RESPONSE LETTER FOR PVLDB VOL 17 PAPER #1785 (REVISION)

We thank all the reviewers for their constructive feedback. In this document, we summarize the revisions that we have performed based on reviewers’ comments, followed by the revised manuscript, where the modified text is highlighted in magenta.

META-REVIEWS


We have significantly improved the writing of our paper by addressing all of these requests. Please see our detailed response to each reviewer below.

V2: New experiments or better explanation of the experiments (R1.D3, R1.D4, R1.D6, R2.W1, R4.D1)

We have performed a major revision to add new experimental results to address all of the requests above. Please see our detailed response to each reviewer below.

REVIEWER 1

All of our references to the appendices refer to our tech report online: https://chenghao.pages.dev/papers/vldb24-lyu-tr.pdf [31]

R1.W1/D1: The proposed approach is complex involving a hybrid compile-time/runtime optimization method and modeling techniques. This complexity could pose implementation challenges in practice. Related to W1, even bringing seemingly simple solutions to a mature product at cloud scale is challenging. I would like to see more discussion around the practical aspects of integrating such a solution.

To respond to the reviewer’s comment, we added a paragraph at the beginning of Section 6 on the implementation details. We implemented different modules of our system as plugins into a Spark cluster. We further reported 1) the time overhead of collecting runtime statistics; 2) the time overhead of compile-time optimization; 3) the time overhead of runtime optimization; 4) model training time. Overall, these overheads appear to be modest enough to permit future large-scale real-world prototyping and deployment.

R1.W2/D2: The paper is quite dense to read. It would be good to provide more intuitive understanding early on before delving deeper. Related to W2, there are a lot of moving parts of the proposed solution primarily because the paper tries to address the optimization problem at a finer-granularity than prior work. This complexity might be mandated however it would be good to provide more intuitive understanding to the readers before jumping straight into the details.

We thank the reviewer for the suggestion. We have significantly improved Section 3 to provide more intuitive understanding of the fine-grained parameter tuning problem and our overall approach, by leaving out details to later sections. Please see the highlighted text in Section 3. In particular,

• we improved the opening paragraphs of the subsections;
• we shortened Section 3.2 to highlight the benefits and complexity of fine-grained parameter tuning, by keeping the main message and a clear example in each paragraph;
• we rewrote Section 3.3.1 to give a more direct description of our compile-time/runtime optimization approach;
• we improved Section 3.3.2 to show a clear distinction between single-objective and multi-objective optimization, before delving into the mathematical definition.

R1.W3/D3: The experimental evaluation did not give the readers a good sense of current limitations of the proposed approach i.e., where it fails and hence what could be some interesting future extensions to address those limitations. Related to W3, I would like to see additional experiments including cases where the proposed approach is either too slow or does not result in any improvement in the workload latency, for example. Plus a presentation of a few insights around such cases.

We appreciate the reviewer’s feedback and have added a new experiment, Expt 8, in Section 6.3 to discuss the limitations of our proposed approach. In particular, Figure 12(a) and 12(b) show the per-query latency using HMOOC3+ or the default configuration. Among the 22 TPC-H and 102 TPC-DS queries, HMOOC3+ significantly outperforms the default for most of the queries, while being similar (within 2s latency improvement) for three short-running queries. HMOOC3+ loses to the default configuration only for one short-running query (h14), mainly due to the model error in predicting its latency. As we use WMAPE as the training loss to give more weight to long-running queries, short queries can have inaccurate predictions and hence experience a suboptimal query plan or insufficient resource allocation. Future work could focus on improving latency prediction for short-running queries by developing specialized models or incorporating adaptive mechanisms based on a learned long-short query classifier.

R1.D4: Related to experiments, it appears 19 parameters were considered, how does the optimization approach scale with the number of parameters?

We significantly extended our paper, including new techniques and new writing, to address the dimensionality of the parameter space.

(1) Initial feature selection: We first clarify that Spark has over 100 parameters and we selected 19 features as follows. (a) We keep query plan parameters $\theta_p$ and partition related parameters $\theta_s$, which can be replicated to have one copy for each query stage in our
(2) Feature importance score (FIS) based parameter filtering: We conducted additional experiments to compute the feature importance scores (FIS) of the 19 parameters in our trained latency and data shuffle models. The FIS of a parameter here is defined to be the increase of the model error if the values of this parameter are randomly shuffled in the training set. Due to space limitations, we include one example in Figure R1, but leave other plots to Figure 21-23 in Appendix B.3. Our main observation is that the FIS values of the parameters, sorted in decreasing order, follow a long tail distribution. The parameters at the tail, with a cumulative FIS less than 5% in WMAPE, do not affect the model performance by much and hence can be filtered out. For instance, the threshold of 5% cumulative FIS means an increment of at most 0.7% in WMAPE for TPC-H and 1.2% for TPC-DS in the subQ model.

(3) Sampling with FIS-based parameter filtering: Then we introduced three sampling methods that could leverage FIS-based selection:
(a) random sampling with or without parameter filtering; (b) Latin hypercube sampling with or without parameter filtering; (c) adaptive grid search with inherent parameter filtering, as we described in the new paragraph starting with “Sampling methods” on Page 8, which is the last paragraph of Section 5.1.1.

(4) New experimental results: We added new results in Figure 11(b)-(c) to show performance metrics (hypervolume and solving time) of the above three sampling methods with parameter filtering. Among them, we report the adaptive grid search to be an overall winner, achieving the highest hypervolume among all FIS-based methods while keeping the solving time under 2s for most queries tested.

R1.D5: Can the paper provide some citation or intuition around why the average runtime of MOO of 0.62-0.83 second acceptable in practice (I understand it's already much better than the other approaches but is it still acceptable or potentially too high?)

We aim to set the solving time to be on the same scale (sub-seconds to seconds) as the VM/Container startup time in the cloud as pointed out by [1, 32, 50]. This is also consistent with our prior experience in solving the resource optimization problem for AliCloud’s serverless computing [32].

To better balance the query execution time and solving time, our revision further introduced a (simple) runtime adaptive scheme to adjust the sampling budget based on the predicted latency. This technique is introduced at the end of Section 5.1.1 and used in our new end-to-end experiments in Section 6.3. Overall, it allows us to improve results a little further while keeping an average solving time of 0.7-0.8s.

R1.D6: Figure 10(g), the paper states that all queries for TPC-H and -DS are used but this figure shows only a subset (perhaps because of space constraints). What was the performance on other queries?

We appreciate the reviewer’s feedback. We have added Expt 8 in Section 6.3 to show the performance of all TPC-H and TPC-DS queries.

R1.D7: Minor: I didn’t understand why in Sec 6.1, experiments were numbered Expt 1, 2, 5, 6, 7, 9 and 10?

We have fixed the numbering of experiments throughout Section 6.

REVIEWER 2
All of our references to the appendices refer to our tech report online: https://chenghao.pages.dev/papers/vldb24-lyu-tr.pdf [31]

R2.W1: The experiments only use the synthetic TPC benchmarks, which rely on very simple data distributions. It is not that difficult to correctly predict selectivity for uniformly distributed data. Hence it is essential that the experiments also include real data.

We agree with the reviewer that real-world datasets would be valuable in the experimental study. However, it is difficult to obtain a large real dataset within the revision window: in our effort to contact a cloud service provider for data sharing, we learned that any dataset involving query plans cannot be shared due to strict privacy constraints.

Nonetheless, we managed to strengthen our experimental study by considering the data skew present in TPC-DS, which has a hybrid domain of synthetic and real-world data [43, 47]. We added a new experiment, Expt3 in Section 6.1, to understand the impact of data skew on our modeling techniques. Our major results are the following.
(1) TPC-DS queries experience skewed data distributions in both base relations and intermediate RDDs.
(2) The correlations between model errors and skewness are close to 0 during runtime optimization. This indicates that once the skewness is captured by runtime statistics, the model can work quite well regardless of the skewness level.

R2.W2: The notion of a subquery is not well-defined in the paper. The authors seem to define it based on shuffle operations and data exchange operators in the physical query plan. However, a rewritten plan can have different shuffle requirements and hence different subqueries. And when both inputs of a join are already partitioned in the same way, then Spark automatically avoids shuffling, even for the shuffled hash join. This needs a more formal treatment in the paper.

To respond to the reviewer’s comment, we made a number of revisions to improve the paper.

(1) We rewrote Section 4.1 to clarify the distinction between runtime models and compile-time models for optimization.
   - For runtime optimization, the Spark optimizer already offers the collapsed logical query plan (LQP) and query stages (QS) in the physical plan. Hence, we collect these data structures and build runtime LQP and QS models to enable fine-grained tuning.
   - For compile-time optimization, however, Spark does not provide any data structures that would suit our goal of fine-grained tuning. For this reason, we introduce the notion of compile-time subquery (subQ) to denote a group of logical operators that will correspond to a query stage (QS) when the logical plan is translated to a physical plan. We build a subQ model for fine-grained tuning at compile time. The output of compile-time optimization is fed as a starting configuration to the runtime optimization, which will subsequently adjust parameters and physical plans as AQE proceeds.

(2) As a further enhancement, we also introduced a compile-time technique that samples multiple physical plans for each query, as described in Section 5.1.3.

(3) We added a new experiment, Expt 10 in Section 6.3, to show that by searching multiple query plans at compile time, several long-running queries with complex query structures have gained additional performance improvements.

R2.W3: Similarly, the optimizer seems to treat subqueries in isolation when exploring optimal configurations for query plan and query stage parameters. In reality, earlier partitioning and implementation decisions affect the cost and running time of downstream stages. For instance, in isolation a downstream subquery plan P1 may look better than alternative P2, but if P2 can reuse a DataFrame that is still cached in memory, or can exploit a specific partitioning established upstream, it may be the better choice overall. The proposed optimizer partially takes this into account through collapsed LQP, but that only gives a "lookahead" of 1 stage. So the problem remains when optimizing across multiple downstream stages. Please discuss and clarify in the paper.

We clarify that our prior compile-time optimization did simplify the problem by only considering one fixed plan structure based on the default configuration, and treat the compile-time subqueries (subQs) independently to enable efficient algorithms.

During the revision, we improved our compile-time optimization with multi-query plan search, as described in Section 5.1.3. This technique is shown to provide additional performance improvements for several long-running queries with complex query structures.

The output of compile-time optimization is then fed as a starting configuration to the runtime optimization, which will subsequently adjust parameters and physical plans as AQE proceeds. We clarified this in the first paragraph of Section 5.2. More specifically, our AQE-based runtime optimization is a progressive procedure: upon the completion of each query stage, the reminder of the query, represented as a collapsed LQP, will be reoptimized. A downstream stage may be optimized multiple times as the collapsed LQP becomes updated. However, at the moment when all of its preceding stages finish execution, the stage gets the chance of being optimized (the last time) with all the latest statistics. Therefore, the characteristics of all the previously completed stages will be considered in this last optimization decision.

R2.W4: It is not completely clear how Spark configuration optimization interacts with optimization of the distributed algorithm itself, for instance the number of partitions [1]. Please discuss this briefly in related work.

We thank the reviewer for the suggestion. We added Li et al. [26] to Section 2. Specifically, Spark AQE is able to take those distributed algorithms as separate query optimization rules, and our framework can then treat the hyperparameters of those algorithms as SQL parameters and optimize those algorithms by tuning their hyperparameters in compile-time/runtime optimization. Due to space limitations, we leave more details to Appendix C.

R2.D1: The paper repeatedly touts the advantages of multi-objective optimization and emphasizes its soundness. Reduce some of that redundancy and reword the misleading soundness claim. Single-objective optimization is also sound.

We modified the text in the introduction to clarify that multi-objective optimization is a theoretically sound approach to adapting between cost and performance. We removed other mentions in the rest of the paper.

R2.D2 It is not clear why optimization time cannot exceed a few seconds. If a stage takes minutes, wouldn’t potentially saving 5 minutes be worth investing more than 2 seconds for optimization? In general, where is the point of diminishing returns when investing more into optimization barely improves the objectives?

We set the solving time to be on the same scale (sub-seconds to seconds) as the VM/Container startup time in the cloud as pointed out by [1, 32, 50]. This is also consistent with our prior experience in solving the resource optimization problem for AliCloud’s serverless computing [32].
That said, we agree with the reviewer on the comment. To better balance the effectiveness of optimization and fast solving time, we introduced two techniques in our revision:

1. We improved our compile-time optimization with multi-query plan search, as described in Section 5.1.3. This technique is shown to provide additional performance improvements for several long-running queries with complex query structures. As we use parallel computing for this extension, it does not add to the increase in wall clock time.

2. To balance query execution time and solving time, our revision further introduced a (simple) runtime adaptive scheme to adjust the sampling budget based on the predicted latency. This technique is introduced at the end of Section 5.1.1 and used in our new end-to-end experiments in Section 6.3. More specifically, we conducted hyperparameter tuning to derive low, medium, and high values for the sampling budget. Our profiling results show that if the predicted latency using the default configuration is over 10s, it is better to use the median sampling rate; otherwise, we use a low sampling rate. Increasing the sampling rate further does not improve the performance but adds the time cost. This is likely due to the model inaccuracies, which we plan to improve in future work.

R2.D3 Technical details: What type of embedding is used (Section 4.3)? The model architecture is also not clear. Does the choice of model have a major impact on prediction accuracy? If so, then the model discussion needs more details.

We thank the reviewer for the comment. As has been discussed in our previous work [32, 51], the choice of model is important, and the graph transformer network (GTN) [68] is shown to be an effective tool to embed query plans with graph structures to a numerical encoding. Besides our previous model analyses in [32], we added additional model discussions in Section 6.1, including Expt3 on the impact of skewness on our models and Expt4 on the model generalizability to unseen workloads.

R2.D4 Both random sampling and grid search for parameter value candidates suffer from the curse of dimensionality. Give some more intuition how the crossover approach could address it.

We significantly extended our paper, including new techniques and new writing, to address the dimensionality of the parameter space.

1. Feature importance score (FIS) based parameter filtering: We conducted additional experiments to compute the feature importance scores (FIS) of the 19 parameters in our trained latency and data shuffle models. The FIS of a parameter here is defined to be the increase of the model error if the values of this parameter are randomly shuffled in the training set. Due to space limitations, we include one example in Figure R1 in this response document, but leave other plots to Figure 21-23 in Appendix B.3. Our main observation is that the FIS values of the parameters, sorted in decreasing order, follow a long tail distribution. The parameters at the tail, with a cumulative FIS less than 5% in WMAPE, do not affect the model performance by much and hence can be filtered out. For instance, the threshold of 5% cumulative FIS means an increment of at most 0.7% in WMAPE for TPC-H and 1.2% for TPC-DS in the subQ model.

2. Sampling with FIS-based parameter filtering: Then we introduced three sampling methods that could leverage FIS-based selection: (a) random sampling with or without parameter filtering; (b) Latin hypercube sampling with or without parameter filtering; (c) adaptive grid search with inherent parameter filtering, as we described in the new paragraph starting with “Sampling methods” on Page 8, which is the last paragraph of Section 5.1.1.

