

# Subsumption Computation

Gruppe 04: Philipp Berger, Thomas Zimmermann

Seminar „Map/Reduce Algorithms on Hadoop“  
13.07.2009



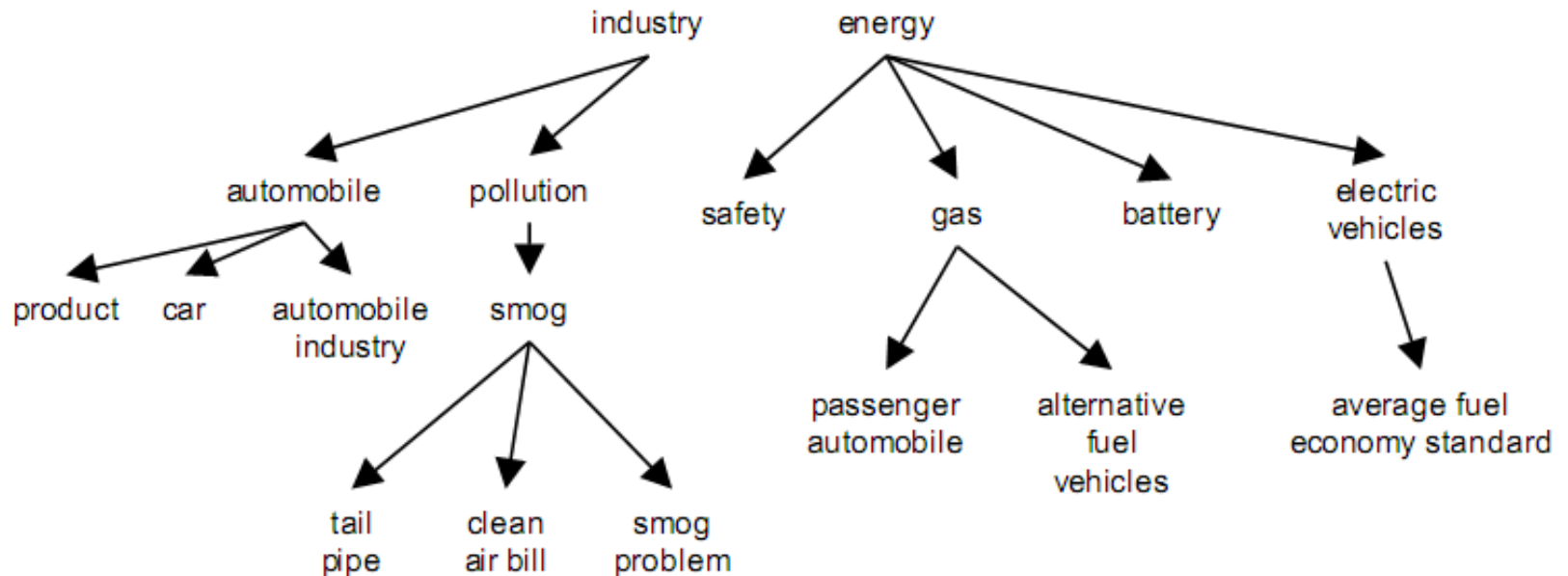
1. Aufgabe
2. Map/Reduce Phasen
3. Laufzeitverhalten

# Aufgabe

3

Subsumption  $\triangleq$  Klassifizierung, Einordnung

→ Falls  $P(x|y) > 0,8$  ist  $x$  wahrscheinlich  $y$  übergeordnet



Quelle: Sanderson, Croft "Deriving concept hierarchies from text"

