AKAMAS

Autonomous Performance Optimization for Apache Spark

Why your Spark applications need AI-driven tuning

Version 1.0

The Apache Spark Problem

The big data analytics market is growing fast, driven by the rapid increase in volume and complexity of enterprise data in virtually every industry. Leading among analytics engines and frameworks, Apache Spark features greater speed and scalability compared to many competitors.

The power and flexibility of Spark make it a notoriously complex tool to tune, with highly sensitive configurations to the kind of application. Given the high level of parallelism usually associated with Spark jobs, misconfigured applications can waste a significant amount of resources and cluster time, inflating the infrastructure costs of operating your big data solutions.

Tuning Spark is complex and laborious

Apache Spark comes with over 300 tunable parameters that significantly impact performance and infrastructure usage. Conventional and established best practices provide only a starting point for the optimization process, which requires significant manual effort to test and analyze the execution of potentially lengthy workflows.

Spark configurations are highly sensitive to data

Spark performance can vary dramatically, given its configuration sensitivity to the nature of each application, and the volume and distribution of data. Even in the same environment, the dataset may change over time with the business context. Frequently, performance engineers are forced to repeat the tuning process to keep up with changing needs.

Allocating the right resources is critical

If the Spark application has insufficient resources it may slow down or even fail, wasting cluster time. As critical is the issue of resources over-allocation, which can cause runaway cloud bills. In both cases, poorly configured sizing can inflate the cost of the cluster or starve other applications in a shared environment.

+300 Spark parameters that impact performance and resources.
Source: Apache Spark

40% Spark cloud instances that are overprovisioned.
Source: 2020 Big Data Performance Report

"Configuring a Spark job is as much an art as a science. Choosing a configuration depends on the size and setup of the data storage solution, the size of the jobs being run, and the kind of jobs"

High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark - H. Karau, R. Warren

AKAMAS

Autonomous Performance Optimization

An AI-driven, automation-oriented platform gives the operations teams the solution they need to ensure they find the optimal Apache Spark configurations, without manually repeating the tuning process. This way, performance engineers are freed from tedious and time-consuming trial-and-error optimization tasks. They can trust that their Spark applications are always running with the optimal amount of resources, achieving higher performance and lower costs.

Autonomously learn from actual performance metrics

An AI-driven platform can correlate the parameters from the whole stack, from the Spark layer to the underlying cloud infrastructure, to the metrics captured from all the configurations under test. This allows the AI engine to learn from empirical data and find the setup that best satisfies your preset goals without any need to dig into the mechanics and execution of application-level optimization.

Scale the tuning process

Autonomous performance optimization removes manual, repetitive and tedious work from the optimization loop. It reduced dramatically the time and effort required to scale the tuning process to multiple Spark environments with different characteristics. It is also continuous, meaning it adapts configurations of the same stack over time as the volume, speed and type of data evolves with the business.

Cut infrastructure footprint and costs

As a goal-driven tool, it can search for Spark stack configurations that use resources more efficiently, reducing the footprint required to meet the desired performance KPIs. It can also alleviate the problem of stalled executors wasting resources, reducing infrastructure cost, or increasing the throughput of multi-tenant clusters while maintaining overall application reliability.

With Autonomous Performance Optimization

-20%

Reduction in average execution time.

-80%

Reduction in manual work for the performance engineers team.

-50%

Less memory footprint for the same execution time.

Akamas

The Autonomous Performance Optimization Platform

Akamas is a new breed of performance optimization technology that helps enterprises, online businesses and SaaS vendors extract unprecedented levels of performance and cost savings from their technology stacks.

Built by veterans in performance engineering and data science, Akamas exploits advanced machine learning techniques to continuously optimize hundreds of interdependent IT configuration parameters.

Akamas is a company of Moviri, a global software and professional services group, and counts BMC Software, Dynatrace, HPE, Neotys and Splunk among its partners. Headquartered in Milan, Akamas has offices in Boston, Los Angeles and Singapore.

Powered by AI

Akamas uses machine learning to solve intractable optimization problems. By shrinking the target configuration space, it delivers massive performance and cost savings in hours instead of weeks.

Automated

Akamas radically outperforms trial-and-error manual optimization by using automation to iteratively design performance experiments, analyze outcomes and deploy settings.

Full-Stack

Akamas is a smart, technology-agnostic optimization platform that understands the interdependencies between operating system, middleware and application configurations.

Goal-Driven

Akamas discovers the best configuration tradeoffs that meet your performance goals. Availability. Throughput. Response time. Cost. You set the goal, Akamas figures out how to achieve it.

Learn more about the future of Apache Spark optimization

For more information about how Akamas can help you optimize your Spark applications, visit akamas.io or contact us at info@akamas.io

Milan

Via Schiaffino,11
20158, Milan Italy

Boston

211 Congress Street
Boston, MA 02110

Los Angeles

12655 W. Jefferson Blvd
Los Angeles, CA 90066

Singapore

5 Temasek Boulevard
Singapore, 038985

AKAMAS