3. New experimental results: We added new results in Figure 11(b)-(c) to show performance metrics (hypervolume and solving time) of the above three sampling methods with parameter filtering. Among them, we report the adaptive grid search to be an overall winner, achieving the highest hypervolume among all FIS-based methods while keeping the solving time under 2s for most queries tested.

We further clarify that the crossover approach is used in the step to enrich the $\theta_c$ candidates in order to address the missing Pareto solutions. Crossover is a genetic operation that tends to inherit promising regions of parameters from the promising candidates, and generate new candidates to potentially improve the performance, without searching the high-dimensional space. Our profiling results show that this technique helps find more global Pareto optimal solutions, especially, in the middle region of the Pareto front. More analyses are in Appendix B.8.

R2.D5 Proposition 5.1 seems obvious, if subqueries are treated as independent. Does it still hold when taking their interactions into account (W2, W3)?

We clarify that Proposition 5.1 is for compile-time optimization, where indeed we made simplifying assumptions to enable efficient algorithms. In particular, it assumes that for a specific $\theta_c$, the subquery structure is fixed and $\theta_p$ can be set for different subqueries independently. The rationale of these assumptions is to recommend a good $\theta_c$ by considering the correlation with $\theta_p$ while achieving high computational efficiency.

In the revision, we introduced a compile-time technique that samples multiple physical plans for each query, as described in Section 5.1.3. Then our runtime optimization further relaxes the assumptions by continuing to adjust the parameters and physical plans as runtime statistics are collected from the completed stages.

R2.D6 The discussion how some initial choice of join algorithm limits options for later reoptimization (Section 5.2) needs to be broadened. Does it only apply to this specific scenario? Or are there other limitations like it in Spark?
To respond to the reviewer’s question, we added concrete examples in Section 5.2 to illustrate why setting join threshold values either too high or too low will lead to suboptimal query plans. It does only apply to this specific scenario due to design decisions made by Spark committers, and so far we have not found other similar limitations in Spark.

**REVIEWER 4**

All of our references to the appendices refer to our tech report online: https://chenghao.pages.dev/papers/vldb24-lyu-tr.pdf [31]

**R4.D1:** In the experiments, the paper says: “To collect training data, we further treat these queries as templates to generate 50k distinct parameterized queries for TPC-H and TPC-DS, respectively”. Thus, it seems that queries are following the same patterns as they use the same template. However, this may not be true in practice where new query workloads may be different from the training workloads. Thus, I would like to see how the proposed approach work on different workloads with different query patterns.

We thank the reviewer for this suggestion. To investigate the generalizability of our learned models to unseen workloads, we conducted a new experiment, Expt 4 in Section 6.1. We investigated two approaches.

1. **Direct Application:** The first approach applies the latency model trained on TPC-DS directly to TPC-H, or vice versa. However, it leads to high errors due to significant differences in the running environments.

2. **Graph Embedding Transfer:** Our second approach transfers the graph embedding through the GTN model and retrains only the regressor for latency prediction, which is relatively fast. Our results indicate that graph embeddings trained on a workload with a broader range of query operators (e.g., TPC-DS) can be transferred effectively, resulting in a modest 0.06 WMAPE increase for the subQ model and a 0.06 and 0.16 WMAPE decrease for QS and LQP models, respectively.

Due to space constraints, we summarized the most significant results in the main paper. Detailed numbers of all our trained models can be found in Appendix B.7. Overall, the results of this new experiment improved our paper with positive evidence on the transfer learning ability of the GTN model, which is the most expensive part of model training.

**R4.D2:** It would be interesting to discuss whether or how to extend the proposed approach to other databases or data systems.

Our work has the potential to apply to other systems that can observe runtime statistics and have the runtime adaptability. These systems include Presto [53], Greenplum Database [33], Microsoft SQL Server [40], MaxCompute [36], etc. However, the exact runtime statistics may vary across these systems: a system with richer statistics allows precise models to be trained which in turn enables effective optimization, whereas a system with limited statistics will experience suboptimal results. Therefore, it will be an interesting study in our future work to extend our approach to some of these systems. We have included this discussion in the conclusions and future work section.
A Spark Optimizer for Adaptive, Fine-Grained Parameter Tuning

Chenghao Lyu† Qi Fan‡ Philippe Guyard‡ Yanlei Diao†‡
† University of Massachusetts, Amherst ‡Ecole Polytechnique
chenghao@cs.umass.edu, {qi.fan,philippe.guyard,yanlei.diao}@polytechnique.edu

ABSTRACT
As Spark becomes a common big data analytics platform, its growing complexity makes automatic tuning of numerous parameters critical for performance. Our work on Spark parameter tuning is particularly motivated by two recent trends: Spark’s Adaptive Query Execution (AQE) based on runtime statistics, and the increasingly popular Spark cloud deployments that make cost-performance reasoning crucial for the end user. This paper presents our design of a Spark optimizer that controls all tunable parameters of each query in the new AQE architecture to explore its performance benefits and, at the same time, casts the tuning problem in the theoretically sound multi-objective optimization (MOO) setting to better adapt to user cost-performance preferences. To this end, we propose a novel hybrid compile-time/runtime approach to multi-granularity tuning of diverse, correlated Spark parameters, as well as a suite of modeling and optimization techniques to solve the tuning problem in the MOO setting while meeting the stringent time constraint of 1-2 seconds for cloud use. Evaluation results using TPC-H and TPC-DS benchmarks demonstrate the superior performance of our approach: (i) When prioritizing latency, it achieves 63% and 65% reduction for TPC-H and TPC-DS, respectively, under an average solving time of 0.7-0.8 sec, outperforming the most competitive MOO method that reduces only 18-25% latency with 2.6-15s solving time. (ii) When shifting preferences between latency and cost, our approach dominates the solutions of alternative methods by a wide margin, exhibiting superior adaptability to varying preferences.

1 INTRODUCTION
Big data query processing has become an integral part of enterprise businesses and many platforms have been developed for this purpose [3, 6, 7, 12, 15, 36, 42, 55, 60, 67, 69, 74, 75]. As these systems are becoming increasingly complex, parameter tuning of big data systems has recently attracted a lot of research attention [23, 25, 27, 28, 51, 66]. Take Apache Spark for example. It offers over 180 parameters for governing a mixed set of decisions, including resource allocation, the degree of parallelism, shuffling behaviors, and SQL-related decisions. Our work on parameter tuning of big data query systems is particularly motivated by two recent trends:

Adaptive Query Execution. Big data query processing systems have undergone architectural changes that distinguish them substantially from traditional DBMSs for the task of parameter tuning. A notable feature is that a SQL query is compiled into a physical plan composed of query stages and a query stage is the granularity of scheduling and execution. The stage-based query execution model enables the system to observe the precise statistics of the completed stages at runtime. Recently, Spark has taken a step further to introduce Adaptive Query Execution (AQE), which upon the completion of each query stage, considers runtime statistics and re-optimizes the logical query plan to a new physical plan using parametric rules. Spark, however, does not support parameter tuning itself and instead, executes AQE based on the default or pre-specified configuration of the parameters, hence suffering from suboptimal performance of AQE when the parameters are set to inappropriate values. On the other hand, recent work on Spark tuning [23, 25, 27, 28, 51, 66] has limited itself to the traditional setting that the parameters are set at query submission time and then fixed throughout query execution, hence missing the opportunity of exploring AQE to improve the physical query plan.

Cost-performance reasoning in cloud deployment. As big data query processing is increasingly deployed in the cloud, parameter tuning in the form of cost-performance optimization [32, 44] has become more critical than ever to end users. Prior work [25, 71, 78] has used fixed weights to combine multiple objectives into a single objective (SO) and solve the SO problem to return one solution. However, the optimization community has established theory [35] pointing out that solving such a SO problem is unlikely to return a solution that balances the cost-performance as the specified weights intend to express (as we will demonstrate in this work). The theoretically sound approach to adapting between cost and performance is to treat it as a multi-objective optimization (MOO) problem [9, 35, 38, 39], compute the Pareto optimal set, and return one solution from the set that best matches the user preference as reflected by the weights set on the objectives [32, 51]. Therefore, our work in this paper aims to design a Spark optimizer that controls all tunable parameters (collectively called a “configuration”) of each Spark application in the new architecture of adaptive query execution to explore its performance benefits and, at the same time, casts the tuning problem in the multi-objective optimization setting to better adapt to user cost-performance needs. This Optimizer for Parameter Tuning (OPT) complements Spark’s cost and rule-based optimization of query plans, where the optimization rules use default or pre-specified values of Spark parameters. Our OPT can be implemented as a plugin in the current Spark optimizer framework and runs each time a query is submitted for execution.

Designing the optimizer for parameter tuning as defined above faces a few salient challenges:

Complex control of a mixed parameter space. One may wonder whether parameter tuning can be conducted solely at runtime, as an augmented AQE process. Unfortunately, Spark parameter tuning is more complex than that due to the need to control a mixed parameter space. More specifically, Spark parameters can be divided into three categories (see Table 1 for examples): the context parameters, \(\theta_c\), initialize the Spark context by specifying (shared) resources to be allocated and later governing runtime behaviors such as shuffling; the query plan parameters, \(\theta_p\), govern the translation from the logical to physical query plan; and the query stage parameters, \(\theta_s\), govern the optimization of the query stages in the
physical plan. The $\theta_p$ and $\theta_s$ parameters are best tuned at runtime to benefit from precise statistics, but they are strongly correlated with the context parameters, $\theta_c$, which control shared resources and must be set at query submission time to initialize the Spark context. How to best tune these mixed parameters, correlated but under different controls in the query lifetime, is a nontrivial issue.

**Stringent MOO Solving time for cloud use.** The second challenge is solving the MOO problem over a large parameter space of Spark while obeying stringent time constraints for cloud use, e.g., under 1-2 seconds to avoid delays in starting a Spark application in serverless computing [1, 32, 50]. Prior work on MOO for Spark tuning [51] has reported the running time of the Evolutional (Evo) method [9] to be about 5 seconds for query-level control of 12 Spark parameters. In our work, when we allow the $\theta_p$ parameters to be tuned separately for different subqueries, the time cost of Evo goes up quickly, exceeding 60 seconds for some TPC-H queries, which is unacceptable for cloud use.

To address the above challenges, we propose a novel approach to multi-granularity tuning of mixed Spark parameters and a suite of modeling and optimization techniques to solve the tuning problem in the MOO setting efficiently and effectively. More specifically, our contributions include the following:

1. **A hybrid approach to OPT (Section 3):** Our OPT is designed for multi-granularity tuning of mixed parameter space: while the context parameters $\theta_c$ configure the Spark context at the query level, we tune the $\theta_p$ and $\theta_s$ parameters at the fine-grained subquery level and query stage level, respectively, to maximize performance gains. To cope with Spark’s different control mechanisms for these parameters, we introduce a new hybrid compile-time/runtime optimization approach to multi-granularity tuning: the compile-time optimization finds the optimal $\theta'_c$, by leveraging the correlation between $\theta_c$ and fine-grained $\{\theta_p\}$ and $\{\theta_s\}$, to construct an ideal Spark context for query execution. Then the runtime optimization adjusts fine-grained $\{\theta_p\}$ and $\{\theta_s\}$ based on the precise statistics of the completed stages. Both compile-time and runtime optimization are cast in the setting of multi-objective optimization.

2. **Modeling (Section 4):** Solving the MOO problem for parameter tuning requires precise models for the objective functions used. Our hybrid approach requires accurate models for both compile-time and runtime optimization, which use different representations of query plans. The Spark execution environment shares resources among parallel stages, which further complicates the modeling problem. To address all of these issues, we introduce a modeling framework that combines a Graph Transformer Network (GTN) embedder of query plans and a regression model that captures the interplay of the tunable parameters (decision variables) and critical contextual factors (non-decision variables) such as query and data characteristics and resource contention. We devise new techniques that construct compile-time and runtime models in this framework.

3. **MOO Algorithms (Section 5):** Solving the MOO problem for multi-granularity tuning needs to conquer the high-dimensionality of the parameter space while obeying the time constraint, which is especially the case at compile-time when we consider the correlation of all the parameters together. We introduce a novel approach for compile-time optimization, named Hierarchical MOO with Constraints (HMOOC): it breaks the optimization problem of a large parameter space into a set of smaller problems, one for each subquery, but subject to the constraint that all subquery-level problems use the same Spark context parameters, $\theta_c$. Since these subproblems are not independent, we devise a host of techniques to prepare a sufficiently large set of candidate solutions for the subproblems and efficiently aggregate them to build global Pareto optimal solutions. Then our runtime optimization runs as part of AQE to adapt $\theta_p$ and $\theta_s$ effectively based on precise statistics.

Evaluation results using TPC-H and TPC-DS benchmarks demonstrate the superior performance of our techniques. (1) Our compile-time MOO algorithm for fine-grained parameter tuning outperforms existing MOO methods with 4.7%-54.1% improvement in hypervolume (the dominated space by the Pareto front) and 81%-98.3% reduction in solving time. (2) Our compile-time/runtime optimization, when asked to prioritize latency, achieves 63% and 65% reduction for TPC-H and TPC-DS, respectively, under an average solving time of 0.7-0.8 sec, whereas the most competitive MOO method reduces only 18-25% latency with high solving time of 2.6-15 sec. When shifting preferences between latency and cost, our approach dominates the solutions of alternative methods by a wide margin, exhibiting superior adaptability to varying preferences.

**2 RELATED WORK**

**DBMS tuning.** Our problem is related to a body of work on performance tuning for DBMSs. Most DBMS tuning systems employ an offline, iterative tuning session for each workload [59, 62, 71, 72], which can take long to run (e.g., 15-45 minutes [59, 71]). OtterTune [59] builds a predictive model for each query by leveraging similarities to past queries, and runs Gaussian Process (GP) exploration to try other configurations to reduce query latency. ResTune [72] accelerates the GP tuning process (with cubic complexity in the number of training samples) by building a meta-learning model to weigh appropriately the base learners trained for individual tuning tasks. CDBTune [71] and QTune [25] use Deep Reinforcement Learning (RL) to predict the reward of a configuration, as the weighted sum of different objectives, and explores new configurations to optimize the reward. These methods can take many iterations to achieve good performance [66], UDO [62] is an offline RL-based tuner for both database physical design choices and parameter settings. OnlineTuner [73] tunes workloads in the online setting by exploring a contextual GP to adapt to changing contexts and safe exploration strategies. Our work on parameter tuning aims to be part of the Spark optimizer, invoked on-demand for each arriving query, hence different from all the tuning systems that require launching a separate tuning session for each target workload.

