Stivo: An Embedded DSL for High Performance Big Data Processing

Stefan Ackermann
ETH Zürich
stefaack@student.ethz.ch

Vojin Jovanovic
EPFL
first.last@epfl.ch

Tiark Rompf
EPFL
first.last@epfl.ch

Martin Odersky
EPFL
first.last@epfl.ch

ABSTRACT

Cluster computing systems today impose a trade-off between generality, performance and productivity. Hadoop and Dryad force programmers to write low level programs that are tedious to compose but easy to optimize. Systems like Dryad/LINQ and Spark allow concise modeling of user programs but do not apply relational optimizations. Pig and Hive restrict the language to achieve relational optimizations, making complex programs hard to express without user extensions. However, these extensions are cumbersome to write and disallow program optimizations.

We present a big data processing library called Stivo. It uses deep language embedding in Scala, staged execution and explicit side effect tracking to analyze the structure of user programs. This analysis is used to apply early projection insertion, which eliminates unused data, and code motion, together with operation fusion to highly optimize the performance critical path of the system. The language embedding and a high-level interface make Stivo programs expressive and resembling regular Scala code. Modular design and a library approach allow users to extend Stivo with modules, which specify an abstract interface and how to generate high performance code for it. Through a modular code generation scheme, Stivo can execute programs on both Spark and Hadoop. On Spark we achieve speedups of up to 149% over naïve implementations, while on Hadoop they can reach up to 59%.

Keywords

Big data, Domain-specific languages, Code Generation, Multi-stage programming, MapReduce

1. INTRODUCTION

In the past decade, numerous systems for big data cluster computing have been studied [11, 27, 22, 26, 14]. Programming models of these systems impose a trade-off between generality, performance and productivity. Systems like Hadoop MapReduce, [3] and Dryad [16] provide a low level general purpose programming model that allows writing fine grained and optimized code. However, low level optimizations greatly sacrifice productivity [8]. Restricted programming models like Pig latin [22] exploit domain knowledge to provide both good performance and productivity for a large number of use cases. However, they support generality only through user defined functions that are cumbersome and difficult to optimize. Finally, models like Spark [14], FlumeJava [8] and Dryad/LINQ [27] provide high level operations and general purpose programming models, but their performance is limited by glue code between high level operations. Also, many relational optimizations are impossible due to the lack of knowledge about the program structure.

The above mentioned trade-off exist due to imperative programming models, run-time binding of methods and open world assumptions commonly used in languages for big data processing. This limits the compiler's ability to optimize the code. Less efficient abstractions like iterators and channels are not removed from the code that connects declarative operations. Side effect free operations like date/time instantiation, regular expression compilation and high precision decimal numbers are often hidden by abstraction and recomputed in the hot path for each piece of data.

Domain specific approaches, like Pig and SQL, have a narrow and side effect free interface and provide good optimizations. However, they come with their own set of limitations. Their programming model is often too simple for a wide range of problems. This requires reverting to external language operations, which are again hard to optimize, or to abandoning the model completely. Moreover, there is the high overhead of learning a new language, and proper tool support, like debugging, type driven correctness checking and IDE support is often limited. It is also hard to extend these frameworks with optimizations for a new domain.

Recently history there have been several solutions that try to make programming big data systems efficient, productive and general at the same time. Steno [21] implements an innovative runtime code generation scheme that eliminates iterators in Dryad/LINQ queries. It operates over flat and nested queries and produces a minimal number of loops without any iterator calls. Manimal [18] and HadoopToSQL [17] apply byte code analysis, to extract information about unused data columns and selection conditions; as a result to

gain enough knowledge to apply common relational database optimizations for projection and selection. However, since these solutions use static byte code analysis they must be safely conservative which can lead to missed optimization opportunities.

This paper presents a new domain specific language Stivo for big data processing that provides a high level declarative interface similar to Dryad/LINQ and Spark. Stivo builds upon language virtualization [20] and lightweight modular staging [23] (LMS) and has the same syntax and semantics as the regular Scala language with only a few restrictions. Because Stivo includes common compiler and relational optimizations, as well as domain specific ones, it produces very fast programs. It is designed in a very modular way - the code generation is completely independent of the parsing and optimizations - and allows extensions for supported operations, optimizations and backends.

Stivo makes following contributions to the state of the art:

- We implement the Stivo framework for big data processing. Stivo has a high level programming model with carefully chosen restrictions that allow relational, domain specific and compiler optimizations but do not sacrifice generality.
- We introduce a novel projection insertion algorithm that operates across general program constructs like classes, conditionals, loops and user defined functions, takes the whole program into account and does not rely on safe assumptions which lead to missed optimization opportunities.
- We show that Stivo allows easy language extension and code portability for big data frameworks.

In section 2 we provide background on LMS, language virtualization and big data frameworks. Then, in section 3 we explain the programming model and present simple program examples. In section 4 we explain the novel projection insertion optimization algorithm in section 4.1 and section 4.2 explains the fusion optimization. We evaluate Stivo in 5 and discuss our approach in section 6. Stivo is compared to state of the art in section 7, future work is in section 8 and we conclude in section 9.

2. BACKGROUND

2.1 Virtualized Scala

Stivo is written in an experimental version of Scala called Virtualized Scala [20] which provides facilities for deep embedding of domain specific languages (DSLs). Deep embedding is achieved by translating regular language constructs like conditionals, loops, variable declarations and pattern matching to regular method calls. For example, for the code `if (c) a else b`, the conditional is not executed but instead a method call is issued to the overrideable method `__ifThenElse(c, a, b)`. In case of deeply embedded DSLs it is used for creation of an intermediate representation (IR) node that represents the `if` statement.

