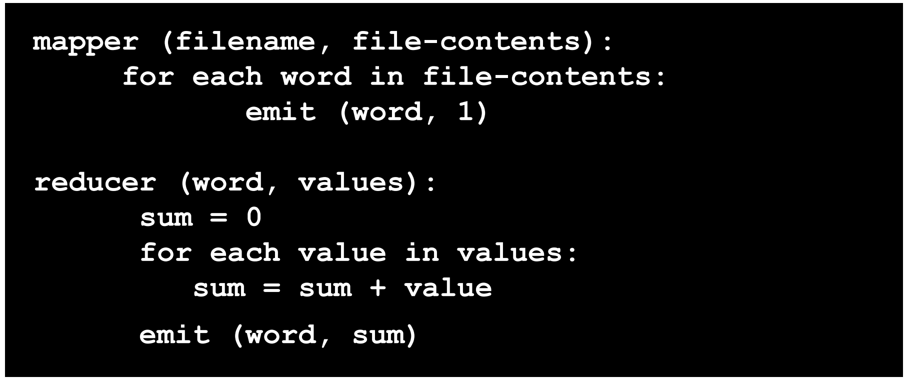
Map Reduce examples

WordCount (Unigram count)

- Take a portion of this online book - http://www.gutenberg.org/cache/epub/35937/pg35937.txt
- Count occurrences of words in the book.
- Consider the following sentence,
  - Split it line by line
  - Split it word by word
  - Count individual words
- Pseudo-Code

Jupiter. offers two peculiar-rities. In its shrunken condition, its diameter, instead of being eleven times that of the Earth, will be not quite seven, and the force of gravity at the surface will be greater than that of the Earth in the same proportion

1.
2.
3.



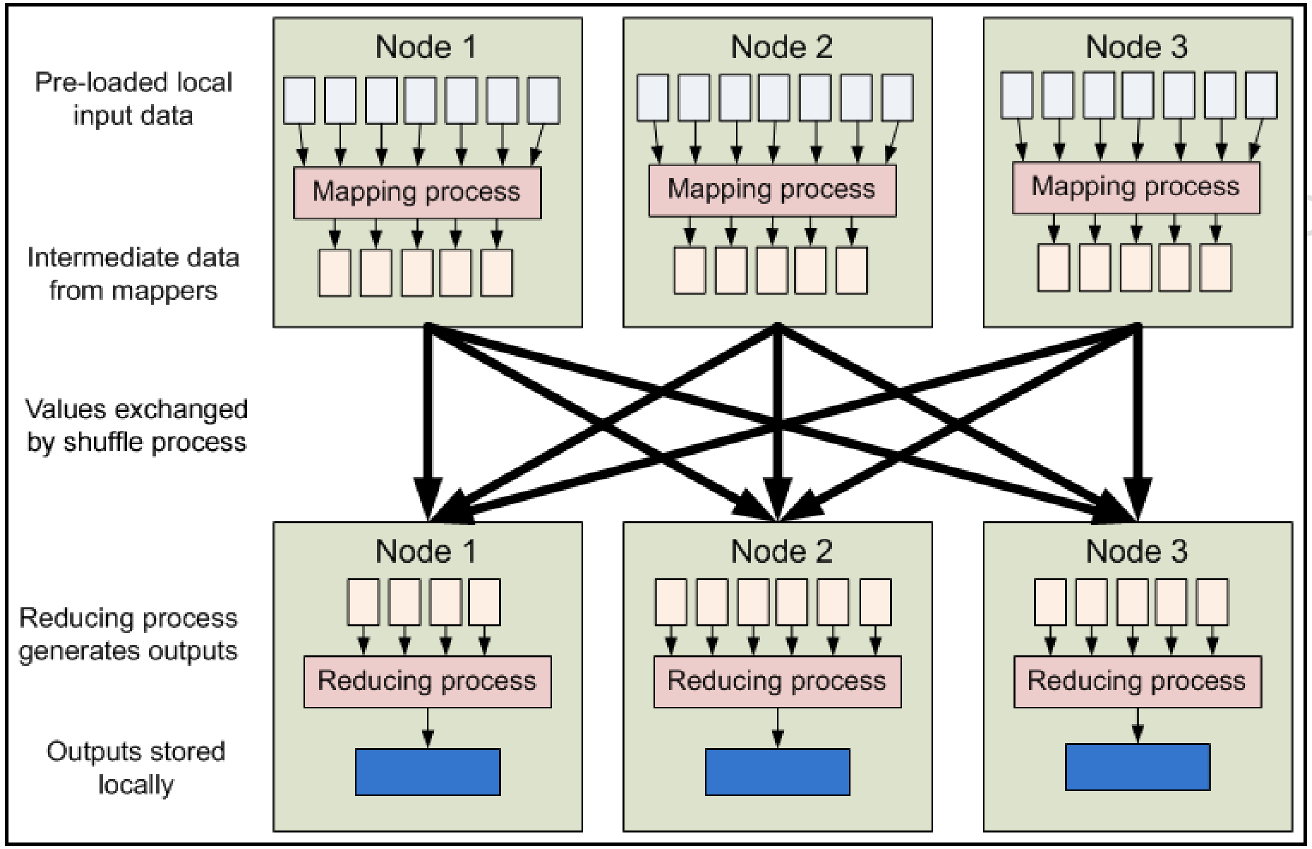
```
mapper (filename, file-contents):
     for each word in file-contents:
            emit (word, 1)

reducer (word, values):
      sum = 0
      for each value in values:
         sum = sum + value
      emit (word, sum)
```

