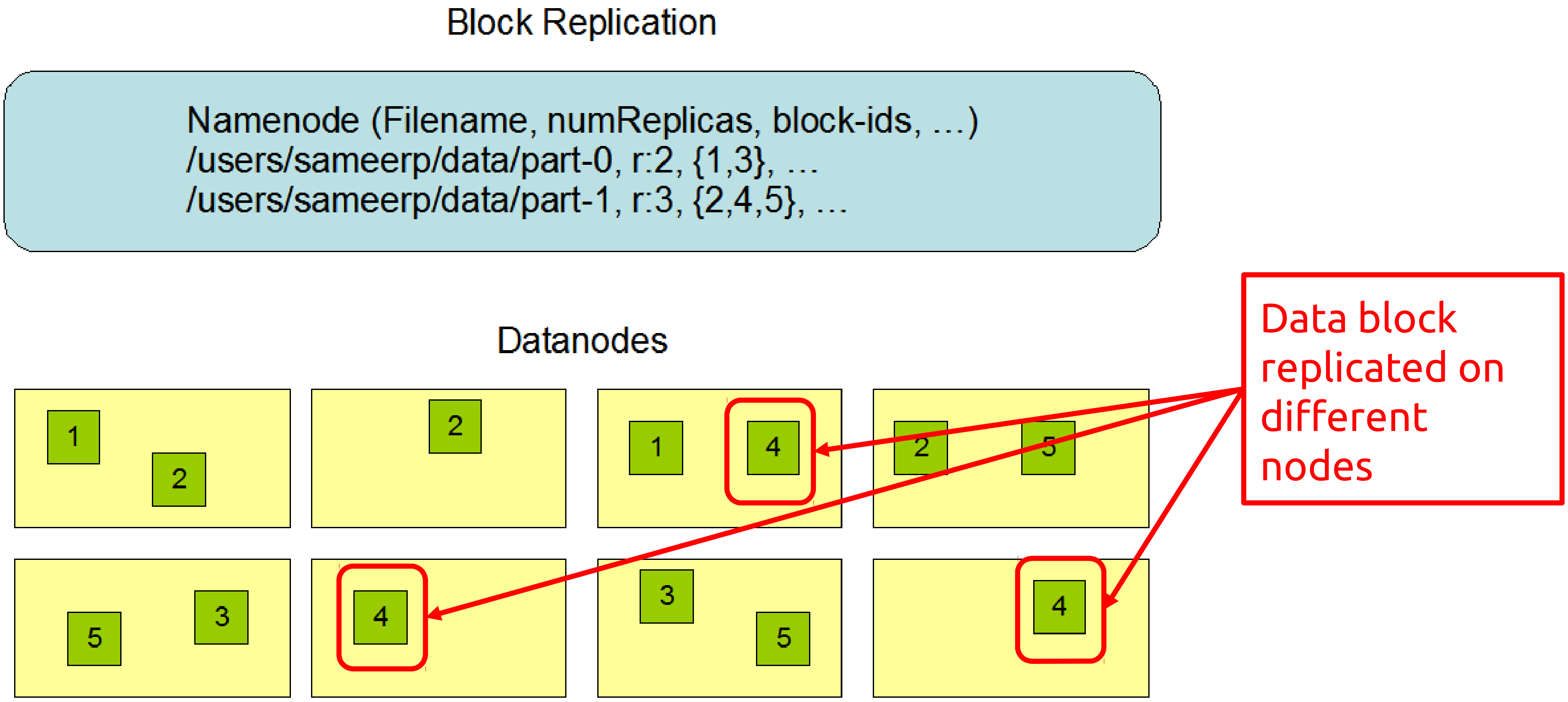
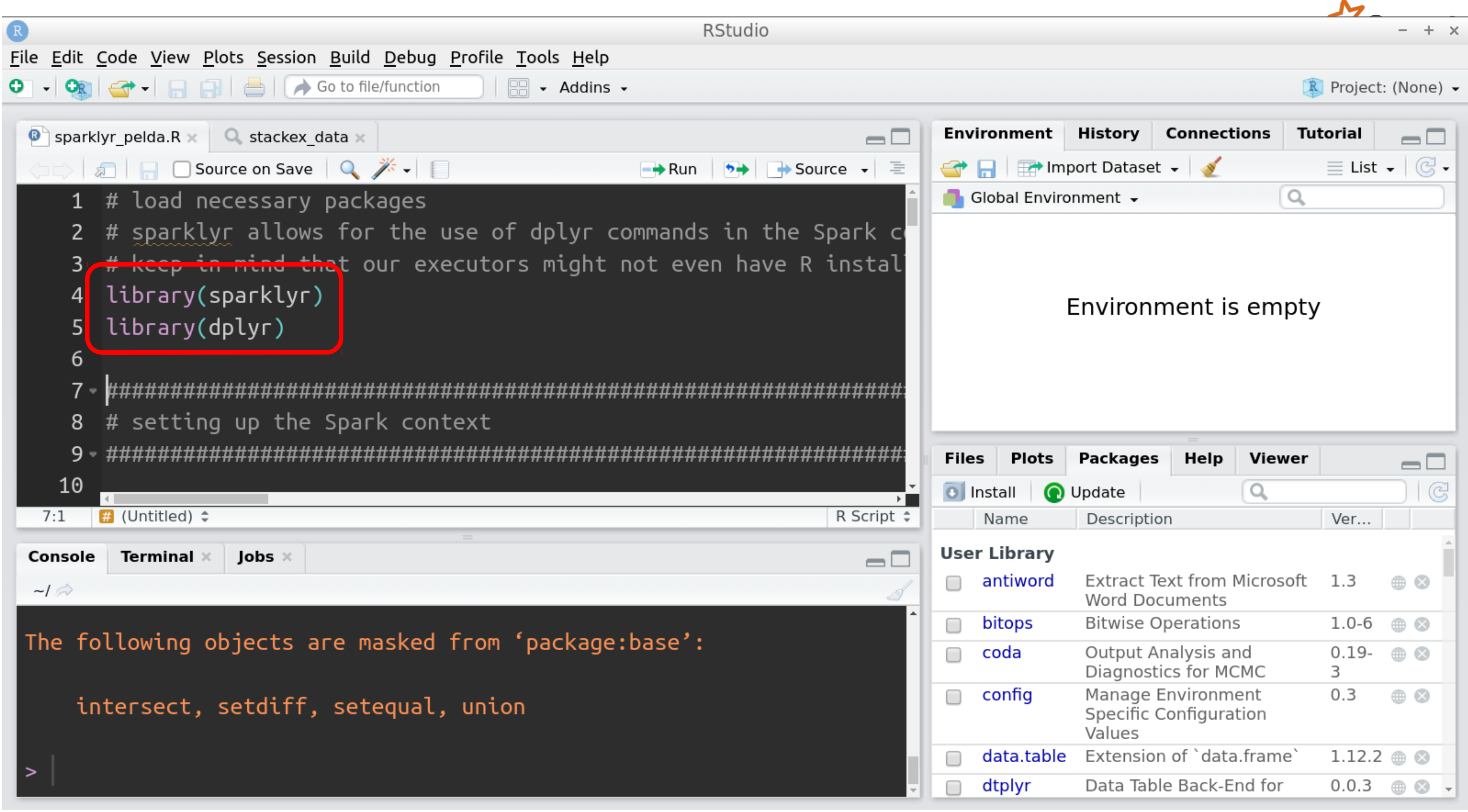
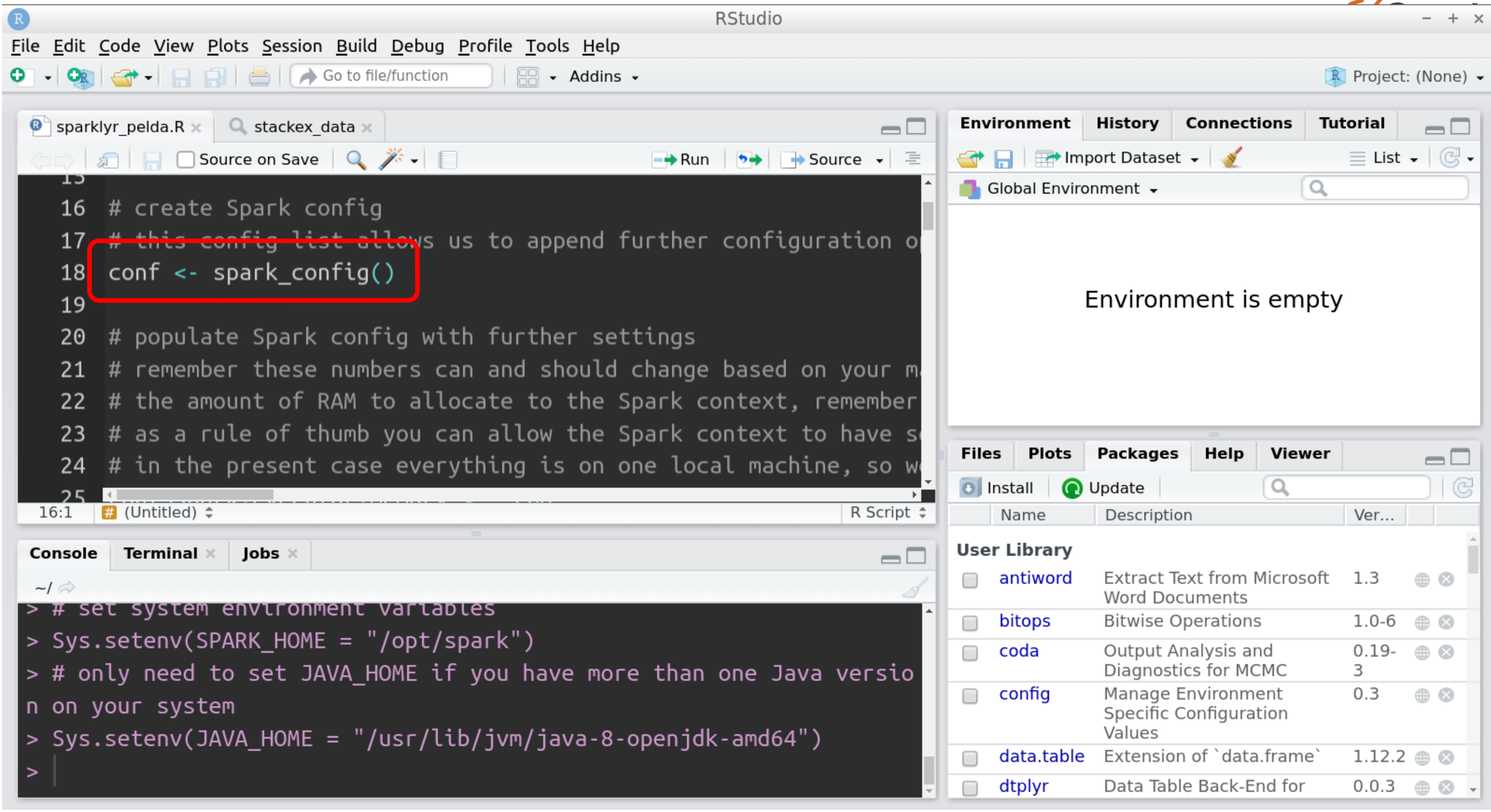