In Virtualized Scala, all embedded DSLs are written within DSL scopes. These special scopes look like method invocations that take one by name parameter (block of Scala code). They get translated to the complete specification of DSL modules that are used in a form of a Scala trait mix-in composition[1]. For example: `stivoDSL{ \\ dsl code }` gets translated into: `new StivoDSL { def main(){...}}` This makes all the DSL functionality defined in StivoDSL visible in the body of the by name parameter passed to Stivo method. Although modified, Virtualized Scala is fully binary compatible with Scala and can use all existing libraries.

2.2 Lightweight Modular Staging

Stivo is built upon the Lightweight Modular Staging (LMS) library [23, 24]. LMS utilizes facilities provided by Virtualized Scala to build a modular compiler infrastructure for developing staged DSLs. It represents types in a DSL with polymorphic abstract data type `Rep[T]`. A term inside a DSL scope that has a type `Rep[T]` declares that once the code is staged, optimized, and generated, the actual result of the term will have type `T`. Since `Rep[T]` is an abstract type, each DSL module can specify concrete operations on it, which are used for building the DSL's intermediate representation.

Since Scala's type system supports type inference and implicit conversions, most of the `Rep[T]` types are hidden from the DSL user. This makes the DSL code free of type information and makes the user almost unaware of the `Rep` types. The only situation where `Rep` types become are visible is in parameters of methods and fields of defined classes. In our experience writing DSLs, `Rep` types do not present a problem but gathering precise and unbiased information on this topic is very difficult.

The modular design of LMS allows the DSL developer to arbitrarily compose the interface, optimizations and code generation of the DSL. LMS provides implementations for most constructs the Scala language and the most common libraries, so that the DSL can be very close to normal Scala. Also, common data parallel patterns are available as modules in project Delite [25]. Module inclusion is simply done by mixing Scala traits together. The correctness of the composition and missing dependencies are checked by the type system. Code generation for a DSL is also modular, so the effort is almost completely spent on domain specific aspects.

Unlike Scala, which does not have an effect tracking mechanism, LMS provides precise information about the effect for each available operation. The DSL developer needs to explicitly specify the effects for each DSL operation he introduces. The LMS effect tracking system then calculates the effects summary for each basic block in the DSL code. This allows the optimizer to apply code motion on the pure (side effect free) parts of the code. All implementations for standard library constructs that LMS provides such as strings, arrays, loops and conditionals, already include effect tracking.

LMS builds a complete intermediate representation of the code, optimizes it, and then generates optimized code for the chosen target language. The generated code has then to be compiled itself, and then it can be invoked. If the

[1]Scala's support for multiple inheritance

```
def parse(st: Rep[String]) = {
  val sp = st.split("\\s")
  Complex(Float(sp(0)), Float(sp(1)))
}
val x = new Array[Complex](input.size)
for (i <- 0 to input.size) {
  x(i) = parse(input(i))
}
for (i <- 0 to x.size) {
  if (x(i).im == 0) println(x(i).re)
}
```

(a) Original program

```
val size = input.size
val x = new Array[Complex](size)
for (i <- 0 to size) {
  val pattern = new Pattern("\\s")
  val sp = pattern.split(input(i))
  x(i) = Complex(Float(sp(0)), Float(sp(1)))
}
for (i <- 0 to x.size) {
  if (x(i).im == 0) println(x(i).re)
}
```

(b) CSE and inlining

```
val size = input.size
val re = new Array[Float](size)
val im = new Array[Float](size)
for (i <- 0 to size) {
  val pattern = new Pattern("\\s")
  val sp = pattern.split(input(i))
  re(i)  =  Float(sp(0)),
  im(i) = Float(sp(1))
}
for (i <- 0 to size) {
  if (x(i).im == 0) println(x(i).re)
}
```

(c) AoS → SoA

```
val size = input.size
val pattern = new Pattern("\\s")
for (i <- 0 to size) {
  val sp = pattern.split(input(i))
  val im = Float(sp(1))
  if (im == 0)  {
    val re = Float(sp(2))
    println(re)
  }
}
```

(d) Loop fusion and code motion

Figure 1: Step by step optimizations in LMS

```
// creates the println statement in the IR
trait PrintlnExp extends BaseExp {
  def println[T](st: Rep[String]) =
    reflectEffect(PrintlnNode(st))
}
trait PrintlnGen extends ScalaGen {
   def emit(node: Rep[Any]) = node match {
       case PrintlnNode(str) =>
          println("println("+str+")")
   }
}
```

Listing 1: Example of how the DSL module is specified. This module is used for measuring a performance of a block of code and can be reused in any other Scala backed DSL.

compilation delay is deemed inappropriate for a certain use case, it is possible to execute the code directly in Scala. A shallow DSL embedding can be achieved this way, instead of building the IR.

In listing 1, we show a simplified version of a reusable DSL module for printing. In trait `PrintlnExp` we define how the `println` operation is linked to the intermediate representation. The `reflectEffect` method defines that the `profile` method has global side effects which signals the compiler that it can not be reordered with respect to other globally effectful statements or be moved across control structures. In the `PrintlnGen` trait, we define how the code for `println` is generated for Scala.

LMS has been used successfully by Brown et al. for heterogeneous parallel computing in project Delite [6], Kossakowski et al. for a JavaScript DSL [13] and in the SIQ project for embedding queries into the Scala language.

2.3 Big Data Frameworks

Stivo generates Scala code for Crunch [2], Scoobi [5] and Spark [14]. Both Crunch and Scoobi use Hadoop as the execution engine and provide an MSCR implementation as presented in [8]. Crunch is implemented in Java and provides a rather low level interface, in which the user must provide implementation for user classes. Scoobi on the other hand is a Scala framework, which features a declarative high level interface and creates efficient serialization for user classes with only a minimal amount of help required.

