THE DATA LAKE SUMMIT

The definitive virtual conference for all things Data Lake

OCT 13-14, 2020

Scaling Data Science with Spark and R

Javier Luraschi

Software Engineer
Author of Mastering Spark with R

Presented by Qubole in collaboration with aws and Google Cloud

RStudio

About RStudio

2009
Founded by
J.J. Allaire

2011
RStudio IDE

2012
Shiny

2014
RStudio Server Pro
Shiny Server Pro

2016
Spark
Tidyverse

2017
RStudio Connect

2018
Databases
Ursa Labs
TensorFlow
Package Manager

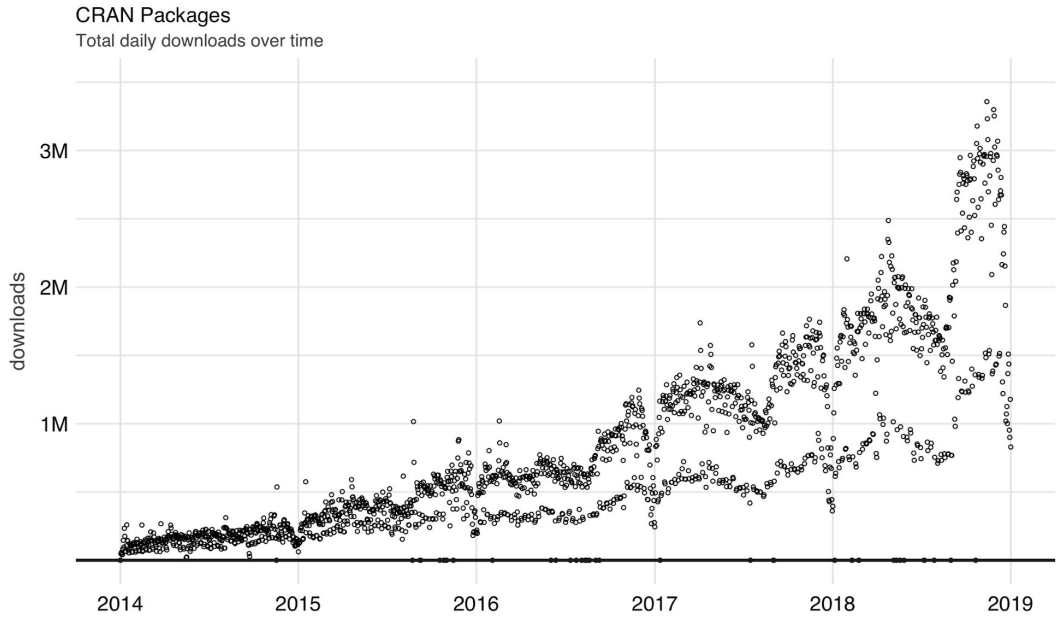
2019
Python & Jupyter
Notebook support

2020
rstudio::conf 2300 attendees
Public Benefit Corporation Announced

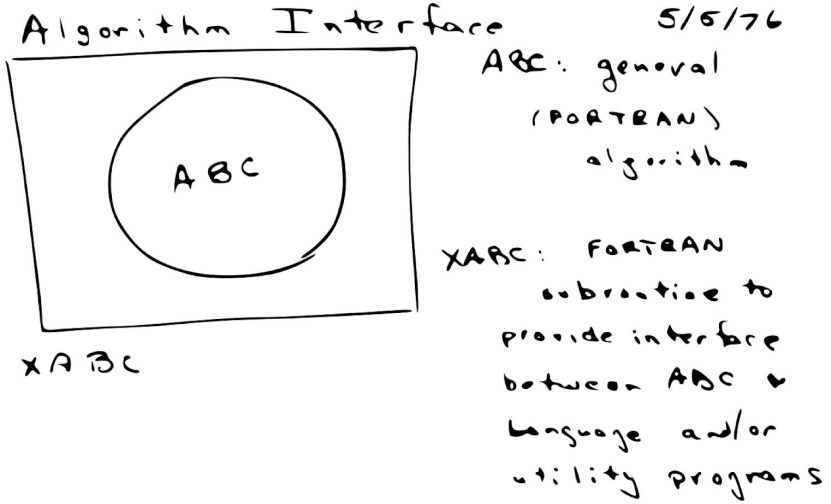
Why R for Data Science?

Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

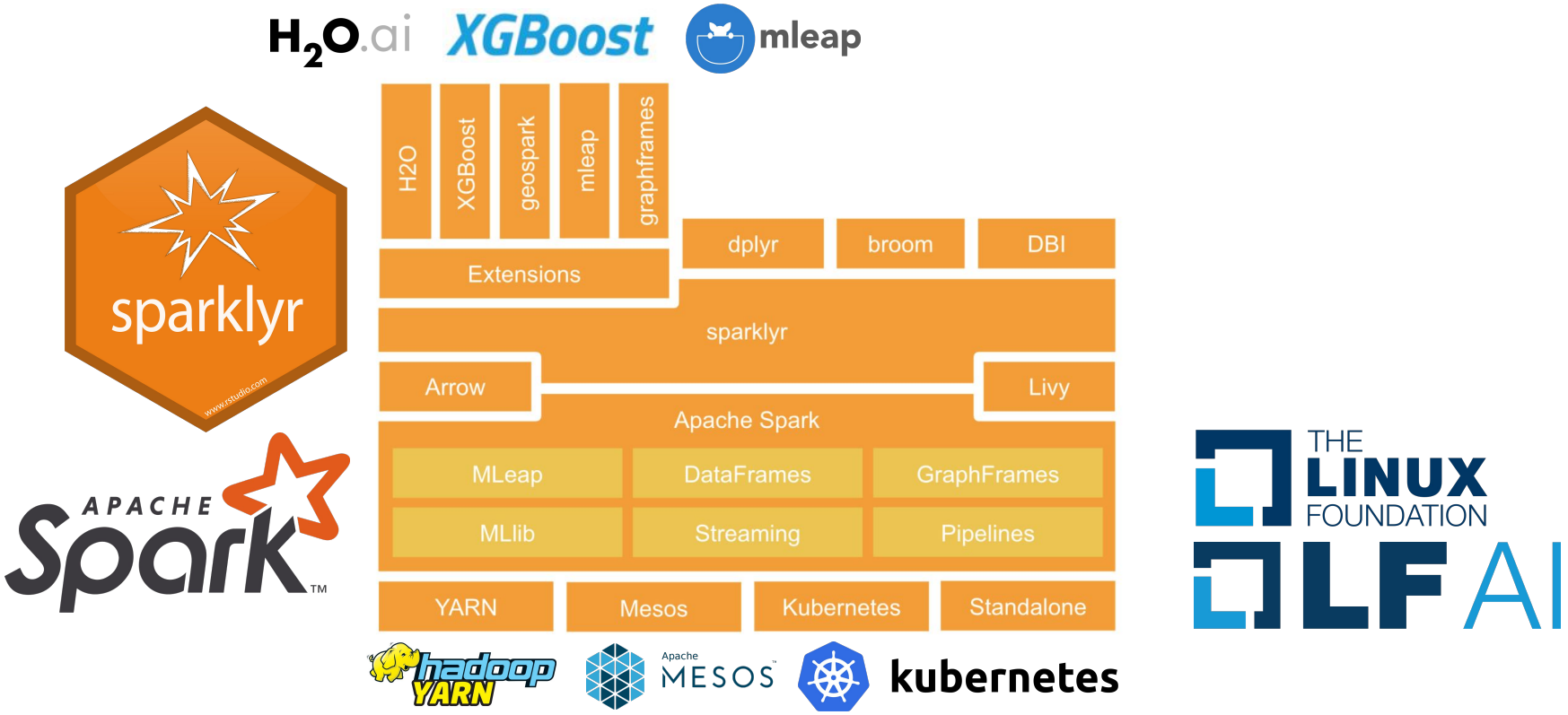


"R is a **programming language** and free software environment for **statistical computing** and graphics."



sparklyr: R Interface to Apache Spark

sparklyr is an open-source and modern interface to scale data science and machine learning workflows using Apache Spark™ and R.



