Data-Trace Types for Distributed Stream Processing Systems

Konstantinos Mamouras  Caleb Stanford  Rajeev Alur
Zachary G. Ives  Val Tannen

Abstract

Distributed architectures for efficient processing of streaming data are increasingly critical to modern information processing systems. The goal of this paper is to develop type-based programming abstractions that facilitate correct and efficient deployment of a logical specification of the desired computation on such architectures. In the proposed model, each communication link has an associated type specifying tagged data items along with a dependency relation over tags that captures the logical partial ordering constraints over data items. The semantics of a (distributed) stream processing system is then a function from input data traces to output data traces, where a data trace is an equivalence class of sequences of data items induced by the dependency relation. This data-trace transduction model generalizes both synchronous data-flow and relational query processors, and can specify computations over data streams with a rich variety of partial ordering and synchronization characteristics. We then describe a set of programming templates for transductions: abstractions corresponding to commonly occurring stream processing tasks. Our system automatically maps these high-level programs to a given topology on the distributed implementation platform Apache Storm while preserving the semantics. Our experimental evaluation shows that (1) when the logical view of the input data stream contains a mix of ordering constraints over data items of different types, our programming abstractions allow a more natural specification of the computation compared to existing approaches, and (2) the throughput of the automatically compiled distributed code is comparable to that of hand-crafted distributed implementations.

1 Introduction

Modern information processing systems increasingly demand the ability to continuously process incoming data streams in a timely manner. Distributed stream processing architectures such as Apache Storm [4] and Twitter Heron [21] provide platforms suitable for efficient deployment of such systems. The focus of the existing systems has been mainly on providing high throughput and fault tolerance. Less developed, however, is the support for high-level programming abstractions for such systems with a clear separation between logical specification of the desired computation and distributed implementation on a given physical platform with correctness guarantees. The goal of this paper is to develop such high-level abstractions for distributed stream processing systems by relying on a type discipline that is suitable for specifying computations and that can be the basis for correct and efficient compilation of specifications on a given platform.

Physically, streams are linearly ordered, of course, and computations consume one item at a time. However, this is only one of many possible logical views of streaming data. Indeed, assuming a strict linear order over input items is not the ideal abstraction for computation specification, for two reasons. First, in an actual implementation, there may be no meaningful logical way to impose a linear ordering among items arriving at different processing nodes. Second, for many computations it suffices to view the input logically as a relation, that is, a bag of unordered data items. Such lack of ordering often has computational benefits for optimization and/or parallelization of the
implementation. Between linear ordering at one extreme and lack of ordering at the other we have
the large space of partial orders and capturing these orders is the main focus of our type discipline.

We use partially ordered multisets (pomsets), a structure studied extensively in concurrency
theory [26]. Pomsets generalize both sequences and bags, as well as sequences of bags and bags of
sequences, etc., and we have found them sufficient and appropriate for our formal development.
To specify the types that capture these partial orders as well as a logical type-consistent semantics
for stream computations, we model —inspired by the definition of Mazurkiewicz traces [25] in
concurrency theory— input and output streams as data traces. We assume that each data item
consists of a tag and a value of a basic data type associated with this tag. The ordering of items is
specified by a (symmetric) dependency relation over the set of tags. Two sequences of data items are
considered equivalent if one can be obtained from the other by repeatedly commuting two adjacent
items with independent tags, and a data trace is an equivalence class of such sequences. A data trace
type is then determined by a tag alphabet, a type of values for each tag, and a dependency relation.

For instance, when all the tags are mutually dependent, a sequence of items represents only itself,
and when all the tags are mutually independent, a sequence of items represents the bag of items
it contains. A suitable choice of tags, along with the associated dependency relation, allows us to
model streams with a rich variety of logical ordering and synchronization characteristics. As another
example, consider a system that implements key-based partitioning by mapping a linearly ordered
input sequence to a set of linearly ordered sub-streams, one per key. To model such a system the
output items corresponding to distinct keys should be unordered. For this purpose, we allow the
output items to have their own tags along with a dependency relation over these tags, and a sequence
of outputs produced by the system is interpreted as the corresponding data trace. This representation
can be easily presented programmatically and is also easily related to physical realizations.

While a system processes the input in a specific order by consuming items one by one in a
streaming manner, it is required to interpret the input sequence as a data trace, that is, outputs
produced while processing two equivalent input sequences should be equivalent. Formally, this
means that a stream processor defines a function from input data traces to output data traces. Such a
data-trace transduction is the proposed semantic model for distributed stream processing systems,
and is a generalization of existing models in literature such as acyclic Kahn process networks [20, 22]
and streaming extensions of database query languages [12, 23]. Our formal model is described in
section 3.

In section 4 we propose a programming model where the overall computation is given as an
acyclic dataflow graph, in which every communication link between two vertices is annotated with
a data-trace type that specifies the ordering characteristics of the stream flowing through the link.
In order to make the type annotation easier for the application developer, we specialize our very
general semantic framework to data-trace types that have two crucial features: (1) the traces contain
linearly ordered periodic synchronization markers for triggering the output of blocking operations
and forcing progress (similar to the punctuations of [23] or the heatbeats of [27]), and (2) the data
items of traces are viewed as key-value pairs in order to expose opportunities for key-based data
parallelism. To ensure that each individual computational element of the dataflow DAG respects
the data-trace types of its input and output channels, we provide a set of operator templates for
constraining the computation appropriately. For example, when the input data-trace type specifies
that the items are unordered, their processing is described by a commutative monoid (a structure
with an identity element and an associative, commutative binary operation), which guarantees that
the output is independent of the order in which the items are processed.

When a programmer uses these typed abstractions to describe an application, she secures
the global guarantee that the overall computation has a well-defined semantics as a data-trace
transduction, and therefore its behavior is predictable and independent of any arbitrary data item
interleavings that are imposed by the network or the distribution system. Moreover, the operator
templates allow for the introduction of data parallelism that always preserves the semantics of
the original specification. As we will explain in section 2, semantics-preserving parallelization is
desirable and not offered by many popular streaming systems.

We have implemented data-trace types, operator templates, and typed dataflow DAGs in Java as an embedded domain-specific language. Our system compiles the specification of the computation into a topology that can be executed using Apache Storm (see section 5). In section 6 we present an experimental evaluation of our system where we address the following questions: (1) Is the code that our system automatically generates as efficient in throughput as a hand-crafted implementation? (2) Does our framework facilitate the development of complex streaming applications? To answer the first question, we compare a hand-coded variant of the Yahoo Streaming Benchmark [8] to an implementation of the same computation that we built using our typed abstractions. The experimental results show near identical performance. This provides evidence that our approach does not impose a computational overhead, while offering superior guarantees of type correctness, predictable behavior, and preservation of semantics when data parallelism is introduced. To address the second question, we consider a significant case study on prediction of power usage. This case study is inspired by the DEBS’14 Grand Challenge, which we slightly adapted to incorporate a more realistic prediction technique based on machine learning. This application requires a mix of ordering constraints over data items. Our system deals automatically with low-level ordering and synchronization, so our programming effort was focused on the power prediction itself.

