An Intermediate Representation for Optimizing Machine Learning Pipelines

Kunft, Andreas; Katsifodimos, Asterios; Schelter, Sebastian; Bress, Sebastian; Rabl, Tilmann; Markl, Volker

DOI
10.14778/3342263.3342633

Publication date
2019

Document Version
Final published version

Published in
Proceedings of the VLDB Endowment

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.
An Intermediate Representation for Optimizing Machine Learning Pipelines

Andreas Kunft*, Asterios Katsifodimos**, Sebastian Schelter†, Sebastian Breß**, Tilmann Rabl**, Volker Markl**

*TU Berlin  **Delft University of Technology  †New York University  ‡DFKI  ††HPI, Universität Potsdam

ABSTRACT
Machine learning (ML) pipelines for model training and validation typically include preprocessing, such as data cleaning and feature engineering, prior to training an ML model. Preprocessing combines relational algebra and user-defined functions (UDFs), while model training uses iterations and linear algebra. Current systems are tailored to either of the two. As a consequence, preprocessing and ML steps are optimized in isolation. To enable holistic optimization of ML training pipelines, we present Lara, a declarative domain-specific language for collections and matrices. Lara’s intermediate representation (IR) reflects on the complete program, i.e., UDFs, control flow, and both data types. Two views on the IR enable diverse optimizations. Monads enable operator pushdown and fusion across type and loop boundaries. Combinators provide the semantics of domain-specific operators and optimize data access and cross-validation of ML algorithms. Our experiments on preprocessing pipelines and selected ML algorithms show the effects of our proposed optimizations on dense and sparse data, which achieve speedups of up to an order of magnitude.

PVLDB Reference Format:
DOI: https://doi.org/10.14778/3342263.3342633

1. INTRODUCTION
Modern data analysis pipelines often include preprocessing steps, such as data cleaning and feature transformation, as well as feature engineering and selection [62, 63, 40, 14, 9]. Once data is in an appropriate shape, machine learning models are trained and evaluated. These training and model evaluation cycles are repeated several times to find the most suitable configuration of different features, machine learning (ML) algorithms, and hyperparameters. To build such training pipelines, data scientists can choose from a variety of tools and languages. Python and R offer popular libraries that are easy to use and provide fast development cycles. These libraries are embedded shallowly [30] in the host language, i.e., they are executed as-is, without any inter-library optimizations and support for large data [53]. General purpose dataflow systems [71, 2] provide second-order functions (e.g., map and reduce) to transform collections via user-defined functions (UDFs). They defer program execution by providing a type-based domain-specific language (DSL) that builds an associated operator graph. This operator graph is optimized and executed on a dedicated dataflow engine. In order to develop ML algorithms in such DSLs, linear algebra operations have to be retrofitted as UDFs on (distributed) collections. As a result, ML algorithms are hardcoded by experts and provided as library functions with fixed data representations and execution strategies. Thus, the semantics of linear algebra operations are concealed behind UDFs, which are treated as black boxes by optimizers [36]. In contrast, dedicated systems for ML, such as SystemML [11] and Tensorflow [1] provide linear algebra operations. However, it is difficult to express pipelines that include preprocessing and data transformation in these systems, as they lack dedicated types for collection processing. In summary, dedicated systems with type-based DSLs provide advantages over shallowly embedded libraries, but still suffer from three major problems in the context of end-to-end pipelines for model training: (i) Development, maintenance, and debugging of end-to-end pipelines is a tedious process in dedicated systems, and limits optimization potential and efficient execution. (ii) Preprocessing and ML are often executed in different systems in practice [60], which prevents optimizations across linear and relational algebra. (iii) Neither shallowly embedded libraries nor type-based DSLs can reason about native UDFs and control flow.

To address these issues, we propose Lara, a DSL that combines collection processing and machine learning.1 Lara is based on Emma [4, 5], a quotation-based DSL [51] for (distributed) collection processing. Emma’s DataBag algebraic data type enables declarative program specification based on for-comprehensions, a native language construct in Scala. In contrast to type-based DSLs, quotation provides access to the abstract syntax tree (AST) of the whole program, allowing us to inspect and rewrite UDFs and native control flow. Emma’s intermediate representation (IR) is based on monad comprehensions [34] and enables operator fusion and implicit caching. Lara extends Emma with Matrix and Vector data types in its API and IR to execute pipelines for model training on single machines. It provides

1A preliminary short version has been published before [42].
two views on the IR to perform diverse optimizations: the monadic view represents operations on both types, DataBag and Matrix, as monad comprehensions in the IR. This common representation enables operator fusion and pushdown of UDFs across type boundaries, e.g., filter pushdown from a Matrix to a DataBag. Access to the control flow allows Lara to reason about operator fusion over loop boundaries, e.g., feature transformations that are iteratively applied over column ranges. The combinator view captures high-level semantics of relational and linear algebra operators as single entities in operator trees similar to relational algebra trees. It enables data dependent selection of specialized physical operators and implicit data layout conversions based on interesting properties [31]: similar to interesting properties of relational operators, e.g., sorted data for joins, linear algebra operators have preferred data access pattern, e.g., row-wise or column-wise.

In summary, we make the following contributions:
1. We propose Lara, a quotation-based DSL for end-to-end model training pipelines (Section 3).
2. We discuss our IR, which has access to the whole AST of the pipeline, and two views on top of it: a view based on monad comprehensions and a view based on the high-level semantics of operators (Section 3 and 4).
3. We discuss the extensibility of our approach by introducing a custom high-level operator and optimizations for k-fold cross-validation, a widely used technique to select hyperparameters for ML models (Section 4.4).
4. We conduct experiments on a typical preprocessing pipeline, and show the effects of data layout and cross-validation optimizations on selected ML algorithms for dense and sparse data. The experiments achieve speedups of up to an order of magnitude (Section 5).

2. BACKGROUND

In this section, we provide a brief overview of approaches to implement domain-specific languages. We refer to Alexandrov et al [3] for an in-depth discussion and comparison.

Shallowly-Embedded Libraries. Many general-purpose programming languages (GPLs) provide dedicated libraries for different domains, such as collection processing and ML, e.g., Python's Pandas [49] and scikit-learn [57]. These libraries are embedded shallowly [30] in the host language (e.g., Python) and executed as is, i.e., without further optimizations except for those hard-coded in the algorithms and performed by the interpreter or compiler of the GPL.

Type-Based DSLs. Type-based or symbolic DSLs use the operations defined on a type (e.g., the RDD in Spark) to construct an operation graph, rather than directly executing the operations. Examples are the APIs of dataflow engines, such as Spark [71] and Flink [2], which mix the application of UDFs in second-order functions (e.g., map) and relational operators. More recent systems, such as MXNet [21] and Tensorflow [1], provide APIs based on tensors for ML. In contrast to shallowly embedded DSLs, type-based DSLs enable optimizations, such as choosing physical operator implementations and operator chaining. However, since the IR only reflects the operations defined on the type, host language UDFs are generally treated as black-boxes [36], and control flow is not visible. This can be partially addressed by specialized loop constructs [70, 26], but still prevents linguistic reuse [58] of the language's native control flow.

