Beispiel: Wortzählung mit Hadoop Streaming API

Voraussetzungen
- Installiere Hadoop auf deinem lokalen System oder verwende einen Hadoop-Cluster.
- Stelle sicher, dass Python installiert ist.

Schritt 1: Mapper und Reducer Skripte erstellen
1. Mapper-Skript(mapper.py):

```
python
#!/usr/bin/env python3
import sys

# Lese die Eingabedaten zeilenweise
for line in sys.stdin:
    # Entferne führende und nachgestellte Leerzeichen
    line = line.strip()
    # Zerlege die Zeile in Wörter
    words = line.split()
    # Gib jedes Wort mit dem Wert 1 aus
    for word in words:
        print(f'{word}\t1')
```

2. Reducer-Skript (reducer.py):

```
python
#!/usr/bin/env python3
import sys

current_word = None
current_count = 0
word = None

# Lese die Eingabedaten zeilenweise
for line in sys.stdin:
    # Entferne führende und nachgestellte Leerzeichen
    line = line.strip()
    # Zerlege die Zeile in Wort und Zähler
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        continue

    if current_word == word:
        current_count += count
    else:
        if current_word:
            print(f'{current_word}\t{current_count}')
        current_count = count
        current_word = word
```

```
    if current_word == word:
        print(f'{current_word}\t{current_count}')
```

Schritt 2: Skripte ausführbar machen
Die Skripte ausführbar machen:
```
bash
chmod +x mapper.py
chmod +x reducer.py
```

Schritt 3: Beispiel-Datendatei erstellen
Erstelle eine Datei `input.txt` mit folgendem Inhalt:

```
Hello Hadoop
Hello MapReduce
Hello World
```

Schritt 4: MapReduce-Job mit Hadoop ausführen
Führe den MapReduce-Job mit Hadoop Streaming API aus:
```
bash
hadoop jar /path/to/hadoop-streaming.jar \
    -input /path/to/input.txt \
    -output /path/to/output \
    -mapper /path/to/mapper.py \
    -reducer /path/to/reducer.py
```

Schritt 5: Ausgabe überprüfen
Nach dem erfolgreichen Abschluss des Jobs können Sie die Ausgabe überprüfen:
```
bash
hadoop fs -cat /path/to/output/part-00000
```

Die Ausgabe sollte die Wortzählung enthalten:

```
Hadoop  1
Hello   3
MapReduce  1
World  1
```

Erklärungen

1. Mapper-Skript (mapper.py)
   - Liest die Eingabezeilen und zerlegt sie in Wörter.
   - Gibt jedes Wort mit einem Wert von `1` aus, getrennt durch ein Tabulatorzeichen (`\t`).

2. Reducer-Skript (reducer.py)
   - Liest die Ausgabe des Mappers und summiert die Werte für jedes Wort.
   - Gibt das Wort und die Gesamtanzahl der Vorkommen aus.