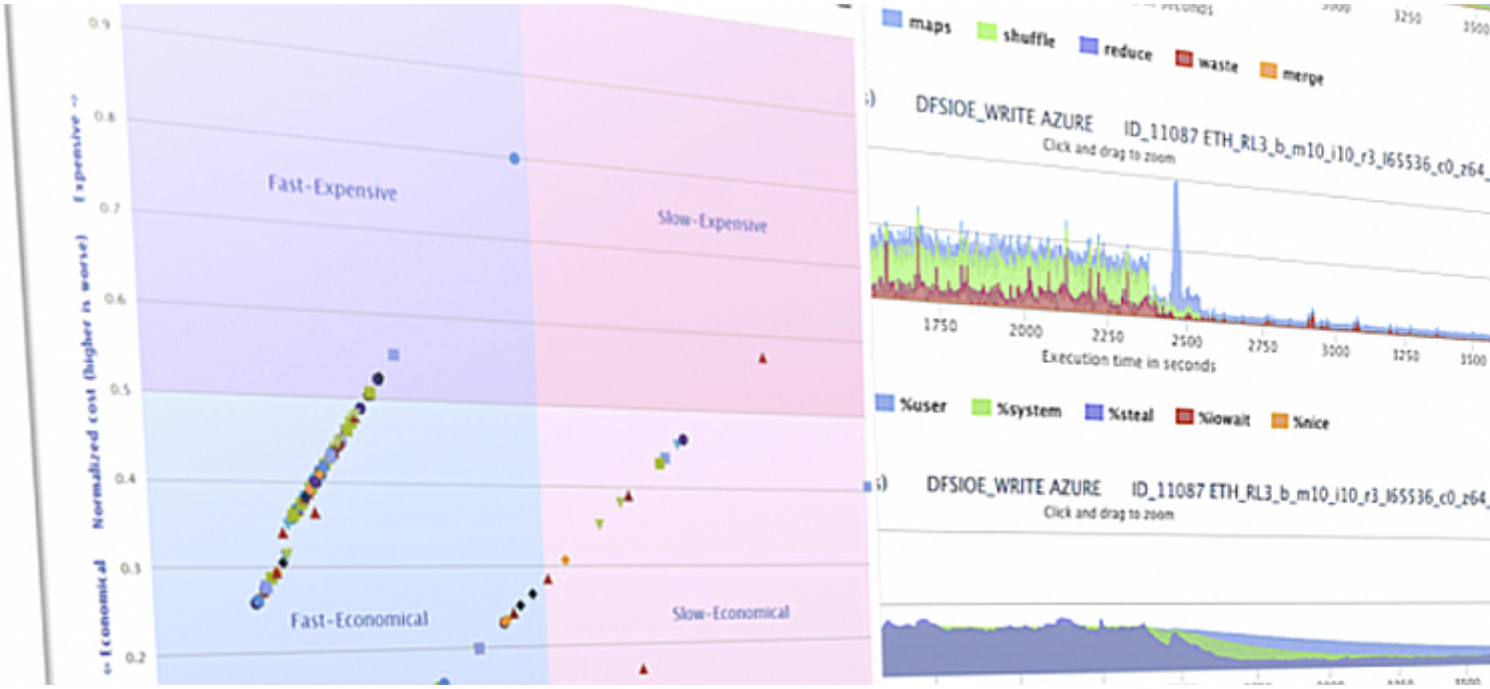
Big Data Frameworks



This research line has developed the ALOJA Project, an open research benchmarking and analysis platform that aims to lower the total cost of ownership (TCO) of Big Data deployments and study their performance characteristics for optimization.

Summary

This line has produced the ALOJA benchmarking project the largest open Big Data performance repository, with over 50,000 runs. The searchable repository features different applications for Hadoop, software configurations, data sizes, and more than 100 different hardware deployment options. Along with the repository, ALOJA provides open source Web Analytics and Machine Learning based tools for the analysis and characterization of results. The Web tools offer both a fine-grain view of runs, as well as a high-level glance of aggregate results, and Predictive Analytics estimations and recommendations of configurations.

ALOJA is a long-term project aiming to automate the characterization of cost-effectiveness on Big Data deployments. The platform initially focused on on-premise Hadoop deployments, but currently covers a wide range of Iaas, PaaS and SaaS Cloud Services. Analytical services like Hive and Spark come pre-configured and ready to use, giving companies a quick entry and fast deployment of ready SQL-like solutions for their Big Data needs. In this context, ALOJA evaluates different provided solutions from main Cloud providers including Microsoft Azure, Amazon Web Services, Google, and Rackspace, from an end-user's evaluation perspective.

The ALOJA project has created an open, vendor-neutral repository, featuring over 50,000 Hadoop job executions and their performance details. The repository is accompanied by a test-bed and tools to deploy and evaluate the cost-effectiveness of different hardware configurations, parameters and Cloud services. Despite early success within ALOJA, a comprehensive study requires automation of modeling procedures to allow an analysis of large and resource-constrained search spaces. The predictive analytics extension, ALOJA-ML, provides an automated system allowing knowledge discovery by modeling environments from observed executions. The resulting models can forecast execution behaviors, predicting execution times for new configurations and hardware choices. That also enables model-based anomaly detection or efficient benchmark guidance by prioritizing executions. In addition, the community can benefit from ALOJA data-sets and framework to improve the design and deployment of Big Data applications.

Objectives

The objective of this line is to explore and improve the scalability and cost-effectiveness of Big Data frameworks such as Hadoop and Spark, to upcoming hardware architectures including Cloud services. With the intent to better understand the performance, therefore the costs of running different data applications through an automated benchmarking and modeling process. By testing different deployment scenarios including: physical servers (commodity, low-power/SoC, appliances, HPC); Cloud IaaS, PaaS, and SaaS; to find optimal framework and deployment configurations and recommending application placement. As well as to provide Analytics and Knowledge Discovery tools, to produce insights that can guide the design of cost-efficient Big Data applications.

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---