

Rover: An Online Spark SQL Tuning Service via Generalized Transfer Learning

Yu Shen*

School of CS, Peking University
ByteDance Inc.
Beijing, China
shenyu@pku.edu.cn

Xinyuyang Ren*

ByteDance Inc.
Beijing, China
renxinyuyang@bytedance.com

Yupeng Lu

School of CS, Peking University
ByteDance Inc.
Beijing, China
xinkelyp@pku.edu.cn

Huaijun Jiang

Center for Data Science, Peking
University
ByteDance Inc.
Beijing, China
jianghuaijun@pku.edu.cn

Huanyong Xu

ByteDance Inc.
Beijing, China
xuhuanyong@bytedance.com

Di Peng

ByteDance Inc.
Beijing, China
pengdi@bytedance.com

Yang Li

School of CS, Peking University
Beijing, China
liyang.cs@pku.edu.cn

Wentao Zhang

Mila - Québec AI Institute
Montréal, Canada
wentao.zhang@mila.quebec

Bin Cui

School of CS, Peking University
Institute of Computational Social
Science, Peking University (Qingdao)
Beijing, China
bin.cui@pku.edu.cn

ABSTRACT

Distributed data analytic engines like Spark are common choices to process massive data in industry. However, the performance of Spark SQL highly depends on the choice of configurations, where the optimal ones vary with the executed workloads. Among various alternatives for Spark SQL tuning, Bayesian optimization (BO) is a popular framework that finds near-optimal configurations given sufficient budget, but it suffers from the re-optimization issue and is not practical in real production. When applying transfer learning to accelerate the tuning process, we notice two domain-specific challenges: 1) most previous work focus on transferring tuning history, while expert knowledge from Spark engineers is of great potential to improve the tuning performance but is not well studied so far; 2) history tasks should be carefully utilized, where using dissimilar ones lead to a deteriorated performance in production.

In this paper, we present Rover, a deployed online Spark SQL tuning service for efficient and safe search on industrial workloads. To address the challenges, we propose generalized transfer learning to boost the tuning performance based on external knowledge, including expert-assisted Bayesian optimization and controlled history transfer. Experiments on public benchmarks and real-world tasks show the superiority of Rover over competitive baselines.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0103-0/23/08...\$15.00
<https://doi.org/10.1145/3580305.3599953>

Notably, Rover saves an average of 50.1% of the memory cost on 12k real-world Spark SQL tasks in 20 iterations, among which 76.2% of the tasks achieve a significant memory reduction of over 60%.

CCS CONCEPTS

- Computing methodologies → Search methodologies;
- Information systems → Data management systems.

KEYWORDS

Spark SQL, Bayesian Optimization, Transfer Learning

ACM Reference Format:

Yu Shen, Xinyuyang Ren, Yupeng Lu, Huaijun Jiang, Huanyong Xu, Di Peng, Yang Li, Wentao Zhang, Bin Cui. 2023. Rover: An Online Spark SQL Tuning Service via Generalized Transfer Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, August 6–10, 2023, Long Beach, CA, USA. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3580305.3599953>

1 INTRODUCTION

Big data query systems, like Hive [47], Pig [15], Presto [40], and Spark SQL [2], have been extensively applied in industry to efficiently process massive data for downstream business, such as recommendation and advertisement. As a module of Apache Spark [61], Spark SQL inherits the benefits of Spark programming model [60] and provides powerful integration with the Spark ecosystem.

However, it's often challenging to set proper configurations for optimal performance when executing Spark SQL tasks [20]. For example, the parameter `spark.executor.memory` specifies the amount of memory for an executor process. A too large value leads to long garbage collection time while a too small value may cause

* Equal contribution.

out-of-memory errors. Therefore, it's crucial to tune the configurations to achieve satisfactory performance (in terms of memory, runtime, etc.). In this paper, we focus on tuning online Spark SQL tasks. Unlike offline tuning where it is tolerable to simulate various configurations in a non-production cluster [5, 21, 27, 58], each configuration can only be evaluated in real production due to high overhead of offline evaluations or security concerns. Thus, the configuration should be carefully selected to achieve high performance quickly (efficient) and low risks of negative results (safe).

Recent studies [1, 10, 56] apply the Bayesian optimization (BO) framework [6, 24, 32, 41] to reduce the required number of evaluations to find a near-optimal configuration. In brief, BO trains a surrogate on evaluated configurations and their performance, and then selects the next configuration by balancing exploration and exploitation. However, BO suffers from the re-optimization issue [31], i.e., BO always starts from scratch when tuning a new task. For periodic tasks that only execute once a day, it may take months until BO finds a good configuration, which makes it impractical to deploy in industry. To accelerate this tuning process, the TLBO (transfer learning for BO) community studies to integrate external knowledge into the search algorithm. But for online Spark SQL tuning, we notice some domain-specific challenges and describe them as two questions below,

C1. What to transfer? Among literature of the TLBO community, almost all methods [13, 17, 30, 31, 35] consider utilizing the tuning history of previous tasks, which is suitable for most scenarios. While for Spark SQL scenarios, we notice that domain-specific knowledge can also be utilized to further improve the tuning performance, e.g., parameter adjustment suggestions from experienced engineers. However, how to translate the knowledge to code and how to combine them with the automatic search algorithm are still open questions.

C2. How to transfer? Most previous work [13, 31, 46, 55, 57] assume that similar tasks always exist in history and consider only one or all tasks (often less than 20). However, in real production, history tasks accumulate over time, and most of them may be dissimilar to the current task. Utilizing dissimilar tasks may lead to deteriorated tuning performance (i.e., negative transfer), which is especially intolerable in online tuning. Therefore, it's also essential to identify similar history tasks beforehand and design a proper solution to combine the knowledge from various similar tasks.

In this paper, we propose Rover, a deployed online Spark SQL tuning service that aims at efficient and safe search on in-production workloads. To address the re-optimization issue with the aforementioned challenges, we propose generalized transfer learning to boost the tuning performance based on external knowledge. For the “what to transfer” question, we first apply expert knowledge to design a compact search space and then integrate expert knowledge into the automatic search algorithm to boost the performance in both the initialization and search phases. For the “how to transfer” question, we design a regression model for history filtering to avoid negative results and build a strong ensemble to enhance the final performance. Our contributions are summarized as follows,

1. We introduce Rover, an online Spark SQL tuning service that provides user-friendly interfaces and performs an efficient and safe search on given workloads. So far, Rover has already been deployed

in ByteDance big data development platform and is responsible for tuning 10k+ online Spark SQL tasks simultaneously.

2. To address the two challenges in algorithm design, we integrate expert knowledge into the automatic search algorithm (**C1**). In addition, we enhance the final performance and avoid negative results by filtering dissimilar history tasks and building a strong ensemble (**C2**).

3. Extensive experiments on public benchmarks and real-world tasks show that Rover clearly outperforms state-of-the-art tuning frameworks used for Spark and databases. Notably, compared with configurations given by Spark engineers, Rover saves an average of 50.1% of the memory cost on 12k real-world Spark SQL tasks within 20 iterations, among which 76.2% of the tasks achieve a significant memory reduction of over 60%.

2 RELATED WORK

Spark SQL [2] is a widely applied module in Apache Spark [61] for high-performance structured data processing. Despite its comprehensive functionality, the performance of a Spark SQL application is controlled by more than 200 parameters. The Spark official website [45], as well as various industry vendors such as Cloudera [49] and DZone [48], provide heuristic instructions to tune those parameters, and previous work [28, 42, 43, 51, 52, 59] also propose to optimize heuristically designed cost models instead of the real objectives. However, these methods still require users of a thorough understanding of the system mechanism.

To reduce the barrier of expert knowledge and further improve the heuristically tuned results, previous methods [4, 8, 18, 37, 58, 67] propose to train a performance model based on a large number of evaluations using an offline cluster. However, they incur a large computational overhead and can not adapt to potential data change in online scenarios.

Recently, a variety of methods propose advanced designs of the tuning framework, which consequently reduce the number of evaluations to find satisfactory configurations and can be tentatively applied to online tuning. Some methods combine the performance model with the genetic algorithm to derive better configurations. RFHOC [5] uses several random forests to model each task and applies the genetic algorithm to suggest configurations. DAC [58] combines hierarchical regression tree models with the genetic algorithm and applies the datasize-aware design to capture workload change. Other methods leverage the Bayesian Optimization (BO) framework [6, 24, 44] and achieve state-of-the-art performance in Spark tuning. CherryPick [1] directly performs BO on a discretized search space. LOCAT [56] further combines dynamic sensitivity analysis and datasize-aware Gaussian process (GP) to perform optimization on important parameters. Despite the competitive converged results, the aforementioned methods suffer from the re-optimization issue [31], which is, the performance model needs retraining and still requires a number of online configuration evaluations for each coming task. To alleviate the issue, Tuneful [10, 36] is a pioneering work that attempts to apply a multi-task GP [46] to utilize the most similar previous task in Spark tuning. However, even considering literature from the TLBO (transfer learning for BO) community, most work [3, 13, 14, 30, 31] only consider history tasks as useful knowledge to transfer. In addition, those methods assume that there

always exists similar history tasks for new tasks. A mechanism to pre-select similar tasks and prevent negative transfer when similar tasks are absent is also required in real production.

In the database community, many ML-based methods [26, 29, 34, 50, 62, 64] have been proposed to search for the best database knobs. While sharing a similar spirit, some of the methods can be directly applied in online Spark tuning. OtterTune [50] and ResTune [64] adopt the BO framework, where ResTune integrates the knowledge of all history tasks to further improve its performance. Due to the difference in research fields, in this paper, we mainly focus on tuning parameters for Spark SQL tasks.

3 OVERVIEW

In this section, we introduce the overview of Rover, and then provide preliminaries of the algorithm design, which are problem definition, Bayesian optimization framework and expert rule trees.

3.1 System Overview

Figure 1 shows the overview of Rover. The framework contains the following four components: 1) the Rover controller interacts with end users and the cloud platform, and controls the tuning process; 2) the knowledge database stores the configuration observations and context information on different Spark SQL tasks, and the expert knowledge concluded from previous tasks; 3) the surrogate model fits the relationship between configurations and observed performance; 4) the configuration generator suggests a promising configuration for the tuning task during each iteration.

To start a tuning task, the end user first specifies the tuning objective and uploads the workload to the controller. Then the iterative workload starts in Rover. During each iteration, the **Rover controller** updates a new configuration to the cloud platform and fetches the specified objective and context information after the workload finishes. The information is sent to the **knowledge database**, where a database stores all the observations and information of previous tuning tasks. Based on expert knowledge and task history, a **surrogate model** is built to fit the relationship between configurations and objectives. The **configuration generator** then samples a new configuration to the Rover controller based on the prediction of the surrogate model. Note that, all configurations are run in the real production environment, which means Rover does not require additional budgets for offline tuning as in previous work [4, 37, 58]. When the tuning task ends, the cloud platform applies the best configuration found during optimization. Rover also provides a set of user interfaces, where the end users can monitor the runtime status and make further decisions on whether to early stop a task or increase the tuning budget.

3.2 Problem Definition

The Spark configuration has a great impact on the performance of a Spark SQL task [1, 10]. A configuration refers to a set of parameters with a certain choice of value. Given a Spark SQL tuning task, the goal of Rover is to find a near-optimal Spark configuration that minimizes a pre-defined objective function. The problem definition is as follows,

$$x^* = \arg \min_{x \in \mathcal{X}} \mathcal{L}(x), \quad (1)$$

Algorithm 1: Pseudo code for Bayesian optimization

Input: the search budget \mathcal{B} , the configuration search space \mathcal{X} , the initial configurations X_{init} .
Output: the best observed Spark configuration.

- 1: evaluate the initial configurations and initialize observations $D = \{(x_{init}, y_{init}) | x_{init} \in X_{init}\}$.
- 2: **while** budget \mathcal{B} does not exhaust **do**
- 3: build the surrogate M based on observations D .
- 4: select the next configuration x_n that maximizes the acquisition function $a(x; M)$.
- 5: evaluate the selected configuration and obtain its performance y_n .
- 6: augment $D = D \cup (x_n, y_n)$.
- 7: **end while**
- 8: **return** the best observed configuration.

where \mathcal{X} is the Spark configuration search space and \mathcal{L} is the optimization target. For targets that need to be maximized, we simply use its negative form in Equation 1. Unlike some previous methods [1] where \mathcal{L} is a fixed target, Rover supports multiple optimization targets, e.g., memory, CPU utilization, and complex targets like costs (runtime \times prices per hours).

The Spark SQL tuning tasks are typically black-box optimization problems, which means the actual objective value $\mathcal{L}(x)$ of a configuration x can not be obtained unless we run it in the real cluster. A straightforward solution is to try a sufficient number of configurations and apply the best one. However, in online industrial scenarios, it's impossible to evaluate a large number of Spark configurations due to the limited budget. To reduce the number of evaluated Spark configurations perform an efficient online search, Rover adopts the Bayesian optimization framework, which will be introduced in the following section.

3.3 Bayesian optimization framework

Bayesian optimization (BO) is a popular optimization framework designed to solve black-box problems with expensive evaluation costs. The main advantage of BO is that it estimates the uncertainty of unseen configurations and balances exploration and exploitation when selecting the next configuration to evaluate. BO follows the framework of sequential model-based optimization, which loops over the following three steps as shown in Algorithm 1 : 1) BO fits a probabilistic surrogate model M based on the observations $D = \{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$, in which x_i is the configuration evaluated in the i^{th} iteration and y_i is its corresponding observed performance; 2) BO uses the surrogate M to select the most promising configuration x_n by maximizing $x_n = \arg \max_{x \in \mathcal{X}} a(x; M)$, where $a(x; M)$ is the acquisition function designed to balance the trade-off between exploration and exploitation; 3) BO evaluates the configuration x_n to obtain y_n (i.e., evaluate the Spark configuration in the real cluster and obtains the objective value), and augment the observations by $D = D \cup \{(x_n, y_n)\}$.

Surrogate. Following popular implementations in the BO community, Rover applies the Gaussian Process (GP) as the surrogate. Compared with other alternatives, Gaussian Process is parameter-free and computes the posterior distributions in closed forms. Given an unseen configuration x , GP outputs a posterior marginal Gaussian

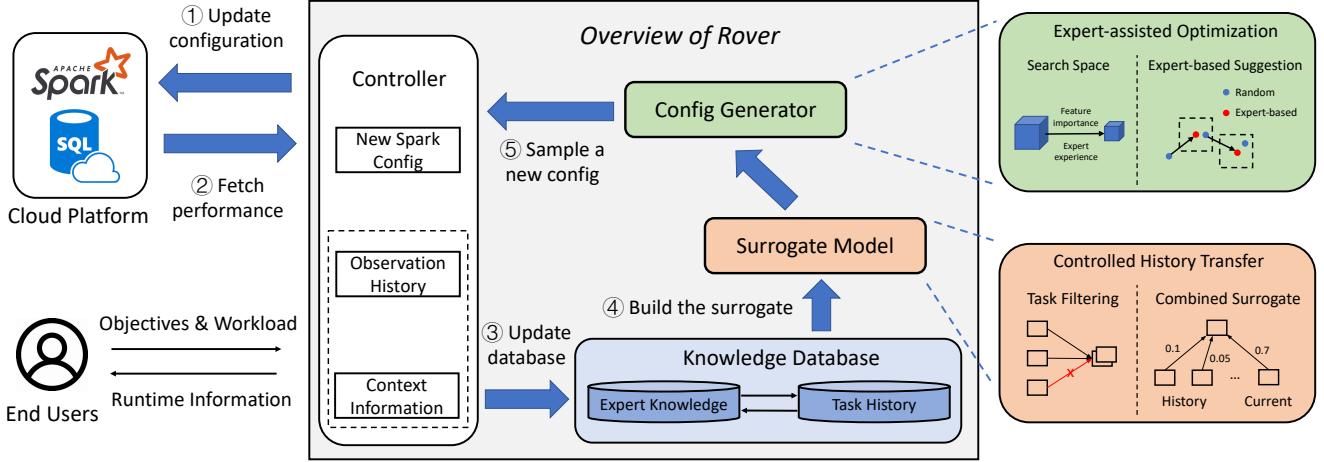


Figure 1: Rover architecture and its interaction with Spark and users.

distribution, whose predictive mean and variance are formulated as follows,

$$\begin{aligned}\mu(x) &= K(X, x)(K(X, X) + \tau^2 I)^{-1}Y, \\ \sigma^2(x) &= K(x, x) + \tau^2 I - (K(X, X) + \tau^2 I)^{-1}K(X, x),\end{aligned}\quad (2)$$

where K is the covariance matrix, X are the observed configurations, Y are the observed performance of X , and τ^2 is the level of white noise. In practice, we use the Matern 5/2 kernel when computing covariance in Gaussian Process.

Acquisition Function. The acquisition function determines the most promising configuration to evaluate in the next iteration. We use the Expected Improvement (EI) function, which measures the expected improvement of a configuration over the best-observed performance, which is formulated as,

$$EI(\mathbf{x}) = \int_{-\infty}^{\infty} \max(y^* - y, 0)p_M(y|\mathbf{x})dy,\quad (3)$$

where y^* is the best performance observed so far.

Context-aware Design. In online scenarios, the running environment (context) of a tuning task often changes due to data shifts or other processes running in the shared cluster. While the optimal configuration slightly differs according to the running environment, we follow the surrogate implementation in previous work [64], which also takes the data size and context vector from SparkEventLog as inputs. Similar to the configuration vector, we also use the Matern 5/2 kernel for the context vector. With this extension, the surrogate can deal with dynamic workloads in online scenarios.

3.4 Expert Rule Trees

Spark experts often adjust the values of specific Spark parameters when the observed performance meets a given condition. To simulate this behavior and translate the knowledge to code, Rover defines a set of actions, which we refer to as expert rule trees. An expert rule includes the following five parts: 1) Parameter name, which determines the Spark parameter to adjust; 2) Direction, which

shows whether to increase or decrease the value of the parameter if a change is needed; 3) Condition, which decides whether to adjust the parameter; 4) Step, which decides how much to change the parameter if a change is needed; 5) Bounds, which ensures that the parameter will not exceed limits for safety consideration.

We take a practical expert rule as an example. To run a Spark SQL task using the MapReduce programming structure, it's essential to decide the number of mappers to process the input data. If the number is set too large, the execution of a mapper finishes too quickly, which may lead to high time overhead to start and delete the mappers, and high memory overhead to maintain those mappers. To prevent creating too many mappers, experts can reduce the number of mappers by increasing the value of `maxPartitionBytes`. An expert rule example can be: Double (`Direction` & `Step`) its value (`Parameter`) if the average execution time of mappers is less than 12 seconds (`Condition`). The minimal and maximal bounds of the value are 16M and 4G, respectively (`Bounds`). In Rover, we design 47 rules in total. Please refer to the Appendix A.5 for more details.

4 ALGORITHM DESIGN

As mentioned in Section 1, Rover designs a generalized transfer learning framework for Spark SQL tuning tasks. To answer the "what to transfer" question, we will first introduce expert-assisted Bayesian optimization to show how Rover integrates expert knowledge in each part of the Bayesian optimization algorithm. And then, we will answer the "how to transfer" question by introducing how we improve previous transfer learning methods from knowledge filtering and combination via controlled history transfer.

4.1 Expert-assisted Bayesian Optimization

While Bayesian optimization requires random initialization to accumulate sufficient observations to fit a surrogate model, the performance may be far from satisfactory at the beginning of the optimization in online tuning scenarios. Meanwhile, previous methods

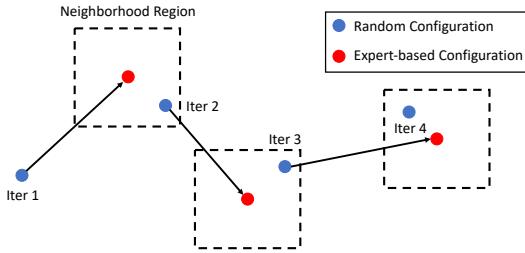


Figure 2: An example of expert-based random initialization. Three configurations are randomly sampled from the neighborhood regions of expert-based configurations one by one.

have shown performance improvement in early iterations by using deliberately designed heuristic rules in Spark tuning scenarios[45]. However, expert knowledge often finds a moderate configuration quickly, but its performance is less promising than BO given a larger search budget. Then, it seems natural and intuitive to speed up Spark SQL tuning by combining the advantages of both Bayesian optimization and expert knowledge, i.e., achieving promising results with fewer evaluations. By analyzing the BO framework, Rover proposes to strengthen the BO framework with heuristic expert knowledge in the following three parts:

4.1.1 Compact Search Space. The size of the search space greatly influences the optimization, as adding a dimension in the search space leads to an exponential growth of possible Spark configurations. To reduce the difficulty of fitting a BO surrogate, Rover tunes a Spark SQL task using a heuristic search space with fixed parameters. Concretely, we collect historical tasks that run on the entire search space, and then use SHAP [33], a toolkit for measuring feature importance, to rank the importance of each Spark parameter over the observed performance. The final search space used in online tuning includes 10 parameters, which are the top-5 important five parameters given by SHAP and 5 extra important parameters selected by Spark SQL experts. The detailed parameters are provided in the Appendix A.4.

4.1.2 Expert-based Random Initialization. While Bayesian optimization uses random start points for initialization, Rover improves this step by integrating the expert rule trees. An example of the initialization phase is shown in Figure 2. During each initialization round, Rover first applies the expert rule trees on the previous configuration and obtains the updated one (the red point). Then, a new configuration (the blue point) is randomly chosen from the neighborhood region of the updated configuration. In practice, the neighborhood region is set to be a region of configurations whose difference is less than $\pm 20\%$ in each parameter.

The initialization design combines both expert rules and random initialization, which has the following advantages: 1) Compared with vanilla random initialization, expert rules tend to find configurations with better performance in early iterations; 2) Compared with using expert rules alone, the random design ensures that each Spark parameter has different values among configurations. Since it is quite often that a Spark parameter dissatisfies all corresponding expert rules so that its value does not change, the random design at

least changes the value of all parameters, which helps the BO surrogate to measure their importance easier; 3) Compared with random initialization, the neighborhood region prevents Rover from evaluating potentially bad configurations sampled in the entire search space, which ensures safe optimization in online scenarios.

4.1.3 Expert-based Configuration Generator. Previous work reports that BO may suffer from over-exploitation in later iterations, i.e., the configurations are generated from a quite small region. To address this issue, mature BO framework [12, 25] often adopts the ϵ -greedy strategy, where there is a probability of ϵ that the configuration is suggested randomly instead of using the acquisition function. Meanwhile, right after the initialization phase, the BO surrogate may not fit well based on scarce observations for initialization, which consequently leads to an unpromising configuration suggested by the BO acquisition function. To address both two issues, we set a dynamic possibility that the expert rule trees generate the next configuration instead of the BO acquisition function. The intuition is that, if the BO surrogate performs worse, the expert rule trees will be applied more frequently.