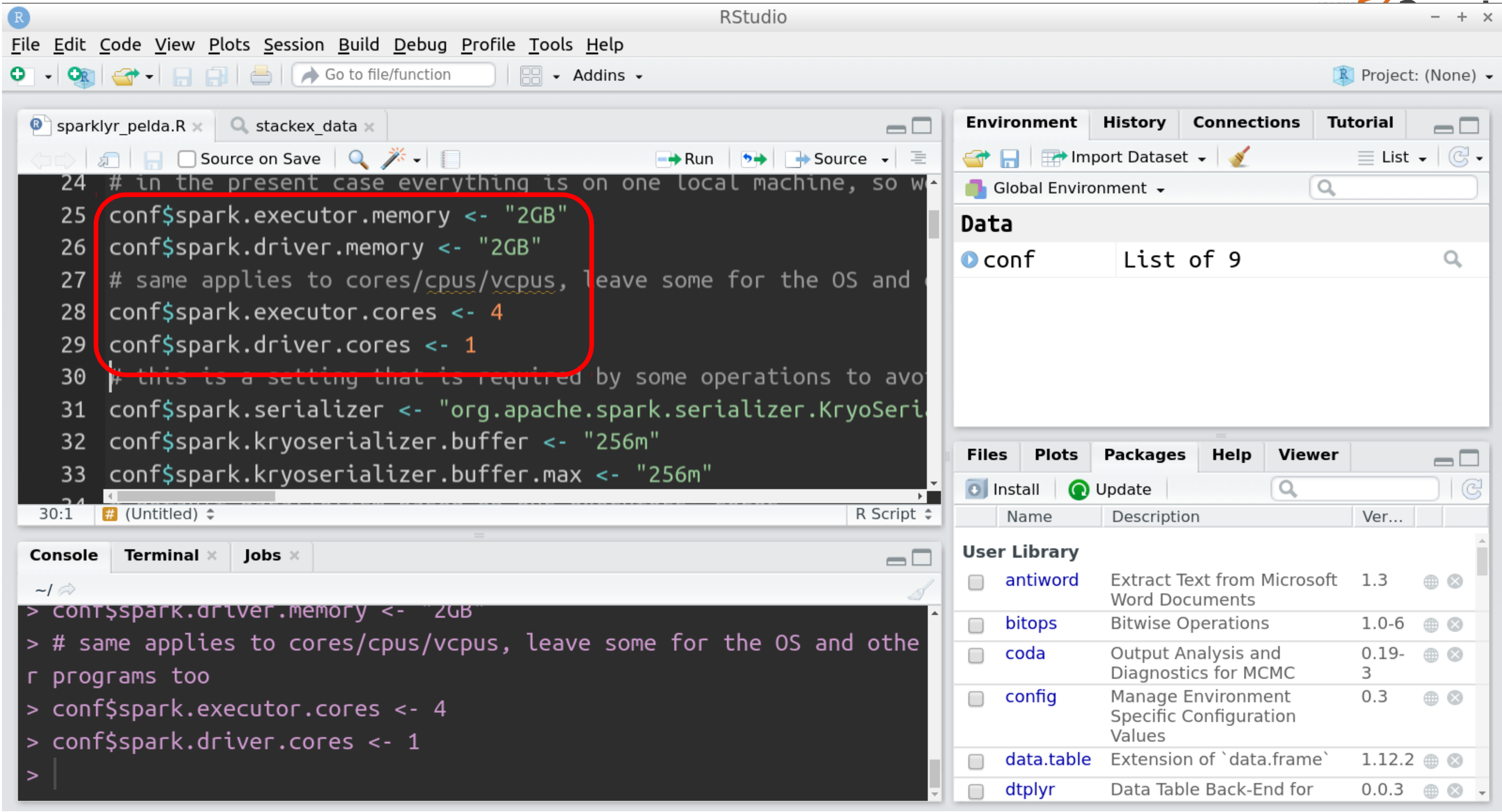
Image source and further details: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Let's get practical!

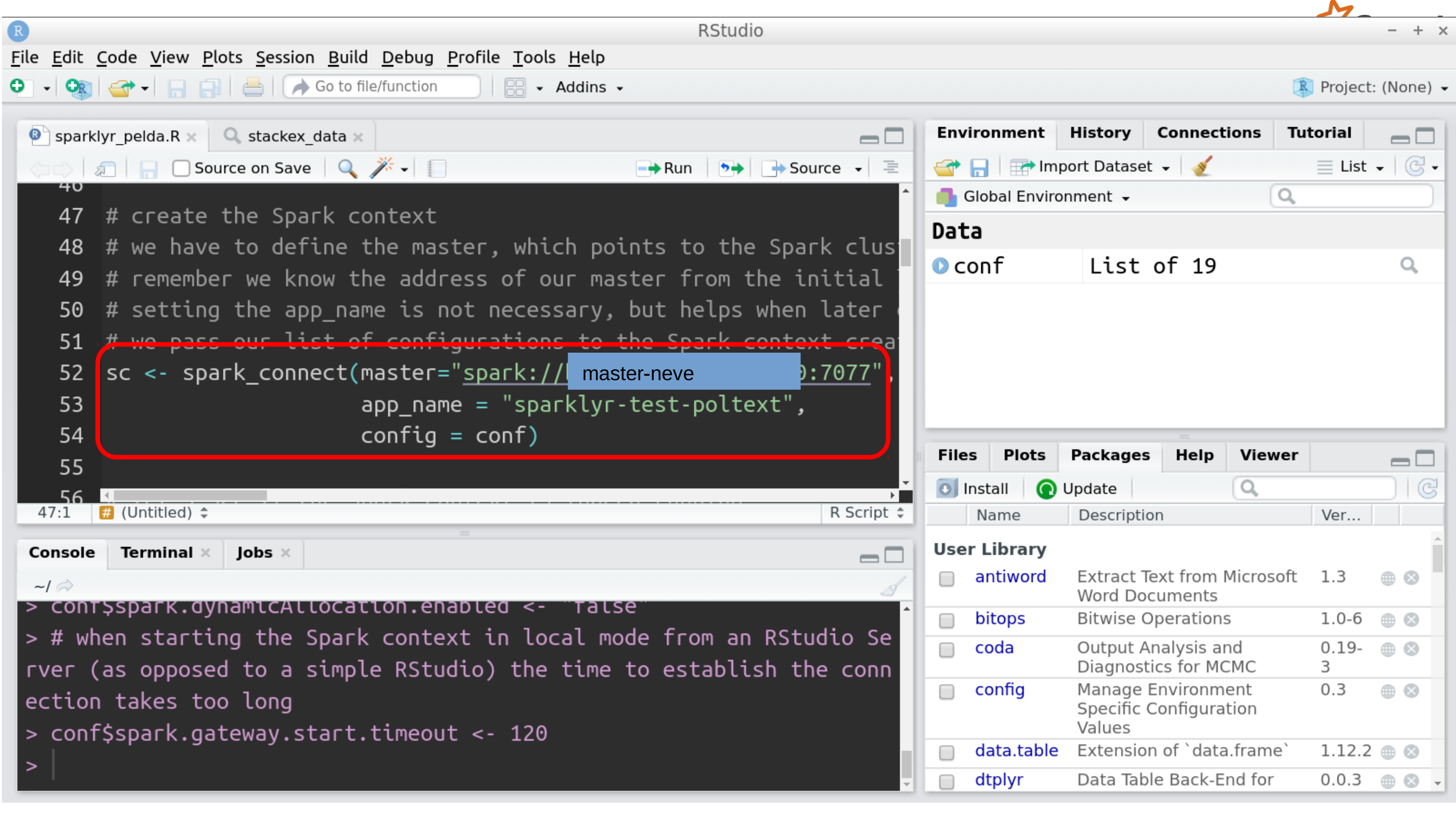
You can find all code and explanations at:
https://github.com/zkpti/poltext2019-sparktutorial

