TUTORIAL

Stream Processing: Issues, Techniques, and Solutions

S. Chakravarthy, Professor
300 Nedderman Hall, 416 Yates Street
Information Technology Laboratory and
Computer Science and Engineering Department
University of Texas at Arlington, Arlington, TX 76019-0019
Ph: 817 272 2082 (0ffice), 817 715 3448 (cell)
Email: sharma@cse.uta.edu
URL: http://itlab.uta.edu/sharma

Sharma Chakravarthy is Professor of Computer and Engineering Department at The University of Texas at Arlington, Texas. He established the Information Technology Laboratory at UT Arlington in Jan 2000 and currently heads it. Sharma Chakravarthy has also established the NSF funded, Distributed and Parallel Computing Cluster (DPCC@UTA) at UT Arlington in 2003. He is the recipient of the university-level "Creative Outstanding Researcher" award for 2003 and the department level senior outstanding researcher award in 2002.

He is well known for his work on semantic query optimization, multiple query optimization, active databases (HiPAC project at CCA and Sentinel project at the University of Florida, Gainesville), and more recently scalability issues in graph mining and its applications. His group at UTA is currently developing MavStream – a QoS driven stream processing system. Other systems under development include InfoSift – a classification system for text, email, and web that uses graph mining techniques. WebVigiL – a web content monitoring system has been developed and released.

His current research includes web technologies, **stream data processing**, mining and knowledge discovery – association, graph and text, active databases, distributed and heterogeneous databases, query optimization, and multi-media databases. He has published over 120 papers in refereed international journals and conference proceedings. He has given tutorial on a number of database topics, such as active, real-time, distributed, object-oriented, and heterogeneous databases in North America, Europe, and Asia. He is listed in Who's Who Among South Asian Americans and Who's Who Among America's Teachers.

Prior to joining UTA, he was with the University of Florida, Gainesville. Prior to that, he worked as a Computer Scientist at the Computer Corporation of America (CCA) and as a Member, Technical Staff at Xerox Advanced Information Technology, Cambridge, MA. Sharma Chakrvarthy received the B.E. degree in Electrical Engineering from the Indian Institute of Science, Bangalore and M.Tech from IIT Bombay, India. He worked at TIFR (Tata Institute of Fundamental Research), Bombay, India for a few years. He received M.S. and Ph.D degrees from the University of Maryland in College park in 1981 and 1985, respectively.

Stream processing: Issues, Techniques, and Solutions

1. Audience

Practitioners and professionals requiring up-to-date information on latest trends in stream processing and how to apply these techniques for various applications, such as sensor data processing, financial applications, and network management applications will benefit from this tutorial. The presenter has been working for a while on data stream management systems (DSMS in contrast to database management systems or DBMS) – on both theoretical and implementation issues. MavStream is a DSMS that has been implemented at UTA. The ubiquitous use of sensors/RFID and generation of large amounts of data in the form of streams in financial and network applications has necessitated a re-examination of assumptions used in traditional DBMSs. In this tutorial, we will present several stream processing systems (Aurora, Stream, Fjord/Telegraph, MavStream) that have been proposed in the literature. Practitioners will benefit from the practical nature of the topics and find the solutions presented applicable to problems they have encountered. Researchers will benefit from the issues that need to be addressed in one of the hot areas currently being revolutionized by increasing amount of information that needs to be processed to satisfy new characteristics.

2. Course Description

Currently, a large class of data-intensive applications, in which data is in the form of continuous streams, has been widely recognized. Not only is the size of the data for these applications unbounded, but the data arrives in a highly bursty mode. Furthermore, these applications have to respond in a timely manner. In other words, these applications have specific Quality of Service (QoS) requirements for query processing. The common QoS requirements include response time, tuple latency, accuracy of the query results, and so on. The amount of computation required and the memory used by a DSMS for processing continuous queries is also very important (for capacity planning). These new characteristics make it infeasible to simply load the arriving data into a traditional database management and use currently available techniques for their processing. Therefore, a data stream management system (DSMS) with its own set of techniques is needed for processing continuous stream data effectively and efficiently.

In this tutorial, we discuss main challenges, techniques, and solutions for building a general-purpose DSMS and present our work in this area as well as the work in the literature. We present work on Aurora, Stream, Fjord/Telegraph, and MavStream (to name a few) covering the major efforts in data stream management systems.

We will cover the following topics in detail during the tutorial: differences between traditional query processing in a DBMS and continuous query processing, operator and query modeling for stream processing, scheduling strategies (for conserving memory and reducing tuple latency), capacity estimation (to determine strategies and to determine when and how much load to shed), and load shedding strategies. The emphasis will be on satisfying QoS requirements as it is extremely important for stream processing applications. Implementation of a stream processing system will also be covered using the implementation of MavStream at UTA. Some of the work on MavStream can be found at http://itlab.uta.edu/sharma under publications (by topic).

3. Material

A set of transparencies (in PDF format) and a collection of papers will be made available to the participants.

4. Significance

Stream processing systems can be viewed as a generalization of the traditional database management systems (DBMSs) as some of the assumptions of a DBMS are relaxed for stream data processing. For example, QoS requirement that is not present (except for throughput requirements) and is not supported in a DBMS. The arrival rate and near-real time processing requirements are also different from the one assumed in a DBMS. Typically, you cannot store a stream data in a database and process it as in a DBMS.

The advent of RFID tags and the use of sensors for environmental and other monitoring (highways, movement of toxic substances) have necessitated the need for data stream processing systems. As the amount of sensor data and the need for real time (or near real time) requirements increase, new techniques have to be developed and incorporated into applications.

This tutorial brings various aspects of stream processing, applications of stream processing, and recent trends in meeting the needs of advanced applications by combining stream processing and event processing.

5. Time Required

For a full-length version, a full day is required. But short versions can be accommodated anywhere from 2 to 4 hours. The tutorial is based on the author's research, teaching, and project experience in these areas.

6. Previous Offerings

The presenter has given a number of tutorials in the past on subjects such as active and real-time databases, distributed databases, object-oriented databases, push vs. pull technologies for information management, and graph mining techniques in US, Europe, and Asia.