Standalone DSLs. Standalone DSLs, such as SQL and SystemML's DML [29] overcome the problems stated above. They provide a full-fledged compiler infrastructure for the DSL. However, a programmer has to develop this infrastructure (including libraries and development tools) from scratch, and support for complex UDFs written in GPLs is hard to achieve [56]. Another problem is user acceptance, i.e., to convince programmers to learn and adapt to the new language. Often, standalone DSLs try to ease the transition by staying close to the syntax of an existing language.

Quotation-Based DSLs. Quotation-based DSLs [51], such as LINQ [50], LMS [59], and Squid [55], reuse the syntax and type system of the host language and gives access to the program's AST during compilation and runtime. In contrast to type-based and shallowly-embedded DSLs, quotation gives access to the entire AST of the GPL, including types, white-box UDFs, and control flow. The AST can then be altered and optimized before execution. Thus, quotation-based DSLs overcome the limitations of the previously presented approaches but require careful design of domain-specific IRs.

3. LANGUAGE AND IR

In this section, we provide an overview of important design decisions for Lara, describe the IR of Lara, and introduce two views on top of the IR used to perform the diverse optimizations showcased in Section 4.

3.1 Language Design Decisions

We identify several shortcomings in current solutions, which led to the design of Lara. It extends the API and IR of Emma [4, 5], a quotation-based DSL for collection processing, with support for matrices and vectors. Users can express preprocessing and successive model training in the same program, which is reflected in a common IR. The following Lara code excerpt highlights these design decisions.

```
1  // Defining a vectorizeComment UDF
2  def vectorizeComment(c: Comment) = { /* UDF */ }
3  def vectorizeUser(u: User) = { /* UDF */ }  
4  optimize {
5    // Join "Comments" and "Users" and vectorize the result
6    val features = for {
7      c <- Comments // DataBag[Comment]
8    } yield vectorizeComment(c) ++ vectorizeUser(u)
9    val X = Matrix(features)
10   // Filter rows that have values > 10 in the third column
11   val M = X.forRows(row => row(2) > 10)
12   // Calculate the mean for each column
13   val means = M.forCols(col => mean(col))
14   // Compute the covariance matrix
15   val C = 1 / (M.nRows - 1) * X ** U.T
16 }
```

Lara enables the declarative specification of relational operators via Emma's DataBag data type with for-comprehensions. Line 6–9 illustrate a join between the datasets Users and Comments. ML pipelines are expressed as high-level linear algebra operators on the Matrix and Vector data type. For example, a matrix multiplication is specified using the ** method in Line 20. Operators of both domains can
be interleaved with calls to user-defined (aggregate) functions: the tuples resulting from the join are converted to vectors in Line 10, followed by a filter predicate, which is applied to the rows of matrix \( X \) in Line 14. Separate types with dedicated syntax in the user-facing API reduce the impedance mismatch between relational and linear algebra, e.g., users do not have to specify linear algebra in terms of for-comprehensions over collections.

Next, we highlight key aspects of Lara’s IR with respect to the code snippet:

1. **UDFs for the second-order functions of the DataBag** (e.g., `map` or `fold`) and **Matrix** (e.g., `forRows` and `forCols`) are defined as closures (Line 14) or provided as library functions (Line 1 and 2). The body of library functions is in-lined in case they are called within a pipeline (e.g., Line 10 and Line 16) and considered during optimization.

2. **DataBag, Matrix, and Vector** types can be used without further optimization. This is useful to debug and test pipelines during development. To enable optimization, the very same pipeline is *quoted* by surrounding the code with an `optimize` macro [17] (Line 4).

3. **Type conversion methods** (Line 12) in the API and the IR track data provenance, i.e., which field of a DataBag element corresponds to a given column in a Matrix and vice versa. It decouples the specification and execution of relational and linear algebra and enables joint optimization.

4. **Type and operator choices** in the user-facing API do not enforce a particular physical execution backend. A unified representation of both types in a common formal representation in the IR enables operator pushdown and fusion of UDFs over type boundaries. For instance, the filter UDF applied on each row of the matrix \( X \) in Line 14 can be pushed to the DataBag and fused with the `vectorize` UDFs applied in a `map` on the join result in Line 10.

5. **White-box UDFs** in the IR enable reasoning about read and write accesses to the processed elements (e.g., fields, rows, and columns). In combination with access to the control flow in the IR, this provides opportunities to fuse UDF applications that are executed iteratively in a loop, if their read/write sets are disjoint. The iterative calculation of the mean in Line 16 can be optimized. Instead of executing the mean function on each column separately, it is executed for all columns at once.

6. **High-level linear algebra operators in the API** (Line 20) and an IR that captures the domain-specific semantics of operators enable the selection of specialized operator implementations, e.g., BLAS [44] instructions for linear algebra.

3.2 Intermediate Representation

In this section, we describe how Lara’s IR facilitates the design decisions described in the previous section. First, we introduce the low-level intermediate representation (LIR) provided by Emma. Then, we present two higher-level *views* on top of the LIR. The **monadic view** represents monad comprehensions [34] over the DataBag and Matrix types. The **combinator view** represents high-level operators (e.g., matrix multiplication) as single entities or *combinators* [33] in an operator tree.

**Low-Level Intermediate Representation.** Emma uses the meta-programming features of Scala [17] to access the AST of a quoted program. Emma transforms the original AST into let-normal form (LNF), a functional representation of static single assignment form (SSA) [8], to overcome several shortcomings of Scala’s AST. LNF offers a normalized representation that encodes dataflow and control flow information directly. LNF guarantees that variables are defined only once, i.e., the *single static assignment* property. Thus, *def-use chains* [6] can be implemented efficiently. This property eases data dependency analysis, most notably the detection of dependencies across control constructs, e.g., between iterations of loops. The LIR is used as basis two views, which we introduce in the next paragraphs. The views combine expressions of the LIR that represent certain operations, e.g., a matrix multiplication, and make their semantics available for reasoning. The LIR augments the views by providing efficient data- and control flow analysis.

