Tensor Query Processing: How to ride the AI investment wave for database analytics?

Carlo Curino[1]

[1]*Microsoft, USA*

Abstract

On 5th July 2023, Carlo Curino delivered a keynote talk at the 31st Symposium on Advanced Database Systems in Galzignano Terme (Padua, Italy). The following is the abstract of his talk and a short biography.

Keywords

Tensor Query Processing, Database Analytics, SEBD 2023 Keynote

Abstract of the Keynote

The successes of modern AI have fueled huge investments in highly parallel computational devices such as GPUs/APUs/TPUs and the development of corresponding tensor-centric runtimes/compilers/optimizers. In this talk, I summarized a significant research investment within Microsoft to embrace this tensor-based world to accelerate database analytics. We have successfully developed a framework to compile SQL to the same tensor abstraction used by popular runtimes such as PyTorch/ONNX/TVM. This allows us to run SQL on GPUs/APUs from multiple vendors and with limited effort we can port to TPUs or other custom tensor-based devices. With careful design of the tensor-relational-operators and leveraging custom kernel-fusion compilers we outperform on perf and price/perf state-of-the-art CPU systems (SQLServer/Spark/Snowflake), and match or outperform GPU-native analytics systems. Even more encouraging is the performance gains we obtain by newer HW generations, and the multitude of obvious optimizations our SW is still lacking. Interestingly, all of this can be done with a mere 20k lines of Python code. We conclude by showcasing other interesting side advantages of this approach that blend SQL+ML in interesting new ways. While the journey towards a production system is still long and treacherous, these initial few steps are very encouraging.

Short Biography

Carlo A. Curino is the lead of Gray Systems Lab (GSL), and applied research group working at the intersection of Databases/Systems/Machine Learning. Before this Carlo was a Principal Scientist in Cloud and Information Services Lab (CISL), working on large-scale distributed systems, with a focus on scheduling for BigData clusters; this line of research was co-developed

SEBD 2023: 31st Symposium on Advanced Database Systems, July 02–05, 2023, Galzignano Terme, Padua, Italy

carlo.curino@microsoft.com (C. Curino)

0000-0003-3712-7358 (C. Curino)

with several team members and open-sourced as part of Apache Hadoop/YARN. Intrinsically, this research work enables us to operate the largest YARN clusters in the world (deployed on 250k + servers within Microsoft). Prior to joining Microsoft was a Research Scientist at Yahoo!; primarily working entity deduplication and scale and mobile+cloud platforms. Carlo spent two years as a Post Doc Associate at CSAIL MIT working with Prof. Samuel Madden and Prof. Hari Balakrishnan, working on relational databases in the cloud. Carlo received a Bachelor in Computer Science at Politecnico di Milano. He participated in a joint project between University of Illinois at Chicago (UIC) and Politecnico di Milano, obtaining a Master Degree in Computer Science at UIC and the Laurea Specialistica (cum laude) in Politecnico di Milano. During the PhD at Politecnico di Milano, Carlo spent two years as a visiting researcher at UCLA.