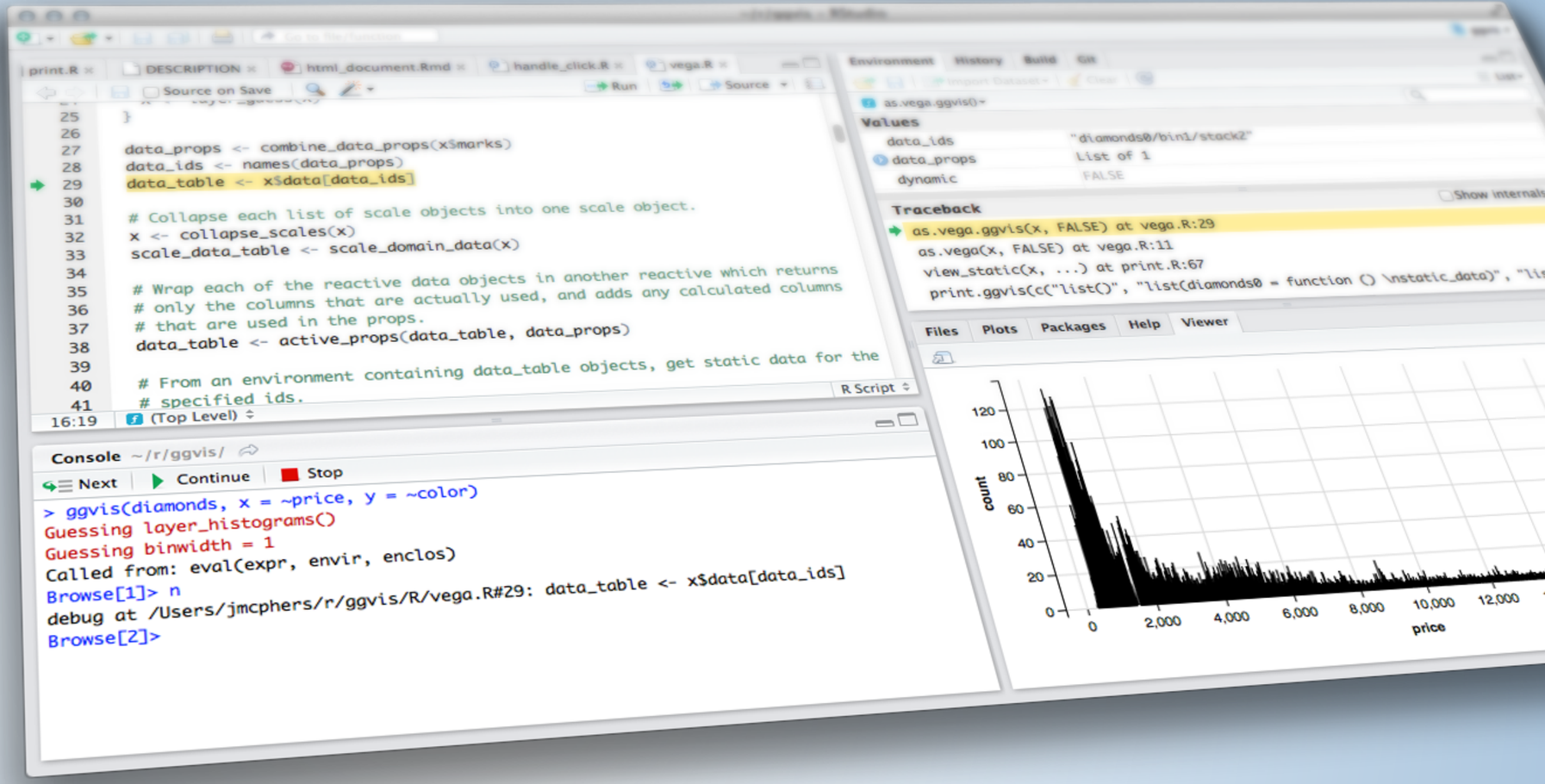


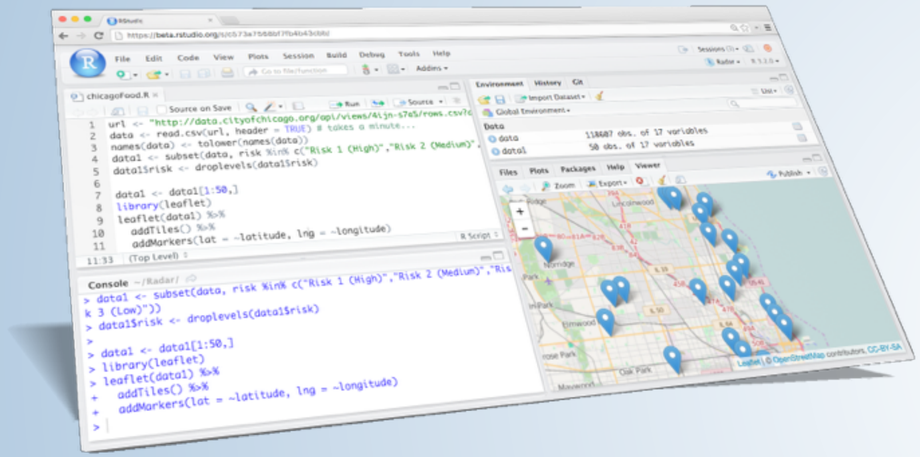
USING APACHE SPARK FROM R

WITH THE SPARKLYR PACKAGE

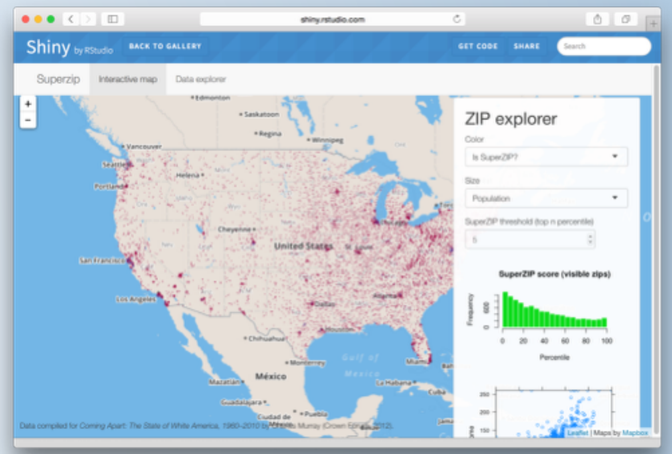
RStudio®

RStudio Products

RStudio IDE



Shiny

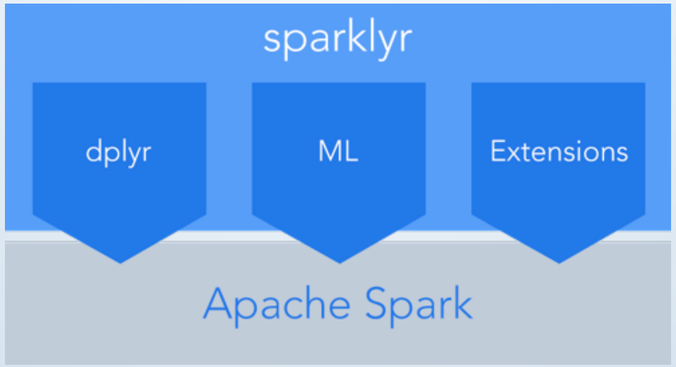


Packages

Introducing…

sparklyr

"SPARK-lee-ARR"



