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

Note

- You should be able to run all of these examples at your leisure.
  - Exercises/solutions/RsparklingInstall.Rmd has code that installs Spark and h2o on a local machine in the (not normally run) “install” block.
    - This gives you a local Spark and h2o cluster.
  - RStudio has tutorials that include the install process (not change to 2.0.0 instead of 1.6.2): http://spark.rstudio.com

What are we going to do?

- Supervised machine learning in SparkML.
- Supervised machine learning in h2o.

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

h2o

- State of the art big data machine learning.
- Has its own storage system.
  - Copies over from Spark lazily through `Sparkling Water` / `rsparkling`.
- Has its own version of many base R commands
  - Prefixed by h2o
    - So h2o equivalent of ls() is h2o.ls().

H2O.ai

H2O.ai, the Company

- Founded in 2012
- Stanford & Purdue Math & Systems Engineers
- Headquarters: Mountain View, California, USA

H2O, the Platform

- Open Source Software (Apache 2.0 Licensed)
- R, Python, Scala, Java and Web Interfaces
- Distributed algorithms that scale to “Big Data”

From: https://github.com/h2oai/h2o-meetups/blob/master/2017_02_23_Metis_SF_Stacked_Ensembles_Deep_Water/stacked_ensembles_in_h2o_feb2017.key

H2O.ai: Scientific Advisory Council

Dr. Trevor Hastie

- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*

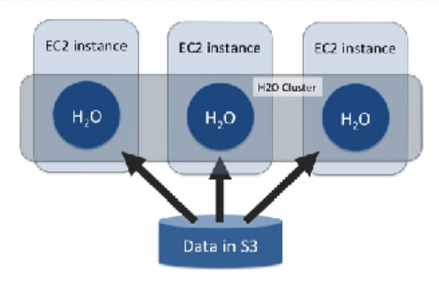
Dr. Steven Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*

From: https://github.com/h2oai/h2o-meetups/blob/master/2017_02_23_Metis_SF_Stacked_Ensembles_Deep_Water/stacked_ensembles_in_h2o_feb2017.key

H2O Distributed Computing

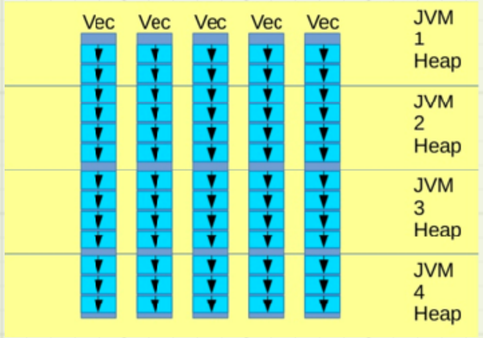
H2O Cluster



- Multi-node cluster with shared memory model.
- All computations in memory.
- Each node sees only some rows of the data.
- No limit on cluster size.

H2O Frame

- Distributed data frames (collection of vectors).
- Columns are distributed (across nodes) arrays.
- Works just like R's data.frame or Python Pandas DataFrame



From: https://github.com/h2oai/h2o-meetups/blob/master/2017_02_23_Metis_SF_Stacked_Ensembles_Deep_Water/stacked_ensembles_in_h2o_feb2017.key

Work through markdowns together

- Exercises/solutions/04a-Spark-ML.Rmd
- Exercises/solutions/04b-Spark-ML-h2o.Rmd
- To keep this interactive, please ask me questions!

H2O Resources

- H2O Online Training: http://learn.h2o.ai
- H2O Tutorials: https://github.com/h2oai/h2o-tutorials
- H2O Meetup Materials: https://github.com/h2oai/h2o-meetups
- H2O Video Presentations: https://www.youtube.com/user/0xdata
- H2O Community Events & Meetups: https://h2o.ai/events

From: https://github.com/h2oai/h2o-meetups/blob/master/2017_02_23_Metis_SF_Stacked_Ensembles_Deep_Water/stacked_ensembles_in_h2o_feb2017.key

Challenge Project: Stacking

- If you have extra time try Dr. Erin LeDell's excellent stacking tutorial.
  - https://github.com/h2oai/h2o-tutorials/tree/master/tutorials/ensembles-stacking