Spark is a recent execution engine which makes better use of the cluster's memory, explicitly allowing the user to cache data. This allows huge speedups on iterative jobs which can reuse the same data multiple times, unlike with Hadoop. It also features a declarative high level interface and has support for multiple serialization frameworks and it also features a shell for low latency interactive data querying.

2.4 LMS Optimizations

When writing DSLs, the DSL author can exploit their domain knowledge to apply high level optimizations and program transformations. Afterwards, the program is usually lowered to a representation closer to the actual generated code. LMS provides a set of common optimizations for the lowered code, which are: common subexpression elimination (CSE), dead code elimination (DCE), constant folding (CF) and function inlining. LMS also applies code motion, which can either: *i)* move independent and side effect free blocks out of hot loops *ii)* move code segments that are used inside conditionals but defined outside, closer to their use site.

Another interesting optimization is the transformation of an array of structural types to a structure of arrays (AoS → SoA), each containing only primitive fields. This transformation removes unnecessary constructor invocations and

enables DCE to collect unused fields of an structure. It can be applied to built-in data structures like tuples as well as immutable user-defined types. It is similar in effect to row storage in databases and it gives great performance and memory footprint improvements.

LMS also provides a very general mechanism for operation fusion that uses standard loops as the basic abstraction. It is better than existing deforestation approaches since it generalizes to loops and can apply both vertical and horizontal fusion. In vertical fusion, the algorithm searches for producer consumer dependencies among loops, and then fuses their bodies together. In horizontal fusion, independent loops of the same shapes are fused together and index variables are relinked to the fused loop's index variable. Fusion greatly improves performance as it removes intermediate data structures and uncovers new opportunities for other optimizations.

All the above mentioned algorithms are repeated for program scopes at one level until a fixed point is reached. Then the whole cycle is applied to the next level of scopes, thereby optimizing the whole program. In listing 1, we present these optimizations on a single example which parses an array of complex numbers and prints only the real parts of them. Step 1a) shows the original program, 1b) shows how CSE extracts `size` and inlining replaces `parse` and `split` invocations with their bodies. In step 1c) the array `x` of complex numbers is split into two arrays of floating points. In 1d) the loops are fused together, which then allows code motion to move the constant pattern out of the loop and move the parsing of the real component into the conditional. The intermediate arrays can then be removed by DCE.

3. PROGRAMMING MODEL

The basic abstraction in our programming model is the interface `DList[T]`. `DList[T]` represents a distributed collection of elements that have type `S` which is a subtype of `T`. The elements of a `DList[T]` collection are immutable, so each operation on the list can only: *i)* produce a new `DList`, *ii)* save it to persistent storage, *iii)* materialize it on the master node or *iv)* return an aggregate value.

`DList` operations are presented in Table 1. In the left column, we show which frameworks support which method. The middle column shows the method name. Finally, the right column contains the type of `DList` that the operation is called on, and return type of the operation.

Operations `DList()` and `save()` are used for loading and storing data to the persistent storage. `map, filter` and `flatMap` are standard list comprehensions for transforming the data by applying the argument function and can also be used with Scala `for` comprehensions. Operations `groupByKey`, `join`, `cogroup`, `cross` and `reduce` are applicable only if the elements of `DList` form a key/value tuple. `reduce` is used for general aggregation after the `groupByKey`, `join, cogroup` and `cross` are different types of relational joins. `sort` sorts the dataset, `partitionBy` defines partitioning among machines, and `cache` signals that data should be kept in cluster memory for faster future accesses. Two `DList`s can be concatenated by the `++` operation. A DList can be materialized on the master node by calling `materialize()`.

```
val read = DList("hdfs://..." + input)
val parsed = read.map(WikiArticle.parse(_))
parsed.flatMap(_.split("\\s"))
  .map(x => (x, 1))
  .groupByKey
  .reduce(_ + _)
  .save("hdfs://..." + output)
```

Listing 2: Example of word count program where type inference removes the need to declare any `Rep` types.

Some methods accept functions as their parameters. Code within these functions can be either written in the Stivo DSL, or by using existing functions from an external library or common JVM libraries. Using JVM libraries requires just one extra line of code per method.

In listing 2, we show an implementation of a simple word count example, in which the code does not have any visible `Rep` types. Since a large subset of the Scala library is implemented as a DSL module, functions like `split` and string concatenation are used the same way as they are in Scala. In the second line, the regular (with arguments wrapped in `Rep`) method `parse` is passed to the `map` method. Pig and Hive do not have functions in their own language, but allow writing user defined functions in other languages which requires a considerable amount of boilerplate code.

All methods except for `cache` and `sort` can be mapped to methods in Scoobi, Spark and Crunch. Other back-ends (including Dryad) provide these primitives as well. The `cache` method currently works with Spark only but it can be added to the interface of other back-ends, in which it would have no effect, such that the code stays portable. From existing frameworks today only HaLoop[7] and Twister [12] can benefit from it, however we did not implement code generation for them. Method `sort` is inconsistent in most of the frameworks so we have not mapped uniformly to all of them. However, with slight modifications to the framework implementations it could be supported as well. `sort` can also be implemented in Stivo itself by using `takeSample` and `partitionBy`.

4. OPTIMIZATIONS

In this section we present the main optimizations implemented in Stivo.

4.1 Projection Insertion

A common optimization in data processing is to remove intermediate values early that are not needed in later phases of the computation. It has been implemented in relational databases for a long time, and has recently been added to the Pig framework. This optimization requires all field accesses in the program to be explicit. A library can provide this, but its usage is more intrusive than if the framework can use compiler support.

