PiCo: High-performance data analytics pipelines in modern C++

This is the author’s manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/1668444 since 2018-12-02T18:57:04Z

Published version:
DOI:10.1016/j.future.2018.05.030

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
PiCo: High-Performance Data Analytics Pipelines in Modern C++

Claudia Misale\textsuperscript{a}, Maurizio Drocco\textsuperscript{b,*}, Guy Tremblay\textsuperscript{c}, Alberto R. Martinelli\textsuperscript{b}, Marco Aldinucci\textsuperscript{b}

\textsuperscript{a}Cognitive and Cloud, Data-Centric Solutions, IBM T.J. Watson Research Center, Yorktown Heights, New York, USA \\
\textsuperscript{b}Computer Science Department, University of Torino, Torino, Italy \\
\textsuperscript{c}Département d’Informatique, Université du Québec à Montréal, Montréal, QC, Canada

Abstract

In this paper, we present a new C++ API with a fluent interface called PiCo (Pipeline Composition). PiCo’s programming model aims at making easier the programming of data analytics applications while preserving or enhancing their performance. This is attained through three key design choices: 1) unifying batch and stream data access models, 2) decoupling processing from data layout, and 3) exploiting a stream-oriented, scalable, efficient C++11 runtime system.

PiCo proposes a programming model based on pipelines and operators that are polymorphic with respect to data types in the sense that it is possible to reuse the same algorithms and pipelines on different data models (e.g., streams, lists, sets, etc.). Preliminary results show that PiCo, when compared to Spark and Flink, can attain better performances in terms of execution times and can hugely improve memory utilization, both for batch and stream processing.

Keywords: Big Data, High Performance Data Analytics, Domain Specific Language, C++, Stream Computing, Fog Computing, Edge Computing.

1. Introduction

The importance of Big Data has been strongly assessed in the last years, and their role is crucial to companies, institutions, and research centres, which keep incorporating their analysis into their strategic vision, using them to make better and faster decisions. In the next years, most of those Big Data will be produced by the Internet of Things (IoT). By 2020, Cisco expects 50 billion of connected devices [1] with an average of almost 7 per person. According to the current IoT+Cloud paradigm for Big Data Analytics (BDA), data sensed from distributed devices ought to be stored in cloud data centres to be analysed. Whilst data-processing speeds have increased rapidly, bandwidth to carry data to/from data centres has not increased equally fast [2]. For this, supporting the transfer of data from billions of IoT devices to cloud to be analysed is becoming increasingly unrealistic due to the volume and geo-distribution of those devices. Moreover, the need to reduce latency to enable applications to react promptly to events is impelling, and it is common to push computation and storage closer to where data is continuously generated [3].

Fog (or Edge) computing actually aims at exploiting many distributed edge nodes (e.g., routers, set-top boxes, mobile devices, micro data centres) to selectively support distributed latency- or bandwidth-sensitive applications also implementing near-data analytics [3]. Fog configures as a powerful enabling complement to the IoT+Cloud scenario, featuring a new layer of cooperating devices that can run services and analytic tasks, independently from and contiguously with existing Cloud systems [4].

Distributing BDA pipelines into Fog systems poses new challenges, stemming from the dynamic and open nature of these systems. For these systems we envision the following desiderata:

1. The ability to process partial data, either a batch’s partition or a stream’s time window. Neither edge devices nor the Cloud will have a global vision of the whole data (no data lake).
2. The ability to switch from stream to batch data model (and vice versa), as well as reusing code and mixing up the two models in large systems. Data will be subject to different degrees of aggregation or coalescing depending on the system condition.
3. The capacity to deploy on heterogeneous computing devices, possibly low-power and with a lightweight framework stack.
4. The availability of a model for users with a clear and compositional semantics. Compositionality at the language level should be reflected at run-time level. An update in the BDA pipeline should not require to recompile and redeploy the whole system.

From the programmability perspective, a common aim in
BDA tools is to allow ease of programming by providing a programming model independent from the data model, resulting in uniform interfaces for batch and stream processing. For instance, such uniformity is the cornerstone of Google Dataflow [5]. Similarly, modifying a batch-oriented Spark [6] program to perform stream processing amounts to wrapping the existing code with stream-oriented constructs.

In this context, we advocate PiCo (Pipeline Composition), a new C++ framework with a fluent API, based on method chaining [7], designed over a conceptual framework of layered Dataflow. PiCo’s programming model aims at facilitating the programming and enhancing the performance of BDA through a unified batch and stream data model (cf. desideratum 2), coupled with a stream-oriented, scalable, efficient C++11 run-time system (cf. desideratum 3).

Moreover, PiCo stands on a functional abstract model, mapping each program to its unambiguous semantics (cf. desideratum 4). This was an explicit design choice, that distinguishes PiCo from most mainstream BDA tools. For instance, Spark, Storm [8], and Flink [9] typically require specializing the algorithm to match the data access and layout. Specifically, data transformation functions exhibit different functional types when accessing data in different ways. For this reason, the source code must often be revised when switching from one data model to another. Some of them, such as Spark, provide the runtime with a module to convert streams into micro-batches (Spark Streaming, a library running on Spark core), but still, different code needs to be written at the user-level.

PiCo is designed as a headers-only C++11 framework and supports headers-only analytics applications. This means that a PiCo application can be compiled to any target platform supporting a modern C++ compiler. Since the PiCo run-time exhibits no dependencies from third-party libraries, this makes it easy to port PiCo applications, and even deploy them on specialised hardware platforms and appliances, e.g., to run high-frequency stream analytics for edge computing. A PiCo pipeline could also be directly compiled into a FPGA bitstream.

PiCo has been designed under a philosophy that can be summarized as “high performance at low cost,” based on a lightweight C++ runtime, in which data collections are streamed among computational nodes to minimize the memory footprint. In particular, PiCo relies on FastFlow [10], a parallel programming framework designed to support streaming applications on cache-coherent multicore platforms. As confirmed by the results presented below, PiCo steadily exhibits low memory consumption with no compromises on the performance side.