Why sparklyr to scale Data Science?

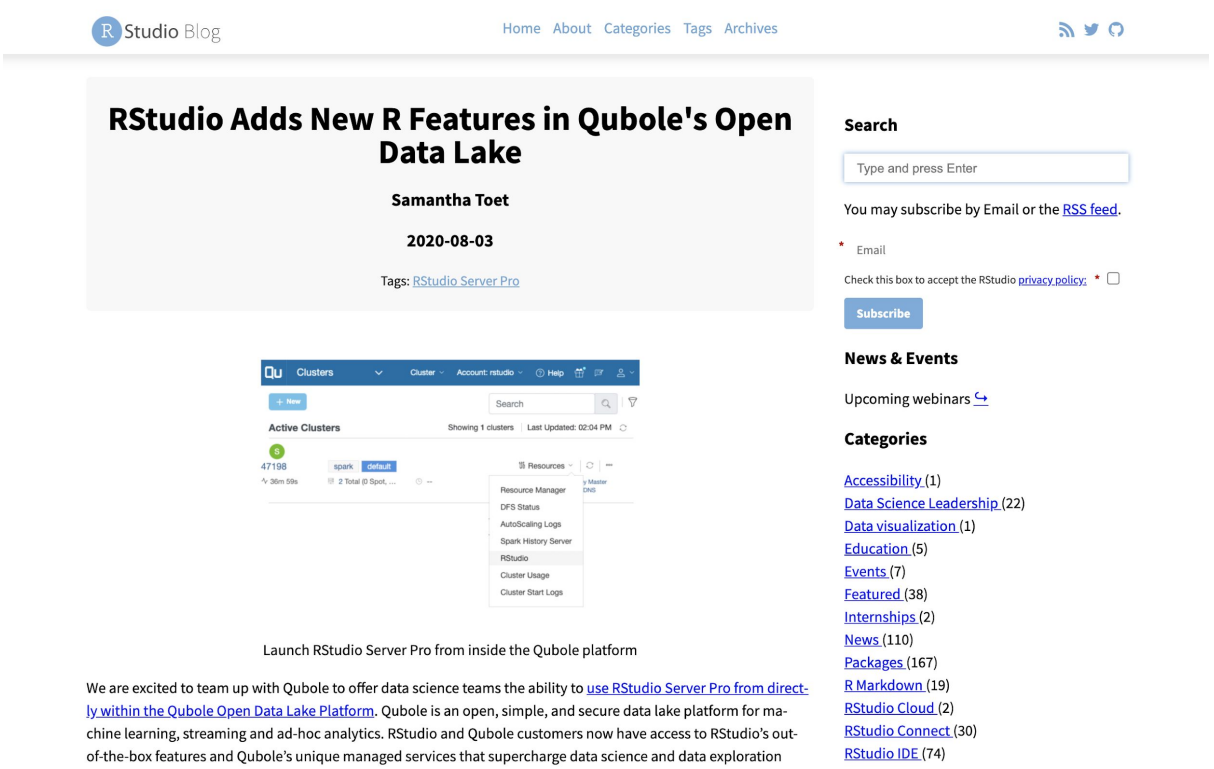
From the very-beginning, way before Pandas, sparklyr was designed to reuse as much R code through packages like dplyr, broom, rlang, tidyr, parsnip, etc. and practices like the pipe operator, formulas, and so on.


```r
library(sparklyr)
sc <- spark_connect(master = "local|yarn|mesos|spark|livy")
flights <- copy_to(sc, flights)
```


```r
library(dplyr)
flights %>%
  group_by(month, day) %>%
  summarise(count = n(), avg_delay = mean(dep_delay)) %>%
  filter(count > 1000)
```

