

## 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.

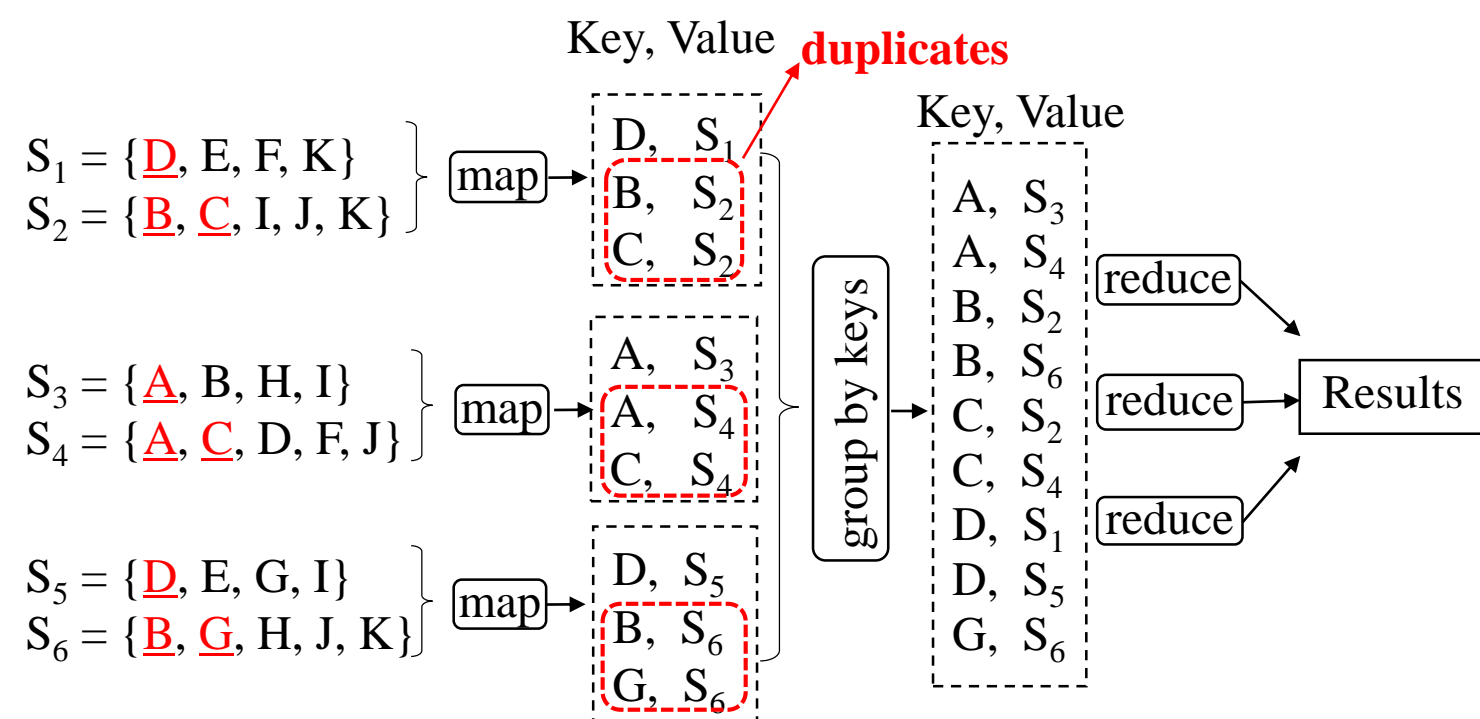
Existing techniques have several limitations.

- 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:  $A \rightarrow B \rightarrow C \rightarrow E \rightarrow F \rightarrow G \rightarrow H \rightarrow I \rightarrow J \rightarrow K$

