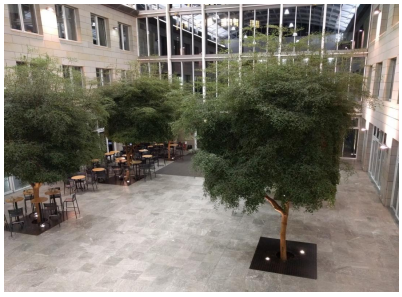
PROGRAM INFORMATION

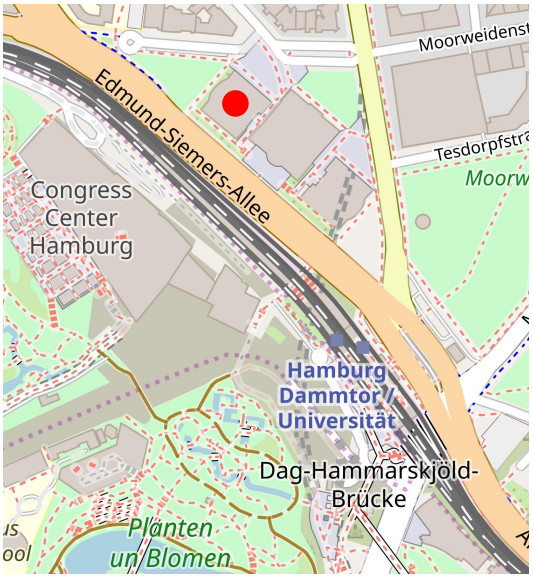
NoSQL Workshop 2020

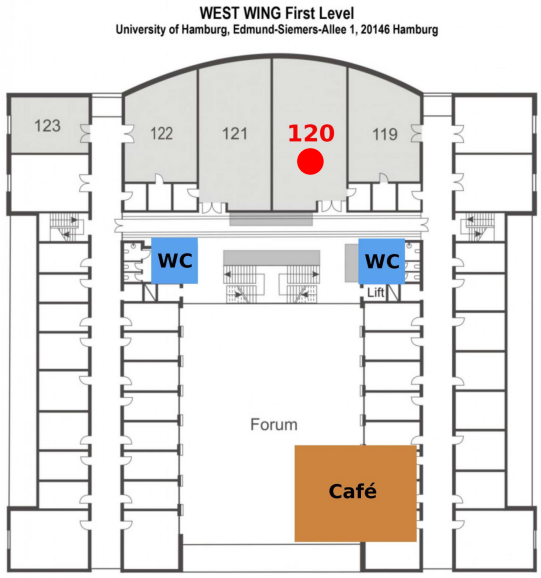
17. - 18. February 2020
Hamburg

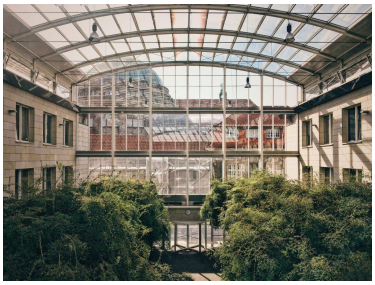
1 Venue

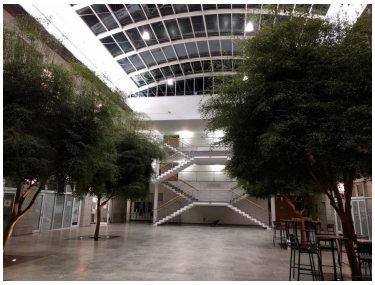
The workshop will take place in room 120 of the west wing of the main building of the University of Hamburg which is located at Edmund-Siemers-Allee 1 (directly opposite to Dammtor Station).







