René Jäkel*, Eric Peukert, Wolfgang E. Nagel, and Erhard Rahm

ScaDS Dresden/Leipzig – A competence center for collaborative big data research

https://doi.org/10.1515/itit-2018-0026
Received October 1, 2018; revised October 16, 2018; accepted October 18, 2018

Abstract: The efficient and intelligent handling of large, often distributed and heterogeneous data sets increasingly determines the scientific and economic competitiveness in most application areas. Mobile applications, social networks, multimedia collections, sensor networks, data intense scientific experiments, and complex simulations nowadays generate a huge data deluge. Nonetheless, processing and analyzing these data sets with innovative methods open up new opportunities for its exploitation and new insights. Nevertheless, the resulting resource requirements exceed usually the possibilities of state-of-the-art methods for the acquisition, integration, analysis and visualization of data and are summarized under the term big data. ScaDS Dresden/Leipzig, as one Germany-wide competence center for collaborative big data research, bundles efforts to realize data-intensive applications for a wide range of applications in science and industry. In this article, we present the basic concept of the competence center and give insights in some of its research topics.

Keywords: big data, data-intensive computing, data integration, collaborative research, knowledge extraction, visualization, high-performance computing

ACM CCS: Information systems → Data management systems, Information systems → Information systems applications, Computer systems organization → Architectures

1 Introduction

Nowadays, digitization permeates all areas of life and produces an ever-increasing amount of digital data. Science and research increasingly contribute to this data deluge [1, 2] through data-intensive experiments, complex simulations, the eruption of interconnected sensor networks, and the opening of new data sources, such as digital archives, serving historical data. With the advent of ubiquitous sensor-based streaming data, especially in the context of the IoT, additional challenges are brought up by handling data with a predominantly temporal component. Furthermore, digitization as a driver of data generation increasingly determines business processes and also starts to influence many aspects in private life of individuals.

However, only if data can be processed efficiently and with intelligent methods, data can be the driving force for gaining knowledge through analysis. In many cases, science and industry face unprecedented challenges to cope with very large and distributed data sets that are in addition very complex in their semantics and heterogeneous in their formats. These challenges are referred to as big data, and it can be observed that capabilities needed to process these data sets exceed well established and state-of-the-art methods of data processing.

To address the data-driven challenges that are usually very different between applications, an intensive and collaborative exchange between domain scientists and data analysis experts is essential to provide insights and solutions for a given challenge. This observation was the starting point to establish a competence center for big data to intensify research activities and connect to application areas for joint developments to address a wide range of big data issues. Since October 2014, the project partners Leibniz Institute of Ecological Urban and Regional Development (IÖR), Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG), Technische Universität Dresden, and University of Leipzig combine their research and application development efforts in a competence center for collaborative big data research funded by the Federal Ministry of Education and Research (BMBF), which operates Germany-wide. Our competence center, the "Competence Center for Scalable Data Services and Solutions – ScaDS Dresden/Leipzig", has been successfully implementing a concept for cooperative research on big data

***Corresponding author: René Jäkel,** Center for Information Services and High Performance Computing (ZIH), Technische Universität Dresden, D-01062 Dresden, Germany, e-mail: rene.jaekel@tu-dresden.de, ORCID: http://orcid.org/0000-0001-6260-5222
Eric Peukert, Big Data Kompetenzzentrum ScaDS Dresden Leipzig, Ritterstraße 9-13, 2.OG, 04109 Leipzig, Germany, e-mail: peukert@informatik.uni-leipzig.de
Wolfgang E. Nagel, Center for Information Services and High Performance Computing (ZIH), Technische Universität Dresden, D-01062 Dresden, Germany, e-mail: wolfgang.nagel@tu-dresden.de
Erhard Rahm, Universität Leipzig, Fakultät für Mathematik und Informatik, Augustusplatz 10, 04109 Leipzig, Germany, e-mail: rahm@informatik.uni-leipzig.de

technologies and their interdisciplinary application in science and industry. To fully comprehend requirements and methods needed for solving big data related challenges, our approach is to establish expert groups with a specific project related focus. This includes data scientists, which are specialized in adopting state-of-the-art technologies for a given application, but additionally, integrate expert knowledge from the methodical and technological computer science perspective. This extends the range of available expertise to tackle given challenges coming from diverse application areas.

