Human-powered sorts and joins
by Adam Marcus Eugene, Wu David Karger and Samuel Madden Robert Miller

Summary:

Focusing on two of the most important database operators, joining and sorting, Eugene et al. extended the previously built declarative query processing system named Qurk to explore the re-usable implementation for common operators and query optimization. Built upon the platform of MTurk, system of this paper proposed three batching user interfaces, estimated several metrics to compute answers and inter-worker agreement and explored the potential different joining and sorting strategies to reduce costs. In particular, for joining task, preconditions through feature filtering are utilized to exclude unusual answers caused by irresponsible human workers lacking compensation, incentives or sufficient time and patience. When combining the responses from every assignment, this paper utilized and compared two methods, MajorityVoting and QualityAdjustment, to achieve better overall performance of the system in terms of accuracy and latency. For sorting task, they examined comparison-based (in pair/group) and rating-based (numerical scores) methods and proposed a hybrid strategy of the two, which reduced the costs of the tasks drastically.

Comments:

Pros:
The paper (and I have been fortunate to have seen Dr. Engene present this materials) is fascinating for convincing outcome it produced in terms of accuracy, latency and cost. Not only did they provide and compare several methods for implementation of the two most common operations (join and sort), they also developed pre-measures. For example, they determined appropriate parameter for reference by virtue of trial runs.

Cons:

This paper claimed sampling 10 elements is an effective way to estimate the two metrics (rate and comparison) without quantitative analysis to obtain such appropriate size of sample. In addition, they neglected to analyze how the extent of ambiguity affected the performances of rating and comparison strategies.

Although this system made several attempts to find desirable batch size, it failed to produce parameters through observation for approximation to the ideal batch size or adaptive algorithms for estimating the ideal batch size.

Furthermore, they need to extend their test set to a wider variety of data types, such as text, videos and audio formats. Specifically, the QA algorithm is constraint on categorical data, not applicable to ordinal and interval data.
Question: Why do they parameterize the algorithm for join implementation to penalize false negatives twice as heavily as false positives?
Executing SQL over encrypted data in the database-service-provider model

By Hacigümüş, Hakan, Bala Iyer, Chen Li, and Sharad Mehrotra

Summary:

This paper proposed a model named “Database as a Service” (DAS), in which the major process of database management is outsourced by enterprises to reduce costs and to concentrate on the core business. To address the issue of data privacy, especially for data stored in the distant server, this paper explored strategies to circumscribe inspection from the service provider on the condition that they might not be trustworthy. Towards this end, Hakan et al. processed the query without having to decrypt the data at the service provider's site, but rather on the client side. They used an algebraic framework to split the query to minimize the computation at the client site.

Comments:

Pros:

This paper has shed profound influence on the protective mechanism in database. The sensitive data are well protected from unauthorized access even from the Database service provider, since such data have been encrypted before being uploaded to the cloud. Besides, instead of shipping entire tables or columns out of the cloud for decryption is bandwidth intensive, the system showed the ability of operating directly on encrypted data so that a smaller amount of data needs to ultimately be shipped elsewhere to be decrypted, which could significantly improve performance.

Cons:

This paper failed to draw alternative encryption algorithms for comparison and contrast and lacked interpretation on why the equi-depth and equi-width histogram are utilized for partitioning task.

Since the paper was published as early as 2002, several other algorithms like homomorphic encryption algorithm emerged and deserves more analysis in terms of performance. Meanwhile, the coarse index is likely to aggravate the overhead of the processing. Finally, other operations like aggregation require special attention due to the intensified complexity.

Question: How does the system resolve the enhanced complexity and intensified security issue for aggregation query in the DAS model?