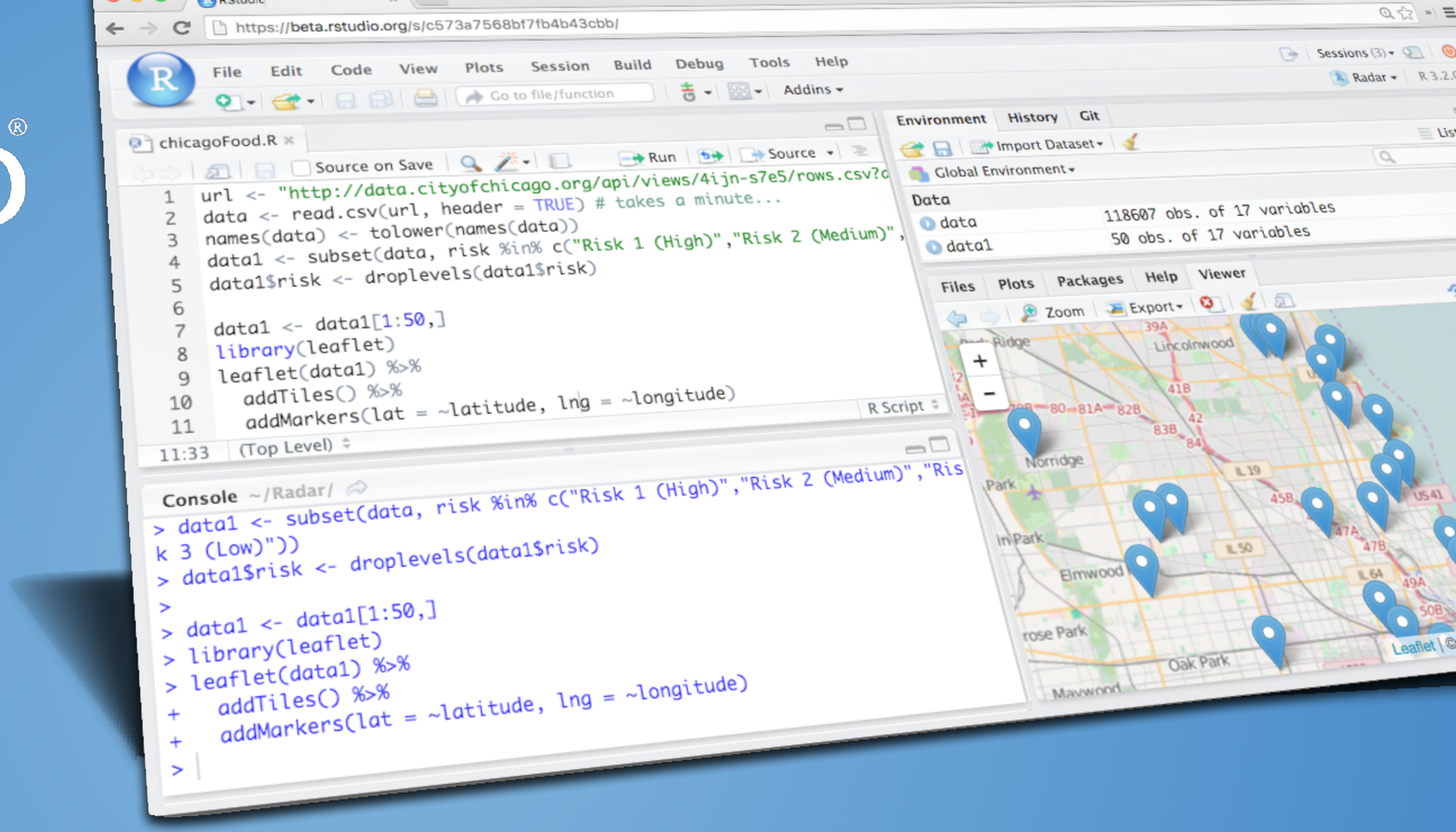
RStudio®



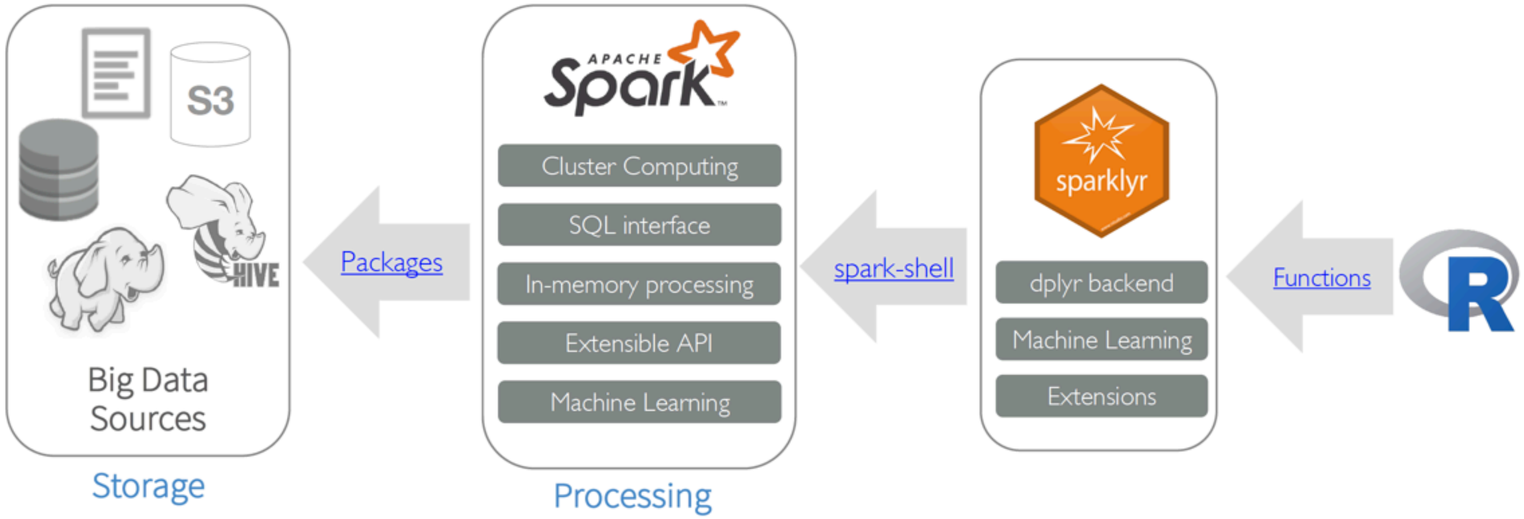
Extending Spark using sparklyr and R

Extending Spark using sparklyr and R

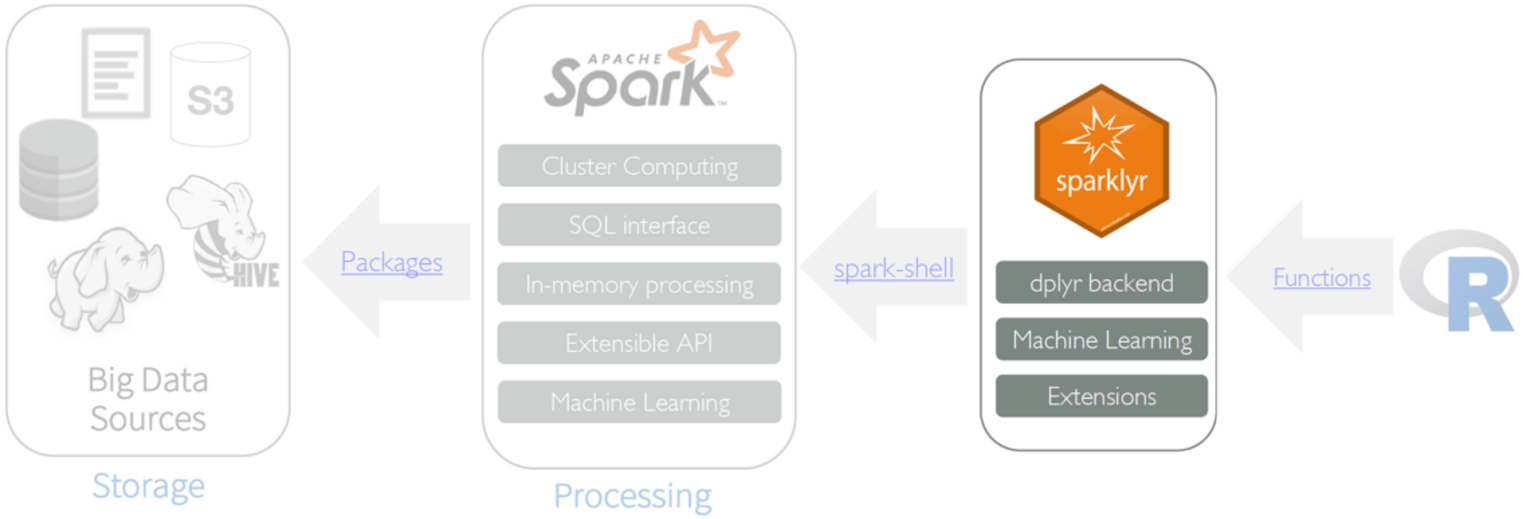
- ✓ Extensions Overview
- ✓ Using Extensions
- ✓ Extensions using R Code
- ✓ Extensions using Scala Code
- ✓ Extensions as R Packages

Overview

OVERVIEW


S3
HIVE
Big Data Sources
Storage
Packages
APACHE Spark
Cluster Computing
SQL interface
In-memory processing
Extensible API
Machine Learning
Processing
spark-shell
sparklyr
dplyr backend
Machine Learning
Extensions
Functions
R

OVERVIEW



Using Extensions

```
1
2  library(rsparkling)
3
```

Using R Code

```
1
2  spark_dataframe(iris_tbl) %>%
3    invoke("isLocal")
4
```

