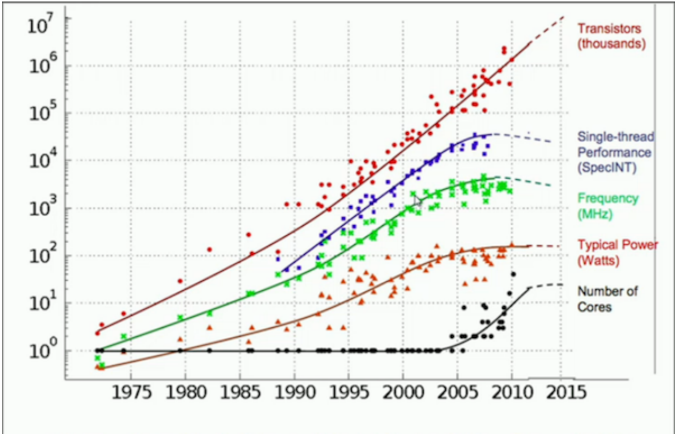
COMPSCI 589
Lecture 13: Introduction to Apache Spark

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).
Created with support from National Science Foundation Award# IIS-1350522.

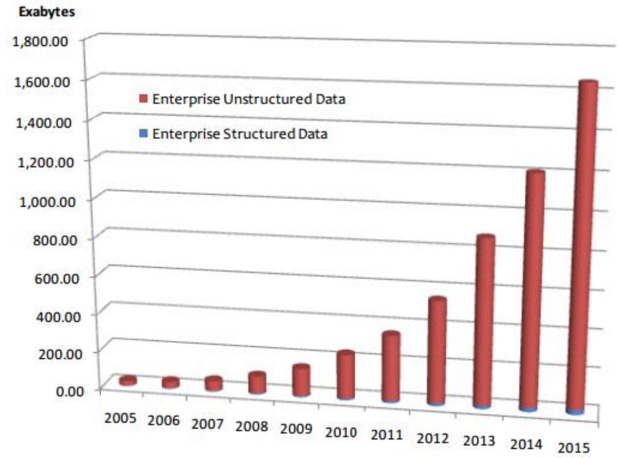
Moore’s Law



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

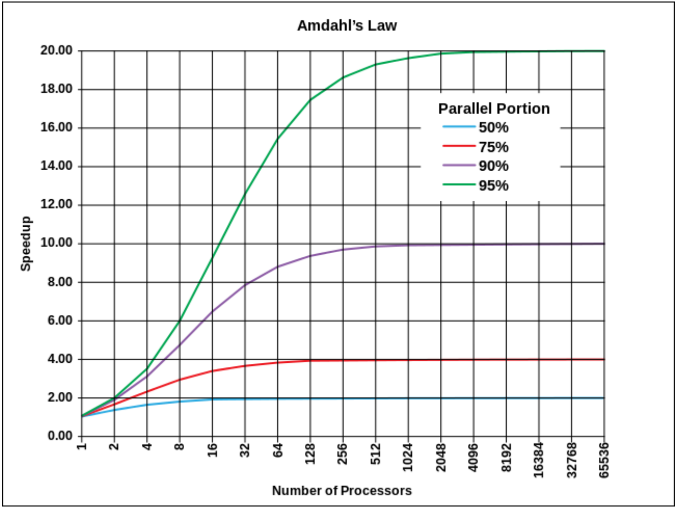
Machine Learning’s free ride ended in about 2005.

Big Data



The amount of data is doubling every two years.

Amdahl's Law


Amdahl's Law
Speedup
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
Number of Processors
Parallel Portion
50%
75%
90%
95%

Functional Programming and Data Parallel Computing

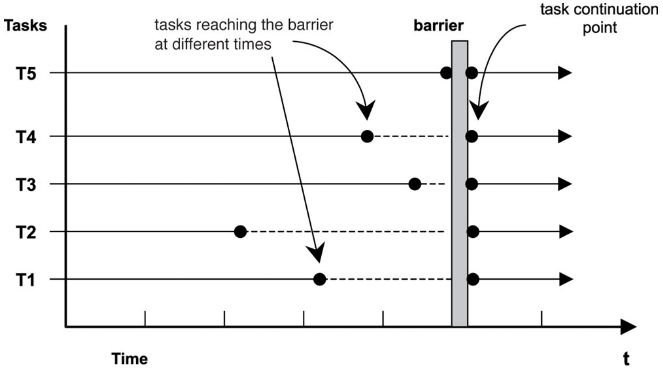
- Functional programming is a natural match for data parallel computing where we want to do things like:
  - Apply the same function to all elements in a data set (Map)
  - Apply a Boolean filter to select only certain data elements (Filter)
  - Aggregate a number of data elements by summing, maxing, etc. (Reduce or Fold).
- It turns out that a small number of such easily parallelizable functional programming primitives are sufficient for creating data-parallel implementations of machine learning algorithms.

MapReduce and Hadoop

- MapReduce is a distributed programming model introduced by Google in the early 2000's where all you can do is apply map and reduce functions to data.
- Hadoop is a widely used open-source implementation of this framework.
- A scheduler breaks up the map computations over a cluster with a data-parallel distributed file system. The results of the map step are written back to the file system.
- The scheduler then schedules the reduce jobs on the cluster, which produce the final output and write it to the file system.
- MapReduce uses a specialization of reduce for key-value pairs called *reduce-by-key*.

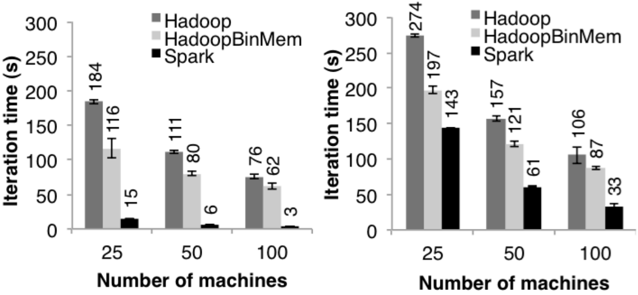
Limitations of MapReduce For ML

- The fact that MapReduce is completely stateless and all communication between processing iterations happens via the file system creates a significant synchronization barrier that negatively affects parallel scalability of iterative computations.



Apache Spark

- Apache Spark is a parallel and distributed programming framework that adds additional parallel abstractions and allows for distributed in-memory caching as well as distributed on-disk data access. This makes it much faster than MapReduce for ML tasks.



(a) Logistic Regression (b) K-Means

Apache Spark

Examples