8.3 MapReduce and Beyond

The capabilities necessary to tackle large-data problems are already within reach by many and will continue to become more accessible over time. By scaling "out" with commodity servers, we have been able to economically bring large clusters of machines to bear on problems of interest. But this has only been possible with corresponding innovations in software and how computations are organized on a massive scale. Important ideas include: moving processing to the data, as opposed to the other way around; also, emphasizing throughput over latency for batch tasks by sequential scans through data, avoiding random seeks. Most important of all, however, is the development of new abstractions that hide system-level details from the application developer. These abstractions are at the level of entire datacenters, and provide a model using which programmers can reason about computations at a massive scale without being distracted by fine-grained concurrency management, fault tolerance, error recovery, and a host of other issues in distributed computing. This, in turn, paves the way for innovations in scalable algorithms that can run on petabyte-scale datasets.

None of these points are new or particularly earth-shattering—computer scientists have known about these principles for decades. However, MapReduce is unique in that, for the first time, all these ideas came together and were demonstrated on practical problems at scales unseen before, both in terms of computational resources and the impact on the daily lives of millions. The engineers at Google deserve a tremendous amount of credit for that, and also for sharing their insights with the rest of the world. Furthermore, the engineers and executives at Yahoo deserve a lot of credit for starting the open-source Hadoop project, which has made MapReduce accessible to everyone and created the vibrant software ecosystem that flourishes today. Add to that the advent of utility computing, which eliminates capital investments associated with cluster infrastructure, large-data processing capabilities are now available "to the masses" with a relatively low barrier to entry.

The golden age of massively distributed computing is *finally* upon us.