Cloud Data Systems: What are the Opportunities for the Database Research Community?

Magdalena Balazinska University of Washington, Seattle

Joseph M. Hellerstein University of California, Berkeley

> Matei Zaharia Stanford University

Surajit Chaudhuri Microsoft Research

Hanuma Kodavalla Microsoft AnHai Doan University of Wisconsin, Madison

> Ippokratis Pandis Amazon Web Services

ABSTRACT

The panel will discuss the research opportunities for the database research community in the context of cloud native data services.

PVLDB Reference Format:

Magdalena Balazinska, Surajit Chaudhuri, AnHai Doan, Joseph M. Hellerstein, Hanuma Kodavalla, Ippokratis Pandis, and Matei Zaharia. Cloud Data Systems: What are the Opportunities for the Database Research Community? PVLDB, 15(12): 3826 - 3827, 2022. doi:10.14778/3554821.3554905

1 PANEL OVERVIEW

Public Infrastructure Clouds have transformed our approach to computing. They have led to emergence of several "Cloud Native" data services that leverage the unique functionality of infrastructure clouds, especially their ability to provision, scale up, and scale down resources elastically. This panel will take an in-depth look at the cloud native data services, their key differences with traditional database software, and the unique technical challenges these systems must address. Specifically, we will consider these aspects in the context of OLTP/HTAP, Data Analytics, Data Integration, and Data Governance.

Infrastructure Clouds continue to evolve. We have seen a transition from provisioned to serverless resource allocation. The disaggregation of storage and compute is common in infrastructure cloud offerings. Currently, disaggregation of memory is being actively explored. With the end of Dennard scaling, there is increasing interest in leveraging accelerators, leading to increasingly heterogeneous compute platforms. Edge IoT devices as endpoints of the infrastructure cloud introduce new opportunities and challenges. Data governance has emerged as a critical need. The industry has proposed data clouds as a value-add abstraction layer on top of multiple infrastructure clouds. The panel will reflect

Proceedings of the VLDB Endowment, Vol. 15, No. 12 ISSN 2150-8097. doi:10.14778/3554821.3554905 on these developments and discuss what innovations will be needed for the next generation cloud native data services.

The panel will also take stock of the contributions the research community has made in the evolution of cloud native data services. While hindsight is always 20:20, the panel will reflect on what opportunities the research community may have missed, if any, for even greater impact. As part of that discussion, the panelists will discuss if there are any structural impediments that need to be removed to ensure greater impact of the community in the future. We will then look forward and ask ourselves what challenges the research community can address to influence the future of cloud native data services. We will end with an interactive Q&A session with the audience.

2 PANEL MEMBERS

Our panel consists of six senior members of the database and systems community. Together, they represent expertise across different facets of data systems and are experienced in cloud native services. Thus, they are uniquely positioned to answer the key questions the panel will be addressing. We now provide a brief introduction to the panelists:

Magdalena Balazinska is the Director of the Paul G. Allen School of Computer Science & Engineering at the University of Washington. Her current research focuses on data management for data science, big data systems, cloud computing, and image and video analytics. She served as a PC co-chair for the VLDB'20 conference. Magdalena is an ACM Fellow. She holds a Ph.D. from the Massachusetts Institute of Technology (2006). She was named a Microsoft Research New Faculty Fellow (2007). Magdalena received the inaugural VLDB Women in Database Research Award (2016) for her work on scalable distributed data systems. She also received an ACM SIGMOD Test-of-Time Award (2017) for her work on fault-tolerant distributed stream processing and a 10-year most influential paper award (2010) from her earlier work on reengineering software clones.

AnHai Doan is Vilas Distinguished Achievement Professor and Gurindar S. Sohi Professor of Computer Science at the University of Wisconsin-Madison. His interests cover data integration, data science, and machine learning. AnHai received the ACM Doctoral Dissertation Award in 2003, a CAREER Award in 2004, and a Sloan Fellowship in 2007. He co-authored Principles of Data

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Integration, a Morgan-Kaufmann textbook in 2012. Data integration technologies from his group have been productionized at several companies, including Google, Walmart, and Informatica. Currently, AnHai also works as Vice President of Technology at Informatica, helping to build a broad range of data integration platforms on the cloud, using databases, machine learning, and big data technologies.