1.1 Project overview

At the two university sites Dresden and Leipzig, the national competence center investigates many different aspects processing large data sets in sciences and industry. New methods and solutions have been developed to investigate large and complex data sets coming from a variety of application areas. Within the competence center, five domain science areas are tied together. They bring in their domain specific requirements to process large data sets, but also advance computer science research on methods for data-intensive applications, see Fig. 1. These applications come from life sciences, material and engineering sciences, digital humanities, environmental and geosciences, as well as from business and industry. Our research is driven by the observation that big data solutions can only be developed by adapting the whole data life cycle, and by providing access to modern data processing and computing architectures. Initially, data requires to be integrated from different sources with high quality, needs to by enriched, and further processed to extract new information. Finally, data has to be presented to the user via methods from visual analysis in order to integrate the domain scientist into the analysis process (see also Fig. 1 for methodical research topics).

Besides the domain and computer science methods, a key success factor of ScaDS Dresden/Leipzig is the foundation of a service center, that bundles interdisciplinary research activities. The service center offers a single point of contact for research and industry and coordinates methodical research and application research at both university sites. In the last years, numerous cooperations could be established with a large number of scientific institutions and companies. The service center does not only rely on experienced data scientists to develop data-driven applications, it also mediates competencies that are already present in the center so that it can support a broad range of application domains in the project. Often, existing and proven solutions, such as the efficient use of data integration techniques or the application of high performance computing for large compute-intense scenarios, can be transferred between application areas. To disseminate developed methods and spread best practices dealing with big data, the service center organizes trainings and workshops. This way, amongst others, we were able to establish an international summer school series, and an industry workshop series called “Big Data in Business” (BIDIB).

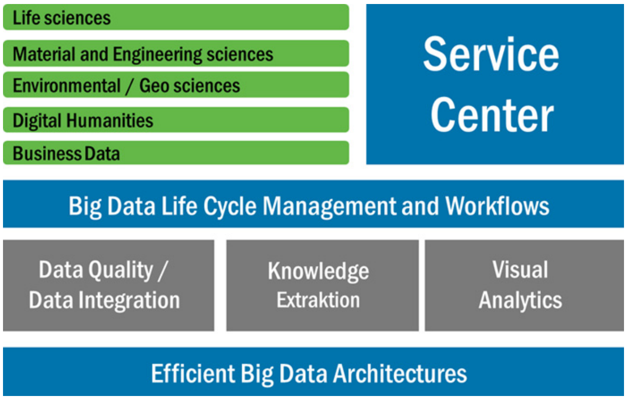
Life sciences
Material and Engineering sciences
Environmental / Geo sciences
Digital Humanities
Business Data
Service Center
Big Data Life Cycle Management and Workflows
Data Quality / Data Integration
Knowledge Extraktion
Visual Analytics
Efficient Big Data Architectures

Figure 1: Overview about application areas and methodical topics of ScaDS Dresden/Leipzig.

The remainder of this contribution is structured as follows. Section 2 discusses requirements and possibilities to provide large-scale architectures for efficient data-driven applications. Section 3 presents the service center approach of ScaDS Dresden/Leipzig and introduces some research highlights, especially focusing on interdisciplinary efforts. Finally, section 4 summarizes and gives a short outlook about future activities of the center.

2 Environments and infrastructures for big data analysis

Processing large volumes of data, most of which come from heterogeneous data sources, requires architectures that are both flexible and secure, enabling fast processing of the data. Therefore, such architectures must support different applications through customized hardware and software configurations and should be designed for parallel processing [3].

Big data applications cover a wide range of requirements. These usually not only affects the analysis, but also for the necessary preprocessing steps in data integration and analysis as well as methods for user interaction and evaluation. These steps are typically modeled as complex task chains (workflows) that need to be mapped and executed within adequate execution environments. In particular in exploratory analyses, where optimal parameter settings (i. e. for machine learning algorithms) are a-priory unknown, new requirements arise, which are different from batch processing. In such scenarios, it is important to provide access to infrastructures that provide a versatile software stack with extended analytics functionality as well as access to fast IO capabilities.

