

KeystoneML: Optimizing Pipelines for Large-Scale Advanced Analytics

Evan R. Sparks*, Shivaram Venkataraman*, Tomer Kaftan*[‡], Michael J. Franklin*[†], Benjamin Recht*
{sparks,shivaram,tomerk11,franklin,brecht}@cs.berkeley.edu

*AMPLab, Department of Computer Science, University of California, Berkeley

[†]Department of Computer Science, University of Chicago

[‡]Department of Computer Science, University of Washington

Abstract—Modern advanced analytics applications make use of machine learning techniques and contain multiple steps of domain-specific and general-purpose processing with high resource requirements. We present KeystoneML, a system that captures and optimizes the end-to-end large-scale machine learning applications for high-throughput training in a distributed environment with a high-level API. This approach offers increased ease of use and higher performance over existing systems for large scale learning. We demonstrate the effectiveness of KeystoneML in achieving high quality statistical accuracy and scalable training using real world datasets in several domains.

I. INTRODUCTION

Today’s advanced analytics applications increasingly use machine learning (ML) as a core technique in areas ranging from business intelligence to recommendation to natural language processing [1] and speech recognition [2]. Practitioners build complex, multi-stage *pipelines* involving feature extraction, dimensionality reduction, data transformations, and training supervised learning models to achieve high accuracy [3]. However, current systems provide little support for automating the construction and optimization of these pipelines.

To assemble such pipelines, developers typically piece together domain specific libraries [4], [5] for feature extraction and general purpose numerical optimization packages [6], [7] for supervised learning. This is often a cumbersome and error-prone process [8]. Further, these pipelines need to be completely re-engineered when the training data or features grow by an order of magnitude—often the difference between an application that provides good statistical accuracy and one that does not [9]. As no broader system has purview of the end-to-end application, only narrow optimizations can be applied. These challenges motivate the need for a system that

- Allows users to specify end-to-end ML applications in a single system using high level logical operators.
- Scales out dynamically as data volumes and problem complexity change.
- Automatically optimizes these applications given a library of ML operators and the user’s compute resources.

While existing efforts in the data management community [10], [11], [7] and in the broader machine learning systems community [6], [12], [13] have built systems to address some of these problems, each of them misses the mark on at least one of the points above.

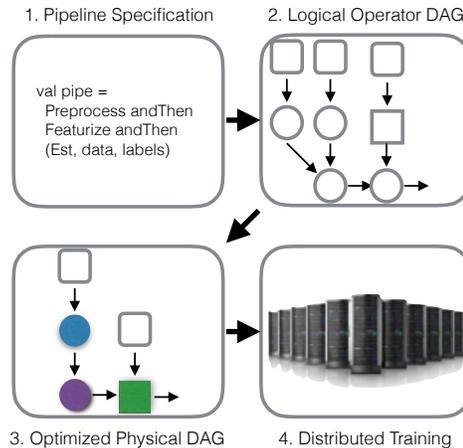


Fig. 1: KeystoneML takes a high-level ML application specification, optimizes and trains it in a distributed environment. The trained pipeline is used to make predictions on new data.

We present KeystoneML, a framework for ML pipelines designed to satisfy the above requirements. Fundamental to the design of KeystoneML is the observation that model training is only one component of an ML application. While a significant body of recent work has focused on high performance algorithms [14], [15], and scalable implementations [16], [7] for model training, they do not capture the featurization process or the logical intent of the workflow. KeystoneML provides a high-level, type-safe API (Figure 1) built around *logical operators* to capture end-to-end applications.

To optimize ML pipelines, database query optimization provides a natural motivation for the core design of such a system [17]. However, compared to relational database query optimization, ML applications present an additional set of concerns. First, ML operators are often *iterative* and may require multiple passes over their inputs, presenting opportunities for data reuse. Second, many ML operators provide only approximate answers to their inputs [15]. Third, numerical data properties such as sparsity and dimensionality are a necessary source of information when selecting optimal execution plans and conventional optimizers do not consider them. Finally, the system should be aware of the computation-vs-communication tradeoffs inherent in distributed processing [11], [6] and choose appropriate distributed execution strategies.

To address these challenges we develop techniques to do both per-operator optimization and end-to-end pipeline optimization for ML pipelines. We use a cost-based optimizer that accounts for both computation and communication costs and our cost model can easily accommodate new operators and hardware configurations. To determine which intermediate states are materialized in memory during iterative execution, we formulate an optimization problem and present a greedy algorithm that works efficiently and accurately in practice.

We measure the importance of cost-based optimization and its associated overheads using real-world workloads from computer vision, speech and natural language processing. We find that end-to-end optimization can improve performance by $7\times$ and that physical operator optimizations combined with end-to-end optimizations can improve performance by up to $15\times$ versus unoptimized execution. We show that in our experiments, poor physical operator selection can result in up to a $260\times$ slowdown. Using an image classification pipeline on over 1M images [3], we show that KeystoneML provides linear performance scalability across various cluster sizes, and statistical performance comparable to recent results [18], [3]. Additionally, KeystoneML can match the performance of a specialized phoneme classification system on a BlueGene supercomputer while using $8\times$ fewer resources.

In summary, we make the following contributions:

- We present KeystoneML, a system for describing ML applications using high level logical operators. KeystoneML enables end-to-end optimization of ML applications at both the operator and pipeline level.
- We demonstrate the importance of physical operator selection in the context of input characteristics of three commonly used logical ML operators, and propose a cost model for making this selection.
- We present and evaluate an initial set of whole-pipeline optimizations, including a novel algorithm that automatically identifies a subset of intermediate data to materialize to speed up pipeline execution.
- We evaluate these optimizations in the context of real-world pipelines in a diverse set of domains: phoneme classification, image classification, and textual sentiment analysis, and demonstrate near-linear scalability over 100s of machines with strong statistical performance.
- We compare KeystoneML with several recent systems for large-scale learning and demonstrate superior runtime from our optimization techniques and scale-out strategy.

KeystoneML is open source software¹ and is being used in scientific applications in solar physics [19] and genomics [20]

II. PIPELINE CONSTRUCTION AND CORE APIS

In this section we introduce the KeystoneML API that can be used to express end-to-end ML pipelines. Each pipeline is composed a number of *operators* that are chained together. For example, Figure 2 shows the KeystoneML source code

```

val textClassifier = Trim andThen
  LowerCase andThen
  Tokenizer andThen
  NGramsFeaturizer(1 to 2) andThen
  TermFrequency(x => 1) andThen
  (CommonSparseFeatures(1e5), data) andThen
  (LinearSolver(), data, labels)
val predictions = textClassifier(testData)

```

Fig. 2: A text classification pipeline is specified using a small set of logical operators.

```

trait Transformer[A, B] extends Pipeline[A, B] {
  def apply(in: Dataset[A]): Dataset[B] = in.map(apply)
  def apply(in: A): B
}

trait Estimator[A, B] {
  def fit(data: Dataset[A]): Transformer[A, B]
}

trait Optimizable[T, A, B] {
  val options: List[(CostModel, T[A,B])]
  def optimize(sample: Dataset[A], d: ResourceDesc): T[A,B]
}

class CostProfile(flops: Long, bytes: Long, network: Long)

trait CostModel {
  def cost(sample: Dataset[A], workers: Int): CostProfile
}

trait Iterative {
  def weight: Int
}

```

Fig. 3: The KeystoneML API consists of two extendable operator types and interfaces for optimization.

for a complete text classification pipeline. We next describe the building blocks of our API.

A. Logical ML Operators

Conventional analytics queries are typically composed using a small number of well studied relational database operators. This well-defined environment enables important optimizations. However, ML applications lack such an abstraction and practitioners typically piece together imperative libraries. Recent efforts have proposed using linear algebra operators such as matrix multiplication [11], convex optimization routines [21] or multi-dimensional arrays as logical building blocks [22].

In contrast, with KeystoneML we propose a design where high-level ML operations (such as `PCA`, `LinearSolver`) are used as building blocks. Our approach has two major benefits: First, it simplifies building applications. Even complex pipelines can be built using just a handful of operators. Second, this higher level abstraction allows us to perform a wider range of optimizations. Our key insight here is that there are usually multiple well studied algorithms for a given ML operator, but that their performance and statistical characteristics vary based on the inputs and system configuration. We next describe the API for operators in KeystoneML.

