Report from Dagstuhl Seminar 13251

Parallel Data Analysis

Edited by

Artur Andrzejak[1], Joachim Giesen[2], Raghu Ramakrishnan[3], and Ion Stoica[4]

1 **Universität Heidelberg, DE,** `artur@uni-hd.de`
2 **Universität Jena, DE,** `joachim.giesen@uni-jena.de`
3 **Microsoft Cloud Information Services Laboratory – Redmond, US,** `raghu@microsoft.com`
4 **University of California – Berkeley, US,** `istoica@cs.berkeley.edu`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13251 "Parallel Data Analysis" which was held in Schloss Dagstuhl – Leibniz Center for Informatics from June 16th 2013 to June 21st 2013. During the seminar, participants presented their current research and ongoing work, and open problems were discussed. The first part of this document describes seminar goals and topics, while the remainder gives an overview of the contents discussed during this event. Abstracts of a subset of the presentations given during the seminar are put together in this paper. Links to extended abstracts or full papers are provided, if available.

Seminar 16.–21. June, 2013 – www.dagstuhl.de/13251

1998 ACM Subject Classification H.2.8 Database Applications (data mining), I.2.6 Learning, I.5 Pattern Recognition, C.2.4 Distributed Systems (Distributed applications), C.4 Performance of systems

Keywords and phrases data analysis, machine learning, parallel processing, distributed computing, software frameworks

Digital Object Identifier 10.4230/DagRep.3.6.67

1 Executive Summary

Artur Andrzejak
Joachim Giesen
Raghu Ramakrishnan
Ion Stoica

Motivation and goals

Parallel data analysis accelerates the investigation of data sets of all sizes, and is indispensable when processing huge volumes of data. The current ubiquity of parallel hardware such as multi-core processors, modern GPUs, and computing clusters has created an excellent environment for this approach. However, exploiting these computing resources effectively requires significant efforts due to the lack of mature frameworks, software, and even algorithms designed for data analysis in such computing environments.

As a result, parallel data analysis is often being used only as the last resort, i.e., when the data size becomes too big for sequential data analysis, and it is hardly ever used for analyzing small and medium-sized data sets though it could be also beneficial for there, i.e., by cutting compute time down from hours to minutes or even making the data analysis process interactive. The barrier of adoption is even higher for specialists from other areas such as sciences, business, and commerce. These users often have to make do with slower, yet much easier to use sequential programming environments and tools, regardless of the data size.

The seminar participants have tried to address these challenges by focusing on the following goals:

- Providing user-friendly parallel programming paradigms and cross-platform frameworks or libraries for easy implementation and experimentation.
- Designing efficient and scalable parallel algorithms for machine learning and statistical analysis in connection with an analysis of use cases.

The program

The seminar program consisted of individual presentations on new results and ongoing work, a plenary session, as well as work in two working groups. The primary role of the focus groups was to foster the collaboration of the participants, allowing cross-disciplinary knowledge sharing and insights. Work in one group is still ongoing and targets as a result a publication in a magazine.

The topics of the plenary session and the working groups were the following ones:

- Panel "From Big Data to Big Money'
- Working group "A": Algorithms and applications
- Working group "P": Programming paradigms, frameworks and software.

2 Table of Contents

3 Abstracts of Selected Talks

3.1 Incremental-parallel Learning with Asynchronous MapReduce

Artur Andrzejak (Universität Heidelberg, DE)

Joint work of Artur Andrzejak, Joos-Hendrik Böse, Joao Bartolo Gomes, Mikael Högqvist
Main reference J.-H. Böse, A. Andrzejak, M. Högqvist, "Beyond Online Aggregation: Parallel and Incremental Data Mining with Online MapReduce," in Proc. of the 2010 Workshop on Massive Data Analytics on the Cloud (MDAC'10), 6 pp., ACM, 2010.
URL http://dx.doi.org/10.1145/1779599.1779602

MapReduce paradigm for parallel processing has turned suitable for implementing a variety of algorithms within the domain of machine learning. However, the original design of this paradigm suffers under inefficiency in case of iterative computations (due to repeated data reads from I/O) and inability to process streams or output preliminary results (due to a barrier / sync operation between map and reduce).