Problem: Bilde  $P(x|y)$  über große Dokumentmengen

# Aufgabe Beispiel Corpus

4

```

C:\Users\Thomas\Desktop\phenomicdb.txt - Notepad++
File Edit Search View Format Language Settings Macro Run TextFX Plugins Window ?
part-00000 phenomicdb.txt
326793 fbcv0000370 male sterile genetic analysis of the male fertility factors on the y chromosome of drosophila melanogaster genetic variations of dro
326794 fbcv0000370 male sterile cytogenetic analysis of a segment of the y chromosome of drosophila melanogaster analysis of y linked mutations to male
326795 analysis of y linked mutations to male sterility in drosophila melanogaster
326796 mutant phenotype involves either kl or kl temperature sensitive mutations in drosophila melanogaster x1 male sterile mutants of the y chr
326797 mutant phenotype involves either kl or kl temperature sensitive mutations in drosophila melanogaster x1 male sterile mutants of the y chr
326798 mutant phenotype involves either kl or kl temperature sensitive mutations in drosophila melanogaster x1 male sterile mutants of the y chr
326799 a cytogenetic analysis of x ray induced male steriles on the y chromosome of drosophila melanogaster
326800 fbcv0000370 male sterile a cytogenetic analysis of x ray induced male steriles on the y chromosome of drosophila melanogaster
326801 fbcv0000370 male sterile cytogenetic analysis of a segment of the y chromosome of drosophila melanogaster cytological and genetic analysis of th
326802 mutant phenotype male sterile fbcv0000370 male sterile cytogenetic analysis of a segment of the y chromosome of drosophila melanogaster cytol
326803 a cytogenetic analysis of x ray induced male steriles on the y chromosome of drosophila melanogaster
326804 the drosophila heterochromatic gene encoding poly adp ribose polymerase parp is required to modulate chromatin structure during development
326805 fbcv0000380 non e mutant phenotype m4 the dro
326806 mutant phenotype the drosophila he development
326807 the drosophila he development
326808 mutant phenotype pupal stage
326809 overexpression of
326810 mutant phenotype on is seen
326811 ecdysone receptor dependent gene regulation mediates histone poly adp ribosyl ation
326812 ecdysone receptor dependent gene regulation mediates histone poly adp ribosyl ation
326813 the drosophila heterochromatic gene encoding poly adp ribose polymerase parp is required to modulate chromatin structure during development
326814 the drosophila heterochromatic gene encoding poly adp ribose polymerase parp is required to modulate chromatin structure during development
326815 mutant phenotype flies expressing parp 4 scer uas under the control of scer gal4 have rough eyes and defects in the abdominal outcicle file
326816 the drosophila heterochromatic gene encoding poly adp ribose polymerase parp is required to modulate chromatin structure during development t
326817 mutant phenotype heterozygotes have a minute bristle phenotype mutant phenotype heterozygous adults have short slender bristles and eclo
326818 a drosophila fragile x protein interacts with components of rna1 and ribosomal proteins a drosophila fragile x protein interacts with components
326819 green fluorescent protein tagging drosophila proteins at their native genomic loci with small p elements
326820 mutant phenotype heterozygotes have a minute bristle phenotype mutant phenotype heterozygous adults eclose with a delay of hours compar
326821 mutant phenotype disrupt ommochrome and peritidine synthesis mutant phenotype drosopterins drastically reduced no maternal effect in homo
326822 mutant phenotype weak allele el tipo mutante pink wing de drosophila melanogaster un problema de localizacion the mutant pink wing in
326823 mutant phenotype rare homozygotes are short lived and sterile like lt in appearance viability and classificability in lt lt is excellen
326824 fbcv0000351 lethal genetic variations of drosophila melanogaster new mutants report
326825 mutant phenotype hemizygotes are viable and adults exhibit a lt phenotype fbcv0000354 visible genetic analysis of the centromeric heterochrom
326826 mutant phenotype lt hdsla lt flies have wild type viability mutant phenotype hemizygotes die at late pupal stages mutant phenotypy
326827 mutant phenotypy hemizygotes die at late pupal stages fbcv0000351 lethal fbcv0000351 lethal fbcv00005349 pupal stage fbcv0000298 recessiv
4
Normal text file nb char:114596369 Ln:169008 Col:148 Sel:0 UNIX ANSI INS
  
```

~320.000 Dokumente

```

C:\Users\Thomas\workspace\Hadoop\phrases\phrases.txt - Notepad++
File Edit Search View Format Language Settings Macro Run TextFX Plugins
part-00000 phenomicdb.txt phrases.txt
1 morphology
2 anatomy
3 abnormal adipose tissue morphology
4 adipose tissue abnormalities
5 adipose tissue dysplasia
6 increased brown adipose tissue amount
7 increased brown fat
8 increased brown fat amount
9 increased white adipose tissue amount
10 increased white fat
11 increased white fat amount
12 abnormal abdominal fat pads
13 loss of subcutaneous adipose tissue
14 abnormal adipose tissue distribution
15 a
16 a
17 b
18 Small ears
19 microtia
20 thick ears
21 scaly ears
22 prominent ears
23 abnormal ear shape
24 abnormal ear distance position
25 lowered ear position
26 otic hypertelorism
27 hypertelorism of ears
28 increased distance between the ears
29 abnormal inner ear morphology
30 inner ear dysplasia
31 horizontal canal defects
32 abnormal pars superior vestibularis morphology
33 abnormal pars superior vestibularis
34 abnormal malleus morphology
35 abnormal malleus
36 abnormal tympanic ring morphology
Normal text file
  
```

~12.000 Phrases

Gesucht:  $P(x|y)$  für Phrases x und y, die gemeinsam vorkommen

# Überblick der 3 Map/Reduce-Phasen

5

1. Phrases in Dokumenten lokalisieren
2. Zählen, wie oft Phrases einzeln und mit anderen Phrases zusammen vorkommen
3. Berechnen der bedingten Wahrscheinlichkeiten