Pipelines are composed of *operators*. Transformers and Estimators are two abstract types of operators in KeystoneML. An operator is a function which operates on zero or more inputs to produce some output. A *logical operator* satisfies

¹<http://www.keystone-ml.org/>

```

trait Pipeline[A,B] {
  def andThen[C](next: Pipeline[B, C]): Pipeline[A, C]
  def andThen[C](est: Estimator[B, C], data: Dataset[A]):
    Pipeline[A, C]
  // Combine the outputs of branches into a sequence
  def gather[A, B](branches: Seq[Pipeline[A, B]]):
    Pipeline[A, Seq[B]]
}

```

Fig. 4: Transformers and Estimators are chained using a syntax designed to allow developers to incrementally build pipelines.

some logical contract. For example, it takes an image and converts it to grayscale. Every logical operator must have at least one *physical operator* associated with it which implements its logic. Logical operators with multiple physical implementations are candidates for *optimization*. They are marked `Optimizable` and have a set of `CostModels` associated with them. Operators that are iterative with respect to their inputs are marked `Iterative`. It is in general difficult to identify a priori the exact number of iterations an iterative ML algorithm will take to converge to a solution. However, in our system each iterative algorithm is parameterized with a *maximum* number of *epochs* over the dataset and we use this as an estimate of the number of times an operator will reuse its input. For the whole-pipeline optimization discussed in Section IV-C the fact that an input *is* reused is more important than *how many* times it is reused.

A *Transformer* is an operator that can be applied to individual data items (or to a collection of items) and produces a new data item (or a collection of data items)—it is a deterministic unary function without side-effects. Examples of Transformers in KeystoneML include basic data transformations, feature extractors and model application. The deterministic and side-effect free properties affords the ability to reorder and optimize the execution of the functions without changing the result.

An *Estimator* is applied to a distributed collection of data items and produces a Transformer—it is a function generating function. ML algorithms provided by the KeystoneML Standard Library are Estimators, while featurizers are Transformers. For example, `LinearSolver` is an Estimator that takes a data set and labels, finds the linear model which minimizes the square loss between the training data and labels, and produces a Transformer that can apply this model to new data.

B. Pipeline Construction

Transformers and Estimators are *chained* together into a `Pipeline` using a consistent set of rules. The chaining methods are summarized in Figure 4. In addition to linear chaining of nodes using `andThen`, KeystoneML’s API allows for pipeline branching. When a developer calls `andThen` a new `Pipeline` object is returned. By calling `andThen` multiple times on the same pipeline, users can create multiple pipelines that branch out. Developers concatenate the output of multiple pipelines of using `gather`. Redundancy is eliminated via common sub-expression optimization (Section IV). We find these APIs are sufficient for a number of ML applications (Section V), but expect to extend them over time.

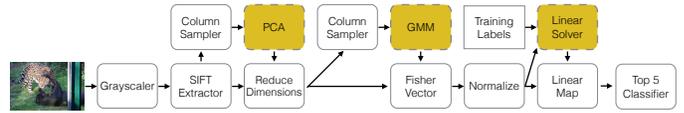


Fig. 5: A pipeline DAG for image classification. Estimators are shaded.

C. Pipeline Execution

KeystoneML is designed to run with large, distributed datasets on commodity clusters. Our high level API and optimizers can be executed using any distributed data-flow engine and we chose Apache Spark as the first execution backend. The execution flow of KeystoneML is shown in Figure 1. First, developers specify pipelines using the KeystoneML APIs described above. As calls to these APIs are made, KeystoneML incrementally builds an operator DAG for the pipeline. An example operator DAG for image classification is shown in Figure 5. Once a pipeline is applied to some data, this DAG is then optimized using a set of optimizations described below—we call this stage *optimization time*. Once the application has been optimized, the DAG is traversed depth-first and operators are executed one at a time, with nodes up until pipeline breakers (i.e. Estimators) packed into the same job—this stage is *runtime*. This lazy optimization procedure gives the optimizer full information about the application in question. We now consider the optimizations made by KeystoneML.

III. OPERATOR-LEVEL OPTIMIZATION

In this section we describe the operator-level optimization procedure used in KeystoneML. Similar to database query optimizers, the goal of the operator-level optimizer is to choose the best physical implementation for every machine learning operator in the pipeline. This is challenging to do because operators in KeystoneML are distributed i.e. they involve computation and communication across the cluster. Operator performance may also depend on statistical properties like sparsity of input data and level of accuracy desired. Finally, as discussed in Section II, KeystoneML consists of a set of high-level operators. The advantage of having high-level operators is that we can perform more wide-ranging optimizations. But this makes designing an optimizer more challenging because unlike relational operators or linear algebra [11], the set of operators in KeystoneML is not closed. We next discuss how we address these challenges.

Approach: The approach we take in KeystoneML is to develop a cost-based optimizer that splits the cost model into two parts: an operator-specific part and a cluster-specific part. The operator-specific part models the computation and communication time given statistics of the input data and number of workers and the cluster specific part is used to weigh their relative importance. More formally, the cost estimate for each physical operator, f can be expressed as:

$$c(f, A_s, R) = R_{exec}c_{exec}(f, A_s, R_w) + R_{coord}c_{coord}(f, A_s, R_w)$$

Where f is the operator in question, A_s contains statistics of a dataset to be used as its input, and R , the *cluster resource*

descriptor represents the cluster computing, memory, and networking resources available. The cluster resource descriptor is collected via configuration data and microbenchmarks. Statistics captured include per-node CPU throughput (in GFLOP/s), disk and memory bandwidth (GB/s), and network speed (GB/s), as well as information about the number of nodes available. A_s is determined through a process we will discuss in Section IV. R_w is the number of cluster nodes available.

The functions, c_{exec} , and c_{coord} are developer-defined operator-specific functions (defined as part of the operator `CostModel`) that describe execution and coordination costs in terms of the longest critical path in the execution graph of the individual operators [23], e.g. the most FLOPS used by a node in the cluster or the amount of data transferred over the most loaded link. Such functions are also used in the analysis of parallel algorithms [24] and are well known for common linear algebra based operators. R_{exec} and R_{coord} are determined by the optimizer from the cluster resource descriptor (R) and capture the relative speed of local and network resources on the cluster.

Splitting the cost model in this fashion allows the the optimizer to easily adapt to new hardware (e.g., GPUs or Infiniband networks) because operator developers only need to implement a `CostModel` and the system accounts for hardware properties. These cost models are approximate and the cost c need not be equal to the actual running time of the operator. As in conventional query optimizers, the goal of the cost model is to *avoid bad decisions*, which a roughly accurate model will do. When two operators have nearly equivalent cost, either should be an acceptable choice. We next illustrate the cost functions for three central operators in KeystoneML and show how input properties affect performance.

Linear Solvers are supervised Estimators that learn a linear map $X \in \mathbb{R}^{d \times k}$ between an input dataset $A \in \mathbb{R}^{n \times d}$ to a labels dataset $B \in \mathbb{R}^{n \times k}$ by finding the X which minimizes the value $\|AX - B\|_F$. In a multi-class classification setting, n is the number of examples or data points, d the number of features and k the number of classes. In the KeystoneML Standard Library we have several implementations of linear solvers, distributed and local, including

- Exact solvers [25] that compute closed form solutions to the least squares loss and return an X to high precision.
- Block solvers that partition the features into a set of blocks and use second-order Jacobi or Gauss-Seidel [26] updates to converge to the right solution.
- Gradient based methods like SGD [15] or L-BFGS [27] which perform iterative updates using the gradient and converge to a globally optimal solution.

Table I summarizes the cost model for each method. Constants are omitted for readability but are necessary in practice.

To illustrate these cost tradeoffs empirically, we vary the number of features generated by the featurization stage of two different pipelines and measure the training time and the training loss. We compare the methods on a 16 node cluster.