In the first part of this talk we propose a framework which modifies the MapReduce paradigm in twofold ways [1]. The first modification removes the barrier / sync operation, allowing reducers to process (and output) preliminary or streaming data. The second change is the mechanism to send any messages from reducers "back" to mappers. The latter property allows efficient iterative processing, as data (once read from disk or other I/O) can be kept in the main memory by map tasks, and reused in subsequent computation phases (usually, each phase being triggered by new messages/data from the reducer). We evaluate this architecture and its ability to produce preliminary results and process streams by implementing several machine learning algorithms. These include simple "one pass" algorithms like linear regression or Naive Bayes. A more advanced example is a parallel – incremental (i.e. online) version of the k-means clustering algorithm.

In the second part we focus on the issue of parallel detection of concept drift in context of classification models. We propose Online Map-Reduce Drift Detection Method (OMR-DDM) [2]. Also here our modified MapReduce framework is used. To this end, we extend the approach introduced in [3]. This is done by parallelizing training of an incremental classifier (here Naive Bayes) and the partial evaluation of its momentarily accuracy. An experimental evaluation shows that the proposed method can accurately detect concept drift while exploiting parallel processing. This paves the way to obtaining classification models which consider concept drift on massive data.

References

1 Joos-Hendrik Böse, Artur Andrzejak, Mikael Högqvist. *Beyond Online Aggregation: Parallel and Incremental Data Mining with Online MapReduce*. ACM MDAC 2010, Raleigh, NC, 2010.
2 Artur Andrzejak, Joao Bartolo Gomes. *Parallel Concept Drift Detection with Online Map-Reduce*. KDCloud 2012 at ICDM 2012, 10 December 2012, Brussels, Belgium.
3 João Gama and Pedro Medas and Gladys Castillo and Pedro Rodrigues. *Learning with drift detection*. Advances in Artificial Intelligence, 2004, pages 66–112, 2004

3.2 Scaling Up Machine Learning

Ron Bekkerman (Carmel Ventures – Herzeliya, IL)

Joint work of Bekkerman, Ron; Bilenko, Mikhail; Langford, John
Main reference R. Bekkerman, M. Bilenko, J. Langford, (eds.), "Scaling Up Machine Learning," Cambridge University Press, January 2012
URL http://www.cambridge.org/us/academic/subjects/computer-science/pattern-recognition-and-machine-learning/scaling-machine-learning-parallel-and-distributed-approaches

In this talk, I provide an extensive introduction to parallel and distributed machine learning. I answer the questions "How actually big is the big data?", "How much training data is enough?", "What do we do if we don't have enough training data?", "What are platform choices for parallel learning?" etc. Over an example of k-means clustering, I discuss pros and cons of machine learning in Pig, MPI, DryadLINQ, and CUDA.

3.3 Efficient Co-Processor Utilization in Database Query Processing

Sebastian Breß (Otto-von-Guericke-Universität Magdeburg, DE)

Joint work of Sebastian Breß, Felix Beier, Hannes Rauhe, Kai-Uwe Sattler, Eike Schallehn, and Gunter Saake
Main reference S. Breß, F. Beier, H. Rauhe, K.-U. Sattler, E. Schallehn, G. Saake, "Efficient Co-Processor Utilization in Database Query Processing," Information Systems, 38(8):1084–1096, 2013.
URL http://dx.doi.org/10.1016/j.is.2013.05.004

Co-processors such as GPUs provide great opportunities to speed up database operations by exploiting parallelism and relieving the CPU. However, distributing a workload on suitable (co-)processors is a challenging task, because of the heterogeneous nature of a hybrid processor/co-processor system. In this talk, we discuss current problems of database query processing on GPUs and present our decision model, which distributes a workload of operators on all available (co-)processors. Furthermore, we provide an overview of how the decision model can be used for hybrid query optimization.

References

1 S. Breß, F. Beier, H. Rauhe, K.-U. Sattler, E. Schallehn, and G. Saake. Efficient Co-Processor Utilization in Database Query Processing. *Information Systems*, 38(8):1084–1096, 2013.
2 S. Breß, I. Geist, E. Schallehn, M. Mory, and G. Saake. A Framework for Cost based Optimization of Hybrid CPU/GPU Query Plans in Database Systems. *Control and Cybernetics*, 41(4):715–742, 2012.

