

Scaling Datalog for Machine Learning on Big Data

Yingyi Bu, Vinayak Borkar, Michael J. Carey
University of California, Irvine

Joshua Rosen, Neoklis Polyzotis
University of California, Santa Cruz

Tyson Condie, Markus Weimer, Raghu Ramakrishnan
Yahoo! Research

ABSTRACT

In this paper, we present the case for a declarative foundation for data-intensive machine learning systems. Instead of creating a new system for each specific flavor of machine learning task, or hard-coding new optimizations, we argue for the use of *recursive queries* to program a variety of machine learning systems. By taking this approach, database query optimization techniques can be utilized to identify effective execution plans, and the resulting runtime plans can be executed on a single *unified* data-parallel query processing engine. As a proof of concept, we consider two programming models—Pregel and Iterative Map-Reduce-Update—from the machine learning domain, and show how they can be captured in Datalog, tuned for a specific task, and then compiled into an optimized physical plan. Experiments performed on a large computing cluster with real data demonstrate that this declarative approach can provide very good performance while offering both increased generality and programming ease.

1. INTRODUCTION

Supported by the proliferation of “Big Data” platforms such as Hadoop, organizations are collecting and analyzing ever larger datasets. Increasingly, machine learning (ML) is at the core of data analysis for actionable business insights and optimizations. Today, machine learning is deployed widely: recommender systems drive the sales of most online shops; classifiers help keep spam out of our email accounts; computational advertising systems drive revenues; content recommenders provide targeted user experiences; machine-learned models suggest new friends, new jobs, and a variety of activities relevant to our profile in social networks. Machine learning is also enabling scientists to interpret and draw new insights from massive datasets in many domains, including such fields as astronomy, high-energy physics, and computational biology.

The availability of powerful distributed data platforms and the widespread success of machine learning has led to a virtuous cycle wherein organizations are now investing in gathering a wider range of (even bigger!) datasets and addressing an even broader range of tasks. Unfortunately, the basic MapReduce framework commonly provided by first-generation “Big Data analytics” platforms like Hadoop lacks an essential feature for machine learning: MapReduce does not support iteration (or equivalently, recursion) or certain key features required to efficiently iterate “around” a MapReduce program. Programmers building ML models on such systems are forced to implement looping in ad-hoc ways outside the core MapReduce framework; this makes their programming task much harder, and it often also yields inefficient programs in the end. This lack of support has motivated the recent development of various specialized approaches or libraries to support iterative programming on large clusters. Examples include Pregel,

Spark, and Mahout, each of which aims to support a particular family of tasks, e.g., graph analysis or certain types of ML models, efficiently. Meanwhile, recent MapReduce extensions such as HaLoop, Twister, and PrItr aim at directly addressing the iteration outage in MapReduce; they do so at the physical level, however.

The current generation of specialized platforms seek to improve a user’s programming experience by making it much easier (relative to MapReduce) to express certain classes of parallel algorithms to solve ML and graph analytics problems over Big Data. Pregel is a prototypical example of such a platform; it allows problem-solvers to “think like a vertex” by writing a few user-defined functions (UDFs) that operate on vertices, which the framework can then apply to an arbitrarily large graph in a parallel fashion. Unfortunately for both their implementors and users, each such platform is a distinct new system that has been built from the ground up. Ideally, a specialized platform should allow for better optimization strategies for the class of problems considered “in scope.” In reality, however, each new system is built from scratch and must include efficient components to perform common tasks such as scheduling and message-passing between the machines in a cluster. Also, for Big Data problems involving multiple ML algorithms, it is often necessary to somehow glue together multiple platforms and to pass (and translate) data via files from one platform to another. It would clearly be attractive if there were a common, general-purpose platform for data-intensive computing available that could simultaneously support the required programming models and allow various domain-specific systems the ability to reuse the common pieces. Also desirable would be a much cleaner separation between the logical specification of a problem’s solution and the physical runtime strategy to be employed; this would allow alternative runtime strategies to be considered for execution, thus leading to more efficient executions of different sorts of jobs.

In this paper, we show that it is indeed possible to provide a declarative framework capable of efficiently supporting a broad range of machine learning and other tasks that require iteration, and then to develop specialized programming models that target specific classes of ML tasks on top of this framework. Hence, much of the effort that is currently involved in building such specialized systems is factored out into a single underlying optimizer and runtime system. We propose the use of Datalog, which allows recursive queries to be naturally specified, as the common declarative language “under the hood”, into which we map high-level programming models.¹ Moreover, Datalog can readily express the

¹We leave open the possibility of exposing Datalog as an “above the hood” user-programmable language; doing so would place a premium upon being able to optimize arbitrary Datalog programs in a “Big Data” cluster environment, which our current results do not yet address.

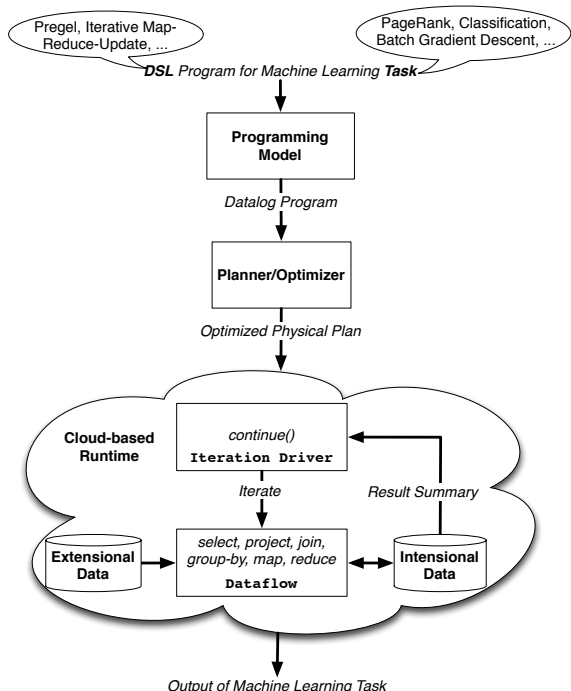


Figure 1: System stack for large-scale Machine Learning.

dataflow aspects of a ML program, which are the main drivers of cost in a Big Data setting. Our approach opens the door to applying the rich body of work on optimizing Datalog programs and identifying effective execution plans, and allows us to execute the resulting plans on a single *unified* data-parallel query processing engine over a large cluster.

Figure 1 sketches the approach advocated here. A domain-specific programming model—such as Pregel or Iterative Map-Reduce-Update—is used to program a specific ML task, e.g., PageRank [25] or Batch Gradient Descent. The task program is then translated to a Datalog program that captures the intended programming model as declarative rule specifications and the task-specific code (e.g., the PageRank algorithm) as UDFs. Subsequently, a planner/optimizer compiles the declarative Datalog program into an efficient execution plan. Lastly, a cloud-based runtime engine—consisting of a dataflow of data-parallel operators, extensional and intensional datasets, and an iteration driver—executes the optimized plan to a fixed point, producing the output of the ML task. Central to our thesis is that by capturing the ML programming model in a high-level declarative language, we can automatically generate physical plans that are optimized—based on hardware configurations and data statistics—for a target class of ML tasks.

As a concrete example of the benefits of our approach, consider Pregel again, a specialized programming model and runtime tuned to graph-oriented computations. Suppose a data scientist wants to train a model through a Batch Gradient Descent (BGD) task using Pregel, which requires them to “think like a vertex.” A possible approach would be to encode each data point as a vertex that computes a (gradient, loss) value and sends it to a global aggregator, which sums all the values to a global statistic and updates the model. This process repeats over a series of “supersteps” until the model converges to some value. There are two problems with this approach. Firstly, it is unnatural to treat the training data—

a set of unrelated feature vectors—as a graph. Secondly, encoding a BGD task into a general purpose graph-oriented programming model is suboptimal, in terms of programming ease and runtime performance. For more appropriate graph-analysis tasks, like PageRank, the Pregel programming model and runtime contains hardcoded features—such as non-monotonic halting conditions—that are often not required. The net result is that Pregel is suitable only for a specific class of ML tasks, and not appropriate or suboptimal for others. In contrast, we propose capturing an ML task as a declarative Datalog program, and then letting a query optimizer translate it to an appropriate physical plan.

To summarize, in our proposed approach to scalable machine learning, programmers need not learn to operate and use a plethora of distinct platforms. Relational database systems separate the conceptual, logical and physical schemas in order to achieve *logical* and *physical data independence*. Similarly, we open the door to principled optimizations by achieving a separation between:

- The user’s program (in a sublanguage of their choice, making use of a library of available templates and user-definable functions) and the underlying *logical query*, expressed in Datalog. This shields the user from any changes in the logical framework, e.g., how the Datalog program is optimized.
- The logical Datalog query and an optimized *physical runtime plan*, reflecting details related to caching, storage, indexing, the logic for incremental evaluation and re-execution in the face of failures, etc. This ensures that any enhancements to the plan execution engine will automatically translate to more efficient execution, without requiring users to re-program their tasks to take advantage of the enhancements.

In essence, the separation identifies “modules” (such as the plan execution engine or the optimization of the logical Datalog program) where localized enhancements lead to higher overall efficiency. To illustrate our approach, we will show here how the Pregel and Iterative Map-Reduce-Update programming models can each be translated into Datalog programs. Second, we will demonstrate that an appropriately chosen data-intensive computing substrate, namely Hyracks [7], is able to handle the computational requirements of such programs through the application of dataflow processing techniques like those used in parallel databases [15]. This demonstration involves the presentation of experimental results obtained by running the sorts of Hyracks jobs that will result from our translation against real data on a large computational cluster at Yahoo!. Our findings indicate that such a declarative approach can indeed provide very good performance while offering increased generality and ease of programming.

The remainder of this paper is organized as follows: Section 2 provides a brief data-centric perspective on machine learning and then reviews two popular programming models in use today for scalable machine learning. Section 3 shows how programs written against these two programming models can be captured in Datalog and then translated into an extended relational algebra, while Section 4 describes the resulting physical plans. Section 5 presents preliminary experimental results obtained by running the physical Hyracks plans for two tasks—Batch Gradient Descent and PageRank—on a large research cluster at Yahoo! against real datasets. Section 6 relates this work to other activities in this area, past and present, and Section 7 presents our conclusions.

