TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science, Database Technology Group

Scalable Construction of a Large IsA-Knowledge Base from Heterogeneous Web Data

Presenter: Muhammad Salman Sadaqat
Supervisor: Dr.–Ing. Dirk Habich
Professor: Prof. Dr.-Ing. Wolfgang Lehner
Database Technology Group, TU Dresden

Outline

- Introduction and Foundations
- Set Similarity Joins
  - Different Similarity Scenarios
- Implementation with Map Reduce
- Analysis and Evaluation
  - Threshold Analysis and Evaluation
  - Efficiency and Scalability Analysis
- Conclusion

Introduction and Foundations

- Background Information
- Motivation
- Foundations

Background Information

- Need for an Information Retrieval System
- Knowledge Base and Taxonomy
- Important Steps to create Taxonomy
  - Extraction
  - Integration

Motivation

- Time and Resources Required
- Extraction: Efficient
- Integration: Relatively Less Efficient

Foundations

- Map Reduce
- Jaccard Coefficient
- Dice Coefficient

Map Reduce

Programming Model

Map Phase

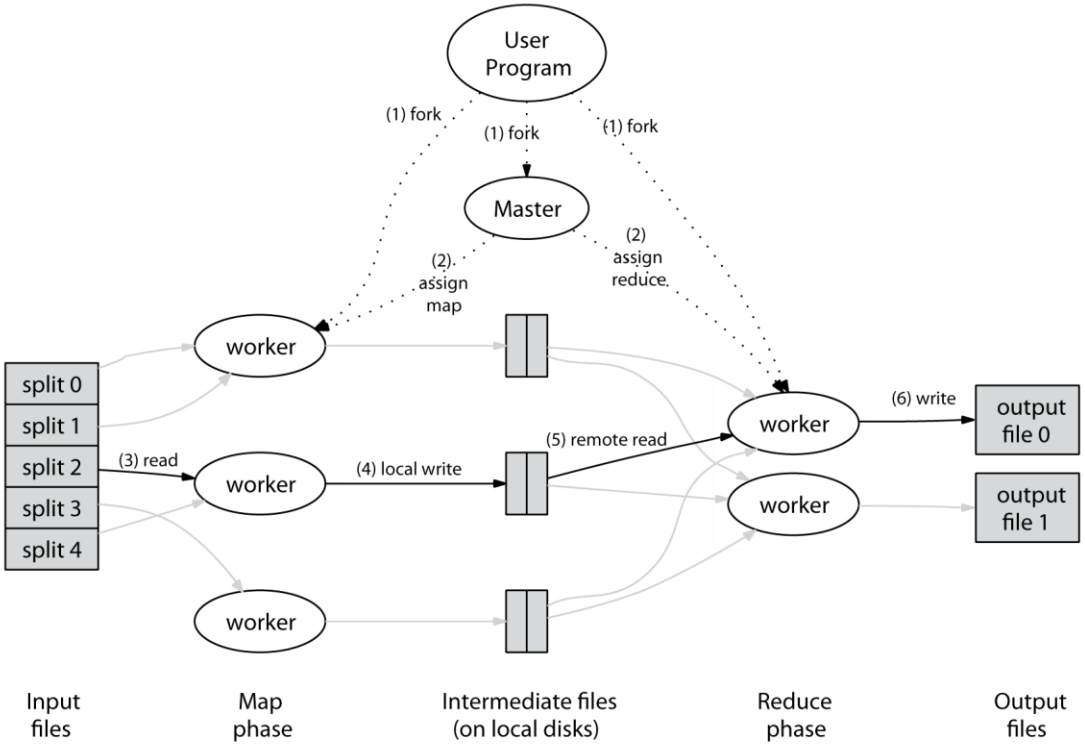
Signatures: Map(k1,v1) → list(k2,v2)

Reduce Phase

Signatures: Reduce(k2, list(v2)) → list(v3)