**Monadic View.** Monad comprehensions [34] on the Bag monad provide a concise and declarative way to specify collection transformations with first-class support for user-defined (aggregation) functions, as shown in the introductory example (Listing 3.1, Line 10). Matrices and vectors can also be represented as a `Set` monad \( \{(i, v)\} \) of index-value tuples, where \( i \) is a singular *index* in case of a vector and a tuple *(row-index, column-index)* in case of a matrix [27]. Writing linear algebra operators as monad comprehensions at the user-level is tedious and error prone. As a consequence, Lara offers high-level operators for linear algebra in its API. Monad comprehensions in the IR enable Lara to examine and optimize applications of UDFs, e.g., fusion and pushdown. These optimizations are not bound to the concrete monad instance, but rely on the general properties of monads. Lara extends the monad representation of the DataBag in Emma with the Matrix and Vector monad. A traversal over the LIR converts all explicit second-order functions on a DataBag to monad comprehensions (e.g., a `map` is converted to the corresponding comprehension). Calls to linear algebra methods and second-order functions of the Matrix type are replaced by their corresponding monad comprehensions. For instance, element-wise addition of two vectors is written as \( x + y \) in the API and represented as following AST nodes in the LIR:

```
App(Select(Ident("x"), "$plus"), Ident("y"))
```

In the monadic view, element-wise addition is represented as following monad comprehension [27]:

```
for { (idx_x, val_x) <- x       // generator
      (idx_y, val_y) <- y       // generator
      if idx_x == idx_y         // guard
    } yield (idx_x, val_x + val_y) // head
```

The comprehension contains two *generators*, which bind each *(index, value)* pair of the two vectors \( x \) and \( y \). The *head* expression is called for each combination of pairs that satisfies the *guard* expression. The values of all pairs that have the same index are added and form a new vector for the result of the addition.

**Combinator View.** The monadic view allows Lara to apply fusion over UDFs based on the properties of monads. In order to apply domain-specific optimizations and trace operator trees, data types and their respective high-level operations need to be represented explicitly in the IR [66]. Comprehension combinators [33] can be leveraged to represent high-level, logical operators whose semantics are not present in the monadic view. At the combinator view, relational operators, such as `join`, and linear algebra operations, such as
a matrix multiplication are captured as single entities and
the program is viewed as a call or operator tree of these
entities. This enables optimizations known from relational
query processing (e.g., join reordering or the choice of differ-
ent physical operator implementations). At the same time,
it enables optimizations on the semantics of linear algebra.

Logical operators can be replaced with specialized physical
implementations, e.g., BLAS sub-routines [44]. Moreover,
operators can propagate interesting properties to their child
nodes, such as row- or column-wise access patterns to their
operands. Based on these properties, different plan variants
are generated. The combinator view for the computation of
the gram matrix $X'X$ is illustrated in Figure 1a. Figure 1b
depicts the operator tree with propagated access patterns –
from the root expression to the sources. The matrix-
multiplication ($\ast \ast$) prefers fast access to the rows (dotted
lines) of its left operand and the columns (snake lines) of
its right operand. The physical row-wise data layout of
the source matrix $X$ is depicted in brackets in the Figure.
Figure 1c represents a physical plan variant. The matrix-
multiplication is replaced by a specialized BLAS instruction
called Dgemm, which requests column-wise partitioned in-
put. An convert enforcer [31] establishes the data layout
desired by the Dgemm instruction for its second operand.
We provide a detailed discussion in Section 4.3 and 4.4.

4. OPTIMIZING END-TO-END PIPELINES

In this section, we showcase how our IR enables differ-
et optimizations and present the LIR’s interplay with the
monadic and combinator view.

Running Example. Listing 1 depicts our running example
of an end-to-end training pipeline, which leverages historical
data about clicks on advertisements to predict the number
of future clicks on other advertisements using a regression
model. The pipeline showcases common user-defined fea-
ture transformations. We omit the implementation of the
transformation UDFs for the sake of space. The categori-
cal features in columns 11 to 15 are dummy-encoded [35] as
sparse vectors in Line 4. The numerical features in columns
1 to 10 are normalized [32] to have zero mean and unit
variance in Line 9. We concatenate the numerical features
in Line 5 and combine them with the dummy-encoded fea-
tures in Line 6, in order to end up with one vector per input
record. After preprocessing, we evaluate different candidates
for the hyperparameter lambda with cross-validation [38]
on the normalized feature matrix $X$. Lara provides cross-
validation as utility method, similar to ML libraries such as
scikit-learn. The learning algorithm supplied to the cross-
validation is executed $k$ times - for $k$ different combina-
tions of training sets and test sets obtained from the feature ma-
trix $X$ and target vector $y$. This example trains a ridge
regression model in Line 17 – 20, and calculates its test er-
ror in Line 22 – 23.
4.2 Operator Fusion

As discussed in Section 4.1, feature transformations apply two consecutive steps: the `fit` step aggregates column values (e.g., the mean and variance for normalization) in a `fold`. The `transform` step changes the column values based on the aggregate in a `map`. If the feature transformation is applied on multiple disjoint columns, Lara can fuse the consecutive `fold` and `map` applications. This allows us to share a given pass over the data, and only requires a single `fold` and `map` operation, independent of the number of transformed columns.

We briefly discuss operator fusion techniques, before we introduce the control flow and dependency analysis, which Lara applies to verify the applicability of operator fusion.

Background: Fold-Fusion. Fusion is based on two core operations of an algebraic data type `T` with element type `A`:

\[
T[A].map[B](f: A -> B): T[B]
\]

and the generic structural recursion

\[
T[A].fold[B](zero: B, plus: (B, B) => B): B
\]

Function composition has been applied on several types [68, 22]. It fuses consecutive applications of UDFs in `map`/second-order functions into a single, composed function call:

\[
T.map(f_1).map(...).map(f_N) = T.map(f_N \circ ... \circ f_1)
\]

Fold-fusion combines multiple fold applications on a type to a single fold. In the following example code, the mean over a `DataBag[Int]` is calculated by computing the sum and the count of its elements:

\[
val \text{sum} = \text{bag.fold(0)(e => e, (s1, c1) => (s1 + s2, c1 + c2))}
val \text{count} = \text{bag.fold(0)(e => 1, (c1, c2) => c1 + c2)}
val \text{mean} = \text{sum} / \text{count}
\]

The banana-split [10] law states that pairs of folds that are applied on the same data type can be fused into a single fold, resulting in following code:

\[
val \text{sum} = \text{bag.fold((0,0)(e => (s, 1), ((s1, c1), (s2, c2)) => (s1 + s2, c1 + c2))}
val \text{mean} = \text{sum} / \text{count}
\]

The cata-fusion [10] law enables us to fuse `map` and `filter` operations into a consecutive `fold` application. Lara leverages the fusion capabilities of Emma [3], and applies them to the monad representations of `Matrix` and `Vector`.

