



Scheinert, D., Zhu, H., Thamsen, L., Geldenhuys, M. K., Will, J., Acker, A. and Kao, O. (2021) Enel: Context-Aware Dynamic Scaling of Distributed Dataflow Jobs using Graph Propagation. In: 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC), 29-31 Oct 2021, ISBN 9781665443319.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/268166/>

Deposited on: 24 May 2022

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

Enel: Context-Aware Dynamic Scaling of Distributed Dataflow Jobs using Graph Propagation

Dominik Scheinert, Houkun Zhu, Lauritz Thamsen, Morgan K. Geldenhuys,
Jonathan Will, Alexander Acker, and Odej Kao
Technische Universität Berlin, Germany, {firstname.lastname}@tu-berlin.de

Abstract—Distributed dataflow systems like Spark and Flink enable the use of clusters for scalable data analytics. While runtime prediction models can be used to initially select appropriate cluster resources given target runtimes, the actual runtime performance of dataflow jobs depends on several factors and varies over time. Yet, in many situations, dynamic scaling can be used to meet formulated runtime targets despite significant performance variance.

This paper presents *Enel*, a novel dynamic scaling approach that uses message propagation on an attributed graph to model dataflow jobs and, thus, allows for deriving effective rescaling decisions. For this, *Enel* incorporates descriptive properties that capture the respective execution context, considers statistics from individual dataflow tasks, and propagates predictions through the job graph to eventually find an optimized new scale-out. Our evaluation of *Enel* with four iterative Spark jobs shows that our approach is able to identify effective rescaling actions, reacting for instance to node failures, and can be reused across different execution contexts.

Index Terms—Scalable Data Analytics, Distributed Dataflows, Dynamic Scaling, Performance Modeling, Runtime Prediction, Resource Management, Graph Neural Networks.

I. INTRODUCTION

Across all domains the amount of data being generated is continually increasing and consequently the need to process and analyse large volumes of data has become an increasingly essential part of any data processing environment. It is here where scalable data-parallel processing jobs are frequently employed to extract valuable information from this data. Nowadays these processing jobs can be developed with the help of distributed dataflow systems like Spark [1] and Flink [2], which take care of the parallelism, distribution, and fault tolerance. As the management of computational resources is often not directly handled by these distributed dataflow systems, they commonly make use of resource management systems such as YARN [3] or Kubernetes [4] to provision resources for the individual processing jobs.

Yet, as the configuration of clusters is traditionally a manual task, finding near optimal resource allocations is difficult as even frequent users and experts do not always fully understand system and workload dynamics [5], [6]. As a consequence, providing methods which automate this process has become a popular area of research [5]–[10]. The need for more sophisticated solutions has been further reinforced by the rise of cloud computing which has led to an increase in the

diversity of resource and infrastructures, as well as, users data processing requirements [11], [12]. In light of this situation, it is necessary to support users in finding a suitable resource configuration that is in line with their respective requirements.

Various solutions have been proposed regarding the optimization of resource configurations. Some methods are designed for specific processing frameworks [13]–[15], others conduct an iterative profiling strategy [8], [16]–[18], and a third line of work builds runtime models for evaluating possible configurations [7], [19]–[24]. However, even with optimized initial resource allocation, the performance of processing jobs can vary greatly due to factors such as data locality, inference between jobs, and failures which impact the performance over time. Consequently, processing jobs should be continuously monitored to ensure any QoS (Quality of Service) requirements involving job runtimes and resource utilization should be adhered to. At the same time, it is desirable to develop reusable solutions in order to reduce the initial expense of finding near-optimal configurations.

In this paper we present *Enel*, a black-box approach to continuously optimizing processing jobs where the underlying graph structure of data-parallel tasks are explicitly taken into consideration. In contrast to related state of the art methods [21], [25], our approach models a distributed dataflow job and its components as an attributed, directed graph and aims to learn the relationship between nodes and their respective predecessor nodes. This is fostered by attaching descriptive properties to the individual nodes that at best describe the respective context of the job execution, e.g. dataset size, algorithm name, or machine type. The learned relationships and attached context information are then used to predict the individual runtimes and runtime statistics. During online inference, data emitted from the current job execution is gathered and used to continuously update runtime predictions. Based on a specified runtime target and the remaining predicted runtime, the current resource allocation is eventually optimized to bear in mind QoS constraints.

Contributions. The contributions of this paper are:

- A novel modeling approach for dynamic scaling that incorporates scale-out information, descriptive properties of the job execution context, and application metrics, while also exploiting the underlying graph structures of data-parallel tasks in distributed dataflow jobs. We show that hereby, we can achieve accurate and robust prediction results and thus a reliable dynamic scaling mechanism.

- An evaluation of our approach to dynamic scaling in various different experiments. We investigate our approach in the presence of failures, and show that our graph model is reusable in different contexts.
- A repository¹ with all relevant source code and data, accompanied by comprehensive information regarding setups, configurations, and application of our method.

Outline. The remainder of the paper is structured as follows. Section II discusses the related work. Section III describes and discusses our modeling approach. Section IV outlines the architecture decisions behind Enel. Section V presents the results of our comprehensive evaluation. Section VI concludes the paper and gives an outlook towards future work.

II. RELATED WORK

Existing solutions for dynamic resource allocation problems of distributed data processing jobs typically involve choosing an initial configuration offline and subsequently adjusting the resources for the running job online, i.e. during the execution.

This section discusses related work that addresses both of these sub-problems and puts them in relation to Enel.

A. Offline Runtime Prediction

Several model-based offline approaches for selecting cluster configurations exist. They often use runtime data to predict the scale-out and runtime behavior of jobs. The runtime data can be gained either from dedicated profiling with a sample of the dataset or previous full executions [19], [20], [23], [24], [26], [27].