2 Motivation

Many popular distributed stream processing systems —such as Apache Storm [4], Twitter’s Heron [6], and Apache Samza [2]— allow the programmer to express a streaming computation as a dataflow graph, where the processing performed by each node is described in a general-purpose language such as Java or Scala. During compilation and deployment, this dataflow graph is mapped to physical nodes and processes.

As a simple example, suppose that we want to process a stream of sensor measurements and calculate every 10 seconds the average of all measurements seen so far. We assume that the sensor generates the data items in increasing timestamp order, but the sensor time series may have missing data points. The processing pipeline consists of three stages: (1) Map deserializes the incoming sensor messages and retains from each message only the scalar value and timestamp (i.e., discards any additional metadata), (2) LI performs linear interpolation to fill in the missing data points, and (3) Avg computes the average historical value and emits an update every 10 seconds.

The above pipeline can be programmed conveniently in Storm by providing the implementations of each node Map, LI, Avg and describing the connections between them.

The implementation described previously exposes pipeline parallelism, and thus suggests a multi-process or distributed execution where each stage of the pipeline computes as an independent process. In the case where the sensor produces messages at a very high rate, the computationally expensive deserialization stage Map becomes a bottleneck. In order to deal with such bottlenecks, Storm provides a facility for data parallelism by allowing the programmer to explicitly specify the creation of several parallel instances of the Map node. It handles automatically the splitting and balancing of the input stream across these instances, as well as the merging of the output streams of these instances.

The problem, however, with this data parallelization transformation is that it does not preserve the semantics of the original pipeline. The issue is that the linear interpolation stage LI relies on
receiving the data elements in increasing order of timestamps. Unfortunately, when Storm merges the output streams of the two Map instances it introduces some arbitrary interleaving that may violate this precondition. This introduces nondeterminism to the system that causes the outputs to be unpredictable and therefore not reproducible without modifications to the computation.

Typically, a practical way to deal with these problems is to generate sequence numbers and attach them to stream elements in order to recover their order later (if they get out of order). However, this increases the size of data items. Moreover, it imposes a linear order, even in cases where a partial order is sufficient. For example, synchronization markers can be used effectively to impose a partial order more efficiently than attaching sequence numbers. In general, many such practical fixes make the programs harder to debug, maintain, and modify correctly and thus less reliable.

In contrast, in order to facilitate semantically sound parallelization transformations and eliminate behaviors that rely on spurious ordering of the data items, our approach relies on data-trace types that classify the streams according to their partial ordering characteristics. For example, we can declare that the connection from Map to LI is “linearly ordered”, and this would indicate that the parallelization transformation of the previous paragraph is not sound because it causes the reordering of data items flowing through that channel. Alternatively, the implementation LI could be replaced by a new implementation, denoted Sort-LI, that can handle a disordered input by first sorting it according to timestamps. Then, the connection channel between Map and Sort-LI can be declared to be “unordered”, which enables sound data parallelization for the Map stage. Assuming that all connections are typed, the problem now arises of whether the computation nodes are consistent with these input/output partial ordering types. We propose later in section 4 a way of structuring the code for each node according to a set of templates, so that it respects the types of its input/output channels.

3 Types for Data Streams

In this section we introduce a semantic framework for distributed stream processing systems, where the input and output streams are viewed as being partially ordered. Under this view, finite prefixes of streams are represented as data traces, and they are classified according to their ordering characteristics using data-trace types. The input/output behavior of a stream processing system is modeled as a data-trace transduction, which is a monotone function from input data traces to output data traces.

3.1 Data Traces

We use data traces to model streams in which the data items are partially ordered. Data traces generalize sequences (data items are linearly ordered), relations (data items are unordered), and independent stream channels (data items are organized as a collection of linearly ordered subsets). The concatenation operation and the prefix order on sequences can be generalized naturally to the setting of data traces.

Definition 1 (Data Type). A data type \( A = (\Sigma, (T_\sigma)_{\sigma \in \Sigma}) \) consists of a potentially infinite tag alphabet \( \Sigma \) and a value type \( T_\sigma \) for every tag \( \sigma \in \Sigma \). The set of elements of type \( A \) is equal to \( \{(\sigma, d) \mid \sigma \in \Sigma \text{ and } d \in T_\sigma\} \), which we will also denote by \( A \). The set of sequences over \( A \) is denoted as \( A^* \).

Example 2. Suppose we want to process a stream that consists of sensor measurements and special symbols that indicates the end of a one-second interval. The data type for this input stream involves the tags \( \Sigma = \{M, \#\} \), where \( M \) indicates a sensor measurement and \( \# \) is an end-of-second marker. The value sets for these tags are \( T_M = \text{Nat} \) (natural numbers), and \( T_\# = \text{Ut} \) is the unit type (singleton).

So, the data type \( A = (\Sigma, T_M, T_\#) \) contains measurements \( (M, d) \), where \( d \) is a natural number, and the end-of-second symbol \( \# \).
Definition 3 (Dependence Relation and Induced Congruence). A dependence relation on a tag alphabet \( \Sigma \) is a symmetric binary relation on \( \Sigma \). We say that the tags \( \sigma, \tau \) are independent (w.r.t. a dependence relation \( D \)) if \( (\sigma, \tau) \notin D \). For a data type \( A = (\Sigma, (T_\sigma)_{\sigma \in \Sigma}) \) and a dependence relation \( D \) on \( \Sigma \), we define the dependence relation that is induced on \( A \) by \( D \) as \( \equiv_D = \{(u, v) \in A^* \times A^* | (u, v) \notin D \} \), which we will also denote by \( D \). Define \( \equiv_D \) to be the smallest congruence (w.r.t. sequence concatenation) on \( A^* \) containing \( \{(a, b) \in A^* \times A^* | (a, b) \notin D \} \).

According to the definition of the relation \( \equiv_D \) above, two sequences are equivalent if one can be obtained from the other by performing commutations of adjacent items with independent tags.

Example 4. For the data type of Example 2, the dependence relation \( D = \{(M, \#), (\#, M), (\#, \#)\} \) says that the tag \( M \) is independent of itself, and therefore consecutive \( M \)-tagged items are considered unordered. For example, the sequences

\[
(M, 5) (M, 5) (M, 8) \# (M, 9) \quad (M, 5) (M, 8) (M, 5) \# (M, 9) \quad (M, 8) (M, 5) (M, 5) \# (M, 9)
\]

are all equivalent w.r.t. congruence \( \equiv_D \).

Definition 5 (Data Traces, Concatenation, Prefix Relation). A data-trace type is a pair \( X = (A, D) \), where \( A = (\Sigma, (T_\sigma)_{\sigma \in \Sigma}) \) is a data type and \( D \) is a dependence relation on the tag alphabet \( \Sigma \). A data trace of type \( X \) is a congruence class of the relation \( \equiv_D \). We also write \( X \) to denote the set of data traces of type \( X \). Since the equivalence \( \equiv_D \) is a congruence w.r.t. sequence concatenation, the operation of concatenation is also well-defined on data traces: \( [u] \cdot [v] = [uv] \) for sequences \( u \) and \( v \), where \( [u] \) is the congruence class of \( u \). We define the relation \( \leq \) on the data traces of \( X \) as a generalization of the prefix partial order on sequences: for data traces \( u \) and \( v \) of type \( X \), \( u \leq v \) iff there are sequences \( u \in u \) and \( v \in v \) such that \( u \preceq v \) (i.e., \( u \) is a prefix of \( v \)).