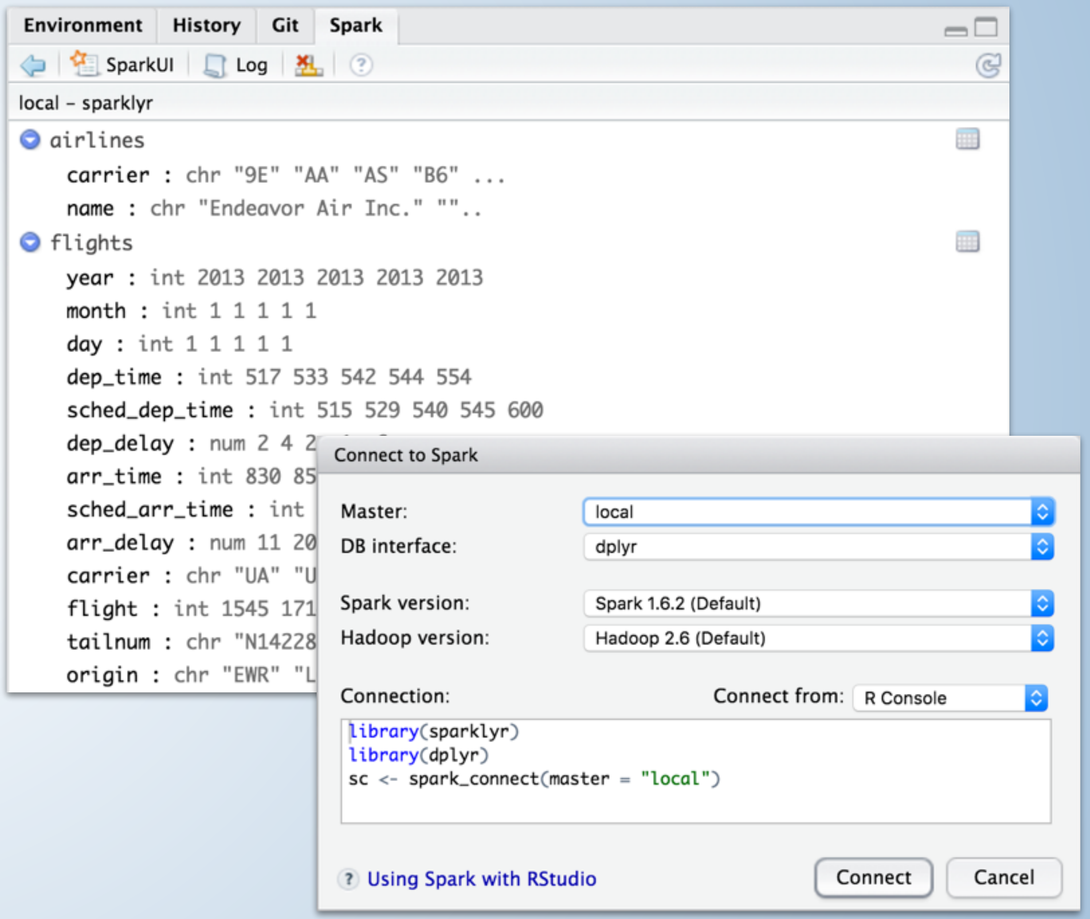
http://spark.rstudio.com/

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

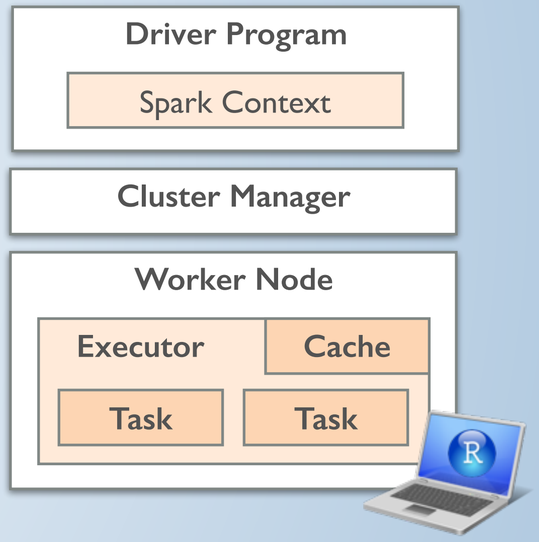
sparklyr

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



Local Mode

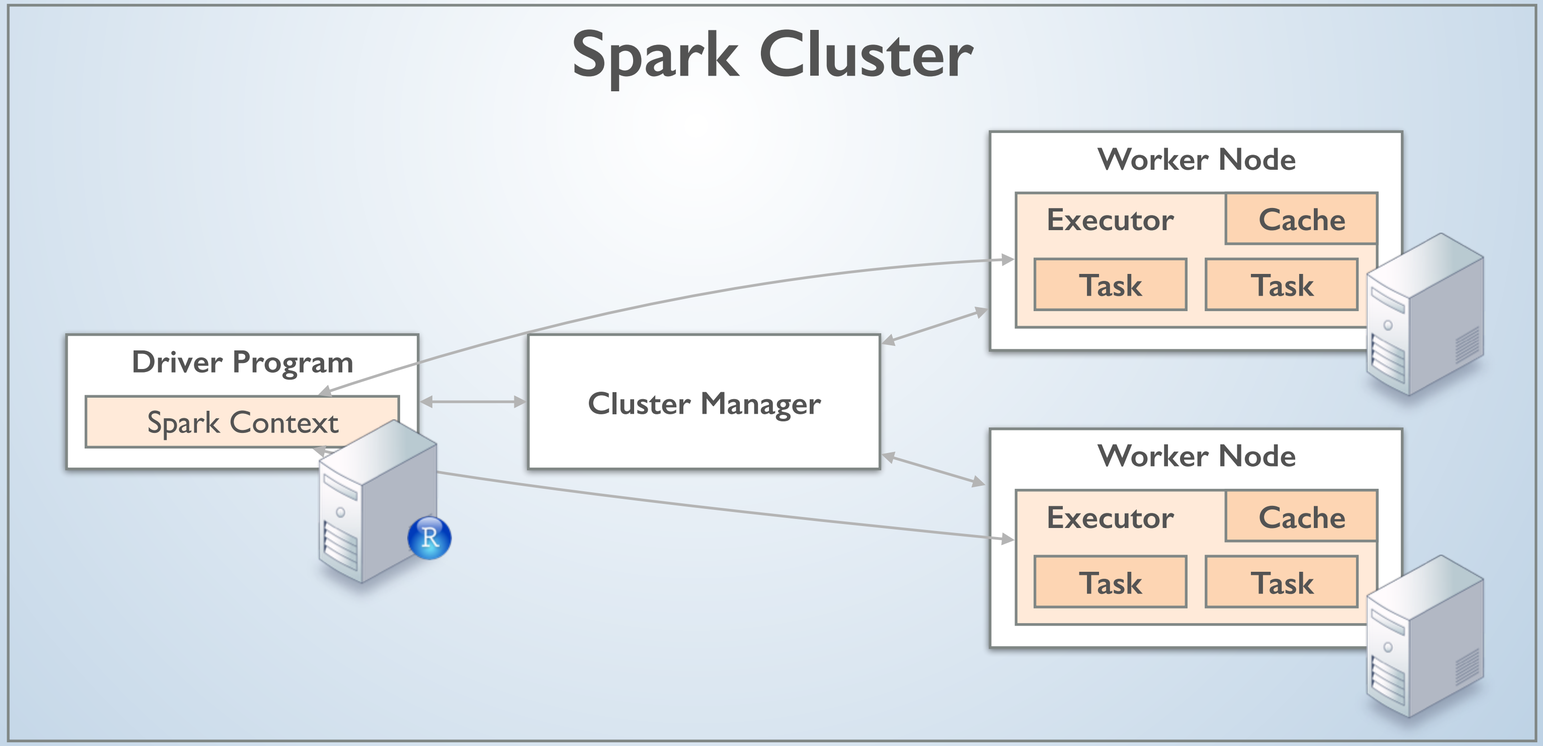
```
library(sparklyr)

spark_install()

sc <- spark_connect("local")

my_tbl <- copy_to(sc, iris)
```



Cluster Mode

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)
```

Demo

sparklyr functionality

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

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

- Rmpi
- Spark

Questions