Using Spark with Shiny and R Markdown

Jeff Allen, RStudio

@trestleJeff

WHAT IS SPARK?

- Open-source Apache computing engine

- Bigger-than-memory data, low-latency distributed computing

- Can integrate with the Hadoop ecosystem

- Built-in machine learning

Introducing…

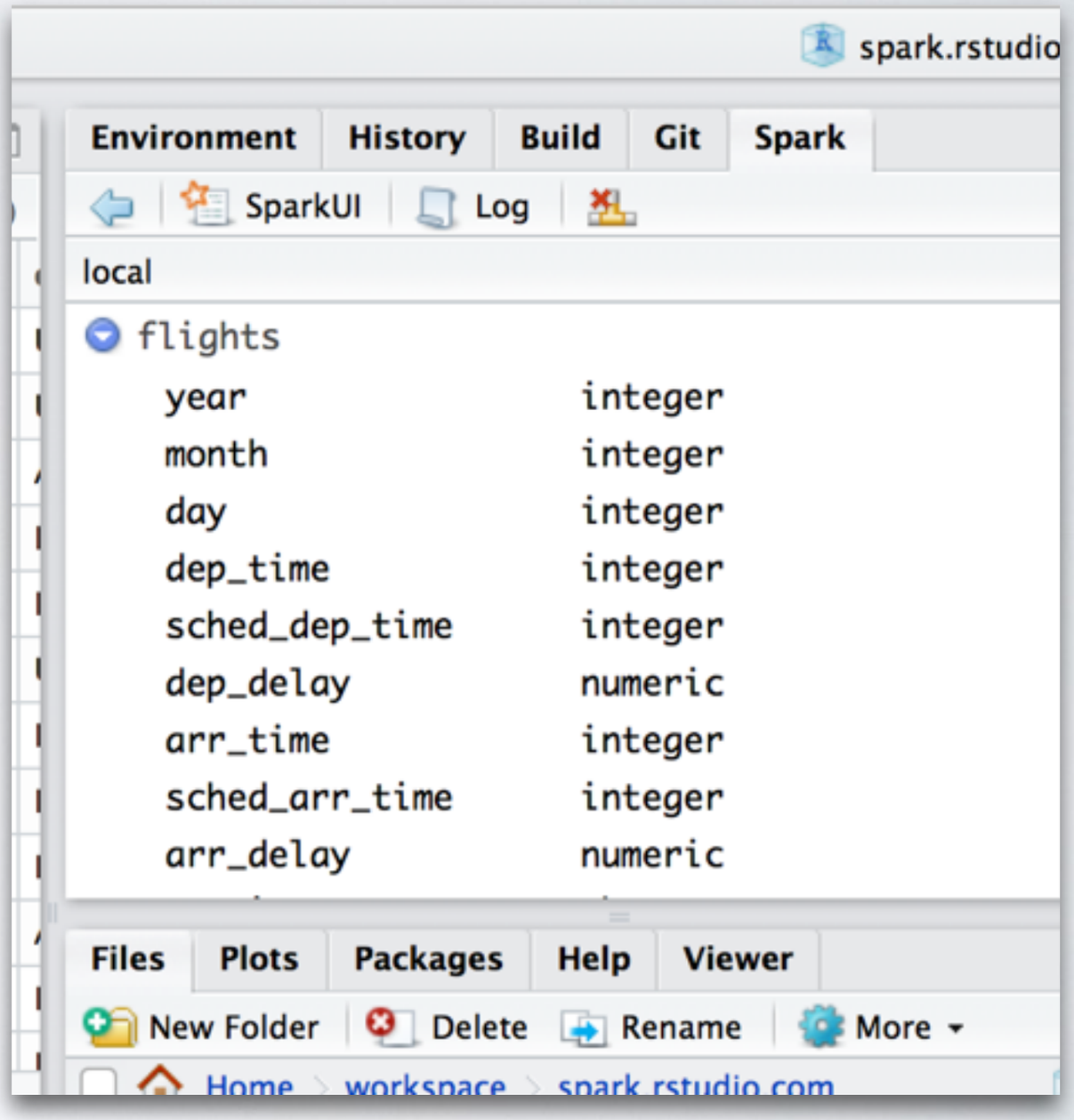
sparklyr

"SPARRK-lee ARR"

