# Data Streams as an Element of Modern Decision Support

**Damianos Chatziantoniou**
*Athens University of Economics and Business, Greece*

**George Doukidis**
*Athens University of Economics and Business, Greece*

## INTRODUCTION

Traditional decision support systems (DSS) and executive information systems (EIS) gather and present information from several sources for business purposes. It is an information technology to help the knowledge worker (executive, manager, analyst) make faster and better decisions. So far, these data were stored statically and persistently in a database, typically in a data warehouse. *Data warehouses* collect masses of operational data, allowing analysts to extract information by issuing decision support queries on the otherwise discarded data. In a typical scenario, an organization stores a detailed record of its operations in a database, which is then analyzed to improve efficiency, detect sales opportunities, and so on. Performing complex analysis on these data is an essential component of these organizations' businesses. Chaudhuri and Dayal (1997) present an excellent survey on decision-making and online analytical processing (OLAP) technologies for traditional database systems.

In many applications however, it may not be possible to process queries within a database management system (DBMS). These applications involve data items that arrive online from multiple sources in a continuous, rapid and time-varying fashion (Babcock et. al., 2002). These data may or may not be stored in a database. As a result, a new class of data-intensive applications has recently attracted a lot of attention: applications in which the data is modeled not as persistent relations but rather as transient *data streams*. Examples include financial applications (streams of transactions or ticks), network monitoring (stream of packets), security, telecommunication data management (stream of calls or call packets), web applications (clickstreams), manufacturing, *wireless sensor networks* (measurements), RFID data, and others. In data streams we usually have "continuous" queries (Terry et. al., 1992; Babu & Widom, 2002) rather than "one-time." The answer to a *continuous query* is produced over time, reflecting the stream data seen so far. Answers may be stored and updated as new data arrives or may be produced as data streams themselves. Continuous queries can be used for monitoring, alerting, security, personalization, etc. Data streams can be either *transactional* (i.e., log interactions between entities, such as credit card purchases, web clickstreams, phone calls), or *measurement* (i.e., monitor evolution of entity states, such as physical phenomena, road traffic, temperature, network).

How to best model, express and evaluate complex queries over data streams is an open and difficult problem. This involves data modeling, rich querying capabilities to support real-time decision support and mining, and novel evaluation and optimization processing techniques. In addition, the kind of decision support over data streams is quite different from "traditional" decision-making: decisions are "tactical" rather than "strategic." Research on data streams is currently among the most active areas in database research community. Flexible and efficient stream querying will be a crucial component of any future data management and decision support system (Abiteboul et al., 2005).

## BACKGROUND

The database research community has responded with an abundance of ideas, prototypes and architectures to address the new issues involved in data stream management systems (DSMS). STREAM is Stanford University's approach for a general-purpose DSMS (Arasu et. al., 2003); *Telegraph* and *TelegraphCQ* (Madden & Franklin, 2002: Chandrasekaran et. al., 2003) are prototypes focused on handling measurements of sensor networks, developed in Berkeley; *Aurora* is a joint project between Brandeis University, Brown University and MIT (Carney et. al., 2002) targeted towards stream monitoring applications; AT&T's Hancock (Cortes et. al., 2000) and *Gigascope* (Cranor et. al., 2003) projects are special-purpose data stream systems for network management; Tribeca (Sullivan, 1996) and NiagaraCQ (Chen, et. al., 2000) are other well-known projects from Telcordia and University of Wisconsin respectively. The objective of all these projects is to develop systems that can support the challenging analysis requirements of streaming applications.

Furthermore, a plethora of articles, papers and tutorials appeared recently in the research literature. Some of the most well known survey articles follow. Faloutsos (2004) discusses indexing and mining techniques over data streams; Koudas and Srivastava (2003), present the state-of-the-art

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

algorithms on stream query processing; Muthukrishnan (2003), reviews data stream algorithms and applications; Babcock et. al. (2002), present an excellent survey on data stream prototypes and issues; Garofalakis et. al. (2002), discuss various existing models and mining techniques over data streams.

## APPLICATIONS

Stream applications span a wide range of everyday life. Real-time analytics is an essential part of these applications and becomes rapidly more and more critical in decision-making. It is apparent from the list of areas below that efficiently querying and processing data streams is a necessary element of modern decision support systems.