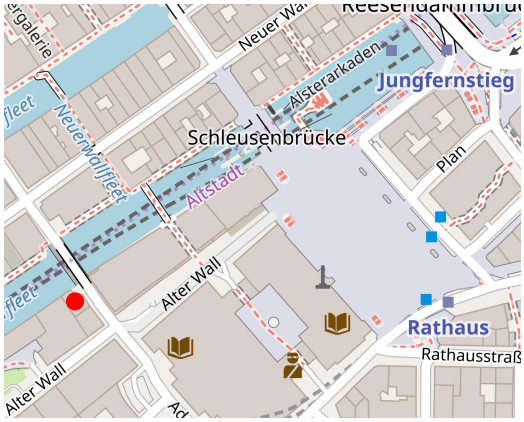


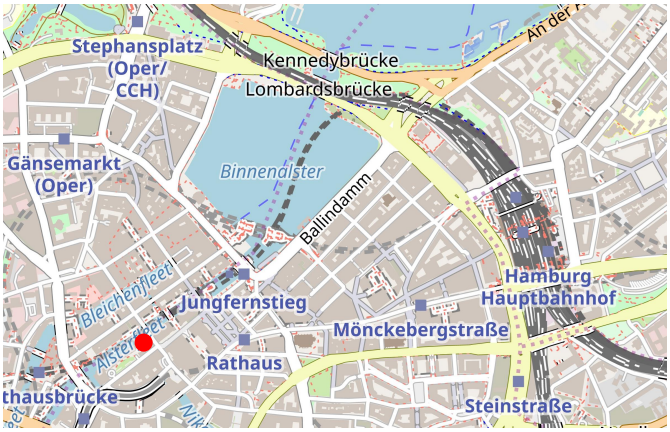


2 Food & Beverages

We will provide water, cookies and some fruits (e.g. bananas, apples) in the workshop room. Coffee and different kinds of food (including warm meals) can be bought in a café located in the forum of the building. Moreover, the mensa is about 500 meters away (we will use it for lunch).

The dinner on Monday evening will take place at the Brauhaus Joh. Albrecht which is located at Adolphsbrücke 7, 20457 Hamburg.





3 Participants

Name	Affiliation	Contact
Thomas Clemen	HAW Hamburg	*thomas.clemen@haw-hamburg.de*
André Düjon	Univ. of Hamburg	*1duejon@informatik.uni-hamburg.de*
Gajendra Doniparthi	Univ. of Kaiserslautern	*donipart@rhrk.uni-kl.de*
Jesús García Molina	Univ. of Murcia	*jesus.gmolina@gmail.com*
Felix Gessert	Baqend	*felix.gessert@baqend.com*
Daniel Glake	Univ. of Hamburg	*glake@informatik.uni-hamburg.de*
Alberto Hernández Chillón	Univ. of Murcia	*alberto.hernandez1@um.es*
Andrea Hillenbrand	Darmstadt Univ.	*andrea.hillenbrand@h-da.de*
Michael Hohenstein	Univ. of Kaiserslautern	*hohenstein@informatik.uni-kl.de*
Felix Kiehn	Univ. of Hamburg	*kiehn@informatik.uni-hamburg.de*
Meike Klettke	Univ. of Rostock	*meike.klettke@uni-rostock.de*
Christoph Langhein	Univ. of Hamburg	*3langhei@informatik.uni-hamburg.de*
Mark Lukas Möller	Univ. of Rostock	*mark.moeller2@uni-rostock.de*
Fabian Panse	Univ. of Hamburg	*panse@informatik.uni-hamburg.de*
Martin Poppinga	Univ. of Hamburg	*poppinga@informatik.uni-hamburg.de*
Dennis Przytarski	Univ. of Stuttgart	*dennis.przytarski@ipvs.uni-stuttgart.de*
Norbert Ritter	Univ. of Hamburg	*ritter@informatik.uni-hamburg.de*
Christopher Rost	Univ. of Leipzig	*rost@informatik.uni-leipzig.de*
Johannes Schildgen	OTH Regensburg	*schildgen@cs.uni-kl.de*
Mareike Schmidt	Univ. of Hamburg	*mschmidt@informatik.uni-hamburg.de*
Heiko Schuldt	Univ. of Basel	*heiko.schuldt@unibas.ch*
Holger Schwarz	Univ. of Stuttgart	*holger.schwarz@ipvs.uni-stuttgart.de*
Diego Sevilla Ruiz	Univ. of Murcia	*dsevilla@ditec.um.es*
Alexander Stiemer	Univ. of Basel	*alexander.stiemer@unibas.ch*
Uta Störl	Darmstadt Univ.	*uta.stoerl@h-da.de*
Marco Vogt	Univ. of Basel	*marco.vogt@unibas.ch*
Rasmus Warrelmann	Univ. of Hamburg	*3warrelm@informatik.uni-hamburg.de*
Lena Wiese	Fraunhofer Institute ITEM	*lena.wiese@item.fraunhofer.de*
Wolfram Wingerath	Baqend	*wolfram.wingerath@baqend.com*
Benjamin Wollmer	Baqend/Univ. of Hamburg	*wollmer@informatik.uni-hamburg.de*