# Ansatz – 1. Durchgang

6

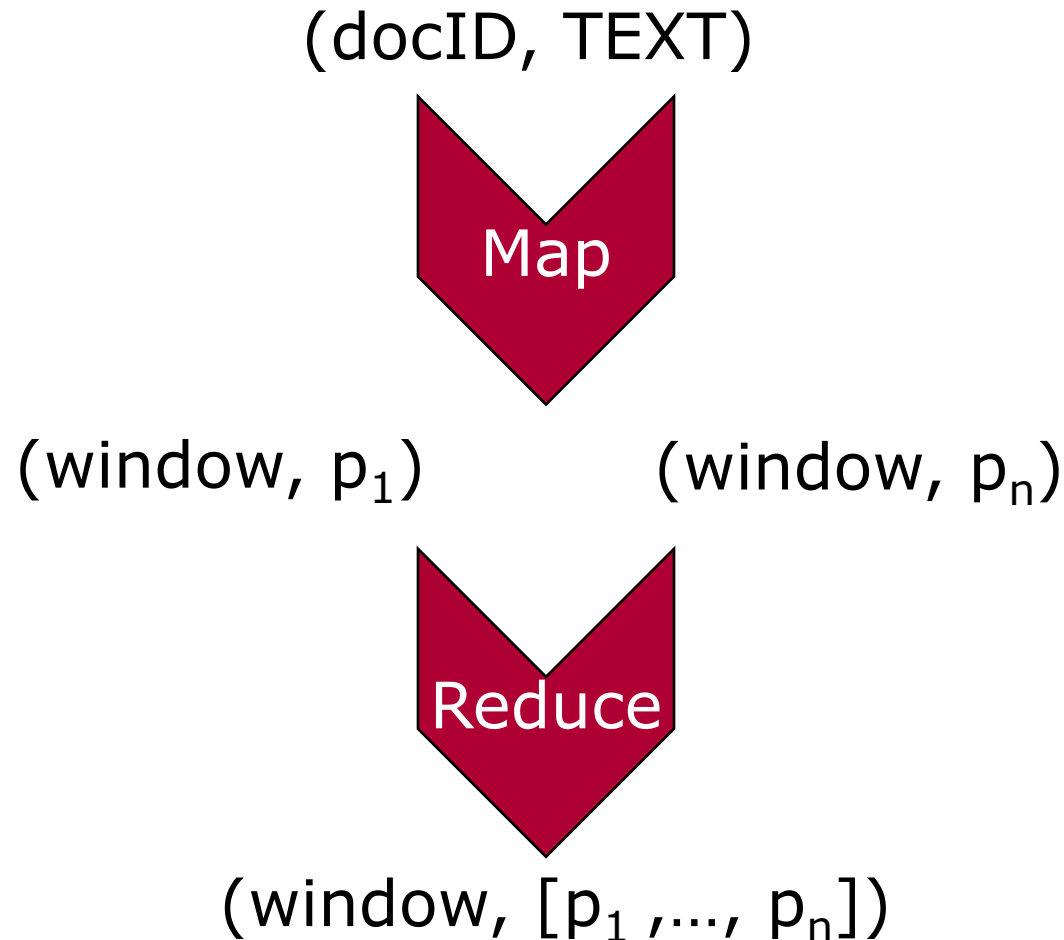
**Wann** tauchen Phrases mit anderen Phrases **zusammen** auf?

Idee: Wenn sie im selben **Sliding Window** vorkommen

```
alopecia areata is a genetically determined immune mediated disorder of the hair follicle with an estimated lifetime
syndrome a microdeletion involving genes on chromosome 1q21 has been found to be necessary but not sufficient to
define a genetic basis of alopecia areata martinez mir et al performed a genomewide search for linkage to fa
c nonsyndromic defects in the urinary tract are the most common cause of end stage renal failure in children and
is mitral valve prolapse mvp is a common disorder characterized by histologic displacement or billowing of the mi
mapping and phenotype information for this qtl its variants and associated markers allele type qtl strain
dmapping and phenotype information for this qtl its variants and associated markers allele type qtl strain
type spontaneous mode of inheritance recessive strain of origin balb cj phenotypic details homeostas
fied mode of inheritance recessive strain of origin balb cj renal urinary system phenotype a mendelian locus
liges allele type targeted knock out strain of origin c57bl phenotypic details life span aging prema
```

# Lösung – 1. Durchgang

7



# Vergleichsalgorithmen

8 **public interface Similarity {**

**public double computeSimilarity(String window,  
String phrase);**

**}**

String.contains-Methode

Aufwand:  $(\text{Länge Window} - \text{Länge Phrase}) * \text{Länge Phrase}$

Levenshtein-Distance (Edit-Distanz)

Aufwand:  $(\text{Länge Window} - \text{Länge Phrase}) * (\text{Länge Phrase})^2$

Levenshtein-Distance auf Wort-Ebene

Aufwand:  $\text{Anzahl Wörter in Window} * (\text{Anzahl Wörter in Phrase})^2$



## Ansatz – 2. Durchgang

9

### **Wie oft** kommen Phrases vor?

Idee: Mapper bildet Phrase-Teilmengen pro Window

1-elementige: Vorkommen einzelner Phrases

2-elementige: Vorkommen von paarweisen Phrases

Reducer summiert Vorkommen auf

# Lösung – 2. Durchgang

10

$(\text{WINDOW}, [p_1, p_2])$



$(p_1, 1), (p_2, 1), ([p_1, p_2], 1)$



$(p_1, x), (p_2, y), ([p_1, p_2], z)$

## Ansatz – 3. Durchgang

11

Wie lauten die **bedingten Wahrscheinlichkeiten**?

