Accepted Manuscript

Using machine learning to optimize parallelism in big data applications

Álvaro Brandón Hernández, María S. Perez, Smrati Gupta,
Victor Muntés-Mulero

PII: S0167-739X(17)31466-8
DOI: http://dx.doi.org/10.1016/j.future.2017.07.003
Reference: FUTURE 3533

To appear in: *Future Generation Computer Systems*

Received date : 16 January 2017
Revised date : 2 April 2017
Accepted date : 1 July 2017

Please cite this article as: Á.B. Hernández, M.S. Perez, S. Gupta, V. Muntés-Mulero, Using machine learning to optimize parallelism in big data applications, *Future Generation Computer Systems* (2017), http://dx.doi.org/10.1016/j.future.2017.07.003

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

- We propose a machine learning based method that recommends optimal parameters for task parallelization in Big Data workloads.
- We leverage these predictions to recommend an optimal configuration to users before launching their workloads in the cluster, avoiding possible failures, performance degradation and wastage of resources.
- We evaluate our method with a benchmark of 15 Spark applications on the Grid5000 testbed. We observe up to a 51% gain on performance when using the recommended parallelism settings.
- The model is also interpretable and can give insights to the user into how different metrics and parameters affect the performance.

ACCEPTED MANUSCRIPT

Download English Version:

https://daneshyari.com/en/article/6873067

Download Persian Version:

https://daneshyari.com/article/6873067

Daneshyari.com