Map Reduce

Execution Structure



Jaccard Coefficient

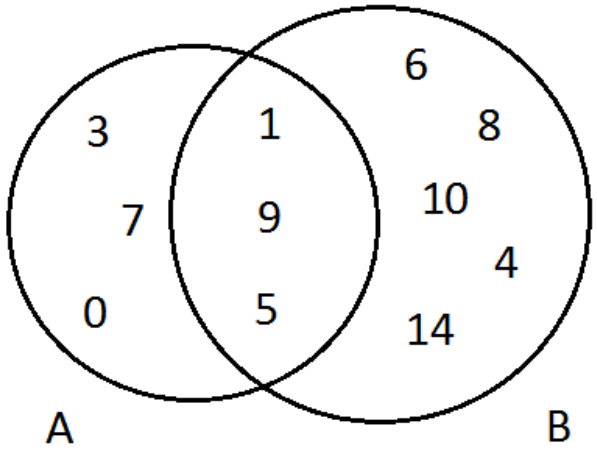
Statistical Measure for Set Similarity

A = {0, 1, 3, 5, 7, 9}

B = {1, 4, 5, 6, 8, 9, 10, 14}

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Jaccard Coefficient = 3 / 11



Dice Coefficient

Statistical Measure for Set Similarity

$$QS = \frac{2C}{A+B} = \frac{2|A \cap B|}{|A| + |B|}$$

QS = Similarity Quotient

C = Common Species

A = {0, 1, 3, 5, 7, 9}

B = {1, 4, 5, 6, 8, 9, 10, 14}

Dice Coefficient = 3 / 7

Dice Coefficient

Measure for String Similarity

Case: Plant, Plants

$$s = \frac{2n_t}{n_x + n_y}$$

nx and ny => No. of elements
in bi-grams of X and Y

X = {pl, la, an, nt}

Y = {pl, la, an, nt, ts}

Similarity= (2 . 4) / (4 + 5) = 8 / 9 = 0.89

Outline

- Introduction and Foundations
- Set Similarity Joins
  - Different Similarity Scenarios
- Implementation with Map Reduce
- Analysis and Evaluation
  - Threshold Analysis and Evaluation
  - Efficiency and Scalability Analysis
- Conclusion

Different Similarity Scenarios

- Set Similarity Identification
- Spelling Difference Identification
- Synonym Identification

Basic Set Similarity Algorithm

Case 1:

plants = {tree, grass, cactus, bamboo}
plants = {cactus, bamboo, holly, tree}

Jaccard Coefficient = 3 / 5 = 0.6

Case 2:

plants = {tree, grass, rye, tea, ferns, keek, neem, osage, holly, cactus}
plants = {cactus, bamboo, holly, tree}

Jaccard Coefficient = 3/11 = 0.27

Issue

What if we had a fixed Jaccard Threshold Value ??
(Let's Say 0.5)

Case 1:

Jaccard Coefficient = 0.6

Case 2:

Jaccard Coefficient = 0.27

Proposed Solution

Important Factors

Size difference of Sets

Percentage of Similarity of smaller set with respect to the larger set

plants = {tree, grass, rye, tea, ferns, keek, neem, osage, holly, cactus}
plants = {cactus, bamboo, holly, tree}

Percentage Similarity = No. Common Elements/Total Elements of smaller set

= 3 /4 = **75%**

Spelling Difference Identification

Why?

Set Similarity Identification→ Exactly Same Keys(SuperConcepts)

What If :

Small Spelling Difference between Super Concepts

But Significant Similarity between the Sets

plants = {tree, grass, flowers}

plant = {grass, sunflower, daisy, tree}

→ Different Reducers?

Solution : Spelling Difference Identification

Spelling Difference Identification

Implementation

Bi-Grams:

plant = {pl, la, an, nt}
plants = {pl, la, an, nt, ts}

Dice Coefficient for String Similarity

s = (2 . 4) / (4 + 5) = 8 / 9 = 0.89

Synonym Identification

What if two Super Concepts are Synonyms ??