Idee: Mapper sammelt für alle Phrases:

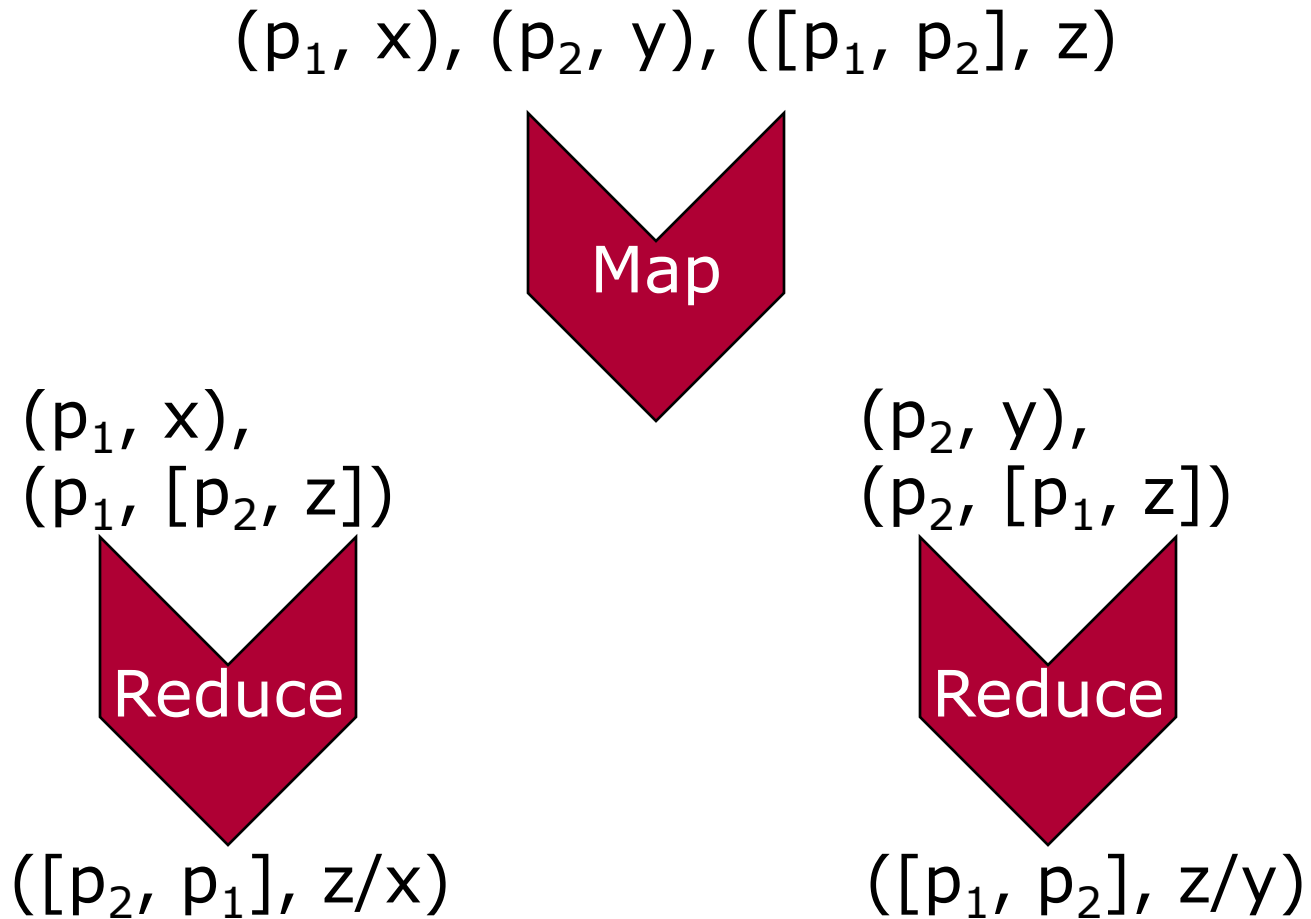
- Absolutes Vorkommen
- Anzahl Vorkommen mit anderen Phrases