In Stivo we support this optimization for algebraic data types, more specifically final immutable Scala classes with a finite level of nesting. Our approach does not require special syntax or access operators and supports method declara-

Framework	Operation	Transformation
All	`DList(uri: Rep[String])`	`String => DList[T]`
	`save(uri: Rep[String])`	`DList[T] => Unit`
	`map(f: Rep[T] => Rep[U])`	`DList[T] => DList[U]`
	`filter(f: Rep[T] => Rep[Boolean])`	`DList[T] => DList[T]`
	`flatMap(f: Rep[T] => Rep[Iter[U]])`	`DList[T] => DList[U]`
	`groupByKey()`	`DList[(K, V)] => DList[(K, Iter[V])]`
	`reduce(f: (Rep[V], Rep[V]) => Rep[V])`	`DList[(K, Iter[V])] => DList[(K, V)]`
	`cogroup(right: Rep[DList[(K, W)]])`	`DList[(K, V)] => DList[(K, (Iter[K], Iter[W]))]`
	`join(right: Rep[DList[(K, W)]])`	`DList[(K, V)] => DList[(K, (V, W))]`
	`++(other: Rep[DList[T]])`	`DList[T] => DList[T]]`
	`partitionBy(p: Rep[Partitioner[T]])`	`DList[T] => DList[T]`
	`takeSample(p: Rep[Double])`	`DList[T] => Iter[T]`
	`materialize()`	`DList[T] => Iter[T]`
Spark	`cache()`	`DList[T] => DList[T]`
	`sort(cmp: Rep[Comparator[T]])`	`DList[T] => DList[T]`
Crunch	`sort(asc: Rep[Boolean])`	`DList[T] => DList[T]`

Table 1: DList operations and their framework support. For clarity reasons, `Iter` represents the Scala `Iterable` and `Rep[_]` types in the rightmost column are omitted.

tions on data types just like methods of regular Scala classes. While implementing our benchmarks we found this to be a reasonably expressive model for big data programming. The DSL user needs to supply class declarations, from which we generate all the necessary code for its use in Stivo. In these cases, LMS describes all field accesses explicitly and we can generate highly specialized code for these types including serialization schemes and other glue code for the back-ends we support.

In section 2.4 we have shown how LMS optimizes these classes within the same program scope. However, in Stivo we endorse a declarative programming model with many short functions - each having its own scope - as these are easier to read. Each operation in our API has an input type and an output type, and they may have a parameter accepting a closure. Operations are chained together to form a data flow graph without cycles. Since types support nesting, we need to define the live values between operations as the sequence of field dereferences respective to a type to access that value. This sequence of field dereferences we call a path, and to represent it we use the scala expression needed to access it on such a type. In figure 2 we visualize the tree which contains all paths and fields for the nested type of t in the example code in 3.

For our projection insertion algorithm we need to analyze the paths on each edge of the data flow graph. We can only compute the paths for one operation when the union of all paths accessed by its successors have been already computed. For this reason we visit the graph in reverse topologically sorted order and process the operations one by one. When the successors of a node have all been processed and their paths have been propagated, that node can be analyzed to compile a list of paths it needs from its predecessors. The result of this analysis can then be used to insert projections before any operation which serializes objects or stores them in memory - we call them barriers - for example `cache`, `groupByKey` or `join`.

The algorithm depends on the analysis of a single operation, which we implemented using following primitives:

```
case class A(id: String, b: B)
case class B(id: String)
val t = ("tuple", A("a", B("b")))
t: scala.Tuple2[String, A]
```

Listing 3: Nested type for paths example.

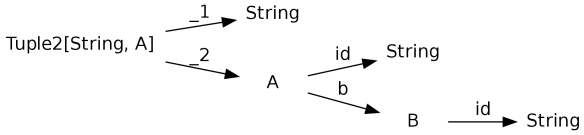

Figure 2: Visualization of the nested type Tuple2[String, A]. The node labels are the types at that path, the edges describe one field dereference. The leafs are always primitive types, and the path to them is formed by concatenating the edge labels. An example path description is "_2.b.id"

- *Paths for type*: Given a type and optionally a path, this primitive creates paths for all the nested fields within. In 3 a `save` with element type A returns the paths `id` and `b.id`.
- *Closure analysis*: This primitive returns a list of all paths accessed in a closure that read (recursively) from the closure's input.
- *Rewrite paths*: Several operations have defined semantics which influence the type and therefore the paths. For example, the `cache` operation will always have the same input and output type and is known to return the same instance, so all paths from it's successors must be propagated to its predecessors. The `groupByKey` operation on the other hand always reads all parts of the key, and has to rewrite all paths of the form `_2.iterable.x` to `_2.x`, as the iterable is introduced by the operation itself and is known to conserve the instances.
- *Narrow closure*: Given a list accessed paths and a closure, this primitive replaces the closure's original output symbol with one that reads from it and creates a new object only containing the fields needed for later stages.

operation	Propagate accessed paths	barrier
filter	All of successor + closure reads	
flatMap	All of the closure with a replaced output	
map	All of the closure with a replaced output	
join	All paths for the key, rewrites accesses to values to the correct predecessor	x
groupByKey	All paths for the key, rewrites accesses to the value's iterable to the value itself	x
reduce	All accesses from the closure are translated to access of the value's iterable	
save	All paths for input type	

Table 2: Access path computation and propagation for selected operations.

For `map` operations which output a nested type we need to combine the narrow closure and the closure analysis primitive. AoS → SoA ensures that the output symbol of a closure is always a constructor invocation for these. We then use the narrow closure primitive to create a new closure, in which the output symbol reads from the old output symbol. LMS recognizes a field read that reads from a constructor invocation in the same scope and optimizes this by reading the value that was used to initialize the field directly. This happens for all the fields, if the corresponding constructor invocation is in the same scope. Therefore the old constructor invocation will not be read anymore, and DCE will pick it up. This means that the field values only it was reading will also not be read anymore, and they too will be eliminated. Then we can analyze this new closure to get the accessed paths from it.