3.4 Analytics@McKinsey

Patrick Briest (McKinsey&Company – Düsseldorf, DE)

To successfully capture value from advanced analytics, businesses need to combine three important building blocks: Creative integration of internal and external data sources and the ability to filter relevant information lays the foundation. Predictive and optimization models striking the right balance between complexity and ease of use provide the means to turn data into insights. Finally, a solid embedding into the organizational processes via simple, useable tools turns insights into impactful frontline actions.

This talk gives an overview of McKinsey's general approach to big data and advanced analytics and presents several concrete examples of how advanced analytics are applied in practice to business problems from various different industries.

3.5 A Data System for Feature Engineering

Michael J. Cafarella (University of Michigan – Ann Arbor, US)

Joint work of Anderson, Michael; Antenucci, Dolan; Bittorf, Victor; Burgess, Matthew; Cafarella, Michael J.; Kumar, Arun; Niu, Feng; Park, Yongjoo; Ré, Christopher; Zhang, Ce
Main reference M. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M.J. Cafarella, A. Kumar, F. Niu, Y. Park, C. Ré, C. Zhang, "Brainwash: A Data System for Feature Engineering," in Proc. of the 6th Biennial Conf. on Innovative Data Systems Research (CIDR'13), 4 pp. , 2013.
URL http://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper82.pdf

Trained systems, such as Web search, recommendation systems, and IBM's Watson question answering system, are some of the most compelling in all of computing. However, they are also extremely difficult to construct. In addition to large datasets and machine learning, these systems rely on a large number of machine learning features. Engineering these features is currently a burdensome and time-consuming process.

We introduce a datasystem that attempts to ease the task of feature engineering. By assuming that even partially-written features are successful for some inputs, we can attempt to execute and benefit from user code that is substantially incorrect. The system's task is to rapidly locate relevant inputs for the user- written feature code with only implicit guidance from the learning task. The resulting system enables users to build features more rapidly than would otherwise be possible.

3.6 Extreme Data Mining: Global Knowledge without Global Communication

Giuseppe Di Fatta (University of Reading, GB)

Joint work of Di Fatta, Giuseppe; Blasa, Francesco; Cafiero, Simone; Fortino, Giancarlo
Main reference G. Di Fatta, F. Blasa, S. Cafiero, G. Fortino. "Fault tolerant decentralised k-Means clustering for asynchronous large-scale networks," Journal of Parallel and Distributed Computing, Vol. 73, Issue 3, March 2013, pp. 317–329, 2013.
URL http://dx.doi.org/10.1016/j.jpdc.2012.09.009

Parallel Data Mining in very large and extreme-scale systems is hindered by the lack of scalable and fault tolerant global communication and synchronisation methods. Epidemic protocols are a type of randomised protocols which provide statistical guarantees of accuracy and consistency of global aggregates in decentralised and asynchronous networks. Epidemic K-Means is the first data mining protocol which is suitable for very large and extreme-scale systems, such as Peer-to-Peer overlay networks, the Internet of Things and exascale supercomputers. This distributed and fully-decentralised K-Means formulation provides a clustering solution which can approximate the solution of an ideal centralised algorithm over the aggregated data as closely as desired. A comparative performance analysis with the state of the art sampling methods is presented.

3.7 Parallelization of Machine Learning Tasks by Problem Decomposition

Johannes Fürnkranz (TU Darmstadt, DE)

Joint work of Fürnkranz, Johannes; Hüllermeier, Eyke

In this short presentation I put forward the idea that parallelization can be achieved by decomposing a complex machine learning problem into a series of simpler problems than can be solved independently, and collectively provide the answer to the original problem. I illustrate this on the task of pairwise classification, which solves a multi-class classification problem by reducing it to a set of binary classification problems, one for each pair of classes. Similar decompositions can be applied to problems like preference learning, ranking, multilabel classification, or ordered classification. The key advantage of this approach is that it gives many small problems, the main disadvantage is that the number of examples that have to be distributed over multiple cores increases n-fold.

3.8 Sclow Plots: Visualizing Empty Space

Joachim Giesen (Universität Jena, DE)

Joint work of Giesen, Joachim; Kühne, Lars; Lucas, Philipp

Scatter plots are mostly used for correlation analysis, but are also a useful tool for understanding the distribution of high-dimensional point cloud data. An important characteristic of such distributions are clusters, and scatter plots have been used successfully to identify clusters in data. Another characteristic of point cloud data that has received less attention are regions that contain no or only very few data points. We show that augmenting scatter plots by projections of flow lines along the gradient vector field of the distance function to the point cloud reveals such empty regions or voids. The augmented scatter plots, that we call sclow plots, enable a much better understanding of the geometry underlying the point cloud than traditional scatter plots.

