Getting Started with Hadoop and R

I am trying to learn about "Big Data" and figured the only way to start is to dive right in. Worth noting is that I am going to use a single machine that I have at home. For context, I have about 700 text files that total about 300gb's of data. Each file contains JSON responses captured from Twitter's firehouse over the last month.

I love R and ultimately what to use it to study my dataset, but I know that I need a way to "store" the data. I hear a lot about Hadoop and HDFS, but can't get my head wrapped around it. Would I simply "copy" the text files to HDFS on my local machine and use the `RHadoop` to write Map/Reduce statements to create datasets?

Lastly, I have MongoDB up and running and was considering storing the data there but I am not sure that the I would capture analytical performance gains, although I know that there is an adaptor for Haddop.

My question: Having successfully captured the data, what is the best way to store this such that I can use R (and other tools) to analyze the data.

r mongodb hadoop

asked Nov 8 '12 at 17:57
Btibert3
3,568 15 51 96
add a comment

2 Answers

If you do not want to do batch processing a lot and do real time queries on tweets, a non relational DB like MongoDB would suit your need very good. So for realtime queries, have a look into MongoDB's Aggregation Framework.

So it comes down to: What you really want to do with the data? Find tweets around places and show avg follower count? Or long term Trend Analysis?

Here is an ruby/mongodb post how someone scraped 3million tweets: how-i-scraped-and-stored-over-3-million-tweets

answered Nov 26 '12 at 16:35
Marc
412 2 5
add a comment

You should definitely not use MongoDB. It is not designed for batch analytics and will not be performant for that purpose.

Your idea of simply copying the files to HDFS and using RHadoop is a good one in general, but using only one machine is not the ideal case. It will certainly make good use of the multiple cores that your one machine has, and it will do a good job processing everything without overflowing memory, but it might not be the ideal tool for the job.

I don't know too much about the R libraries that are out there, but I would guess there might be a better one out there for processing of large data sets, but not so large that multiple machines are needed. Hell, even just putting your data in a more traditional analytical database might be better.

answered Nov 8 '12 at 22:55

Joe K
4,915 5 16

add a comment

Not the answer you're looking for? Browse other questions tagged r mongodb hadoop or ask your own question.