Datenbank Spektrum
https://doi.org/10.1007/s13222-018-00303-6

FACHBEITRAG

Big Data Competence Center ScaDS Dresden/Leipzig: Overview and selected research activities

Erhard Rahm[1] · Wolfgang E. Nagel[2] · Eric Peukert[1] · René Jäkel[2] · Fabian Gärtner[1] · Peter F. Stadler[1] · Daniel Wiegreffe[1] · Dirk Zeckzer[1] · Wolfgang Lehner[2]

Received: 5 November 2018 / Accepted: 17 December 2018

Abstract
Since its launch in October 2014, the Competence Center for *Sca*lable *D*ata Services and *S*olutions (ScaDS) Dresden/Leipzig carries out collaborative research on Big Data methods and their use in challenging data science applications of different domains, leading to both general, and application-specific solutions and services. In this article, we give an overview about the structure of the competence center, its primary goals and research directions. Furthermore, we outline selected research results on scalable data platforms, distributed graph analytics, data augmentation and integration and visual analytics. We also briefly report on planned activities for the second funding period (2018-2021) of the center.

Keywords Big Data · Data science · Data management

1 Introduction

The Competence Center for *Sca*lable *D*ata Services and *S*olutions (ScaDS) Dresden/Leipzig [25] is one of two German Big Data competence centers that the Federal Ministry of Education and Research (BMBF) established in 2014 after a competitive selection process (the second center is the Berlin Big Data Center [4]). Initial funding has been for four years and in 2018 the BMBF extended the funding for a second phase until Sep. 2021. The funded partners in phase 1 are the two Saxonian universities TU Dresden and University of Leipzig as well as two application partners, the Max Planck Institute for Molecular Cell Biology and Genetics (MI-CBG) and the Leibniz Institute for Ecological Urban and Regional Development (IÖR), Dresden. ScaDS Dresden/Leipzig also has about 20 associated partner organizations from science and economy (see *www.scads.de*).

The mission of ScaDS Dresden/Leipzig is to advance research on Big Data methods in key areas (Big Data platforms, data integration, visual analytics) and to apply these methods in diverse scientific and business applications to achieve novel solutions and services. For this purpose, ScaDS Dresden/Leipzig builds on the broad computer science expertise at the two universities and domain-specific knowledge of the application partners. To achieve a close and focused cooperation, the ScaDS-financed coworkers work together in labs at both universities. These labs also implement the ScaDS service center to coordinate computer science research and application development

This work was supported by the German Federal Ministry of Education and Research (BMBF, Grant No.: 01IS14014A-D) by funding the competence center for Big Data "ScaDS Dresden/Leipzig"

Erhard Rahm
rahm@informatik.uni-leipzig.de

Wolfgang E. Nagel
wolfgang.nagel@tu-dresden.de

✉ Eric Peukert
peukert@informatik.uni-leipzig.de

René Jäkel
rene.jaekel@tu-dresden.de

Fabian Gärtner
fabian@bioinf.uni-leipzig.de

Peter F. Stadler
studla@bioinf.uni-leipzig.de

Daniel Wiegreffe
daniel@informatik.uni-leipzig.de

Dirk Zeckzer
zeckzer@informatik.uni-leipzig.de

Wolfgang Lehner
wolfgang.lehner@tu-dresden.de

[1] Leipzig University, Leipzig, Germany

[2] Technische Universität Dresden, Dresden, Germany

Published online: 28 December 2018

and to serve new cooperation requests. Phase 1 of ScaDS Dresden/Leipzig was highly productive leading to more than 150 publications and more than a dozen prototypes and service implementations. Furthermore, a significant amount of additional funding could be acquired to broaden the research and application activities.

In this paper, we provide an overview about the structure and work areas of ScaDS Dresden/Leipzig (Sect. 2). In the following, we describe selected research results in Sect. 3, in particular on Big Data platforms, data integration, and visual analytics, and briefly discuss activities for the second phase of the center. Further articles in this issue provide in-depth views about the graph analytic platform GRADOOP [49], the analysis of time series data [17], scalable approaches for privacy-preserving data integration [11], and Big Data support for Digital Humanities [20].

2 ScaDS Overview

Fig. 1a shows the gross structure of ScaDS Dresden/Leipzig in phase 1 with five main research areas on Big Data methods, five application areas as well as the service center. For phase 2, the successful structure has largely been retained but with a stronger research focus in three areas and additional application fields (Fig. 1b). In the following, we will give an overview of the research areas, application areas, and the service center. We will also discuss activities in education and qualification.

2.1 Research areas

In its first phase, computer science research has focused on five core areas (shown in Fig. 1a), which we introduce here briefly.

2.1.1 Efficient Big Data architectures:

Big Data applications need a powerful and efficient infrastructure to meet the demanding processing requirements for data preparation and analysis. For example, biomedical microscopy typically leads to extremely high data production rates, while other applications require interactive user involvement that brings state-of-the-art platforms to their limit [24, 41]. Furthermore, data transfer is often a problem if the data is produced in experimental labs far away from the central computing infrastructure where data analysis has to be performed. We therefore investigated methods for efficient data reduction and data transfer as well as flexible storage technologies and intelligent I/O for HPC-based data analysis (see Sect. 3.1). The research also resulted in extending tools for performance analysis of parallel applications [5] into the Big Data domain investigating the performance of data-intensive workloads [12].

2.1.2 Data quality and data integration:

Good data quality and scalable methods for data integration are crucial to obtain meaningful analysis results. Big Data applications create new challenges due to the usually very high data volume, far greater data variety and the often large number of data sources to be integrated. Moreover, data privacy requirements introduce additional data integration challenges. To address these challenges, we investigate scalable methods for data integration, in particular methods for parallel execution of data integration and graph-analytics workflows [31, 50, 51], holistic methods for data integration from many sources [47], and privacy-preserving record linkage techniques [11]. Moreover, novel techniques for extracting and processing data from widely used spreadsheets have been developed [33, 34]. In Sects. 3.3 and 3.4, we give further details on our research on data extraction/augmentation and data integration, respectively.

2.1.3 Knowledge extraction:

The broad area of knowledge extraction deals with the scalable preprocessing and analysis of heterogeneous kinds of data to derive new knowledge and insights. We considered knowledge extraction for different kinds of data (texts, biological sequence data, graph data, measurements, images, and 3D microscopy datasets) and different analysis techniques including deep learning methods. The analysis of text data focuses on relationship discovery, classification through clustering or topic detection, to support its interpretation. For image data, methods for segmentation and classification are investigated. This research resulted in a novel technique for semantic segmentation of microscopy images (of developing zebrafishs) that is based on a mapping of the two commonly applied techniques of decision forests and deep convolutional networks [48]. Further scalable segmentation techniques have been developed for detecting settlements in historical geographic maps using binary segmentation [53] and for identifying land use of topographical maps [19]. The segmentation tasks are run on a parallel HPC infrastructure to achieve sufficiently fast training and execution times. A special focus has been the analysis of networked or graph data which resulted in the development of the GRADOOP platform for distributed graph analysis (Sect. 3.2) and generic and scalable techniques for graph pattern mining [45], grouping [30] and pattern matching [28]. We also investigated knowledge extraction techniques to analyze multiple genomes and their alignments by building a so-called supergenome [14] as described in Sect. 3.5.

