Billion Row Benchmark

Minimum Specs  
1. CentOS or Redhat  
2. Yum properly configured to RHEL/EPEL repos

Installation Instructions  
Software  
cp /net/share/software/Revolution/Revo-Ent-6.1.0-RHEL6.tar.gz .  
tar –xvf Revo-Ent-6.1.0-RHEL6.tar.gz  
cd RevolutionR\_6.1.0  
./install.py

Use all the defaults, press [Y] at all prompts. Default installation is /etc (if I’m not mistaken)

Data  
wget –c   
<http://www.revolutionanalytics.com/subscriptions/datasets/AirlineData87to08.tar.gz>  
tar –xvf AirlineData87to08.tar.gz  
The data should be in AirlineData87to08\_83010/AirlineData87to08.xdf

Inflating the Data to 1Billion Rows  
ReplicateXdfFile.R  
1. Edit Lines 35 and 36 to indicate your datainput folder and dataoutputfolder  
2. R CMD BATCH ReplicateXdfFile.R &

Running the Benchmark  
BillionLogistic.R  
1. Edit Line 10, to indicate the folder name and file name of the 1BillionRow dataset  
2. R CMD BATCH BillionLogistic.R &  
  
A File BenchMarktest.txt will be created. That is the results of the Benchmark.