BigDataX 2018: Database System Project

Aaron J. Elmore, University of Chicago

Database systems provide the ability to intelligently manage and query large amounts of structured data and are the backbone of Big Data platforms. Our research focuses on methods for large scale adaptive, elastic, and collaborative database systems. Such systems are needed to support intensive mixed workloads (e.g. high data ingest, online analytics, and transactional updates), without relying on expert administrators to tune the underlying system and to prevent the system to becoming brittle under non-ideal workload scenarios. Our current work addresses three critical emerging problems in areas of compression, dealing with incomplete data, and collaborative data analytics and versioning.

Data Compression for Distributed Databases

Analytical systems benefit from column-oriented storage, where data is organized by attributes instead of records, to enable efficient disk I/O and support directly querying over losslessly compressed data. Example of such encodings include run-length encoding, dictionary encodings, and delta encodings. However, as many systems rely on administrators to select the encoding type manually, a critical challenge being explored is how to select an ideal compression type to minimize storage and the impact on queries. Summer projects in this space will related to identifying and generating efficient encoding techniques for distributed data-intensive systems (e.g. Spark and Impala), supporting lossless compression of non-relational data types (e.g. lists, maps, time-series), and use of deep learning to support a mix of lossy and lossless compression for massive archival datasets based on a corpus of documents.

Dealing with Incomplete Data

Due to an explosion of data, many applications need to support near-real time queries over high-ingest data (e.g. sensors, financial data, high energy physics experiments). Often the complexity of the queries and volume of data to ingest may require too many resources or too long to accurately answer. This is often due to the overheads of parsing, indexing, and executing complex queries over larger than memory amounts of data. To address this issue we are building a system that allows for approximate answers (e.g. answers with confidence intervals) that can incrementally repair an answer to the accurate result or a more progressive sample without needing to recalculate the entire result. Summer projects for this system will involve building interfaces for users to explore incomplete data to specify where queries should be repaired, building models to share repair work for multiple ongoing queries, and changes to a data loading process to minimize ingested data and index non-ingested data records.

Collaborative Analytics and Versioning

Teams often perform data collection, wrangling, and analytics in a collaborative way, where several users contribute to cleaning, modeling, analyzing, and integrating new data. Current best-practices either force everyone to use the same database for the analysis, allowing only one user to make changes at a time, or to export and copy the dataset multiple times. However, a system will ideally allow users to work on these tasks in isolation and selectively share back the results. To support these goals we have actively developing systems to support lightweight dataset versioning, that is similar to software control systems like Git, but for large-scale structured datasets. Summer projects supporting this work will involve building components for a distributed versioning system, such as identify conflicts for distributed modifications, tools for visualizing stored datasets and dataset evolution, and approximate cloning strategies based on workload patterns.

Mentor

Dr. Aaron J. Elmore is an assistant professor in computer science at the University of Chicago. Aaron was previously a Postdoctoral Associate at MIT working with A.M. Turing award recipient Mike Stonebraker on elastic database systems, and Sam Madden on the DataHub project. Aaron's thesis on Elasticity Primitives for Database-as-a-Service was completed at the University of California, Santa Barbara under Divy Agrawal and Amr El Abbadi. Prior to receiving a PhD, Aaron received his MS from UChicago and spent several years in industry. Both as a post-doc and professor, Aaron has supervised many undergraduate students that has lead to publications in top database conferences, including SIGMOD and VLDB.

References

Anil Shanbhag, Alekh Jindal, Samuel Madden, Jorge-Arnulfo Quiané-Ruiz, Aaron J. Elmore:
A robust partitioning scheme for ad-hoc query workloads. SoCC 2017: 229-241

Silu Huang, Liqi Xu, Jialin Liu, Aaron J. Elmore, Aditya G. Parameswaran:
OrpheusDB: Bolt-on Versioning for Relational Databases. PVLDB 10(10): 1130-1141 (2017)

Yuanwei Fang, Chen Zou, Aaron J. Elmore, Andrew A. Chien:
UDP: a programmable accelerator for extract-transform-load workloads and more. MICRO 2017: 55-68

Michael Maddox, David Goehring, Aaron J. Elmore, Samuel Madden, Aditya G. Parameswaran, Amol Deshpande:
Decibel: The Relational Dataset Branching System. PVLDB 9(9): 624-635 (2016)

John Meehan, Stan Zdonik, Shaobo Tian, Yulong Tian, Nesime Tatbul, Adam Dziedzic, Aaron J. Elmore: Integrating real-time and batch processing in a polystore. HPEC 2016: 1-7

Anant P. Bhardwaj, Amol Deshpande, Aaron J. Elmore, David R. Karger, Sam Madden, Aditya G. Parameswaran, Harihar Subramanyam, Eugene Wu, Rebecca Zhang: Collaborative Data Analytics with DataHub. PVLDB 8(12): 1916-1919 (2015)