Database Systems Research in the Arab World: A Tradition that Spans Decades

By Ashraf Aboulnaga, Azza Abouzied, Karima Echihabi, and Mourad Ouzzani

From Hammurabi's stone tablets to papyrus rolls and leather-bound books, the Arab region has a rich history of records keeping and transactional systems that closely matches the evolution of data storage mediums. Even modern-day data management concepts like data provenance and lineage have historic roots in the Arab world; generations of scribes meticulously tracked Islamic prophetic narrations from one narrator to the next, forming lineage chains that originated from central Arabia.

Database systems research has been part of the academic culture in the Arab world since the 1970s. High-quality computer science and database education was always available at several universities within the Arab region, such as Alexandria University in Egypt. Many students who went through these programs were drawn to database systems research and became globally prominent, such as Ramez Elmasri (professor at University of Texas, Arlington), Amr El Abbadi (professor at University of California, Santa Barbara), and Walid Aref (professor at Purdue University). Some have commented that it was easy to get into database research because it is a microcosm of computer science. The big tent of database research encompasses all: from systems to theory to languages and query optimization.

Nowadays, many prominent databases researchers are settling within the region and furthering its database research culture. For example, Ahmed Elmagarmid at the Qatar Computing Research Institute (QCRI) received the 2019 SIGMOD Contributions Award, Azza Abouzied at New York University Abu Dhabi (NYUAD) in the U.A.E. received the 2019 VLDB Test of Time Award, and of the six ACM Distinguished Members and Fellows in the region at the time of this writing, three are database systems researchers: Ashraf Aboulnaga, Ahmed Elmagarmid, and Mohamed Mokbel (all at QCRI). While the reasons for establishing their research groups within the region may vary, there are a few main influencing factors. First, there is a growing, collaborative nucleus of researchers spread throughout the region. Second, there is a thirst for establishing global research universities locally or in partnership with international universities. Third, there are funding opportunities and supportive environments for high-impact research. Finally and most importantly, there are many eager graduate students and young researchers interested in data research.

We describe the works of a few researchers to illustrate the diversity and richness of database systems research in the region. We begin our tour by exploring research on data preparation and cleaning, followed by research on graph data management and advanced analytics. We conclude with research that aims to protect individuals' privacy as our society's reliance on data-driven systems continues to grow.

The research we present addresses problems of broad interest to the global database community and is not necessarily specific to the Arab world. Nevertheless, much of the work has regional relevance. For example, data preparation and cleaning are required for local datasets, and Arab societies are paying increasing attention to issues of fairness and data privacy. Like their predecessors, today's researchers in the Arab region also are training future generations of database systems researchers, who can use their knowledge to address pressing data management problems both regionally and globally.

Data Preparation and Cleaning

Data preparation, including chiefly data discovery, integration, and cleaning, consumes a disproportionate amount of time in data science tasks. In particular, data cleaning is hard to formalize and as a result, in practice, most data cleaning is carried out in an ad-hoc and non-reproducible manner.

During the last decade, a team at QCRI has done influential work in this area. In particular, the team built NADEEF,[7] an early precursor of several contemporary data cleaning systems that uses generic data-quality rules that a given dataset must satisfy to identify errors and repair them. NADEEF was later extended[16] to handle extremely large datasets by judiciously taking the data quality rules into a series of transformations that enable distributed computations and several optimizations, such as shared scans and specialized joins operators.

Another data preparation challenge relates to entity resolution, identifying records that refer to the same real-world entity. A critical challenge is that records come in different shapes and forms, which makes matching them hard even for humans. DeepER[9] was the first to use deep learning to solve this problem by capturing the semantics of records using distributed representations of words, also known as word embeddings. DeepER converts each record into a distributed representation (that is, a vector) using uni- and bi-directional recurrent neural networks (RNNs) with long-short-term-memory (LSTM) hidden units effectively capturing similarities between records.

Finally, the team built, in collaboration with MIT, Data Civilizer,[8] an end-to-end system to support the entire life cycle of data preparation while tying it to downstream analytic applications, especially machine learning applications. The premise is that data in any organization is scattered among a multitude of databases, data lakes, and so on, and there is a need for a system to find the data of interest, integrate it, and clean it in a scalable fashion. Data Civilizer provides several modules to address these different steps.

Graph Data Management

Graph data is ubiquitous, from the Web to advertising to biology, and the development of algorithms and systems for graph management and analytics has spawned a large amount of research in the region.

Panos Kalnis and his group at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia have built several graph systems. They recently proposed to use sparse matrix algebra as a design paradigm for graph query engines. They built MAGiQ,[14] a system that represents an RDF graph as a sparse matrix and uses matrix algebra to execute queries on graphs with hundreds of billions of edges, scaling to thousands of compute nodes.

Arabesque,[18] developed by a team at QCRI, was perhaps the first system to make it possible to carry out large-scale graph data mining tasks, like frequent subgraph mining, in a principled fashion. Arabesque reconceptualized graph analysis by introducing the paradigm of “think like an embedding” instead of “think like a vertex,” which had been the standard approach for graph analysis. This new way of perceiving graph analysis made it possible to create a succinct API for many graph mining tasks and a scalable implementation of this API in a cluster setting.