Using Scala Code

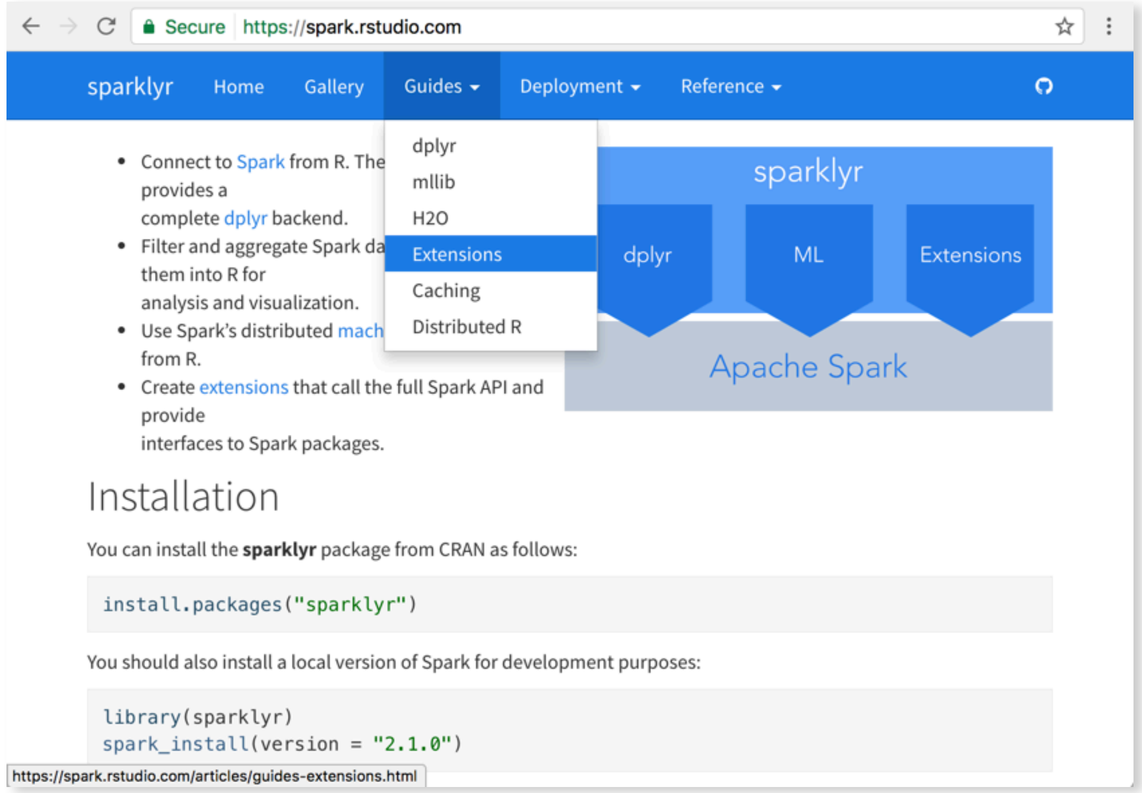
```
1
2  package SparkHello
3
4  object HelloWorld {
5    def hello() : String = {
6      "Hello, world! - From Scala"
7    }
8  }
9
```