Table 2 shows how these primitives are combined to form the rules for the most important operations in our API.

4.2 Operation Fusion

In section 3 we have shown that the `DList` provides declarative higher order operations. Anonymous functions passed to these operations do not share the same scope of variables. This reduces the number of opportunities for optimizations described in 2.4. Moreover, each piece of data needs to be read, passed to and returned from the higher order function. In both the push data-flow model and the pull model this enforces virtual method calls [21] for each data record. To overcome this performance penalty we have implemented fusion of operations `map`, `flatMap` and `filter` through the underlying loop fusion algorithm described in section 2.4.

The loop fusion optimization described in section 2.4 supports horizontal and vertical fusion of loops as well as fusion of nested loops. Also, it provides a very simple interface to the DSL developer for specifying loop dependencies and for writing fusable loops. We decided to extend the existing mechanism to the `DList` operations although they are not strictly loops. We could have taken the path of Murray et al. in project Steno [21] by generating an intermediate language which can be used for simple fusion and code generation. Also, we could use the Coutts et al. [10] approach of converting `DList` to streams and applying equational transformation to remove intermediate results. After implementing the algorithm by reusing loop fusion we are confident that it required significantly less effort than reimplementing existing approaches.

$$
\begin{aligned}
&out = map(n, op) \rightarrow loop(shape\_dep(n), \{ \\
&\quad yield(out, op(iterator\_value(n)) \\
&\}) \\
&out = filter(map, op) \rightarrow loop(shape\_dep(n), \{ \\
&\quad if(op(iterator\_value(in)) \\
&\qquad yield(out, iterator\_value(in)) \\
&\}) \\
&out = in.flatMap(op) \rightarrow loop(shape\_dep(n), \{ \\
&\quad w = op(iterator\_value(in)) \\
&\quad loop(w.size)\{yield(out, w(i))\} \\
&\})
\end{aligned}
$$

Listing 4: Lowering transformations.

Before fusion optimization, the program IR represents an almost one to one mapping to the operations in the programming model. Each operation is represented by the corresponding IR node which carries its data and control dependencies. On these IR nodes we first apply a lowering transformation which maps operations `map`, `flatMap` and `filter` to the corresponding loop representation. Described transformation is achieved by the program translation described in 4. These rules introduce two new nodes: *i)* $shape_dep$ that takes the place of the loop shape variable and carries the explicit information about its vertical predecessor and *ii)* $iterator_value$ that represents reading from an iterator of the preceding `DList`. The `yield` operation represents storing to the successor collection and is afterwards replaced by the bodies of fused loops.

After the lowering transformation the loop fusion is applied. It vertically fuses pairs of loops until a fixed point is reached. In each fusion iteration all other LMS optimizations are applied as well. To avoid generating actual `while` loops we include a modified loop generation module for every back-end. This module emits the most general operation (equivalent of the Hadoop `Mapper` class) that the framework provides. With this approach we could also generate code directly for Hadoop MapReduce which would result in a single highly optimized loop per `Mapper`. After prototype experiments we concluded that the gain is not significant compared to using higher level back-ends. Therefore, as an alternative, we used Scoobi and Crunch.

Unlike MapReduce based back-ends, Spark's design uses the pull data-flow model, implemented through iterators. Generating code for the pull data-flow model from the loop based (push data-flow) model proved to be non-trivial. After evaluating different types of queues and array buffers we have decided to buffer intermediate results in a 4 MB array.

5. EVALUATION

We evaluate Stivo optimizations by comparing the performance of three programs: *i)* A word count with prior text parsing, *ii)* TPCH [9] query 12 and a *iii)* k-means application. To evaluate the extensibility of the framework we introduce a new DSL component that represents vectors in the k-means benchmark.

All experiments were performed on the Amazon EC2 Cloud,

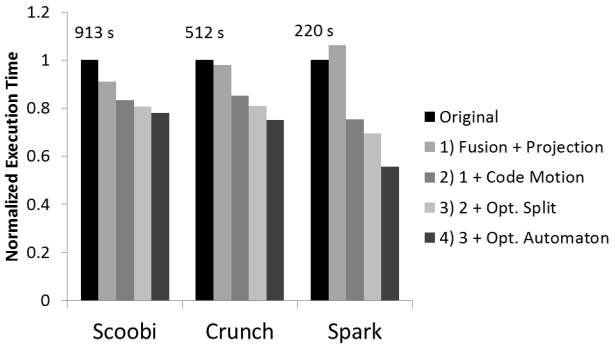

Figure 3: Word Count benchmark.

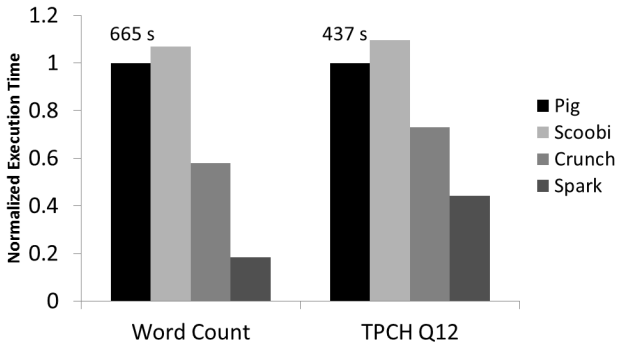

Figure 4: Comparison between Pig, Scoobi, Spark and Crunch.