4 Schedule

Monday, 17.02.2020 (Room ESA W120)

Start	End	Title	Speaker
11:30	–	Begin of Registration	DBIS Group *University of Hamburg*
12:00	13:15	Joint Lunch in Mensa	
13:15	13:30	Welcome & Introduction	Everybody
13:30	14:15	*Mobile Site Speed and the Impact on E-Commerce*	Felix Gessert, Wolfram Wingerath *Baqend*
14:15	14:50	*Webcaching: Moving Away from a Binary Decision*	Benjamin Wollmer *Baqend / University of Hamburg*
14:50	15:25	*Using Triples as the Data Model for Blockchain Systems*	Dennis Przytarski *University of Stuttgart*
15:25	15:45	Coffee Break	
15:45	16:20	*Integration of Medical Data with Neo4J*	Lena Wiese *Fraunhofer Institute ITEM*
16:20	16:55	*Distributed Analyzing of Temporal Graphs*	Christopher Rost *University of Leipzig*
16:55	17:05	Coffee Break	
17:05	17:40	*Leveraging Bloom-Filters for Interactive Exploration of Data from Hierarchical Data Models*	Gajendra Doniparthi *University of Kaiserslautern*
17:40	18:15	*Querying Geometric Patterns in Protein Structures: Spatial Search Beyond Geoinformatics*	Martin Poppinga *University of Hamburg*
19:00	23:00	Joint Dinner in Brauhaus Joh. Albrecht	

Thuesday, 18.02.2020 (Room ESA W120)

Start	End	Title	Speaker
09:00	09:45	*Towards Polyglot Persistence: Overview and Vision*	Daniel Glake, Felix Kiehn, Mareike Schmidt *University of Hamburg*
09:45	10:20	*Polypheny: Wenn me dr Batze und s' Weggli ka ha*	Marco Vogt *University of Basel*
10:20	10:55	*Cost-based Combination of Fragmentation, Replication, Allocation & Migration in a Distributed Database*	Alexander Stiemer *University of Basel*
10:55	11:10	Coffee Break	
11:10	11:45	*A Unified Schema for Polyglot Persistence*	Jesús García Molina, Diego Sevilla Ruiz *University of Murcia*
11:45	12:20	*A DSL Family to Define and Manage Schemas and Data on NoSQL Databases*	Alberto Hernández Chillón *University of Murcia*
12:20	13:20	Lunch Break	
13:20	13:55	*Towards Self-Adapting Data Migration in the Context of Schema Evolution in NoSQL Databases*	Andrea Hillenbrand *Darmstadt University*
13:55	14:30	*Considerations towards a Multi-Model NoSQL Schema Evolution Benchmark*	Mark Lukas Möller *University of Rostock*
14:30	15:05	*Generating Large and Heterogenous Test Data for Duplicate Detection*	Fabian Panse *University of Hamburg*
15:05	15:20	Coffee Break	
15:20	15:55	*Leveraging Approximate Query Processing to Realize Progressive Visual Analytics*	Michael Hohenstein *University of Kaiserslautern*
15:55	16:30	*Chronos - The Swiss Army Knife for Systems Evaluations*	Marco Vogt *University of Basel*
16:30	–	Farewell & Departure	