http://spark.rstudio.com

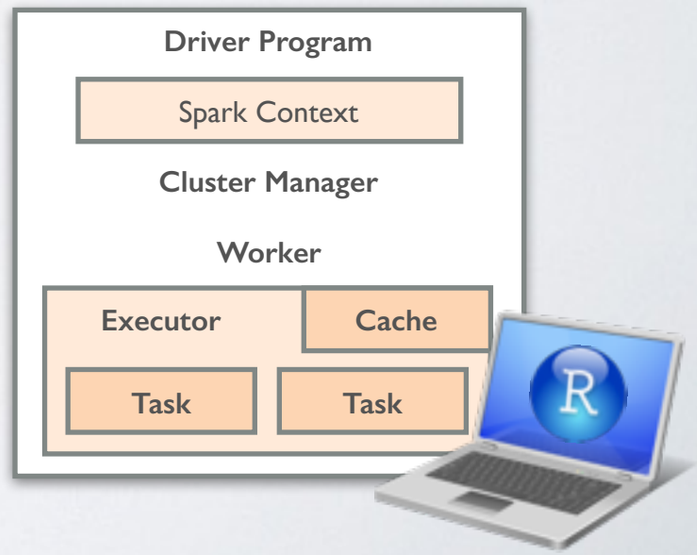
sparklyr

- New open-source R package from RStudio
- Complete dplyr back-end for Spark
- Integrated with the RStudio IDE
- Extensible foundation for Spark + R



KICK THE TIRES

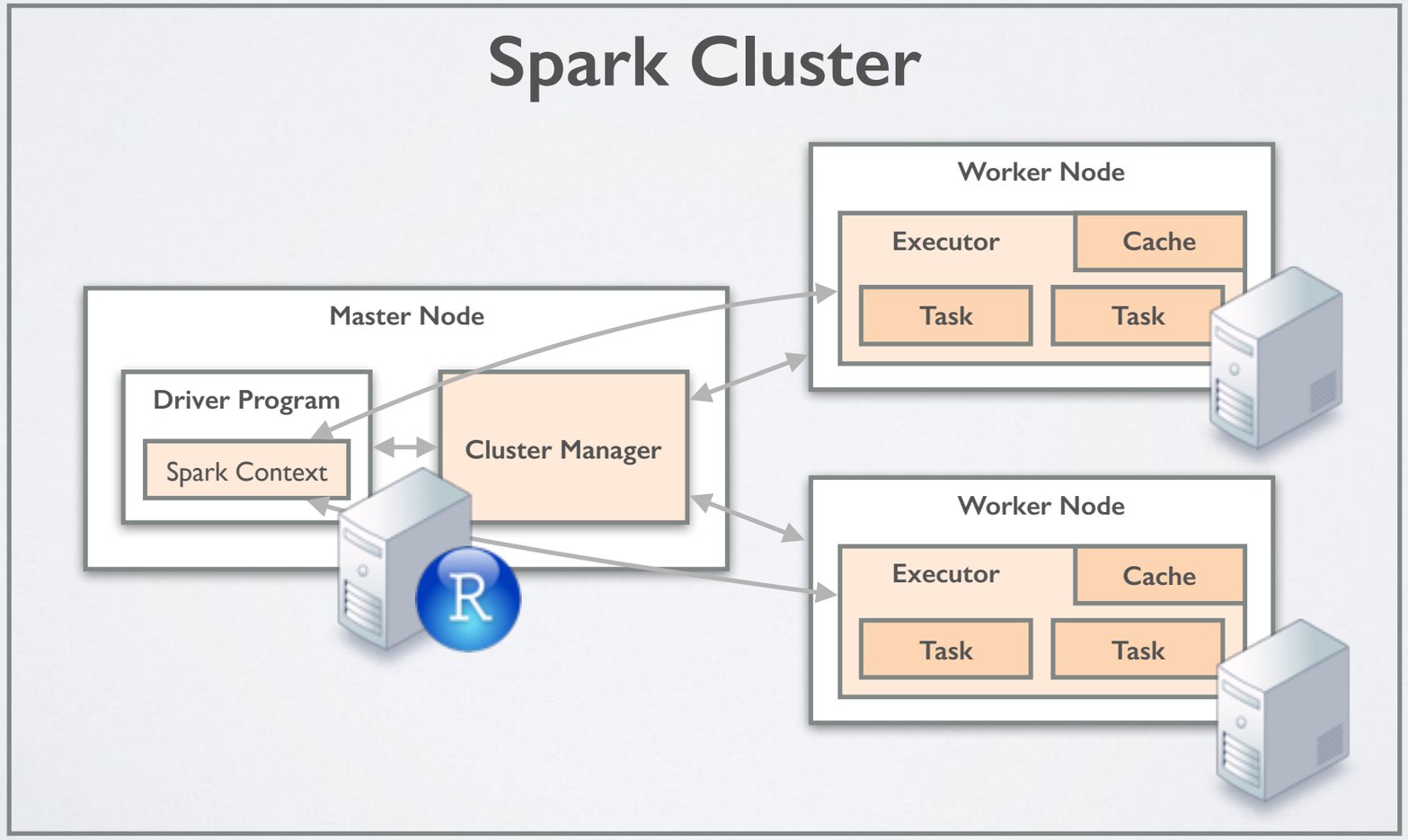
```
library(sparklyr)
spark_install()
sc <- spark_connect(“local")
my_tbl <- copy_to(sc, iris)
```



