Cloud Computing

WordCount

Dell Zhang

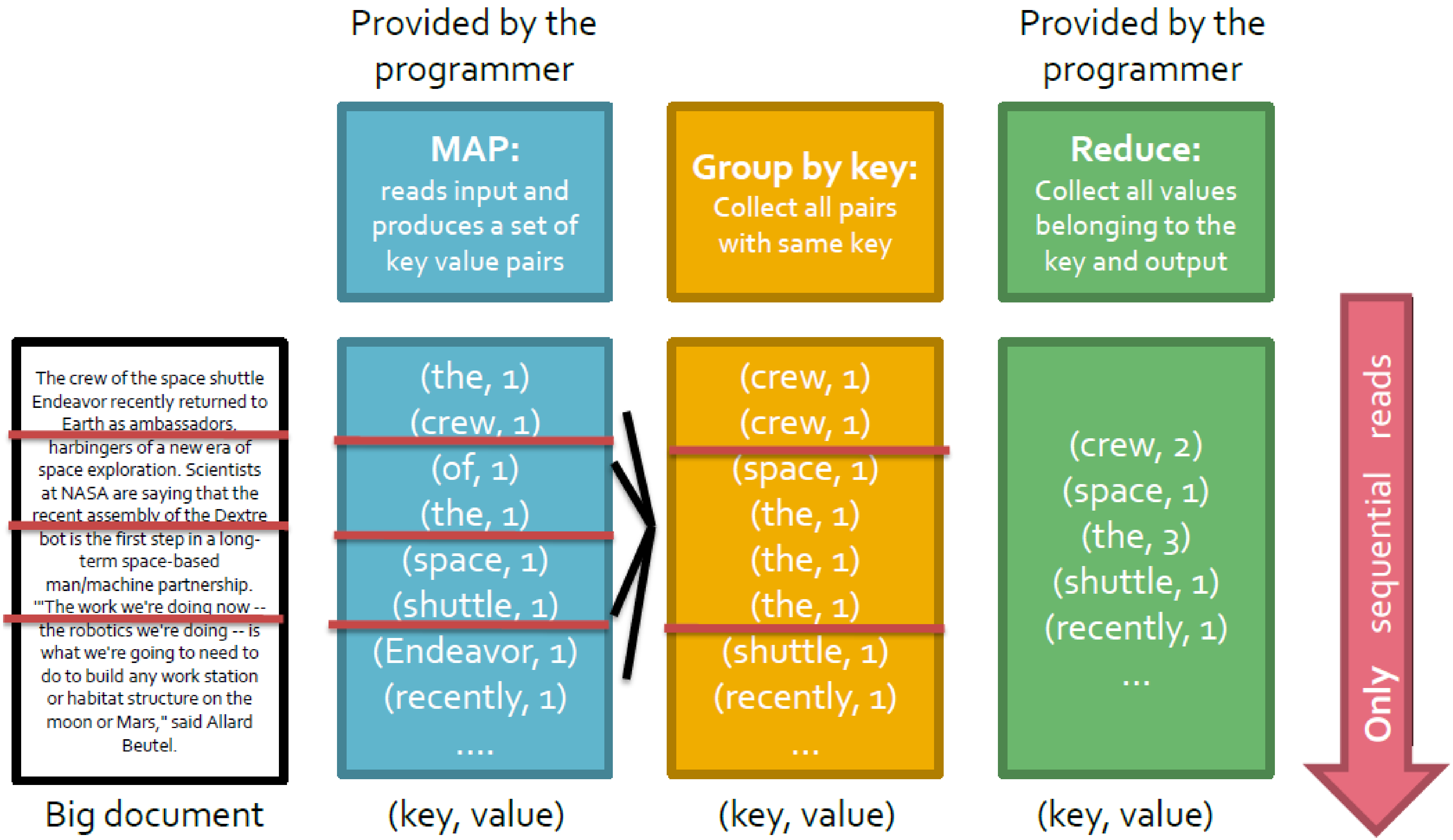
Birkbeck, University of London

2018/19

Warm-Up

- The task:
  - We have a huge text document
  - Count the number of times each distinct word appears in the file
- Sample application:
  - Analyse web server logs to find popular URLs

“Hello World”: Word Count



Provided by the programmer

Provided by the programmer

MAP: reads input and produces a set of key value pairs

Group by key: Collect all pairs with same key

Reduce: Collect all values belonging to the key and output

The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership. '"The work we're doing now -- the robotics we're doing -- is what we're going to need to do to build any work station or habitat structure on the moon or Mars," said Allard Beutel.