The present paper is an extension of a previous workshop paper [11]. In particular, the PiCo API has been extended to include iterative and binary pipelines, making it possible to implement join-based applications like Page Rank. The experimental evaluation has also been enriched with the use of larger batch and stream datasets, along with Page Rank. Additionally, the PiCo abstract model is now presented—viz., formal syntax and semantics.

This paper thus proceeds as follows. In Section 2, we provide a brief survey over the state-of-the-art data analytics frameworks. In Sections 3 and 4, we introduce the abstract programming model and its current implementation, thus discussing the C++ API exposed by PiCo. In Section 5, we provide some experimental evaluation. Finally, we conclude the paper in Section 6 and present some future work.

2. Related Work

In recent years, a plethora of BDA tools has been proposed. Although each tool claims to provide better programming, data, and execution models—for which only informal semantics are generally provided—all share some characteristics at different levels [12].

Apache Spark design is intended to address iterative computations by reusing the working dataset by keeping it in memory [13, 6]. For this reason, Spark represents a landmark in Big Data tools history, having a strong success in the community. The overall framework and parallel computing model of Spark is similar to MapReduce, while the innovation is in the data model, represented by the Resilient Distributed Dataset (RDD). An RDD is a read-only collection of objects partitioned across a cluster of computers that can be operated on in parallel. A Spark program can be characterized by the two kinds of operations applicable to RDDS: transformations and actions. Those transformations and actions compose the directed acyclic graph (DAG) representing the application. For stream processing, Spark implements an extension through the Spark Streaming module, providing a high-level abstraction called discretized stream or DStream [13]. Such streams represent results in continuous sequences of RDDS of the same type, called micro-batches. Operations over DStreams are “forwarded” to each RDD in the DStream, thus the semantics of operations over streams is defined in terms of batch processing according to the following simple translation: Given a DStream \( a = [a_1, a_2, \ldots] \)—i.e., a possibly unbounded ordered sequence of items, where each item \( a_i \) is a micro-batch of type RDD—then \( \text{op}(a) = [\text{op}(a_1), \text{op}(a_2), \ldots] \). All RDDS in a DStream are processed in order, whereas data items inside an RDD are processed in parallel without any ordering guarantees.

Formerly known as Stratosphere [14], Apache Flink [9] focuses on stream programming. The abstraction used is the DataStream, a representation of a stream as a single object. Operations are composed (i.e., pipelined) by calling operators on DataStream objects. Flink also provides
ory management in modern C++11 has been considerably 210
that memory has to be managed explicitly, although mem-

gle one (called pipeline). A consequence of using C++ is 260
lowing to apply optimizations such as pipelining or chain-

tions is permitted—i.e., it is only possible to apply oper-

processes the data. Bounded PCollections can be pro-

elements. A DIA remains an abstract entity flowing between two concrete DIA operations, al-


to create logical windows of finite size. 

rifying between epochs. 

StreamBox [16] is another recent stream processing en-

gene, also in C++, specifically designed to exploit the 320
pipelines transformed over records, expressed using a 330
program similar to Google Dataflow (viz., PCollection, 340
ner. To enhance parallelism, records as well as epochs may be processed out-of-order. However, 360
while an epoch is run on different machines. 

Apache Storm is a framework targeting only stream pro-

form. Here, the term “Dataflow” refers to the “Dataflow 390
programming model, to describe the processing and programming 400
Storm’s programming model is based on three key notions: 410
Spouts, Bolts, and Topologies. A Spout is a source of a 420
stream, that is (typically) connected to a data source or 430
that can group its own stream. A Bolt is a processing 440
element, so it processes any number of input streams and 450
produces any number of new output streams. A topology 460
is a composition of Spout and Bolts. 

Google Dataflow SDK [5] is part of the Google Cloud Plat-

form. Here, the term “Dataflow” refers to the “Dataflow 490
model,” to describe the processing and programming 500
model of the Cloud Platform. This framework aims at 510
providing a unified model for stream, batch, and micro-

batch processing. The base entity is the Pipeline, repre-

senting a data processing job consisting of a set of opera-

tions that can read a source of input data, transform that 550
data, and write out the resulting output. The data model 560
in Google Dataflow is represented by PCollections, repre-

senting potentially large, immutable bags of elements, that 590
can be either bounded or unbounded. The bounded (or unbounded) nature of a PCollection affects how Dataflow 610
processes the data. Bounded PCollections can be pro-

cessed using batch jobs, that might read the entire data set 630
once and perform processing in a finite job. Unbounded 640
PCollections must be processed using streaming jobs, as 650
the entire collection may never be available for processing 660
at any one time, and they can be grouped by using 670
windowing to create logical windows of finite size. 

Thrill [15] is a prototype of a general purpose big data 690
batch processing framework with a dataflow style program-

ming interface implemented in C++ and exploiting tem-

plate meta-programming. Thrill’s data model is the Dis-

tributed Immutable Array (DIA), an array of items dis-

tributed over the cluster, to which no direct access to ele-

tments is permitted—i.e., it is only possible to apply opera-

tions to the array as a whole. A DIA remains an abstract 750
entity flowing between two concrete DIA operations, al-

lowing to apply optimizations such as pipelining or chain-

ing, combining the logic of multiple functions into a sin-

gle one (called pipeline). A consequence of using C++ is 790
that memory has to be managed explicitly, although mem-

ory management in modern C++11 has been considerably 800
simplified—for instance, Thrill uses reference counting ex-

tensively. Thrill provides a SPMD (Single Program, Mul-

tiple Data) execution model, similar to MPI, where the 810
same program is run on different machines. 

StreamBox [16] is another recent stream processing en-

machine similar to Google Dataflow (viz., PCollection, 840
ner, etc.). Records can be grouped into epochs deline-

ated by watermarks. To enhance parallelism, records as 850
well as epochs may be processed out-of-order. However, 860
all records within an epoch are ensured to be consumed 870
before the watermarks, thus ensuring an appropriate or-

dering between epochs. 

Despite the fact that the design of both the C++ API and 880
the underlying abstract model focuses on minimality (à la 890
RISC), PiCo is expressive enough to almost subsume all 900
the mentioned frameworks. Thus, starting from a program 910
expressed in one of those tools’ API, any function can eas-

ily be mapped to PiCo functions, yielding a PiCo Pipeline 920
with a similar but more abstract semantics. 

