Technische Universität Berlin

Contributor Names and Short CVs

1. Asterios Katsifodimos (asterios.katsifodimos@tu-berlin.de) is a Postdoctoral Researcher working on the *Stratosphere* Research Project (getstratosphere.org) in the Database Systems and Information Management (DIMA) Group at the Technische Universität Berlin (TUB). He received his PhD in 2013 from INRIA Saclay and Université Paris-Sud under the supervision of Ioana Manolescu. His PhD thesis focused on materialized view-based techniques for the management of web data. He was a member of the High Performance Computing Lab at the University of Cyprus, where he obtained his B.Sc. and M.Sc. degrees. His research interests include query optimization, large-scale distributed data management, and big data analytics.

2. Volker Markl (volker.markl@tu-berlin.de) is a Full Professor and Chair of the DIMA Group and Speaker for the Data Analytics Lab (www.analytics.tu-berlin.de) at TUB. In addition, he is the Speaker of a German National Science Foundation (DFG) funded Research Unit called *Stratosphere*, which continues to (further) develop a next-generation big data analytics platform. Earlier this year, under his leadership, a study on big data challenges and opportunities was conducted for the German Federal Ministry of Economics & Technology (BMWi). To date, he has given over 200 invited talks and published over 80 research papers at world-class scientific venues. His research interests include new hardware architectures for information management, scalable processing and optimization of data programming languages, information processing, and information modeling.

3. Juan Soto (juan.soto@tu-berlin.de) is a Senior Project Manager in the DIMA Group at TUB. His responsibilities include grant writing, conducting research in data science related topics (e.g., large scale data generation and validation, numerical issues in data analytics), among others. He holds an M.S. in Applied Mathematics from Stony Brook University, an M.S. in Computer Science from the University of Delaware, and a Graduate Certificate in Federal Statistics from George Mason University. His areas of expertise include cyber security, numerical analysis, and statistical computing. Over his 18 year career, he has worked in varying domains, including cyber security and mathematical/statistical science.

4. Kostas Tzoumas (kostas.tzoumas@tu-berlin.de) is a Postdoctoral Researcher at TUB, where he is working on *Stratosphere*, an open source next-generation platform for big data analytics. Prior to joining TUB, Kostas was at the National Technical University of Athens (Greece), Aalborg University (Denmark), Microsoft Research (USA), and the University of Maryland, College Park (USA). Kostas has co-authored several publications, regularly serves as a PC member at top data management conferences, and initiated the DanaC (Data Analytics in the Cloud) series of workshops held in conjunction with ACM SIGMOD. His research interests include architectures and programming models for parallel data management systems, query processing, and query optimization.

Technische Universität Berlin

Type of the Presentation Proposed: Research Contribution

Title of the Presentation: *"The Stratosphere Platform for Big Data Analytics"*

Summary of the Presentation (100 words): In this talk, we will present *Stratosphere* (getstratosphere.org), a European developed open source software stack for complex big data analytics currently available for download. *Stratosphere* covers a wide-variety of big data use cases, including data warehousing, information extraction/integration, data cleansing, graph analysis, and statistical analysis. Its unique set of features enables easy, efficient, and expressive programming that is well suited for the development of scalable data analytics. *Stratosphere's* features include "in situ" data processing, a declarative query language, treatment of user-defined functions as first-class citizens, automatic program parallelization and optimization, support for iterative programs, and an efficient, scalable execution engine.

Extended Abstract of the Presentation (1-4 pages in 11pt A4 format)

The Big Data Era. The last decade was marked by the digitization of virtually all aspects of our daily lives. Due to the decline in the price of disk storage and the increasing popularity of cloud storage, businesses, citizens, and public institutions, among others, face an avalanche of digital data on a daily basis. What looks like a huge gain in information for the networked society at a first glance, turns out to be a curse. Data, per se is neither information nor knowledge. On the contrary, only intelligent questions concerning data will produce real information and thus economic and social value.

Big data creates new demands on data management and data analysis systems that cannot be met by the current state of the art. Big data analysis systems will need to handle huge amounts of data in the Terabyte and Petabyte range, and beyond (the **data volume** challenge). Data analysis systems will need to provide accurate analyses with low latency despite potentially high data rates (the **data velocity** challenge). At the same time, this data is often heterogeneous (the **data variety** challenge) and exists in several data formats (e.g., relational, hierarchical, as a graph, in the form of text, image, audio, or video). Due to inherent uncertainty in experiments, simulations, and algorithms, the confidence associated with the analysis results, also plays an important role (the **data veracity** challenge). Furthermore, the complexity of the analysis incurred by using machine-learning algorithms or applying signal, voice, image, and video processing methods has increased dramatically. Thus, a data analysis system has to promptly process complex algorithms of linear algebra, statistics or mathematical optimization, which often consist of user-defined functions and iterative, stateful algorithms, far beyond the usual operations of relational algebra (and thus of traditional SQL database systems) and the current big data solutions like Hadoop.

