The Need for Tabular Representation Learning: An Industry Perspective

Alexandra Savelieva, Andreas Mueller, Avrilia Floratou, Carlo Curino, Hiren Patel, Jordan Henkel, Joyce Cahoon, Markus Weimer, Nellie Gustafsson, Richard Wydrowski, Roman Batoukov, Shaleen Deep, Venkatesh Emani
`first_name.last_name@microsoft.com`

Abstract

The total addressable market for data applications has been estimated at $70B. This includes the $11B market for data integration, which is estimated to grow at 25% in the coming year; $35B market for analytics, growing at 11%; and $19B market for business intelligence, growing at 8% [1]. Given this data-driven future and the scale at which Microsoft operates, we survey PMs, engineers and researchers and synthesize their opinions around extracting insights from tabular data at-scale. We see three main areas where tabular representation learning (TRL) can be leveraged:

Data insights. Enabling real-time analytics is one of the key priorities for Microsoft's new intelligence platform [6] now that a converged environment exists to house any type of data. TRL models can help expose column and table-level semantic annotations, relationships between columns and between tables, and advanced data patterns such as semantic-aware denial constraints [5].

Data management. From our internal workload telemetry, we know that 17.8% of tabular data across our virtual clusters remain unaccessed [9]. From an external perspective, leveraging telemetry from Azure Observability Platform, we observed that out of the 10B+ metrics generated, less than 0.1% is used [3]. Bringing this data to light requires sophisticated data discovery, data understanding and data integration capabilities. We believe TRL models can play an important role on tasks such as entity detection and deduplication, schema mapping, and data imputation.

Data movement. It is well-known that data movement remains a key bottleneck in analytics [10]. In order to ensure that our users receive the best performance possible, investments have been made in smart caching policies, like those involving materialized views [4], as well as predicate operator pushdown [2]. Recent work [8] predicts various structural and performance properties of queries by pre-training encoder models with database workloads; but, the application of these strategies fail to consider the underlying tabular data. With TRL models, we can jointly pre-train tables with their query plans to enhance our understanding and ability to characterize workloads, and thus further efforts in reducing data movement.

Challenges and opportunities. Existing tabular models are mostly trained on Wikipedia tables and/or spreadsheets. However, in an enterprise setting, both the customer data and their associated schema is often industry specific. Access to the customer data is typically not possible due to privacy regulations [11, 7], thus training TRL models on such data is often not possible. It remains an open question, whether the existing TRL models can be successfully used on domain-specific data. Along the same lines, the existence of large language models (LLMs) and Microsoft's exclusive license to them, allows rapid prototyping of many applications even on top of tabular data. We encourage the community to provide studies that compare the performance of LLMs and TRL on some of the tasks mentioned above. Such systematic studies will be useful to application developers and product teams that are looking to incorporate more ML-based capabilities.

Table Representation Learning Workshop at NeurIPS 2022.

References

[1] Analytics and business intelligence platforms. https://www.gartner.com/reviews/market/analytics-business-intelligence-platforms. Accessed: 2022-09-15.

[2] Using S3 select pushdown with Presto to improve performance. https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-presto-s3select.html. Accessed: 2022-11-06.

[3] Souren Aghajanyan, Roman Batoukov, and Jian Zhang. Signal Fabric—An AI-assisted platform for knowledge discovery in dynamic system. Santa Clara, CA, May 2019. USENIX Association.

[4] Rana Alotaibi, Bogdan Cautis, Alin Deutsch, Moustafa Latrache, Ioana Manolescu, and Yifei Yang. Estocada: Towards scalable polystore systems. *Proceedings of the VLDB Endowment*, 13(12):2949–2952, 2020.

[5] Xu Chu, Ihab F Ilyas, and Paolo Papotti. Discovering denial constraints. *Proceedings of the VLDB Endowment*, 6(13):1498–1509, 2013.

[6] Rohan Kumar. Introducing the Microsoft Intelligent Data Platform. https://azure.microsoft.com/en-us/blog/introducing-the-microsoft-intelligent-data-platform/. Accessed: 2022-09-15.

[7] State of California Dept of Justice. California consumer privacy act. https://oag.ca.gov/privacy/ccpa. Accessed: 2022-11-10.

[8] Debjyoti Paul, Jie Cao, Feifei Li, and Vivek Srikumar. Database workload characterization with query plan encoders. *Proceedings of the VLDB Endowment*, 15(4):923–935, 2021.

[9] Conor Power, Hiren Patel, Alekh Jindal, Jyoti Leeka, Bob Jenkins, Michael Rys, Ed Triou, Dexin Zhu, Lucky Katahanas, Chakrapani Bhat Talapady, et al. The Cosmos big data platform at Microsoft: Over a decade of progress and a decade to look forward. *Proceedings of the VLDB Endowment*, 14(12):3148–3161, 2021.

[10] Karla Saur, Tara Mirmira, Konstantinos Karanasos, and Jesús Camacho-Rodríguez. Containerized execution of UDFs: An experimental evaluation. *Proceedings of the VLDB Endowment*, 15(12):3158–3171, 2022.

[11] European Union. General data protection regulation. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex. Accessed: 2022-11-10.