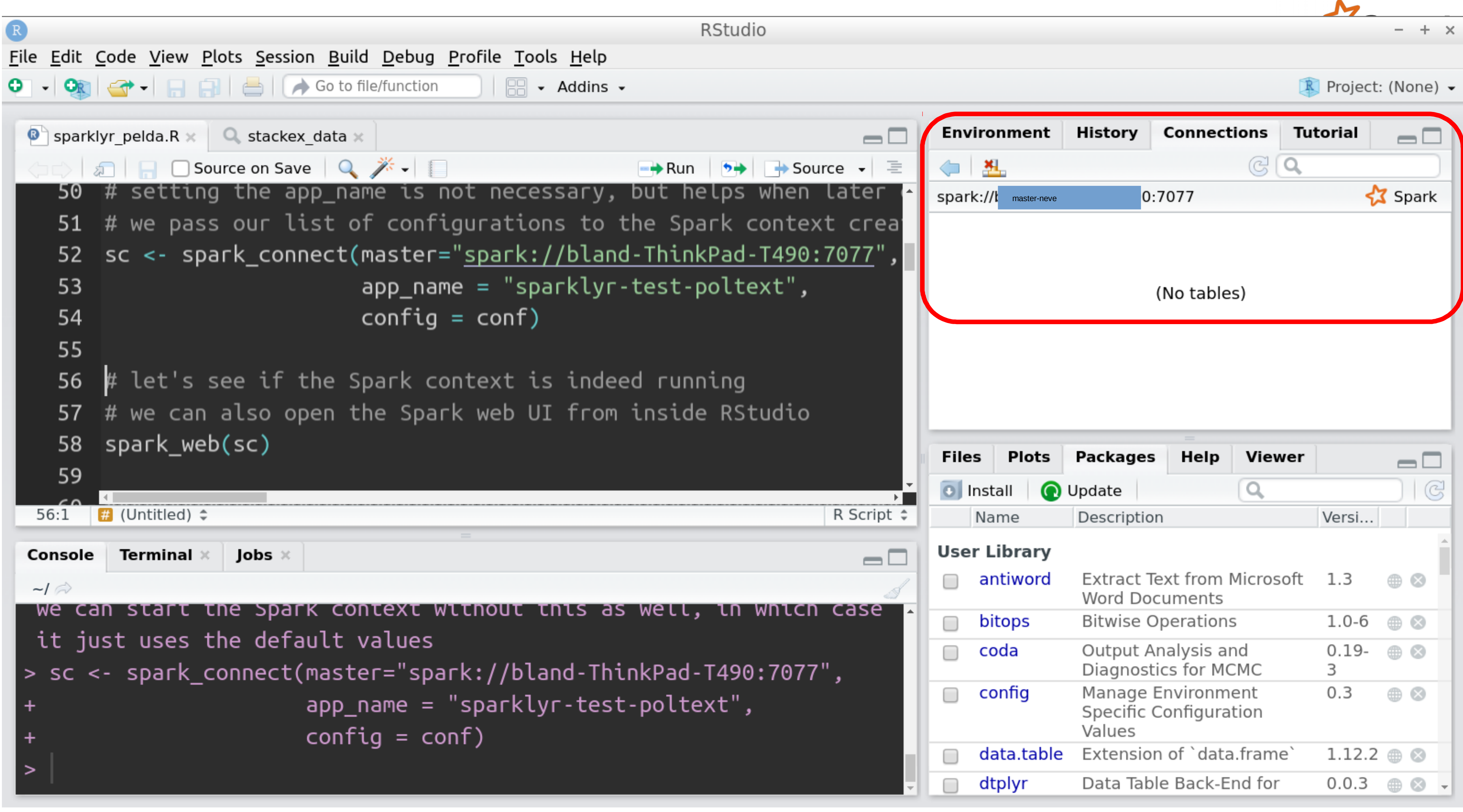

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
1 # load necessary packages
2 # sparklyr allows for the use of dplyr commands in the Spark c
3 # keep in mind that our executors might not even have R instal
4 library(sparklyr)
5 library(dplyr)
6
7 ##
8 # setting up the Spark context
9 ##
10
7:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
The following objects are masked from 'package:base':
intersect, setdiff, setequal, union
>
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Ver...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

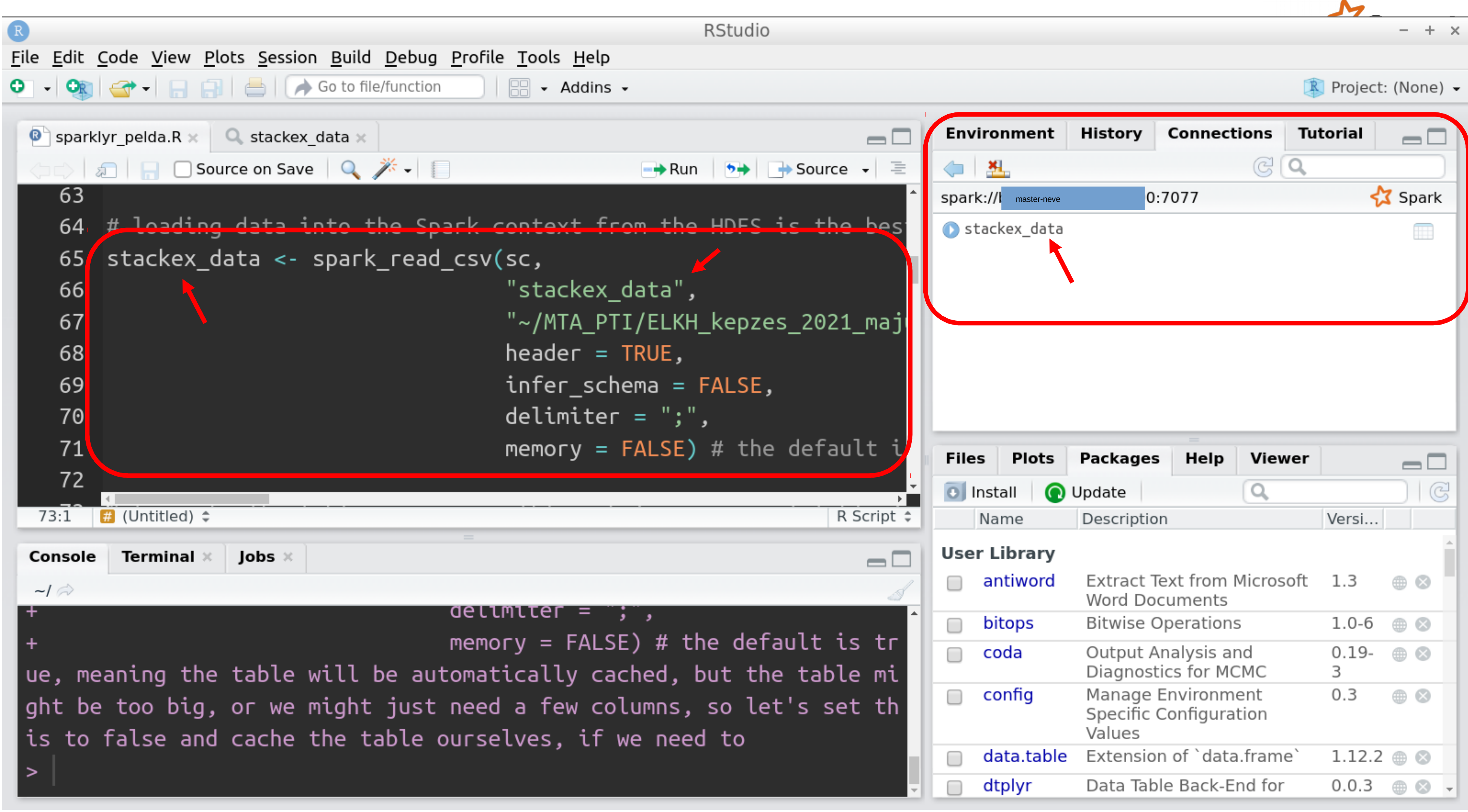

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
# create Spark config
# this config list allows us to append further configuration o
conf <- spark_config()
# populate Spark config with further settings
# remember these numbers can and should change based on your m
# the amount of RAM to allocate to the Spark context, remember
# as a rule of thumb you can allow the Spark context to have se
# in the present case everything is on one local machine, so w
16:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
> # set system environment variables
> Sys.setenv(SPARK_HOME = "/opt/spark")
> # only need to set JAVA_HOME if you have more than one Java version on your system
> Sys.setenv(JAVA_HOME = "/usr/lib/jvm/java-8-openjdk-amd64")
>
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Ver...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