The Apache Hadoop System. Hadoop programmers are limited to two functional operators, namely, *map* and *reduce*, which are scheduled in two execution stages on a set of clusters. The MapReduce programming model and the Hadoop system are suitable for a number of use cases (e.g., for information retrieval), however, they have been considered too rigid for implementing iterative algorithms and important database operators in an efficient manner. The MapReduce

paradigm forces programmers to specify their programs in terms of an initial *map*, which is then subsequently followed by a *reduce* operator. Furthermore, data exchange must be conducted using the Hadoop Distributed File System (HDFS), which seriously limits data flow throughput in Hadoop clusters. These shortcomings result in Hadoop being inefficient for both iterative computations (e.g., K-Means and other machine learning algorithms) and complex analytical database queries.

The Stratosphere System. The Stratosphere system implements a more general execution model represented as a directed acyclic graph (DAG), where the vertices of the graph represent data-parallel operators (e.g., *Map*, *Reduce*, *Join*, *GroupBy*) and the edges between the operators represent the data flow between them. Data exchange does not have to take place through a file system; instead, it is streamed between operators, achieving very high performance. Unlike the Hadoop system, Stratosphere processes complex data analysis programs containing iterations, complex UDFs, and a more declarative specification of data flows.

System Architecture and Features. Stratosphere offers two programming language APIs, one in Java and another in Scala. It also features a program optimizer, which makes low-level decisions concerning data transfer between machines or intermediate data processing. Stratosphere's optimizer raises the level of abstraction, freeing programmers from having to make low-level decisions (e.g., concerning data flow execution). The programming abstraction and optimizer build on key research results involving static code analysis, compiler construction, and database optimization. In addition, Stratosphere features a high-performance parallel runtime that executes optimized programs across numerous machines. Its runtime was designed to be efficient and it was developed from scratch to support advanced analytics use cases. Finally, Stratosphere was designed to be compatible with an existing Hadoop installation (e.g., YARN), while offering additional benefits, namely, programming abstraction and runtime capabilities.

Value to European Businesses and Use Cases. Data analysis is at the heart of many IT applications and simultaneously a powerful tool to increase efficiency in various business settings. Stratosphere was designed with generality and expressiveness in mind. It is apt at helping European businesses solve a large variety of data processing problems.

Currently, a number of companies around Europe are conducting pilot studies using Stratosphere. Such companies include Deutsche Telekom in Germany and Internet Memory Research (IMR) in France. IMR is currently developing tools using Stratosphere in order to collect and extract information at large scale (millions of sites and social networks) and then perform data analysis for web intelligence. In addition, companies taking part in the DOPA project (dopa-project.eu) are also using Stratosphere in order to establish a repository of valuable business data, and make them available to data seekers. Finally, other use cases where Stratosphere is used involve customer risk management, targeted advertising, text analytics/natural language processing, log file analysis to optimize online presence, training recommenders, topic classification, and training fraud detection models.

Technische Universität Berlin

Project Status and Future. Stratosphere was developed by the Technische Universität Berlin together with researchers from Humboldt University in Berlin and the Hasso Plattner Institute in Potsdam. It is an open-source system and is distributed under the Apache 2.0 license. Stratosphere was awarded with an IBM SUR Award and a Hewlett Packard Open Innovation Award and serves as the flagship project for big data analytics for the EIT (European Institute of Innovation and Technology) ICT Labs.

Recently, the Stratosphere research team received additional funding, which enables us to continue to further develop the platform. Greater still, the open source community is also complementing our efforts with their contributions. The next generation of the platform (Stratosphere II) will focus on the declarative specification of data analysis programs, on supporting streaming data processing, and on conducting deep analysis for big data, while automatically optimizing machine learning and iterative data analysis programs. Currently, development efforts are concentrated on implementing a machine-learning library on top of Stratosphere. Additionally, support for the R programming language is planned, as well as the development of a workflow system based on the Scala programming language.

For more information about *Stratosphere* we invite you to visit the project's website at **getstratosphere.org**.