Reducer kann pro Phrase  $x$  alle  $P(x|y)$  für die  $y$ , mit denen  $x$  auftaucht, berechnen

$$P(x|y) = \frac{P(x \cap y)}{P(y)}$$

# Lösung – 3. Durchgang

12



# Problem beim 3. Reducer

13

$$P(x|y) = \frac{P(x \cap y)}{P(y)}$$

**worst case:**

- (p<sub>2</sub>, [p<sub>1</sub>, z])
- (p<sub>2</sub>, [p<sub>3</sub>, z])
- (p<sub>2</sub>, [p<sub>2</sub>, z])
- (p<sub>2</sub>, [p<sub>8</sub>, z])
- (p<sub>2</sub>, [p<sub>4</sub>, z])
- (p<sub>2</sub>, [p<sub>9</sub>, z])
- (p<sub>2</sub>, [p<sub>11</sub>, z])
- (p<sub>2</sub>, y)

12000 phrases x

(4 Byte phraseID + 4 Byte relative Häufigkeit)

= 96 KByte

## Secondary Sort

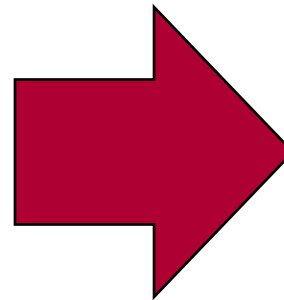
Eigene Pair-Klasse implementiert

→ Sortierung nach beiden Werten

→ Verteilung auf Reducer nur nach erstem Wert

Output des Mappers

$([p_2, p_1], [p_1, z_1])$   
 $([p_2, p_{11}], [p_{11}, z_{11}])$   
 $([p_2, p_3], [p_3, z_3])$   
 $([p_2, p_9], [p_9, z_9])$   
 **$([p_2, " "], [" ", y])$**



Input des Reducers

**$([p_2, " "], [" ", y])$**   
 $([p_2, p_1], [p_1, z_1])$   
 $([p_2, p_3], [p_3, z_3])$   
 $([p_2, p_9], [p_9, z_9])$   
 $([p_2, p_{11}], [p_{11}, z_{11}])$

# Theoretischer Aufwand

16

$\#windows = (doccount * (avg-doc-length - win-size))$

$\#computeSimilarity\text{-Aufrufe} = \#phrases * \#windows$

$Aufwand = Similarity\text{-Aufwand} * \#computeSimilarity\text{-Aufrufe}$

# Theoretischer Aufwand - Beispiel

17

Levenshtein-Distance auf Wort-Ebene

Aufwand: Anzahl Wörter in Window \* (Anzahl Wörter in Phrase)<sup>2</sup>

$$\#windows = (320.000 * (66 - 20)) = 14.720.000$$

$$\# \text{ computeSimilarity-Aufrufe} = 12.000 * 14.720.000 = 176.640.000.000$$

$$\text{Aufwand} = (20 * 3^2) * 176.640.000.000 = \mathbf{3,17952 \times 10^{13}}$$



