Reflections On My Data Management Research Journey (VLDB Women in Database Research Award Talk)

Fatma Özcan
Google LLC
Mountain View, CA
fozcan@google.com

ABSTRACT

Data-driven decision making is critical for all kinds of enterprises, public and private. It has been my mission to find more efficient, and effective ways to store, manage, query and analyze data to drive actionable insights. Throughout my career, I worked on many different technologies and systems, including semi-structured query processing, and large-scale data analytics. In this talk, I will talk about lessons learned, both technical and non-technical, using two of these systems as examples.

PVLDB Reference Format:
Fatma Özcan. Reflections On My Data Management Research Journey (VLDB Women in Database Research Award Talk). PVLDB, 15(12): 3821 - 3822, 2022.
doi:10.14778/3554821.3554903

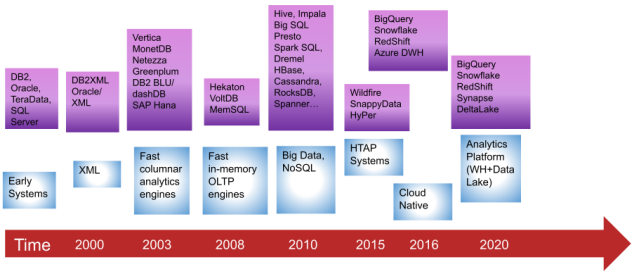

Figure 1: Evolution of Databases

1 REFLECTIONS

The impedance mismatch between the applications and the relational database has been a long standing problem, with multiple attempts from the database side at closing the gap. My first project after PhD was adding XML support to Db2. One on hand, managing nested data in a database would make databases easier to use for the applications. Semi-structured data provided schema flexibility and was a better match for applications than flat relational tables. In Db2XML[2, 3], we set out to handle schema evolution and did not enforce a strict schema. Instead, we would allow storing any shaped XML data in the same column. On the other hand, XML ecosystem, particularly XML types, as well as the XQuery language with its strict semantics on document order, were unnecessarily complex, which was in conflict with the ease of use sought by application developers. As a result, XML in the database and XQuery as a query language never took off. Reflecting back, Db2XML was one of the first noSQL databases, but with the wrong format and the wrong language.

Databases have gone through many cycles of innovation, fueled by changes in use-cases as well as hardware and platform shifts. Figure 1 shows a timeline with multiple rounds of new systems during my career.

MapReduce[4] was an inflection point in data analysis, leading to the rise of large-scale data analytics platforms like Hadoop and Spark. These systems enabled processing large amounts of data that was otherwise too messy or too big to handle in traditional warehouse and stored large amounts of structured, semi-structured and unstructured data. Although in the initial days, large scale analysis was dominated by user-defined code and new programming paradigms, like Pig, SQL processing became vital as enterprises started to adapt big data platforms like Hadoop and Spark. Recognizing the importance of large-scale SQL processing, we set out to build Big SQL[1, 5, 6]. In this new setting, open data formats, like ORC and Parquet, was gaining traction, and no one wanted to lock in their data into proprietary data formats. This triggered the design of new systems that decoupled storage and query processing. In this new world of open formats, we built query processors, not databases. While building Big SQL, we realized the need for decomposing the monolithic database architecture. The query processor did no longer owned the data, and the notion of buffer pools were obsolete. In addition, we needed a new component, a *scheduler*, to assign data blocks to workers. In traditional shared nothing databases, the allocation of data partitions to workers are fixed by where the data is stored. While running queries over the data that is stored in a distributed file system like HDFS, all nodes could access and process all storage blocks. A new scheduler bridged this gap when the query processor was decoupled from the storage layers.

