When Massive Data Never becomes Big Data

Steve Conway, IDC

The recent PRACE Days conference in Barcelona provided powerful reminders that massive data doesn't always become big data — mainly because moving and storing massive data can cost massive money. PRACE is the Partnership for Advanced Computing in Europe, and the 2014 conference was the first to bring together scientific and industrial users of PRACE supercomputers located in major European nations.

Tom Lange of Procter & Gamble, which has used high performance computer (HPC) systems to help design consumer products including Pampers diapers and Pringles potato chips, said that although P&G manufactures billions of Pampers a year and they all have data sensors affixed to them, all this data is deleted once the diapers exit the manufacturing line. "P&G doesn't have any big data," he stated, explaining that "businesses typically won't pay for storing data that doesn't make them money." Someone has to make the economic argument for storing data, especially when it comes in petabyte-size chunks.

This revelation sparked a lively discussion on who gets to decide which data gets saved when the future value of the data isn't known. For businesses like P&G, the decision process is clear: if no one steps forward to pay for storage, the data probably won't be saved. For massive data collected by governments, decisions can be more serendipitous and comprehensive policies are often lacking today.

Because the HPC community confronts very massive data, much of it meeting the "four Vs" definitional criteria for big data (volume, variety, velocity, and value), leading HPC users are already wrestling with issues that may not affect mainstream IT markets for a while. In addition:

- The Large Hadron Collider at CERN generates 1 PB of data per second when it's operating and the upcoming Square Kilometer Array telescope will disgorge 1 EB of data (1,000 PB) per day. In both cases, only a small fraction of the data will be stored. Storing it all would break the budget.
- China's Internet-of-Things initiative is targeting a 10,000-fold data reduction to avoid having to confront 100ZB of data from home sensors alone by 2030. A major strategy of the effort, spearheaded by the Chinese Academy of Sciences, is to create a single mega-sensor that will do the work of the 40-50 device sensors that would otherwise be expected to inhabit the average Chinese home at that time.
- The HPC community has already figured out how to exploit the benefits of Hadoop while evading Hadoop's requirement to store three copies of data. The workaround is to decouple Hadoop from the Hadoop Distributed File System (HDFS) and attach it instead to a fully POSIX-compliant parallel file system.

Storage is not the only daunting expense associated with massive data. Moving data can also be very costly. Technical experts say that a single computation typically requires 1 picojoule of energy, but moving the results of the computation may cost as much as 100 picojoules. At an average $1 million per megawatt, skyrocketing energy costs have become the number two issue in the worldwide HPC community, right after the perennial desire for bigger budgets. Hence, curtailing data movement is a bulls-eye target for cost-conscious HPC users in government, academia, and industry.

The HPC community was arguably the original home of big data and still operates at the leading edge of many big data developments, but costs for data movement and storage will continue to act as a brake on the growth of HPC big data. Even with this partial brake, IDC forecasts that the server market for high performance data analysis (big data using HPC) will grow rapidly (23.5% CAGR) to reach $2.7 billion in 2018 and the related storage market will expand to about $1.6 billion in the same year.

Steve Conway is Research VP, IDC High Performance Computing Group.