# Laufzeitverhalten

18

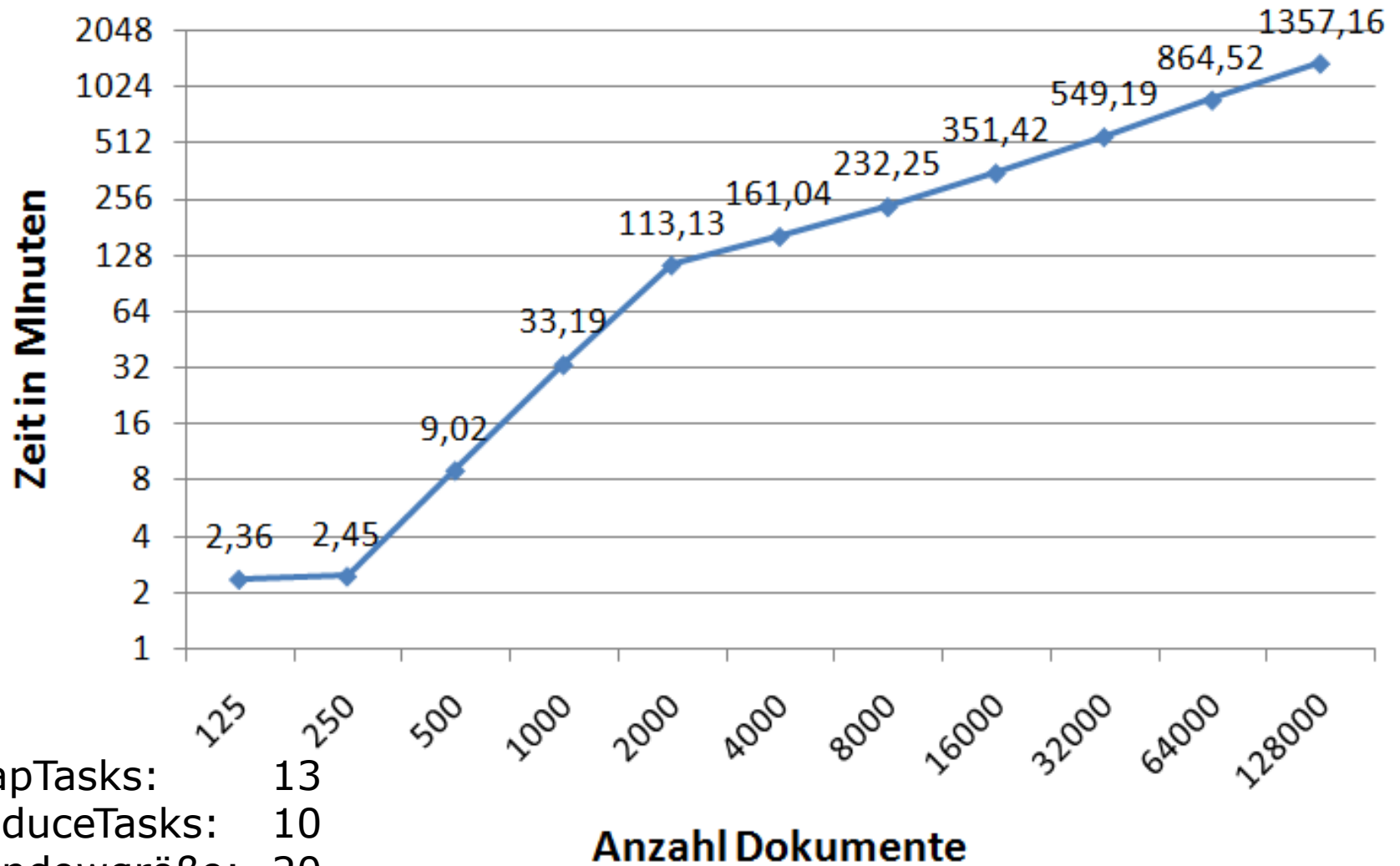
Messungen für folgende Größen:

- Anzahl der Dokumente
- Anzahl MapTasks
- Anzahl Reducer
- Größe des Windows (in Worten)

