Scientific Big Data Analytics by HPC

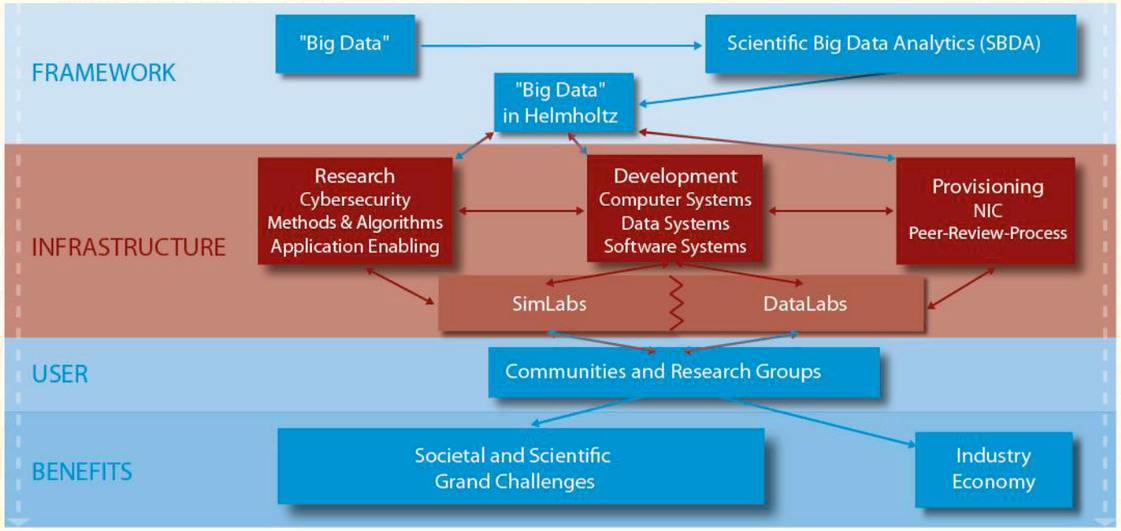

Figure 1: 'Scientific Big Data Analytics' by HPC and the role of large-scale research

Managing, sharing, curating and especially analyzing the ever increasing large quantities of data face an immense visibility and importance in industry and economy as well as in science and research. Commercial applications exploit "Big Data" for predictive analysis, to increase the efficiency of infrastructures, customer segmentation, and tailored services. In contrast, scientific big data allows for addressing problems with complexities that were impossible to deal with so far. This article offers the reader a view of how HPC addresses some of the exponentially growing data challenges that are evident in all areas of science with a particular focus on "Big Data Analytics".

Scientific Big Data Analytics

There is a wide variety of parallel and scalable "Big Data Analytics" technologies and approaches in the field. Researchers are constantly distracted from exploring new and partly innovative technologies and in too many cases lack a sound infrastructure for developing solutions beyond small-scale technology testbeds. The notion "Scientific Big Data Analytics (SBDA)" comprises the work of researchers and engineers, that is based on the creation and analysis of big data sets, and thus relies on the availability of appropriately sized infrastructures as highlighted in red in Figure 1. This is required in order to be competitive in the respective scientific domain.

It is necessary to provide large-scale infrastructures to scientists and engineers of university and research institutes, who perform projects with highest demands on storing, transfer, and processing of data sets. We therefore foresee the importance of establishing a systematic provisioning process in the field of SBDA. This provisioning should be done similar to the provisioning of HPC resources as done for the simulation sciences. This constitutes a key element in the SBDA building block in our framework shown in Figure 1 and explained by Lippert et al. in [2].