Big document

(the, 1)
(crew, 1)
(of, 1)
(the, 1)
(space, 1)
(shuttle, 1)
(Endeavor, 1)
(recently, 1)
....

(key, value)

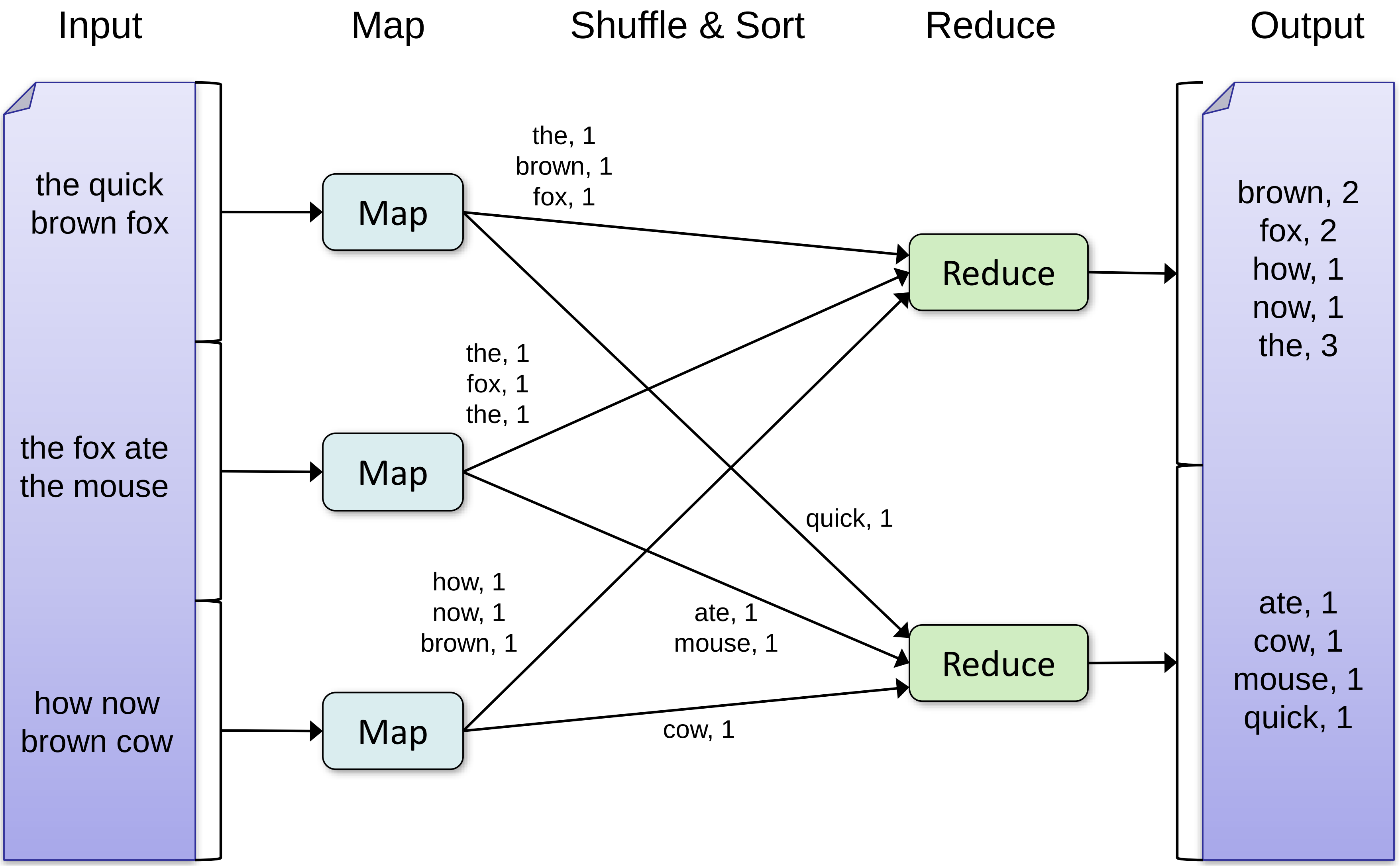
(crew, 1)
(crew, 1)
(space, 1)
(the, 1)
(the, 1)
(the, 1)
(shuttle, 1)
(recently, 1)
…

(key, value)

(crew, 2)
(space, 1)
(the, 3)
(shuttle, 1)
(recently, 1)
…

(key, value)