Concretely, Rover assigns two weights to the expert rule trees and BO surrogate to control the probability of applying each component. For expert rule trees, the weight is decreasing and is only controlled by the number of iterations. The weight is formulated as $w_e = 0.5^T + 0.2$, where T is the current number of iterations. This matches the intuition that 1) the expert rule trees should be applied less frequently when the BO surrogate generalizes better in later iterations, and 2) there's always a probability that the expert rules are applied to avoid over-exploitation in Bayesian optimization.

For the BO surrogate, to measure its generalization ability, we define the weight as the ratio of **concordant pairs**. Given two configurations x_1 and x_2 , the pair is concordant if the sort order of $(M(x_1), M(x_2))$ and (y_1, y_2) agrees, where $M(x_1)$ is the predictive mean of x_1 by the BO surrogate and y_1 is the ground-truth performance of x_1 . Since the BO surrogate is directly trained on the observations, using observations again cannot reflect its generalization ability on unseen configurations. Therefore, we apply cross-validation and the weight is then calculated as $w_s = \frac{2}{|D|(|D|-1)} n_s$, where D is the current observations, and n_s is the number of concordant pairs. Denote f as the mapping function that maps an observed configuration x_i to its corresponding fold index as $f(x_i)$. n_s is calculated as,

$$n_s = \sum_{j=1}^{|D|} \sum_{k=j+1}^{|D|} \mathbb{1}((M_{-f(x_j)}(x_j) < M_{-f(x_k)}(x_k)) \otimes (y_j < y_k)) \quad (4)$$

where \otimes is the exclusive-nor operator, and $M_{-f(x_j)}$ refers to predictive mean of the surrogate trained on observations D with the $f(x_j)$ -th fold left out. In this way, x_j is not used when generating $M_{-f(x_j)}$, thus the definition is able to measure the generalization ability of the BO surrogate only by using the observations D . Finally, during each iteration, there's a probability of $p_e = w_e/(w_e+w_s)$ that Rover applies the expert rule trees to select the next configuration instead of the BO acquisition function.

4.2 Controlled History Transfer

In this subsection, we explain when and how Rover transfers knowledge from history tasks, i.e., previously evaluated tasks. As mentioned above, history knowledge may not be properly utilized in previous methods by simply transferring all tasks or the most similar one. In the following, we will answer two questions: 1) how to decide the set of history tasks which is potentially beneficial to the current task and 2) how to combine the knowledge of history tasks when optimizing the current task.

4.2.1 Task Filtering. In industrial scenarios, as the tuning requirements arrive and end continuously, the history database accumulates the tuning history of a large number of history tasks. While it is almost impossible to ensure that all history tasks are similar to the current task, it is essential to filter the history tasks before optimization to ensure safe transfer.

Previous work [38] proposes an intuitive method that generates a meta-feature for each task based on the outputs from SparkEventLog and defines the distance between two tasks as the Euclidean distance between two vectors. Then, the most similar history task with the lowest distance is selected to aid the optimization of the current task. However, the meta-features are often of significantly different scales, where the Euclidean distance is dominated by those dimensions with large value ranges. In addition, more complex relationships may hide in the meta-feature space, e.g., tasks with quite different meta-features may also be similar in tuning behavior.

To address the above two issues, we propose to apply a regression model to learn the similarity between two given tuning tasks. Following [38], we vectorize each task based on 17 runtime metrics in the event log. The regression model takes two task meta-features as inputs and outputs their similarity, which is $(v_1, v_2) \mapsto s$, where $s \in [0, 1]$ and a larger s indicates the two tasks are more similar.

To train the regression model, we first generate training data based on history tasks. Denote the BO surrogate built on the i^{th} task as M^i . Similar to Section 4.1.3, for each given pair of history tasks i and j , we define the similarity as the ratio of **concordant prediction pairs**. Given two configurations x_1 and x_2 , the prediction pair is concordant if the sort order of $(M^i(x_1), M^i(x_2))$ and $(M^j(x_1), M^j(x_2))$ agrees, where $M^i(x_1)$ is the predictive mean of x_1 by the surrogate M^i . The similarity $S(i, j)$ between the i^{th} and j^{th} task can be computed as,

$$\begin{aligned} S(i, j) &= \frac{2}{|D_r|(|D_r| - 1)} F(i, j), \\ F(i, j) &= \sum_{k=1}^{|D_r|} \sum_{l=k+1}^{|D_r|} \mathbb{1} \left(\left(M^i(x_k) < M^i(x_l) \right) \otimes \left(M^j(x_k) < M^j(x_l) \right) \right), \end{aligned} \quad (5)$$

where D_r is a set of randomly selected configurations on the search space, and \otimes is the exclusive-nor operator. While $F(i, j)$ computes the number of concordant prediction pairs, we divide $F(i, j)$ by the number of pairs and scale it to $[0, 1]$. For each pair of history tasks i and j , the training sample is a triple represented as $< v_i, v_j, S(i, j) >$. To ensure symmetry, $< v_j, v_i, S(i, j) >$ is also included in the training data. Consider a history database with N history tasks, about N^2 training samples are generated. We use a Catboost regressor [9] as the machine learning model.

After training the regressor, given a new tuning task, we first generate the meta-feature of the new task and compute its similarity with all history tasks using the regressor. Then we filter the tasks using a threshold so that most of the irrelevant tasks are ignored. In practice, we set a quite high threshold (0.65) to perform conservative transfer learning and avoid negative transfer caused by dissimilar history tasks. To reduce the overhead of computation, we choose the top-5 tasks with the largest similarity as history knowledge if there are a large number of valid history tasks. Note that, if no similar tasks are found, history transfer is disabled and the algorithm reverts to expert-assisted BO as introduced in Section 4.1.

4.2.2 Knowledge Combination. Given a set of similar history tasks, the next step is to combine those tuning knowledge with the BO framework. While a simple warm-starting does not fully utilize the potential of history [12], previous methods [13, 64] propose to directly combine the predictions of different task surrogates. However, the scale and distribution of surrogate outputs across different tasks are often different, e.g., the memory cost given the same configuration depends on the data size and running environment of a workload. Then, the combined surrogate output in previous methods will be dominated by the surrogate where the output values are larger. To alleviate this issue, we propose to combine the ranking output of different surrogates, as the tuning procedure only aims to find the optimum in the search space rather than getting its value.

Concretely, given a set of candidate configurations by random sampling and mutation [24], we rank the configurations based on the EI value of each task surrogate. We denote the ranking value of the configuration x_i based on surrogate M^j on the j^{th} task as $R_{M^j}(x_i)$. The combined ranking of x_i is computed as,

$$CR(x_i) = \sum_{j=1}^{K+1} w_j R_{M^j}(x_i), \quad (6)$$

where we assume that there are K history tasks available and the current task is the $K + 1^{th}$ task. w_j controls the importance of each task, which is defined as,

$$\begin{aligned} w_j &= \frac{\exp(G(j)/\tau)}{\sum \exp(G(\cdot)/\tau)}, \text{ where } \tau = \frac{\tau_0}{1 + \log T}, \\ G(j) &= \sum_{k=1}^{|D|} \sum_{l=k+1}^{|D|} \mathbb{1} \left(\left(M^j(x_k) < M^j(x_l) \right) \otimes \left(y_k < y_l \right) \right), \end{aligned} \quad (7)$$

where D is the current observations, τ_0 is a hyper-parameter that controls the softmax temperature τ , and T is the current number of BO iterations. $G(j)$ measures whether the surrogate M^j fits the ground-truth objective values y . $G(K + 1)$ is obtained by cross-validation similar to Equation 4. In each iteration, Rover selects the configuration with the lowest ranking value to evaluate. To compute the weight (Section 4.1.3) for the combined surrogate when history transfer is enabled, we simply take $w_s = \sum_j w_j \frac{2*G(j)}{|D|(|D|-1)}$. The tuning procedure will be discussed in Section 5.

5 SYSTEM IMPLEMENTATION

To integrate the designs introduced in Section 4, the tuning procedure of Rover is split into two phases: the initialization and the search phase. To make a better understanding of the tuning procedure, we also provide the flow chart of both two phases in Figure 3. The pseudo-code is provided in Appendix A.1.

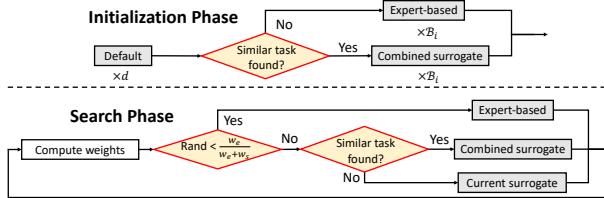


Figure 3: The flow chart of the tuning procedure.

Initialization Phase. During the initialization phase, Rover first evaluates 5 default configurations. Then, it obtains the meta-feature from the running log and selects similar tasks from the history database. If no similar tasks are found, Rover applies the expert-based initialization to generate configurations. Else, Rover applies the combined surrogate instead until the initialization budget exhausts. **Search Phase.** During the search phase, Rover first computes the weight w_s for BO depending on whether there are similar history tasks and w_e for expert rule trees. Then, based on random sampling, there's a probability that the next configuration is suggested by expert knowledge. Else, Rover samples a new configuration by optimizing the EI acquisition function or the combined ranking function when history transfer is disabled and enabled, respectively. **Special Cases.** When the search budget \mathcal{B}_s exhausts, Rover stops the tuning procedure and returns the best-observed configuration. By monitoring the running information from the user interface, the end users can also early stop a tuning task if they find there's no significant improvement as expected. In addition, as the workload data may change in online scenarios, if Rover detects a significant workload change, it will consider it as a new tuning task and restarts from the initialization phase. The demonstration of Rover functionality and user interfaces is provided in Appendix A.3.

6 EXPERIMENTAL EVALUATION

To show the practicality and effectiveness of Rover, we will investigate the following insights: 1) Compared with state-of-the-art tuning algorithms, Rover achieves superior performance on benchmarks and real-world tasks; 2) Rover achieves positive performance in production; 3) The algorithm framework of Rover is more reasonable compared with other design alternatives.

6.1 Setups

Benchmarks. For end-to-end comparison, we follow LOCAT [56] and use three SQL-related tasks from the widely used Spark benchmark HiBench [23]: (1) ‘Join’ is a query that executes in two phases: Map and Reduce; (2) ‘Scan’ is a query that only executes Map operations; (3) ‘Aggregation’ is a query that executes the Map and Reduce operation alternately. In addition, we also prepare 200 relatively small workloads that run internally in ByteDance, which are collected following data privacy rules. The input sizes of workloads range from 200MB to 487GB, and the workloads involve comprehensive combinations of Map and Reduce operations.

Baselines. For end-to-end comparison, we compare Rover with the following SOTA BO-based tuning algorithms: (1) CherryPick [1]: A method with a discretized search space; (2) Tuneful [10]: A method that explores significant parameters and applies a multi-task GP to use the most similar previous task; (3) LOCAT [56]: A method

that identifies important parameters and dynamically reduces the search space; (4) ResTune [64]: A transfer learning method that uses all the history knowledge to accelerate the tuning process.