LiveGraph,[19] developed by QCRI and Tsinghua University, is a system designed to simultaneously support transactions and complex analytics on graphs. The key innovation in LiveGraph is a novel data structure that stores edges contiguously and supports efficient edge scans and multi-version concurrency control. This allows high-speed, concurrent processing of transactional and analytical workloads on the primary graph store without the need for expensive extract-transform-load processes.

Mohammad Hammoud and his team at Carnegie Mellon University in Qatar (CMU Qatar) have developed an open-source, cloud-based distributed system for graph analytics called LA3.[1] Like KAUST’s MAGiQ system, LA3 uses a highly optimized linear algebra-based execution engine. It provides a familiar vertex-based programming model and was later extended with a novel architecture-aware parallelism model for high-performance computing platforms, substantially outperforming the state of the art.

Kamel Boukhalfa and his group at the University of Science and Technology Houari Boumediene in Algeria have worked on the related area of databases and cloud computing. They recently proposed an approach to optimize data placement in a hybrid storage system.[3] The approach considers several sub-costs, such as occupancy cost, durability cost, and migration cost.

Karam Gouda and his students at Benha University in Egypt have made several contributions to graph edit-based similarity search queries. Due to the hardness of computing sub-graph isomorphism and graph edit distance, these queries are often executed using a filter-and-verify approach: an efficient approximate algorithm is first used to identify a set of candidate results, then an expensive verification process on the candidates is applied to compute the final results. The group proposed an efficient verification method that can work as a stand-alone graph edit-based similarity search and outperforms the state of the art by over two orders of magnitude.[13]

Prescriptive Analytics

Azza Abouzied at NYUAD is trying to shift the data analytics paradigm from descriptive and predictive to prescriptive. Currently, database systems do not natively support the many data processing needs of data-driven decision-making, leaving experts to develop their own custom, ad-hoc application-level solutions that are difficult to scale and may produce sub-optimal results. While many systems provide support for scalable descriptive analytics (like statistics and summaries of the raw data) and even some predictive analytics (such as forecasts), there is little support for prescriptive analytics, which searches for the best course of action given the available data. As we move from “what is the data?” to “what to do with it?,” Abouzied and her colleagues at the University of Massachusetts Amherst are augmenting database systems with efficient computational problem-solving capabilities[4, 5] that take into consideration the inherent uncertainty of data and models.[6] In particular, they are integrating state-of-the-art solvers within the DBMS to scalably solve stochastic constrained optimization problems with tight approximation guarantees. Abouzied’s research has won several awards, such as the CACM 2019 Research Highlight Award, and she is applying it to current problems facing the region: her team is building a tool that advises policymakers on how to cost-effectively control epidemics like the current COVID-19 pandemic with data-driven prescriptive analytics.

Scalable, Accurate Analytics

Karima Echihabi at Mohammed VI Polytechnic University in Morocco is tackling fundamental problems to facilitate scalable and accurate analytics, focusing on supporting efficient similarity search for very large collections of high-dimensional vectors. One particular work demonstrates that it is possible to design efficient high-dimensional vector similarity search algorithms with theoretical guarantees on the quality of the answers,[11] and thus offers a more promising alternative than the current state of the art. She has conducted the two most extensive experimental evaluations in the area of similarity search for data series and generic high-dimensional vectors, offering novel insights into this challenging problem and identifying new promising research directions.[10, 11] Also, her work on progressive query answering, in collaboration with colleagues from Université de Paris and Inria, has led to techniques that support interactive exploration and fast decision-making on massive data series collections.[12]

Top-k Information Retrieval Queries

Information retrieval is an important data management area, with contributions from several research groups in the Arab region. Shady Elbassuoni at the American University of Beirut in Lebanon and his collaborators have worked on quantifying and addressing fairness in online job search platforms. They have built various frameworks to reveal and compare the fairness of different jobs at different locations for different demographic groups. With the aid of top-k style query-processing algorithms, they were also able to retrieve the jobs, locations, or groups for which a job search platform is the least or most fair.[2]

Hammoud's group at CMU Qatar with Tamer Elsayed from Qatar University also worked on top-k queries in information retrieval systems. They recently conducted the first extensive comparison between 10 effective information retrieval strategies under five representative ranking models. Based on this comparison, they proposed LazyBM,[15] a simple query evaluation strategy that consistently and robustly outperforms all considered strategies.