3.9 Financial and Data Analytics with Python

Yves J. Hilpisch (Visixion GmbH, DE)

Main reference Y.J. Hilpisch, "Derivatives Analytics with Python – Data Analysis, Models, Simulation, Calibration, Hedging," Visixion GmbH.
URL http://www.visixion.com/?page_id=895

The talk illustrates, by the means of concrete examples, how Python can help in implementing efficient, interactive data analytics. There are a number of libraries available, like pandas or PyTables, that allow high performance analytics of e.g. time series data or out-of-memory data. Examples shown include financial time series analytics and visualization, high frequency data aggregation and analysis and parallel calculation of option prices via Monte Carlo simulation. The talk also compares out-of-memory analytics using PyTables with in-memory analytics using pandas.

Continuum Analytics specializes in Python-based Data Exploration & Visualization. It is engaged in a number of Open Source projects like Numba (just-in-time compiling of Python code) or Blaze (next-generation disk-based, distributed arrays for Python). It also provides the free Python distribution Anaconda for scientific and enterprise data analytics.

3.10 Convex Optimization for Machine Learning Made Fast and Easy

Soeren Laue (Universität Jena, DE)

Joint work of Giesen, Joachim; Mueller, Jens; Laue, Soeren

In machine learning, solving convex optimization problems often poses an efficiency vs. convenience trade-off. Popular modeling languages in combination with a generic solver allow to formulate and solve these problems with ease, however, this approach does typically not scale well to larger problem instances. In contrast to the generic approach, highly efficient solvers consider specific aspects of a concrete problem and use optimized parameter settings. We describe a novel approach that aims at achieving both goals at the same time, namely, the ease of use of the modeling language/generic solver combination, while generating production quality code that compares well with specialized, problem specific implementations. We call our approach a generative solver for convex optimization problems from machine learning (GSML). It outperforms state-of-the-art approaches of combining a modeling language with a generic solver by a few orders of magnitude.

3.11 Interactive, Incremental, and Iterative Dataflow with Naiad

Frank McSherry (Microsoft – Mountain View, US)

Joint work of McSherry, Frank; Murray, Derek; Isaacs, Rebecca; Isard, Michael
URL http://research.microsoft.com/naiad/

This talk will cover a new computational frameworks supported by Naiad, differential dataflow, that generalizes standard incremental dataflow for far greater re-use of previous results when collections change. Informally, differential dataflow distinguishes between the multiple reasons a collection might change, including both loop feedback and new input data, allowing a system to re-use the most appropriate results from previously performed work when an incremental update arrives. Our implementation of differential dataflow efficiently executes queries with multiple (possibly nested) loops, while simultaneously responding with low latency to incremental changes to the inputs. We show how differential dataflow enables orders of magnitude speedups for a variety of workloads on real data, and enables new analyses previously not possible in an interactive setting.

3.12 Large Scale Data Analytics: Challenges, and the role of Stratified Data Placement

Srinivasan Parthasarathy (Ohio State University, US)

Joint work of Parthasarathy, Srinivasan; Wang, Ye; Chakrabarty, Aniket; Sadayappan, P
Main reference Y. Wang, S. Parthasarathy, P. Sadayappan, "Stratification driven placement of complex data: A framework for distributed data analytics," in Proc. of IEEE 29th Int'l Conf. on Data Engineering (ICDE'13), pp. 709–720, IEEE, 2013.
URL http://dx.doi.org/10.1109/ICDE.2013.6544868

With the increasing popularity of XML data stores, social networks and Web 2.0 and 3.0 applications, complex data formats, such as trees and graphs, are becoming ubiquitous. Managing and processing such large and complex data stores, on modern computational eco-systems, to realize actionable information efficiently, is daunting. In this talk I will begin with discussing some of these challenges. Subsequently I will discuss a critical element at the heart of this challenge relates to the placement, storage and access of such tera- and peta-scale data. In this work we develop a novel distributed framework to ease the burden on the programmer and propose an agile and intelligent placement service layer as a flexible yet unified means to address this challenge. Central to our framework is the notion of stratification which seeks to initially group structurally (or semantically) similar entities into strata. Subsequently strata are partitioned within this eco-system according to the needs of the application to maximize locality, balance load, or minimize data skew. Results on several real-world applications validate the efficacy and efficiency of our approach.

3.13 Big Data @ Microsoft

Raghu Ramakrishnan (Microsoft CISL, Redmond, WA, US)

Joint work of Raghu Ramakrishnan; CISL team at Microsoft

The amount of data being collected is growing at a staggering pace. The default is to capture and store any and all data, in anticipation of potential future strategic value, and vast amounts of data are being generated by instrumenting key customer and systems touchpoints. Until recently, data was gathered for well-defined objectives such as auditing, forensics, reporting and line-of-business operations; now, exploratory and predictive analysis is becoming ubiquitous. These differences in data scale and usage are leading to a new generation of data management and analytic systems, where the emphasis is on supporting a wide range of data to be stored uniformly and analyzed seamlessly using whatever techniques are most appropriate, including traditional tools like SQL and BI and newer tools for graph analytics and machine learning. These new systems use scale-out architectures for both data storage and computation.

Hadoop has become a key building block in the new generation of scale-out systems. Early versions of analytic tools over Hadoop, such as Hive and Pig for SQL-like queries, were implemented by translation into Map-Reduce computations. This approach has inherent limitations, and the emergence of resource managers such as YARN and Mesos has opened the door for newer analytic tools to bypass the Map-Reduce layer. This trend is especially significant for iterative computations such as graph analytics and machine learning, for which Map-Reduce is widely recognized to be a poor fit. In this talk, I will examine this architectural trend, and argue that resource managers are a first step in re-factoring the early implementations of Map-Reduce, and that more work is needed if we wish to support a variety of analytic tools on a common scale-out computational fabric. I will then present REEF, which runs on top of resource managers like YARN and provides support for task monitoring and restart, data movement and communications, and distributed state management. Finally, I will illustrate the value of using REEF to implement iterative algorithms for graph analytics and machine learning.

3.14 Berkeley Data Analytics Stack (BDAS)

Ion Stoica (University of California – Berkeley, US)

One of the most interesting developments over the past decade is the rapid increase in data; we are now deluged by data from on-line services (PBs per day), scientific instruments (PBs per minute), gene sequencing (250GB per person) and many other sources. Researchers and practitioners collect this massive data with one goal in mind: extract "value" through sophisticated exploratory analysis, and use it as the basis to make decisions as varied as personalized treatment and ad targeting. Unfortunately, today's data analytics tools are slow in answering even simple queries, as they typically require to sift through huge amounts of data stored on disk, and are even less suitable for complex computations, such as machine learning algorithms. These limitations leave the potential of extracting value of big data unfulfilled.

To address this challenge, we are developing BDAS, an open source data analytics stack that provides interactive response times for complex computations on massive data. To achieve this goal, BDAS supports efficient, large-scale in-memory data processing, and allows users and applications to trade between query accuracy, time, and cost. In this talk, I'll present the architecture, challenges, early results, and our experience with developing BDAS. Some BDAS components have already been released: Mesos, a platform for cluster resource management has been deployed by Twitter on +6,000 servers, while Spark, an in-memory cluster computing frameworks, is already being used by tens of companies and research institutions.

3.15 Scalable Data Analysis on Clouds

Domenico Talia (University of Calabria, IT)

URL http://gridlab.dimes.unical.it

This talk presented a Cloud-based framework designed to program and execute parallel and distributed data mining applications: The Cloud Data Mining Framework. It can be used to implement parameter sweeping applications and workflow-based applications that can be programmed through a graphical interface and trough a script-based interface that allow to compose a concurrent data mining program to be run on a Cloud platform. We presented the main system features and its architecture. In the Cloud Data Mining framework each node of a workflow is a service, so the application is composed o a collection of Cloud services.

3.16 Parallel Generic Pattern Mining

Alexandre Termier (University of Grenoble, FR)

Joint work of Termier, Alexandre; Negrevergne, Benjamin; Mehaut, Jean-Francois; Rousset, Marie-Christine
Main reference B. Negrevergne, A. Termier, M.-C. Rousset, J.-F. Méhaut, "ParaMiner: a generic pattern mining algorithm for multi-core architectures," iData Mining and Knowledge Discovery, April 2013, Springer, 2013.
URL http://dx.doi.org/10.1007/s10618-013-0313-2

Pattern mining is the field of data mining concerned with finding repeating patterns in data. Due to the combinatorial nature of the computations performed, it requires a lot of computation time and is therefore an important target for parallelization. In this work we show our parallelization of a generic pattern mining algorithm, and how the pattern definition influes on the parallel scalability. We also show that the main limiting factor is in most cases the memory bandwidth, and how we could overcome this limitation.