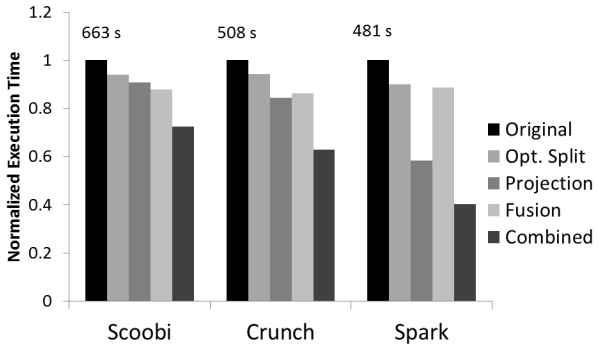

Figure 5: TPCH query 12 benchmark.

using 20 "m1.large" nodes as slaves and one as a master. They each have 7.5 GB of memory, 2 virtual cores with 2 EC2 compute units each, 850 GB of instance storage distributed over 2 physical hard drives and have 1 Gb/s network interface. Prior to the experiments we have measured up to 50 MB/s between two nodes. For the Hadoop experiments we used the cdh3u4 Cloudera Hadoop distribution. On top of it we used Crunch version 0.2.4 and Scoobi 0.4.0. We did not tweak Hadoop configuration beyond the default settings set by the Whirr 0.7.1 [1] tool. For benchmarking Spark we used the Mesos [15] EC2 script to start a cluster, and the most recent version of Spark for our tests. For Spark we changed the default parallelism level to the number of cores in the cluster and increased the maximum memory to 6GB.

While doing preliminary benchmarking we found some easy tweaks focused on regular expressions that we needed to include in Stivo in order to have a fair comparison against Pig, which contains them. We implemented a fast splitter, which uses an efficient character comparison whenever the regular expression allows this. Additionally we select based on the regular expression between Java's implementation and the dk.brics.automaton library [19].

For serialization of data we used LMS code generation to achieve minimal overhead for both Crunch and Scoobi frameworks because they outperformed the Kryo [4] library by a small margin. For Spark we used the standard serialization mode which uses Kryo. All benchmarks were run three times and in the figures we present the average value. We also computed the standard deviations but we omitted them since they are smaller than 3% in all the experiments.

We made the Stivo code, as well as the generated code, available on `https://github.com/stivo/Distributed`.

5.1 Parsing and Word Count

In this benchmark we evaluate the performance of Stivo compiler optimizations without focusing on projection insertion. We choose a word count application that, prior to the inexpensive network shuffle, parses the input with 5 regular expressions making this job CPU bound. For this evaluation we start with an the original version of the program and add optimizations one by one. We first add the operation fusion and projection insertion optimizations. We then include code motion that removes regular expression compilation out of hot loops. Next we add the fast splitter and for the fully optimized version we use the optimized automaton regular expression library.

Our input is a 62 GB set of plain text version of Freebase Wikipedia articles. Our regular expressions are used clean up articles from strings that are not text words. This benchmark does not benefit from projection insertion but we include it for the comparison with the Pig framework in subsection 5.3.

In figure 3 we show the job times for these versions normalized to the original program version. Performance improvements of all optimizations combined are from 29% for Scoobi, 33% for Crunch and 79% for Spark. The base performance of the frameworks differ by a large margin for this benchmark. Scoobi profits the most in this case from the fusion which indicates that the framework imposes additional overhead for declarative operations. In Spark, we notice larger benefits from our optimizations. We argue that it has significantly smaller IO overhead so that the optimizations have a bigger impact. Also, we notice that fusion optimization with projection insertion is slower than the original program for Spark. This result does not match our experiments in a smaller cluster setup, we believe that it could be caused by a straggler node in the cloud environment.

5.2 TPCH Query 12

This benchmark evaluates all optimizations combined but emphasizes the projection insertion. We chose the TPCH query 12 which includes an expensive `join` operation after which only two columns of the original data are used, thus giving projection insertion opportunity to eliminate unused columns. We compare the original program to each optimizations separately and all of them combined. As the data set we use a 100 GB plain text input generated by the DbGen tool [9].

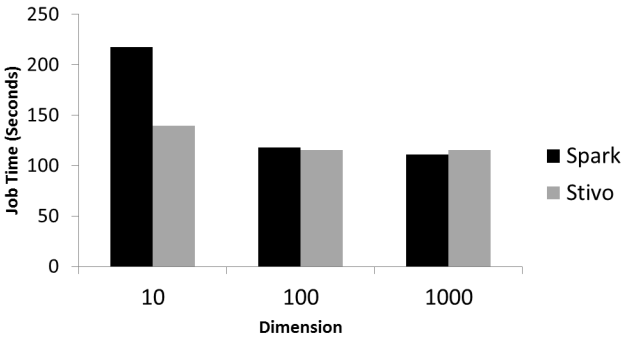

Figure 6: K-means benchmark.

In figure 5 we show job times for different optimizations normalized to the original program version on different frameworks. We notice that projection insertion gives 20% percent better performance on Crunch and 11% on Scoobi. On Spark, the projection insertion improves the performance by 40%, significantly more than for the Hadoop based frameworks. We believe the difference is caused by Hadoop spills of the data to disk earlier, while Spark tries to keep it in the memory before it spills it. Spark is therefore very sensitive to memory overhead of the large `join`.

In this benchmark the optimizations interact with each other. The absolute performance gain for combined optimizations is 3% greater for Crunch, equal for Scoobi and 9% smaller for Spark than the sum of the absolute individual gains.

5.3 Comparison with Pig

In figure 4 we compare the most optimal versions of benchmarks to equivalent Pig programs. The figure is normalized to the Pig execution time and overall job time is stated above the bar. We notice that for TPCH query 12 the combination of fusion, code motion and field reduction outperform Pig when the Crunch framework is used.