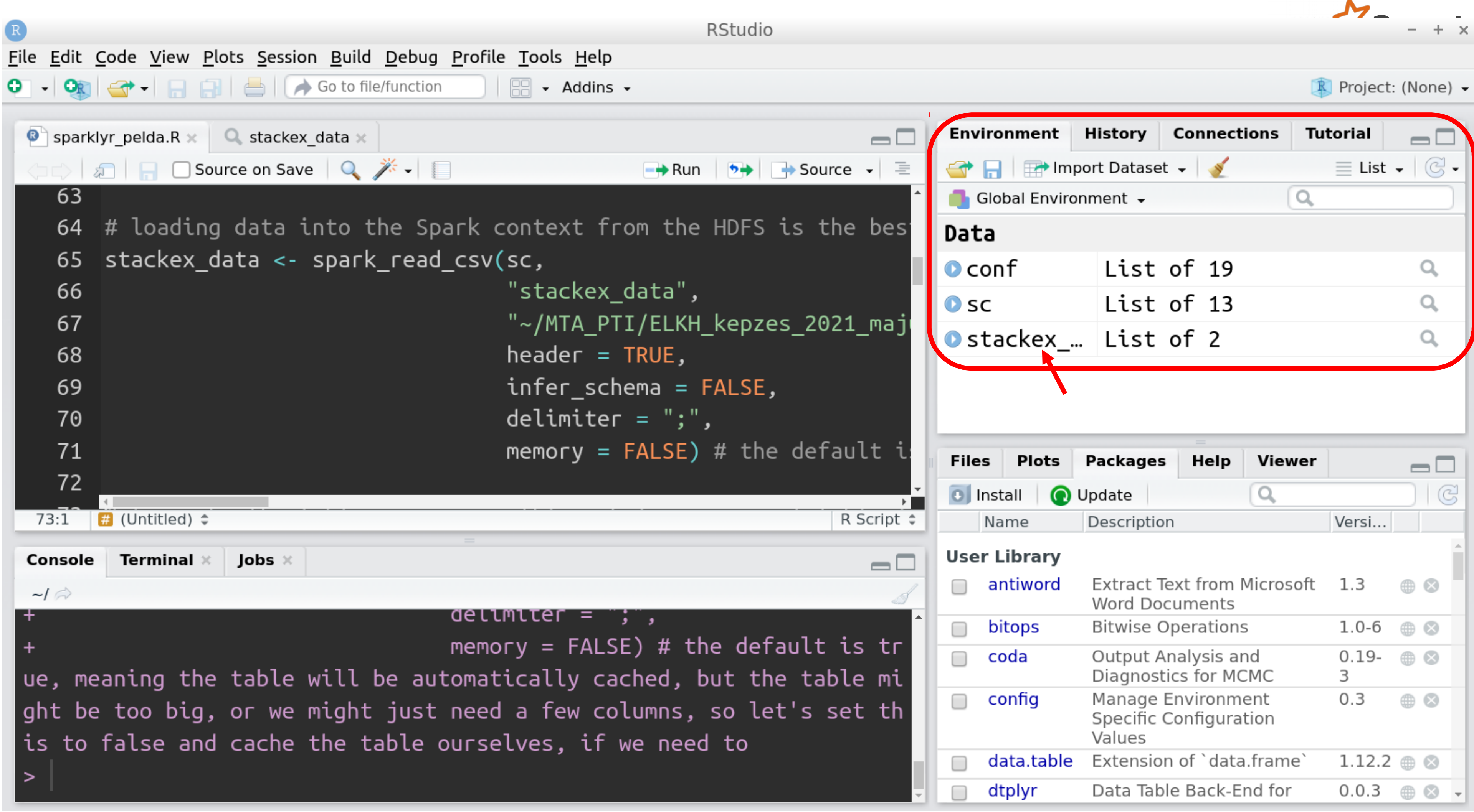


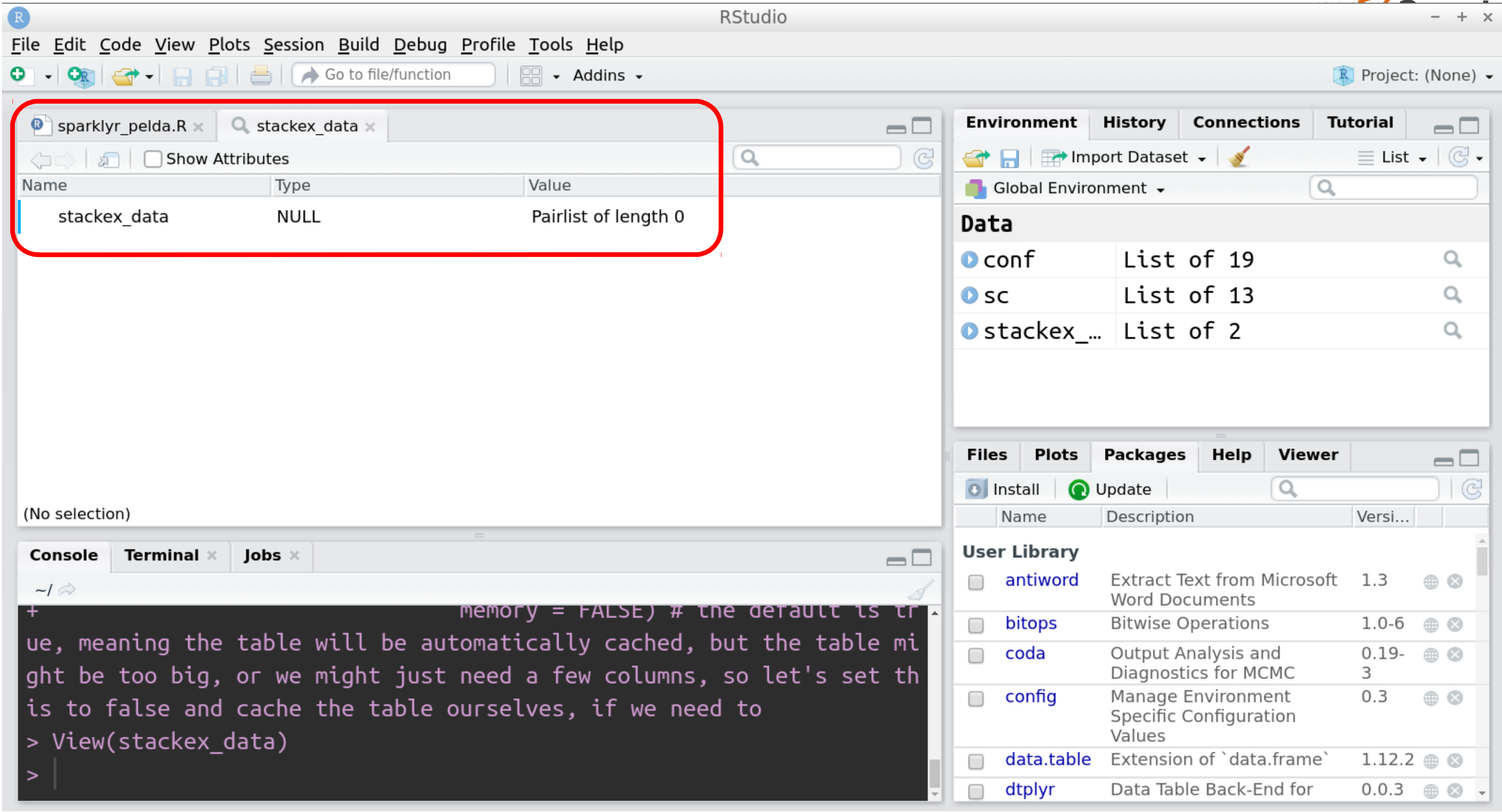
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
24 # in the present case everything is on one local machine, so we
25 conf$spark.executor.memory <- "2GB"
26 conf$spark.driver.memory <- "2GB"
27 # same applies to cores/cpus/vcpus, leave some for the OS and
28 conf$spark.executor.cores <- 4
29 conf$spark.driver.cores <- 1
30 # this is a setting that is required by some operations to avo
31 conf$spark.serializer <- "org.apache.spark.serializer.KryoSeri
32 conf$spark.kryoserializer.buffer <- "256m"
33 conf$spark.kryoserializer.buffer.max <- "256m"
30:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
> conf$spark.driver.memory <- "2GB"
> # same applies to cores/cpus/vcpus, leave some for the OS and othe
r programs too
> conf$spark.executor.cores <- 4
> conf$spark.driver.cores <- 1
>
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
Data
conf
List of 9
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Ver...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