3.17 REEF: The Retainable Evaluator Execution Framework

Markus Weimer (Microsoft CISL, Redmond, WA, US)

Joint work of Chun, Byung-Gon; Condie, Tyson; Curino, Carlo; Douglas, Chris; Narayanamurthy, Shravan; Ramakrishnan, Raghu; Rao, Sriram; Rosen, Joshua; Sears, Russel; Weimer, Markus

The Map-Reduce framework enabled scale-out for a large class of parallel computations and became a foundational part of the infrastructure at Web companies. However, it is recognized that implementing other frameworks such as SQL and Machine Learning by translating them into Map-Reduce programs leads to poor performance.

This has led to a refactoring the Map-Reduce implementation and the introduction of domain-specific data processing frameworks to allow for direct use of lower-level components. Resource management has emerged as a critical layer in this new scale-out data processing stack. Resource managers assume the responsibility of multiplexing fine-grained compute tasks on a cluster of shared-nothing machines. They operate behind an interface for leasing *containers*—a slice of a machine's resources (e.g., CPU/GPU, memory, disk)—to computations in an elastic fashion.

In this talk, we describe the Retainable Evaluator Execution Framework (REEF). It makes it easy to retain state in a container and reuse containers across different tasks. Examples include pipelining data between different operators in a relational pipeline; retaining state across iterations in iterative or recursive distributed programs; and passing state across different types of computations, for instance, passing the result of a Map-Reduce computation to a Machine Learning computation.

REEF supports this style of distributed programming by making it easier to: (1) interface with resource managers to obtain containers, (2) instantiate a runtime (e.g., for executing Map-Reduce or SQL) on allocated containers, and (3) establish a control plane that embodies the application logic of how to coordinate the different tasks that comprise a job, including how to handle failures and preemption. REEF also provides data management and communication services that assist with task execution. To our knowledge, this is the first approach that allows such reuse of dynamically leased containers, and offers potential for order-of-magnitude performance improvements by eliminating the need to persist state (e.g., in a file or shared cache) across computational stages.

4 Group Composition and Schedule

4.1 Participants

The seminar has brought together academic researchers and industry practitioners to foster cross-disciplinary interactions on parallel analysis of scientific and business data. The following three communities were particularly strongly represented:

- researchers and practitioners in the area of frameworks and languages for data analysis
- researchers focusing on machine learning and data mining
- practitioners analysing data of various sizes in the domains of finance, consulting, engineering, and others.

In summary, the seminar gathered 36 researchers from the following 10 countries:

Country	Number of participants
Canada	1
France	1
Germany	13
Israel	1
Italy	1
Korea	1
Portugal	1
Singapore	1
UK	1
USA	15

Most participants came from universities or state-owned research centers. However, a considerable fraction of them were affiliated with industry or industrial research centers – altogether, 13 participants. Here is a detailed statistic of the affiliations:

Industry	Institution	Country	# Participants
	Argonne National Laboratory	USA	1
	Brown University – Providence	USA	1
Yes	Carmel Ventures – Herzeliya	Israel	1
	Freie Universität Berlin	Germany	1
Yes	Institute for Infocomm Research (I2R)	Singapore	1
Yes	McKinsey & Company	Germany	1
Yes	Microsoft and Microsoft Research	USA	6
	Ohio State University	USA	1
	Otto-von-Guericke-Universität Magdeburg	Germany	1
Yes	SAP AG	Germany	2
Yes	SpaceCurve	USA	1
	Stony Brook University / SUNY Korea	USA / Korea	1
	TU Berlin	Germany	1
	TU Darmstadt	Germany	1
	Universidade do Porto	Portugal	1
	Universität Heidelberg	Germany	2
	Universität Jena	Germany	3
	University of Alberta	Canada	1
	University of Calabria	Italy	1
	University of California – Berkeley	USA	3
	University of Grenoble	France	1
	University of Michigan	USA	1
	University of Minnesota	USA	1
	University of Reading	UK	1
Yes	Visixion GmbH / Continuum Analytics	Germany	1