Joseph M. Hellerstein is the Jim Gray Professor of Computer Science at the University of California, Berkeley, whose work focuses on data-centric systems and the way they drive computing. He is an ACM Fellow, an Alfred P. Sloan Research Fellow and the recipient of three ACM-SIGMOD "Test of Time" awards for his research. Hellerstein is a co-founder of Aqueduct, which is bringing new open source technology for Prediction Infrastructure and cloud services to market. Previously he co-founded Trifacta, the pioneering company in Data Preparation based on the Wrangler research project. Hellerstein has served on the technical advisory boards of a number of computing and Internet companies including DellEMC, Greenplum, SurveyMonkey, Acryl Data, Brim Data, and Datometry.

Hanuma Kodavalla is a Technical Fellow in the Azure Databases group at Microsoft where he has been for twenty years. He previously worked at Data General, Digital Equipment Corporation, Oracle, Sybase and Asera. For more than three decades, Hanuma worked on many aspects of Relational Database Systems and has been instrumental in architecting multiple commercial database systems for high performance and high availability. Hanuma received BTech in Electronics and Communications in 1981 from National Institute of Technology, Warangal, India, MTech in Computer Science in 1983 from Indian Institute of Technology, Chennai, India, and MS in Computer Science in 1988 from University of Massachusetts, Amherst, USA. He has many patents related to novel implementation techniques for online transaction processing and data warehousing in the areas of concurrency control, recovery, high-availability, query processing and security.

Ippokratis Pandis is a Senior Principal Engineer at Amazon Web Services. He spends most of his time on AWS's Analytics services, especially Amazon Redshift. Redshift is Amazon's fully managed, petabyte-scale data warehouse service. Previously, Ippokratis has held positions as software engineer at Cloudera where he worked on the Impala SQL-on-Hadoop query engine, and as member of the research staff at the IBM Almaden Research Center, where he worked on IBM DB2 BLU. Ippokratis received his PhD from the Electrical and Computer Engineering department at Carnegie Mellon University. He is the recipient of Best Demo awards at ICDE 2006 and SIGMOD 2011, and Test-of-Time award at EDBT 2019. He has served as PC chair of DaMoN 2014, DaMoN 2015, CloudDM 2016, HPTS 2019 and ICDE Industrial 2022. He is the General Chair of SIGMOD 2023 and the president of HPTS.

Matei Zaharia is a Coufounder and Chief Technologist at Databricks as well as an Assistant Professor of Computer Science at Stanford. He started the Apache Spark project during his PhD at UC Berkeley in 2009, and has worked broadly on other widely used data and machine learning software, including MLflow, Delta Lake and Apache Mesos. He works on a wide variety of projects in data management and machine learning at Databricks and Stanford. Matei's research was recognized through the 2014 ACM Doctoral Dissertation Award, an NSF CAREER Award, and the US Presidential Early Career Award for Scientists and Engineers (PECASE).

Surajit Chaudhuri (Moderator) is a Distinguished Scientist at Microsoft Research and leads the Data Systems Research Group. His current areas of interest are data analytics for big data platforms, self-manageability, and cloud database services. Working with his colleagues in Microsoft Research, he helped incorporate the Index Tuning Wizard (and subsequently the Database Engine Tuning Advisor) and Data Cleaning technology in Microsoft SQL Server. Surajit is an ACM Fellow, a recipient of the ACM SIGMOD Edgar F. Codd Innovations Award, ACM SIGMOD Contributions Award, a VLDB 10-year Best Paper Award, and an IEEE Data Engineering Influential Paper Award. Surajit received his Ph.D. from Stanford University.