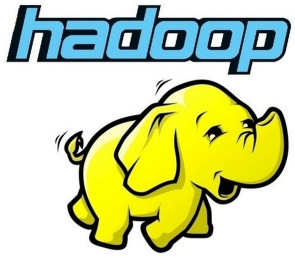
Spark

Jake Schefrin
2/20/2020

Spark, an introduction

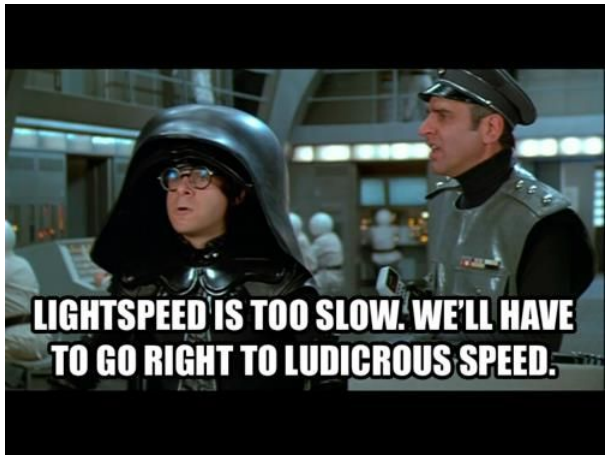
- Spark is open source unified analytics engine that assists with data processing and machine learning.
- Derived from Hadoop (we will talk about differences in a second)
- Developed at UC Berkeley in 2009
- Apache is its caretaker now, checks out some of their other open source programs!

What is Hadoop?



- Allows for the processing of big data across clusters of computers
- Your data and applications are safe with Hadoop thanks to distributed computing
- Every machine spun up offers local computation and storage to disk

So why Spark?



- SPEED! Spark was designed to be much faster than Hadoop. It does this by saving fewer operations to disk, and using RAM.
- Spark is available in multiple programming languages (SQL, Java, Python, Scala)
- Spark uses Hadoop for storage only, uses seperate clusters for computation

Other benefits of Spark

- Real time processing, which allows for processing of streaming data. This also provides instantaneuous results.
- Spark is extremely useful for machine learning. (Sit in on Ed's class next year!!! With his permission of course)
- Spark is really useful for graph algorithms (GraphX) and unites exploratory graph analysis, iterative graph computation, and ETL (extract, transform, load) under one roof.

Who cares?

- Many major companies like Yahoo!, Amazon, and ebay use Spark
- Saves time, that's pretty precious
- Grant will say you should care so you should, otherwise face Grant Moff Tarkin's wrathful glare

Other possibilities

-Monet DB was recommended by Ed