UMA ABORDAGEM ONLINE PARA DEDUPLICAÇÃO DE GRANDES BASES DE DADOS

Pedro Beschorner Marin[1]

Bolsista PRH-PB 217, pbmarin@inf.ufrgs.br, [1]Instituto de Informática, Universidade Federal do Rio Grande do Sul

R E S U M O

MOTIVAÇÃO: A necessidade de manutenção de bases de dados cada vez maiores motiva a criação de ferramentas capazes de lidar com problemas que esse grande volume de informações pode gerar. Uma das dificuldades neste processo é a identificação de duplicatas, ou seja, a identificação de registros que correspondam ao mesmo objeto, durante, por exemplo, a integração de bases de dados. Esta tarefa, chamada deduplicação, torna-se complexa a medida que o volume de dados é acentuado.

OBJETIVO: Tradicionalmente, o processo de deduplicação de bases de dados é executado em lote (computação *batch*). Isso implica, na maioria das vezes, aplicá-lo a uma versão estática de uma base de dados que talvez esteja em constante mudança. A proposta deste trabalho é desenvolver um modelo de deduplicação para bases de dados capaz de atuar em tempo real.

APLICAÇÃO NA INDÚSTRIA DO PETRÓLEO: A deduplicação de bases de dados pode contribuir de diversas maneiras à indústria do petróleo. De uma maneira geral, qualquer processo que envolva o armazenamento ou a análise de um grande volume de dados que esteja sucetível a inserção de informações redundantes pode ser beneficiado pelo processo de deduplicação. Como resultado, a deduplicação pode promover uma economia de espaço e custo computacional e até mesmo análises mais apuradas de dados coletadas em campo, por exemplo.

RESULTADOS OBTIDOS: O trabalho está sendo desenvolvido através do framework para computação distribuída em tempo real *Storm*. Este framework de código aberto oferece um ambiente tolerante a falhas baseado no paradigma de computação distribuída *MapReduce*. Ao usuário, o sistema promove um modelo abstrato para a gerência do cluster capaz de atribuir tarefas e controlar a comunicação necessária entre os nodos. A determinação da tarefa de pareamento de registros para comparação é baseada na abordagem de maximização de eficiência e escalabilidade de deduplicação *All Pairs* e o processo de filtragem implementa uma combinação de filtros do algoritmo *ppjoin+*.

A implementação está na fase final de testes e manutenção. Para o auxilio nesta etapa, está sendo utilizada uma base de 100 mil registros e 5 mil duplicatas gerada pelo *Freely extensible biomedical record linkage*. O resultado desse processo de deduplicação é avaliado por uma rotina de validação final.