**Tuning of big data systems.** Among search-based methods, Best-Config [78] searches for good configurations by dividing high-dimensional configuration space into subspaces based on samples, but it cold-starts each tuning request. ClassyTune [77] solves the optimization problem by classification, and Li et al. [26] prunes searching space with a running environment independent cost model, both of which cannot be easily extended to the MOO setting. A new line of work considered Spark parameter tuning for recurring workloads. ReM [23] tunes memory management decisions online by guiding the GP approach using manually-derived memory models. Locat [66] is a data-aware GP approach for tuning
Spark queries that repeatedly run with the input data size changing over time. While it outperforms prior solutions such as Tuneful [11], ReM [23], and QTune [25] in efficiency, it still needs hours to complete. Li et al. [27] further tune periodic Spark jobs using a GP with safe regions and meta-learning from history. LITE[28] tunes parameters of non-SQL Spark applications and relies on stage code analysis to derive predictive models, which is impractical as cloud providers usually have no access to application code under privacy constraints. These solutions do not suit our problem as we cannot afford to launch a separate tuning session for each query or target workload, and these methods lack support of adaptive runtime optimization and are limited to single-objective optimization.

**Resource optimization in big data systems.** In cluster computing, a resource optimizer (RO) determines the optimal resource configuration on demand and with low latency as jobs are submitted. Morpheus [18] codifies user expectations as multiple Service-Level Objectives (SLOs) and enforces them using scheduling methods. However, its optimization focuses on system utilization and predictability, but not cost and latency of Spark queries. PerOrator [49] optimizes latency via an exhaustive search of the solution space while calling its model for predicting the performance of each solution. WiseDB [34] manages cloud resources based on a decision tree trained on minimum-cost schedules of sample workloads. ReLocag[16] presents a predictor to find the near-optimal number of CPU cores to minimize job completion time. Recent work [24] proposes a heuristic-based model to recommend a cloud instance that achieves cost optimality for OLAP queries. This line of work addresses a smaller set of tunable parameters (only for resource allocation) than the general problem of Spark tuning with large parameter space, and is limited to single-objective optimization.

**Multi-objective optimization** (MOO) computes a set of solutions that are not dominated by any other configuration in all objectives, aka, the Pareto-optimal set (or Pareto front). Theoretical MOO solutions suffer from various performance issues in cloud optimization: Weighted Sum [35] is known to have poor coverage of the Pareto front [38]. Normalized Constraints [39] lacks in efficiency due to repeated recomputation to return more solutions. Evolutionary methods [9] approximately compute a Pareto set but suffer from inconsistent solutions. Multi-objective Bayesian Optimization [5, 14] extends the Bayesian approach to modeling an unknown function with an acquisition function for choosing the next point(s) that are likely to be Pareto optimal. But it is shown to take long to run [51] and hence lacks the efficiency required by a cloud optimizer.

In the DB literature, MOO for SQL queries [17, 20, 56–58] finds Pareto-optimal query plans by efficiently searching through a large set of plans. The problem, essentially a combinatorial one, differs from MOO for parameter tuning, which is a numerical optimization problem. TEMPO [54] considers multiple SLOs of SQL queries and guarantees max-min fairness when they cannot be all met. MOO for workflow scheduling [20] assigns operators to containers to minimize total running time and money cost, but is limited to searching through 20 possible containers and solving a constrained optimization for each option.

The closest work to ours is UDAO [51, 70] that tunes Spark configurations to optimize for multiple objectives. It Progressive Frontier (PF) method [51] provides the MOO solution for spark parameter tuning with good coverage, efficiency, and consistency. However, the solution is limited to coarse-grained query-level control of parameters. Lyu et al. extended the MOO solution to serverless computing [32] by controlling machine placement and resource allocation to parallel tasks of each query stage. However, its solution only guarantees Pareto optimality for each individual stage, but not the entire query (with potentially many stages).

### 3 PROBLEM STATEMENT AND OVERVIEW

In this section, we provide background on Spark including its adaptive query execution extension and present initial results illustrating the benefits and complexity of fine-grained tuning. We then formally define our Spark parameter tuning problem and provide an overview of our compile-time/runtime optimization approach.

#### 3.1 Background on Spark

Apache Spark [69] is an open-source distributed computing system for large-scale data processing and analytics. The core concepts of Spark include jobs, representing computations initiated by actions, and stages, which are organized based on shuffle dependencies, serving as boundaries that partition the computation graph of a job. Stages comprise sets of tasks executed in parallel, each processing a specific data partition. Executors, acting as worker processes, execute these tasks on individual cluster nodes.

Spark SQL seamlessly integrates relational processing into the Spark framework [2]. A submitted SQL query undergoes parsing, analysis, and optimization to form a logical query plan (LQP). In subsequent physical planning, Spark transforms the LQP to one or more physical query plans (PQP), using physical operators provided by the Spark execution engine. Then it selects one PQP using a cost model, which mostly applies to join algorithms. The physical planner also performs rule-based optimizations, such as pipelining projections or filters into one map operation. The PQP is then divded into a directed acyclic graph (DAG) of query stages (QSs) based on data exchange dependencies such as shuffling or broadcasting. These query stages are then executed in a topological order.

The execution of a Spark SQL query is configured by three categories of parameters, as shown in Table 1, providing different controls in query lifetime. As Figure 1(a) shows, query plan parameters \( \theta_p \) guide the translation from a logical query plan to a physical query plan, influencing the decisions such as the bucket
size for file reading and the join algorithms to use via Spark’s parametric optimization rules. Figure 1(b) shows a concrete example of translating a LQP to PQP, where each logical operator is instantiated by specific algorithms (e.g., the first join is implemented by sorting both input relations and then a merge join of them), additional exchange operators are injected to realize data exchanges, and query stages are identified at the boundaries of exchange operators. Further, query stage parameters \( \theta_s \) control the optimization of a query stage via parametric rules, such as rebalancing data partitions. Finally, context parameters \( \theta_c \), specified on the Spark context, control shared resources, shuffle behaviors, and memory management throughout query execution. While they are in effect only at runtime, \( \theta_c \) must be specified at the query submission time when the Spark context is initialized.

Adaptive Query Execution (AQE). Cardinality estimation [13, 29, 30, 45, 48, 52, 61, 63–65, 76] has been a long-standing issue that impacts the effectiveness of the physical query plan. To address this issue, Spark introduced Adaptive Query Execution (AQE) that enables runtime optimization based on precise statistics collected from completed stages [10]. Figure 2 shows the life cycle of a SQL query with the AQE mechanism turned on. At compile time, a query is transformed to a LQP and then a PQP through query optimization (step 3). Query stages (QSs) that have their dependencies cleared are then submitted for execution. During query runtime, Spark iteratively updates LQP by collapsing completed QSs into dummy operators with observed cardinalities, leading to a so-called collapsed query plan LQP (step 5) and re-optimizes the LQP (step 7) and the QSs (step 10), until all QSs are completed. At the core of AQE are runtime optimization rules. Each rule internally traverses the query operators and takes effect on them. These rules are categorized as parametric and non-parametric, and each parametric rule is configured by a subset of \( \theta_p \) or \( \theta_s \) parameters. The details of those rules are left to Appendix B.1.2.

3.3 Our Parameter Tuning Approach

We next introduce our approach that supports multi-granularity parameter tuning using hybrid compile-time/runtime optimization and formally define the optimization problem in the MOO setting.

3.3.1 Hybrid, Multi-Granularity Tuning. The goal of our work is to find, for each Spark query, the optimal configuration of all the \( \theta_p \), \( \theta_s \), and \( \theta_c \) parameters under multi-granularity tuning. While the context parameters \( \theta_c \) configure the Spark context at the query level, we tune other parameters at fine granularity to maximize performance gains, including setting the query plan parameters \( \theta_p \) distinctly for each collapsed query plan and the query stage parameters \( \theta_s \) for each query stage in the physical plan.

To address the correlation between the context parameters \( \theta_c \) (set at query submission time) and \( \theta_p \) and \( \theta_s \) parameters (best tuned at runtime), we introduce a hybrid compile-time / runtime optimization approach, as depicted by the two red boxes in Figure 2.

Compile-time: Our goal is to (approximately) derive the optimal \( \theta^*_p \) by leveraging the correlation of all the parameters, to construct an ideal Spark context for query execution. Our compile-time optimization uses cardinality estimates by Spark’s cost optimizer.

Runtime: With Spark context fixed, our runtime optimization runs as a plugin of AQE, invoked each time the collapsed query plan (LQP) is updated from a completed query stage, and adjusts \( \theta_p \) for the collapsed query plan based on the latest runtime statistics.
Then AQE applies $\theta_p$ to its parametric rules to generate an updated physical query plan (PQP). For the query stages in this new physical plan, our runtime optimization kicks in to optimize $\theta_i$ parameters based on precise statistics. Then AQE applies parametric rules with the tuned $\theta_i$ to optimize data partitions of these stages.

3.3.2 Multi-Objective Optimization. Targeting cloud use, our optimization problem concerns multiple user objectives such as query latency and cloud cost in terms of CPU hours or a weighted combination of CPU, memory, and IO resources.

Prior work [25, 71, 78] used fixed weights to combine multiple objectives into a single objective (SO) and solved it to return one solution, denoted as the SO-FW method. It is a special case of a classical MOO algorithm, weighted sum (WS) [35], that repeatedly applies $n$ weight vectors to create a set of SO problems and returns a solution for each, denoted as the MO-WS method (i.e., $n = 1$ in SO-FW). It is known from the theory of WS that trying different weights to create SO problems is unlikely to return points that evenly cover the Pareto front unless the objective functions have a very peculiar shape [35]. Empirically, MO-WS has been reported with sparse coverage of the Pareto front in prior work [51]. Figure 4 illustrates this for TPCH-Q2 in the 2D space of query latency and cloud cost: 11 SO problems generated from evenly spaced weight vectors return only two distinct solutions (marked by the blue dots), where 10 of them collide to the same bottom point. Increasing to 101 weight vectors still returns only 3 distinct points. Such sparse coverage of the Pareto front leads to poor adaptability when the user shifts preference (e.g., from favoring latency to favoring cost) because there are not enough points on the Pareto front to capture the tradeoffs between the objectives.

For this reason, our work casts the optimization problem in multi-objective optimization (MOO) framework [9, 35, 38, 39], which computes the Pareto front properly to capture the tradeoffs and later allows us to recommend one that best matches the user preference.

Formally, a MOO problem aims to minimize multiple objectives simultaneously, where the objectives are represented as functions $f = (f_1, ..., f_k)$ on all the tunable parameters $\theta$.

Definition 3.1. Multi-Objective Optimization (MOO).

$$\arg \min_{\theta} f(\theta) = \{ f_1(\theta), f_2(\theta), \ldots, f_k(\theta) \}$$

$s.t.$

$$\begin{cases} f(\theta) \in \Phi \subseteq \mathbb{R}^k \, \land \, L_i \leq f_i(\theta) \leq U_i, \quad i = 1, \ldots, k \end{cases}$$

where $\theta$ is the configuration with $d$ parameters, $\Sigma \subseteq \mathbb{R}^d$ denotes all possible configurations, and $\Phi \subseteq \mathbb{R}^k$ denotes the objective space. If an objective favors larger values, we add the minus sign to the objective function to transform it into a minimization problem.

Definition 3.2. Pareto Optimal Set. In the objective space $\Phi \subseteq \mathbb{R}^k$, a point $F'$ Pareto-dominates another point $F''$ if $\forall i \in [1,k], f'_i \leq f''_i$ and $\exists j \in [1,k], f'_j < f''_j$. For a given query, solving the MOO problem leads to a Pareto set (Front) $\mathcal{F}$ that includes all the Pareto optimal solutions $(F, \theta)$, where $F$ is a Pareto point in the objective space $\Phi$ and $\theta$ is its corresponding configuration in $\Sigma$.

Figure 4 shows a Pareto front for TPCH-Q2. Most configurations, depicted by the grey dots, are dominated by the Pareto optimal solutions, depicted by the red dots, in both objectives. Hence, the MOO solution allows us to skip the vast set of dominated configurations. Furthermore, the Pareto points themselves represent tradeoffs between the two competing objectives. The optimizer can recommend one of them based on the user preference, e.g., favoring latency to cost in peak hours with weights 0.9 to 0.1 and vice-versa in off-peak hours. The recommendation can be made based on the Weighted Utopia Nearest (WUN) distance [51] of the Pareto points from the Utopia point $U$, which is the hypothetical optimum in all objectives, marked by the orange dot in the figure. Figure 4 shows a few recommendations when we run WUN on the Pareto front with different weighted preferences on the objectives.

We next define the MOO problem for Spark parameter tuning.

Definition 3.3. Multi-Objective Optimization for Spark SQL.

$$\arg \min_{\theta, \{ \theta_p \}, \{ \theta_s \}} f(\theta, \{ \theta_p \}, \{ \theta_s \}) = \left[ \begin{array}{c} f_1(LQP, \theta_c, \{ \theta_p \}, \{ \theta_s \}, \alpha, \beta, \gamma) \\ \vdots \\ f_k(LQP, \theta_c, \{ \theta_p \}, \{ \theta_s \}, \alpha, \beta, \gamma) \end{array} \right]$$

$s.t.$

$$\begin{cases} \{ \theta_p \} = \{ \theta_{p1}, \theta_{p2}, \ldots, \theta_{pt}, \ldots \}, \forall \theta_{pt} \in \Sigma_p \\ \{ \theta_s \} = \{ \theta_{s1}, \theta_{s2}, \ldots, \theta_{st}, \ldots \}, \forall \theta_{st} \in \Sigma_s \end{cases}$$

where $LQP$ denotes the logical query plan with operator cardinality estimates, and $\theta_c, \{ \theta_p \}, \{ \theta_s \}$ represent the decision variables configuring Spark context, LQP transformations, and query stage (QS) optimizations, respectively. More specifically, $\{ \theta_p \}$ is the collection of all LQP parameters, and $\theta_{pt}$ is a copy of $\theta_p$ for the $t$-th transformation of the collapsed query plan $LQP$. Similarly, $\{ \theta_s \}$ is the collection of QS parameters, and $\theta_{st}$ is a copy for optimizing query stage $i$. $\Sigma_c, \Sigma_p, \Sigma_s$ are the feasible space for $\theta_c, \theta_p$ and $\theta_s$, respectively. Finally, $\alpha, \beta, \gamma$ are the non-decision variables (not tunable, but crucial factors that affect model performance), representing the input characteristics, the distribution of partition sizes for data exchange, and resource contention status during runtime.

3.3.3 Comparison to Existing Approaches. We finally summarize our work in relation to existing solutions to Spark parameter tuning in terms of the coverage of the mixed parameter space, adaptive runtime optimization, multi-granularity tuning, and multi-objective
We first devise models for fine-grained parameter tuning at both when the logical plan is translated to a physical plan. In other words, we introduce the notion of TPCH-Q3, which can be divided into five subQs, each corresponding to different subQ structures. We collect all of these subQ structures, and develop a predictive model for each subQ to enable fine-grained tuning of \( \theta_p \) and \( \theta_s \) at compile-time.

