An Efficient Spark-based Adaptive Windowing for Entity Matching

Demetrio Gomes Mestre, Carlos Eduardo Santos Pires,
Dimas Cassimiro Nascimento, Andreza Raquel Queiroz,
Veruska Borges Santos, Tiago Brasileiro Araujo

PII: S0164-1212(17)30055-9
DOI: 10.1016/j.jss.2017.03.003
Reference: JSS 9935

To appear in: *The Journal of Systems & Software*

Received date: 16 August 2016
Revised date: 28 February 2017
Accepted date: 3 March 2017

Please cite this article as: Demetrio Gomes Mestre, Carlos Eduardo Santos Pires, Dimas Cassimiro Nascimento, Andreza Raquel Queiroz, Veruska Borges Santos, Tiago Brasileiro Araujo, An Efficient Spark-based Adaptive Windowing for Entity Matching, *The Journal of Systems & Software* (2017), doi: 10.1016/j.jss.2017.03.003

An Efficient Spark-based Adaptive Windowing for Entity Matching

Demetrio Gomes Mestre[a,b], Carlos Eduardo Santos Pires[a], Dimas Cassimiro Nascimento[a,c], Andreza Raquel Queiroz[a], Veruska Borges Santos[a], Tiago Brasileiro Araujo[a]

[a] *Data Quality Research Group - Federal University of Campina Grande (UFCG), Aprigio Veloso Street, 882, Campina Grande, Paraiba, Brazil*
[b] *Coordination of Information Technology and Communication (CTIC) - State University of Paraiba (UEPB), Baraunas Street, 351, Campina Grande, Paraiba, Brazil*
[c] *Federal Rural University of Pernambuco (UFRPE), Garanhuns, Pernambuco, Brazil*

Abstract

Entity Matching (EM), i.e., the task of identifying records that refer to the same entity, is a fundamental problem in every information integration and data cleansing system, e.g., to find similar product descriptions in databases. The EM task is known to be challenging when the datasets involved in the matching process have a high volume due to its pair-wise nature. For this reason, studies about challenges and possible solutions of how EM can benefit from modern parallel computing programming models, such as Apache Spark (Spark), have become an important demand nowadays [1][2]. The effectiveness and scalability of Spark-based implementations for EM depend on how well the workload distribution is balanced among all workers. In this article, we investigate how Spark can be used to perform efficiently (load balanced) parallel EM using a variation of the Sorted Neighborhood Method (SNM) that uses a varying (adaptive) window size. We propose Spark Duplicate Count Strategy (S-DCS++), a Spark-based approach for adaptive SNM, aiming to increase even more the performance of this method. The evalu-

[*] Corresponding author
Email addresses: demetriogm@gmail.com (Demetrio Gomes Mestre), cesp@dsc.ufcg.edu.br (Carlos Eduardo Santos Pires), dimascnf@gmail.com (Dimas Cassimiro Nascimento), andreza.queiroz@ccc.ufcg.edu.br (Andreza Raquel Queiroz), veruska.santos@ccc.ufcg.edu.br (Veruska Borges Santos), tiagobrasileiro@copin.ufcg.edu.br (Tiago Brasileiro Araujo)