

## SparkR (R on Spark)

- [Overview](#)
  - [SparkDataFrame](#)
    - [Starting Up: SparkSession](#)
    - [Starting Up from RStudio](#)
    - [Creating SparkDataFrames](#)
      - [From local data frames](#)
      - [From Data Sources](#)
      - [From Hive tables](#)
    - [SparkDataFrame Operations](#)
      - [Selecting rows, columns](#)
      - [Grouping, Aggregation](#)
      - [Operating on Columns](#)
      - [Applying User-Defined Function](#)
        - [Run a given function on a large dataset using dapply or dapplyCollect](#)
          - [dapply](#)
          - [dapplyCollect](#)
        - [Run a given function on a large dataset grouping by input column\(s\) and using gapply or gapplyCollect](#)
          - [gapply](#)
          - [gapplyCollect](#)
        - [Run local R functions distributed using spark.lapply](#)
          - [spark.lapply](#)
  - [Running SQL Queries from SparkR](#)
- [Machine Learning](#)
  - [Algorithms](#)
    - [Classification](#)
    - [Regression](#)
    - [Tree](#)
    - [Clustering](#)
    - [Collaborative Filtering](#)
    - [Frequent Pattern Mining](#)
    - [Statistics](#)
  - [Model persistence](#)
- [Data type mapping between R and Spark](#)
- [Structured Streaming](#)
- [R Function Name Conflicts](#)
- [Migration Guide](#)



```
1 iris_tbl <- copy_to(sc, iris)
2 flights_tbl <- copy_to(sc, nycflights13::flights, "flights")
3 batting_tbl <- copy_to(sc, Lahman::Batting, "batting")
4 src_tbls(sc)
5
6 # filter by departure delay and print the first few records
7 flights_tbl %>% filter(dep_delay == 2)
8
9 delay <- flights_tbl %>%
10   group_by(tailnum) %>%
11   summarise(count = n(), dist = mean(distance), delay = mean(arr_delay)) %>%
12   filter(count > 20, dist < 2000, !is.na(delay)) %>%
13   collect
14
15 # plot delays
16 library(ggplot2)
17 ggplot(delay, aes(dist, delay)) +
18   geom_point(aes(size = count), alpha = 1/2) +
19   geom_smooth() +
20   scale_size_area(max_size = 2)
```

20:32 (Top Level) R Script

Console

```
> delay <- flights_tbl %>%
+   group_by(tailnum) %>%
+   summarise(count = n(), dist = mean(distance), delay = mean(arr_delay)) %>%
+   filter(count > 20, dist < 2000, !is.na(delay)) %>%
+   collect
Warnmeldung:
Missing values are always removed in SQL.
Use `mean(x, na.rm = TRUE)` to silence this warning
This warning is displayed only once per session.
>
> # plot delays
> library(ggplot2)
> ggplot(delay, aes(dist, delay)) +
+   geom_point(aes(size = count), alpha = 1/2) +
+   geom_smooth() +
+   scale_size_area(max_size = 2)
'geom_smooth()' using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