4.2 Modeling Objectives

With the objective of optimizing latency and cost, our modeling work seeks to make these metrics more robust and predictable.

Table 2: Comparison of Spark parameter tuning methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLocag [16]</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>BestConfig [78]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>ClassyTune [27]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>LITE [28]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>LOCAT [66]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Li et al. [27]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>UDAO [51]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Query latency in Spark, defined as the end-to-end duration to execute a query, benefits from Spark’s cluster manager that ensures a dedicated allocation of cores and memory to the entire query. Such resource isolation enhances the predictability of latency, making it a suitable target for optimizing a query or a collapsed query plan.

However, within a query, Spark shares resources among parallel query stages, raising two issues in modeling latency at the stage level. First, the end-to-end latency of a set of parallel stages often leads to a longer latency than their maximum due to resource contention. Prior work [32] assumed ample resources in industry-scale clusters and simplified this issue by taking the max latency among parallel tasks, which is not applicable in the Spark environment of shared resources. Second, predicting the latency of each stage directly is very hard due to its variability in a shared-resource setting, where performance fluctuates based on resource contention.

To address these issues, we propose the concept of analytical latency, calculated as the sum of the task latencies across all data partitions divided by the total number of cores. This approach yields two significant advantages. Firstly, it establishes a direct link between the latency of a query and its constituent stages, enabling the computation of query-level latency at compile time through a sum aggregator over the task latencies of all subQs. Secondly, it enhances the predictability of query stage latency by excluding the variability introduced by resource wait times, thus offering a more consistent basis for latency prediction. To validate the efficacy of analytical latency, we conducted a comparison with actual latencies at the query level using the TPC-H and TPC-DS benchmarks under the default Spark configuration. The results demonstrate a robust correlation between analytical and actual latencies, with Pearson correlation coefficients of 97.2% for TPC-H and 87.6% for TPC-DS. Furthermore, Figure 5 shows that the analytical latency closely mirrors the actual execution time, with the ratio close to 1 for most of the tested queries.

Cloud costs are mainly based on the consumption of resources, such as CPU-hour, memory-hour, I/O and shuffle sizes [4]. We model all of these costs to support multi-objective optimization.

To summarize, our models capture 1) end-to-end latency and cost for collapsed query plans (LQP) at runtime, 2) analytical latency and cost for query stages (QS) at runtime, in the face of resource sharing, and 3) analytical latency and cost for subQs at compile time—the latter two ensure both to be robust targets for modeling.

4.3 Model Formulation for Optimization

We now introduce the methodology for building models for subQ, LQP, and QS which will enable their respective fine-graining later.

Feature Extraction. We extract features to capture the characteristics of queries and the dynamics of their execution environment,
configured by decision variables and non-decision variables. First, we extract the query plan as a DAG of vectors, where each query operator is characterized through a composite encoding that integrates i) the operator type via one-hot encoding, ii) its cardinality, represented by row count and size in bytes, and iii) an average of the word embeddings [41] computed from its predicates, providing a rich, multidimensional representation of the operator’s functional and data characteristics. Second, we capture critical contextual factors as non-decision variables, including i) input characteristics $a$, aggregated from the statistics of leaf operators, ii) data distribution $\beta$, quantifying the size distribution of input partitions with metrics like standard deviation-to-average ratio ($\frac{s}{\mu}$), skewness ratio ($\frac{\mu - \mu}{\mu}$), and range-to-average ratio ($\frac{\max - \min}{\mu}$), and iii) runtime contention $y$, capturing the statistics of parallel stages in a numeric vector, tracking their tasks in running and waiting states, and aggregating statistics of completed tasks to characterize their behaviors.

Third, we convert the tunable parameters as decision variables into a numeric vector to represent the Spark behavior.

**Model Structures.** The hybrid data structure of the query plan, with a DAG of operator encodings, and other tabular features, poses a challenge in model formulation. To tackle this, we adopt a multi-channel input framework [32] that incorporates a Graph Transformer Network (GTN) [8] and a regressor to predict our objectives, as shown in Figure 6. We first derive the query embedding using a GTN model [8], which can handle the non-linear and non-sequential relationships and employ attention mechanisms and Laplacian positional encoding to capture operator correlations as well as positional information. These encodings are then concatenated with other tabular data and processed through a regressor, allowing us to capture the interplay among the query characteristics, critical contextual factors, and tunable parameters. Figure 6 illustrates the architecture of the LQP model, which has the largest number of feature factors.

5 **COMPILE-TIME/RUNTIME OPTIMIZATION**

In this section, we present our hybrid compile-time/runtime optimization approach to multi-granularity parameter tuning in the multi-objective optimization setting.

5.1 **Hierarchical MOO with Constraints**

Our compile-time optimization finds the optimal configuration $\theta_c^*$ of the context parameters to construct an ideal Spark context for query execution. We do so by exploring the correlation of $\theta_c$ with fine-grained $\theta_p$ and $\theta_t$ parameters for different subqueries (subQs), under the modeling constraint that the cardinality estimates are based on Spark’s cost-based optimizer. Despite the modeling constraint, capturing the correlation between the mixed parameter space allows us to find a better Spark context for query execution.

The multi-objective optimization problem in Def. 3.3 provides fine-grained control of $\theta_p$ and $\theta_t$, at the subQ/query stage level, besides the query level control of $\theta_c$. This leads to a large parameter space, linear in the number of query stages in the plan, which defeats most existing MOO methods when the solving time must be kept under the constraint of 1-2 seconds for cloud use.

To combat the high-dimensional parameter space, we propose a new approach named Hierarchical MOO with Constraints (HMOOC).

It follows a divide-and-conquer framework to break a large optimization problem on $\theta_c$, $\theta_p$, $\theta_t$ to a set of smaller problems on $\theta_c$, $\theta_p$, $\theta_t$, one for subQ of the logical query plan. However, these smaller problems are not independent as they must obey the constraint that all the subproblems must share the same $\theta_c$ value. More specifically, the problem for HMOOC is defined as follows:

**Definition 5.1. Hierarchical MOO with Constraints (HMOOC)**

$$\arg \min_{\theta} f(\theta) = \begin{cases} f_1(\theta) = \Lambda(\phi_1(LQP_1, \theta_c, \theta_p, \theta_t)) & \ldots \\ \ldots & \ldots \\ f_k(\theta) = \Lambda(\phi_k(LQP_k, \theta_c, \theta_p, \theta_t)) & \ldots \\ \end{cases}$$

s.t. $\theta_c \in \Sigma_c \subseteq \mathbb{R}^d_c$, $\theta_p \in \Sigma_p \subseteq \mathbb{R}^d_p$,

where $LQP_i$ denotes the $i$-th subQ of the logical plan query, $\theta_i = (\theta_c, \theta_p, \theta_t)$ denotes its configuration, with $i = 1, \ldots, m$, and $m$ is the number of subQs. Most notably, all the subQs share the same $\theta_c$, but can use different values of $\theta_p$ and $\theta_t$. Additionally, $\phi_j$ is the subQ predictive model of the $j$-th objective, where $j = 1, \ldots, k$.

The function $\Lambda$ is the mapping from subQ-level objective values to query-level objective values, which can be aggregated using sum based on our choice of analytical latency and cost metrics.

The main idea behind our approach is to tune each subQ independently under the constraint that $\theta_c$ is identical among all subQ’s. By doing so, we aim to get the local subQ-level solutions, and then recover the query-level Pareto optimal solutions by composing these local solutions efficiently. In brief, it includes three sequential steps: (1) subQ tuning, (2) DAG aggregation, and (3) WUN recommendation.

Figure 7 illustrates an example of compile-time optimization for TPCH-Q3 under the latency and cost objectives. For simplicity, we show only the first three subQ’s in this query and omit $\theta_t$ in this example. In subQ-tuning, we obtain subQ-level solutions with only the Pareto optimal values of latency and cost are retained. Subsequently, in the DAG aggregation step, the query-level latency and cost are computed as the sum of the three subQ-level latency and cost values, and only the Pareto optimal values of latency and cost are retained.

Finally, in the third step, we use the WUN (weighted Utopia nearest) policy to recommend a configuration from the Pareto front.

5.1.1 **Subquery (subQ) Tuning.** Subquery (subQ) tuning aims to generate an effective set of local solutions of $(\theta_c, \theta_p, \theta_t)$ for each subQ while obeying the constraint that all the subQs share the same $\theta_c$. For simplicity, we focus on $(\theta_c, \theta_p)$ in the following discussion as $\theta_t$ is treated the same way as $\theta_p$.

One may wonder whether it is sufficient to generate only the local Pareto solutions of $(\theta_c, \theta_p)$ of each subQ. Unfortunately, this will lead to missed global Pareto optimal solutions due to the constraint on $\theta_c$. Figure 9 illustrates an example with 3 subQs, where solutions sharing the same index fall under the same $\theta_c$ configuration and have achieved optimal $\theta_p$ under that $\theta_c$ value. The first row in Figure 9 showcases subQ-level solutions, where triangle points represent subQ-level optima and circle points denote dominated solutions. The second row in Figure 9 displays the global
set the query values. Then, we enrich the \( \Theta \) computation complexity. It is based on the hypothesis that, within each subQ, only subQ-level values contribute to the query-level \( \Theta \) cost becomes better than solution 4 (a subQ-level Pareto optimal solution) and is a query-level Pareto point.

1. Effective \( \Theta_c \) Candidates. To minimize the chance of missing global solutions, we seek to construct a diverse, effective set of \( \Theta_c \) configurations to be considered across all subQs. \( \Theta_c \) can be initialized by random sampling or grid-search over its domain of values. Then, we enrich the \( \Theta_c \) set using a few methods. Drawing inspiration from the evolutionary algorithms [9], we introduce a crossover operation over the existing \( \Theta_c \) population to generate new candidates. If crossover cannot generate more candidates, e.g., for some grid search methods used for initial sampling of \( \Theta_c \) candidates, then we add random sampling to discover new candidates.

2. Optimal \( \Theta_p \) Approximation. Next, under each \( \Theta_c \) candidate, we show that it is crucial to keep track of the local Pareto optimal \( \Theta_p \) within each subQ. The following proposition explains why.

**Proposition 5.1.** Under any specific value \( \Theta_c^i \), only subQ-level Pareto optimal solutions \( \Theta_p^{i,j} \) contribute to the query-level Pareto optimal solutions.

In the interest of the space, all the proofs in this paper are deferred to Appendix A.1.

The above result allows us to restrict our search of \( \Theta_p \) to only the local Pareto optimal ones. However, given the large, diverse set of \( \Theta_c \) candidates, it is computationally expensive to solve the MOO problem for \( \Theta_p \) repeatedly, once for each \( \Theta_c \) candidate. We next introduce a clustering-based approximation to reduce the computation complexity. It is based on the hypothesis that, within the same subQ, similar \( \Theta_c \) candidates entail similar optimal \( \Theta_p \) values in the tuning process. By clustering similar \( \Theta_c \) values into a small number of groups (based on their Euclidean distance), we then solve the MOO problem of \( \Theta_p \) for a single \( \Theta_c \) representative of each group. We then use the optimized \( \Theta_p \) as the estimated optimal solution for other \( \Theta_c \) candidates within each group.

**Algorithm.** Algorithm 1 describes the steps for obtaining an effective solution set of \( \{ \Theta_c, \Theta_p \} \) for each subQ. Line 1 initiates by generating \( C \times P \) samples, where \( C \) and \( P \) are the numbers of distinct values of \( \Theta_c \) and \( \Theta_p \), respectively. The \( \Theta_c \) candidates are then grouped using a clustering approach (Line 2), where \( \text{rep}_c \text{_list} \) constitutes the list of \( \Theta_c \) representatives for the \( n \) groups, \( \text{C}_\text{list} \) includes the members within all \( n \) groups, and \( k \) represents the clustering model. In Line 3, \( \Theta_p \) optimization is performed for each representative \( \Theta_c \) candidate (using the samples from Line 1). Subsequently, the optimal \( \Theta_p \) of the representative \( \Theta_c \) is assigned to all members within the same group and is fed to the predictive models to get objective values (Line 4). After that, the initial effective set is obtained, where \( \Omega^{(0)} \) represents the subQ-level objective values under different \( \Theta_c \), and \( \Theta^{(0)} \) represents the corresponding configurations. Line 5 further enriches \( \Theta_c \) using the crossover method or random sampling, which expands the initial effective set to generate new \( \Theta_c \) candidates. Afterwards, the cluster model \( \kappa \) assigns the new \( \Theta_c \) candidates with their group labels (Line 6). The previous optimal \( \Theta_p \) values are then assigned to the new members within the same group, resulting in their corresponding subQ-level values as the enriched set (Line 7). Finally, the initial set and the enriched set are combined as the final effective set of subQ tuning (Line 8).

**Sampling methods.** We next detail the sampling methods used in Line 1 of the algorithm. (1) We include basic sampling methods, including random sampling and Latin-hypercube sampling (LHS) [37] as a grid-search method. (2) To reduce dimensionality, we introduce feature importance score (FIS) based parameter filtering; we sort the parameters by the FIS value from the trained model and leverage the long-tail distribution to drop the parameters at the tail. Precisely, many parameters at the tail have a low cumulative FIS and we apply a threshold (e.g., 5% model loss) to remove them from sampling. (3) We further propose an adaptive grid search method with FIS-based parameter filtering. Given the sample budget \( (C \text{ or } P) \), it goes down the FIS ranking list and progressively covers one more parameter, including its min, median and max values. If it reaches the budget before adding all parameters, it ignores those
uncovered ones. Otherwise, it loops over the list again to add more sampled values of each parameter. (4) We conduct hyperparameter tuning to derive low, medium, and high values for C and P. We also employ a simple runtime adaptive scheme to adjust the sampling budget based on the predicted latency under the default configuration. See Appendix A.4.1 and B.3 for more details.

5.1.2 DAG Aggregation. DAG aggregation aims to recover query-level Pareto optimal solutions from subQ-level solutions. It is a combinatorial MOO problem, as each subQ must select a solution from its non-dominated solution set while satisfying the constraint of sharing the configuration among all subQs. The complexity of this process can be exponential in the number of subQs. Our approach below addresses this challenge by providing optimality guarantees and reducing the computation complexity.

Simplified DAG. A crucial observation that has enabled our efficient methods is that our optimization problem over a DAG structure can be simplified to an optimization problem over a list structure. This is due to our choice of analytical latency and cost metrics, where the query-level objective can be computed as the sum of subQ-level objectives. The MOO problem over a DAG can be simulated with a list structure for computing query-level objectives.