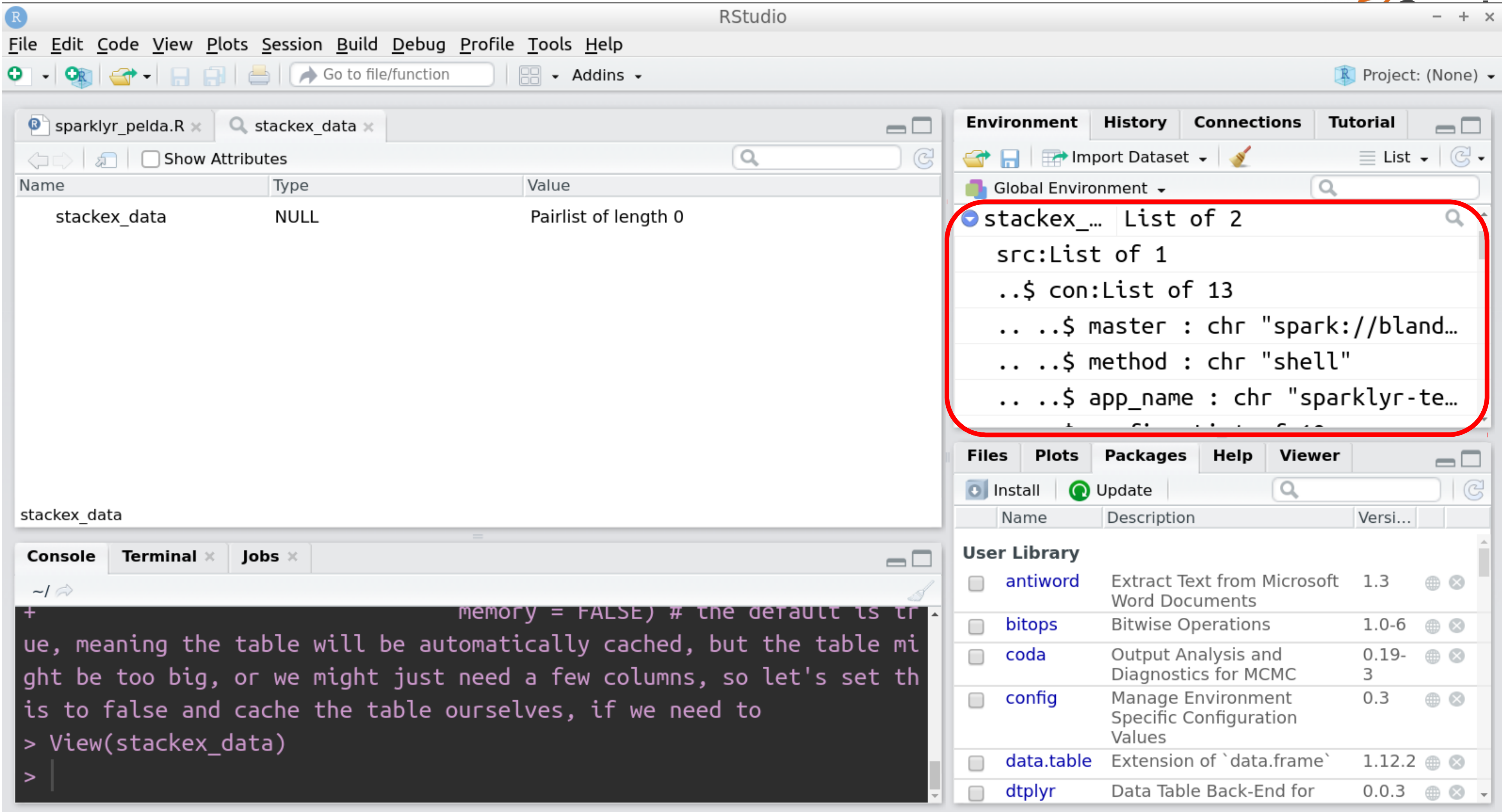

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
47 # create the Spark context
48 # we have to define the master, which points to the Spark clus
49 # remember we know the address of our master from the initial
50 # setting the app_name is not necessary, but helps when later
51 # we pass our list of configurations to the Spark context crea
52 sc <- spark_connect(master="spark://
master-neve
:7077",
53 app_name = "sparklyr-test-poltext",
54 config = conf)
55
56
47:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
> conf$spark.dynamicAllocation.enabled <- "false"
> # when starting the Spark context in local mode from an RStudio Server (as opposed to a simple RStudio) the time to establish the connection takes too long
> conf$spark.gateway.start.timeout <- 120
>
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
Data
conf
List of 19
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Ver...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

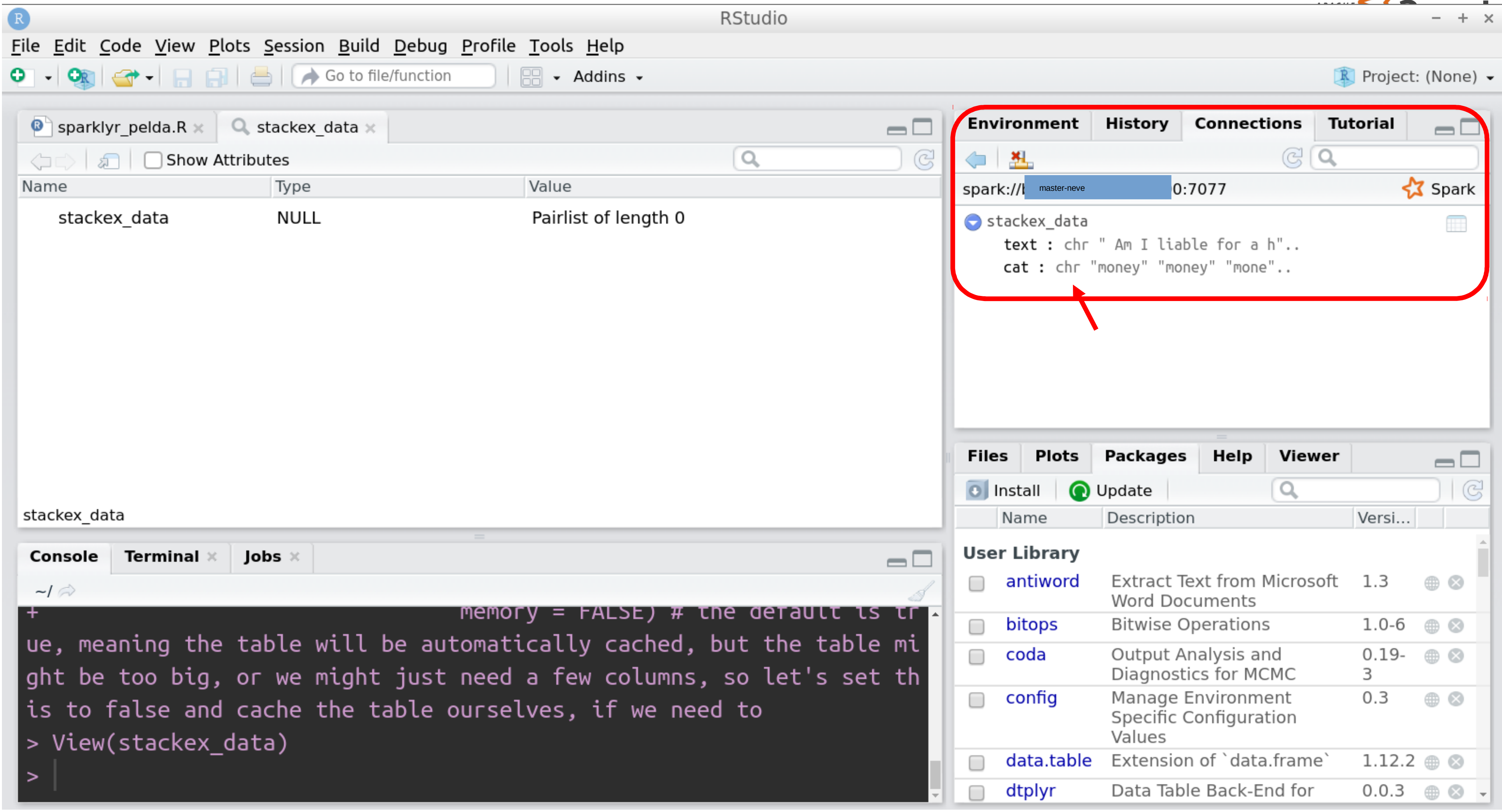

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
# setting the app_name is not necessary, but helps when later
# we pass our list of configurations to the Spark context crea
sc <- spark_connect(master="spark://bland-ThinkPad-T490:7077",
app_name = "sparklyr-test-poltext",
config = conf)
# let's see if the Spark context is indeed running
# we can also open the Spark web UI from inside RStudio
spark_web(sc)
56:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
we can start the Spark context without this as well, in which case
it just uses the default values
> sc <- spark_connect(master="spark://bland-ThinkPad-T490:7077",
+ app_name = "sparklyr-test-poltext",
+ config = conf)
>
Environment
History
Connections
Tutorial
spark://b
master-neve
0:7077
Spark
(No tables)
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

