Stream Processing in the Context of CTS - course (90min) as part of a lecture on CTS

Liebig, Thomas; Ilarri, Sergio; Liang, Steve; Geers, Glenn

The recent development of innovative technologies related to mobile computingcombined with smart city infrastructures is generating massive, heterogeneousdata and creating opportunities for novel applications in transportationalcomputation science. The heterogeneous data sources provide streams ofinformation that can be used to create smart cities. The knowledge on streamanalysis is thus crucial and requires collaboration of people working inlogistics, city planning, transportation engineering and data science.

We provide a list of materials for a course on stream processing forcomputational transportation science. The objectives of the course are:

- Motivate data stream and event processing, its model and challenges.
- Acquire basic knowledge about data stream processing systems.
- Understand and analyze their application in the transportation domain.

Since the subject is large and comprises many aspects, we propose that thecourse should start with an examplary application which is familiar to theaudience. The chosen example expands through the whole course and illustrates aparticular aspect in each section.

Topics to be covered:

1. Introduction
   - Literature:
     Models and Issues in Data Stream Systems[4, 13]
     Data Stream Management: [21]
     Transportation and Data Streams: [40]
     Smart Cities and Heterogeneous Data Streams: [36]
   - Event stream examples in transportation:
     -linear ordered sequence events, e.g., bus arrival times
     -an event cloud consists of many event streams, e.g., travelerarrival time and bus arrival time at interchange
     -moving car trajectories
   - Challenges in stream processing
   - OGC standards and interfaces [15]
2. Data Stream Management Systems (DSMSs)
   - Lambda Architecture for Stream Processing [31]
   - Speed Layer: STREAM [1], Aurora [6], Borealis[39], Storm [38], streams [9], S4[32], Kafka [19]
   - Batch Layer: MapReduce [14], Hadoop [41], Spark[42], Disco [16]
   - Distributed NoSQL Databases: Cassandra [24], MongoDB[35]
3. Data Analysis
   - Query Languages: Esper [30], NiagaraCQ [10], andothers [5, 22, 25]
   - Complex Event Processing (CEP): [7, 12, 27]
   - Learning: streams [9], Mahout [33], MOA [8]
   - Distributed streams: [37]
   - Sketches: [11, 18] privacy with sketches [23]
4. Example applications in the transportation domain:[2, 3, 17, 20, 26, 28, 29, 34, 43]

Possible home assignment:

- Study a certain DSMS and summarize its features in a report.

References

1 Arvind Arasu, Brian Babcock, Shivnath Babu, Mayur Datar, Keith Ito, Rajeev Motwani, Itaru Nishizawa, Utkarsh Srivastava, Dilys Thomas, Rohit Varma, and Jennifer Widom. STREAM: The Stanford stream data manager. *IEEE Data Eng. Bull.*, 26(1):19–26, 2003.

2 Arvind Arasu, Mitch Cherniack, Eduardo Galvez, David Maier, Anurag S. Maskey, Esther Ryvkina, Michael Stonebraker, and Richard Tibbetts. Linear Road: A stream data management benchmark. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 480–491. VLDB Endowment, 2004.

3 Alexander Artikis, Matthias Weidlich, Francois Schnitzler, Ioannis Boutsis, Thomas Liebig, Nico Piatkowski, Christian Bockermann, Katharina Morik, Vana Kalogeraki, Jakub Marecek, Avigdor Gal, Shie Mannor, Dimitrios Gunopulos, and Dermot Kinane. Heterogeneous stream processing and crowdsourcing for urban traffic management. In *Proceedings of the 17th International Conference on Extending Database Technology*, page (to appear), 2014.

4 Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.

5 Shivnath Babu and Jennifer Widom. Continuous queries over data streams. *SIGMOD Rec.*, 30(3):109–120, September 2001.

6 Hari Balakrishnan, Magdalena Balazinska, Don Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Eddie Galvez, Jon Salz, Michael Stonebraker, Nesime Tatbul, Richard Tibbetts, and Stan Zdonik. Retrospective on Aurora. *The VLDB Journal*, 13(4):370–383, December 2004.

7 Tim Bass. Mythbusters: Event stream processing versus complex event processing. In *Proceedings of the 2007 Inaugural International Conference on Distributed Event-based Systems*, DEBS '07, pages 1–1, New York, NY, USA, 2007. ACM.

8 Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. MOA: Massive online analysis. *J. Mach. Learn. Res.*, 11:1601–1604, August 2010.

9 Christian Bockermann and Hendrik Blom. The streams framework. Technical report, TU Dortmund University, 2012.

10 Jianjun Chen, David J. DeWitt, Feng Tian, and Yuan Wang. NiagaraCQ: A scalable continuous query system for internet databases. *SIGMOD Rec.*, 29(2):379–390, May 2000.

11 Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Found. Trends databases*, 4:1–294, January 2012.

12 Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv.*, 44(3):15:1–15:62, June 2012.

13 Lukasz Golab and M. Tamer Özsu. Issues in data stream management. *SIGMOD Rec.*, 32, 2:5–14, June 2003.

14 Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

15 Johannes Echterhoff and Thomas Everding. OpenGIS Sensor Event Service Interface Specification. Technical Report OGC 08-133, http://portal.opengeospatial.org/files/?artifact_id=29576, October 2008.

16 Jared Flatow, Prashanth Mundkur, and Ville Tuulos. Disco: A computing platform for large-scale data analytics. In *Proceedings of the 10th ACM SIGPLAN workshop on Erlang - Erlang '11*, Tokyo, Japan, 2011. ACM Press, ACM Press.