5 Presentations

Monday, 17.02.2020

13:30 - 14:15

Mobile Site Speed and the Impact on E-Commerce

Felix Gessert and Wolfram Wingerath, Baqend

Abstract: If you don't invest into page speed now, you will pay for it later – because poor web performance drags down your search rank. But how do you measure web performance? What can you do to improve it? And why are we talking about this at a data management workshop? This talk will answer these questions – and more! First, we present Speed Kit as an opt-in approach to integrate state-of-the-art database caching with content delivery mechanisms in the web. To this end, we recap the database research that went into Baqend's DBaaS platform and also describe how we use our technology for website acceleration. We then go into detail on our real-user monitoring pipeline for collecting production data on accelerated websites and analyzing the technical and business performance uplift we achieve. We also cover the differences between technical and user-centric performance metrics and compare our own approach against state-of-the-art competitors. Arguing for a correlation between web performance and business success, we conclude with key insights gained from a large web performance study conducted in 2019 as a joint effort by Baqend, Google, and the University of Hamburg.

14:15 - 14:50

Webcaching: Moving Away from a Binary Decision

Benjamin Wollmer, Baqend / University of Hamburg

Abstract: Caching is an indispensable means to accelerate content delivery on the web. While standard HTTP caching has been designed for static resources such as files, advanced approaches such as Orestes have also made it applicable to frequently changing data like database query results or server-generated HTML files. However, whether or not cached resources can be used for acceleration has always been a binary decision: a cached response is either valid and can be used or has been invalidated and must be avoided. In this talk, we present a research plan for an early-stage Ph.D. project and suggest a novel scheme for delta encoding that allows partial usage of cached resources to minimize the amount of data to be retrieved from the origin. We discuss related work on

the topic and analyze why delta encoding has not been established as a standard in the industry so far, despite significant gains such as reduced bandwidth usage and loading times for end-users. We finally sketch an end-to-end architecture that makes delta encoding feasible to implement in practice, and we close with the most critical challenges and potential use cases.

14:50 - 15:25

Using Triples as the Data Model for Blockchain Systems

Dennis Przytarski, University of Stuttgart

Abstract: Current permissioned blockchain systems utilize the key-value data model to store and query the ledger. As the key-value pairs are not sufficiently expressive to represent relationships between data, we present a proposal for the utilization of triples as the data model for blockchain systems. This approach enables a powerful query engine and reduces the number of data stores that have to be maintained.

15:45 - 16:20

Integration of Medical Data with Neo4J

Lena Wiese, Fraunhofer Institute ITEM

Abstract: Genome analysis is a major precondition for future advances in the life sciences. The complex organization of genome data and the interactions between genomic components can often be modeled and visualized in graph structures. In this paper we propose the integration of several data sets into a graph database. We study the aptness of the database system in terms of analysis and visualization of a genome regulatory network (GRN) by running a benchmark on it. Major advantages of using a database system are the modifiability of the data set, the immediate visualization of query results as well as built-in indexing and caching features.

16:20 - 16:55

Distributed Analyzing of Temporal Graphs

Christopher Rost, University of Leipzig

Abstract: Real-world graphs naturally change over time. The analysis of such temporal graphs is an important requirement in many domains, e.g., social and communication networks, financial transactions or biological networks. Since current graph databases and graph processing systems focus on a static perspective of a graph, the maintenance and analysis of the evolution of real-world graphs is hardly supported. We, therefore, extended the distributed graph analytics framework Gradoop for time-related graph analysis by introducing a new temporal property graph data model. Our model supports bitemporal time dimensions for vertices and edges to represent both rollback and historical information. In addition to the data model, we introduce several operators (e.g., Snapshot, Diff and Grouping) that natively support the time dimensions of the graph. Since this is an extension of Gradoop, the temporal operators are compatible and can be combined with the already known operators to build complex analytical tasks in a declarative way.

