Adaptive Indexing for In-situ Visual Exploration and Analytics

Stavros Maroulis
Nat. Techn. Univ. of Athens &
ATHENA Research Center
Greece

Nikos Bikakis
ATHENA Research Center
Greece

George Papastefanatos
ATHENA Research Center
Greece

Panos Vassiliadis
University of Ioannina
Greece

Yannis Vassiliou
Nat. Techn. Univ. of Athens
Greece

ABSTRACT

In-situ processing has received a great deal of attention in recent years. In in-situ scenarios, large data files which do not fit in main memory, must be efficiently handled on-the-fly using commodity hardware, without the overhead of a preprocessing phase or the loading of data into a database system. In this work, we present an indexing scheme and adaptive techniques that enable efficient visual exploration and analytics over large raw data files. Beyond visual interactions and numeric statistics, here we also support categorical-based group-by and filter operations. The proposed scheme combines a *tile-based structure* that offers efficient exploratory operations over the 2D space, with a *tree-based structure* that organizes a tile's objects based on their categorical values, enabling efficient visual analytics and the support of advanced visualization methods. The index resides in main memory and is built progressively as the user explores parts of the raw file, whereas its structure and level of granularity are adjusted to the user's exploration areas and type of analysis. We conduct experiments using real and synthetic datasets, and demonstrate that the proposed approach, is in most cases more than 40× faster compared to the existing solutions, and performs around 3 orders of magnitude less I/O operations.

ACM Reference Format:
Stavros Maroulis, Nikos Bikakis, George Papastefanatos, Panos Vassiliadis, and Yannis Vassiliou. 2021. Adaptive Indexing for In-situ Visual Exploration and Analytics. In *Proceedings of 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, Nicosia, Cyprus, 2021 (DOLAP 2021),* 2 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INFO

This paper will appear in 23rd *International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data* (DOLAP 2021). [Camera Ready version is coming soon]

REFERENCES

[1] MySQL: The CSV Storage Engine. https://dev.mysql.com/doc/refman/8.0/en/csv-storage-engine.html.

[2] Oracle: External Table Enhancements in Oracle Database 12c Release 1. https://oracle-base.com/articles/12c/external-table-enhancements-12cr1.

[3] PostgreSQL: Foreign Data. https://www.postgresql.org/docs/current/ddl-foreign-data.html.

[4] SciPy: Open Source Scientific Tools for Python. http://www.scipy.org.

[5] Wolfram : Descriptive Statistics. https://reference.wolfram.com/language/tutorial/DescriptiveStatistics.html.

[6] I. Alagiannis, R. Borovica, M. Branco, S. Idreos, and A. Ailamaki. Nodb: Efficient Query Execution on Raw Data Files. In *SIGMOD*, 2012.

[7] K. Alexiou, D. Kossmann, and P. Larson. Adaptive Range Filters for Cold Data: Avoiding Trips to Siberia. *PVLDB*, 6(14), 2013.

[8] L. Battle, R. Chang, and M. Stonebraker. Dynamic Prefetching of Data Tiles for Interactive Visualization. In *SIGMOD*, 2016.

[9] N. Bikakis, J. Liagouris, M. Krommyda, G. Papastefanatos, and T. Sellis. Towards Scalable Visual Exploration of Very Large Rdf Graphs. In *ESWC*, 2015.

[10] N. Bikakis, J. Liagouris, M. Krommyda, G. Papastefanatos, and T. Sellis. Graphvizdb: A Scalable Platform for Interactive Large Graph Visualization. In *ICDE*, 2016.

[11] N. Bikakis, S. Maroulis, G. Papastefanatos, and P. Vassiliadis. RawVis: Visual Exploration over Raw Data. In *ADBIS*, 2018.

[12] N. Bikakis, S. Maroulis, G. Papastefanatos, and P. Vassiliadis. In-situ visual exploration over big raw data. *Information Systems*, 95, 2021.

[13] N. Bikakis, G. Papastefanatos, M. Skourla, and T. Sellis. A Hierarchical Aggregation Framework for Efficient Multilevel Visual Exploration and Analysis. *SWJ*, 2017.

[14] C. A. de Lara Pahins, S. A. Stephens, C. Scheidegger, and J. L. D. Comba. Hashedcubes: Simple, Low Memory, Real-time Visual Exploration of Big Data. *IEEE TVCG*, 23(1), 2017.

[15] M. El-Hindi, Z. Zhao, C. Binnig, and T. Kraska. Vistrees: Fast Indexes for Interactive Data Exploration. In *HILD*, 2016.

[16] V. Gaede and O. Günther. Multidimensional Access Methods. *ACM Comput. Surv.*, 30(2), 1998.