2. PROGRAMMING MODELS FOR ML

The goal of machine learning (ML) is to turn observational data into a *model* that can be used to predict for or explain yet unseen data. While the range of machine learning techniques is broad,

most can be understood in terms of three complementary perspectives:

- **ML as Search:** The process of training a model can be viewed as a *search problem*. A domain expert writes a program with an objective function that contains, possibly millions of, unknown parameters, which together form the *model*. A runtime program *searches* for a good set of parameters based on the objective function. A *search strategy* enumerates the parameter space to find a model that can correctly capture the known data and accurately predict unknown instances.
- **ML as Iterative Refinement:** Training a model can be viewed as iteratively closing the gap between the model and underlying reality being modeled, necessitating iterative/recursive programming.
- **ML as Graph Computation:** Often, the interdependence between the model parameters is expressible as a graph, where nodes are the parameters (e.g., statistical/random variables) and edges encode interdependence. This view gives rise to the notion of *graphical models*, and algorithms often reflect the graph structure closely (e.g., propagating refinements of parameter estimates iteratively along the edges, aggregating the inputs at each vertex).

Interestingly, each of these perspectives lends itself readily to expression as a Datalog program, and as we argue in this paper, thereby to efficient execution by applying a rich array of optimization techniques from the database literature. The fact that Datalog is well-suited for iterative computations and for graph-centric programming is well-known [26], and it has also been demonstrated that Datalog is well-suited to search problems [12]. The natural fit between Datalog and ML programming has also been recognized by others [6, 16], but not at “Big Data” scale. It is our goal to make the optimizations and theory behind Datalog available to large-scale machine learning while facilitating the use of established programming models. To that end, we advocate *compilation* from higher order programming models to Datalog and subsequently physical execution plans.

To study the feasibility of this approach, we show how two major programming models supporting distributed machine learning today—Pregel and Iterative Map-Reduce-Update—can be expressed in Datalog and subsequently efficiently executed. In the remainder of this Section, we describe these two programming models in more detail with the goal of isolating the user code in terms of user-defined functions (UDFs). This will set the stage for the next section, in which we present concise Datalog representations for these two ML programming models, reusing the UDFs introduced here. As we will then see, the structure of many ML problems is inherently recursive.

2.1 Pregel

Pregel [22] is a system developed at Google for supporting graph analytics. It exposes a message-passing interface in the form of two per-vertex UDFs:

update the per-vertex update function. It accepts the current vertex state and inbound messages and produces outbound messages as well as an updated state.

combine (optional) aggregates messages destined for a vertex.

We omit other aspects of Pregel—graph mutation and global aggregators—because they are not necessary in many graph algorithms [22], and the machinery for global aggregators is captured later when we address Iterative Map-Reduce-Update.

The Pregel runtime executes a sequence of iterations called *supersteps* through a bulk-synchronous processing (BSP) model. In a single superstep, the Pregel runtime executes the `update` UDF on all *active* vertices exactly once. A vertex is active in the current superstep if there are messages destined for it or if the `update` UDF indicates—in the previous superstep—its desire to execute. The output of a superstep can be materialized for fault tolerance before executing the subsequent superstep. The runtime halts when no vertices are active.

Example: PageRank [25] is a canonical example of a graph algorithm that is concisely captured by Pregel. Websites are represented by vertices and hyperlinks form the edges of the graph. In a single superstep, the `update` UDF receives the PageRank of the current vertex and its neighbors. It emits the updated PageRank for this vertex and if its PageRank changed sufficiently, its new value is sent to its neighbors. The system converges when no more such updates occur or a maximum number of supersteps is reached.

2.2 Iterative Map-Reduce-Update

A large class of machine learning algorithms are expressible in the statistical query model [20]. Statistical queries (e.g. `max`, `min`, `sum`, ...) themselves decompose into a data-local `map` function and a subsequent aggregation using a `reduce` function [10], where `map` and `reduce` refer to the functions by the same name from the functional programming literature. We support the resulting Iterative Map-Reduce-Update programming model through the following three UDFs:

map receives read-only global state as side information and is applied to all training data points in parallel.

reduce aggregates the `map`-output. This function is commutative and associative.

update receives the combined aggregated results and produces a new global state for the next iteration or indicates that no additional iteration is necessary.

An Iterative Map-Reduce-Update runtime executes a series of iterations, each of which first calls `map` with the required arguments, then performs a global `reduce` aggregation, and lastly makes a call to `update`. We assume the runtime terminates when `update` returns the same model that it was given.² It is interesting to point out here that Google’s MapReduce [13] programming model is not an ideal fit: it contains a group-by key component that is not needed for many statistical queries from the machine learning domain. Additionally, we require an iterative looping construct and an update step that fall outside the scope of Google’s MapReduce framework.

Example: Convex Optimization A large class of machine learning—including Support Vector Machines, Linear and Logistic Regression and structured prediction tasks such as machine translation—can be cast as convex optimization problems, which in turn can be solved efficiently using an Iterative Map-Reduce-Update approach [1, 28]. The objective is to minimize the sum over all data points of the divergences (the loss) between the model’s prediction and the known data. Usually, the loss function is convex and differentiable in the model, and therefore the *gradient* of the loss function can be used in iterative optimization algorithms such as Batch Gradient Descent.³ Each model update step is a single Map-Reduce-Update iteration. The `map` UDF computes (loss, gradient) tuples for all data points, using the current model as side

²Alternatively, a vote to halt protocol could be simulated through a boolean value in the model that signals the termination.

³A more detailed discussion can be found in the Appendix A.

input. The `reduce` UDF sums those up and `update` updates the model. The updated model becomes the input of the next iteration of the optimization algorithm.

2.3 Discussion

From a data flow perspective, a key differentiator between different types of ML models is the relationship of the model to the observational data; both in size and structure. Certain models (e.g., regression, classification and clustering) are *global* to all observation data points and are relatively small in size (think MB vs. GB). In others (e.g., topic models and matrix factorization), the model consists of independent parameters that are *local* to each observation and are therefore on the same order-of-magnitude in terms of size as the observational data. Any system that seeks to support both classes of ML models efficiently must recognize the nature of the task—global or local—and be able to optimize accordingly.

The two programming frameworks that we consider span a wide area of machine learning and graph analytics. Pregel is a well-known graph analytics platform that can be used to develop local models. Iterative Map-Reduce-Update is gaining traction as an ideal framework for producing global models. It is important to point out that both frameworks can express each other—Pregel can be implemented on top of an Iterative Map-Reduce-Update system and vice versa—but that each system was designed and optimized for a specific “native” application type. Abusing one to solve the other would incur significant performance overheads. Instead, in Section 3, we unify these frameworks and the intended semantics as declarative specifications written in the Datalog language. We then show a direct translation from the Datalog specifications to a data-parallel recursive runtime that is able to retain the performance gains offered by runtimes specifically tuned to a given framework.

3. DECLARATIVE REPRESENTATION

This section presents a translation of the two programming models into declarative Datalog programs and a formal analysis of the semantics and correctness of this translation. In doing so, we expose information about the semantics and structure of the underlying data operations, which in turn allows us to reason about possible optimizations (e.g., reordering the operators or mapping logical operators to different possible implementations) and thus generate efficient execution plans over a large range of configurations. Datalog is a natural choice for this intermediate logical representation, as it can encode succinctly the inherent recursion of the algorithms.

Before diving into the details of the translation of the two programming models, we present a short overview of the main concepts in Datalog. A Datalog program consists of a set of *rules* and an optional *query*. A Datalog rule has the form $p(\mathbf{Y}) :- q_1(\mathbf{X}_1), \dots, q_n(\mathbf{X}_n)$, where p is the head predicate of the rule, q_1, \dots, q_n are called the body predicates, and $\mathbf{Y}, \mathbf{X}_1, \dots, \mathbf{X}_n$ correspond to lists of variables and constants. Informally, a Datalog rule reads “if there is an assignment of values $\mathbf{v}, \mathbf{v}_1, \dots, \mathbf{v}_n$ corresponding to $\mathbf{Y}, \mathbf{X}_1, \dots, \mathbf{X}_n$ such that $q_1(\mathbf{v}_1) \wedge \dots \wedge q_n(\mathbf{v}_n)$ is true **then** $p(\mathbf{v})$ is true.” In the rules that we consider, a predicate can be one of three types:

- An extensional predicate, which maps to the tuples of an existing relation. An extensional predicate $q_i(\mathbf{v})$ is true if and only if the tuple \mathbf{v} is present in the corresponding relation.
- An intensional predicate, which corresponds to the head p of a rule. Intensional predicates essentially correspond to views.
- A function predicate, which corresponds to the application of a function. As an example, consider a function f that receives as input three datums and outputs a tuple of two datums, and

assume that the corresponding function predicate is q_f . We say that $q_f(v_1, v_2, v_3, v_4, v_5)$ is true if and only if the result of $f(v_1, v_2, v_3)$ is (v_4, v_5) . By convention, we will always designate the first attributes of the predicate as the inputs to the function and the remaining attributes as the output.

We allow group-by aggregation in the head in the form $p(Y, \text{aggr}\langle Z \rangle)$. As an example, the rule $p(Y, \text{SUM}\langle Z \rangle) :- q_1(Y, Z)$ will compute the sum of Z values in q_1 grouped-by Y . We also allow variables to take set values and provide a mechanism for member iteration. As an example, the rule $p(X, Y) :- q_1(X, \{Y\})$ implies that the second attribute of q_1 takes a set value, and binds Y to every member of the set in turn (generating a tuple in p per member, essentially unnesting the set).

Recursion in Datalog is expressed by rules that refer to each other in a cyclic fashion. The order that the rules are defined in a program is semantically immaterial. Program evaluation proceeds bottom-up, starting from the extensional predicates and inferring new facts through intensional and function predicates. The evaluation of a Datalog program reaches a *fixpoint* when no further deductions can be made based on the currently inferred facts [26].

