Accepted Manuscript

Building a fault tolerant framework with deadline guarantee in big data stream computing environments

Dawei Sun, Guangyan Zhang, Chengwen Wu, Keqin Li, Weimin Zheng

ISSN 0022-0000
ELSEVIER
JOURNAL OF COMPUTER AND SYSTEM SCIENCES
Available online at www.sciencedirect.com
ScienceDirect

PII: S0022-0000(16)30114-3
DOI: http://dx.doi.org/10.1016/j.jcss.2016.10.010
Reference: YJCSS 3035

To appear in: *Journal of Computer and System Sciences*

Received date: 14 January 2016
Revised date: 21 October 2016
Accepted date: 29 October 2016

Please cite this article in press as: D. Sun et al., Building a fault tolerant framework with deadline guarantee in big data stream computing environments, *J. Comput. Syst. Sci.* (2016), http://dx.doi.org/10.1016/j.jcss.2016.10.010

Building a Fault Tolerant Framework with Deadline Guarantee in Big Data Stream Computing Environments

Dawei Sun[1, 2], Guangyan Zhang[1, *], Chengwen Wu[1], Keqin Li[3], Weimin Zheng[1]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P.R. China
[2] School of Information Engineering, China University of Geosciences, Beijing, 100083, P.R. China
[3] Department of Computer Science, State University of New York, New Paltz, New York 12561, USA

***Abstract*—Big data stream computing systems should work continuously to process streams of on-line data. Therefore, fault tolerance is one of the key metrics of quality of service in big data stream computing. In this paper, we propose a fault tolerant framework with deadline guarantee for stream computing called FTDG. First, FTDG identifies the critical path of a data stream graph at a given data stream throughput, and quantifies the system reliability of a data stream graph. Second, FTDG allocates tasks by the fault tolerance aware heuristic and critical path scheduling mechanism. Third, FTDG online optimizes the task scheduling by reallocating the critical vertices on the critical path of the data stream graph to lower the response time and reduce system fluctuations. Theoretical as well as experimental results demonstrate that the FTDG makes a desirable trade-off between high fault tolerance and low response time objectives in big data stream computing environments.**

Keywords- big data; critical path; deadline guarantee; fault tolerance; online applications; stream computing

1. Introduction

1.1 *Background and motivation*

Big data stream computing, the long-held dream of high-throughput computing which uses programs that compute continuous data streams, has opened up the new era of future computing due to big data, which are datasets that are too large, too fast, too dispersed, and too unstructured, and thus beyond the ability of available hardware and software facilities to undertake their acquisition, access, analysis and/or applications in a reasonable amount of time and space. Some popular features of big data are described by *n*Vs, high Volume, high Velocity, high Variety, high Veracity, high Validity, high Value, and so on. The rise of big data presents big opportunities and big challenges. Stream computing is an effective way to support big data and cloud computing by providing extremely low-latency velocities with massively parallel processing architectures, and is becoming the fastest and most efficient way to obtain useful knowledge from big data, allowing organizations to react quickly when problems appear or to predict new trends in the near future [1] [2] [3] [4].

Big data stream computing can be employed in many different scenarios, such as stock market analysis, click streams analysis, traffic stream analysis, and emergency response, to name but a few. Usually, when compared with batch data, big data stream is difficult to be processed in real time with traditional data computing infrastructures, as it has the following distinctive characteristics [2] [5] [6] [7]. (1) The data are not all available at once as they arrive one tuple by one tuple in continuous data stream form. (2) The order of arrival of each data tuple cannot be controlled, when the same data tuples are to be recomputed, the order of those tuples is always different with that of before. (3) The input data stream rate is often at a high speed level and might fluctuate with time, or some statistical properties might change and increase computing and communication demands, failing to deal with such cases may result in performance bottlenecks, and at worst, data loss. So the big data stream computing system needs to online adjust and elastically adapt to the change of input data stream. (4) Timely analysis of the data stream is very important as the life cycle of most of the data is very short, all the data tuples will be processed in real time, and each data tuple can be processed only once. (5) The scale of data is infinite, and the infinite scale of data need to be processed under tight constraints, furthermore, the volume of data is so high that there is not enough space for storage, and not all data need to be stored, so the batch computing model with the feature of store then computing is not fit all. Nearly all the data in big data environments have the feature of stream, and thus stream computing has appeared to solve the dilemma of big data computing by computing data online within real time constraints. So the stream computing model will be a new trend for high-throughput computing in big data era, and it is urgent to investigate the challenges in big data stream computing systems.

The issue of high fault tolerance is one of the major obstacles for opening up the new era of reliable stream computing in big data environments. While in most current stream computing environments, high performance and service level objectives are under consideration, high fault tolerance is ignored. In big data stream computing environments, the data centers are

[*] Corresponding author.

E-mail addresses: sundaweicn@ cugb.edu.cn (D. Sun), gyzh@tsinghua.edu.cn (G. Zhang), wcw14@mails.tsinghua.edu.cn (C. Wu), lik@newpaltz.edu (K. Li), zwm-dcs@tsinghua.
edu.cn (W. Zheng).