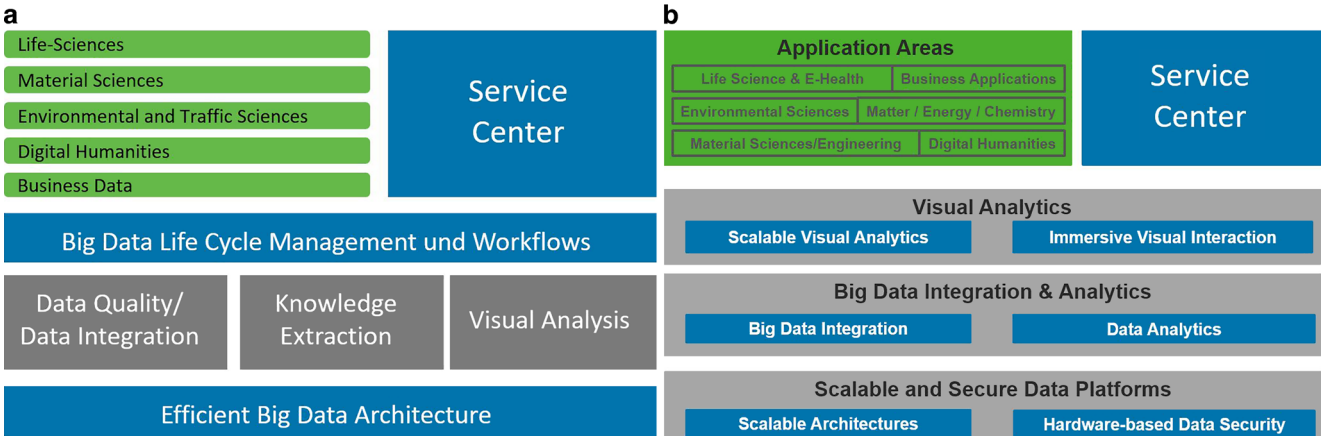

Fig. 1 Schematic overview about the research areas of ScaDS Dresden/Leipzig. **(a)** Phase 1, **(b)** Phase 2

2.1.4 Visual analytics:

Large amounts of data not only reach the limits of computation, but above all the limits of comprehensibility. In order to facilitate improved understanding, ScaDS Dresden/Leipzig investigates novel methods to interactively visualize Big Data with a particular focus on scientific visual analysis. This includes methods for intuitive navigation both in space and time as well as at different data abstraction levels. Furthermore, methods for semi-automatic adjustment of the multiplicity of parameters for analysis and visualization techniques have been investigated. In ScaDS1, a number of interactive visual analysis methods for complex simulation datasets were proposed that include focus and context techniques for particle-based simulations [57], for large amounts of point data [58] or for time-dependent data sets [59]. Further results on visual analytics in bioinformatics are described in Sect. 3.6.

2.1.5 Data life cycle management and workflows:

In the first phase of ScaDS Dresden/Leipzig, we have investigated flexible workflow and novel data lifecycle management (DLCM) methods tho record, store, extract, transform, and analyze data that is stored in data lakes. In particular, lightweight ETL (data extraction, transformation, loading) methods could be developed that support an optimization regarding different quality dimensions such as maintainability and performance [60]. ETL workflows are further extended to perform data integration and data augmentation with semi- or unstructured data at query time [8, 9]. Another research topic was the extension of the popular KNIME data analysis tool [3] to allow users to model graphical workflows in KNIME and to execute these transparently on a HPC infrastructure [16]. This approach strongly reduces the user effort to develop data-intensive workflows for HPC (high-performance computing) platforms. Moreover, ScaDS Dresden/Leipzig could contribute with extensions to the widely adopted UNICORE workflow model [2].

For phase 2, we continue research in most of these areas with a stronger focus on three research fields as shown in Fig. 1b. In the field of visual analysis, we focus more on interactive visualization with so-called immersive interaction techniques for large display walls and virtual and augmented reality. For data integration and analysis, we will address novel methods for information extraction, graph-based data integration, privacy-preserving data integration and the analysis of dynamic and temporal graph data. We furthermore extend the research on scalable and secure data platforms and investigate how to combine different Big Data frameworks with HPC infrastructures. The primary goal is to develop semi-automatic methods to provide powerful working environments for different applications.

2.2 Application areas

In phase 1, we addressed applications in five broad areas: life sciences, material sciences, environmental and traffic sciences, digital humanities, and business data. In the following, we briefly discuss the work in these areas:

2.2.1 Life science applications:

Biomedical research is a dynamic area characterized by the massive use of highly data-intensive technologies. At the ScaDS Dresden/Leipzig partner institutes, a main focus is in molecular data analysis with so-called omics technologies as well as the analysis of large-scale image data. For example, a close cooperation with MPI-CBG helped to segment 3D microscopy images of zebrafish development. The resulting publication [48] received the Best Science Paper Award of the British Machine Vision Conference (BMVC 2016). In another cooperation, ScaDS Dresden/Leipzig supported the development team of the Segemehl sequence

alignment tool [44] to identify parallelization opportunities for faster processing. In that context, ScaDS Dresden/Leipzig developed an FPGA-based accelerator for the Myers bit-vector algorithm which is used at the core of Segemehl [22]. The accelerator is able to significantly increase processing time of the algorithm with very low energy consumption. Further research contributions related to life sciences are outlined in Sects. 3.5 and 3.6.

2.2.2 Material and engineering science applications:

Research in material science generates a large number of structured and unstructured data that relates to materials, construction, simulation, production and the operation of the finished components. In ScaDS Dresden/Leipzig, we cooperate with the Institute of Lightweight Engineering and Polymer Technology of the TU Dresden to investigate methods to extract and analyze reliable material properties (e.g. life, strength, deformation) to achieve a better understanding of materials. The cooperation resulted in user-friendly scripts to run programs for finite element analysis and for simulations (Abaqus and LS-Dyna) on HPC infrastructure with SLURM. Moreover, new approaches for online and offline multi-scale visualization could be developed [21].

2.2.3 Environmental and traffic sciences applications:

Environmental sciences cope with a large set of data processing, modeling and simulation tasks that connect to environmental aspects. In that context we deal with all kinds of data such as climate, geographical, satellite or remote sensing data. We observe that vehicles are increasingly becoming data transmitters and receivers of environmental signals (e.g., GPS, WLAN, GSM, DAB). Based on these diverse data sources, methods for climate modeling, land use detection, impact modeling, and traffic control are investigated in ScaDS Dresden/Leipzig. For example, in a close collaboration with IÖR advanced methods for settlement and land use detection are developed [19, 53], that can be applied to a large corpus of historical maps from the Saxonian state and university library (SLUB).

2.2.4 Digital humanities:

In the last years, the humanities and social sciences gained access to a large amount of data (e.g., from large-scale digitization programs) for data-driven analysis. The challenge to tackle is the linking and interplay of quantitative, data-driven analysis with qualitative interpretations. In ScaDS Dresden/Leipzig, we developed methods for structuring and annotating text collections based on the so-called CTS (canonical text services) protocol [61, 62]. Further details are given within this issue in contribution [20].

