OPEN DATA SCIENCE CONFERENCE

Burlingame I November 2nd 2017

@ODSC

@WinVectorLLC

@RStudio

Nov 02
2:00 PM
Room T2

Modeling big data with R, sparklyr, and Apache Spark

BIG DATARINTERMEDIATE

Dr. John Mount
Consulting Algorithmist/Researcher/Principal at Win-Vector LLC and
Co-author of Practical Data Science with R

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.