Spark 2.4.3 Overview Programming Guides API Docs Deploying More

SparkR (R on Spark)

```
iris_tbl <- copy_to(sc, iris)
flights_tbl <- copy_to(sc, nycflights13::flights, "flights")
batting_tbl <- copy_to(sc, Lahman::Batting, "batting")
src_tbls(sc)

# filter by departure delay and print the first few records
flights_tbl %>% filter(dep_delay == 2)

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

# 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)
```

```
> 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")'
>
```