The relation \( \leq \) on data traces of a fixed type is easily checked to be a partial order. Since it generalizes the prefix order on sequences (when the congruence classes of \( \equiv_D \) are singleton sets), we will call \( \leq \) the prefix order on data traces.

Example 6 (Data Traces). Consider the data-trace type \( X = (A, D) \), where \( A \) is the data type of Example 2 and \( D \) is the dependence relation of Example 4. A data trace of \( X \) can be represented as a sequence of multisets (bags) of natural numbers and visualized as a pomset. For example, the data trace that corresponds to the sequence

\[
(M, 5) (M, 7) \# (M, 9) (M, 8) (M, 9) \# (M, 6)
\]

can be visualized as the following labeled partial order

\[
\begin{array}{c}
(M, 5) \\
(M, 7) \\
(M, 8) \\
(M, 9) \\
(M, 9) \\
(M, 6)
\end{array}
\]

where a line from left to right indicates that the item on the right must occur after the item on the left. The end-of-second markers \# separate multisets of natural numbers. So, the set of data traces of \( X \) has an isomorphic representation as the set \( \text{Bag}(\text{Nat})^+ \) of nonempty sequences of multisets of natural numbers. For a set \( X \), we write \( \text{Bag}(X) \) to denote the set of all finite multisets (bags) over \( X \). In particular, the empty sequence \( \varepsilon \) is represented as \( \emptyset \) and the single-element sequence \# is represented as \( \emptyset \).

Observe that a singleton tag alphabet can be used to model sequences or multisets over a basic type of values. For the data type given by \( \Sigma = \{\sigma\} \) and \( T_\sigma = T \) there are two possible dependence relations for \( \Sigma \), namely \( \emptyset \) and \( \{(\sigma, \sigma)\} \). The data traces of \( (\Sigma, T, \emptyset) \) are multisets over \( T \), which we denote as \( \text{Bag}(T) \), and the data traces of \( (\Sigma, T, \{(\sigma, \sigma)\}) \) are sequences over \( T \).
Example 7 (Multiple Input and Output Channels). Suppose we want to model a streaming system with multiple independent input and output channels, where the items within each channel are linearly ordered but the channels are completely independent. This is the setting of (acyclic) Kahn Process Networks [22] and the more restricted synchronous dataflow models [22]. We introduce tags \( \Sigma_I = \{I_1, \ldots, I_m\} \) for \( m \) input channels, and tags \( \Sigma_O = \{0_1, \ldots, 0_n\} \) for \( n \) output channels. The dependence relation for the input consists of all pairs \((I_i, I_j)\) with \( i = 1, \ldots, m \). This means that for all indexes \( i \neq j \) the tags \( I_i \) and \( I_j \) are independent. Similarly, the dependence relation for the output consists of all pairs \((0_i, 0_j)\) with \( i = 1, \ldots, n \). Assume that the value types associated with the input tags are \( T_1, \ldots, T_m \), and the value types associated with the output tags are \( U_1, \ldots, U_n \). The sets of input and output data traces are (up to a bijection) \( T_1^* \times \cdots \times T_m^* \) and \( U_1^* \times \cdots \times U_n^* \) respectively.

3.2 Data-String Transductions

In a sequential implementation of a stream processor the input is consumed in a sequential fashion, i.e. one item at a time, and the output items are produced in a specific linear order. Such sequential semantics is formally described by data-string transductions, which we use as a precursor to defining data-trace transductions.

Definition 8 (Data-String Transductions). Let \( A \) and \( B \) be data types. A data-string transduction with input type \( A \) and output type \( B \) is a function \( f : A^* \rightarrow B^* \). Let \( \mathcal{Y}(A, B) \) be the set of all string transductions from \( A \) to \( B \).

A data-string transduction \( f : A^* \rightarrow B^* \) describes a streaming computation where the input items arrive in a linear order. For an input sequence \( u \in A^n \) the value \( f(u) \) gives the output items that are emitted right after consuming the sequence \( u \). In other words, \( f(u) \) is the output that is triggered by the arrival of the last data item of \( u \). We say that \( f \) is a one-step description of the computation because it gives the output increment that is emitted at every step.

The lifting \( \bar{f} \) of a data-string transduction \( f : A^* \rightarrow B^* \) is the function \( \bar{f} : A^* \rightarrow B^* \) that maps a sequence \( a_1a_2 \ldots a_n \in A^* \) to

\[
\bar{f}(a_1a_2 \ldots a_n) = f(\epsilon) \cdot f(a_1) \cdot f(a_1a_2) \cdots f(a_1a_2 \ldots a_n).
\]

In particular, the definition implies that \( \bar{f}(\epsilon) = f(\epsilon) \). That is, \( \bar{f} \) accumulates the outputs of \( f \) for all prefixes of the input. Notice that \( \bar{f} \) is monotone w.r.t. the prefix order: \( u \leq v \) implies that \( \bar{f}(u) \leq \bar{f}(v) \) for all sequences \( u, v \in A^* \).

For strings \( x, y \in A^* \) with \( x \leq y \), we write \( x^{-1}y \) for the unique string \( z \in A^* \) such that \( y = xz \). Suppose now that \( \varphi : A^* \rightarrow B^* \) is a monotone function. The derivative \( \partial \varphi : A^* \rightarrow B^* \) of \( \varphi \) is defined as follows: \( (\partial \varphi)(\epsilon) = \varphi(\epsilon) \) and

\[
(\partial \varphi)(ua) = \varphi(u)^{-1}\varphi(ua) \text{ for all } u \in A^* \text{ and } a \in A.
\]

The lifting and derivative operators witness a bijection between the class of functions from \( A^* \) to \( B^* \) and the monotone subset of this class. That is, \( \partial \bar{f} = f \) for every \( f : A^* \rightarrow B^* \) and \( \bar{\partial} \varphi = \varphi \) for every monotone \( \varphi : A^* \rightarrow B^* \).

Example 9. Suppose the input is a sequence of natural numbers, and we want to define the transformation that outputs the current data item if it is strictly larger than all data items seen so far. This computation is modeled by the string transduction \( f : \text{Nat}^* \rightarrow \text{Nat}^* \), given by \( f(\epsilon) = \epsilon \) and

\[
f(a_1 \ldots a_{n-1} a_n) = \begin{cases} a_n & \text{if } a_n > a_i \text{ for all } i = 1, \ldots, n-1; \\ \epsilon & \text{otherwise.}\end{cases}
\]

The table below gives the values of \( f \) and its lifting \( \bar{f} \) on input prefixes:
### 3.3 Data-Trace Transductions

Data-trace transductions are useful for giving the meaning (semantics) of a stream processing system. Consider the analogy with a functional model of computation: the meaning of a program consists of the input type, the output type, and a mapping that describes the input/output behavior of the program. Correspondingly, the semantics for a stream processing systems consists of: (1) the type $X$ of input data traces, (2) the type $Y$ of output data traces, and (3) a monotone mapping $\beta : X \rightarrow Y$ that specifies the cumulative output after having consumed a prefix of the input stream. The monotonicity requirement captures the idea that output items cannot be retracted after they have been omitted. Since a transduction is a function from trace histories, it can model stateful systems, where the output that is emitted at every step depends potentially on the entire input history.

