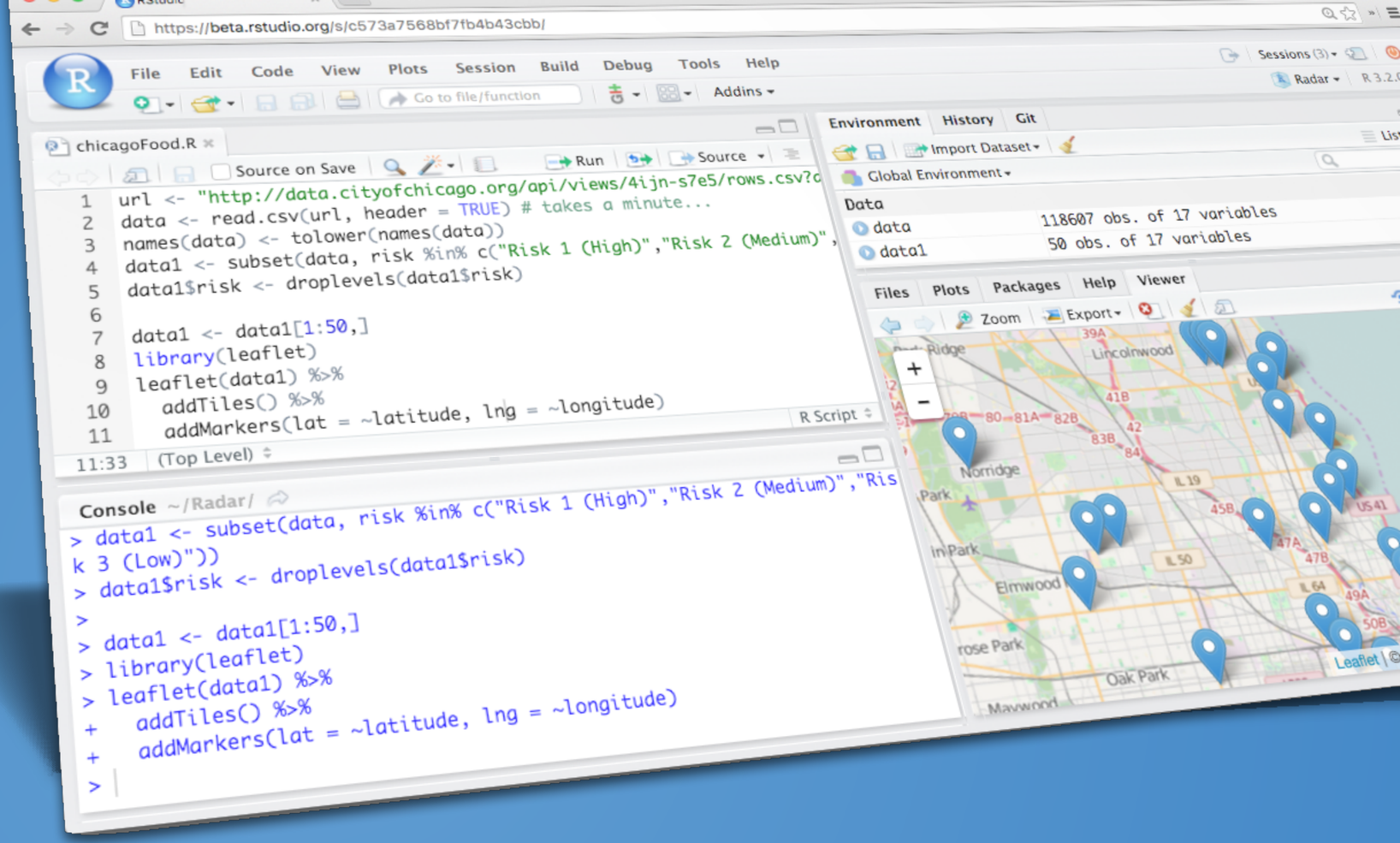
RStudio®



sparklyr

Introducing an R interface for Apache Spark

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

RStudio®

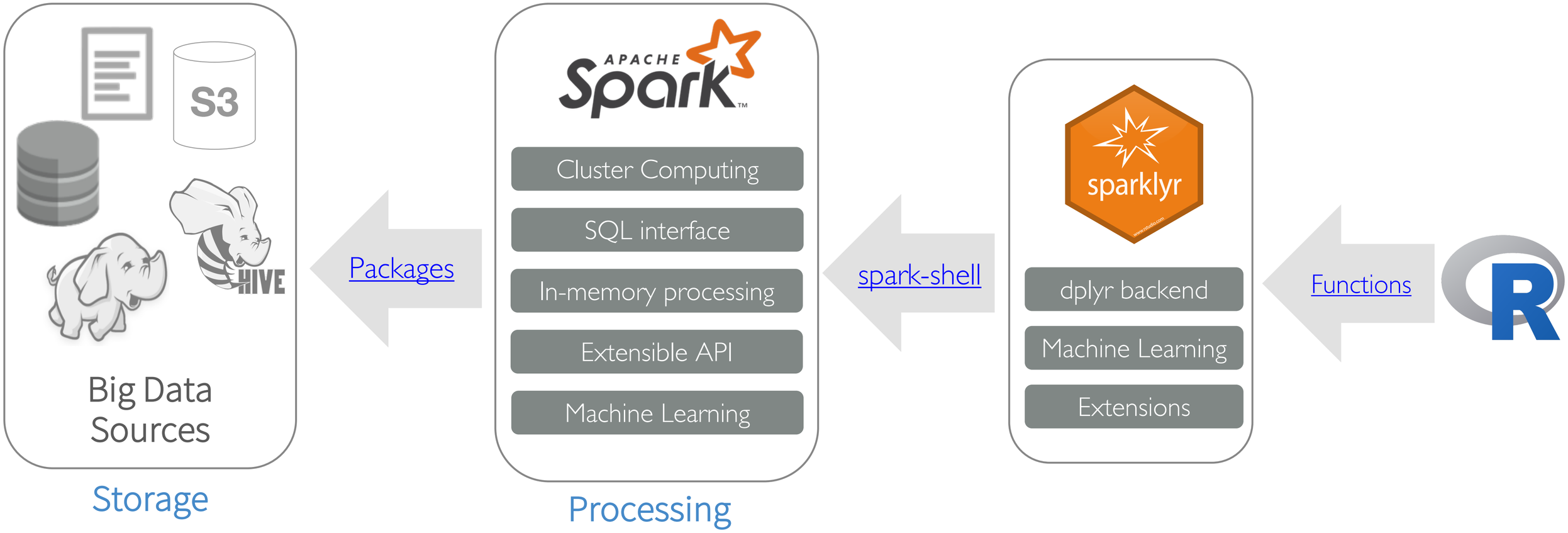
What to expect

