

Data streams have become an important class of datasets. Performance measurements in network monitoring and traffic management, call details records in telecommunications, transactions in retail chains, ATM operations in banks, log records generated by web servers, and sensor network data are some specific examples. In all of these applications, the data volume is massive, up to several terabytes. Data volume increases even further with the rapid arrival of new tuples. Traditional DBMS's are ill-equipped to process data streams in real time and to support continuous queries, which are typical for data streams. The main concern in a DBMS is to provide exact answers, while *approximation* and *adaptivity* are key features for a Data Stream Management System (DSMS).

### Summary of Past Research

My doctoral thesis focuses on models and issues towards designing an efficient DSMS called "Stardust". In this context, I have considered several problems:

- **Summarizing Data Streams:** I proposed a new wavelet-based approximation scheme called "SWAT" that maintains multiple levels of information about a dynamic stream. The scheme keeps more precise approximations for more recent data and less precise approximations for older data. This information is maintained incrementally using a unique time-varying tree structure. I have received a **Best Paper** award for this work [3].
- **Replicating a Data Stream in Large Networks:** In centralized stream processing models, a stream is summarized at a central site, and all the queries over the stream values are processed at this site. In data and query intensive environments, the central site will become a bottleneck. Therefore, I considered adaptive stream replication algorithms to disseminate stream summaries: the summaries computed at the central site are cached adaptively at the clients. The access pattern, i.e. reads and writes, over the stream results in multiple replication schemes at different resolutions. Each replication scheme expands as the corresponding read rate increases, and contracts as the corresponding write rate increases. This adaptive scheme minimizes the total communication cost, i.e., the number of inter-site messages [3].
- **Monitoring Multiple Data Streams:** Continuous queries that run indefinitely, unless a lifetime has been specified, fit naturally into the mold of data stream applications. Examples of these queries include monitoring a set of conditions or events to occur, detecting a certain trend in the underlying raw data, or in general discovering relations between various components of a large real time system. I identified three different kinds of queries that are of interest from an application point of view: (1) monitoring aggregates, (2) monitoring or finding patterns, and (3) detecting correlations. A common aspect of these queries is that each of them requires data management over some history of values, and not just over the most recently reported values. Therefore, the system has to maintain historical data along with the current data in order to be able to answer these queries. I envisioned that all these queries are interconnected in a monitoring infrastructure. For example, an unusual volatility of a stream of data may trigger an in-depth trend analysis to discover the concepts hidden in the data. In order to realize this vision, I proposed a general scheme that

accommodates all these monitoring tasks in a single body, a unified system solution called “Stardust” [4].

- **Distributed Data Stream Networks:** Data stream systems such as sensor networks, network monitoring systems, and security sensors in military applications consist of a multitude of streams at different locations. The core of such system architectures is formed by *data centers*, which coordinate their data handling for providing collaborative data mining and fusion, and for responding to various types of user queries. Typically, data centers can be dispersed over significant distances and communication between them is expensive. I exploited the scalability and load balancing of communication as well as adaptivity in presence of dynamic changes provided by content-based routing schemes, and proposed an adaptive and scalable middleware for data stream processing in a distributed environment. This work has been done in collaboration with a researcher from **IBM Research Haifa**. The solution relies on the standard distributed hash table interface, and can be used on top of virtually any existing content-based routing implementation tailored to a specific data stream environment. The proposed architecture handles the most popular types of queries in complex data stream applications, while minimizing the amount of network and computational resources consumed by data centers and network links [6, 7].
- **Network Management Architectures:** I designed an event detection framework to be used in network management for monitoring a set of network elements. This work has been done in collaboration with researchers from **AT&T Research Labs**. The set of distributed network elements are coordinated by a central monitoring station. Each network element monitors its stream of packets flowing in, and identifies a given set of individual race conditions. The goal is to detect potentially interesting instances specified by users in terms of a multitude of race conditions across a set of network elements while maintaining a low monitoring overhead, i.e., low bandwidth usage, fast processing time, and small space consumption. The approach is to formulate the problem as a multi-query optimization problem, and to provide a continuous, adaptive, and a cost-effective scheme. The solution is general in the sense that it can be used in general purpose networks such as Telecommunications networks and Internet, as well as special purpose networks such as sensor networks [5].
- **Structural Health Monitoring via Sensor Networks:** I integrated Stardust into a real-time decision support system for nondestructive health monitoring. This work has been done in collaboration with researchers from **San Diego Supercomputer Center**. The system is instrumented by an integrated network of wireless sensors mounted on civil infrastructures such as bridges, highways, and commercial and industrial facilities. To address scalability and power consumption issues related to sensor networks, a three-tier system that uses wavelets to adaptively reduce the streaming data spatially and temporally is proposed. At the sensor level, measurement data is temporally compressed before being sent upstream to intermediate communication nodes. There, correlated data from multiple sensors is combined and sent to the operation center for further reduction and interpretation. At each level, the compression ratio can be changed adaptively via wavelets [2].

This multi-resolution approach is useful in optimizing total resources in the system. At the operation center, Support Vector Machines (SVMs) are used to detect the

location of potential damage from the reduced data. The approach is validated using a finite element model of the Humboldt Bay Bridge [1].