Modern computer systems, and in particular high-performance computing (HPC) architectures, are an ideal basis for providing customized and high-performance work environments for various application requirements [4]. In the past, resource provisioning was done by manual interaction between the user and the resource manager of the HPC system. Especially in the big data ecosystem, with its often explorative analysis and heterogeneous requirements regarding the analysis tools, the HPC system integration presented an additional challenge from the users perspective [5, 6]. Unlike shared-nothing systems, which were predominantly used in the first and second development stages of big data frameworks as standard execution environments, modern HPC systems offer high-performance environments for fast and scalable execution of data-intensive applications. In particular, the use of fast network connections to exchange data between compute nodes and the access to different storage technologies with a corresponding IO hierarchy is crucial to support complex applications with iterative analysis models.

On the other hand, the big data ecosystem was driven by generic frameworks such as Apache Flink, Spark, or Hadoop, which are widely used in industry to develop big data applications, mainly because of better APIs, with which the basic building blocks and integrated parallelism can be combined. Partners of the competence center have access to various clusters providing different architectures, such as an Shared-Nothing cluster operated by the center, cloud-based resources for service-orientated developments, and HPC systems. This allows to investigate sophisticated software stacks as well as computing environments for complex analysis scenarios. Especially in interdisciplinary settings, where scientists with different backgrounds work on application specific challenges, the ability to conduct explorative studies on flexible hardware architectures is essential to gain fast insights in data and compare different analytical scenarios.

The efficiency and scalability of data-driven analysis has been significantly improved by further extending and improving the existing HPC infrastructure in Saxony operated by the ZIH of TU Dresden. A further enhancement of the HPC infrastructure to strengthen the compute and storage capabilities is currently ongoing. Significant investments have been done to equip the system with hierarchical storage components with more than 2 PB flash-based memory with a bandwidth of about 2 TB/s, where the flash memory is flexibly configurable and can be used at all existing computer nodes. Large amounts of research data can be stored in an object store of 10 PB over a longer period of time. The existing compute capacity will be extended by a large number of Power 9-nodes, especially suited for scalable machine learning tasks, which will also be connected to the flexible storage system [7]. Together with the existing components, this creates a system, in which different technologies can be interconnected in a flexible way to set up efficient and customizable research infrastructures.

3 Big data services and solutions

In the first phase of ScaDS Dresden/Leipzig (10/2014 – 09/2018), the service center was successfully established and acts as single-point-of-contact for users in the field of data-intensive computing / big data. A multitude of R&D activities have already been successfully managed by the service center in various scientific disciplines as well as brought in by customers from industry. The service activities follow basically three lines of action:

- evaluation, analysis, and deployment support to realize big data developments
- assistance in utilizing analytics software stacks and access to clusters
- training and education activities for science and industry

In the following, some selected research and application project results of ScaDS Dresden/Leipzig are shortly introduced to give an overview about the competence spectrum and to demonstrate the flexibility of its interdisciplinary service center approach.

3.1 Detection of settlement structures in topographic maps

Topographic maps are an important source of information for geoscience research and for many practical applications in the field of spatial, urban, and landscape planning [9]. An important aspect is the distribution of open land and settlement areas, as well as their historical development, which can be investigated by the analysis of current and historical map material.

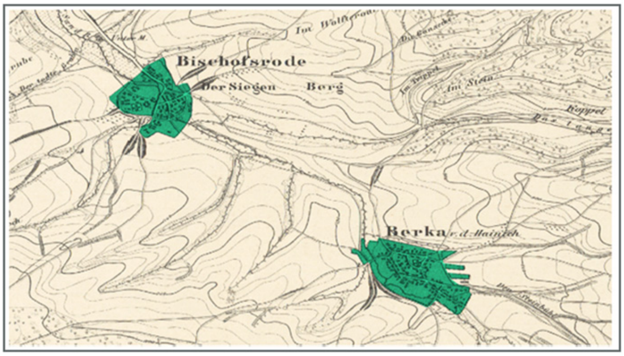

Figure 2: Automatic settlement detection in historical geographic maps [8] using binary segmentation techniques. The recognized settlements structures are visualized as overlay in green.