2.2.5 Business applications:

There have been numerous cooperations with companies in the first phase of ScaDS Dresden/Leipzig, some of which resulted in additional projects that develop new Big Data methods for specific problems. The BMBF project InnoPlan, for example, focused on operational (decision-) processes for monitoring and control in health care. The cooperation resulted in novel techniques for phase detection in real-time medical device data [54] and the prediction of the remaining times during surgeries [56]. The approaches could also be transferred to the logistics domain [55]. In a further exemplary cooperation with a producer of renewable energy, ScaDS Dresden/Leipzig investigated novel techniques for anomaly detection in wind parks. The focus has been on methods to reduce the configuration effort to set suitable parameters and to analyze correlations in sensor data streams to provide early warnings in case of anomalies [6]. There are also projects to bring graph analytics into practical applications, in particular, by integrating GRADOOP into the KNIME data analysis tool [49].

In phase 2 of ScaDS Dresden/Leipzig, we will investigate further applications in the mentioned areas as well as in the new areas of matter, energy, chemistry and electronic health. The new applications will be addressed in cooperation with two project partners that are newly funded in ScaDS2, the Helmholtz Center for Environmental Research (UFZ) Leipzig and the Helmholtz Center Dresden Rossendorf (HZDR).

2.3 Service Center

ScaDS Dresden/Leipzig successfully established a so-called service center at both sites in Dresden and Leipzig to coordinate methodical and application researchers and to serve as a single point of contact for interested parties and cooperation partners (see Fig. 2). To foster the ScaDS-internal cooperation, the service center organizes several project meetings per year, alternately in Dresden and Leipzig, as well as dedicated workshops for Ph.D. students and also for specific applications. The service center also handled a large number of cooperation requests from science and industry that partly resulted in third-party funded projects to bring research results of ScaDS Dresden/Leipzig into practice or to address new research topics. Such additional projects relate to a wide spectrum of areas including data security, machine learning and graph analytics.

The service center is also responsible for teaching and qualification activities such as the organization of public workshops and international summer schools as well as for providing training for users who need to use Big Data technologies in their applications (see below). Furthermore, the service center helps with technical services and the pro-

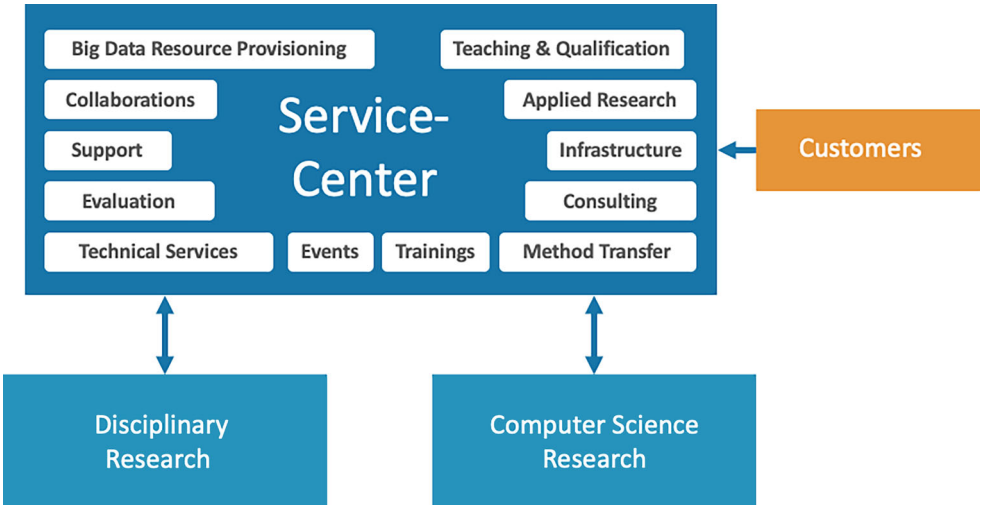

Fig. 2 Role of the ScaDS Dresden/Leipzig Service Center

vision of storage and processing infrastructure, typically in the form of Shared Nothing clusters (e.g., utilizing Hadoop) or HPC resources, for its cooperation partners. Finally, the service center helps to make developed research methods publicly available either as generic or domain-specific services. Such services include solutions for data deduplication, graph analytics, time series analytics, canonical text services, or peak calling in life sciences.

In the second funding phase, the service portfolio and the developed services will be retained structurally and extended in content. To deal with the increasing number of cooperation requests, we will follow a stream-lined model to quickly decide whether a problem can be handled with existing techniques and resources or whether it requires a new project with additional funds.

2.4 Education and qualification

There is still a significant lack of experts in Big Data and data science approaches so that a key goal of ScaDS Dresden/Leipzig is to provide extensive support for qualification and knowledge dissemination as well as to improve the study programs at the universities in Dresden and Leipzig. We thus developed a portfolio of advanced training modules in key technologies and methods that are offered to application scientists and industry partners. The goal is to lower the entry barrier when applying big data technologies. Application researchers should be enabled to evaluate and optimize existing algorithms and tools and to efficiently apply Big Data infrastructures and tools.

Great efforts have also been made in organizing scientific events that attracted many participants. Since 2015, we organize yearly international summer schools on Big Data and Machine Learning with talks and tutorials by many renowned experts. Furthermore, we established the highly successful BIDIB workshop series (Big Data in Business) in Leipzig focusing on business applications.

ScaDS Dresden/Leipzig members have further brought new Big Data-related lectures, practical exercises and seminars into the bachelor and master programs in computer science at both universities. At the University of Leipzig, a new study track on “Big Data” has been introduced that is the basis for an upcoming master program in data science.

3 Selected Research Areas

To illustrate some of the research activities of ScaDS Dresden/Leipzig, we outline results in six areas: scalable data platforms, distributed graph analytics, data augmentation, large-scale data integration, graph-based analysis of multiple genomes using a so-called supergenome, and the visual analysis of sequence-related data in bioinformatics.

3.1 Scalable data platforms

Data-driven scientific applications are often complex workflows with several compute and analysis tasks [1]. These workflows can be sequential but may also include cycles, e.g., to enrich the input data by additional steps. Such analytics workflows impose special requirements and challenges for the analytics platform [13]:

- provisioning of storage and compute resources for fast analysis
- access to flexible tool set or even complete analytics environments
- easy-to-use and accessible web-based interface for user interaction, e.g., to find optimal parameters to build a machine learning model.

Due to the high diversity of application requirements and research areas, we consider different kinds of data platforms, especially relatively inexpensive Shared Nothing clusters running Hadoop and parallel processing frameworks like Apache Flink and Spark, as well as HPC platforms. HPC systems are used in data-intensive scientific applications, but the developed solutions are typically rather static and community-bound making it difficult to support new requirements such as the incorporation of temporal varying (sensor) data. From the systems perspective, the challenges to support a broad range of usage scenarios are many-fold. The need to support exploratory analysis and the use of heterogeneous analysis tools make the use of HPC systems difficult for users [23]. Furthermore, users should be able to use different storage and processing technologies in their applications, e.g., to benefit from a high-performance I/O infrastructure or accelerator components such as GPUs. This is further complicated by the current tendency in the HPC community towards an even more complex storage hierarchy including fast storage layers (hot storage), e.g., by NVM-based[1] storage solutions or classical parallel file system access [38].