Experiment Settings. Rover implements BO based on OpenBox [32], a toolkit for black-box optimization. The other baselines are implemented following their original papers. Rover uses the 10-d search space (Section 4.1.1), while the space for the other baselines follows the 30-d search space in Tuneful [10]. For transfer learning, Rover takes the first 25 iterations of previous tasks as history knowledge. The meta-feature for each task is computed as the average runtime vectors of default configurations (Section 4.2.1). In all experiments, we optimize the **accumulated memory cost**, which is computed as the average memory cost multiplied by the time cost (GB·h). We use the ratio relative to the average result of default configurations as the metric so that the result on each task is of the same scale. In Section 6.4, we report the mean **best-observed** result during each iteration with 5 repetitions. The total tuning budget for each method is set to 50 iterations (5 for initialization and 45 for search).

Environment. All the experiments are conducted using a virtual cluster of 24 nodes. Each node is equipped with 4 AMD EPYC 7742 64-core processors and 2T memory. We use Spark 3.2 as our computing framework.

6.2 End-to-end Comparison

We first provide an end-to-end comparison on three HiBench tasks and the internal tasks. In Figure 4, we show the memory reduction ratio after tuning 10 and 50 iterations relative to the default configuration. We observe that: 1) Among Spark tuning baselines, CherryPick does not reduce the search space, thus it can not handle the tasks well. Though LOCAT selects important parameters dynamically during the tuning process, it requires about 15 iterations before it can shrink the space. Its performance is worse than Rover where a compact space is used from the beginning; 2) Tuneful and Restune generally outperforms LOCAT in 10 iterations due to the use of similar history tasks. However, as they do not utilize history well, there's almost no improvement in 50 iterations; 3) The performance of Rover is not significant on Scan, where no similar tasks are found among 202 tasks. Even though transfer learning is not enabled, Rover still outperforms other baselines in early and late iterations. 4) Among the competitive baselines, Rover achieves the largest memory reduction on all the tasks. Concretely, on internal tasks, Rover further reduces **9.36%** and **6.76%** the memory cost relative to the baselines Tuneful and LOCAT, respectively. We also provide results on another objective (CPU costs) in Appendix A.6.

6.3 In-production Performance

To demonstrate the practicality of Rover, we deploy the system in a production cluster and evaluate it on 12000 real-world Spark SQL tasks. Each task is collected from actual business in ByteDance (e.g., recommendation, advertisement, etc.), which executes once a day and processes massive data generated from billions of users every day. The input sizes of workloads range from 200MB to 37.3TB. Since the evaluation runs in actual business and is computationally expensive, we only compare Rover with default Spark configurations suggested by Spark engineers. The tuning budget for all tasks is set to 20 days (tuning once per day).

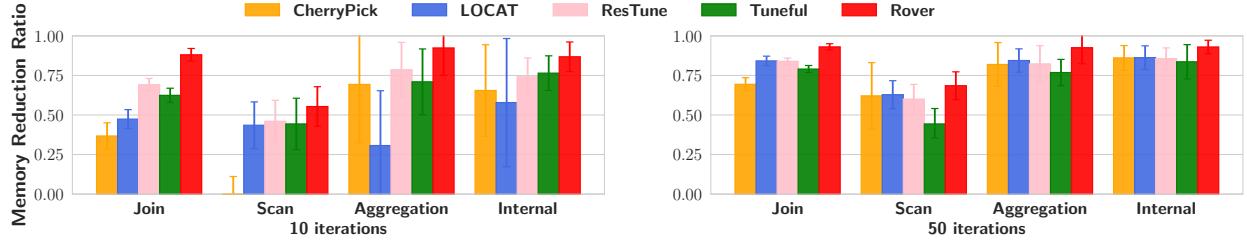


Figure 4: Memory reduction ratio (the higher, the better) with standard deviation on 3 HiBench tasks and internal tasks.

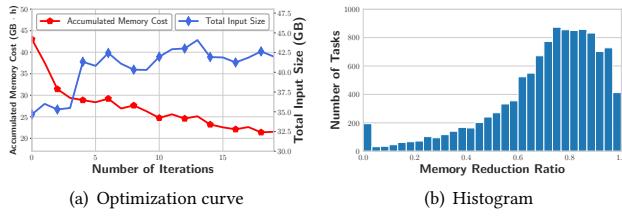


Figure 5: In-production performance of Rover.

Figure 5(a) shows the mean accumulated memory cost during optimization. We observe that Rover reduces the memory cost by 32.8% in the initialization phase due to expert-based initialization and controlled history transfer. After that, though the online workload gradually changes, the memory cost continuously reduces with the help of the context-aware surrogate. The improvement is not as significant as Section 6.2 because the in-production workloads are changing frequently and the default configurations for some tasks have been already tuned well by experts before optimization. Figure 5(b) further shows the number of tuning tasks with different improvements, where 76.2% of the tasks get a significant memory reduction of over 60%. Moreover, 97.7% of the tasks get an improvement of over 10% compared with the default configurations. With the above improvement within 20 iterations, Rover has already saved about **\$1.1M** of the annual computing resource expense for the above 12k tasks. While Rover currently tunes a small proportion of tasks on the platform, it is estimated that Rover may lead to a significant expense reduction when we apply it to more and larger tasks (over 320k) in near future.

6.4 Detailed Analysis

In this part, we provide a detailed analysis of the effects of both two algorithm designs and ablation study on the algorithm framework.

6.4.1 Analysis of Expert-assisted BO. In this part, we will analyze the performance of the pre-defined search space and expert-based optimization. We first run Bayesian optimization over three different search spaces: 1) Rover space (10 params); 2) the Tuneful [10] space (30 params); 3) the Tuneful space with dynamic shrinking used in LOCAT [56]. The results are shown in Figure 7(a). We observe that the Rover space achieves clear improvement over the Tuneful space due to the use of important parameters selected by experts. In addition, though LOCAT performs better than Tuneful, dynamic shrinking requires a number of evaluations and performs

worse than BO that starts from a more compact space. The observations show that the compact space is more practical than strategies used in state-of-the-art frameworks.

Then, we directly compare expert-based BO with other baselines using the same space in Figure 7(b). Due to initialization with expert knowledge, expert-assisted BO and expert rules achieve significant improvement over vanilla BO, LOCAT and Tuneful in the initialization phase. After that, the memory cost of expert-assisted BO gradually decreases and consistently outperforms other baselines. In general, at the 10th iteration, expert-assisted BO reduces the memory cost of vanilla BO, Tuneful, and LOCAT by 2.96%, 1.38% and 7.00%. And at the 50th iteration, expert-assisted BO further reduces the memory cost of vanilla BO, Tuneful, and LOCAT by 0.74%, 3.05% and 3.80%. The reason that Tuneful and LOCAT performs worse than vanilla BO in our scenarios may be that 1) multi-task BO used in Tuneful may not be effective to transfer history knowledge; 2) space shrinking in LOCAT may discard important parameters when the space is already compact. We also compute the average memory ratio of all configurations for expert-assisted BO and vanilla BO, which are 37.07% and 45.43%. Vanilla BO leads to higher average cost because of unsafe configurations suggested by random initialization and ϵ -greedy strategy. The observations show that Rover achieves better results and performs safer than vanilla BO and using expert rules alone.

6.4.2 Analysis of Controller History Transfer. We first treat 25 from 200 tasks as the test set and analyze the generalization ability of the regression model. In Figure 6(a), we plot the mean squared error (MSE) of three regression models using a different number of history tasks. Generally, we observe that Catboost achieves the best performance compared with the other two alternatives, thus we use Catboost as the regressor. In addition, the MSE continuously drops when the number of history tasks used for training increases. When the number reaches 175, we also take one task from the test set and plot the predicted and ground-truth similarity in Figure 6(b). The Spearman correlation between predictions and ground truths is 0.889, while the correlation between Euclidean distances and ground truths is only 0.120. This shows that the regression model can measure the similarity of unseen tasks relatively well.

Then, we compare the controlled history transfer in Rover with other intuitive solutions: 1) One: transfer only the most similar task by the model; 2) All [64]: transfer all history tasks. Since Rover only performs transfer learning when similar tasks are found, we plot the results on those tasks (43 out of 200 tasks) in Figure 6(c). We observe

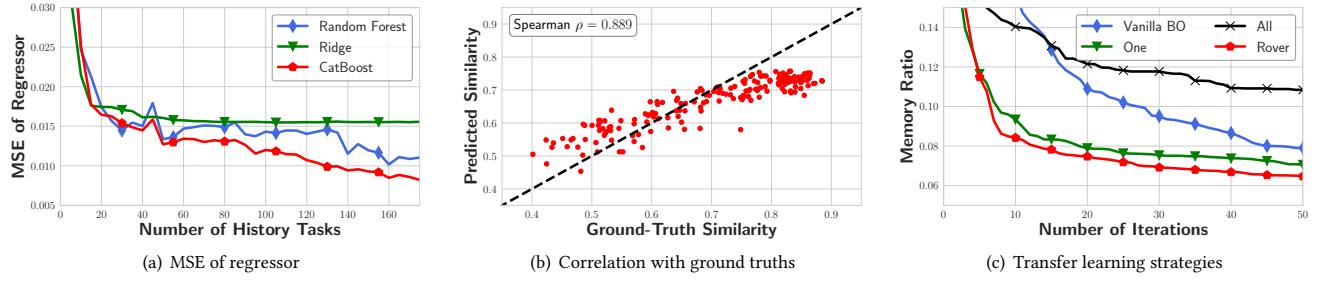


Figure 6: Analysis of controlled history transfer.

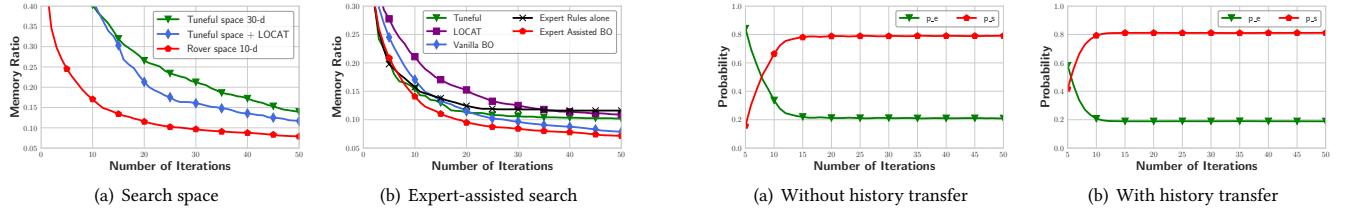


Figure 7: Analysis of expert-assisted BO.

that **All** performs worse than vanilla BO at later iterations due to the use of too many dissimilar tasks in history. In addition, the mean result of **Rover** is slightly better than **One**, but the variance of **Rover** is much lower (0.013 compared with 0.021 at the 50th iteration). We attribute this to the use of more than one similar tasks, and the ensemble is known to produce better results and reduce variance [7, 65]. In addition, we also compare **One** with vanilla BO at the 10th iteration on the 157 tasks when transfer learning is **disabled** by **Rover**. **One** performs worse than BO on 23% of them, while the value is only 2 out of 43 tasks when **Rover** decides to use history transfer. This shows that task filtering is essential to avoid negative transfer in production.

Finally, since history transfer is only enabled when similar tasks are found, we perform an ablation study on those tasks. Table 1 shows the performance gap of **Rover** without certain components relative to **Rover**. We observe that, controlled history transfer further improves expert-based optimization about 8.42% and 2.86% in the first 5 and 15 iterations. In addition, without history transfer, **Rover** still outperforms vanilla BO due to expert-based designs.

Table 1: Reduction ratio and gaps (%) at different iterations relative to **Rover** on tasks where history transfer is enabled.