In this talk, I will give a brief overview of the Gradoop system, the temporal property graph model and how it supports the time-dependent analysis of large historical graphs. I will show the expressiveness and flexibility of the temporal operators based on a real-world use case.

17:05 - 17:40

Leveraging Bloom-Filters for Interactive Exploration of Data from Hierarchical Data Models

Gajendra Doniparthi, University of Kaiserslautern

Abstract: Through this work, we address some of the big-data challenges in bio-informatics, particularly, with high through-put technologies such as Mass Spectrometry. We use the abstract specifications and file formats from existing standard frameworks such as ISA, Research Objects etc., to model the curated meta-data of a bio-science experiment and the raw-data generated from the instruments. We propose a segmented approach and leverage bloom filters to index the massive number of instrument parameters and contextual information. We also use natural language style querying to decouple the database operations performed on the curated meta-data from the predicate expressions evaluated on the massive sets of raw-data.

17:40 - 18:15

Querying Geometric Patterns in Protein Structures: Spatial Search Beyond Geoinformatics

Martin Poppinga, University of Hamburg

Abstract: The use and production of spatial data is not limited to the field of Geoinformatics. Another important field is Bioinformatics which also uses large amounts of 3D spatial data, but on much smaller spatial scales. This data must be queriable for research purposes, as it is relevant for, e.g., drug discovery and repositioning. One way to search in a dataset containing thousands of protein structures is to use spatial geometric queries to find structures, described by properties, distances, and angles between various atoms as well as the interactions between atoms. The current scenario is connected to the PELIKAN and GeoMine software developed at the Center for Bioinformatics at the University of Hamburg, where the goal is to mine for spatial patterns in large collections of protein structure as the Protein Data Bank (PDB). I will introduce the field in which I will be working in my Ph.D. thesis and give an outlook on various use cases and problems that need to be addressed there.

Thuesday, 18.02.2020

9:00 - 9:45

Towards Polyglot Persistence: Overview and Vision

Daniel Glake, Felix Kiehn and Mareike Schmidt, University of Hamburg

Abstract: Nowadays, data-intensive applications face one main problem during their development: Handling data with multiple, differing requirements regarding consistency, availability or data store functionalities. Thus, these applications require a set of different data stores (RDBMS, NewSQL, NoSQL, etc.), each optimized for a specific kind of data and task. However, the wide spectrum of data store interfaces makes it difficult to access and integrate data from multiple sources and incorporate their provided storage features (e.g., provided consistency and scaling). This severe problem has motivated the design of a new generation of systems, called multi- or polystores. Providing an integrated, transparent access to serve different data stores by using one or more query languages, these systems can benefit from each of the stores' respective functionalities and characteristics. We describe representative real-world use cases for data-intensive applications (OLTP and OLAP) and derive a set of requirements for polyglot data stores. Subsequently, we discuss the properties of selected Multi- and Polystores and evaluate them based on our requirements, grouped into functional features, query processing technique, architecture and adaptivity. Our analysis reveals a severe lack of several capabilities, especially adaptivity. We outline the benefits and drawbacks of the surveyed systems and propose future research directions, including our vision of a polyglot persistence management system.

9:45 - 10:20

Polypheny: Wenn me dr Batze und s' Weggli ka ha

Marco Vogt, University of Basel

Abstract: In the last few years, the "one-size-fits-all" paradigm, according to which database systems have been designed for several decades, has come to an end. The reason for this lies in the very broad spectrum and heterogeneity of applications a database system needs to support, ranging from business over science to the private life of individuals. Therefore, demands are becoming more and more heterogeneous and range, for example, from frequently updated to immutable data, from highly structured data to unstructured data, or from applications which demand always consistent data to applications that are fine with lower levels of consistency.

Hence, many applications feature heterogeneous data and/or workloads as they in-

trinsically come with data (sub-collections) having different requirements for storage, access, consistency, etc. which have to be dealt with in the same system. This development can either be coped with a large zoo of specialized systems (for each subset of data with different properties), or by a new type of flexible database system that automatically adapts to the needs and characteristics of applications.

