Information Sciences 319 (2015) 171–188

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Towards building a data-intensive index for big data computing – A case study of Remote Sensing data processing

Yan Ma [a], Lizhe Wang [a,*], Peng Liu [a], Rajiv Ranjan [b]

[a] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, PR China
[b] Computational Informatics, CSIRO, Australia

ARTICLE INFO

Article history:
Received 20 June 2014
Received in revised form 2 October 2014
Accepted 5 October 2014
Available online 25 October 2014

Keywords:
Big data
Parallel computing
Data-intensive computing
Remote Sensing data processing

ABSTRACT

With the recent advances in Remote Sensing (RS) techniques, continuous Earth Observation is generating tremendous volume of RS data. The proliferation of RS data is revolutionizing the way in which RS data are processed and understood. Data with higher dimensionality, as well as the increasing requirement for real-time processing capabilities, have also given rise to the challenging issue of "Data-Intensive (DI) Computing". However, how to properly identify and qualify the DI issue remains a significant problem that is worth exploring. DI computing is a complex issue. While the huge data volume may be one of the reasons for this, some other factors could also be important. In this paper, we propose an empirical model (DI_{RS}) of DI index to estimate RS applications. DI_{RS} here is a novel empirical model (DI_{RS}) that could quantify the DI issues in RS data processing with a normalized DI index. Through experimental analysis of the typical algorithms across the whole RS data processing flow, we identify the key factors that affect the DI issues mostly. Finally, combined with the empirical knowledge of domain experts, we formulate DI_{RS} model to describe the correlations between the key factors and DI index. By virtue of experimental validation on more selected RS applications, we have found that DI_{RS} model is an easy but promising approach.

1. Introduction

The advent of the high-resolution global Earth Observation era is revolutionizing the way in which Remote Sensing (RS) data are generated, processed, and analyzed. The latest generation of space-borne sensors are generating nearly continuous streams of massive RS imageries. These data streams are transmitted through downlink channels at a rate of several gigabits per second. The amount of RS data acquired by a single satellite data center is dramatically increasing by several TB per day [25,12]. The advances in the spatial, temporal, and spectral resolution of sensors also lead to the high dimensionality of RS imagery pixels. Meanwhile, large-scale RS applications [10,31,27,9,19,30] are exploiting multi-temporal RS data with regional to global coverage as input for processing. Obviously, Remote Sensing applications are overwhelmed with the amount of high-dimensional RS data.

With the proliferation of data, Remote Sensing data processing turns out to be extremely challenging. Issues to be considered include efficient management and rapid processing of massive RS data, pixels of high dimensionality, and especially various multi-staged algorithms of RS applications with higher complexity and intensive irregular data access patterns [21].

* Corresponding author.
E-mail address: Lizhe.Wang@gmail.com (L. Wang).

http://dx.doi.org/10.1016/j.ins.2014.10.006
0020-0255/

The situation has become even worse because of the increasing real-time or near-real-time processing requirements of many time-critical applications like hazard monitoring and tracking [11,24]. Generally, Remote Sensing data processing, especially for large-scale environmental monitoring and research, is regarded as typical Data-Intensive (DI) problems [23].

DI issues occur when the volumes and rates of data emerge as the rate-limiting factor of processing, and promise a revolutionary change in the way we seek solutions [18,14]. However, most of the existing DI issues are qualitatively defined and tend to focus on problems related to the huge amount of data [4]. Relying on these obscure and qualitative definitions to determine whether a problem is data-intensive is rather difficult. For a deeper insight into DI problems, it is critical important to quantify the meaning of data intensiveness beyond a qualitative framework. The requirements and challenges for DI problems are totally different from those related to traditional computing-intensive issues, where the computation capability is the main concern. As a result, the peak computing performance TFLOPS and Linpack benchmark are no longer applicable to DI computing, where the huge data processing capability turns out to be the main problem. Recently, plenty of benchmarks have emerged [2,13,5] for DI computing, designed for evaluating the performance of a platform rather than analyzing the DI characteristics of specific applications. Plenty of related factors need to be considered in identifying a problem as data-intensive [18]. The requirements and challenges posed by DI computing vary across different applications. Thus, it is difficult to give a definition covering the full scope of diverse DI applications. Focusing on the large-scale RS data processing, the DI issues are not well defined and analyzed, except for awareness of the huge volumes of RS data.