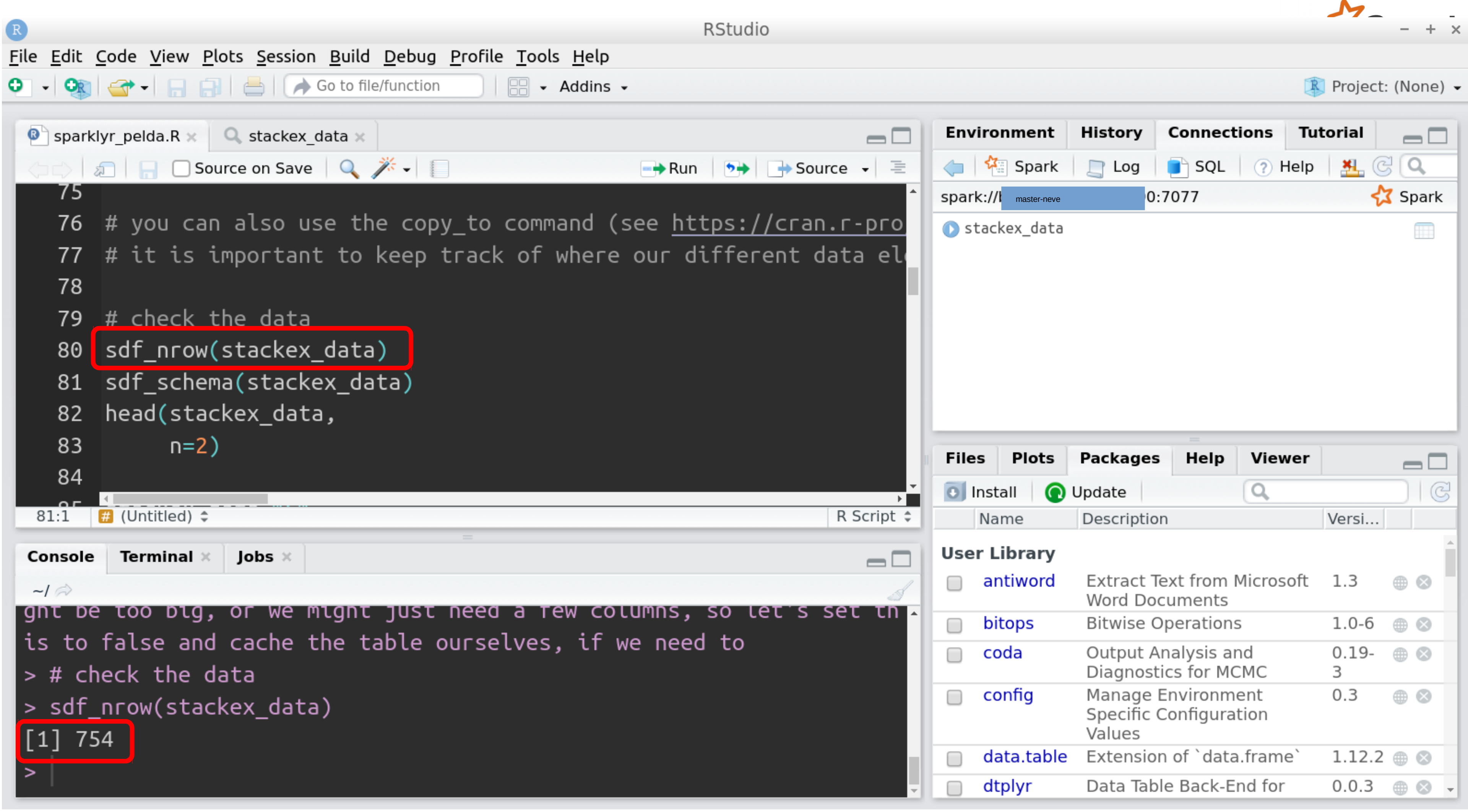

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
63
64 # loading data into the Spark context from the HDFS is the bes
65 stackex_data <- spark_read_csv(sc,
66 "stackex_data",
67 "~/MTA_PTI/ELKH_kepzes_2021_maj
68 header = TRUE,
69 infer_schema = FALSE,
70 delimiter = ";",
71 memory = FALSE) # the default i
72
73:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
+ delimiter = ";",
+ memory = FALSE) # the default is tr
ue, meaning the table will be automatically cached, but the table mi
ght be too big, or we might just need a few columns, so let's set th
is to false and cache the table ourselves, if we need to
>
Environment
History
Connections
Tutorial
spark://
master-neve
0:7077
Spark
stackex_data
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

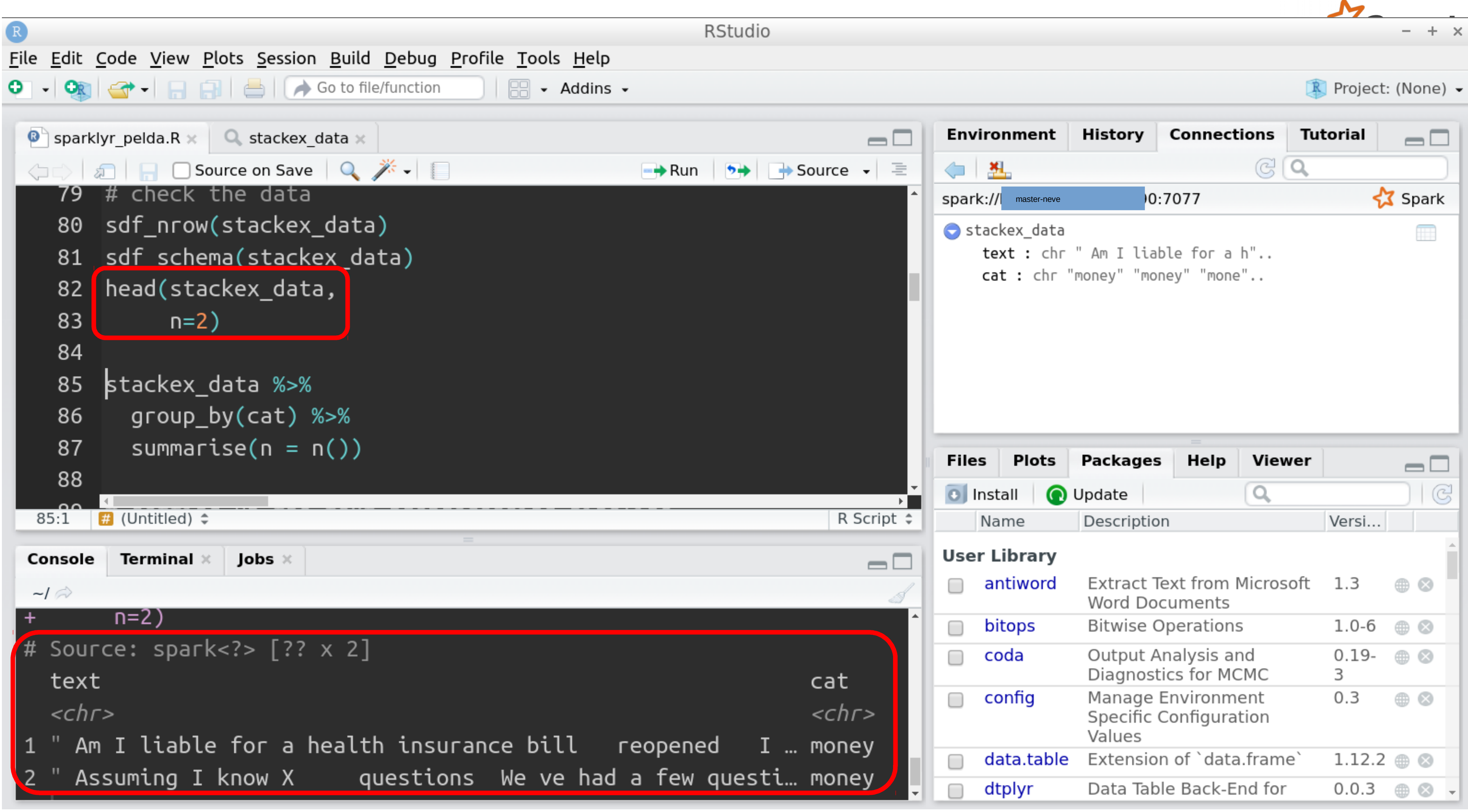

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
63
64 # loading data into the Spark context from the HDFS is the bes
65 stackex_data <- spark_read_csv(sc,
66 "stackex_data",
67 "~/MTA_PTI/ELKH_kepzes_2021_maj
68 header = TRUE,
69 infer_schema = FALSE,
70 delimiter = ";",
71 memory = FALSE) # the default i
72
73:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
+ memory = FALSE) # the default is tr
ue, meaning the table will be automatically cached, but the table mi
ght be too big, or we might just need a few columns, so let's set th
is to false and cache the table ourselves, if we need to
>
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
Data
conf List of 19
sc List of 13
stackex_… List of 2
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword Extract Text from Microsoft Word Documents 1.3
bitops Bitwise Operations 1.0-6
coda Output Analysis and Diagnostics for MCMC 0.19-3
config Manage Environment Specific Configuration Values 0.3
data.table Extension of `data.frame` 1.12.2
dtplyr Data Table Back-End for 0.0.3

