Big Data Challenges in Big Science

Andreas Heiss¹

Received: 2 August 2019 / Accepted: 9 November 2019 / Published online: 19 November 2019
© Springer Nature Switzerland AG 2019

The term “big data” is currently on everyone’s lips. Is it just a new buzz word that describes something already familiar, or are we really talking about new challenges in the field of computing and data analytics?

“Big data” generally describes large amounts of data, high data rates, or particularly complicated or unstructured data. There are, however, different views on what the “big data” challenge is. In business and industry, overcoming the barriers of proprietary interfaces and protocols to merge data from different sources and generate value from them can be a challenge, e.g., pooling of customer data and data from social networks to gain valuable information for marketing. In science, the sheer volumes of data, extreme data rates, and making the data available to countless scientists scattered around the world in a scalable way are typical challenges of today.

In many scientific disciplines, amounts of data are now produced which no one believed possible a few years ago. For example, genome sequencers and high-throughput microscopes used by biologists or the high-speed cameras at synchrotron beam lines can generate several terabytes of data per day. Users of these instruments can no longer rely on commercial commodity hardware, such as USB hard disks, desktop PCs, or laptops, to store and process the volume of data they produce. Efficient storage and processing instead requires large-scale data management and computing systems, well-thought-out workflows, as well as complex data management and workload management software.

In order to enable scientists to use such large-scale IT infrastructures, and to develop sophisticated software and workflows, a collaboration of computer scientists and domain specialists has proven to be very efficient. In Germany, for example, platforms for such collaborations are the Data Life Cycle Labs [1] (DLCL) and Simulation Labs

implemented by the Helmholtz Association. Computer-affine domain scientists connect their particular scientific disciplines with the computer scientists and operators of corresponding computing and storage systems, providing the necessary domain-specific knowledge to link the two. These successful collaborations frequently lead to joint publications.

Many physicists, however, might think “big data” is just another buzzword, a lot of hype about things they mastered a long time ago, with the enormous data volumes of the LHC experiments and the Worldwide LHC Computing Grid in mind. But physics also faces new challenges that cannot be met using existing methods and established computing models.

In a few years, the high-luminosity LHC (HL-LHC), for example, will deliver an annual data volume of approximately one exabyte, and the antennas of the Square Kilometer Array (SKA) will produce more than 100 terabytes per second on site and virtually online. A special High Performance Computing system will be required to preprocess the data from the antennas before the resulting data stream can be transferred to globally distributed data centres for further processing and analysis. The anticipated technical advances and associated decrease in the price of IT technology in the coming years will not be sufficient to meet the HL-LHC and SKA requirements. Additional large investments in IT resources will also be necessary. Further optimisation of computing models, algorithms and software is essential in order to keep the additional costs for the data processing of these experiments within limits.

The cooperation of computer scientists and domain sciences can lead to major improvements in algorithm engineering and software design, and significantly speed up data processing and analysis. Bringing the right brains together is often more promising than buying more or faster computers.

Other big data challenges for science include the long-term archiving of scientific data (data preservation) and the “open data” concept. Data preservation is crucial for unrepeatable measurements, for example an atmospheric or geological measurement at a certain point in time. If the

✉ Andreas Heiss
andreas.heiss@kit.edu

¹ Karlsruhe Institute of Technology, Steinbuch Centre for Computing, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

measured data is lost or can no longer be read or interpreted, the information is gone forever. “Open data” means scientific data is made freely available to everyone in a form that enables non-experts to use and obtain scientific results from it. This requires an open-data format as well as accurate annotation of data and metadata.

Funding agencies and science ministries have recognised these challenges. Initiatives have been launched and funding is provided to develop and build data and analysis infrastructures on national and European levels.

An example is the Helmholtz Data Federation (HDF) [2], founded in 2017 in Germany. The HDF, in turn, will be a national building block of the European Open Science Cloud (EOSC) [3], an envisaged pan-European federation of computing and data infrastructures and services for science.

R&D projects are promoted and funded with the aim of developing the cross-disciplinary methods, techniques, and software required to utilise these dedicated infrastructures, and to make use of other computing resources, such as commercial clouds and HPC centres. For these R&D projects, the data management, data preservation, and the operation and further development of the data infrastructures requires IT specialists with in-depth knowledge of the various scientific domains.

Particularly in physics, there are many young people who have the necessary knowledge and interest in working on such interdisciplinary projects on the border between physics and computer science. There is, unfortunately, a significant lack of attractive career opportunities in this area.

A classical scientific career in scientific domains like physics or computer science is almost impossible for these interdisciplinary researchers. Their research topics are usually too thematically distant from the respective scientific field and, on the other hand, too specific and application-driven to be recognised in computer science.

Even publishing their work and results is difficult, since there are very few suitable, established journals or conferences.

Many talents, therefore, prefer to go into industry instead of hoping for one of the rare permanent positions at universities or research centres. Large companies and various start-ups are currently looking for big data experts, data scientists and developers, and offer interesting career opportunities. Facing the challenges ahead, it would be fatal if academia was not able to compete in this contest with industry and to bind excellent people and their expertise.

This journal, *Computing and Software for Big Science*, is a step in the right direction and offers IT experts the opportunity publish their valuable work and to establish themselves as scientists. The lack of long-term career opportunities at the interface between computer science and domain sciences is an on-going problem. It is currently the biggest big data challenge in science!

References

1. Data life cycle labs, a new concept to support data-intensive science. [arXiv:1212.5596](https://arxiv.org/abs/1212.5596)
2. <https://www.helmholtz.de/en/research/information-data-science/helmholtz-data-federation-hdf/>. Accessed Sept 2019
3. Realising the European open science cloud. <https://doi.org/10.2777/940154> and <https://ec.europa.eu/research/openscienc/index.cfm?pg=open-science-cloud>. Accessed Sept 2019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.