To properly solve the above issues in the RS domain, we propose DI_{RS}, an empirical model of a data-intensive index for Remote Sensing applications. Our main contribution in this work is a novel model to quantify DI issues in RS data processing with a normalized DI index. RS data processing is normally carried out as a multi-staged workflow that corresponds to the concept of on-the-flow processing [6]. For a thorough analysis, we choose typical algorithms covering the entire processing flow from satellite data acquisition to thematic applications for study. At each processing stage, we specify an empirical DI_{RS} index for each typical algorithm according to experts experience. Here, the possible factors that may influence the DI issues are taken into account for analysis. These factors include data volumes, the data rate, data throughput, complexity of algorithms, and so on. Then, the correlation between these factors and the DI_{RS} index, as well as the contribution of each factor to the total DI_{RS} index, are used for experimentation and quantified. Thereafter, the significant factors are distinguished to model the DI index mathematically. Accordingly, by combining empirical knowledge, experimentation, and quantitative analysis, we construct an empirical (DI_{RS}) model to quantify the DI issues for RS data processing.

The rest of this paper is organized as follows. The Section 2 reviews some related work, and the problem definition is addressed in Section 3. Section 4 presents the analysis of the RS data processing flow as a whole. In Section 5, we go into the detail concerning the construction of the empirical model (DI_{RS}) for quantitatively estimating DI issues in RS applications. The Section 6 discusses the experimental validation and analysis of the DI_{RS} model, and finally the Section 7 summarizes this paper.

2. Related works

The growing amount of datasets is outstripping the current capacity to explore and interpret them [18]. As stated in DOE-sponsored report [1], "we are entering a new era: data-intensive computing". Many organizations and researchers have proposed qualitative definitions of emerging DI issues by emphasizing the huge volumes of data [17,23].

Driven by the widening of application requirements, the definition of data-intensive computing is shifted to a broader realm to focus on the time it takes to reach a solution as a key factor [18]. One definition is as follows: "*computational task where data availability is the rate-limiting factor to producing time-critical solutions*" [18]. Another more promising and comprehensive definition has been put forward by [14]: "*data-intensive computing is managing, analyzing, and understanding data at volumes and rates that push the frontiers of current technologies*" [18,28]. However, both these definition of DI issues have a qualitative focus. Solely depending on these vague definitions will make it almost impossible to classify an application as data-intensive. Thus, some more specific standards for determination are essential. These standards could be estimation indexes or benchmarks. Therefore, the point is that a more quantitative approach to measure and analyze the realm of DI issues turns out to be valuable.

Benchmarks are commonly accepted for performance evaluation. In contrast to compute-intensive problems where Peak Flops, Linpack [8], and TOP500 are accepted as widespread benchmarks, DI issues are much more complicated to measure. The reason for this may be the complexities arising from the extremely large scale of data and the actual geographical distribution characteristics. Recently, a sort of DI benchmarks have emerged where Malstone [2] designed DI computing of data mining in the Cloud, the Sort Benchmark is used for massive data sorting in the Cloud, and Graph 500 [5] uses DI graph computing for estimation. However, all of these benchmarks are used for performance estimation of platforms with given DI requirements of applications, but not for analyzing applications themselves.

Reagan [23] defined DI computing as "*Applications that are I/O bound, and devote the largest fraction of execution time to movement of data*". He also proposed "*Computational Bandwidth[the] number of bytes of data processed per floating-point operation*" as a quantitative way of identify DI problems. Actually, the computational bandwidth describes the data throughput requirement of applications on a platform with a computation capability of certain flops. The mismatch between the evaluated computational bandwidth of applications and the fixed rate of the system (disk bandwidth divided by Peak Flops) would lead to imbalance in the system. From a system perspective, this is an excellent way of quantitatively evaluating DI