For the sake of showing comparison between the Hadoop based frameworks and the Spark framework we include the Spark results in the graph. We see that in all cases except for unoptimized TPCH query 12 it significantly outperforms the Hadoop based frameworks.

In the word count benchmark Crunch outperforms Pig even without any optimizations applied. We believe that Pig uses inefficiency string processing primitives. With all optimizations Crunch is 73% faster than Pig. We explain this by the more optimal regular expressions processing support included in the Stivo. In regular expressions used in the benchmark Pig falls back to default Java regular expressions while Stivo uses optimized automaton library. Scoobi framework performs slower than Pig in both benchmarks even with all optimizations applied.

5.4 Extensibility and Modularity

To evaluate modularity and extensibility of Stivo we decided to extend with a `Vector` abstraction that has abstract methods for arithmetic vector operations that get compiled into loops over arrays. We also chose this benchmark to emphasize the extensibility of Stivo.

We took a version of Spark k-means program [14] application and ported it to Stivo. This application can neither benefits from projection insertion reduction nor from operation fusion. We extended our DSL for this program with a highly optimized vector type that has all its operations iterating over the dimensions compiled into while loops. We only evaluate this benchmark on Spark, since it uses operations only defined in Spark and since it is known to outperform Hadoop by a large margin. As input we use synthetic data with 10 to 1000 dimensions, 100 centers and we keep the $dimensions * points$ factor constant so that each input file is around 20Gb.

Our results are similar to those described by Murray et al. in [21]. In lower dimensions our optimization shows large speedup while for 1000 dimensions our version performs slightly worse. We believe that the iterator overhead is quite high in case of 10 dimensions, such that our loops which removes it performs much better. At higher dimensions it's possible that the JVM can do a better job optimizing if the code is smaller, such that our pre optimized and larger code becomes slightly slower. In any case our implementation seems favorable as it performs more consistently for different dimensions.

6. DISCUSSION

The language we provide is the same as Scala in its basic constructs, however it does not support all of the functionalities. The following functionalities are not available:

- Projection optimization can only be applied to final immutable classes. This somewhat limits the language, but in large big data processing, data records are often not polymorphic.
- Polymorphism is supported only in a limited form. The limitation is that all possible implementations need to be known at staging time and currently it prevents optimization of the polymorphic method call.
- The whole Scala library is not available in its DSL form. This however does not limit the user since JVM methods can be used in the native form, but they can not be optimized. Due to language embedding, for using JVM methods there is no boilerplate code.

One of the caveats of the staged DSL approach is that the program staging, compilation, generation and compilation of the generated code increases the startup time for the task. For the benchmarks we have evaluated that this process takes from 4 to 14 seconds. Although this can be significant, it needs to be only done once on a single machine so we believe it is not a limiting factor for batch jobs.

The only case where compile time becomes relevant is with back-ends that support interactive data analytics, like the Spark framework. Spending more than a couple of seconds for compilation would affect the interactivity.

We see two ways to overcome issues with delay in execution:

- We can implement a version which does not generate the IR, but executes the original code straight away.

In this case all optimizations we perform would be disabled but the user would gain original Spark interactivity. This feature is not implemented in Stivo, but Kossakowski et al. [13] have done this for the Javascript DSL.

- If optimizations are required, we can build and optimize the intermediate representation after each user input. This gives the compiler time to do the work while the last command is being typed. The overall delay in this case the delay would be $delay = IR\_building - user\_delay) + generated\_code\_compilation$. Since we did not optimize the compiler and do not have data about the time it takes to do interactive commands we can not speculate on the final result.

Each job requires a framework specific configuration for its optimal execution (e. g. the number of mappers, reducers, buffer sizes etc.). Our current API does not include tuning of these parameters, but in the future work we want to introduce a configuration part of the DSL to unify configuration of different backends. With the current programming model it is not possible to tune these parameters in a unified way.

7. RELATED WORK

Pig [22] is a framework for writing jobs for the Hadoop platform using an imperative domain specific language called Pig latin. Pig latin's restricted interface allows the Pig system to apply relational optimizations that include operator rewrites, early projection and early filtering to achieve good performance. It has extensive support for relational operations and allows the user to choose between different join implementations with varying performance characteristics. Pig latin users need to learn a new language which is not the case with frameworks such as Hadoop, Hive and Crunch. It does not include user defined functions, the user needs to define them externally in another language, which will often prevent optimizations as Pig can not analyze those. Pig latin is not Turing complete as it does not include control structures itself. The language is not statically type checked so runtime failures are common and time consuming. Also, pure Java approaches benefit from a rich ecosystem of productivity tools.

Even though Stivo also adopts a domain specific approach, it is deeply embedded in Scala, Turing complete and allows the user to easily define functions which do not disable the optimizations. Currently Stivo does not have support for many relational optimizations. However, it includes compiler optimizations and it is well extensible.

Steno [21] is a .NET library that, through runtime code generation, effectively removes abstraction overhead of the LINQ programming model. It removes all iterator calls inside LINQ queries and provides significant performance gains in CPU intensive jobs on Dryad/LINQ. Queries that use Steno do not limit the generality of the programming model but optimizations like code motion and early projection are not possible. Stivo also removes excess iterator calls from the code but during operation fusion it enables other optimizations, especially when combined with early projection. The drawback of Stivo is that it is slightly limited in generality.

Manimal [18] and HadoopToSQL [17] perform static byte code analysis on Hadoop jobs to infer different program properties that can be mapped to relational optimizations. They both use the inferred program properties to build indexes and achieve much more efficient data access patterns. Manimal can additionally organize the data into columnar storage. These approaches are limited by the incomplete program knowledge which is lost by compilation and runtime determined functions. They both do not restrict the programming model at all. Stivo shares the idea of providing code generality to these approaches. However, it currently does not include data indexing schemes which could enable big performance improvements. We believe that the full type and program information available in Stivo will enable us to build better data indexing schemes for a larger set of user programs.