Algorithm 1: Effective Set Generation

| Require: Q, φi, n, ∀i ∈ [1,k], α, β, γ, C, P. |
| 1: Θ(0), Θp(0) = sampling(C, P) |
| 2: rep_c_list, C_list, k = cluster(Θ(0), n) |
| 3: Gp = optimize_p_moo(Θp(0), rep_c_list, φ, α, β, γ, Q) |
| 4: Θ(0), G0(0) = assign_opt_p(C_list, rep_c_list, Gp, φ, α, β, γ, Q) |
| 5: Θ(0) = enrich_c(Θ(0), G(0)) |
| 6: C_list(new) = assign_cluster(Θc(new)), rep_c_list, k |
| 7: Θ(new), Gnew = assign_opt_p(C_list(new), rep_c_list, Gp, φ, α, β, γ, Q) |
| 8: Θ, Θ = union(Θ(0), G(0), G(new), G(new)) |
| 9: return Θ, Θ |

HMOOC1: Divide-and-Conquer. Under a fixed Θp, i.e., satisfying the constraint inherently, we propose a divide-and-conquer method to compute the Pareto set of the simplified DAG, which is reduced to a list of subQs. The idea is to (repeatedly) partition the list into two halves, solve their respective subproblems, and merge their solutions to global optimal ones. The merge operation enumerates all the combinations of solutions of the two subproblems, sums up their objective values, and retains only the Pareto optimal ones. Our proof (available in Appendix A.1) shows that this method returns a full set of query-level Pareto optimal solutions.

HMOOC2: WS-based Approximation. Our second technique approximates the MOO solution over a list structure. For each fixed Θp, we apply the weighted sum (WS) method to generate evenly spaced weight vectors. Then for each weight vector, we obtain the (single) optimal solution for each subQ and sum the solutions of subQ’s to get the query-level optimal solution. It can be proved that this WS method over a list of subQs guarantees to return a subset of query-level Pareto solutions (see Appendix A.1).

HMOOC3: Boundary-based Approximation. Our next approximate technique stems from the idea that the objective space of DAG aggregation under each Θp can be approximated by k extreme points, where k is the number of objectives. In our context, the extreme point under a fixed Θp is the Pareto optimal point with the best query-level value for any objective. The rationale behind this approximation lies in the observation that solutions from different Θp candidates correspond to distinct regions on the query-level Pareto front. This arises from the fact that each Θp candidate determines the total resources allocated to the query, and a diverse set of Θp candidates ensures good coverage across these resources. Varying total resources, in turn, lead to different objectives of query performance, hence resulting in good coverage of the Pareto front of cost-performance tradeoffs.

Therefore, we consider the degenerated extreme points to symbolize the boundaries of different (resource) regions within the query-level Pareto front. Figure 8 illustrates an example. Here, the dashed rectangles with their extreme points under different colors represent the objective space of query-level solutions under various Θp candidates. The brown dashed line represents the approximate query-level Pareto front derived by filtering the dominated solutions from the collection of extreme points. The star solution indicates a missed query-level Pareto solution, as it cannot be captured from the extreme points.

The algorithm works as follows. For each Θp candidate, for each objective, we select the subQ-level solution with the best value for that objective for each subQ, and then sum up the objective values of such solutions from all subQs to form one query-level extreme point. Repeating this procedure will lead to a maximum of kn query-level solutions, where k is the number of objectives and n is the number of Θp candidates. An additional filtering step will retain the non-dominated solutions from the kn candidates, using an existing method of complexity O(kn log(kn)) [22].

Proposition 5.2. Under a fixed Θp candidate, the query-level objective space of Pareto optimal solutions is bounded by its extreme points in a 2D objective space.

Proposition 5.3. Given subQ-level solutions, our boundary approximation method guarantees to include at least k query-level Pareto optimal solutions for a MOO problem with k objectives.

5.1.3 Multiple Query Plan Search. Our compile-time MOO algorithm so far has considered only one physical query plan (with the corresponding subQs) based on the default configuration. Since at runtime, AQE may generate a very different physical plan, we further enhance our compile-time optimization by considering multiple physical plans. Initially, we sample plan-related parameters (s1 and s2) in Θp to collect different physical query plans. We then rank these plans based on their predicted query latency under the default configuration. Subsequently, we trigger compile-time optimization (the HMOOC algorithm) for each of the top-k fastest query plans, run them in parallel, and merge their Pareto solutions into the same objective space to obtain the final Pareto set.

5.2 Runtime Optimization.

The compile-time optimization relied on the estimated cardinality and assumption of uniform data distributions. However, it true value
is to recommend the optimal context parameters $\theta^c$ by considering the correlations with $\theta_p$ and $\theta_s$. Then, our runtime optimization addresses the remaining problems, adapting $\theta_p$ and $\theta_s$ based on actual runtime statistics and plan structures.

To start the runtime process, we need to first suit the constraint that Spark accepts only one copy of $\theta_p$ and $\theta_s$ at the query submission time. To do so, we intelligently aggregate the fine-grained $\theta_p$ and $\theta_s$ from compile-time optimization to initialize the runtime process. In particular, Spark AQE can convert a sort-merge join (SMJ) to a shuffled hash join (SHJ) or a broadcast hash join (BHJ), but not vice versa. Thus, imposing high thresholds ($s_1, s_4$ in Table 1) to force SHJ or BHJ based on inaccurate compile-time cardinality can result in suboptimal plans. In the example of Figure 3(b), when the cardinality of Join4 is underestimated at the compile time, MO-WS returns a query plan broadcasting the output of Join4. At runtime, when the actual output size of Join4 is observed to be 4.5GB, Spark cannot switch the BHJ back to other joins but broadcasting the 4.5GB data, leading to suboptimal performance. On the other hand, setting these thresholds to zeros initially might overlook opportunities for applying BHJs, especially for joins on base tables with small input sizes. To mitigate this, we initialize $\theta_p$ with the smallest threshold among all join-based subQs, enabling more effective runtime decisions. Other details of aggregating $\theta_p$ and $\theta_s$ are in Appendix A.3.

Runtime optimization then operates within a client-server model. The client, integrated with the Spark driver, dispatches optimization requests—including runtime statistics and plan structures—when a collapsed logical query plan (LQP) or a runtime query stage (QS) necessitates optimization (Steps 6, 9 in Figure 2). The server, hosted on a GPU-enabled node and supported by the learned models and a library of MOO algorithms [51], processes these requests over a high-speed network connection.

Complex queries can trigger numerous optimization requests every time when a collapsed logical plan or a runtime QS is produced, significantly impacting overall latency. For instance, TPCCDS queries, with up to 47 subQs, may generate up to nearly a hundred requests throughout a query’s lifecycle. To address this, we established rules to prune unnecessary requests based on the runtime semantics of parametric rules as detailed in Appendix A.3. By applying these rules, we substantially reduce the total number of optimization calls by 86% and 92% for TPCC-H and TPCC-DS respectively.

### 6 Experimental Evaluation

In this section, we evaluate our modeling and fine-grained compile-time/runtime optimization techniques. We further present an end-to-end evaluation against the SOTA tuning methods.

**Spark setup.** We perform SQL queries at two 6-node Spark 3.5.0 clusters with runtime optimization plugins. Our optimization focuses on 19 parameters, including 8 for $\theta_p$, 9 for $\theta_s$, and 2 for $\theta_c$, selected based on feature selection profiling [19] and best practices from the Spark documentation. More details are in Appendix B.1.

**Workloads.** We generate datasets from the TPCC-H and TPCC-DS benchmarks with a scale factor of 100. We use the default 22 TPCC-H and 102 TPCC-DS queries for the optimization analyses and end-to-end evaluation. To collect training data, we further treat these queries as templates to generate 50k distinct parameterized queries for TPCC-H and TPCC-DS, respectively. We run each query under one configuration sampled via Latin Hypercube Sampling [37].

#### Implementation

We implement our optimizer using plugins into Spark. (1) The trace collector is implemented as a customized Spark listeners to track runtime plan structures and statistics, costing an average of 0.02s overhead per query. (2) The compile-time optimizer operates as a standalone module, costing an average of 0.4s. (3) The runtime optimizer operates in the server-client model, where the client is wrapped in customized Spark query plan rules, and the server side is on a standalone machine, costing an average of 0.3-0.4s per query. (4) The model server maintains up-to-date models on the same server as the optimizer. The TPCC-H and TPCC-DS traces were split into 8:1:1 for training, validation, and testing. It took 6-12 hours to train one model and 2 weeks for hyperparameter tuning on a GPU node with 4 NVIDIA A100 cards.

#### 6.1 Model Evaluation

We trained separate models for subQ, QS, and LQP to support compile-time/runtime optimization. We evaluate each model using the weighted mean absolute percentage error (WMAPE), median and 90th percentile errors (P50 and P90), Pearson correlation (Corr), and inference throughput (Xput).

**Expt 1: Model Performance.** We present the performance of our best-tuned models for TPCC-H and TPCC-DS in Table 3. First, our models can provide highly accurate prediction in latency and analytical latency for Spark queries for different compile-time and runtime targets, achieving WMAPEs of 13-28%, P50 of 3-10%, and P90 of 29-65%, alongside a correlation range of 93-99% with the ground truth. Second, the shuffle size is more predictable than latency, evidenced by a WMAPE of 0.2-11% and almost perfect 99-100% correlation with the actual shuffle size, attributed to its consistent performance across configurations. Third, the models show high inference throughput, ranging from 60-462K queries per second, which enables efficient solving time of our compile-time and runtime optimizations. Overall, these results demonstrate the robust performance of our models in predicting cost performance metrics for Spark queries, while enabling efficient optimization.

**Expt 2: Comparison of Compile-Time and Runtime Results.** We then look into the performance differences between the runtime QS and its corresponding subQ at compile time. First, the latency performance in runtime QS is slightly inferior to its corresponding subQ at compile time. This disparity can be attributed to the runtime QS’s exposure to more varied and complex query graph structures, which complicates the prediction process. Second, the runtime QS consistently surpasses the subQ in shuffle prediction. This superior performance is linked to the direct correlation between shuffle and input size; the runtime QS benefits from access to actual input sizes.
Impact of Skewness.

Benefits over Default Configuration.

-14% / 44% -55% / 3%
-34% / 139% -59% / 55%
-1% / 25% -46% / -3%

Model Generalizability.

Comparison with SOTA MOO methods.

Table 4: Correlation between model errors and skewness ratios in intermediate RDDs.

<table>
<thead>
<tr>
<th>Lat. range</th>
<th>TPC-H</th>
<th>TPC-DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1s</td>
<td>0.001</td>
<td>-0.017</td>
</tr>
<tr>
<td>&gt;Lpq</td>
<td>0.019</td>
<td>-0.015</td>
</tr>
<tr>
<td>&gt;Lpq0.5</td>
<td>-0.026</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Table 5: Latency reduction with a strong speed preference

<table>
<thead>
<tr>
<th>Method</th>
<th>TPC-H</th>
<th>TPC-DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MO-WS HMOOC3</td>
<td>18%</td>
<td>18%</td>
</tr>
<tr>
<td>HMOOC3</td>
<td>61%</td>
<td>61%</td>
</tr>
<tr>
<td>HMOOC3*</td>
<td>63%</td>
<td>63%</td>
</tr>
<tr>
<td>Avg Solving Time</td>
<td>2.6</td>
<td>2.3</td>
</tr>
<tr>
<td>P50 Solving Time</td>
<td>0.41</td>
<td>0.38</td>
</tr>
<tr>
<td>Max Solving Time</td>
<td>0.81</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Figure 10: CDF of TPC-DS skewness

6.2 Compile-time MOO Methods

We now extend our compile-time MOO methods against existing MOO methods. The objectives include query latency and cloud cost at a weighted sum of cpu hours, memory hours, shuffle sizes.

Expt 5: DAG Aggregation methods.

Expt 6: Sampling Methods.

Table 6: HD latency and cost adapting to preferences

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Lat/Cost</th>
<th>SO-FW</th>
<th>SO-FW</th>
</tr>
</thead>
<tbody>
<tr>
<td>MO-WS</td>
<td>HMOOC3</td>
<td>HMOOC3*</td>
<td></td>
</tr>
<tr>
<td>HMOOC3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0, 1.0</td>
<td>20% / -11%</td>
<td>-15% / -10%</td>
<td>-6% / 64%</td>
</tr>
<tr>
<td>0.1, 0.9</td>
<td>1% / 1%</td>
<td>-1% / 25%</td>
<td>-6% / -5%</td>
</tr>
<tr>
<td>0.5, 0.5</td>
<td>-1% / 25%</td>
<td>-16% / -3%</td>
<td>-28% / 128%</td>
</tr>
<tr>
<td>0.9, 0.1</td>
<td>-13% / 27%</td>
<td>-51% / -9%</td>
<td>-34% / 139%</td>
</tr>
<tr>
<td>1.0, 0.0</td>
<td>-14% / 44%</td>
<td>-55% / -3%</td>
<td>-26% / 144%</td>
</tr>
</tbody>
</table>

report on HV and solving time, in the first 5 bars in each group. In terms of HV, H3-A is the best for TPC-H, and H3-L and H3-R slightly outperform the other three with parameter filtering for TPC-DS. Considering solving time, we see that H3-A achieves the lowest solving time, finishing all TPC-H queries in 1s and 100/102 TPC-DS queries in 2s. The methods without parameter filtering can have the solving time exceeding 6 seconds for some TPC-DS queries. Overall, H3-A achieves a good balance between HV and solving time.

Expt 7: Comparison with SOTA MOO methods.

We next extend our best compile-time optimization method HMOOC3 with runtime optimization, denoted as HMOOC3*, which includes the adaptive sampling rate scheme introduced in Section 5.1 to achieve a better balance between accuracy and efficiency. We compare it with existing methods in actual execution time when Spark AQE is enabled. To account for model errors, we refine the search range for each Spark parameter by avoiding the extreme values of the parameter space that could make the predictions less reliable.

6.3 End-to-End Evaluation

We now extend our best compile-time optimization method HMOOC3 with runtime optimization, denoted as HMOOC3*, which includes the adaptive sampling rate scheme introduced in Section 5.1 to achieve a better balance between accuracy and efficiency. We compare it with existing methods in actual execution time when Spark AQE is enabled. To account for model errors, we refine the search range for each Spark parameter by avoiding the extreme values of the parameter space that could make the predictions less reliable.

Expt 8: Benefits over Default Configuration.

Figure 12(a) and 12(b) show the per-query latency using HMOOC3* or the default configuration. Among the 22 TPC-H and 102 TPC-DS queries, HMOOC3* significantly outperforms the default configuration for most of the queries, while being similar (within 2s latency improvement) for three short-running queries (h6, ds8, and ds20). HMOOC3* loses to the default configuration only for one short-running query (h14), mainly due to the model error in predicting its latency. As we use WMAPE as the training loss to give more weight to long-running
queries, short queries can have inaccurate predictions and hence experience a suboptimal query plan or insufficient resource allocation.