Ernest [19] uses a parametric model for the scale-out behavior of jobs, which is trained on the results of sample runs on reduced input data. This succeeds for programs which exhibit intuitive scale-out behaviors. Ernest uses optimal experiment design to select configurations to test in the sample runs.

Another example is *Vanir* [26], which firstly finds a good-enough initial configuration for profiling runs by using a heuristics method, and then utilizes Mondrian forest based performance model and transfer learning to progressively improve the configuration in the production runs. In addition, Vanir employs transfer learning in the profiling phase to eliminate the cost of profiling runs for similar jobs.

Bell [20] is an approach which can continue to learn the job’s scale-out behavior from historical full executions besides initial runs. It automatically chooses via cross-validation between a parametric model based on that of Ernest and a non-parametric model, which leads to an overall increase in prediction accuracy. Enel uses Bell’s runtime model as part of its process of finding a good initial resource allocation.

Gaining training data for the profiling leads to overhead in both time and to some extent also cost. Meanwhile, metrics from previous executions of a job are not always available. As a possible remedy for this issue, we previously proposed a system [27] that facilitates the global sharing of context aware runtime models, allowing for runtime prediction based

on historical executions of a job by different users [23], [24]. Enel assumes a recurring job and thus, the initial profiling cost can be amortized over time. However, even the most well-trained runtime models cannot anticipate failures in the cluster, which could endanger reaching the runtime target. Likewise, data locality will also have an impact on the runtimes of frequently performed jobs as tasks are not guaranteed to be scheduled on the same nodes where the data files reside. Enel therefore additionally monitors the execution and dynamically scales the cluster as required to reach runtime targets.

B. Progress Estimation and Dynamic Scaling

There are several works that attempt to estimate the progress of iterative distributed data processing jobs [21], [28], [29]. These estimates can then be used to continuously re-evaluate the current resource configuration and make adjustments towards a given runtime target for the job [21]. Besides meeting deadlines, progress estimation can also be used to support scheduling decisions in shared clusters [30]–[32].

Sahni and Vidyarthi [28] propose a dynamic and cost-effective algorithm to schedule deadline-constrained jobs in public clouds. It incorporates information on performance variability among virtual machines and instance acquisition delay. To achieve the cost-effective goal, it scales in a just-in-time schedule manner.

Thonglek and et al. [29] utilizes model-based deep reinforcement learning (RL) to achieve an auto-scaling system for real time processing applications. To address the problem of dynamic input data size and unpredictable availability of resources, it applies deep RL to continuously optimize scaling by observing and learning from the previous settings.

Ellis [21] uses progress estimates as a basis for dynamic re-allocations of resources to meet a runtime target. Progress is estimated by continuously monitoring scale-out behavior of the job’s individual stages and comparing it to that of previous executions. It predicts the remaining runtime of jobs and then scales out to meet the runtime target if needed.

Most previous works either utilize no information on resource availability and cluster loads, or do not consider the directed acyclic graph (DAG) of jobs, which makes it harder to satisfy user constraints in complex environments, e.g. running in a multi-tenant cluster. Enel, however, uses message propagation on an attributed graph to model the progress of distributed dataflow jobs and only uses a single global model, thus being more robust to complex situations like resource competitions and small changes of the execution context.

III. APPROACH

This section presents our approach Enel and its usage to select resources according to formulated runtime targets. As it is designed for reusability, context-awareness, and operates on graphs, we will discuss the individual aspects in the following.

A. Preliminaries

Distributed dataflows are commonly modeled as graphs to represent the temporal relationships between consecutive or

¹<https://github.com/dos-group/enel-experiments>

parallel actions. A directed, weighted, and attributed graph $G^{(k)} = (V^{(k)}, E^{(k)})$ consists of a set of vertices $V^{(k)} = \{v_1, \dots, v_n\}$ and a set of edges $E^{(k)} \subseteq \{(v_i, v_j) | v_i, v_j \in V^{(k)}\}$. One can also write $V(G^{(k)})$ and $E(G^{(k)})$. An edge $e_{ij} \Leftrightarrow (v_i, v_j) \in E^{(k)}$ describes a directed connection between vertex v_i and v_j , and $|e_{ij}|$ is the corresponding edge weight. Thus, the node v_j is then called a neighbor of node v_i , formally written as $j \in \mathcal{N}(i)$. Throughout this paper, we use the notion of $G^{(k)}$ to refer to the k -th component of a distributed dataflow job, e.g. a concrete iteration in the case of iterative dataflows. Consequently, we define a dataflow job as a sequence $D = (G^{(1)}, G^{(2)}, \dots, G^{(n-1)}, G^{(n)})$ of n graphs, each corresponding to a component of the job. This is depicted in Figure 1 and is of importance to our approach.

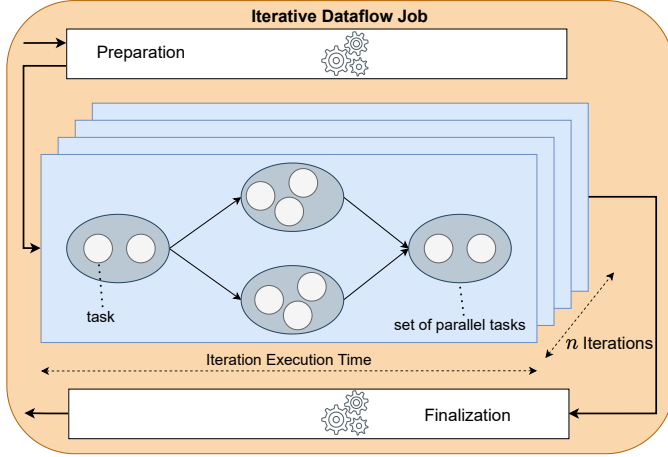


Fig. 1. An iterative distributed dataflow job. Typically, an initial preparation phase is followed by a number of actual iterations of an algorithm, with a finalization phase eventually concluding the computation.