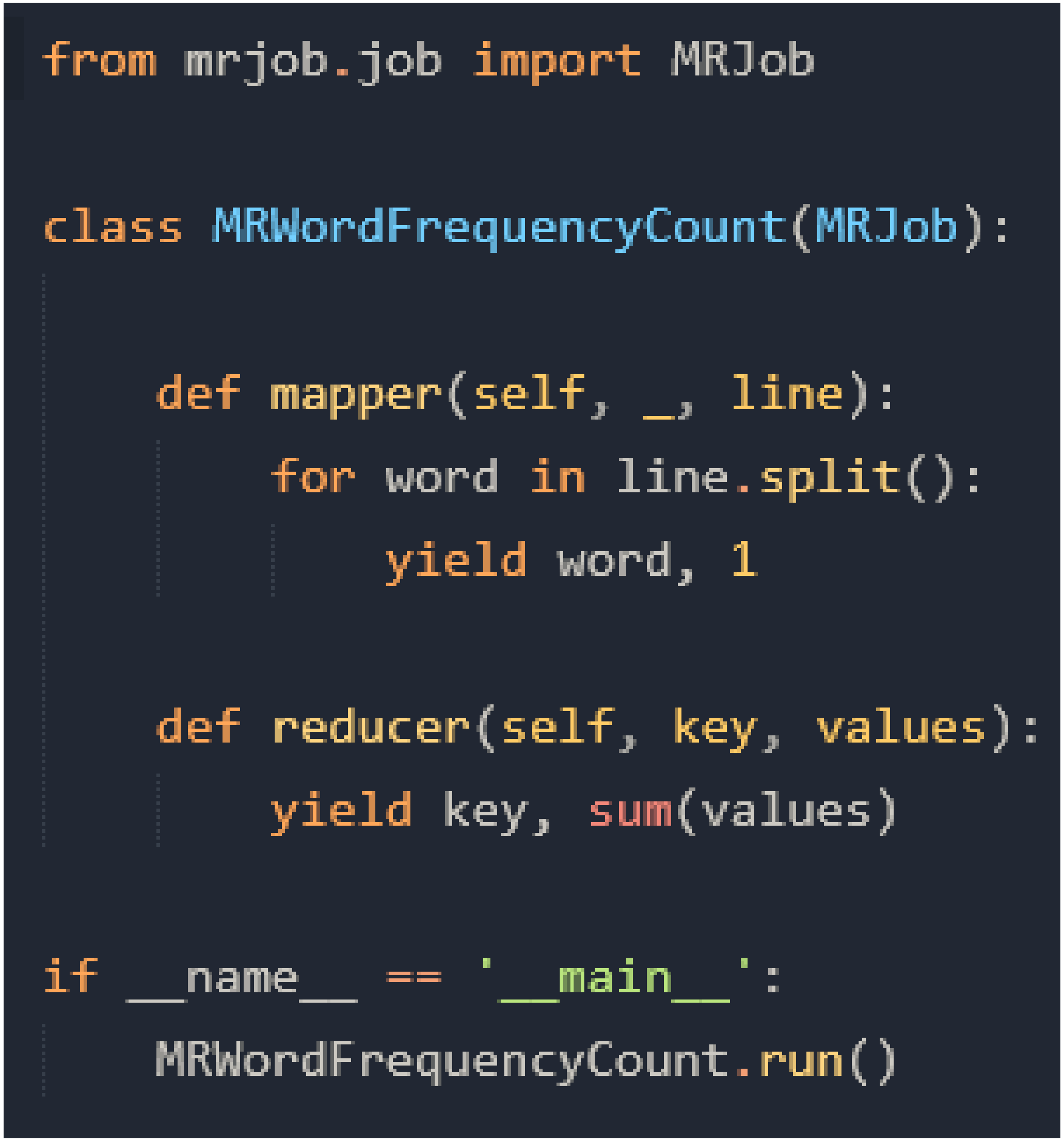
“Hello World”: Word Count



“Hello World”: Word Count

```
map(key, value):
// key: document name; value: text of the document
    for each word w in value:
        emit(w, 1)

reduce(key, values):
// key: a word; values: an iterator over counts
    result = 0
    for each count v in values:
        result += v
    emit(key, result)
```



```python
from mrjob.job import MRJob

class MRWordFrequencyCount(MRJob):

    def mapper(self, _, line):
        for word in line.split():
            yield word, 1

    def reducer(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    MRWordFrequencyCount.run()
```