TopN(Unigram count)

- Take a portion of this online book - http://www.gutenberg.org/cache/epub/35937/pg35937.txt
- Count occurrences of words in the book.
- Consider the following sentence,
  - Split it line by line
  - Split it word by word
  - Count individual words
- Sort the output words
- Print top 15 occurances
- Hints – avoid punctuations in words, use the regex replace bad stuff
  - "[_|$#<>\\^=\\[\\]*/\\\\,;,.\\-:()?!\""]";
  - Sort a Map on values - http://javarevisited.blogspot.it/2012/12/how-to-sort-hashmap-java-by-key-and-value.html

Under the hoods



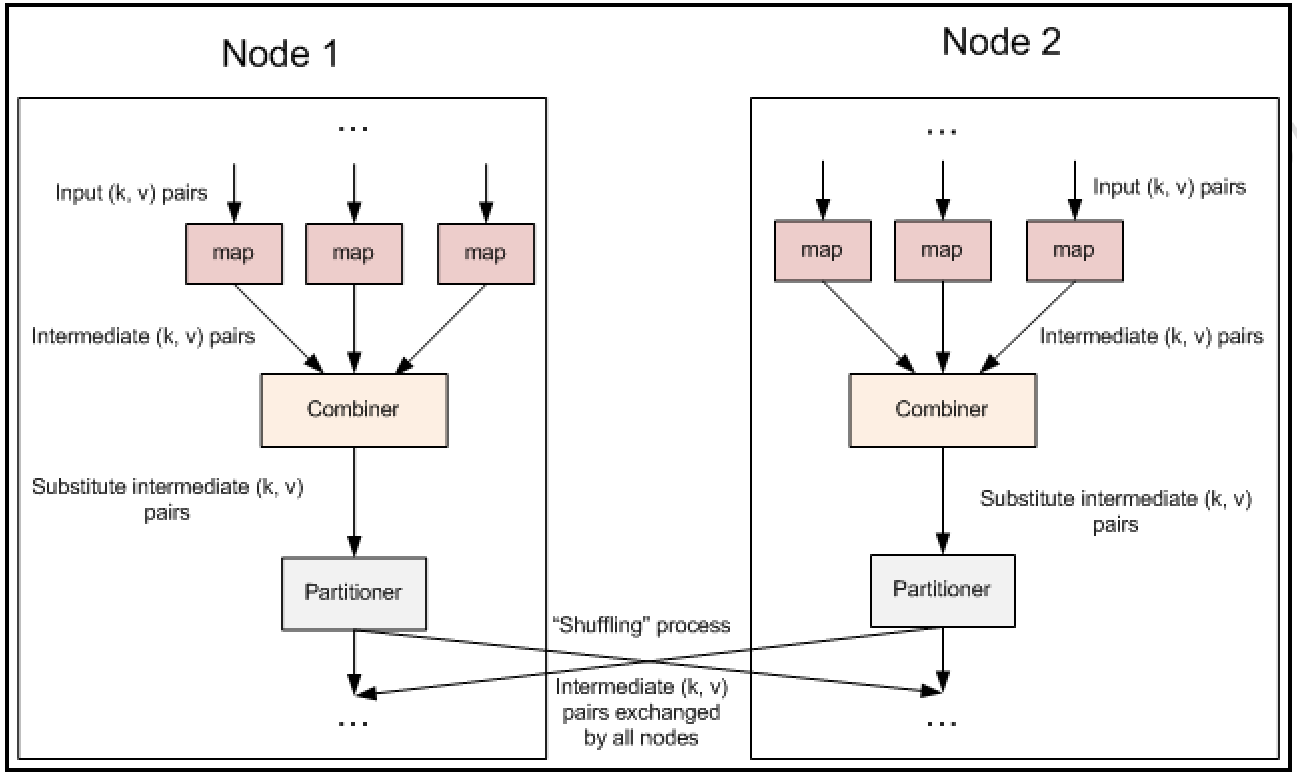
Goals:

To reduce the bytes sent across

Activities:

1. Optimize the object creation
2. Optimize Mapper
3. Use Combiner

Combiner


Node 1
Node 2
...
Input (k, v) pairs
map
map
map
Intermediate (k, v) pairs
Combiner
Substitute intermediate (k, v) pairs
Partitioner
"Shuffling" process
Intermediate (k, v) pairs exchanged by all nodes
...

- Combiner, is an optional phase that runs after the Mapper and before the Reducer. Usage of the Combiner is optional. If this pass is suitable for your job, instances of the Combiner class are run on every node that has run map tasks.
- The Combiner will receive as input all data emitted by the Mapper instances on a given node. The output from the Combiner is then sent to the Reducers, instead of the output from the Mappers.
- **The Combiner is a "mini-reduce" process which operates only on data generated by one machine.**

- **conf.setCombinerClass(Reduce.class);**

Combiner - Caution

- Combiner is an optimization technique, do not drive functionality out of a Combiner
- Combiners can be called 0 times or n number of a times per mapper
- Reduceres can be used in place of Combiners when the function you want to apply is both commutative and associative .
  - commutative ==> f(a, b) = f(b, a)
  - associativity ==> f(f(a, b), c) = f(a, f(b, c))
  - Example – sum of count of words in commutative and associative, but avg (or) concatenation of words is not.

No Combiner:

```
Mapper1                       Mapper2                        Mapper3
    A=2,2,2,2,2,2,2,2           A=4                            A=2,3,4
AvgReducer
           A=(2,2,2,2,2,2,2,2,4,2,3,4)/12 = 2.14
```

With Combiner:

```
 Mapper1                      Mapper2                        Mapper3
    A=2,2,2,2,2,2,2,2           A=4                            A=2,3,4
 AvgCombiner                  AvgCombiner                    AvgCombiner
    A=2                         A=4                            A=3
 AvgReducer
                A=(2,4,3)/3 = 3
```

Inverted Index

- Take a portion of this online book - http://www.gutenberg.org/cache/epub/35937/pg35937.txt
- Create a map of words and the files in which they occur.
  - Hint: use FileSplit class
    - FileSplit fs = (FileSplit) context.getInputSplit();
    - fs.getPath().getName() ➔ would give the file name

Pending Assignment

- Pick real time data for All India seasonal Annual Min/Max temperatures series from 1901 – 2014) from below linkhttps://data.gov.in/catalog/all-india-seasonal-and-annual-minmax-temperature-series
- The data is layed out year wise, with min and max, but not averages.
- Compute average temperature year wise and spit data in the format of Year, Average.
- We will use this result for some analytics.

Some more pending assignments

- Some more MR patterns
  - ~~Find how many Unigrams are there in a file~~
  - How many unique bigrams are there in your text file? A bigram is a N-gram of two words.
    - Consider the text "A cat jumped over a wall to catch a rat with no fat"
    - Bigrams are two words, for the above sentence the bigrams are
      - a cat
      - cat jumped
      - jumped over
      - over a
      - a wall
      - wall to
      - ....
  - Find how many anagrams occur in a file. Anagram is a word formed by rearranging the letters of another, such as spar, formed from rasp.