**R Hadoop Demo**

1. Intro
   1. R is the gold standard Data Scientist’s toolkit
   2. But R is intended for single-user and is an in-memory system so scalability limitation
   3. We are also looking at Python (pandas & numPy) but it has similar scalability limitations anyway
   4. The objective is to avoid having to re-write in an enterprise language like Java that scales great but has terrible support for Data Science.
   5. Hadoop and Apache Spark/Storm look capable of getting the best of both worlds – Data Science gold standard algorithms in R with the scalability of Java
2. Show PMT paradata files
3. Show For-loop R code
4. Run For-loop R code to show how slow and serial it is
5. Reset RStudio
6. Show RHadoop R code
7. Show empty Hadoop HDFS
8. Copy From Local to Hadoop
9. Show Hadoop HDFS populated
10. Run RHadoop code
    1. Use top to show the Hadoop running
    2. Show PMT-out appearing in HDFS
11. Show the PMT-out HDFS at the end
    1. The SUCCESS file and the out files
12. Show the output data frame
13. Explain pros and cons:
    1. Pro – can handle entire year of data in a single batch job
    2. Pro – scales R
    3. Con – need to duplicate data in HDFS so could be costly
    4. Con – not real-time – say that we will demo Apache Spark streaming in an upcoming demo day