To support these requirements, we have investigated different directions to increase usability of our available data platforms. On our HPC systems, users can directly deploy analytics frameworks, such as Apache Hadoop and Spark or TensorFlow and Keras for machine learning. Furthermore, we currently extend our HPC infrastructure with additional hardware to improve IO capabilities by providing fast access to NVM-based storage and compute nodes specially suited for deep learning applications. To lower the entrance hurdles for users, we investigate virtualization and container technologies for well-defined resource provisioning and to support portability of analytics scenarios to alternative compute architectures.

For centralized systems, the transfer of data from heterogeneous, external sources (e.g., a sensor network) is still a challenging task. We have therefore developed a service-based and easy-to-use system to provide basic management functionalities [13]. The service consists of a REST (representational state transfer) web interface for connectivity to data acquisition services. The same REST interface can be used directly from analytics frameworks to exchange data or use analysis services. The general idea is that the system manages data streams from various sources and sends the data to a storage and to an analysis component. This way, the user of the system can control the creation or deletion of datasets, can derive data from it or enrich the data stream with new information, e.g., coming from other sources [13]. We have investigated an analysis infrastructure for monitoring data from our computing systems. Some of the results from [13] are summarized in Table 1.

Table 1 Scalability tests for runtime (R) of an analytics workflow on varying resources (from [13])

# nodes	# cores	R [s]	speed-up
2 (VM)	30	2110	1.00
2	30	930	2.27
4	80	430	4.91
8	160	322	6.55
12	240	242	8.72
16	320	238	8.87
20	400	199	10.60

The scalability test shows a comparison between a cloud-based system (first row in table) and an HPC system, porting the analytics workflows to the alternative system and increasing the hardware resources. This demonstrates the portability, which is interesting from the users perspective, and the reasonably good performance gain by speeding up the execution of the workflow.

3.2 Distributed Graph Analytics

Many scientific and business applications have to process and analyze highly inter-related data that can naturally be represented as graphs or networks. Examples include social networks, citation networks, transport networks or protein interaction networks in bioinformatics. These graphs can become very big with millions to billions of data entities (represented as graph vertices) and relationships (represented as graph edges) leading to high performance and scalability requirements for graph management and analysis. Such graphs can be heterogeneous with multiple kinds of entities and relationships. A large spectrum of analysis capabilities is typically needed ranging from focused queries to graph-wide mining algorithms, e.g., to determine clusters or frequent subgraphs. In the last decade, many approaches for graph data management and analysis have been developed, including graph database systems and distributed graph processing systems such as Google Pregel and similar approaches [39]. The known approaches, however, still suffer from significant restrictions such as limited scalability and mining support for graph database systems or insufficient support of heterogeneous graph data and querying in graph processing systems [29].

At ScaDS Dresden/Leipzig, we have therefore developed a new distributed graph processing platform called GRADOOP (*Gr*aph *A*nalytics on Had*oop*) [26, 27, 31] that aims at combining the strengths of graph database and distributed graph processing systems. In particular, it is based on the property graph data model and query capabilities of graph database systems, but also supports scalable, distributed processing and graph mining.

[1] NVM: Non Volatile Memory.

GRADOOP implements an extended version of property graphs supporting graph collections so that not only vertices and edges, but also individual graphs can be of different types and have different properties (attributes). The processing of graphs and graph collections is supported by a number of declarative graph operators, in particular, for filtering, subgraph selection, aggregation, and pattern matching that determines all occurrences of a graph pattern. There is also query support for a subset of Cypher, the query language of the graph database system Neo4J [28]. A powerful graph grouping operator is supported that allows a structural grouping and aggregation of graphs [30]. Additional structural transformation operations support the construction of vertices and edges from property values and the restructuring of graphs, e.g. to determine a co-authorship graph from a network of publications and their authors [37]. A call operator allows the execution of graph mining algorithms, e.g., for finding all frequent subgraphs within graph collections [46].

GRADOOP runs in distributed Hadoop-based Shared Nothing clusters and Fig. 3 shows the main architectural components. The graph data can be permanently stored either in HDFS (Hadoop file system), HBase or Accumulo. All operators and algorithms are implemented on the basis of Apache Flink and can therefore be executed in parallel and in-memory at the machines of the cluster, thereby supporting scalability to large amounts of data. A domain-specific language called GrALa is provided to make all GRADOOP operators and graph mining algorithms available within workflow programs for graph analysis. GRADOOP also offers a Java API with the GrALa operations to process and analyze extended property graphs. Since GRADOOP is implemented in Apache Flink, it represents a Flink extension that is available for the development of Flink applications.

The GRADOOP implementation is available as Open Source under Apache licence (www.gradoop.org). It has been integrated into the KNIME data analysis platform such that a visual definition of graph analysis workflows as well as a visual exploration of graphs and analysis results are possible [49]. In ScaDS 2, we will extend the functionality of GRADOOP to facilitate data preparation for graph analytics, in particular, the integration of heterogeneous sources into a combined property graph. Furthermore, we will explore support for dynamically changing graph data and temporal graph analysis.

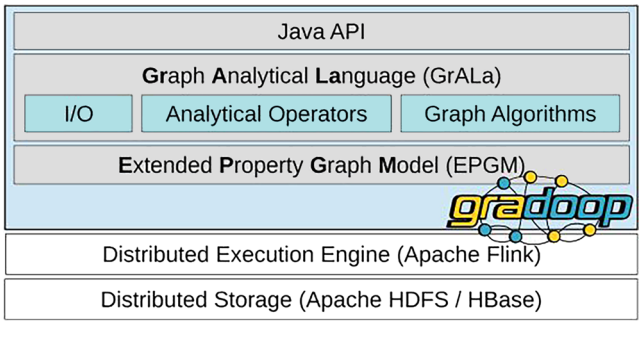

Fig. 3 High-level architecture of GRADOOP

3.3 Database Augmentation

In the era of Big Data, the number and variety of data sources is increasing every day. However, not all of this new data is available in well-structured databases or warehouses. Instead, heterogeneous collections of individual datasets such as data lakes are becoming more prevalent. This new wealth of data, though not integrated, has enormous potential for generating value in ad-hoc analysis processes, which are becoming more and more common with increasingly agile data management practices. However, in today's database management systems there is a lack of support for ad-hoc data integration of such heterogeneous data sources.

Within ScaDS Dresden/Leipzig, we therefore developed the entity augmentation system REA [9] that, given a set of entities and a large corpus of possible data sources, automatically retrieves the missing attributes. Due to the inherent uncertainty of the data sources and the matching process in general, REA produces not one, but k different augmentations which the user can choose from. To this end, we developed an extended version of the Set Cover problem, called Top-k Consistent Set Covering, onto which we map our requirements.

Moreover, we built DrillBeyond [8] by integrating REA with PostgreSQL that allows us to combine structured and unstructured query processing and enables seamless SQL queries over both RDBMS and the Web of Data. Therefore, we designed a novel plan operator that encapsulates the retrieval part and allows direct integration of such systems into relational query processing. The operator is placed in a cost-based manner to create query plans that are optimized for large invariant intermediate results which can be reused between multiple query evaluations.

In order to provide rich datasets for the augmentation process of well-structured databases, we devise multiple research activities. The most promising direction resulted in the publication of the "Dresden WebTable Corpus" (DWTC)[2] [9] based on the freely available web crawl "CommonCrawl". The DWTC corpus consists of 125 million tables extracted from the web and enriched by information surrounding a specific table like heading, extraction of the description, page title, etc. A second branch of research activities explored the challenges of using spreadsheets as data source. Spreadsheets in general are one of the most successful content generation tools, used in almost every

[2] https://wwwdb.inf.tu-dresden.de/misc/dwtc/.