3.1 Pregel for Local Models

We begin with the Datalog program in Listing 1, which specifies the Pregel programming model as it pertains to deriving local models. A special temporal argument (the variable J) is used to track the current superstep number, which will be passed to the `update` UDF invocation in Rule *L6* discussed below. Rule *L1* invokes an initialization UDF `init_vertex` (which accepts the `(Id, Datum)` variables as argument and returns the `(State)` variable) on every tuple in the input: referenced by the `data` predicate. Rule *L2* then initializes a `send` predicate with an activation message to be delivered to all vertices in iteration zero. Rule *L3* implements the combination of messages that are destined for the same vertex. It performs a group-by aggregation over predicate `send`, using the `combine` aggregate function (which is itself a proxy for `combine`, explained in Section 2.1). In Pregel, a vertex may forgo updating the state of a given vertex or global aggregator for some period of supersteps. Rules *L4* and *L5* maintain a view of the most recent vertex state via the `local` predicate.

Rule *L6* implements the core logic in a superstep by matching the collected messages with the target local state and then evaluates the function predicate `update` (which corresponds to UDF `update`). The $(J, Id, InState, InMsgs)$ variables represent the arguments and the $(OutState, OutMsgs)$ variables hold the return values: a new state and set of outbound messages.

Finally, rules *L7* and *L8* stage the next superstep: *L7* updates the state of each vertex, and *L8* forwards outbound messages to the corresponding vertices. Note that the body of *L7* is conditioned on a non-null state value. This allows vertices to forgo state updates in any given superstep. Finally, the vote to halt protocol is implemented in the `update` UDF, which produces a special “self” message that activates the vertex in the next superstep.

3.2 Iterative Map-Reduce-Update

The Datalog program in Listing 2 specifies the Iterative Map-Reduce-Update programming model. Like before, a special temporal variable (J) is used to track the iteration number. Rule *G1* performs initialization of the global model at iteration 0 through function predicate `init_model`, which takes no arguments and returns the initial model in the (M) variable. Rules *G2* and *G3* implement the logic of a single iteration. Let us consider first rule *G2*. The evaluation of `model(J, M)` and `training_data(Id, R)` binds

Listing 1: Datalog program for the Pregel programming model. The temporal argument is defined by the J variable.

```

1  % Initialize vertex state
2  L1: vertex(0, Id, State) :-
3     data(Id, Datum),
4     init_vertex(Id, Datum, State).

6  % Initial vertex message to start iteration 0.
7  L2: send(0, Id, ACTIVATION_MSG) :-
8     vertex(0, Id, _).

10 % Compute and aggregate all messages.
11 L3: collect(J, Id, combine<Msg>) :-
12     send(J, Id, Msg).

14 % Most recent vertex timestamp
15 L4: maxVertexJ(Id, max<J>) :-
16     vertex(J, Id, State).

18 % Most recent vertex local state
19 L5: local(Id, State) :-
20     maxVertexJ(Id, J), vertex(J, Id, State).

22 % new state and outbound messages.
23 L6: superstep(J, Id, OutState, OutMsgs) :-
24     collect(J, Id, InMsgs),
25     local(Id, InState),
26     update(J, Id, InState, InMsgs, OutState, OutMsgs).

28 % Update vertex state for next superstep.
29 L7: vertex(J+1, Id, State) :-
30     superstep(J, Id, State, _),
31     State != null.

33 % Flatten messages for the next superstep.
34 L8: send(J+1, Id, M) :-
35     superstep(J, _, _, {(Id,M)}).

```

(M) and (R) to the current global model and a data record respectively. Subsequently, the evaluation of $\text{map}(M, R, S)$ invokes the UDF that generates a data statistic (S) based on the input bindings. Finally, the statistics from all records are aggregated in the head predicate using the reduce UDF (defined in Section 2.2).

Rule $G3$ updates the global data model using the aggregated statistics. The first two body predicates simply bind (M) to the current global model and (AggrS) to the aggregated statistics respectively. The subsequent function predicate $\text{update}(J, M, \text{AggrS}, \text{NewM})$ calls the update UDF; accepting (J, M, AggrS) as input and producing an updated global model in the (NewM) variable. The head predicate records the updated global model at time-step $J+1$.

Program termination is handled in rule $G3$. Specifically, update

Listing 2: Datalog runtime for the Iterative Map-Reduce-Update programming model. The temporal argument is defined by the J variable.

```

1  % Initialize the global model
2  G1: model(0, M) :- init_model(M).

4  % Compute and aggregate all outbound messages
5  G2: collect(J, reduce<S>) :- model(J, M),
6     training_data(Id, R), map(R, M, S).

8  % Compute the new model
9  G3: model(J+1, NewM) :-
10     collect(J, AggrS), model(J, M),
11     update(J, M, AggrS, NewM), M != NewM.

```

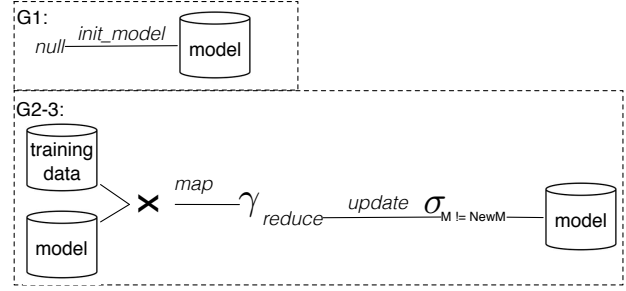


Figure 2: Logical query plan for Iterative Map-Reduce-Update.

is assumed to return the same model when convergence is achieved. In that case, the predicate $M \neq \text{NewM}$ in the body of $G3$ becomes false and we can prove that the program terminates. Typically, update achieves this convergence property by placing a bound on the number of iterations and/or a threshold on the difference between the current and the new model.

3.3 Semantics and Correctness

Up to this point, we have argued informally that the two Datalog programs faithfully encode the two programming models. However, this claim is far from obvious. The two programs contain recursion that involves negation and aggregation, and hence we need to show that each program has a well-defined output. Subsequently, we have to prove that this output corresponds to the output of the target programming models. In this section, we present a formal analysis of these two properties.

The foundation for our correctness analysis is based on the following theorem, which determines that the two Datalog programs fall into the specific class of XY-stratified programs.

Theorem 1 *The Datalog programs in Listing 1 and Listing 2 are XY-stratified [31].*

The proof can be found in Appendix B and is based on the machinery developed in [31]. XY-stratified Datalog is a more general class than stratified Datalog. In a nutshell, it includes programs whose evaluation can be stratified based on data dependencies even though the rules are not stratified.

We describe the semantics of the two Datalog programs by translating them (using standard techniques from the deductive database literature [26]) into an extended relational algebra. The resulting description illustrates clearly that the Datalog program encodes faithfully the corresponding machine-learning task. Moreover, we can view the description as a logical plan which can become the input to an optimizing query processor, which we discuss in the next section.

To facilitate exposition, we examine first the Datalog program for Iterative Map-Reduce-Update (Listing 2). Following XY-stratification, we can prove that the output of the program is computed from an initialization step that fires $G1$, followed by several iterations where each iteration fires $G2$ and then $G3$. By translating the body of each rule to the corresponding relational algebra expression, and taking into account the data dependencies between rules, it is straightforward to arrive at the logical plan shown in Figure 2. The plan is divided into two separate dataflows, each labeled by the rules they implement in Listing 2. The dataflow labeled $G1$ initializes the global model using the init_model UDF, which takes no input, and produces the initial model. The $G2$ – 3 dataflow

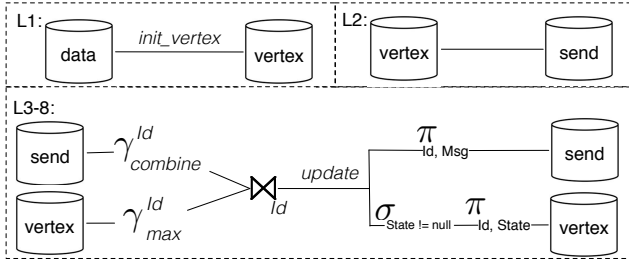


Figure 3: Logical query plan for Pregel.

executes one iteration. The `cross-product` operator combines the model with each tuple in the training dataset, and then calls the `map` UDF on each result. (This part corresponds to the body of rule $G2$.) The mapped output is passed to a `group-all` operator, which uses the `reduce` aggregate (e.g., `sum`) to produce a scalar value. (This part corresponds to the head of rule $G2$.) The aggregate value, together with the model, is then passed to the `update` UDF, the result of which is checked for a new model. (This part corresponds to the body of $G3$.) A new model triggers a subsequent iteration, otherwise the `update` output is dropped and the computation terminates. (This part corresponds to the head of $G3$.)

We can use the same methodology to analyze the Datalog program in Listing 1. Here we provide only a brief summary, since the details are more involved. *XY*-stratification prescribes that the output of the program is computed from an initialization step with rules $L1$ and $L2$, followed by several iterations where each iteration fires rules in the order $L3, \dots, L8$. Figure 3 shows the corresponding logical plan in relational algebra. Data flows $L1$ and $L2$ initialize the computation, as follows. In $L1$, each tuple from the training data is passed to the `init_vertex` UDF before being added to the `vertex` dataset. This triggers $L2$ to generate an initial `send` fact that is destined for each initial `vertex`.

The dataflow $L3$ - $L8$ encodes a single Pregel superstep. The `send` dataset is first grouped by the destination vertex identifier, and each such group of messages is aggregated by the `combine` UDF. The `vertex` dataset is also grouped by the vertex identifier, and its most recent state is selected by the `max` aggregate. The two results form the `collect` and `local` IDB predicates (rules $L3$, $L4$, and $L5$), which are joined along the vertex identifier attribute to produce the set of vertices with outstanding messages. The join result is passed to the `update` function to produce the `superstep` view, which is subsequently projected along two paths (rule $L6$). The bottom path checks for a non-null state object before projecting any resulting state objects onto the `vertex` dataset (rule $L7$). The top path projects the set of messages for the next superstep onto the `send` dataset (rule $L8$).

