CS 8803-MDS

Human-in-the-loop Data Analytics

Lecture 6

09/12/22

Today's class

Overview of paper review grading rubrics

Experiences with Approximating Queries in Microsoft's Production Big-Data Clusters

Authors: Ashmita, Aniruddha

Reviewer: Myna

Researcher: Abhinav

Archaeologist: Andrew

Paper Review Feedback

- What problem is the paper trying to solve?
  - Don’t copy paste the abstract!
  - 1-2 sentences
- Why is the problem important?
  - Motivating examples
- What sets it apart from prior work?
  - Be technical!
  - How is novelty from previous work justified?
- What are the key technical ideas?
  - Straightforward
- General:
  - Minor details – offline sample creation, online sample selection
  - Length doesn’t matter – Few sentences per question

Grading Rubric

- Negative grading (from a total of 6)
- **Scope/Relevance/Importance** (-2)
  - What problem is the paper trying to solve?
  - Why is the problem important?
- **Contributions/Strengths** (-2)
  - What are the key technical ideas?
  - What sets it apart from prior work?
- **Improvements/Weaknesses** (-2)
  - What are the main areas of improvements and open questions?
    - Could talk about weaknesses/loose ends/scope for improvement

VLDB 2019 | General Information | Conference Program | Call for Contributions | Dates & Guidelines | Sponsorship | Participant Information

VLDB 2019: Overview

VLDB is a premier annual international forum for data management and database researchers, vendors, practitioners, application developers, and users. The VLDB 2019 conference will feature research talks, tutorials, demonstrations, and workshops. It will cover issues in data management, database and information systems research, since they are the technological cornerstones of the emerging applications of the 21st century.

VLDB 2019 will take place in Los Angeles, California, from August 26th to August 30th, 2019.

Tweets from @VLDB2019

VLDB 2019 @VLDB... · Sep 2, 2019

Sending a special shout-out to VLDB 2019's organizing committee and sponsors for making VLDB 2019 a grand success. And thanks to everyone for being a part of it. #VLDB2019 #VLDBSponsors

Contribution/Strengths

Production measurements

- The experiments were performed on production queries and the product was used by thousands of developers over six months - thus it inspires more belief and shows the relevance in the real world.
- They provide insights into some users' resistance to ceding control on query accuracy (even with only the tiniest bit of decrease in performance).
- Case studies substantiated in the paper suggest that batched log analysis which is common in production big data systems can highly benefit from query-time sampling.
- Additional use-cases for sampling have been reported here - construction of training data for ML models, explicit sampling and output sampling

Contribution/Strengths

Technical Ideas

- Introduction of universe and distinct samplers.
- Transformation rules which pushdown samplers have been substantially simplified which is possible because script owners specify the sampler to use.
- Previous techniques that sampled the input could not handle large complex queries and had significant computation overhead.
- What sets this paper apart is their focus which is directed towards the practicality of approximate query processing.

Limitation/Weaknesses

Presentation

- The clarity of the figures can be improved.
- The structure of the paper is very non-intuitive.
- Even the case study has been stripped away of a lot of context.

- TPC-H results could have been compared with other existing systems.
- How the distinct sampler is implemented without recording the occurrences of every distinct value?ß
- Figure 8: what is the height on the x-axis?
- In Table 3, there is no explanation given for the basis of the transformation rules.

Comparison with BlinkDB [1]

Setup: 64 queries from TPC-DS, varying budget

Storage Budget	Coverage	Median Perf. gain: All	Median Perf. gain: Covered	Median Error
Default parameters (specifically, K=M=10^5).				
0.5×	0/64	0%	–	–
1×	0/64	0%	–	–
4×	9/64	0%	27%	6%
10×	14/64	0%	24%	5%
Tuned for small group size (K=M=10^1).				
0.5×	8/64	0%	35%	6%
1×	7/64	0%	35%	6%
4×	11/64	0%	32%	6%
10×	12/64	0%	24%	6%

Table 6: BlinkDB's performance on TPC-DS.

[1] Kandula, Srikanth, et al. "Quickr: Lazily approximating complex adhoc queries in bigdata clusters." SIGMOD'16.

Limitation/Weaknesses

Barriers to adoption

- The user must "know" which parameters to pass and which sampler best fits their needs.
- The query writer has complete control over the sampling and parameters required, which means the large-scale adoption of this method is difficult.
- … does not really lower the barrier for users - it only shifts it from being able to understand statistical outputs to being able to choose suitable samplers with the correct parameters.
- .. companies and users may not have enough time to update existing queries/scripts. Also, there is a risk that the changes may break working code.
- Tolerance for error is very low in a production setting since the results of queries need to be consumed by other teams and any type of change in results needs to be well communicated and agreed upon.

Limitation/Weaknesses

Error

- There is no guarantee on error for unseen inputs.
- Losing the accuracy control to query writers will deteriorate the situation since it requires users to identify the samplers and parameters by themselves.
- While the QO performs well on Microsoft clusters, it does not significantly improve the answer quality on TPC-H.
- I am left with a general sense that the workloads are complex which makes me wonder if these proposed improvements are over fitting to Microsoft's workloads and practices or are truly general and applicable to other systems.

Improving adoption

- How can one better explain the implications of sampling to these people or people in general (perhaps through the output of a model, or explanation of a system)
- …authors in this field do not consider what people (users of interactive systems) do with the error bounds. If there is no way to handle confidence intervals in the downstream tasks then no matter how good an approximate system is, it can never be used

Future Work ideas

- Automatically choosing sampler type and probability value based on data statistics
- Is it possible to have a precomputed (sketched) version of a universe or distinct sampler?
- Building a lightweight recommendation engine to help users explore their workloads that can benefit from AQP.
- A system that can self-learn by analyzing the results of the queries (with perhaps feedback from users about the effectiveness of the approximation). Database Learning: Toward a Database that Becomes Smarter Every Time
- Data is increasingly being stored in data partitions for systems like SPARK SQL where reading a small subset of samples from the partition forces an entire disk read. Could these operators be extended to process data partitions? Approximate Partition Selection for Big-Data Workloads using Summary Statistics

Next class

Interactive Visualization

[Database Benchmarking for Supporting Real-Time Interactive Querying of Large Data]

Author: Hamsika

Reviewer: Cangdi

Archaeologist: Jingfan