Umgebung:  
Cluster mit 9 Knoten

# Variable Größe: Dokumentenanzahl

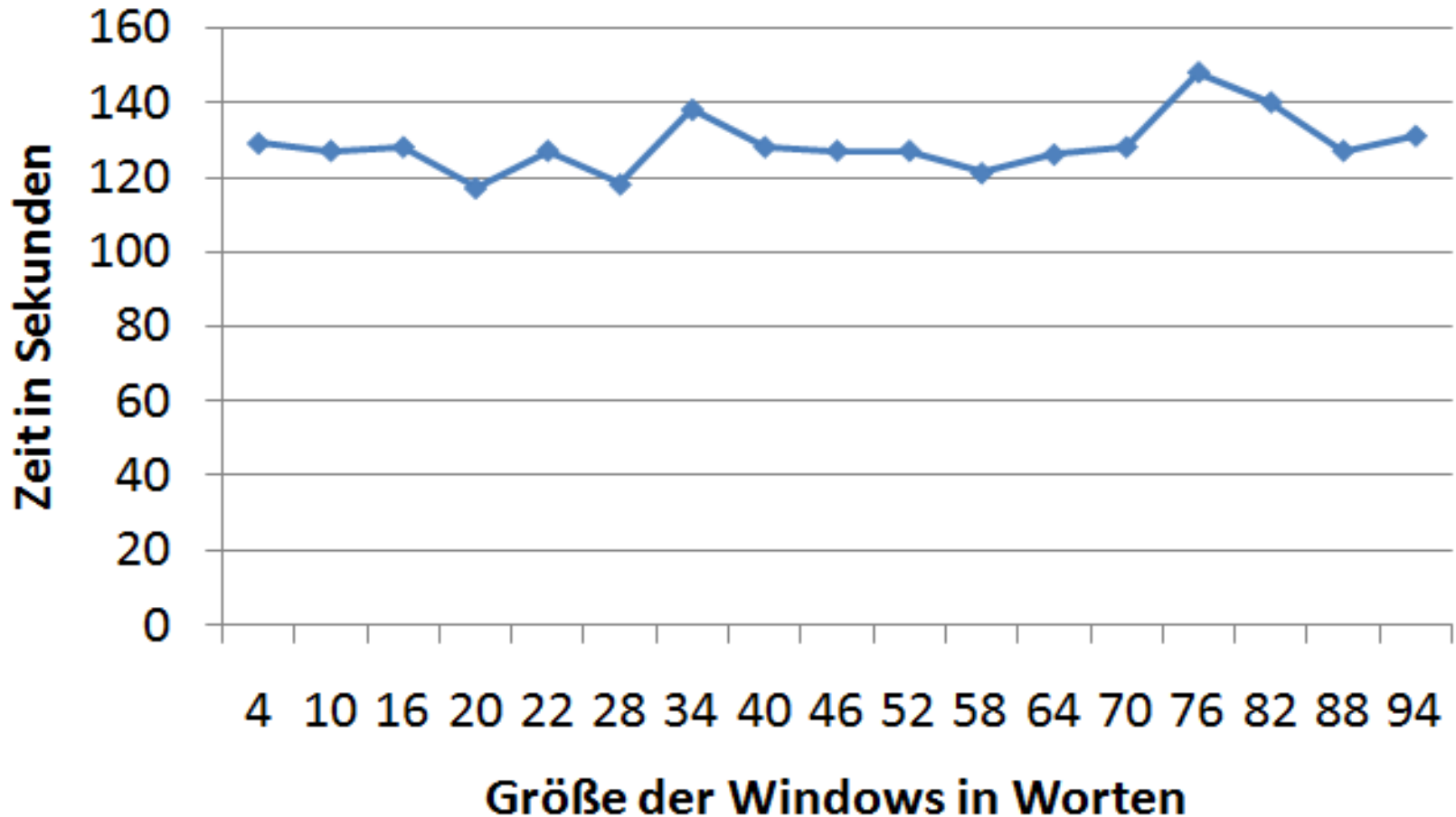
19



MapTasks: 13  
 ReduceTasks: 10  
 Windowgröße: 20

# Variable Größe: Window-Größe

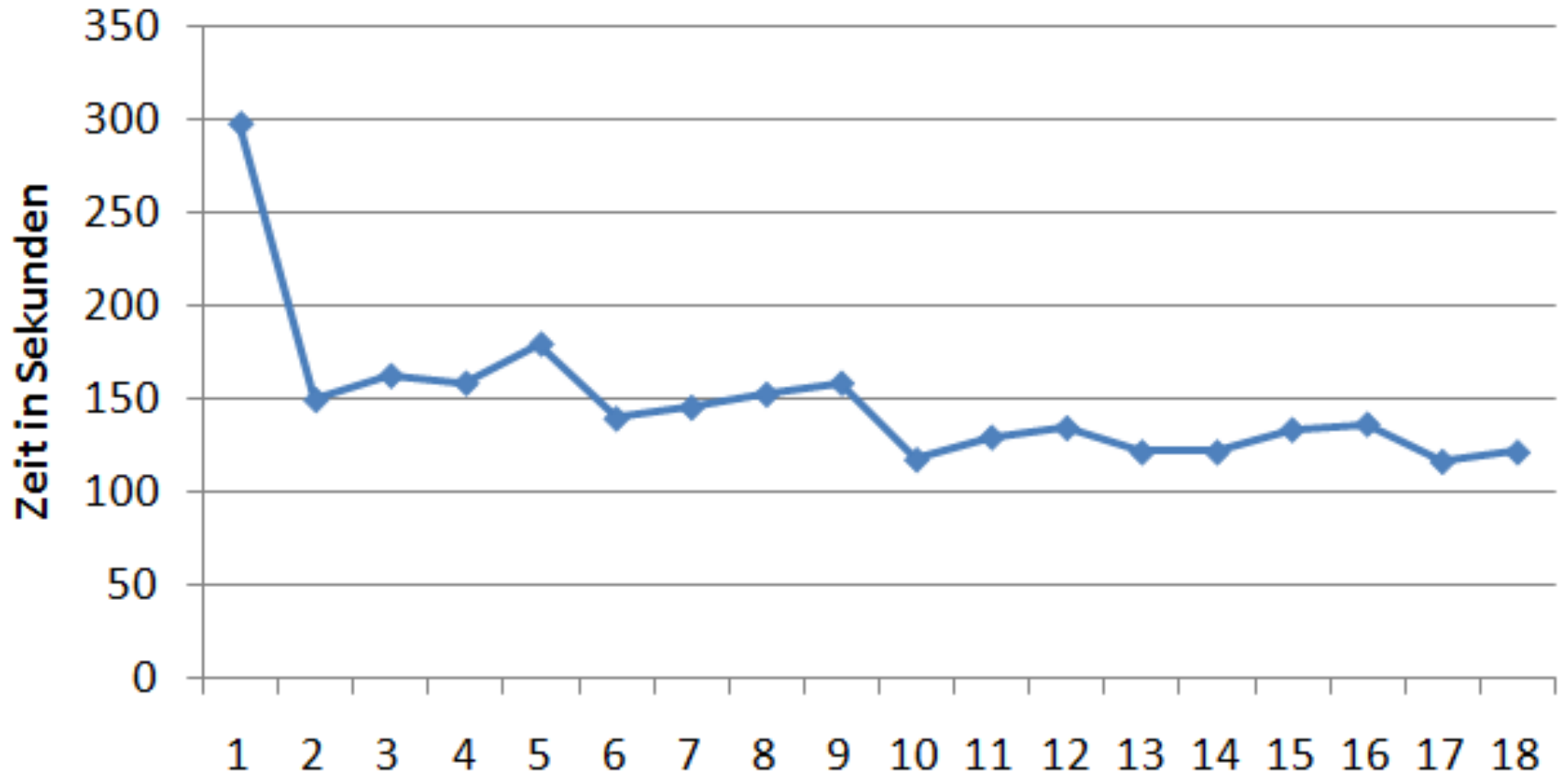
20



MapTasks: 10  
 ReduceTasks: 4  
 Dokumentenanzahl: 125

