S

Stream Processing Languages and Abstractions

Martin Hirzel and Guillaume Baudart
IBM Research, Yorktown Heights, NY, USA

Synonyms

Continuous dataflow language; Streaming language; Stream processing language

Definition

A *stream processing language* is a programming language for specifying streaming applications. Here, a *stream* is an unbounded sequence of data items, and a *streaming application* is a computer program that continuously consumes input streams and produces output streams. This article surveys recent streaming languages designed around the user's mental model, the stream data model, or the execution model, as illustrated in Fig. 1. In addition to specific languages, this article also discusses *abstractions* for stream processing, which are high-level language constructs that make it easy to express common stream processing tasks.

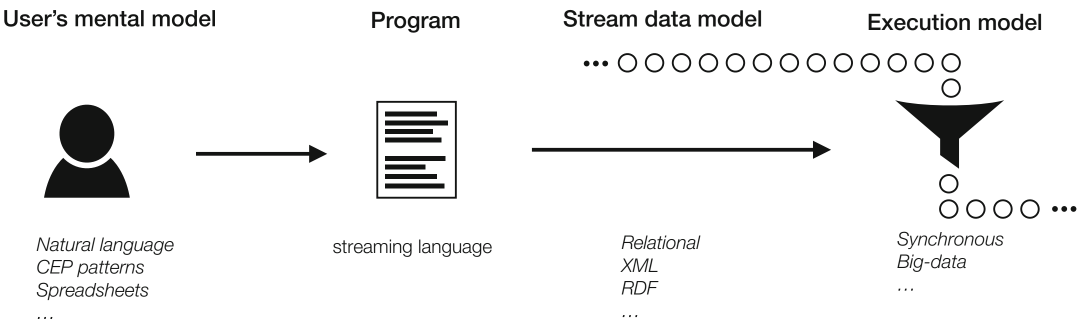
Overview

Continuous data streams arise from many directions, including sensors, communications, and commerce. Stream processing helps when low-latency responses are of the essence or when streams are too big to store for offline analysis. Programmers can of course write streaming applications in a general-purpose language without resorting to a dedicated domain-specific language (DSL) for streaming. However, using a streaming language makes code easier to read, write, understand, reason about, modularize, and optimize. Indeed, a suitable streaming language can help developers conceive of a solution to their streaming problems.

This article provides definitions, surveys concepts, and offers pointers for more in-depth study of recent streaming languages. The interested reader may also want to refer to earlier papers for historic perspective: the 1997 survey by Stephens focuses on streaming languages (Stephens 1997), the 2002 survey by Babcock et al. focuses on approximate streaming algorithms (Babcock et al. 2002), and the 2004 survey by Johnston et al. addresses dataflow languages, where streaming is a special case of dataflow (Johnston et al. 2004).

The central abstractions of stream processing are streams, operators, and stream graphs. A *stream* is an unbounded sequence of data items, for example, position readings from a delivery truck. A streaming *operator* is a stream transformer that transforms input streams to output streams. From the perspective of a streaming

S. Sakr, A. Zomaya (eds.), *Encyclopedia of Big Data Technologies*,
https://doi.org/10.1007/978-3-319-63962-8_260-1



Stream Processing Languages and Abstractions, Fig. 1 Stream processing languages

application, an operator can also have zero input streams (*source*) or zero output streams (*sink*). Finally, a *stream graph* is a directed graph whose nodes are operators and whose edges are streams. Some literature assumes that the shape of stream graphs is restricted, e.g., acyclic, but this article makes no such assumption. While only some streaming languages make the stream graph explicit, others use it as an intermediate representation. For example, the query plan generated from streaming SQL dialects is a stream graph.

The field of streaming languages is diverse and fast-moving. To understand where that diversity comes from, it is instructional to classify streaming languages by their raison d'être. Some streaming languages are based on the attitude that since streams are data in motion, data is most central, and the language should be built around a data model (relational, XML, RDF). Other streaming languages focus on the execution model for processing the dataflows efficiently, by enforcing timing constraints or exploiting distributed hardware (synchronous, big-data). A third class of streaming languages focus more on enabling the end user to develop streaming applications in high-level or familiar abstractions (complex events, spreadsheets, or even natural language). Section "Findings" surveys languages in each of these classes, and section "Examples" gives concrete examples for two languages.

Findings

This section gives an overview of the field of stream processing languages by surveying eight prominent approaches. Each approach is exemplified by one concrete language. The approaches are grouped along the lines of the previous section into approaches driven by the data model, by the execution model, or by the target user.