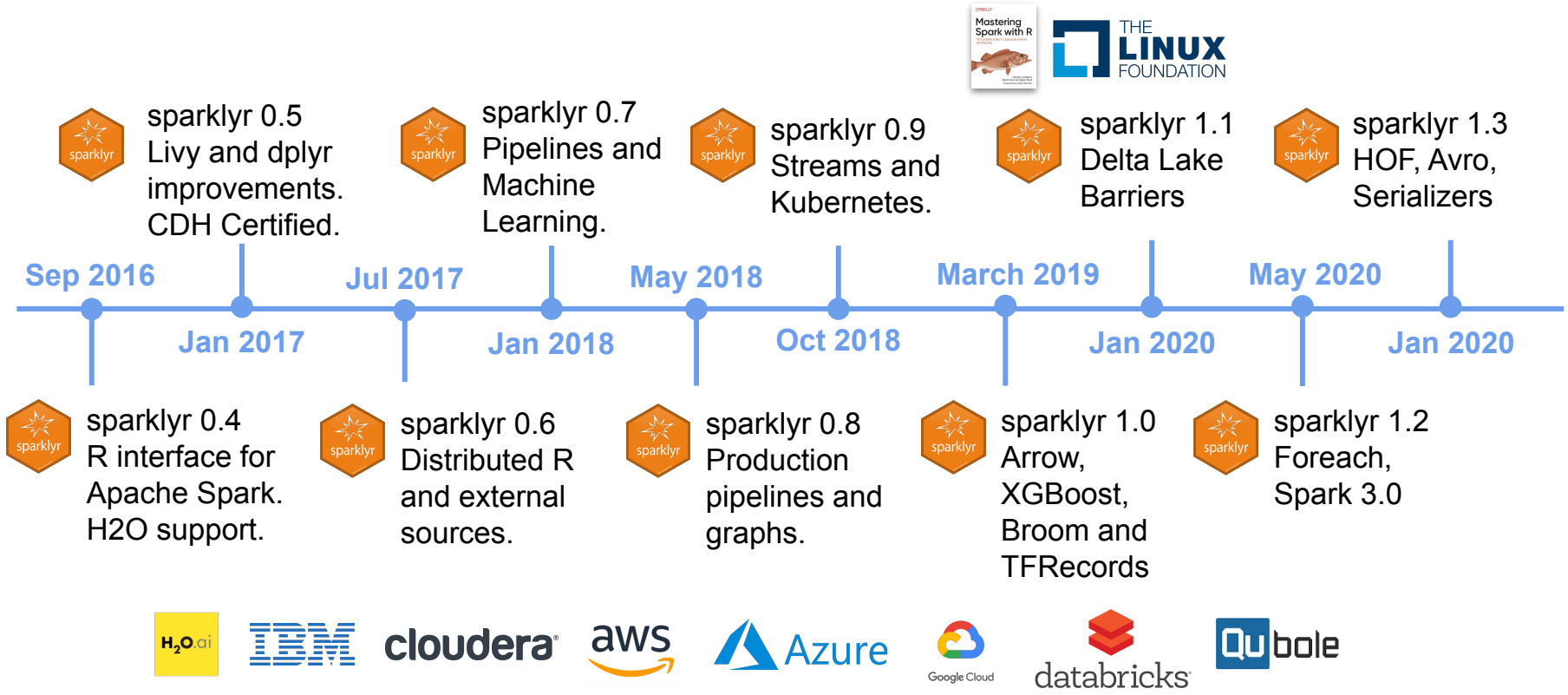
How do I use Spark with R in Qubole?