Creating R Packages

```
1
2  install.packages(rsparkling)
3
```

Extensions

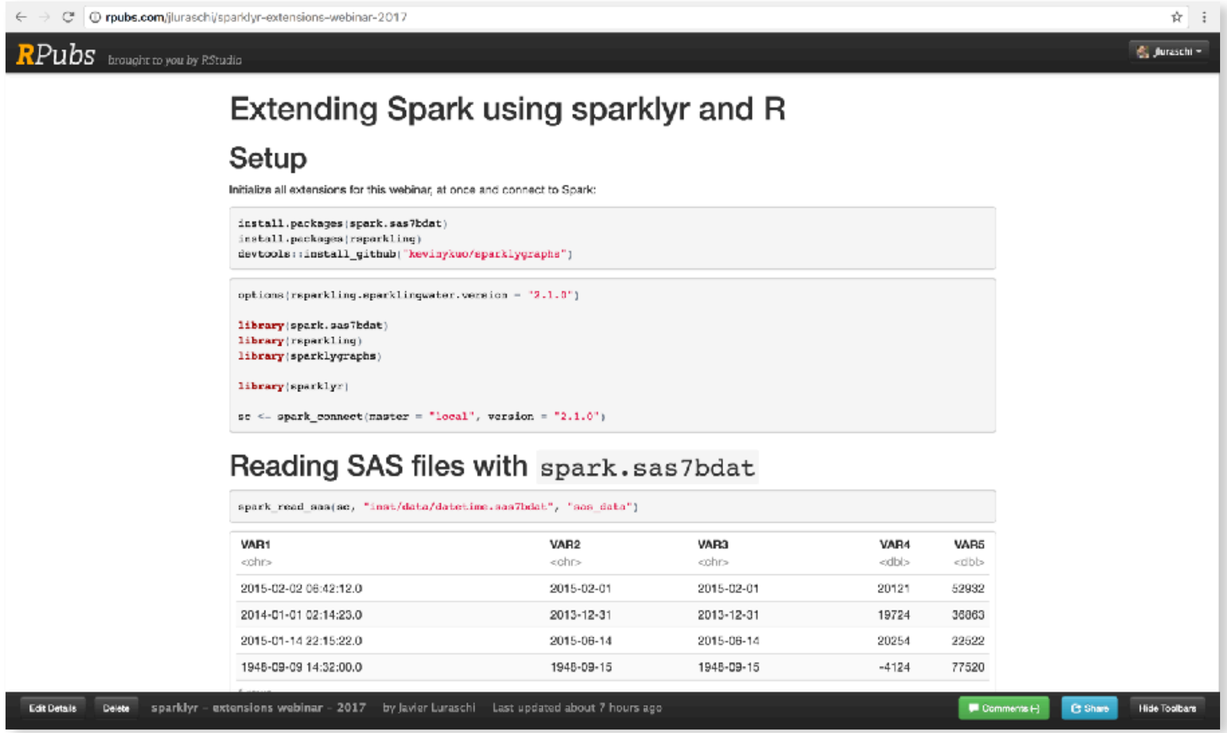
Using Extensions

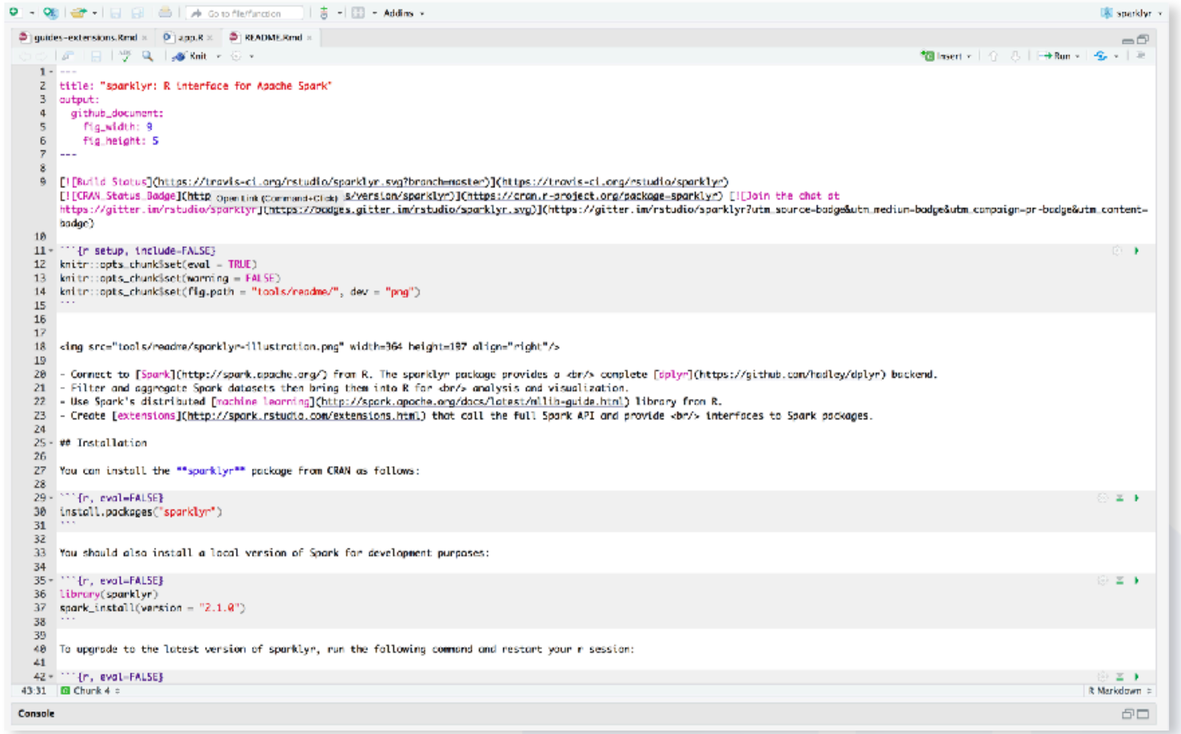


Package	Description
spark.sas7bdat	Read in SAS data in parallel into Apache Spark.
rsparkling	Extension for using H2O machine learning algorithms against Spark Data Frames.
sparkhello	Simple example of including a custom JAR file within an extension package.
rddlist	Implements some methods of an R list as a Spark RDD (resilient distributed dataset).
sparkwarc	Load WARC files into Apache Spark with sparklyr.
sparkavro	Load Avro data into Spark with sparklyr. It is a wrapper of spark-avro
crassy	Connect to Cassandra with sparklyr using the `Spark-Cassandra-Connector`.
sparklygraphs	R interface for GraphFrames which aims to provide the functionality of GraphX.

```
1
2  install.packages("spark.sas7bdat")
3  library(spark.sas7bdat)
4
```