plant = {grass, sunflower, spinach, tree}
green = {tree, spinach, flower, grass}

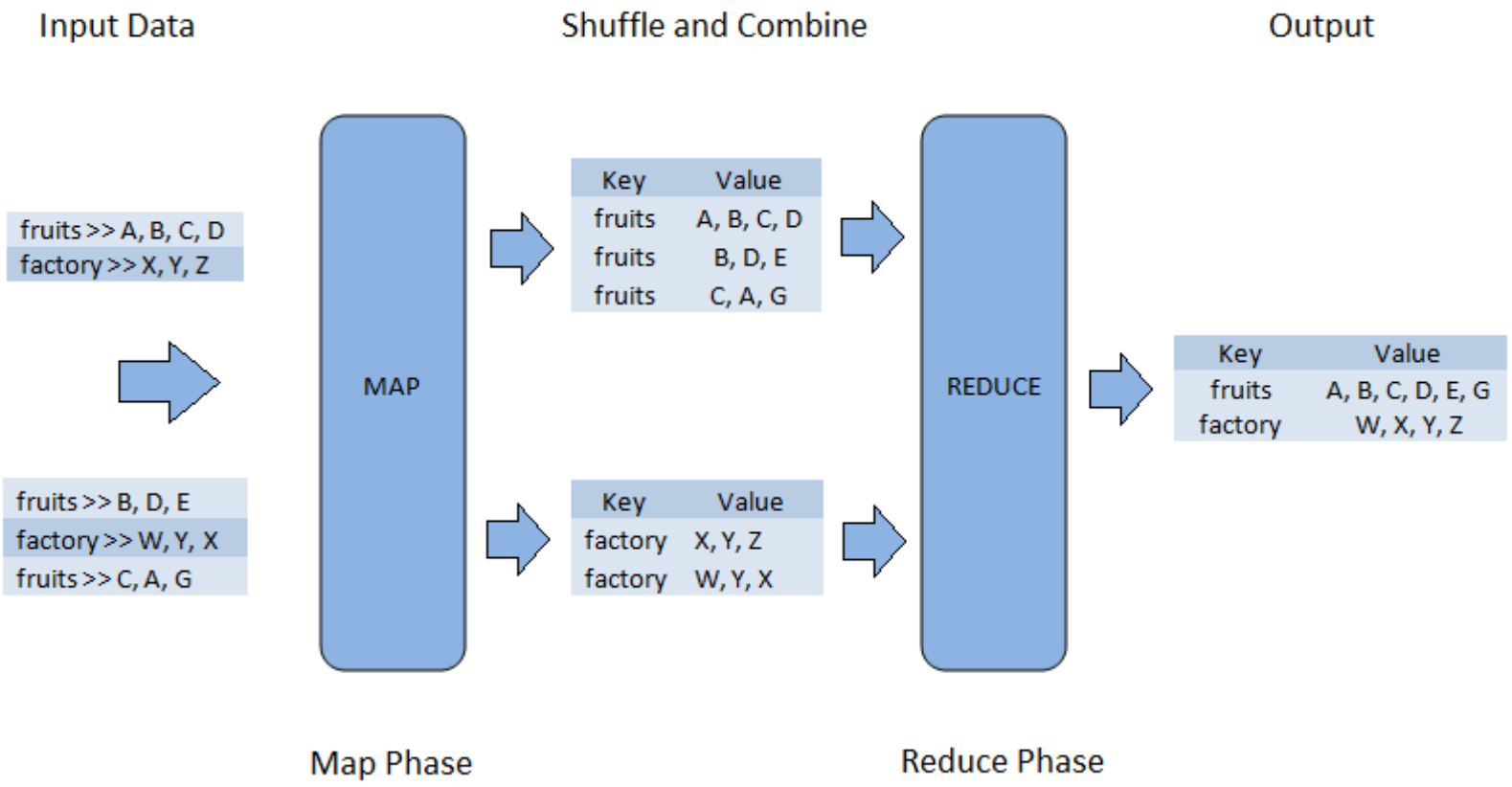
→ Different Reducers?

Solution: Synonym Identification

- Introduction and Foundations
- Set Similarity Joins
  - Different Similarity Scenarios
- Implementation with Map Reduce
- Analysis and Evaluation
  - Threshold Analysis and Evaluation
  - Efficiency and Scalability Analysis
- Conclusion