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

Goal: *Share something new and useful*

RStudio®

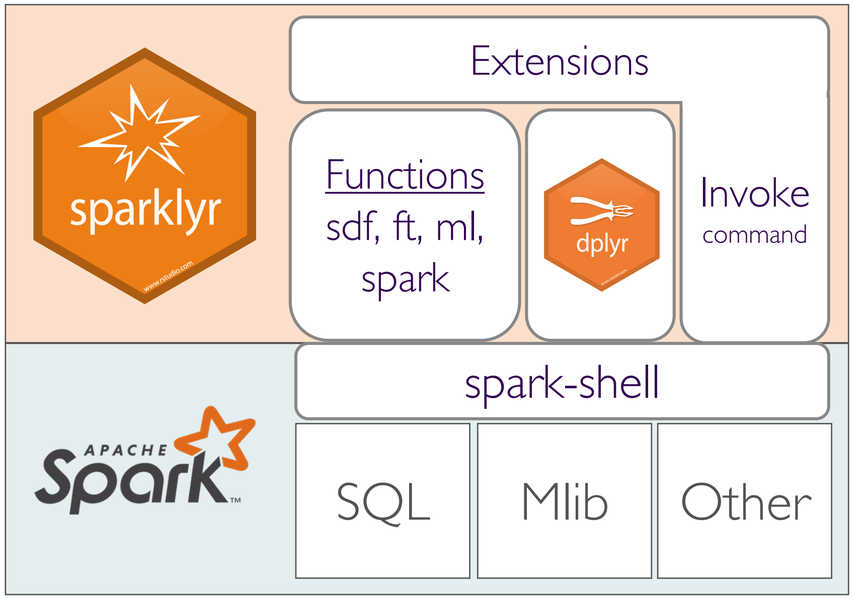
SPARKLYR – AN R INTERFACE FOR SPARK



RStudio®

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.



