

# R Studio®

sparklyr  
Introducing an  
R interface for Apache Spark

The screenshot shows the R Studio interface. In the top-left corner, there's a large white circle containing the blue R logo. The main area features a large title "sparklyr" followed by a subtitle "Introducing an R interface for Apache Spark". To the right of this text is a screenshot of the R Studio environment.

The R Studio window has a toolbar at the top with File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help menus. Below the toolbar is a menu bar with "chicagoFood.R x" selected. The code editor pane contains R code for reading a CSV file from a Chicago government API and creating a Leaflet map. The console pane shows the same R code being run. The right side of the interface includes a "Data" pane showing two datasets: "data" (118607 obs. of 17 variables) and "data1" (50 obs. of 17 variables). The bottom-right pane displays a map of Chicago with several blue location markers placed on it, representing data points from the "data1" dataset.

```
1 url <- "http://data.cityofchicago.org/api/views/4ijn-s7e5/rows.csv?c=1&q=1"
2 data <- read.csv(url, header = TRUE) # takes a minute...
3 names(data) <- tolower(names(data))
4 data1 <- subset(data, risk %in% c("Risk 1 (High)", "Risk 2 (Medium)", "Risk 3 (Low)"))
5 data1$risk <- droplevels(data1$risk)
6
7 data1 <- data1[1:50,]
8 library(leaflet)
9 leaflet(data1) %>%
10   addTiles() %>%
11   addMarkers(lat = ~latitude, lng = ~longitude)
```

```
> data1 <- subset(data, risk %in% c("Risk 1 (High)", "Risk 2 (Medium)", "Risk 3 (Low)"))
> data1$risk <- droplevels(data1$risk)
>
> data1 <- data1[1:50,]
> library(leaflet)
> leaflet(data1) %>%
+   addTiles() %>%
+   addMarkers(lat = ~latitude, lng = ~longitude)
```

# Edgar Ruiz

Edgar has a background in deploying enterprise reporting and Business Intelligence solutions. He has posted multiple articles and blog posts sharing analytics insights and server infrastructure for Data Science.