From the implementation perspective, along with the same 930
line as StreamBox, PiCo is based on a C++ stream-based 940
runtime, thus we expect PiCo and StreamBox to exhibit 950
similar performance. Unfortunately, it was not possible to 960
deploy StreamBox on our testing platform, therefore an 970
effective comparison has to be relegated to future work. 

3. PiCo Abstract Model 

In this section, we present the abstract model underly-

ing PiCo [17, 18]. We remark that providing the possibil-

ity of reasoning about programs in abstract—rather than 980
operational—terms is a primary goal of PiCo, in particular 990
with respect to the Fog context, where compositionality is a 1000
key requirement (cf. desideratum 4). 

The building blocks of the PiCo syntax are Pipelines (Sect. 3.4), that are composed by Opera-

ators (Sect. 3.3) according to a type system supporting partial polymorphism. Conversely, Collections (Sect. 3.2) 1030
are considered only at the semantic level, where a PiCo Pipeline is mapped to a Dataflow graph, with tokens, 1040
nodes, and edges representing, respectively, Collections, 1050
Operators, and data dependencies. But first, we present a 1060
brief overview of dataflow concepts. 

3.1. The Dataflow Model 

Dataflow is a model of computation where a program is 1070
described as a set of concurrent processes—aka. actors— 1080
communicating with each other by sending/receiving data 1090
tokens through channels [19].
A set of firing rules is associated with each actor. Processing then consists of “repeated firings of actors” based on the availability of tokens. The Dataflow model is inherently parallel as all actors whose data tokens are available can be fired simultaneously.

More precisely, a Dataflow actor consumes input tokens when it fires and then produces output tokens. The function mapping input to output tokens is called the kernel. Typically, kernels are purely functional, meaning that firings have no side effects and that output tokens are pure functions of the input tokens. However, the model can be extended to allow stateful actors. As for output tokens, they can be replicated and placed onto all output channels (i.e., broadcasting) or sent to specific channels (e.g., switch, scatter). Extensions to the basic model also allow to express arbitrary stream consuming policies (e.g., gathering from any channel).

3.2. Collections

In PiCo, collections are semantic entities that flow across semantic graphs, in the form of tokens. A collection is either bounded or unbounded; moreover, it is also either ordered or unordered. A combination of the mentioned characteristics defines the structure type of a collection. We refer to the key structure types with a mnemonic name:

- a bounded, ordered collection is a list;
- a bounded, unordered collection is a (bounded) bag;
- an unbounded, ordered collection is a stream.

A collection type is characterized by a structure type and a data type, namely the type of the collection elements. Formally, a collection type has form \( T_\sigma \), where \( \sigma \in \Sigma \) is the structure type, and \( T \) is the data type, and where \( \Sigma = \{ \text{bag}, \text{list}, \text{stream} \} \) is the set of all structure types. We also partition \( \Sigma \) into \( \Sigma_b \) and \( \Sigma_u \), defined as the sets of bounded and unbounded structure types. Moreover, we define \( \Sigma_o \) as the set of ordered structure types, thus \( \Sigma_b \cap \Sigma_o = \{ \text{list} \} \) and \( \Sigma_u \cap \Sigma_o = \{ \text{stream} \} \). Finally, we allow the void type \( \emptyset \).

At the semantic level, collection are Dataflow tokens. Unordered collections are mapped to multi-sets, whereas ordered collections are mapped to sequences, in which each item is associated with a numeric timestamp, representing its temporal coordinate. In the following, for the sake of simplicity, we restrict the presentation to unordered, bounded collections (i.e., multi-sets), although the model also covers the other collection types [17].

3.3. Operators

Operators are the building blocks composing a Pipeline. They are categorized according to the following grammar of core operator families:

\[
\begin{align*}
\text{core-operator} &::= \text{core-unary-operator} \\
&\mid \text{core-binary-operator} \\
\text{core-unary-operator} &::= \langle \text{map} \rangle \mid \langle \text{combine} \rangle \\
&\mid \langle \text{emit} \rangle \mid \langle \text{collect} \rangle \\
\text{core-binary-operator} &::= \langle \text{b-map} \rangle
\end{align*}
\]

In addition to core operators, generalized operators can decompose their input collections by:

- partitioning the input collection according to a user-defined grouping policy (e.g., group by key);
- windowing the ordered input collection according to a user-defined windowing policy (e.g., sliding windows).

The complete grammar of operators follows:

\[
\begin{align*}
\text{operator} &::= \langle \text{core-operator} \rangle \\
&\mid \langle \text{w-operator} \rangle \mid \langle \text{p-operator} \rangle \mid \langle \text{w-p-operator} \rangle
\end{align*}
\]

where \( \text{w-} \) and \( \text{p-} \) denote decomposition by windowing and partitioning.

Operator types have form \( T_\sigma \rightarrow U_\sigma \), where \( T_\sigma \) and \( U_\sigma \) are collection types. All operators are polymorphic with respect to data types. Moreover, all operators but emit and collect are polymorphic with respect to structure types. Conversely, each emit and collect operator deals with one specific structure type. For example, an emit for a finite text file (i.e., a from-file Operator) would generate a bounded collection of strings, whereas an emit for a stream of tweets would generate an unbounded collection of tweet objects. The depicted polymorphic type system is the essence of the unified processing model for batch and streaming, that we identified as a primary requirement for the Fog context (cf. desideratum 2).

3.3.1. Data-Parallel Operators

Operators in the map family process data collections on a per-element basis, according to a kernel function. They are defined by the following grammar:

\[
\begin{align*}
\langle \text{map} \rangle &::= \text{map } f \mid \text{flatmap } f
\end{align*}
\]

where \( f \) is the kernel function. The former produces exactly one output element from each input element, whereas the latter produces a (possibly empty) bounded set of output elements for each input element and the output collection is the merging of the output sets. Formally:

\[
\text{map } f \ m = \{ f(m_i) \mid m_i \in m \}
\]

\[
\text{flatmap } f \ m = \bigcup \{ f(m_i) \mid m_i \in m \}
\]

Operators in the map family have type \( T_\sigma \rightarrow U_\sigma \), for \( \sigma \in \Sigma \).
Operators in the `combine` family synthesize the elements from an input collection into a single value. They are defined by the following grammar:

\[
\text{(combine)} ::= \text{reduce} \oplus | \text{fold+reduce} \oplus_1 z \oplus_2
\]

The former corresponds to a classical reduction with a binary operator, whereas the latter is a two-phase aggregation that uses \(\oplus_2\) to reduce a set of partial accumulated states obtained using \(z\) and \(\oplus_1\).\(^2\) Formally, for the `reduce` operator—we do not discuss further the `fold+reduce` for the sake of simplicity:

\[
\text{reduce} \oplus m = \left\{ \bigoplus \{ m_i \in m \} \right\}
\]

Operators in the `combine` family deal only with bounded collections, thus they have type \(T_a \rightarrow U_\sigma\), for \(\sigma \in \Sigma_b\).

### 3.3.3. Windowing Operators

In semantic terms, a windowing operator takes an ordered collection, produces a collection (with the same structure type as the input one) of windows (i.e., lists), according to the windowing policy \(\omega\), and applies the subsequent operation to each window.

We also rely on windowing to extend bounded operators\(^3\) and have them deal with unbounded collections; for instance, given a (bounded) windowing `combine` operator \(op\), the semantics of its unbounded variant is a direct extension of the bounded one. Formally, using \(+\) as an infix operator for list concatenation, where \(s^{(\omega)}\) denotes the result of windowing the stream \(s\) according to policy \(\omega\):

\[
w-op \omega s = op s^{(\omega)}_0 + \ldots + op s^{(\omega)}_i + \ldots
\]

For the sake of simplicity, we omit some parts of the formalization, including windowed collections and unbounded map operators. We refer to the complete PiCo formalization \cite{17} for further details.

### 3.3.4. Partitioning Operators

Partitioning operators are defined according to the following grammar, where \(\pi\) is the partitioning policy:

\[
\text{(p-operator)} ::= \text{p-(core-operator)} \pi
\]

Operators in the `combine` and `b-map` families support partitioning, so, for instance, a `p-combine` produces a multi-set, in which each value is the synthesis of one group; also the natural join operator from the relational algebra is a particular case of per-group joining.

In semantic terms, a partitioning policy \(\pi : T \rightarrow K\) defines how to group the elements from a \(T\)-typed multi-set \(m\), such that each element \(m_i\) is mapped to a sub-collection, depending on the corresponding key \(\pi(m_i) \in K\). The sub-collection containing all the elements from \(m\) mapped to \(k\) by \(\pi\) is denoted by \(\sigma_k^{(\pi)}(m)\), whereas the collection containing all such sub-collections is denoted by \(\epsilon^{(\pi)}(\pi)\).

Given a `combine` operator \(op\) and a partitioning policy \(\pi\), the semantics of operator \(p-op\ \pi\) is as follows:

\[
p-op \pi c = \left\{ \text{op } c' : c' \in \epsilon^{(\pi)}(\pi) \right\}
\]

As for binary operators, given an operator \(op\) in the `b-map` family and a partitioning policy \(\pi\), the semantics of `p-op \pi\` is defined as follows, for all sub-collections mapped to the same key \(k \in K\):

\[
p-op \pi (m_1, m_2) = \bigcup \{ \text{op } (\sigma_k^{(\pi)}(m_1), \sigma_k^{(\pi)}(m_2)) \}
\]

The common `group-by-key` partitioning, with \(\pi_1\) being the left projection,\(^4\) can be identified by collections with data type \(K \times V\) and \(\pi = \pi_1\).

\(^2\)More precisely, \(z \in S\) is the initial value, \(\oplus_1 : S \times T \rightarrow S\) is the fold (how each input item affects the accumulated state) and \(\oplus_2 : S \times S \rightarrow S\) is the reduce (how the various states are combined into a final value).

\(^3\)An operator is `bounded` if it only deals with bounded collections.

\(^4\)\(\pi_1(x, y) = x\)
3.4. Pipelines

PiCo’s cornerstone concept is the Pipeline. The grammar for Pipelines is essentially a graph-composition of Operators (Sect. 3.4.1), that allows to map each Pipeline to a corresponding Dataflow graph, representing its semantic counterpart (Sect. 3.4.2).

3.4.1. Pipeline Syntax

Fig. 1 presents the grammar for Pipelines. Pipelines in Figs. 1a, 1b, and 1c, respectively, produce, consume, and process data, through the operator op. Fig. 1d is the to composition, in which the output from Pipeline $P$ is sent as input to every downstream Pipeline—a particular case is when $n = 1$, which then represents basic Pipeline chaining. Fig. 1e is the merge composition, in which the outputs from two Pipelines are merged. Fig. 1f is the pair-with composition, in which the input is paired with the output of an internal Pipeline $P$, through the binary Operator op. Finally, Fig. 1g is the iterate constructor, that represents iterative processing by feeding the output of the internal Pipeline $P$ back as input to $P$ itself, until some termination condition (denoted by $h$ in Fig. 1g) holds.

Note that in Fig. 1, a dashed line means the path may be void. We refer to such input- or output-less Pipelines as, respectively, sources and sinks. In particular, a Pipeline is executable only if it is both input-less and output-less. The type system for Pipelines, which we omit for the sake of simplicity, defines legal compositions with respect to both data and structure types—see the whole formalization for further details [17]. We remark that, to ease the implementation of composition and typing, we prohibit legal pipelines to have multiple input or output paths.

3.4.2. Pipeline Semantics

The semantics of a Pipeline is obtained by mapping its syntactic tree to a Dataflow graph, such that Operators are mapped to Dataflow nodes, Dataflow edges are inferred from the Pipeline structure, and Dataflow tokens represent either synchronization signals (for sources, e.g., a firing signal for an emit Operator that allows to read from a file: cf. sect. 3.3) or whole collections. Therefore, the proposed mapping is denotational rather than operational, in the sense that it models the (abstract) functional transformations applied to (possibly unbounded) data collections, rather than any specific data processing.

The most basic mapping is for a single-Operator Pipeline (Figs. 1a, 1b, and 1c), resulting in a Dataflow graph with a single node that, when fired upon the arrival of an input token, produces and emits a token according to its Operator semantics; moreover, if the Operator is input-less, a firing (synchronization) token is preloaded on the incoming edge, so that the node is activated at the beginning of the (logical) execution.

The semantic mapping for compound Pipelines is defined by structural induction. Some mappings introduce additional Dataflow nodes, with respect to syntactic nodes; for instance, mapping to and merge Pipelines introduces gathering nodes at the downstream end of the resulting Dataflow graph. Moreover, some mappings require to modify the internal behavior of Dataflow nodes; for instance, mapping a to Pipeline requires modifying the most upstream node, so that it emits $n$ copies of the produced tokens, one for each downstream sub-graph (resulting from mapping each of the downstream Pipeline in Fig. 1d).

Finally, mapping iterate Pipelines is the most complex case, since it requires to both: 1. introduce an additional upstream (resp. downstream) switch node to control the input source (resp. output destination) for the iteration “box”; 2. introduce additional edges from the downstream switch to each input-less node within the iteration box, so that firing tokens are emitted to trigger the next iteration.

4. PiCo Implementation

In this section, we present an implementation of the model discussed in Section 3. The presented implementation is available as open-source code.\(^5\)

The logical path from an (abstract) PiCo Pipeline to the corresponding runtime-level support is illustrated in

\(^5\)https://github.com/alpha-unito/PiCo
As a fundamental design choice, we adopted a functional fluent-style interface, exploiting method cascading (aka. method chaining) to relay the instruction context to a subsequent call. According to the functional design, each function call returns a fresh object (either Pipeline or Operator), so that natural composition is supported. More-
instance, a BinaryMap object represents a PiCo Operator in the b-map family (e.g., join-map), that processes pairs of elements coming from two different input collections and produces a single output for each pair. A BinaryMap object is passed as parameter to Pipeline objects built by calling the pairWith member function (cf. Table 1). Table 2 summarizes the constructors for the subclasses of the Operator class, whose instances represent PiCo Operators. For each constructor, the corresponding abstract operator is described.

The constructors Map, FlatMap, and Reduce produce core data-parallel Operators (cf. Sect. 3.3.1). The FlatMap constructor takes as input a FlatMapCollector object, whose add method is used to build the output collection to be emitted by a kernel instance. Similarly, the JoinMap constructor produces a (core) binary join-map Operator (cf. Sect. 3.3.2).

In addition to core Operators, some constructors are provided that produce partitioning Operators (cf. Sect. 3.3.4). For instance, the ReduceByKey constructor produces a p-reduce Operator, that reduces a key-value collection by applying the kernel function to each value mapped to a given key. Similarly, a JoinMapByKey produces a p-join-map Operator, that applies the kernel function to each pair generated by joining two key-value collections, on a per-key basis. Moreover, although not shown in Table 2, the window function can be invoked on any supported Operator (e.g., a combine Operator), by passing as input a windowing policy, yielding a windowing Operator.

Constructors for source and sink Operators are also provided. For instance, the ReadFromFile constructor produces a from-file Operator, that reads data from an input file on a per-line basis and returns each line as a std::string. Dually, the WriteToFile constructor produces a to-file Operator, that writes data to an output file through the << operator. Finally, ReadFromSocket and WriteToSocket are the socket-based sources and sinks.

### 4.1.3. Polymorphism

One distinguishing feature of the PiCo programming model, compared to other state-of-the-art frameworks, is that the same syntactic object (viz., a Pipeline or an Operator) can be used to process data collections having different types. As discussed in Sect. 3, Pipelines and Operators are polymorphic with respect to both data types and structure types (i.e., the “shape” of the collections, such as bag or stream).

In the proposed C++ API, the data type polymorphism is expressed at compile-time by implementing operators as template classes. As shown in Table 2, each operator takes a template parameter representing the data type of the collections processed by the operator. Namely, any copy-constructible\(^7\) data type is supported. Moreover, each Pipeline object is decorated by the set of supported structure types.

In the proposed type system, all polymorphism is dropped in executable Pipelines. Therefore, executable Pipe objects have a unique type. By construction, source operators (i.e., objects of the SourceOperator class) play the role of specifying the unique structure type processed by the (executable) pipe they belong to. For instance, a pipe starting with a ReadFromFile type operator will only process multi-sets, whereas a pipe starting with a ReadFromSocket type operator will only process streams. The types supported by a Pipe object are not exposed by the C++ API, but they are exploited at run-time to

---

<table>
<thead>
<tr>
<th>Function</th>
<th>Abstract Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipe()</td>
<td>Create an empty Pipeline</td>
</tr>
<tr>
<td>template&lt;typename OpType&gt; Pipe(OpType&amp;)</td>
<td>Create a Pipeline from an initial operator</td>
</tr>
<tr>
<td>template&lt;typename OpType&gt; Pipe add(OpType&amp;)</td>
<td>Create a Pipeline by adding a new stage to the subject</td>
</tr>
<tr>
<td>Pipe to(Pipe&amp;)</td>
<td>Create a Pipeline by appending the argument Pipeline to the subject</td>
</tr>
<tr>
<td>Pipe to(std::vector&lt;Pipe*&gt;&amp;)</td>
<td>Create a Pipeline by appending a set of Pipelines, all fed by the subject</td>
</tr>
<tr>
<td>template&lt;typename BinaryOpType&gt; Pipe pairWith(Pipe&amp;, BinaryOpType&amp;)</td>
<td>Create a Pipeline by pairing the subject with an argument Pipeline, through a binary operator</td>
</tr>
<tr>
<td>Pipe merge(Pipe pipe)</td>
<td>Create a Pipeline by merging the subject with the argument Pipeline</td>
</tr>
<tr>
<td>template&lt;typename TerminationPolicy&gt; Pipe iterate(TerminationPolicy&amp;)</td>
<td>Create a Pipeline that iterates the subject until the termination condition is met</td>
</tr>
<tr>
<td>void run()</td>
<td>Execute the subject Pipeline</td>
</tr>
</tbody>
</table>

Table 1: The Pipe class API. Although const annotations are omitted for readability, all the arguments are const and all the functions are marked as const, yielding a purely functional API.