Data-Model Driven Streaming Languages

The success of the **relational** data model for database systems has inspired streaming dialects of the SQL database language. These dialects benefit from developers' familiarity with SQL and from its relational algebra underpinnings. A prominent example is the CQL language, which complements standard relational operators with operators to transform streams into relations and vice versa (Arasu et al. 2006). CQL lends itself to strong static typing, cf. Figure 10 of Soulé et al. (2016). Efforts toward standardizing streaming SQL focused on clarifying semantic corner cases (Jain et al. 2008).

The success of **XML** as a universal exchange format for events and messages has inspired XML-based streaming languages. These languages take advantage of a rich ecosystem of XML tools and standards and of the fact that XML documents are self-describing. The languages come in different flavors, from view maintenance over XML updates in NiagaraCQ (Chen et al. 2000) to languages that process streams where each data item is a (part of an) XML document (Diao et al. 2002; Mendell et al. 2012).

The Resource Description Framework (**RDF**) is a versatile data format for integration and reasoning, based on triples of the form ⟨subject, predicate, object⟩. A popular language for querying static RDF knowledge bases is SPARQL (Prud'hommeaux and Seaborne 2008), and C-SPARQL (Barbieri et al. 2009) extends SPARQL for continuous queries, just like CQL extends SQL. A stream is a sequence of timestamped triples, but a query can also return a graph by emitting multiple triples with the same timestamp.

Execution-Model Driven Streaming Languages

Dataflow **synchronous** languages (Benveniste et al. 2003) were introduced in the late 1980s as domain-specific languages for the design of embedded control systems. A dataflow synchronous program executes in a succession of discrete steps, and each step is assumed to be instantaneous (the synchronous hypothesis). A programmer writes high-level specifications in the form of stream functions specifying variable values at each step or instant. Section "Synchronous Dataflow in Lustre" illustrates this approach with the language Lustre (Caspi et al. 1987).

Streaming **big data** is motivated by the "4 Vs": a lot of data (volume) streams quickly (velocity) into the system, which must deal with diverse data and functionality (variety) and with uncertainty (veracity). Languages for big-data streaming let users specify an explicit stream graph that can be easily distributed with minimal synchronization and are extensible by operators in widely adopted general-purpose languages. Section "Big-Data Streaming in SPL" elaborates on this for the concrete example of SPL (Hirzel et al. 2017).

Target-User Driven Streaming Languages

Complex event processing, or **CEP**, lets users compose events hierarchically to span the gap between low-level and high-level concepts. There are various pattern languages for CEP that compile to finite automatons. Recognizing this, the MATCH-RECOGNIZE SQL extension proposal simply adopts familiar regular expressions as the CEP pattern language (Zemke et al. 2007). While the SQL basis focuses on a relational model, regular expressions can also be used for CEP in big-data streaming (Hirzel 2012).

Since there are many more **spreadsheet** users than software developers, a spreadsheet-based streaming language could reach more target users. Furthermore, spreadsheets are reactive: changes trigger updates to dependent formulas. ActiveSheets hooks up some spreadsheet cells to input or output streams, with normal spreadsheet formulas in between (Vaziri et al. 2014). When the two-dimensional spreadsheet data model is too limiting, it can be augmented with windows and partitioning (Hirzel et al. 2016).

A streaming language based on **natural language** might reach the maximum number of target users. However, since natural language is ambiguous, a controlled natural language (CNL) is a better choice (Kuhn 2014). For instance, the language for META is a CNL for specifying event-condition-action rules, temporal predicates, and data types (Arnold et al. 2016). The data model includes events and entities with nested concepts and can be shown to be equivalent to the nested relational model (Shinnar et al. 2015).

S

Examples

This section gives details and concrete code examples for two out of the eight approaches for languages surveyed in the previous section: the synchronous dataflow approach exemplified by Lustre (Caspi et al. 1987) and the big-data streaming approach exemplified by SPL (Hirzel et al. 2017).