17 Simona Florescu, Christine Körner, Michael Mock, and Michael May. Efficient mobility pattern stream matching on mobile devices. In *Proc. of the Ubiquitous Data Mining Workshop (UDM 2012)*, pages 23–27, 2012.

18 Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. Mining data streams: A review. *SIGMOD Rec.*, 34(2):18–26, June 2005.

19 N. Garg. *Apache Kafka.* Packt Publishing, 2013.

20 Sandra Geisler, Christoph Quix, Stefan Schiffer, and Matthias Jarke. An evaluation framework for traffic information systems based on data streams. *Transportation Research Part C: Emerging Technologies*, 23(0):29–55, 2012.

21 L. Golab, T. Ozsu, and T. Özsu. *Data Stream Management.* Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2010.

22 Namit Jain, Shailendra Mishra, Anand Srinivasan, Johannes Gehrke, Jennifer Widom, Hari Balakrishnan, Uğur Çetintemel, Mitch Cherniack, Richard Tibbetts, and Stan Zdonik. Towards a streaming SQL standard. *Proc. VLDB Endow.*, 1(2):1379–1390, August 2008.

23 Michael Kamp, Christine Kopp, Michael Mock, Mario Boley, and Michael May. Privacy-preserving mobility monitoring using sketches of stationary sensor readings. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip eleznÜ, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 370–386. Springer Berlin Heidelberg, 2013.

24 Avinash Lakshman and Prashant Malik. Cassandra: Structured storage system on a P2P network. In *Proceedings of the 28th ACM Symposium on Principles of Distributed Computing*, PODC '09, pages 5–5, New York, NY, USA, 2009. ACM.

25 Yan-Nei Law, Haixun Wang, and Carlo Zaniolo. Query languages and data models for database sequences and data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 492–503. VLDB Endowment, 2004.

26 Thomas Liebig, Nico Piatkowski, Christian Bockermann, and Katharina Morik. Predictive trip planning - smart routing in smart cities. In *Proceedings of the Workshop on Mining Urban Data at the International Conference on Extending Database Technology*, page (to appear), 2014.

27 Gérard Ligozat, Zygmunt Vetulani, and Jedrzej Osinski. Spatiotemporal aspects of the monitoring of complex events for public security purposes. *Spatial Cognition & Computation*, 11(1):103–128, 2011.

28 Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1010–1018, New York, NY, USA, 2011. ACM.

29 Ying Liu, Alok N. Choudhary, Jianhong Zhou, and Ashfaq A. Khokhar. A scalable distributed stream mining system for highway traffic data. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *PKDD*, volume 4213 of *Lecture Notes in Computer Science*, pages 309–321. Springer, 2006.

30 Floyd Marinescu. Esper: High volume event stream processing and correlation in Java. Online article, July 2006.

31 Nathan Marz. *Big Data: Principles and best practices of scalable realtime data systems.* O'Reilly Media, 2013.

32 Leonardo Neumeyer, Bruce Robbins, Anish Nair, and Anand Kesari. S4: Distributed stream computing platform. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 170–177, Washington, DC, USA, 2010. IEEE Computer Society.

33 Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman. *Mahout in Action.* Manning Publications Co., Manning Publications Co. 20 Baldwin Road PO Box 261 Shelter Island, NY 11964, first edition, 2011.
34 Oliver Pawlowski, JÌrgen Dunkel, Ralf Bruns, and Sascha Ossowski. Applying event stream processing on traffic problem detection. In Luis Seabra Lopes, Nuno Lau, Pedro Mariano, and Luis M. Rocha, editors, *Progress in Artificial Intelligence*, volume 5816 of *Lecture Notes in Computer Science*, pages 27–38. Springer Berlin Heidelberg, 2009.
35 Eelco Plugge, Tim Hawkins, and Peter Membrey. *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing.* Apress, Berkely, CA, USA, 1st edition, 2010.
36 IBM Research. System requirements spec, standards and guidelines for development and architecture. Technical Report FP7-318225 D2.1, INSIGHT Consortium, August 2013.
37 Izchak Sharfman, Assaf Schuster, and Daniel Keren. A geometric approach to monitoring threshold functions over distributed data streams. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 301–312, New York, NY, USA, 2006. ACM.
38 Storm. Storm. *Storm - Distributed and fault-tolerant realtime computation*, Available: http://storm-project.net/ [Last accessed: 27 June 2013], 2013.
39 Nesime Tatbul, Yanif Ahmad, Ugur Cetintemel, Jeong-Hyon Hwang, Ying Xing, and Stan Zdonik. Load management and high availability in the Borealis distributed stream processing engine. In Silvia Nittel, Alexandros Labrinidis, and Anthony Stefanidis, editors, *GeoSensor Networks*, volume 4540 of *Lecture Notes in Computer Science*, pages 66–85. Springer Berlin Heidelberg, 2008.
40 Piyushimita Vonu Thakuriah and D. Glenn Geers. *Transportation and Information: Trends in Technology and Policy.* Springer Publishing Company, Incorporated, 2013.
41 Tom White. *Hadoop: The Definitive Guide.* O'Reilly Media, Inc., 1st edition, 2009.
42 Matei Zaharia, N. M. Mosharaf Chowdhury, Michael Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. Technical Report UCB/EECS-2010-53, EECS Department, University of California, Berkeley, May 2010.
43 Jiadong Zhang, Jin Xu, and Stephen Shaoyi Liao. Aggregating and sampling methods for processing GPS data streams fortraffic state estimation. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1629–1641, 2013.