Execution Environments for Big Data: Challenges for User Centric Scenarios

Wolfgang E. Nagel, René Jäkel, Ralph Müller-Pfefferkorn

Center for Information Services and High Performance Computing
Technische Universität Dresden

The past decade has witnessed a tremendous increase in the availability of data for sophisticated scientific analysis. Not only large-scale research projects are able to produce large amounts of raw data but also the availability of affordable instruments for studies of different types has fed the scientific data deluge. Technical improvements on the instrumental level and the availability of inexpensive and ubiquitous sensors give rise to a new level of research opportunities that take massive data sets into account. Even small research communities are now facing the challenge of organizing the ever-growing data bases in their daily work routine. Furthermore, sophisticated studies require the constant refinement and improvement of scientific models and, therefore, largely increase the storage as well as computing requirements.

On the other hand comprehensive scientific investigations require taking the full data life cycle into account, the creation of large data sets towards the generation of results, including the reusability of data for further investigation, and the circumstances under which data are generated. From an architectural viewpoint this includes the consideration of different technical systems and how data are processed on different levels and for different purposes. All these aspects further increase the requirements scientists face in using high performance computing (HPC) systems.

To enable users from different research domains to efficiently use HPC systems for scientific analysis they need assistance in their daily work for processing Big Data scenarios. Usually these massive data are generated close to experiments or sensors in distributed environments and need to be transferred to or close to the computing system. Nowadays the simple transfer approaches of massive raw data towards the HPC system are outdated, since the I/O and network capabilities may be limiting factors even in large scale HPC systems. Early data reduction close to the source of data generation is one of many options for realizing efficient Big Data workflows over the full data life cycle. In the past, special large-scale approaches were highly successful in realizing a massive drop in the amount of data to be processed in the analysis; the high energy physics community is just one example where data reduction in the data acquisition is an inevitable task that renders the experiments possible. Such specialized approaches cannot, however, easily be adapted or be simply reused in other scenarios due to the fact that they are highly target-oriented for the given data acquisition schema and usage scenario. There are manifold scenarios that would highly profit from an adaptable data reduction or feature extraction scheme early on in the analysis chain. For example, in life sciences a new generation of instruments allows the observation of life processes in organisms with sophisticated imaging techniques. Similar techniques from the technical perspective allow in materials science the observation of stress tests of components during experiments to infer new material properties.

Other Big Data scenarios require the integration of variable data sources in the analysis chain, e.g. coming from sensor data or from data sources, whose content may alter over time or with the context of scientific hypotheses. The potential for new insights is high by including publicly available data resources in the scientific analysis, such as coming from Linked Open Data resources or large text corpora. In these distributed environments difficulties arise often from incompatibilities of data formats or meta-data schemes describing the data.

For the different scientific questions pertaining to the different communities the analysis chains might be similar, while a lack of easy to use tools for large scale data processing still hinders the fast development of Big Data scenarios for individual scientific communities.

Regarding the large potential of Big Data scenarios a transformation from static and highly adapted processing chains towards more services, user control, and interaction, including iterative approaches, would substantially increase the outcome of large-scale scientific analysis. This would require the development of easy to use tools on various levels, from data acquisition on the low-end to high-end analysis. From a user perspective, the lack of sophisticated tools and interfaces in the data processing chain is still a major hurdle for realizing complex Big Data scenarios. This is an even more severe phenomenon in the case of small research groups, which have the expertise and the domain-specific knowledge to analyze their data, but lack the knowledge to efficiently integrate larger amounts and/or heterogeneous data sets into their analysis chains and process.

The high potential of easy to use approaches can, for instance, already be observed in the fast dissemination of the MapReduce model. Even though the applicability of MapReduce is still limited to a particular class of analysis approaches, quick insights in large data sets were possible with a reasonable amount of development and adoption time in the analysis. Since then, numerous improvements have been suggested to overcome the limitation of the original MapReduce approach (such as SQL-like extensions, join operations, support for iterative processing). Yet no standards arose from those developments to this day.

To enable the transformation from static approaches to highly adaptable data processing chains, domain experts and HPC experts need to work closely together. This includes primarily the development of:

- Easy to use tools in the data analysis chain,
- Standards for the interoperability of data processing steps over the full data life cycle, from data acquisition to long term storage of results,
- Generalized methods for data and meta-data access and combination
- Integration into easy to use workflows supporting the reusability for other purposes or in different scenarios, and
- Joint developments and collaborative work towards real Big Data services.

A requirement for the successful implementation and provision of such Big Data services is the joint development and collaborative work of computer scientists from the HPC world, data scientists, and the application domain experts. Such an approach may not result in Exascale applications immediately. However, it has a high potential for easy to use services, leading to better usability and will lay the foundation for scalability in the future.

To support such essential collaborative work for developing adaptable Big Data processing techniques and for providing a working environment for domain experts and computer scientists, the Technische Universität Dresden has formed as a leading partner a competence center for Big Data, namely the “Competence Center for Scalable Data Services and Solutions” (ScaDS Dresden/Leipzig). This center is one of two national competence centers for Big Data research in Germany. Domain experts from various research fields bring in their requirements for large-scale data processing and analysis and work closely together with data analysts from computer science to extend current methods of data reduction, the extraction of knowledge from the broad data bases (data mining, machine learning, visual analytics), and aim at a service oriented approach to generalize methods development towards Big Data services.