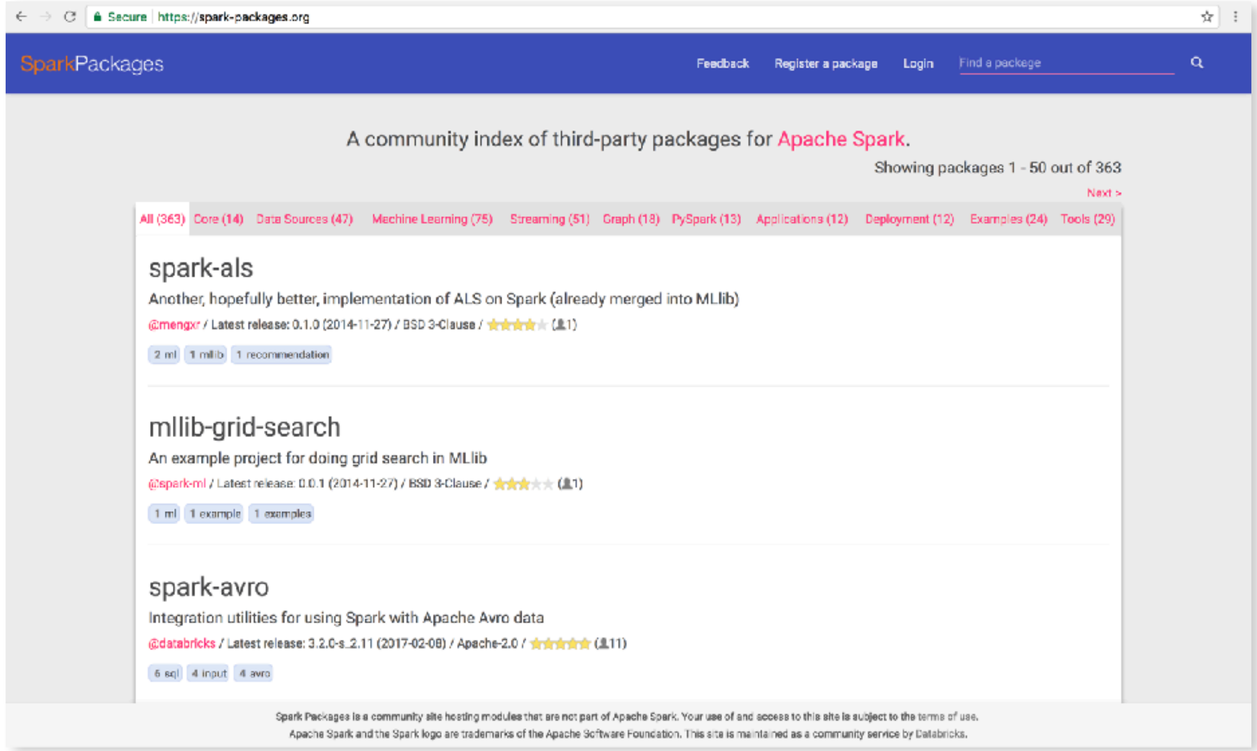
Using Extensions





R Code

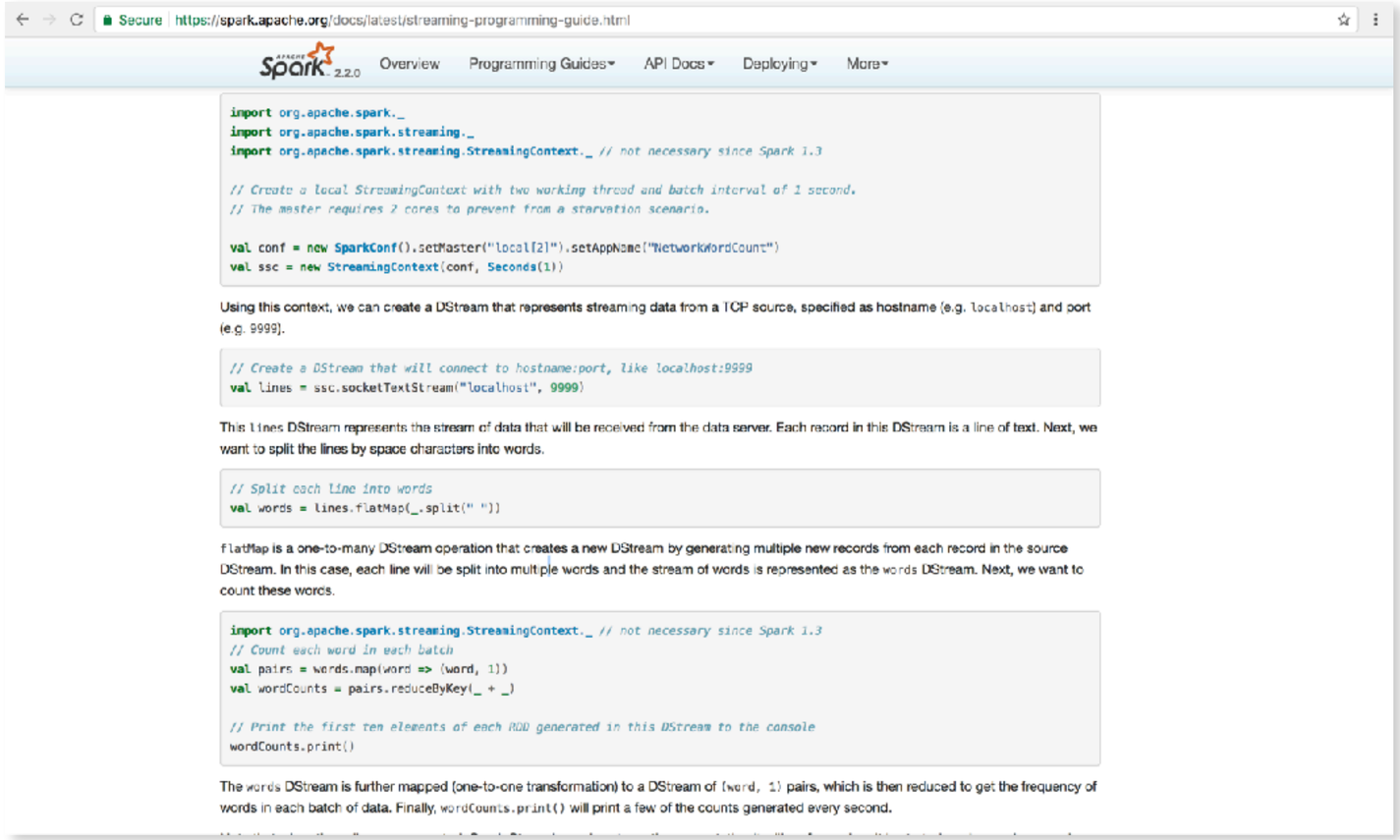
SPARK PACKAGES




```
1
2  library(sparklyr)
3
4  config <- spark_config()
5  config$sparklyr.defaultPackages <- c(
6    "datastax:spark-cassandra-connector:2.0.0-RC1-s_2.11"
7  )
8
9  sc <- spark_connect(master = "local", config = config)
10
```

```
1
2  spark_read_source(
3    sc,
4    "emp",
5    "org.apache.spark.sql.cassandra",
6    list(keyspace = "dev", table = "emp")
7  )
8
```

INVOKE() - INTRO




```
1
2  ssc <- invoke_new(
3    sc,
4    "org.apache.spark.streaming.StreamingContext",
5    spark_context(sc),
6    invoke_new(
7      sc,
8      "org.apache.spark.streaming.Duration",
9      10000L
10   )
11 )
12
```