We have already discussed how (monotone) functions from $A^\ast$ to $B^\ast$ model sequential stream processors. We will introduce now the formal notion of consistency, which captures the intuition that a sequential implementation does not depend on the relative order of any two elements unless the stream type considers them to be relatively ordered.

**Definition 10 (Consistency)**. Let $X = (A, D)$ and $Y = (B, E)$ be data-trace types. We say that a string transduction $f \in S(A, B)$ is $(X, Y)$-consistent if $u \equiv_d v$ implies that $\bar{f}(u) \equiv_e \bar{f}(v)$ for all sequences $u, v \in A^\ast$.

Let $f \in S(A, B)$ be a $(X, Y)$-consistent data-string transduction. The function $\beta : X \rightarrow Y$, defined by $\beta([u]) = [\bar{f}(u)]$ for all $u \in A^\ast$, is called the $(X, Y)$-denotation of $f$.

**Definition 11 (Data-Trace Transductions)**. Let $X = (A, D)$ and $Y = (B, E)$ be data-trace types. A data-trace transduction with input type $X$ and output type $Y$ is a function $\beta : X \rightarrow Y$ that is monotone w.r.t. the prefix order on data traces: $u \leq v$ implies that $\beta(u) \leq \beta(v)$ for all traces $u, v \in X$. We write $\mathcal{S}(X, Y)$ for the set of data-trace transductions from $X$ to $Y$.

Definition 10 essentially says that a string transduction $f$ is consistent when it gives equivalent cumulative outputs for equivalent input sequences. It is shown in [16] that the set $\mathcal{S}(X, Y)$ of trace transductions is equal to the set of $(X, Y)$-denotations of all $(X, Y)$-consistent string transductions.

### 3.4 Examples of Data-Trace Transductions

In this section we present examples that illustrate the concept of a data-trace transduction for several streaming computations on streams of partially ordered elements. We start by considering examples that fit into the model of process networks [20, 22], where the inputs and outputs are organized in collections of independent linearly ordered channels.

**Example 12 (Deterministic Merge)**. Consider the streaming computation where two linearly ordered input channels are merged into one. More specifically, this transformation reads cyclically items from the two input channels and passes them unchanged to the output channel. Recall from Example 7.
that an input data trace is essentially an element of $T^* \times T^*$, and the set of output data traces is essentially $T^*$. The trace transduction $\text{merge} : T^* \times T^* \rightarrow T^*$ is given by:

$$\text{merge}(x_1 \ldots x_m, y_1 \ldots y_n) = \begin{cases} x_1 y_1 \ldots x_m y_m, & \text{if } m \leq n; \\ x_1 y_1 \ldots x_n y_n, & \text{if } m > n. \end{cases}$$

**Example 13** (Copy). The copy transformation creates two copies of the input stream by reading each item from the input channel and copying it to two output channels. An input data trace is an element of $T^*$, and an output data trace is an element of $T^* \times T^*$. The trace transduction for this computation is given by $\text{copy}(u) = (u, u)$ for all $u \in T^*$.

One possible sequential implementation of copy is modeled by the data-string transduction $f : T^* \rightarrow B^*$, where $B = ((\sigma) \times T) \cup (\tau) \times T)$. We put $f(\varepsilon) = \varepsilon$ and $f(ua) = (\sigma, a) (\tau, a)$ for all $u \in A^*$. Notice that the implementation makes the arbitrary choice to emit $(\sigma, a)$ before $(\tau, a)$, but it is also possible to emit the items in reverse order.

**Example 14** (Key-based Partitioning). Consider the computation that maps a linearly ordered input sequence of data items of type $T$, each of which contains a key, to a set of linearly ordered sub-streams, one per key. The function $\text{key} : T \rightarrow K$ extracts the key from each input value. An input trace is represented as a $K$-indexed tuple, that is, a function $K \rightarrow T^*$. The trace transduction $\text{partition}_{\text{key}} : T^* \rightarrow (K \rightarrow T^*)$ describes the partitioning of the input stream into sub-streams according to the key extraction map $\text{key}$:

$$\text{partition}_{\text{key}}(u)(k) = u|_k \text{ for all } u \in T^* \text{ and } k \in K,$$

where $u|_k$ denotes the subsequence of $u$ that consists of all items whose key is equal to $k$. The implementation of this transduction can be modeled as a data-string transduction $f : T^* \rightarrow (K \times T)^*$ given by $f(\varepsilon) = \varepsilon$ and $f(wx) = (\text{key}(x), x)$ for all $w \in T^*$ and $x \in T$.

The computation of aggregates (e.g., sum, max, and min) is meaningful for static unordered input (i.e., a bag), but becomes meaningless in the streaming setting. Any partial aggregate depends on a particular linear order for the input items, which is inconsistent with the notion of unordered input. So, for a computation of relational aggregates in the streaming setting we must assume that the input contains linearly ordered markers that trigger the emission of output (see [23] for a generalization of this idea). The input can then be viewed as an ordered sequence of bags (each bag is delineated by markers), and it is meaningful to compute at every marker occurrence the aggregate over all items seen so far. Our definition of data-trace transductions captures these subtle aspects of streaming computation with relational data.

**Example 15.** Suppose the input stream consists of unordered natural numbers and linearly ordered markers #. We will specify the computation that emits at every # occurrence the maximum of all numbers seen so far. More specifically, the input type is given by $\Sigma = \{\sigma, \tau\}$, $T_\sigma = \text{Nat}$, $T_{\tau} = \text{Unit}$ (unit type), and $D = \{(\sigma, \tau), (\tau, \sigma), (\tau, \tau)\}$. So, an input data trace is essentially a nonempty sequence of multisets of numbers, i.e. an element of $\text{Bag(Nat)}^*$. The correspondence between traces and elements of $\text{Bag(Nat)}^*$ is illustrated below:

$$\begin{align*}
\varepsilon & \rightarrow \emptyset \\
1 & \rightarrow \{1\} \\
1 \rightarrow \{1\} & \rightarrow \{1, 2\} \\
1 \# & \rightarrow \{1, 2\} \emptyset \\
1 \# & \rightarrow \{1, 2\} \emptyset \\
1 \# & \rightarrow \{1, 2\} \emptyset \\
1 \# & \rightarrow \{1, 2\} \emptyset \\
1 \# & \rightarrow \{1, 2\} \emptyset
\end{align*}$$

The streaming maximum computation is described by the trace transduction $\text{smax} : \text{Bag(Nat)}^* \rightarrow \text{Nat}^*$, defined as follows: $\text{smax}(\varepsilon) = \varepsilon$, $\text{smax}(R_{1}R_{2}) = \max(R_{1})$, and

$$\text{smax}(R_{1} \ldots R_{n}) = \max(R_{1}) \max(R_{1} \cup R_{2}) \cdots \max(R_{1} \cup R_{2} \cup \cdots \cup R_{n-1}).$$
The last relation $R_n$ of the input consists of the elements after the last occurrence of a # symbol, and therefore they are not included in any calculation of max. The implementation of smax is modeled as a string transduction $f : (\text{Nat} \cup \{\#\})^* \rightarrow \text{Nat}^*$, which outputs at every # occurrence the maximum number so far. That is, $f(\varepsilon) = \varepsilon$ and

$$f(a_1 \ldots a_n) = \begin{cases} \varepsilon, & \text{if } a_n \in \text{Nat}; \\ \text{maximum of } \{a_1, \ldots, a_n\} \setminus \{\#\}, & \text{if } a_n = \#.
\end{cases}$$

for all sequences $a_1 a_2 \ldots a_n \in (\text{Nat} \cup \{\#\})^*$.