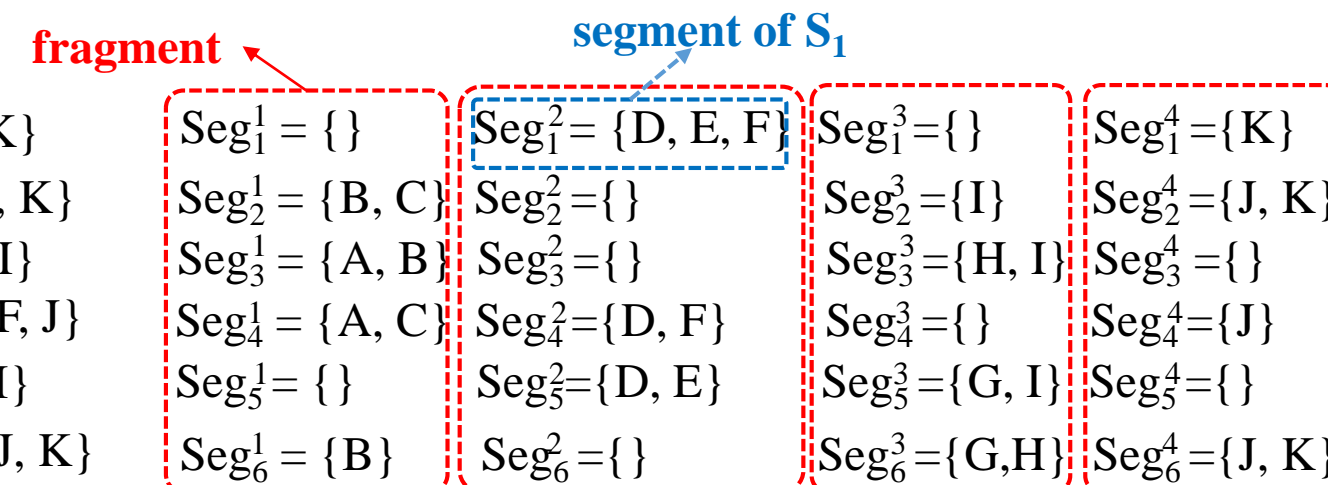
Pivots: { C, F, I }

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

Re-ordered sets:  
 $S_1 = \{D, E, F, K\}$   
 $S_2 = \{B, C, I, J, K\}$   
 $S_3 = \{A, B, H, I\}$   
 $S_4 = \{A, C, D, F, J\}$   
 $S_5 = \{D, E, G, I\}$   
 $S_6 = \{B, G, H, J, K\}$

(a) Original sets

(b) Re-ordered sets



(c) Partitioning based on pivots {C, F, I}

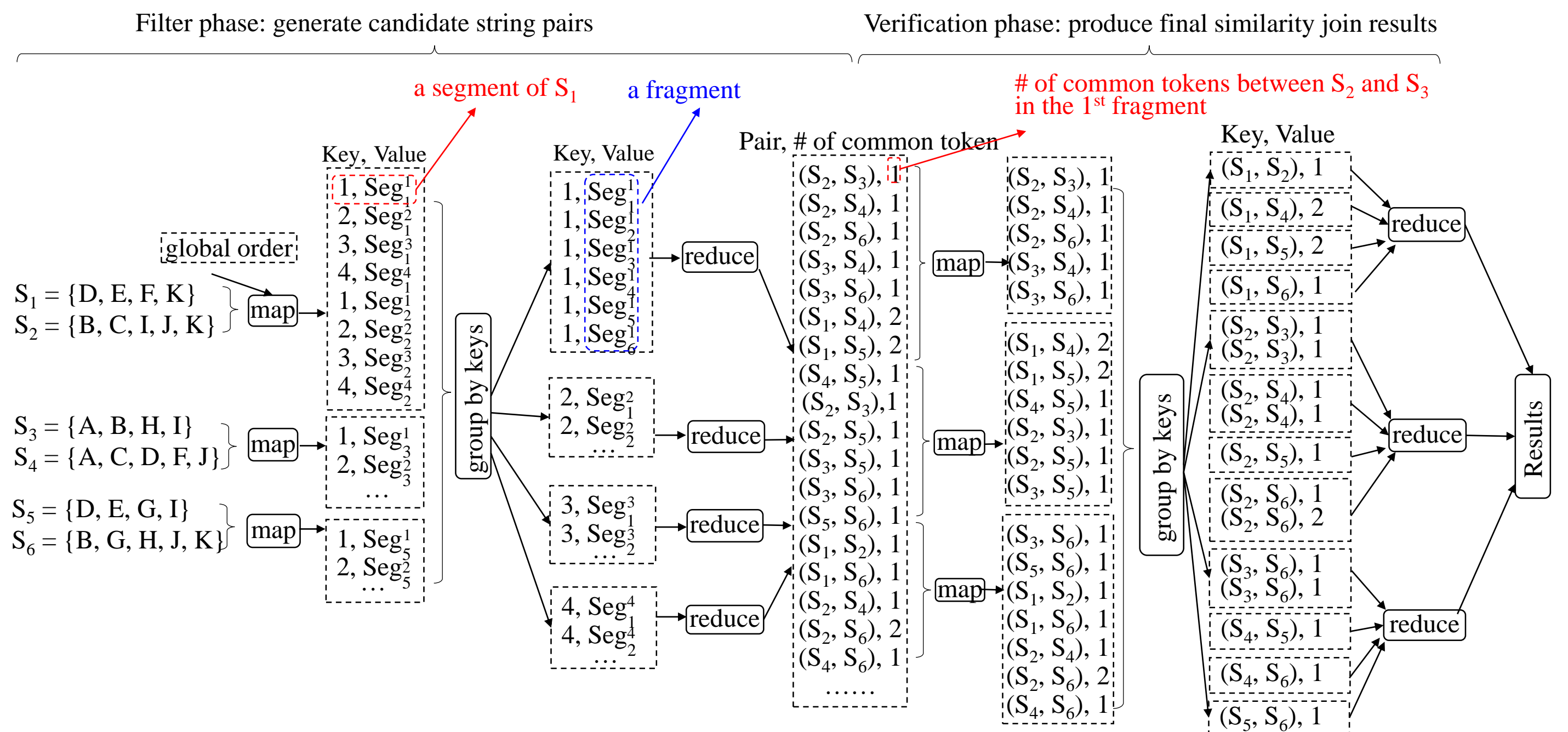
## Pivots Selection

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

## Filtering Methods

- String Length Filtering(StrL-Filter)
- Segment Length Filtering(SegL-Filter)
- Segment Intersection Filtering(SegI-Filter)
- Segment Difference Filtering(SegD-Filter)