In this talk, we introduce Polypheny-DB, a novel self-adaptive polystore that provides cost- and workload aware access to heterogeneous data. As a polystore, Polypheny-DB seamlessly combines different underlying data storage engines (for instance relational databases, key-value stores, etc.) to provide good query performance independent of the type of workload.

10:20 - 10:55

Cost-based Combination of Fragmentation, Replication, Allocation & Migration in a Distributed Database

Alexander Stiemer, University of Basel

Abstract: The various problems and challenges distributed database have to deal with have led to multiple data management protocols. There are, for example , data fragmentation protocols tackling the response latency demands, or data replication protocols which increase data availability. To tackle their challenge, each of these protocols uses multiple data management techniques which can be categorized into data fragmentation, replication, allocation, and migration. A typical fragmentation protocol, for example, uses data fragmentation to split the data, data allocation to place the data onto nodes, and data migration to relocate the data in case the fragmentation scheme has changed.

While such a tight coupling provides a complete protocol, it lacks the flexibility of altering the underlying techniques, for example, the migration strategy. Therefore, we propose a more modular approach which, first, provides such flexibility, second, helps with development of new protocols since each module is easy to grasp, and third, is based on a defined cost model which integrates all four techniques.

In this talk, we will take a look at these four different data management techniques. Afterwards, we will derive generic cost models from the techniques which are then integrated into one meta cost model. At the end of the talk, we will briefly introduce a prototype implementation and sketch out its architecture.

11:10 - 11:45

A Unified Schema for Polyglot Persistence

Jesús García Molina and Diego Sevilla Ruiz, University of Murcia

Abstract: The ModelUM research group (University of Murcia, Spain) has been working on NoSQL data engineering since 2015. We have implemented strategies to infer schemas from NoSQL data and application code. Schemas extracted include variations of each entity inferred and relationship between entities: reference, aggregation and inheritance. Here, we will present the unified metamodel we have designed to represent schemas for the most widely used NoSQL paradigms: document, graph, columnar, key-value, as well as relational schemas. In addition, we will comment our work in progress around the unified metamodel. It is worth noting that we use MDE (Model-Driven Engineering) techniques to implement the created database tooling.

11:45 - 12:20

A DSL Family to Define and Manage Schemas and Data on NoSQL Databases

Alberto Hernández Chillón, University of Murcia

Abstract: Many actual NoSQL systems are schemaless, that is, the structure of the data is not defined beforehand in any schema, but it is implicit in the data itself. This characteristic is very convenient when the data structure suffers frequent changes. However, advantages of having an explicit schema such as assuring that the data stored fits the database schema are lost in favor of greater agility and flexibility. In previous work, a model-based reverse engineering approach to infer schema models from NoSQL data was proposed. Model-driven engineering (MDE) techniques can be used to take advantage of extracted models with different purposes, such as schema visualization or automatic code generation. In this work we focus on creating a family of DSLs aimed at defining and operating over NoSQL schemas, on one hand, and defining and manipulating data for NoSQL databases, on the other. This objective is divided into four tasks. The first task will be to design a NoSQL schema definition language independent of any database. Then another DSL will be designed in order to define and execute operations over this defined abstract schema. The third task will be focused on designing a data definition language used to generate content on already existing database schemas, and finally a language that allows manipulating and operating this newly generated data. These last two languages will be used to define a process which automates the generation of large datasets from an previously inferred schema. The existence of this process will be key to prove other tools to develop or extend in the future, such as the creation of Object-document mappers or

the study of identification of subtypes on inferred database schemas.