Importance of Peer-Review

In order to guarantee that the data analysis achieves the highest scientific quality, it is necessary to apply the principle of autonomous controlling of resource allocation by science in a competitive peer-review process, like it is common practice for international large-scale infrastructures. In addition, the scientific controlling of resource allocations will allow to focus on problem areas that are highly relevant for science and society. The steering process prevents that science gets lost in the details of this industry-driven topical area, with many technologies that are highly relevant only for commercial applications (e.g. recommender engines, shopping basket association rule mining, etc.). Instead, new approaches will be developed and, subsequently have to be translated to economy and industry. Scientific technologies and approaches in this field will mature, leading eventually to community-approved codes to tackle an ever increasing amount of research questions. Hence, the effective usage of the SBDA infrastructure as illustrated in Figure 1 is ensured by a scientific peer-review. These science-led processes not only guarantee the most beneficial usage of the infrastructure, but also steer their evolution and focus through the involvement of research communities in key areas of science and engineering.

Role of the NIC

The John von Neumann Institute for Computing (NIC) has established a scientific peer-review process for the provision of supercomputing cycles at the Jülich Supercomputing Centre (JSC).

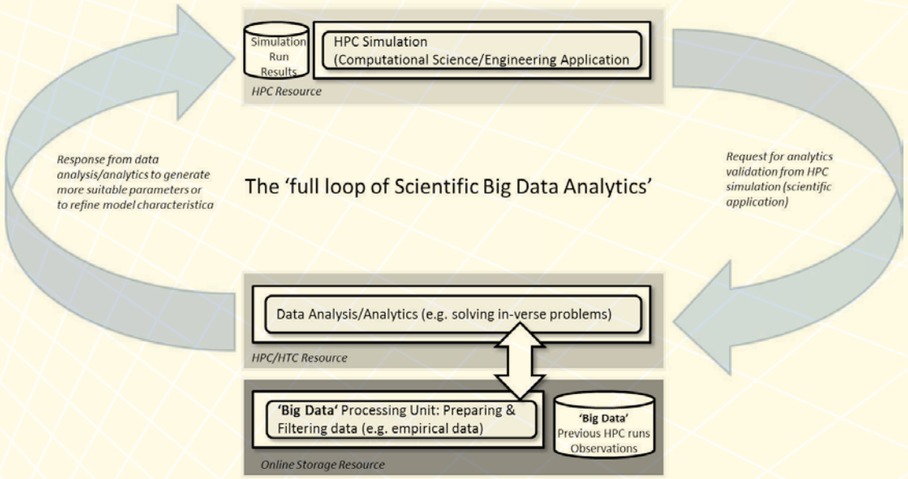

Figure 2: The full productive loop of scientific big data analytics by HPC

Scientists and researchers who apply for computing time on infrastructure resources are supported by a continuously growing number of domain-specific simulation laboratories (SimLabs) at JSC. The SimLabs offer support on a high level and push forward research and development in co-design together with specific scientific communities that take advantage of HPC techniques in parallel applications.