The execution of distributed dataflows is characterized by descriptive properties. These are different based on the concrete dataflow system and execution environment. Examples are job parameters, hardware specifications of the underlying infrastructure, the size of the dataset to be processed, or versions of utilized software. Properties are furthermore available on multiple levels, i.e. a node $v_i \in V(G^{(k)})$ is characterized by different properties than the corresponding component k . We estimate the *execution context* of a job using the entirety of its descriptive context properties.

B. General Idea

While an initial resource allocation for a target runtime can be computed using historical data or profiling, there are a variety of factors influencing the actual runtime, e.g. data locality or failures. In order to meet given runtime targets despite these factors, it is advisable to continuously monitor the progress of a dataflow job and to make changes in the resource allocation, when jobs do not progress as expected.

With our approach *Enel*, we aim at learning to predict the scale-out behavior of jobs by explicitly incorporating the job graph for modeling distributed dataflows. As each graph node is described by its descriptive context properties and also

runtime statistics, we are enabled to make predictions based on locally available information and the respective predecessor nodes. Estimated runtimes and metrics are eventually propagated through the graph to successor nodes, which in turn use them to compute own predictions. Runtime predictions are accumulated in the process of propagation, providing a runtime prediction for the respective graph, as illustrated in Figure 2.

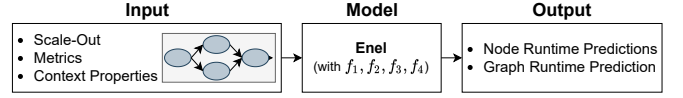


Fig. 2. High-level overview of approach for a single graph.

For training our graph-based model, it is required to present it with a sufficient amount of graphs. Therefore, our approach is especially suitable for iterative dataflows as the number of available graphs to learn from is numerous, since each iteration results in a new component modeled as graph. Later, during monitoring of a job execution, our trained graph model can be used to predict the remaining runtime of the job, and potentially adapt the scale-out to meet the given runtime target.

C. Context Encoding

In order to make use of descriptive properties of a job execution context, we require an efficient, yet robust way of representing these properties. In a first step, we transform each property p to a vector of fixed-size length $\vec{p} \in \mathbb{R}^N$, i.e.

$$\vec{p} = [\lambda, \vec{q}_1, \vec{q}_2, \dots, \vec{q}_{L-1}, \vec{q}_L]^\top, \quad (1)$$

where $\vec{q} \in \mathbb{R}^L$ with $L = N - 1$ is a vector and $\vec{q}_1, \dots, \vec{q}_L$ its elements, obtained from an encoding method as

$$\vec{q} = \begin{cases} \text{hasher}(p) & p \notin \mathbb{N}_0 \\ \text{binarizer}(p) & p \in \mathbb{N}_0 \end{cases} \quad (2)$$

and $\lambda \in \{0, 1\}$ is a binary prefix indicating the utilized method, i.e. hasher or binarizer.

The *hasher* method operates on textual properties (e.g. job parameters) and combines character cleansing and n-gram extraction to derive terms. The occurrence of each resulting term t_s is then counted and inserted at a specific position in the output vector, such that $\vec{q}_j = |t_s|$, where the index j is calculated by a hash function that realizes the term to index mapping. Lastly, the resulting vector \vec{q} is projected on the euclidean unit sphere, such that $\sum_{j=1}^L (\vec{q}_j)^2 = 1$ is ensured.

The *binarizer* method takes a natural number and converts the respective value into its binary representation. As a consequence, each property $p \in \mathbb{N}_0$ (e.g. number of CPU cores) can be encoded as long as $p \leq 2^L$ holds true. This allows for uniquely encoding any number of reasonable size.

As many of the hereby computed vectors can be expected to be sparse, we further employ an auto-encoder to obtain dense, low-dimensional representations. These so called *embeddings* are later used in downstream prediction tasks. Given a vector $\vec{p} \in \mathbb{R}^N$, a decoder function h will try to reconstruct the

original vector from the embedding $\vec{e} \in \mathbb{R}^M$ calculated by the encoder function g , such that $\min \|\vec{p} - h(\vec{e})\|_2^2$ and $M \ll N$.

D. Prediction and Propagation

For rescaling decisions, we mainly focus on three sources of information that can be gathered for each node $v_i \in V(G^{(k)})$:

- *Scale-out*, i.e. the number of workers available at the start and at the end of the respective set of parallel tasks, i.e. a_i and z_i . We from here on use the enriched vector representations $\vec{a}_i, \vec{z}_i \in \mathbb{R}^3$ altered from the parametric model discussed in [19], with e.g. \vec{a}_i derivable as $\vec{a}_i = [1 - \frac{1}{a_i}, \log(a_i), a_i]^\top$.
- *Metrics*, which compactly describe the state of the execution environment and the special characteristics of individual sets of parallel tasks, denoted as $\vec{m}_i \in \mathbb{R}^+$. Examples are CPU or memory utilization.
- *Context properties*, which capture the overall context of the job execution and thus allow for differentiating between related contexts. We compute a context vector $\vec{c}_i \in \mathbb{R}^{3N}$, which is the result of a vector concatenation $\vec{c}_i = \vec{u}_i \parallel \vec{v}_i \parallel \vec{w}_i$, with $\vec{u}_i, \vec{v}_i, \vec{w}_i \in \mathbb{R}^N$. The latter denote mean vectors of context embeddings computed in Subsection III-C, where \vec{u}_i covers all always available properties (e.g. job signature), \vec{v}_i accounts for properties which are not necessarily uniformly recorded (e.g. software versions), and \vec{w}_i is the mean vector of all properties unique to the respective set of parallel tasks (e.g. number of tasks, attempt ID).