---

\(^7\)http://en.cppreference.com/w/cpp/concept/CopyConstructible
Table 2: Operator constructors. As for the Pipe API in Table 1, const annotations are omitted for readability but the API is purely functional.

<table>
<thead>
<tr>
<th>Constructor</th>
<th>Abstract Operator</th>
</tr>
</thead>
</table>
| template<
| typename T, typename U>
| Map(std::function<U(TK)>) f)                    | map f, with kernel function f : T -> U |
| template<
| typename T, typename U>
| Flatten(std::function<void(TK, FMapCollector<U>)> f) | flatmap f, with kernel function f : T -> p(U) |
| template<T>                                    | reduce f, with kernel function f : T x T -> T |
| template<X, typename V>
| ReduceByKey(std::function<V(TK, VK)> f)         | p-reduce π1 f', with kernel function f' : V x V -> V, that applies f to the values and leaves the K-typed keys unchanged (a slight variant of the semantics in Sect. 3.3.4) |
| template<T1, typename T2>
| JoinMapByKey(std::function<U(T1, T2)> f)        | join-map f, with kernel function f : T1 x T2 -> U |
| template<T1, typename T2>
| JoinMapByKey(std::function<U(T1, T2)> f)        | p-join-map π1 f, with kernel function f : T1 x T2 -> U, with T1 = K x T1' and T2 = K x T2' |
| template<T>                                    | to-file, writing data to the argument filename |
| template<
| SocketType>
| ReadFromSocket(SocketType &, char)              | from-socket, reading data from the argument socket, where lines are separated by the argument delimiter |
| template<
| typename T>
| WriteToSocket(std::string &)                    | to-socket, writing data to the argument socket through the << operator for type T |

The usage of the proposed API is illustrated in Listing 1, with the source code for the PageRank algorithm.

In the example, the top-level executable Pipeline is the pageRank object (line 60). At the input end, the generateLinks Pipeline (line 61) builds the input graph by: 1. reading lines from the input text file; 2. parsing each line, by mean of the map Operator parseLinks (line 14), into a src-links pair, where src is a page identifier and links is the set of pages cited by src. From the generated graph, the map Operator generateInitialRanks (line 23) generates the initial ranking by associating the default rank 1.0 to each page. The initial approximation is fed as input to the core Pipeline, built by iterating the improveRanks Pipeline (line 54) for 20 iterations. Finally, the WriteToFile Operator (line 64) writes the computed ranks to the output text file.

The improveRanks Pipeline, whose iteration is the core processing in the example, combines a ranking with the input graph (i.e., the output from generateLinks), through the binary Operator computeContributions (line 29). At the first iteration, the input ranking is the output from generateInitialRanks, whereas at iteration i+1 it is the output from the previous i-th iteration. Internally, computeContributions joins the input collections (i.e., the input graph and the most recent ranking) based on page identifiers, so that, for a given page p, the pairwise kernel receives as input both p's neighbors (nl1) and the rank associated to p (nr); for each of such pairs, the Operator generates a contribution from each of p's neighbors, according to the PageRank formula, by calling the add function on the collector (cf. Sect. 4.1.2). The contributions are reduced on a per-key basis, so that all the contributions for a given page, generated by different instances of the binary Operator, are summed up through the p-reduce Operator sumContributions (line 36). Finally, each contribution is converted into an updated rank through the normalization provided by the map Operator normalize (line 42). Note that, differently for instance from the reference Spark implementation of Page Rank, there is no need to specify which data should be cached, since in PiCo only the runtime system is in involved in taking this kind of decisions.

4.2. Intermediate Optimizations

We now show how a PiCo program is mapped to a graph of parallel processing nodes. Logically, as shown in Fig. 2, the mapping takes as input a Semantic Dataflow graph representing a PiCo program (cf. 3.4.2) and produces a Parallel Execution (PE) Graph, representing a possible parallelization of the Semantic graph. The resulting PE graph is a classical macro-Dataflow network [19], in which tokens represent portions of data collections and nodes are persistent processing units, mapping input to output tokens, according to a pure functional behavior.
Dataflow networks naturally express some basic forms of parallelism. For instance, non-connected nodes (i.e., independent nodes) may execute independently from each other, exploiting embarrassing parallelism. Moreover, connected nodes (i.e., data-dependent nodes) may process different tokens independently, exploiting pipeline parallelism (also known as task parallelism). Finally, each PiCo operator is compiled into a Dataflow (sub-)graph of nodes, each processing different portions of the data collection at hand, exploiting data parallelism.

We also provide a set of rewriting rules for optimizing PE graphs, similarly to what is done by optimizing compilers over intermediate representations. For instance, Fig. 3 shows the rewriting for a map/p-reduce composition. This optimization pattern is commonly referred as shuffle*: between the map and p-reduce operators, the data is moved from the map workers (ui in the figure) to the reduce workers (vi in the figure) by following a partitioning criterion. By shuffling data, only the data belonging to a given partition can be assigned to each worker, and the reduce operator produces a single value for each partition. Typically, data shuffling produces an all-to-all communication pattern among map and reduce workers, highlighted by the dotted box in Fig. 3c. As a further optimization, part of the reducing phase can be moved into the map workers, so that each reduce worker computes the final result for each key by combining partial results coming from map workers.

4.3. FastFlow Runtime

The bottom-level runtime support for PiCo is implemented on top of FastFlow [10], a programming framework in which lock-free parallel applications can be expressed as arbitrary networks of ff_node objects (i.e., threads). FastFlow provides a set of preset core patterns, capturing some common networks such as ff_pipeline, representing the pipeline pattern, and ff_farm, representing the farm pattern. We exploit ff_node, ff_pipeline, and ff_farm as building blocks for implementing PE graphs.