Seamless, non-disruptive RStudio Server Pro integration with Qubole — combining the power of the RStudio Server Pro with enhanced Apache Spark capabilities and near-zero cluster administration benefits provided by Qubole.



Catching up with sparklyr?

Since 2016, RStudio and the R community have been developing sparklyr. Starting with support for dplyr, Spark's MLlib and H2O; to structured streaming, Apache Arrow, XGBoost, Delta Lake and joining the Linux Foundation.

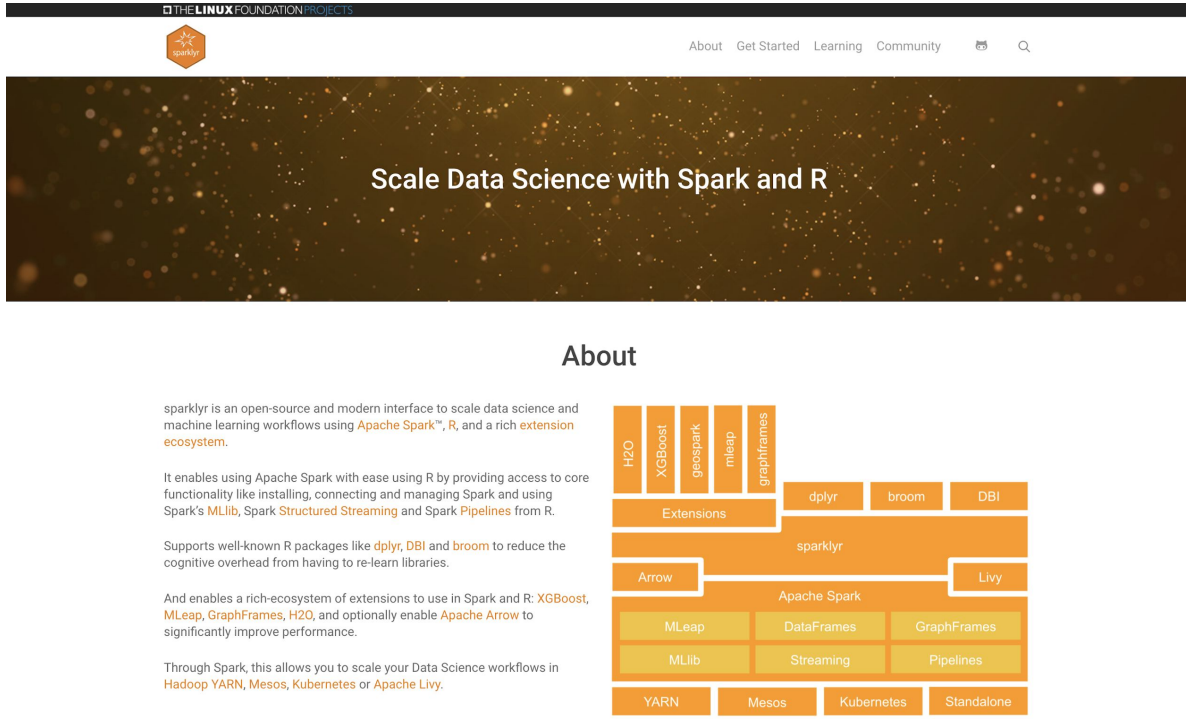


How do I use Spark from R?