Algorithm	Compute	Network	Memory
Local QR	$O(nd(d+k))$	$O(n(d+k))$	$O(d(n+k))$
Dist. QR	$O(\frac{nd(d+k)}{w})$	$O(d(d+k))$	$O(\frac{nd}{w} + d^2)$
L-BFGS	$O(\frac{inskw}{w})$	$O(idk)$	$O(\frac{ns}{w} + dk)$
Block Solve	$O(\frac{ind(b+k)}{w})$	$O(id(b+k))$	$O(\frac{nb}{w} + dk)$

TABLE I: Resource requirements for linear solvers. w is the number of workers in the cluster, i the number of passes over the dataset. For the sparse solvers s is the the average number of non-zero features per example, and b is the block size for the block solver. Compute and Memory requirements are per-node, while network requirements are in terms of the data sent over the most loaded link.

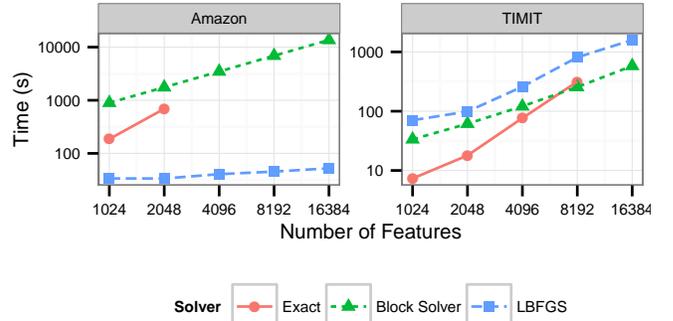


Fig. 6: A poor choice of solver can mean orders of magnitude difference in runtime. Runtime for exact solve grows quadratically in the number of features and times out with 4096 features for Amazon and 16384 features for TIMIT running on 16 c3.4xlarge nodes.

On an Amazon Reviews dataset (see Table IV) with a text classification pipeline, as we increase the number of features from 1k to 16k we see in Figure 6 that L-BFGS performs 5-20 \times faster than the exact solver and 26-260 \times faster than the block-wise solver. Additionally the exact solver crashes for greater than 4k features as the memory requirements are too high. The reason for this speedup is that the features generated in text classification problems are sparse and the L-BFGS solver exploits the sparse inputs to calculate gradients cheaply.

The optimal solver choice does not always stay the same as we increase the problem size or as sparsity changes. For the TIMIT dataset, which has dense features, we see that the exact solver is 3-9 \times faster than L-BFGS for smaller number of features. However when the number of features goes beyond 8k we see that the exact solver becomes slower than the block-wise solver which is also 2-3 \times faster than L-BFGS.

Principal Component Analysis (PCA) is an Estimator used for tasks ranging from dimensionality reduction to whitening to visualization. PCA takes an input dataset A in $\mathbb{R}^{n \times d}$, and a value k and produces a Transformer which can apply a matrix P in $\mathbb{R}^{d \times k}$, where P consists of the first k eigenvectors of the covariance matrix of A . The P matrix can be found using Singular Value Decomposition (SVD) or via an approximate algorithm, Truncated SVD [28]. Both methods can be distributed, and their resource requirements are shown in Table II.

Algorithm	Compute	Network	Memory
SVD	$O(nd^2)$	$O(nd)$	$O(nd + d^2)$
TSVD	$O(ndk + ink^2)$	$O(nd)$	$O(nd + dk)$
Dist. SVD	$O(\frac{nd^2}{w})$	$O(d^2)$	$O(\frac{nd}{w} + d^2)$
Dist. TSVD	$O(\frac{ndk + ink^2}{w})$	$O(i(nk + dk))$	$O(\frac{nd}{w} + dk)$

TABLE II: Resource requirements for PCA Algorithms. For these algorithms, k is the number of principal components requested. All other quantities are defined as in Table I.

	$d = 256$			$d = 4096$		
	$k = 1$	16	64	$k = 16$	64	1024
$n = 10^4$						
SVD	0.1	0.1	0.1	26	26	26
TSVD	0.2	0.3	0.4	3	6	34
Dist. SVD	1.7	1.7	1.7	106	106	106
Dist. TSVD	4.9	3.8	5.3	6	22	104
$n = 10^6$						
SVD	11	11	11	x	x	x
TSVD	14	30	65	x	x	x
Dist. SVD	2	2	2	260	260	260
Dist. TSVD	16	59	262	75	1,326	8,310

TABLE III: Comparison of runtimes (in seconds) for approximate and exact PCA operators across different dataset sizes. A dataset has n examples and d features. k is an algorithm input. An x indicates that the operation did not complete.

To better illustrate how the choice of a PCA implementation affects the run time, we construct a micro-benchmark that varies problem size along n , d , and k , and execute both local and distributed implementations of the approximate and exact algorithm on a 16-node cluster. In Table III, we can see that as data volumes increase in n and d it makes sense to run PCA in a distributed fashion, while for relatively small values of k , it can make sense to use the approximate method.

Convolution is a critical building block of Signal, Speech, and Image Processing pipelines. In image processing, the Transformer takes in an Image of size $n \times n \times d$ and applies a bank of b filters (each of size $k \times k$, where $k < n$) to the Image and returns the b resulting convolved images of size $m \times m$, where $m = n - k + 1$. There are three main ways to implement convolutions: via a matrix-vector product scheme when convolutions are separable, using BLAS matrix-matrix multiplication [29], or via a Fast Fourier Transform (FFT) [30].

The matrix-vector product scheme takes $O(dbk(n - k + 1)^2 + bk^3)$ time. The matrix-matrix multiplication scheme has a cost of $O(dbk^2(n - k + 1)^2)$. The FFT based scheme takes $O(6dbn^2 \log n + 4dbn^2)$, and does not depend on k .

To illustrate the tradeoffs between these methods, in Figure 7, we vary the size of the convolution filter, k , and use representative input images and batch sizes. For small values of k , we see that BLAS is fastest operator. However, as k grows, the algorithm’s dependence on k^2 makes this approach inappropriate. If the filters are separable, it is faster to use the

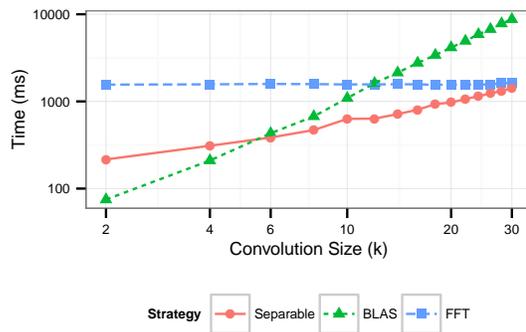


Fig. 7: Time to perform 50 convolutions on a 256x256 3-channel image. As k increases, the optimal method changes.

matrix-vector algorithm. The FFT algorithm does not depend on k and thus performs the same regardless of k .

Cost Model Evaluation: To evaluate how well our cost-model works, we compared the physical operator chosen by our optimizer against the best choice from empirically measured values for linear solvers (Figure 6) and PCA (Table III). We found that our optimizer made the right choice 90% of the time for linear solvers and 84% of the time for PCA. In both cases we found that the wrong choices were made when the running time of two operators were close to each other and thus the approximate cost model did not severely impact overall performance. For example, for the linear solver with 4096 dense features, the optimizer chooses the BlockSolver but empirically the Exact solver is about 30% faster.

As seen from the three examples above, the choice of optimal physical execution depends on hardware properties and on properties of the input data. Thus, choices made in support of operator-level optimization depend on upstream processing and cheaply estimating data properties at various points in the pipeline is an important problem. We next discuss how operator chaining semantics can help in achieving this.

IV. WHOLE-PIPELINE OPTIMIZATION

A. Execution Subsampling

Operator optimization in KeystoneML requires the collection of statistics about input data at each pipeline stage. For example, a text featurization operator might map a string into a 10,000-dimensional sparse feature vector. Without statistics about the input (e.g. vector sparsity) after featurization, a downstream operator will be unable to make its optimization decision. As such, dataset statistics (A_s) are determined by first estimating the size of the initial input dataset (in records), and optimizing the first operator in the pipeline with statistics derived from a sample of the input data. The optimized operator is then executed on the sample, and subsequent operators are optimized. This procedure continues until all nodes have been optimized. Along the way, we form a *pipeline profile*, which includes not just the information needed to form A_s at each step, but also information about operator execution time and memory consumption of each operator’s execution on the sample. We use the pipeline profile to inform the

Automatic Materialization optimization described below. We also evaluate the overheads from profiling in Section V-C.