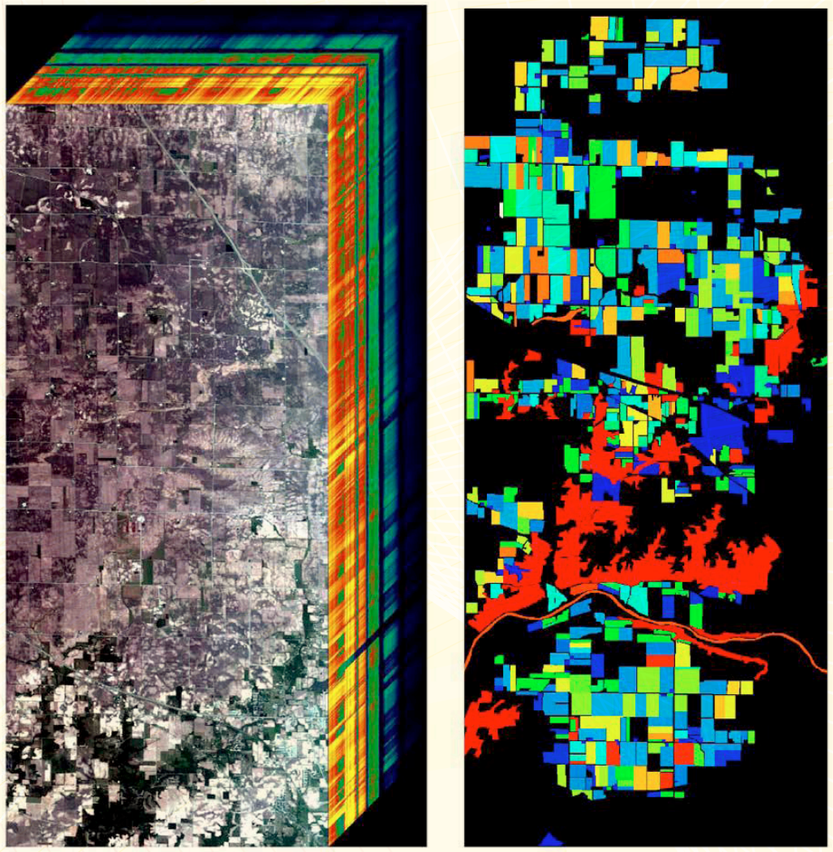

Figure 3: AVIRIS Indian Pines (a) image cube representation and (b) ground reference

The NIC allocation principles served also as a blueprint for the allocation of computing time in the Gauss Centre for Supercomputing (GCS). Given NIC's strong experience and its scientific advancement, it is natural to apply the NIC provisioning concept to data infrastructures and analysis resources as well. This kind of provisioning together with the activities in the Helmholtz programme "Supercomputing & Big Data", will promote research and development for supercomputing and data infrastructures. It is the heart of truly innovative SBDA and ensures scientific advancement.

In order to gain better insights into the demand by communities and requirements, the JSC has performed an initial step towards implementing principles as proposed in this article. The importance of data analytics, management, sharing, and preservation of very big, heterogeneous or distributed data sets from experiments, observations and simulations is of increasing significance for science, research and industry. This has been recognized by many research institutions, among them leading HPC centers. They want to advance their support for researchers and engineers using SBDA by HPC.

First Experience revealed

For the first time NIC invited Expressions of Interest (EoI) for SBDA projects using HPC to identify and analyze the needs of scientific communities. The goal is to extend and optimize the HPC and data infrastructures and services in order to provide optimal methodological support. As described above, a peer-review was performed on the EoI submissions in a similar manner as known from HPC calls. First experiences were presented as posters at the last NIC symposium in Jülich [1]. They clearly demonstrate in which areas SBDA by HPC is of major importance.

EoI submissions have been received from the field of Biology (e.g. mining of molecular structure data) and Neurosciences (e.g. deep learning analysis of brain images), from the field of statistical turbulence research and from traditional HPC simulation communities in the context of earth sciences (e.g. SimLab Climate and SimLab Terrestrial Systems). One identified common scheme we refer to as the "full loop of SBDA" is illustrated in Figure 2. Solving a wide variety of inverse problems can actually lead to better algorithms for simulation sciences that in turn then deliver more accurate models to understand our world. In order to establish this strong "productive loop" between HPC simulations and more data-intensive applications, a strong foundational infrastructure is required.

One Example of HPC Impact

Recent advances in remote sensor and computer technology are substituting the traditional sources and collection methods of data, by revolutionizing the way remotely sensed data are acquired, managed, and analyzed. The term "remote sensing" refers to the science of measuring, analyzing, and interpreting information about a scene acquired by sensors mounted on board of different platforms for Earth and planetary observation.

Our motivation is driven by the needs of a specific remote sensing application dataset as shown in Figure 3 based on data from AVIRIS. It raises the demand for technologies that are scalable with respect to big data and thus this application represents one example of a SBDA project that requires HPC. One of the main purposes of satellite remote sensing is to interpret the observed data and classify meaningful features or classes of land-cover types. In hyperspectral remote sensing, images are acquired with hundreds of channels over contiguous wavelength bands, providing measurements that we consider as concrete big data in this example. The reasoning includes not only a large data volume but also a large number of dimensions (i.e., spectral bands). A deeper introduction to the field and its application to HPC using parallel and scalable support vector machines on HPC resources are in G. Cavallaro et al. in [3].