	5	Δ	15	Δ	50	Δ
Vanilla BO on compact space	24.39	13.15	12.78	5.04	7.86	1.43
Rover w/o controlled history transfer	19.66	8.42	10.60	2.86	7.32	0.88
Rover	11.24	/	7.74	/	6.44	/

6.4.3 Analysis of dynamic weights. To show how **Rover** make decisions during each iteration, we show the probability of using expert rules (p_e) or BO (p_s) during optimization on 200 internal tasks in Figure 8, which are computed as $p_e = w_e/(w_e + w_s)$ and $p_s = w_s/(w_e + w_s)$.

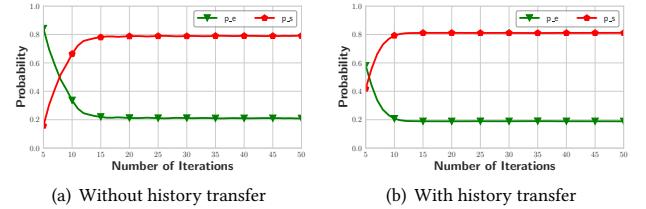


Figure 8: Possibility of components during each iteration.

As mentioned in Section 5, we use expert rules (no similar tasks found) or surrogate ensemble (similar tasks found) in the first 5 iterations, so we display the probability after 5 iterations. If history transfer is disabled, the BO surrogate may be under-fitted with limited observations. Therefore, it takes BO another 3 iterations to get a higher probability than expert rules and it dominates the selection (over a probability of 75%) after the 12th iterations. Expert rules still have a low probability due to the constant bias as mentioned in Section 4.1.3. If similar tasks are found, we find that the probability of BO dominates the optimization process at the 9th iteration, which shows that the surrogate ensemble can model the relationship between configurations and performance relatively well in the beginning.

7 CONCLUSION

In this paper, we propose **Rover**, a deployed online Spark SQL tuning service that provides user-friendly interfaces and performs efficient and safe search on in-production workloads. To address the re-optimization issue and enhance the tuning performance, **Rover** integrates expert knowledge into the search algorithm and build a strong surrogate ensemble on elaborately selected history tasks. Extensive experiments on public benchmarks and real-world tasks show that **Rover** clearly outperforms competitive tuning frameworks used for Spark and databases. Notably, when deployed in ByteDance, **Rover** achieves significant memory reduction on over 12k real-world Spark SQL tasks.

ACKNOWLEDGMENTS

This work is supported by NSFC (No. 61832001 and U22B2037) and ByteDance-PKU joint program. Wentao Zhang and Bin Cui are the corresponding authors.

REFERENCES

- [1] Omid Alipourfard, Hongqiang Harry Liu, Jianshu Chen, Shivaram Venkataraman, Minlan Yu, and Ming Zhang. 2017. {CherryPick}: Adaptively Unearthing the Best Cloud Configurations for Big Data Analytics. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. 469–482.
- [2] Michael Armbrust, Reynold S Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K Bradley, Xiangrui Meng, Tomer Kaftan, Michael J Franklin, Ali Ghodsi, et al. 2015. Spark sql: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 1383–1394.
- [3] Tianyi Bai, Yang Li, Yu Shen, Xinyi Zhang, Wentao Zhang, and Bin Cui. 2023. Transfer Learning for Bayesian Optimization: A Survey. *arXiv preprint arXiv:2302.05927* (2023).
- [4] Liang Bao, Xin Liu, and Weizhao Chen. 2018. Learning-based automatic parameter tuning for big data analytics frameworks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 181–190.
- [5] Zhendong Bei, Zhibin Yu, Huijing Zhang, Wen Xiong, Chengzhong Xu, Lieven Eeckhout, and Shengzhong Feng. 2015. RFHOC: A random-forest approach to auto-tuning hadoop’s configuration. *IEEE Transactions on Parallel and Distributed Systems* 27, 5 (2015), 1470–1483.
- [6] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*. 2546–2554.
- [7] Yijun Bian and Huanhuan Chen. 2021. When does diversity help generalization in classification ensembles? *IEEE Transactions on Cybernetics* (2021).
- [8] Guoli Cheng, Shi Ying, and Bingming Wang. 2021. Tuning configuration of apache spark on public clouds by combining multi-objective optimization and performance prediction model. *Journal of Systems and Software* 180 (2021), 111028.
- [9] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
- [10] Ayat Fekry, Lucian Carata, Thomas Pasquier, Andrew Rice, and Andy Hopper. 2020. Tuneful: An online significance-aware configuration tuner for big data analytics. *arXiv preprint arXiv:2001.08002* (2020).
- [11] Yu-Jing Feng, De-Jian Li, Xu Tan, Xiao-Chun Ye, Dong-Rui Fan, Wen-Ming Li, Da Wang, Hao Zhang, and Zhi-Min Tang. 2022. Accelerating Data Transfer in Dataflow Architectures Through a Look-Ahead Acknowledgment Mechanism. *Journal of Computer Science and Technology* 37, 4 (2022), 942–959.
- [12] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in neural information processing systems*. 2962–2970.
- [13] Matthias Feurer, Benjamin Letham, and Eytan Bakshy. 2018. Scalable meta-learning for bayesian optimization using ranking-weighted gaussian process ensembles. In *AutoML Workshop at ICML*.
- [14] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. 2015. Initializing Bayesian Hyperparameter Optimization via Meta-Learning.. In *AAAI*. 1128–1135.
- [15] Alan F Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M Narayananmurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, and Utkarsh Srivastava. 2009. Building a high-level dataflow system on top of MapReduce: the Pig experience. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1414–1425.
- [16] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. 2014. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607* (2014).
- [17] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. 2017. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1487–1495.
- [18] Anastasios Goumaris, Georgia Kouka, Ruben Tous, Carlos Tripiana Montes, and Jordi Torres. 2017. Dynamic configuration of partitioning in spark applications. *IEEE Transactions on Parallel and Distributed Systems* 28, 7 (2017), 1891–1904.
- [19] Lihua Guo, Dawu Chen, and Kui Jia. 2022. Knowledge transferred adaptive filter pruning for CNN compression and acceleration. *Science China Information Sciences* 65, 12 (2022), 229101.
- [20] Herodotos Herodotou, Yuxing Chen, and Jiaheng Lu. 2020. A survey on automatic parameter tuning for big data processing systems. *ACM Computing Surveys (CSUR)* 53, 2 (2020), 1–37.
- [21] Herodotos Herodotou, Harold Lim, Gang Luo, Nedyalko Borisov, Liang Dong, Fatma Bilgen Cetin, and Shivnath Babu. 2011. Starfish: A self-tuning system for big data analytics.. In *Cidr*, Vol. 11. 261–272.
- [22] Chun Kit Jeffery Hou and Kamran Behdinan. 2022. Dimensionality Reduction in Surrogate Modeling: A Review of Combined Methods. *Data Science and Engineering* (2022), 1–26.
- [23] Shengsheng Huang, Jie Huang, Jinquan Dai, Tao Xie, and Bo Huang. 2010. The HiBench benchmark suite: Characterization of the MapReduce-based data analysis. In *2010 IEEE 26th International conference on data engineering workshops (ICDEW 2010)*. IEEE, 41–51.
- [24] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2011. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*. Springer, 507–523.
- [25] Huaijun Jiang, Yu Shen, and Yang Li. 2021. Automated Hyperparameter Optimization Challenge at CIKM 2021 AnalyticCup. *arXiv preprint arXiv:2111.00513* (2021).
- [26] Mayuresh Kunjir and Shivnath Babu. 2020. Black or White? How to develop an autotuner for memory-based analytics. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1667–1683.
- [27] Palden Lama and Xiaobo Zhou. 2012. Aroma: Automated resource allocation and configuration of mapreduce environment in the cloud. In *Proceedings of the 9th international conference on Autonomic computing*. 63–72.
- [28] Hai Lan, Zhifeng Bao, and Yuwei Peng. 2021. A survey on advancing the dbms query optimizer: Cardinality estimation, cost model, and plan enumeration. *Data Science and Engineering* 6 (2021), 86–101.
- [29] Guoliang Li, Xuanhe Zhou, Shifu Li, and Bo Gao. 2019. Qtune: A query-aware database tuning system with deep reinforcement learning. *Proceedings of the VLDB Endowment* 12, 12 (2019), 2118–2130.
- [30] Yang Li, Yu Shen, Huaijun Jiang, Tianyi Bai, Wentao Zhang, Ce Zhang, and Bin Cui. 2022. Transfer Learning based Search Space Design for Hyperparameter Tuning. *arXiv preprint arXiv:2206.02511* (2022).
- [31] Yang Li, Yu Shen, Huaijun Jiang, Wentao Zhang, Zhi Yang, Ce Zhang, and Bin Cui. 2022. TransBO: Hyperparameter Optimization via Two-Phase Transfer Learning. *arXiv preprint arXiv:2206.02663* (2022).
- [32] Yang Li, Yu Shen, Wentao Zhang, Yuanwei Chen, Huaijun Jiang, Mingchao Liu, Jiawei Jiang, Jinyang Gao, Wentao Wu, Zhi Yang, Ce Zhang, and Bin Cui. 2021. OpenBox: A Generalized Black-box Optimization Service. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021).
- [33] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.
- [34] Lin Ma, Dana Van Aken, Ahmed Hefny, Gustavo Mezerhane, Andrew Pavlo, and Geoffrey J Gordon. 2018. Query-based workload forecasting for self-driving database management systems. In *Proceedings of the 2018 International Conference on Management of Data*. 631–645.
- [35] Valerio Perrone, Rodolphe Jenatton, Matthias W Seeger, and Cédric Archambeau. 2018. Scalable hyperparameter transfer learning. *Advances in neural information processing systems* 31 (2018).
- [36] Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987* (2019).
- [37] Panagiotis Petridis, Anastasios Goumaris, and Jordi Torres. 2016. Spark parameter tuning via trial-and-error. In *INNS Conference on Big Data*. Springer, 226–237.
- [38] David Buchaca Prats, Felipe Albuquerque Portella, Carlos HA Costa, and Josep Lluis Berral. 2020. You only run once: spark auto-tuning from a single run. *IEEE Transactions on Network and Service Management* 17, 4 (2020), 2039–2051.
- [39] Alexandre Scotto Di Perrotolo. 2018. A Theoretical Framework for Bayesian Optimization Convergence.
- [40] Raghav Sethi, Martin Traverso, Dain Sundstrom, David Phillips, Wenlei Xie, Yutian Sun, Nezih Yigitbasi, Haozhen Jin, Eric Hwang, Nileema Shingte, et al. 2019. Presto: SQL on everything. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1802–1813.
- [41] Yu Shen, Yupeng Lu, Yang Li, Yaofeng Tu, Wentao Zhang, and Bin Cui. 2022. DivBO: Diversity-aware CASH for Ensemble Learning. *Advances in Neural Information Processing Systems* 35 (2022), 2958–2971.
- [42] Subhajit Siddhanta, Wojciech Golab, and Supratik Mukhopadhyay. 2019. Deadline-aware cost optimization for spark. *IEEE Transactions on Big Data* 7, 1 (2019), 115–127.
- [43] Rekha Singhal and Praveen Singh. 2018. Performance assurance model for applications on SPARK platform. In *Performance Evaluation and Benchmarking for the Analytics Era: 9th ITC Technology Conference, TPCTC 2017, Munich, Germany, August 28, 2017, Revised Selected Papers* 9. Springer, 131–146.
- [44] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.
- [45] SparkConf. 2022. Configuration - Spark 3.2.1 Documentation. <https://spark.apache.org/docs/latest/configuration.html>
- [46] Kevin Swersky, Jasper Snoek, and Ryan P Adams. 2013. Multi-task bayesian optimization. *Advances in neural information processing systems* 26 (2013).
- [47] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. 2009. Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1626–1629.
- [48] Apache Spark Tuning. 2017. Apache Spark Tuning - DZone. <https://dzone.com/articles/apache-spark-performance-tuning-degree-of-parallel>
- [49] Cloudera Spark Tuning. 2018. Cloudera Performance Management - Tuning Spark Applications. <https://www.cloudera.com/documentation/enterprise/5-9>

