VII JORNADA DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA
HABITATS DE INOVAÇÃO E DESENVOLVIMENTO
UFFS - CAMPUS ERECHIM
23 E 24 DE OUTUBRO DE 2017
UNIVERSIDADE FEDERAL DA FRONTEIRA SUL

BLOCAGEM DE GRANDES BASES DE DADOS EM TEMPO REAL

LUAN FÉLIX PIMENTEL[1*], GUILHERME DAL BIANCO[1]

[1]Universidade Federal da Fronteira Sul, c*ampus* Chapecó; Grupo de Estudos e Pesquisas em Inovação e Desenvolvimento Tecnológico da Universidade Federal da Fronteira Sul

*Autor para correspondência: Luan Félix Pimentel (luanfelixpimentel@gmail.com)

1 Introdução

A integração de dados tem como objetivo facilitar o acesso a informações a partir da consolidação de diferentes fontes de dados em um único repositório. Serviços como bibliotecas virtuais, *media streaming* e redes sociais dependem de um processo de integração com uma alta qualidade. Para isto, uma tarefa fundamental é a identificação de entidades (registros, documentos, textos, etc.) que já estão armazenadas na base de dados, portanto não devem ser novamente inseridas. Tal etapa, é conhecida como deduplicação.

A deduplicação online, diferente da versão estática, deve ser capaz de lidar com picos de processamento sem que sejam evidenciados gargalos e ao mesmo tempo deve ser capaz de se adaptar a possíveis alterações nos padrões dos dados. A deduplicação de dados envolve três etapas principais: blocagem, comparação e a classificação [1]. A blocagem corresponde ao processo de geração de pares candidatos. Ou seja, todos os registros devem ser analisados em busca de potenciais duplicatas. Somente registros pertencentes a um mesmo bloco são utilizados para a criação dos pares candidatos com custo quadrático de processamento. Por isso, é importante que o processo de blocagem, que representa a maior fatia de processamento [3], seja suficientemente eficiente para não resultar em atrasos de processamento. Dentro deste contexto, esta pesquisa propõe uma nova abordagem para a blocagem online através do desenvolvimento de um protótipo.

2 Objetivo

O objetivo principal deste trabalho é de propor uma nova abordagem para a blocagem online. Para tal, trabalhos publicados envolvendo a deduplicação estática foram estudados devido a ampla variedade de pesquisas já existentes. Além disso, a plataforma distribuída *Apache Storm,* para processamento online, foi integrada devido a alta eficiência no processamento de grandes volumes de dados.

3 Metodologia

Para atingir o objetivo proposto, foi utilizada a seguinte metodologia dividida em quatro etapas apresentadas na figura abaixo.

