Title: Data Analytics in the Cloud with Flexible MapReduce Workflows

Author(s): Goncalves, Carlos [1]; Assunção, Luis [1]; Cunha, José C.

Source: 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)

Published: 2012

Conference: 4th IEEE International Conference on Cloud Computing Technology and Science (CloudCom) **Location:** Taipei, Taiwan **Date:** DEC 03-06, 2012

Sponsor(s): IEEE; Quanta Comp; MediaTek; Microsoft; Inst Informat Ind; Ind Technol Res Inst; HareDB.com; IEEE Comp Soc; IEEE Tech Comm Scalable Comp (TCSC); Natl Tsing Hua Univ

Document Type: Proceedings Paper

Language: English

Abstract: Data analytic applications are characterized by large data sets that are subject to a series of processing phases. Some of these phases are executed sequentially but others can be executed concurrently or in parallel on clusters, grids or clouds. The MapReduce programming model has been applied to process large data sets in cluster and cloud environments. For developing an application using MapReduce there is a need to install/configure/access specific frameworks such as Apache Hadoop or Elastic MapReduce in Amazon Cloud It would be desirable to provide more flexibility in adjusting such configurations according to the application characteristics. Furthermore the composition of the multiple phases of a data analytic application requires the specification of all the phases and their orchestration. The original MapReduce model and environment lacks flexible support for such configuration and composition. Recognizing that scientific workflows have been successfully applied to modeling complex applications, this paper describes our experiments on implementing MapReduce as sub-workflows in the A WARD framework (Autonomic Workflow Activities Reconfigurable and Dynamic). A text mining data analytic application is modeled as a complex workflow with multiple phases, where individual workflow nodes support MapReduce computations. As in typical MapReduce environments, the end user only needs to define the application algorithms for input data processing and for the map and reduce functions. In the paper we present experimental results when using the A WARD framework to execute MapReduce workflows deployed over multiple Amazon EC2 (Elastic Compute Cloud) instances.

Author Keywords: MapReduce; Workflow; Text Mining; Cloud

Reprint Address: Gonçalves, C (reprint author) - Univ Nova Lisboa and Inst Super Engn Lisboa, Lisbon, Portugal.

E-mail Addresses: cgoncalves@deetc.isel.pt; lass@isel.pt; jcc@fct.unl.pt

Addresses:
[1] Univ Nova Lisboa, Inst Super Engn Lisboa, Lisbon, Portugal

Publisher: IEEE
Publisher Address: 345 E 47TH ST, New York, NY 10017 USA

ISBN: 978-1-4673-4510-1

Citation: GONÇALVES, Carlos; ASSUNÇÃO, Luis; CUNHA, José C. - Data Analytics in the Cloud with Flexible MapReduce Workflows. 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom). ISBN 978-1-4673-4510-1. (2012).