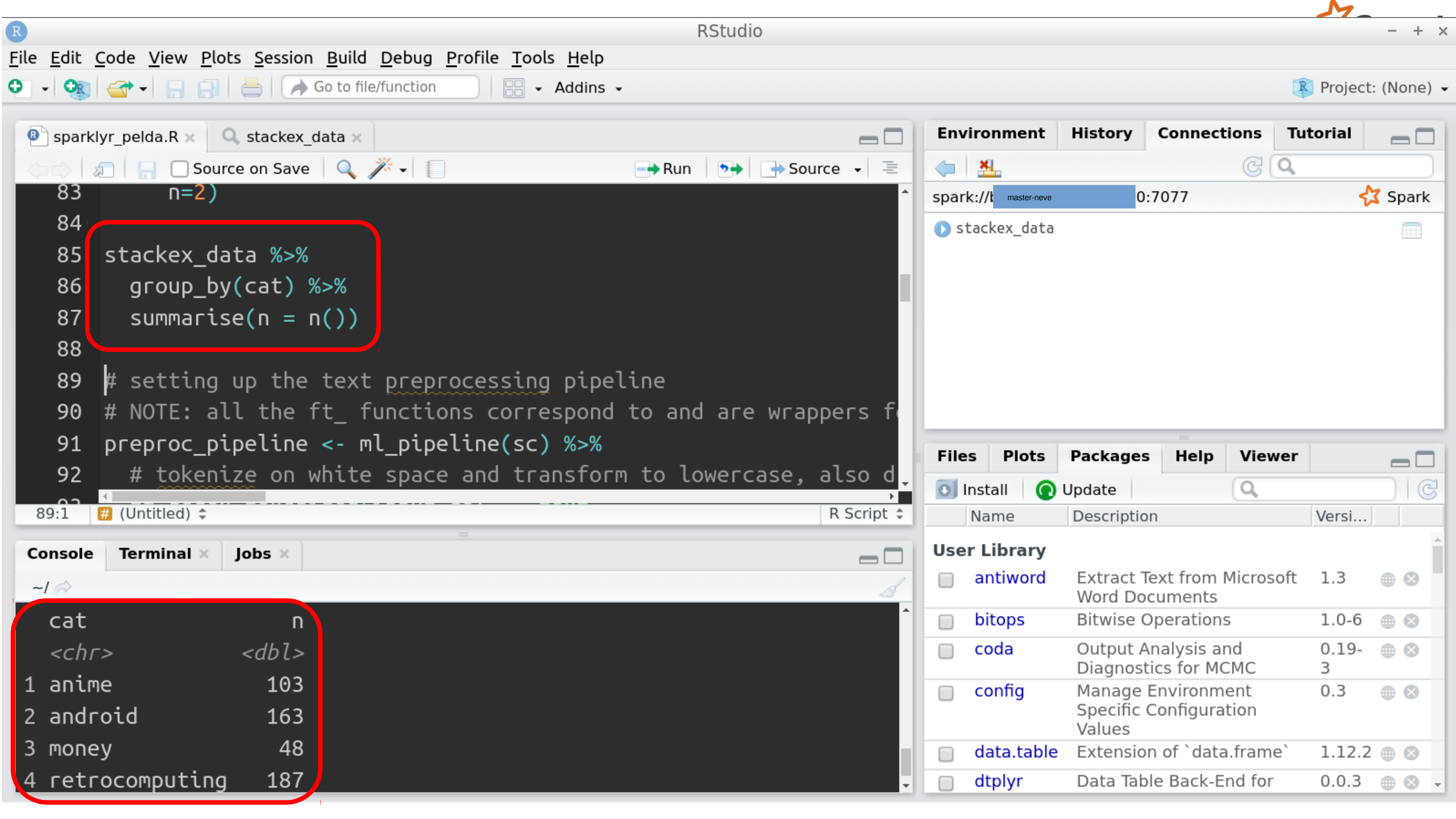

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Show Attributes
Name Type Value
stackex_data NULL Pairlist of length 0
(No selection)
Console Terminal Jobs
~/
+ memory = FALSE) # the default is tr
ue, meaning the table will be automatically cached, but the table mi
ght be too big, or we might just need a few columns, so let's set th
is to false and cache the table ourselves, if we need to
> View(stackex_data)
>
Environment History Connections Tutorial
Import Dataset
List
Global Environment
Data
conf List of 19
sc List of 13
stackex_… List of 2
Files Plots Packages Help Viewer
Install Update
Name Description Versi…
User Library
antiword Extract Text from Microsoft Word Documents 1.3
bitops Bitwise Operations 1.0-6
coda Output Analysis and Diagnostics for MCMC 0.19-3
config Manage Environment Specific Configuration Values 0.3
data.table Extension of `data.frame` 1.12.2
dtplyr Data Table Back-End for 0.0.3


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Show Attributes
Name
Type
Value
stackex_data
NULL
Pairlist of length 0
stackex_data
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
stackex_… List of 2
src:List of 1
..$ con:List of 13
.. ..$ master : chr "spark://bland…
.. ..$ method : chr "shell"
.. ..$ app_name : chr "sparklyr-te…
Console
Terminal
Jobs
~/
+ memory = FALSE) # the default is tr
ue, meaning the table will be automatically cached, but the table mi
ght be too big, or we might just need a few columns, so let's set th
is to false and cache the table ourselves, if we need to
> View(stackex_data)
>
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Show Attributes
Name
Type
Value
stackex_data
NULL
Pairlist of length 0
stackex_data
Environment History Connections Tutorial
spark://
master-neve
0:7077
Spark
stackex_data
text : chr " Am I liable for a h"..
cat : chr "money" "money" "mone"..
Files Plots Packages Help Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3
Console Terminal Jobs
~/
+ memory = FALSE) # the default is tr
ue, meaning the table will be automatically cached, but the table mi
ght be too big, or we might just need a few columns, so let's set th
is to false and cache the table ourselves, if we need to
> View(stackex_data)
>


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
75
76 # you can also use the copy_to command (see https://cran.r-pro
77 # it is important to keep track of where our different data el
78
79 # check the data
80 sdf_nrow(stackex_data)
81 sdf_schema(stackex_data)
82 head(stackex_data,
83 n=2)
84
81:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
ght be too big, or we might just need a few columns, so let's set th
is to false and cache the table ourselves, if we need to
> # check the data
> sdf_nrow(stackex_data)
[1] 754
>
Environment
History
Connections
Tutorial
Spark
Log
SQL
Help
spark://
master-neve
0:7077
Spark
stackex_data
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
79 # check the data
80 sdf_nrow(stackex_data)
81 sdf_schema(stackex_data)
82 head(stackex_data,
83 n=2)
84
85 stackex_data %>%
86 group_by(cat) %>%
87 summarise(n = n())
88
85:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
+ n=2)
# Source: spark<?> [?? x 2]
text cat
<chr> <chr>
1 " Am I liable for a health insurance bill reopened I … money
2 " Assuming I know X questions We ve had a few questi… money
Environment
History
Connections
Tutorial
spark://
master-neve
0:7077
Spark
stackex_data
text : chr " Am I liable for a h"..
cat : chr "money" "money" "mone"..
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
83 n=2)
84
85 stackex_data %>%
86 group_by(cat) %>%
87 summarise(n = n())
88
89 # setting up the text preprocessing pipeline
90 # NOTE: all the ft_ functions correspond to and are wrappers f
91 preproc_pipeline <- ml_pipeline(sc) %>%
92 # tokenize on white space and transform to lowercase, also d
89:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
cat n
<chr> <dbl>
1 anime 103
2 android 163
3 money 48
4 retrocomputing 187
Environment
History
Connections
Tutorial
spark://l
master-neve
0:7077
Spark
stackex_data
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Versi...
User Library
antiword
Extract Text from Microsoft Word Documents
1.3
bitops
Bitwise Operations
1.0-6
coda
Output Analysis and Diagnostics for MCMC
0.19-3
config
Manage Environment Specific Configuration Values
0.3
data.table
Extension of `data.frame`
1.12.2
dtplyr
Data Table Back-End for
0.0.3

Spark Master at spar × | sparklyr-test-poltext × | +

Not secure | 192.168.0.16:4040/jobs/

» Reading list

Spark 2.4.3 | Jobs | Stages | Storage | Environment | Executors | SQL | **sparklyr-test-poltext** application UI

Spark Jobs (?)