4.2 Complete list of talks

Monday, June 17th 2013

S1: Applications

Krishnaswamy, Shonali	Mobile & Ubiquitous DataStream Mining
Broß, Jürgen	Mining Customer Review Data
Will, Hans-Martin	Real-time Analysis of Space and Time

S2: Frameworks I

Peterka, Tom	Do-It-Yourself Parallel Data Analysis
Joseph, Anthony D.	Mesos
Zaharia, Matei	The Spark Stack: Making Big Data Analytics Interactive and Real-time

Tuesday, June 18th 2013

S3: Overview & Challenges I

Bekkerman, Ron	Scaling Up Machine Learning: Parallel and Distributed Approaches
Ramakrishnan, Raghu	Big Data @ Microsoft

S4: Overview & Challenges II

Briest, Patrick	Analytics @ McKinsey
Parthasarathy, Srinivasan	Scalable Analytics: Challenges and Renewed Bearing

S5: Frameworks II

Stoica, Ion	Berkeley Data Analytics Stack (BDAS)
Hilpisch, Yves	Financial and Data Analytics with Python
Cafarella, Michael J.	A Data System for Feature Engineering

Wednesday, June 19th 2013

S6: Visualisation and Interactivity

Giesen, Joachim	Visualizing empty space
McSherry, Frank	Interactive, Incremental, and Iterative Data Analysis with Naiad

S7: Various

Müller, Klaus	GPU-Acceleration for Visual Analytics Tasks
Laue, Soeren	Convex Optimization for Machine Learning made Fast and Easy
Di Fatta, Giuseppe	Extreme Data Mining: Global Knowledge without Global Communication

Thursday, June 20th 2013

S8: Frameworks III

Talia, Domenico	Scalable Data Analysis workflows on Clouds
Weimer, Markus	REEF: The Retainable Evaluator Execution Framework
Termier, Alexandre	Prospects for parallel pattern mining on multicores

S9: Efficiency

Andrzejak, Artur	Incremental-parallel learning with asynchronous MapReduce
Fürnkranz, Johannes	Parallelization of machine learning tasks via problem decomposition
Breß, Sebastian	Efficient Co-Processor Utilization in Database Query Processing

Participants

- Artur Andrzejak
Universität Heidelberg, DE
- Ron Bekkerman
Carmel Ventures – Herzeliya, IL
- Joos-Hendrik Böse
SAP AG – Berlin, DE
- Sebastian Breß
Universität Magdeburg, DE
- Patrick Briest
McKinsey&Company – Düsseldorf, DE
- Jürgen Broß
FU Berlin, DE
- Lutz Büch
Universität Heidelberg, DE
- Michael J. Cafarella
University of Michigan – Ann Arbor, US
- Surajit Chaudhuri
Microsoft Res. – Redmond, US
- Tyson Condie
Yahoo! Inc. – Burbank, US
- Giuseppe Di Fatta
University of Reading, GB
- Rodrigo Fonseca
Brown University, US
- Johannes Fürnkranz
TU Darmstadt, DE
- Joao Gama
University of Porto, PT
- Joachim Giesen
Universität Jena, DE
- Philipp Große
SAP AG – Walldorf, DE
- Max Heimel
TU Berlin, DE
- Yves J. Hilpisch
Visixion GmbH, DE
- Anthony D. Joseph
University of California – Berkeley, US
- George Karypis
University of Minnesota – Minneapolis, US
- Shonali Krishnaswamy
Infocomm Research – Singapore, SG
- Soeren Laue
Universität Jena, DE
- Frank McSherry
Microsoft – Mountain View, US
- Jens K. Müller
Universität Jena, DE
- Klaus Mueller
Stony Brook University, US
- Srinivasan Parthasarathy
Ohio State University, US
- Tom Peterka
Argonne National Laboratory, US
- Raghu Ramakrishnan
Microsoft Res. – Redmond, US
- Ion Stoica
University of California – Berkeley, US
- Domenico Talia
University of Calabria, IT
- Alexandre Termier
University of Grenoble, FR
- Markus Weimer
Microsoft Res. – Redmond, US
- Hans-Martin Will
SpaceCurve – Seattle, US
- Matei Zaharia
University of California – Berkeley, US
- Osmar Zaiane
University of Alberta, CA