Abstract

This tutorial provides a comprehensive and cohesive overview of the key research results in the area of data stream query processing, both for SQL-like and XML query languages.

1. Motivation

Measuring and monitoring complex, dynamic phenomena – traffic evolution in internet and telephone communication infrastructures, usage of the web, email and newsgroups, movement of financial markets, atmospheric conditions – produces highly detailed stream data, i.e., data that arrives as a series of “observations”, often very rapidly. With traditional data feeds, one modifies and augments underlying databases and data warehouses: complex queries over the data are performed in an offline fashion, and real time queries are typically restricted to simple filters. However, the monitoring applications that operate on modern data streams require sophisticated real time queries (often in an exploratory mode) to identify, e.g., unusual/anomalous activity (such as network intrusion detection or telecom fraud detection), based on intricate relationships between the values of the underlying data streams.

Stream data are also generated naturally by (message-based) web services, in which loosely coupled systems interact by exchanging high volumes of business data (e.g., purchase orders, retail transactions) tagged in XML (the lingua franca of web services), forming continuous XML data streams. A central aspect of web services is the ability to efficiently operate on these XML data streams executing queries (expressed in some XML query language) to continuously match, extract and transform parts of the XML data stream to drive legacy back-end business applications.

Manipulating stream data presents many technical challenges. This is an active research area in the database community, involving new stream operators, SQL extensions, query optimization methods, operator scheduling techniques, etc., with the goal of developing general-purpose (e.g., NiagaraCQ, Stanford Stream, Telegraph, Aurora, Nile) and specialized (e.g., Gigascope) data stream management systems.

The objective of this tutorial is to provide a comprehensive and cohesive overview of the key research results in the area of data stream query processing, both for SQL-like and XML query languages. It will cover specific applications that require and motivate use of this technology, issues related to stream SQL query processing, such as data models, algebra, implementation of specific operators, optimization issues as well as several query processing issues arising due to specific requirements of streaming applications. We will review state of the art in current stream data management system design and implementation. Finally, we will discuss several problems and proposed solutions in the context of stream XML query processing, and highlight the connections between stream XML query processing and stream SQL query processing. This tutorial is an enhanced version of our VLDB 2003 tutorial on the same topic.

2. Professional Biographies

**Nick Koudas** is a Principal Technical Staff Member at AT&T Labs–Research. He holds a Ph.D. from the University of Toronto, an M.Sc. from the University of Maryland at College Park, and a B.Tech. from the University of Patras in Greece. He serves as an associate editor for the Information Systems journal and the IEEE TKDE journal. His research interests include core database management, metadata management and its applications to networking.

**Divesh Srivastava** is the head of the Database Research Department at AT&T Labs–Research. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech. from the Indian Institute of Technology, Bombay, India. He was a vice-chair of ICDE 2002, and is on the editorial board of the ACM SIGMOD Digital Review. His current research interests include XML databases, IP network data management, and data quality.