# Variable Größe: MapTasks

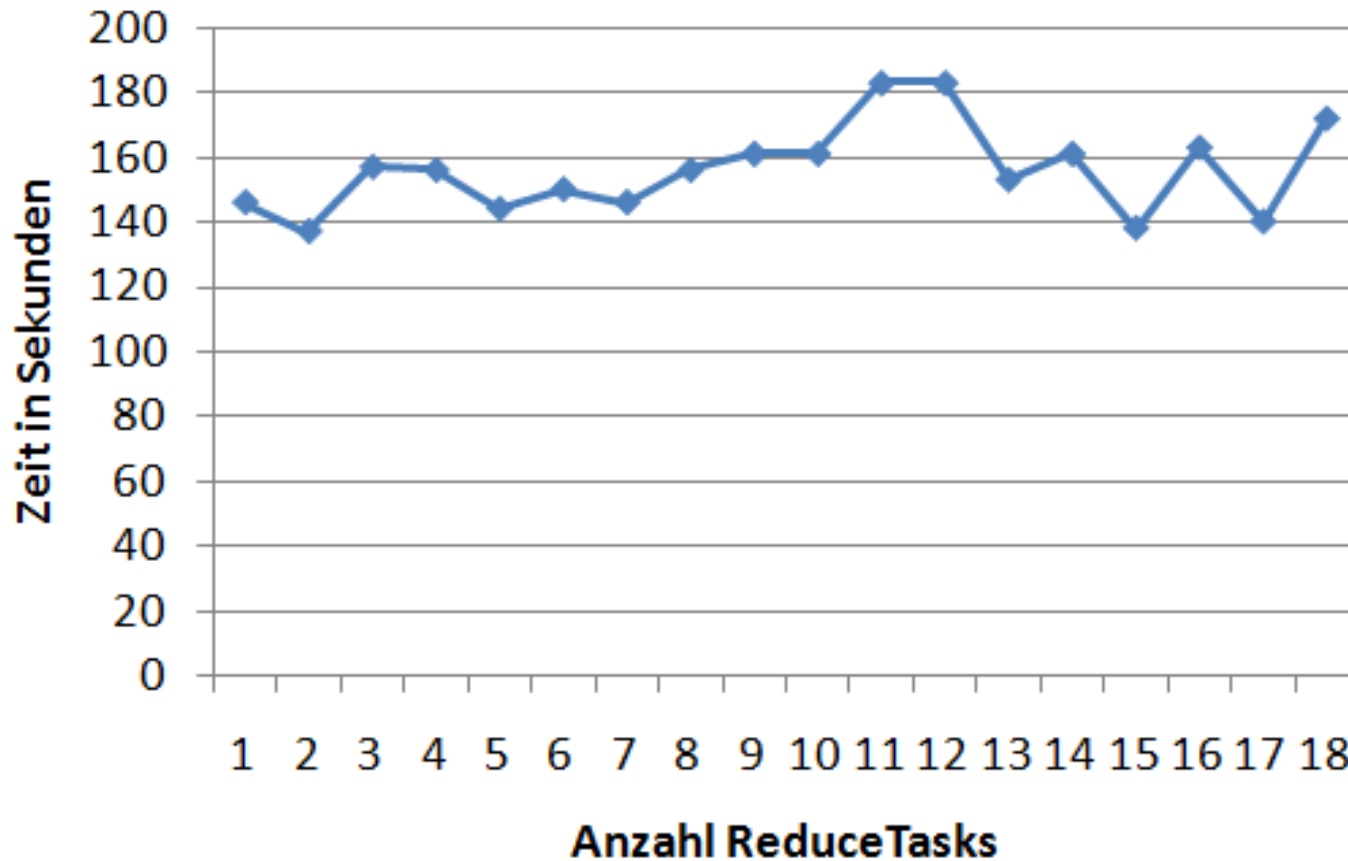
21



ReduceTasks: 1                      **Anzahl MapTasks**  
 Dokumentenanzahl: 125  
 Windowgröße: 20

# Variable Größe: ReduceTasks

22



MapTasks: 13  
 Windowgröße: 20  
 Dokumentenanzahl: 125

**Danke** ? ? ? ? ? ? **keit!**