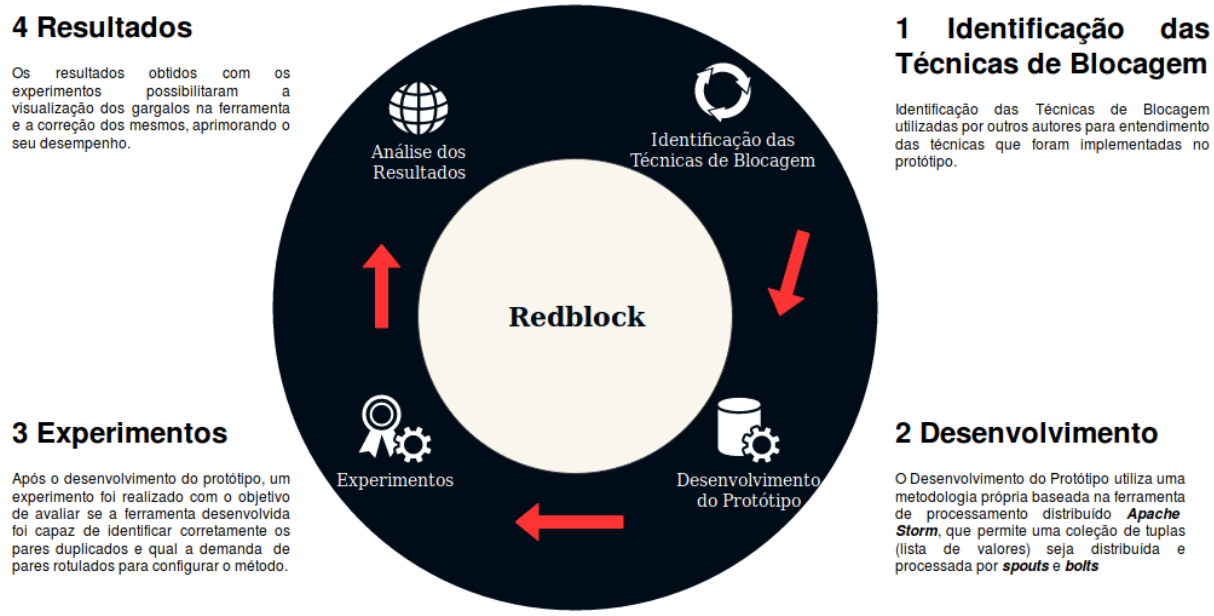

Figura 1. Metodologia de Pesquisa

Para o protótipo, foi proposta uma metodologia própria que permite que uma coleção de tuplas (lista de valores) seja distribuída e processada por *spouts* e *bolts*. O banco de dados não relacional *Redis*, foi integrado na plataforma *Apache Storm* como um *bolt, salvando as tuplas recebidas pelo Line Spout através do Line-Saver Bolt,* em tempo ágil, e permitindo uma posterior recuperação das tuplas recebidas para os demais processamentos, mantendo a estrutura inicial da informação recebida, evitando possíveis gargalos.

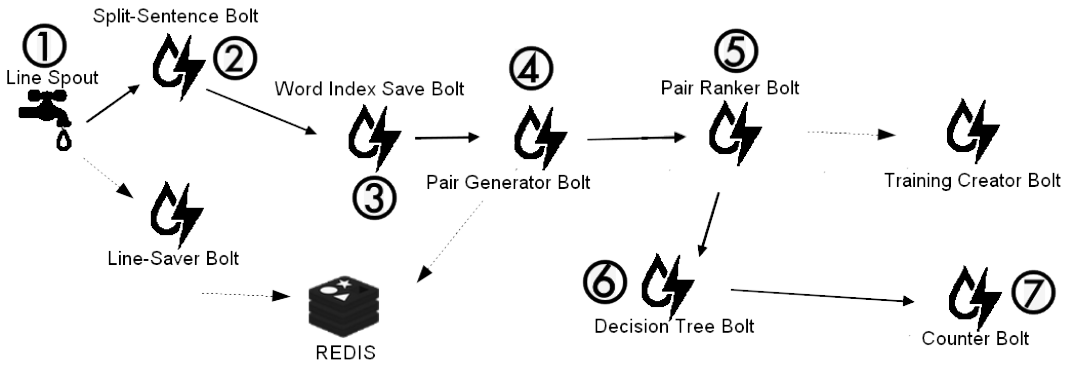

Figura 2. Topologia da Redblock (fluxo contínuo do protótipo)

As linhas pontilhadas representam processos que ocorrem em paralelo no funcionamento da ferramenta. Uma breve descrição de seu funcionamento de suas 7 etapas se encontra no quadro a seguir:

Nome	Função
Line Spout	Desempenhar a leitura da base de dados.
Line-Saver Bolt	Salvar a tupla recebida no banco de dados *Redis* no formato [ID][Linha].
Split-Sentence Bolt	Promover o processo de fragmentação da tupla.
Word Index Save Bolt	Utilizar o método do índice invertido (para blocagem dos dados).
Pair Generator Bolt	Construir pares baseados no conjunto de palavras que foram salvos.
Pair Ranker Bolt	Computar a similaridade de cada linha recebida pelo bolt anterior, mensurando o grau de semelhança de cada atributo a partir de uma função de similaridade.
Training Creator Bolt	Construir um modelo de treinamento a partir do algoritmos de árvore de decisão.
Decision Tree Bolt	Utiliza o modelo de classificação previamente criado para identificar os pares como duplicatas ou não duplicatas.
Counter Bolt	Visualizar a contagem de pares classificados durante a execução da Redblock.

Quadro 1. Descrição do fluxo contínuo do protótipo

4 Resultados e Discussão

Devido a importância para idenfificar se a Redblock é capaz de agrupar corretamente os pares e de posteriormente construir os pares candidatos, um experimento foi realizado com métricas tradicionais. A Precisão, avalia dos pares recuperados, a taxa de pares que foram

corretamente identificados. A Revocação, mede a taxa de pares recuperados comparando com o total de pares duplicados que estão presentes na base de dados.

