Abstract:
Data warehouses (DWs) have traditionally been loaded with data at regular time intervals, e.g., monthly, weekly, or daily, using fast bulk loading techniques. Recently, the trend is to insert all (or only some) new source data very quickly into DWs, called near-realtime DWs (right-time DWs). This is done using regular INSERT statements, resulting in too low insert speeds. There is thus a great need for a solution that makes inserted data available quickly, while still providing bulk-load insert speeds. This paper presents RiTE (“Right-Time ETL”), a middleware system that provides exactly that. A data producer (ETL) can insert data that becomes available to data consumers on demand. RiTE includes an innovative main-memory based catalyst that provides fast storage and offers concurrency control. A number of policies controlling the bulk movement of data based on user requirements for persistency, availability, freshness, etc. are supported. The system works transparently to both producer and consumers. The system is integrated with an open source DBMS, and experiments show that it provides “the best of both worlds”, i.e., INSERT-like data availability, but with bulk-load speeds (up to 10 times faster).