In an interdisciplinary working group of environmental scientists, computer vision, and the HPC experts, a pipeline had been established to identify settlement structures as regions of interest in historical topographic maps by incorporating binary segmentation techniques into the analysis workflow.

In a first attempt random forest and conditional random field methods from machine learning were used to segment large amounts of data (scanned maps) efficiently on parts of an HPC cluster. By mapping the analysis requirements to the parallel infrastructure, the training and analysis of typical input sets (5700 maps with in total 800 GB image data) was significantly faster to be executed, in three hours compared to days on typical workstations [8]. The outcome are highlighted areas in the maps visualizing the segmented settlement structures (see Fig. 2). In principle this approach can be extended to other unstructured information encoded in the maps. Currently this approach is extended with further machine learning methods to improve the quality of training data on different variants of maps and on automatic detection of writings and their meaning on the historic maps.

3.2 Canonical text services for textual analysis

The Digital Humanities represent an important area of application since interactive text analysis and annotation procedures for typical eHumanities applications are of particular importance for the text-oriented intellectual sciences. The field of application is typically very broad and include the linking of multi-lingual and multimodal resources, network analyses with named entities and their visualization, identification of topics and sentiments about space and time, and citation analyses over different text corpora, to name only a few prominent areas. Adopting those methods enables a spatio-temporal contextualization as well as a linguistic evolutionary classification of works, which adds a completely new dimension to the interpretation of individual works and discuss them in their contextual frameworks.

For new big data applications in the Digital Humanities, the hierarchical structure of texts (chapters, sections, sentences, phrases, words) is used as a basis to enable distributed processing, the representation of data at different levels, and to efficiently compare data at different levels. For this purpose, data is organized following the CTS (Canonical Text Services) standard, and in the competence center a novel CTS service was introduced for mass processing of text data. Based on the existing CTS protocol, a high-performance implementation of these services was implemented. For evaluation and practical operation, a large number of data sets were collected and made publicly available as CTS instances for further use. Among the highlights are data of the German Text Archive in two different variations, the Parallel Bible Corpus, Digital Muqtabas, the data set of the Textgrid Project, Adrien Barbaresis German Political Speeches Corpus, and the multilingual transcripts of the TED Talks. The data of the PBC, DTA and Textgrid alone reach a volume of more than half a billion words, and the total volume of the repository is about one billion words.

The approach developed in ScaDS Dresden/Leipzig was also integrated in the tool set of the “European Research Infrastructure for Language Resources and Technology”, the CLARIN[1] project [10]. CLARIN makes digital language resources available to scholars, researchers, students, and citizen-scientists from all disciplines, especially in the humanities and social sciences. It offers longterm solutions and technology services for deploying, connecting, analyzing, and sustaining digital language data and tools and supports scholars who want to contribute to a truly multilingual European Research Area.

3.3 Visualization techniques for large-scale data

Data analysis guided by visualization techniques is a powerful tool to provide insights into complex data structures of very large volume. Depending on the type of data and application area, adapted visual representations of data are often crucial. In particular, if data needs to be aggre-

[1] CLARIN project homepage: https://www.clarin.eu/

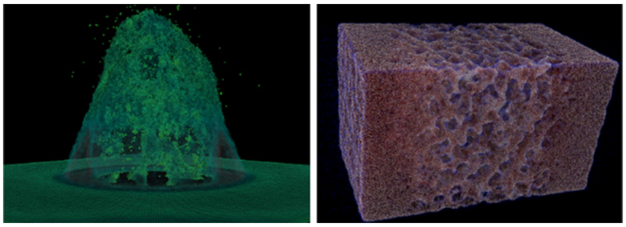

Figure 3: Visulization results [11] using extended transparency calculations to display underlaying structures.

gated for certain analyses or if the quantity of elements to be visualized is overlaying relevant information, which can not directly be presented to the user, specific visualizations need to be developed.