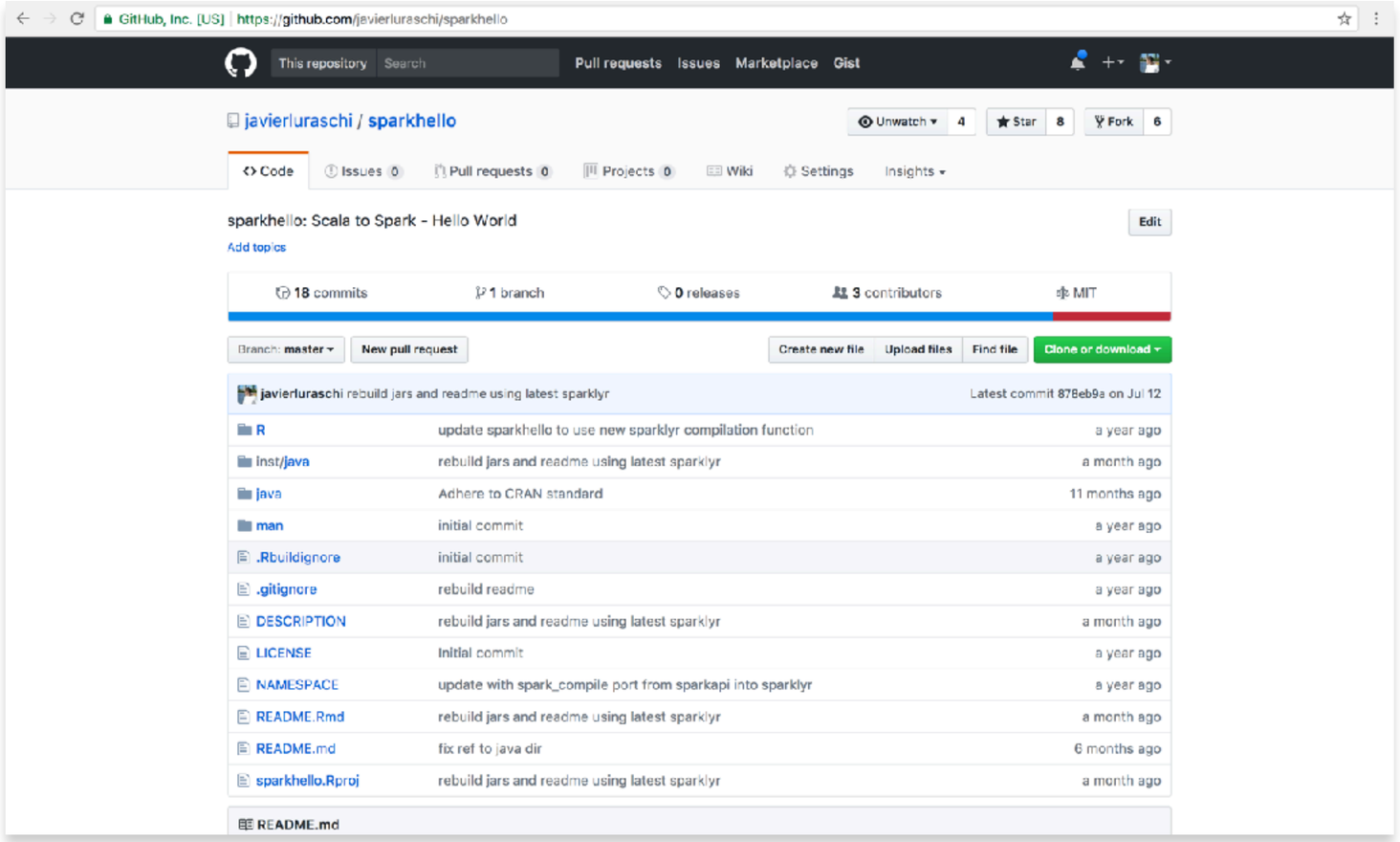
INVOKE() - MAPPINGS

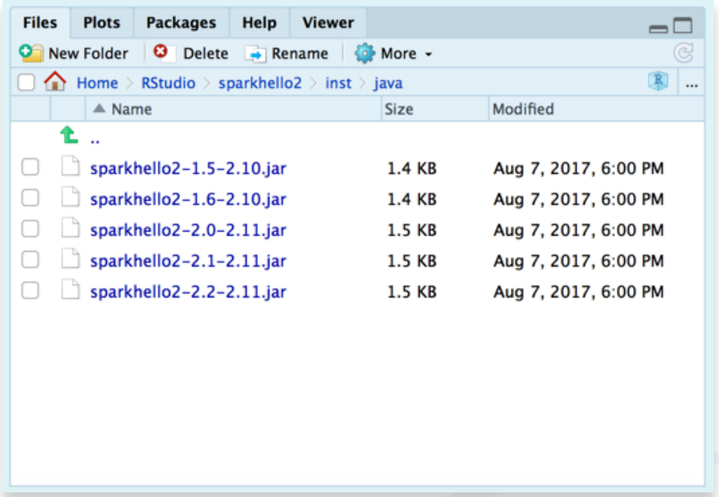
From R	Scala	To R
NULL	void	NULL
integer	Int	integer
character	String	character
logical	Boolean	logical
double	Double	double
numeric	Double	double
	Float	double
	Decimal	double
	Long	double
raw	Array[Byte]	raw
Date	Date	Date
POSIXlt	Time	
POSIXct	Time	POSIXct
list	Array[T]	list
environment	Map[String, T]	
jobj	Object	jobj

```
1
2 ensure_scalar_integer()
3 ensure_scalar_double()
4 ensure_scalar_boolean()
5 ensure_scalar_character()
6
```

Scala Code

COMPILE_PACKAGE_JARS()