These information will be used to compute messages and propagate them through the graph. We design multiple functions to achieve our goal of graph-based runtime prediction, which we then use to derive reasonable rescaling recommendations. Given a node $v_i \in V(G^{(k)})$, we predict its rescaling *overhead* using a function $f_1 : \mathbb{R}^+ \rightarrow \mathbb{R}$ that takes into account the enclosing context, the metrics recorded for this node, the start and end scale-out, and the fraction of time the node spent in different scale-outs. Formally, we write

$$\hat{o}_i = f_1(\vec{c}_i, \vec{m}_i, \vec{a}_i, \vec{z}_i, r_i), \quad (3)$$

where $\hat{o}_i \in \mathbb{R}$ denotes the predicted overhead, and r_i is the aforementioned fraction of time. Taking the overhead into account, e.g. due to changed data I/O reflected in captured metrics, we subsequently predict the node *runtime* with

$$\hat{t}_i = f_2(\vec{c}_i, \vec{m}_i, \vec{z}_i, \hat{o}_i), \quad (4)$$

using a function $f_2 : \mathbb{R}^+ \rightarrow \mathbb{R}$, with \hat{t}_i being the predicted runtime. Note that f_2 only considers the end scale-out, but incorporates the predicted overhead to accurately model the runtime. While \hat{t}_i is the predicted runtime for a particular node, we also compute the total prospective runtime up to a certain node by accumulating the propagated predictions. Thus, for any node v_i , the accumulated runtime is established with

$$\hat{t}t_i = \hat{t}_i + \max_{j \in \mathcal{N}(i)} \hat{t}t_j. \quad (5)$$

Consequently, for the last node in a graph, $\hat{t}t_i$ is also the predicted total runtime of the graph.

In order to predict the remaining runtime of a running dataflow job, we need to ensure that the respective function inputs reflect the current job execution as good as possible. As the enclosing execution context is static by definition after a single dataflow job execution, the start and end scale-out for future iterations are ideally equal (i.e. we assume a non-changing scale-out after we have adapted it), an important remaining thing we need to take care of are *metrics*. Thus, we aim at learning the metrics of a node based on its context, scale-out, and the metrics of its predecessor nodes. This would allow us to predict metrics for nodes and propagate these through the graph in order to effectively calculate the desired runtimes. In a first step, we aim at learning the relevance of predecessor nodes and their metrics in light of the enclosing context. Inspired by [33], we write $\vec{x}_i = \vec{a}_i \parallel \vec{c}_i \parallel \vec{z}_i$ and compute

$$|e_{ij}| = \frac{\exp(\vec{a}^\top \sigma(f_3(\vec{x}_i, \vec{x}_j)))}{\sum_{k \in \mathcal{N}(i)} \exp(\vec{a}^\top \sigma(f_3(\vec{x}_i, \vec{x}_k)))}, \quad (6)$$

i.e. the edge weights, where $f_3 : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a function that transforms the respective context vectors, σ is a non-linear activation, and \vec{a} is a learnable attention weight vector. The scalar result is subsequently normalized using softmax, such that the sum of all edge weights for incoming edges of a node equals one. Eventually, metrics are predicted with

$$\vec{m}_i = \sum_{j \in \mathcal{N}(i)} |e_{ij}| \cdot f_4(f_3(\vec{x}_i, \vec{x}_j), \vec{m}_j), \quad (7)$$

where $f_4 : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a function that transforms predecessor metrics given the enclosing context. At last, we compute \vec{m}_i as a weighted average of vectors. We are thus estimating and propagating the metric vectors of successor nodes for which actual metrics are not necessarily available yet. The predictions can in turn be used to estimate the corresponding runtimes, which effectively allows us to make predictions for remaining iterations of a respective dataflow job.

In a last step, we explicitly establish a relationship between multiple iteration graphs with the use of summary nodes. For a graph $G^{(k)}$, the corresponding summary nodes $P^{(k)}$ and $H^{(k)}$ are installed as predecessor nodes to the respective root node(s) of the next graph $G^{(k+1)}$. The summary node $P^{(k)}$ belongs to the component k of the current dataflow job execution, and carries information about the start and end scale-out of the component, and mean vectors of the respective context and metric vectors of graph nodes. The summary node $H^{(k)}$ acts as a historical reference point, where the previously mentioned node information are each averaged across the β most similar historical summary nodes of the target component. These β nodes are selected based on scale-out proximity. The purpose of $H^{(k)}$ is to stabilize the inference process during the propagation of predictions. The summary nodes are only considered for the metric vector prediction.