- **Telecommunications:** The telecommunications sector is undoubtedly one of the prime beneficiaries of such data management systems due to the huge amount of data streams that govern voice and data communications over the network infrastructure. Examples of stream analysis include fraud detection, real-time billing, dynamic pricing, network management, traffic monitoring and so on. Streams (calls, packets) have to be mined at real-time to discover outliers and patterns (fraud detection); correlated, joined and aggregated to express complex business rules (billing, dynamic pricing); and monitored─computing averages and min, max values over periods of time─to uncover unusual traffic patterns (network management).
- **Sensors:** Sensor technology becomes extremely widespread and it will probably be the next killer application: large number of cheap, wireless sensors attached to products, cars, computers, even sport players and animals, tracking and digitizing behavior, traffic, location and motion. Examples involve electronic property stickers (super markets, libraries, shopping carts, etc.), vehicle sensors (to electronically pay tolls, route traffic, set speed, etc.), and location-identification sensors (to report location, serve content, detect routes, etc.) A "sensor" world leads to a "stream" world. Millions of input data every few seconds need to be analyzed: aggregate (what is the average traffic speed), correlate (two products sell together), alert (quantity of a product is below a threshold), localize and monitor.
- **Finance:** Financial data streams come in many different forms: stock tickers, news feeds, trades, etc. Financial companies want to analyze these streams at real-time and take "tactical" business decisions (opposed to "strategic" decisions, associated to OLAP or data mining). For example, Charles Schwab wants to compute commission on each trade at real-time; Fidelity would like to route content in trades at real-time. Traderbot (www.traderbot.com) is a Web-based financial search engine that evaluates queries (both traditional and continuous) over real-time streaming data (e.g., "find all stocks between €20 and €200 where the spread between the high tick and the low tick over the past 30 minutes is greater than 3% of the last price and in the last 5 minutes the average volume has surged by more than 300%.")
- **Web management:** Large Web sites monitor Web logs (clickstreams) online to enable applications such as personalization, performance monitoring, and load balancing. Some web sites served by widely distributed web servers (e.g., Yahoo) may need to coordinate many distributed clickstream analyses, e.g. to track heavily accessed Web pages (e.g., CNN, BBC) as part of their real-time performance monitoring (Babcock et al., 2002}.
- **Network management:** Network traffic management systems monitor a variety of continuous data streams at real-time, such as packet traces, packet flows and performance measurements in order to compute statistics, detect anomalies, adjust routing, etc. The volume of data streams can be humongous and thus, query processing must be done very carefully.
- **Military:** One of the most interesting applications in military is battalion monitoring—where sensors are installed on every vehicle, human, etc.—having thousands of sensors reporting state in real-time. In these applications we want to know each time where vehicles and troops are. Examples include queries such as "tell me when three of my four tanks have crossed the front line" and "tell me when someone is pointing a gun at me."

## ISSUES AND CHALLENGES

Performing decision-making queries on top of data streams is a major challenge. For example, only one-pass algorithms are allowed (because data can be seen only once) and memory has to be managed very carefully (what to keep and what to discard). The need for a data stream management system comes in two forms: either the volume of data is huge and can not be stored in persistent relations (e.g., packet network traffic)—but still some data analysis has to be carried out, or an answer is required for a report at real-time (e.g., monitoring, alerting, fraud-detection.) As a result, DSMS are quite different in nature from traditional DBMS. Data is transient instead of persistent; queries may be "continuous" instead of "one-time"; processing techniques differ significantly, primarily due to main-memory management requirements. In Table 1 we list the differences between traditional database and data stream management systems (Babcock et al., 2002).

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/data-streams-element-modern-decision/13688

## Related Content

The Expert’s Opinion
Michael O. Leavitt (1993). *Information Resources Management Journal (pp. 45-47).*
www.irma-international.org/article/expert-opinion/50982

Fit Between Strategy and IS Specialization: A Framework for Effective Choice and Customization of Information System Application Modules
Marc N. Haines, Dale L. Goodhueand Thomas F. Gattiker (2008). *Innovative Technologies for Information Resources Management (pp. 95-109).*
www.irma-international.org/chapter/fit-between-strategy-specialization/23848

Using Diffusion of Innovations Theory to Encourage Workers to Make Healthy Food Choices and Engage in Physical Activity
Debra N. Weiss-Randall (2020). *Information Diffusion Management and Knowledge Sharing: Breakthroughs in Research and Practice  (pp. 13-26).*
www.irma-international.org/chapter/using-diffusion-of-innovations-theory-to-encourage-workers-to-make-healthy-food-choices-and-engage-in-physical-activity/242121

Assessing the Introduction of Electronic Banking in Egypt Using the Technology Acceptance Model
Sherif Kameland Ahmed Hassan (2003). *Annals of Cases on Information Technology: Volume 5  (pp. 1-25).*
www.irma-international.org/article/assessing-introduction-electronic-banking-egypt/44530

Toward a Working Definition of Digital Literacy
Margaret-Mary Sulentic Dowell (2019). *Advanced Methodologies and Technologies in Library Science, Information Management, and Scholarly Inquiry (pp. 118-129).*
www.irma-international.org/chapter/toward-a-working-definition-of-digital-literacy/215917