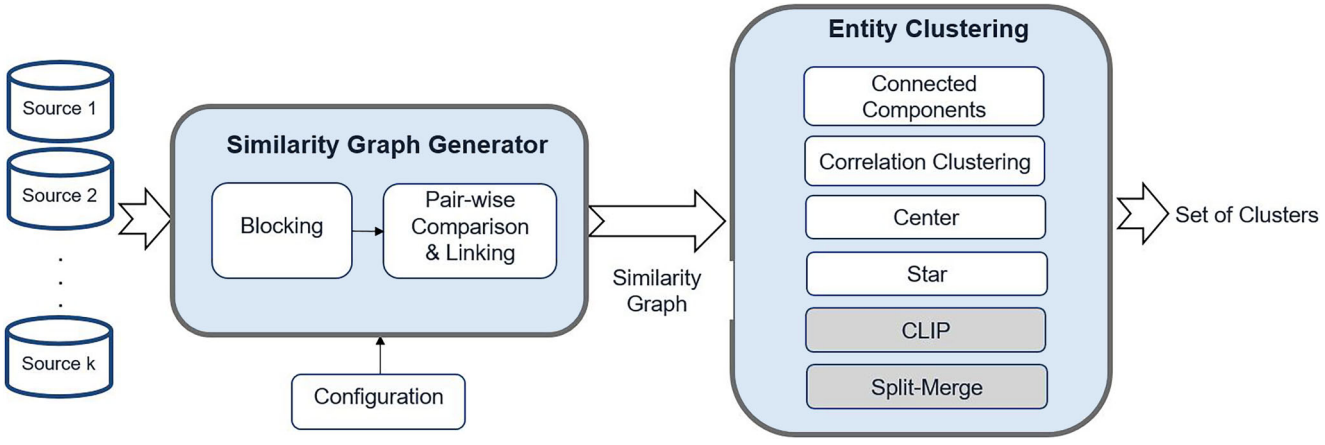

Fig. 4 Overview of FAMER

enterprise to perform data transformation, visualization, and analysis. The high degree of freedom provided by these tools results in very complex sheets, intermingling the actual data with formatting, formulas, layout artifacts, and textual metadata. To unlock the wealth of data contained in spreadsheets, a human analyst will often have to understand and transform the data manually. To overcome this cumbersome process, we proposed the DeExcelerator approach [7] that is able to automatically infer the structure and extract the data from these documents in a canonical form [33, 34].

3.4 Large-scale data integration

The effective analysis of Big Data depends on the provision of the relevant high-quality data from multiple sources. The data preparation steps for this purpose are even more complex than in traditional data analysis platforms (e.g., data warehouses), since Big Data involves a potentially much higher data volume as well as a higher diversity of data including streamed data and heterogeneous and only partly structured data, e.g., from websites or social networks. At ScaDS Dresden/Leipzig, we built on our previous research on data integration like the Map-Reduce-based entity resolution system DeDoop [35, 36] and focused on parallel matching and clustering of entities from many data sources. We further studied scalable approaches for privacy-preserving record linkage [10, 63] that are described in a separate paper of this issue [11].

To find and integrate matching entities (e.g., customers or products) from many data sources, we developed a new parallel entity resolution tool called FAMER (FAst Multi-source Entity Resolution system) [50, 51]. In contrast to previous approaches, it is not limited to only one or two data sources, but can holistically integrate data from more than two sources by clustering all matching entities. All steps in FAMER are implemented in Apache Flink, thereby supporting parallel execution on Shared Nothing clusters and scalability to large data volumes. Fig. 4 shows the main workflow of FAMER. In the first phase, a so-called similarity graph is built where vertices represent the entities from the input sources and edges represent similarity links for similar entities. These links are computed by a pairwise comparison using a configurable set of similarity functions. A configurable blocking step is employed to restrict comparisons to entities of the same block or data partition, e.g. costumers with the same birth-year or products of the same manufacturer.

The research focus has been on entity clustering that uses the similarity graph as input and groups together all matching entities from all sources within clusters. These clusters can then be used to fuse the different entity representations for further data analysis. FAMER supports several known techniques including connected components as well as two new approaches called Split-Merge [43] and CLIP [52]. The new approaches have been shown to clearly outperform the previous cluster algorithms and optimize the case where individual data sources are duplicate-free so that each cluster should contain at most one entity per data source. The CLIP approach is especially versatile since it can also be applied to repair clusters determined by other approaches [52]. For clustering, CLIP favors so-called strong links (e_A, e_B) in the similarity graph where entity e_A from source A is the most similar entity in A for entity e_B from source B and e_B is the most similar entity in B for e_A. By contrast, weak links, where none of the two linked entities is the most similar one in a source, are ignored thereby helping to reach high-quality clusters. In [42], we further investigate how to incrementally update a set of entity clusters when new entities or new data sources are added.

In ScaDS2, we plan to investigate novel learning-based methods using word embeddings and neural networks to improve entity categorization and matching with reduced configuration effort. Furthermore, we will extend FAMER

for graph data as needed to integrate several data sources for graph analytics, e.g., using GRADOOP. This requires the combined integration of entities as well as relationships of multiple types, which has not been sufficiently solved in previous work.

3.5 Graph-based analysis of multiple genome alignments

The rapid progress in high throughput technologies like sequencing technologies poses an increasing number of Big Data challenges in bioinformatics [15]. For example, comparative genomics applications typically make use of multiple genome-wide sequence alignments leading to high computational processing demands. The analysis demands increase further by the fast-growing availability of additional data for many species, such as transcript and protein sequences, binding sites of proteins, chemical modifications of the DNA, and chromatin features. All these data can modulate as intervals on the genome.

In ScaDS1, we investigated the use of a common coordinate system – known as a supergenome [18] – to permit the comprehensive comparison of such data between different genomes. The considered alignments consist of usually short blocks of similar sequences from different species. The partitioning accrues because genomes of different species are co-linear ("syntenic") only in local regions that are interrupted by breakpoints arising from the accumulation of genomic rearrangements during evolution. The construction of supergenome coordinate systems is, therefore, a challenging optimization problem where one has to find the "best possible" sorting of the local alignment blocks.

To find a solution, we transform the problem into a graph problem [14]. Every alignment block forms a vertex. Each chromosome (longest coherent parts of the genome) of each of the input genomes defines a path, and, for each adjacent block in a genome, we add a corresponding edge to the graph. The situation is further complicated by the fact that DNA is double-stranded and hence every sequence can legitimately be read in both directions. As a consequence, only betweenness but not the absolute order of blocks is well-defined for each genome. The task is therefore to find a vertex order of the alignment graph that respects the betweenness information for all chromosomes as far as possible.

Not surprisingly, this combinatorial optimization problem is NP-complete [14] making exact solutions infeasible for the problem sizes of practical interest – graphs with hundred thousands or millions of vertices. We, therefore, developed heuristic approximations that interactively reduce the set of vertices to simplify the vertex sorting problem. While some of these simplified approaches have asymptotically quadratic running time, they apply in practice only to subgraphs of limited size, which makes them feasible in practice. Nevertheless, this is still a time-consuming endeavor for datasets with more than four billion edges.