### 4 Templates for Type-Consistent Programming

Complex streaming computations can be naturally described as directed acyclic graphs (DAGs), where the vertices represent simple operations and the edges represent streams of data. Such a representation explicitly exposes task and pipeline parallelism, and suggests a distributed implementation where every vertex is an independent process and inter-process communication is achieved via FIFO channels.

The semantic framework of section 3, which includes the notions of data-trace types and data-trace transductions, will serve a dual purpose. First, it will allow us to give a formal denotational semantics for streaming computation DAGs that respect the input/output stream types. Second, it will enable reasoning about equivalence and semantics-preserving transformations, such as data parallelization. We will focus here on a subset of data-trace types (recall Definition 5) that emphasizes two crucial elements that are required by practical streaming computations: (1) a notion of synchronization markers, and (2) viewing the data items as key-value pairs in order to expose opportunities for data parallelization.

The synchronization markers can be thought of as events that are periodically generated by the input sources. The period is configurable and can be chosen by the application programmer depending on the time-granularity requirements of the computation (e.g., 1 msec, 1 sec, etc). The purpose of the markers is similar to the punctuations of [23] or the heartbeats of [27]. They are used for triggering the output of nonmonotonic operations (e.g., the streaming aggregation of Example 15) and making overall progress, as well as for merging streams in a predictable way by aligning them on corresponding markers. These synchronization markers are always assumed to be linearly ordered, and they occur in order of increasing timestamp.

Let $K$ be a set of keys and $V$ be a set of values. We write $U(K, V)$ to denote the data-trace type with alphabet $K \cup \{\#\}$, values $V$ for every key, values $\text{Nat}$ for the # tag (i.e., marker timestamps), and dependence relation $\{(\#, \#)\} \cup \{(k, \#), (\#, k) \mid k \in K\}$. In other words, the type $U(K, V)$ consists of data traces where the marker tags # are linearly ordered and the elements between two such tags are of the form $(k, v)$, where $k \in K$ and $v \in V$, and are completely unordered. The data-trace type $O(K, V)$ is defined similarly, with the difference that the dependence relation contains additionally $\{(k, k) \mid k \in K\}$. That is, in a data trace of $O(K, V)$ the elements with the same key are linearly ordered between # markers, but there is no order across keys.

A transduction DAG is a tuple $(S, N, T, E, \rightarrow, \lambda)$ which represents a labelled directed acyclic graph, where in particular: $S$ is the set of source vertices, $T$ is the set of sink vertices, $N$ is the set of processing vertices, $E$ is the set of edges (i.e., connections/channels), $\rightarrow$ is the edge relation, and $\lambda$ is a labelling function. The function $\lambda$ assigns: (1) a data-trace type to each edge, (2) a data-trace transduction to each processing vertex that respects the input/output types, and (3) names to the source/sink vertices. We require additionally that each source vertex has exactly one outgoing edge, and each sink vertex has exactly one incoming edge.

A transduction DAG $(S, N, T, E, \rightarrow, \lambda)$ with source vertices named $S_1, \ldots, S_m$ and sink vertices named $T_1, \ldots, T_n$ denotes a data-trace transduction, where the set of input traces is (up to a bijection) $\prod_{i=1}^{m} \{S_i\} \times \lambda(S_i)$ and the set of output traces is (up to a bijection) $\prod_{i=1}^{n} \{T_i\} \times \lambda(T_i)$. Given an
input data trace, we will describe in an operational way how to obtain the output data trace of the
transduction DAG. We will gradually label every edge $e$ of the DAG with a data trace $u(e)$. First,
label every edge emanating from a source vertex with the corresponding input trace. Then, consider
in any order the processing vertices whose incoming edges have already been labeled. For such
a vertex $n$, apply the transduction $\lambda(n)$ to the input traces and label the outgoing edges with the
corresponding output traces. After this process ends, the output of the transduction DAG can be
read off from the data-trace labelling of the edges that point to sinks.

**Example 16 (Time-Series Interpolation)**. Consider a home IoT system where temperature sensors
are installed at a residence. We wish to analyze the sensor time series to create real-time notifications
for excessive energy loss through the windows. The sensor time series sometimes have missing
data points, and therefore the application requires a pre-processing step to fill in any missing
measurements using linear interpolation. We assume that the sensors first send their measurements
to a hub, and then the hub propagates them to the stream processing system. The stream that
arrives from the hub does not guarantee that the measurements are sent in linear order (e.g., with
respect to a timestamp field). Instead, it produces synchronization markers every 10 seconds with
the guarantee that all elements with timestamps $< 10 \cdot i$ have been emitted by the time the $i$-th
marker is emitted. That is, the $i$-th marker can be thought of as a watermark with timestamp $10 \cdot i$.
So, the input stream is a data trace of $U(U_t, M)$, where $M$ is the type of measurements consisting of a
sensor identifier $id$, a scalar value $value$, and a timestamp $ts$. Figure 1 shows a transduction DAG
that describes the pre-processing computation:

- The stage Join-Filter-Map (JFM) joins the input stream with a table that indicates the location of
each sensor, filters out all sensors except for those that are close to windows, and reorganizes the
fields of the input tuple.

- Recall the guarantee for the synchronization markers, and notice that it implies the following
property for the input traces: for any two input measurements that are separated by at least one
marker, the one on the left has a strictly smaller timestamp than the one on the right. So, the
sorting stage SORT sorts for each sensor the measurements that are contained between markers.

- The linear interpolation stage LI considers each sensor independently and fills in any missing
data points.

The preceding paragraphs describe informally the data-trace transductions JFM, SORT and LI. The
transduction DAG of Fig. 1 denotes a data-trace transduction $U(U_t, M) \rightarrow O(ID, V)$.