Overall, it is straightforward to verify that the logical plans match precisely the logic of the two programming models, which in turn proves our claim that the Datalog programs are correct. An equally important fact is that the logical plan captures the entirety of the computation, from loading the training data in an initial model to refining the model through several iterations, along with the structure of the underlying data flow. As we discuss in the next section, this holistic representation is key for the derivation of an efficient execution plan for the machine learning task.

4. PHYSICAL DATAFLOW

In this Section, we present the physical parallel dataflow plans that execute the Pregel and Iterative Map-Reduce-Update programming models. We choose the Hyracks data-parallel runtime [7] as

the target platform to develop and execute these physical plans. We first give a very brief overview of Hyracks (Section 4.1), then describe the physical plans for Pregel (Section 4.2) and Iterative Map-Reduce-Update (Section 4.3). The physical plans illustrated in this section are used to produce the experimental results presented in Section 5.

4.1 Hyracks Overview

Hyracks is a data-parallel runtime in the same general space as Hadoop [3] and Dryad [18]. Jobs are submitted to Hyracks in the form of directed acyclic graphs that are made up of *operators* and *connectors*. Operators are responsible for consuming input partitions and producing output partitions. Connectors perform redistribution of data between operators. Operators are characterized by an algorithm (e.g., filter, index-join, hash group-by) and input/output data properties (e.g., ordered-by or partitioned-by some attribute). Connectors are classified by a connecting topology and algorithm (e.g., one-to-one connector, aggregate connector, m-to-n hash partitioning connector, or m-to-n hash partitioning merging connector) as well as by a materialization policy (e.g., fully pipelining, blocking, sender-side or receiver-side materializing).

4.2 Pregel

Figure 4 shows the optimized physical plan for Pregel. The top dataflow executes iteration 0 and is derived from the logical plans $L1$ and $L2$ in Figure 3. The file scan operator ($O3$) reads partitions of the input (graph) data, repartitions each datum by its vertex identifier, and feeds the output to a projection operator ($O1$) and a sort operator ($O4$). Operator $O1$ generates an initial activation message for each vertex that and writes that result to the `send` dataset ($O2$). The sorted tuples from $O4$ are passed to the `init_vertex` function, the output of which is then bulk loaded into a B-Tree structure ($O5$).

The bottom dataflow executes iterations until a fixed point: when the `send` dataset becomes empty (no messages). This dataflow corresponds to the logical plan $L3$ – 8 in Figure 3. An iteration starts by scanning the message lists in the `send` dataset (using operator $O11$), which is consumed by an index inner-join operator ($O7$) that “joins” with the vertices in the B-Tree along the identifier attribute. The join result is passed to the `update` UDF ($O8$), which produces a new vertex state object and set of messages. Operator $O9$ forwards non-null state objects to update the B-Tree ($O10$), which occurs locally as indicated by the one-to-one connector in the physical plan. The messages are sorted by operator ($O12$), and subsequently fed to a pre-clustered group-by operator ($O15$), which groups messages by the destination vertex ID and uses the `combine` function to pre-aggregate the messages destined for the same vertex. A hash partitioning merging connector shuffles the tuples (using the vertex ID as the shuffle key and a list of messages as the value) over the network to a consumer (pre-clustered) group-by operator $O14$, which again applies the `combine` aggregate function to the final result before writing to (via $O14$) the new `send` dataset; occurring on the local machine that also holds the target vertex in the local B-Tree. All `send` partitions will report to the driver program, which determines if another iteration is required: when the `send` dataset is not empty.

Our translation of the Pregel physical plan from the corresponding logical plan included a number of physical optimizations:

Early Grouping: Applies the `combine` function to the sender-side transfer of the message data in order to reduce the data volume.

Storage Selection: In order to efficiently support primary key updates and avoid the logical `max` aggregation in Figure 3, a B-Tree index was chosen over raw files.

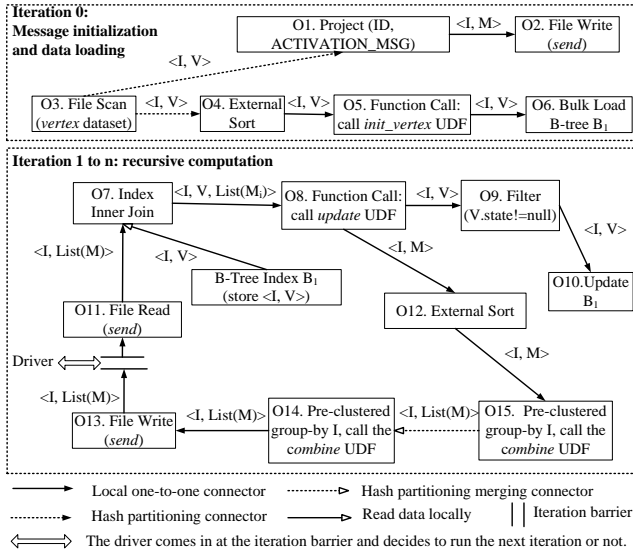


Figure 4: The physical plan for Pregel (operators are labeled O_n , where n is a number). A connector indicates a type (identified by a line and arrow) and a labeled data schema (I: vertex ID; V: vertex data object; M: message.)

Join Algorithm Selection: Inputs at operator O_7 are sorted by the join key; allowing an efficient (ordered) probing strategy.
Order Property: We selected an order-based group-by strategy at operator O_{14} since the input is already sorted.
Shared Scan: Operators O_1 and O_4 share one file scan.

4.3 Iterative Map-Reduce-Update

We now describe the construction and optimization of a physical plan for the Iterative Map-Reduce-Update programming model, which is tuned to run Batch Gradient Descent (BGD). Figure 5 describes a physical plan produced by translating the logical query plan in Figure 2 to a physical dataflow. Iteration 0 executes in the top dataflow, and corresponds to the logical plan G_1 . Here, we simply write the initial model to HDFS in operator O_2 . The bottom dataflow executes subsequent iterations until the driver detects termination. This dataflow corresponds to the logical plans G_2 – G_3 . At the start of each iteration, the current model is read from HDFS and paired with the record dataset (O_7 , O_3 and O_4). A `map` function call operator (O_5) is passed each record and the model, and produces a `(gradient, loss)` vector. The output vectors are then aggregated through a series of operators (O_6 , O_8 , O_{11}) that apply the `reduce` UDF, which in our experiments is a `sum`. The final aggregate result is passed to the `update` function call operator (O_{10}), along with the existing model, to produce the next model. The new model is written to HDFS (O_9), where it is read by the driver to determine if more iterations should be performed.

Two important physical optimization rules were considered when translating the logical plan into the physical plan:

Early aggregation: For commutative and associative `reduce` UDFs (like `sum`), the aggregation should be performed map-local to reduce the shuffled data volume; thus, O_6 is included in the plan.
Model volume property: Large objects (e.g., vectors in BGD) may saturate a single aggregator’s network connection, resulting in poor performance. In this case, a layered aggregation tree must be used to improved performance. Therefore, O_8 is included in the plan.

5. EXPERIMENTS

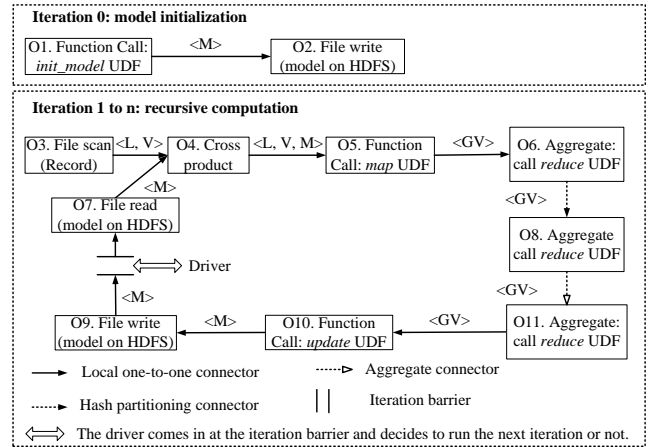


Figure 5: The physical plan for Iterative Map-Reduce-Update (operators are labeled with O_n , where n is a number). A connector indicates a type (identified by a line and arrow) and a labeled data schema (L: classification label; M: model vector; V: feature vector; GV: (gradient, loss) vector.)

In this section, we present experiments comparing the Datalog-derived physical plans of Section 4 to implementations of the same tasks on two alternative systems: Spark [30] for BGD and Hadoop [3] for Pregel. The purpose of these experiments is to demonstrate that a declarative approach, in addition to shielding ML programmers from physical details, can provide performance and scalability competitive with current “best of breed” approaches.

All experiments reported here were conducted on a 6-rack, 180-machine Yahoo! Research Cluster. Each machine has 2 quad-core Intel Xeon E5420 processors, 16GB RAM, 1Gbps network interface card, and four 750GB drives configured as a JBOD, and runs RHEL 5.6. The machines are connected to a top rack Cisco 4948E switch. The connectivity between any pair of nodes in the cluster is 1Gbps. We discuss system-specific configuration parameters in the relevant subsections. In Section 5.1, we compare our approach against a Spark implementation on a Batch Gradient Descent task encoded in the Iterative Map-Reduce-Update programming model. Section 5.2 presents a PageRank experiment that runs on a full snapshot of the World-Wide Web from 2002 and compares our approach to an implementation based on Hadoop.

5.1 Batch Gradient Descent

We begin with a Batch Gradient Descent (BGD) task on a real-world dataset drawn from the web content recommendation domain. The data consists of 16,557,921 records sampled from Yahoo! News. Each record consists of a feature vector that describes a `(user, content)` pair and a label that indicates whether the user consumed the content. The goal of the ML task is to learn a linear model that predicts the likelihood of consumption for a yet unseen `(user, content)` pair. The total number of features in the dataset is 8,368,084,005, and each feature vector is sparse: users are only interested in a small subset of the content. The dataset is stored in HDFS and, before running the job, it is perfectly balanced across all machines used in the experiment. That is, each machine is assigned an equal number of records.