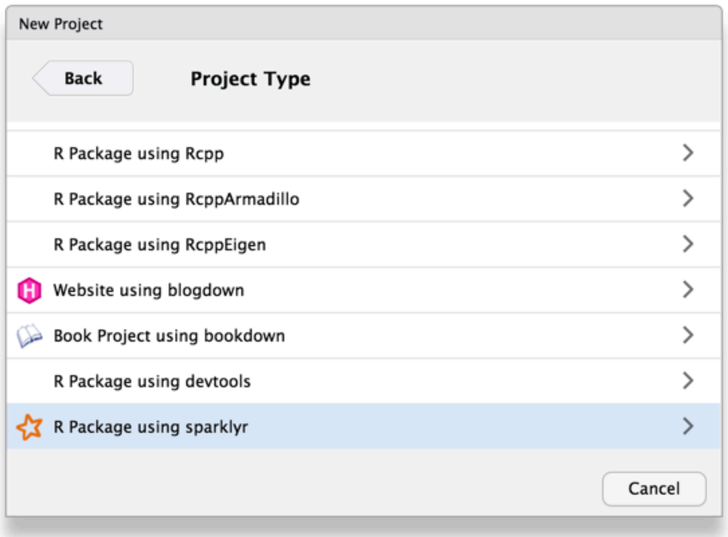


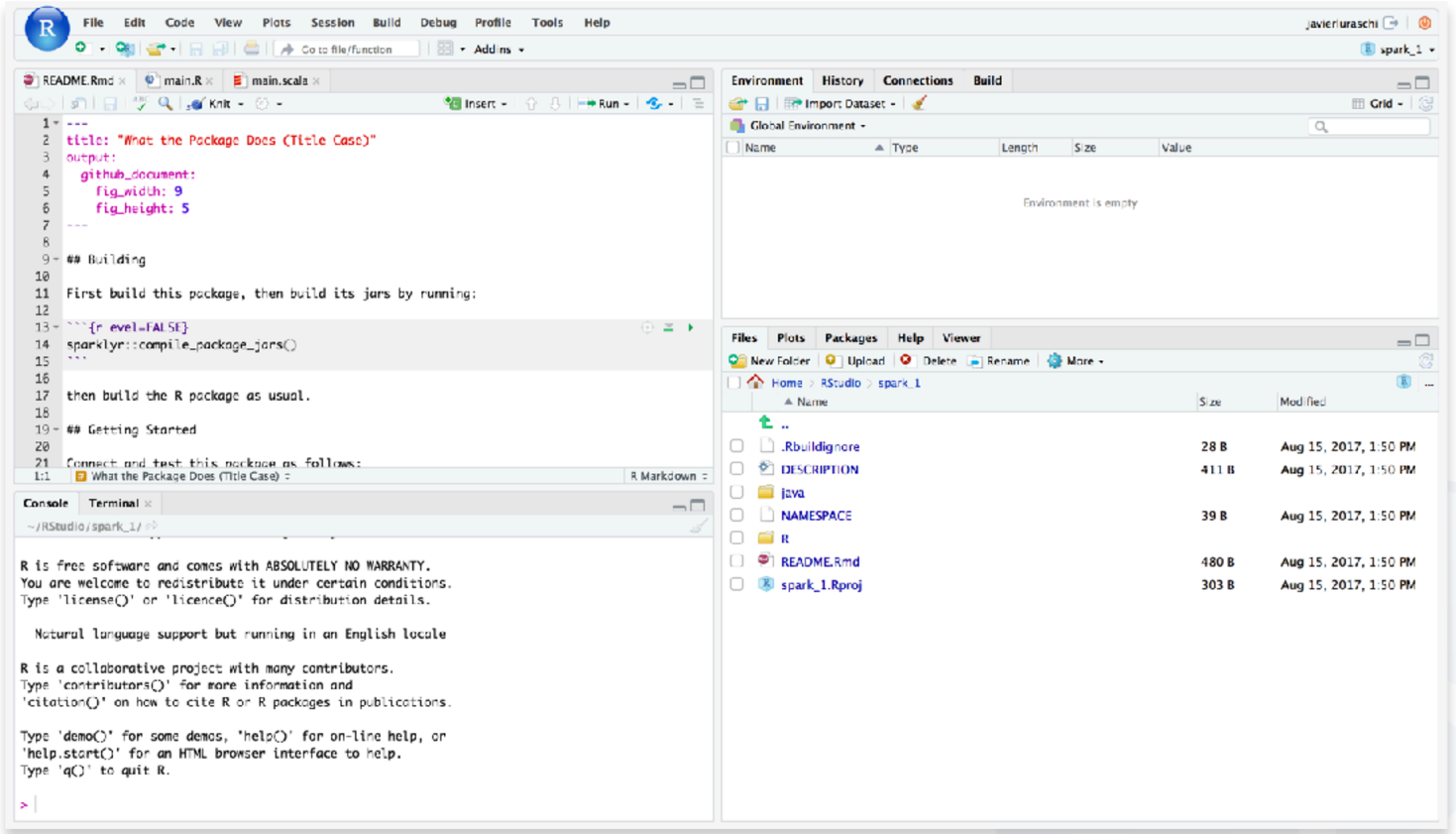


R Packages

R PACKAGES

```
devtools::install_github("rstudio/sparklyr")
```





RStudio New Project integration requires sparklyr devel or sparklyr 0.7 (unreleased)

Useful Websites

RStudio and Spark:
spark.rstudio.com

RStudio Webinar Resources:
rstudio.com/resources/webinars/

Sparklyr Github:
github.com/rstudio/sparklyr

Spark Documentation:
spark.apache.org/docs/2.1.0/api/scala/

Questions

Open Source & Free

Desktop: http://www.rstudio.com/products/rstudio/download/
RStudio Server: http://www.rstudio.com/products/rstudio/download-server/
Shiny Server: http://www.rstudio.com/products/shiny/download-server/
shinyapps.io beta: https://www.shinyapps.io/admin/#/signup

45 Day Evaluation of Pro Products

RStudio Server Pro: http://www.rstudio.com/products/rstudio-server-pro/evaluation/
Shiny Server Pro: http://www.rstudio.com/products/shiny-server-pro/evaluation/

PLEASE STAY IN TOUCH


Blog - http://rviews.rstudio.com/


Blog - http://blog.rstudio.org/


Twitter - @rstudio #rstats http://twitter.com/rstudio/


GitHub - https://github.com/rstudio/


LinkedIn - https://linkedin.com/company/rstudio-inc


Facebook - https://www.facebook.com/pages/RStudio-inc


Google+ - https://plus.google.com/110704473211154995841/posts