RUN IN PRODUCTION

Use RStudio Server on the Spark cluster master node

```
> spark_connect("spark://spark.company.org:7077")
> my_tbl <- tbl(sc, "tblname")
```



USE DPLYR TO WRITE SPARK SQL

```
library(dplyr)

# use standard verbs to filter and aggregate
select(
  filter(my_tbl, Petal_Width < 0.3),
  Petal_Length, Petal_Width
)

# use magrittr pipes for a cleaner syntax
my_tbl %>%
  filter(Petal_Width < 0.3) %>%
  select(Petal_Length, Petal_Width)
```


print.R
DESCRIPTION
html_document.Rmd
handle_click.R
vega.R
Source on Save
Run
Source
25 }
26
27 data_props <- combine_data_props(x$marks)
28 data_ids <- names(data_props)
29 data_table <- x$data[data_ids]
30
31 # Collapse each list of scale objects into one scale object.
32 x <- collapse_scales(x)
33 scale_data_table <- scale_domain_data(x)
34
35 # Wrap each of the reactive data objects in another reactive which returns
36 # only the columns that are actually used, and adds any calculated columns
37 # that are used in the props.
38 data_table <- active_props(data_table, data_props)
39
40 # From an environment containing data_table objects, get static data for the
41 # specified ids.
16:19
(Top Level)
R Script
Console ~/r/ggvis/
Next
Continue
Stop
> ggvis(diamonds, x = ~price, y = ~color)
Guessing layer_histograms()
Guessing binwidth = 1
Called from: eval(expr, envir, enclos)
Browse[1]> n
debug at /Users/jmcphers/r/ggvis/R/vega.R#29: data_table <- x$data[data_ids]
Browse[2]>
Environment
History
Build
Git
as.vega.ggvis()
Values
data_ids "diamonds0/bin1/stack2"
data_props List of 1
dynamic FALSE
Traceback
as.vega.ggvis(x, FALSE) at vega.R:29
as.vega(x, FALSE) at vega.R:11
view_static(x, ...) at print.R:67
print.ggvis(c("list()", "list(diamonds0 = function () \nstatic_data)"
Files
Plots
Packages
Help
Viewer
count
price
0
20
40
60
80
100
120
0
2,000
4,000
6,000
8,000
10,000
12,000