Operator Fusion over Loops. We demonstrate operator fusion on the `dummyEncode` method of our running example (Listing 1, Line 4), which is implemented as follows. We hide the implementation details of the fit and transform UDFs and indicate their data dependencies with colored boxes.

```scala
def dummyEncode(bag: DataBag[Array[Any]], columns: Seq[Int]) = {
  var encoded = bag
  for ((columnIndex, elem) <- columns) {
    // Fit: build a dictionary of column values
    val dictionary = encoded.fold(('fit-UDAF))
    // Transform: create encoding in sparse vector
    encoded = encoded.map(('transform-UDF))
  }
  encoded
}
```

Line 5 and 7 depict the dummy encoding for a single column, which is applied iteratively for all columns defined by the `columns` parameter. The `fold` creates a dictionary that maps each distinct column value to a unique index. The consecutive `map` replaces the categorical values by sparse vectors.
containing a single non-zero entry at the index obtained by a dictionary lookup. A naïve execution of the code (i.e., independently on each column) is suboptimal, as it requires two passes over the data per column. In order to fuse the UDFs and save multiple passes over the data, Lara performs (i) a dependency analysis of the loop variable columnIndex, and (ii) an analysis of the UDFs to determine their access patterns to the elements of the DataBag.

Dependency Analysis. At first glance, it is not obvious that the operators can be fused. The fold in Line 5 appears to be a fusion barrier, as the previously built dictionary is required to perform a lookup in the map operator in Line 7. A closer look reveals that the code accesses only a distinct feature column within each iteration. Thus, when we would unroll the loop, we could fuse the consecutive fold applications (Background: Fold-Fusion), if disjoint columns are accessed. Analogously, we could fuse all map applications. Lara analyses the loop based on the direct style control flow representation of the LIR. It validates that no loop-carried dependencies exist and that the read and write accesses to the array elements inside the fit and transform UDFs are conducted with the columnIndex loop variable. Thus, Lara can verify that each consecutive iteration step reads and writes on disjoint columns, if the values of the loop variable are known at compile time (a constant sequence for instance, e.g., 11 to 15) or the function semantics guarantee disjoint access (e.g., all Bulk-Operators in Table 1).

Fusion. After the dependencies have been evaluated successfully, Lara unrolls the loop to enable operator fusion. First, the banana-split [10] rule combines the UDFs executed in the fold operations, as they are applied on the same dataset; all dictionaries are created by executing a single combined fold only (Background: Fold-Fusion). Second, the successive transformations to sparse vectors in the map UDFs are fused, in order to apply all transformations in a single map operation. Thus, the optimized code executes a single fold and a single map only, independent of the number of transformed columns.

In general, operator fusion is always limited by pipeline breakers [52], i.e., (aggregated) data, which is required by an successive operator and thus, has to be materialized. While Lara can not overcome this inherent limitation, it can fuse multiple folds that are applied on the same data and thus, reduce the cost to a single pass over the data. Similar fusion techniques have been proposed in the Stubby [46] optimizer for Hadoop [7]. Lara leverages white-box UDFs and control flow analysis to ensure disjoint field access an thus enables these techniques over loop boundaries. In the moment, Lara requires direct access to the loop variable in its dependency analysis and does not support complex index expressions.

Type-based DSLs (e.g., in Spark and Flink) must execute the loop as-is, which is suboptimal as it prevents pipelining and fusion. Their IR can only reason about the operators, e.g., the fold and map higher-order functions in the example. Control flow and UDFs are not visible, which prevents the required dependency analysis.

4.3 Choosing a Data Layout

Choosing efficient physical operators for linear algebra operations, such as matrix-matrix multiplications, can have a huge impact on the runtime of ML pipelines [67]. We leverage the combinator view to choose appropriate physical implementations of operators based on the layout of the data.

<table>
<thead>
<tr>
<th>Bulk-Operations</th>
<th>Databag (row-wise)</th>
<th>Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>map(a:Vector =&gt; Vector)</td>
<td>forRows: (m:Vector =&gt; Vector)</td>
</tr>
<tr>
<td></td>
<td>map(a:Vector =&gt; Double)</td>
<td>forRows: (m:Vector =&gt; Double)</td>
</tr>
<tr>
<td></td>
<td>.withFilter(f:Vector =&gt; Boolean)</td>
<td>forCols: (n:Vector =&gt; Vector)</td>
</tr>
<tr>
<td></td>
<td>.flatMap(split in (colIdx, (rowIdx, value)))</td>
<td>.forCols: (n:Vector =&gt; Double)</td>
</tr>
<tr>
<td></td>
<td>.map: (rowIdx, value) : values as vector</td>
<td>.forCols: (f:Vector =&gt; Boolean)</td>
</tr>
<tr>
<td></td>
<td>.map(a:Vector =&gt; Vector)</td>
<td>.withFilter(f:Vector)</td>
</tr>
<tr>
<td></td>
<td>.map(m:Vector =&gt; Vector)</td>
<td>.withFilter(index: Int)</td>
</tr>
<tr>
<td></td>
<td>.map(m)</td>
<td>.row(index: Int)</td>
</tr>
</tbody>
</table>

Figure 2 depicts three plan variants for the ridge regression algorithm (Listing 1, Line 17 – 20) in the combinator view. Operators are represented as single entities in an operator graph (e.g., ** denotes matrix multiplication). Similarly to query optimization on relational algebra trees [31], Lara applies the following optimizations: (i) expressions (i.e., subgraphs) are transformed into equivalent expressions based on algebraic rules, (ii) logical operators are replaced by physical operator implementations and (iii) the desired physical data layouts are established by enforcers [31] based on interesting properties.

Transformation Rules. Lara provides an extensible set of rules to check for the applicability of backend specific operators. We define transformation rules to replace our default implementations of linear algebra operations in Scala. Lara applies these transformations for BLAS level 2 (i.e., matrix-vector) and level 3 (i.e., matrix-matrix) operations on dense data. For instance Lara replaces the whole subtree for $X^T X + I * \lambda$ with a general level 3 BLAS matrix-matrix multiplication DGEMM (Figure 2b). Similarly, the general BLAS matrix-vector multiplication DGEMV is used to multiply $X^T y$.

Physical Properties. Access patterns (row-, column-, or element-wise) of linear algebra operators and implementations can differ per operand. For example, a sparse matrix multiplication has fast access to the rows of the left operand and fast column-wise access to the right operand in the best case. Matching those access patterns has a large impact on the performance. For instance, suboptimal access pattern to Compressed Sparse Column (CSC) or CSR formats increases the asymptotic complexity from constant to logarithmic in the number of non-zero values.

To overcome this problem, we annotate the edges of plan variants with the access pattern of operators in a top-down traversal, similar to interesting properties in Volcano [31]. For instance, the default Scala implementation of a matrix-matrix multiplication (** in Figure 2a) yields the best performance in case of fast row-wise access to the left and fast column-wise access to the right operand. In contrast, BLAS sub-routines expect column-wise partitioned inputs, shown in Figure 2b for DGEMM and DGEMV. Lara considers the initial data format of matrices (depicted in brackets next to the matrices) to create plan variants by implicitly inserting conversion operators. Enforcers establish a certain data format, if the sources do not match the propagated access pattern. Furthermore, certain operator implementations can produce different output formats, which allows us to choose the for-

\[http://www.netlib.org/blas/#_blas_routines\]
Language Integration. We implement cross-validation as a utility function (Listing 1, Line 14). In our running example, we use cross-validation to select a regularization constant lambda for the ridge regression model. Its execution is illustrated on the left side of Figure 3. Step 1 is independent of the validation algorithm and partitions the specified feature matrix X and target vector y into k splits. In Step 2, the ridge regression algorithm is executed k times. For k = 3, the algorithm is executed once with X1 as test set and X2,3 as training set, next with X2 test set and X1,3 as training set, and finally with X3 as test set and X1,2 as training set.