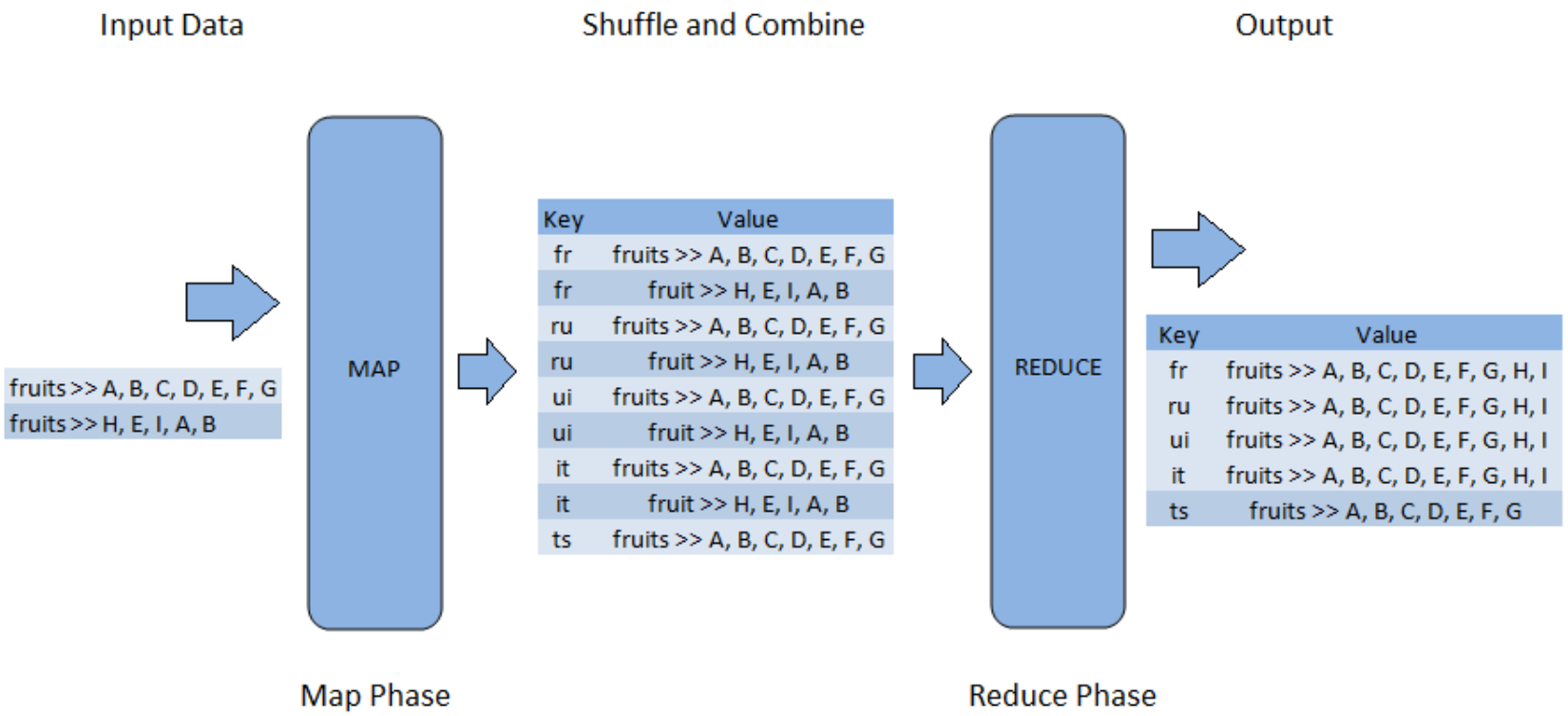
Implementation with Map Reduce

Set Similarity Identification



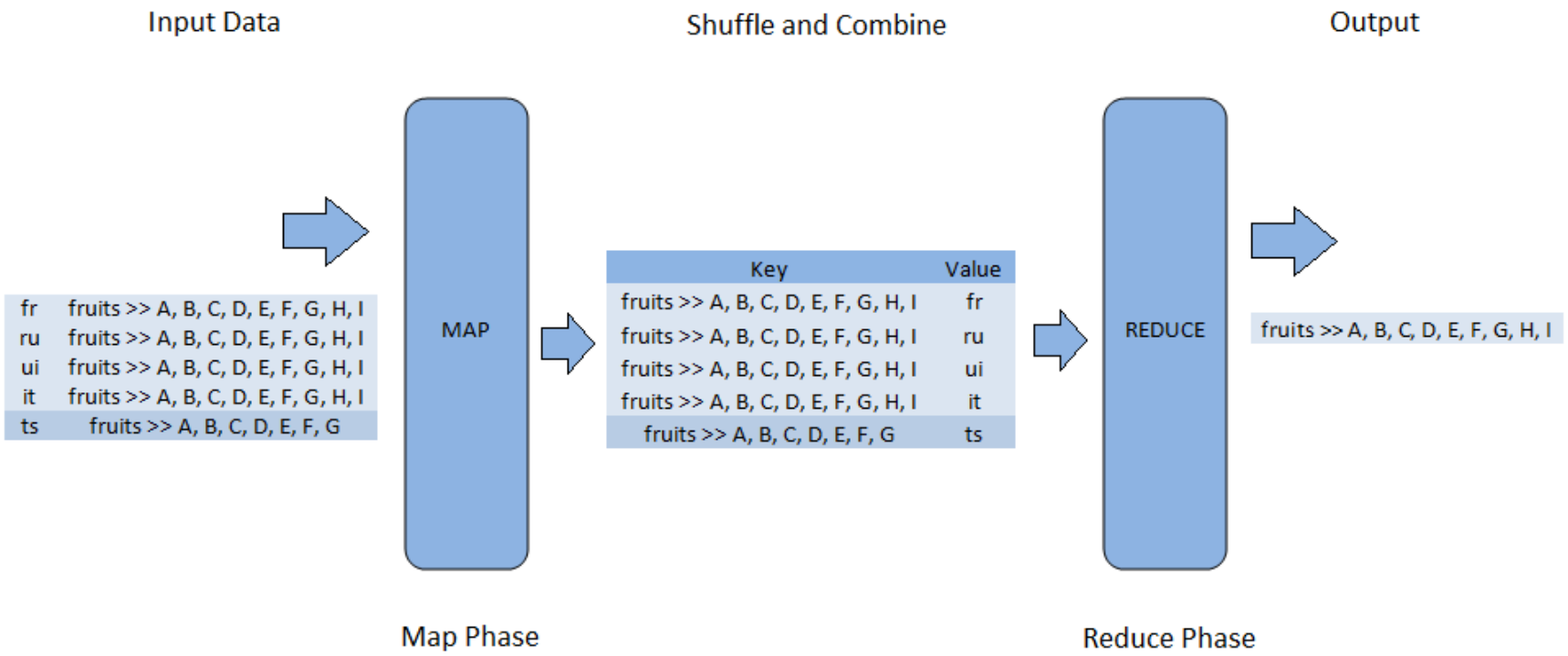
Implementation with Map Reduce

Spelling Difference Identification – First Job



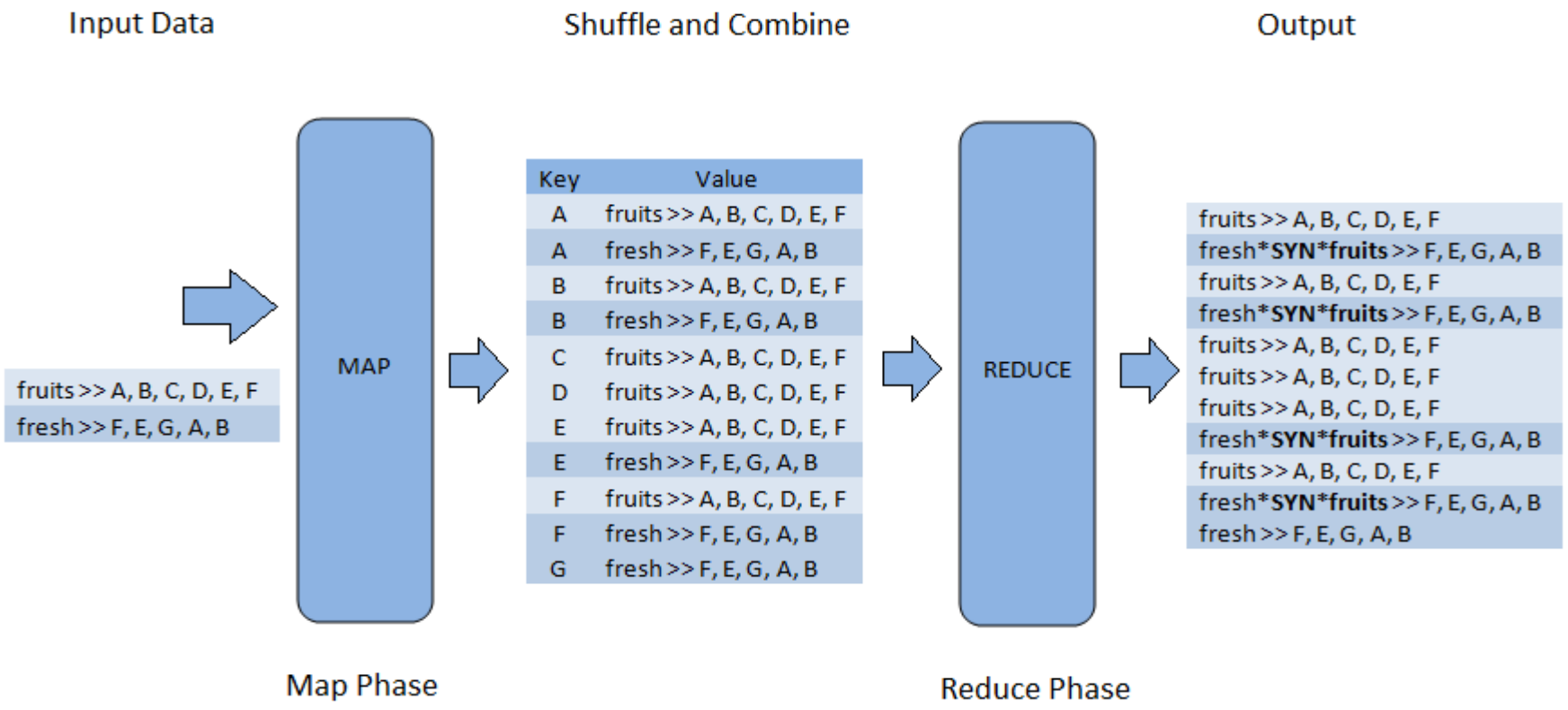