We highlight in this article the impact of HPC since parallelization techniques lead to significant speed-ups for the cross-validation and for each training and testing process of the aforementioned dataset. To provide an example during cross-validation, HPC SBDA techniques have been able to reduce the time from 14.41 minutes (Matlab) to 1.02 minutes using just the best parameter set. Since this "best parameter" needs to be searched a typical grid search is performed consisting of many runs with different parameter sets. In this context, parallel and scalable SBDA by HPC methods have been able to reduce the time to solution from roughly 9 hours to 35 minutes.

More notably, this is achieved by maintaining the same accuracy as achieved when performing the processes with serial tools (e.g. Matlab, R, libSVM). In the majority of cases, the minimal training and testing time was around 1 min that still can be considered as an interactive experience thus enabling remote sensing scientists to perform easier and faster experiment with different techniques (e.g., applying quick parameter variations of feature-extraction techniques). The parallel and scalable community code for support vector machines and its feature engineering approach based on mathematical morphology are now maintained by JSC and starting to be used for other application areas in science (e.g. neuroscience images) and industry (e.g. welding image analysis).

References

[1] http://www.john-von-neumann-institut.de/nic/EN/News/Symposium/NIC-Symposium-2016/PosterSession/_node.html

[2] **Lippert, Th., Mallmann, D., Riedel, M.,** Scientific Big Data Analytics by HPC, in: NIC Symposium 2016, Publication Series of the John von Neumann Institute for Computing (NIC), NIC Series 48, ISBN 978-3-95806-109- 5, pp. 1 - 10

[3] **Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J.A., Plaza, A.,** On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Issue 99, pp. 1-13, 2015

contact:
Morris Riedel,
m.riedel@fz-juelich.de

- Morris Riedel[1,2]
- Thomas Lippert[1]
- Daniel Mallmann[1]
- Gabriele Cavallaro[2]

[1]Jülich Supercomputing Centre (JSC), Germany

[2]University of Iceland, Reykjavik

Improved Safety for disabled People

In order to help disabled people gain easier access, many buildings, e.g. sheltered workshops, as well as other venues are now accessible barrier-free. But how can such places be evacuated effectively if disabled people are involved? How can stakeholders be trained, evacuation plans be adjusted, and facilities be designed to make these sites safer? To answer these questions, the joint project "Safety for people with physical, mental or age-related disabilities" (SiME) has been funded in the context of the "Research Programme for Civil Security" by the German Federal Ministry of Education and Research (BMBF). The three-year project began in February 2016 and is coordinated by the Federal Institute for Materials Research and Testing (BAM). Other partners are Otto-von-Guericke University Magdeburg, Hochschule Niederrhein, Werkstatt Lebenshilfe, PTV Transport Consult GmbH, and Forschungszentrum Jülich.

The simulation of the process of evacuating a building enables the identification of bottlenecks or lack of assistance and the calculation of the evacuation time. For this purpose, parameters of the realistic movement of persons involved are needed, but such data are as yet only available for people with unrestricted mobility. In SiME, the team from JSC will execute parameter studies for mixed traffic, i.e. for people with and without disabilities, and also analyse the process of movement of disabled people, e.g. transfer from a wheelchair to an evacuation chair, during an evacuation process.

SiME
Sicher miteinander leben

The intended parameter studies have two focuses:

1. Evacuation of disabled people from their daily environment like sheltered workshops or residences

2. Collective movement of people with and without disabilities in venues for large public events.

For the extraction of trajectories of individual participants methods for visual sensors developed during previous projects like Hermes and BaSiGo can be used. But people with a low height inside a dense gathering like persons using wheel chairs will often be occluded so that sensors have to be examined and new methods be developed to track also covered people. After sensor fusion the movement of every single person will be available for further analysis.

With the collected data, more reliable models could be developed to simulate the evacuation of sheltered workshops or homes for people with disabilities. A simulated forecast of the dynamic inside gatherings including people with limited mobility will be more realistic.

contacts:
Maik Boltes,
m.boltes@fz-juelich.de,
Stefan Holl,
st.holl@fz-juelich.de

- Maik Boltes
- Stefan Holl

Jülich Supercomputing Centre (JSC), Germany