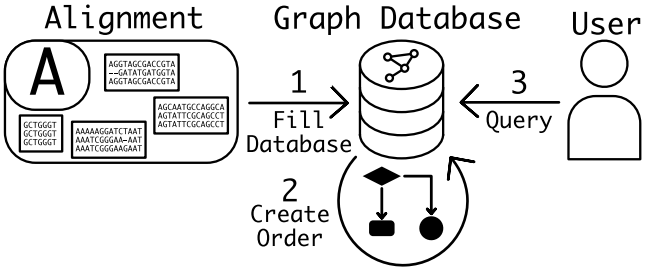

Fig. 5 First, we transform the alignments into a graph and save them in the graph database. Then the computation of the order updates the graph database. After this precomputing, users can query the graph database in real time

The goal is, therefore, to pre-compute and store the supergenome and then only query the stored data in real time. As shown in Fig. 5, we first store the alignment blocks in the database (currently Neo4J). We then determine the ordering and thus the supergenome that is also stored in the database as a basis for user queries. For graph interaction, we use the Tinkerpop framework that allows efficient programming with the possibility of changing the graph database. The graph database with the alignments and supergenome can be queried in many ways, e.g., for a comparison between a genome and the coordinate system.

In ScaDS2, we will investigate variations of the heuristics to create different coordinate systems. Furthermore, we will investigate extended analysis possibilities by using GRADOOP as the graph processing platform. Furthermore, we create a service to make the approach publicly available for the research community.

3.6 Big Data Visual Analytics

Visual support for analyzing data puts the human into the data analysis loop by providing visual encodings of the data and analysis results, e.g., obtained using statistical measures and data mining approaches, together with interaction facilities. Moreover, visual analytics extends these approaches by making the selection of analysis methods and their parameters explicit thereby allowing the user to change and adapt them to the problem and data under analysis [32]. Visual analytics for Big Data leverages this approach by taking the size of the data and its distribution among many compute cores into account.

In phase 1, we developed new Big Data visual analytics approaches for biology and bioinformatics applications, in particular, for studying diseases that are caused by genetic defects and heritable cell changes. For our purposes, the genome can be considered as a long sequence of molecules,

the so-called nucleotides. To fit it into a cell's nucleus, it is wrapped around a complex of eight proteins, the so-called histones. Together they are called nucleosome. The proteins can be modified, e.g., by methylation, where a methyl group is added to a specific location at the protein. One important field of study is which and how histones are modified, and how this is related to different cell types, expression of genes, and also to diseases. The analysis of these histone modifications uses a special algorithm called peak calling. However, the previous implementations could not process multiple replicates with high performance in terms of error rate and computing time.

We therefore developed the tool 'Sierra Platinum' for solving the problem of peak calling for replicated ChIP-Seq experiments [40, 64] as well as several applications. The tool provides new visualizations together with appropriate interaction to support the visual analysis of segmented peak-calling data of replicated ChIP-Seq experiments [65–67]. One of these visualizations provided is shown in Fig. 6.

The data set from which the peaks are extracted consists of two measurements: a background signal describing where are histones and an experiment signal describing where are *modified* histones. Both the background and the experiment have a signal with ridges and valleys. The comparison of the experiment signal with the background signal, can lead to three basic results, namely that the experiment signal is either lower, equal, or larger than the background signal. For repeated measurements, several pairs of background experiment measurements are available. However, state-of-the-art tools were not capable of handling those in a satisfactory manner.

The 'Sierra Platinum' multi-replicate peak caller is based on a new approach that allows to handle these replicated measurements. It combines the computation of the final result – the list of peaks – with additional statistics about the data based on mathematical models suitable for describing the distribution underlying the data. Moreover, this statistical information can be analyzed using adapted visualizations: the analyst can literally see whether or not one of the data sets has a problem. Further, the correlation among the replicates and the contribution of each data set to the final result (list of peaks) can be analyzed using these visualizations. This allows steering the computation: if a data set has low quality it can either be completely excluded from the determination of the peaks, or its influence on the result can be reduced.

Tests with artificial data sets (known ground truth) and real data showed that the quality of the resulting list of peaks is high, and thus that our method is reliable. Moreover, our expectation with respect to finding "bad" data sets visually were fulfilled. Due to our efforts of reducing time and space requirements of the computations, the amount of resources needed (computation time and space) is acceptable.

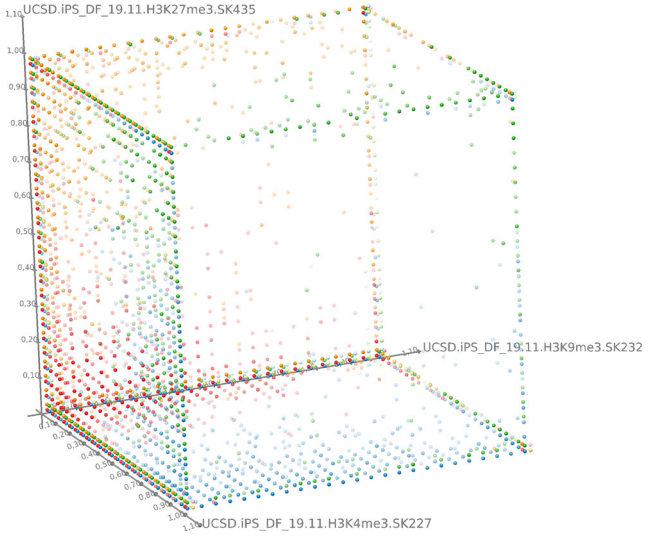

Fig. 6 A 3D tiled-binned scatter plot supporting comparing histone modifications of two different experiments [66]

4 Conclusions and Outlook

The Big Data Competence Center ScaDS Dresden/Leipzig follows a unique collaborative approach to fundamental and applied research on Big Data and data analysis methods. Since 2014, it combines the Big-Data-related computer science expertise at the two universities in Dresden and Leipzig with the domain-specific knowledge of its scientific and business application partners. A service center at both locations coordinates the activities and serves a central point of contact. The research results have been widely published and resulted in several generic and domain-specific tools and services, e.g., for graph analytics (GRADOOP), deduplication (FAMER) or HPC integration for KNIME workflows. The visibility of ScaDS Dresden/Leipzig has been further promoted by a series of well-attended workshops and international summer schools. Furthermore, Big Data and data science now play a major role in the computer science programs at the two universities so that there is a fast growing number of bachelor and master students with strong Big Data competences. In the second phase of ScaDS Dresden/Leipzig, the successful cooperation model will be continued to address further research and application challenges and to increase the outreach to industry. We will also extend the educational efforts, e.g., by introducing a new study program on data science.

Acknowledgements ScaDS Dresden/Leipzig is funded by the German Federal Ministry of Education and Research under grant BMBF 01IS14014B.