Implementation with Map Reduce

Spelling Difference Identification – Second Job



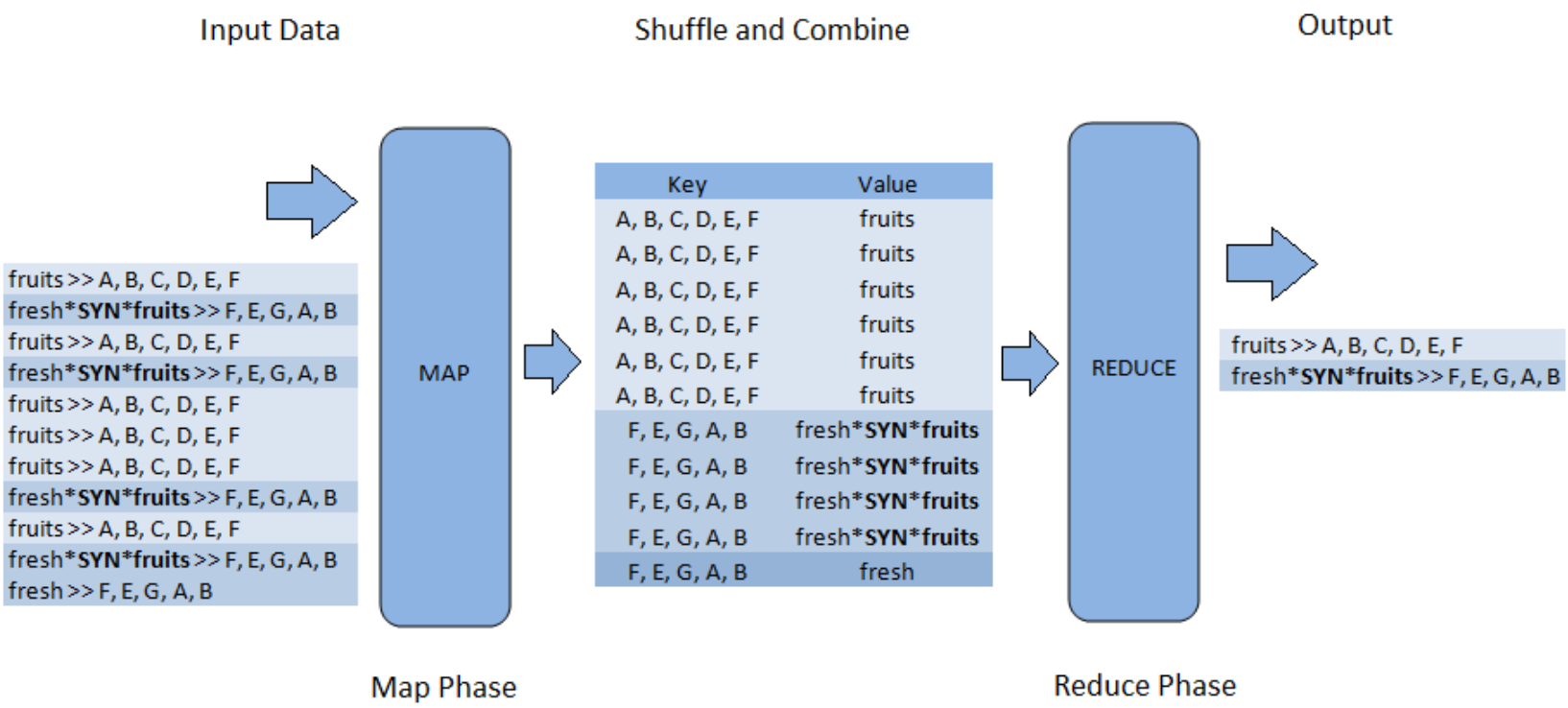
Implementation with Map Reduce

Synonym Identification – First Job

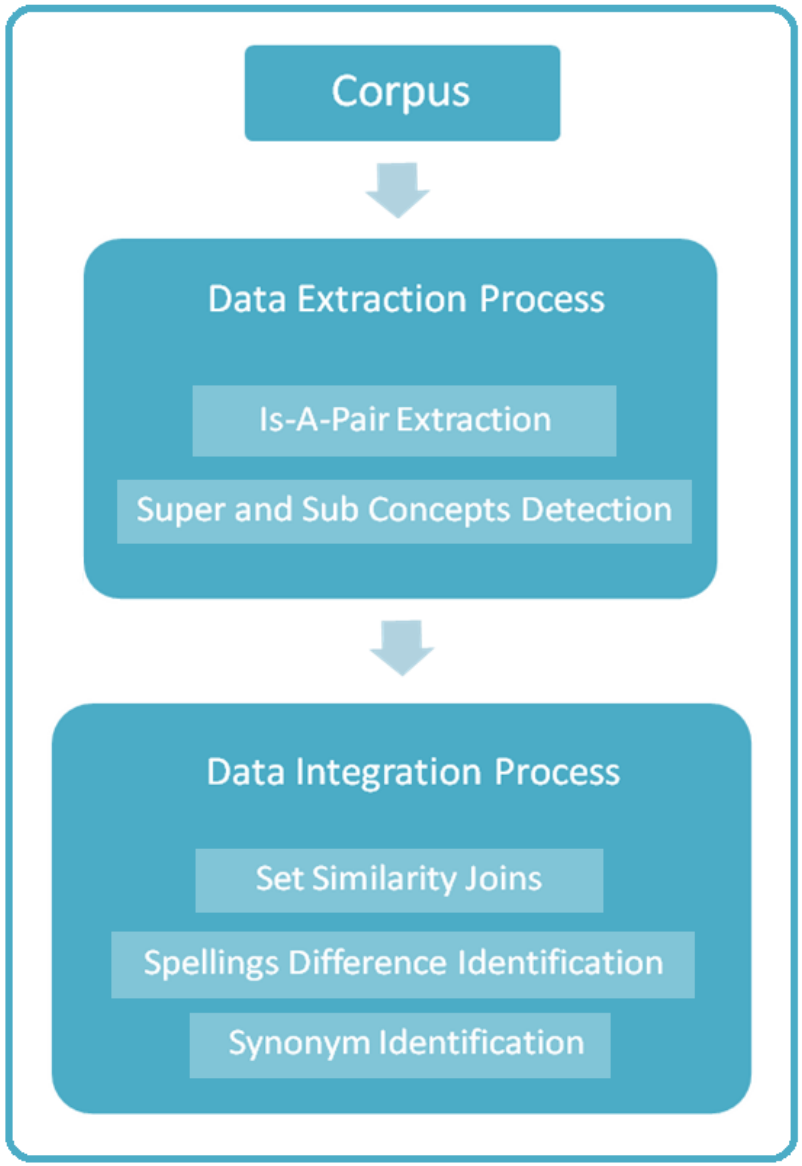


Implementation with Map Reduce

Synonym Identification – Second Job



System Architecture



Outline

- Introduction and Foundations
- Set Similarity Joins
  - Different Similarity Scenarios
- Implementation with Map Reduce
- Analysis and Evaluation
  - Threshold Analysis and Evaluation
  - Efficiency and Scalability Analysis
- Conclusion

Threshold Analysis and Evaluation

Set Similarity Joins: Optimal Threshold Values

Optimal Threshold Values for Set Similarity		
Size Difference	Jaccard Coefficient	Percentage Similarity
1 - 2 Times	>0.15	>30%
2 - 3 Times	>0.1	>50%
3 - 4 Times	>0.08	>60%
>4 Times	>0.05	>65%