The computation performed by a processing node is given in a structured fashion, by completing
function definitions of a specified *operator template*. Table 1 shows the three templates that are

---

**Figure 1**: A pipeline that uses linear interpolation to fill in missing data points of sensor time series.

---

**Table 1**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUB</td>
<td>The source of sensor measurements.</td>
</tr>
<tr>
<td>$M$</td>
<td>Type of measurements of the form $(id, value, ts)$.</td>
</tr>
<tr>
<td>JFM</td>
<td>The Join-Filter-Map stage retains the measurements of sensors that are placed next to windows.</td>
</tr>
<tr>
<td>ID</td>
<td>Type of sensor identifiers.</td>
</tr>
<tr>
<td>$V$</td>
<td>Type of timestamped values $(value, ts)$.</td>
</tr>
</tbody>
</table>
| SORT   | For every sensor, it sorts the measurements between synchronization markers in increasing
timestamp order. |
| LI     | For every sensor, it fills in missing data points by applying linear interpolation. |
| SINK   | The destination of the output stream. |
Table 1: Operator templates for data-trace transductions.

<table>
<thead>
<tr>
<th>Operator Type</th>
<th>Description</th>
<th>Function Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OpStateless</strong></td>
<td>transduction $U(K,V) \rightarrow U(L,W)$</td>
<td>onItem(key, value) { } onMarker(Marker m) { }</td>
</tr>
<tr>
<td><strong>OpKeyedOrdered</strong></td>
<td>transduction $O(K,V) \rightarrow O(K,W)$</td>
<td>initialState() { } onItem(state, key, value) { } onMarker(state, key, Marker m) { }</td>
</tr>
<tr>
<td><strong>OpKeyedUnordered</strong></td>
<td>transduction $U(K,V) \rightarrow U(L,W)$</td>
<td>initialState() { } id() { } in(key, value) { } combine(x, y) { } onItem(lastState, key, value) { } onMarker(oldState, key, agg, Marker m) { }</td>
</tr>
</tbody>
</table>

- $U(K,V)$: unordered key-value pairs between markers
- $O(K,V)$: for every key, ordered values between markers
- Type parameters: $K$, $V$, $L$, $W$, $S$, $A$

Supported, which encompass both ordered and unordered input streams. Each operator is defined by a sequential implementation, which we describe informally below. This means that each operator can be modeled as a data-string transduction (recall Definition 9). It can then be proved formally that these string transductions are consistent w.r.t their input/output data-trace types (Definition 10). It follows that each operator that is programmed according to the template conventions has a denotation (semantics) as a data-trace transduction of the appropriate type.

**OpStateless**: The simplest template concerns stateless computations, where only the current input event—not the input history—determines the output. The programmer fills in two function definitions: (1) onItem for processing key-value pairs, and (2) onMarker for processing synchronization markers. The functions have no output (the output type is $Ut$, i.e. the unit type) and their only side-effect is emitting output key-value pairs to the output channel by invoking emit(outputKey, outputValue).

**OpKeyedOrdered**: Assuming that the input is ordered per key, this template describes a stateful computation for each key independently that is order-dependent. The programmer fills in three function definitions: (1) initialState for obtaining the initial state, (2) onItem for processing a key-value pair and updating the state, and (3) onMarker for processing a synchronization marker and updating the state. The functions have output $S$, which is the type of the data structure for representing the state. As for stateless computations, the functions allow the side-effect of emitting output key-value pairs to the output channel. This template requires a crucial restriction for maintaining the order for the output: every occurrence of emit must preserve the input key. If this restriction is violated, e.g. by projecting out the key, then the output cannot be viewed as being ordered.

**OpKeyedUnordered**: Assuming that the input is unordered, this template describes a stateful computation for each key independently. Recall that the synchronization markers are ordered, but the key-value pairs between markers are unordered. In order to guarantee that the computation does not depend on some arbitrary linear ordering of the key-value pairs, their processing does not update the state. Instead, the key-value pairs between two consecutive markers are aggregated using the
operation of a monoid \(A\): the programmer specifies the identity \(id()\) of the monoid, and a binary operation \(combine()\) that must be associative and commutative. Whenever the next synchronization marker is seen, \(onMarker\) is invoked to incorporate the aggregate (of type \(A\)) into the state (of type \(S\)). The behavior \(onItem\) may depend on the last snapshot of the state, i.e. the one that was formed at the last marker. The functions \(onItem\) and \(onMarker\) are allowed to emit output, but the rest of the functions must be pure (i.e., no side-effects).

Table 2 shows how some streaming computations (which are based on the setting of Example 16) can be programmed using the operator templates of Table 1. The first example is the stateless computation \(joinFilterMap\), which retains the measurements of temperature sensors that are placed near windows. The second example is the per-sensor ordered stateful computation \(linearInterpolation\), which fills in the missing data points of a sensor time series by performing linear interpolation. The last example is the per-sensor unordered (between markers) stateful computation that takes the average of the measurements between markers and reports the maximum over all the averages so far.

The templates of Table 1 define data-trace transductions with only one input channel and output
Figure 2: A transduction DAG that is equivalent to the one of Figure 1 and allows data parallelism.

channel. The operation \textit{merge}, which we denote by \textsc{MrG}, combines several input streams into one by aligning them on synchronization markers and taking the union of the key-value pairs that are in corresponding blocks. We consider two variants of \textit{merge}, which we will not distinguish notionally. The first one has unordered input channels with the same input keys and values, i.e. \textsc{MrG} : \textit{U}(\textit{K}, V) \times \cdots \times \textit{U}(\textit{K}, V) \to \textit{U}(\textit{K}, V). The second variant of \textit{merge} has ordered input channels with pairwise disjoint sets of input keys \textit{K}_1, \textit{K}_2, \ldots, \textit{K}_n, so we write \textsc{MrG} : \textit{O}(\textit{K}_1, V) \times \cdots \times \textit{O}(\textit{K}_n, V) \to \textit{O}(\textit{K}_1 \cup \cdots \cup \textit{K}_n, V).

To enable parallelization, we also need to consider operations that split one into stream into several output stream. The \textit{round-robin} splitter, denoted \textsc{rr} : \textit{U}(\textit{K}, V) \to \textit{U}(\textit{K}, V) \times \cdots \times \textit{U}(\textit{K}, V), sends every input key-value pair to one output channel by cycling through them and sends a synchronization marker to all output channels. The \textit{hash-n} splitter, denoted \textsc{h} : \textit{U}(\textit{K}, V) \to \textit{U}(\textit{K}_0, V) \times \cdots \times \textit{U}(\textit{K}_{n-1}, V) sends a key-value pair \textit{(k, v)} with \textit{k} \in \textit{K}_i with the \textit{i}-th output channel where \textit{K}_i = \{ \textit{k} \in \textit{K} \mid \text{hash}(\textit{k}) = \textit{i} (\text{mod} \textit{n})\}. As for the round-robin splitter, the transduction \textsc{h} sends a synchronization marker to all output channels.

In order to convert an unordered input trace of type \textit{U}(\textit{K}, V) to an ordered trace of \textit{O}(\textit{K}, V), we also consider the \textit{sorting} transduction \textsc{sort}^< : \textit{U}(\textit{K}, V) \to \textit{O}(\textit{K}, V). The transformation \textsc{sort}^< uses the linear order \textless to impose a total order for every key separately on the key-value pairs between synchronization markers. Even when the stream source is ordered, the parallelization of intermediate processing stages can reorder the key-value pairs between markers in an arbitrary way. So, if a later stage of the processing requires the original ordered view of the data, sorting must be used.