13:20 - 13:55

Towards Self-Adapting Data Migration in the Context of Schema Evolution in NoSQL Databases

Andrea Hillenbrand, Universität Darmstadt

Abstract: When NoSQL database systems are used in an agile software development setting, data model changes occur frequently and thus, data is routinely stored in different versions. This leads to an overhead affecting the software development and in particular, the management of data accesses. In this context, different data migration strategies exist, which are characterized by certain advantages and disadvantages. Using exactly that strategy whose characteristics match the according migration scenario, depends on the query workload, the changes in the data model caused by schema evolution, and the requirements for the application in terms of migration costs and latency during data accesses. In this paper we present a methodology of self-adapting data migration, which automatically adjusts migration strategies and its parameters accordingly, thereby supporting the agile software development.

13:55 - 14:30

Considerations towards a Multi-Model NoSQL Schema Evolution Benchmark

Mark Lukas Möller, University of Rostock

Abstract: To evaluate and compare the performance of various systems, using benchmarks is the common way-to-go. While benchmarks in the context of databases often measure the execution performance of transactional or analytical queries, benchmarks measuring the efficiency or correctness of schema evolution operations are very rare.

Our aim is to develop a toolkit which allows to benchmark and compare several schema evolution approaches for NoSQL databases whereby multiple NoSQL avors, such as document stores or key-value stores, shall be supported. For this, characteristics and methodologies of already existing benchmarks were investigated in a study, upon which we present basic considerations regarding an appropriate data model, proper queries and metrics deliberations towards designing a state-of-the-art multi-model NoSQL schema evolution benchmark.

14:30 - 15:05

Generating Large and Heterogeneous Test Data For Duplicate Detection

Fabian Panse, University of Hamburg

Abstract: The number of data sources has grown enormously in recent years and will continue to grow due to ongoing digitization. As the number of sources grows, so does the need and benefit of integrating them. The integration of data sources includes a detection of duplicate records. An evaluation of duplicate detection algorithms requires test data which contain the true duplicate status of the individual records. Current test data sets and generators, however, are limited to single relational tables of small sizes. Nevertheless, in times of Big and NoSQL data, data sources are becoming larger, more heterogeneous and more complex than ever before. Moreover, existing generators lack in generating realistic error patterns as they result from outdated values and copying processes. The goal of our research is to develop a new test data generator which is capable of generating test data of an arbitrary size, data model and schema complexity while producing data values and patterns that are realistic as possible. In our talk we present our vision of such a generator and describe the respective research challenges that come along with the individual generation steps.

15:20 - 15:55

Leveraging Approximate Query Processing to Realize Progressive Visual Analytics

Michael Hohenstein, University of Kaiserslautern

Abstract: Progressive Visual Analytics is a relatively new paradigm in the realm of visualization. It's main objective is to develop algorithms and infrastructure to support analysts in exploratory ad hoc data analysis. This means each query should return an (approximate) result in an upper time bound, so that the data exploration can be considered a real-time process. Additionally, the analyst should be able to steer the query by tuning parameters of the computation. The keystone of the progressive paradigm is to instantly return an approximate result which is (progressively) updated in the background. Ideally some notion of (partial) re-use of earlier results that intersect with a live query should be in place to further reduce the time to return of later queries. PVA is closely related to approximate query processing and streaming applications with a focus on real-time interactivity, which is usually reached by (partial) re-use of prior results, data- or process chunking, sampling and using fast algorithms that replace exact computations with approximate results.

We want to observe the paradigm of progressive data science through the lens of database systems, trying to improve and tailor existing approaches to create an efficient infrastructure for PVA systems.

15:55 - 16:30

Chronos - The Swiss Army Knife for Systems Evaluations

Marco Vogt, University of Basel

Abstract: Chronos is an open source system for the automation of the entire (database) systems evaluation workflow including the set-up of the evaluation, its execution, and the subsequent analysis of the results. It allows to evaluate a wide range of applications from simple databases to complex polystores or from plain text retrieval to complex multimedia search engines and thus fosters the reproducibility of evaluations.

In this talk, we will introduce Chronos and demonstrate how it supports developers in defining, executing, monitoring, and analyzing reproducible evaluations.


M72
100101001,00101
83,343434
M72
100101001,00101