B. Common Sub-expression Elimination

One of the whole-pipeline rewrites done by KeystoneML is a form of common sub-expression elimination. It is common for training data or the output of featurization stages to be used in several pipeline stages. KeystoneML identifies and merges such common sub-expressions to enable computation reuse.

C. Automatic Materialization

Cache management and automatic selection of materialized views are important optimizations used by database management systems [31] and they have been studied in the context of analytical query systems [32], and feature selection [33]. For ML workloads, materialization of intermediate data is very important for performance because the iterative nature of these workloads means that recomputation costs are multiplied across iterations. By capturing the iterative nature of the pipelines in the DAG, our optimizer is capable of identifying opportunities for reuse, eliminating redundant computation. We next describe a formulation for the materialization problem in iterative pipelines and propose an algorithm to automatically select a good set of intermediate objects to materialize in order to speed up ML pipeline execution.

Given the depth-first execution model and the deterministic and side-effect free nature of KeystoneML operators, a natural strategy is materialization of operator outputs that are visited multiple times during the execution. This optimization works well in the absence of memory constraints.

In many applications we have built with KeystoneML, intermediate output can grow to multiple terabytes in size, even for modestly sized inputs. On current hardware, this output is too big to fit in memory, even with hundreds of GB of memory per machine. Commonly used caching policies such as LRU can result in suboptimal run times because the decision to cache a large object (e.g. intermediate features) may evict a smaller object that is needed later in the pipeline and may be expensive to recompute (e.g. image features).

Our goal is an algorithm that automatically selects the items to materialize in the presence of memory constraints, given that we know how often the objects will be accessed, that we can estimate their size, and that we can estimate the runtime associated with materializing them.

We formulate the problem as follows: Given a memory budget, we want to find the set of outputs to include in the cache set that minimizes total execution time.

Let v be our node of interest in a pipeline G , $t(v)$ is the time taken to do the computation that is local to node v per iteration, $C(v)$ is the number of times a node will be called by its direct successors during execution, and w_v is the number of times a node iterates over its inputs. $T(n)$, the total execution time of the pipeline up to and including node v is:

$$T(v) = \frac{w_v(t(v) + \sum_{c \in \chi(v)} T(c))}{C(v)^{\kappa_v}}$$

```

1 Algorithm GreedyOptimizer
   input : G, t, size, memSize
   output: cache
   cache  $\leftarrow \emptyset$ ;
   memLeft  $\leftarrow$  memSize;
   next  $\leftarrow$  pickNext (G, cache, size, memLeft, t);
   while nextNode  $\neq \emptyset$  do
     cache  $\leftarrow$  cache  $\cup$  next;
     memLeft  $\leftarrow$  memLeft - size(next);
     next  $\leftarrow$  pickNext (G, cache, size, memLeft, t);
   end
10 return cache;
1 Procedure pickNext ()
   input : G, cache, size, memLeft, t
   output: next
   minTime  $\leftarrow \infty$ ;
   next  $\leftarrow \emptyset$ ;
   for  $v \in$  nodes(G) do
     runtime  $\leftarrow$  estRuntime (G, cache  $\cup$  v, t);
     if runtime < minTime & size(v) < memLeft then
       next  $\leftarrow$  v;
       minTime  $\leftarrow$  runtime;
     end
10 end
11 return next;

```

Algorithm 1: The caching algorithm in KeystoneML builds a cache set by finding the node that will maximize time saved subject to memory constraints. $estRuntime$ is a procedure that computes $T(v)$ for a given DAG, cache set, and node.

where $\kappa_v \in \{0, 1\}$ is a binary indicator variable signifying whether a node is cached or not, and $\chi(v)$ represents the direct predecessors of v in the DAG.

Where $C(v)$ is defined as follows:

$$C(v) = \begin{cases} \sum_{p \in \pi(v)} w_p C(p)^{\kappa_p}, & |\pi(v)| > 0 \\ 1, & \text{otherwise} \end{cases}$$

where $\pi(v)$ represents the direct successors of v in the DAG. Because of the DAG structure of the pipeline graph, we are guaranteed to not have any cycles in this graph, thus both $T(v)$ and $C(v)$ are well-defined.

We can state the problem of minimizing pipeline execution time formally as an optimization problem as follows:

$$\begin{aligned} & \min_{\kappa} T(sink(G)) \\ & s.t. \sum_{v \in V} size(v) \kappa_v \leq memSize \end{aligned}$$

Where $sink(G)$ is the pipeline terminus, $size(v)$ the size of v 's output, and $memSize$ the memory constraint.

This problem can also be thought of as problem of finding an optimal cache schedule. It is tempting to reach for classical results [34], [35] in the optimal paging literature to identify an optimal or near-optimal schedule for this problem. However, neither of these results matches our problem setting fully. In particular, Belady's algorithm is only optimal when each item has a fixed cost to bring into cache (as is common in reads from a two-level memory hierarchy), while in our

problem these costs are variable and depend heavily on the computation time to materialize them. e.g. recomputing may be two orders of magnitude faster than reading from disk but an order of magnitude slower than reading from memory, and each operator has a different computational profile. Second, algorithms for the weighted paging problem don't take into account weights that are dependent on the current cache state.

However, it is possible to rewrite the optimization problem above as a mixed-integer linear program (ILP), but in our experiments the cost of solving these problems was prohibitive for practical use at optimization time. Instead, we implement the greedy Algorithm 1. Given an unoptimized pipeline DAG, the algorithm chooses to cache the node which will lead to the largest savings in terms of execution time but whose output fits in available memory. This process proceeds iteratively until either no benefit to additional caching is possible or all available memory has been used.

V. EVALUATION

To evaluate the effectiveness of KeystoneML, we explore its ability to efficiently support large scale ML applications in three domains. We also compare KeystoneML with other systems for large scale ML and show how our high-level operators and optimizations can improve performance. Following that we break down the end-to-end benefits of the previously discussed optimizations. Finally, we assess the system's ability to scale and show that KeystoneML scales well by enabling the development of scalable, composable components.

Implementation: We implement KeystoneML on top of Apache Spark, a cluster computing engine that has been shown to have good scalability and performance for many iterative ML algorithms [7]. In KeystoneML we added an additional cache-management layer that is aware of the multiple Spark jobs that comprise a pipeline, and implemented ML operators in the KeystoneML Standard Library that are absent from Spark MLlib. Porting KeystoneML to work with other distributed computing systems would require that the presence of a distributed collections API that supports the MapReduce paradigm as well as predictable communication via broadcast, all-to-one reduce operations, and aggregation trees. These primitives also exist in systems like Apache Tez, DryadLINQ, FlumeJava, Apache Flink and we plan to study the complexity of porting KeystoneML to these runtimes in the future.

Experiments are run on Amazon EC2 `r3.4xlarge` instances. Each machine has 8 physical cores, 122 GB of memory, and a 320 GB SSD, and was running Apache Spark 1.3.1, Scala 2.10, and HDFS from the CDH4 distribution of Hadoop. We have also run KeystoneML on Apache Spark 1.5, 1.6 and not encountered any performance regressions. We use OpenBLAS for numerical operations. We compare KeystoneML with the distributed version of Vowpal Wabbit (VW) [6], [36] (v8.0) and SystemML [11] (v0.9) running on the same Spark version. If not otherwise specified, we run on a 16-node cluster. Comparisons among KeystoneML, SystemML, VW, and TensorFlow were performed on identical EC2 instance types.

A. End-to-End ML Applications

To demonstrate the flexibility and generality of the KeystoneML API, we implemented end-to-end machine learning pipelines in several domains including text classification, image classification and speech recognition. We next describe these pipelines and compare statistical accuracy and performance results obtained using KeystoneML to previously published results. We took every effort to recreate these pipelines as they were described by their authors, and made sure that our pipelines achieved comparable or better statistical results than those reported by each benchmark's respective authors.