User: [redacted]
Total Uptime: 4.9 min
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 11, only showing 10

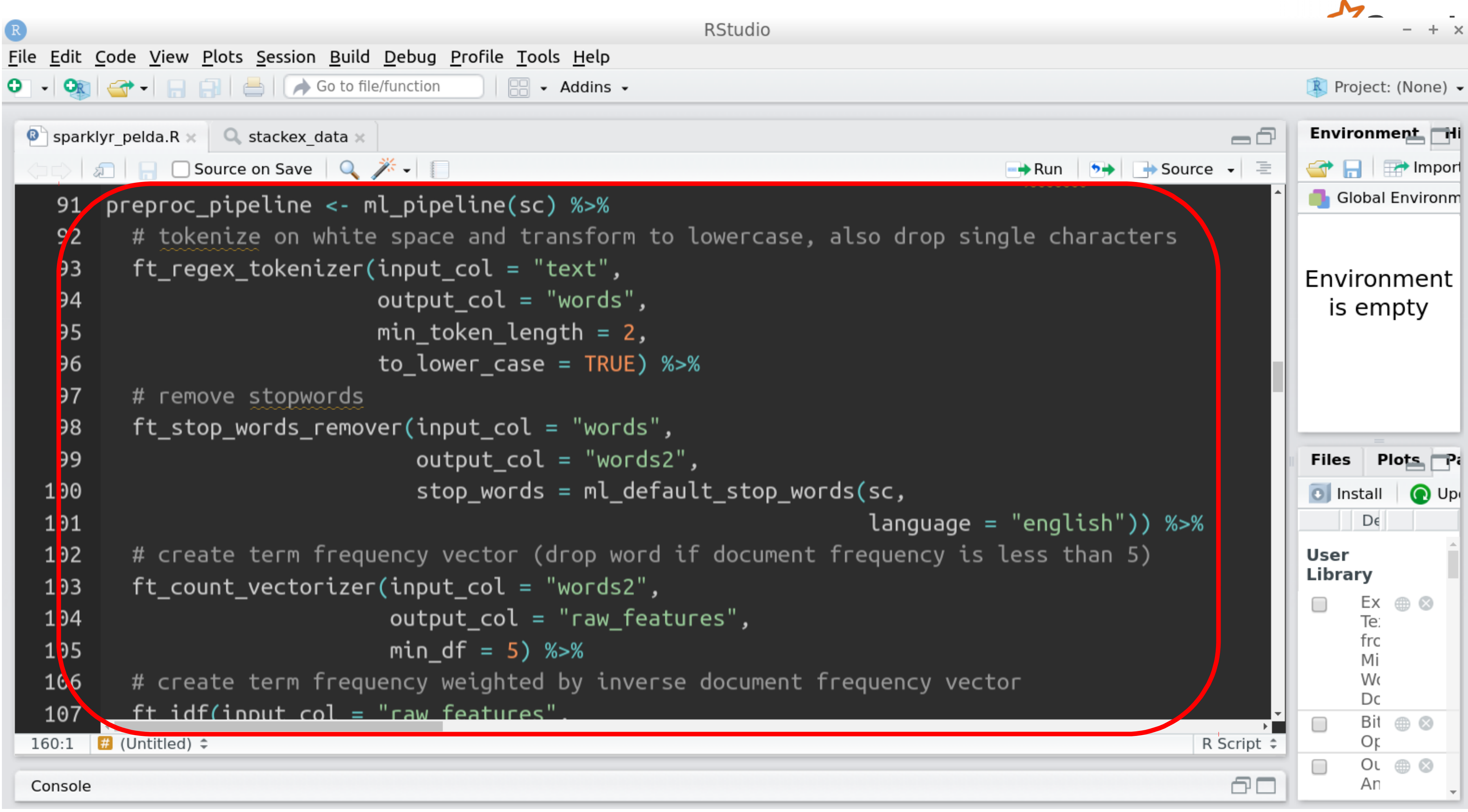
▸ Event Timeline

▾ Active Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
10	collect at utils.scala:204 collect at utils.scala:204 (kill)	2021/05/24 16:52:29	5 s	1/2	70/78 (5 running)

▾ Completed Jobs (11, only showing 10)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
9	collect at utils.scala:204 collect at utils.scala:204	2021/05/24 16:52:21	7 s	1/1 (1 skipped)	100/100 (4 failed) (3 skipped)


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
sparklyr_pelda.R
stackex_data
Source on Save
Run
Source
preproc_pipeline <- ml_pipeline(sc) %>%
# tokenize on white space and transform to lowercase, also drop single characters
ft_regex_tokenizer(input_col = "text",
output_col = "words",
min_token_length = 2,
to_lower_case = TRUE) %>%
# remove stopwords
ft_stop_words_remover(input_col = "words",
output_col = "words2",
stop_words = ml_default_stop_words(sc,
language = "english")) %>%
# create term frequency vector (drop word if document frequency is less than 5)
ft_count_vectorizer(input_col = "words2",
output_col = "raw_features",
min_df = 5) %>%
# create term frequency weighted by inverse document frequency vector
ft_idf(input_col = "raw_features",
160:1
(Untitled)
R Script
Console
Environment
Import
Global Environment
Environment is empty
Files
Plots
Install
User Library

```
                      output_col = "raw_features",
                      min_df = 5) %>%
  # create term frequency weighted by inverse document frequency vector
  ft_idf(input_col = "raw_features",
         output_col = "features")

# call pipeline to transform the data
stackex_data <- ml_fit_and_transform(preproc_pipeline,
                                     stackex_data)

# check the data again
sdf_nrow(stackex_data)
sdf_schema(stackex_data)
head(stackex_data,
     n=2)

```

Spark ? ... !

```r
# NOTE: all the ml_ functions correspond to and are wrappers for Spark MLlib functions, ju
nb_pipeline <- ml_pipeline(sc) %>%
  # have to change the text labels to numeric for the classifier to be able to handle it
  ft_string_indexer(input_col = "cat",
                    output_col = "label") %>%
  # add the naive bayes classifier to the pipeline
  ml_naive_bayes()

# parameter grid for parameter tuning (this is very small and simple on purpose for the tu
# the name of the list element that corresponds to a pipeline element (in this case "naive
# you can add more than one parameter to tune at the same time, note how the parameter nam
param_grid <- list(
  naive_bayes = list(
    smoothing = c(1.0,
                  0.5)
  )
)
```

```r
# set up cross validator for parameter tuning
# note: the default metric for ml_multiclass_classification_evaluator is f1, we could als
# see https://cran.r-project.org/web/packages/sparklyr/sparklyr.pdf
cv <- ml_cross_validator(sc,
                         estimator = nb_pipeline,
                         estimator_param_maps = param_grid,
                         evaluator = ml_multiclass_classification_evaluator(sc),
                         num_folds = 3,
                         parallelism = 4 # this is based on the number of cores/cpus/vcpus
)

# create train-test split
split_data <- sdf_random_split(stackex_data,
                               training = 0.7,
                               test = 0.3)
```

Apache Spark? ... !

```r
# remember previously we had ml_fit_and_transform, but here we just need our model, so we
cv_model <- ml_fit(cv,
                   split_data$training)

# check the metrics of the parameter tuning results
ml_validation_metrics(cv_model)

# apply model to the test set (now we call only ml_transform, since our model is already f
test_with_pred <- ml_transform(cv_model,
                               split_data$test)

# check evaluation metrics for the test set
ml_multiclass_classification_evaluator(test_with_pred,
                                       label_col = "label",
                                       prediction_col = "prediction",
                                       metric_name = "f1")
```

APACHE Spark ? ... !

```r
# since this can be too much to handle, let's just collect the columns we need for now
test_with_pred_redux <- select(test_with_pred,
                               cat,
                               pred_cat)
test_df <- sdf_collect(test_with_pred_redux)

install.packages("tidyr") # we need this for the spread function
library(tidyr)

# now we can use the spread function on our dataframe in the R session to create the con
test_df %>%
  group_by(cat,
           pred_cat) %>%
  summarise(n = n()) %>%
  spread(key = cat,
         value = n)
```

APACHE Spark ? … !

```r
# (just as it is important to keep track of where our data is, it is also important to kee
test_with_pred %>%
  group_by(cat,
           pred_cat) %>%   # this happens in the Spark context
  summarise(n = n()) %>%   # this happens in the Spark context
  collect() %>%            # this is where we move our data from the Spark context to the R
  spread(key = cat,
         value = n)        # this happens in the R session

# hmmm, maybe we should check our metrics without the gardening category
test_with_pred %>%
  filter(label != 0) %>%
  ml_multiclass_classification_evaluator(label_col = "label",
                                         prediction_col = "prediction",
                                         metric_name = "f1")

# let's try the linear SVC model
```

Thank you for your attention!

You can find all code and explanations at:
https://github.com/zkpti/poltext2019-sparktutorial