Within the framework of ScaDS, novel focus and context methods for large particle data sets have been developed. Such data is often generated by simulation for example in thermodynamics, molecular dynamics, or material sciences. The interaction of a large number of independent elements, the particles, is simulated in several time steps. Individual particles, for example, represent atoms or grains of sand. They form macro-structures composed of thousands to millions of particles in each time step. Adapted efficient visualizations are necessary for the understanding of these complex structures and processes. In particular, the problem of masking was addressed and the presented solution is based on an extended transparency calculation as well as the determination of approximative global illumination for particles in three-dimensional space [11], as shown as an example in Fig. 3. A focus and context method for multidimensional, abstract particle and point data was derived from this initial approach and is based on a generalization of depth blur for the multidimensional case [12].

3.4 GRADOOP – A distributed graph processing framework

The analysis and integration of very large network datasets is becoming increasingly valuable, for example to gain insights from logistics, business processes, social networks, or protein interaction networks in bioinformatics. By representing network data as a graph, complex relationships between heterogeneous data objects can be analyzed. In big data applications, efficient analysis and mining on such graphs is challenging and is currently not well supported by traditional graph databases or distributed graph processing systems. For this reason, ScaDS Dresden/Leipzig developed a comprehensive new graph analysis system GRADOOP (Graph Analytics on Hadoop) that is available as Open Source.

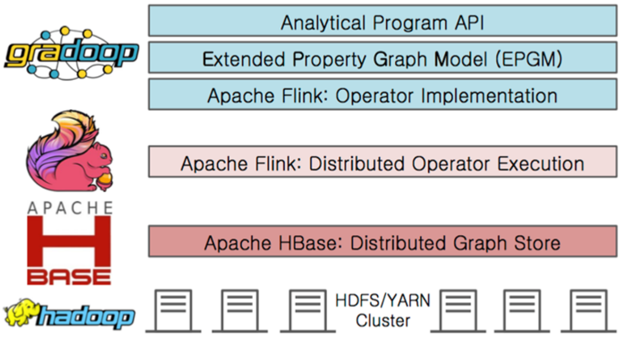
gradoop
Analytical Program API
Extended Property Graph Model (EPGM)
Apache Flink: Operator Implementation
Apache Flink: Distributed Operator Execution
APACHE
HBASE
Apache HBase: Distributed Graph Store
hadoop
HDFS/YARN
Cluster

Figure 4: Fundamental concepts of GRADOOP and its connection to the Apache software stack.

GRADOOP is characterized by a flexible graph data model based on extended property graphs and provides a variety of powerful operators (including pattern matching, graph grouping and aggregation) [13] as well as a library of graph mining algorithms [14]. The GRADOOP framework (see architectural overview in Fig. 4) allows data scientists and analysts to express complex graph analysis tasks using simple and intuitive analytical workflows as we recently presented at VLDB [15]. The operators and algorithms are implemented on top of Apache Flink and can therefore be executed on shared-nothing clusters to be able to process large amounts of data in parallel. The GRADOOP system has already been put into practice in the context of two running third-party funded projects with SMEs. In the second funding phase of ScaDS, which has started just now, the technology for graph data analysis will be expanded, especially in the area of data integration as well as for supporting dynamic graphs. Various operators for data integration, duplicate detection and transformation are currently developed on the basis of GRADOOP and Apache Flink. These operators make it possible to implement complex and scalable Graph ETL pipelines that are modeled as workflows and that are executed within Apache Flink.

The Grouping concepts of GRADOOP were contributed to the Graph-Library of the Apache Flink project, and GRADOOP is listed as a third party extension on the Apache Flink pages.

3.5 Automatic execution of data-intensive workflows on HPC infrastructures

In a scientific cooperation with the KNIME project, methods have been developed in the field of Life Sciences that enable the connection of the KNIME workflow modeling environment [17] to the HPC infrastructure. Thus, a direct possibility for interaction out of the workflow environment with the high performance computer has been realized. This makes it possible to execute entire workflows or parts of the workflow on the connected HPC infrastructure without any HPC knowledge of one's own. The connection was made via the HPC middleware UNICORE and its service interfaces [18]. As an example for real analyses of the partner MPI-CPB it could be shown by employees of the service center and the ZIH of the TU Dresden that a high data parallel execution of the workflow on HPC is possible. Compared to the usual execution of the application workflow limited to one workstation, the automatic execution on HPC enabled an acceleration by a factor of $>$ 200 [16], whereby a special example application processed a total of 1.8 TB of input image data and more than 7.5 million individual files were processed automatically.