The operators used to implement these applications are outlined in Table V, and the datasets used to train them are described in Table IV. In each case, the datasets significantly increase in size as part of the featurization process, so at model fitting time the size is substantially larger than the raw data, as shown in the last two columns of the table. The Solve Size is the size of the dataset that is input to a Linear Solver. This may be too large for available cluster memory, as is the case for TIMIT. Accuracy results on each dataset achieved with KeystoneML as well as those achieved with the original authors code or (where code was unavailable) as reported in their respective works, are reported in Table VI.

Text Analytics: KeystoneML makes it simple for developers to scale their text pipelines to large datasets. Combined with libraries like CoreNLP [37], KeystoneML allows for scalable implementations of many text classification pipelines such as the one shown in Figure 2. We evaluated a text classification pipeline based on [1] on the Amazon Reviews dataset of 65m product reviews [38] with 100k sparse features. We find that KeystoneML matches the statistical performance of a Vowpal Wabbit [6] pipeline when run on identical resources with the same solver, finishing in 440s.

Kernel SVM for Speech Recognition: Kernel SVMs can be used in many classification scenarios as they can approximate any function. Often their performance has been shown to be much better than simpler generalized linear models [39]. Kernel evaluations can be efficiently approximated using random feature transformations [40], [41] and pipelines are a natural way to specify such transformations. Statistical operators like FFTs and cosine transformations and APIs to merge features help us succinctly describe the pipeline in KeystoneML. We evaluated a kernel SVM solver on the TIMIT dataset with 528k features. Using KeystoneML this pipeline runs in 138 minutes on 64 machines. By contrast, a 256 node IBM Blue Gene machine with 16 cores per machine takes around 120 minutes [41]. In this case, while KeystoneML may be 11% slower, it is using only $\frac{1}{8}$ the number of cores to solve this computationally demanding problem.

Image Classification: Image classification systems are useful in many settings. As images carry local information (i.e. information specific to where in the image a feature appears), locality sensitive techniques, e.g. convolutions or spatially-pooled fisher vectors [3], can be used to generate training features. KeystoneML makes it easy to use modular,

Dataset	Train Size (GB)	Num Train	Test Size (GB)	Num Test	Classes	Type	Solve Features	Solve Size (GB)
Amazon	13.97	65000000	3.88	18091702	2	text	100000 (0.1% sparse)	89.1
TIMIT	7.5	2251569	0.39	115934	147	440-dim vector	528000 (dense)	8857
ImageNet	74	1281167	3.3	50000	1000	10k pixels image	262144 (dense)	2502
VOC	0.428	5000	0.420	5000	20	260k pixels image	40960 (dense)	1.52
CIFAR-10	0.500	500000	0.001	10000	10	1024 pixels image	135168 (dense)	62.9
Youtube8m	22.07	5786881	6.3	1652167	4800	1024-dim vector	1024 (dense)	44.15

TABLE IV: Dataset Characteristics. While raw input sizes may be modest, intermediate state may grow by orders of magnitude before being input to a solver.

Task	Type	Operators Used
Amazon Reviews Classification	Text	LowerCase, Tokenize NGrams, TermFrequency LogisticRegression
TIMIT Kernel SVM	Speech	RandomFeatures, Pipeline.gather LinearSolver
ImageNet Classification	Image	GrayScale, SIFT, LCS, PCA, GMM FisherVector, LinearSolver
VOC Classification	Image	GrayScale, SIFT, PCA, GMM FisherVector, LinearSolver
CIFAR-10 Classification	Image	Windower, PatchExtractor ZCAWhitener, Convolver, LinearSolver SymmetricRectifier, Pooler

TABLE V: Operators used in constructing pipelines for datasets in Table IV.

efficient implementations of image processing operators like SIFT [42] and Fisher Vectors [3], [18]. Many of the same operators we consider here are necessary components of “deep-learning” pipelines [43] which typically train neural networks via stochastic gradient descent and back-propagation.

Using the VOC dataset, we implement the pipeline described in [18]. This pipeline executes end-to-end on 32 nodes using KeystoneML in just 7 minutes. Using the authors original source code the same workload takes 1 hour and 27 minutes to execute on a single 16-core machine with 256 GB of RAM—KeystoneML achieves a 12.4 \times speedup with 16 \times the cores. We evaluated a Fisher Vector based pipeline on ImageNet with 256k features. The KeystoneML pipeline runs in 4.5 hours on 100 machines. The original pipeline takes four days [44] to run using a highly specialized codebase on a 16-core machine, a 21 \times speedup on 50 \times the cores.

In summary, using KeystoneML we achieve one to two orders of magnitude improvement in end-to-end throughput versus a single node, and equivalent or better performance over cluster systems running similar workloads. These improvements mean much quicker ML application development which leads to higher developer productivity. Next we compare KeystoneML to other large scale learning systems.

B. KeystoneML vs. Other Systems

We compare runtimes for the KeystoneML solver with both a specialized system, **Vowpal Wabbit** [6], built to estimate linear models, and **SystemML** [11], a general purpose ML

²We report accuracy on 64k features for ImageNet, while time is reported on 256k features due to lack of consistent reporting by the original authors. The workloads are otherwise similar.

Dataset	KeystoneML		Reported	
	Accuracy	Time (m)	Accuracy	Time (m)
Amazon [1]	91.6%	3.3	-	-
TIMIT [2]	66.06%	138	66.33%	120
ImageNet [3] ²	67.43%	270	66.58%	5760
VOC 2007 [18]	57.2%	7	59.2%	87
CIFAR-10 [45]	84.0%	28.7	84.0%	50.0

TABLE VI: Time to Accuracy with KeystoneML obtained on ML pipelines described in the relevant publication.

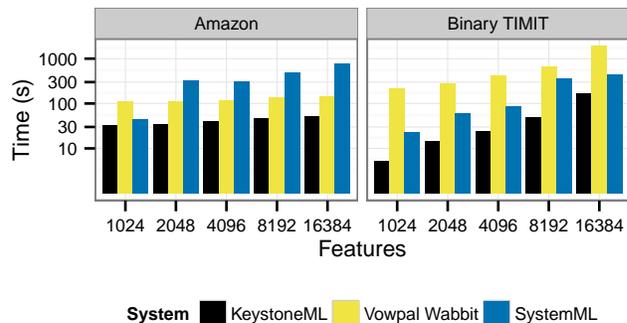


Fig. 8: KeystoneML’s optimizing linear solver outperforms both a specialized and optimizing ML system for two problems across feature sizes. Times are on log scale.

system, which optimizes the implementation of linear algebra operators used in specific algorithms (e.g., Conjugate Gradient Method), but does not choose among logically equivalent algorithms. We compare solver performance across different feature sizes for two binary classification problems: Amazon and a binary version of TIMIT. The systems were run with identical inputs and objective functions, and we report end-to-end solve time. For this comparison, we solve binary problems because SystemML does not include a multiclass linear solver.

The results are shown in Figure 8. The optimized solver in KeystoneML outperforms both Vowpal Wabbit and SystemML because it selects an *appropriate algorithm* to solve the logical problem, as opposed to relying on a one-size fits all operator. At 1024 features for the Binary TIMIT problem, KeystoneML chooses to run an exact solve, while from 2048 to 32768 features it chooses a Dense L-BFGS implementation. At 65536 features (not pictured), KeystoneML finishes in 17 minutes, while SystemML takes 1 hour and 40 minutes to converge to *worse* training loss over 10 iterations, a speedup of 5.5 \times .

The reasons for these performance differences are twofold: first, since KeystoneML raises the level of abstraction to the logical level, the system can automatically select, for example,

Machines	1	2	4	8	16	32
TensorFlow (strong)	184	90	57	67	122	292
TensorFlow (weak)	184	135	135	114	xxx	xxx
KeystoneML	235	125	69	43	32	29

TABLE VII: Time, in minutes, to 84% accuracy on the CIFAR-10 dataset with KeystoneML and TensorFlow configured for both strong and weak scaling. In large weak scaling regimes TensorFlow failed to converge to a good model.

a sparse solver for sparse data or an exact algorithm when the number of features is low, or a block solver when the features are high. In the middle, particularly for KeystoneML vs. SystemML on the Binary TIMIT dataset, the algorithms are similar in terms of complexity and access patterns. In this case KeystoneML is faster because feature extraction is *pipelined* with the solver, while SystemML requires a conversion process for data to be fed into a format suitable for the solver. If we *only* consider the solve step of the pipeline, KeystoneML is roughly $1.5\times$ faster than SystemML for this problem.

