Incremental Detection of Inconsistencies in Distributed Data

Abstract:

This paper investigates incremental detection of errors in distributed data. Given a distributed database (D) , a set (Sigma) of conditional functional dependencies (CFDs), the set ({ {mathsf {V}}}) of violations of the CFDs in (D) , and updates (Delta {D}) to (D) , it is to find, with minimum data shipment, changes (Delta { {mathsf {V}}}) to ({ {mathsf {V}}}) in response to (Delta {D}) . The need for the study is evident since real-life data is often dirty, distributed and frequently updated. It is often prohibitively expensive to recompute the entire set of violations when (D) is updated. We show that the incremental detection problem is NP-complete for database (D) that is partitioned either vertically or horizontally, even when (Sigma) and (D) are fixed. Nevertheless, we show that it is bounded: there exist algorithms to detect errors such that their computational cost and data shipment are both linear in the size of (Delta {D}) <- inline-formula> and (Delta { {mathsf {V}}}) , independent of the size of the database (D) . We provide such incremental algorithms for vertically partitioned data and horizontally partitioned data, and show that the algorithms are optimal. We further propose optimization techniques for the incremental algorithm over vertical partitions to reduce data shipment. We verify experimentally, using real-life data on Amazon Elastic Compute Cloud (EC2), that our algorithms substantially outperform their batch counterparts.