**Solutions Engineer**  
- Pass Christian, Mississippi

# SPARKLYR - AUTHORS



Javier Luraschi – author , creator



JJ Allaire – author



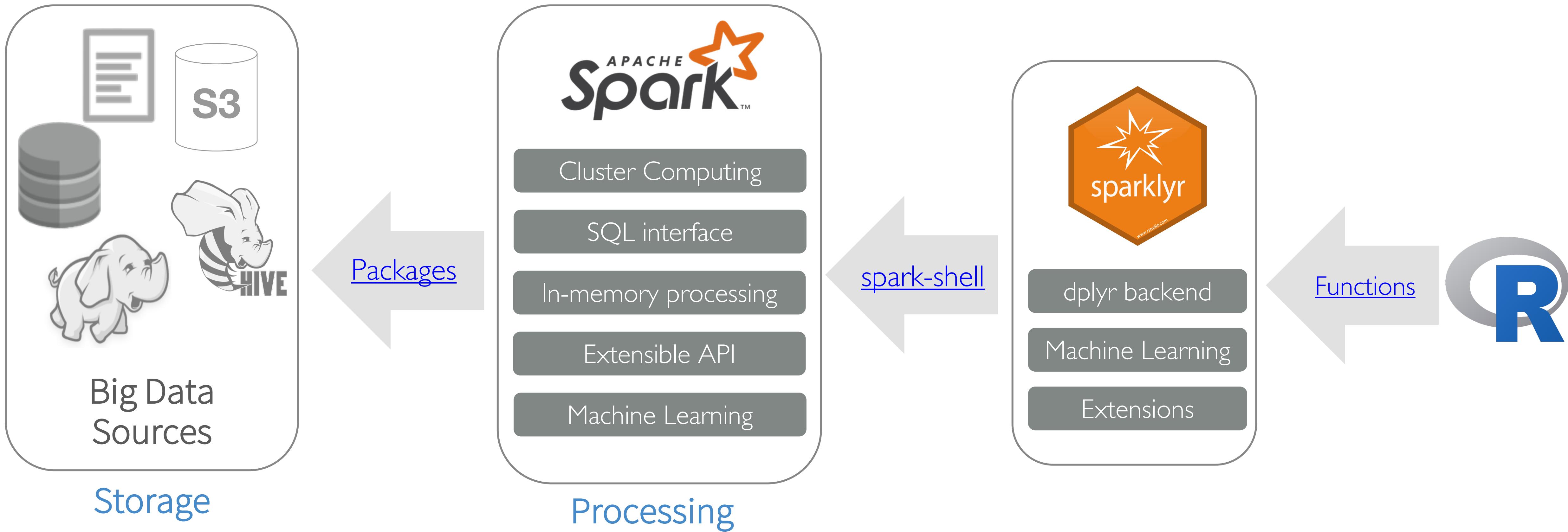
Kevin Ushey – author

# WHAT TO EXPECT

- New slides and illustrations
- Reproducible demo code
- Highlight some new features in sparklyr

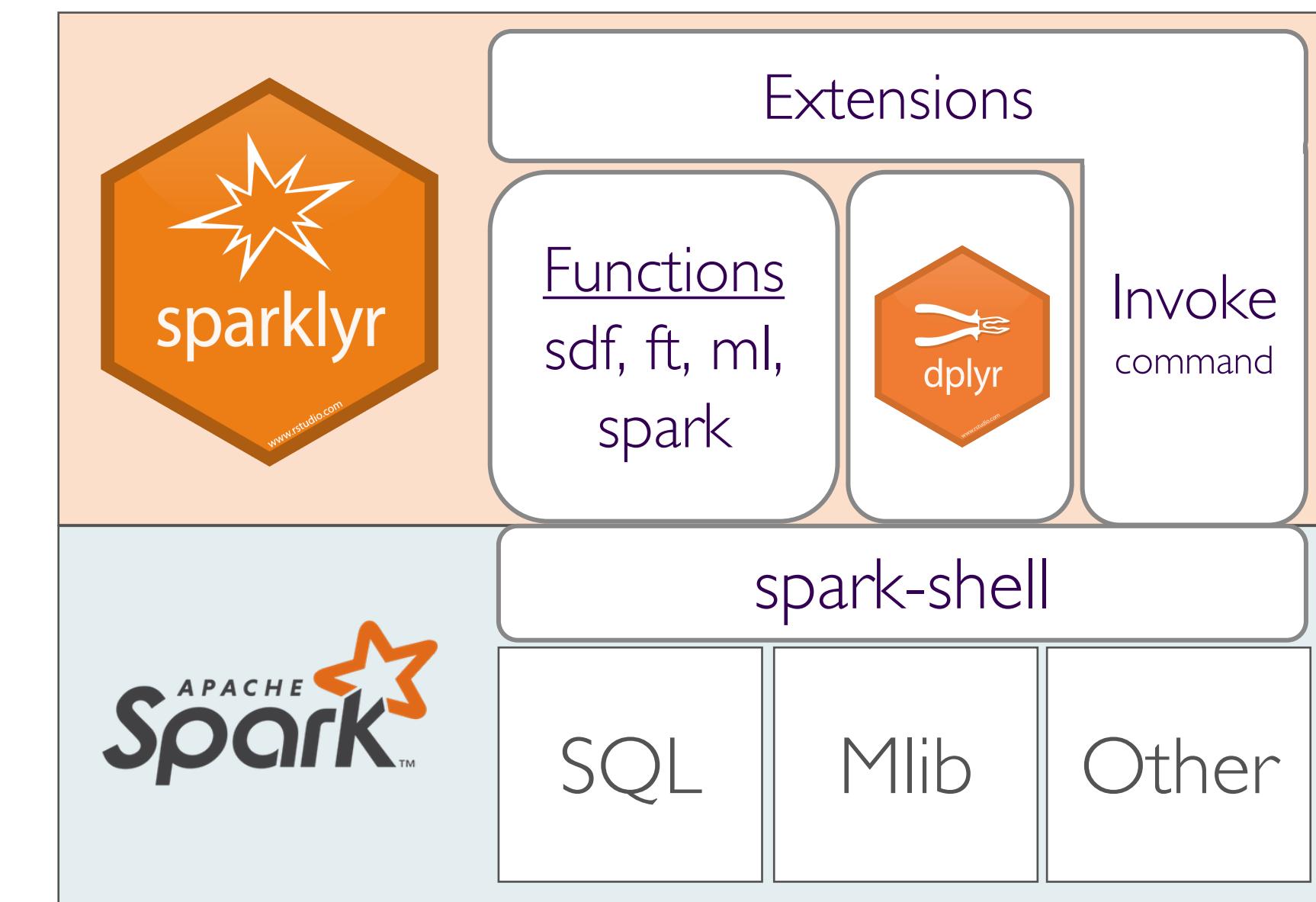
**Goal:** *Share something new and useful*

# SPARKLYR – AN R INTERFACE FOR SPARK



# SPARKLYR OPENS ACCESS TO THE SPARK API

**sparklyr** uses the spark-shell to gain access to Spark. In addition to providing SQL access, sparklyr also opens access to Spark API functions like Feature Transformers, Machine Learning models and others.



# DPLYR BACKEND



With **dplyr** as an interface to manipulating Spark DataFrames, you can:

- Select, filter, and aggregate data
- Use window functions (e.g. for sampling)
- Perform joins on DataFrames
- Collect data from Spark into R

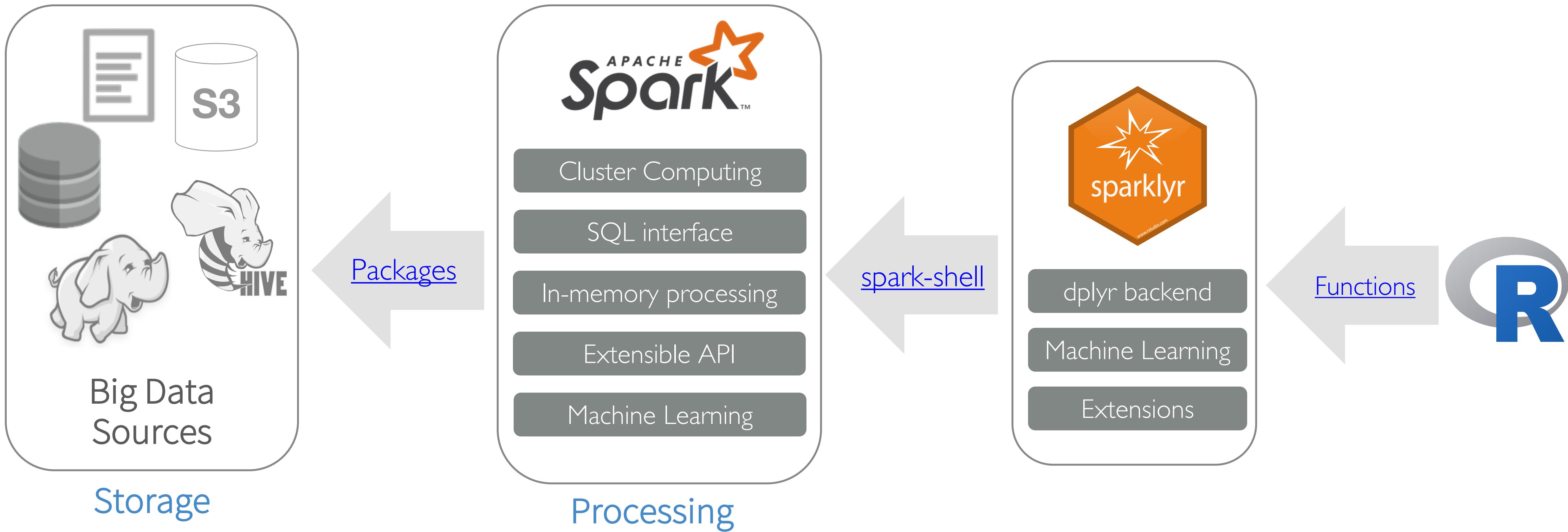
Under the hood, **sparklyr** translates the **dplyr** verbs and vector expressions into Spark SQL

```
tbl(sc, "flights") %>%  
  select(Origin)
```

Translates to

```
SELECT `Origin` AS `Origin`  
FROM `flights`
```

# SPARKLYR – AN R INTERFACE FOR SPARK



# READING DATA FROM SPARK

Option 1 – Read data live

Spark will retrieve data from the source every time a new request from R is sent



---

Option 2 – Cache data in Spark

Spark will retain data in-memory while the session is active



# DEMO



Review of sparklyr functions

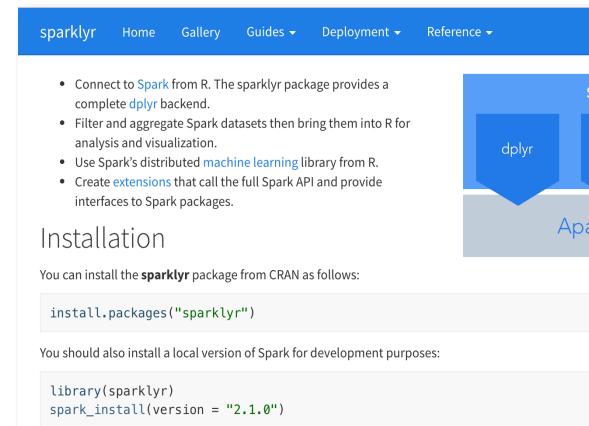
Use the Spark UI during analysis

Read & cache data in Spark

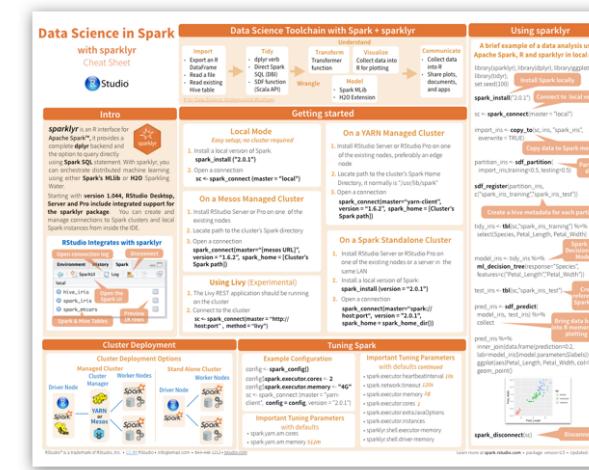
What's new in sparklyr

Tips n' tricks

# USEFUL LINKS



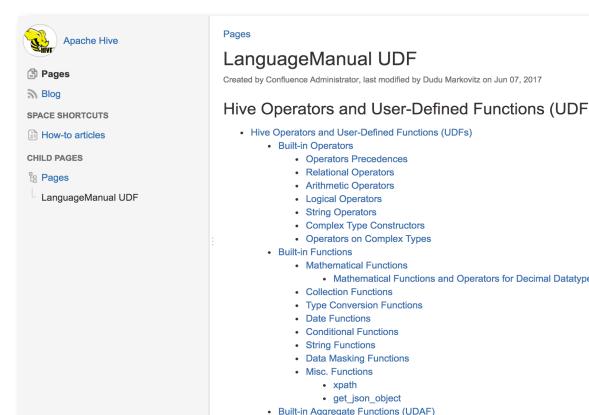
sparklyr's official website  
[spark.rstudio.com](http://spark.rstudio.com)



sparklyr, IDE, RMarkdown cheatsheets  
[www.rstudio.com/resources/cheatsheets](http://www.rstudio.com/resources/cheatsheets)



Spark documentation  
<https://spark.apache.org/docs/latest/>



Hive SQL UDF  
<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF>



## **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>