- x/topics/admin_spark_tuning.html
- [50] Dana Van Aken, Andrew Pavlo, Geoffrey J Gordon, and Bohan Zhang. 2017. Automatic database management system tuning through large-scale machine learning. In *Proceedings of the 2017 ACM international conference on management of data*. 1009–1024.
 - [51] Shivararam Venkataraman, Zongheng Yang, Michael Franklin, Benjamin Recht, and Ion Stoica. 2016. Ernest: Efficient performance prediction for large-scale advanced analytics. In *13th {USENIX} symposium on networked systems design and implementation ({NSDI} 16)*. 363–378.
 - [52] Kewen Wang and Mohammad Maifi Hasan Khan. 2015. Performance prediction for apache spark platform. In *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*. IEEE, 166–173.
 - [53] Ziyu Wang and Nando de Freitas. 2014. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758* (2014).
 - [54] Zi Wang, Beomjoon Kim, and Leslie P Kaelbling. 2018. Regret bounds for meta bayesian optimization with an unknown gaussian process prior. *Advances in Neural Information Processing Systems* 31 (2018).
 - [55] Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme. 2015. Sequential model-free hyperparameter tuning. In *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 1033–1038.
 - [56] Jinhan Xin, Kai Hwang, and Zhibin Yu. 2022. LOCAT: Low-Overhead Online Configuration Auto-Tuning of Spark SQL Applications [Extended Version]. *arXiv preprint arXiv:2203.14889* (2022).
 - [57] Dani Yogatama and Gideon Mann. 2014. Efficient transfer learning method for automatic hyperparameter tuning. In *Artificial Intelligence and Statistics*. 1077–1085.
 - [58] Zhibin Yu, Zhendong Bei, and Xuehai Qian. 2018. Datasize-aware high dimensional configurations auto-tuning of in-memory cluster computing. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*. 564–577.
 - [59] Nikos Zacheilas, Stathis Maroulis, and Vana Kalogeraki. 2017. Dione: Profiling spark applications exploiting graph similarity. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 389–394.
 - [60] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster computing with working sets. In *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*.
 - [61] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. 2016. Apache spark: a unified engine for big data processing. *Commun. ACM* 59, 11 (2016), 56–65.
 - [62] Ji Zhang, Yu Liu, Ke Zhou, Guoliang Li, Zhili Xiao, Bin Cheng, Jiashu Xing, Yangtao Wang, Tianheng Cheng, Li Liu, et al. 2019. An end-to-end automatic cloud database tuning system using deep reinforcement learning. In *Proceedings of the 2019 International Conference on Management of Data*. 415–432.
 - [63] Xinyi Zhang, Zhuo Chang, Yang Li, Hong Wu, Jian Tan, Feifei Li, and Bin Cui. 2022. Facilitating database tuning with hyper-parameter optimization: a comprehensive experimental evaluation. *Proceedings of the VLDB Endowment* 15, 9 (2022), 1808–1821.
 - [64] Xinyi Zhang, Hong Wu, Zhuo Chang, Shuowei Jin, Jian Tan, Feifei Li, Tieying Zhang, and Bin Cui. 2021. Restune: Resource oriented tuning boosted by meta-learning for cloud databases. In *Proceedings of the 2021 International Conference on Management of Data*. 2102–2114.
 - [65] Zhi-Hua Zhou. 2012. *Ensemble methods: foundations and algorithms*. CRC press.
 - [66] Dan-Hao Zhu, Xin-Yu Dai, and Jia-Jun Chen. 2021. Pre-Train and Learn: Preserving Global Information for Graph Neural Networks. *Journal of Computer Science and Technology* 36, 6 (2021), 1420–1430.
 - [67] Yuqing Zhu, Jianxun Liu, Mengying Guo, Yungang Bao, Wenlong Ma, Zhuoyue Liu, Kunpeng Song, and Yingchun Yang. 2017. Bestconfig: tapping the performance potential of systems via automatic configuration tuning. In *Proceedings of the 2017 Symposium on Cloud Computing*. 338–350.

A APPENDIX

A.1 Algorithm Details

We provide the pseudo-code of the initialization and search phase in Algorithm 2 and 3, respectively.

A.2 Discussions

A.2.1 Relationship with previous work. According to [20], Rover belongs to **machine learning-based methods**. The main drawback of this type of method is that fitting the machine learning model often requires a large number of observations, so the optimization performance at early iterations is far from satisfactory. We observe the same challenge in our industrial scenario, where we need to find near-optimal configurations given limited observations. To tackle the challenge, Rover proposes to integrate expert rules into the machine learning-based method. According to the survey, using expert rules alone belongs to **rule-based methods**, which finds good configurations in the very beginning but the performance is less competitive than machine learning-based methods. Therefore, Rover can be briefly summarized as a **combination of rule-based methods and machine learning-based methods** to complement each other, where we improve the early performance by using expert rules and the later performance by the machine learning model. As shown in Section 6.2, Rover performs better than state-of-the-art machine learning-based methods like Tuneful and LOCAT.

In addition, Rover applies similar history tasks to **further improve machine learning-based methods**. The intuition is that the history surrogates are well-fitted and the surrogates of similar tasks may suggest good configurations in the current task. In Section 6.4, we also provide the ablation study to show that machine learning-based methods can be improved by both integrating expert rules and knowledge from history tasks.

A.2.2 Advantages of Gaussian Process. In Rover, we use the combination of Gaussian Process (GP) and Matern 5/2 kernel due to the following four reasons:

a) **Theoretical guarantee.** The theoretical characteristics of Bayesian optimization (BO) have been widely studied in previous work. For example, [54] and [53] provide the regret bounds of BO with Gaussian Process and Matern 5/2 kernel. [39] studies different aspects of the convergence of BO with Gaussian Process. While Rover aims at applying BO to solve real-world data science problems, we do not make theoretical contributions to BO. Please refer to the mentioned literatures for more theoretical results.

b) **Empirical superiority.** The combination of Gaussian Process and Matern 5/2 kernel has achieved state-of-the-art in different black-box optimization problems. For example, this combination has been successfully applied in automatic algorithm hyperparameter tuning [44], database knob tuning [64], Spark parameter tuning [10, 56]. More specifically, an empirical paper [63] on database knob tuning compares different implementations of BO. It shows that GP+Matern achieves comparable results with SMAC and better results than TPE. In addition, the combination of GP and Matern kernel is the winning solution [25] of a recent competition (2021 CIKM AnalyticCup Track 2), where the algorithm is tested on synthetic functions, real-world A/B testing, and automatic algorithm hyperparameter tuning.

Algorithm 2: Pseudo code for the initialization phase.

Input: the total initialization budget \mathcal{B}_i , the compact search space X (Sec. 4.1.1), the default configurations $\{x_1, \dots, x_d\}$.

- 1: initialize $x_{pre} = x_d$, observations $D = \{(x_i, y_i) | i \in \{1, \dots, d\}\}$.
- 2: Select similar history tasks and fetch the corresponding history surrogate set \mathcal{M} (Sec. 4.2.1).
- 3: **while** \mathcal{B}_i does not exhaust **do**
- 4: **if** $\mathcal{M} = \emptyset$ **then**
- 5: apply the expert rule trees to x_{pre} and sample x_{new} from the neighborhood region (Sec. 4.1.2).
- 6: **else**
- 7: select x_{new} by minimizing Equation 6 of the combined surrogate (Sec. 4.2.2).
- 8: **end if**
- 9: evaluate x_{new} and augment $D = D \cup (x_{new}, y_{new})$.
- 10: set $x_{pre} = x_{new}$.
- 11: **end while**

c) **Specialized requirements.** In our scenarios, the combination of GP and Matern kernel is more appropriate than other variants of BO. Concretely, as mentioned in Section 1, we focus on finding good configurations with limited observations. We take TPE [6] as a possible alternative to GP, where TPE uses kernel density estimation to model the density of configurations with good and bad performance. However, in practice, TPE requires a large number of observations for initialization, which is twice the number of parameters in the search space. Specifically, for a 30-d search space, TPE requires 60 random observations before it starts working. Therefore, we use the combination of GP and Matern kernel so that it can model the relationship between configurations and performance even with scarce observations.

d) **Fair comparison.** The last reason is that we want to make a relatively fair comparison with baselines used in Spark tuning. Tuneful and LOCAT also use the combination of GP and Matern kernel. In this way, we conclude that the improvement shown in our experiments is brought by generalized transfer learning rather than the implementation of BO surrogate.

A.2.3 Support for constrained optimization. While Rover only supports Spark tuning without constraints so far, we discuss how to extend Rover to constrained optimization in two steps: a) how Bayesian optimization supports constraints and b) how the generalized transfer learning supports constraints.

a) As Rover does not modify the core of Bayesian optimization, constraints can be supported by applying a variant of BO using the Expected Improvement with Constraints (EIC) [16] as the acquisition function. For problems without constraints, BO fits a performance surrogate that maps from a configuration to its estimated objective value. For problems with constraints, for each constraint, BO fits an extra constraint surrogate that maps from a configuration to the probability of satisfying this constraint. The ground-truth label is 1 for satisfying and 0 for not satisfying.

To choose the next configuration, we select the configuration with the largest expectation of the predicted improvement multiplied by the predicted probability of satisfying all constraints, which is, $E((y_{pred} - y^*) \cdot \Pr(c_1) \cdot \dots \cdot \Pr(c_n))$. Here y_{pred} is the prediction of the performance surrogate, y^* is the best performance observed

Algorithm 3: Pseudo code for the search phase.

Input: the search budget \mathcal{B}_s , the compact search space \mathcal{X} (Sec. 4.1.1), the history surrogate set \mathcal{M} , the observations D .
Output: the best observed Spark configuration.

```

1: while  $\mathcal{B}_s$  does not exhaust do
2:   if  $\mathcal{M} \neq \emptyset$  then
3:     Build the combined surrogate using  $\mathcal{M}$  (Sec. 4.2.2).
4:     Compute the weight  $w_s$  for the combined surrogate.
5:   else
6:     Compute the weight  $w_s$  for the BO surrogate.
7:   end if
8:   Compute the weight  $w_e$  for expert rule trees.
9:   if  $\text{rand}() < w_e / (w_e + w_s)$  then
10:    apply the expert rule trees to  $x_{pre}$  and obtain the updated
      configuration  $x_{new}$ .
11:   else
12:     if  $\mathcal{M} \neq \emptyset$  then
13:       select  $x_{new}$  by optimizing Equation 6.
14:     else
15:       select  $x_{new}$  by optimizing Equation 3.
16:     end if
17:   end if
18:   evaluate  $x_{new}$  and augment  $D = D \cup (x_{new}, y_{new})$ .
19:   set  $x_{pre} = x_{new}$ .
20: end while
```

so far, and $\Pr(c_i)$ is the probability prediction of the i^{th} constraint surrogate. In this way, Bayesian optimization can be extended to support constraints. We refer to [16] for more implementation details.

b) For expert rule trees, since there is a risk of suggesting configurations that violate the constraints, we can add more rules to support other targets like runtime. For example, as the parameter **sparkDaMaxExecutors** is highly related to runtime, we can simply increase its value when the runtime is large.