[17] G. Graefe and H. A. Kuno. Self-selecting, self-tuning, incrementally optimized indexes. In *EDBT*, 2010.

[18] F. Halim, S. Idreos, P. Karras, and R. H. C. Yap. Stochastic Database Cracking: Towards Robust Adaptive Indexing in Main-Memory Column-Stores. *PVLDB*, 5(6), 2012.

[19] P. Holanda, S. Manegold, H. Mühleisen, and M. Raasveldt. Progressive Indexes: Indexing for Interactive Data Analysis. *PVLDB*, 12(13), 2019.

[20] S. Idreos, I. Alagiannis, R. Johnson, and A. Ailamaki. Here Are My Data Files. Here Are My Queries. Where Are My Results? In *CIDR*, 2011.

[21] S. Idreos, M. L. Kersten, and S. Manegold. Database Cracking. In *CIDR*, 2007.

[22] S. Idreos, M. L. Kersten, and S. Manegold. Self-organizing tuple reconstruction in column-stores. In *SIGMOD*, 2009.

[23] S. Idreos, S. Manegold, H. A. Kuno, and G. Graefe. Merging What's Cracked, Cracking What's Merged: Adaptive Indexing in Main-Memory Column-Stores. *PVLDB*, 4(9), 2011.

[24] A. Kalinin, U. Çetintemel, and S. B. Zdonik. Interactive Data Exploration Using Semantic Windows. In *SIGMOD*, 2014.

[25] M. Karpathiotakis, I. Alagiannis, and A. Ailamaki. Fast Queries Over Heterogeneous Data Through Engine Customization. *PVLDB*, 9(12), 2016.

[26] M. Karpathiotakis, M. Branco, I. Alagiannis, and A. Ailamaki. Adaptive Query Processing on Raw Data. *PVLDB*, 7(12), 2014.

[27] L. D. Lins, J. T. Klosowski, and C. E. Scheidegger. Nanocubes for Real-Time Exploration of Spatiotemporal Datasets. *IEEE TVCG*, 19:2456–2465, 2013.

[28] C. Liu, C. Wu, H. Shao, and X. Yuan. Smartcube: An adaptive data management architecture for the real-time visualization of spatiotemporal datasets. *IEEE TVCG*, 26(1), 2020.

[29] F. Miranda, L. Lins, J. T. Klosowski, and C. T. Silva. TopKube: A Rank-Aware Data Cube for Real-Time Exploration of Spatiotemporal Data. *IEEE TVCG*, 2017.

[30] K. Morton, M. Balazinska, D. Grossman, and J. D. Mackinlay. Support the Data Enthusiast: Challenges for Next-generation Data-analysis Systems. *PVLDB*, 7(6), 2014.

[31] V. Nathan, J. Ding, M. Alizadeh, and T. Kraska. Learning multi-dimensional indexes. In *SIGMOD*, 2020.

[32] M. Olma, M. Karpathiotakis, I. Alagiannis, M. Athanassoulis, and A. Ailamaki. Slalom: Coasting through Raw Data Via Adaptive Partitioning and Indexing. *PVLDB*, 2017.

[33] M. Olma, M. Karpathiotakis, I. Alagiannis, M. Athanassoulis, and A. Ailamaki. Adaptive partitioning and indexing for in situ query processing. *VLDBJ*, 2019.

[34] M. Pavlovic, D. Sidlauskas, T. Heinis, and A. Ailamaki. QUASII: query-aware spatial incremental index. In *EDBT*, 2018.

[35] M. Pavlovic, E. Tzirita Zacharatou, D. Sidlauskas, T. Heinis, and A. Ailamaki. Space odyssey: efficient exploration of scientific data. In *ExploreDB*, 2016.

[36] E. Petraki, S. Idreos, and S. Manegold. Holistic Indexing in Main-memory Column-stores. In *SIGMOD*, 2015.

[37] S. Richter, J. Quiané-Ruiz, S. Schuh, and J. Dittrich. Towards zero-overhead static and adaptive indexing in Hadoop. *VLDBJ*, 23(3), 2014.

[38] Y. Tian, I. Alagiannis, E. Liarou, A. Ailamaki, P. Michiardi, and M. Vukolic. Dinodb: An Interactive-speed Query Engine for Ad-hoc Queries on Temporary Data. *IEEE Transactions on Big Data*, 2017.

[39] Z. Wang, N. Ferreira, Y. Wei, A. S. Bhaskar, and C. Scheidegger. Gaussian cubes: Real-time modeling for visual exploration of large multidimensional datasets. *IEEE TVCG*, 23(1), 2017.

[40] A. Wasay, X. Wei, N. Dayan, and S. Idreos. Data Canopy: Accelerating Exploratory Statistical Analysis. In *SIGMOD*, 2017.