RStudio®

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

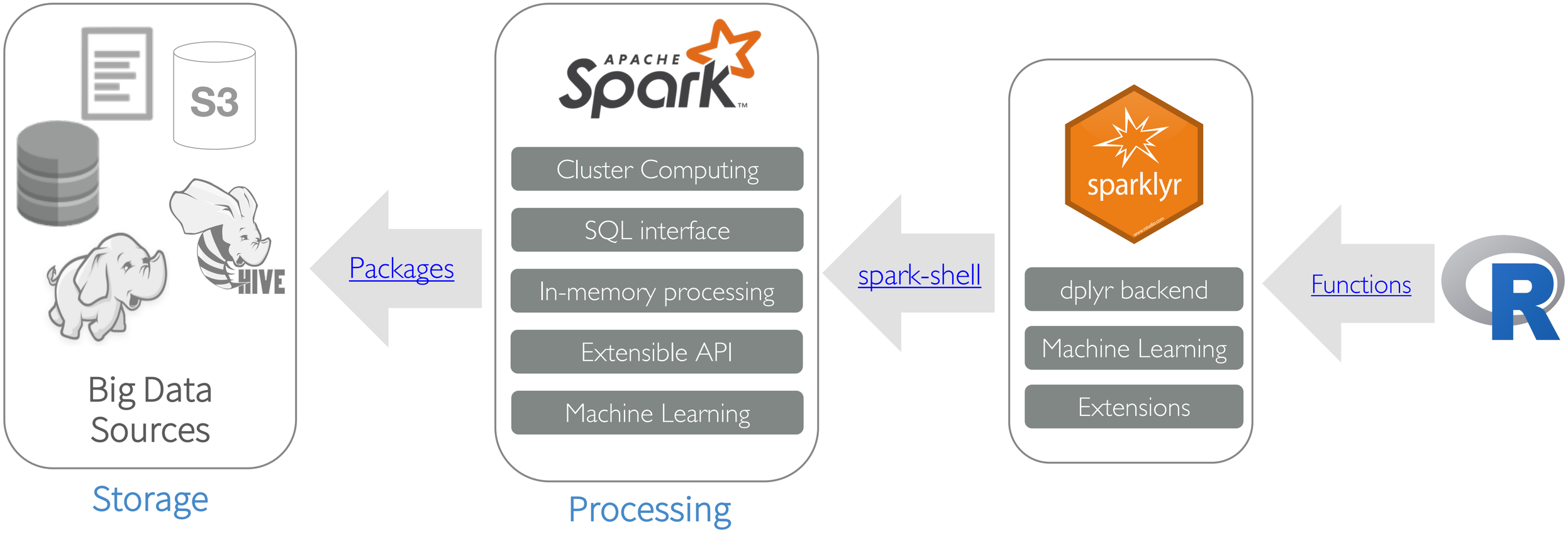
```
tbl(sc,"flights") %>%
  select(Origin)
```

Translates to

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

RStudio®

SPARKLYR – AN R INTERFACE FOR SPARK



RStudio®

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

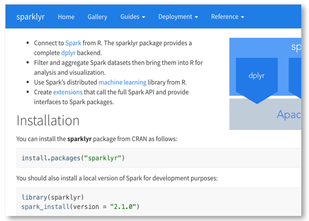
RStudio®

DEMO

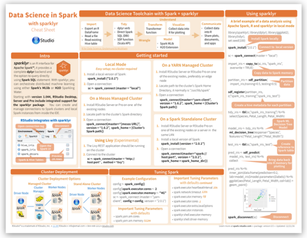
- Review of sparklyr functions
- Use the Spark UI during analysis
- Read & cache data in Spark
- What's new in sparklyr
- Tips n' tricks

RStudio®

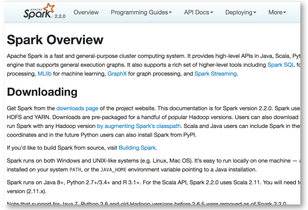
Useful Links



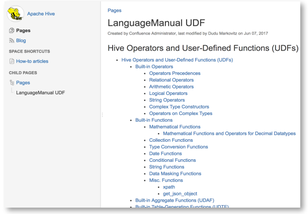
sparklyr's official website
spark.rstudio.com



sparklyr, IDE, RMarkdown cheatsheets
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

RStudio®

RStudio®

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

RStudio®