Expt 9: Benefits over Query-level MOO. We next show the advantages of our methods (HMOOC3 and HMOOC3+*) over the best-performing MOO method identified in the previous study, i.e., WS for query-level MOO, denoted as MO-WS. Here, we prioritize latency over cost with a preference vector of (0.9, 0.1) on latency and cost. The results in Table 5 show the improvement over the default Spark configuration. First, fine-grained tuning significantly enhances performance (major result R1), cutting latency by 61% for both benchmarks with compile-time optimization (HMOOC3), and by 63-65% with runtime optimization (HMOOC3*). They both outperform MO-WS with only 18-25% reductions and in some cases, worse than the default configuration. Second, MO-WS, even limited to query-level tuning, suffers in efficiency, with 14s and 26s as the P50 and P90 solving time, respectively, for TPC-DS. In contrast, our approach solves MOO with a P50 solving time of 0.65-0.69s and P90 time of 0.9-1.4s for the two benchmarks.

Expt 10: Benefits for Long-running Queries. We further consider long-running queries that are hard to optimize and often suffer from suboptimal plans with $\theta_p$ tuned based on compile-time cardinality estimates. Figure 12(c) shows the top long-running queries from TPC-H and TPC-DS. While HMOOC3 already offers significant latency reduction, HMOOC3+ achieves up to a 22% additional latency reduction over the default configuration. We finally turned on multi-query plan search, running up to 3 query plan structures in parallel at compile-time optimization. This method, HMOOC3*+, further improves $h_9$, $h_21$, ds58, etc. Overall, HMOOC3*+ reduces latency by 64-85% for long-running queries (R2).

Expt 11: Adaptability Comparison to SO with fixed weights. As MO-WS is too slow for cloud use due to its inefficiency, we now compare the adaptability of our approach against the common, practical approach that combines multiple objectives into a single objective using fixed weights [25, 71, 78], denoted as SO-FW. Table 6 shows the average reduction rates in latency and cost relative to the default configurations across a range of preference vectors. First, HMOOC3*+ dominates SO-FW with more latency and cost reductions in most cases (R3), achieving up to 55-59% latency reduction and 10-22% cost reduction in both benchmarks, while SO-FW gets at most 1-34% average latency reduction and in most cases, increases the cost compared to default. Second, our approach demonstrates superior adaptability to varying preferences (R4), enhancing latency reductions progressively as preferences shift towards speed. In contrast, SO-FW does not make meaningful recommendations: Under a cost-saving preference of (0.0, 1.0), SO-FW struggles to lower costs in TPC-DS, instead increasing the average cost by 64% across all queries. Despite this cost increase, it achieves a merely 6% reduction in latency. In contrast, HMOOC3*+ achieves a 45% reduction in latency alongside a 22% cost saving, underscoring its effectiveness and adaptability to the specified cost performance preference.

7 CONCLUSIONS

This paper presented a Spark optimizer for fine-grained parameter tuning in the new AQE architecture based on a hybrid compile-time/runtime optimization approach. Our approach employed sophisticated modeling techniques to capture different compile-time and runtime modeling targets, and a suite of techniques tailored for multi-objective optimization (MOO) while meeting the stringent solving time constraint of 1-2 seconds. Evaluation results using TPC-H and TPC-DS benchmarks show that (i) when prioritizing latency, our approach achieves 63% and 65% latency reduction on average for TPC-H and TPC-DS, respectively, under the solving time of 0.7-0.8 sec, outperforming the most competitive MOO method with 18-25% latency reduction and high solving time of 2.6-15 sec; (ii) when shifting preferences between latency and cost, our approach dominates the solutions from alternative methods by a wide margin. In the future, we plan to extend our tuning approach to support diverse (e.g., machine learning) workloads. In addition, we plan to extend our approach to other big data/DBMS systems like Presto [33], Greenplum [33], and MaxCompute [36], which can observe runtime statistics and support runtime adaptability.


This means that \( N \) denoted as a hypothesis. Therefore, a Pareto optimal solution for the query-level \( F \) can be expressed as the query-level Pareto optimal solutions.

Proposition 5.1. Below, we provide the proof of A.1.1 Subquery (subQ) Tuning.

**Algorithm 2: General_Divide_and_conquer**

```
Require: subQ-level values \( \Omega, \) subQ-level configurations \( \Theta. \)
1: if \(|\Omega| = 1 \) then
2: return \( \Omega, \Theta \)
3: else
4: \( \Omega^h, \Theta^h = \text{first_half}(\Omega, \Theta) \)
5: \( \Omega', \Theta' = \text{second_half}(\Omega, \Theta) \)
6: \( F^h, C^h = \text{General_Divide_and_conquer}(\Omega^h, \Theta^h) \)
7: \( F', C' = \text{General_Divide_and_conquer}(\Omega', \Theta') \)
8: return merge\( (F^h, C^h, F', C') \)
9: end if
```

**Algorithm 3:** merge

```
Require: \( F^h, C^h, F', C' \)
1: \( F, C = \emptyset, \emptyset \)
2: for \((F_1, F_2), (c_1, c_2) \in (F^h, C^h) \) do
3: for \((F_1', F_2'), (c_1', c_2') \in (F', C') \) do
4: \( F = F \cup \{(F_1 + F_1', F_2 + F_2')\} \)
5: \( C = C \cup \{(c_1 + c_1', c_2 + c_2')\} \)
6: end for
7: end for
8: return \( F, C = \text{filter}_{\text{dominated}}(F, C) \)
```

### A ADDITIONAL MATERIALS FOR MOO

#### A.1 Algorithms and Proofs

In this section, we include the algorithms, proofs and complexity analysis of our optimization techniques.

**A.1.1 Subquery (subQ) Tuning.** Below, we provide the proof of Proposition 5.1.

**Proposition 5.1** Under any specific value \( \theta_c \), only subQ-level Pareto optimal solutions \((\theta_c^q, \theta_p^q)\) for the i-th subQ contribute to the query-level Pareto optimal solutions \((\theta_c, \theta_p^q)\).

**Proof.** Let \( F^q_i \) be a query-level Pareto optimal solution for \( \theta_c^q \). It can be expressed as \( F^q_i = \sum_{j=1}^{m} F^q_{i,j} \). Assume that there exists at least one \( i \), e.g., \( i_1 \), such that \( F^q_{i,1} \) is not optimal for the i-th subQ. Let \( F_{q, i_1}^q = F_{q, i_1}^r + \sum_{j=1, i \neq i_1}^{m} F^q_{i,j} \) where \( F^q_{i,j} \) is Pareto optimal for the i-j-th subQ.

We have

\[
F^q_{i} - F^q_{i_1} = F^q_{i,1} - F^q_{i_1}
\]

Since \( F^q_{i,1} \) is optimal for the i-j-th subQ, we have \( F^q_{i,1} > F^q_{i,1} \). This means that \( F^q_{i,1} \) is dominated by \( F^q_{i,1} \), which contradicts our hypothesis. Therefore, a Pareto optimal solution for the query-level can only contain subQ-level Pareto optimal solutions under a fixed \( \theta_c^q \).

#### Complexity analysis of model inference related to the sampling rate

The sampling rate includes the number of \( \theta_c \) and \( \theta_p \) samples, denoted as \( N_c \) and \( N_p \) respectively. The model is called three times in line 3, 4 and 7 respectively.

Line 3 calls the model to provide the predictions of \( C \) cluster-representative \( \theta_c \) and \( N_p \) \( \theta_p \) samples. For \( m \) subQs, there is \( m \times C \times N_p \) predictions. And it filters the dominated solutions of each \( \theta_c \) under each subQ to ensure the \( \theta_c \) are identical among all subQs while keeping the optimality of \( \theta_p \) as illustrated in Proposition 5.1.

Line 4 assigns the optimal \( \theta_p \) solutions to \( N_c \) initialized \( \theta_c \) samples for all subQs and provides the corresponding predictions. Since the number of optimal solutions of a \( \theta_c \) sample for different subQs could be different, for simplicity, we use \( p_{\text{aug}} \) to denote the average number of optimal solutions among all \( \theta_c \) samples and all subQs. The total number of predictions is \( N_c \times p_{\text{aug}} \times m \).

Line 7 functions the same as line 4, where the only difference is to assign the optimal \( \theta_p \) solutions (returned from line 3) to the newly generated \( \theta_c \) samples for all subQs. Assume \( N_c^\text{new} \) new \( \theta_c \) samples are generated, the total number of predictions is \( N_c^\text{new} \times p_{\text{aug}} \times m \).

#### A.1.2 DAG Aggregation

**HMOOC1: Divide-and-Conquer**. This DAG aggregation method is described in Algorithm 2. The \( \Omega, \Theta \) are the effective set of all subQs, where \( \Omega \) represents subQ-level objective values, and \( \Theta \) represents the corresponding configurations, including \( \{\theta_c, \{\theta_p\}, \{\theta_p^q\}\} \). If there is only one subQ, it returns the \( \Omega, \Theta \) (lines 1-2). Otherwise, it follows a divide-and-conquer framework (lines 4-8).

The main idea is a merging operation, which is described in Algorithm 3. The input includes the subQ-level objective values (e.g. \( F^h \) is a Pareto frontier) and its configurations (\( C^* \)) for the two nodes to be merged, where \( h \) and \( r \) denote they are two different nodes. It merges two nodes into a pseudo node by enumerating all the combinations of solutions in the two nodes (lines 2-3), summing up their objective values (lines 4-5) and taking its Pareto frontier as the solutions of this pseudo node (line 8).

From the view of optimality, Algorithm 2 is proved to return a full set of the query-level Pareto optimal solutions as it enumerates over all subQ-level solutions. The complexity of merge function is \( O(M^2 + N) + O((M^2 + N) \log(M + N)) \) if there are \( M \) and \( N \) solutions in two nodes, where the enumeration takes \( O(M^2 + N) \) and filtering dominated solutions takes \( O((M + N) \log(M + N)) \). While after merging several times, \( M \) and \( N \) could be highly. Thus the total complexity could be high.

The core operation in HMOOC1 is the merge, which enumerates over all subQ-level solutions. The following are the theoretical proof. For the sake of simplicity, we consider the case with two nodes.

**Proposition A.1.** Algorithm 2 always output the full Pareto front of the simplified DAG.

**Proof.** Let \( D \) and \( G \) be two nodes (e.g., subQs or aggregated subQs). Let \( \oplus \) be the Minkowski sum, i.e., \( D \oplus G = \{F_D + F_G, F_D \in D, F_G \in G\} \). Let \( PF \) denote the Pareto Front of a node.

\[
P_F(P_F(D) \oplus P_F(G)) = P_F(D \times G)
\]

Let \( E_D : C_D \rightarrow \mathbb{R} \) be the evaluation function of node \( D \), where \( C_D \) is the set of configurations for node \( D \). We define \( E_G \) in a similar manner. We also define \( E : (C_D, C_G) \mapsto E_D(C_D) + E_G(C_G) \), where \( C_D \) and \( C_G \) are one configuration in \( C_D \) and \( C_G \) respectively.
Algorithm 4: Compressing_list_nodes

Require: subQ_list, ws_pairs.
1: PO = [ ], conf = [ ]
2: for [ w, w ] in ws_pairs do
3:   po_n = [ ], conf_n = []
4: for subQ-po, subQ-conf in subQ_list do
5:   subQ-po-norm = normalize_per_subQ(subQ-po)
6:   opt_ind = minimize_ws([ w, w ], subQ-po-norm)
7:   po_n.append(subQ-po[opt_ind])
8:   conf_n.append(subQ-conf[opt_ind])
9: end for
10: end for
11: return filter-dominated(PO, conf)

Let \( p \in Pf(Pf(D) \oplus Pf(G)) \). Then \( p \) can be addressed as a sum of two terms: one from \( Pf(D) \) and the other from \( Pf(G) \), i.e.,

\[
p = PD + PG,
\]

\( PD \in Pf(D), PG \in Pf(G) \)

Let \( cD \in Pf^{-1}(PD) \) and \( cG \in Pf^{-1}(PG) \), i.e., \( cD \) is chosen such as \( E_P(c) = PD \), and likewise for \( cG \). Then we have \( p = E(c_D, c_G) \), so \( p \) belongs to the objective space of \( D \times G \). Suppose \( p \) doesn’t belong to \( Pf(D \times G) \). This means that there exists \( p' \) in \( Pf(D \times G) \) that dominates \( p \). \( p' \) can be expressed as \( p' = P_D' + P_G' \) with \( P_D' \in Pf(D) \) and \( P_G' \in Pf(G) \). \( p' \) dominates \( p \) can be expressed as \( P_D' \leq PD \) and \( P_G' \leq PG \) or \( P_D' \leq PD \) and \( P_G' \leq PG \) in our optimization problem with minimization, which contradicts the definition of \( p \) as belonging to the Pareto Front. Therefore \( p \) must belong to \( Pf(D \times G) \) and \( Pf(Pf(D) \oplus Pf(G)) \). Let us now suppose that \( p \in Pf(D \times G) \). Then there exists \( c \in D \times G \), i.e., \( c = (c_D, c_G) \) with \( c_D \in D, c_G \in G \), such that \( p = E(c) \). By setting \( PD = E_D(c_D) \) and \( PG = E_G(c_G) \), we have \( p = PD + PG \). By definition, \( PD \) belongs to \( Pf(D) \) and \( PG \) belongs to \( Pf(G) \). Hence, \( p = Pf(D \times G) \).

Suppose that \( p \) doesn’t belong to \( Pf(D \times G) \) and that there exists \( p' \) in \( Pf(D \times G) \) that dominates \( p \). We showed above that \( p' \) must belong to \( Pf(D \times G) \). Thus both \( p \) and \( p' \) belong to \( Pf(D \times G) \) and \( p' \) dominates \( p \), which is impossible. Therefore \( p' \) must belong to \( Pf(Pf(D) \oplus Pf(G)) \) and \( Pf(Pf(D) \oplus Pf(G)) \subseteq Pf(D \times G) \).

By combining the two inclusions, we obtain that \( Pf(Pf(D) \oplus Pf(G)) = Pf(D \times G) \), where the left side is the Pareto optimal solution from the Algorithm 2 and the right side is the Pareto optimal solution over the whole configuration space of two nodes. Thus, Algorithm 2 returns a full set of Pareto solutions.

HMOOC2: WS-based Approximation. We propose a second technique to approximate the MOO solution over a list structure.

For each fixed \( \theta_S \), we apply the weighted sum (WS) method to generate evenly spaced weight vectors. Then for each weight vector, we obtain the (single) optimal solution for each subQ and sum the solutions of subQ’s to get the query-level optimal solution. It can be proved that this WS method over a list of subQs guarantees to return a subset of query-level Pareto solutions.

Algorithm 4 describes the full procedures. The input includes a subQ_list, which includes both subQ-level objective values and the corresponding configurations of all subQs. ws_paris are the weight pairs, e.g. \([0, 0.5, 0.5], [0.5, 0.5, 0.5], \ldots\) for latency and cost. Line 1 initializes the query-level objective values and configurations. Lines 3-9 address the Weighted Sum (WS) method to generate the query-level optimal solution for each weight pair. Specifically, Lines 5-8 apply the WS method to obtain the optimal solution choice for each subQ and sum all subQ-level values to get the query-level values. Upon iterating through all weights, a Pareto solution set is derived after the necessary filtering (Line 12).