Lara leverages the semantics of cross-validation to execute certain linear algebra operations more efficiently. We concentrate on our detailed discussion on model training and, without loss of generality, leave the calculation of the test error in Line 22 – 23 out of the discussion.

Detecting Redundant Computations. If we look carefully at the execution of the learning algorithm in Figure 3, we observe that each feature matrix and target vector split is used twice (k – 1 times in general) as part of the training set. Thus, the training algorithm is executed k – 1 times on each split. This overlap poses potential for optimization: we can avoid redundant computations by (partially) pre-computing the algorithm on the individual splits outside of the validation loop.

The semantics of cross-validation guarantee that the individual splits do not overlap. Therefore, we can treat the splits as block-wise partitioning of the feature matrix X with size m × n into k matrix blocks X1 ... Xk with size m ÷ k × n. Block-partitioned matrices can be multiplied when they have conformable partitions [25], i.e., the block matrix itself and the individual blocks obey the rules of matrix multiplication. For example, X T X can be calculated as sum over its k splits (i.e., matrix blocks X1 ... Xk) using the distributive law: ∑ki=1 X T i X i. This allows us to pre-compute the matrix multiplication of the individual splits X T i X i once outside of the cross-validation loop (Figure 3, Step 3). In each particular iteration, the pre-computed results of the test-set splits have to be added to calculate the final result (Figure 3, Step 4). The time complexity for X T X for a matrix with m rows and n columns is O(n * m * n). Under regular execution, the multiplication has to be executed for each training set with m ÷ k rows and n columns. Thus, the overall complexity for all k iterations in the regular execution is O(k [n * (m ÷ k) * n]) = O(k · (k–1)m · n). In the optimized execution, the multiplication is performed once for all individual splits in Step 4. Each split has m ÷ k rows and n columns, resulting in the complexity of O(k [n ÷ k] · n) = O(mn²) for all k splits. In each iteration k we only have to add (k–1) pre-

3Ridge regression is used for presentation reasons only – other algorithms, e.g., generalized linear models are supported as well.
calculated partitions of size $n \times n$, resulting in the overall complexity of $O(k((k-1)mn))$ for all iterations.

Cost Improvements. ML algorithms are composed of several operators for which we also exploit the data redundancy introduced by cross-validation. Table 2 (Time Complexity) shows a cost comparison for matrix-matrix, matrix-vector and element-wise operations. The calculation depicts the overall cost of executing the operator $k$ times during cross-validation on a feature matrix $X$ with $m$ rows and $n$ columns. $X$ is divided into $k$ splits and each split has $\frac{m}{k}$ rows and $n$ columns. This results in $m - \frac{m}{k}$ rows per training set. Baseline shows the combined cost to execute the operator for all training sets, i.e., the overall cost for using the operator in standard $k$-fold cross-validation. Static is the one-time cost for the part of the operator that can be pre-computed statically on each split. Loop is the combined cost to calculate the final result based on the statically pre-computed values for all cross-validation iterations for the particular operator. Ratio depicts the ratio between the complexity of the baseline and the optimized version, e.g., $1 : (k-1)$ means the baseline has $(k-1) \times$ more time complexity. Such cost computations can also be applied to estimate the additional memory required to store pre-computed results.

Eliminating Redundant Computations. We introduced the cross-validation function in Lara’s IR. Lara pattern matches calls to the cross-validation function and uses the combinator view to represent the linear algebra operations of the cross-validation UDF. It then traverses the combinator view of the UDF in post-order (i.e., from the sources) and checks if rewrites can be applied based on the rules of the currently traversed operator node. Matching operator nodes are split and the algorithm extracts operator trees for the Pre-Computation. The former sub-trees in the original tree are replaced with calls to their results, as depicted in Figure 4 for the ridge regression example. Two trees for the Pre-Computation are created, which calculate the matrix-matrix multiplication for $X \times X$ and matrix-vector multiplication $X \times y$ for the individual splits (Table 2 (Operation)). The extracted sub-trees from the Per-Iteration tree are replaced with the operation of the corresponding sub-tree root node (Table 2 (Per-Iteration)). An additional optimization pass eliminates remaining shortcomings once the optimization for the cross-validation body is finished, e.g., dead code elimination and common subexpression elimination (CSE)\[6\]. For instance, after the two trees for the Pre-Computation are created, the transposition ($\tau$) of the splits $X_{[1,4]}$ would be calculated for each tree separately. After CSE, the transposition is executed only once. Often, cross-validation is also part of an outer loop (e.g., when different hyperparameter candidates are validated). In such cases, Lara moves loop-invariant Pre-Computations out of the loop. As a result, the Pre-Computations are computed only once in our example and we improve the performance by a factor of $1 : (k-1) \times h$ compared to naive execution, where $h$ denotes the number of hyperparameter candidates. Thus, larger hyperparameter spaces can be explored in the same time, which potentially finds better models.

5. EVALUATION

We implemented Lara and the outlined optimizations in Scala, based on Emma v0.2.3. Operations on collections are backed by Scala Streams. The matrix-vector types apply netlib-java v1.1.2 for native BLAS routines in case of dense data, and ScalaNLP Breeze v0.13.1 for sparse data.

Experiment Setup. We conducted our experiments on a server node with an Intel E5530 processor (2.4 GHz, 8 cores), and 48 GB main memory. We run on Oracle Java 8 VM (build 1.8.0_72-b15, -Xmx40G) and use Scala version 2.11.11.

Datasets. We conducted our experiments on synthetic data and two real world datasets: (i) The Criteo\[4\] dataset contains click feedback of display ads. (ii) The Reddit\[5\] dataset contains comments of the news aggregator website reddit.com.

Overview. We evaluate the optimizations for the preprocessing pipelines based on the introductory example (Section 3.1) and our running example (Listing 1) and validate the importance of operator pushdown in Section 5.1 and 5.2. We measure the impact of data layout optimizations on ML algorithms for dense and sparse data in Section 5.3. We benchmark the optimizations for cross-validation in Section 5.4 and hyperparameter tuning in Section 5.5.