The templates of Table 1 not only enforce the data-trace type discipline on the input and output channels, but they also expose explicitly opportunities for parallelization and distribution. Computations that are described by the templates \textsc{OpKeyedOrdered} and \textsc{OpKeyedUnordered} can be parallelized on the basis of keys, and stateless computations can be parallelized arbitrarily. For the sensor input stream of Example 16, the processing pipeline shown in Figure 1 can be parallelized. Figure 2 shows an equivalent processing graph, where every stage of the pipeline is instantiated two times. The input for the JFM stage is partitioned in a round-robin fashion, and the input for the \textsc{sort} and \textsc{li} stages is partitioned based on the key (sensor identifier). All the vertices of the graph have a formal denotational semantics as data-trace transductions, which enables a rigorous proof of equivalence for the DAGs of Figure 1 and Figure 2.

5 Implementation in Storm

In the previous section we proposed an abstraction for describing a distributed streaming computation as a transduction DAG, where each processing element is programmed using one of three predefined...
templates. This principled manner of defining computations enforces a data-trace type discipline that disallows operations which depend on a spurious ordering of the data items. Additionally, it enables a number of equivalence-preserving parallelization transformations.

We have implemented a compilation procedure that converts a transduction DAG into a deployment for the distributed streaming framework Apache Storm [4]. In Storm, a computation is structured as a DAG (called topology) of logical source vertices called spouts and processing/sink vertices called bolts. Each vertex (bolt or spout) may be instantiated multiple times, across different physical nodes or CPU threads. In such settings, the connections between vertices always specify a data partitioning strategy, which is employed when the vertices are instantiated multiple times. These connections are called groupings in Storm’s terminology, and the most useful ones are: (1) shuffle grouping, which randomly partitions the stream in balanced substreams, (2) fields grouping, which partitions the stream on the basis of a key, and (3) global grouping, which sends the entire stream to exactly one instantiation of the target bolt. We refer the reader to [5] for more information on the streaming model of computation supported by Storm.

Our framework ensures that the data-trace types of input, output and intermediate streams are respected. Our compilation procedure constructs automatically the glue code for propagating synchronization markers throughout the computation, merging input channels, partitioning output channels, and sorting input channels to enforce a per-key total order on the elements between markers. We use Storm’s built-in facilities for the parallelization of individual processing vertices, but we have replaced Storm’s “groupings” because they inhibit the propagation of the synchronization markers. For efficiency reasons, we always fuse the merging (MRG) transduction and the sorting (SORT) transduction with the operator that follows in order to eliminate unnecessary communication delays.

6 Experimental Evaluation

In this section we experimentally evaluate our data-trace type-based framework. We address two questions:
— Can our system automatically generate code that is as efficient as a hand-crafted implementation, while automatically adapting to whatever levels of parallelism are available?
— Does our framework facilitate the development of complex streaming applications?

To answer the first question, we hand-coded an extension of the widely-used Yahoo Streaming Benchmark [15] and we compare it to an implementation built using our framework. To address the second question, we consider a significant case study: a variant of the Smart Homes Benchmark [1] used in the Grand Challenge of the DEBS 2014 conference, which we have modified to include a more realistic power prediction technique based on a machine learning model.

Experimental Setup
Our focus in the experiments is to establish how well stream applications scale, independent of network performance. To do this, we ran our implementation on top of Apache Storm on a single server machine with 64 GB of memory and 20 cores clocked at 2.20GHz. Across multiple trials and configurations, we measured maximum throughput for each configuration, up to the point where all cores were maxed out at 100% utilization.

6.1 Yahoo Streaming Benchmark
The Yahoo Benchmark [15] defines a stream of events that concern the interaction of users with advertisements, and suggests an analytics pipeline to process the event stream. There is a fixed set of campaigns and a set of advertisements, where each ad belongs to exactly one campaign. The static map from ads to campaigns is computed ahead-of-time and stored in memory. Each element of the data stream is of the form (userId, pageId, adId, eventType, eventTime), and it records the
interaction of a user with an advertisement, where eventType is one of \{view, click, purchase\}. The component eventTime is the timestamp of the event.

The basic benchmark query (as described in [15]) computes, at the end of each second, a map from each campaign to the number of views associated with that campaign within the last 30 seconds. For each event tuple, this involves a lookup to determine the campaign associated with the advertisement viewed. The reference implementation published with the Yahoo benchmark involves a multi-stage pipeline: (i) stage 1: filter view events, (ii) stage 2: project the ad id from each view tuple, (iii) stage 3: lookup the campaign id of each ad, (iv) stage 4: compute for every window the number of events (views) associated with each campaign. The query involves key-based partitioning on only one property, namely the derived campaign id of the event.

We have streamlined the reference implementation [8] by fusing the filtering, projection and lookup operations into one stage (since filtering and projection are very cheap computationally), and by eliminating the use of the Redis [7] in-memory database engine. The input stream is generated by several independent processes, and each process generates events in increasing timestamp order. When we refer to the “reference implementation” from now on we will mean this simplified version. In this implementation, events may arrive at the sliding-window counting stage completely out of order, and the implementation has to account for this disorder by maintaining and updating the counts for all windows that have been opened so far.

To compare the effectiveness of our framework, we next re-implemented the analytics pipeline as a transduction DAG, where every processing vertex is programmed using the templates of Table 1. This is shown in the top diagram of Figure 3, where YItem is the type of input tuples and CID is the type of campaign identifiers. The system is configured so that the stream sources emit synchronization markers at 1 second intervals, i.e. exactly when the timestamps of the tuples cross 1 second boundaries.

Figure 4 shows the experimental comparison of the reference implementation (blue line) and the transduction-based implementation (orange line). We have varied the degree of parallelization for every processing stage of the pipeline from 1 up to 6 (shown in the horizontal axis). Recall that the server has 20 cores clocked at 2.20GHz. The vertical axis shows the maximum throughput that can be attained for each configuration. Observe that the hand-written reference implementation and the implementation generated automatically from the transduction DAG have very similar performance. We conclude that our framework indeed achieves excellent performance — despite the higher-level specification and additional typing information in the transduction-based scheme.
Reference Implementation
Data-Trace Transduction

Figure 4: Yahoo Streaming Benchmark: Reference implementation (dark gray line) vs transduction-based implementation (light gray line). The horizontal axis shows the parallelization for each stage of the pipeline, and the vertical axis shows the throughput in million events/sec. The higher-level implementation achieves equal performance.

6.2 Case Study: Smart Homes Benchmark

To examine the suitability of our framework for more expressive stream processing applications, we consider a variant of the benchmark used for the “Smart Homes” Internet of Things (IoT) competition of the DEBS 2014 conference [1]. In this benchmark, the input stream consists of measurements produced by smart power plugs that are deployed in private households. A smart plug is connected to a wall power outlet, and then an electrical device is connected to the plug. This allows the plug sensors to measure quantities that are relevant to power consumption.

