

Data Stream Management

Synthesis Lectures on Data Management

Editor

M. Tamer Özsu, *University of Waterloo*

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series will publish 50- to 125 page publications on topics pertaining to data management. The scope will largely follow the purview of premier information and computer science conferences, such as ACM SIGMOD, VLDB, ICDE, PODS, ICDT, and ACM KDD. Potential topics include, but not are limited to: query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Data Stream Management

Lukasz Golab and M. Tamer Özsu

2010

Access Control in Data Management Systems

Elena Ferrari

2010

An Introduction to Duplicate Detection

Felix Naumann and Melanie Herschel

2010

Privacy-Preserving Data Publishing: An Overview

Raymond Chi-Wing Wong and Ada Wai-Chee Fu

2010

Keyword Search in Databases

Jeffrey Xu Yu, Lu Qin, and Lijun Chang

2009

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2010

Data Stream Management

Lukasz Golab and M. Tamer Özsu

ISBN: 978-3-031-00709-5 paperback

ISBN: 978-3-031-01837-4 ebook

DOI 10.1007/978-3-031-01837-4

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #5

Series Editor: M. Tamer Özsu, *University of Waterloo*

Series ISSN

Synthesis Lectures on Data Management

Print 2153-5418 Electronic 2153-5426

Data Stream Management

Lukasz Golab

AT&T Labs—Research, USA

M. Tamer Özsu

University of Waterloo, Canada

SYNTHESIS LECTURES ON DATA MANAGEMENT #5

ABSTRACT

In this lecture many applications process high volumes of streaming data, among them Internet traffic analysis, financial tickers, and transaction log mining. In general, a data stream is an unbounded data set that is produced incrementally over time, rather than being available in full before its processing begins. In this lecture, we give an overview of recent research in stream processing, ranging from answering simple queries on high-speed streams to loading real-time data feeds into a streaming warehouse for off-line analysis.

We will discuss two types of systems for end-to-end stream processing: Data Stream Management Systems (DSMSs) and Streaming Data Warehouses (SDWs). A traditional database management system typically processes a stream of ad-hoc queries over relatively static data. In contrast, a DSMS evaluates static (long-running) queries on streaming data, making a single pass over the data and using limited working memory. In the first part of this lecture, we will discuss research problems in DSMSs, such as continuous query languages, non-blocking query operators that continually react to new data, and continuous query optimization. The second part covers SDWs, which combine the real-time response of a DSMS by loading new data as soon as they arrive with a data warehouse's ability to manage Terabytes of historical data on secondary storage.

KEYWORDS

Data stream Management Systems, Stream Processing, Continuous Queries, Streaming Data Warehouses

Contents

1	Introduction	1
1.1	Overview of Data Stream Management	1
1.2	Organization	6
2	Data Stream Management Systems	9
2.1	Preliminaries	9
2.1.1	Stream Models	9
2.1.2	Stream Windows	11
2.2	Continuous Query Semantics and Operators	12
2.2.1	Semantics and Algebras	12
2.2.2	Operators	13
2.2.3	Continuous Queries as Views	17
2.2.4	Semantics of Relations in Continuous Queries	18
2.3	Continuous Query Languages	19
2.3.1	Streams, Relations and Windows	19
2.3.2	User-Defined Functions	21
2.3.3	Sampling	22
2.3.4	Summary	22
2.4	Stream Query Processing	23
2.4.1	Scheduling	23
2.4.2	Heartbeats and Punctuations	24
2.4.3	Processing Queries-As-Views and Negative Tuples	26
2.5	Stream Query Optimization	29
2.5.1	Static Analysis and Query Rewriting	29
2.5.2	Operator Optimization - Join	30
2.5.3	Operator Optimization - Aggregation	31

	2.5.4 Multi-Query Optimization	33
	2.5.5 Load Shedding and Approximation	34
	2.5.6 Load Balancing	35
	2.5.7 Adaptive Query Optimization	35
	2.5.8 Distributed Query Optimization	36
3	Streaming Data Warehouses	39
	3.1 Data Extraction, Transformation and Loading	39
	3.2 Update Propagation	40
	3.3 Data Expiration	42
	3.4 Update Scheduling	43
	3.5 Querying a Streaming Data Warehouse	45
4	Conclusions	47
	Bibliography	49
	Authors' Biographies	65