We report results for running this task in Hyracks (using the physical plan in Figure 5) to Spark. The Spark code was organized similarly and verified by the system author (Matej Zaharia).

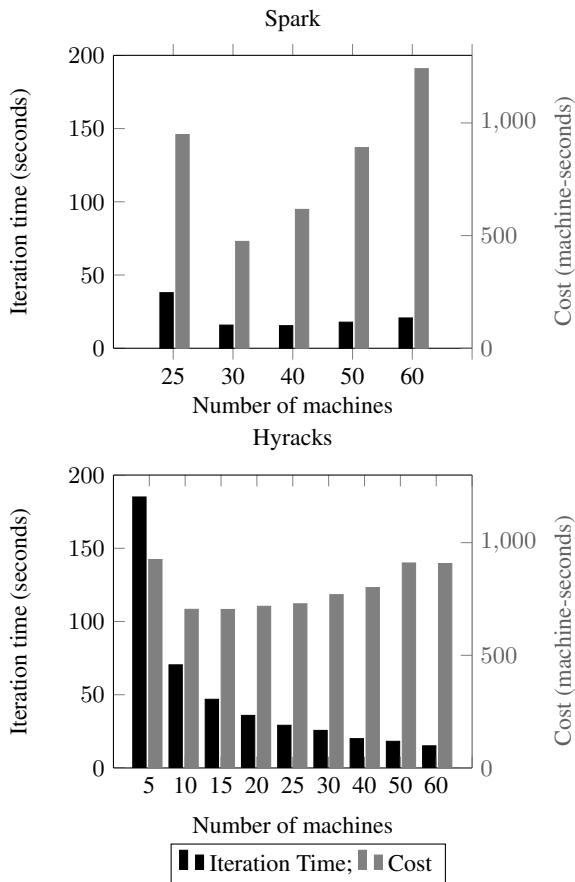


Figure 6: BGD speed-up of Hyracks and Spark on Yahoo! News dataset

Specifically, in the first step, we read each partition from HDFS and convert it to an internal record: sparse vectors are optimized in a compact form. In Hyracks, each partition is converted into a binary file representation and stored in the local file system. For Spark, we make an explicit cache call that “pins” the records in memory. Both systems then execute a fixed number of iterations, each of which executes a single `map`, `reduce`, and `update` step.

In the `map` step, we make a pass over all the records in a given internal partition and compute a single $(\text{gradient}, \text{loss})$ value based on the current model, which resides in HDFS. The `reduce` step sums all the $(\text{gradient}, \text{loss})$ values produced by individual `map` tasks to a single aggregate value. We use pre-aggregators in both systems to optimize the computation of these sums. In Spark, we use a single layer of $\sqrt{\text{num partitions}}$ pre-aggregators. Hyracks performs local pre-aggregation on each machine (holding four `map` partitions) followed by a single layer of $\sqrt{\text{num map machines}}$ pre-aggregators. We also evaluate an alternative (more optimal) Hyracks configuration, which again performs a local pre-aggregation but then uses a 4-ary aggregation tree (a variable-height aggregation tree where each aggregator receives at most 4 inputs). The Spark API did not allow us to use a `(map)` machine local pre-aggregation strategy, and there is no system support for such optimizations. The final `update` step takes the aggregated result and the current model and produces a new model that is written to HDFS⁴ for use by the next iteration’s `map` step.

We now present two sets of experiments. First, we identify the

⁴Spark exposes this operation through a “broadcast” variable.

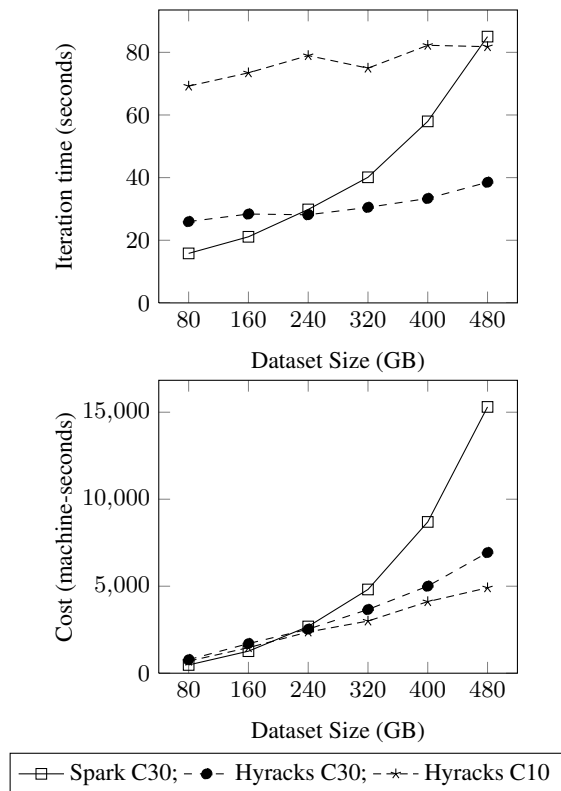


Figure 7: BGD scale-up of Hyracks vs. Spark

cost-optimal number of machines that each system should use to process a fixed-size dataset. Using the cost-optimal configuration, we measure the scalability of each system by proportionately increasing the dataset size and number of machines used.

5.1.1 Cost-optimal configuration for fixed-size data

The goal of this experiment is to determine the optimal number of machines that should be used to perform the BGD task for a fixed-size dataset on Spark and Hyracks. We measure cost in terms of machine-seconds (*number of machines × average iteration time*) and look for a cluster size that minimizes it. Figure 6 reports both time and cost, averaged over five iterations, as we increase the number of machines while keeping the total dataset size fixed at ~80GB. Increasing the number of machines generally improves the iteration time, but diminishing returns due to increasing overhead make it cost-inefficient. From this experiment, we identify the cost optimal configurations to be 30 machines for Spark and 10 machines (giving preference to fewer machines) for Hyracks. Note that Hyracks could use an arbitrarily small number of machines since it supports out-of-core computations. Spark, however, is restricted to main-memory, and as a result, requires at least 25 machines to run this experiment.

5.1.2 Scalability

Given the cost-optimal configuration, we now explore the scalability of each system as we proportionally increase the input training data size and number of machines. To scale up the data, we duplicated and randomly shuffled the original data. The cost-optimal settings are captured by the following two cluster-size-for-data-size configurations:

- C10** : 10 nodes per 80GB (Hyracks cost-optimal)
- C30** : 30 nodes per 80GB (Spark cost-optimal)

We executed both Hyracks and Spark on configuration C30, but we only ran Hyracks on C10 since Spark was unable to retain this much data in the given amount of main memory. Figure 7 reports the results of this scalability experiment, showing iteration time at the top and cost at the bottom. The x-axis for both graphs range over increasing data sizes. Each configuration adds the baseline number of nodes to match the data size. Example: At data size 160GB, configuration C10 uses 20 machines and C30 uses 60 machines.

As we scale up, we expect the `map` part of the iteration to scale perfectly. However, as we add more partitions, we create more intermediate results that need to be transferred over the network to the `reduce` aggregation. It turns out that the amount of network traffic between the `map` nodes and the intermediate pre-aggregators is linear in the number of `map` nodes, and the work done in reducing the intermediate results is proportional to the square root of the `map` nodes. Thus, we expect a growth in completion time as we scale up the number of `map` nodes. We clearly see this trend for the execution time of Spark. However, the growth in completion time for Hyracks is much slower, benefiting from the machine-local aggregation strategy. Hyracks also uses a packet-level fragmentation mechanism that achieves better overlap in the network transfer and aggregation of intermediate results; receiving aggregators can immediately start reducing each fragment independently while other fragments are in transit. Spark on the other hand, waits for the complete `(gradient, loss)` result—a $\sim 16\text{MB}$ size vector—to be received before incorporating it into the running aggregate. Additionally, Spark faces other system-level bottlenecks due to its use of a stock data-transfer library to move data between processes, while Hyracks has an efficient custom networking layer built for low-latency high-throughput data transfer.

For data sizes 80GB and 160GB, Spark finishes slightly earlier than Hyracks. There are two factors that contribute to this phenomenon. The first is that Hyracks currently has slightly higher overhead in the `map` phase in how it manages the input data. As mentioned earlier, Hyracks uses the local file system to store a binary form of the data and relies on the file system cache to avoid disk I/O. Accessing data through the file system interface on each iteration, and the data-copies across this interface, account for slightly larger `map` times. Spark, on the other hand, loads the input data into the JVM heap and no data copies are necessary during each iteration. A second cause of slow down for Hyracks is that the local aggregation step in Hyracks adds latency, but does not help much in lowering the completion time of an iteration for the 80GB case. In the 160GB case, the benefit of local aggregation is still out-weighed by the latency introduced. In our experimental setup, each rack has 30 machines resulting in rack-local computation in the 80GB case. In the 160GB case, the computation is spread across two racks. Since enough bandwidth is available within a rack, the local aggregation does not appear to pay off in these two cases. Our planned solution to the first problem is to take on more of the buffer-management in Hyracks to reduce or eliminate data copies for the data that resides in memory, but still use the file system so that we can scale to data larger than main memory. The second problem motivates the need for a runtime optimizer that decides when it is appropriate to use local combiners to solve the problem for a given physical cluster. In the future, the optimizer that generates the Hyracks job for the Batch Gradient Descent task will be expected to make the correct choice with regards to local aggregation.