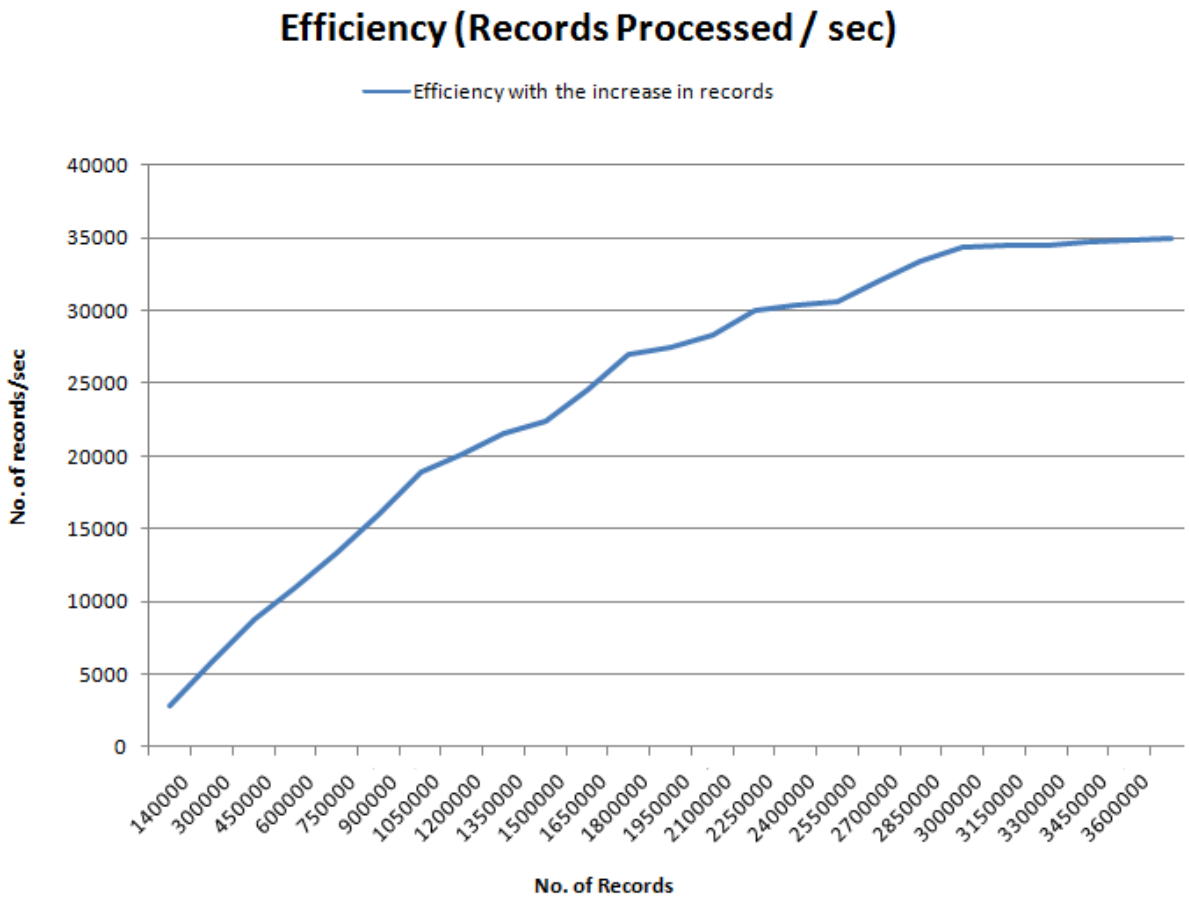
Threshold Analysis and Evaluation

String Similarity: Analysis for Optimal Threshold Value

Threshold Analysis of Dice Coeffient			
No.	String A	String B	Dice Coefficient
1	plant	plants	0.88
2	plant	planted	0.8
3	plant	slant	0.75
4	plant	chant	0.5
5	plant	plan	0.85

Optimal Threshold Value = **0.7**

Efficiency Analysis of Set Similarity Joins



Scalability Analysis of Set Similarity Joins

Details of Experiments

Input Size (MB)	Input Records	Elapsed Time(s)	Output Records	Mappers	Reducers
≈ 2000	142688	51	2220	2	2
≈ 4000	299638	51	4519	4	3
≈ 6000	445584	51	6467	6	4
≈ 8000	587346	54	8430	8	5
≈ 10000	722591	54	10249	10	6
≈ 44000	3171830	92	37955	44	23
≈ 46000	3303678	95	39349	46	24
≈ 48000	3455075	99	40947	48	25
≈ 50000	3596664	103	42493	50	26

Conclusion

- Jaccard Coefficient for Set Similarity
- Size difference and Percentage of Similarity of smaller set with respect to the larger set
- Spelling Difference Identification is important for further reduction of records

Thank You!
&
Questions