When it comes to implementing streaming languages, there is a spectrum from basic to sophisticated techniques. At the basic end of the spectrum are configuration files in some existing markup format such as XML. The streaming engine interprets the configuration file to construct and then execute a stream graph. An intermediate

```
1  node counter (init, incr: int; reset: bool) returns (n: int);
2  var pn: int;
3  let
4    pn = init -> pre n;
5    n = if reset then init else pn + incr;
6  tel
7
8  node tracker (speed, limit: int) returns (t: int);
9  var x: bool; cpt: int when x;
10 let
11   x = (speed > limit);
12   cpt = counter((0, 1, false) when x);
13   t = current(cpt);
14 tel
```

speed	28	29	32	30	44	53	58	48	33	28	29	...
limit	30	30	30	30	55	55	55	30	30	30	30	...
x	F	F	T	F	F	F	T	T	T	F	F	...
cpt			1				2	3	4			...
t	0	0	1	1	1	1	2	3	4	4	4	...

Stream Processing Languages and Abstractions, Fig. 2 Lustre code example with a possible execution

point is a domain-specific embedded language (EDSL or sometimes DSEL) (Hudak 1998). As the name implies, an EDSL is a domain-specific language (DSL) that is embedded in some host language, typically a general-purpose language (GPL). The line between simple libraries and EDSLs is blurred, but in general, EDSLs encourage a more idiomatic programming style. Recently, EDSLs have gained popularity as several GPLs have added features that make them more suitable for hosting EDSLs. At the sophisticated end of the spectrum are full-fledged, stand-alone DSLs with their own syntax, compiler, and other tools. While stand-alone streaming DSLs are not embedded in a GPL, they often interface with a GPL, e.g., for user-defined operators.

For clarity of exposition, the following examples use stand-alone streaming languages. Stand-alone languages are the norm for synchronous dataflow, because self-contained code is easier to reason about. On the other hand, for big-data streaming, EDSLs that specify an explicit stream graph are popular, because they are easier to implement. But as the example below illustrates, once implemented, stand-alone languages also have advantages for big-data streaming.

Synchronous Dataflow in Lustre

Synchronous dataflow languages were introduced to ease the design and certification of embedded systems by providing a well-defined mathematical framework that combines a logical notion of time and deterministic concurrency. It is then possible to formally reason about the system, simulate it, prove safety properties, and generate embedded code. The synchronous dataflow language Lustre is the backbone of the industrial language and compiler Scade (Colaco et al. 2017) routinely used to program embedded controllers in many critical applications.

In Lustre a program is a set of equations defining streams of values. Time proceeds by discrete logical steps, and at each step, the program computes the value of each stream depending on its inputs and possibly previously computed values. Consider the example of Fig. 2 adapted from Bourke et al. (2017). The function *counter* takes three input streams, two integer streams *init* and *incr*, and one boolean stream *reset*. It returns the cumulative sum of the values of *incr* initialized with *init* and similarly reset when *reset* is *true* (Line 5). The variable *pn* (Line 4) stores the value of the counter *n* at the previous step using the initialization operator (−>) and the non-initialized delay **pre**.

A stream is not necessarily defined at each step. The clock of a stream is a boolean sequence giving the instants where it is defined. Streams with different clocks can be combined via sampling (**when**) or stuttering (**current**). For instance, the *tracker* function of Fig. 2 tracks the number of times the speed of a vehicle exceeds the speed limit. The **when** operator samples a stream according to a boolean condition. The function *counter* is thus only activated when *x* is *true* (Line 12). The **current** operator completes a stream with the last defined value when it is not present (Line 13). The value of *t* is thus sustained when *x* is *false*. The execution of such a program can be represented as a timeline, called a chronogram (illustrated in Fig. 2), showing the sequence of values taken by its streams at each step.

Specific compilation techniques for synchronous languages exist to generate efficient and reliable code for embedded controllers. Compilers produce imperative code that can be executed in a control loop triggered by external events or on a periodic signal (e.g., every millisecond). The link between logical and real time is left to the designer of the system.

Since the seminal dataflow languages Lustre (Caspi et al. 1987) and Signal (Le Guernic et al. 1991), multiple extensions of the dataflow synchronous model were proposed. Lucid Synchrone (Pouzet 2006) combines the dataflow synchronous approach with functional features à la ML, the n-synchronous model (Mandel et al. 2010) relaxes the synchronous hypothesis by allowing communication with bounded buffers, and Zélus (Bourke and Pouzet 2013) is a Lustre-like language extended with ordinary differential equations to define continuous-time dynamics.