The following mapping holds for the API from Sect. 4.1,

- Pipe objects (i.e., PiCo Pipelines) are implemented as ff_pipeline objects; in particular, iterative Pipelines are implemented as FastFlow pipelines with a feedback channel;
- All operators but sources and sinks are implemented as ff_farm objects, in which multiple workers are spawned to exploit data parallelism;
- Input and output operators are implemented as ff_node objects, thus only sequential input/output is provided in the current implementation.

In the prototypical implementation we present, the internal parallelism of each Operator is set manually (i.e., no automatic elasticity is provided). Although omitted for simplicity in Table 2, each Operator constructor takes as

---

The shuffle pattern is sometimes referred as “parallel-sorting.”

Listing 1: PageRank example in PiCo. Input and output filenames are taken as first and second command-line parameters, respectively.
5. Experiments

In this section, the performance sustained by PiCo in a shared-memory setting is evaluated in comparison to two mainstream tools, Spark (v2.1.0) and Flink (v1.2.0).

The machine used for the experiments is the Occam Supercomputer\(^9\) (Open Computing Cluster for Advanced data Manipulation) [20], designed and managed by the University of Torino and the National Institute for Nuclear Physics. We used one node having the following characteristics: Hardware: 4x Intel\(^\circledR\)Xeon\(^\circledR\)Processor E7-4830 v3 12 core/2.1Ghz, 768GB/1666MHz (48 x 16GB) DDR4 RAM, 1x SSD 800GB + 1x HDD 2TB/7200rpm, InfiniBand 56Gb + 2x Ethernet 10Gb. Software: Linux CentOS v7.3 with Linux kernel 3.10, gcc v4.8.5 compiler (PiCo was compiled with o3 optimization flag), and OpenJDK Server v1.8 Java runtime.

For the sake of fairness, the comparison is carried on the configuration leading to the best performance, for each tool.\(^10\) For Spark and Flink, as both rely on master-workers run-time systems, this means finding the optimal number of workers. For PiCo, each program is mapped to a different network of threads, requiring to explore a broader space of configurations. Although a general solution for automating this is beyond the current scope, some simple strategies are discussed for each benchmark.

The observed results are summarized in Table 3 (p. 12). Each reported measurement is the average of 10 repetitions. Since only negligible deviations were observed, no discussion about variance is included in the analysis.

This section proceeds as follows. Sect. 5.1 discusses two benchmarks conforming to the simple MapReduce pattern, namely Word Count and Stock Pricing. Sect. 5.2 discusses a streaming variant of Stock Pricing. Finally, Sect. 5.3 discusses Page Rank, a more complex application, typical of a real world application.

5.1. MapReduce-like Benchmarks

In Big Data analytics, Word Count is considered the “Hello, World!” application. In its batch flavour, Word Counts finds the number of occurrences of each word from a text file. Fig. 4 shows the semantic graph (cf. Sect. 3.4.2) for a PiCo Pipeline implementing Word Count.\(^11\) The input is read line by line from the text file foo; each line is tokenized into a sequence of words using a FlatMap node (as each line contains a varying number of words) and each word \(w\) is mapped to a key-value pair \((w, 1)\); the generated pairs are grouped by-word and then the values (i.e., the 1s) are summed up by a ReduceByKey node; finally, the result (i.e., one pair per word along with its number of occurrences in the text) is written to the text file bar.

The second benchmark is Stock Pricing, an application from the stock market domain that computes the maximum price for each option from a bunch of stock-option

\(^9\)http://c3s.unito.it/index.php/super-computer
\(^10\)Initialization/termination excluded.
\(^11\)The semantic graph for a PiCo application can be generated in dot format by the to_dotfile function from the Pipe class API.
data. Its semantic graph is omitted since it is analogous to the Word Count graph (Fig. 4): each stock-option entry from a text file is parsed and processed by a Map node, to compute the stock-price associated to the entry, by means of the Black & Scholes formula; finally, a ReduceByKey extracts the maximum price for each stock name and the output is written to a text file.

5.1.1. Settings

For the Word Count benchmark, an input file of 6 GB is considered, formed by text lines of various lengths, each composed by random words from a dictionary of 1024 words. For the Stock Pricing benchmark, an input file of 10 GB is considered, generated by replicating the largest dataset included in the PARSEC suite for the blackscholes benchmark.

The parallel configuration for PiCo can be controlled by setting the degree of parallelism for parsing and processing stages. Indeed, in the current implementation, both benchmarks are implemented at the FastFlow level as two pipelined farms, one for reading the input file and parsing lines into strings (ReadFromFile, Fig. 4), the other for applying the map-reduce processing (FlatMap and ReduceByKey, Fig. 4). Therefore, the optimal configuration is observed with the minimal parallelism degree for parsing that prevents a bottleneck at parsing side, leaving as many resources as possible to the heavier processing stage.

As we discuss in Sect. 6, more refined solutions can be envisioned based on a larger set of intermediate optimizations (cf. Sect. 4.2), leading to improved FastFlow networks. However, in the current implementation, the simple heuristic already discussed yields good performance, as we show in the following.

5.1.2. Results

Table 3 shows the comparison on minimum execution time, whereas Fig. 5 shows the relative speedup for the Word Count benchmark, with respect to the parallelism degree of the processing stage.

For Word Count, with the optimal parallelism for parsing found at 4, PiCo features the best performance with parallelism for processing at 36, that is also the maximum degree in Fig. 5 not causing overbooking. For the Stock Pricing benchmark, 8 parallel readers are needed to avoid the bottleneck, then lowering the optimal parallelism for processing to 32. With these configurations, PiCo outperforms both Spark (48 workers) and Flink (24 workers).

5.2. Streaming Benchmark

We now consider a streaming variant of the Stock Pricing benchmark discussed above. In this variant (referred as “Streaming S.P.” in Table 3), the data comes from a socket (rather than a file) and more complex processing is done on each stock-option entry—Binomial Tree and Explicit Finite Difference are computed, in addition to the Black & Scholes formula, and the result is averaged. Finally, a per-window reduction is performed under fixed-size windowing with size 8. The semantic graph for this benchmark is again analogous to the graph in Fig. 4.