TensorFlow is an open source ML system developed by Google [13]. Developed concurrently to KeystoneML, TensorFlow also represents pipelines as graph of dataflow operators. However, the design goals of the two systems are fundamentally different. KeystoneML is designed to support horizontally scalable workloads to offer good scale out performance for conventional machine learning applications consisting of featurization and model estimation, while TensorFlow is designed to support neural network models trained via mini-batch SGD with back-propagation. We compare against TensorFlow v0.8 and adapt a multi-GPU example [45] to a distributed setting in a procedure similar to [46].

To illustrate the differences, we compare the systems’ time to achieve a set accuracy on the CIFAR-10 dataset. While the learning tasks are identical (i.e., make good predictions on a test dataset, given a training dataset), the workloads are not identical. Specifically, TensorFlow implements a model similar to the one presented in [43], while in KeystoneML we implement a version of the model similar to [47]. TensorFlow was run with default parameters and we experimented with strong scaling (fixed 128 image batch size) and weak scaling (batch size of $128 \times \text{Machines}$).

For this workload, TensorFlow achieves its best performance on 4-node cluster with 32 total CPU cores, running in 57 minutes. Meanwhile, KeystoneML surpasses its performance at 8 nodes and continues to improve in total runtime out to 32 nodes, achieving a minimum runtime of 29 minutes, or a $1.97\times$ speedup. These results are summarized in Table VII. We ran TensorFlow on CPUs for the sake of comparability. Prior benchmarks [45] have shown that the speed of a single multi-core CPU is comparable to a single GPU; thus the same pipeline finishes in 50 minutes on a 4 GPU machine.

TensorFlow’s lack of scalability on this task is fundamental to the chosen model and the algorithm being used to fit it. Minimizing a non-convex loss function via minibatch Stochastic Gradient Descent (SGD) requires coordination of the model

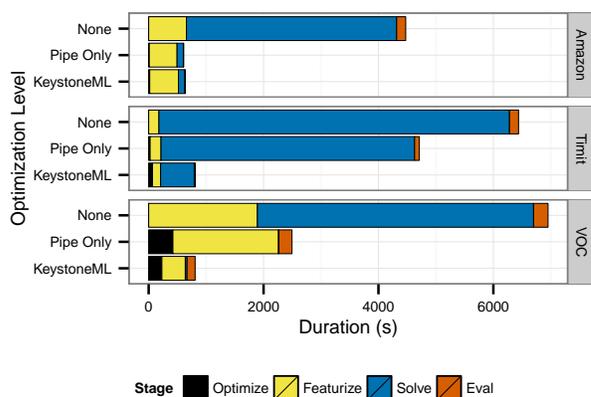


Fig. 9: Impact of optimization levels on three applications, broken down by stage.

parameters after a small number of examples are seen. In this case, the coordination requirements surpass the savings from parallelism at a small number of nodes. While TensorFlow has better scalability on some model architectures [48], it is not scalable for other architectures. By contrast, by using a communication-avoiding solver we are able to scale out KeystoneML’s performance on this task significantly further.

Finally, a recent benchmark dataset from YouTube [49] describes learning pipelines involving featurization with a neural network [48] followed by a logistic regression model or SVM. The authors claim that “models train to convergence in less than a day on a single machine using the publicly-available TensorFlow framework.” We performed a best-effort replication of this pipeline using KeystoneML. We are unable to replicate the author’s claimed accuracy—our pipeline achieves 21% mAP while they report 28% mAP. KeystoneML trains a linear classifier on this dataset in 3 minutes, and a converged logistic regression model with worse accuracy in 90 minutes (31 batch gradient evaluations) on a 32-node cluster. The ability to choose an appropriate solver and readily scale out are the key enablers of KeystoneML’s performance.

We now study the impact of KeystoneML’s optimizations.

C. Optimization Levels

The end-to-end results reported earlier in this section are achieved by taking advantage of the complete set of optimizations available in KeystoneML. To understand how important the per-operator and whole-pipeline optimizations described in Sections III and IV are we compare three different levels of optimization: a default unoptimized configuration (None), a configuration where only whole-pipeline optimizations are used (Pipe Only) and a configuration with operator-level and whole-pipeline optimizations (KeystoneML).

Results comparing these levels, with a breakdown of stage-level timings on the VOC, Amazon and TIMIT pipelines are shown in Figure 9. For the Amazon pipeline the whole-pipeline optimizations improve performance by $7\times$, but the operator optimizations do not help further, because the Amazon pipeline uses CoreNLP featurizers which do not have statistical optimizations associated with them, and the default

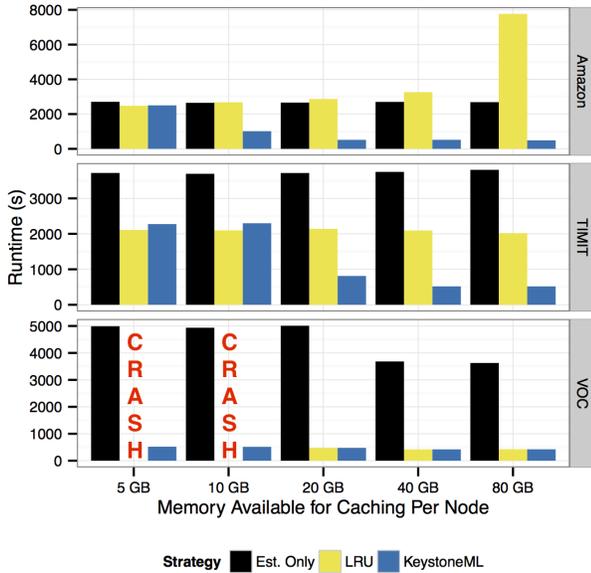


Fig. 10: The KeystoneML caching strategy outperforms a rule-based and LRU caching strategy at many levels of memory constraints and responds well to memory pressure.

L-BFGS solver turns out to be optimal. The performance gains come from caching intermediate features just before the L-BFGS solve. For the TIMIT pipeline, run with 16k features, we see that the end-to-end optimizations only give a $1.3\times$ speedup but that selecting the appropriate solver results in a $8\times$ speedup over the baseline. Finally in the VOC pipeline the whole pipeline optimization gives around $3\times$ speedup. Operator-level optimization chooses good PCA, GMM and solver operators resulting in a $12\times$ improvement over the baseline, or $15\times$ if we amortize the optimization costs across many runs of a similar pipeline. Optimization overheads are insignificant except for the VOC pipeline. This dataset has relatively few examples, so the sampling strategy takes more time relative to the other datasets.

D. Automatic Materialization Strategies

As discussed in Section IV, one key optimization enabled by KeystoneML’s ability to capture the complete application DAG to dynamically determine where to materialize reused intermediate objects, particularly in the presence of memory constraints. In Figure 10 we demonstrate the effectiveness of the greedy caching algorithm proposed in Section IV. Since the algorithm needs local profiles of each node’s performance, we measured each node’s running time on two samples of 512 and 1024 examples. We then extrapolate the node’s memory usage and runtime to full scale. We found that memory estimates from this process are highly accurate and runtime estimates were within 15% of actual runtimes. If estimates are inaccurate, we fall back to an LRU replacement policy for the cache set determined by this procedure. This process is imperfect, but is adequate at identifying relative running times and is sufficient for our purpose of resource management.

We compare this strategy with two alternatives—the first is a simple rule-based approach which only caches the results of Estimators. This is a sensible rule to follow, as the result of an Estimator (a Transformer or model) is computationally expensive to acquire and typically holds a small memory footprint. However, this is not sufficient for most practical pipelines because if a pipeline contains more than one Estimator, often the input to the first Estimator will be used downstream, thus presenting an opportunity for reuse. The second approach is the standard Least Recently Used (LRU) policy.

