Approximate Query Processing for Interactive Data Science

Tim Kraska
Brown University
tim_kraska@brown.edu

ABSTRACT

Unleashing the full potential of big data requires a paradigm shift in the algorithms and tools used to analyze data towards more interactive systems with highly collaborative and visual interfaces. Ideally, a data scientist and a domain expert should be able to make discoveries together by directly manipulating, analyzing, and visualizing data on the spot, for example, using an interactive whiteboard like the recently released Microsoft Surface Hub.

While such an interactive pattern would democratize data science and make it more accessible to a wider range of users, it also requires a rethinking of the full analytical stack. Most importantly, it necessitates the next generation of approximate query processing (AQP) techniques to guarantee (visual) results at interactive speeds during the data exploration process. The first generation of AQP focused on online aggregation for simple OLAP queries; a small subset of the functionality needed for data science workflows. The second generation widened the scope to more complex workflows mainly by taking advantage of pre-computed samples at the cost of assuming that most or all queries are known upfront; again a bad fit as it is rarely the case that all exploration patterns are known in advance. The next, the third, generation of AQP has to give up on this assumption, that most queries are known upfront, but instead can leverage that data exploration pipelines are incrementally created by the user through a visual interface.

In this talk, I will present some of our recent results from building a third-generation AQP system, called IDEA. IDEA is the first Interactive Data Exploration Accelerator and allows data scientists to connect to a data source and immediately start exploring without any preparation time while still guaranteeing interactive latencies largely independently of the type of operation or data size. IDEA achieves this through novel AQP- and result reuse-techniques, which better leverage the incremental nature of the exploration process. Most importantly, IDEA automatically creates stratified samples based on the user interaction and is able to reuse approximate intermediate results between interactions.

The core idea behind our approximation and reuse technique is a reformulation of the AQP model itself based on the observation that most visualizations convey simple statistics over the data. For example, the commonly used count histograms can be seen as visualizations of the frequency statistic over the value range of an attribute. To leverage this, we propose a new AQP model that treats the aggregate query results as random variables. Surprisingly, this new model makes it not only easier to reuse results and to reason formally about the error bounds but also enables a completely new set of query rewrite rules based on probability theory.

Finally, it turns out that online aggregation, which is typically used to approximate results without a pre-computed index, stratified samples, or sketches, struggles to provide high-quality results for rare sub-populations. At the same time, as one of our user studies revealed, it is quite common for users to explore rare sub-populations, as they often contain the most interesting insights (e.g., the habits of the few highly valued customers, the suspicious outliers, etc.). We, therefore, propose a new data structure, called *tail index*, which is a low-overhead partial index that is created on the fly based on the user interactions. Together with our new AQP model, *tail indexes* enable us to provide low approximation errors, even on increasingly small sub-population, at interactive speeds without any pre-computation or an upfront known workload.

SIGMOD'17 May 14-19, 2017, Chicago, IL, USA

ACM ISBN 978-1-4503-4197-4/17/05.

DOI: http://dx.doi.org/10.1145/3035918.3056099