4 Summary and outlook

As the previous sections have shown by a few examples, the developed applications span a wide range of different scientific areas and use different methods in order to gain new insight into available data. Although the developed solutions of data-intensive applications are different in their implementation and use of methods, they share some common features. Data-intensive applications represent complex workflows rather than individual applications. In future, stronger interactive and exploratory use will characterize such workflows, which will be enriched partly with machine learning techniques as well as modern analytics functionalities, such as e. g. implementing deep learning methods into the whole analytics chain.

After its initial successful four years phase and convincing results, the competence center was able to receive full funding for additional three years starting in October 2018. In the next stage, novel big data research topics will be worked on and the concept of strong cooperation between domain and computer science research mediated through a service center will be extended with the goal to achieve long-term continuation.

Acknowledgment: We thank the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for generous allocations of computer time. Furthermore the authors would like to express their gratitude to Hendrik Herold, Florian Jug, Jochen Tiepmar, Joachim Staib, Peter Winkler, Richard Grunzke and Bernd Schuller for providing insights into their research fields.

Funding: This work was supported by the German Federal Ministry of Education and Research (BMBF, 01IS14014A-D) by funding the competence center for Big Data "ScaDS Dresden/Leipzig".

References

1. D. Gershon, *Dealing with the data deluge*. Nature 416 (2002), no. 6883, p. 889–891.
2. G. Bell, T. Hey, and A. Szalay, *Beyond the data deluge*, Science 323 (2009). no. 5919, p. 1297–1298.
3. M. Asch et al.*Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry*. The International Journal of High Performance Computing Applications, vol. 32, (2018), no. 4, p. 435–479.
4. G. Fox, J. Qiu, S. Jha, S. Ekanayake, and S. Kamburugamuve, *Big Data, Simulations and HPC Convergence*. In: 7th Workshop on Big Data Benchmarking, 2015.
5. R. Jäkel, R. Müller-Pfefferkorn, M. Kluge, R. Grunzke, and W. E. Nagel, *Architectural implications for exascale based on big data workflow requirements*. In: Big Data and High Performance Computing, vol. 26, Advances in Parallel Computing, IOS Press, 2015, p. 101–113.
6. W. E. Nagel, R. Jäkel, and R. Müller-Pfefferkorn. *Execution Environments for Big Data: Challenges for User Centric Scenarios*, BDEC white paper, Barcelona 2015.
7. Press release (German, July 2018): *Fusion von HPC und Data Analytics*, https://tu-dresden.de/zih/die-einrichtung/news/fusion-von-hpc-und-data-analytics-hpc-da.
8. D. Schemala, D. Schlesinger, P. Winkler, H. Herold, and G. Meinel. *Semantic segmentation of settlement patterns in gray-scale map images using RF and CRF within an HPC environment*. In: Proceedings of the GEOBIA 2016, Enschede, Holland.
9. H. Herold, R. Hecht, and G. Meinel. *Old maps for land use change monitoring – analysing historical maps for long-term land use change monitoring*. In: Proceedings of the International Workshop Exploring Old Maps (EOM 2016), University of Luxembourg, 2016, p. 11–12.
10. J. Tiepmar, T. Eckart, D. Goldhahn, C. Kuras. *Integrating Canonical Text Services into CLARIN's Search Infrastructure*, Linguistics and Literature Studies, vol. 5, (2017), p. 99–104.
11. J. Staib, S. Grottel, and S. Gumhold. *Visualization of Particle-based Data with Transparency and Ambient Occlusion*, Computer Graphics Forum, vol. 34, p. 151–160.
12. J. Staib, S. Grottel, and S. Gumhold. *Enhancing Scatterplots with Multi-Dimensional Focal Blur*, Computer Graphics Forum, vol. 35, p. 11–20.
13. M. Junghanns, A. Petermann, K. Gomez, E. Rahm. *Distributed Grouping of Property Graphs with GRADOOP*. In: Proc. Datenbanksysteme für Business, Technologie und Web (BTW) 2017, 3 2017.