Recent efforts focus on the compilation, verification, and test of dataflow synchronous programs. New techniques have been proposed to compile Lustre programs for many-core systems (Rihani et al. 2016) or improve the computation of the worst-case execution time (WCET) of the compiled code (Bonenfant et al. 2017; Forget et al. 2017). Kind2 (Champion et al. 2016) is a verification tool based on SMT solvers to model-check Lustre programs, and the Vélus compiler (Bourke et al. 2017) tackles the problem of verifying the compiler itself using a proof assistant. Lutin (Raymond and Jahier 2013) (and its industrial counterpart, the Argosim Stimulus tool Argosim 2015) is a DSL to design non-deterministic test scenarios for Lustre programs.

Big-Data Streaming in SPL

Big-data streaming languages are designed to handle high-throughput streams while at the same time being expressive enough to handle diverse data formats and streaming operators. A popular way to address the requirement of high throughput is to make it easy to execute the streaming application not just on a single core or even a single computer, but on a cluster of computers. And a popular way to address the requirement of high expressiveness is to make it easy for programmers to define new streaming operators, possibly using a different programming language than the stream processing language they use for composing operators into a graph.

SPL is a big-data streaming language designed for distribution and extensibility (Hirzel et al. 2017). It was invented in 2009 and is being actively used in industry (IBM 2008). An SPL program is an explicit stream graph of streams and operators. Unlike dataflow synchronous languages, and like other big-data streaming languages, SPL uses only minimal synchronization: an operator can fire whenever there is data available on any of its input ports, following semantics formalized in the Brooklet calculus (Soulé et al. 2010). Since synchronization across different cores and computers is hard to do efficiently, reducing synchronization simplifies distribution, giving the runtime system more flexibility for which operators to co-locate in the same core or computer. There is no assumption of simultaneity between different operator firings. When downstream operators cannot keep up with the data rate, they implicitly throttle upstream operators via back-pressure.

Figure 3 shows an example SPL program. Line 1 defines a stream *Calls* as the output of invoking an operator *CallsSource*. In SPL, streams

```
1 stream<float64 len, rstring caller> Calls = CallsSource() { }
2
3 stream<float64 len, int32 num, rstring who> CallStats = Aggregate(Calls) {
4   window Calls:     sliding, time(24.0 * 60.0 * 60.0), time(60.0);
5   output CallStats: len = Max(Calls.len),
6                     num = MaxCount(Calls.len),
7                     who = ArgMax(Calls.len, Calls.caller);
8 }
```

Stream Processing Languages and Abstractions, Fig. 3 SPL code example

carry tuples that are strongly and statically typed and whose fields can hold simple numbers or strings as in the example but can also hold nested lists and tuples. The *CallsSource* operator has no further configuration (empty curly braces); for this example, assume it is user-defined elsewhere. Programmers can define their own operators either in SPL or in other languages such as C++ or Java. Lines 3–8 define a stream *CallStats* by invoking an operator *Aggregate*. The code configures the operator with an input stream *Calls*, with a **window** clause for a 24-h sliding window with 1-minute granularity, and with an **output** clause. While many operators support these and other clauses, they can also contain code restricted to the operator at hand. The *Aggregate* operator in SPL's standard library supports intrinsic functions for *Max*, *MaxCount*, and various other aggregations. Programmers can extend SPL with new operators that, like *Aggregate*, support various configurations and operator-specific intrinsic functions.

SPL was influenced by earlier big-data streaming systems such as Borealis (Abadi et al. 2005) and TelegraphCQ (Chandrasekaran et al. 2003), generalizing them to be less dependent on relational data and more extensible. Various other streaming systems after SPL, such as Storm (Toshniwal et al. 2014) and Spark Streaming (Zaharia et al. 2013), also target big-data streaming. Like SPL, they use explicit stream graphs as their core abstraction, but unlike SPL, they use embedded (not stand-alone) domain-specific languages.

While the examples of Lustre and SPL draw a stark contrast between synchronous dataflow and big-data streaming, there are also commonalities. For instance, the StreamIt language is synchronous, but like big-data streaming languages, it uses an explicit stream graph as its core abstraction (Thies et al. 2002). And Soulé et al. show how to reduce the dependence of StreamIt on synchrony (Soulé et al. 2013).