5.1 Preprocessing

In this experiment, we evaluate the presented optimizations on the preprocessing pipeline of our running example depicted in Listing 1, Line 3 – 9. We conducted the benchmark on differently sized versions of the Criteo dataset. We evaluate the impact of operator fusion and pushdown on each of the three preprocessing steps in Figure 5. Figure 5a depicts the runtime without pushing the normalization to the array representation of the data, while the pushdown is performed in the experiment represented in Figure 5b. Encode includes reading and converting the lines of the raw datafile to an array representation, as well as dummy encoding the categorical features to sparse vectors (Listing 1, Line 4). Normalize transforms the numerical features to have zero mean and unit variance (Listing 1, Line 9). Concat combines the numerical and the dummy encoded features in a single sparse vector (Listing 1, Line 5 – 6). In Figure 6, we compare Lara to scikit-learn, Spark, and Tensorflow Transform on different data sizes.

Discussion. Baseline shows the runtime for the pipeline executed without any optimizations. Lara depicts the results

\[3\]http://labs.criteo.com/2013/12/download-terabyte-click-logs-2/

\[5\]http://files.pushshift.io/reddit/comments/
with operator fusion and function composition (Section 4.2). Figure 5a depicts the results without operator pushdown for the normalization. Thus, the normalization is executed on sparse vectors. Under baseline execution, (i) the encoding is conducted for each column separately, requiring 5 × 2 passes over the data. (ii) the value concatenation is executed in two separate map operators, and (iii) the normalization is applied to each column separately, requiring 10 × 2 passes over the data. Lara enables the following optimizations: (i) Instead of separately encoding each column, a single fold creates all column dictionaries and then leverages these for encoding the column values in a single map operation. This reduces the complexity to two passes, independent of the number of encoded columns. (ii) Lara fuses the two map UDFs for concatenating the vector into a single map, which reduces the number of functions calls. (iii) The normalization benefits in the same way as the encoding, and reduces the number of passes to two. As the normalization is not pushed to the array representation, access to the numerical features in the UDFs suffers from the slow element-wise access of the sparse vectors: element-wise access requires a binary search with a cost that is logarithmic in the number of non-zero values of the row. Figure 5b depicts the results with operator pushdown in the baseline and Lara. As the elements are still stored in an array, read and write access to the features has constant cost. The scaling benchmarks in Figure 6 show that Lara and Spark scale linear with the increasing data size. Both execute in a streaming fashion and thus, are not affected by growing data sizes. Scikit-learn loads the whole dataset in memory, which leads to degrading performance for larger data sizes.

**Results.** The baseline without pushdown of the feature normalization takes 5.75× longer than Lara without pushdown and 16.1× than the completely optimized version. Overall, the baseline with pushdown is 7.3× slower than Lara. Lara improves the runtime for encoding by 4.7× compared to the baseline as the number of passes over the data is independent from the number of encoded features. Normalization with pushdown is 12.5× faster. This is roughly twice as much improvement compared to the encoding. This is expected as twice as much columns are normalized. Even though the access to the columns in logarithmic time degrades the overall runtime of Lara without pushdown, it is still 6.2× faster than the baseline without pushdown. Concatenation of the features to a single sparse vector requires no data materialization, as only map operators are used. Thus, the baseline and Lara can both stream data, and the function composition applied by Lara does not yield significant benefits. Figure 6 shows that scikit-learn initially outperforms single core Spark but degrades heavily and fails to execute for the 25GB sample due to out-of-memory errors. It already uses 15GB of memory for the smallest sample. For the 15GB sample, scikit-learn uses the whole 48GB main-memory available on the cluster node, which leads to 10.7× worse performance compared to the initial data sample. Lara outperforms scikit-learn by 2.1 and 7.3×. Lara consistently executes around 3× faster than single threaded execution in Spark. Spark with 8 parallel executors outperforms Lara (on a single core) by a factor of 2.2×. Tensorflow runs on Apache Beam, but only supports Google Dataflow and the unoptimized DirectRunner as backends in the moment. We ran Tensorflow Transform (TFT) on our cluster node with the DirectRunner, which failed with an realloc error after successfully applying the preprocessing for the 5GB sample. It fails to execute on the larger samples.

### 5.2 Operator Pushdown

In this experiment series, we evaluate the effects of operator pushdown based on Line 6 – 14 in the introductory example in Section 3.1, which applies a filter on the third column of the vectorized join result. The baseline executes the pipeline as specified, applying the filter on the matrix type. Lara pushes the filter application to the DataBag representation, before it is converted to a matrix. We conduct the experiments on a normalized version of the Reddit dataset with 1.4 million users and 31 million comments. The vectorize UDFs extract the id, down-votes, up-votes, and perform feature-hashing [69] of the n-grams obtained from the user-name \((n = 2)\) and comment-text \((n = 10)\) to a fixed, sparse vector space of 10000 and 50000.

**Discussion.** As described in Section 5.1, the element type of the DataBag representation (product types for user and comment) provides constant time access. The filter UDF of the forRows method is called for each row of the CSR matrix. Element-wise access to a particular value in the sparse row vector has logarithmic complexity. In a CSC matrix, the filter UDF could be evaluated for all non-zero values of the vector that represents the filtered column, but would require a conversion beforehand. It is important to note that the pushdown is only possible, because the filter is applied on the numerical feature down-votes. An inherent barrier for the pushdown of a function \(f\), applied on columns \(c\), is any previously applied function \(g\), which is applied on the same columns \(c\) and has no inverse function. The feature hashing transformation has no inverse and thus, prevents operator pushdown.

**Results.** Lara takes 120.60 seconds to create the matrix representation of the filtered join result. Without filter pushdown, the execution is 7.3× slower and takes 881.41 seconds.

### 5.3 Data Layout

In this experiment series, we benchmark the impact of the matrix data layout on the performance of ML algorithms. We first evaluate ridge-regression as shown in Listing 1, Line 17 – 20. It calculates the solution directly using a
Table 3: Benchmarks on data access patterns.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Variant</th>
<th>Feature Layout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Regression</td>
<td>Scala</td>
<td>column-wise</td>
</tr>
<tr>
<td></td>
<td>BLAS</td>
<td>row-wise</td>
</tr>
<tr>
<td></td>
<td>BLAS+Convert</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression w/ BGD</td>
<td>Breeze</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BLAS+Convert</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression w/ BGD</td>
<td>Breeze+Convert</td>
<td></td>
</tr>
</tbody>
</table>

1 var weights = Vector(...) // initialize weight vector
2 for (_ <- 0 until Iterations) {
3   val hyp = X ** w
4   val exp = hyp.map(value => 1 / (1 + math.exp(-1 * value)))
5   val loss = exp - y
6   weights = weights - alpha * ((X.t ** loss) / X.numRows)
7 }

Listing 2: Logistic regression with BGD in Lara.

