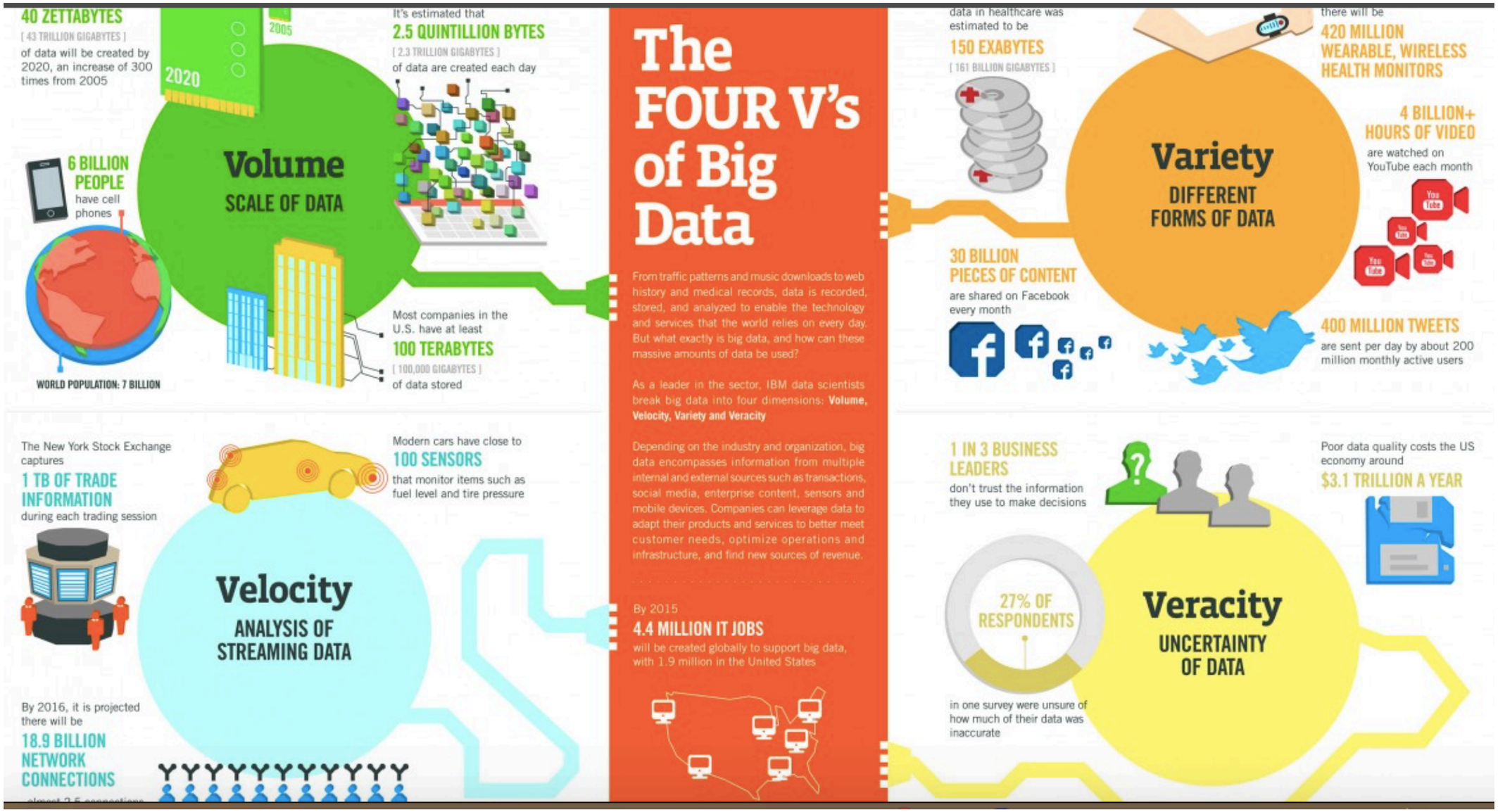
Big Data

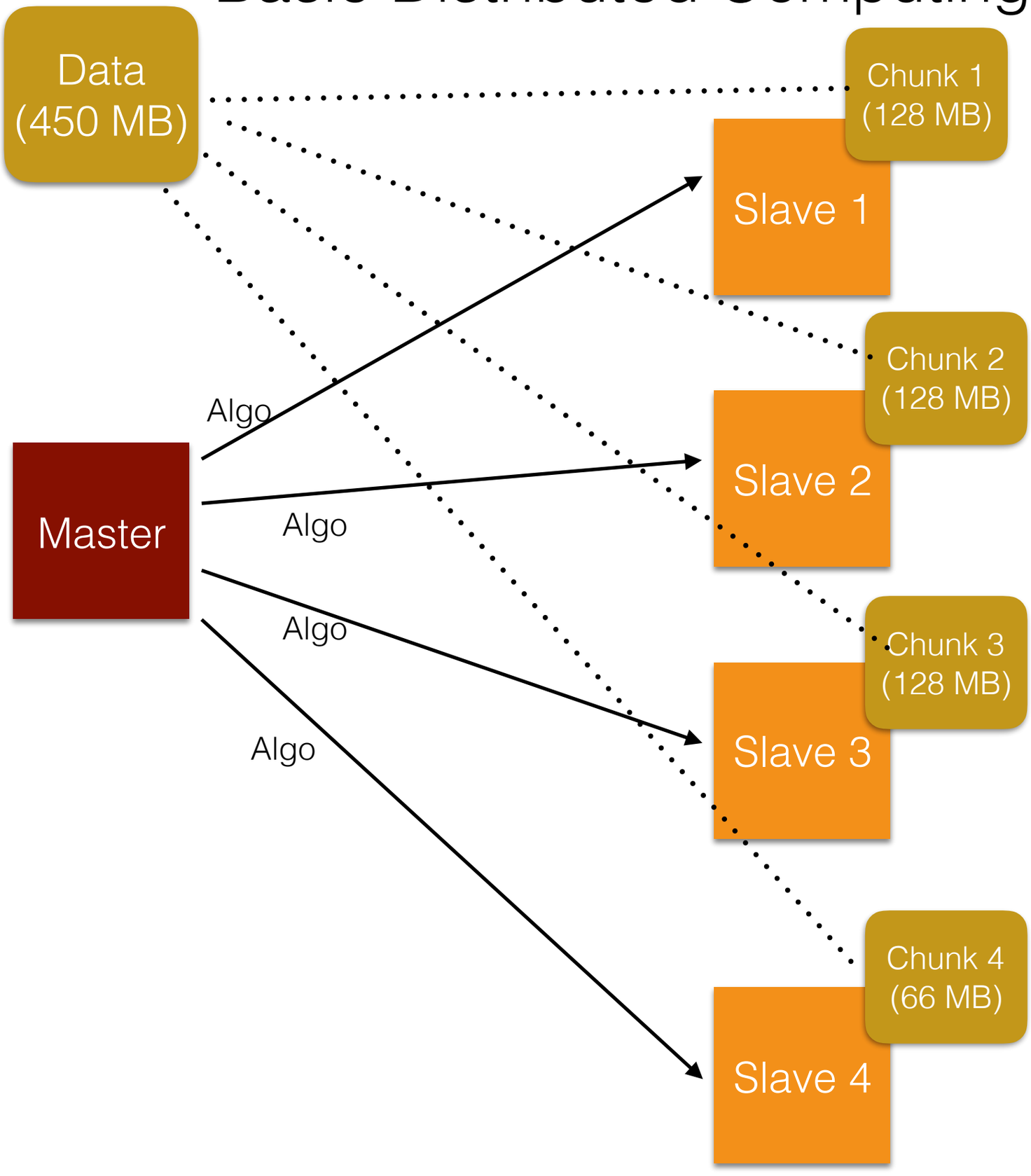


- Source: https://www.ibmbigdatahub.com/infographic/four-vs-big-data

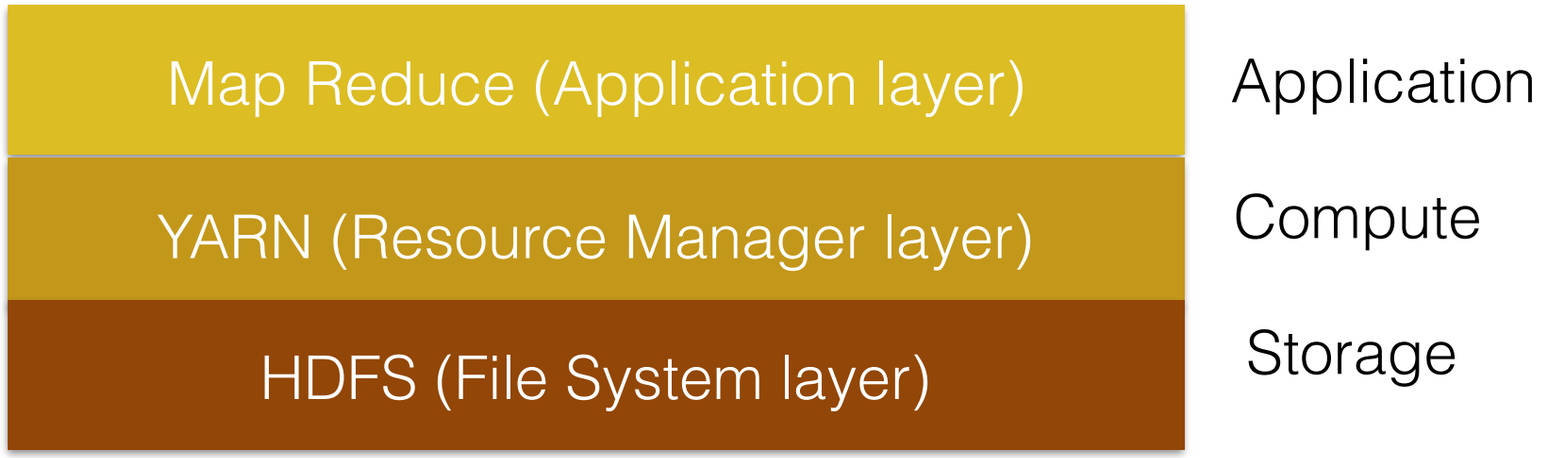
Hadoop

- http://hadoop.apache.org/docs/current/index.html
- Distributed Processing / Parallel Computing
- Breakdown data into chunks and distribute it across the cluster and send the same program to execute on each chunk
- Send logic to where data is
- In case of breakdown of node(s) in the cluster use a replica on another node to accomplish job
- Fault tolerant

Basic Distributed Computing Architecture



Components of Hadoop Infrastructure



Running mrjob

- Mrjob package needs to be installed and it has its built in Hadoop installation so we start with it
- Later on we will see how to run mrjob on a remote Hadoop cluster (the real thing!)
- `pip install mrjob`
- write a word count program and run

WordCount Program

```
from mrjob.job import MRJob

class MyMRWC(MRJob):
    def mapper(self, key, line):
        words = line.split(' ')
        for word in words:
            yield word, 1
            self.increment_counter('word','no of words',1)

    def reducer(self, word, count_one):
        yield word, sum(count_one)
        self.increment_counter('word','no of unique words',1)

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

- python MyMRWC.py mytextfile > output1
- cat output1