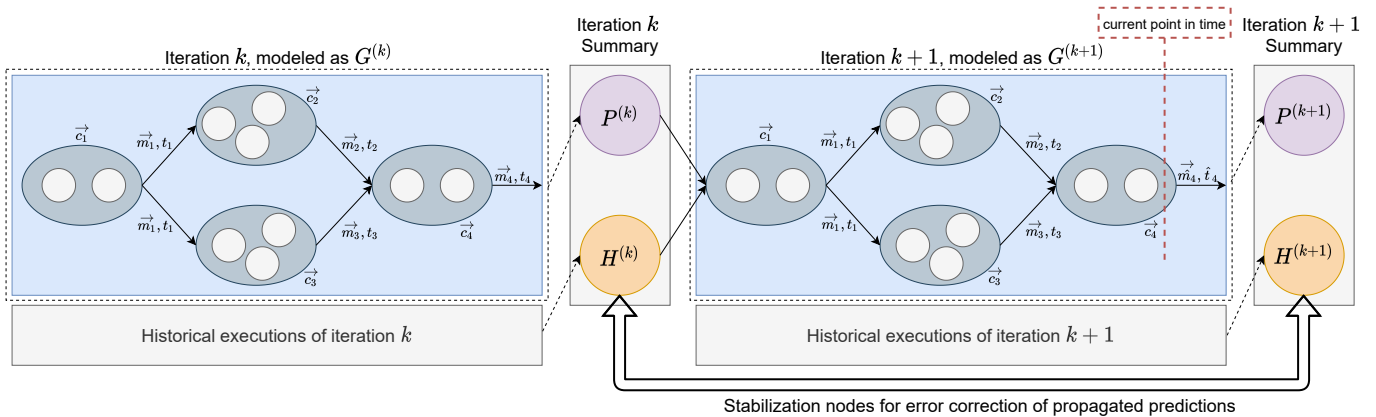


Fig. 3. The general idea of Enel. The task sets of each iteration are modeled as graph, and predictions of runtimes and metric vectors are propagated through the graph. We further establish an explicit relationship between multiple iteration graphs by using summary nodes of an iteration, both of the current iteration and its historical executions, as predecessor nodes of the root node(s) of the respective successor iteration graph.

IV. ARCHITECTURE

Enel is devised as a blackbox approach, i.e. it is designed as an external component that does not make assumptions about the availability of specific statistics or monitoring data. A user can decide which information to use for representing the job execution context and runtime statistics. Our approach only requires information about the scale-out at any particular point in time and the fraction of time spent in either start scale-out or end scale-out (e.g. for a set of tasks), which can be fairly assumed to be present in each dataflow system.

A. Interaction and Usage

Given a distributed dataflow job, the idea follows that a respective dataflow system can consult our external service and receive a scale-out recommendation for a user-specified runtime target. Upon each request, Enel tunes a pre-trained model for the targeted job with the most recent runtime information. It then constructs the remaining component graphs using the static component characteristics gathered from historical job executions, and eventually uses the tuned model for inference on the connected component graphs. In this process, predictions are made and pushed through the graphs, until the final component is reached. These operations are repeated for any valid scale-out over a defined range and the accumulated runtime predictions are then used to select the configurations which best comply with the runtime targets. If used for finding a good initial resource allocation, Enel adheres to this process with slight modifications. As the first component of a graph has no predecessor nodes which could be exploited, Enel uses a simple runtime model, namely Bell [20], on the historical data of the first graph component, and adds the predicted values to its own given the remaining graph components.

B. Prototypical Implementation

For our experiments, we developed our prototype with a lightweight web service implemented with Python. We choose Apache Spark as the distributed dataflow system and schedule Spark applications using the Kubernetes resource manager.

In this scenario, we implement a listener within Spark that gathers runtime statistics and facilitates communication with our service. Regarding metrics used in our optimization process, five were selected directly from within the Spark listener, namely CPU utilization, Shuffle R/W, Data I/O, the fraction of time spent in garbage collection, and the ratio of memory spilled to disk to peak execution memory. Enel uses these metrics together with context properties to predict the remaining runtime and, if required, recommends an optimized scale-out. This scale-out recommendation is then used within our Spark listener to renegotiate resources with Kubernetes.

C. Prediction Model

For modeling we employ soft computing, i.e. we realize each of our differentiable functions f_1 , f_2 , f_3 , and f_4 as a two-layer feed-forward neural network. The latter are good function estimators with two layers being sufficient to distinguish data that is not linearly separable. Operating on graphs and propagating messages eventually makes our model a spatial graph neural network, as we are defining our calculations in the vertex domain by leveraging the graph structure and aggregating node information. With a total of 5155 learnable parameters, our model allows for training even using a CPU.

V. EVALUATION

This section presents our infrastructure setup and experiment design, followed by a discussion of the obtained results. Further model implementation details and technical information are provided in our repository².

A. Infrastructure Setup

All experiments were done on a multi-tenant Kubernetes cluster spanned over 50 machines, which makes each of our scheduled workloads compete with other workloads in the cluster about resources. Of the machines used for deploying the cluster, 42 were also used to roll out a Hadoop Distributed File System (HDFS), where the datasets and spark jobs

²<https://github.com/dos-group/enel-experiments>

TABLE I
HARDWARE & SOFTWARE SPECIFICATIONS

Hardware	
CPU, vCores	Intel(R) Xeon(R) CPU @ 3.30 GHz, 8
Memory	16 GB RAM
Network	connected by a single 1 GB switch
Software	
Linux	Kernel 4.15.0
Others	Python 3.8.0, Kubernetes 1.18.10 Hadoop 2.8.3, Scala 2.12.11 Spark-Operator ³ 1.1.3, Spark 3.1.1 PyTorch 1.8.0, PyTorch Geometric 1.7.2
Spark	
Driver CPU	6 cores
Driver Memory	10240 MB (+ 2048 MB)
Executor CPU	6 cores
Executor Memory	10240 MB (+ 2048 MB)

TABLE II
OVERVIEW OF BENCHMARK JOBS

Job	Dataset	Input Size	Parameters
LR	Multiclass	27 GB	20 iterations
MPC	Multiclass	27 GB	20 iterations, 4 layers with 200-100-50-3 perceptrons
K-Means	Points	48 GB	10 iterations, 8 clusters
GBT	Vandermonde	35 GB	10 iterations, "Regression" configuration

were located. Each Spark application scheduled within the Kubernetes cluster used a scale-out range from 4 to 36 Spark executors. Further details are reported in Table I.