Data Privacy

The final area we cover is data privacy. A team of researchers at QCRI, in collaboration with Yin "David" Yang at Hamad Bin Khalifa University in Qatar, is investigating local differential privacy, which allows data to be collected from users while providing strong privacy protection, even when the data collector cannot be trusted. Some of this team's recent work focuses on protecting sensitive relationship data in decentralized social networks, in which each user only has a view of their immediate neighborhood. For example, the contact lists in a group of users' phones form a decentralized social network. The team designed a suite of techniques to collect local views from users with local differential privacy. These techniques can be used for various graph analysis tasks, including link prediction and computing subgraph statistics.[17]

Conclusion

In this article, we offered a bird's-eye view of database systems research in the Arab world, spanning two continents from the Atlantic Ocean in the west to the Gulf in the east. We highlighted hubs of excellence, committed to tackling key research challenges with regional and global impact, training and empowering the next generation of researchers, and supporting technology transfer that can propel their societies forward.

References

1. Ahmad, M.Y., Khattab, O., Malik, A., Musleh, A., Hammoud, M., Kutlu, M., Shehata, M. and Elsayed, T. LA3: A scalable link- and locality-aware linear algebra-based graph analytics system. In *Proceedings of VLDB Endow*. *11*, 8 (2018).

2. Amer-Yahia, S., Elbassuoni, S., Ghizzawi, A., Borromeo, R.M., Hoareau, E., and Mulhem, P. Fairness in online jobs: A case study on TaskRabbit and Google. In *Proc. Int. Conf. on Extending Database Technology*, 2020.

3. Boukhelef, D., Boukhobza, J., Boukhalfa, K., Ouarnoughi, H., and Lemarchand, L. Optimizing the cost of DBaaS object placement in hybrid storage systems. *Future Gen. Comput. Sys. 93,* 2019.

4. Brucato, M., Abouzied, A., and Meliou, A. Package queries: Efficient and scalable computation of high-order constraints. *The VLDB J. 27*, 5 (2018).

5. Brucato, M., Abouzied, A., and Meliou, A. Scalable computation of high-order optimization queries. *Commun. ACM 62*, 2 (2019).

6. Brucato, M.,Yadav, N., Abouzied, A., Haas, P.J., and Meliou, A. Stochastic package queries in probabilistic databases. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2020.

7. Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A.K., Ilyas, I.F., Ouzzani, M., and Tang, N. NADEEF: a commodity data cleaning system. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2013.

8. Deng, D., Fernandez, R.C., Abedjan, Z., Wang, S. Stonebraker, M., Elmagarmid, A.K., Ilyas, I.F., Madden, S., Ouzzani, M., and Tang, N. The Data Civilizer system. In *Proc. Conf. on Innovative Data Systems Research*, 2017.

9. Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., and Tang, N. Distributed representations of tuples for entity resolution. *Proc. VLDB Endow*. *11,* 11 (2018).

10. Echihabi, K., Zoumpatianos, K., Palpanas, T., and Benbrahim, H. The Lernaean hydra of data series similarity search: An experimental evaluation of the state of the art. *Proc. VLDB Endow 12,* 2 (2018).

11. Echihabi, K., Zoumpatianos, K., Palpanas, T., and Benbrahim, H. Return of the Lernaean hydra: Experimental evaluation of data series approximate similarity search. *Proc. VLDB Endow*. *13*, 3 (2019).

12. Gogolou, A., Tsandilas, T., Echihabi, K., Bezerianos, A., and Palpanas, T. Data series progressive similarity search with probabilistic quality guarantees. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2020.

13. Gouda, K., and Hassaan, M. CSI_GED: An efficient approach for graph edit similarity computation. In *Proc. IEEE Int. Conf. on Data Engineering*, 2016.

14. Jamour, F.T., Abdelaziz, I., Chen, Y., and P. Kalnis, P. Matrix algebra framework for portable, scalable and efficient query engines for RDF graphs. In *Proc. EuroSys Conf.*, 2019.

15. Khattab, O., Hammoud, M., and Elsayed, T. Finding the best of both worlds: Faster and more robust top-k document retrieval. In *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2020.

16. Khayyat, Z., Ilyas, I.F., Jindal, A., Madden, S., Ouzzani, M., Quiane-Ruiz, J.-A., Papotti, P., Tang, N., and Yin, S. BigDansing: a system for big data cleansing. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2015.

17. Sun, H., Xiao, X., Khalil, I., Yang, Y., Qin, Z., Wang, W.H., and Yu, T. Analyzing subgraph statistics from extended local views with decentralized differential privacy. In *Proc. ACM SIGSAC Conf. on Computer and Communications Security*, 2019.

18. Teixeira, C.H., Fonseca, A.J., Serafini, M., Siganos, G., Zaki, M.J., and Aboulnaga, A. Arabesque: a system for distributed graph mining. In *Proc. Symp. on Operating Systems Principles*, 2015.

19. Zhu, X., Serafini, M., Ma, X., Aboulnaga, A., Chen, W., and Feng, G. LiveGraph: A transactional graph storage system with purely sequential adjacency list scans. In *Proc. VLDB Endow*. *13*, 7(2020).

Ashraf Aboulnaga is a Senior Research Director at the Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar.

Azza Abouzied is an assistant professor of computer science at New York University, Abu Dhabi, U.A.E.

Karima Echihabi is an assistant professor of computer science at Mohammed VI Polytechnic University, Morocco.

Mourad Ouzzani is a Principal Scientist at the Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar.