Accepted Manuscript

EDAWS: A distributed framework with efficient data analytics workspace towards discriminative services for critical infrastructures

Renke Wu, Linpeng Huang, Peng Yu, Haojie Zhou

PII: S0167-739X(17)31118-4
DOI: https://doi.org/10.1016/j.future.2017.11.009
Reference: FUTURE 3801

To appear in: *Future Generation Computer Systems*

Received date : 29 May 2017
Revised date : 18 September 2017
Accepted date : 5 November 2017

Please cite this article as: R. Wu, L. Huang, P. Yu, H. Zhou, EDAWS: A distributed framework with efficient data analytics workspace towards discriminative services for critical infrastructures, *Future Generation Computer Systems* (2017), https://doi.org/10.1016/j.future.2017.11.009

EDAWS: a distributed framework with efficient data analytics workspace towards discriminative services for critical infrastructures

Renke Wu[a,*], Linpeng Huang[a], Peng Yu[b], Haojie Zhou[c]

[a] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Minhang, Shanghai, 200240, China*
[b] *School of Software, Shanghai Jiao Tong University, 800 Dong Chuan Road, Minhang, Shanghai, 200240, China*
[c] *The State Key Laboratory of Mathematic Engineering and Advance Computing, Jiangnan Institute of Computing Technology, 699 Shanshui East Road, Wuxi, 214000, China*

Abstract

Critical infrastructure systems which are interrelated with people's daily life perform functions in multiple domains. However, with the explosion of specialized textual information in such systems, providing discriminative services for users through potential knowledge discovery becomes an essential and technical concern. Once massive data analytics is conducted in standalone server, the performance will degenerate tremendously. Alternatively, people cannot conveniently get such discriminative (self-caring) services. To address these concerns, we propose the general solution of **EDAWS**: *a Novel Distributed Framework with Efficient Data Analytics Workspace towards Discriminative Service for Critical Infrastructures*, through leveraging the state-of-the-art software technologies and computing paradigms. We argue it from the following aspects: Firstly, the server-side platform facilitates native data capture, storage, index and data mining with a systematic organization. Secondly, a text-mining approach with index building in parallel is conducted for various functional business, by exploiting the potential of *Lucene-based* distributed cluster. Thirdly, with the widespread usage of tiny but powerful mobile devices, the server-side platform could be accessed by mobile-side clients remotely in a more convenient way. To demonstrate our solution, a case study of *smart residence prototype* towards discriminative services in terms of information retrieval, personalized information push, and hot topic discovery is thoroughly discussed. The extensively experimental studies are conducted for the prototype over various real-world datasets. Experimental results indicate that, data processing which runs on computing nodes has good scalability with data sizes and computing nodes, and the prototype passes from data to discriminative services successfully.

Keywords: Critical infrastructure systems, Parallel processing, Distributed environment, Lucene, Discriminative services, *EDAWS* framework;
2010 MSC: 00-01, 99-00

1. Introduction

Critical infrastructure systems, which refer to the systems that tightly integrate computation, physical devices (e.g., smart phones) and networking, are omnipresent to support useful applications or effective services [1]. Such systems based on data analytics have been directly applied to smart residence, smart health-care, smart building, etc. Massive data (e.g., textual information) is produced in physical devices through communication networks from individuals. With the rapidly growing scale and utmost importance of data, the increasing demand of data analytics becomes inevitable. Moreover, people are eager to obtain discriminative services in a more convenient way. Hence, providing a distributed framework with efficient data analytics workspace towards discriminative services reinforces a substantial challenge to critical infrastructures.

Big data produced by critical infrastructure systems poses serious challenges to cyber operations. Meanwhile, scalable methods are required to capture, store, manage and process the collected big data. Nowadays, various data analytical techniques have been proposed to mine out useful information. They provide innovative services for enterprises and city users in different domains of commerce, government and society, etc. Among currently advanced techniques, information retrieval libraries (e.g., Apache Lucene [2], Egothor [3]) are able to provide an effective solution. Specifically, words or sentences could be built as inverted indexes by utilizing the advanced tools mentioned above.

Researchers have devoted fundamental work of conducting text-mining to providing sophisticated services [4]. Nevertheless, there exist inherent limitations: (a) *Unsuccessfully tackle the ever-increasing volume of data because of burdensome computing operations.* (b) *Lack of a comparatively mature approach of text-mining for knowledge discovery.* (c) *Ignore of providing no convenient information interaction manner.* Consequently, the design of such distributed framework is a non-trivial work.

Originally loose even incompatible computer equipment (standalone) is now often evolved into a highly integrated data parallel processing system. It leads to reconsidering the way of coping with such data by leveraging the potential of distributed systems. The purpose is to explore higher efficiency while processing massive data volume, variety and veracity. At the same time, the well-known approaches about machine learning, artificial intelligence or complicated data mining are migrated to distributed or cloud environment [5, 6]. Due to the salient characteristics of distributed computing as the elastic computing power, data analytics capacities keep pace with the requirement of continuously increasing system computation in terms

*Corresponding author. Tel.: +86 15869027198.
Email addresses: sjtuwrk@sjtu.edu.cn (Renke Wu), lphuang@sjtu.edu.cn (Linpeng Huang), yuzhiyu3@sjtu.edu.cn (Peng Yu), zhou.haojie@meac-skl.cn (Haojie Zhou)