Theoretical Analysis

Lemma 1. For a DAG aggregation problem with \( k \) objectives that use the sum operator only, Algorithm 4 guarantees to find a non-empty subset of query-level Pareto optimal points under a specified \( \theta_S \) candidate.

In proving Lemma 1, we observe that Algorithm 4 is essentially a Weighted Sum procedure over Functions (WSF). Indeed we will prove the following two Propositions: 1) each solution returned by WSF is Pareto optimal; 2) the solution returned by the Algorithm 4 is equivalent to the solution returned by WSF. Then it follows that the solution returned by Algorithm 4 is Pareto optimal.

To introduce WSF, we first introduce the indicator variable \( x_{ij} \), \( i \in [1, m], j \in [1, p_i] \), to indicate that the \( j \)-th subQ in \( i \)-th subQ list is selected to contribute to the query-level solution. \( \sum_{j=1}^{p_i} x_{ij} = 1 \) means that only one solution is selected for each subQ. Then \( x = [x_{1j_1}, \ldots, x_{m_j_m}] \) represents the 0/1 selection for all \( m \) subQs to construct a query-level solution. Similarly, \( f = [f_{i_1j_1}, \ldots, f_{k_mj_m}] \) represents the value of the objectives associated with \( x \).

So for the \( v \)-th objective, its query-level value could be represented as the function \( H_v \) applied to \( x \):

\[
F_v = H_v(x; f) = \sum_{i=1}^{m} \sum_{j=1}^{p_i} x_{ij} \cdot f_{ij}^v,
\]

where \( \sum_{j=1}^{p_i} x_{ij} = 1, i \in [1, m], v \in [1, k] \)

For simplicity, we refer to \( H_0(x; f) \) as \( H_0(x) \) when there is no confusion. Now we introduce the Weighted Sum over Functions (WSF) as:

\[
\arg\min_x \left( \sum_{v=1}^{k} w_v \cdot H_v(x) \right)
\]

s.t. \( \sum_{v=1}^{k} w_v = 1, \quad w_v \geq 0 \quad \text{for} \quad v = 1, \ldots, k \)

Where \( w_v \) is the weight value for objective \( v \). Next, we prove for Lemma 1. As stated before, it is done in two steps.

Proposition A.2. The solution constructed using \( x \) returned by WSF is Pareto optimal.

Proof.

Assume that \( x^* \) (corresponding to \( [f_{i_1}^*, \ldots, f_{k}^*] \)) is the solution of WSF. Suppose that another solution \( [f_{i_1}, \ldots, f_{k}^*] \) (corresponding to \( x' \)) dominates \( [f_{i_1}^*, \ldots, f_{k}^*] \). This means that \( \sum_{v=1}^{k} w_v \cdot H_v(x^*) \) is smaller than that of \( x^* \).
This contradicts that \( x' \) is the solution of WSF. So there is no \([F_1',...,F_k']\) dominating \([F_1',...,F_k']\). Thus, \([F_1',...,F_k']\) is Pareto optimal.

**Proposition A.3.** The optimal solution returned by the Algorithm 4 is equivalent to the solution constructed using \( x \) returned by WSF.

**Proof.** Suppose \( x' \) is returned by WSF. The corresponding query-level solution is \([F_1',...,F_k']\)

\[
x' = \arg \min_x \sum_{q=1}^k w_q \times H_q(x)
\]

\[
= \arg \min_x \sum_{q=1}^k w_q \times \left( \sum_{i=1}^m p_i x_{ij} \times f_{ij}^q \right)
\]

\[
= \arg \min_x \sum_{i=1}^m \left( \sum_{j=1}^k (w_q \times f_{ij}^q) \times x_{ij} \right)
\]

For the solution \([F_1'',...,F_k'']\) returned by Algorithm 4, \( x'' \) represents the corresponding selection. It is achieved by minimizing the following formula:

\[
\sum_{i=1}^m \min_{j \in \{1,...,k\}} (WS_{ij})
\]

\[
= \sum_{i=1}^m \min_{j \in \{1,...,k\}} \left( \sum_{q=1}^k w_q \times f_{ij}^q \right)
\]

\[
= \sum_{i=1}^m \min_{j \in \{1,...,k\}} \left( \sum_{q=1}^k \sum_{l=1}^m p_l \times f_{ij}^q \times x_{ij} \right)
\]

where \( WS_{ij} = \sum_{q=1}^k w_q \times f_{ij}^q \). Given a fixed \( i \), \( x_{ij} \) can only be positive (with value 1) for one value of \( j \). So, \( x'' \) must solve:

\[
x'' = \left( \sum_{i=1}^m \arg \min_{x_{ij}} \left( \sum_{j=1}^k (w_q \times f_{ij}^q) \times x_{ij} \right) \right)
\]

\[
= \arg \min_x \left( \sum_{i=1}^m \left( \sum_{j=1}^k (w_q \times f_{ij}^q) \times x_{ij} \right) \right)
\]

Here, optimizing for each \( subQ \) is independent of the optimization of the other \( subQs \), so we can invert the sum over \( i \) and the arg min. Thus, \( x'' = x'''' \). Therefore, WSF and Algorithm 4 are equivalent. □

With these two propositions, we finish the proof of Lemma 1. Algorithm 4 varies \( w \) weight vectors to generate multiple query-level solutions. And under each weight vector, it takes \( O(m \cdot p_{max}) \) to select the optimal solution for each LQP-subtree based on WS, where \( p_{max} \) is the maximum number of solutions among \( m \) \( subQs \). Thus, the overall time complexity of one \( \theta_c \) candidate is \( O(w \cdot (m \cdot p_{max})) \).

**HMOOC3: Boundary-based Approximation.**

We now show the proof of Proposition 5.2.

**Proposition 5.2** Under a fixed \( \theta_c \) candidate, the query-level objective space of Pareto optimal solutions is bounded by its extreme points in a 2D objective space.

**Proof.**

Assume that \( F_1' = [F_1'^1, F_1'^2] \) and \( F_2' = [F_2'^1, F_2'^2] \) are two extreme points under a fixed \( \theta_c \), recalling that the extreme point under a fixed \( \theta_c \) is the Pareto optimal point with the best query-level value for any objective. Here the superscript \( \{1\} \) means it achieves the best in objective 1 and \( \{2\} \) means it achieves the best in objective 2. The two extreme points form an objective space as a rectangle.

Suppose that an existing query-level Pareto optimal solution \( F_q' = [F_q'^1, F_q'^2] \) is outside this rectangle, which includes 2 scenarios.

In scenario 1, it has \( F_q' < F_q'^1 \) or \( F_q' < F_q'^2 \), which is impossible as extreme points already achieves the minimum values of two objectives. In scenario 2, it has \( F_q' > F_q'^1 \) or \( F_q' > F_q'^2 \), which is impossible as \( F_q' \) is dominated by any points inside the rectangle.

So there is no Pareto optimal solution \( F_q' \) existing outside the rectangle and it concludes the proof. □

We next show the proof of Proposition 5.3.

**Proposition 5.3** Given subQ-level solutions, our boundary approximation method guarantees to include at least \( k \) query-level Pareto optimal solutions for a MOO problem with \( k \) objectives.

**Proof.**

Assume that \( F_1'^1,...,F_k'^k \) are \( k \) extreme points, which are Pareto optimal and achieve the best (e.g. the lowest in the minimization problem) query-level values of objectives 1,...,\( k \) among all \( \theta_c \) configurations. Suppose that an existing Pareto optimal solution \( F_q' \), distinct from the extreme points, dominates any point in \( F_1'^1,...,F_k'^k \). \( F_q' \) must achieve a better value than \( F_1'^1,...,F_k'^k \) in any objectives 1,...,\( k \), where the superscript \( \{1\} \) means it achieves the best in objective 1 and \( \{k\} \) means it achieves the best in objective \( k \), which is impossible as the extreme points already achieves the best.

So these \( k \) extreme points cannot be dominated by any other solutions. Thus, they are Pareto optimal and this concludes the proof. □

**A.2 Compile-time optimization**

**A.2.1 \( \theta_c \) enrichment.** The following theoretical result sheds light on \( subQ \) tuning, indicating that \( \theta_c \) derived from the \( subQ \)-level Pareto optimal solutions can serve as a useful warm-start for generating new \( \theta_c \) configurations.
In Figure 14, the others.

Proposition A.4. As shown in Figure 13, for all sub-Qs, solutions with the same θ_e located in the red region cannot contribute to query-level Pareto optimal solutions.

Proof. In any sub-Q of an arbitrary DAG, the dominated solution f^* (red area) is dominated by f^** with any arbitrary θ_e.

Given query-level solution F built from f^*, f^** built from f^***, supposing F is non-dominated with f^* or dominates f^*, there must be at least one f^* with lower latency or cost than f^**, which is impossible and concludes the proof.

To uncover unexplored θ_e configurations, drawing inspiration from the crossover operation in evolutionary algorithms, we introduce a heuristic method termed θ_e crossover. The core concept involves utilizing the Cartesian Product (CP) operation to generate new θ_e candidates from the existing pool. This operation is executed by randomly selecting a crossover location to divide the initial θ_e configurations into two parts. The following example illustrates this process.

Example In Figure 14, θ_e1 and θ_e2 represent initial θ_e candidates obtained from the sub-Q-level tuning, each consisting of 8 variables (e.g., k1, ..., k8). The values within the boxes denote the configurations of θ_e. The blue point denotes a randomly generated crossover location. Based on this location, θ_e3 and θ_e6 are divided into two parts: k1, k2, k3 and k4, k5, k6, k7, k8, delineated by a blue rectangle. A Cartesian Product (CP) is then applied to these two parts. In this example, there are two distinct configurations for k1, k2, k3 and two distinct configurations for k4, k5, k6, k7, k8. Consequently, the CP generates four configurations, represented as θ_e3 to θ_e6. It is noteworthy that θ_e crossover generates new θ_e configurations (e.g., θ_e4, θ_e5) without discarding the initial θ_e candidates (e.g., θ_e3, θ_e6).

Analysis on Boundary-based Approximation

Since different θ_e values result in diverse total resources and cover various regions of the Pareto frontier, it’s noteworthy that all DAG optimization methods produce the same effective set, as confirmed by Figure 15. In scenarios with multiple θ_e candidates, the boundary-based method achieves comparable hypervolume to the others.

Figure 15: Objective space under different resources

A.3 Additional Materials for Runtime Optimization

A.3.1 More on θ_p and θ_e Aggregation. Ideally, one could copy θ_p and θ_e from the initial sub-Q, allowing the runtime optimizer to adjust them by adapting to the real statistics.

Given the constraint that Spark takes only one copy of θ_p and θ_e at query submission time, we intelligently aggregate the fine-grained θ_p and θ_e from compile-time optimization to initialize the runtime process. In particular, Spark AQE can convert a sort-merge join (SMJ) to a shuffling hash join (SHJ) or a broadcast hash join (BHJ), but not vice versa. Thus, imposing high thresholds (s_3, s_4 in Table 1) to force SHJ or BHJ based on inaccurate compile-time cardinality can result in suboptimal plans (as shown in Figure 3(b)). On the other hand, setting these thresholds to zero at SQL submission might overlook opportunities for applying BHJs, especially for joins rooted in scan-based sub-Qs with small input sizes. To mitigate this, we initialize θ_p with the smallest threshold among all join-based sub-Qs, enabling more informed runtime decisions. In addition, we cap these thresholds (25 MB for broadcast threshold s_3 and 0 MB for shuffle hash threshold s_5) at their default values to ensure BHJs are not missed for small scan-based sub-Qs.

A.3.2 More on Pruning Optimization Requests. To address this, we established rules to prune unnecessary requests based on the runtime semantics of parametric rules: LQP parametric rules are used to decide join algorithms and QS parametric rules are used to re-balance the data partitions in a post-shuffle QS. Therefore, we bypass requests for non-join operations and defer requests for LQP-containing join operators until all input statistics are available, thereby avoiding decisions based on inaccurate cardinality estimations. Additionally, we skip all the scan-based QSs and only send the requests when the input size of a QS is larger than the target partition size (configured by s_2). By applying the above rules, we substantially reduce the total number of optimization calls by 86% and 92% for TPC-H and TPC-DS respectively.

A.4 Scalability Issues

The dimensionality of the parameter space is considered in three aspects in this work.
We clarify that Spark has over 100 parameters and we selected 19 features as follows: (a) We keep query plan parameters $\theta_q$ and partition related parameters $\theta_p$, which can be replicated to have one copy for each query stage in our fine-grained tuning approach. (b) Context parameters: The dimensionality of the context parameter space $\theta_c$ can reach up to hundreds [66]. To address the issue, we use the LASSO-path feature selection approach (as in our prior work [51] and in Ottertune [59]) and integrate domain knowledge from Spark documentation, resulting in a list of 8 context parameters. More details are in Appendix B.2.

Feature importance score (FIS) based parameter filtering: We conducted additional experimental to compute the feature importance scores (FIS) of the 19 parameters in our trained latency and data shuffle models. The FIS of a parameter here is defined to be the increase of the model error if the values of this parameter are randomly shuffled in the training set [46]. Results in Figure 21-23 show the FIS of each parameter in a long tail distribution. The parameters at the tail, with a cumulative FIS less than 5% in WMAPE, do not affect the model performance by much and hence can be filtered out. For instance, the threshold of 5% cumulative FIS means an increment of at most 0.7% in WMAPE for TPC-H and 1.2% for TPC-DS in the subQ model.

Sampling with FIS-based parameter filtering: Then we introduced three sampling methods that could leverage FIS-based selection: (a) random sampling with or without parameter filtering; (b) Latin hypercube sampling with or without parameter filtering; (c) adaptive grid search with inherent feature filtering, as we described in the new paragraph starting with "Sampling methods" in Section 5.1.1.

Experimental results: Results in Figure 11(b)-11(c) show performance metrics (hypervolume and solving time) of the above three sampling methods with parameter filtering. Among them, we report the adaptive grid search to be an overall winner, achieving the highest hypervolume among all FIS-based methods while keeping the solving time under 2s for most queries tested.

Runtime adaptive scheme for adaptive grid search: If the predicted latency under the default configuration is over 10s, we use the median sampling rate; otherwise, a low sampling rate. Our analyses show that further increasing the sampling rate does not improve performance but increases solving time, hence negatively impacting query performance.

B.1.1 Hardware. We use two 6-node Spark 3.5.0 clusters with runtime optimization plugins. Each node is CentOS-based with 2 16-core Intel Xeon Gold 6130 processors, 768 GB of RAM, and RAID disks, connected with 100 Gbps Ethernet.

B.1.2 Parametric Optimization Rules. Figure 19 illustrates the transformations of a collapsed LQP and a runtime Q5, which both pass through a pipeline of parametric rules (blue) and non-parametric rules (gray).

B.1.3 Spark Parameters Details. We list our 19 selected Spark parameters in Table 1, which are categorized into three groups: context parameters ($\theta_c$), logical query plan parameters ($\theta_p$), and query stage parameters ($\theta_q$). The default configuration is set to Spark’s default values shown in the table.