As we scale up the data beyond 160GB, we see that the Hyracks system shows better scale up characteristics for the reasons mentioned above. The cost curve in the bottom graph shows a similar trend to that of the time curve. As we linearly increase the data and

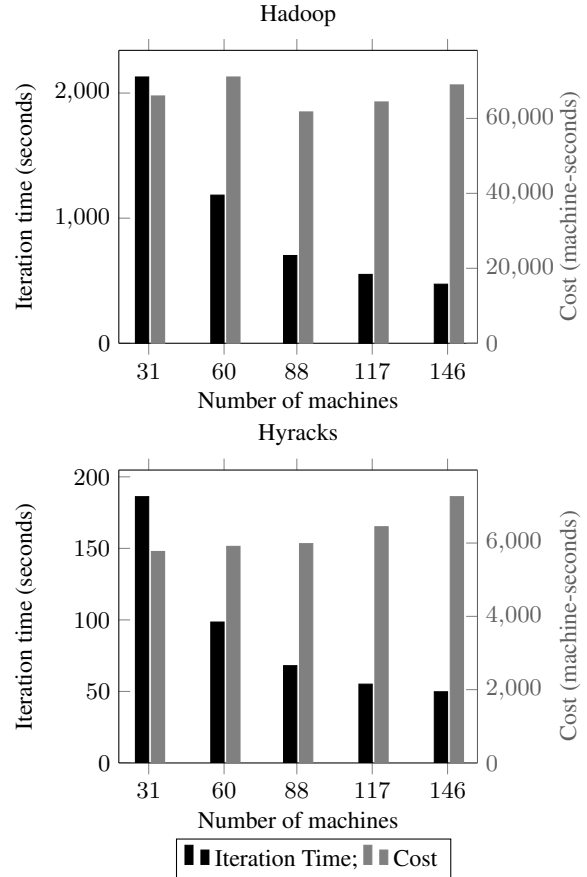


Figure 8: PageRank speed-up of Hyracks vs. Hadoop

cluster size, the cost to solve the problem with Spark grows much faster than the cost to solve the problem with Hyracks.

5.2 PageRank

The goal of PageRank is to rank all web pages by their relative importance. We perform our experiments on Yahoo!’s publicly available webmap dataset [29], which is a snapshot of the World-Wide Web from 2002. The data consists of 1, 413, 511, 393 vertices, each of which represents a web-page. Each record consists of a source vertex identifier and an array of destination vertex identifiers forming the links between different webpages in the graph. The size of the decompressed data is 70GB, which is stored in HDFS and evenly balanced across all participating machines.

We compare the performance of Hyracks (using the physical plan in Figure 4) to an implementation of PageRank in Hadoop. The Hadoop code for PageRank consists of a job that first joins the ranks with the corresponding vertices. This is followed by a grouping job that combines contribution rank values from “neighboring” vertices to compute the new rank. Hyracks executes the whole iteration of the PageRank algorithm in a single job. For both systems we perform 10 iterations.

We follow the same methodology as above: First, we identify the cost-optimal number of machines for each system using a fixed-size dataset (70GB). Then we explore the scalability of Hyracks and Hadoop by running PageRank against proportionately increasing dataset sizes and number of machines, using the cost-optimal machine configurations.

5.2.1 Cost-optimal configuration for fixed-size data

Configuration	Dataset Size(GB)	Iteration Time(s)	Cost
Hyracks-C88	70	67.993	5983.394
Hadoop-C88	70	701.411	61724.153
Hyracks-C88	140	84.970	14869.750
Hadoop-C88	140	957.727	167602.196
Hyracks-C31	70	186.137	5770.240
Hyracks-C31	140	208.444	12506.658

Table 1: PageRank scale-up of Hyracks vs. Hadoop

In this experiment, we determine the cost-optimal number of machines to be used for a fixed-size (70GB) dataset on Hadoop and Hyracks. Figure 8 reports the average iteration time and the cost in terms of machine-seconds ($number\ of\ machines \times average\ iteration\ time$) for different number of machines. The iteration time in both systems is decreases as we add more machines. Hadoop’s iteration cost fluctuates as we increase the number of machines, whereas Hyracks’ cost increases slowly. Also, we note the following effects: 1) As we add more machines, the benefit obtained from local combiners gradually diminishes, and 2) the repartitioning step becomes the bottleneck in Hadoop. Hadoop’s implementation of PageRank needs to shuffle both the graph data (which is invariant across iterations) and the rank contributions, leading to far more data movement over the network than the PageRank plan in Hyracks. Hyracks moves around (shuffles) only the rank contributions over the network, while caching the loop-invariant graph data at the same nodes across iterations. This extra data movement accounts for most of the order-of-magnitude increase in iteration time experienced by Hadoop when comparing to Hyracks.

The cost-optimal configuration is 31 machines for Hyracks and 88 machines for Hadoop per 70GB of data as per Figure 8.

5.2.2 Scalability

To scale up the data size, we duplicated the original graph data and renumbered the duplicate vertices by adding each identifier with the largest vertex identifier in the original graph. Thus, one duplication creates a graph that has twice as many vertices in two disconnected subgraphs. The nodes in the resulting graphs were randomly shuffled before loading the data onto the cluster. While we recognize that this does not follow the structure of the web, this experiment is concerned with the behavior of the dataflow rather than the actual result of the PageRank algorithm.

Based on the cost-optimal results from the speed-up we derive the following two configurations:

C31 : 31 machines per 70GB (Hyracks cost-optimal)

C88 : 88 machines per 70GB (Hadoop cost-optimal)

Table 1 shows that Hyracks PageRank performance for data sizes 70GB and 140GB is an order-of-magnitude faster and cheaper than Hadoop (in Hadoop’s optimal configuration C88) owing to more data movement over the network, as described above. Both systems scale similarly as we grow the graph data and the size of the cluster.

5.2.3 Comparing Different Hyracks Plans

To further investigate performance differences associated with alternate physical data movement strategies, we tried rerunning Hyracks with a slight variation in the connector used to redistribute the messages from the message combiners (O15) to the message reducers (O14) in the plan shown in Figure 4. We replaced the hash partitioning merging connector with a simpler hash partitioning connector. While the original merging connector maintained the sorted order of messages as they were received from each combiner, the hash partitioning connector merges data from any sender

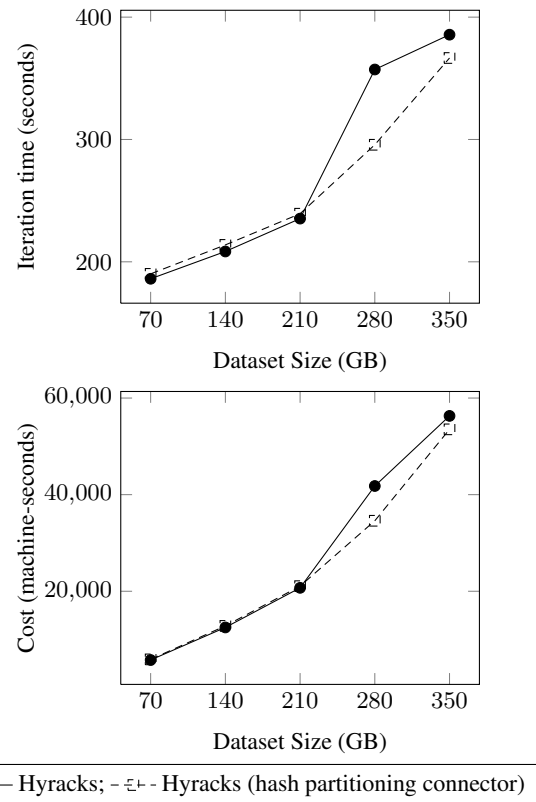


Figure 9: PageRank scale-up of Hyracks alternative plans

in the order it is received, thus destroying the sorted property. In order to get the sorted property back, we added an explicit sorter before feeding the messages into O14. Figure 9 shows the iteration times and the cost of iterations of the two Hyracks plans as we scale up the graph size and the number of machines used to compute the PageRank using configuration C31.

We see that for smaller data and cluster sizes (70GB to 210GB), the Hyracks plan with the hash partitioning merging connector runs faster than the one with the hash connector with explicit sorting. This is because the former plan does less work in maintaining the sort order because the merge exploits the sorted input property to merge the incoming data at receiver using a priority queue, much like the merge phase used in external sorting. However, each receiver of the merge process selectively waits for data to arrive from a specific sender as dictated by the priority queue. Temporary slowness on behalf of a sender at a time when a receiver needs data from it leads to a stall in the merge pipeline. Although other senders are willing to send data, they have to wait for the specific sender that the receiver is waiting for to make data available. The resulting degradation in the iteration time is observable as the size of the cluster grows to data sizes of 280GB and 350GB. At these sizes, the savings in work achieved by the hash partitioning merging connector are far outweighed by the coordination overhead introduced by the merge process. This tradeoff is evidence that an optimizer is ultimately essential to identify the best configuration of the runtime plan to use in order to solve the Pregel problem.

5.3 Discussion

One might wonder why Hadoop was the chosen for the reference implementation of the Pregel runtime plan. Before we compared our system with Hadoop, we also tried to compare it with

three other “obvious candidate” systems, namely Giraph [2], Mahout [5], and Spark [30]. What we discovered (the hard way!) is that none of those systems was able to run PageRank for the Yahoo! webmap dataset, even given all 6 racks (175 machines), due to design issues related either to memory management (Giraph and Spark) or to algorithm implementation (Mahout).

An interesting observation regarding the Spark user model was the process involved in implementing a 1-level aggregation tree. In order to perform a pre-aggregation in Spark we had to explicitly write—in user facing code—an intermediate “reduceByKey” step that subsequently feeds the final (global) reduce step. We assigned a random number (using `java.lang.Random.nextInt` (modulo the number of pre-aggregators) as the key to the `(gradient, loss)` record from the map step. Ideally, such an optimization should be captured by the system, and not in user code.

6. RELATED WORK

Our work builds upon and extends prior results from a number of different research areas.

Parallel database systems such as Gamma [14], Teradata [27], and GRACE [17] applied partitioned-parallel processing to data management, particularly query processing, over two decades ago. The introduction of Google’s MapReduce system [13], based on similar principles, led to the recent flurry of work in MapReduce-based data-intensive computing. Systems like Dryad [18] and Hyracks [7] have successfully made the case for supporting a richer set of data operators beyond map and reduce as well as a richer set of data communication patterns.