Another big disruption in recent past has been the move to the cloud. Cloud warehouses freed the user from the burden of maintaining and running on-premise hardware and software, and offered elasticity with their abundance of resources. This decoupling of storage and compute enabled storage and compute resources to scale independently, and introduced new concepts to store and manage data. For example, with the data stored on object stores, the notion of fixed size disk pages are no longer applicable. While early

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.
Proceedings of the VLDB Endowment, Vol. 15, No. 12 ISSN 2150-8097.
doi:10.14778/3554821.3554903

cloud warehouses were built on the notion of clusters, the introduction of serverless computing was another disruption that enabled full advantages of available resources on the cloud. Serverless data analytics provides extreme scale out, and enables pay-as-you-go models, but it comes at the cost of data locality. The move to the cloud also facilitated the convergence of data lake and classic warehouse workloads, giving rise to the next analytics platforms. In these new platforms, enterprises execute all their analytics tasks, from data preparation, SQL analytics, to machine learning and graph analytics. While the demand for data analytics is increasing, with the end of Moore's law and slowing down of Dennard's scaling, horizontal scaling alone will not be sufficient. It is time to step back and think for new software and hardware architectures to meet the demand. Building this next generation of analytics platforms on the cloud is my current research focus.

I am humbled and grateful to receive the VLDB Women in Database Research Award. I am lucky enough to work with the brightest minds and learn from the pioneers of the field.

SHORT BIOGRAPHY

Fatma Özcan is a Principal Engineer at Google in the Systems Research Group. Before that, she was a Distinguished Research Staff Member and a senior manager at IBM Almaden Research Center. Her current research focuses on platforms and infra-structure for large-scale data analysis, knowledge graphs, democratizing analytics via NLQ and conversational interfaces to data, and query processing and optimization of semi-structured data. Dr Özcan got her PhD degree in computer science from University of Maryland, College Park, and her BSc degree in computer engineering from METU, Ankara. She has over 20 years of experience in industrial research, and has delivered core technologies into various IBM products. She has been a contributor to various SQL standards, including SQL/XML, SQL/JSON and SQL/PTF. She is the co-author of the book "Heterogeneous Agent Systems", and co-author of numerous conference papers and patents. She is an ACM Distinguished Member, and the vice chair of ACM SIGMOD. She has been serving on the board of directors of CRA (Computing Research Association) since 2020, and is a steering committee member of the CRA-Industry.

REFERENCES

[1] Daniel Abadi, Shivnath Babu, Fatma Ozcan, and Ippokratis Pandis. 2015. Tutorial: SQL-on-Hadoop Systems. *Proc. VLDB Endow.* 8, 12 (2015), 2050–2051. https://doi.org/10.14778/2824032.2824137

[2] Kevin S. Beyer, Roberta Cochrane, Vanja Josifovski, Jim Kleewein, George Lapis, Guy M. Lohman, Robert Lyle, Fatma Özcan, Hamid Pirahesh, Normen Seemann, Tuong C. Truong, Bert Van der Linden, Brian Vickery, and Chun Zhang. 2005. System RX: One Part Relational, One Part XML. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, Fatma Özcan (Ed.). ACM, 347–358. https://doi.org/10.1145/1066157.1066197

[3] Kevin S. Beyer, Fatma Özcan, Sundar Saiprasad, and Bert Van der Linden. 2005. DB2/XML: designing for evolution. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, Fatma Özcan (Ed.). ACM, 948–952. https://doi.org/10.1145/1066157.1066299

[4] Jeffrey Dean and Sanjay Ghemawat. 2010. MapReduce: a flexible data processing tool. *Commun. ACM* 53, 1 (2010), 72–77. https://doi.org/10.1145/1629175.1629198

[5] Avrilia Floratou, Umar Farooq Minhas, and Fatma Özcan. 2014. SQL-on-Hadoop: Full Circle Back to Shared-Nothing Database Architectures. *Proc. VLDB Endow.* 7, 12 (2014), 1295–1306. https://doi.org/10.14778/2732977.2733002

[6] IBM. 2022. IBM Db2 Big SQL. https://www.ibm.com/products/db2-big-sql?utm_content=SRCWW&p1=Search&p4=43700068092182375&p5=p&gclid=CjwKCAjw6fyXBhBgEiwAhhiZsj0qPdl8FFe31mo2j_Wz97I-xYE4bnWdsesMwIwqlX1p5P9noxEg0BoCLGgQAvD_BwE&gclsrc=aw.ds