B.2 Tunable Parameter Selection

We first select parameters for each query based on the LASSO path approach as our prior work [51] and Ottertune [59] did. For each query, we prepared 50 context parameters to form a tunable Spark configuration in high dimension, and run each query over 1000 different configurations. Based on the traces, we trained LASSO regression models for each query and generate the LASSO Path in each model to evaluate the importance of each parameter. Figure 20 shows the LASSO path for different queries. As shown in the figure, important parameters from different queries can vary a lot among each other. Therefore, we further integrate the domain knowledge from the Spark official documentation into the results of the LASSO Path approach to agree on a final list parameters as $\theta_c$.

B.3 Adapting Important Parameters

We show the feature importance score (FIS), in terms of the permutation importance [46] of all dimensions of the graph embedding (shown as the non-decision variables), and the selected tunable parameters, shown as the context parameters ($\theta_c$) and SQL parameters ($\theta_p$ and $\theta_q$) in the latency subQ model for TPC-H in Figure 21. The importance score indicates the performance (WMAPE) drop when the corresponding feature is randomly shuffled. A larger number indicates a more critical feature.

Figure 22 shows the importance score of tunable parameters in TPC-H and -DS. We show the importance score of the context parameters ($\theta_c$), SQL parameters ($\theta_p$ and $\theta_q$) in the latency and shuffle size subQ models for TPC-H and -DS.

To select important parameters from the four diverse models, we first reserve all query structure related parameters (e.g., $s_3$ and $s_4$). Then we normalize the importance score of each parameter in each model by the WMAPE values, and aggregate the normalized scores to show a global importance score of each parameter. Figure 23 shows the normalized importance scores of all parameters in a long-tail distribution. By further filtering the long tail of parameters whose cumulative FIS is less than a threshold (e.g., 5%), we get a
list of important parameters, including $k_7$, $k_1$, $k_3$, $k_2$ for context parameters, and $s_3$, $s_4$, $s_5$, $s_8$, $s_9$, $s_1$ for the SQL parameters.

B.4 Handle query plan related parameters

As explained in Section 5.2, the join threshold parameters $s_4$ cannot be set either too small or too big. Therefore, it is set as the max$(25MB, \min_i s_4^{(i)})$ where $i$ is the indicator of all subQs with at least one join operator.

B.5 Specific Knob Concerns

spark.sql.adaptive.enable=true: The parameter was introduced in Spark 1.6 and has been set to true by default since Spark 3.2. We have chosen to enable it for two main reasons. First, enabling adaptive query execution (AQE) allows us to perform parameter tuning at the stage level. Specifically, our runtime optimizer could fine-tune the runtime parameters (at the stage level) while AQE re-optimizes the query plan based on the actual runtime statistics in the middle of query execution. Second, enabling AQE improves the robustness of query latency. When AQE is disabled, the DAGScheduler asynchronously converts the entire query to a DAG of stages. Consequently, parallel stages can be randomly interleaved during query execution, leading to unpredictable query latencies. For instance, Figure 24 demonstrates that disabling AQE can result in different stage interleaving patterns, causing a significant 46% increase in query latency. When AQE is enabled, stages are wrapped in query stages (QSs) that are synchronously created. As a result, the stage interleaving patterns are consistent for running one query, making the query performance more stable and predictable.

spark.locality.wait=0s: The default value of the parameter is 3s, which specifies the wait time for launching a data-local task before giving up and launching it on a less-local node. However, the waiting behavior can introduce instability in query performance due to the randomness of locality detection. As demonstrated in Figure 25(a) and Figure 25(b), the latency of a stage can vary significantly (changing from 11.7s to 18.4s) when different locality detection methods are employed. To ensure a stable query performance in our workload, we have fixed the spark.locality.wait
Figure 18: Actual latency under different sampling rates for ($\theta_c$ and $\theta_p$)

Table 7: (Selected) Spark parameters in three categories

<table>
<thead>
<tr>
<th>$\theta_c$ Context Parameters</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$ spark.executor.cores</td>
<td>1</td>
</tr>
<tr>
<td>$k_2$ spark.executor.memory</td>
<td>1G</td>
</tr>
<tr>
<td>$k_3$ spark.executor.instances</td>
<td></td>
</tr>
<tr>
<td>$k_4$ spark.default.parallelism</td>
<td>48M</td>
</tr>
<tr>
<td>$k_5$ spark.reducer.maxSizeInFlight</td>
<td>200</td>
</tr>
<tr>
<td>$k_6$ spark.shuffle.sort.bypassMergeThreshold</td>
<td>True</td>
</tr>
<tr>
<td>$k_7$ spark.shuffle.compress</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_p$ Logical Query Plan Parameters</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$ spark.sql.adaptive.advisoryPartitionSizeInBytes</td>
<td>64M</td>
</tr>
<tr>
<td>$s_2$ spark.sql.adaptive.nonEmptyPartitionRatioForBroadcastJoin</td>
<td>0.2</td>
</tr>
<tr>
<td>$s_3$ spark.sql.adaptive.maxShuffledHashJoinLocalMapThreshold</td>
<td>0b</td>
</tr>
<tr>
<td>$s_4$ spark.sql.adaptive.autoBroadcastJoinThreshold</td>
<td>10M</td>
</tr>
<tr>
<td>$s_5$ spark.shuffle.partitions</td>
<td>200</td>
</tr>
<tr>
<td>$s_6$ spark.sql.adaptive.skewedPartitionThresholdInBytes</td>
<td>256M</td>
</tr>
<tr>
<td>$s_7$ spark.sql.adaptive.skewedPartitionFactor</td>
<td>5.0</td>
</tr>
<tr>
<td>$s_8$ spark.sql.files.maxPartitionBytes</td>
<td>128M</td>
</tr>
<tr>
<td>$s_9$ spark.sql.files.openCostInBytes</td>
<td>4M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_s$ Query Stage Parameters</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_{10}$ spark.sql.adaptive.rebalancePartitionsSmallPartitionFactor</td>
<td>0.2</td>
</tr>
<tr>
<td>$s_{11}$ spark.sql.adaptive.coalescePartitions.minPartitionSize</td>
<td>1M</td>
</tr>
</tbody>
</table>

Figure 20: LASSO path for different queries, with weights of different parameters against the L1 penalty (varied from $10^{-3}$ to $10^0$). The importance of each parameter is ordered by the timing of its weight deviating from zero.
Figure 21: Feature importance scores of all parameters in the latency subQ model for TPC-H

Figure 22: Feature importance scores of tunable parameters in TPC-H and TPC-DS

Figure 23: Normalized feature importance scores of all parameters across four models

parameter to 0s, thereby avoiding the waiting time for locality and achieving consistent and better query performance as shown in Figure 25(c). It is worth noting that in the production environment [32], the impact of locality is mitigated due to the high-speed network cards, which aligns with a near-zero waiting time.

spark.sql.adaptive.coalescePartitions.parallelismFirst=false: We respect the recommendation on the Spark official website and set
this parameter to false such that the advisory partition size will be respected when coalescing contiguous shuffle partitions. 

spark.sql.adaptive.forceOptimizeSkewedJoin=false: We follow the default setting for the parameter and avoid optimizing the skewed joins if it requires an extra shuffle. However, if needed, we can set it to true and adjust our tuning process to consider $s_5$-$s_7$ as three additional plan-dependent parameters.

### B.6 More Skewness Analyses

Figure 28(a) and Figure 28(d) show the CDF of query stage (QS) latencies in TPC-H and TPC-DS, respectively, showing the distribution of the QS latencies and the important percentiles. The error distributions in different skewness ranges across different latency ranges are shown in Figure 28(b)-28(c) and Figure 28(e)-28(f), respectively, indicating the skewness does not significantly affect the error distribution. Interestingly, we also observed that longer running QSs have higher errors in general, which is consistent with the intuition that longer running queries are more complex.

### B.7 More Model Generalization

Table 8 shows the model performance on TPC-H dataset by (1) a model end-to-end trained over TPC-H, (2) a model end-to-end trained over TPC-DS, and (3) a model retrained over TPC-H based on the graph transformer network (GTN) model trained over TPC-DS. Our results show that applying the DS model directly to TPC-H leads to a significant performance drop, while retraining the model based on the DS-GTN model can achieve a better performance, indicating a good generalization capability of the GTN model.

### B.8 More Integration Evaluation

The framework of compile-time and runtime optimization is shown in Figure 26. We show per query latency comparison with a strong speed preference in TPC-H in Figure 27.

### B.9 More experimental results in compile-time optimization

#### B.9.1 The sampling rate of sampling methods

There are three options of sampling methods, i.e. adaptive grid-search, Latin Hypercube Sampling (LHS) and random sampling. All sampling choices follow the same sampling rate for a fair comparison of performance.

Our experiments in compile-time optimization fix the sampling rate for both TPC-H and TPC-DS queries by selecting the one with higher HV under the constraints of solving time (e.g. $1-2s$) among different sampling rates. Finally, the sampling rate is set as $54 \theta_c$ and 243 $\theta_p$ samples for both TPC-H queries and TPC-DS queries.

#### B.9.2 Analysis on new $\theta_c$ extension

Figure 29 compares HV and solving time with and without new $\theta_c$ extension for all sampling methods, where $h3-\Lambda$, $h3-\tilde{L}$ and $h3-\tilde{R}$ denote results without new $\theta_c$ extension from adaptive grid search, LHS and random sampling. For adaptive grid search, it extends new $\theta_c$ by random sampling. Figures 29(a) and 29(d) show the HV and solving time of all queries in TPC-H and TPC-DS. Without new $\theta_c$ extension, HV reduces slightly compared to the HV with new $\theta_c$ extension ($\leq 0.6\%$). This is because the adaptive grid search uses random sampling to extend new $\theta_c$, which works to extend more global Pareto optimal $\theta_c$ while with limit capability due to the high-dimensional space.

Figures 29(b) and 29(e) measure the performances of LHS with and without new $\theta_c$ extension by crossover. It shows the HV and solving time of all queries in TPC-H and TPC-DS. Without new $\theta_c$ extension, HV reduces up to 2.6% compared to the HV with crossover. This is because the boundary solutions (i.e. solutions with minimum latency or minimum cost) impact largely on HV and boundary solutions of both methods are similar, where the boundary solutions are mainly decided by the total resources that are set initially rather than the process of new $\theta_c$ extension. The solving time without crossover decreases as it also reduces procedures after new $\theta_c$ extension. Similar observations in Figures 29(c) and 29(f) are shown for extending new $\theta_c$ for random sampling. [CH: a bit hard to follow the writing]

Figure 30 shows Pareto frontier of example queries with LHS and random-sampling. Green points are solutions with crossover operation, and orange points are solutions without crossover. For both LHS and random sampling as initial the sampling method, solutions with crossover result in better solutions in the middle region of the objective space to achieve slightly higher HV. [CH: a bit hard to follow the writing]

#### B.9.3 Analysis on Query-level control

It’s noteworthy that query-control cannot achieve a higher upper-bound than the finer-control. To verify this, we implemented a smaller search space (each parameter having only 2 values) for WS to fully explore query-control, where WS performs the best among all baselines for both TPC-H and TPC-DS. Figure 31 displays the hypervolumes (HVs) of WS under different numbers of samples, with blue and orange bars representing the HVs of finer-control and query-control, respectively. It is observed that as the number of samples increases, the

---

**Figure 25:** The end-to-end stage latency comparison over different settings for spark locality.wait.
HV of query-control plateaus at 1M samples (89%), while the HV of finer-control continues to improve (90%), illustrating the necessity of finer-control in our problem. [CH: hard to convince the readers that 1% of improvements from 89% to 90% is significant enough to have the fine-grained control. Any other metrics with a larger gap between query-control and finer-control to be shown?]
Table 8: Model performance with Graph+Regressor via transform learning

<table>
<thead>
<tr>
<th>Target</th>
<th>Ana-Latency/Latency (s)</th>
<th>Shuffle (MB)</th>
<th>Xput K/s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMAPE</td>
<td>P50</td>
<td>P90</td>
</tr>
<tr>
<td>TPC-H (Trained on TPC-H)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subQ</td>
<td>0.131</td>
<td>0.029</td>
<td>0.292</td>
</tr>
<tr>
<td>QS</td>
<td>0.149</td>
<td>0.027</td>
<td>0.353</td>
</tr>
<tr>
<td>LQP</td>
<td>0.164</td>
<td>0.060</td>
<td>0.337</td>
</tr>
<tr>
<td>TPC-H (Directly applying the DS model)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subQ</td>
<td>0.991</td>
<td>0.173</td>
<td>2.503</td>
</tr>
<tr>
<td>QS</td>
<td>0.907</td>
<td>0.282</td>
<td>2.418</td>
</tr>
<tr>
<td>LQP</td>
<td>1.573</td>
<td>1.156</td>
<td>3.277</td>
</tr>
<tr>
<td>TPC-H (Retrain 3 hours based on DS-GTN)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subQ</td>
<td>0.139</td>
<td>0.018</td>
<td>0.346</td>
</tr>
<tr>
<td>QS</td>
<td>0.143</td>
<td>0.017</td>
<td>0.346</td>
</tr>
<tr>
<td>LQP</td>
<td>0.148</td>
<td>0.046</td>
<td>0.292</td>
</tr>
</tbody>
</table>

Figure 29: Comparison of Hypervolume and solving time with/without new $\theta_C$ extension for three sampling methods

Figure 30: Comparison of Pareto frontiers with/without crossover in LHS and random-sampling

B.9.4 More results on DAG aggregation methods. Figure 32 compares the HVs of three DAG aggregation methods with adaptive grid-search. The HVs of three methods are close with lower than 0.1% differences.

B.9.5 More results on Query-level tuning. Figure 33 shows results of query-level tuning, where TPC-H queries take over an average of at least 2.7s and much lower HV (at most 87%) than HMOOC3 (93%). All of TPCD queries lose to HMOOC3 in HV (at most 84% v.s. 89%) and in solving time (the average exceeding 14s v.s. 0.8s). [CH: a bit hard to read.]

C MORE RELATED WORK

Relationship with other distributed algorithms for system optimization. Li et al. [26] tunes parameters that determine the computation graph, as well as the executor numbers and degree of parallelism to efficiently minimize the execution time by cutting the search space with a runtime environment independent cost model. However, its parameters are not adaptable during the execution, and its search strategy is hard to be adapted to the MOO setting. Having said that, Spark AQE is able to take those distributed algorithms as separate query optimization rules, and our framework can then treat the hyperparameters of those algorithms as SQL parameters and optimize those algorithms by tuning their hyperparameters in compile-time/runtime optimization.
Figure 31: Comparison of query-control and finer-control with smaller searching space

Figure 32: Average Hypervolume with different DAG aggregation methods

Figure 33: Analytical performance of our algorithm, compared to the state-of-the-art (SOTA) methods with query-level tuning