High-level language abstractions like Pig [24], Hive [4], and DryadLINQ [19] reduce the accidental complexity of programming in a lower-level dataflow paradigm (e.g., MapReduce). However, they do not support iteration as a first class citizen, instead focusing on data processing pipelines expressible as directed acyclic graphs. This forces the use of inefficient external drivers when iterative algorithms are required to tackle a given problem.

Iterative extensions to MapReduce like HaLoop [9] and PrIter [32] were the first to identify and address the need for runtime looping constructs. HaLoop uses a “sticky scheduling” policy to place map and reduce tasks in downstream jobs on the same physical machines with the same inputs. In Hyracks, the job client is given control over the task placement, which we use to implement a similar policy. PrIter uses a key-value storage layer to manage its intermediate MapReduce state, and it also exposes user-defined policies that can prioritize certain data to promote fast algorithmic convergence. However, those extensions still constrain computations to “map” and “reduce” functions, while Hyracks allows more flexible computations and forms of data redistribution for optimizing machine learning tasks.

Domain-specific programming models like Pregel [22], GraphLab [21], and Spark [30], go beyond one-off implementations for specific algorithms (e.g. [1, 28]), to general purpose systems that capture a specific class of ML tasks. Of these, Spark is the most general, but it lacks a runtime optimizer and support for out-of-core operators, making it hard to tune. GraphLab and Pregel expose a graph-oriented programming model and runtime that is very appropriate for some ML tasks but suboptimal for others. GraphLab supports asynchronous execution, which lends itself to graphical model machine learning.

RDBMS extensions have been proposed that provide direct support for ML tasks. In Tuffy [23], Markov Logic Networks are represented as declarative rules in first-order-logic, and from there, optimized into an efficient runtime plan by a RDBMS. MadLib [11]

maps linear algebra operations, such as matrix multiplies, to SQL queries that are then compiled and optimized for a parallel database system. These approaches are limited to single pass algorithms (i.e., closed form solutions) or require the use of an external driver for iterative algorithms.

Datalog extensions have also been proposed for implementing ML tasks. Atul and Hellerstein [6] use Overlog—a distributed Datalog-like declarative language—to elegantly capture probabilistic inference algorithms; among them, a Junction Tree Running Intersection Property expressed in a mere seven Overlog rules. Dyna uses a Datalog extension to capture statistical Artificial Intelligence algorithms as systems of equations, which relate intensional and extensional data to form structured prediction models [16]. Dyna compiles such model specifications into efficient code.

Our approach shares many aspects with various of the aforementioned systems, yet it is unlike any one of those. To the best of our knowledge, this paper has proposed the first distributed, out-of-core-capable runtime for Datalog aimed at supporting several end-user programming models at once, thereby unifying machine learning and ETL processes within a single framework and on a single, and scalable, runtime platform.

7. CONCLUSION

The growing demand for machine learning is pushing both industry and academia to design new types of highly scalable iterative computing systems. Examples include Mahout, Pregel, Spark, Twister, HaLoop, and PrIter. However, today’s specialized machine learning platforms all tend to mix logical representations and physical implementations. As a result, today’s platforms 1) require their developers to rebuild critical components and to hardcode optimization strategies and 2) limit themselves to specific runtime implementations that usually only (naturally) fit a limited subset of the potential machine learning workloads. This leads to the current state of practice, wherein the implementation of new scalable machine learning algorithms is very labor-intensive and the overall data processing pipeline involves multiple disparate tools hooked together with file- and workflow-based glue.

In contrast, we have advocated a declarative foundation on which specialized machine learning workflows can be easily constructed and readily tuned. We verified our approach with Datalog implementations of two popular programming models from the machine learning domain: Pregel, for graphical algorithms, and Iterative Map-Reduce-Update, for deriving linear models. The resulting Datalog programs are compact, tunable to a specific task (e.g., Batch Gradient Descent and PageRank), and translated to optimized physical plans. Our experimental results show that on a large real-world dataset and machine cluster, our optimized plans are very competitive with other systems that target the given class of ML tasks. Furthermore, we demonstrated that our approach can offer a plan tailored to a given target task and data for a given machine resource allocation. In contrast, in our large experiments, Spark failed due to main-memory limitations and Hadoop succeeded but ran an order-of-magnitude less efficiently.

The work reported here is just a first step. We are currently developing the ScalOps query processing components required to automate the remaining translation steps from Figure 1; these include the Planner/Optimizer as well as a more general algebraic foundation based on extending the Algebricks query algebra and rewrite rule framework of ASTERIX [8]. We also plan to investigate support for a wider range of machine learning tasks and for a more asynchronous, GraphLab-inspired programming model for encoding graphical algorithms.

8. REFERENCES

- [1] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *CoRR*, abs/1110.4198, 2011.
- [2] Giraph: Open-source implementation of Pregel. <http://incubator.apache.org/giraph/>.
- [3] Hadoop: Open-source implementation of MapReduce. <http://hadoop.apache.org>.
- [4] The Hive Project. <http://hive.apache.org/>.
- [5] The Mahout Project. <http://mahout.apache.org/>.
- [6] Ashima Atul. Compact implementation of distributed inference algorithms for network. Master’s thesis, EECS Department, University of California, Berkeley, Mar 2009.
- [7] Vinayak R. Borkar, Michael J. Carey, Raman Grover, Nicola Onose, and Rares Vernica. Hyracks: A flexible and extensible foundation for data-intensive computing. In *ICDE*, pages 1151–1162, 2011.
- [8] Vinayak R. Borkar, Michael J. Carey, and Chen Li. Inside “Big Data Management”: Ogres, Onions, or Parfaits? In *EDBT*, 2012.
- [9] Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. HaLoop: Efficient iterative data processing on large clusters. *PVLDB*, 3(1):285–296, 2010.
- [10] Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary R. Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. In *NIPS*, pages 281–288, 2006.
- [11] Jeffrey Cohen, Brian Dolan, Mark Dunlap, Joseph M. Hellerstein, and Caleb Welton. Mad skills: New analysis practices for big data. *PVLDB*, 2(2):1481–1492, 2009.
- [12] Tyson Condie, David Chu, Joseph M. Hellerstein, and Petros Maniatis. Evita raced: metacompilation for declarative networks. *PVLDB*, 1(1):1153–1165, 2008.
- [13] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI*, pages 137–150, 2004.
- [14] David J. DeWitt, Shahram Ghandeharizadeh, Donovan A. Schneider, Allan Bricker, Hui-I Hsiao, and Rick Rasmussen. The gamma database machine project. *IEEE Trans. Knowl. Data Eng.*, 2(1):44–62, 1990.
- [15] David J. DeWitt and Jim Gray. Parallel database systems: The future of high performance database systems. *Commun. ACM*, 35(6):85–98, 1992.
- [16] Jason Eisner and Nathaniel W. Filardo. Dyna: Extending Datalog for modern AI. In Tim Furche, Georg Gottlob, Giovanni Grasso, Oege de Moor, and Andrew Sellers, editors, *Datalog 2.0*, Lecture Notes in Computer Science. Springer, 2011. 40 pages.
- [17] Shinya Fushimi, Masaru Kitsuregawa, and Hidehiko Tanaka. An overview of the system software of a parallel relational database machine grace. In *VLDB*, pages 209–219, 1986.
- [18] Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In *EuroSys*, pages 59–72, 2007.
- [19] Michael Isard and Yuan Yu. Distributed data-parallel computing using a high-level programming language. In *SIGMOD Conference*, pages 987–994, 2009.
- [20] Michael Kearns. Efficient noise-tolerant learning from statistical queries. In *Journal of the ACM*, pages 392–401. ACM Press, 1993.
- [21] Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein. GraphLab: A new framework for parallel machine learning. In *UAI*, pages 340–349, 2010.
- [22] Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ian Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: a system for large-scale graph processing. In *SIGMOD Conference*, pages 135–146, 2010.
- [23] Feng Niu, Christopher Ré, AnHai Doan, and Jude W. Shavlik. Tuffy: Scaling up statistical inference in markov logic networks using an RDBMS. *PVLDB*, 4(6):373–384, 2011.
- [24] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig Latin: a not-so-foreign language for data processing. In *SIGMOD Conference*, pages 1099–1110, 2008.
- [25] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
- [26] Raghu Ramakrishnan and Jeffrey D. Ullman. A survey of research on deductive database systems. *Journal of Logic Programming*, 23:125–149, 1993.
- [27] Jack Shermer and Philip M. Neches. The genesis of a database computer. *IEEE Computer*, 17(11):42–56, 1984.
- [28] Markus Weimer, Sriram Rao, and Martin Zinkevich. A convenient framework for efficient parallel multipass algorithms. In *LCCC : NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*, December 2010.
- [29] Yahoo! Webscope Program. <http://webscope.sandbox.yahoo.com/>.
- [30] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: cluster computing with working sets. HotCloud’10, page 10, Berkeley, CA, USA, 2010.
- [31] Carlo Zaniolo, Natraj Arni, and KayLiang Ong. Negation and aggregates in recursive rules: the LDL++ approach. In *DOOD*, pages 204–221, 1993.
- [32] Yanfeng Zhang, Qixin Gao, Lixin Gao, and Cuirong Wang. PrIter: a distributed framework for prioritized iterative computations. In *SOCC*, pages 13:1–13:14, New York, NY, USA, 2011.

APPENDIX

A. BATCH GRADIENT DESCENT

Much of supervised machine learning can be cast as a convex optimization problem. In supervised machine learning, we are given a database of pairs (x, y) , where x is a data point and y is a label. The goal is to find a function $f_w(x)$ that can predict the labels for the yet unseen examples x . Depending on the type of y , this definition specializes into many machine learning tasks: regression, (binary and multi-class) classification, logistic regression and structured prediction are some examples.

