

MapReduce

Seminar

Large scale data mining mit Apache Mahout
Wintersemester 2009/2010



Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

<http://www.dima.tu-berlin.de/>

Oleg Mayevskiy

Betreuer: Isabel Drost



Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

<http://www.dima.tu-berlin.de/>

- Einleitung
- Programmiermodel
- Verarbeitungsablauf
- Beispiele
- Diskussion

- Funktionales Programmiermodell
- Nebenläufigkeit
- map & reduce - Methoden
- unkompliziert

- Java, C++, C#
- Cluster, Shared Memory, NUMA
Multiprozessorsysteme
- Apache Hadoop MapReduce, Google
MapReduce

- funktional
- beschränkt auf nur zwei Methoden:
 - map (in_key, in_value) ->
 (out_key, intermediate_value) list
 - reduce (out_key,intermediate_value list) ->
 out_value list

```
map(String key, String value):
```

```
// key: document name
```

```
// value: document contents
```

```
for each word w in value:
```

```
    EmitIntermediate(w, "1");
```

```
reduce(String key, Iterator values):
```

```
// key: a word
```

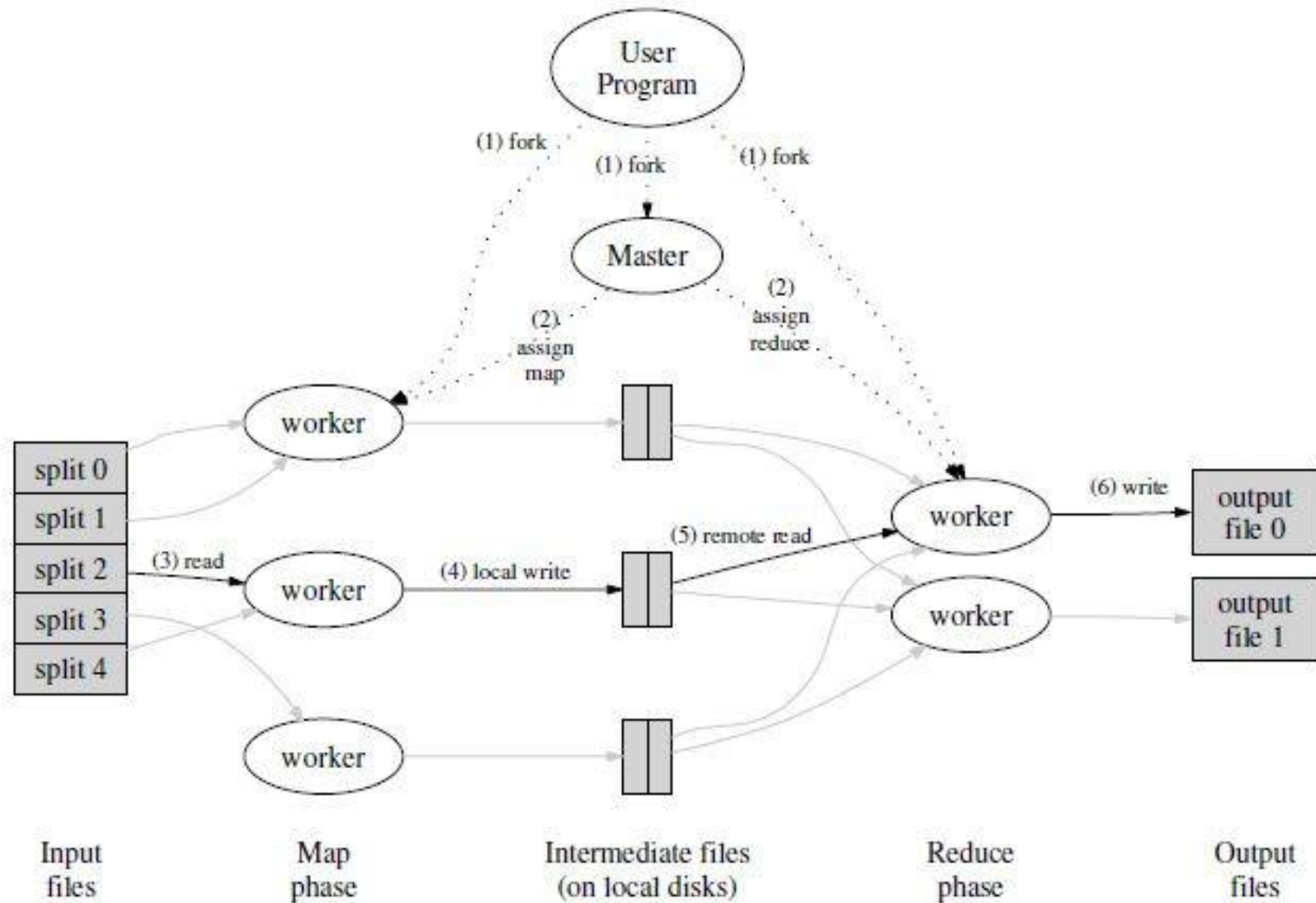
```
// values: a list of counts
```

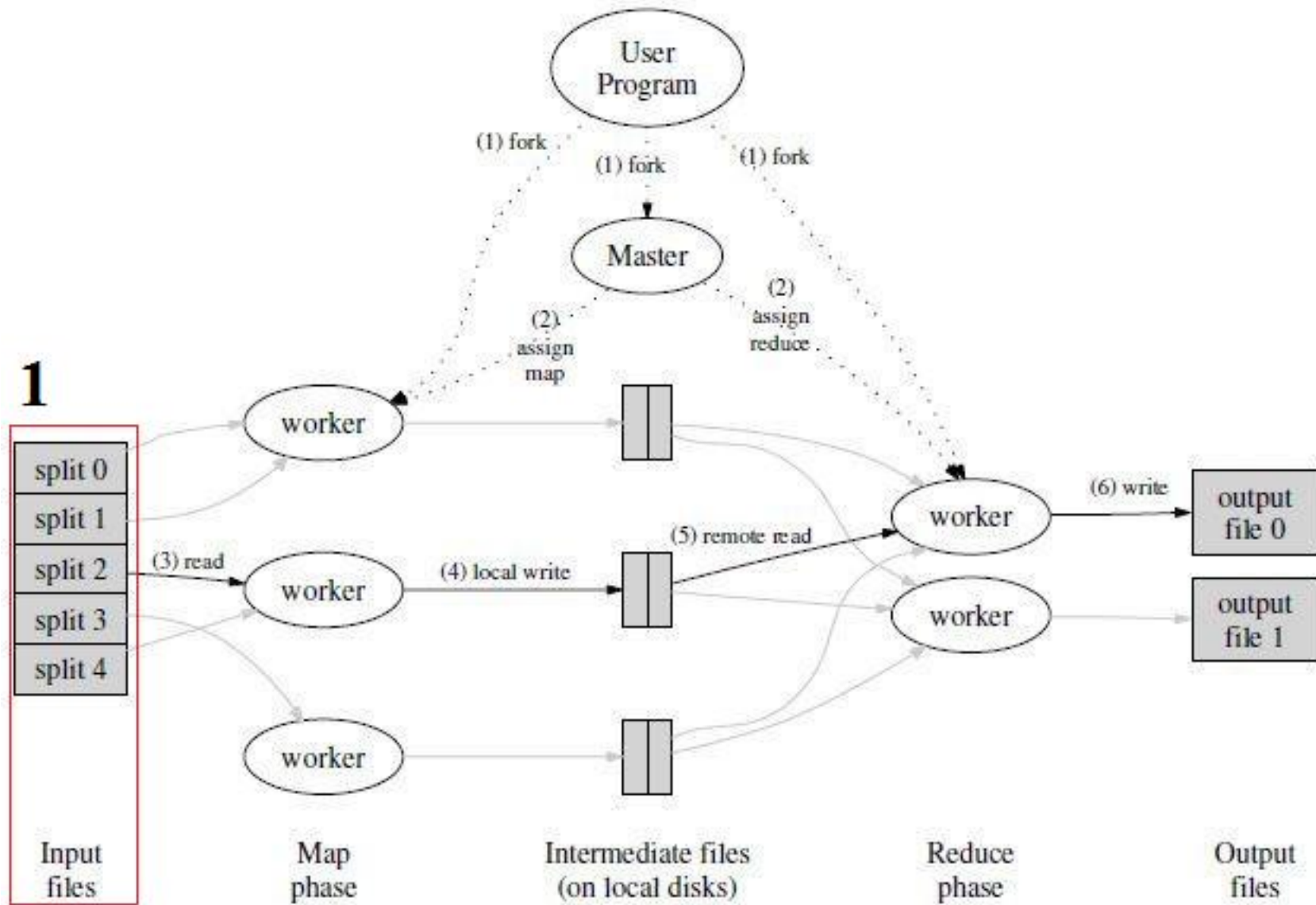
```
int result = 0;
```

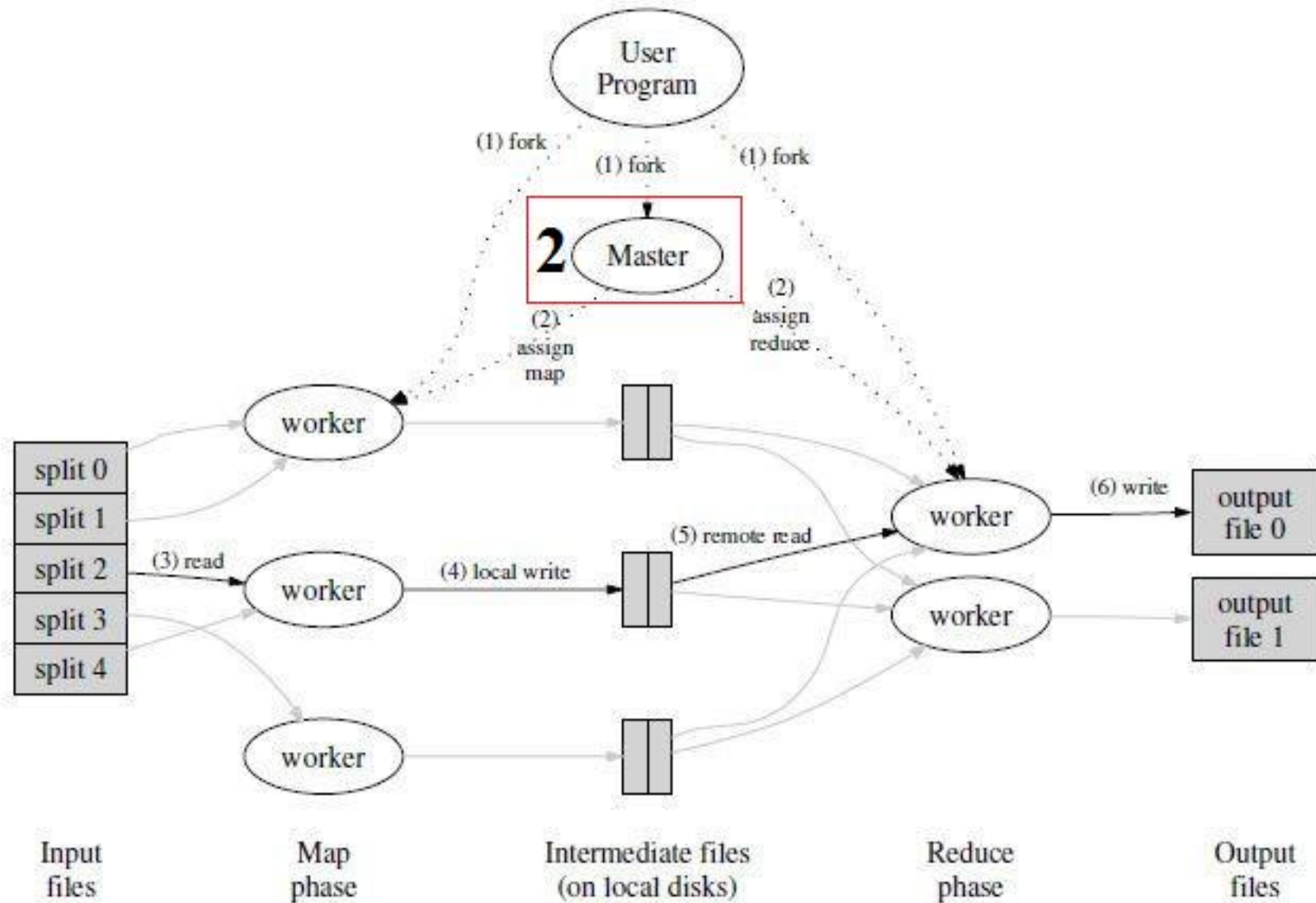
```
for each v in values:
```

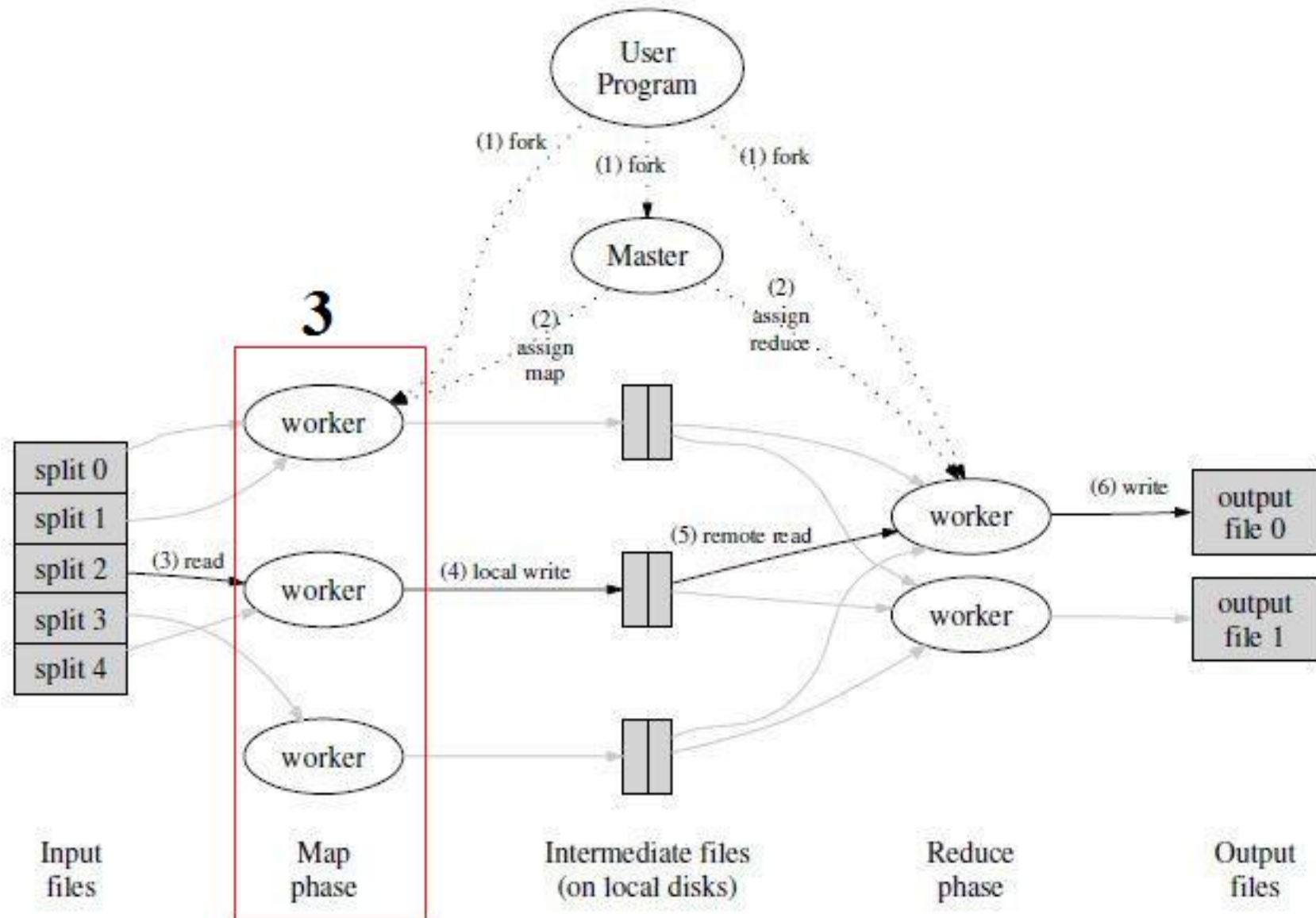
```
    result += ParseInt(v);
```

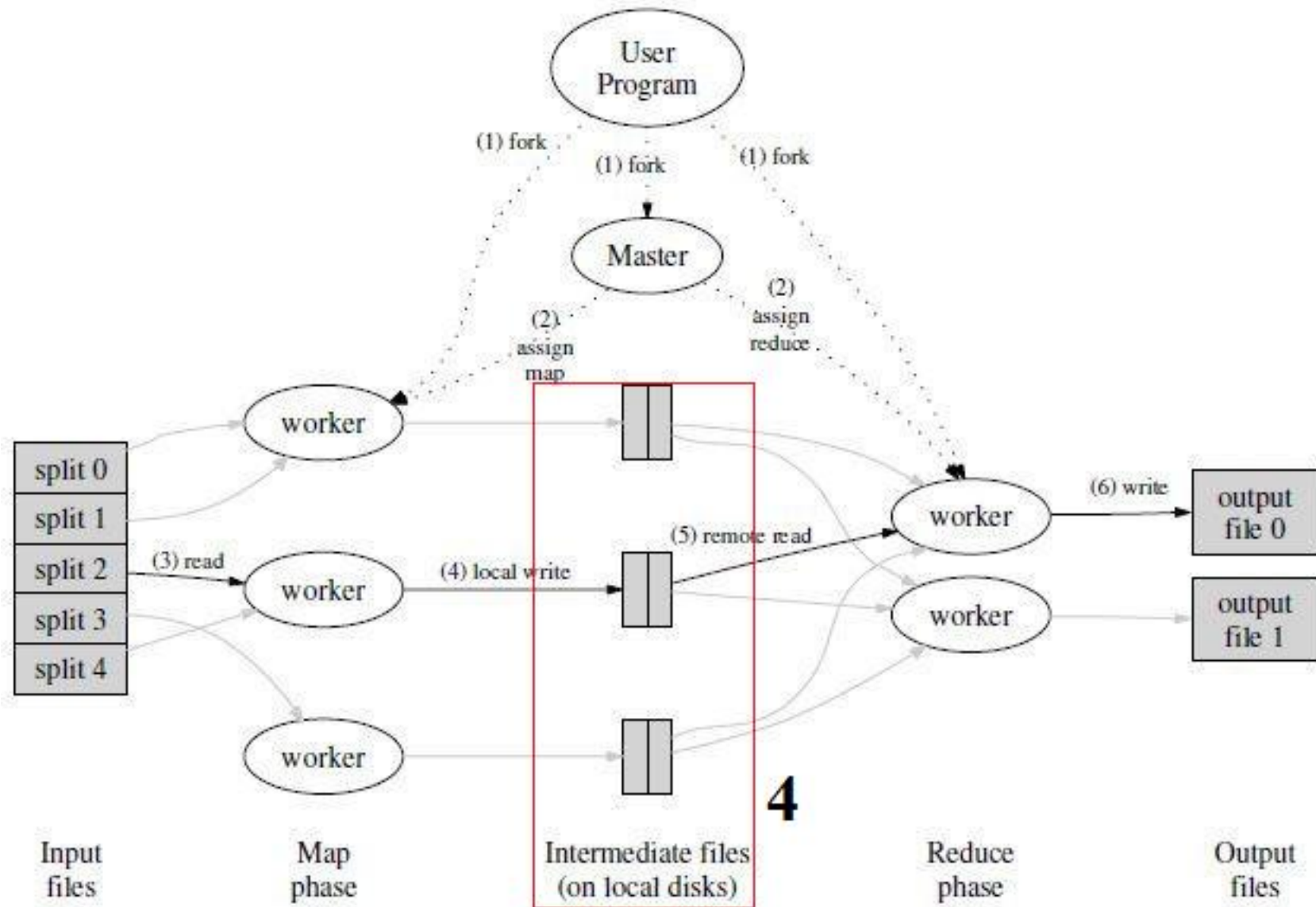
```