For controlled history transfer, to train the regressor, we define the similarity as the ratio of concordant prediction pairs of prediction under constraints. The prediction under constraints is computed as $y_{mean} \cdot \Pr(c_1) \cdot \dots \cdot \Pr(c_n)$, where y_{mean} is the mean prediction of the performance surrogate, and $\Pr(c_i)$ is the probability prediction of the i^{th} constraint surrogate. Then we combine different history tasks by summing up the rank of their EIC values instead of EI values.

A.3 User Interface

We also show parts of the Rover user interfaces in Figure 9. In Figure 9(a), we plot the dashboard of all tuning tasks managed by the same user. The dashboard includes two parts: 1) the overview that includes the overall cluster information, an overall optimization curve, and a histogram of the number of tasks with different improvements; 2) the summary of all tuning tasks, including reminders of failed, abnormal and finished tasks. By clicking on a certain task in the summary list, Rover jumps to a diagnosis panel with all runtime information collected during optimization so far. In Figure 9(b), we demonstrate the diagnosis panel. The task information (including Spark DAGs, running environment, etc.) is listed

on the left side. On the right side, Rover plots the performance curve of the chosen task and all configurations evaluated so far. The users can then early stop the current task in the task center by analyzing the information provided by the diagnosis panel.

A.4 Search space

In Table 2, we list the 10 parameters tuned by Rover, along with their types and value ranges. Among the 10 parameters, 3 of them are categorical, which contains 3 valid values. The others are numerical, which follows a given lower and upper bound.

Table 2: The 10 parameters tuned by Rover.

Name	Type	Value Range
spark.sql.files.maxPartitionBytes	Numerical	[16777216, 8589934592]
spark.sql.adaptive.maxNumPostShufflePartitions	Numerical	[20, 20000]
spark.dynamicAllocation.maxExecutors	Numerical	[5, 10000]
spark.driver.cores	Categorical	{1, 2, 4}
spark.driver.memory	Numerical	[1, 48]
spark.driver.memoryOverhead	Numerical	[512, 10240]
spark.executor.cores	Categorical	{1, 2, 4}
spark.executor.memory	Numerical	[1, 64]
spark.executor.memoryOverhead	Numerical	[512, 12288]
spark.vcore.boost.ratio	Categorical	{1, 2, 3}

We also provide a detailed discussion on why we optimize these 10 parameters. As mentioned in Section 4.1, we calculate the SHAP importance value of 30 parameters beforehand, and the results are shown in Table 3. The SHAP importance is computed based on the tuning history of 1000 internal tasks before we design Rover. Each task is tuned with over 50 iterations. The parameters in bold and italic type are selected based on SHAP importance rank and human experts, respectively.

Concretely, the top-5 parameters with the largest SHAP importance are **sparkExecutorMemory**, **sparkExecutorMemoryOverhead**, **sparkExecutorCores**, **sparkDriverCores**, and **sparkDaMaxExecutors**. Then we discuss why human experts choose the other five parameters.

a) **sparkDriverMemory** & **sparkDriverMemoryOverhead**. We observe that for Spark executors, the number of cores (**sparkExecutorCores**), memory (**sparkExecutorMemory**), and memory overhead (**sparkExecutorMemoryOverhead**) have been already included in the search space. We also include these two parameters to perform a find-grained search for Spark drivers.

b) **sparkFilesMaxPartitionBytes** & **sparkAeMaxShufflePartitions**. Since Map-Reduce operations are the foundation of Spark, we choose these two parameters to control the number of Mappers (**sparkFilesMaxPartitionBytes**) and Reducers (**sparkAeMaxShufflePartitions**).

c) **sparkVcoreBoostRatio**. This parameter is designed specifically for ByteDance infrastructure platform, which controls the number of RDD partitions a single core can process simultaneously. We suggest removing this parameter on other platforms for reproduction.

In addition, **sparkDaInitialExecutors** and **sparkSqlAutoBroadcastJoinThreshold** have relatively large SHAP importance. They are not included in the search space because,

a) **sparkDaInitialExecutors**. The parameter **sparkDaMaxExecutors** has a similar effect with **sparkDaInitialExecutors**. Since the former is already in the search space, we do not include the latter.

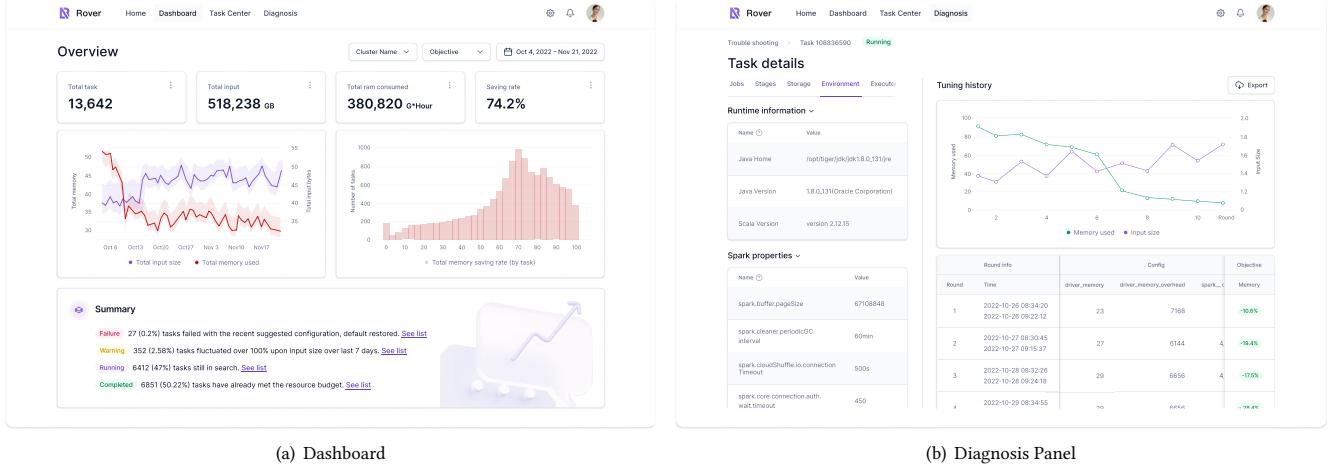


Figure 9: Demonstration of the Rover user interface.

b) `sparkSqlAutoBroadcastJoinThreshold`. This parameter controls the threshold of table size for whether we need to store the tables in memory during Map Join. While a larger value leads to better performance, we simply set a relatively large value (40M) and do not include it in the search space.

A.5 Expert Rules

In Tables 4 and 5, we list all the expert rules designed for the 10 parameters. In each table, we show the parameter name, adjustment condition, direction, step, and bounds.

We also provide instructions on how to design expert rules for each parameter in the search space. We highlight the following two important parts among all five parts mentioned in Section 3.4:

a) Condition. To design the condition of an expert rule, we should answer two questions: 1) What is the direct effect of a certain parameter on Spark? and 2) How can we measure this effect using running metrics from Spark? Then, we tune the parameter based on the observed metric values.

We take the parameter `sparkFilesMaxPartitionBytes` as an example. This parameter directly controls the maximum partition size of input data during mapping, and it influences the number of mappers that process the input data (Question 1). The metric `stage_max_avg_input_run_time` records the runtime of a mapper (Question 2), thus it can be used to check whether the parameter `sparkFilesMaxPartitionBytes` is properly set. Concretely, a large `stage_max_avg_input_run_time` indicates that the number of mappers is insufficient, so we should increase the number of mappers by reducing `sparkFilesMaxPartitionBytes`.

b) Step. In Rover, we update the parameter by multiplying it by a constant. The constant value influences the speed to reach the region of near-optimal configurations. i) We apply a large step for parameters where the upper bound is significantly higher than the lower bound. For example, the constant value is 2 or 0.5 for `sparkFilesMaxPartitionBytes` where the parameter ranges from 16M to 8G. ii) In addition, we suggest using a small step for sensitive parameters. Remind that we use SHAP importance to build

Table 3: The SHAP values of 30 parameters tuned by Tuneful.

Parameter Name	SHAP Value
<code>sparkFilesMaxPartitionBytes</code>	1.4118541
<code>sparkAeMaxShufflePartitions</code>	1.161272
<code>sparkDaMaxExecutors</code>	2.1909058
<code>sparkDriverCores</code>	2.2878344
<code>sparkDriverMemory</code>	0.7766007
<code>sparkDriverMemoryOverhead</code>	1.0949618
<code>sparkExecutorCores</code>	2.8089051
<code>sparkExecutorMemory</code>	9.722165
<code>sparkExecutorMemoryOverhead</code>	4.940778
<code>sparkVcoreBoostRatio</code>	0.8479438
<code>sparkDaInitialExecutors</code>	1.6789923
<code>sparkDaMinExecutors</code>	0.7285232
<code>sparkAeTargetPostShuffleInputSize</code>	0.5051892
<code>sparkAeSkewPartitionFactor</code>	0.6474083
<code>sparkAeSkewPartitionSizeThreshold</code>	0.8660434
<code>sparkAeBroadcastJoinThreshold</code>	0.38758022
<code>sparkSqlAutoBroadcastJoinThreshold</code>	2.1695352
<code>sparkSqlFilesOpenCost</code>	0.5295473
<code>sparkSpeculationMultiplier</code>	0.5726126
<code>sparkSpeculationQuantile</code>	0.2905913
<code>sparkDaExecutorIdleTimeout</code>	0.4301605
<code>sparkNetworkTimeout</code>	0.59897256
<code>sparkShuffleIoConnectionTimeout</code>	0.5292596
<code>sparkShuffleHighlyMapStatusThreshold</code>	0.34495476
<code>sparkMaxRemoteBlockSizeFetchToMem</code>	0.64484304
<code>sparkShuffleHdfsThreads</code>	0.34853762
<code>sparkShuffleIoMaxRetries</code>	0.49766314
<code>sparkShuffleIoRetryWait</code>	0.7191459
<code>sparkSqlInMemoryColumnarStorageBatchSize</code>	0.21765669
<code>sparkKryoserializerBufferMax</code>	0.5383134

the search space. We provide the SHAP importance of 30 parameters above in Table 3. For parameters with large SHAP importance

(e.g., 9.72 for `sparkExecutorMemory`), we set the constant value as 1.1 or 0.9 for most cases.

A.6 Additional Results

A.6.1 Comparison with expert-assisted methods. As mentioned in Section 2, the rule-based methods only provide what parameters to tune rather than how to tune them. In this part, we compare Rover with cost model-based methods. We take Ernest [51] as a representative, which builds a heuristic parametric cost model on some Spark parameters. The cost model is trained on previous observations and then suggests the configuration that maximizes the output of the cost model. As the total tuning budget of Rover is 50 iterations, we also fit the cost model in Ernest using observations of 50 random configurations.

To ensure a fair comparison, we disable controlled history transfer in Rover. The configuration suggested by Ernest achieves an average memory cost ratio of 21.37% on 200 internal tasks. As an iterative tuning method, it takes Rover only 5 iterations to achieve that result. Moreover, with the same budget (50 iterations), Rover achieves an average memory cost ratio of 7.15%, which is 14.22% lower than Ernest. We attributed this gain to a well-designed set of expert rules and the use of BO algorithm, which models the complex relationship between configurations and performance rather than assuming a simple relationship based on heuristic knowledge.