Learning the function f_w amounts to searching the space of parameterized functions. The parameters are also called model and are typically referred to as w . Hence, the search for f_w is the search for w . This search problem is guided by a *loss function* $l(f_w(x), y)$ that measures the divergence between a prediction $f_w(x)$ and a known label y . A large class of machine learning problems also include a *regularizer* function $\Omega(w)$ that measures the complexity of f_w . Following Occam’s razor—all things equal favor a simpler model—the regularizer is added to the loss to form the template of a supervised machine learning problem:

$$\hat{w} = \operatorname{argmin}_w \left(\lambda \Omega(w) + \sum_{(x,y) \in D} l(f_w(x), y) \right) \quad (1)$$

The loss function is sometimes referred to as (empirical) risk, and therefore the above optimization problem is known as *regularized risk minimization* in the literature. From a dataflow perspective, *evaluating* a given model w is easily parallelized, since the sum of the losses decompose over the data points (x, y) .

Example: A regularized linear regression⁵ is a *linear model*, hence $f_w(x) = \langle w, x \rangle$ is the inner product between the data point x and a weight vector w . Choosing the quadratic distance $l(f(x), y) = \frac{1}{2}(f(x) - y)^2$ as the loss function leads to linear regression. Finally, we select the squared norm of w as the regularizer $\Omega(w) = \frac{1}{2}|w|_2^2$:

$$\hat{w} = \operatorname{argmin}_w \left(\frac{\lambda}{2}|w|_2^2 + \sum_{(x,y) \in D} \frac{1}{2} (\langle w, x \rangle - y)^2 \right) \quad (2)$$

In most instances, the loss $l(f_w(x), y)$ is *convex* in w , which guarantees the existence of a minimizer \hat{w} and *differentiable*. This facilitates efficient search strategies that use the gradient of the cost function with respect to w . Different choices for the optimization algorithm are possible. Here, we restrict ourselves to the iterative procedure *Batch Gradient Descent* (BGD), as it embodies the core dataflow of a wide variety of optimization algorithms. Until convergence, batch gradient descent performs the following step:

$$w_{t+1} = w_t - \left(\lambda \partial_w \Omega(w) + \sum_{(x,y) \in D} \partial_w l(f_w(x), y) \right) \quad (3)$$

Here, ∂_w denotes the gradient with respect to w . Just as in the case of evaluating a model w above, the sum decomposes per data point (x, y) , which facilitates efficient parallelization and distribution of the computation of each gradient descent step.

The beauty of this approach lies in its generality: Different choices for the loss l , the prediction function f_w and the regularizer Ω yield a wide variety of machine learning models: Support

⁵a.k.a., linear support vector regression or ridge regression

Vector Machines, LASSO Regression, Ridge Regression and Support Vector novelty detection to name a few. All of which can be efficiently learned through BGD or similar algorithms.

BGD can be captured in Iterative Map-Reduce-Update quite easily. In fact, the sum in (3) can be efficiently captured by a single MapReduce step where each `map` task computes gradients for its local data points while the `combine` sums them up and `reduce` applies them and the gradient of the regularizer Ω to the current model w . The user needs to supply the UDFs mentioned in section 2.2:

map computes a gradient for the current data point, using the current model w_t

reduce aggregates a set of gradients into one.

update accepts a current model w_t and the aggregated gradients and produces a new predictor w_{t+1} after applying the regularizer Ω .

B. MODEL(ING) SEMANTICS

Datalog least-fixedpoint semantics tells us that a program without aggregation and negation has a unique minimal model. In other words, the result we get from evaluating the rules to fixpoint is always the same and consistent with the logic program. A Datalog program that includes aggregates and negated subgoals—like those in Section 3—may have several minimal models. There are more general classes of Datalog semantics that can decide which one minimal model is consistent with the intent of the programmer. In Section B.1, we show that the programs in Section 3 are in the class of locally stratified Datalog programs. In Section B.2, we argue that our runtime selects the one minimal model that is consistent with locally stratified Datalog semantics and our conditions for program termination.

B.1 Program Stratification

Stratified Datalog semantics extend least-fixedpoint semantics with a method for organizing predicates into a hierarchy of strata; using a process called stratification. If some predicate A depends on an aggregated or negated result of another predicate B then A is placed in a higher stratum than B . A runtime that supports Stratified Datalog evaluates rules in lower strata first. Intuitively, this forces the complete evaluation of predicate B before predicate A is allowed to view the result. Stratification fails when there are cycles through negation or aggregation in the (rule/goal) dependency graph. Intuitively, if A and B depend on each other, perhaps even indirectly, then we can not evaluate one to completion while isolating the other.

Program stratification fails in Listings 1 and 2 (Section 3) since they both contain cycles through a stratum boundary (i.e., aggregation or negation). Therefore, we look to another class of Datalog semantics called locally stratified programs, which is defined in terms of a data dependent property. Intuitively, these programs are not necessarily stratified according to the syntax of the rules, but rather according to the application of those rules on a specific data collection. The following definition follows from Zaniolo et al., [31].

Definition 1 *A program is locally stratifiable iff the Herbrand base can be partitioned into a (possibly infinite) set of strata S_0, S_1, \dots , such that for each rule r with head h and each atom g in the body of r , if h and g are, respectively, in strata S_i and S_j , then*

1. $i \geq j$ if g is a positive goal, and
2. $i > j$ if g is a negative goal.

Listing 3: Listing 2 after XY-Stratification.

```

1 % Initialize the global model
2 G1: new_model(M) :- init_model(M).
3
4 % Compute and aggregate all outbound messages
5 G2: new_collect(reduce<S>) :- new_model(M),
6     training_data(Id, R), map(R, M, S).
7
8 % Compute the new model
9 G3: new_model(NewM) :-
10     old_collect(AggrS), old_model(M),
11     old_update(M, AggrS, NewM), M != NewM.

```

Intuitively, a program is locally stratifiable if the model data—formed from the initial facts and rule derivations—is stratifiable. The key to proving that the programs in Listings 1 and 2 are locally stratified lies in the temporal argument of our recursive predicates. The values of the temporal argument are taken from a discrete temporal domain that is monotonic. This allows us to use another program stratification technique called XY-Stratification [31].

Definition 2 Let P be a program with a set rules defining mutually recursive predicates. P is an XY-Stratified program if it satisfies the following conditions:

1. Every recursive predicate has a distinguished temporal argument.
2. Every recursive rule is either an X-rule or a Y-rule.

In an X-rule, the temporal arguments of every recursive predicate must refer to the current temporal state (e.g., J). A Y-rule has the following constraints.

1. The head predicate temporal argument value contains a successor state (e.g., $J + 1$).
2. Some positive goal in the body has a temporal argument of the current state (e.g., J).
3. The remaining recursive goals have a temporal argument that contains either the current state (e.g., J) or the successor state (e.g., $J + 1$).

Intuitively, an X-rule reasons within the current state and a Y-rule reasons from the current state to the next.

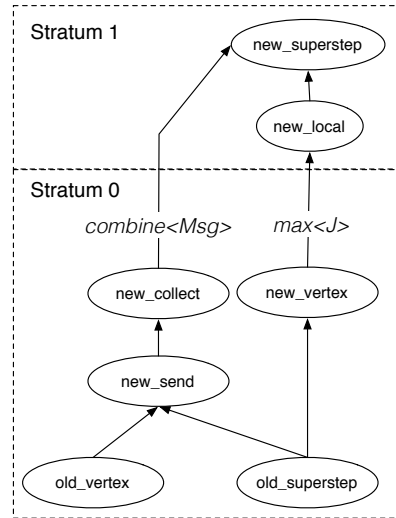
It is known that if a program is XY-stratified then it is locally stratified [31]. We now show that the programs in Section 3 are XY-stratified programs using the following construction applied to each recursive rule r .

1. Rename all recursive predicates that have the same temporal argument as the head with a prefix **new**.
2. Rename all other occurrences of recursive predicates with the prefix **old**.
3. Drop the temporal arguments from all recursive predicates.

If the resulting program following this construction can be stratified then the original program is locally stratified [31].

Theorem 2 Listing 2 is in the class of XY-stratified programs.

PROOF. The program in Listing 3 follows from applying XY-Stratification to the program in Listing 2. Listing 3 is trivially stratified by placing `new_collect` in the highest stratum. Therefore, evaluating the rules in Listing 3 produces a locally stratified model that is consistent with the programmer’s intent in Listing 2. \square

**Figure 10: Dependency graph for XY-stratified Listing 1.**

Theorem 3 Listing 1 is in the class of XY-stratified programs.

PROOF. Figure 10 contains the dependency graph for the predicates appearing in Listing 1 after the XY-stratified transformation. The graph shows that the program is stratified into two strata. We further note that naming `new_local` comes from using `max` aggregation applied to the temporal argument of base predicates `new_vertex`, and `new_superstep` comes from using `new_local` and `combine` UDF applied to `new_collect`. \square

B.2 Stratified Evaluation and Termination

So far, we have applied XY-Stratification to our programs and to produce new programs that are stratifiable. The data in the i^{th} time-step treated data from previous time-steps $j < i$ as the extensional database (EDB). This allowed us to break dependency cycles at Y-rules, which, by definition, derive data for the subsequent time-step. These XY-Stratified programs formed the basis of the template physical plans described in Section 4.

We now conclude with a discussion of termination of our Datalog programs. The runtime terminates when the Datalog program reaches a fixpoint. We have already shown that the result of a fixpoint is a locally stratified model. However, this model could be infinite, in which case it would never terminate. Therefore, termination depends solely on a finite fixpoint solution. Under Datalog semantics this occurs when derivations range over a finite domain. Intuitively, if the range is finite then we will eventually derive all possible values since Datalog is monotonic and set-oriented.

For the programs listed in Section 3, this can occur in two possible ways. First, when the temporal argument ranges over a finite time domain. Since this argument is monotonic and finite, we are guaranteed to reach an upper bound, and hence terminate. A second possible termination condition comes from the range of state values given by the `update` function. Recall that this function produces new state objects when given the (current) state object and list of messages. The runtime will consider the `update` function predicate to be false if the new state object does not differ from the previous. Therefore, if there are a finite number of possible state objects, and each state object is produced exactly once, then we are also guaranteed to terminate. In other words, there are a finite number of state objects and the `update` UDF enumerates them in a monotonic fashion.