Discussion. All results are depicted in Table 3. Ridge regression is conducted on dense data with row- and column-wise formats of feature matrix X. Scala depicts the result for our own Scala implementations of dense linear algebra operators. BLAS depicts the result with BLAS instructions. BLAS+Convert depicts the result with BLAS and enforcers that establish the desired data layout for the dgemm instruction. Scala executes the plan depicted in Figure 2a for row- and column-wise features. BLAS executes the variant shown in Figure 2b for both layouts. For row-wise features, BLAS+Convert executes the plan shown in Figure 2c. For column-wise features, BLAS+Convert executes a plan variant that converts the input to the transpose (T), both for the dgemm and dgemv instruction, to achieve a compliant data layout. The results for logistic regression are conducted on sparse data (10 percent non-zero values). Analogous to the ridge regression example, we conducted the experiments on row- and column-wise features X. Breeze depicts the results for Scala, which internally uses the Breeze library to execute sparse linear algebra on matrices in compressed sparse row (CSR) and column (CS) layout. Breeze+Convert depicts the results when an enforcer establishes the desired data layouts of the operators (Listing 2): for row-wise features, Breeze+Convert converts the feature matrix X used in the multiplication with the loss vector X.t ** loss; for column-wise features, Breeze+Convert introduces an enforcer for X read in the multiplication with the weight vector X ** w.

Results. The benchmarks for ridge regression on dense data show the importance of specialized physical operators. BLAS is 51.8x faster for column-wise and 122x faster for row-wise features compared to the Scala implementation. For the Scala implementation, the column-wise feature layout matches the properties of the operators (Figure 2a) and is 2.4x faster than the row-wise feature layout. Converting the matrix for the BLAS instructions (BLAS+Convert) introduces a performance overhead of 1.43x for column and 1.39x for row-wise features: faster execution of the BLAS instruction can not overcome the overhead of the conversion.

The experiments for sparse data show the importance of choosing the best-suited data format. The initial feature layout only satisfies the access pattern of one of the two matrix-vector multiplications. Column-wise partitioned features are slightly faster, as the loss vector used in X.t ** loss is larger than the weight vector (by factor 10 in our experiments). The variants that introduce an enforcer to satisfy the desired access pattern of the operators increase the performance by a factor of 15.1x for 1 and up to 141.7x for 100 iterations in case of column-wise partitioned features. A enforcer for row-wise partitioned features brings the execution time on par with the column-wise features and achieves an up to 194x performance improvement for 100 iterations. This is due to the asymptotic access cost, which changes from logarithmic to constant.

5.4 Cross-Validation

In this experiment series, we benchmark the impact of the proposed rewrites for cross-validation. We evaluate ridge regression as shown in Listing 1, Lines 17–20 and logistic regression with batch gradient descent (BGD) (Listing 2). Each figure depicts the results for synthetic data with a fixed number of columns (1000) and folds (5 and 10). The number of rows in the dataset is scaled on the x-axis.

Ridge Regression Discussion. Figure 7 depicts the results for dense data. The Baseline implementation executes the algorithm without the proposed optimizations for cross-validation described in Section 4.4. Lara executes the matrix-matrix and matrix-vector multiplications in a Pre-Computation step on each split before the cross-validation iterations are executed.

Ridge Regression Results. Lara is up to 65x faster than the Baseline for five folds and up 136x faster for ten folds. This heavily exceeds the expected ratio from the cost estimation in Table 2. We relate this to the very small intermediate result for the Pre-Computation of X’X. The intermediate results for the individual splits have the size n x n, where n is the number of columns in the training matrix X.

Logistic Regression Discussion. Figure 8a and 8b depict the results on dense data, while Figure 8c and 8d depict the results for sparse data. The experiment setup matches the previous experiment series on ridge regression. In the Baseline, the two most expensive operations are the multiplication Xw of the feature matrix X with the weight vector w, and the multiplication X’loss of the transposed feature matrix with the loss vector loss to calculate the gradient. Lara is able to extract these operations to lower the computational complexity, as described in Section 4.4. The Pre-Computation of the hypothesis hyp can now be calculated for all k weight vectors wk in a single matrix multiplication XkW, by stacking all weight vectors into a matrix W.

Logistic Regression Results. On dense data, Lara is up to 4.8x faster for 5 folds and 8.6x faster for 10 folds than...
the baseline (Figure 8a and 8b). The impact of the redundant computations in the baseline grows with the number of rows in the training set. Additionally, Lara benefits from the more efficient execution that leverages a single matrix-matrix multiplication instead of multiple matrix-vector multiplications in the baseline. On sparse data, Lara achieves a speedup of up to 1.2× for 5 folds and 1.4× for 10 folds compared to the baseline (Figure 8c and 8d). The cross-validation optimization is only beneficial once the number of rows is larger than 50,000 rows. In contrast to the dense implementation, the Pre-Computations for sparse data cannot leverage more efficient instructions, and the speedup is solely based on the cross-validation optimization.

5.5 Hyperparameter Tuning

In this experiment series, we benchmark the performance impact of our proposed rewrites for the cross-validation utility function with hyperparameter tuning. We evaluate ridge regression with different \( \lambda \) values for the regularization matrix, as shown in the running example. Next, we evaluate logistic regression with BGD with different initializations of the weight vector \( w \). The feature matrix has 1000 columns in both experiment series, while we scale the number of rows. We tune for 5 different hyperparameters and validate them with 5-fold cross validation. The logistic regression with BGD runs for a fixed amount of 100 iterations.

Ridge Regression Discussion. Figure 9 depicts the results of the benchmark. The Baseline executes the cross-validation and hyperparameter loop without rewrites, but uses BLAS instructions. We provide experiments for two optimization variants: Lara (CV only) depicts the results of optimized cross-validation without removing loop invariant code. Lara depicts the results after the loop invariant code is pulled out of the hyperparameter loop.

Ridge Regression Results. As expected, the baseline implementation takes 5× longer than the single cross-validation (Section 5.4). Lara (CV only) is up to 141× faster than the baseline, which is analogous to the improvements for a single cross-validation. Lara with all optimizations achieves up to 800× speedups compared to the baseline and is up to 8× faster than Lara (CV only).