8. FUTURE WORK

From the wide range of relational optimizations, Stivo currently supports only early projection. In the future work we plan to introduce early filtering which will push filter operations before the expensive operators that require a barrier. Also, we plan to include program analysis phase which will allow building of indexes.

Text and XML processing are often processed in cluster computing and efficient processing of it can greatly reduce cost and energy consumption. With that in mind we plan to integrate Stivo with other text parsing DSLs that are deeply embedded into the standard library. If prototyping shows that performance gains are significant we will add DSL modules for regular expressions, Scala parser combinators and XML library.

Stivo currently only operates on distributed datasets so programs written in it can not be used for in memory data. We plan to integrate Stivo with Delite [6] collections DSL which supports very efficient execution of batch operations. Delite also allows running queries on heterogeneous hardware architectures where jobs are scheduled for execution on both CPU and GPU processors.

9. CONCLUSION

We have presented the big data processing library Stivo that provides an expressive high level programming model. Stivo uses language virtualization, lightweight modular staging and side effect tracking to analyze user programs at runtime. This allows Stivo to apply projection insertion, code motion as well as operation fusion optimizations to achieve high performance for declarative programs. Through modular code generation Stivo allows execution on Spark, Crunch and Scoobi. Presented optimizations result in speedups of 148% in Spark, 59% in Crunch and 38% in Scoobi.

Unlike existing domain specific approaches Stivo provides high performance, general and expressive programming model which is integrated into the Scala language. It allows high performance user extensions and code portability between different big data processing frameworks.

10. REFERENCES

[1] Apache whirr, http://whirr.apache.org/.

[2] The crunch framework, https://github.com/cloudera/crunch.

[3] The hadoop framework, http://hadoop.apache.org/.

[4] Kryo: Fast, efficient java serialization and cloning, http://code.google.com/p/kryo/.

[5] The scoobi framework, https://github.com/nicta/scoobi.

[6] K. J. Brown, A. K. Sujeeth, H. J. Lee, T. Rompf, H. Chafi, M. Odersky, and K. Olukotun. A heterogeneous parallel framework for domain-specific languages. In *Parallel Architectures and Compilation Techniques (PACT), 2011 International Conference on*, page 89–100, 2011.

[7] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst. HaLoop: efficient iterative data processing on large clusters. *Proceedings of the VLDB Endowment*, 3(1-2):285–296, 2010.

[8] C. Chambers, A. Raniwala, F. Perry, S. Adams, R. R. Henry, R. Bradshaw, and N. Weizenbaum. FlumeJava: easy, efficient data-parallel pipelines. *ACM SIGPLAN Notices*, 45(6):363–375, 2010.

[9] T. P. Council. Tpc benchmark$^{\text{TM}}$, http://www.tpc.org/tpch/.

[10] D. Coutts, R. Leshchinskiy, and D. Stewart. Stream fusion: From lists to streams to nothing at all. In *ACM SIGPLAN Notices*, volume 42, page 315–326, 2007.

[11] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[12] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. Bae, J. Qiu, and G. Fox. Twister: a runtime for iterative MapReduce. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, HPDC '10, page 810–818, New York, NY, USA, 2010. ACM.

[13] G. K. et al. Javascript as an embedded dsl. In *To appear in ECOOP*, 2012.

[14] M. Z. et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of NSDI*, 2012.

[15] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. Katz, S. Shenker, and I. Stoica. Mesos: A platform for fine-grained resource sharing in the data center. In *Proceedings of NSDI*, 2011.

[16] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, page 59–72, 2007.

[17] M. Iu and W. Zwaenepoel. HadoopToSQL: a mapReduce query optimizer. In *Proceedings of the 5th European conference on Computer systems*, EuroSys '10, page 251–264, New York, NY, USA, 2010. ACM.

[18] E. Jahani, M. J. Cafarella, and C. Ré. Automatic optimization for MapReduce programs. *Proc. VLDB Endow.*, 4(6):385–396, Mar. 2011.

[19] A. Møller. dk. brics. automaton–finite-state automata and regular expressions for java, 2005.

[20] A. Moors, T. Rompf, P. Haller, and M. Odersky. Scala-virtualized. In *Proceedings of the ACM SIGPLAN 2012 workshop on Partial evaluation and program manipulation*, page 117–120, 2012.

[21] D. G. Murray, M. Isard, and Y. Yu. Steno: automatic optimization of declarative queries. In *Proceedings of the 32nd ACM SIGPLAN conference on Programming language design and implementation*, PLDI '11, page 121–131, New York, NY, USA, 2011. ACM.

[22] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, page 1099–1110, 2008.

[23] T. Rompf and M. Odersky. Lightweight modular staging: a pragmatic approach to runtime code generation and compiled DSLs. In *Proceedings of the ninth international conference on Generative programming and component engineering*, page 127–136, 2010.

[24] T. Rompf and M. Odersky. Lightweight modular staging: a pragmatic approach to runtime code generation and compiled DSLs. *Communications of the ACM*, 55(6):121–130, 2012.

[25] T. Rompf, A. Sujeeth, H. Lee, K. Brown, H. Chafi, M. Odersky, and K. Olukotun. Building-blocks for performance oriented dsls. *Arxiv preprint arXiv:1109.0778*, 2011.

[26] A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy. Hive - a petabyte scale data warehouse using hadoop. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 996 –1005, Mar. 2010.

[27] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. Gunda, and J. Currey. DryadLINQ: a system for general-purpose distributed data-parallel computing using a high-level language. In *Proceedings of the 8th USENIX conference on Operating systems design and implementation*, page 1–14, 2008.