B. Experiment Design

In order to assess the applicability of our approach, we conduct a series of experiments, in which we compare ourselves to related works and simulate the prediction robustness by injecting synthetic failures during spark application executions. We are primarily interested in investigating if and how good our single model, which is reusable due to its context-awareness, performs in comparison to related work that utilizes an ensemble of specialized models.

1) *Jobs*: We used four Spark jobs as benchmarks, namely Logistic Regression (LR), Multilayer Perceptron Classifier (MPC), K-Means, and Gradient Boosted Trees (GBT). Table II shows the jobs and the respective input parameters. The implementation are based on Spark MLLib⁴, a library for implementing distributed machine learning algorithms.

2) *Datasets*: We used three different datasets for the benchmark jobs. All datasets were generated synthetically.

- **Multiclass**: A classification dataset with 3 classes and 200 features. The dataset was generated using the classification generator from scikit-learn⁵.

- **Vandermonde**: A regression dataset with 19 features. The dataset was generated using our own generator by explicitly computing the Vandermonde matrix. For this, data points were randomly generated following a polynomial of degree 18 with added Gaussian noise.
- **Points**: A two-dimensional dataset with points sampled from a Gaussian Mixture Model of eight normal distributions with random cluster centers and equal variances.

3) *Objective: Prediction Performance*: For each dataflow job, we first perform an initial ten runs without dynamic scaling. The execution data of these ten runs are gathered and later used to train models for initial resource allocation and dynamic scaling. For dynamic scaling, we compare our approach to our own previous work *Ellis* [21], as it is most related. For initial resource allocation, we rely on the method presented in *Ellis* to guarantee an equal and fair starting point for both dynamic scaling approaches. After initial model fitting, we conduct various adaptive runs and hereby investigate the performance of our approach. While *Ellis* is by design fitting and using a new set of models for every new run, we train with *Enel* a new model from scratch after every fifth run and fine-tune this model then on each of the subsequent five runs.

4) *Objective: Prediction Robustness*: In order to test the reliability of our method in the presence of partial failures, i.e. the loss of executors while a job is executing, we created a failure injector to inject failures into the running system. Our failure experiments were structured in such a way that they executed across the full time duration of each job execution, where a single failure was injected at any random second within a 90 second interval as long as at least 4 Spark executors were available. Injections resulted in the immediate termination of Spark executor pods, which in turn required Spark, if configured accordingly, to start up new or additional executors to cover for that. We conduct multiple runs which are subject to failure injections, distributed across two phases and interrupted by a handful of normal runs.

5) *Evaluation Metrics*: We assess the performance of our approach using multiple metrics. First, we utilize metrics introduced in [21], namely CVC (runtime constraint violation count) and CVS (runtime constraint violation sum). Moreover, we investigate the time to train required for our approach.

C. Results

The general prediction performance over all runs is illustrated in Figure 4. It is observable that *Enel*, due to its complexity and the limited training data, is initially often volatile in its prediction performance. This is especially evident for runs with MPC or K-Means. Over time though, the predictions tend to stabilize and converge, with the exception of single outliers. In Table III, we report mean and median values of CVC and CVS for each algorithm over multiple ranges. The results support our previous finding of improved prediction performance over time. In general, the predicted runtime is many times only slightly off and regularly below the respective formulated runtime target. From this we conclude that achieving results comparable to state-of-the-art solutions is

⁴<http://spark.apache.org/mllib/>, accessed 2021-06-08

⁵<https://scikit-learn.org/>, accessed 2021-06-08

TABLE III
EVOLUTION OF PREDICTION PERFORMANCE OVER RUNS

		Runs 11-22		Runs 22-33		Runs 33-44		Runs 44-55		Runs 55-65	
		\bar{x} ¹	\tilde{x} ²	\bar{x}	\tilde{x}	\bar{x}	\tilde{x}	\bar{x}	\tilde{x}	\bar{x}	\tilde{x}
LR	CVC	0.55	1.00	0.36	0.00	0.73	1.00	1.00	1.00	0.91	1.00
	CVS	0.33m	0.11m	0.12m	0.00m	1.40m	0.31m	1.44m	1.35m	1.58m	0.92m
MPC	CVC	0.73	1.00	0.27	0.00	0.45	0.00	0.82	1.00	0.27	0.00
	CVS	2.41m	1.70m	1.12m	0.00m	1.95m	0.00m	2.46m	1.22m	1.23m	0.00m
K-Means	CVC	0.55	1.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CVS	5.89m	0.60m	4.03m	0.00m	0.00m	0.00m	0.00m	0.00m	0.00m	0.00m
GBT	CVC	0.36	0.00	0.36	0.00	0.82	1.00	0.73	1.00	0.27	0.00
	CVS	0.55m	0.00m	0.85m	0.00m	5.21m	2.05m	2.85m	1.04m	0.27m	0.00m

¹ The *mean* value of the metric over the respective runs.
² The *median* value of the metric over the respective runs.

