Title: Cost Efficient Scheduling of MapReduce Applications on Public Clouds

Authors: Xuezhi Zeng, Saurabh Kumar Garg, Zhenyu Wen, Peter Strazdins, Albert Y. Zomaya, Rajiv Ranjan

PII: S1877-7503(17)30854-2
DOI: http://dx.doi.org/doi:10.1016/j.jocs.2017.07.017
Reference: JOCS 732

To appear in:

Received date: 1-12-2016
Revised date: 27-6-2017
Accepted date: 25-7-2017

Please cite this article as: Xuezhi Zeng, Saurabh Kumar Garg, Zhenyu Wen, Peter Strazdins, Albert Y.Zomaya, Rajiv Ranjan, Cost Efficient Scheduling of MapReduce Applications on Public Clouds, Journal of Computational Sciencehttp://dx.doi.org/10.1016/j.jocs.2017.07.017

Cost Efficient Scheduling of MapReduce Applications on Public Clouds

Xuezhi Zeng[1], [2]Saurabh Kumar Garg, [3]Zhenyu Wen, [1]Peter Strazdins, [4]Albert Y. Zomaya, [5]Rajiv Ranjan
[1]Australian National University, Australia
[2]University of Tasmania, Australia
[3]School of Informatics, The University of Edinburg, United Kingdom
[4]School of Information Technologies, University of Sydney, Australia
[5]Newcastle University, United Kingdom

Abstract—MapReduce framework has been one of the most prominent ways for efficient processing large amount of data requiring huge computational capacity. On-demand computing resources of Public Clouds have become a natural host for these MapReduce applications. However, the decision of what type and in what amount computing and storage resources should be rented is still a user's responsibility. This is not a trivial task particularly when users may have performance constraints such as deadline and have several Cloud product types to choose with the intention of not spending much money. Even though there are several existing scheduling systems, however, most of them are not developed to manage the scheduling of MapReduce applications. That is, they do not consider things such as number of map and reduce tasks that are needed to be scheduled and heterogeneity of Virtual Machines (VMs) available. This paper proposes a novel greedy-based MapReduce application scheduling algorithm (MASA) that considers the user's constraints in order to minimize cost of renting Cloud resources while considering Service Level Agreements (SLA) in terms of the user given budget and deadline constraints. The simulation results show that MASA can achieve 25-50% cost reduction in comparison to current SLA agnostic methods and there is only 10% performance disparity between MASA and an exhaustive search algorithm.
Keywords- Big Data; MapReduce; Cloud Computing; Service Level Agreement; Scheduling; Cross Layer

1. INTRODUCTION

The efficient processing of Big Data has become a predominant challenge in several emerging application domains including (but not limited to) enterprise computing, smart cities, remote healthcare, high energy physics, bio-informatics, and astronomy[1]. For example, online retail companies are required to analyze click stream data and up-to-the-minute inventory status for offering dynamically priced and customized product bundles. Similarly, banks are looking to detect and react to frauds in based on analyzing transactional data. On the other hand, cities are evolving into smart cities by fusing and analyzing data from several sources (e.g., traffic cameras, social media, remote sensing data, GPS data[2-4]. With the push towards more automation for faster business strategy adaptation, most enterprises are moving towards the next generation Business Intelligence (BI) systems that can support data-driven decision making [5]. Such organizations often utilize MapReduce-based applications for efficient and effective large-scale processing of their Big Data. This requires either installation of a private Hadoop Cluster or deployment of MapReduce application on Public Cloud. Given the on-demand, vast and scalable computing and storage resources provided by Clouds, they are becoming more and more preferable deployment infrastructure.

Public Cloud providers such as Amazon Web Services have started to offer on-demand Hadoop clusters (PaaS), referred to as Elastic MapReduce, on its EC2 datacenters (IaaS) on pay-as-you-go basis[6]. However, current scheduling techniques and systems for deploying Hadoop clusters[7] on public IaaS Clouds are incapable of supporting Service Level Agreement (SLA)-driven data processing application management. Important SLA constraints include: (i) Deadline: upper bound on the time finishing the data processing task and (ii) Budget: upper bound on the monetary limit for completing the data processing task. In the current practice, public Cloud providers require users (MapReduce application administrators) to