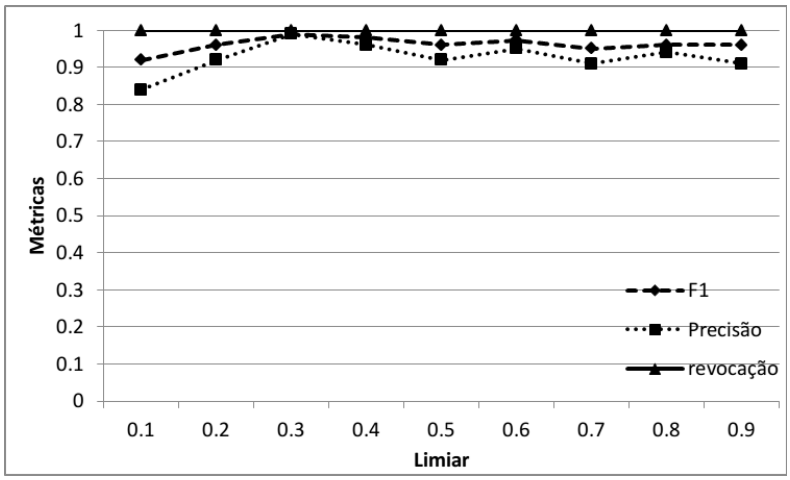

Figura 3. Experimento realizado em uma base de dados sintética.

Por fim, o F1 combina a precisão e a revocação em uma medida única. Além disso, foi utilizada uma base de dados sintética contendo 10.000 registros sendo 1.000 deles duplicatas. A base de dados foi gerada com a ferramenta *Febrl* [2]. A Figura 3 apresenta os resultados das métricas F1, precisão e revocação obtidos com a *Redblock*. O eixo X define o limiar que determina o tamanho do conjunto de treinamento. O limiar 0.1 resultou um valor de F1 de 92%, aumentando para 0.3 o valor é melhorado em 7% atingindo um valor máximo. Percebe-se que quanto mais pares rotulados, mais preciso é o treinamento do método de classificação.

É importante notar que a revocação se mantém no valor máximo em todos os limiares, demostrando que a Redblock foi capaz de encontrar todos os pares duplicados da base de dados. Tais resultados obtidos com a ferramenta *Redblock* foram publicados na Escola Regional de Banco de dados do Sul do país [4] e na Revista Brasileira de Computação Aplicada (RBCA) [5], atribuindo menção honrosa aos autores.

5 Conclusão

Como resultado desta pesquisa, foi possível obter um novo protótipo para a deduplicação online com foco no processo de blocagem. A ferramenta, denominada *Redblock*, combina o *framework Apache Storm* juntamente com o banco de dados *Redis* para possibilitar

um processamento massivo de dados. No experimento, foi possível constatar que a *Redblock* manteve uma alta qualidade (alta eficácia), ou seja, as etapas de blocagem e a classificação foram capazes de recuperar um alto número de pares duplicados sem perdas substanciais de registros positivos.

Referências

[1] CHRISTEN, P. A survey of indexing techniques for scalable record linkage and deduplication. **IEEE transactions on knowledge and data engineering**, IEEE, v. 24, n. 9, p. 1537–1555, 2012.

[2] CHRISTEN, P. Febrl: a freely available record linkage system with a graphical user interface. In: **HDKM '08:Proceedings of the second Australasian workshop on Health data and knowledge management**. Darlinghurst,Australia, Australia: Australian Computer Society, Inc., 2008. p. 17–25. ISBN 978-1-920682-61-3.

[3] BIANCO, G. D. et al. A practical and effective sampling selection strategy for large scale deduplication. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 27, n. 9, p. 2305–2319, 2015.

[4] PIMENTEL, L. F. et al. Redblock: Uma ferramenta para a deduplicação de grandes bases de dados em tempo real. **In: Escola Regional de Banco de Dados** 2017, Passo Fundo. ERBD, 2017. (Best Paper).

[5] PIMENTEL, L. F. et al. Redblock: A Tool for Online Deduplication on Large Datasets. **Revista Brasileira de Computação Aplicada,** RBCA, v. 9, n. 2, 2017. (Aprovado e aguardando publicação).

Palavras-chave: blocagem de dados; deduplicação; integração de dados.

Fonte de Financiamento: PIBITI, edital 384 – UFFS.