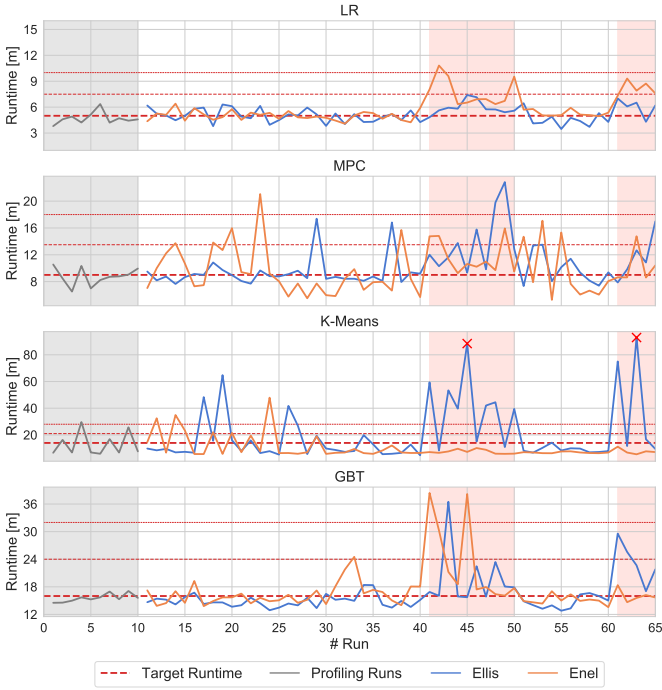


Fig. 4. Initial profiling (highlighted in grey), followed by alternating phases of normal runs and anomalous runs (highlighted in red). Our graph model Enel is improving over time and shows a certain robustness against failures.

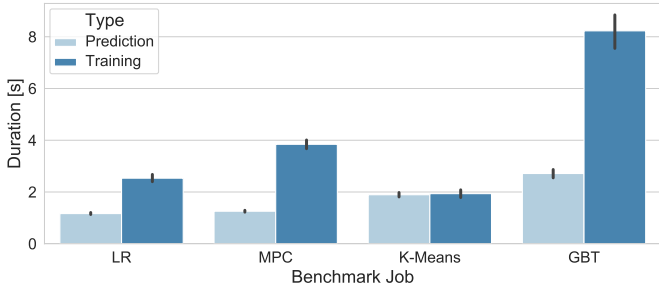


Fig. 5. Time in seconds required for fine-tuning an Enel model and eventually using it for predictions. The variance for some jobs stems from the fact that training takes differently long dependent on the number of graphs to process.

possible given a certain amount of training data to work with. This is interesting, as we only utilize a single model operating on individual graphs and perform training only occasionally, whereas Ellis trains a new set of models after every run, with each model being specialized for a concrete job component. From this we can deduce that the enclosing execution contexts were sufficiently represented using descriptive properties, and that our model was able to make context-aware predictions with the computed embeddings. Moreover, learning the relationships between graphs appears to be beneficial.

In terms of prediction robustness, e.g. in the presence of node failures, Enel reveals a certain stability, where especially the results on K-Means are worth mentioning. In case of initial problems during anomalous runs, the prediction error is significantly corrected with the next scheduled retraining, as can be seen for GBT after iteration 45. For LR, the decisions of Enel tend to lead to a slightly longer runtime, however, as the runtime target is in general quite low, the magnitude of violation is tolerable. Meanwhile, the comparative method Ellis faces difficulties reacting to certain failures, which maximizes for MPC and K-Means. The problems persist also for the second phase of anomalous runs. As indicated by the red crosses, certain runs even had to be terminated manually because a timely application end was not realistic. We thus conclude that Enel is able to react more appropriately to failures due to its consideration of runtime statistics.

Lastly, we discuss the required time for training and prediction. Figure 5 summarizes the measured times for each job class. It can be seen that inference is a fast and robust process, whereas the fine-tuning of a model depends on the number of provided graphs. As a consequence, GBT requires more time as the job is internally decomposed into many components. Either way, we argue that the cost of a few seconds is acceptable given that dataflow jobs usually run for longer periods of times. Note that we did not list the time required for the initial training of the models, as this is a task that can be scheduled for execution periodically.

VI. CONCLUSION

This paper presented *Enel*, a novel dynamic scaling approach for distributed dataflows. It explicitly incorporates the graph structure of distributed dataflow jobs as well as descriptive properties of job execution contexts to develop a better understanding of how scale-out behaviors are impacted during execution. At any point in time, *Enel* is able to estimate the remaining runtime of a targeted job and then compares it to projected runtime targets. The resource allocation is then accordingly adapted to at best comply with existing constraints. With the consideration of descriptive context properties, *Enel* can be reused across multiple iterations of an iterative dataflow job instead of training individual models. Moreover, it is a black-box approach and thus can be used with different resource managers and for different dataflow systems.

We implemented *Enel* as a graph neural network with multiple task-specific components, which for instance realized the prediction of runtimes and runtime statistics. Our experiments showed that as more data becomes available, *Enel* is able to improve its predictions and thus achieves comparable results to related works and additionally is able to handle failures in a more robust manner. The advantage of our approach is especially significant for iterative dataflow jobs which incorporate small changes in terms of configurations or inputs. This is as a result of the reusability of our model and thus knowledge from previous related executions is recycled which saves on resource and energy usage.

In the future, we want to further investigate possibilities of exploiting the dataflow graph information for optimized resource allocation, dynamic scaling, and scheduling of multiple dataflow jobs. As this is a common scenario in shared clusters, sophisticated solutions are required and desired.

ACKNOWLEDGMENTS

This work has been supported through grants by the German Federal Ministry of Education and Research (BMBF) as BI-FOLD (funding mark 01IS18025A) and WaterGridSense 4.0 (funding mark 02WIK1475D).