A.6.2 Optimization on another target. To show the generality of Rover, we evaluate Rover using another target (accumulated CPU cost). The accumulated CPU cost is computed as the CPU usage multiplied by the runtime, which is also an important metric used in industry.

To support different targets, the users do not need to alter any algorithm parts of Rover. Instead, they only need to collect the corresponding performance metric of the configuration, and Rover will automatically model the relationship between those configurations and performance metric. The expert rules still work when the target changes from memory cost to CPU cost. For history transfer, the users should provide the tuning history on CPU cost instead of memory cost.

Then, we show the results of Rover using accumulated CPU cost as the target. We compare Rover with vanilla BO on 200 internal tasks using the same compact space, and the average results (CPU cost ratio relative to default configurations, the lower the better) are shown in Figure 10.

We observe that Rover still achieves clear improvement compared with Bayesian optimization (4.22% and 2.92% at the 10-th and 50-th iteration), which demonstrates the generality of Rover on different requirements.

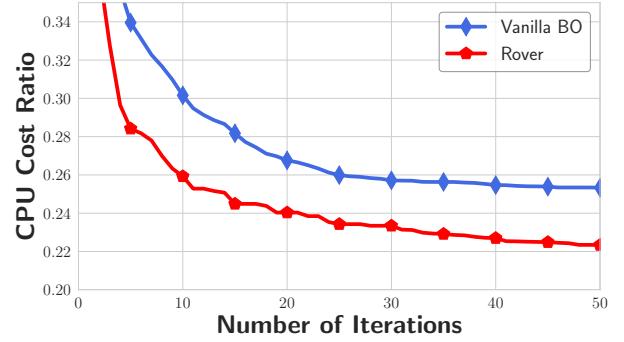


Figure 10: CPU cost ratio of Rover and vanilla BO.

Table 4: All expert rules used in the main experiments of Rover (1/2).

Name	Condition	Direction	Step	Bounds
spark.sql.files.maxPartitionBytes	stage_max_avg_input_run_time ≤ 0.2	↑	*2	Lower: 16M, Upper: 4G
spark.sql.files.maxPartitionBytes	stage_max_avg_input_run_time ≥ 0.4	↓	*0.5	Lower: 16M, Upper: 4G
spark.vcore.boost.ratio	ifnull(spark.vcore.boost.ratio, 1) ≤ 2 and ((0 < max_mem_usage ≤ 0.55 and 0 < avg_mem_usage ≤ 0.45) or avg_mem_usage = 0) and stage_max_avg_tasks_run_time ≤ 0.2	↑	/	Lower:3, Upper:3
spark.vcore.boost.ratio	spark.vcore.boost.ratio = 3 and (max_mem_usage ≥ 0.30 or avg_mem_usage = 0) and stage_max_avg_tasks_run_time ≥ 0.4	↓	/	Lower:2, Upper:2
spark.executor.cores	spark.executor.cores = 2 and stage_max_avg_tasks_run_time ≤ 0.1667	↓	/	Lower:1, Upper:1
spark.executor.cores	spark.executor.cores = 1 and stage_max_avg_tasks_run_time ≥ 0.3334 and not(0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0)	↑	/	Lower:2, Upper:2
spark.executor.cores	spark.executor.cores = 1 and stage_max_avg_tasks_run_time ≥ 0.3334 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	/	Lower:2, Upper:2
spark.executor.memory	spark.executor.cores = 2 and stage_max_avg_tasks_run_time ≤ 0.1667	↓	*0.5	Lower:1, Upper:64
spark.executor.memory	spark.executor.cores = 2 and 0 < max_mem_usage ≤ 0.72 and 0 < avg_mem_usage ≤ 0.60	↓	*0.9	Lower:1, Upper:64
spark.executor.memory	spark.executor.cores = 1 and 0 < max_mem_usage ≤ 0.72 and 0 < avg_mem_usage ≤ 0.60	↓	*0.9	Lower:1, Upper:64
spark.executor.memory	spark.executor.cores = 2 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	*1.1	Lower:1, Upper:64
spark.executor.memory	spark.executor.cores = 1 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	*1.1	Lower:1, Upper:64
spark.executor.memory	spark.executor.cores = 1 and stage_max_avg_tasks_run_time >= 0.3334 and not (0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0)	↑	*2	Lower:1, Upper:64
spark.executor.memory	spark.executor.cores = 1 and stage_max_avg_tasks_run_time >= 0.3334 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	spark.executor.memory = spark.executor.memory * 2.2 + spark.executor.memoryOverhead / 1024	Lower:1, Upper:64
spark.executor.memoryOverhead	spark.executor.cores = 2 and stage_max_avg_tasks_run_time ≤ 0.1667	↓	*0.5	Lower:512, Upper:12K
spark.executor.memoryOverhead	spark.executor.cores = 2 and 0 < max_mem_usage ≤ 0.72 and 0 < avg_mem_usage ≤ 0.60	↓	*0.9	Lower:512, Upper:12K
spark.executor.memoryOverhead	spark.executor.cores = 1 and 0 < max_mem_usage ≤ 0.72 and 0 < avg_mem_usage ≤ 0.60	↓	*0.9	Lower:512, Upper:12K
spark.executor.memoryOverhead	spark.executor.cores = 2 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	*1.1	Lower:512, Upper:12K
spark.executor.memoryOverhead	spark.executor.cores = 1 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	*1.1	Lower:512, Upper:12K
spark.executor.memoryOverhead	spark.executor.cores = 1 and stage_max_avg_tasks_run_time >= 0.3334 and not (0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0)	↑	*2	Lower:512, Upper:12K
spark.executor.memoryOverhead	spark.executor.cores = 1 and stage_max_avg_tasks_run_time >= 0.3334 and 0.90 ≤ max_mem_usage < 2.0 and 0.74 ≤ avg_mem_usage < 2.0	↑	*2.2	Lower:512, Upper:12K

Table 5: All expert rules used in the main experiments of Rover (2/2).

Name	Condition	Direction	Step	Bounds
spark.dynamicAllocation.maxExecutors	$0 \leq \text{total_memory} \leq 3$	/	/	Lower: 5, Upper: 5
spark.dynamicAllocation.maxExecutors	$3 < \text{total_memory} \leq 5$	/	/	Lower: 6, Upper: 6
spark.dynamicAllocation.maxExecutors	$5 < \text{total_memory} \leq 8$	/	/	Lower: 8, Upper: 8
spark.dynamicAllocation.maxExecutors	$8 < \text{total_memory} \leq 15$	/	/	Lower: 10, Upper: 10
spark.dynamicAllocation.maxExecutors	$15 < \text{total_memory} \leq 20$	/	/	Lower: 15, Upper: 15
spark.dynamicAllocation.maxExecutors	$20 < \text{total_memory} \leq 40$	/	/	Lower: 20, Upper: 20
spark.sql.adaptive.maxNumPostShufflePartitions	spark.sql.adaptive.enabled = 'true' and spark.sql.adaptive.maxNumPostShufflePartitions ≥ 40 and $0 \leq \text{stage_max_avg_shuffle_read_run_time} \leq 0.1667$	\downarrow	*0.5	Lower:40, Upper:400
spark.sql.adaptive.maxNumPostShufflePartitions	spark.sql.adaptive.enabled = 'true' and $\text{stage_max_avg_shuffle_read_run_time} \geq 1.2$	\uparrow	*2	Lower:80, Upper:800
spark.driver.cores	$0 < \text{max_driver_mem_usage} \leq 0.35$ and $0 < \text{avg_driver_mem_usage} \leq 0.25$	/	/	Lower:1, Upper:1
spark.driver.cores	$0 < \text{max_driver_mem_usage} \leq 0.45$ and $0 < \text{avg_driver_mem_usage} \leq 0.35$	/	/	Lower:1, Upper:1
spark.driver.cores	$0 < \text{max_driver_mem_usage} \leq 0.60$ and $0 < \text{avg_driver_mem_usage} \leq 0.50$	/	/	Lower:1, Upper:1
spark.driver.cores	$0.92 \leq \text{max_driver_mem_usage} \leq 2.0$ and $0.92 \leq \text{avg_driver_mem_usage} \leq 2.0$	/	/	Lower:1, Upper:1
spark.driver.cores	$0.8 \leq \text{max_driver_mem_usage} \leq 1.0$ and $0.72 \leq \text{avg_driver_mem_usage} \leq 1.0$	/	/	Lower:1, Upper:1
spark.driver.cores	$0.9 \leq \text{max_driver_mem_usage} \leq 1.0$ and $0.6 \leq \text{avg_driver_mem_usage} \leq 1.0$	/	/	Lower:1, Upper:1
spark.driver.memory	$0 < \text{max_driver_mem_usage} \leq 0.35$ and $0 < \text{avg_driver_mem_usage} \leq 0.25$	\downarrow	*0.8	Lower:1, Upper:48
spark.driver.memory	$0 < \text{max_driver_mem_usage} \leq 0.45$ and $0 < \text{avg_driver_mem_usage} \leq 0.35$	\downarrow	*0.85	Lower:1, Upper:48
spark.driver.memory	$0 < \text{max_driver_mem_usage} \leq 0.60$ and $0 < \text{avg_driver_mem_usage} \leq 0.50$	\downarrow	*0.9	Lower:1, Upper:48
spark.driver.memory	$0.92 \leq \text{max_driver_mem_usage} \leq 2.0$ and $0.92 \leq \text{avg_driver_mem_usage} \leq 2.0$	\uparrow	*1.2	Lower:1, Upper:48
spark.driver.memory	$0.8 \leq \text{max_driver_mem_usage} \leq 1.0$ and $0.72 \leq \text{avg_driver_mem_usage} \leq 1.0$	\uparrow	*1.1	Lower:1, Upper:48
spark.driver.memory	$0.9 \leq \text{max_driver_mem_usage} \leq 1.0$ and $0.6 \leq \text{avg_driver_mem_usage} \leq 1.0$	\uparrow	*1.1	Lower:1, Upper:48
spark.driver.memoryOverhead	$0 < \text{max_driver_mem_usage} \leq 0.35$ and $0 < \text{avg_driver_mem_usage} \leq 0.25$	\downarrow	*0.8	Lower:512, Upper:10K
spark.driver.memoryOverhead	$0 < \text{max_driver_mem_usage} \leq 0.45$ and $0 < \text{avg_driver_mem_usage} \leq 0.35$	\downarrow	*0.85	Lower:512, Upper:10K
spark.driver.memoryOverhead	$0 < \text{max_driver_mem_usage} \leq 0.60$ and $0 < \text{avg_driver_mem_usage} \leq 0.50$	\downarrow	*0.9	Lower:512, Upper:10K
spark.driver.memoryOverhead	$0.92 \leq \text{max_driver_mem_usage} \leq 2.0$ and $0.92 \leq \text{avg_driver_mem_usage} \leq 2.0$	\uparrow	*1.2	Lower:512, Upper:10K
spark.driver.memoryOverhead	$0.8 \leq \text{max_driver_mem_usage} \leq 1.0$ and $0.72 \leq \text{avg_driver_mem_usage} \leq 1.0$	\uparrow	*1.1	Lower:512, Upper:10K
spark.driver.memoryOverhead	$0.9 \leq \text{max_driver_mem_usage} \leq 1.0$ and $0.6 \leq \text{avg_driver_mem_usage} \leq 1.0$	\uparrow	*1.1	Lower:512, Upper:10K