From Figure 10 we notice several important trends. First, the KeystoneML strategy is nearly always better than either of the other strategies. In the unconstrained case, the algorithm is going to remember all reused items as late in their journey through the pipeline as possible. In the constrained case, it will do as least as well as remembering the (small) estimators which are by definition reused later in the pipeline. Additionally, the strategy degrades effectively, mixing between the best performance of the limited-memory rule-based strategy and the LRU based “cache everything” strategy which works well in unconstrained settings. As we increased the memory available to caching per-node, the LRU strategy performed *worse* for the Amazon pipeline. Upon further investigation, this is because LRU does not take into account the cost of materializing an object and so computationally expensive objects may be evicted by larger objects at larger cache sizes.

To give a concrete example of the optimizer in action, consider the VOC pipeline shown in Figure 5 in Section II. When memory is not constrained, the outputs from the SIFT, ReduceDimensions, Normalize and TrainingLabels are cached. When memory is restricted, only the output from Normalize and TrainingLabels are cached. These results show the importance of both per-operator and whole-pipeline optimizations.

E. Scalability

As discussed in previous sections, KeystoneML’s API design encourages the construction of scalable operators. However, some estimators like linear solvers need coordination [25] among workers to compute correct results. In Figure 11 we demonstrate the scaling properties from 8 to 128 nodes of the text, image, and Kernel SVM pipelines on the Amazon, ImageNet (with 16k features) and TIMIT datasets (with 65k features) respectively. The ImageNet pipeline exhibits near-perfect horizontal scalability up to 128 nodes, while the Amazon and TIMIT pipeline scale well up to 64 nodes.

To understand why the Amazon and TIMIT pipeline do not scale linearly to 128 nodes, we further analyze the breakdown of time take by each stage. We see that each pipeline is dominated by a different part of its computation. The TIMIT pipeline is dominated by its solve stage, while featurization dominates the Amazon and ImageNet pipelines. Scaling linear solvers is known to require coordination [25], which leads directly to sub-linear scalability of the whole pipeline. Similarly, in the Amazon pipeline, one of the featurization steps uses an aggregation tree which does not scale linearly.

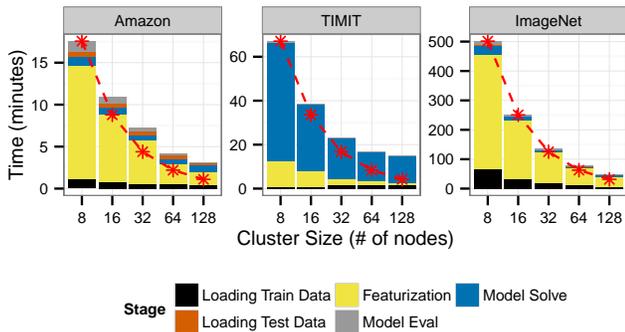


Fig. 11: Time breakdown of workloads by stage. The red line indicates ideal strong scaling performance over 8 nodes.

VI. RELATED WORK

ML Frameworks: ML researchers have traditionally used MATLAB or R packages to develop ML routines. The importance of feature engineering has led to tools like scikit-learn [12] and KNIME [50] adding support for featurization for small datasets. Further, existing libraries for large scale ML [51] like Vowpal Wabbit [6], GraphLab [52], MLlib [7], RIOT [53], DimmWitted [14] focus on efficient implementations of learning algorithms like regression, classification and linear algebra routines. In KeystoneML, we focus on pipelines that include featurization and show how to optimize performance with end-to-end information. Work in Parameter Servers [54] has studied how to share model updates. In KeystoneML we implement a high-level API for linear solvers and can leverage parameter servers in our architecture.

Closely related to KeystoneML is SystemML [11] which also uses an optimization based approach to determine the physical execution strategy of ML algorithms. However, SystemML places less emphasis on support for UDFs and featurization, while instead focusing on linear algebra operators which have well specified semantics. To handle featurization we develop an extensible API in KeystoneML which allows for cost profiling of arbitrary nodes and uses these cost estimates to make node-level and whole-pipeline optimizations. Other work [33], [29] has looked at optimizing caching strategies and operator selection in the regime of feature selection and feature generation workloads. KeystoneML considers similar problems in the context of distributed ML operators and end-to-end learning pipelines. Developed concurrently to KeystoneML is TensorFlow [13]. While designed to support different learning workloads the optimizations that are a part of KeystoneML can also be applied to systems like TensorFlow.

Projects such as Bismarck [21], MADLib [10], and GLADE [55] have proposed techniques to integrate ML algorithms inside database engines. In KeystoneML, we develop a high level API and show how we can achieve similar benefits of modularity and end-to-end optimization while also being scalable. These systems do not present cross-operator optimizations and do not consider tradeoffs at the operator level that we consider in KeystoneML. Finally, Spark ML [56] represents an early design of a similar high-level API for

machine learning. We present a type safe API and optimization framework for such a system. The version we present in this paper differs in its use of type-safe operations, support for complex data flows, internal DAG representation and optimizations discussed in Sections III and IV.

Query Optimization, Modular Design, Caching: There are several similarities between the optimizations made by KeystoneML and traditional relational query optimizers. Even the earliest relational query optimizers used multiple physical implementations of equivalent logical operators, and like many relational optimizers, the KeystoneML optimizer is cost-based. However, KeystoneML supports a much richer set of data types than a traditional relational query system, and our operators lack some relational algebra semantics, such as commutativity, limiting the system’s ability to perform certain optimizations. Further, KeystoneML switches among operators that provide exact answers vs approximate ones to save time due to the workload setting. Data characteristics such as sparsity are not traditionally considered by optimizers.

The caching strategy employed by KeystoneML can be viewed as a form of view selection for materialized view maintenance over queries with expensive user-defined functions [31], we focus on materialization for intra-query optimization, as opposed to inter-query optimization [57], [32], [58], [59]. While much of the related work focuses on the challenging problem of view maintenance in the presence of updates, KeystoneML the immutable properties of this state.

VII. FUTURE WORK AND CONCLUSION

KeystoneML represents a significant first step towards easy-to-use, robust, and efficient end-to-end ML at massive scale. We plan to investigate pipeline optimizations like node reordering to reduce data transfers and also look at how hyperparameter tuning [60] can be integrated into the system. The existing KeystoneML operator APIs are synchronous and our existing pipelines are acyclic. In the future we plan to study how algorithms like asynchronous SGD [54] or back-propagation can be integrated with the robustness and scalability that KeystoneML provides.

We have presented the design of KeystoneML, a system that enables the development end-to-end ML pipelines. By capturing the end-to-end application, KeystoneML can automatically optimize execution at both the operator and whole-pipeline levels, enabling solutions that automatically adapt to changes in data, hardware, and other environmental characteristics.

Acknowledgements: We would like to thank Xiangrui Meng, Joseph Bradley for their help in design discussions and Henry Milner, Daniel Brucker, Gylfi Gudmundsson, Zongheng Yang, Vaishaal Shankar for contributions to the KeystoneML source code. We would also like to thank Florent Perronin, Jorge Sanchez, Thomas Mensink for their help in developing an image classification pipeline and Vikas Sindhwani for providing us with the TIMIT dataset. Peter Alvaro, Peter Bailis, Joseph Gonzalez, Eric Jonas, Sanjay Krishnan, Nick Lanham, Kay Ousterhout, Aurojit Panda, Ameet Talwarkar, Stephen Tu, Eugene Wu provided feedback on earlier drafts. This research

is supported in part by NSF CISE Expeditions Award CCF-1139158, DOE Award SN10040 DE-SC0012463, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, IBM, SAP, The Thomas and Stacey Siebel Foundation, Adatao, Adobe, Apple, Inc., Blue Goji, Bosch, Cisco, Cray, Cloudera, EMC2, Ericsson, Facebook, Guavus, HP, Huawei, Informatica, Intel, Microsoft, NetApp, Pivotal, Samsung, Schlumberger, Splunk, Virdata and VMware.