Future Directions for Research

The landscape of streaming languages is far from consolidating on any dominant approach. New languages keep coming out to address a variety of open issues. One active area of research is the interaction between streams (data in motion) and state (data at rest). While CQL gave a conceptually clean answer (Arasu et al. 2006), people are debating alternative approaches, such as the Lambda architecture (Marz 2011) and the Kappa architecture (Kreps 2014). Another active area of research is how to handle uncertainty, such as out-of-order data, missing fields, erroneous sensor readings, approximate algorithms, or faults. On this front, streaming languages have not yet reached the clarity of databases with their ACID properties. When it comes to implementation strategies, there has been a recent surge in embedded domain-specific languages. But while EDSLs have fewer tooling needs and are less intimidating for users familiar with their host language, they are less self-contained and offer less static optimization and error-checking than stand-alone languages. We hope this article inspires innovation in streaming languages that are well-informed by those that came before.

Cross-References

- ► Continuous Queries
- ► Languages for Big Data Analysis

References

Abadi DJ, Ahmad Y, Balazinska M, Cetintemel U, Cherniack M, Hwang JH, Lindner W, Maskey AS, Rasin A, Ryvkina E, Tatbul N, Xing Y, Zdonik S (2005) The design of the Borealis stream processing engine. In: Conference on innovative data systems research (CIDR), pp 277–289

Arasu A, Babu S, Widom J (2006) The CQL continuous query language: semantic foundations and query execution. J Very Large Data Bases (VLDB J) 15(2):121–142

Argosim (2015) Stimulus. http://argosim.com/. Retrieved Nov 2017

Arnold M, Grove D, Herta B, Hind M, Hirzel M, Iyengar A, Mandel L, Saraswat V, Shinnar A, Siméon J, Takeuchi M, Tardieu O, Zhang W (2016) META: middleware for events, transactions, and analytics. IBM J Res Dev 60(2–3):15:1–15:10

Babcock B, Babu S, Datar M, Motwani R, Widom J (2002) Models and issues in data stream systems. In: Symposium on principles of database systems (PODS), pp 1–16

Barbieri DF, Braga D, Ceri S, Della Valle E, Grossniklaus M (2009) C-SPARQL: SPARQL for continuous querying. In: Poster at international World Wide Web conferences (WWW-poster), pp 1061–1062

Benveniste A, Caspi P, Edwards SA, Halbwachs N, Le Guernic P, De Simone R (2003) The synchronous languages 12 years later. Proc IEEE 91(1):64–83

Bonenfant A, Carrier F, Cassé H, Cuenot P, Claraz D, Halbwachs N, Li H, Maiza C, De Michiel M, Mussot V, Parent-Vigouroux C, Puaut I, Raymond P, Rohou E, Sotin P (2016) When the worst-case execution time estimation gains from the application semantics. In: European congress on embedded real-time software and systems (ERTS2)

Bourke T, Pouzet M (2013) Zélus: a synchronous language with odes. In: Conference on hybrid systems: computation and control (HSCC), pp 113–118

Bourke T, Brun L, Dagand P, Leroy X, Pouzet M, Rieg L (2017) A formally verified compiler for Lustre. In: Conference on programming language design and implementation (PLDI), pp 586–601

Caspi P, Pilaud D, Halbwachs N, Plaice JA (1987) LUSTRE: a declarative language for real-time programming. In: Symposium on principles of programming languages (POPL), pp 178–188

Champion A, Mebsout A, Sticksel C, Tinelli C (2016) The Kind 2 model checker. In: Conference on computer aided verification (CAV), pp 510–517

Chandrasekaran S, Cooper O, Deshpande A, Franklin MJ, Hellerstein JM, Hong W, Krishnamurthy S, Madden S, Raman V, Reiss F, Shah MA (2003) TelegraphCQ: continuous dataflow processing for an uncertain world. In: Conference on innovative data systems research (CIDR)

Chen J, DeWitt DJ, Tian F, Wang Y (2000) NiagaraCQ: a scalable continuous query system for internet databases. In: International conference on management of data (SIGMOD), pp 379–390

Colaco JL, Pagano B, Pouzet M (2017) Scade 6: a formal language for embedded critical software development. In: International symposium on theoretical aspect of software engineering (TASE), pp 1–10

Diao Y, Fischer PM, Franklin MJ, To R (2002) YFilter: efficient and scalable filtering of XML documents. In: Demo at international conference on data engineering (ICDE-Demo), pp 341–342

Forget J, Boniol F, Pagetti C (2017) Verifying end-to-end real-time constraints on multi-periodic models. In: International conference on emerging technologies and factory automation (ETFA)

Hirzel M (2012) Partition and compose: parallel complex event processing. In: Conference on distributed event-based systems (DEBS), pp 191–200

