Introduction

Set similarity join is an essential operation in big data analytics, e.g., data integration and data cleaning, that finds similar pairs from two collections of sets. Multiple techniques have been proposed to perform similarity joins using MapReduce in recent years.

Existing techniques have several limitations.
- Generation of many duplicates
- Skewness problem
- Expensive verification processing

To address these problems, we have made the following contributions in our work:
- We proposed a vertical-partitioning based algorithm, called FS-Join, to support parallel set similarity joins without generating duplicates. In addition, it guarantees load balancing in both map and reduce phases.
- We introduced three new segment-based filtering methods, which significantly reduce the number of candidates.
- We proposed an optimization method by integrating horizontal data partitioning with vertical data partitioning to achieve higher scalability.

Existing Work

Vertical Partitioning

Global Ordering: A \rightarrow B \rightarrow C \rightarrow E \rightarrow F \rightarrow G \rightarrow H \rightarrow I \rightarrow J \rightarrow K

Pivots: \{C, F, I\}

S_1 = \{D, E, F\}
S_2 = \{A, B, C, I, K\}
S_3 = \{D, E, G\}
S_4 = \{B, C, I, J\}
S_5 = \{G, I, D, E\}
S_6 = \{A, C, D, F\}

(a) Original sets
(b) Re-ordered sets
(c) Partitioning based on pivots \{C, F, I\}

Computation Framework of FS-Join

Filter phase: generate candidate string pairs
Verification phase: produce final similarity join results

Filtering Methods

Random Selection (Random)
Even Interval (Even-Interval)
Even Token Frequency (Even-TF)

Comparison with Existing Methods

Scalability Tests