MapReduce

- MapReduce is a programming model for performing parallel processing on large data sets.

- Imagine we have a collection of items we'd like to process somehow.

- For instance, the items might be website logs, the texts of various books, image files, or anything else.

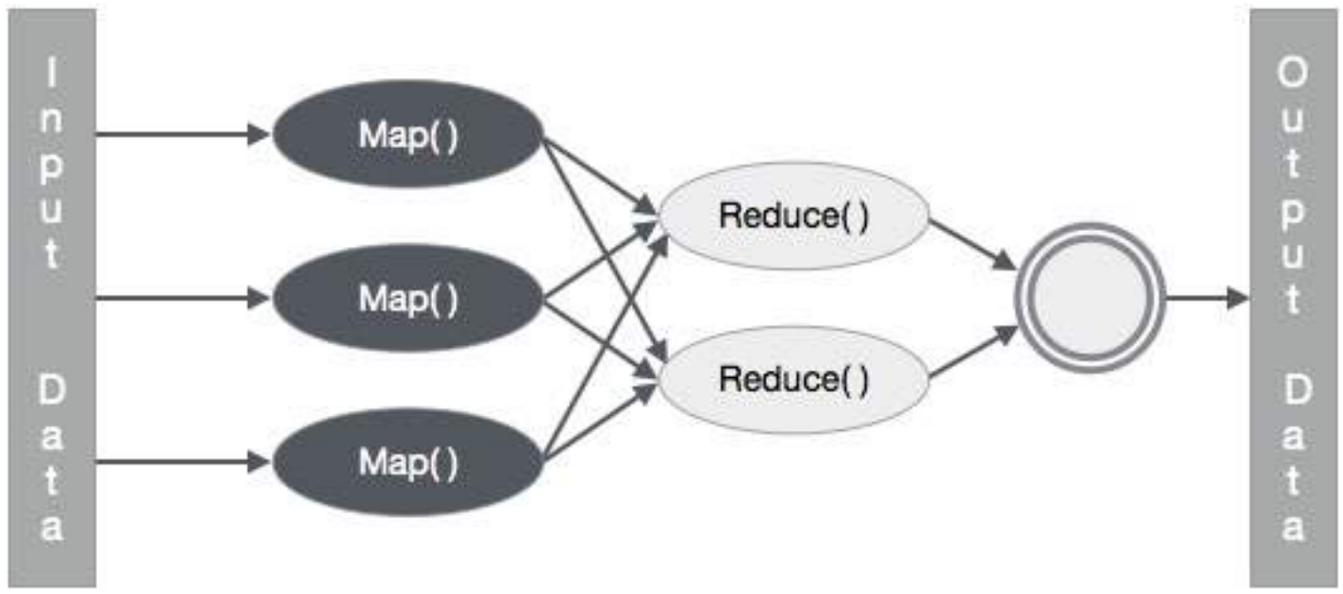
https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

MapReduce Algorithm

- Use a `mapper` function to turn each item into zero or more key-value pairs.

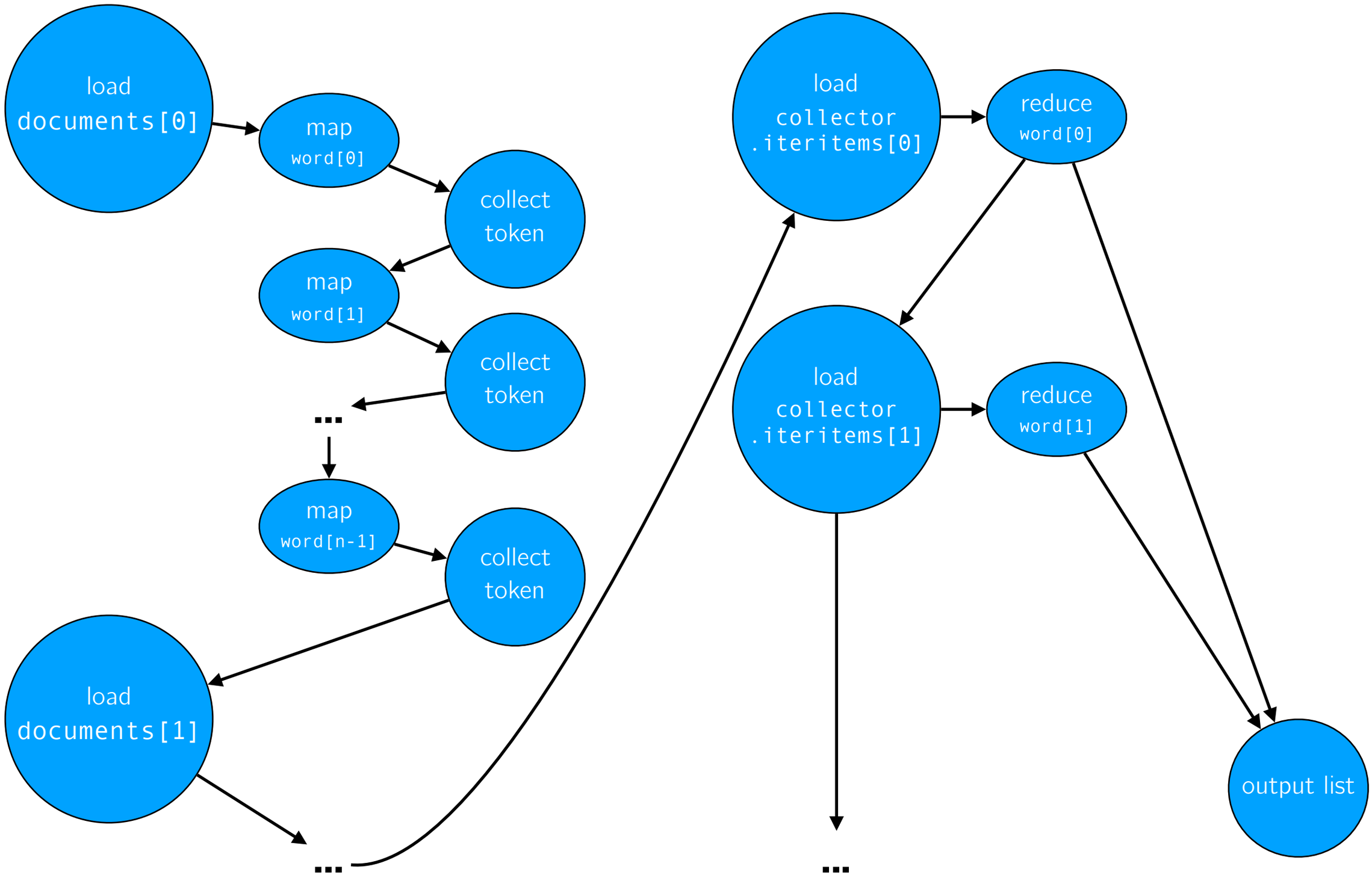
- Collect together all the pairs with identical keys.

- Use a `reducer` function on each collection of grouped values to produce output values for the corresponding key.



Word Count MapReduce

In a single notebook



Word Count Map Reduce

In parallel

- Use a `mapper` function to turn each item into zero or more key-value pairs.

- Collect together all the pairs with identical keys.

```
for document in documents_test:
    Q.enqueue(worker.collector, document, 'test_vocab_1')
```

- Use a `reducer` function on each collection of grouped values to produce output values for the corresponding key.

```
vocabulary = [word.decode() for word in REDIS.smembers('test_vocab_1')]
```

Word Count Map Reduce

In parallel