REFERENCES

- [1] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *HotCloud*. USENIX, 2010.
- [2] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache flink™: Stream and batch processing in a single engine," *IEEE Data Eng. Bull.*, vol. 38, no. 4, 2015.
- [3] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache hadoop YARN: yet another resource negotiator," in *SOCC*. ACM, 2013.
- [4] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, "Large-scale cluster management at google with borg," in *EuroSys*. ACM, 2015.
- [5] K. Rajan, D. Kakadia, C. Curino, and S. Krishnan, "Perforator: eloquent performance models for resource optimization," in *SoCC*. ACM, 2016.
- [6] P. Lama and X. Zhou, "AROMA: automated resource allocation and configuration of mapreduce environment in the cloud," in *ICAC*. ACM, 2012.
- [7] A. Verma, L. Cherkasova, and R. H. Campbell, "ARIA: automatic resource inference and allocation for mapreduce environments," in *ICAC*. ACM, 2011.
- [8] C. Hsu, V. Nair, T. Menzies, and V. W. Freeh, "Micky: A cheaper alternative for selecting cloud instances," in *CLOUD*. IEEE, 2018.
- [9] M. Casimiro, D. Didona, P. Romano, L. E. T. Rodrigues, and W. Zwaenepoel, "Lynceus: Tuning and provisioning data analytic jobs on a budget," *CoRR*, vol. abs/1905.02119, 2019.
- [10] A. Fekry, L. Carata, T. F. J. Pasquier, A. Rice, and A. Hopper, "Tuneful: An online significance-aware configuration tuner for big data analytics," *CoRR*, vol. abs/2001.08002, 2020.
- [11] M. Bux and U. Leser, "Parallelization in scientific workflow management systems," *CoRR*, vol. abs/1303.7195, 2013.
- [12] E. Deelman, K. Vahi, M. Rynge, R. Mayani, R. F. da Silva, G. Papadimitriou, and M. Livny, "The evolution of the pegasus workflow management software," *Comput. Sci. Eng.*, vol. 21, no. 4, 2019.
- [13] A. Verma, L. Cherkasova, and R. H. Campbell, "Resource provisioning framework for mapreduce jobs with performance goals," in *Middleware*. Springer, 2011.
- [14] A. D. Ferguson, P. Bodík, S. Kandula, E. Boutin, and R. Fonseca, "Jockey: guaranteed job latency in data parallel clusters," in *EuroSys*. ACM, 2012.
- [15] H. Al-Sayeh and K. Sattler, "Gray box modeling methodology for runtime prediction of apache spark jobs," in *ICDEW*. IEEE, 2019.
- [16] O. Alipourfard, H. H. Liu, J. Chen, S. Venkataraman, M. Yu, and M. Zhang, "Cherry-pick: Adaptively unearthing the best cloud configurations for big data analytics," in *NSDI*. USENIX, 2017.
- [17] C. Hsu, V. Nair, V. W. Freeh, and T. Menzies, "Arrow: Low-level augmented bayesian optimization for finding the best cloud VM," in *ICDCS*. IEEE, 2018.
- [18] C. Hsu, V. Nair, T. Menzies, and V. W. Freeh, "Scout: An experienced guide to find the best cloud configuration," *CoRR*, vol. abs/1803.01296, 2018.
- [19] S. Venkataraman, Z. Yang, M. J. Franklin, B. Recht, and I. Stoica, "Ernest: Efficient performance prediction for large-scale advanced analytics," in *NSDI*. USENIX, 2016.
- [20] L. Thamsen, I. Verbitskiy, F. Schmidt, T. Renner, and O. Kao, "Selecting resources for distributed dataflow systems according to runtime targets," in *IPCCC*. IEEE, 2016.
- [21] L. Thamsen, I. Verbitskiy, J. Beilharz, T. Renner, A. Polze, and O. Kao, "Ellis: Dynamically scaling distributed dataflows to meet runtime targets," in *CloudCom*. IEEE, 2017.
- [22] S. Shah, Y. Amannejad, D. Krishnamurthy, and M. Wang, "Quick execution time predictions for spark applications," in *CNSM*. IEEE, 2019.
- [23] J. Will, L. Thamsen, D. Scheinert, J. Bader, and O. Kao, "C3o: Collaborative cluster configuration optimization for distributed data processing in public clouds," in *IC2E*. IEEE, 2021.
- [24] D. Scheinert, L. Thamsen, H. Zhu, J. Will, A. Acker, T. Wittkopp, and O. Kao, "Bellamy: Reusing Performance Models for Distributed Dataflow Jobs Across Contexts," in *Cluster*. IEEE, 2021.
- [25] J. Koch, L. Thamsen, F. Schmidt, and O. Kao, "SMiPE: Estimating the progress of recurring iterative distributed dataflows," in *PDCAT*. IEEE, 2017.
- [26] M. Bilal, M. Canini, and R. Rodrigues, "Finding the right cloud configuration for analytics clusters," in *SoCC*. ACM, 2020.
- [27] J. Will, J. Bader, and L. Thamsen, "Towards collaborative optimization of cluster configurations for distributed dataflow jobs," in *Big Data*. IEEE, 2020.
- [28] J. Sahni and D. P. Vidyarthi, "A cost-effective deadline-constrained dynamic scheduling algorithm for scientific workflows in a cloud environment," *IEEE Trans. Cloud Comput.*, vol. 6, no. 1, 2018.
- [29] K. Thonglek, K. Ichikawa, C. Sangkeetrakarn, and A. Piyatunrong, "Auto-scaling system in apache spark cluster using model-based deep reinforcement learning," in *Heuristics for Optimization and Learning*. Springer, 2021.
- [30] T. Renner, L. Thamsen, and O. Kao, "Coloc: Distributed data and container collocation for data-intensive applications," in *BigData*. IEEE Computer Society, 2016.
- [31] G. Xu, C. Xu, and S. Jiang, "Prophet: Scheduling executors with time-varying resource demands on data-parallel computation frameworks," in *ICAC*. IEEE Computer Society, 2016.
- [32] J. Shi and J. Lu, "Performance models of data parallel DAG workflows for large scale data analytics," in *ICDE*. IEEE, 2021.
- [33] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *CoRR*, vol. abs/2105.14491, 2021.