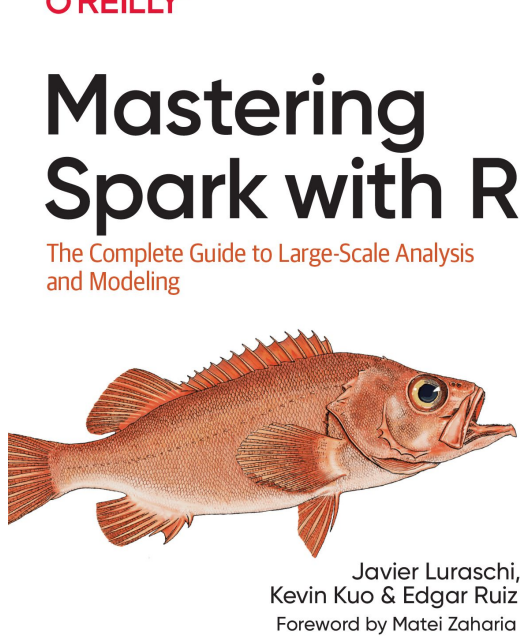
```
spark_install()                                  # Install local Spark
sc <- spark_connect(master = "local")            # Connect to Spark cluster

cars <- spark_read_csv(sc, "cars", "input/")     # Read data in Spark

summarize(cars, n = n())                         # Count records with dplyr
dbGetQuery(sc, "SELECT count(*) FROM cars")      # Count records with DBI

ml_linear_regression(cars, mpg ~ wt + cyl)       # Perform linear regression

ml_pipeline(sc) %>%                              # Define Spark pipeline
  ft_r_formula(mpg ~ wt + cyl) %>%               # Add formula transformation
  ml_linear_regression()                         # Add model to pipeline

spark_context(sc) %>% invoke("version")          # Extend sparklyr with Scala
spark_apply(cars, nrow)                          # Extend sparklyr with R

stream_read_csv(sc, "input/") %>%                # Define Spark stream
  filter(mpg > 30) %>%                           # Add dplyr transformation
  stream_write_json("output/")                   # Start processing stream
```

Where can I learn more about Spark and R?

Since 2016, RStudio and the R community have been developing sparklyr. Starting with support for dplyr, Spark's MLlib and H2O; to structured streaming, Apache Arrow, XGBoost, Delta Lake and joining the Linux Foundation.



sparklyr.ai



therinspark.com

THE DATA LAKE SUMMIT

The definitive virtual conference for all things Data Lake

OCT 13-14, 2020

Thank you!

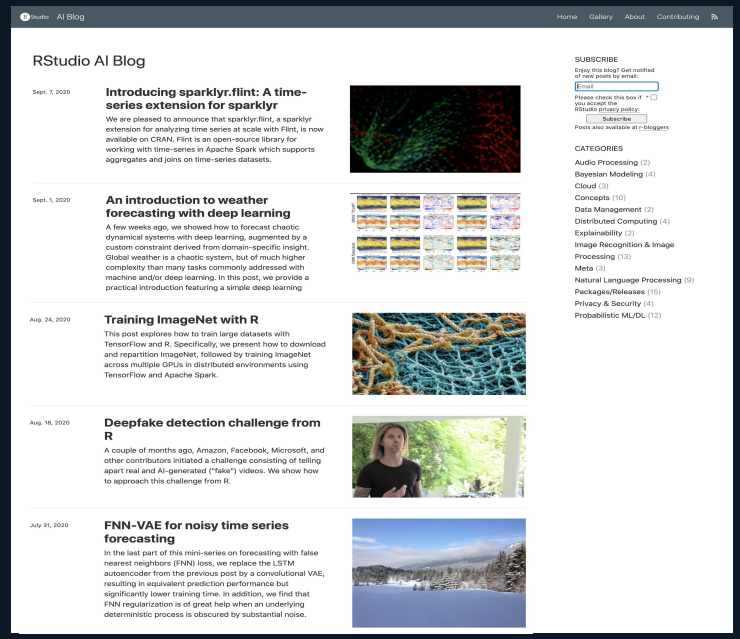

@dfalbel


@yl790


@javierluraschi


@zkajdan


@samantha_toet



blogs.rstudio.com/ai

Qubole presents
DATA LAKE SUMMIT

[Section Divider Title]

in collaboration with aws and Google Cloud

RStudio

Thank you

- Share this session on social media using **#DataLakeSummit**
- Interact with other attendees on Slack at **datalake-summit.slack.com**