Mining Massive Complex Datasets on Emerging Architectures

Chandrika Kamath
Lawrence Livermore National Laboratory
kamath2@llnl.gov
April 30, 2013

The size and complexity of modern datasets makes it impossible to gain insights without assistance from analysis techniques from fields such as machine learning, statistics, signal and image processing, and data mining. This is especially true of data from experiments, observations, and high-fidelity simulations conducted by the applied programs in the Office of Science, Department of Energy. We will face several challenges as this problem of data overload will grow dramatically in the future. On one hand, we will need to modify existing algorithms to address the increase in the volume and complexity of the data. On the other hand, we will need to design and develop new algorithms as the data being collected prompt new types of analysis. For example, simulations are being used to design experiments, resulting in a need for techniques to efficiently sample the high-dimensional input space while providing adequate coverage of the output space. Data fusion techniques will enable greater insights by combining multi-modal, multi-resolution data, for example, from different experiments which analyze a material using different modalities or different astronomical surveys observing the sky at different frequencies. The large size of the data will require more automation in the analysis, prompting the need for adaptive algorithms that can account for the spatial and temporal variation in the data. Finally, statistical techniques will be used to assign probabilities to the results as the analysis is used in inference and in decision support.

Our efforts to gain insights into massive, complex datasets are enabled by the easy availability of computational power, ranging from the desktop to systems with moderate to high degree of parallelism. However, the characteristics of emerging architectures, with their multi-core processors, limited memory and bandwidth, and high cost of data movement, will make it challenging to efficiently apply existing analysis techniques to larger, more complex, datasets. Further, as we develop techniques to meet the requirements of the new types of analysis prompted by our improved data gathering capabilities, we will need to design these new algorithms to be suitable for modern architectures. The extreme-scale systems themselves will provide new opportunities for the application of analysis algorithms. We next discuss these challenges and opportunities in the context of two broad sources of data – simulations of complex phenomena run on high-performance systems and experimental/observational datasets.

Analysis of simulation data: As we move towards extreme-scale computing, the transfer of simulation output to persistent storage will become impractical, both due to the large size of the data and the limited I/O bandwidth that will be supported by the systems. The current approach to this problem is to write the output every few time steps instead of every time step. However, we will also need to re-think the traditional post-processing approach to analysis, where all the analysis is done after the data are written out. A common solution proposed is *in-situ analysis*, where the analysis is moved into the simulation and only the results written out. However, this will require the development of algorithms that are robust to the parameters used in the algorithms and can adapt to the changing statistics of the data as the simulation progresses. We will also need to address the challenges that arise when the simulations and analysis algorithms co-exist on a system with limited memory and high cost of data movement. Simulations run for the purpose of scientific discovery pose a different problem as we may not have even formulated all the questions to ask of the data before we run the simulations [ExaDM13]. Further, we may not know a-priori what information may be in the data and which algorithm/parameters would be the most suitable for extracting this information from the data. In such cases, we will need to develop mathematical techniques that can significantly reduce the volume of data, while still retaining useful information. This reduced dataset could then be output for analysis off-line.

Efficient use of extreme-scale systems: In addition to analysis of simulation output, there are two other opportunities for analysis algorithms to help address the challenges of extreme-scale computing. Given the power requirements of the extreme-scale systems, running a simulation will be expensive. When a number of these simulations are run, for example, as an ensemble for sensitivity analysis or to explore the design space of input parameters, we will have to carefully select the sample points in the input space of parameters so we can obtain the most insights using as few simulations as possible. Analysis techniques from the field of design and analysis of computer experiments [FLS06] will play an important role in addressing this concern. Another opportunity for analysis algorithms arises in mining the data from the monitoring of the extreme-scale systems. Algorithms that analyze multi-variate streams of data will be helpful in providing insights into what conditions might cause components to fail, or even to predict when they are likely to fail, so parts of a long-running computation can be transferred to other components without stopping the simulation.

Effective use of experimental/observational facilities and data: The experimental and observational facilities of the DOE Office of Science will enable new science through the analysis of their data and the validation of simulations. The easy availability of compute power in emerging architectures provides several opportunities to make better use of the data from the facilities. We can use more sophisticated statistical techniques to analyze data with noise and missing values, and assign a probability indicating to what extent we can trust the results. In analysis of streaming data, where real-time response is important, we can replace low-flop-count algorithms, which may suffer from numerical instability, by more compute-intensive, but stable, methods. However, there are challenges as well. We will need to develop concurrent algorithms for problems where large amounts of data are analyzed simultaneously, for example, in inverse problems, such as the reconstruction of a three-dimensional object from a multitude of two-dimensional images and in data fusion, where data from several modalities are combined for greater insights. In addition, analysis algorithms will have to be modified to make better use of the characteristics of the architectures, especially the smaller memory, high degree of parallelism, and high cost of data movement.

References:

[ExaDM13] ExaDM: Enabling Scientific Discovery in Exascale Simulations, Web page, ASCR Exascale project, https://computation.llnl.gov/casc/StarSapphire/ExaDM.html. Last accessed April 2013.

[FLS06] K.T. Fang, R. Li, and A. Sudjianto, *Design and Modeling for Computer Experiments*, Chapman and Hall/CRC Press, 2006.