References

1. Asch M et al (2018) Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry. Int J High Perform Comput Appl 32(4):435–479
2. Benedyczak K, Schuller B, Petrova-El Sayed M, Rybicki J, Grunzke R (2016) Unicore 7 middleware services for distributed and federated computing. Proc High Perform Comput Simul (hpcs) Ieee Pp. https://doi.org/10.1109/HPCSim.2016.7568392
3. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME-the Konstanz information miner: version 2.0 and beyond. Acm Sigkdd Explor Newsl 11(1):26–31
4. Boden C, Rabl T, Markl V (2018) The Berlin Big Data Center (BBDC). it Inf Technol 60(5-6):321–326
5. Brunst H, Knüpfer A (2011) Vampir. Encyclopedia of Parallel Computing. Springer, pp 2125–2129. https://doi.org/10.1007/978-0-387-09766-4_60
6. Dienst S, Beseler J (2016) Automatic anomaly detection in offshore wind SCADA data. ProcWindEurope Summit, Hamburg
7. Eberius J, Werner C, Thiele M, Braunschweig K, Dannecker L, Lehner W (2013) DeExcelerator: a framework for extracting relational data from partially structured documents. In: CIKM, pp 2477–2480, https://doi.org/10.1145/2505515.2508210
8. Eberius J, Thiele M, Braunschweig K, Lehner W (2015a) Drill-Beyond: processing multi-result open world SQL queries. Proc 27th Int Conf on Scientific and Statistical Database. Management. https://doi.org/10.1145/2791347.2791370
9. Eberius J, Thiele M, Braunschweig K, Lehner W (2015b) Top-k entity augmentation using consistent set covering. Proc 27th Int Conf on Scientific and Statistical Database. Management. https://doi.org/10.1145/2791347.2791353
10. Franke M, Sehili Z, Rahm E (2018) Parallel Privacy Preserving Record Linkage using LSH-based blocking. Proc 3rd Int. Conf.on Internet of Things, Big Data and Security (IoTBDS), pp 195–203. https://doi.org/10.5220/0006682701950203
11. Franke M, Gladbach M, Sehili Z, Rohde F, Rahm E (2019) ScaDS research on scalable privacy-preserving record linkage. Datenbank Spektrum 19(1)
12. Frenzel J, Feldhoff K, Jäkel R, Müller-Pfefferkorn R (2018) Tracing of multi-threaded Java applications inScore-P using bytecode instrumentation, Proc. ARCS Workshop, pp 1–8
13. Frenzel J, Sastri Y, Lehmann C, Lazariv T, Jäkel R, Nagel W (2018) A generalized service infrastructure for data analytics. In: Proc. IEEE 4th Int. Conf. on Big Data Computing Service and Applications (BigDataService), pp 25–32, https://doi.org/10.1109/BigDataService.2018.00013
14. Gärtner F, zu Siederdissen C, Müller L, Stadler PF (2018) Coordinate systems for supergenomes. Algorithms for Molecular Biology 13(1):15
15. Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. Nat Rev Genet 17(3):175–188
16. Grunzke R, Jug F, Schuller B, Jäkel R, Myers G, Nagel WE (2016) Seamless HPC integration of data-intensive KNIME workflows via UNICORE. In: European Conf. on Parallel Processing, Springer, pp 480–491. https://doi.org/10.1007/978-3-319-58943-5_39
17. Hahmann M, Hartmann C, Kegel L, Lehner W (2019) Large-scale time series analytics – novel approaches for generation and prediction. Datenbank Spektrum 19(1)
18. Herbig A, Jäger G, Battke F, Nieselt K (2012) GenomeRing: alignment visualization based on SuperGenome coordinates. Bioinformatics 28(i7):i15
19. Herold H, Hecht R, Meinel G (2016) Old maps for land use change monitoring – analysing historical maps for long-term land use change monitoring. Proc Int Workshop Exploring Old Maps. EOM 201(6):11–12
20. Heyer G, Tiepmar J (2019) A Big Data case study in Digital Humanities: Creating a performance benchmark for Canonical Text Services. Datenbank Spektrum 19(1)
21. Hoehne R, Staib J (2016) Multi-scale visualisation – key to an enhanced understanding of materials. Carbon Compos Mag 4:20–21 (ISSN 2366-8024)
22. Hoffmann J, Zeckzer D, Bogdan M (2016) Using FPGAs to accelerate Myers bit-vector algoriththm. In: XIV Mediterranian Conf. Med Biol Eng Comput, pp 529–535. https://doi.org/10.1007/978-3-319-32703-7_104
23. Jäkel R, Müller-Pfefferkorn R, Kluge M, Grunzke R, Nagel WE (2014) Architectural implications for Exascale based on Big Data workflow requirements. In: High Performance Computing Workshop, IOS Press, Advances in Parallel Computing, vol 26, pp 101–113
24. Jäkel R, Müller-Pfefferkorn R, Kluge M, Grunzke R, Nagel WE (2015) Architectural implications for Exascale-based on Big Data workflow requirements. Advances in Parallel Computing vol 26, pp 101–113
25. Jäkel R, Peukert E, Nagel WE, Rahm E (2018) ScaDS Dresden/Leipzig – a competence center for collaborative Big Data research. it Inf Technol 60(5-6):327–334
26. Junghanns M, Petermann A, Gómez K, Rahm E (2015) GRADOOP: scalable graph data management and analytics with Hadoop. Arxiv Prepr Arxiv 150600548
27. Junghanns M, Petermann A, Teichmann N, Gómez K, Rahm E (2016) Analyzing extended property graphs with Apache Flink. In: Proc. ACM, SIGMOD Workshop on Network Data Analytics https://doi.org/10.1145/2980523.2980527
28. Junghanns M, Kießling M, Averbuch A, Petermann A, Rahm E (2017a) Cypher-based graph pattern matching in GRADOOP. In: Proc. 5th Int. Workshop on Graph Data Management Experiences & Systems (GRADES), https://doi.org/10.1145/3078447.3078450
29. Junghanns M, Petermann A, Neumann M, Rahm E (2017b) Management and analysis of big graph data: current systems and open challenges. In: Handbook of Big Data Technologies. Springer, Cham, pp 457–505 https://doi.org/10.1007/978-3-319-49340-4_14
30. Junghanns M, Petermann A, Rahm E (2017c) Distributed grouping of property graphs with GRADOOP. Proc Database systems for Business, Technology and Web (BTW)
31. Junghanns M, Kießling M, Teichmann N, Gómez K, Petermann A, Rahm E (2018) Declarative and distributed graph analytics with GRADOOP. Proc VLDB Endowment. PVLDB 11(12):2006–2009. https://doi.org/10.14778/3229863.3236246
32. Keim D, Andrienko G, Fekete JD, Görg C, Kohlhammer J, Melançon G (2008) Visual analytics: Definition, process, and challenges. In: Information visualization. Springer, Berlin, Heidelberg, pp 154–175. https://doi.org/10.1007/978-3-540-70956-5_7
33. Koci E, Thiele M, Romero O, Lehner W (2016) A machine learning approach for layout inference in spreadsheets. In: Proc. KDIR '16. https://doi.org/10.5220/0006052200770088
34. Koci E, Thiele M, Romero O, Lehner W (2017) Table identification and reconstruction in spreadsheets. In: Proc. 29th Int. Conf. on Advanced Information Systems Engineering (CAiSE), https://doi.org/10.1007/978331959536833
35. Kolb L, Rahm E (2013) Parallel entity resolution with DeDoop. Datenbank Spektrum 13(1):23–32
36. Kolb L, Thor A, Rahm E (2012) DeDoop: efficient deduplication with Hadoop. PVLDB 5(12). https://doi.org/10.14778/2367502.2367527
37. Kricke M, Peukert E, Rahm E (2019) Graph data transformations in GRADOOP. Proc. BTW, conf