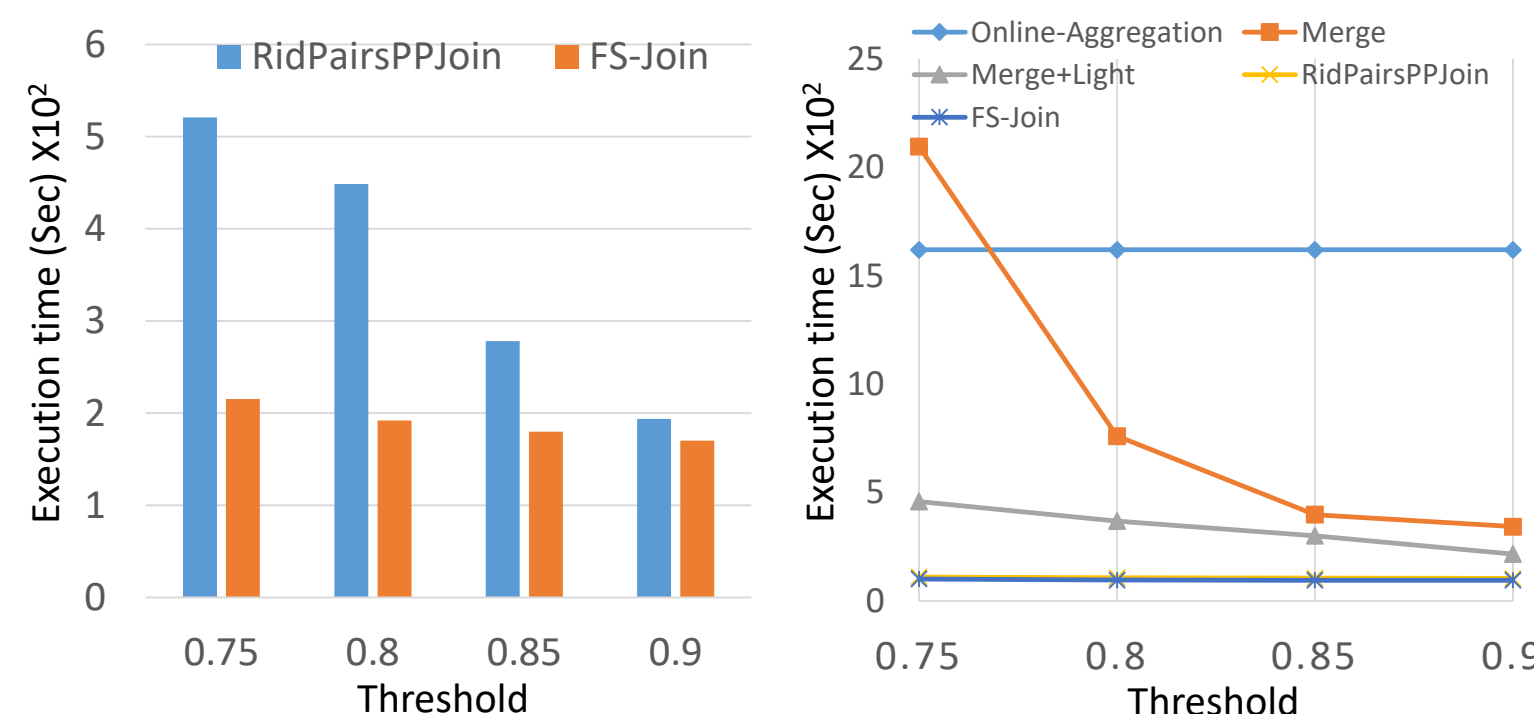
## Computation Framework of FS-Join



## Filters And Their Effectiveness

Filter	Candidates Number		
	Email(10%)	Wiki(1%)	PubMed(1%)
StrL	271,385,025	1,473,167,384	1,403,760,351
StrL + SegL	233,063,886	1,449,842,593	1,399,927,097
StrL + SegI	1,164,102	2,287,718	31,498
StrL + SegD	1,143,783	1,236,775	8,342
StrL + Prefix	1,011,428	1,147,016	792,185
All	493,644	515,664	6,840

## Comparison with Existing Methods



## Scalability Tests