Logistic Regression Discussion. We evaluate the hyperparameter tuning for logistic regression on dense and sparse matrix representation with 10 percent non-zero values. Figure 10a depicts the results for dense matrix representation and Figure 10b for sparse matrix representation. We provide results for an implementation that uses batching as additional baseline (similar to the approach presented in Tu-PaQ [63]) to provide a reference point for our optimization. Batching reduces the number of times the dataset has to be read from \( n \) times (i.e., for each hyperparameter individually) to one time. It batches the model training by combining the weight vectors \( w \) (i.e., the hyperparameters) into a matrix \( W \), and thereby replaces \( n \) matrix-vector multiplications \( h = Xw \) with a single matrix-matrix multiplication \( Y = XW \). Lara applies the optimizations presented in Section 4.4. The hyperparameter loop adds an additional nesting layer: we do not use a single weight vector \( w \) per split \( k \), but a matrix \( W \), which contains all the weight candidates as columns, thus \( h_{\text{test}} = [W_0, ..., W_k] \). Therefore, the resulting Pre-Computation for the individual splits involves \( k^2 \) iterations, as shown in the following code snippet:

```scala
for (k <- 0 until k) {
  for (j <- 0 until k) { // W_j from hyp-list
    val b = X_train ** W_j
    val exp = h.map(value => 1 / (1 + math.exp(-1 * value)))
    val loss = exp - y_train
    val w_{i,j} = X_train,t ** loss
  }
}
```

Logistic Regression Results. Batching outperforms the baseline by up to 5× for 10 hyperparameters. Lara achieves speedups of up to 8× compared to the baseline. Up to 10,000 rows, Lara and batching provide comparable performance. For larger number of rows, Lara outperforms batching by up to 1.8×. For the sparse matrix representation we can observe that the baseline outperforms batching and Lara until a scaling factor of 50,000 rows. For small number of rows, the baseline is up to 1.5× faster. Batching outperforms Lara by 1.1× and the baseline by 1.2×. We account the loss in Laras performance for smaller data sizes to the management and access cost of the weight matrices for the different hyperparameters. In future work, we plan to integrate batching as a rewrite rule for sparse data.
6. RELATED WORK

ML Libraries & Languages. SystemML [15] and Mahout Samsara [61] have R-like linear algebra abstractions and execute locally or distributed on Hadoop and Spark. They apply pattern-based rewrites and inter-operator optimizations such as operator fusion, and SystemML’s execution strategy is based on cost estimates. Mahout Samsara does not provide substantial relational algebra capabilities. SystemML provides a transform function to apply predefined feature engineering methods, such as dummy encoding, binning, and missing value imputation, to raw datasets. SystemML can fuse the specified transformations, as their semantics guarantee disjoint column access. In contrast to Lara, users can not specify their own transformations UDFs in SystemML. OptiML [65] is a DSL for machine learning based on the Delite [18] framework. It shares a lot of ideas with Lara, as it provides pattern-based rewrites for linear algebra operations and operator fusion to avoid intermediate results. OptiML does not provide optimizations based on control flow analysis and is restricted to linear algebra operations. KeystoneML [64] executes ML pipelines on Apache Spark, automatically chooses solvers, and selects data materialization strategies. Due to its type-based DSL, KeystoneML cannot apply operator re-ordering and fusion. To the best of our knowledge, Lara is the first language abstraction to combine linear and relational algebra, which is at the same time able to reason and optimize across the two algebraic abstractions, control flow and UDFs.

ML Specific Optimizations. Kumar et al. [41] propose learning of linear models on data in relational databases, which was later extended to linear algebra operators [20]. In this work, linear algebra operations can be pushed down to relations in databases, similar to [19]. Control flow, UDFs and general preprocessing pipelines are not considered. SystemML provides a ParFOR [13] primitive that executes the body of the loop in parallel or distributed. The operator fusion over loops, presented in Section 4.2, detects independent tasks (e.g., encoding of distinct columns), but fuses them instead of executing them in parallel. SystemML also performs operator fusion [23, 12] and generates linear algebra kernels based on skeleton classes. During a cost-based selection, the best plan with regards to fusion and caching for pipeline breakers is chosen. While the fusion techniques used in SystemML are superior to those presented in this work, SystemML does not consider collection processing for fusion. Yuan Yu et al. [70] extend TensorFlow with support for dynamic control flow, but, to the best of our knowledge, do not perform control flow and UDFs analysis to apply rewrites such as operator fusion. TuPaQ [63] is a framework for automated model training and supports custom optimizations such as batching to train multiple hyperparameters for linear models in parallel, which can be integrated in Lara. MLBase [39] provides high-level abstractions for ML tasks and basic support for relational operators. Its optimizer can choose between different ML algorithm implementations. In contrast to Lara, it does not consider relational operators during optimization and thus provides no capabilities for holistic optimizations.

Execution Engines. Weld [54] focuses on efficient data movement of data-parallel operators between different libraries. Domain-specific optimizations such as reordering linear algebra and operator pushdown are not supported. Scalable linear algebra on a relational database system [47] proposes a system to efficiently execute and optimize linear algebra over a parallel relational DBMS. It uses the foreign function interface of the DBMS to execute UDFs and complex linear algebra operations, which prohibits holistic optimizations and requires data movement in case of end-to-end pipelines. Meta-Dataflows [28] proposes a framework for exploratory execution of dataflows. It provides a high-level API with ML algorithms as function calls and does not focus on optimizing pipelines including UDFs.

7. CONCLUSION

In this paper, we present Lara, a DSL and IR for ML training pipelines. We based Lara on three key requirements that a DSL design should adhere to, in order to enable holistic optimizations: (i) The user-facing API should be declarative and provide dedicated types for both domains — the execution order and operator implementation is independent of the program specification. (ii) The complete pipeline should be visible by the optimizer — next to the data types and operations, UDFs and control flow have to be analyzed to perform certain optimizations. (iii) The IR should provide different levels of abstraction for diverse optimizations — a unified representation of types is required to reason about operator fusion and pushdown, while domain-specific optimizations require a high-level representation of operator semantics. We showcase such a DSL and IR and presented concrete optimizations for ML training pipelines. Our evaluation shows that the proposed optimizations can yield speedups of up to an order of magnitude.

Limitations & Future Work. Our prototype is not integrated in a dedicated runtime nor uses code-generation in the moment. This would alleviate several shortcomings of our current implementation: we did not yet implement a robust caching mechanism, e.g., to test different models on the same feature set. Memory-safe caching requires runtime support; simply caching data in the Java Virtual Machine (JVM) heap is subject to out-of-memory errors. Emma supports caching for the DataBag type, but Lara misses a robust implementation for matrices in the moment. The common view as monads enables fusion of linear algebra operators with applications of UDFs. Lara currently does not apply fusion of linear algebra operators and UDFs applications, as our current dense (BLAS) and sparse (Breeze) backends do not support fused operators. Future work could extend our optimizations on data layout access patterns to generate kernels for sparse linear algebra operations with UDF support and hardware-efficient code by integrating ideas from recent work [37, 12, 43, 16]. Furthermore, one could extend the combinator view by integrating more data representations (e.g., block-wise or compressed [24]). Finally, to support data layout optimizations for intermediate results, we plan to extend Lara with runtime code analysis and an cost-based optimizer. This also enables reasoning about data layout decisions that are dependent on dynamic control flow, e.g., for conditional operations that are dependent on predicates that have to be evaluated at runtime.

Acknowledgments. This work has been supported by the EU project E2Data (ref. 780245), the German Ministry for Education and Research as BBDC 2 (ref. 01IS18025A) and BZML (ref. 01IS18037A), and the Moore-Sloan Data Science Environment at New York University.
8. REFERENCES