14. A. Petermann, M. Junghanns, S. Kemper, K. Gomez, N. Teichmann, and E. Rahm, *Graph Mining for Complex Data Analytics*. In: ICDM, 2016.
15. M. Junghanns, M. Kießling, N. Teichmann, K. Gomez, A. Petermann, E. Rahm, *Declarative and distributed graph analytics with GRADOOP*, PVLDB, vol. 11, (2018), no. 12, p. 2006–2009.
16. R. Grunzke, F. Jug, B. Schuller, R. Jäkel, G. Myers, and W. E. Nagel. *Seamless HPC Integration of Data-intensive KNIME Workflows via UNICORE*. In: Desprez F. et al., (eds), Euro-Par 2016: Parallel Processing Workshops, Euro-Par 2016. Lecture Notes in Computer Science, vol. 10104. Springer, Cham.
17. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. *KNIME – the Konstanz information miner: version 2.0 and beyond*. SIGKDD Explor. Newsl. 11 (November 2009), no. 1, p. 26–31.
18. K. Benedyczak, B. Schuller, M. Petrova-ElSayed, J. Rybicki, R. Grunzke. *UNICORE 7 – Middleware Services for Distributed and Federated Computing*. In: International Conference on High Performance Computing & Simulation, HPCS2016, Innsbruck, Austria, IEEE 2016, p. 613–620.

Bionotes

Dr. René Jäkel
Center for Information Services and High Performance Computing (ZIH), Technische Universität Dresden, D-01062 Dresden, Germany
rene.jaekel@tu-dresden.de

René Jäkel is Management Director of the national Big Data Competence Center ScaDS Dresden/Leipzig. He studied physics and finished his PhD in hadron physics at the TU Dresden. His research interests cover analytics pipelines for big data applications on High Performance Computing systems and the performance characteristics of analytics applications, especially in presence of data-intensive settings. He heads the Service Center of the Competence Center as a central contact point for research requests and collaborations from industry and science and is active in numerous activities in education.

Dr. Eric Peukert
Big Data Kompetenzzentrum ScaDS Dresden Leipzig, Ritterstraße 9-13, 2.OG, 04109 Leipzig, Germany
peukert@informatik.uni-leipzig.de

Eric Peukert coordinates service center activities at the University of Leipzig. He studied Computer Science and Media at the TU Dresden and worked at SAP Research in the field of data integration and schema mapping within various BMBF and EU research projects. After completing his doctorate at the University of Leipzig and two more years with SAP, he now coordinates the activities of the center in Leipzig with a special focus on industry contacts and cooperations. His research includes big data technologies, data integration and learning-based duplicate detection methods.

Prof. Dr. Wolfgang E. Nagel
Center for Information Services and High Performance Computing (ZIH), Technische Universität Dresden, D-01062 Dresden, Germany
wolfgang.nagel@tu-dresden.de

Wolfgang E. Nagel holds the chair for computer architecture at TU Dresden and is director of the Center for Information Services and HPC (ZIH). His research covers programming concepts and software tools to support the development of scalable and data intensive applications, analysis of computer architectures, and development of efficient parallel algorithms and methods. Prof. Nagel is chairman of the Gauß-Allianz e.V. and member of the international Big Data and Extreme-scale Computing (BDEC) project. He is leading the Big Data competence center ScaDS – Competence Center for Scalable Data Services and Solutions Dresden/Leipzig.

Prof. Dr. Erhard Rahm
Universität Leipzig, Fakultät für Mathematik und Informatik, Augustusplatz 10, 04109 Leipzig, Germany
rahm@informatik.uni-leipzig.de

Erhard Rahm is full professor for databases at the computer science institute of the University of Leipzig, Germany. His current research focuses on big data and data integration. His research on data integration and schema matching has been awarded several times, in particular with the renowned 10-year best-paper award of the conference series VLDB (Very Large Databases) and the Influential Paper Award of the conference series ICDE (Int. Conf. on Data Engineering). Prof. Rahm is deputy scientific coordinator of the new German center of excellence on Big Data ScaDS Dresden/Leipzig.