REFERENCES

- [1] C. Manning and D. Klein, "Optimization, maxent models, and conditional estimation without magic," in *HLT-NAACL, Tutorial Vol 5*, 2003.
- [2] P.-S. Huang, H. Avron, T. N. Sainath *et al.*, "Kernel methods match deep neural networks on timit," in *ICASSP*. IEEE, 2014, pp. 205–209.
- [3] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.
- [4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget *et al.*, "The Kaldi speech recognition toolkit," 2011.
- [5] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.
- [6] J. Langford, L. Li, and A. Strehl, "Vowpal wabbit online learning project," 2007.
- [7] X. Meng, J. K. Bradley, B. Yavuz, E. R. Sparks *et al.*, "MLlib: Machine Learning in Apache Spark," *CoRR*, vol. abs/1505.06807, 2015.
- [8] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.
- [9] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *Intelligent Systems, IEEE*, vol. 24, no. 2, pp. 8–12, 2009.
- [10] J. M. Hellerstein, C. Ré, F. Schoppmann, D. Z. Wang *et al.*, "The MADlib analytics library: or MAD skills, the SQL," *PVLDB*, vol. 5, no. 12, pp. 1700–1711, 2012.
- [11] A. Ghoting, R. Krishnamurthy, E. Pednault, B. Reinwald *et al.*, "SystemML: Declarative machine learning on MapReduce," in *ICDE*, 2011.
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.
- [13] M. Abadi, A. Agarwal, P. Barham, E. Brevdo *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <http://tensorflow.org/>
- [14] C. Zhang and C. Ré, "Dimmwit: A study of main-memory statistical analytics," *PVLDB*, vol. 7, no. 12, pp. 1283–1294, 2014.
- [15] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," in *NIPS*, 2011, pp. 693–701.
- [16] A. Crotty, A. Galakatos, and T. Kraska, "Tupeware: Distributed machine learning on small clusters," *IEEE Data Eng. Bull.*, vol. 37, no. 3, 2014.
- [17] T. Kraska, A. Talwalkar, J. C. Duchi, R. Griffith *et al.*, "MLbase: A Distributed Machine-learning System," *CIDR*, 2013.
- [18] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference*, 2011.
- [19] E. Jonas, V. Shankar, M. Bobra, and B. Recht, "Flare prediction using photospheric and coronal image data," *AGU Fall Meeting*, 2016.
- [20] "ENCODE-DREAM in-vivo Transcription Factor Binding Site Prediction Challenge," <https://www.synapse.org/#!Synapse:syn6131484>, 2016.
- [21] X. Feng, A. Kumar, B. Recht, and C. Ré, "Towards a unified architecture for in-rdbms analytics," in *SIGMOD*, 2012.
- [22] M. Stonebraker, J. Becla, D. J. DeWitt, K.-T. Lim, D. Maier, O. Ratzesberger, and S. B. Zdonik, "Requirements for science data bases and scidb," in *CIDR*, vol. 7, 2009, pp. 173–184.
- [23] S. Williams, A. Waterman, and D. Patterson, "Roofline: An Insightful Visual Performance Model for Multicore Architectures," *CACM*, 2009.
- [24] G. M. Ballard, "Avoiding communication in dense linear algebra," Ph.D. dissertation, University of California, Berkeley, 2013.
- [25] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou, "Communication-optimal parallel and sequential QR and LU factorizations," *SIAM Journal on Scientific Computing*, vol. 34, no. 1, pp. A206–A239, 2012.
- [26] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., 1989.
- [27] W. Chen, Z. Wang, and J. Zhou, "Large-scale l-bfgs using mapreduce," in *NIPS*, 2014, pp. 1332–1340.
- [28] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Review*, 2011.
- [29] F. Abuzaid, S. Hadjis, C. Zhang, and C. Ré, "Caffe con Troll: Shallow Ideas to Speed Up Deep Learning," *CoRR abs/1504.04343*, 2015.
- [30] M. Mathieu, M. Henaff, and Y. LeCun, "Fast Training of Convolutional Networks through FFTs," *ICLR*, 2014.
- [31] R. Chirkova and J. Yang, "Materialized Views," *Foundations and Trends in Databases*, 2012.
- [32] D. C. Zilio, C. Zuzarte, G. M. Lohman, H. Pirahesh *et al.*, "Recommending Materialized Views and Indexes with IBM DB2 Design Advisor," in *ICAC 2004*, May 2004.
- [33] C. Zhang, A. Kumar, and C. Ré, "Materialization optimizations for feature selection workloads," in *SIGMOD*, 2014.
- [34] L. A. Belady, "A study of replacement algorithms for a virtual-storage computer," *IBM Systems journal*, vol. 5, no. 2, pp. 78–101, 1966.
- [35] P. Raghavan and M. Snir, "Memory versus randomization in on-line algorithms," in *ICALP*. Springer, 1989, pp. 687–703.
- [36] A. Agarwal, O. Chapelle, and J. Langford, "A reliable effective terascale linear learning system," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1111–1133, 2014.
- [37] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel *et al.*, "The Stanford CoreNLP natural language processing toolkit," in *ACL*, 2014.
- [38] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *KDD*, 2015, pp. 785–794.
- [39] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," <https://goo.gl/m68USr>, 2003.
- [40] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *NIPS*, 2007, pp. 1177–1184.
- [41] V. Sindhvani and H. Avron, "High-performance kernel machines with implicit distributed optimization and randomization," *CoRR*, vol. abs/1409.0940, 2014.
- [42] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, vol. 2. IEEE, 1999, pp. 1150–1157.
- [43] A. Krizhevsky and G. Hinton, "Convolutional Deep Belief Networks on CIFAR-10," *Unpublished manuscript*, 2010.
- [44] J. Sanchez, F. Perronnin, and T. Mensink, "Improved fisher vector for large scale image classification," http://image-net.org/challenges/LSVRC/2010/ILSVRC2010_XRCE.pdf.
- [45] "TensorFlow CIFAR-10 Performance as reported in TensorFlow Source Code," <https://git.io/v2b4J>.
- [46] J. Chen, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous sgd," *arXiv preprint arXiv:1604.00981*, 2016.
- [47] A. Coates and A. Y. Ng, "Learning Feature Representations with K-Means," in *Neural Networks: Tricks of the Trade*, 2012.
- [48] C. Szegedy, V. Vanhoucke *et al.*, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.
- [49] S. Abu-El-Hajja, N. Kothari *et al.*, "YouTube-8M: A Large-Scale Video Classification Benchmark," *arXiv preprint arXiv: 1609.08675*, 2016.
- [50] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel *et al.*, "KNIME: The Konstanz information miner," in *Data analysis, machine learning and applications*. Springer, 2008, pp. 319–326.
- [51] Z. Cai, Z. J. Gao, S. Luo, L. L. Perez *et al.*, "A comparison of platforms for implementing and running very large scale machine learning algorithms," in *SIGMOD 2014*, 2014, pp. 1371–1382.
- [52] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin *et al.*, "Distributed graphlab: a framework for machine learning and data mining in the cloud," *PVLDB*, vol. 5, no. 8, pp. 716–727, 2012.
- [53] Y. Zhang, H. Herodotou, and J. Yang, "RIOT: I/O-Efficient Numerical Computing without SQL," in *CIDR*, 2009.
- [54] M. Li, D. G. Andersen, J. W. Park, A. J. Smola *et al.*, "Scaling Distributed Machine Learning with the Parameter Server." *OSDI*, 2014.
- [55] C. Qin and F. Rusu, "Scalable I/O-bound Parallel Incremental Gradient Descent for Big Data Analytics in GLADE," in *DanaC*, 2013.
- [56] X. Meng, J. Bradley, E. Sparks, and S. Venkataraman, "ML Pipelines: A New High-Level API for MLlib," <https://goo.gl/pluhq0>, 2015.
- [57] S. Chaudhuri and V. R. Narasayya, "AutoAdmin 'What-if' Index Analysis Utility." *SIGMOD*, 1998.
- [58] I. Elghandour and A. Aboulnaga, "ReStore: reusing results of MapReduce jobs," in *PVLDB*, 2012.
- [59] L. Perez and C. Jermaine, "History-aware query optimization with materialized intermediate views," in *ICDE*, March 2014, pp. 520–531.
- [60] E. R. Sparks, A. Talwalkar, D. Haas, M. J. Franklin *et al.*, "Automating model search for large scale machine learning," in *SoCC '15*, 2015.