

Applied Mathematics for Data Analysis in the Exascale

Aydın Buluç (abuluc@lbl.gov)
Lawrence Berkeley National Laboratory

Abstract

Simulation and modeling is one aspect of scientific discovery, the other being data analysis. Exascale computing will generate tremendous amounts of scientific data that can not be analyzed using existing algorithms. The research areas that will provide the necessary mathematical tools for big data analysis are (sparse) linear algebra, machine learning, and graph theory. Among those, DOE has significant expertise only in linear algebra. We argue that research efforts in data should target all three areas; with a special focus on scalability to large data sets and large concurrencies.

Introduction

Modeling and simulation has traditionally been the workhorse of a majority of DOE funded computing research. As scientific data sets get larger and more complex, especially as we move into exascale era, the challenges are unlikely to be addressed with existing data analysis methods. Data mining, processing, and inference requires a combination of revamping some of our assumptions about input characteristics and introducing different kinds of mathematics into the DOE computing portfolio. This is not to say that simulation and modeling efforts will become obsolete. On the contrary, it is the simulation and modeling that creates most of the data to be fed into big data analysis systems. Computation is viewed as the third pillar of science. We argue for a refinement of that view in which modeling and simulation is the third and data analysis is the fourth paradigm [3].

Data are the fundamental sources of insight for all experimental and computational sciences. Our mathematical models for understanding the diversity and expressiveness of complex data sets, however, are in their infancy. Some of the challenges arise due to the scale of the data. In the past, a method was computationally feasible if it had polynomial complexity in input size. It is now prohibitive to do more than a constant number of passes over petascale data sets. We need methods that can sample intelligently without losing important potential outliers in the data [5].

DOE areas of interest that urgently require big data analytics include material science, computational cosmology, bioinformatics, climate science, and cybersecurity. All these application areas need automated prediction, classification, detection, and identification of patterns in their data sets. Visual inspection and manual browsing of scientific data is no longer possible, even with the help of sophisticated visualization techniques and human-computer interaction tools. The data tools of exascale should enable exploratory data analytics by guiding the scientist into interesting patterns in the data set.

Area to evolve: Applied linear algebra

Applied linear algebra is at the heart of modeling and simulation. Surprisingly, linear algebra is also the main kernel in many data mining algorithms such as spectral clustering, dimensionality reduction, and classification. DOE has already been funding fundamental applied mathematics research that will lead us to a better understanding of the complex data sets we have. Several

techniques and trade-offs that has been well known in scientific computing literature has just recently been brought up to the attention of the data mining community [2]. What is different in terms of linear algebraic methods in simulation/modeling and data-driven science is the characteristics of inputs. Data from experimental facilities especially tend to be noisy and not well-structured. The nonzero distribution patterns of sparse matrices that arise in simulation/modeling and data sciences can be vastly different as well. The former has good separators in general (notable exceptions include MFDn), while the latter usually follows a power law distribution. Therefore, how much of the existing literature is directly applicable to big data domain is subject to further investigation.

Areas to revolve: Machine learning and graph analysis

Mathematical needs of big data analysis are not solely based on linear algebra. There are at least two major areas whose focused DOE funding would result in significant impact to a broad range of science problems. First is statistics and machine learning, and second is graph theory/analysis. Both research areas have been funded by agencies like NSF for decades. However, this produced mainly theoretical research that are either never tested on real data sets, or just applicable only to very small data sets. In order to bring these much needed techniques to exascale, DOE should refocus the research on methods that are scalable both in terms of the input size and parallelism.

The graph abstraction provides a natural way to represent relationships among data in systems biology, genomics, chemistry, ecology, and astrophysics. Machine learning, on the other hand, use sophisticated statistical techniques to automatically classify data, detect patterns, and extract results from it. Computational kernels for large-scale graph processing and machine learning is hard to optimize and scale. Ideally, all these complexities should be encapsulated and hidden from the domain scientist. A promising approach is employed by the Knowledge Discovery Toolbox, which is easy to use through its high-level Python interface, flexible and customizable through user defined operations and filters, scalable and fast through its high-performance backend engine and selective JIT specialization of its user defined operations [1].

Scaling and speed, however, are not the only challenges of machine learning and graph processing. The complexity of the data, the uncertainties associated with it, and its potentially dynamic nature pose significant challenges to scientists that apply machine learning and graph theoretical methods to their datasets. For example, description of textbook graph algorithms typically start with the graph itself as the input while the cost of constructing this graph is overlooked. Even though the final graph is sparsified for faster processing, the construction has to look at every pair of data points to decide whether the edge should be retained or removed due to sparsification. Graph theoretical methods that are faster than $O(n^2)$ should ideally be applied directly to the data set itself without explicitly constructing its graph. One promising approach is the “path folding” method applied to power iteration clustering [4]. Such techniques, when coupled with sampling methods, has the potential to bring the full power of machine learning and graph theory to datasets that will be generated in exascale and beyond.

Conclusion

To enable real scientific discoveries through computing in the exascale, data analytics should be seen as important as simulation and modeling. The unique challenges of exascale will require algorithmic advances that enable faster and more scalable data analytics. DOE already has expertise in some areas that are fundamental to data analytics, such as applied linear algebra. We argued that this expertise needs to be re-tailored for special characteristics of the datasets. We also argued that significant research is urgently needed to bring the power of machine learning and graph theory to DOE’s big data problems.

References

- [1] Aydın Buluç, Erika Duriakova, Armando Fox, John Gilbert, Shoaib Kamil, Adam Lugowski, Leonid Oliker, and Samuel Williams. High-productivity and high-performance analysis of filtered semantic graphs. In *Proceedings of the IPDPS*. IEEE Computer Society, 2013.
- [2] Jie Chen and Yousef Saad. Lanczos vectors versus singular vectors for effective dimension reduction. *IEEE Trans. Knowl. Data Eng.*, 21(8):1091–1103, 2009.
- [3] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. The fourth paradigm: data-intensive scientific discovery. 2009.
- [4] Frank Lin and William W Cohen. A very fast method for clustering big text datasets. In *Proceedings of the 2010 conference on ECAI*, pages 303–308, 2010.
- [5] Michael W Mahoney. Randomized algorithms for matrices and data. *Advances in Machine Learning and Data Mining for Astronomy*, CRC Press, Taylor & Francis Group, Eds.: Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, Ashok N. Srivastava, p. 647-672, 1:647–672, 2012.