ABSTRACT

With the exponential growth in the amount of data generated, research on processing and analyzing large scale data is gaining importance. Significant research has been done on effectively finding aggregates and patterns from massive data sets in real time. We consider three different models and use these models for online identification of aggregates/patterns from large data. These models are infinite and sliding window model, centralized and distributed streaming model, and historical and streaming data model.

We aim to address the following problems. First we provide the first detailed experimental evaluation of streaming algorithms over sliding window for distinct counting, which is a fundamental aggregation problem widely applied in database query optimization and network monitoring. Next, we present the first communication-efficient distributed algorithm for tracking persistent items in a distributed data stream from both infinite and sliding windows. We present theoretical analysis on communication cost and accuracy, and provide experimental results to validate the guarantees. Finally, we are working towards designing and evaluating a low cost algorithm that would identify quantiles from a union of historical and streaming data. Our goal is to propose an algorithm that identifies quantiles with significantly improved accuracy when compared to the results given by the streaming quantile algorithms that have a similar memory requirement.