DEMO

SPARKLYR FUNCTIONALITY

- Full dplyr back-end for Spark DataFrames
- R wrappers for all MLlib functions
- Easily leverage from R Markdown, Shiny, etc.
- IDE integration
- Windows, too!

RELATIONSHIP TO SPARKR

- Working together to establish a common extension API

- Some differences in approach:
  - CRAN distribution
  - dplyr compatibility

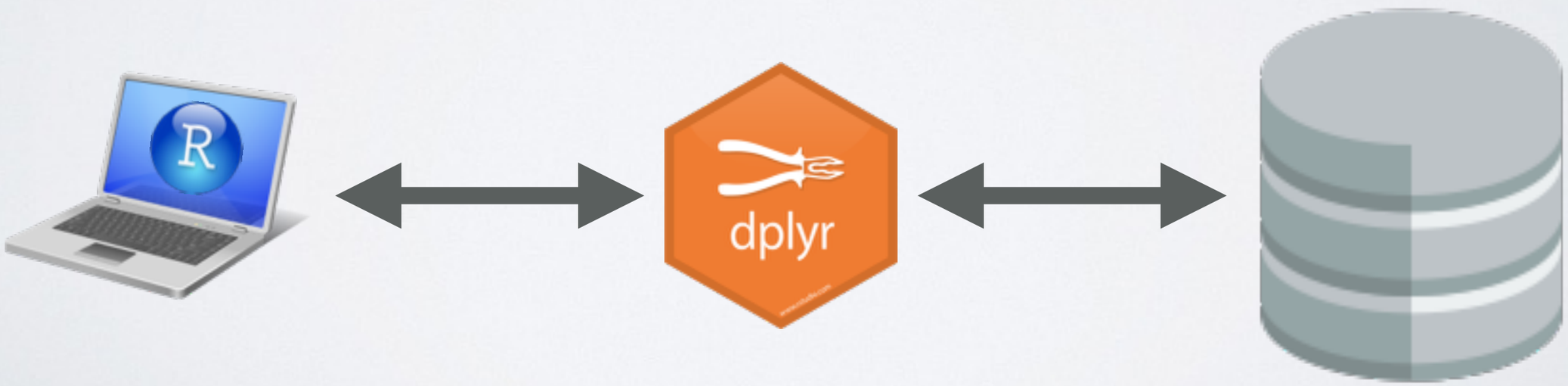
SPARKR DPLYR

Console ~/spark/R/pkg/

```
> library(SparkR)

Attaching package: 'SparkR'

The following objects are masked from 'package:dplyr':

    arrange, between, collect, contains, count, cume_dist,
    dense_rank, desc, distinct, explain, filter, first, group_by,
    intersect, lag, last, lead, mutate, n, n_distinct, ntile,
    percent_rank, rename, row_number, sample_frac, select, sql,
    summarize, union

The following objects are masked from 'package:stats':

    cov, filter, lag, na.omit, predict, sd, var, window

The following objects are masked from 'package:base':

    as.data.frame, colnames, colnames<-, drop, endsWith,
    intersect, rank, rbind, sample, startsWith, subset, summary,
    transform, union

>
```

APPROACHES

- Load a subset of data at a time
- Use dplyr to connect to external DB
- SQLite, PostgreSQL, MySQL, BigQuery, Redshift



APPROACHES

- dplyr + DB
- Rcpp
- parallel
- bigmemory

- RStudio Server in the Cloud
- Up to 2TB of RAM

- Rmpi
- Spark

QUESTIONS?

http://spark.rstudio.com