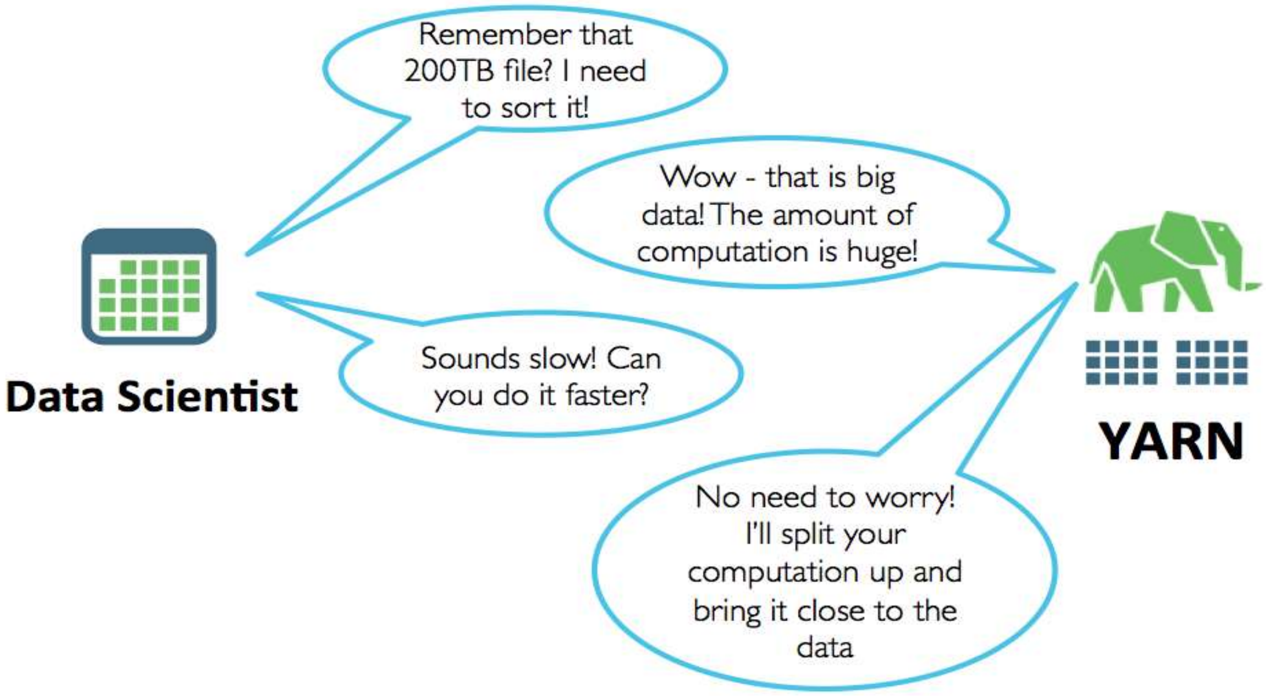
What Is YARN?

Remember our data scientist storing the most important data set in HDFS? Now, she is back and she wants to find the top 1000 records for a given sorting criteria. Before Hadoop, this whole process required writing an algorithm, running it on a single node, and making it act as a single client to HDFS data, essentially sifting through all of it record-by-record. As you may imagine, if your data set qualifies to really be called ***big*** data, this will be an extremely slow process.


Remember that 200TB file? I need to sort it!
Wow - that is big data! The amount of computation is huge!
Data Scientist
Sounds slow! Can you do it faster?
YARN
No need to worry! I'll split your computation up and bring it close to the data

Storing and Reading Data with HDFS (by Hortonworks, Inc.)

First of all, this process is limited by running just one copy of the algorithm on a single server. Second of all, you are making all the required data be transferred over the networks (remember: you do not need all the data, just the top 1000 records).

When Hadoop came onto the scene, it radically improved the scalability of these types of computation by:

- Allowing computation to run on each node in the cluster
- Using the same nodes for computation that HDFS was using for storing blocks
- Making sure that pieces of computation got scheduled on the nodes that hosted blocks required for computation locally, thus avoiding the need to send data over the network.

The mantra of Hadoop has always been: "*Don't bring data to compute, bring compute to data*", and the part of Hadoop that made this mantra a reality is YARN (Yet Another Resource Negotiator).

Learn About Verified Certificates