Strata+Hadoop WORLD

PRESENTED BY

O'REILLY cloudera

RStudio

Win-Vector LLC

strataconf.com

#StrataHadoop

Modeling big data with R, sparklyr, and Apache Spark

1:30pm–5:00pm Tuesday, March 14, 2017
Data science & advanced analytics
Location: LL21 C/D
Level: Intermediate
Secondary topics: R

John Mount *(Win-Vector LLC)*
Steve Nolen *(RStudio)*
Edgar Ruiz *(RStudio)*

url: https://github.com/WinVector/BigDataRStrata2017

Our current project

- Data manipulation in Spark.

Work through markdowns together

- Exercises/03a-Spark-SQL.Rmd
- Exercises/solutions/03b-Spark-SQL.Rmd

Exercise

- Please complete Exercises/03a-Spark-SQL.Rmd

Open Discussion Activity

- Let’s step trough Exercises/solutions/03b-Spark-SQL.Rmd together.
    - Allows us to cover more material than a formal exercise.
- Please interrupt me (but not each other).

Challenge Project: more functions

- If you have extra time try implementing something like.
    - `dplyr::bind_rows()`
        - hint: select to re-order columns before union.
    - `tidyr::complete()`
        - hint: anti-join or right-join nearly do this.
- Not sure if possible to do reasonably and efficiently:
    - `tidyr::gather()`
    - `tidyr::spread()`