Hirzel M, Rabbah R, Suter P, Tardieu O, Vaziri M (2016) Spreadsheets for stream processing with unbounded windows and partitions. In: Conference on distributed event-based systems (DEBS), pp 49–60

Hirzel M, Schneider S, Gedik B (2017) SPL: an extensible language for distributed stream processing. Trans Program Lang Syst (TOPLAS) 39(1):5:1–5:39

Hudak P (1998) Modular domain specific languages and tools. In: International conference on software reuse (ICSR), pp 134–142

IBM (2008) IBM streams. https://ibmstreams.github.io. Retrieved Nov 2017

Jain N, Mishra S, Srinivasan A, Gehrke J, Widom J, Balakrishnan H, Cetintemel U, Cherniack M, Tibbets R, Zdonik S (2008) Towards a streaming SQL standard. In: Conference on very large data bases (VLDB), pp 1379–1390

Johnston WM, Hanna JRP, Millar RJ (2004) Advances in dataflow programming languages. ACM Comput Surv (CSUR) 36(1):1–34

Kreps J (2014) Questioning the Lambda architecture. http://radar.oreilly.com/2014/07/questioning-the-lambda-architecture.html. Retrieved Nov 2017

Kuhn T (2014) A survey and classification of controlled natural languages. Comput Linguist 40(1):121–170

Le Guernic P, Gautier T, Le Borgne M, Le Maire C (1991) Programming real-time applications with SIGNAL. Proc IEEE 79(9):1321–1336

Mandel L, Plateau F, Pouzet M (2010) Lucy-n: a n-synchronous extension of Lustre. In: International conference on mathematics of program construction (MPC), pp 288–309

Marz N (2011) How to beat the CAP theorem. http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html. Retrieved Nov 2017

Mendell M, Nasgaard H, Bouillet E, Hirzel M, Gedik B (2012) Extending a general-purpose streaming system for XML. In: Conference on extending database technology (EDBT), pp 534–539

Pouzet M (2006) Lucid synchrone, version 3, Tutorial and reference manual

Prud'hommeaux E, Seaborne A (2008) SPARQL query language for RDF. W3C recommendation. http://www.w3.org/TR/rdf-sparql-query/. Retrieved Nov 2017

Raymond P, Jahier E (2013) Lutin reference manual version Trilby-1.54

Rihani H, Moy M, Maiza C, Davis RI, Altmeyer S (2016) Response time analysis of synchronous data flow programs on a many-core processor. In: Conference on real-time networks and systems (RTNS), pp 67–76

Shinnar A, Siméon J, Hirzel M (2015) A pattern calculus for rule languages: expressiveness, compilation, and mechanization. In: European conference on object-oriented programming (ECOOP), pp 542–567

Soulé R, Hirzel M, Grimm R, Gedik B, Andrade H, Kumar V, Wu KL (2010) A universal calculus for stream processing languages. In: European symposium on programming (ESOP), pp 507–528

Soulé R, Gordon MI, Amarasinghe S, Grimm R, Hirzel M (2013) Dynamic expressivity with static optimization for streaming languages. In: Conference on distributed event-based systems (DEBS), pp 159–170

Soulé R, Hirzel M, Gedik B, Grimm R (2016) River: an intermediate language for stream processing. Softw Pract Exp (SP&E) 46(7):891–929

Stephens R (1997) A survey of stream processing. Acta Informatica 34(7):491–541

Thies W, Karczmarek M, Gordon M, Maze D, Wong J, Hoffmann H, Brown M, Amarasinghe S (2002) StreamIt: a compiler for streaming applications. Technical report LCS-TM-622, MIT

Toshniwal A, Taneja S, Shukla A, Ramasamy K, Patel JM, Kulkarni S, Jackson J, Gade K, Fu M, Donham J, Bhagat N, Mittal S, Ryaboy D (2014) Storm @Twitter. In: International conference on management of data (SIGMOD), pp 147–156

Vaziri M, Tardieu O, Rabbah R, Suter P, Hirzel M (2014) Stream processing with a spreadsheet. In: European conference on object-oriented programming (ECOOP), pp 360–384

Zaharia M, Das T, Li H, Hunter T, Shenker S, Stoica I (2013) Discretized streams: fault-tolerant streaming computation at scale. In: Symposium on operating systems principles (SOSP), pp 423–438

Zemke F, Witkowski A, Cherniak M, Colby L (2007) Pattern matching in sequences of rows. Technical report, ANSI Standard Proposal