### Future Research Directions

We are witnessing the blurring of the traditional boundaries between Networks and Databases, especially in the emerging areas of sensor and peer-to-peer networks. Data stream processing in these application domains requires networked data management, solutions of which borrow ideas from both disciplines. I believe that researchers from these two communities should share their expertise, results, terminologies, and contributions. This exchange can promote ideas that will influence and foster continued research in the areas of sensor and peer-to-peer networks. I am planning to focus mainly on the following three research avenues for the future: (1) Distributed monitoring systems (e.g., network management architectures and emergency response systems), (2) Probabilistic modeling of sensor networks, and (3) Publish-subscribe systems (e.g., content delivery/distribution networks).

- **Distributed Monitoring Systems:** The application of data mining techniques to telecommunications management is critical for building the next generation of network management systems. For this purpose, extending Stardust functionality in order to realize a distributed system composed of Stardust monitoring stations is an important step to utilize data mining solutions for network management. I provide an outline of two immediate research problems in this context below.

Each peer monitoring station characterizes its stream of data in terms of a model (signature) and transmits this information to a central site using an adaptive communication protocol. The abstraction levels of signatures collected at the server can be quite different. A higher level corresponds to coarser statistics. Therefore, it contains less representative information, and incurs smaller transmission cost. A lower level corresponds to finer statistics. Therefore, it has more characteristics information; however it incurs larger transmission cost. Naturally, there is an interplay of opposing factors, i.e., accuracy vs. overhead. At the server, we execute tasks that involve information from multiple clients. The question is to find an optimal data acquisition strategy that maximizes the conservation of limited system resources.

A more specific scenario arises in a sensor-net: anomalous event detection is a collaborative task, which involves aggregation of measurements from a number of sensors. Only if a certain number of these sensors signify an alarm and a consensus is reached, we should perform a drill-down analysis to collect more information or take affirmative steps. After computing a local fingerprint on its stream of data, each sensor needs to diffuse this fingerprint into the net in order to reach a consensus on the alarming incidents. This work will introduce “reactive monitoring” into sensor networks.

- **Probabilistic Modeling of Sensor Networks:** Embedded low-power sensing devices revolutionize the way we collect and process information for building emergency response systems. Miniature sensors are deployed to monitor ever-changing conditions in their surroundings. Statistical models such as stochastic models and multivariate regression enable capturing intra-sensor and inter-sensor dependencies in order to model sensor network data accurately. Such models can be used in backcasting missing sensor values, forecasting future data values, and guiding efficient data acquisition. Current mathematical models allow decomposing the main research problem into subproblems.

This in turn leads to a natural way for computing model components incrementally and distributively. The system will be used to (1) model the behavior of ecological systems in multiple resolutions in order to understand spatio-temporal processes, (2) capture cross modality correlations in order to enable power-aware sampling algorithms, and (3) detect episodial incidents in real-time for ecological insight.

- **Content Distribution Networks:** Publish-and-subscribe services provide the ability to create persistent queries or subscriptions to new content. In a typical content based pub-sub system, content providers send structured content to instances of *pub-sub service*, which are responsible for sending messages to the subscribers of each particular content. The pub-sub system forms a semantic layer on the top of a monitoring infrastructure by providing a query interface: events of interest are specified using an appropriate continuous query language. Furthermore, it realizes the reactive part of the whole infrastructure by sending notifications about events of interest to users. Recent advances in application layer multicast for content delivery address the scalability issues that usually arise in data stream applications with large receiver sets. However, the problem of providing real-time guarantees for time-critical tasks under stringent constraints still needs exploration.

I am fully motivated to collaborate with researchers in a leading institution, which I believe is extremely important for me to broaden my horizon, solve research problems of practical value, and contribute to the cutting-edge technology.

## References

- [1] A. Bulut, A. Singh, P. Shin, H. Jasso, T. Fountain, L. Yan, and A. Elgamal. Real-time non-destructive structural health monitoring using support vector machines and wavelets. In *Proceedings of Advanced Sensor Technologies for Nondestructive Evaluation and Structural Health Monitoring (NDE)*, 2005.
- [2] A. Bulut and A. K. Singh. Stardust: Data stream indexing for sensor networks (demo). In *ACM Conference on Embedded Networked Sensor Systems*, 2003.
- [3] A. Bulut and A. K. Singh. SWAT: Hierarchical stream summarization in large networks. In *Proceedings of 19th International Conference on Data Engineering*, pages 303–314, 2003.
- [4] A. Bulut and A. K. Singh. A unified framework for monitoring data streams in real time. In *Proceedings of 21st International Conference on Data Engineering*, 2005.
- [5] A. Bulut, A. K. Singh, N. Koudas, and D. Srivastava. Adaptive reactive network monitoring. In *submitted for publication*.
- [6] A. Bulut, R. Vitenberg, F. Emekci, and A. K. Singh. An adaptive and scalable middleware for distributed indexing of data streams. In *Proceedings of International Workshop On Databases, Information Systems and Peer-to-Peer Computing*, pages 123–137, 2003.
- [7] A. Bulut, R. Vitenberg, and A. K. Singh. Distributed data streams indexing using content-based routing paradigm. In *Proceedings of 19th International Parallel and Distributed Processing Symposium*, 2005.