38. Lüttgau J, Kuhn M, Duwe K, Alforov Y, Betke E, Kunkel J, Ludwig T (2018) A Survey of Storage Systems for High-Performance Computing. Supercomputing Frontiers and. Innovations:31–58. https://doi.org/10.14529/jsfi180103
39. McCune RR, Weninger T, Madey G (2015) Thinking like a vertex: a survey of vertex-centric frameworks for large-scale distributed graph processing. ACM Comput Surv 48(2):25
40. Müller L, Gerighausen D, Farman M, Zeckzer D (2016) Sierra Platinum: A Fast and Robust Multiple-Replicate Peak Caller With Visual Quality-Control and -Steering. BMC Bioinformatics 17(1):1–13
41. Nagel WE, Jäkel R, Müller-Pfefferkorn R (2015) Execution environments for Big Data: Challenges for user centric scenarios. In: BDEC white paper BDEC. Proc. Int. Workshop on Extreme Scale Scientific Computing (Big Data and Extreme Computing, BDEC), Barcelona, 2015
42. Nentwig M, Rahm E (2018) Incremental clustering on linked data. In: Proc. IEEE, Int. Conf. on Data Mining Workshops (ICDMW)
43. Nentwig M, Groß A, Rahm E (2016) Holistic entity clustering for linked data. In: Proc. Data Mining Workshops (ICDMW), IEEE, pp 194–201, https://doi.org/10.1109/ICDMW.2016.0035
44. Otto C, Stadler PF, Hoffmann S (2014) Lacking alignments? The next-generation sequencing mapper Segemehl revisited. Bioinformatics 30(13), pp 1837–1843. https://doi.org/10.1093/bioinformatics/btu146
45. Petermann A, Junghanns M, Kemper S, Gómez K, Teichmann N, Rahm E (2016) Graph mining for complex data analytics. In: Data Mining Workshops (ICDMW), IEEE, pp 1316–1319, https://doi.org/10.1109/ICDMW.2016.0193
46. Petermann A, Junghanns M, Rahm E (2017) DIMSpan: Transactional frequent subgraph mining with distributed in-memory dataflow systems. In: Proc. 4th IEEE/ACM Int. Conf. on Big Data Computing, Applications and Technologies (BDAT), pp 237–246, https://doi.org/10.1145/3148055.3148064
47. Rahm E (2016) The case for holistic data integration. Proc ADBIS, LNCS 9809:11–27. https://doi.org/10.1007/978-3-319-44039-2_2
48. Richmond D, Kainmüller D, Yang M, Myers E, Rother C (2016) Mapping auto-context decision forests to deep convnets for semantic segmentation. Proc British Machine Vision Conference. BMVC. https://doi.org/10.5244/C.30.144
49. Rostami A, Kricke M, Peukert E, Kühne S, Dienst S, Rahm E (2019) BIGGR: Bringing GRADOOP to applications. Datenbank Spektrum 19(1)
50. Saeedi A, Peukert E, Rahm E (2017) Comparative evaluation of distributed clustering schemes for multi-source entity resolution. In: Advances in Databases and Information Systems. Springer, Cham, pp 278–293 https://doi.org/10.1007/978-3-319-66917-5_19
51. Saeedi A, Nentwig M, Peukert E, Rahm E (2018a) Scalable matching and clustering of entities with FAMER. Complex Syst Informatics Model Q (CSIMQ) 16:61–83. https://doi.org/10.7250/csimq.2018-16.04
52. Saeedi A, Peukert E, Rahm E (2018b) Using Link Features for Entity Clustering in Knowledge Graphs. In: Proc. ESWC, LNCS 10843, Springer, pp 576–592, https://doi.org/10.1007/978-3-319-93417-4_37
53. Schemala D, Schlesinger D, Winkler P, Herold H, Meinel G (2016) Semantic segmentation of settlement patterns in gray-scale map images using RF and. CRF, within an HPC environment. Proc GEOBIA
54. Spangenberg N, Augenstein C, Franczyk B, Wagner M, Apitz M, Kenngott H (2017a) Method for intrasurgical phase detection by using real-time medical device data. Proc Int Conf Comput Med Syst. https://doi.org/10.1109/CBMS.2017.65
55. Spangenberg N, Roth M, Mutke S, Franczyk B (2017b) Big Data in der Logistik – ein ganzheitlicher Ansatz für die datengetriebene Logistikplanung, -überwachung und -steuerung. In: Industrie 4.0 Management 33(4):43–47
56. Spangenberg N, Wilke M, Franczyk B (2017c) A big data architecture for intra-surgical remaining time predictions. Proc Int Conf Curr Future Trends Inf Commun Technol Healthc (icth). https://doi.org/10.1016/j.procs.2017.08.332
57. Staib J, Grottel S, Gumhold S (2015) Visualization of particle-based data with transparency and ambient occlusion. Comput Graph Forum 34:151–160
58. Staib J, Grottel S, Gumhold S (2016) Enhancing Scatterplots With Multi-dimensional Focal Blur. Comput Graph Forum 35:11–20. https://doi.org/10.1111/cgf.12877
59. Staib J, Grottel S, Gumhold S (2017) Temporal focus+context for clusters in particle data. In: Vision, Modeling and Visualization (VMV17)
60. Theodorou V, Abelló A, Thiele M, Lehner W (2015) Poiesis: a tool for quality-aware ETL process redesign. Proc 18th Int Conf on Extending Database Technology. EDBT. https://doi.org/10.5441/002/edbt.2015.54
61. Tiepmar J (2014) Release of the MySQL-based implementation of the CTS protocol. In: Proc. 3rd LREC Workshop on Challenges in the Management of Large Corpora, pp 35–43
62. Tiepmar J (2016) CTS text miner – text mining framework based on the canonical text service protocol. In: Proc. 4th LREC Workshop on Challenges in the Management of Large Corpora, pp 1–7
63. Vatsalan D, Sehili Z, Christen P, Rahm E (2017) Privacy-preserving record linkage for Big Data: Current approaches and research challenges. Handb Big Data Technol, pp 851–895. https://doi.org/10.1007/978-3-319-49340-4_25
64. Wiegreffe D, Müller L, Steuck J, Zeckzer D, Stadler PF (2018) The Sierra Platinum Service for generating peak-calls for replicated ChIP-seq experiments. BMC Res Notes. https://doi.org/10.1186/s13104-018-3633-x
65. Zeckzer D, Gerighausen D, Steiner L, Prohaska SJ (2014) Analyzing Chromatin Using Tiled Binned Scatterplot Matrices. IEEE, Symp on Biological Data Visualization (BioVis)
66. Zeckzer D, Gerighausen D, Müller L (2016) Analyzing Histone Modifications in iPS Cells Using Tiled Binned 3D Scatter Plots. In: Proc. Big Data Visual Analytics (BDVA), pp 1–8, https://doi.org/10.1109/BDVA.2016.7787042
67. Zeckzer D, Wiegreffe D, Müller L (2018) Analyzing Histone Modifications Using Tiled Binned Clustering and 3D Scatter Plots. J Wscg 26:1–10