The deployment of these smart plugs is done across several buildings, each of which contains several units. A smart plug is uniquely identified by three numbers: a building identifier, a unit identifier (which specifies a unit within a building), and a plug identifier (which specifies a plug within a unit). For simplicity, we assume here that the plugs only generate load measurements, i.e. power in Watts. More specifically, every stream event is a tuple with the following components:

- **timestamp**: Timestamp of the measurement, given as the number of seconds since January 1, 1970, 00:00:00 GMT.
- **value**: The value of the load measurement (in Watts).
- **plugId**: Identifier that specifies the plug (within a unit) that produced the measurement.
- **unitId**: Identifier that specifies the unit (within a building) where the plug is located.
- **buildingId**: Identifier that specifies that building where the plug is located.

A plug generates roughly one load measurement for every 2 seconds, but the measurements are not uniformly spaced. There are generally gaps in the measurements, as well as cases with several measurements for the same timestamp.

We now describe our implementation of a load prediction pipeline in the framework of data-trace transductions. We assume here that we have a static map (table) at our disposal that determines what kind of electrical device is connected to each smart plug (e.g., A/C unit, lights, fridge, oven, washer, heater, etc.). The load prediction will be performed separately for each device type. The top diagram of Figure 5 is a transduction DAG that describes the computation.

- **Stage JFM** U(Ut, SItem) → U(Plug, VT):
  This “join-filter-map” stage joins the input streams with the static table that specifies the type
of electrical device connected to each plug. Only the devices that generally have a large power consumption (e.g., A/C, heater, etc.) are retained, and the rest (e.g., small electronic devices) are discarded. This stage also reorganizes the fields of the input tuple, and separates them into a plug key (unique global identifier of a plug) of type Plug and a timestamped value of type VT.

- **Stage SORT** $U(\text{Plug}, \text{VT}) \rightarrow O(\text{Plug}, \text{VT})$:
  For every plug key, it sorts the input items (between consecutive markers) according to their timestamp.

- **Stage LI** $O(\text{Plug}, \text{VT}) \rightarrow O(\text{Plug}, \text{VT})$:
  For every plug key, it fills in any missing data points using linear interpolation. This guarantees that for each plug there is exactly one data point per second.

- **Stage Map** $O(\text{Plug}, \text{VT}) \rightarrow U(\text{DType}, \text{VT})$:
  Projects every input key (global identifier of a plug) to the kind of device it is connected to.

- **Stage SORT** $U(\text{DType}, \text{VT}) \rightarrow O(\text{DType}, \text{VT})$:
  For every device type, it sorts the input items (between consecutive markers) according to their timestamp.

- **Stage AVG** $U(\text{DType}, \text{VT}) \rightarrow O(\text{DType}, \text{VT})$:
  Data items with the same key (device type) and same timestamp are averaged. This amounts to computing the average load for the devices of the same type.

- **Stage Predict** $O(\text{DType}, \text{VT}) \rightarrow O(\text{DType}, \text{VT})$:
  For every device type and every input value (there is exactly one value per second), it predicts the total power consumption over the next 10 minutes. The features used for this prediction are: current time, current load, power computation over the past 1 minute. A decision/regression tree is used to perform the prediction (REPTree of [18]), which has been trained on a subset of the data.

Figure 6 shows that by varying the degree of parallelism for each stage of the pipeline the computation scales up almost linearly until level 6, which finally maxes out CPU utilization. As before, we conducted the experiment on a server with 64GB of memory and 20 cores clocked at 2.20GHz; at level 6 parallelism, our framework has generated six concurrent threads for each of the operators in the benchmark. We conclude from these results that our framework indeed can scale out to high levels of concurrency, even for complex operations such as machine learning inference over streams.

Overall, our experiments have demonstrated both that our framework can express the complex computations required in both enterprise and IoT streaming applications, and that it can be compiled to an efficient implementation comparable to hand-coded solutions.
7 Related Work

Dataflow Computation Models: Data-trace transductions are a generalization of what is considered by acyclic Kahn process networks [20], where the interface consists of a finite number of independent linearly ordered input and output channels. A process network consists of a collection of processes, where each process is a sequential program that can read from the input channels and write to the output channels. The input/output channels are modeled as first-in first-out queues. A specialization of process networks is the model of Synchronous Dataflow [22], where each process reads a fixed finite number of items from the input queues and also emits a fixed finite number of items as output. We accommodate a finite number of independent input or output streams, but also allow more complicated independence relations on the input and output, and in particular, viewing the input or output stream as a bag of events is not possible in Kahn process networks.

Streaming extensions of database query languages: There is a large body of work on streaming query languages and database systems such as Aurora [10] and its successor Borealis [9], STREAM [12], CEDR/StreamInsight [14][11], and System S [19]. The query language supported by these systems (for example, CQL [12]) is typically a version of SQL with additional constructs for sliding windows over data streams. This allows for rich relational queries, including set-aggregations (e.g. sum, maximum, minimum, average, count) and joins over multiple data streams, but requires the programmer to resort to user-defined functions in another language for richer computations such as machine learning classification. A precise semantics for how to merge events from different streams has been defined using the notion of synchronization markers [23]. A recent proposal is the query language of [24], called StreamQRE, which proposes a combination of sequence-based computations with static relational operations. The pomset view — central to the formalism of data-trace transductions — is strictly more general, and gives the ability to view the stream in many different ways, for example: as a linearly ordered sequence, as a relation, or even as a sequence of relations. This is useful for describing streaming computations that combine relational operations with sequence-aware operations. Extending relational query languages to pomsets has been studied in [17], though not in the context of streaming. The problem of extending a query language like StreamQRE to the setting of pomsets remains open.

Distributed stream processing engines: A variety of popular stream processing engines, including Apache Samza [2] and Storm [4], Twitter Heron [6], and Apache Spark Streaming [3], have achieved widespread use. Spark Streaming supports SQL-style queries (as above) or, equivalently, lower-level operations roughly corresponding to the relational algebra underlying SQL. The other stream engines provide much lower-level abstractions in which the programmer writes event handlers that
take tuples, combine the data with windows, and emit results. As with the manually coded Storm implementation used in our experiments, this provides great power but does not aid the programmer in reasoning about correctness.

8 Conclusion

We have proposed a type discipline for classifying streams according to their partial ordering characteristics using data-trace types. These types are used to annotate the communication links in the dataflow graph that describes a streaming computation. Each vertex of this typed dataflow graph is programmed using a pre-defined set of templates, so as to ensure that the code respects the types of the input and output channels. We have implemented this framework in Java and we have provided an automatic procedure for deployment on Apache Storm. We have shown experimentally that our framework can express complex computations required in IoT streaming applications, and that it can produce efficient implementations comparable to hand-coded solutions.

A direction for further work is to enrich the set of type-consistent operator templates with common patterns of stream transformations. For example, our templates can already express sliding-window aggregation, but a specialized template for that purpose would relieve the programmer from the burden of re-discovering and re-implementing efficient sliding-window algorithms (e.g., [30, 13, 29, 28]). Other avenues for future research are to extend the compilation procedure to target streaming frameworks other than Storm, and to automatically perform optimizations that take the underlying hardware infrastructure into account.

References