5.2.1. Settings

The same 10GB input file as in Stock Pricing is used, but streaming it to a network socket using the netcat tool.

![Figure 4: Semantic graph for Word Count in PiCo.](image)

![Figure 5: Scalability of Word Count in PiCo.](image)

<table>
<thead>
<tr>
<th></th>
<th>Spark</th>
<th>Flink</th>
<th>PiCo</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Min. Execution Time (s)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word Count</td>
<td>10.07</td>
<td>69.29</td>
<td>5.95</td>
</tr>
<tr>
<td>Stock Pricing (S.P.)</td>
<td>19.99</td>
<td>42.80</td>
<td>14.20</td>
</tr>
<tr>
<td>Page Rank</td>
<td>226.52</td>
<td>82.11</td>
<td>81.08</td>
</tr>
<tr>
<td><strong>Max. Throughput (MB/s)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streaming S.P.</td>
<td>18.63</td>
<td>37.89</td>
<td>112.39</td>
</tr>
</tbody>
</table>

Table 3: Outcome of the performance experiments. All measurements refer to the best-performing parallel configuration. Execution time and throughput are showed for batch and streaming benchmarks, respectively. The memory footprint for Spark and Flink is represented by the difference in memory usage with respect to PiCo.

---

First, let us note that the we could not do a totally fair comparison with Spark, since Spark does not provide fixed-size windowing.

As for PiCo, although the streaming variant is semantically similar to the Stock Pricing batch benchmark, its parallel configuration is different. First, in the current implementation, no parallelism is exploited when reading from/writing to sockets. Second, no fusion is applied between Map (or FlatMap) and per-window reductions, when performing intermediate optimizations. Therefore, in the streaming benchmark, the parallelism degrees for the two stages can be set independently.

5.3.2. Results

Table 3 shows the comparison on maximum sustained throughput. In terms of stock-option entries processed per second, PiCo processes more than 1.7M entries per second, outperforming Flink and Spark, that process approx. 600K and 300K entries per second, respectively.

The optimal parallel configuration for PiCo was found when not exploiting any parallelism on the reduce-side while setting the map-side parallelism to 24; Spark and Flink also attain the best performance with 24 workers.

In addition to poor throughput, Table 3 shows a striking memory explosion for Spark. To understand this poor performance, an analysis of bottlenecks in streaming platforms would need to be performed, which would also help determine the overhead induced by the JVM [21].

5.3. Page Rank

In this section, we now examine the Page Rank application, described in Sect. 4.1.4. Fig. 6 shows the semantic graph for a PiCo implementation, which is a slight variant of the code in Listing 1. In particular, this variant takes as input the list of pages and the list of links from two separate files (like the Flink implementation) and the adjacency lists are generated by per-page reduction of the links (the leftmost ReduceByKey node in Fig. 6).

5.3.1. Settings

The largest graph from the SNAP repository, representing links between users of the LiveJournal on-line community, is given as input to the Page Rank implementations. The graph has approx. 4.8M nodes and 69M edges (i.e., edge factor is 14.37), requiring approx. 1GB of memory.

The parallel configuration for PiCo is relevant only for the critical portion of the application—the highlighted box in Fig. 6. Indeed, the execution time is dominated by the iterative processing. Within the critical portion, the optimal parallelism degree for the ReduceByKey stage was found at 6, whereas various degrees were tested for the (binary) FlatMap stage.

In the current implementation, as for the benchmarks in Sect. 5.1, only some map-reduce fusion is performed as intermediate optimizations (e.g., between FlatMap and ReduceByKey within the critical portion). Therefore, the performance will unavoidably suffer from significant resource overbooking, as discussed in more detail in Sect. 6.

5.3.2. Results

Table 3 shows the comparison for minimum execution time. The results show that, although suffering from limited scalability due to the factors discussed above, PiCo performs as well as Flink, reaching the optimal performance with parallelism degree set to 16 for the FlatMap stage, whereas Spark and Flink attain the best performance with 32 workers. Again, Spark exhibits the highest execution time and significant memory explosion.

6. Conclusions

In this paper, we presented PiCo, a new C++ API with a fluent interface for generic Big Data Analytics pipelines, based on a streaming runtime that provides high performance at low cost of resource consumption.

PiCo provides an expressive programming model backed by a functional abstract semantics. The abstract model is coupled with a concrete API expressed using modern C++, thus ensuring good code portability. One distinguishing feature of PiCo is the polymorphic pipelines, that allow uniform programming for different data models (i.e., stream or batch processing).

The experiments performed, where we compared execution times in shared memory for both batch and stream applications, showed that PiCo attained the best execution time when compared to two state-of-the-art frameworks, Spark and Flink. Moreover, the C++ streaming runtime provides efficiency also in terms of low memory footprint, thus showing it is possible to do high performance analytics with low resource consumption.

The presented results allow us to advocate the exploitation of PiCo for emerging Fog scenarios, as it addresses some fundamental challenges inherent in these systems: batch and stream programs can be assembled by composition (enabled by the abstract model) and executed over a lightweight, high-performance runtime. In this respect, we remark that both Spark and Flink are systems supporting distributed memory hardware holistically, and both the higher memory footprint and the longer execution time may be strongly influenced by the support for system-wide fault tolerance and data distribution.

https://snap.stanford.edu/data/
As future work, we plan to design and implement a broader class of intermediate optimisations for runtime-level networks, with the aim of minimising both the number of threads—thus alleviating the performance penalties due to overbooking—and the amount of inter-thread communication. In the same perspective, we will consider adopting some strategy to automate (at least partially) the process of finding optimal parallel configurations. Finally, we plan to investigate a distributed-memory implementation on top of some distributed-shared memory system. In that context, we will also address the issues of fault-tolerance (e.g., automatic restores in case of failures).

Acknowledgments

We would like to thank the C3S supercomputing facility for the support with the OCCAM SuperComputing Ph.D. Scholarship program.

This work has been partially supported by the EU-H2020 RIA project “Toreador” (no. 688797), the EU-H2020 RIA project “Rephrase” (no. 644235), and the 2015-2016 IBM Ph.D. Scholarship program.

References


