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

Note

- You should be able to run all of these examples at your leisure.
  - This gives you a local Spark cluster.
  - RStudio has tutorials that include the install process (note: change to 2.2.0 instead of 1.6.2): http://spark.rstudio.com

What are we going to do?

- Supervised machine learning in SparkML.

Spark ML

- Machine learning on Spark

Spark ML (continued)

- MLlib is Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as:
  - ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering
  - Featurization: feature extraction, transformation, dimensionality reduction, and selection
  - Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
  - Persistence: saving and load algorithms, models, and Pipelines
  - Utilities: linear algebra, statistics, data handling, etc.

From: http://spark.apache.org/docs/latest/ml-guide.html

SparkML (continued)

- “Spark ML” is not an official name but occasionally used to refer to the MLlib DataFrame-based API. This is majorly due to the org.apache.spark.ml Scala package name used by the DataFrame-based API, and the “Spark ML Pipelines” term we used initially to emphasize the pipeline concept.
- MLlib switching to the DataFrame-based API
    - DataFrames provide a more user-friendly API than RDDs. The many benefits of DataFrames include Spark Datasources, SQL/DataFrame queries, Tungsten and Catalyst optimizations, and uniform APIs across languages.

From: http://spark.apache.org/docs/latest/ml-guide.html

Work through markdowns together

- Exercises/solutions/04-Spark-ML.Rmd
- To keep this interactive, please ask me questions!