Emit(AsString(result));
```

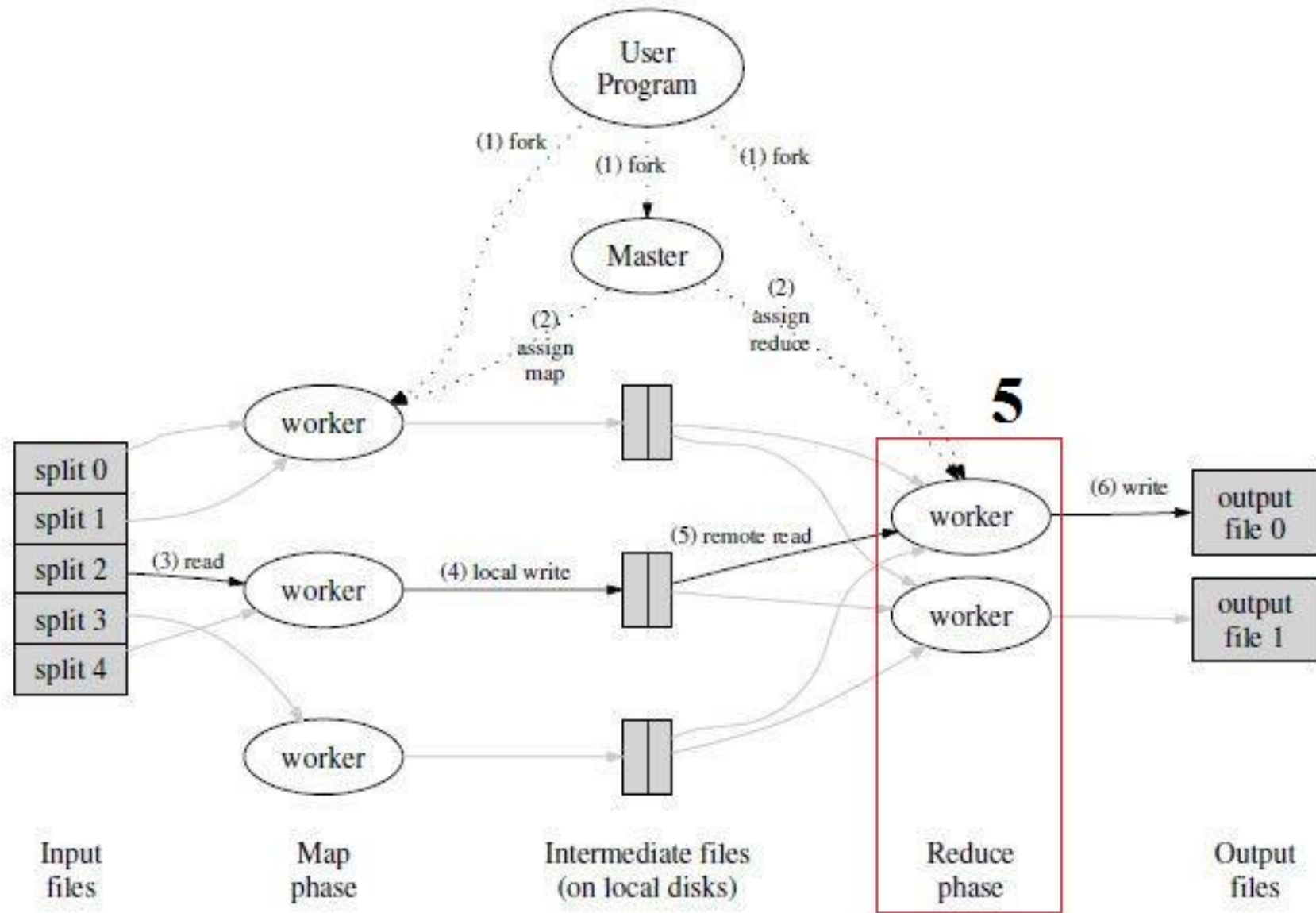


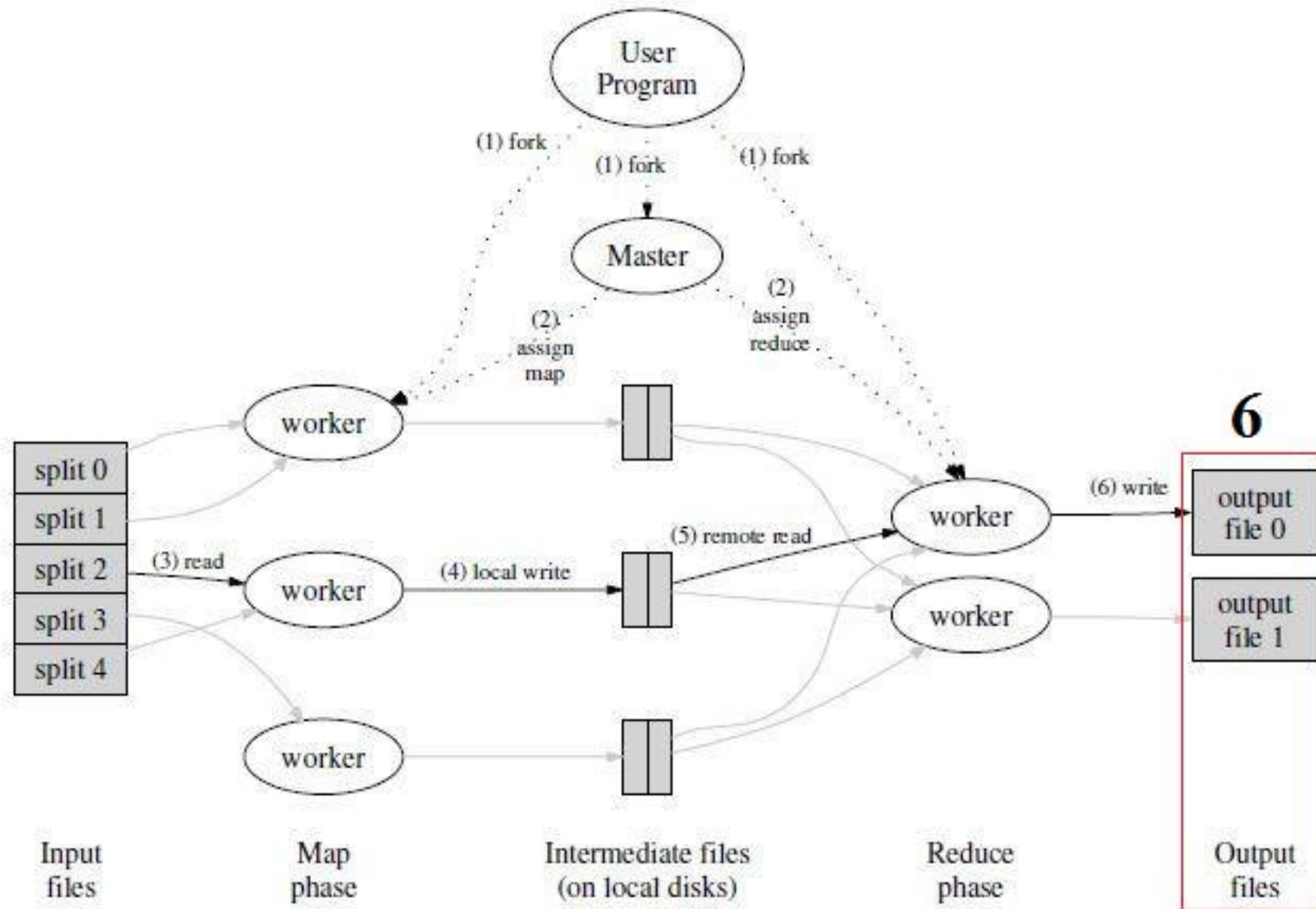












Verteiltes Grep

- Map Funktion liefert Zeilen zurück, die einem gesuchten Pattern entsprechen
- Reduce Funktion ist eine Identitätsfunktion, die das Ergebnis in eine Datei schreibt

Count of URL Access Frequency

- Map Funktion bearbeitet Logdateien eines Webservers und liefert $\langle \text{Url}, 1 \rangle$ zurück
- Reduce Funktion aggregiert die Ergebnisse für jede Url und liefert $\langle \text{Url}, \text{gesamtAnzahl} \rangle$ zurück.

Web-Link Graph

- Die Map Funktion parst eine Webseite und liefert $\langle \text{target}, \text{source} \rangle$ zurück, wobei "target" ein Link, der in der Webseite "source" gefunden wurde, ist.
- Die Reduce Funktion fasst die Ergebnisse zusammen und liefert $\langle \text{target}, \text{list}(\text{source}) \rangle$

- Begriff MapReduce
- Funktionales Programmiermodell
- Interner Verarbeitungsablauf
- Problembeispiele

Haben Sie noch Fragen?

Fragen!